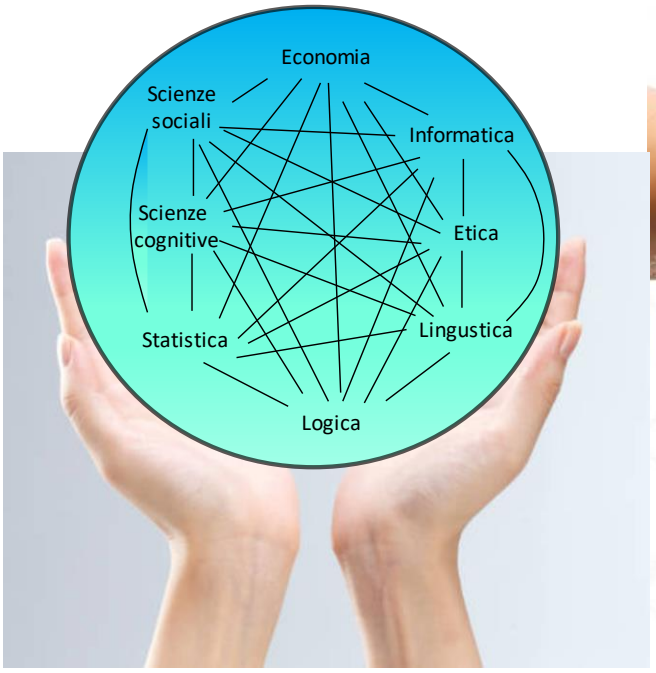
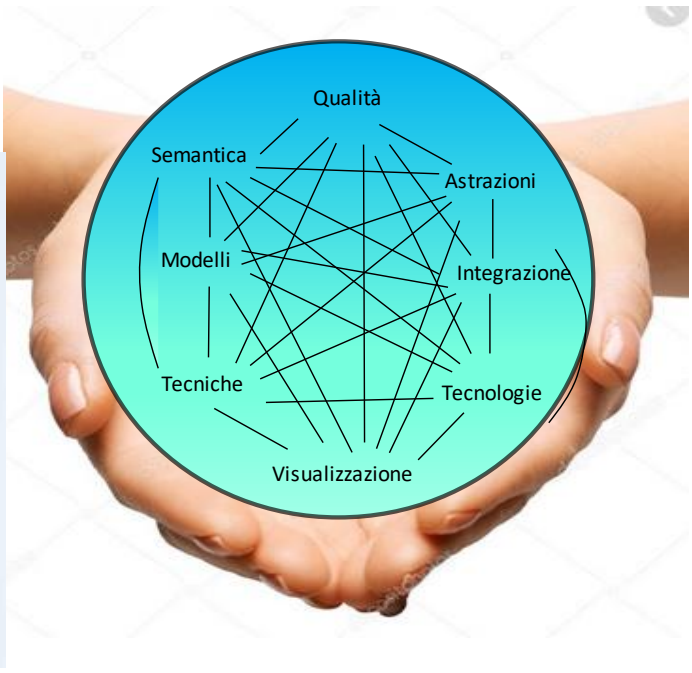
La Scienza dei Dati

Carlo Batini, Federico Cabitza, Paolo Cherubini, Anna Ferrari,
Andrea Maurino, Roberto Masiero, Matteo Palmonari, Fabio Stella





La Scienza dei Dati

Carlo Batini (*), Federico Cabitza (*), Paolo Cherubini (*),
Anna Ferrari (*), Andrea Maurino (*), Roberto Masiero (+),
Matteo Palmonari (*), Fabio Stella (*)

(*) Università di Milano-Bicocca
(+) Dirigente di azienda

Prologo

Trentacinque anni fa, nel 1984, uno degli autori di questo testo scrisse un libro per gli Editori Riuniti nella collana Libri di Base, intitolato *Le basi della Informatica*. La copertina del libro era quella che vedete qui sotto in Figura 1.

Figura 1: La copertina del libro "Le basi della informatica"

Non è un caso che la prima figura del libro (riprodotta in Figura 2) rappresenti dati in formato di tabella. Questi dati rappresentavano una pagina immaginaria di una agendina di carta, un documento che è sempre più spesso sostituito da molti con una agenda elettronica, ma che molti altri, compreso l'autore del libro, continuano ad usare in formato cartaceo.

Figura 2: La mia agendina

Un'altra figura del primo capitolo del libro è un esempio di orario grafico (vedi Figura 3), che veniva usato tanto tempo fa, in cui le linee rappresentano gli spostamenti nel tempo dei treni; quanto più una linea è vicina alla verticale, tanto più il treno va veloce.

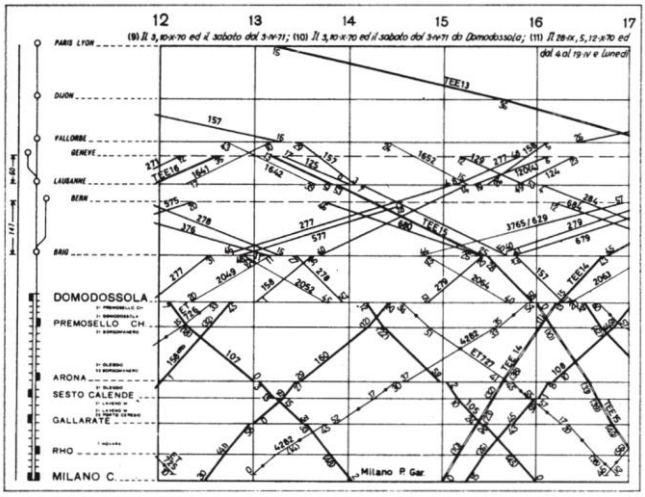


Figura 3: Un orario grafico

La Figura 3 mostra dati che non sono rappresentati attraverso una tabella, ma per mezzo di un simbolismo grafico; il simbolismo fornisce una rappresentazione meno compatta dei viaggi dei treni rispetto alle tabelle, ma visivamente più efficace nel mostrarci le diverse velocità dei treni.

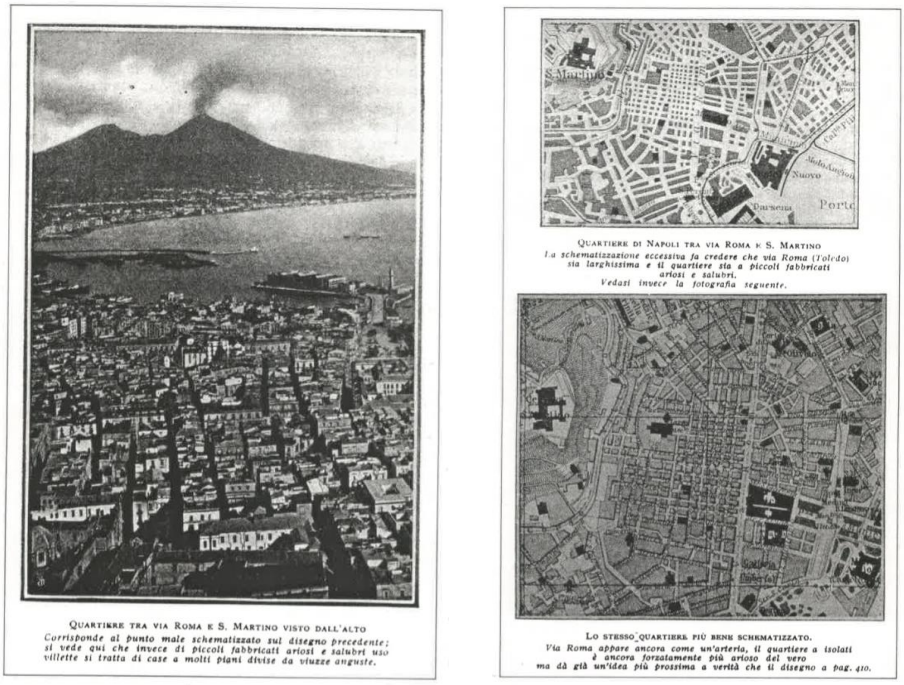


Figura 4: Una delle figure finali del libro “Le basi della informatica”

Nella parte finale del libro *Le basi dell’informatica*, dopo aver indagato concetti come la complessità e la correttezza dei programmi, si torna a parlare di dati, mostrando come una mappa possa dare una visione molto distorta del territorio che rappresenta.

Nella Figura 4 compare a sinistra una vecchia foto che mostra la zona centrale di Napoli, quella dei quartieri spagnoli, e a destra due mappe della stessa zona che compaiono in successive edizioni della guida del Touring su Napoli; nella prima, quella in alto, i quartieri spagnoli sono così stilizzati che sembrano una amena zona di villette, la seconda fornisce una rappresentazione più realistica.

Le mappe in Figura 4 non sono *dati* nel senso tradizionale del termine. Ma quale è, appunto, il senso tradizionale del termine *dato*? Da quando Ted Codd ha proposto una rappresentazione relazionale dei dati, almeno nel mondo della informatica la parola *dati* è stata a lungo riferita alle tabelle relazionali.

Perché sono così utili e diffuse le tabelle relazionali? La Figura 5 mostra un testo in linguaggio naturale che descrive tre studenti di una ipotetica Università, gli esami che hanno superato, e i corsi a cui si riferiscono gli esami. Proviamo, ad esempio, a cercare visivamente nel testo e nella tabella i nomi dei corsi per i quali Batini ha superato l'esame. Credo si possa concludere che, in virtù della maggiore strutturazione, sia più facile rispondere alla domanda "navigando" nelle tre tabelle piuttosto che nel testo in linguaggio naturale.

Supponi di voler rappresentare i seguenti fatti mediante **tre** tabelle.
Ci sono tre studenti universitari, Batini con matricola 13242, che è nato in Italia, Xu con matricola 24195 nato In Cina, e Smith, con matricola 32845 nato in USA. Ci sono anche tre corsi, Analisi con codice 27 al primo anno, Algoritmi con codice 49 al primo anno e Logica con codice 77 al secondo anno. La matricola 13242 ha sostenuto l'esame del corso 27 con voto 25, la matricola 24195 il corso con codice 77 con 28 e la matricola 32845 il corso con codice 27 con voto 30

Una possibile rappresentazione

Studente

Matricola	Cognome	Stato Estero
13242	Batini	-
24195	Xu	Cina
32845	Smith	USA

Esame

Matricola	Codice Corso	Voto
13242	27	25
13242	77	28
32845	27	30

Corso

Codice	Nome Corso	Anno
27	Analisi	1
49	Algoritmi	1
77	Logica	2

Figura 5: Una frase che descrive alcuni studenti, esami e corsi,
e la sua rappresentazione con tre tabelle relazionali

Coloro che nel mondo della ricerca informatica hanno lavorato nell'area dei dati, e anche chi con i dati ha lavorato professionalmente, hanno spesso avuto negli scorsi decenni un complesso di inferiorità verso quei ricercatori e professionsiti che lavoravano nel campo del software e concepivano metodologie di progettazione dei programmi software. Questi ultimi guardavano talvolta i dati dall'alto in basso, come un sottoprodotto dei loro programmi.

Questa sudditanza dei dati rispetto ai programmi si poteva ritrovare anche nei linguaggi di programmazione. In uno dei linguaggi più utilizzati di tutti i tempi, il Cobol, i dati sono una parte del programma chiamata Data Division. In un approccio, molto diffuso a quei tempi, in cui ogni programma definiva i propri dati, come si poteva mai pensare che due programmi potessero utilizzare gli stessi dati, e questi dati avessero una propria vita, indipendente dai programmi?

Così andava l'Informatica a quei tempi, e il concetto di base di dati, introdotto tra gli altri da Ted Codd, voleva superare questa parcellizzazione e sudditanza verso i programmi, separando i dati gestiti nella base di dati dai programmi software che ne facevano uso. In una base di dati, ogni dato è di interesse di più programmi, ed è logicamente rappresentato una volta sola. Tutto risolto? Sfortunatamente no: se in una base di dati due unità di una organizzazione condividono logicamente gli stessi dati, allora spesso vengono create due basi di dati, una per la prima unità e una perla seconda, tornando in questo modo alla vecchia frammentazione.

Per quasi quaranta anni nell'Informatica i dati sono stati obiettivamente figli di serie B. C'era un'altra ragione per questo fenomeno: i dati in formato digitale erano pochi, ed erano pochi perché le tecnologie che permettevano di trasformare una informazione del mondo esterno in un dato digitale, ad esempio, l'ora in cui vorrei partire da Milano domani in treno per raggiungere Roma, erano poche, nel nostro caso la tastiera di un calcolatore. Negli ultimi quaranta anni sono state scoperte e prodotte tante tecnologie che trattano dati digitali, l'elenco è impressionante: il personal computer, Internet, il Web, il telefono cellulare, tanti tipi di sensori, il cloud, le reti sociali, per citarne solo alcune. Per fornire solo due elementi: ogni anno e mezzo raddoppia la quantità di dati digitali scambiata nel Web, e nel 2.025 si prevede che ci saranno mille sensori di dati digitali per ogni essere umano.

Tutto ciò ha creato negli ultimi anni il fenomeno dei cosiddetti big data, le grandi quantità di dati digitali. I dati hanno avuto uno sviluppo così incredibilmente veloce, che oramai, come ha scritto qualche autore, e come commenteremo criticamente, "parlano da soli". Sono state sviluppate moltissime applicazioni che utilizzano dati digitali, applicazioni che possiamo attivare con una semplice pressione del dito sul nostro smart phone, e che ci dicono: quanto dobbiamo aspettare per un autobus, quanti passi abbiamo fatto da stamattina, quale viaggio aereo costa meno per raggiungere un aeroporto in un certo giorno.

Dunque, i dati digitali sono *arrivati*; e sono arrivati mentre uno degli autori del libro si avvicinava alla pensione; questo autore, con un misto di sindrome da accerchiamento e autoironia, ha pensato che gli stesse succedendo la stessa cosa del capitano Giovanni Drogo nel libro di Dino Buzzati "Il Deserto dei Tartari", vedi Figura 6.

Figura 6: Il Deserto dei Tartari

Drogo ha aspettato i Tartari alla Fortezza Bastiani tutta la vita, e proprio quando arrivano, se ne deve andare perché va in pensione ed è oramai diventato un peso. Naturalmente il caso dell'autore e quello di Giovanni Drogo sono molto diversi….

Questo nuovo libro parla di una nuova Scienza, la Scienza dei dati. Attribuire il nome di Scienza ai dati può forse apparire azzardato, ma da dversi anni orami nelle Università Italiane e prima ancora negli Stati Uniti e in Cina, sono attivi corsi di laurea magistrali in Scienza dei Dati, e tra questi il Corso di Laurea Magistrale in Data Science della Università di Milano Bicocca, che al momento attuale (2020) entra nel suo quarto anno di vita. Dunque è tempo di iniziare a capire cosa è questa nuova Scienza.

E per capirlo, è stato opportuno scrivere un libro a più mani. Accanto all'autore della maggior parte dei Capitoli, Carlo Batini, sono coautori del libro Federico Cabitza in tema di visualizzazione, Anna Ferrari per la Statistica, Paolo Cherubini sulle Scienze cognitive, Roberto Masiero in tema di economia dei dati digitali, Andrea Maurino sulle piattaforme tecnologiche per big data, Matteo Palmonari per la semantica dei dati, Fabio Stella sulle tecniche di learning. Quasi tutti gli autori sono docenti del Corso di Laurea Magistrale su menzionato, e quindi hanno dovuto in questi anni approfondire le tematiche di cui qui scrivono, per poterle trasferire agli studenti attraverso un corpo di conoscenze strutturato.

Questo libro è dunque un cantiere in cui l'artefatto complessivo, l'edificio costituito dalla Scienza dei dati, è in febbrile, rapida costruzione. Va perciò visto come una prima esplorazione dei tanti temi che compongono la Scienza dei dati, e come uno stimolo a continuare a lavorare per costruire le fondamenta di questa nuova Scienza.

Il Capitolo 1 introduce a tutte le tematiche trattate nel libro; preghiamo perciò il lettore di investire un po' di tempo nella lettura del Capitolo 1, al termine del quale sono proposti diversi percorsi di approfondimento; auguriamo a tutti di ritrovarsi in uno di questi percorsi, ovvero di crearsene uno proprio, così da proseguire la lettura del libro secondo i propri interessi e inclinazioni.

Gli autori

La Scienza dei Dati

Indice

Capitolo 1 – Introduzione alla Scienza dei Dati

Carlo Batini

A partire da una certa età, i nostri ricordi sono così intrecciati fra di loro che la cosa cui pensiamo, il libro che leggiamo non hanno quasi più importanza. Abbiamo messo dovunque un po' di noi stessi, tutto è fecondo, tutto è pericoloso, e possiamo fare scoperte altrettanto importanti nei "Pensieri" di Pascal quanto nella pubblicità di una saponetta.

Marcel Proust, Alla Ricerca del Tempo Perduto

1. Introduzione

Questo pensiero di Proust è un'ottima introduzione al percorso che intraprendiamo in questo libro. La nostra vita, la conoscenza che accumuliamo ed elaboriamo si amplia nel tempo e copre spazi sempre più ampi; allo stesso tempo si ampliano gli intrecci tra informazioni apparentemente lontane.

In questa continua crescita, I dati digitali stanno diventando sempre più importanti nella nostra vita, e possono darci tanto in termini di conoscenza del mondo, se, ricomponendoli, siamo in grado di fare le scoperte di cui al pensiero di Proust. Allo stesso tempo, i dati possono anche deformare la nostra immagine del mondo, creando una realtà virtuale che rende meno nitida e deformata la nostra conoscenza del mondo sensibile.

L'etimologia del termine "dato" [Borgman, 2015] deriva dal Latino *data*, nominativo plurale di *datum* ("che è dato, che è fornito"). Se ci pensiamo, dato è anche il participio passato del verbo dare; un dato riguarda dunque il passato, qualcosa che è già successo, che guarda al passato. Ma nella nostra vita noi non guardiamo sempre al passato, immaginiamo anche il futuro; torneremo su questa osservazione nel Capitolo 16 sui limiti della Scienza dei dati.

La diffusione dei dati digitali sta crescendo negli ultimi anni a ritmi sempre più intensi; per fornire un solo indicatore, i dati scambiati sul Web raddoppiano in dimensione ogni anno e mezzo, dando luogo ad una crescita esponenziale dei dati digitali nel tempo, quale mai si è verificata per la informazione e la conoscenza nella storia della umanità. Ciò dà luogo alla utilizzazione dei dati digitali in tutti i settori produttivi, in particolare nei servizi e nella ricerca scientifica, e, allo stesso tempo, ad una presenza sempre più intensa e a volte intrusiva nella vita della singole persone e delle comunità sociali.

Dati, informazioni, conoscenza sono tre forme diverse degli stessi artefatti. I dati sono rappresentazioni digitali a cui non è in genere associato un significato (ad esempio il numero 38,5 letto su un termometro, vedi Figura 1); le informazioni si ottengono applicando ai dati la nostra conoscenza pregressa sui fenomeni osservati (38,5 è una temperatura in gradi Celsius), la conoscenza è ciò che noi estraiamo

dalla informazione mediante elaborazioni o ragionamenti (so che 37 è la temperatura corporea normale, quindi 38,5 è indicazione di febbre).

Figura 1 - Dati, informazione, conoscenza quando usiamo un termometro

Il Capitolo è organizzato come segue. Partiamo nella Sezione 2 da problemi concreti, e cerchiamo di capire come questi problemi possano essere risolti, e perché alcuni sono risolti da tempo e altri sono stati risolti recentemente avendo a disposizione tanti dati descrittivi del problema e tecnologie digitali per risolverlo. La Sezione 3 comincia a descrivere la grande trasformazione del nostro mondo dall'epoca in cui erano disponibili pochi dati all'epoca attuale dei big data. A questo punto, la Sezione 4 ci dice come è organizzato il libro, nella sua doppia articolazione in *fasi* del ciclo di vita del dato, ad esempio la fase di analisi o la fase di visualizzazione, e *discipline* preesistenti alla Scienza dei dati, su cui è rilevante indagare per il loro forte legame con la Scienza dei dati, ad esempio l'Economia digitale o l'Etica dei dati. Un'appendice descrive alcune classificazioni relative a diverse tipologie dei dati.

2. I problemi che riusciamo a risolvere con i dati digitali

Il grande, rapido sviluppo dei dati digitali permea la nostra vita, modifica la comunicazione pubblica e privata, influenza e modifica in modo radicale la economia, fornisce alla ricerca scientifica materiale prezioso, presenta grandi opportunità e grandi rischi. Indaghiamo alcuni esempi di problemi la cui soluzione comporta la osservazione e misurazione di fenomeni della realtà attorno a noi, la rappresentazione di tali fenomeni mediante dati digitali e la successiva elaborazione di tali dati per risolvere il problema.

Problema 1: *Prevedere le eclissi* - Fin dalla antichità, l'umanità ha cercato di interpretare e prevedere le eclissi del sole e della luna. Questi fenomeni grandiosi hanno sempre suscitato forti emozioni di paura o stupore, per cui già nelle antiche civiltà Babilonesi si svilupparono osservazioni che portarono a scoprire cicli temporali nella evoluzione delle eclissi.

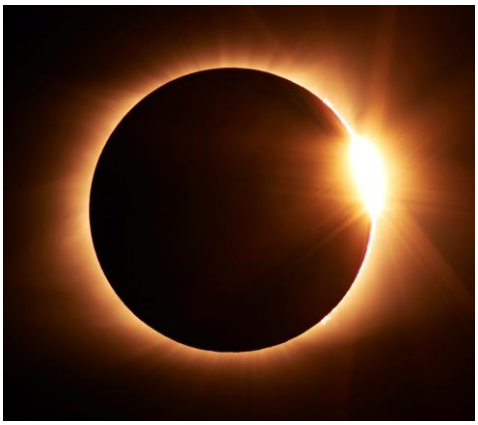


Figura 2 – La previsione delle eclissi

I geografi/astronomi dovettero trascrivere gli anni in cui le eclissi si verificarono, producendo in tal modo semplici serie temporali; analizzando queste serie temporali, furono in grado di trovare delle regolarità e di prevedere le future eclissi.

Problema 2: *Trovare il biglietto aereo più economico* - Vi sarà capitato di acquistare un biglietto aereo, o che un vostro parente o amico abbia acquistato un biglietto aereo. Oramai sul Web abbiamo tanti siti che ci permettono di risolvere il seguente problema: fissato il giorno del viaggio, la città di partenza e quello di arrivo, trovare l'elenco dei possibili voli (o combinazione di voli) disponibili, ordinati dal meno costoso al più costoso, vedi Figura 3.

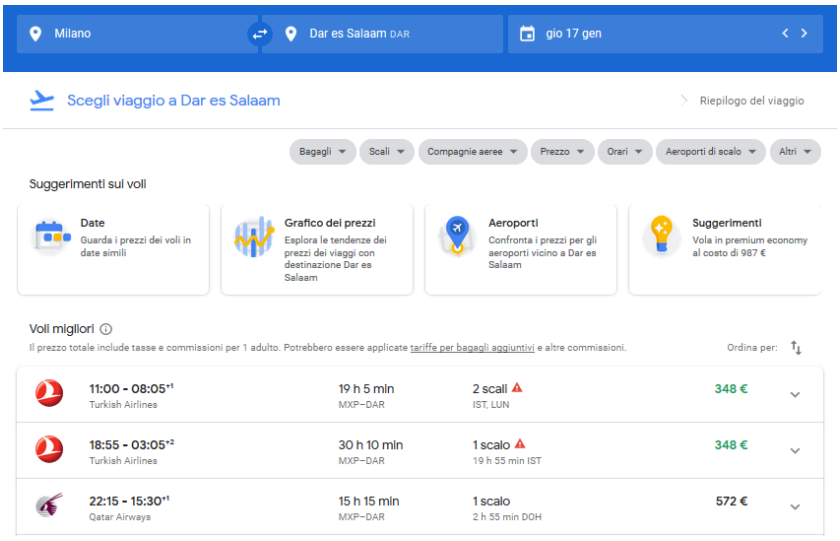


Figura 3 – Prenotare e acquistare un volo (dal sito eDreams)

In questo caso, per risolvere il problema dobbiamo poter accedere agli orari dei voli di tutte le compagnie aeree, che possiamo rappresentare mediante delle tabelle, e a questo punto, noti l'aeroporto di partenza A e aeroporto di arrivo B e il giorno del volo, dobbiamo collegare in tutti i modi

possibili aeroporti di partenza e di arrivo dei voli che ci permettono di costruire itinerari con 0, 1, 2 ecc., scali intermedi, che complessivamente permettono di andare dall'aeroporto A all'aeroporto B. Sommando i prezzi dei biglietti delle varie tratte componiamo il prezzo complessivo di ogni viaggio e a questo punto ordiniamo i voli dal più economico al più costoso.

A ben vedere, a noi spesso interesserebbe avere una risposta a un'altra esigenza, espressa dal problema seguente.

Problema 3: *Predire il giorno in cui conviene acquistare un biglietto aereo*, cioè, fissato il giorno del viaggio, l'aeroporto di partenza e quello di arrivo, conoscere il giorno in cui l'acquisto del biglietto sia più conveniente.

La nostra esigenza deriva dal fatto che le compagnie aeree applicano leggi che determinano il prezzo del biglietto a noi ignote. Ad esempio, abbiamo tutti notato che il prezzo tende ad aumentare negli ultimi giorni prima del viaggio, per poi ridursi molto nella imminenza della partenza (ammesso che ci siano ancora posti disponibili nella classe che abbiamo scelto), le cosiddette offerte last minute. Ebbene, questo problema è stato risolto solo pochi anni fa, ed è stato risolto perché solo pochi anni fa sono stati inventate tecniche algoritmiche che operando su grandi quantità di dati sono in grado di individuare le regolarità sui dati che permettono di predire cosa accadrà in futuro.

Problema 4: *Scoprire in quale zona di una città c'è meno inquinamento* - Se facciamo jogging nella nostra città, oppure in una città dove siamo temporaneamente per un viaggio, siamo interessati a sapere in quale ora del giorno ci conviene correre per respirare aria meno inquinata e dove conviene dirigerci per trovare meno inquinamento, vedi Figura 4.

Figura 4 – Dove andare a correre senza farsi avvelenare dallo smog?

Anche in questo caso il problema è stato risolto solo recentemente; per affrontare questo problema, dobbiamo avere a disposizione diverse informazioni, con il loro andamento nel tempo, riguardanti i fattori che influenzano l'inquinamento, come il traffico, il riscaldamento delle abitazioni, il tempo

atmosferico, la velocità del vento. Dobbiamo quindi capire da quale fonte acquisire queste informazioni e come metterle in relazione con il tasso di inquinamento.

Problema 5: *Tradurre una frase da una lingua a un'altra* – Quante volte abbiamo la necessità di tradurre frasi da una lingua a un'altra (vedi Figura 5). I traduttori di libri fanno questo per mestiere, e hanno il problema di adattare le frasi da un linguaggio che ha un determinato lessico, sintassi dei periodi e significato delle parole ed espressioni composte, ad un altro linguaggio con lessico, sintassi e significato spesso diversissimi, perché frutto di secoli e talvolta millenni di mutamenti e adattamenti; la traduzione deve preservare al massimo possibile il senso, e, talvolta, le emozioni che l'autore ha voluto attribuire alla frase.

Figura 5 – Come tradurre dall'italiano al cinese (da Google Translator)

I linguisti da tempo studiano il problema della traduzione automatica da lingua naturale a lingua naturale, con risultati per molte lingue non soddisfacenti, nell'ambito della disciplina del Natural Language Processing. Partendo dalla grande disponibilità di dati sul Web, ad esempio nelle enciclopedie del tipo di Wikipedia in cui lo stesso testo (o testi simili) compare in diverse lingue, è possibile "imparare" a tradurre, usando tecniche che verranno descritte nel Capitolo 10 sul machine learning.

Problema 6 – Prevedere il churn (l'abbandono di un cliente che ha deciso di passare alla concorrenza)

Account Length	VMail Message	Day Mins	Churn	Intl Calls	Intl Charge	State	Area Code	Phone
128	25	265,1	n	3	2,7	KS	415	382-4657
107	26	161,6	n	3	3,7	OH	415	371-7191
137	0	243,4	n	5	3,29	NJ	415	358-1921
84	0	299,4	y	7	1,78	OH	408	375-9999
75	0	166,7	n	3	2,73	OK	415	330-6626
118	0	223,4	n	6	1,7	AL	510	391-8027
121	24	218,2	n	7	2,03	MA	510	355-9993
147	0	157	n	6	1,92	MO	415	329-9001
117	0	184,5	n	4	2,35	LA	408	335-4719
141	37	258,6	n	5	3,02	WV	415	330-8173

Figura 6 – Prenotare e acquistare un volo

Le società telefoniche e in genere quelle che erogano servizi in regime di concorrenza hanno il problema di capire quando un cliente sta per lasciarle, perché non più soddisfatto del servizio che riceve, o perché la concorrenza ha lanciato nuove offerte. Se riescono a intercettare queste intenzioni, possono cercare

di adottare tattiche preventive per dissuadere il cliente; questo fenomeno è chiamato churn. Anche in questo caso, la disponibilità di grandi quantità di dati storici del tipo di quelli di Figura 6 permette di adottare tecniche che apprendono e sono in grado di costruire modelli predittivi.

I problemi precedenti (giorno ottimale di acquisto di un biglietto, andamento dell'inquinamento nella nostra città, tradizione da lingua a lingua, previsione del churn) possono essere affrontati in due modi completamente diversi.

Metodo 1 – In questo metodo noi cerchiamo di capire la legge, o le leggi, che regolano il problema che intendiamo risolvere. Si dice anche che questo tipo di metodi cerca di capire il *perché*. Il Metodo 1, ad esempio nel caso del Problema 3 relativo al giorno in cui acquistare il biglietto, è complesso o addirittura impossibile da applicare, in quanto dovremmo acquisire tanti dati sulle politiche di pricing che sono segreti, perché fanno parte del 'business' aziendale.

Metodo 2 – Raccogliere dati su come il fenomeno si è manifestato nel passato, ad esempio sui voli e i prezzi dei voli, e, senza "incaponirsi" sul perchè, cercare di imparare dai dati passati sul fenomeno l'andamento nel tempo del prezzo. Si dice anche che questo tipo di metodi cerca di capire il *cosa*.

I problemi che abbiamo descritto hanno la necessità di analizzare automaticamente grandi quantità di dati digitali, e richiedono lo sviluppo di tecniche algoritmiche basate su un paradigma nuovo, quello dell'apprendimento.

Possiamo collocare la nascita delle tecniche utilizzate nella Scienza dei dati a Londra, nel 1854. In quell'anno si diffuse a Londra una epidemia di colera. Un medico, John Snow, per cercare di comprendere le cause del la epidemia, iniziò a produrre mappe, come quella di Figura 7, che fa riferimento all'area di Broad Street.

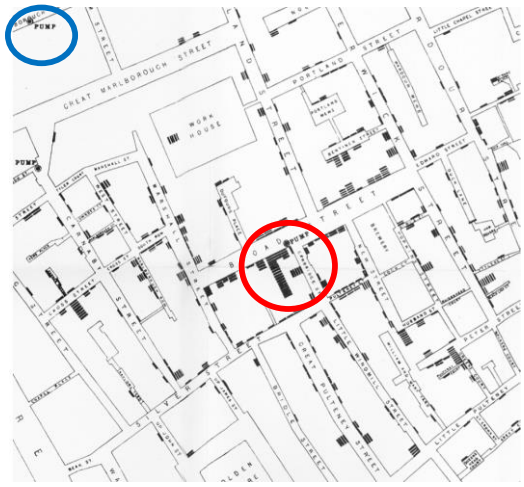


Figura 7 – Osservazioni sul colera a Londra

Snow ebbe la idea di mettere in relazione nelle mappe due fenomeni, la distribuzione delle pompe e dei decessi, che apparentemente non avevano niente in comune. Ogni quadratino nero nella mappa

rappresenta un decesso dovuto al colera, i cerchi rossi e blu rappresentano le aree attorno alle pompe (Pump nella mappa) dell'acqua potabile erogata dalle diverse compagnie. Se osserviamo i due cerchi, notiamo che nel cerchio rosso i morti sono molti di più. La visualizzazione della correlazione tra i due fenomeni fu decisiva per Snow per formulare la conclusione; la sospensione della erogazione dell'acqua nelle pompe della compagnia "rossa" portò ad una rapida riduzione dei morti.

L'insieme di tecniche, metodi, modelli, tecnologie, linguaggi dell'informatica e della statistica, insieme, per estensione, agli studi sulle conseguenze sociali dell'uso dei dati digitali, gli studi di scienze cognitive, le nuove leggi della economia digitale, l'insieme delle norme giuridiche ed etiche da seguire nell'uso dei dati digitali, viene oramai in tutto il mondo chiamato Scienza dei dati. Vedremo nel Capitolo 17 come le precedenti affermazioni trovino riscontro con l'offerta formativa di molte Università nel mondo.

La Scienza dei dati è nata pochi anni fa, si è infatti iniziato ad attribuire al concetto di *dato* la dignità di Scienza a partire dal nuovo millennio. Il motivo principale per cui si usa questo termine è la recente, grande diffusione dei dati digitali; i dati digitali sono i dati prodotti e scambiati nella rete Internet, nelle reti sociali, e in generale in ogni sistema che rappresenti la informazione mediante fenomeni fisici che adottano un alfabeto binario. Nel passato, i dati venivano scambiati mediante testi su carta, o a voce; a partire dai primi anni del secolo scorso si sono diffuse le schede perforate, e dagli anni 50' si sono diffuse le memorie magnetiche e in seguito le memorie a stato solido e le memorie ottiche, che in virtù della diminuzione dei costi e dell'aumento della capacità hanno permesso di memorizzare sempre più grandi quantità di dati. Oramai, oltre il 95% dei dati è prodotto e scambiato nel mondo utilizzando tecnologie digitali (i libri cartacei sono nel restante 5%, e speriamo che vivano a lungo…); ogni anno e mezzo, come abbiamo detto, raddoppiano i dati prodotti e scambiati sul World Wide Web (Web nel seguito), dando luogo ad una crescita esponenziale dei dati digitali nel tempo quale mai si è verificata nella storia della umanità.

3. Dati, piccoli dati, grandi dati

John Snow ebbe bisogno di pochi dati per formulare la propria ipotesi sulla diffusione del colera. I dati disponibili, inoltre, rappresentavano una piccola porzione della città di Londra; questa piccola porzione e i dati sui decessi disponibili per l'area di Broad Street potevano considerarsi un campione della intera città; la ipotesi sul ruolo delle pompe, formulata sul campione, venne estesa alla intera città. A lungo la statistica ha lavorato su pochi dati, e su campioni rappresentativi di un universo molto più ampio.

Negli anni recenti sono state prodotte tecnologie che, producendo o operando su dati digitali, hanno accresciuto la loro disponibilità nella descrizione dei fenomeni fisici e nella produzione di servizi, dando luogo al fenomeno detto dei Big data. Le più importanti tra tali tecnologie sono:

- I *social media* come Twitter o Facebook, che permettono, accedendo a funzionalità facili da usare, la comunicazione tra persone.
- *L'Internet delle cose*, o Internet of Things (IoT), che attraverso sensori distribuiti nel mondo fisico permettono di Integrare il mondo fisico attorno a noi con il mondo virtuale dei dati digitali.
- Il *cloud computing*, che rende facilmente accessibili e condivise grandi risorse di calcolo.

- Il *mobile computing* o *telefoni cellulari*, che ci rendono connessi sempre e ovunque e ci permettono di usare una miriade di applicazioni basate sui dati digitali.

Diverse definizioni sono state date del termine Big data. La definizione del McKinsey Global Institute è: "i big data si riferiscono a dataset (insiemi di dati) il cui volume è talmente grande che eccede la capacità dei sistemi di basi di dati tradizionali di catturare, immagazzinare, gestire ed analizzare".

Le principali caratteristiche che caratterizzano i big data sono:

- il volume, con riferimento alla dimensione dei dati nell'ambito delle tre coordinate di Figura 8. Le tre coordinate fanno riferimento a
  1. *L'ampiezza* della conoscenza sulla realtà osservata; si pensi al genoma umano, il cui sequenziamento è disponibile per un insieme sempre più ampio di persone.
  2. La *profondità* della conoscenza sulla realtà osservata; il sequenziamento del genoma umano fornisce informazioni sul corpo umano molto più ampie di quelle disponibili nel passato.
  3. Il *tempo*, alcune parti del genoma di una persona cambiano nel tempo, e l'analisi della evoluzione del genoma permette di comprendere, ad esempio, l'evoluzione delle malattie nella vita della persona.
- la *velocità*, intesa come tasso di generazione e trasmissione dei dati nella unità di tempo; pensiamo al valore dei titoli azionari di una borsa, che evolve con costanti di tempo di frazioni di secondo.
- la *varietà*, in termini di eterogeneità dei tipi di dati. Nel passato, i dati rappresentati nei calcolatori elettronici erano dotati di una struttura tabellare; questi tipi di dati sono detti dati strutturati. Successivamente i dati digitali sono evoluti verso formati semistrutturati o non strutturati come i documenti, che utilizzano come forma espressiva prevalente il linguaggio naturale, ovvero verso forme visuali di rappresentazione di dati, come le mappe geografiche, le immagini, i video, i suoni.

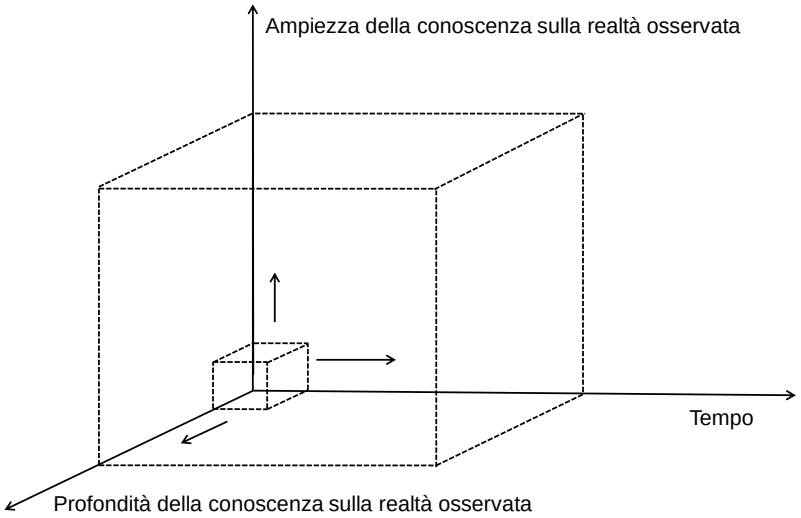


Figura 8 – Lo spazio dei big data

Accanto alle precedenti "v", altri aspetti caratteristici dei dati digitali sono:

- Il *valore*, inteso come utilità che il dato ha per il soggetto che lo elabora; l'utilità può riguardare una decisione che dobbiamo prendere, un processo produttivo di una azienda, una attività

amministrativa di una pubblica amministrazione. Il valore, come detto in [Beltram 2017] può riguardare lo sviluppo di nuove scoperte scientifiche, l'economia, la società.
- La *veridicità*, intesa come esattezza del dato nel rappresentare un fatto o un fenomeno del mondo reale, si pensi al diffondersi del fenomeno delle "fake news".
- La *viscosità* [Desouza 2014] intesa come la resistenza opposta da organizzazioni o tecnologie al fluire dei dati.
- La *volatilità* [Desouza 2014], l'intervallo temporale in cui i dati sono utilizzabili.

Nel seguito del libro, utilizzeremo volta a volta i termini "big data" e "grandi dati", per sottolineare che nell'ambito della Scienza dei dati possiamo iniziare a utilizzare accanto alla terminologia inglese anche una terminologia italiana, e che oggi siamo un po' in mezzo al guado.

4. I big data, la società la ricerca scientifica

Come evidenziato in [Beltram 2017], riferimento cui ci ispiriamo nella presente sezione, i dati digitali sono diventati talmente pervasivi da essere la principale risorsa utilizzata nella ricerca scientifica, nella economia e nella società. Man mano che cresce la disponibilità di dati digitali, e di tecniche per la loro analisi, la scienza dei dati può fornire in tutti i precedenti ambiti modelli di varia natura per la risoluzione di problemi come quelli descritti all'inizio del capitolo:

- *Modelli descrittivi*, che forniscono una risposta a domande del tipo: quali sono le caratteristiche salienti del paese, della società, del lavoro, della salute, della cultura, in questo momento della vita, mia, della mia famiglia, della comunità a cui appartengo? Per fare un solo esempio: quale percentuale degli italiani legge almeno un libro all'anno?
- *Modelli diagnostici*, o *interpretativi*, che forniscono una risposta alla domanda: perché è accaduto un certo evento, come, ad esempio, i risultati di una elezione politica, la crescita o il calo del "Prodotto interno lordo", l'aumento di fenomeni metereologici estremi?
- Modelli *predittivi*, che ci forniscono previsioni sul futuro, come ad esempio i modelli relativi ai problemi 3,4, e 6.
- Modelli *prescrittivi*, che ci forniscono strategie e decisioni che ci permettono di far accadere ciò che vogliamo (per esempio la mossa migliore in una partita a scacchi).

Riguardo all'impatto dei dati digitali sulla ricerca e sulla società, vi è una importante differenza tra dati utilizzati nella ricerca e dati utilizzati a scopo sociale. I primi sono generati secondo procedure controllate (ad esempio i risultati di un esperimento), i secondi sono usualmente ottenuti come risultato di interazioni tra esseri umani (ad esempio messaggi Twitter) ovvero tra esseri umani e tecnologie (es. transazioni economiche), e in tal modo sono caratterizzati da mancanza di controllo sulle modalità di generazione dei dati e più scarsa comprensione e possibilità di modellazione precisa del significato.

Con riferimento alla ricerca scientifica, in tutte le discipline scientifiche si sta affermando un approccio centrato sui dati (data-centric), affrontando problemi che nel passato sono stati considerati difficili o impossibili da affrontare. Così gli astronomi di tutto il mondo utilizzano lo Sloan Digital Sky Survey, che ha creato le mappe tridimensionali più dettagliate dell'Universo mai realizzate, con immagini

multicolori di un terzo del cielo e immagini spettrali per oltre tre milioni di oggetti astronomici. Un altro esempio riguarda Landsat, una costellazione di satelliti per telerilevamento che osservano la Terra; i dati raccolti, caratterizzati da volumi e precisione crescenti, sono usati per studiare l'ambiente, le risorse, e i cambiamenti naturali e artificiali avvenuti sulla superficie terrestre. Nella biologia e nella medicina, i dati digitali stanno rivoluzionando la ricerca; le tecnologie più recenti forniscono ai ricercatori grandi quantità di dati genomici, che, integrati con dati comportamentali, epidemiologici, ambientali, sociali, permettono di comprendere le basi genetiche della risposta alle medicine e permettono di affinare le strategie di cura delle malattie, nell' ambito della disciplina della medicina personalizzata. La meteorologia, l'agricoltura, la geologia, l'ambiente sono altri settori per cui l'approccio data driven sta potenziando profondamente la ricerca scientifica.

Anche la nostra vita di ogni giorno è profondamente influenzata dai dati digitali, crescono continuamente i momenti della nostra vita in cui accediamo a servizi e applicazioni basate sui dati digitali; d'altra parte i sistemi di raccomandazione, sulla base delle nostre interazioni e scelte di acquisito o di selezione di servizi, come nel caso di Amazon o Netflix, sono oramai in grado di costruire profili personali per ciascuno di noi, proponendoci prodotti da acquistare attraverso cross-selling ovvero avvisi pubblicitari.

La Scienza dei dati è anche un veicolo di innovazione fondamentale per le società, fornendo ai singoli cittadini e ai decisori pubblici una migliore comprensione dei sistemi socio economici, metodi per la comprensione di processi globali, per la pianificazione nello sviluppo del territorio e delle città, nel trasporto pubblico, nel consumo di energia, e strumenti di partecipazione inclusiva alle decisioni su base locale o globale. Allo stesso tempo, crescono le minacce di un uso distorto dei dati digitali per influenzare le opinioni e scelte politiche delle comunità, e sul rischio che i dati digitali aumentino, invece che ridurre, il divario sociale. Su questi aspetti torneremo nel Capitolo 14 e nel Capitolo 15.

5. Come è organizzato questo libro

I dati digitali stanno creando nuove professioni, come quella sempre più diffusa dello scienziato dei dati (data scientist), e stanno modificando le relazioni sociali e le leggi della economia. In questa sezione vediamo come abbiamo organizzato questo libro sulle Basi della Scienza dei dati, avvertendo peraltro il lettore che molte analisi, tecniche, metodi, modelli, leggi che fanno riferimento ai dati digitali sono ancora nella loro infanzia.

I punti vista che possiamo adottare in un percorso di comprensione del fenomeno dei dati digitali sono molteplici. In questo libro osserviamo questo fenomeno secondo due punti di vista tra di loro complementari.

Il primo punto di vista è centrato sul *ciclo di vita dei dati*, cioè su quell'insieme di fasi e attività che vengono compiute quando si vuole analizzare i dati per risolvere un problema o prendere una decisione. Fanno parte del ciclo di vita un insieme di fasi o attività che introdurremo nel Capitolo 2 e svilupperemo nei capitoli successivi, che riguardano le tecnologie, i modelli di rappresentazione dei dati, la qualità, la semantica, la integrazione, le tecniche di analisi, la visualizzazione, le astrazioni, e infine il valore.

Il secondo punto di vista riguarda le grandi *discipline scientifiche* che forniscono concetti e strumenti per affrontare e per comprendere criticamente il fenomeno dirompente dei dati digitali nella nostra epoca, e che a loro volta sono influenzate profondamente da questo fenomeno. Appartengono a queste tematiche, anzitutto, la Informatica e la Statistica, e, accanto ad esse, l'Economia, le Scienze sociali, le Scienze cognitive, la Linguistica, la Logica, l'Etica.

Questi due punti di vista sono messi in evidenza nella copertina del libro, insieme alle molteplici relazioni che sussistono tra le fasi del ciclo di vita e tra le discipline scientifiche. Fasi del ciclo di vita e discipline scientifiche sono anche messe in relazione nella Figura 9; le relazioni potranno risultare più chiare quando saremo entrati nel merito nei capitoli successivi. Forniamo ora nel resto del capitolo ulteriori elementi sulle fasi del ciclo di vita e sulle discipline scientifiche che approfondiremo maggiormente nel seguito.

									Valore
								Astrazioni	
							Visualizzazione		
						Tecniche			
					Integrazione				
				Semantica					
			Qualità						
		Modelli							
	Tecnologie								
Informatica	x	x	x	x	x	x	x	x	
Statistica		x				x			
Economia		x							x
Scienze sociali	x						x	x	x
Scienze cognitive			x	x			x	x	
Lingustica			x						
Logica		x		x	x	x		x	
Etica	x	x	x						

Figura 9 – Fasi del ciclo di vita dei dati digitali e discipline scientifiche messe in relazione con le fasi

Il Ciclo di vita

Ogni fenomeno che noi osserviamo ha una nascita, una evoluzione, una maturità e una decadenza. Così pure accade per i dati digitali, che sono caratterizzati da un ciclo di vita con una nascita, quando vengono acquisiti dal mondo fisico, una evoluzione e maturità, in cui vengono elaborati e analizzati, e infine una decadenza quando non ci servono più.

In questo ciclo di vita, i dati possono seguire tante storie. Ad esempio, il prezzo corrente di una azione ci serve per decidere su comprarla o no. Ma i prezzi passati di quell'azione ci forniscono una serie storica su cui ragionare per capirne l'evoluzione nel tempo e fare delle inferenze rispetto al futuro. Il prezzo corrente decade in fretta, ma quel prezzo non più valido diventa parte di una serie storica e quindi riacquista valore.

Se vogliamo usare i dati per le nostre esigenze e per rispondere alle nostre domande (es. i problemi introdotti nella Sezione 1), occorre conoscere bene come è organizzato il ciclo di vita, le tecniche a nostra disposizione e i metodi e i linguaggi informatici utilizzabili per esprimere e applicare tali tecniche. E occorre capire come affrontare le tematiche legate al volume, la velocità, la varietà, il valore, la veridicità dei big data. Analizzeremo il ciclo di vita dei dati digitali nel Capitolo 2.

I Modelli

Per poter analizzare e ragionare sui dati, abbiamo la necessità di rappresentarli mediante un modello. Guardiamo la Figura 10; nella parte superiore è descritto un insieme di studenti universitari che hanno superato degli esami riguardanti alcuni corsi universitari. Provate a leggere il testo; se vi chiedo di trovare i nomi dei corsi superati da Batini, oppure il voto medio ottenuto negli esami da Smith, dovete leggere più volte il testo individuando con fatica dove si annidano i dati necessari per rispondere alle domande.

Supponi di voler rappresentare i seguenti fatti mediante **tre** tabelle.
Ci sono tre studenti universitari, Batini con matricola 13242, che è nato in Italia, Xu con matricola 24195 nato In Cina, e Smith, con matricola 32845 nato in USA. Ci sono anche tre corsi, Analisi con codice 27 al primo anno, Algoritmi con codice 49 al primo anno e Logica con codice 77 al secondo anno. La matricola 13242 ha sostenuto l'esame del corso 27 con voto 25, la matricola 24195 il corso con codice 77 con 28 e la matricola 32845 il corso con codice 27 con voto 30

Una possibile rappresentazione

Studente

Matricola	Cognome	Stato Estero
13242	Batini	-
24195	Xu	Cina
32845	Smith	USA

Esame

Matricola	Codice Corso	Voto
13242	27	25
24195	77	28
24195	27	30

Corso

Codice	Nome Corso	Anno
27	Analisi	1
49	Algoritmi	1
77	Logica	2

Figura 10 – Dati rappresentati mediante tabelle

Nella figura viene mostrato come possiamo trasformare il testo in un insieme di tabelle che rappresentano in modo strutturato e ordinato i dati descritti nel testo. Se come suggerito utilizziamo tre tabelle, possiamo rappresentare nelle tabelle rispettivamente gli studenti, gli esami e i corsi, con le rispettive proprietà. Questo ci permette di rappresentare in ogni riga delle tre tabelle uno studente, un esame, un corso. Rispondere alle due domande iniziali è ora un pò più semplice, perché i dati hanno una struttura che ci aiuta a ritrovare quelli di nostro interesse, e intuiamo che avendo a disposizione un linguaggio per esprimere interrogazioni non sia complicato esprimere le due domande nel linguaggio.

Nel Capitolo 3 parleremo di modelli dei dati, arrivando anche a discutere dei modelli di dati utilizzati nel Web. Il Web è una immensa prateria in cui ciascuno di noi può condividere ciò che vuole; quando il Web viene utilizzato per condividere dati digitali, c'è la necessità di collegare tra loro dati prodotti da

diverse fonti. Intuitivamente ciò è possibile solo con una struttura diversa dalle tabelle, una tale struttura è un grafo, in cui i nodi rappresentano i singoli dati, e i rami collegano tra di loro dati diversi.

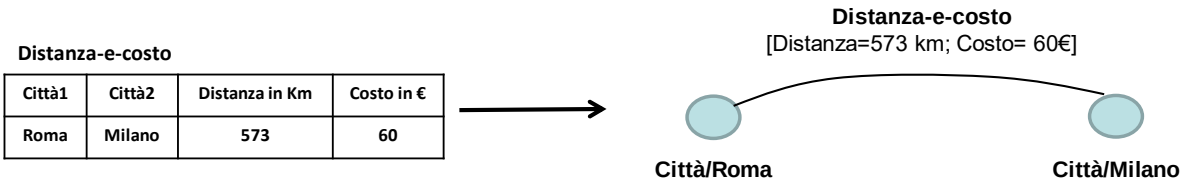


Figura 11 – Dati rappresentati mediante grafi

Le Tecnologie

Abbiamo brevemente visto nella precedente sezione quali siano le grandi tecnologie che hanno maggiore rilevanza nei dati digitali. Tali tecnologie avrebbero scarsa efficacia se la struttura stessa degli strumenti di calcolo, i calcolatori elettronici, non stesse vivendo a sua volta una profonda e rapida evoluzione.

La struttura classica di un calcolatore elettronico è quella della architettura di Von Neumann, in cui possiamo distinguere:

- una unità centrale di calcolo, dove vengono eseguiti i programmi;
- una memoria, dove vengono rappresentati i dati;
- componenti di input/output (ad esempio la tastiera o una stampante) che fanno comunicare il calcolatore con l'ambiente esterno.

Quando un calcolatore viene utilizzato per eseguire applicazioni software che operano su un insieme di dati, è necessario utilizzare un programma prodotto una volta per tutte da una azienda, che si chiama sistema di gestione di basi di dati, e che funge da interfaccia tra i dati e le applicazioni software, chiamate in Figura 9 funzioni software. L'insieme dei dati elaborati viene anche chiamato base di dati, concetto che approfondiremo nel Capitolo 3. Quando i dati sono tanti e cambiano velocemente, una tale architettura, in cui tutte le funzioni software utilizzano una sola unità centrale di calcolo e un solo insieme di dati, è intuitivamente inefficiente.

Nel Capitolo 4 descriveremo l'evoluzione delle architetture dati verso architetture distribuite, come quella rappresentata nella parte destra di Figura 12. Come si vede nella architettura distribuita, tante funzioni software vengono eseguite in parallelo su dati diversi, incrementando il numero di funzioni che possono essere eseguite nell'unità di tempo.

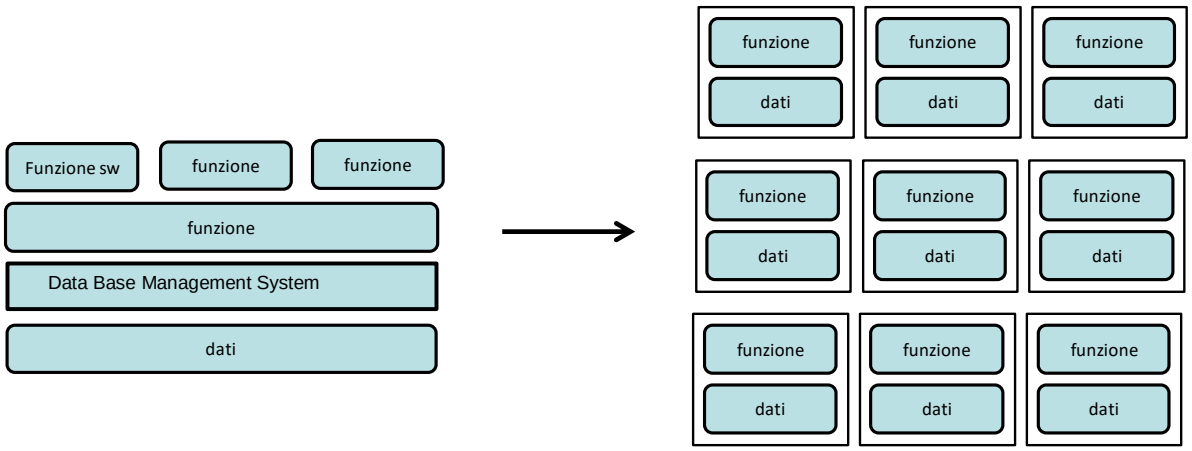


Figura 12 - Evoluzione delle architetture

La qualità dei dati

Supponete di avere la esigenza di fare un viaggio in treno da Milano a Roma, e di consultare i siti di Trenitalia e di Italo per conoscere gli orari di partenza e arrivo dei treni, in modo da poter scegliere l'orario più comodo. E' chiaro che gli orari riportati sui due siti devono essere precisi al minuto; se c'è un treno che parte alle 8 la mattina, deve partire proprio alle 8 in punto; magari ci potrà essere un ritardo, e il treno partirà più tardi, ma certamente non può partire prima. Anche le informazioni sui ritardi dovrebbero essere precise, ma sappiamo che spesso sono solo approssimate, e talvolta vengono aggiornate in aumento quando il ritardo si accumula. Insomma, i dati che noi utilizziamo devono rispettare determinate dimensioni di qualità, tra esse gli esempi precedenti fanno riferimento alla accuratezza, o precisione.

Nel Capitolo 5 discuteremo di qualità dei dati, e vedremo che la qualità può riguardare diverse caratteristiche del dato; ad esempio, oltre la accuratezza, la completezza, la consistenza, ecc. Vedremo inoltre che il problema della qualità dei dati diventa molto più complesso quando si passa dalle basi di dati utilizzate, per esempio, per l'orario ferroviario, alle informazioni pubblicate e scambiate nel Web. Ciò è naturale: mentre la pubblicazione dell'orario ferroviario è preceduta da una accurata serie di verifiche, sul Web ognuno è libero di pubblicare e di dire e scrivere ciò che vuole.

Umberto Eco diceva che il Web è talvolta simile al Bar Sport, in cui ciascuno dice ciò che gli viene in mente, spesso senza nessun filtro o verifica delle affermazioni fatte. E' per questo ordine di ragioni che abbiamo rappresentato nella Figura 13 il fenomeno della progressiva estensione dei dati digitali con una sfera opaca, per rappresentare il fatto che quanti più dati digitali vengono prodotti tanto più noi estendiamo sì la nostra conoscenza sul mondo, ma allo stesso tempo rischiamo di rappresentare il mondo in modo complessivamente meno nitido rispetto a quando avevamo a disposizione pochi dati. Nel Capitolo 5 parleremo della qualità dei dati, sia nelle basi di dati che nei dati digitali sul Web.

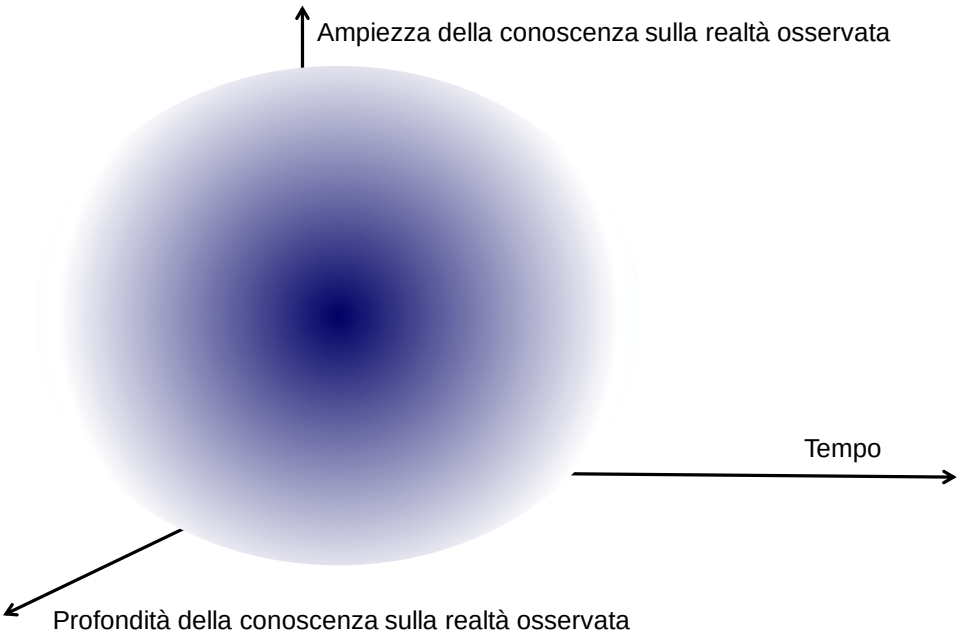


Figura 13 - La grande sfera opaca

L'integrazione dei dati

L'informazione sul Web è pubblicata da una miriade di soggetti; non meraviglia che essa risulti estremamente frammentata, e che solo molto raramente ci si preoccupi di collegarla. Anche l'informazione nelle basi di dati è frammenta; ad esempio, le basi di dati più importanti della Pubblica Amministrazione centrale italiana sono molte centinaia, con ciascuna centinaia di tabelle diverse. Le stesse informazioni sono ripetute spesso in diverse basi di dati, con formati diversi. Quando lavoravo all'Aipa abbiamo scoperto che gli indirizzi toponomastici hanno almeno una decina di formati diversi nelle basi di dati della Pubblica Amministrazione centrale! Nel Capitolo 6 parleremo di come i dati possono essere integrati e fusi in una unica versione, arrivando ad una rappresentazione riconciliata della realtà rappresentata.

Il significato dei dati

Supponiamo che le tre tabelle di Figura 10 siano rappresentate come in Figura 13, con nomi dunque meno espressivi per le tabelle e per le proprietà rappresentate nelle colonne. Non bisogna meravigliarsi del fatto che i nomi siano incomprensibili, non è una forzatura dell'esempio, spesso gli analisti per urgenze nello sviluppo di basi di dati vanno di fretta e non sono accurati nella documentazione e nella scelta dei nomi. Quel che è certo è il fatto che il significato delle tre tabelle e, di conseguenza, dei dati rappresentati è molto meno chiaro di prima.

T1

Mat	Att2	Att3
13242	Batini	-
24195	Xu	Cina
32845	Smith	USA

T2

Matr	Cod1	V
13242	27	25
24195	77	28
32845	27	30

T3

Cod2	NC	Y
27	Analisi	1
49	Algoritmi	1
77	Logica	2

Figura 14 – Tre tabelle dal significato poco comprensibile

Durante il ciclo di vita dei dati, può essere utile arricchire i dati di significato, per esempio, scegliere nomi più espressivi per le tabelle di Figura 14. In genere per capire il significato di un dato, chiediamo a qualcuno, magari a chi ce lo ha fornito, oppure ai tempi nostri facciamo delle ricerche sul Web, tramite Google o qualche altri motore di ricerca. L'arricchimento di significato dei dati sempre più spesso viene effettuato in modo automatico sfruttando le informazioni e la conoscenza disponibili nel Web. Ad esempio, se abbiamo due documenti in cui è citata la parola Roma, parola a cui possiamo far corrispondere almeno due significati, capitale d'Italia e quartiere di Buenos Aires, per capire se le due parole hanno lo stesso significato o no possiamo procedere come segue: consideriamo le parole "vicine" a Roma nel primo e nel secondo documento, e cerchiamo le voci di Wikipedia corrispondenti ai due gruppi di parole; mediante un opportuna tecnica su cui non entriamo qui nel merito, la distanza tra le due parole "Roma" può essere ricondotta alla distanza tra i due insiemi di voci, arrivando così ad una loro disambiguazione. Per potere arricchire di significato i dati, abbiamo bisogno di modelli cosiddetti semantici, in cui il significato sia espresso in modo tale da poter essere elaborato attraverso inferenze. Le tematiche riguardanti i modelli semantici saranno trattate nel Capitolo 7.

La trasformazione dei dati

Nei due capitoli precedenti abbiamo parlato di integrazione e di significato dei dati. Il processo di integrazione per poter essere condotto efficacemente ha bisogno di elaborare conoscenza sui dati; l'esempio che abbiamo fatto poco fa sul termine Roma può essere visto come un processo di integrazione tra due concetti; per disambiguarne il significato (sono la stessa cosa o cose diverse?) abbiamo bisogno di accedere a nuova conoscenza che troviamo in Wikipedia. Nel Capitolo 8 vediamo come la integrazione dei dati possa essere effettuata in modo più efficace se prima di integrare effettuiamo una trasformazione di modello, da tabelle in grafi semantici. La trasformazione ha dunqe lo scopo di rappresentare i dati in un modello che ci facilita la risoluzione di un problema.

La statistica

Nel Capitolo 9 parleremo della evoluzione delle tecniche statistiche. Abbiamo visto che John Snow ha analizzato i decessi a Londra, rappresentandone la intensità correlata alla presenza di pompe di acqua. Questo è un esempio di correlazione, una misura statistica che ha lo scopo di valutare la vicinanza di fenomeni. Le tecniche di correlazione sono diventate sempre più generali per rappresentare un insieme sempre più vasto di problemi. Il Capitolo 9, che parla di Statistica, è molto diverso dagli altri, è una sorta di storia della statistica vista in soggettiva, come se la statistica fosse una persona che ricorda il proprio passato e le proprie vicissitudini. Alla fine del capitolo sono suggeriti testi per una trattazione più sistematica dei concetti della Statistica.

Le tecniche basate su apprendimento

Se riconsideriamo il Problema 3 della introduzione, che ha lo scopo di predire il giorno in cui è conveniente acquistare un biglietto per un viaggio aereo, possiamo sviluppare tecniche che predicono il futuro sulla base delle regolarità riscontrate sui dati riguardanti i biglietti e i loro prezzi disponibili sul passato. Per capire la legge che lega il prezzo al giorno di acquisto, possiamo considerare le diverse tratte, il giorno della settimana, la compagnia aerea, e altre caratteristiche dei biglietti aerei, cercando delle regolarità e apprendendo le relazioni tra biglietto e prezzo che possono applicarsi ai nuovi biglietti. Nel Capitolo 10 parleremo delle tecniche di apprendimento, comunemente chiamate di machine learning.

Le visualizzazioni

Quando vogliamo rappresentare i dati contenuti in una tabella o più in generale dati che sono il risultato della applicazione di una tecnica di calcolo, spesso utilizziamo una rappresentazione visuale. Consideriamo l'esempio di Figura 15, tratto da [Tufte 2001]. La tabella a sinistra descrive per un certo numero di anni i consumi di riferimento stabiliti da una agenzia federale USA per le auto a benzina. Questa tabella può essere visualizzata mediante la rappresentazione visuale a destra, che attraverso la metafora di una strada fa corrispondere l'andamento nel tempo delle miglia per gallone alla larghezza crescente della strada. Peccato che la proporzione con cui cresce la larghezza della strada sia molto superiore alla proporzione di aumento dei valori numerici!

Anno	Miglia per gallone
1978	18
1979	19
1980	20
1981	22
1982	24
1983	26
1984	27
1985	27,5

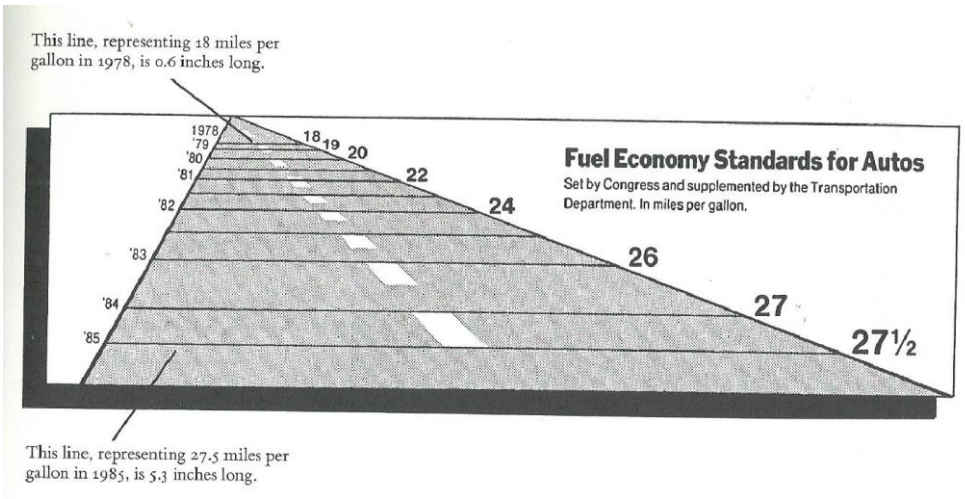


Figura 15 - Visualizzazione dei dati (tratta da E. Tufte – The Visual Display of Quantitative Information)

Nel Capitolo 11 parleremo delle visualizzazioni, mostrando quali grandi vantaggi comportino in termini di comprensione intuitiva del significato dei dati, e allo stesso tempo quali trappole possano presentarsi nelle rappresentazioni visuali, trappole cui dobbiamo fare attenzione per non essere ingannati nella comprensione.

Le astrazioni

Consideriamo la Figura 16; nella parte bassa della figura sono rappresentati un diagramma concettuale, una mappa geografica, un grafo.

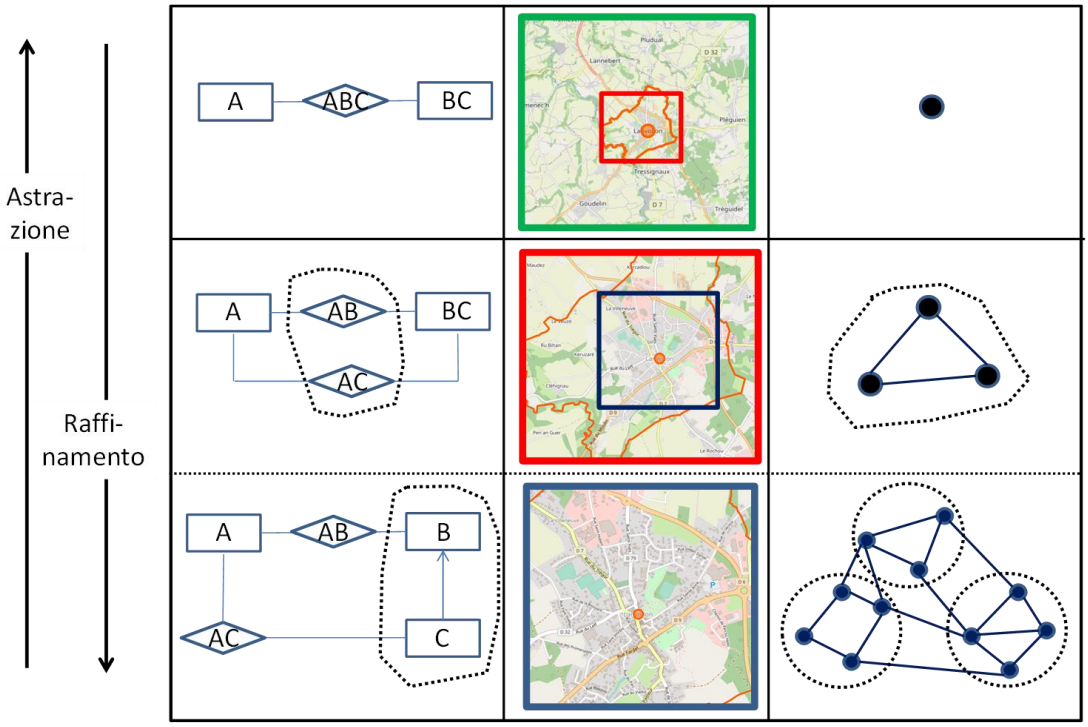


Figura 16 - Livelli di astrazione in un diagramma concettuale, una mappa, un grafo

Il diagramma concettuale rappresenta concetti come Persona, Lavoratore e Luogo, legami di sottoinsieme tra concetti (i Lavoratori sono un sottoinsieme delle Persone) e relazioni tra concetti (es. i Lavoratori lavorano in Luoghi). La mappa geografica rappresenta un frammento della rete viaria francese, il grafo è formato da nodi e archi, e per esso non è indicato nel disegno un significato particolare.

Per tutte e tre le rappresentazioni vengono mostrati nella parte superiore della figura delle rappresentazioni più compatte, nel seguito diremo più *astratte*. Il diagramma concettuale viene rappresentato con un numero inferiore di concetti, eliminando via via dettagli; la mappa è descritta rappresentando meno dettagli sulla rete di strade e sugli edifici nei centri abitati; il grafo viene via via semplificato fondendo gruppi di nodi e rami.

Possiamo dire che in tutti e tre i casi applichiamo nella figura trasformazioni di astrazione, eliminando via via dettagli, ovvero, inversamente, procedendo dall'alto in basso attraverso trasformazioni di raffinamento che, al contrario, introducono dettagli. Nella vita di ogni giorno noi facciamo spesso uso di astrazioni, quando vogliamo mettere in evidenza gli aspetti più rilevanti di un artefatto o fenomeno, e escludere dettagli non rilevanti; nei big data questa operazione di astrazione spesso diventa una necessità. Nel Capitolo 12 parleremo delle astrazioni.

Il valore dei dati

I dati digitali che abbiamo imparato a utilizzare tramite le applicazioni degli smart phone ci forniscono orami una miniera di servizi, dalla possibilità di sapere tra quanti minuti arriverà il tram che ci porta a casa, alle previsioni sull'inquinamento di cui al Problema 3, gli orari dei treni che questo pomeriggio partiranno per una città che intendiamo visitare, l'itinerario più rapido per andare a Budapest, ecc.; tutte queste informazioni hanno per noi un *valore*, e la misura di questo valore è spesso soggettiva, ed è legata allo scopo o decisione che dobbiamo prendere a seguito della disponibilità del dato. Il valore può essere un valore d'uso, un valore economico, come scopriremo nel Capitolo 13, un valore sociale, come vedremo nel Capitolo 14.

L'economia dei dati digitali

I dati digitali stanno cambiando le leggi della Economia; la Figura 17 mostra un libro cartaceo e un eBook.

Figura 17 - Come cambiano le leggi che regolano la economia

Produrre due copie di un libro cartaceo costa circa il doppio del costo di una copia, perchè dobbiamo utilizzare il doppio della carta e dobbiamo effettuare due rilegature. Analogamente, spedire due copie di un libro cartaceo a due indirizzi differenti costa il doppio che spedirne una copia. Nel caso della copia digitale dell'eBook, duplicare la copia digitale ha un costo trascurabile. La disponibilità di dati digitali con costi di riproduzione e di trasferimento su rete praticamente nulli, insieme ad altre caratteristiche dei dati digitali, trasforma la Economia classica dei beni e servizi; nel Capitolo 13 parleremo della Economia digitale.

Dati digitali e società

L'estensione raggiunta dalle reti sociali, l'utilizzo di dati digitali in settori come l'affitto di abitazioni o il noleggio di autovetture, l'uso di rappresentazioni digitali al posto delle vecchie rappresentazioni analogiche su carta nelle mappe stanno profondamente modificando, al di là della economia, le stesse

società, incrementando enormemente la comunicazione diretta tra esseri umani ma allo stesso tempo rischiando di distorcere profondamente i rapporti sociali e la comunicazione interpersonale. Pensiamo a come sta cambiando la comunicazione politica, che usa sempre più spesso canali sociali come Twitter o Facebook, che ampliano enormemente la platea dei lettori rispetto, ad esempio, a una intervista su un giornale, semplificando il messaggio e allo stesso tempo facendo appello alle emozioni piuttosto che al ragionamento e all'approfondimento della analisi. La Figura 18 mostra i risultati di uno studio che ha analizzato la diffusione in una rete sociale dei sentimenti di rabbia (nella parte superiore, avverto che tutte le figure a colori sono state riprodotte in scale di grigio per non far lievitare troppo i costi), allegria (a sinistra in basso), tristezza (a destra in basso) e disgusto (in fondo); i sentimenti di rabbia sono prevalenti. nel Capitolo 14 analizzeremo in maniera approfondita alcune di queste problematiche.

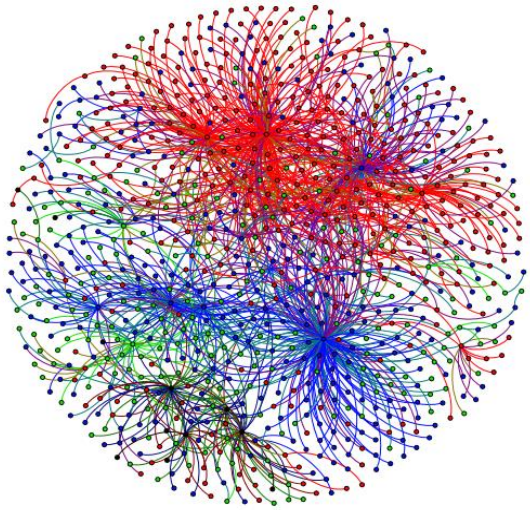


Figura 18 – I sentimenti nelle reti sociali

L'etica dei dati digitali

Tutti noi abbiamo un concetto di comportamento etico, anche se forse non sapremmo dare su due piedi una definizione di etica. La diffusione dei dati digitali impone di riconsiderare i temi legati all'etica, come il diritto alla privacy, la trasparenza nell'accesso ai dati pubblici, l'equità di trattamento delle tecniche predittive. Riguardo a questo ultimo punto, e per fare un solo esempio, si stanno diffondendo tecniche di predictive policing che, sulla base delle informazioni disponibili sui reati commessi nel passato, prevedono il luogo e il tempo in cui possono essere commessi in futuro nuovi reati. In diversi casi, è stato mostrato che tali tecniche sono orientate a sovrastimare la previsione per determinati gruppi sociali o etnici. Di tutti questi problemi parleremo nel Capitolo 15.

Il cosa e il perché: i limiti della scienza dei dati

Abbiamo visto nella precedente introduzione al Capitolo 10 sul machine learning che analizzando dati disponibili sul passato relativi a un fenomeno, possiamo prevedere eventi futuri (ad es. la zona della città con minor inquinamento) ovvero prendere decisioni o eseguire azioni (ad es. produrre la traduzione migliore di un testo da una lingua naturale a un'altra). Alcuni autori hanno ipotizzato che ormai non interessi più capire le cause dei fenomeni, come le leggi che regolano il prezzo dei biglietti

aerei (si dice anche: il perché), ormai le tecniche di machine learning producono modelli previsionali che analizzano i dati a "scatola nera" (si dice anche: il cosa). Ci sono dei limiti a questo metodo analitico che analizza il cosa senza chiedersi il perché?. Nel capitolo 16 approfondiremo questa problematica ed altre, che evidenziano i limiti della scienza dei dati; alcuni di questi limiti sono intrinseci, altri pongono sfide per il futuro della ricerca).

I dati digitali e le scienze cognitive

Abbiamo visto che la disponibilità di dati digitali sta profondamente innovando la ricerca in diverse discipline scientifiche; allo stesso modo, una tesi di questo libro riguarda il fatto che molte discipline scientifiche stanno dando contributi importanti al formarsi della Scienza dei dati. Nel Capitolo 17 noi esamineremo questa doppia influenza nell'ambito di una disciplina tra le più rilevanti nelle moderne società, le Scienze cognitive. Vedremo come la disponibilità di big data stia influenzando le metodologie di ricerca nelle scienze cognitive, e come le scienze cognitive possano dare un contributo sulla problematica delle cosiddette fake news e della post-verità.

La datacy

La cultura delle società è tradizionalmente misurata da due indicatori, la numeracy, che esprime la capacità di usare metodi matematici per risolvere problemi, e la literacy, che caratterizza la capacità di interpretare un testo scritto o un discorso, e di esprimere un pensiero attraverso un testo scritto. I temi trattati nel testo sono così vasti, nel cercare di dimostrare l'impatto dei dati digitali sulla nostra vita e sulla evoluzione delle società e delle attività economiche, che possiamo ipotizzare la nascita di un nuovo indicatore, che chiamiamo con il termine *datacy*. Il Capitolo 18 è dedicato a discutere quali siano i contenuti di questa nuova cultura, traendo spunto dalla esperienza in atto nel Corso di Laurea Magistrale in Scienza dei dati presso la Università degli Studi di Milano Bicocca, ed estendendola a molte altre offerte di contenuti nelle Università di diversi paesi del mondo in tema di Scienza dei dati.

6. Percorsi di lettura

La lettura di questo libro è sicuramente impegnativa, come per tutti i libri di 500 pagine e oltre. Nel tentativo di alleviare questo impegno, ho rappresentato in Figura 19 diverse figure in cui il lettore si può ritrovare, con i relativi possibili percorsi di lettura.

Il "tuttologo" è il caso più semplice, è il lettore che è interessato a tutto, ed è disposto ad investire il proprio tempo per capire ogni dettaglio di una disciplina: per lui/lei non ci sono problemi, è invitato a leggere tutti i capitoli.

L'interdisciplinare in genere vuole raggiungere un livello di comprensione adeguato nella disciplina discussa nel libro, ma è interessato soprattutto ai legami della disciplina con altre scienze, per capire quali aspetti di queste scienze stanno influenzando, nel nostro caso, il formarsi della nuova scienza dei dati, e quali aspetti siano invece influenzati.

L’esperto, chiamato anche data scientist, dovendo scegliere è probabilmente poco interessato a contenuti (che ritiene) di contorno o approfondimento, e privilegia le tecnologie, le metodologie e tecniche di gestione del dato, insieme alle tecniche statistiche e informatiche per la analisi del dato, e per la visualizzazione dei risultati.

Il manager ha un interesse complementare al precedente, e tende a trascurare gli aspetti tecnici e tecnologici, per focalizzarsi sui temi di economia e di scienze sociali.

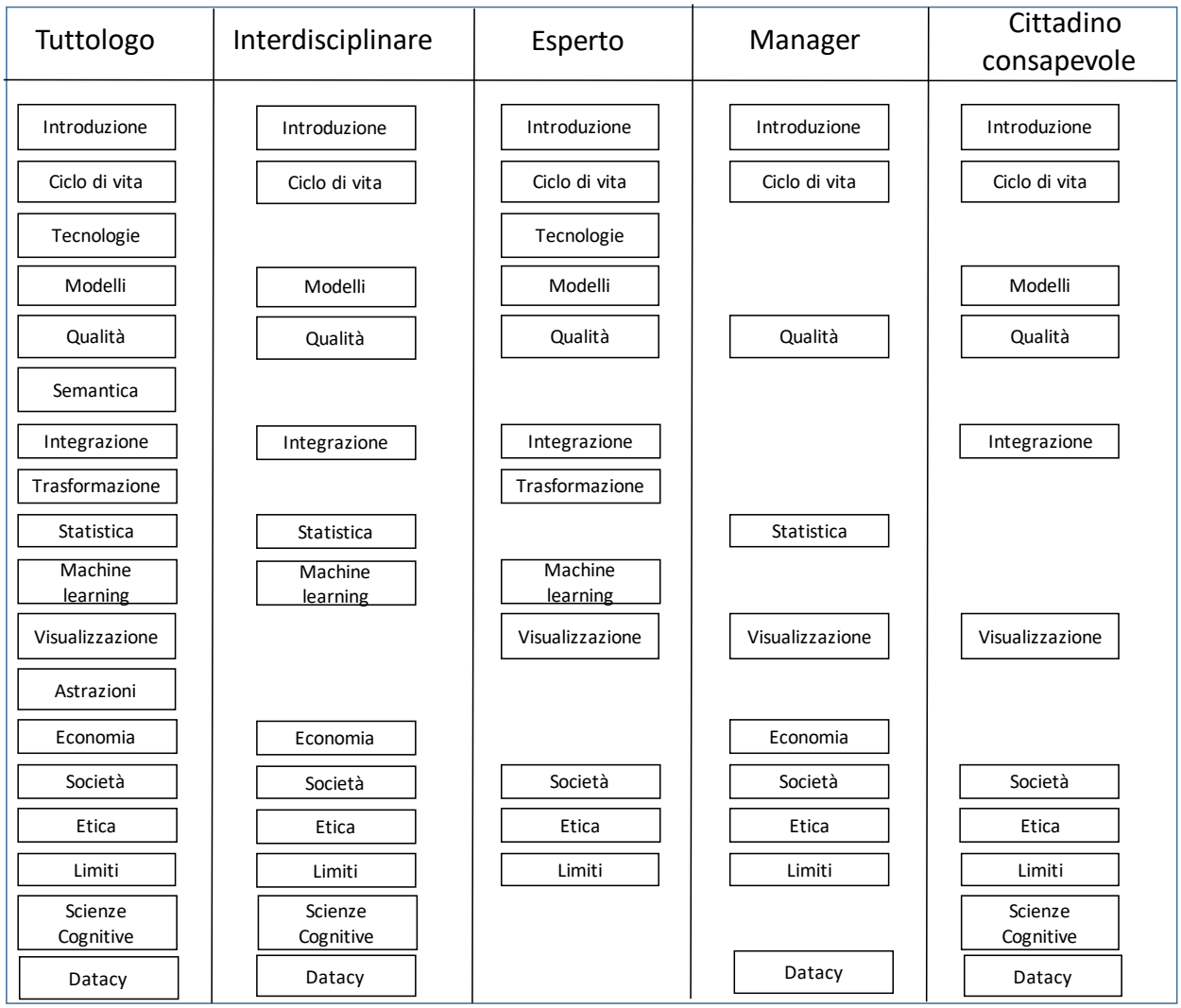


Figura 19 – Tipi di lettori e percorsi di lettura

E infine il caso più complesso, quello che abbiamo chiamato il cittadino consapevole, con interessi e livello culturale molto variegati nella nostra società. Credo che un sempre crescente numero di *cittadini* avranno il desiderio di capire ed essere consapevoli degli aspetti più rilevanti di questa nuova scienza, perché sentono che sta cambiando il nostro modo di vivere, ma sono ancora un po' confusi sul percorso culturale da compiere, i capitoli indicati sono un primo pacchetto da cui partire.

Notate che tutti i profili hanno in comune i capitoli finali su società, etica, limiti della scienza dei dati e sul concetto di datacy; questo perchè mi risulterebbe difficile toglierne qualcuno, sono tutti per me importanti, e sono importanti per tutti.

Riferimenti

AIPA - I dati pubblici: linee guida per l'accesso, la comunicazione e la diffusione, Febbraio 2002

F. Beltram, F. Giannotti, D. Pedreschi – Joint statement on new economic growth: the role of Science, Technology, Innovation and Infrastructure, Positio Paer on Data Science – G7 Academia meeting, 2017

C. Borgman – Big Data, Little Data, No Data, The MIT Press, 2017

C. Cesouza & K. Smith – Big data for Social Innovation – Stanford Social Innovation Review, 2014

E. Tufte – The visual display of quantitative information, 2001.

Appendice 1 – Tipologie di dati

In questa appendice riporto una sintesi di diverse definizioni e indicazioni alle Pubbliche Amministrazioni Centrali che l'Autorità per l'Informatica per la Pubblica Amministrazione formulò nel documento [Aipa 2002].

I soggetti pubblici, in virtù dei loro compiti istituzionali, raccolgono e trattano grandi quantità di informazioni, codificate in forma di dati, sui cittadini, le imprese, le istituzioni, il territorio e i principali fenomeni della vita del paese. La *conoscibilità* di tali dati è costituita dall'insieme di regole che disciplinano la fruibilità in favore dei soggetti interessati da parte dei soggetti pubblici che li raccolgono e li trattano. La conoscibilità comporta due fondamentali *qualità* dei dati: la *sicurezza*, intesa come insieme delle misure tese ad assicurare l'accesso ai soli dati conoscibili a un soggetto interessato, e l'*usabilità*, intesa come la facilità con cui un soggetto interessato ed abilitato a conoscere il dato riesce ad accedervi con le tecnologie disponibili, tenuto conto della sua situazione fisica, psichica e culturale.

Un dato conoscibile può essere reso noto al soggetto interessato mediante tre *modalità di scambio: accesso*, *comunicazione*, *diffusione*. L'accesso permette al soggetto interessato di fruire direttamente del dato; la comunicazione consiste nel far pervenire il dato ad uno o più destinatari abilitati predeterminati; la diffusione consiste nel rendere i dati disponibili ad una platea indeterminata di soggetti, anche tramite la loro pubblicazione, in forma tradizionale o su internet.

Con il termine *dato* si intende la rappresentazione di un fenomeno osservabile in un *formato codificato*, tale da essere memorizzabile ed elaborabile. Tale formato codificato appartiene in generale a un insieme di possibili *elementi*, chiamato *dominio di definizione* del dato. Il dominio è in genere denotato da un *nome*. Il nome del dominio può essere associato anche al dato. In tale accezione, che assumeremo nel seguito, il dato è una coppia <Nome, elemento di dominio>. Ad esempio il dato <ETÀ, 37>, che rappresenta attraverso un formato codificato (appunto, "37") l'età di una persona fisica, ha come dominio dei possibili elementi l'insieme dei numeri interi da 0 a 150.

La locuzione *dati pubblici* può essere interpretata secondo diverse accezioni:
1. dati accessibili pubblicamente: questa definizione fa riferimento all'assenza di requisiti di riservatezza, e riflette quindi un aspetto legato alla legittimità della consultazione da parte di soggetti comunque interessati.
2. dati detenuti da un soggetto pubblico: fa riferimento alla natura pubblica del titolare del trattamento che può eventualmente esserne anche il produttore.
3. dati di interesse di un soggetto pubblico: fa riferimento alla natura pubblica del fruitore nell'interesse della collettività.

I dati e in particolare i dati pubblici possono essere classificati secondo diverse caratteristiche: Quelle che rilevano qui sono:
1. identificabilità;
2. presenza in registri pubblici o simili;
3. aggregazione e generalizzazione;
4. grado di elaborazione;

5. utilità per i soggetti interessati ad accedervi.

Identificabilità - I dati possono essere riferiti a soggetti, persone fisiche o giuridiche, identificati o identificabili, nel qual caso sono *dati personali*, oppure non essere riconducibili a singole persone fisiche o giuridiche, nel qual caso sono *dati anonimi*. Alcuni dati personali, per la delicatezza che li caratterizza, sono definiti *sensibili*, idonei cioè a rivelare l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni od organizzazioni a carattere religioso, filosofico, politico o sindacale, lo stato di salute e la vita sessuale.

Presenza in registri pubblici - In alcuni casi i dati sono *provenienti da pubblici registri, elenchi, atti o documenti conoscibili da chiunque*, senza condizioni. Essi sono tenuti o formati da uno o più soggetti pubblici, in virtù di una norma di legge o di regolamento; la norma che è alla base della conoscibilità in questi casi può prevedere particolari modalità di accesso o limiti temporali che vanno rispettati anche in caso di comunicazione o di diffusione dei dati. Il solo fatto di poter rinvenire un dato personale, ad esempio l'indirizzo di posta elettronica, in uno spazio pubblico di internet, non comporta che esso sia conoscibile da chiunque e quindi non autorizza l'uso libero del dato.

Aggregazione e generalizzazione - I dati sono detti *elementari* quando rappresentano un aspetto della realtà non ulteriormente riconducibile, date le ipotesi adottate, ad aspetti più semplici, *statistici* quando sono il risultato di una elaborazione attraverso funzioni di aggregazione su dati elementari. Un esempio di dato elementare è l'età di una persona; esempi di dato statistico sono il numero di soggetti con quell'età in un territorio specifico e l'età media della popolazione di un comune.

Secondo un diverso processo di astrazione, usualmente detto di generalizzazione, si distinguono i *metadati* e gli *schemi di dati*. I metadati sono proprietà di un insieme di dati. Uno schema di dati consiste nella descrizione di un insieme di classi di dati e delle relazioni che intercorrono tra di essi.

Ad esempio:
1. un quadro di un orario ferroviario è un insieme di dati;
2. la proprietà per cui gli orari di arrivo e partenza siano rappresentati mediante due cifre per l'ora e due cifre per il minuto è un metadato;
3. la conoscenza della relazione tra treni e città collegate da treni attraverso gli orari di arrivo e di partenza dei treni in o rispettivamente da tali città è lo schema dei dati del quadro. I metadati e gli schemi di dati sono in genere anonimi, perché non fanno riferimento ad uno specifico soggetto fisico o giuridico.

Grado di elaborazione - Questa classificazione, tratta dall'importante rapporto Mandelkern1 commissionato dal governo francese, tiene conto delle elaborazioni cui i dati sono stati sottoposti. Secondo tale classificazione, i dati pubblici possono essere così suddivisi:

a. *dati grezzi*: sono i dati raccolti ma non ancora sottoposti ad elaborazioni significative e che quindi si trovano sostanzialmente nella forma in cui sono stati acquisiti, quali:

1. dati acquisiti tramite digitazione da moduli di carta (oppure via internet tramite moduli elettronici), prima delle verifiche tese a confrontarne il contenuto con altri dati contenuti in archivi di riferimento o raccolti contestualmente;
2. messaggi ricevuti per posta elettronica prima di operazioni di marcatura del testo le quali ne caratterizzino specifiche parti o informazioni presenti;
3. dati geografici prima delle operazioni di normalizzazione;

b. *dati di base*: sono i dati già sottoposti alle elaborazioni necessarie a renderli elaborabili al di fuori di un singolo sistema o di una singola tecnologia, solitamente da parte di soggetti diversi da quello che li ha raccolti, ad esempio:
1. indirizzi postali normalizzati;
2. nominativi di professionisti iscritti ad un albo;

c. *dati arricchiti* (o *elaborati*): sono i dati risultanti da operazioni di ricerca e di confronto con informazioni di differente provenienza, ma collegate ad uno stesso fenomeno; la categoria contiene anche i dati aggregati in senso statistico, come medie, indici ed altro. Esempi di dati arricchiti sono:
1. la posizione fiscale di un'impresa come risulta da diverse basi di dati del Ministero dell'economia e finanze;
2. i dati contenuti nello stato di famiglia di un cittadino;
3. l'indice dei prezzi al consumo collegato ad una città;
4. una carta geografica ricavata da dati fotogrammetrici grezzi.

I *dati essenziali* sono i dati pubblici dei quali i cittadini, le imprese e altri operatori privati devono poter disporre per esercitare i propri diritti. I dati essenziali possono essere sia anonimi, come le norme e molti dati statistici, sia personali (la maggior parte dei dati amministrativi); per questi ultimi è necessario mettere a punto regole che ne assicurino la protezione senza ostacolarne la conoscibilità da parte degli aventi diritto. Pur non esistendo una definizione giuridicamente rilevante dei dati essenziali, si possono fare rientrare in questa categoria i dati la cui conoscibilità è sancita da leggi, ad esempio quella relativa alla trasparenza amministrativa. La seguente è una lista, indicativa e non esaustiva, di dati da considerare essenziali:
1. le leggi e i regolamenti vigenti;
2. i dati statistici nazionali più importanti o necessari per le decisioni individuali o collettive;
3. i dati personali in possesso dei soggetti pubblici e che riguardano il richiedente;
4. le indicazioni necessarie ad usufruire dei servizi erogati da soggetti pubblici e a verificare lo stato dell'iter amministrativo (portali e sportelli di accesso unificati ai servizi, organigrammi, indirizzi postali, numeri di telefono, eccetera).

Secondo un'interpretazione estensiva, si potrebbe includere tra i dati essenziali per un cittadino, o un'impresa o un altro operatore privato, tutti i dati ad essi relativi in possesso di soggetti pubblici: ne consegue che potrebbe essere garantito non soltanto il diritto di conoscerli, ma anche quello di aggiornarli, sia pure secondo procedure individuate da ciascun soggetto pubblico o concordate fra più soggetti pubblici, comunque rese note e, se necessario, sancite da norme.

Sempre estensivamente, potrebbero essere considerati essenziali per un soggetto pubblico i dati, di qualsiasi natura, in possesso di altri soggetti pubblici e necessari al primo per adempiere ai propri compiti istituzionali.

Capitolo 2 - Il ciclo di vita del dato digitale

Carlo Batini

1. Il ciclo di vita dei dati nei sistemi informativi tradizionali

Nel passato, diciamo dagli anni 50 agli anni 90 del secolo scorso, i dati venivano rappresentati nei sistemi informativi delle organizzazioni sotto forma di tabelle, un cui esempio è un orario dei viaggi che partono da Milano per Roma, organizzati da una azienda di trasporti, vedi Figura 1. L'orario può essere utilizzato per permettere a un utente di prenotare un posto su un treno e acquistare un biglietto. Un insieme di tabelle utilizzato da un insieme di programmi informatici è anche chiamato *base di dati*.

Partenza	Arrivo	Durata	Treno	Prezzo
Milano Centrale 18:00 →	Roma Termini 20:55	2h 55'	Frecciarossa 1000 **9657**	
Milano Centrale 18:20 →	Roma Termini 21:40	3h 20'	Frecciarossa 1000 **9559**	da 107,00 €
Milano Centrale 18:30 →	Roma Termini 21:29	2h 59'	Frecciarossa 1000 **9659**	da 92,00 €
Milano Centrale 19:00 →	Roma Termini 21:59	2h 59'	Frecciarossa 1000 **9663**	da 92,00 €

Figura 1 - Esempio di tabella (tratta dal sito eDreams)

Ogniqualvolta noi consultiamo un sito di una azienda di trasporti ferroviari per prenotare un posto su un treno, effettuiamo una operazione che già nel Capitolo 1 abbiamo chiamato transazione, consistente nella scrittura in alcune tabelle dei dati relativi alla prenotazione. Il precedente è un esempio di elaborazione chiamata di On Line Transaction Processing (OLTP), un tipo di elaborazione in cui le operazioni prevalenti su una base di dati sono le transazioni. Accanto alle transazioni, le basi di dati della azienda possono essere interrogate per acquisire informazioni sui ritardi o inviare richieste di rimborsi, vedi Figura 2.

Accanto alla prenotazione e acquisto biglietti, un altro servizio fornito dalla azienda è il vero e proprio viaggio, durante il quale possono essere fornite o richieste informazioni sui ritardi, sulla velocità del treno e così via. Finito il viaggio, la relazione tra cliente e azienda potrebbe proseguire con la raccolta dei punti fedeltà, oppure con una richiesta di rimborso (vedi ancora Figura 2) o con la denuncia di un furto. Tutti i precedenti processi sono gestiti in genere da diverse basi di dati, e richiedono la esecuzione di interrogazioni o transazioni su tabelle.

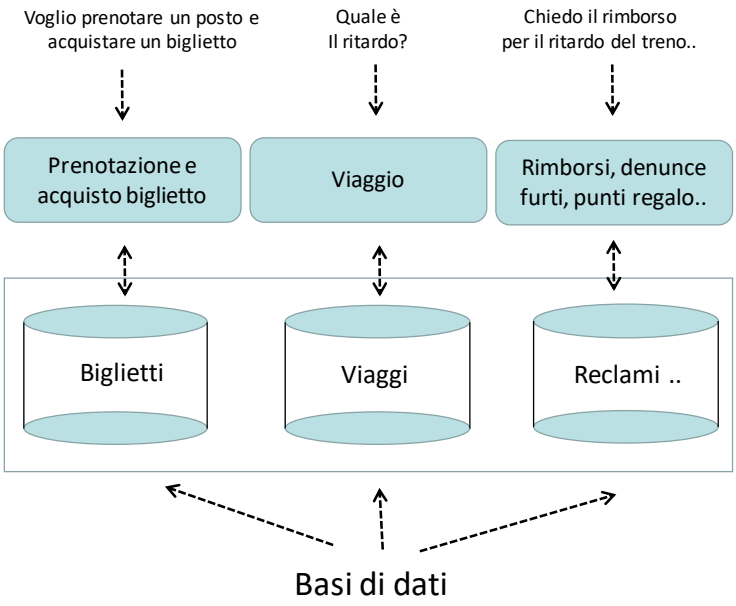


Figura 2 – Interrogazioni e transazioni

L' azienda di trasporti può essere interessata mensilmente o annualmente a costruire informazioni di sintesi, richieste attraverso interrogazioni, due delle quali sono rappresentare nella prossima Figura 3, parte destra. Le interrogazioni, a differenza delle transazioni, non modificano le tabelle, ma estraggono dati; se ci pensate un attimo, e osservate la Figura 3, per rispondere alle due interrogazioni è necessario mettere insieme informazioni che compaiono in diverse basi di dati.

Ecco perché, accanto all'OLTP, le organizzazioni hanno sempre avuto bisogno di un secondo tipo di elaborazioni, chiamate On Line Analytical Processing (OLAP), in cui ad esempio si vuole avere risposta alle due domande mostrate in Figura 3.

Il ciclo di vita dei dati elaborati a fini di OLAP nei sistemi informativi classici è composto da tre fasi, chiamate sinteticamente ETL a partire dai loro nomi inglesi:

- una fase di estrazione (Extract) dei dati dalle basi di dati utilizzate dai processi operativi, cioè nel nostro esempio le tre basi di dati Biglietti, Viaggi, Reclami.
- una fase di trasformazione (Transform) e integrazione dei dati, per creare ad esempio una unica tabella, e
- una fase di caricamento (Load) dei dati nel sistema di analisi OLAP.

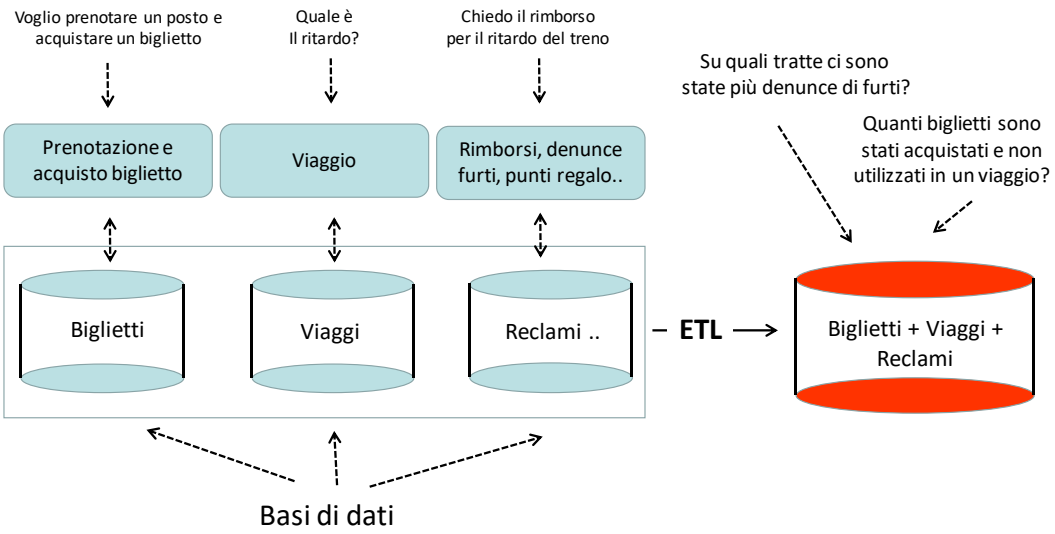


Figura 3 – On line transaction processing e On line analytical processing

Esercizio 1 - Secondo te, a quale o quali delle tre basi di dati si deve accedere per poter dare risposta alle due interrogazioni:
D1. Su quali tratte ci sono state più denunce di furti?
D2. Quanti biglietti sono stati acquistati e non utilizzati nei viaggi?
Rispondi sulla base del significato intuitivo dei contenuti delle tre basi di dati Biglietti, Viaggi, Reclami.
Vedi la risposta in Appendice 1.

Questo capitolo descrive come il ciclo di vita dei dati sta evolvendo dal vecchio mondo in cui i dati su un fenomeno erano pochi e spesso noti solo per un campione dell'universo, al nuovo mondo in cui i dati a disposizione sono diventati tanti. La Sezione 2 parla del ciclo di vita in generale, le sezioni successive sono dedicate alla fase di Definizione del problema, Sezione 3, la fase di Scelta e acquisizione dei dati digitali, la Sezione 4 e la Sezione 5 per la fase di Gestione, che come vedremo è molto articolata al suo interno, la Sezione 6 sulla fase di Analisi, e la Sezione 7 dedicata alla Visualizzazione.

2. Il ciclo di vita nei big data – introduzione

Rispetto al ciclo di vita tradizionale dei dati, quello dei big data è più articolato e complesso. Infatti:

1. i dati possono essere acquisiti da una varietà di fonti, dalle basi di dati dei sistemi tradizionali, alle basi documentali, ai sensori dell'Internet delle cose (Internet of Things, IoT), le frasi in linguaggio naturale nelle reti sociali, le foto e altri tipi di immagini di Instagram, i file di dati risultato di esperimenti, e molte altre fonti.
2. Come conseguenza, i dati digitali sono rappresentati con una varietà di formati, come dati strutturati, testi, mappe, immagini, video, suoni.
3. Le fonti possono offrirci dati che non sono adeguati al processo di analisi, e può quindi essere necessario esplorarne di nuove, per trovare i dati utili, ovvero modificare l'obiettivo per mancanza di dati adeguati.

4. A seconda della fonte e delle modalità di misurazione, i dati sono in genere disponibili in un formato che esprime il loro significato in modo spesso parziale e incompleto.
5. In virtù della grande varietà di fonti, i dati rappresentano spesso solo una porzione delle informazioni utili per il problema, ed è quindi necessario "metterli insieme" per produrne una versione integrata.
6. A seguito della imprecisione dei sensori utilizzati nell'Internet delle cose, le ambiguità del linguaggio naturale, gli errori che possono essere commessi nei processi di produzione delle versioni digitali, la qualità dei dati, intesa, ad esempio, come la loro accuratezza nel descrivere fenomeni e la completezza di rappresentazione, può essere non adeguata alle esigenze della analisi.

Per tutte le precedenti ragioni, il ciclo di vita dei dati digitali nell'epoca dei big data richiede la esecuzione di un insieme di attività più esteso rispetto al ciclo di vita tradizionale, attività che sono descritte con una breve definizione del loro scopo nella Figura 4. Approfondiremo la presentazione delle attività, utilizzando due esempi, riportati nelle due colonne finali di Figura 4:

- la applicazione Breezometer (www.breezometer.com), prodotta da una startup israeliana per risolvere il Problema 4 del Capitolo 1, con lo scopo di fornire un servizio che ci dica quali sono le condizioni di inquinamento in un certo luogo in una certa ora, e
- una anagrafe di un gruppo di professionisti europei, che si vuole analizzare per produrre statistiche sulla distribuzione dei professionisti per aree geografiche e per caratteristiche anagrafiche.

		Breezometer	EU Profess.
Formulazione del problema	Fase in cui si decidono gli scopi per cui I dati vengono raccolti e analizzati	x	x
Scelta e acquisizione dei dati	Vengono individuate e acquisite le fonti che forniranno gli insiemi di dati (dataset) oggetto di analisi	x	x
Scelta delle fonti	Tra tutte le fonti potenziali, vengono scelte quelle più utili sulla base del costo, qualità, livello di aggiornamento, ecc.		
Acquisizione dei dati grezzi	A seconda delle fonti, I dati vengono acquisiti, ovvero misurati con sensori, organizzati in dataset di dati grezzi	x	
Gestione (o preparazione) dei dati	I dati sono osservati e elaborati con lo scopo di prepararli alla fase di analisi		
Fasi sulla semantica dei dati	Attività che si riferiscono al significato dei dati		
Modellazione	I dati sono rappresentati con un modello formato da strutture di rappresentazione e regole per il loro utilizzo		x
Profiling	I dati vengono analizzati, sono prodotte statistiche (es. su frequenze/distribuzioni) al fine di attrribuire un significato ai dati		x
Arricchimento semantico	Viene arricchita la descrizione del significato dei dati		x
Normalizzazione	I dati sono trasformati in un formato standard		x
Gestione dei metadati	Vengono associati valori ai metadati, quali la provenienza, l'owner dei dati, la data di acquisizione, ecc.	x	
Trasformazione di modello	Processo che opera una trasformaizone di modello sui dati, effettuata tramite opportune operazioni primitive		
Controllo di qualità	Gesitione delle proprietà desiderate dei dati nella rappresentazione del mondo a cui I dati si riferiscono		
Valutazione della qualità	Vengono definite le caratteristiche di qualità (es. Accuratezza) e misurate mediante metriche		
Miglioramento della qualità	Vengono indiviuduate tecniche per il miglioramento dei valori delle caratteristiche di qualità		x
Fasi sulla integrazione dei dati			
Deduplicazione	In ogni dataset si cercano I gruppi di n-ple che fanno riferimento allo sesso oggetto e si integrano in una unica tupla		x
Integrazione	Attività in cui dataset diversi vengono integrati e fusi in un data set unico	x	
Fusione	Se due o più n-ple vengono fuse in una unica n-pla, si decide quale valore scegliere nella nuova n-pla		x
Implementazione della architettura tecnologica	Attività che fanno riferimento alla scelta delle tecnologie informatiche per la elaborazione e memorizzazione dei dati		
Scelta della architettura tecnologica	Scelta dei componenti tecnologici e delle interazioni tra essi		
Scelta della architettura di memorizzazione	Characteristics and techhnological implementation of data that outlives the process that created it		
Analisi dei dati	Applicazione di modelli e tecniche sui dati per raggiungere l'obiettivo inizialmente definito		
Scelta delle features	Scelta delle caratteristiche dei dati (es. Età per persone, data scadenza per prodotti, ecc.) su cui basare la analisi	x	
Definizione del modello statistico	Construction or application of techniques that can learn from and make predictions on data		
Defnizione della tecnica di learning	Learning is performed inferring target properties from labeled training data		
Test di qualità	Verifica su dati di test del grado dsi raggiungimento degli obiettivi		
Visualizzazione dei dati	Processo di rappresentazione dei dati mediante diagrammi, simbolismi e metafore che ne facilitino la comprensione		
Scelta del tipo di visualizzazione	Vengono scelti I tipi di diagrammi, ovvero le metafore e I simboli visuali utilizzati per la visualizzazione	x	
Produzione della visualizzazione	Vengono scelte e applicate le tecniche algoritmiche per produrre la visualizzazine	x	
Pubblicazione dei dati	Pubblicazione sul Web dei dati in formato utilizzabile da chi sia interessato, e collegabile con altri dataset		

Figura 4 – Attività del ciclo di vita dei big data, con indicazione dei passi esemplificati con i due esempi.

Nella Figura 4 sono riportate le attività in cui utilizzeremo i due esempi. E' bene osservare che l'elenco di Figura 4 non rappresenta un ciclo di vita in cui a ogni attività segue necessariamente la successiva; le attività nella figura vanno viste come una cassetta degli attrezzi, che va volta per volta utilizzata per capire quali attrezzi (attività) ci servono, e in quale ordine.

La Figura 5 mostra la versione iniziale del dataset che costituisce il punto di partenza della analisi sui professionisti europei.

3. Fase di formulazione del problema

In questa fase si definisce nella maniera più precisa possibile il problema da affrontare. Accade spesso che la grande disponibilità di dati digitali porti i responsabili delle aziende o delle amministrazioni pubbliche a dire: abbiamo tanti dati, e adesso che cosa ci facciamo? Questo punto di vista è pericoloso,

Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20…	Ital
Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Figura 5 – I dati per la applicazione "professionisti europei" sono inizialmente disponibili così

perché rischia di far perdere tempo e soldi nel tentativo di fare qualcosa di utile con i dati. Quali classi di problemi possiamo risolvere con i dati? Abbiamo già dato una prima risposta a questo quesito quando nel Capitolo 1 abbiamo descritto i modelli che possono essere prodotti dalla attività di analisi. Riprendiamo quell'elenco, esemplificando in maniera più concreta alcune classi di problemi senza pretendere di essere esaustivi:

1. Interpretare un fenomeno, essendo in grado di produrre statistiche di varia natura, a partire dai valori medi di determinate variabili descrittive del fenomeno, le varianze, valori aggregati di variabili in funzione di determinate dimensioni (es. persone per età e sesso che abitano in una città): questi temi sono anche chiamati di statistica descrittiva. Corrispondono ai modelli descrittivi.

2. predire un evento, ad esempio quando convenga sostituire un pneumatico o cambiare la frizione, ovvero, cambiando scenario, quando convenga sostituire un componente in una rete telefonica o di erogazione di energia (questa attività è chiamata di manutenzione predittiva), ovvero ancora predire quando un utente di una società di servizi è tentato di cambiare passando alla concorrenza (abbiamo già chiamato *churn* questo fenomeno nel Capitolo 1). Corrispondono ai modelli predittivi.
3. trovare correlazioni tra variabili descrittive di un fenomeno, come ad esempio per una malattia al polmone quale sia la correlazione con il fumo e con i livelli di inquinamento in cui ha vissuto un campione della popolazione. Corrispondono sia ai modelli diagnostici, sia ai modelli predittivi.
4. ricostruire le cause di un fenomeno, ad esempio trovare le cause di un guasto in un prodotto, ovvero di un attacco alla sicurezza, ovvero di un fermo di sistema, ovvero di un comportamento anomalo di un cliente di un circuito di carte di credito per intercettare furti di carte o furti di identità. Ricostruire le cause è più ambizioso rispetto a trovare le correlazioni, perché la relazione causale esprime una relazione asimmetrica tra variabili; ci può essere alta correlazione tra consumo di dentifricio e uso di apparecchi per la igiene dentaria, senza che vi sia nessuna relazione di causa effetto tra i due fenomeni. Corrispondono ai modelli diagnostici.
5. prendere una decisione, ad esempio, su un appartamento che vogliamo acquistare o prendere in affitto, ovvero, per una azienda produttrice di beni da immettere nel mercato, decidere quale prodotto vendere e con quali caratteristiche, ovvero per una azienda di trasporti urbani pianificare le partenze degli autobus in modo da massimizzare l'incontro con le esigenze degli utenti. In genere, in questi problemi la decisione deve essere ottimizzata rispetto ad un obiettivo. Corrispondono ai modelli prescrittivi.

Nei nostri due esempi i problemi possono essere formulati nel seguente modo:

Breezometer – essere in grado di predire per i dintorni di un determinato luogo specificato tramite coordinate spaziali, in un determinato giorno dell'anno, in una determinata ora, il livello di inquinamento delle polveri sottili e di altri agenti inquinanti, vedi Figura 6. Siamo nell'ambito del Problema 2, predire un evento.

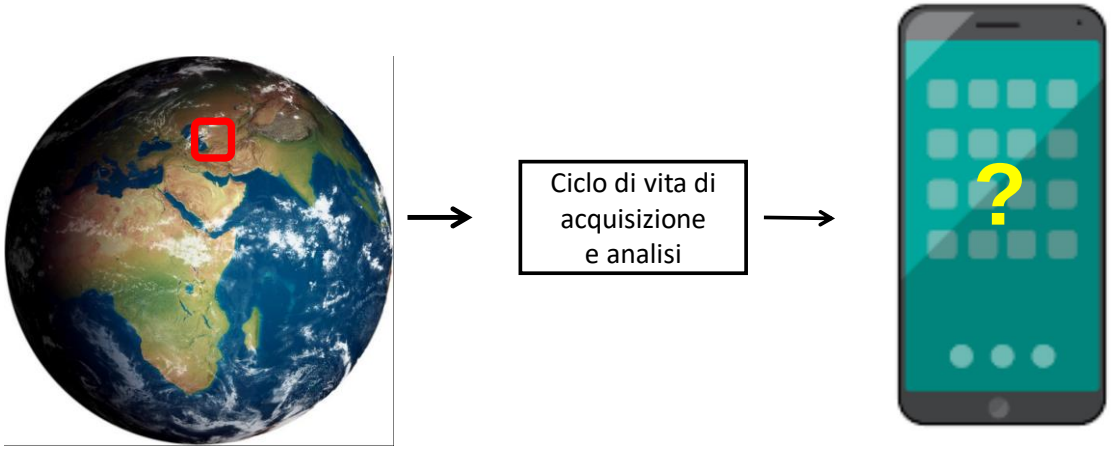


Figura 6 – Obiettivi dello strumento Breezometer

Professionisti europei – analizzare il fenomeno dei professionisti europei, individuando diverse statistiche descrittive (es. l'età media, l'andamento della età media nel tempo, il numero di professionisti per fascia di età e nazione, ecc.) ovvero correlazioni tra diverse variabili (es. quale sia la correlazione tra titoli di studio, nazione di provenienza e salario dichiarato).

E' questo unque il momento in cui occorre capire quale è il fenomeno (o variabile) in output al problema, e quali sono le tipologie di dati in input alla funzione che esprime la decisione, interpretazione o predizione (ad esempio la pressione, l'altitudine, il periodo dell'anno, ecc.)

4. Scelta e acquisizione dei dati

Nel nuovo mondo dei big data, le fonti di dati sono un numero infinitamente più grande che nel passato, e sono caratterizzate da più elevati volume, velocità, varietà. Questa situazione di grande disponibilità di dati, peraltro, richiede maggiore attenzione rispetto al mondo dei sistemi informativi tradizionali nei due passi di cui si compone questa attività. Vediamoli:

1. Scelta delle fonti, in cui dobbiamo selezionare tra le tante fonti disponibili quelle che fanno al caso nostro, tenendo conto della loro eventuale disponibilità come dati aperti, e dunque gratuiti, del loro livello di aggiornamento, del formato in cui sono resi disponibili, dell'eventuale costo derivante dal fatto che i dati non risultano essere già disponibili; in questo caso, ad esempio, in una partita di pallacanestro in cui siamo interessati ad analizzare il tipo di gioco dei singoli giocatori e della squadra nel suo complesso, occorre dispiegare una rete di sensori sul pallone e sulle scarpe e le braccia dei giocatori, per essere in grado di misurare la velocità con cui le due squadre fanno girare la palla, le traiettorie dei giocatori, ecc.

2. Acquisizione dei dati grezzi, in cui si acquisiscono e si memorizzano i dati scelti nel passo precedente. In questo caso, qualora si acquisiscano dati caratterizzati da un flusso continuo, occorre dimensionare la memoria e gli elaboratori disponibili per essere in grado di elaborare dati per un determinato periodo temporale al termine del quale i dati non servono più, e fare spazio a nuovi dati in streaming che vengono caricati in pipeline rispetto ai precedenti. Alcune di queste problematiche saranno esaminate nel Capitolo 4.

Le fonti individuate e i dati acquisiti nei due passi precedenti potrebbero non essere sufficienti a descrivere le variabili del problema; in questo caso dobbiamo individuare nuove fonti, ovvero un nuovo problema che approssima il precedente.

Nell'esempio di Breezometer, le fonti riguardano tutti i fenomeni che influenzano l'inquinamento, e perciò il traffico, il riscaldamento, l'orografia, il vento, la presenza di arre verdi, tutti fenomeni che vanno monitorati nel tempo. Occorre poi acquisire una rappresentazione del territorio su cui "mappare" tutte le precedenti fonti (vedi Figura 7 in cui diamo una rappresentazione figurativa delle fonti).

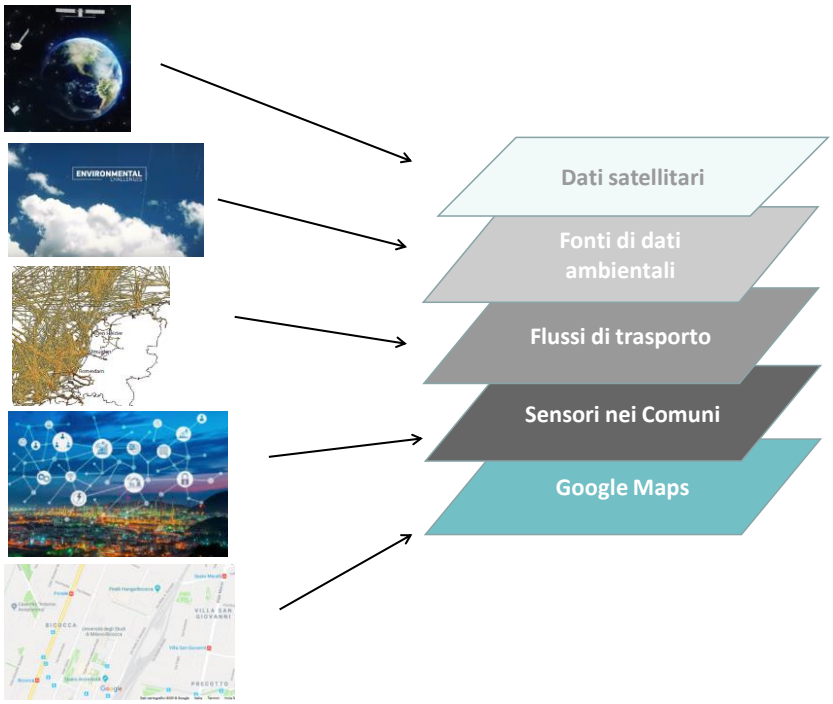


Figura 7 – Fonti di dati per Breezometer

Nell'esempio dei professionisti, supponiamo che la anagrafe esista già e venga messa a nostra disposizione per effettuare le analisi attraverso un trasferimento di file, ovvero su supporto magnetico. Il relativo file è stato mostrato in Figura 5, che rappresenta un frammento della anagrafe nel suo complesso.

5. Gestione

La fase di gestione è la fase in cui occorre governare i problemi connessi alla varietà ed eterogeneità delle rappresentazioni dei dati digitali. Nella fase di gestione (anche chiamata di preparazione), i dati sono analizzati per renderne più comprensibile il significato (passi di Modellazione, Trasformazione, Arricchimento semantico), per «metterli insieme» (passo di Integrazione), per migliorarne la accuratezza (passo di Valutazione e miglioramento della qualità), per ridurne la dimensione. Questa è usualmente la fase più dispendiosa del ciclo di vita, e può corrispondere fino all'80 per cento dello sforzo complessivo.

5.1 Modellazione

In questo passo si decide il modello con cui rappresentare i dati grezzi. Nel prossimo Capitolo 3 vedremo diversi modelli di rappresentazione, il più semplice è il modello relazionale, che rappresenta i dati con tabelle; a ogni colonna nella tabella viene assegnato un nome della colonna, detto attributo.

Per il momento non abbiamo ancora molti elementi per poter assegnare nomi a colonne (mentre il nome della tabella può essere, in inglese European professionals, o, al singolare, European professional), per cui l'output del passo di modellazione è la tabella di Figura 8, in cui abbiamo inserito

una prima colonna che dà la posizione della riga nella tabella e una prima riga che fornisce la posizione della colonna nella tabella. Non riportiamo qui e nel seguito il nome della tabella.

	1	2	3	4	5	6	7	8	9	10	11
Tuple #											
1	Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
2	Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
3	Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20…	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
10	Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Figura 8 – Output del passo di modellazione per la tabella *European professional*

	1	2	3	4	5	6	7	8	9	10	11
Tuple #			Year	Month	Day						
1	Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
2	Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
3	Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20…	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
10	Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Figura 9 – Output del passo di profilazione

5.2 Profilazione

In questo passo vengono prodotti dati statistici sulle caratteristiche dei dati, che forniscano elementi per ricostruirne il significato. Ad esempio, una analisi della terza colonna mostra che i valori sono tutti distribuiti tra 1949 e 1978, mentre una analisi della colonna successiva mostra che i valori numerici sono distribuiti tra 7 e 11 e i valori testuali sono relativi a mesi. Dall'esito di queste analisi possiamo assegnare i nomi alle colonne dalla terza alla quinta, rispettivamente Year, Month, e Day, vedi Figura 9 della pagina precedente. Si noti che per il momento possiamo soltanto dare nomi generici per i tre riferimenti temporali, non abbiamo elementi per specializzarli ulteriormente, come, ad esempio "Year of first employment" o altro.

5.3 Arricchimento semantico

In questo passo dobbiamo cercare di arricchire il significato dei dati. Nello studio di caso dei professionisti europei, l'esigenza è quella di completare i nomi degli attributi, aggiungendo eventualmente altre proprietà. Se ci concentriamo ad esempio sulla colonna 9, e confrontiamo i valori con quelli di domini (cioè insiemi di definizione di valori) di cui abbiamo una descrizione sul Web, possiamo facilmente scoprire che si tratta di Città europee. Osservando le colonne Year, Month, Day, scopriamo che le date sono compatibili con una interpretazione compatibile con le date di nascita delle persone descritte nella tabella; arriviamo perciò alla conclusione di sostituire i nomi generici delle tre colonne con nomi facenti riferimento alla data di nascita.

Tuple #	F.Name	Last Name	Y.of Birth	M. Of Birth	D. Birth	Toponym	Name of toponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
2	Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
3	Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20…	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
10	Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Figura 10 – Output del passo di arricchimento semantico

Più difficile è la interpretazione del significato delle altre colonne. Ad esempio, la colonna 11 ha valori che in parte troviamo nel dominio dei nomi in italiano delle nazioni europee, mentre altri valori come

"Slovensko" sono valori di nazioni, ma espressi in una lingua diversa dall'italiano. Poiché i vari domini compatibili con i valori nella colonna fanno tutti riferimento a Nazioni, possiamo arrivare alla conclusione che la colonna rappresenti Nazioni, chiamandola così Country. Vediamo anche qui che per il ciclo di vita del dato possiamo sfruttare risorse disponibili nel Web; avevamo già fatto un esempio di questo uso del Web nel Capitolo 1 a proposito del termine Roma. L'esito del passo di arricchimento semantico è mostrato in Figura 10 della pagina precedente.

5.4 Normalizzazione

Il passo di normalizzazione ha lo scopo di uniformare i valori presenti nella tabella rispetto a versioni descritte attraverso domini o grammatiche standardizzate e condivise. Ad esempio, la colonna Month of Birth è espressa da valori che sono stati riconosciuti essere relativi a mesi, ma che sono eterogenei tra loro, in quanto in parte numerici e in parte testuali. In questo caso possiamo convertire i valori testuali in valori numerici, ottenendo la nuova colonna di Figura 11. Possiamo anche sostituire i valori Slovensko e Slov nella colonna Country con il valore Slovakia.

Tuple #	F.Name	Last Name	Y.of Birth	M. Of Birth	D. Birth	Toponym	Name of toponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prage	412776	null
2	Martin	Necasky	1975	7	8	Square	Vienna	null	Bratislava	101278	**Slovakia**
3	Miroslav	Konecny	null	null	null	Street	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	**6**	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	10	Square	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	**9**	7	Street	Sesto	null	Mian	20…	Ital
7	Anita	Rula	1982	9	7	Street	Seso	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Street	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	Carla	Botni	1949	6	7	Avenue	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	**Slovakia**

Figura 11 – Esito del passo di normalizzazione

Possiamo poi operare una normalizzazione sulla colonna Name of Toponym, in cui ad esempio compaiono i valori

- St. John e
- Saint Jon

E' frequente che per fretta o per imprecisione quando si inseriscono valori di parole composte si abbrevino le parole facenti riferimento a nomi comuni frequenti come ad esempio

- Saint → St.

- Year of Birth → Y. Of Birth

Nel passo di normalizzazione si devono anzitutto scoprire pattern di parole appartenenti a queste forme lessicali, parole che vanno successivamente ricondotte alla forma lessicale standard. E' quello che abbiamo fatto con St. John nella prima riga della tabella di Figura 11. Si noti che non siamo intervenuti in questa fase sul termine Saint Jon perché, pur essendo visibilmente mancante di un carattere, risponde al formato standard dei nomi di toponimi.

5.5 Metadatazione

I metadati sono i dati sui dati; ad esempio i nomi degli attributi nella tabella di Figura 11 sono metadati, perché esprimono il significato dei dati nella colonna. Altri tipici metadati sono rappresentati in Figura 12, riferiti al caso Breezometer; in questo caso facciamo riferimento a dati di tipo remote sensing (vedi [Barsi 2019], inviati da aereo o satellite. Per alcuni metadati di una immagine vengono riportati il nome del metadato e il tipo dei valori. In particolare:

- la fonte è un dato fondamentale, perché permette di risalire al soggetto o tecnologia che ha inizialmente prodotto l'immagine, ad es. un aereo o un satellite.
- Il tempo di inizio e fine forniscono informazioni sull'intervallo di tempo in cui la immagine è stata prodotta e acquisita;
- le longitudini, insieme alle latitudini, esprimono le coordinate spaziali dei quattro vertici della immagine, che supponiamo sia inviata in formato rettangolare.

Metadato	**Tipo di dato**
Fonte	Stringa di caratteri
Tempo di inizio	Timestamp
Tempo di fine	Timestamp
Longitudine punto in alto a sinistra	Floating point
Longitudine punto in basso a sinistra	Floating point
Longitudine punto in alto a destra	Floating point
Longitudine punto in basso a destra	Floating point

Figura 12 – Metadati per il caso Breezometer

Esistono molti standard per la definizione di metadati, si veda ad esempio lo standard Dublin core, www.dublincore.org.

5.6 Trasformazione di modello

In questo passo si opera sui dati per trasformare il modello con cui sono rappresentati. L'esigenza connessa alla trasformazione di modello può derivare dal fatto che l'attuale modello non è più adatto a rappresentare i dati, in virtù del loro progressivo aumento, oppure che si vuole adottare un modello che descrive la semantica dei dati in modo più ricco, oppure ancora per uniformare il modello di un insieme di dataset inizialmente rappresentati con modelli diversi. In Figura 13 è mostrato il primo caso; inizialmente abbiamo un grafo rappresentato per mezzo di un insieme di nodi e archi; successivamente, man mano che il grafo cresce, è modellato inizialmente con una matrice di adiacenza, in cui i nodi

corrispondono alle righe e colonne della matrice e le celle rappresentano le coppie di nodi connessi, ed è modellato infine con una rappresentazione ibrida tra le due precedenti, che riprenderemo nel Capitolo 12 sulle astrazioni.

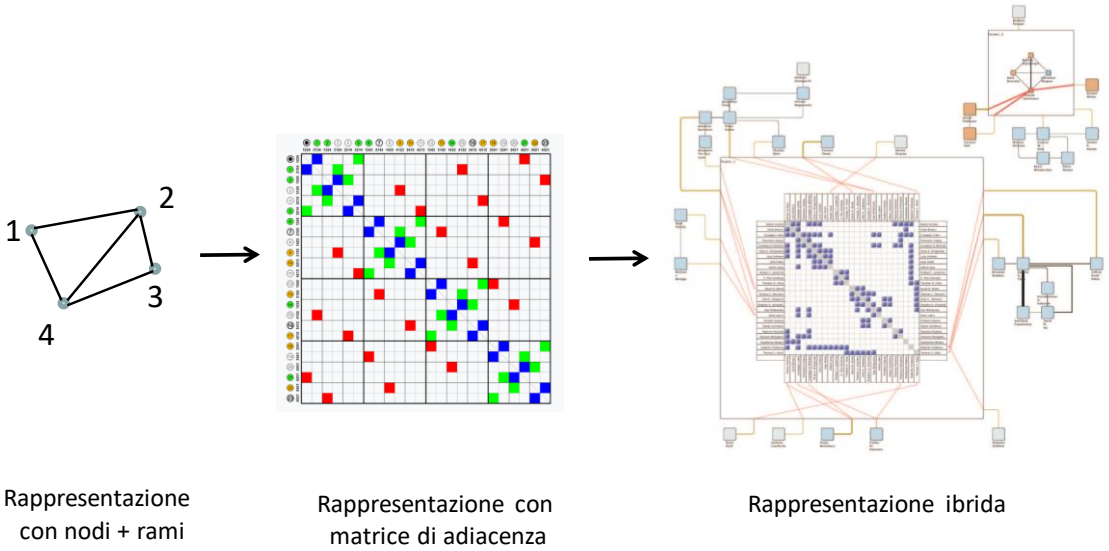


Figura 13 – Tre diversi modelli di rappresentazione di un grafo che cresce nel tempo (in parte tratta da V. Batagelij - Visual analysis of large graphs using (x, y)-clustering and hybrid visualizations, 2010)

5.7 Controllo di qualità

I dati che vengono rilevati nel ciclo di vita sono spesso soggetti ad errori o approssimazioni. Ad esempio, se vengono rilevati dati da un sensore di temperatura, è possibile che per qualche ragione la misurazione sia imprecisa, o errata. E' anche possibile che nel caso i dati siano rilevati senza soluzione di continuità, il sensore si guasti, e quindi i dati siano non definiti per un intervallo temporale che dura fin quando il sensore non sia riparato. Il tema della qualità dei dati è molto vasto, come vedremo nel Capitolo 9. Qui ci focalizziamo su due proprietà dei dati nelle tabelle, la accuratezza e la completezza.

Consideriamo la prima riga della tabella, e in particolare il nome della città, "Prage". Questo nome non compare tra le città europee, e quindi possiamo fare la assunzione che sia errato. Per correggerlo, possiamo confrontare "Prage" con i nomi delle città europee, e cercare quella con il nome più vicino, che è "Prague". Siccome disponiamo dello Zip code, ma non del Country, possiamo verificare per riscontro se effettivamente quello zip code sia associabile alla città di Praga. Dopo queste verifiche possiamo sostituire "Prage" con "Prague"; possiamo inoltre confrontare "Saint Jon" con le strade nello stradario di Praga, cercando quella più vicina come stringa di caratteri, che assumiamo sia "Saint John", ottenendo infine la nuova tabella di Figura 14.

Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	**Prague**	412776	null
2	Martin	Necasky	1975	7	8	Square	Vienna	null	Bratislava	101278	Slovakia
3	Miroslav	Konecny	null	null	null	Street	Saint John	49	Prague	412776	null
4	Crlo	Btini	1949	6	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	10	Square	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	9	7	Street	Sesto	null	Mian	20…	Ital
7	Anita	Rula	1982	9	7	Street	Seso	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Street	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	Carla	Botni	1949	6	7	Avenue	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovakia

Figura 14 - Output del passo di verifica di accuratezza

Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czeck
2	Martin	Necasky	1975	7	8	Square	Vienna	**null**	Bratislava	101278	Slovackia
3	Miroslav	Konecny	**null**	**null**	**null**	Street	Saint John	49	Prague	412776	Czeck
4	Carlo	Btini	1949	6	7	Street	Dessiè	15	Roma	00198	Italy
5	Miroslav	Knecy	1978	7	10	Square	Budapest	23	Wien	k2345	Austria
6	Anisa	Rula	1982	9	7	Street	Sesto	**null**	Milano	20127	Italy
7	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	Anna	Rla	**null**	**null**	**null**	Street	Sarca	336	Milano	20126	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	Carla	Botni	1949	6	7	Avenue	Charles	**null**	Prague	412733	Czeck
11	Marta	Necasky	1976	11	8	**null**	**null**	**null**	Bratislava	112234	Slovakia

Figura 15 - Output del passo di verifica di completezza

La seconda caratteristica di qualità dei dati che prendiamo in considerazione è la completezza, e fa riferimento alla presenza di valori nulli, che corrispondono ad assenza di informazione. E' buona norma cercare di sostituire i valori nulli con valori espliciti e noti. Nel nostro caso, con riferimento alla prima e terza riga, possiamo sostituire i due valori null con il nome del paese in cui è localizzata Prague, da cui

come risultato la tabella di Figura 15. Non abbiamo altre forme di conoscenza che ci permettano di sostituire altri valori nulli con valori espliciti.

A conclusione possiamo affermare che migliorare la qualità dei dati è fondamentale, in questo modo il problema che vogliamo risolvere sarà anche esso caratterizzato da maggiore qualità. Si veda anche [Barsi 2019] per una discussione sulle qualità dei dati acquisiti da remote sensing nel caso Breezometer.

5.8 Integrazione

La realtà che noi rileviamo acquisendo fonti di dati distinte, spesso da fornitori di dati diversi, è frammentata; per analizzare un problema, noi auspicabilmente dobbiamo disporre di dati integrati, che omogeneamente facciano riferimento al fenomeno da analizzare. Per ottenere ciò, occorre effettuare un passo di integrazione, il cui scopo è mettere insieme dati da diverse fonti, caratterizzate spesso da eterogeneità e differenze, ottenendo un nuovo insieme omogeneo di dati in cui tali eterogeneità sono risolte e le diverse rappresentazioni dei dati sono riconciliate. Ad esempio, la data di nascita potrebbe essere rappresentata con tre attributi <giorno, mese, anno> in una fonte e un unico attributo <data> in un'altra fonte. Oppure, gli impiegati di una organizzazione potrebbero essere identificati in una fonte da un numero progressivo, e in un'altra dal codice fiscale.

L'attività di integrazione si diversifica in due casi:
1. Quello in cui noi vogliamo integrare due o più fonti; è il passo di integrazione vero e proprio
2. Quello in cui la fonte sia unica, ma si ha il fondato sospetto che vi siano diversi record nella fonte che rappresentano la stessa persona, o luogo, o evento, a seconda del significato della fonte. Questo passo è chiamato di deduplicazione.

In entrambi i precedenti casi, l'attività di integrazione è completata solo quando tutti i gruppi di dati che fanno riferimento ad una stessa persona, luogo, evento, ecc. sono fusi in un'unica versione (si parla di unica versione della verità); ciò avviene con un passo chiamato di fusione. Vediamo ora i tre passi di integrazione, deduplicazione e fusione esemplificandoli nei nostri due studi di caso.

<u>Passo di integrazione e fusione in Breezometer</u>

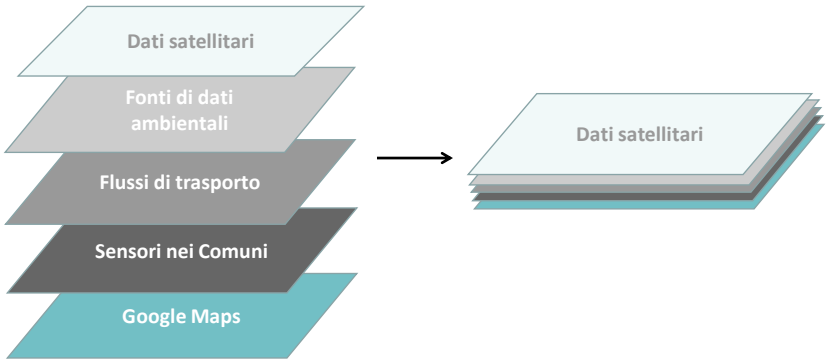


Figura 16 – Integrazione di diversi tipi di rappresentazioni del territorio

Nel caso Breezometer le diverse fonti individuate in precedenza (vedi Figura 16) fanno tutte riferimento al territorio, ma potrebbero avere diverse scale e potrebbero essere in alcuni casi georeferenziate, cioè riferite ad un sistema di coordinate di riferimento, in altri no. La letteratura scientifica ha affrontato questo problema, chiamato nei sistemi informativi geografici con il termine di *coalescenza*. Vediamo in Figura 17 un esempio di coalescenza di due mappe, una georeferenziata e l'altra no. Per procedere alla integrazione e fusione, è necessario dapprima trovare un insieme di punti caratterizzati dalle stesse coordinate geografiche nelle due mappe, e poi, tramite algoritmi di matching, unificare le due mappe rispetto allo stesso sistema di riferimento (questo, nella nostra terminologia, è il passo di fusione).

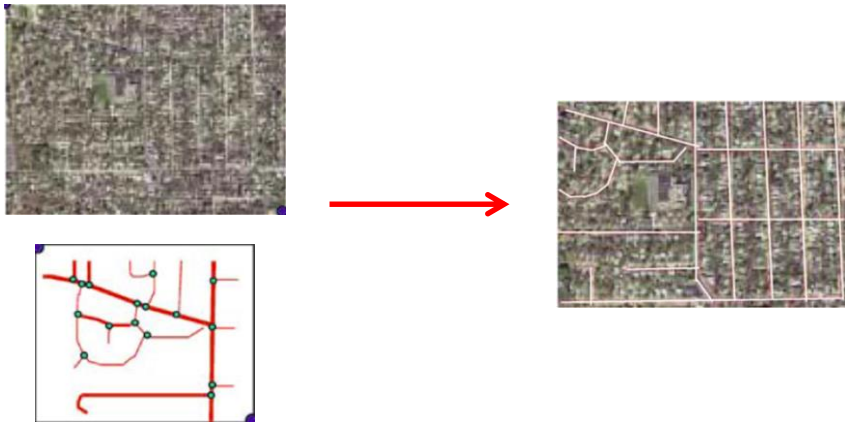


Figura 17 – Integrazione di diversi tipi di rappresentazioni del territorio (tratta da Chen C.C. et al. – Automatically and efficiently matching road networks with spatial attributes in unknown geometry systems, 2006)

<u>Passo di deduplicazione nella tabella dei professionisti europei</u>

Tuple #	Candidate tuples	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
5	**M1**	Miroslav	Knecy	1978	7	null	Square	Budapest	23	Wien	k2345	Austria
1	**M1**	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
3	**M1**	Miroslav	Konecny	null	null	null	Street	Saint John	49	Prague	412776	Czech
10	**NM**	Carla	Botni	1949	6	7	Avenue	Charles	null	Prague	412733	Czeck
4	**M3**	Carlo	Btini	1949	6	7	Street	Dessiè	15	Roma	00199	Italy
6	**M2**	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	20127	Italy
7	**M2**	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	**NM**	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	**M3**	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
2	**NM**	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovakia
11	**NM**	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovakia

Figura 18 – Output del passo di deduplicazione della tabella

Assumendo che i dati siano organizzati in tabelle, come nel caso dei professionisti europei, nel passo di deduplicazione si individuano le differenti righe della tabella che fanno riferimento agli stessi oggetti del mondo reale. La identificazione può farsi in modo approssimato confrontando a due a due le righe, e definendo una funzione di distanza tra i valori alfanumerici dei diversi attributi. Per esempio, in Figura 15, considerando le righe 1 e 5 vediamo che nei campi definiti le due righe differiscono nel cognome e nell'indirizzo. Se consideriamo come rilevanti nel confronto nome, cognome e data di nascita, e trascuriamo l'indirizzo, che potrebbe essere cambiato nel tempo, allora le due righe differiscono solo di due caratteri nel cognome, e perciò possono essere dichiarate come matching, cioè fare riferimento alla stessa persona. Vedi Figura 18.

Altre righe invece possono essere "lontane" da tutte le altre, in questo caso le dichiareremo non matching con nessun'altra, ad indicare che sono le uniche a rappresentare una singola persona.

Passo di fusione nella tabella dei professionisti europei

Resta un ultimo passo per concludere il processo di deduplicazione nella tabella dei professionisti europei, in cui partendo dai diversi gruppi di righe matching, scegliamo per i diversi valori degli attributi i valori rappresentativi nella riga risultato della fusione.

Tuple #	Candidate tuples	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name Toponym	Number T.	City	Zip Code	Country
1	**M1**	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
10	**NM**	Carla	Botni	1949	6	7	Avenue	Charles	null	Prague	412733	Czeck
7	**M2**	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	**NM**	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	**M3**	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
2	**NM**	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovakia
11	**NM**	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovakia

Figura 19 – Output del passo di fusione delle righe matching nella tabella

Per effettuare la fusione tra valori, possiamo seguire diverse strategie; ad esempio per le righe 1 e 5 già considerate in precedenza, possiamo scegliere come valore del cognome quello tra i due più frequente tra i cittadini della Repubblica Ceca. Il risultato finale della attività di fusione è mostrato in Figura 19.

5.9 Implementazione della architettura tecnologica

Il passo di implementazione della architettura tecnologica ha lo scopo di scegliere la architettura della memoria e dei processori di calcolo che meglio si adatta alla natura del problema e alle caratteristiche di volume dei dati disponibili. Diverse architetture sono disponibili per memoria e processori, la cui caratteristica comune è di permettere l'accesso in memoria e le elaborazioni eseguibili in parallelo su

diverse partizioni dei dati, per ottimizzare il tempo di esecuzione e la produzione dei dati intermedi del calcolo. Nel Capitolo 4 discuteremo approfonditamente le architetture tecnologiche per big data.

6. Analisi dei dati

L'analisi dei dati ha lo scopo, terminate le attività legate alla gestione, di costruire il modello risolutivo del problema posto nella fase iniziale di formulazione. Il modello dove dirci in qual modo le variabili di input sono legate alla variabile di output. Nelle analisi statistiche classiche, i dati disponibili sono in genere pochi, e quindi spesso non rappresentativi del fenomeno. E' inoltre troppo costoso o impossibile acquisire ed elaborare dati che riguardano l'intero universo sotto osservazione, e ci si limita ad analizzare un campione. In questi casi è necessario valutare la adeguatezza del campione a inferire, a partire dal modello individuato, l'applicabilità del modello all' intero universo osservato.

Oggi in molti fenomeni, avendo a disposizione le infrastrutture di calcolo parallele di recente generazione, è possibile analizzare l'intero universo sotto osservazione. Inoltre, il progressivo aumento della dimensione dei dataset permette di estrarre e analizzare molte più proprietà del fenomeno sotto osservazione; ad esempio, l'analisi per sondare gli umori dei mercati e per effettuare profilazioni di utenti, può essere arricchita acquisendo e analizzando, accanto a dati di natura anagrafica e professionale, anche dati georeferenziati, dati provenienti dalle interazioni con i siti o ottenuti attraverso canali social (effettuando la cosiddetta sentiment analysis). Corrispondentemente, sono evoluti i metodi statistici e, successivamente, sono stati sviluppati metodi basati sul machine learning.

6.1 Metodi statistici

I metodi statistici per trovare la relazione che lega una variabile in output con un insieme di variabili in input, sono basati sui concetti di correlazione e regressione. In statistica, una correlazione [Wikipedia 2018] è una relazione tra due variabili tale che a ciascun valore della prima corrisponda, con una certa regolarità, un valore della seconda. Contrariamente a quanto si potrebbe inferire, la correlazione non esprime una relazione di causa effetto tra le variabili, quanto un legame più debole, che caratterizza la tendenza di una variabile a cambiare valore in relazione ad un'altra. Ad esempio, nel diagramma a sinistra di Figura 19 (tratto da [Pearl 2018]) si evidenzia il risultato di una indagine sul consumo congiunto di dentifricio e di filo interdentale; possiamo dire che esiste un'alta correlazione tra i due consumi, ma non abbiamo nessun elemento per esprimere un rapporto di causa effetto tra essi.

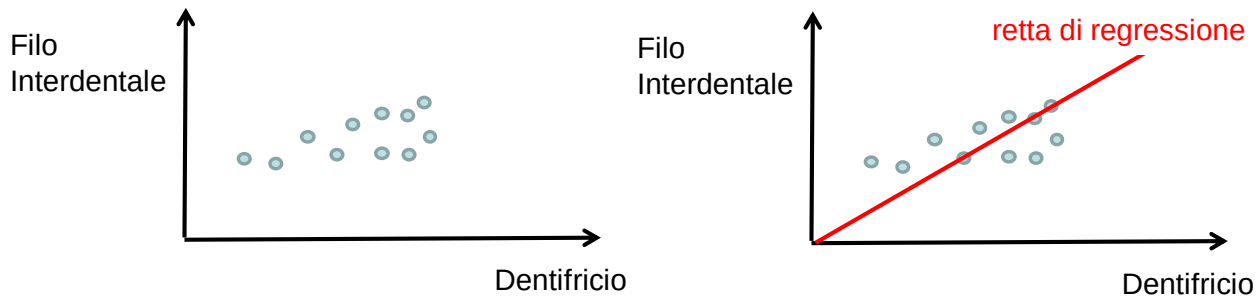


Figura 20 – Esempi di correlazione e regressione

Quando si voglia esprimere un legame funzionale tra due variabili, come nel caso di Figura 20, si usano tecniche cosiddette di regressione per individuare la funzione che meglio approssima i dati correlati (si veda il diagramma a destra di Figura 20).

6.2 Metodi basati sul machine learning

Il Machine Learning si può definire come un processo di creazione di un algoritmo o programma che consente di svolgere un compito sulla base di informazioni che non forniscono un'esplicita descrizione di quel programma. Al contrario di quanto accade usualmente nella progettazione del software, in cui esseri umani codificano il programma sulla base di specifiche testuali, è la tecnica di machine learning che apprende dai dati disponibili, auto-generando il programma che fornisce il modello risolutivo del problema. Ma cosa si intende per *apprende*? Ciò che viene chiamato apprendimento, può assumere diverse forme, come si argomenta nel seguito.

- *Apprendimento supervisionato*: all'algoritmo di apprendimento sono forniti, come esperienza, dati di input e rispettivi dati di output ad essi collegati tramite una funzione che associa a ciascun input il corrispondente output; in altre parole, per quei dati, la funzione tra dati di input e dati di output è nota. Vedi un esempio in Figura 21, in cui si mostra la variabile Qualità dell'aria espressa come funzione di diverse variabili in input, quali le coordinate spaziali, il tempo, il traffico e l'inquinamento; il processo che porta l'algoritmo a costruire un modello predittivo è chiamato *processo di classificazione*. Tipicamente, come esempi di algoritmi di apprendimento supervisionato vi sono quelli che si basano sugli Alberi di decisione e le Reti Neurali. Una volta che l'algoritmo ha appreso la funzione a partire dai dati (*training set*), segue una fase di test per valutare la sua capacità di generalizzazione, la capacità, cioè, di fornire output corretti a partire da input non contenuti nel training set.

- *Apprendimento non supervisionato*: all'algoritmo di apprendimento sono forniti, come esperienza, dati di input senza che questi siano associati a valori di output, ed esso ha il compito di riconoscere schemi/strutture nell'insieme dei dati fornito. Rientrano nella classe degli algoritmi di apprendimento non supervisionato quelli di clustering, ossia quegli algoritmi che hanno il compito di raggruppare i dati sulla base di strutture, schemi o affinità che essi devono riconoscere.

- *Apprendimento per rinforzo*: l'algoritmo di apprendimento apprende sulla base della risposta dell'ambiente alle sue azioni; in tal caso, infatti, vi è un'interazione con un ambiente nel quale esso deve svolgere un compito e dal quale ottiene una risposta in termini di "ricompensa" (il rinforzo) che consiste nella valutazione della prestazione. Come esempi, citiamo gli algoritmi che imparano le strategie da usare in un gioco attraverso partite contro un avversario e quelli che consentono ad un robot di muoversi all'interno di un ambiente con un certo fine.

In Breezometer, ad esempio, sono usate tecniche di Machine Learning per predire l'andamento nel tempo del livello di inquinamento in una data area, sulla base di dati come quelli di Figura 21.

latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscalda-mento	Qualità dell'aria
308.40	215.20	2017	6	12	8	10	0.7	0.7	47
308.40	215.20	2017	6	12	8	20	0.7	0.8	49
308.40	215.20	2017	6	12	8	30	0.8	0.8	51
308.40	215.20	2017	6	12	8	40	0.9	0.7	54
..	..	..	..	..	..	.			..
308.45	215.25	2017	6	12	8	10	0.7	0.7	54
308.45	215.25	2017	6	12	8	20	0.7	0.7	57
308.45	215.25	2017	6	12	8	30	0.8	0.8	60
308.45	215.25	2017	6	12	8	40	0.9	0.7	65
308.45	215.25	2017	6	12	8	50	0.8	0.6	70

Figura 21 – La qualità dell'aria come funzione di caratteristiche misurate nei dataset disponibili

7. Visualizzazione

Nella fase di visualizzazione occorre decidere come rappresentare dati in output in modo tale da risultare i più comprensibili possibile per l'utente finale. Lo scopo della visualizzazione è far percepire il contenuto di un insieme di dati mediante rappresentazioni basate sul senso della visione. Le visualizzazioni possono essere utili anche prima della analisi, in cui può essere opportuno osservare una rappresentazione visuale dei dati per capire quali attività di preparazione vanno intraprese, oppure quali modelli di analisi vanno considerati.

Un tipo di visualizzazione che è sempre più usata sono le mappe di calore, o heat map. Le mappe di calore rappresentano l'intensità dei valori assunti dai dati all'interno di diverse aree di una superficie. Con l'aumento dei valori dei dati, la mappa di calore visualizza i valori mediante un indicatore di colore passando in genere da colori deboli, come un azzurro o verde leggero, a colori forti corrispondenti ad un rosso scuro, in questo libro da un grigio chiaro a un grigio scuro.

In Figura 22 mostriamo un esempio di mappa di calore prodotta da Breezometer, il grigio scuso corrisponde a maggiore inquinamento. E' evidente la maggiore espressività della heat map rispetto al freddo insieme di dati nella Figura 21, che conferma il detto "una figura è meglio di mille parole" rifrasato questa volta nel detto "una figura è meglio di mille dati".

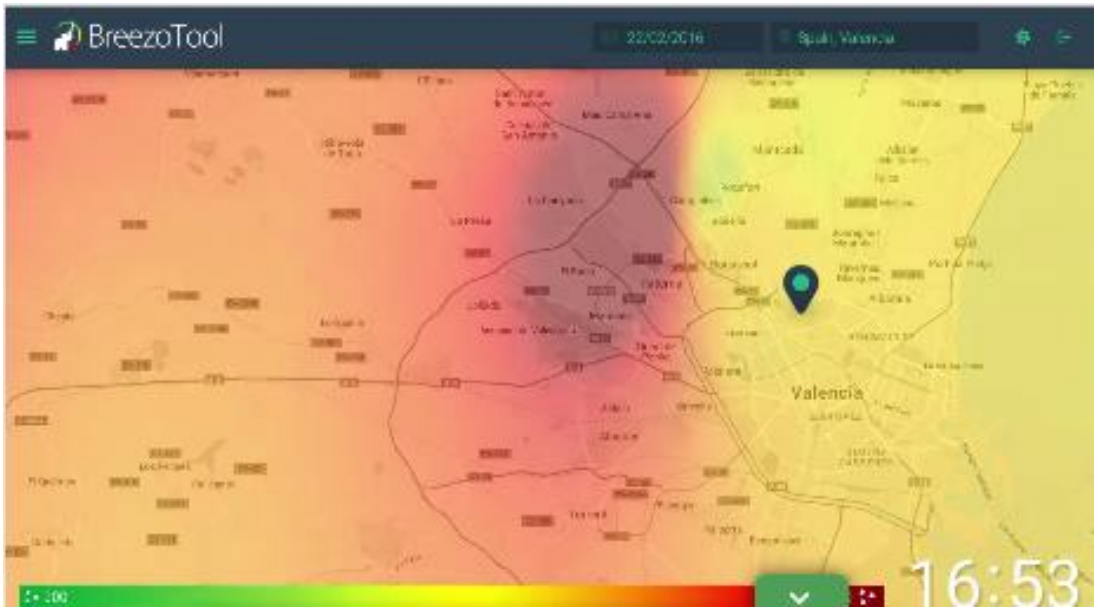


Figura 22 – Visualizzazione mediante heat map adottata da Breezometer
(tratta dal sito www.breezometer.com, 2017)

Come ulteriore esempio di visualizzazione, in Figura 23 è mostrato un grafo nella sua evoluzione nel tempo, attraverso due diverse visualizzazioni che corrispondono a diverse rappresentazioni sul piano (o layout) dei nodi e degli archi. In entrambe le visualizzazioni, il layout (cioè la disposizione sul piano del disegno) del grafo è basato sul clustering dei nodi, l'operazione di aggregazione di nodi in un unico nodo rappresentante.

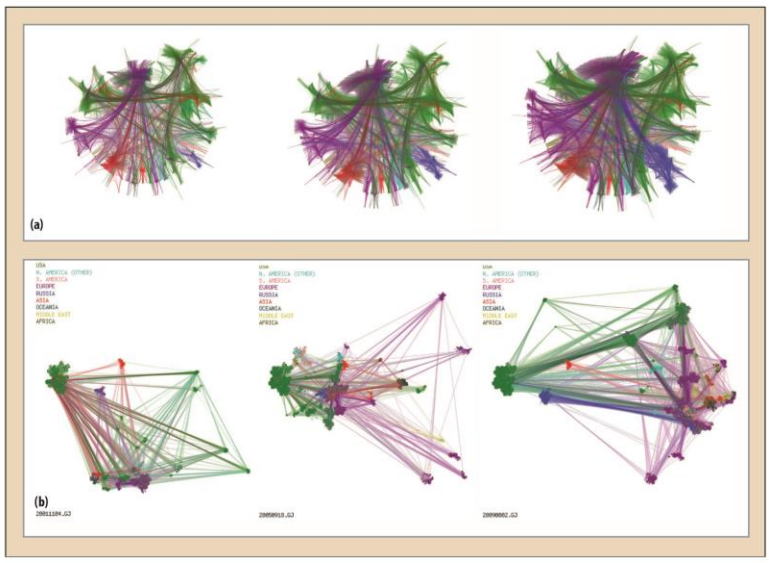


Figura 23 - Due diverse rappresentazioni dello stesso grafo che si modifica dinamicamente nel tempo
(tratta da Kwan Lu et al. Large-Scale Graph Visualization and Analytics, 2013

Nella parte superiore della figura il metodo di layout produce un primo clustering, quindi modifica in modo incrementale il clustering nel tempo, aggiornandosi appunto nel tempo. La visualizzazione risultante inizia con la generazione di un layout ideale (visualizzazione a sinistra), che nel tempo, per garantire la stabilità del layout, degrada la qualità, intesa come leggibilità del grafo.

Un metodo di layout basato su clustering globale (vedi parte inferiore della figura) utilizza l'intero intervallo di tempo nella scelta del layout, e mira a soddisfare in modo ottimale la qualità del layout per ogni passo temporale e la stabilità del layout nel tempo.

Il precedente esempio dimostra che vi sono in generale diverse visualizzazioni che possono essere associate a un insieme di dati, anche all'interno di un fissato insieme di regole di visualizzazione.

Riferimenti

Inseriamo qui e nel seguito anche lavori di ricerca ulteriori rispetto a quelli citati nel testo, per fornire al lettore possibili approfondimenti sugli argomenti trattati nel capitolo.

Á. Barsi, Z. Kuglera, C. Batini et al. Remote sensing data quality model: from data sources to lifecycle phases, Intl. Journal of Image and Data Fusion, 2019.

IBM - Wrangling big data: Fundamentals of data lifecycle management How to maintain data integrity across production and archived data, 2013

Y. Demchenko, C.de Laat; P. Membrey - Defining architecture components of the Big Data Ecosystem, 2014

Happiest Minds – Big Data, Creating Real Value form the data life cycle, 2014.

Bloomberg - 7 phases of a data life cycle, Information Management July 14, 2015

K. Cagle, Understanding the Big Data Life-Cycle, 2015.

L. Pouchard - Revisiting the Data Lifecycle with Big Data Curation – International Journal of Data Curation, 2015

Fujitsu – The white book of big data, 2016

J.L. Kourik et al. - The Intersection of Big Data and the Data Life Cycle: Impact on Data Management Wang International Journal of Knowledge Engineering, Vol. 3, No. 2, December 2017

J. Bladh e K. Hertzmann - Big Data and Product Lifecycle Management Case Studies from the Automotive Industry, 2017

Texas A&M Transportation Institute - Data Management Life Cycle -, 2018

G. Nelson - The Analytics life cycle toolkit, 2018

M. Arass - Data Lifecycle: From Big Data to Smart Data, 2018 IEEE.

M. Kaufmann - Big Data Management Canvas: A Reference Model for Value Creation from Data Big Data and Cognitive Computing

Gartner - Market Guide for Data Preparation Tools, 2019

Capitolo 3 - Come rappresentare i dati: i modelli

Carlo Batini

1. Introduzione

Siete mai stati a Portofino? Supponiamo di essere a Portofino, magari sul promontorio che domina il porto e di voler rappresentare in qualche modo delle informazioni associate al piccolo paese di Portofino. In Figura 1 mostriamo quattro modi diversi di rappresentare "informazioni su Portofino": mediante una tabella, una foto, una mappa e un messaggio consistente in un testo. Certo, la foto è più espressiva delle altre rappresentazioni, e solo il testo riesce a trasmettere emozioni.

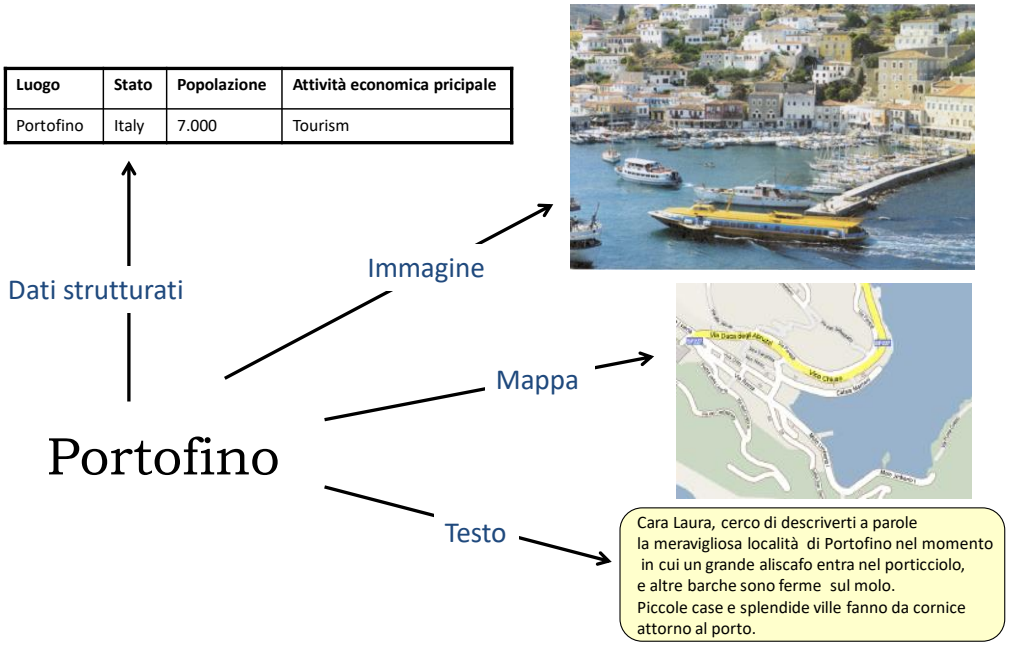


Figura 1 – In quanti modi possiamo descrivere Portofino?

In Figura 2 rappresentiamo ciò che accade quando vogliamo rappresentare un panorama di montagna, ripreso al tramonto; utilizziamo strutture linguistiche ovvero strutture basate sui sensi per tradurre le nostre sensazioni in un testo o una immagine. L'aspetto comune alle sei rappresentazioni di Figura 1 e Figura 2 sta nel fatto che noi in tutti questi casi procediamo ad una attività che chiameremo di modellazione, e che può dar luogo a tante rappresentazioni diverse. Se, ad esempio, la rappresentazione è una immagine prodotta con una fotocamera digitale, il numero di pixel a disposizione influenza la precisione con cui la immagine è resa. Questo capitolo è dedicato a investigare alcuni dei modelli che noi possiamo utilizzare nel rappresentare un frammento di realtà per mezzo di dati digitali.

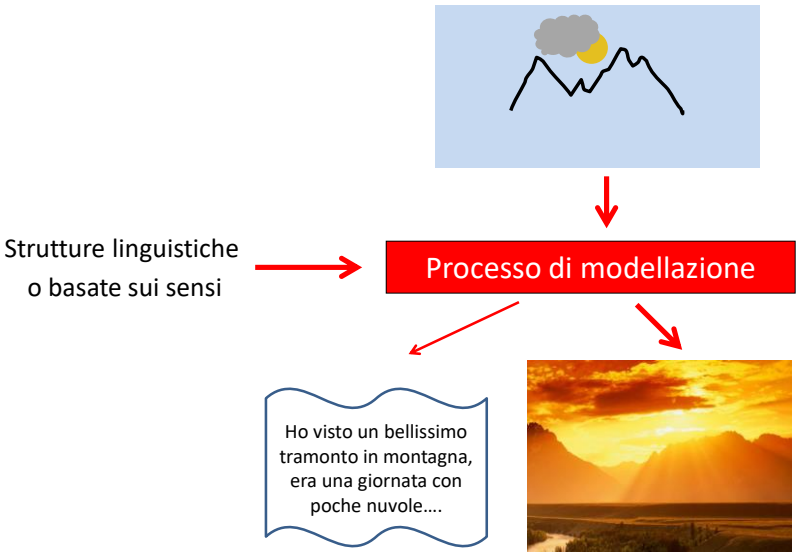
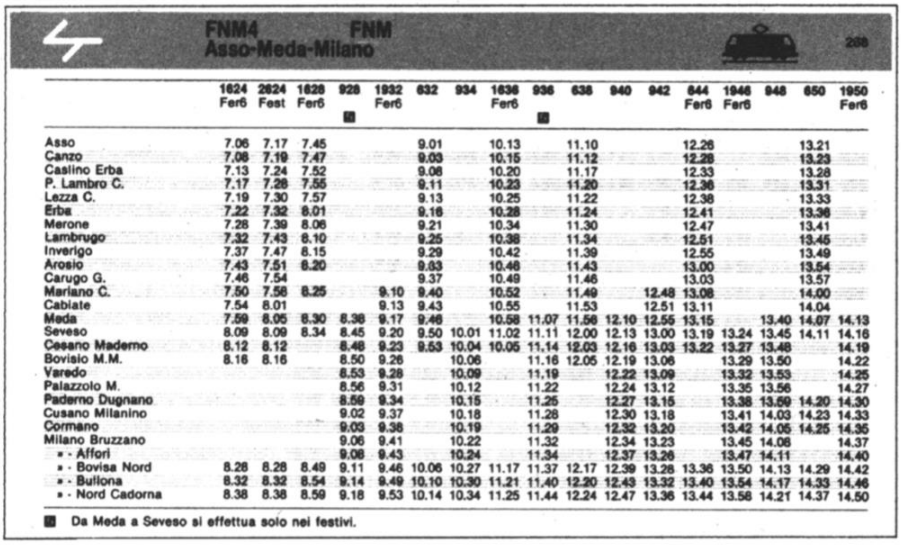


Figura 2 – Il processo di modellazione

All'inizio del libro "Le basi della informatica" di trentacinque anni fa immaginavo di organizzare un viaggio dalle parti di Milano (dove effettivamente sarei approdato diciotto anni dopo…). Questo viaggio è composto di diverse tratte, che utilizzano diversi mezzi di trasporto. Una delle tappe del viaggio avviene in treno, e quindi mostro nel libro una pagina di un orario ferroviario cartaceo (vedi Figura 3).



FNM4 FNM
Asso-Meda-Milano

	1624 Fer6	2624 Fest	1628 Fer6	928 ■	1932 Fer6	632	934	1636 Fer6	936 ■	638	940	942	844 Fer6	1946 Fer6	948	650	1950 Fer6
Asso	7.06	7.17	7.45			9.01		10.13		11.10			12.26			13.21	
Canzo	7.08	7.19	7.47			9.03		10.15		11.12			12.28			13.23	
Caslino Erba	7.13	7.24	7.52			9.08		10.20		11.17			12.33			13.28	
P. Lambro C.	7.17	7.28	7.56			9.11		10.23		11.20			12.36			13.31	
Lezza C.	7.19	7.30	7.57			9.13		10.25		11.22			12.38			13.33	
Erba	7.22	7.32	8.01			9.16		10.28		11.24			12.41			13.36	
Merone	7.28	7.39	8.06			9.21		10.34		11.30			12.47			13.41	
Lambrugo	7.32	7.43	8.10			9.25		10.38		11.34			12.51			13.45	
Inverigo	7.37	7.47	8.15			9.29		10.42		11.39			12.55			13.49	
Arosio	7.43	7.51	8.20			9.33		10.48		11.43			13.00			13.54	
Carugo G.	7.46	7.54				9.37		10.49		11.46			13.03			13.57	
Mariano C.	7.50	7.58	8.25		9.10	9.40		10.52		11.49		12.48	13.08			14.00	
Cabiate	7.54	8.01			9.13	9.43		10.55		11.53		12.51	13.11			14.04	
Meda	7.59	8.05	8.30	8.38	9.17	9.46		10.58	11.07	11.56	12.10	12.55	13.15		13.40	14.07	14.13
Seveso	8.09	8.09	8.34	8.45	9.20	9.50	10.01	11.02	11.11	12.00	12.13	13.00	13.19	13.24	13.45	14.11	14.16
Cesano Maderno	8.12	8.12		8.48	9.23	9.53	10.04	10.05	11.14	12.03	12.16	13.03	13.22	13.27	13.48		14.19
Bovisio M.M.	8.16	8.16		8.50	9.26		10.06		11.16	12.05	12.19	13.06		13.29	13.50		14.22
Varedo				8.53	9.28		10.09		11.19		12.22	13.09		13.32	13.53		14.25
Palazzolo M.				8.56	9.31		10.12		11.22		12.24	13.12		13.35	13.56		14.27
Paderno Dugnano				8.59	9.34		10.15		11.25		12.27	13.15		13.38	13.59	14.20	14.30
Cusano Milanino				9.02	9.37		10.18		11.28		12.30	13.18		13.41	14.03	14.23	14.33
Cormano				9.03	9.38		10.19		11.29		12.32	13.20		13.42	14.05	14.25	14.35
Milano Bruzzano				9.06	9.41		10.22		11.32		12.34	13.23		13.45	14.08		14.37
» - Affori				9.08	9.43		10.24		11.34		12.37	13.26		13.47	14.11		14.40
» - Bovisa Nord	8.26	8.26	8.49	9.11	9.46	10.06	10.27	11.17	11.37	12.17	12.39	13.28	13.36	13.50	14.13	14.29	14.42
» - Bullona	8.32	8.32	8.54	9.14	9.49	10.10	10.30	11.21	11.40	12.20	12.43	13.32	13.40	13.54	14.17	14.33	14.46
» - Nord Cadorna	8.38	8.38	8.59	9.18	9.53	10.14	10.34	11.25	11.44	12.24	12.47	13.36	13.44	13.58	14.21	14.37	14.50

■ Da Meda a Seveso si effettua solo nei festivi.

Figura 3 - Una pagina di un orario ferroviario

Nell'orario sono mostrate le corse dei treni che percorrono la linea Asso Milano, terminando la corsa alla stazione di Milano Cadorna. Sulle righe sono mostrate tutte le stazioni in cui il treno ferma, sulle colonne sono mostrati tutti i treni che percorrono la linea nella prima parte della giornata. Se ho deciso in quale ora partire da Asso, devo scorrere le varie colonne fino a trovare un treno che parta circa a

quell'ora, e se a questo punto voglio sapere l'orario di arrivo, non resta che scorrere gli orari fino a trovare quello di Milano Cadorna. I dati sono rappresentati nell'orario in modo simile alle tre tabelle che abbiamo visto nel Capitolo 1, per righe e per colonne (quando troviamo uno spazio vuoto, significa che il treno non ferma nella stazione associata alla riga), con una differenza importante.

Le righe e le colonne di Figura 3 hanno entrambe un significato, sulle righe sono rappresentate le stazioni, sulle colonne i treni. Nelle tabelle dei Capitolo 1 il significato dei dati è espresso nella prima riga, tutte le righe successive hanno nei vari campi lo stesso significato, espresso nelle colonne. La struttura di Figura 3 è chiamata matrice righe per colonne.

Come già notavo nel prologo, nel passato gli itinerari e orari dei viaggi venivano rappresentati mediante orari grafici, ne riproduco nuovamente una pagina nella Figura 4. Quali differenze ci sono tra le struttura di dati di Figura 3 e Figura 4? Provate a pensarci su.

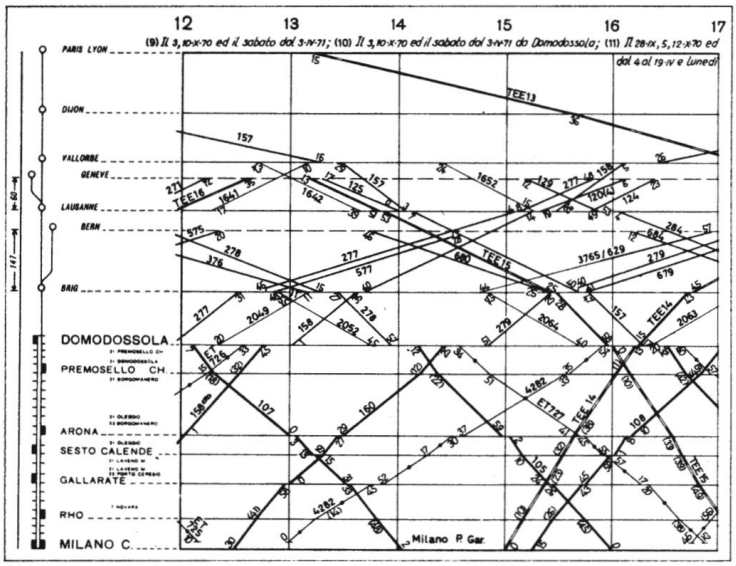


Figura 4 - Una pagina di un orario ferroviario grafico

Risposta – In entrambe le figure le righe rappresentano le stazioni; per quanto riguarda le colonne, in Figura 3 esse descrivono gli orari di arrivo (e di partenza) *in modo discreto* (ore 7.25, 7.48, ecc.), mentre in Figura 4 le colonne rappresentano *in modo continuo* le ore del giorno. Nel disegno di Figura 4 è rappresentato il movimento dei treni, oltre che il loro numero identificativo. Intersecando i segmenti corrispondenti ai viaggi dei treni con le righe corrispondenti alle stazioni siamo in grado di determinare l'ora e il minuto di partenza, che peraltro è rappresentato anche da un numero vicino al segmento. Spesso sulla riga vi sono due numeri, corrispondenti a ora di arrivo e di partenza.

La matrice di Figura 3 e il grafico di Figura 4 sono esempi di *dati strutturati*, dati cioè che sono organizzati mediante strutture; nel caso della Figura 3 la struttura è formata dalle righe e dalle colonne, a cui è associato un significato, per le righe le stazioni e per le colonne i treni, così che anche i valori numerici che compaiono nelle celle della matrice hanno un significato, corrispondente agli orari di arrivo e partenza dei treni nelle stazioni. Osservando ancora le *strutture di dati* di Figura 3 e 4 scopriamo che la

stessa informazione può essere descritta con due strutture di dati che hanno rappresentazioni diverse. Chiameremo nel seguito tali rappresentazioni con il termine di *modelli*.

……I saw them not long ago I love flowers Id love to have the whole place swimming in roses God of heaven theres nothing like nature the wild mountains then the sea and the waves rushing then the beautiful country with the fields of oats and wheat and all kinds of things and all the fine cattle going about that would do your heart good to see rivers and lakes and flowers all sorts of shapes and smells and colours springing up even out of the ditches primroses and violets nature it is as for them saying theres no God I wouldnt give a snap of my two fingers for all their learning why dont they go and create something I often asked him atheists or whatever they call themselves go and wash the cobbles off themselves first then they go howling for the priest and they dying and why why because theyre afraid of hell on account of their bad conscience ah yes I know them well who was the first person in the universe before there was anybody that made it all who ah that they dont know neither do I so there you are they might as well try to stop the sun from rising tomorrow the sun shines for you he said the day we were lying among the rhododendrons on Howth head in the grey tweed suit and his straw hat the day I got him to propose to me yes first I gave him the bit of seedcake out of my mouth and it was leapyear like now yes 16 years ago my God after that long kiss I near lost my breath yes he said I was a flower of the mountain yes so we are flowers all a womans body yes that was one true thing he said in his life and the sun shines for you today yes that was why I liked him because I saw he understood or felt what a woman is and I knew I could always get round him and I gave him all the pleasure I could leading him on till he asked me to say yes and I wouldnt answer first only looked out over the sea and the sky I was thinking of so many things he didnt know of Mulvey and Mr Stanhope and Hester and father and old captain Groves and the sailors playing all birds fly and I say stoop and washing up dishes they called it on the pier and the sentry in front of the governors house with the thing round his white helmet poor devil half roasted and the Spanish girls laughing in their shawls and their tall combs and the auctions in the morning the Greeks and the jews and the Arabs and the devil knows who else from all the ends of Europe and Duke street and the fowl market all clucking outside Larby Sharons and the poor donkeys slipping half asleep and the vague fellows in the cloaks asleep in the shade on the steps and the big wheels of the carts of the bulls and the old castle thousands of years old yes and those handsome Moors all in white and turbans like kings asking you to sit down in their little bit of a shop and Ronda with the old windows of the posadas 2 glancing eyes a lattice hid for her lover to kiss the iron and the wineshops half open at night and the castanets and the night we missed the boat at Algeciras the watchman going about serene with his lamp and O that awful deepdown torrent O and the sea the sea crimson sometimes like fire and the glorious sunsets and the figtrees in the Alameda gardens yes and all the queer little streets and the pink and blue and yellow houses and the rosegardens and the jessamine and geraniums and cactuses and Gibraltar as a girl where I was a Flower of the mountain yes when I put the rose in my hair like the Andalusian girls used or shall I wear a red yes and how he kissed me under the Moorish wall and I thought well as well him as another and then I asked him with my eyes to ask again yes and then he asked me would I yes to say yes my mountain flower and first I put my arms around him yes and drew him down to me so he could feel my breasts all perfume yes and his heart was going like mad and yes I said yes I will Yes. Trieste-Zurich-Paris 1914-1921

Figura 5 – Esempio di testo non strutturato….

Nella Figura 5 vediamo un esempio di testo formato da parole e periodi non interrotti da simboli di interpunzione; siamo di fronte all'ultimo capitolo dell'Ulisse di Joice, e il testo non strutturato vuole rendere il flusso di pensieri di Molly Bloom nel momento in cui si sta addormentando. Il precedente è un esempio di dato, costituito da un testo, di tipo *non strutturato*, dove gli unici meccanismi di strutturazione sono costituiti dagli spazi bianche tra parole e dalla sintassi della lingua inglese; per i dati non strutturati, il modello è un concetto più labile rispetto ai dati strutturati, e ciò rende molto più complessa la loro interpretazione e utilizzo.

Tornando agli orari, gli orari grafici sono scomparsi da tempo, gli orari cartacei sono ancora in vendita in alcune edicole (2019), ma stanno via via scomparendo, come stanno scomparendo le edicole. La loro consultazione è stata sostituita nei siti Web delle aziende di trasporto con orari digitali; l'esame dell'orario è facilitato nel senso che occorre fornire la stazione di partenza e di arrivo, il giorno e l'ora della partenza, e a questo punto vengono mostrate tutte le combinazioni di treni che arrivano a destinazione, inclusa l'ora di arrivo e il tempo di percorrenza del viaggio.

Per poter comprendere e poter condividere i dati, occorre dunque descriverli mediante modelli; nel nostro caso, nella Figura 6 del Capitolo 1 è stato utilizzato il modello relazionale che riprenderemo tra poco, nella precedente Figura 3 è stata utilizzata una matrice, nella Figura 4 un modello misto formato da dati numerici e grafici costituiti da segmenti e linee. Nel seguito del capitolo discuteremo quattro modelli usati per la rappresentazione dei dati nelle basi di dati e nel Web, il modello relazionale nella Sezione 2, il modello Entità Relazione nella Sezione 3 e i modelli property graph e RDF usati nel Web nella Sezione 4; successivamente nella Sezione 5 parleremo di modelli per mappe geografiche.

2. Il modello relazionale

Il modello relazionale rappresenta i dati mediante *tabelle*, chiamate anche *relazioni*. Le tabelle sono le strutture di dati adottate da molto tempo nei sistemi di gestione di basi di dati tradizionali, e furono introdotte come già detto da Ted Codd per offrire all'utente una rappresentazione di "spartana semplicità", superando le rappresentazioni gerarchiche dei primi sistemi di gestione di basi di dati. Riproduciamo in Figura 6 qui sotto la Figura 10 del Capitolo 1, tre tabelle che rappresentano tre studenti di una Università, gli esami che hanno superato e i corrispondenti corsi.

Studente

Matricola	Cognome	Stato Estero
13242	Batini	-
24195	Xu	Cina
32845	Smith	USA

Esame

Matricola	Codice Corso	Voto
13242	27	25
24195	77	28
32845	27	30

Corso

Codice	Nome Corso	Anno
27	Analisi	1
49	Algoritmi	1
77	Logica	2

Figura 6 - Tre tabelle per rappresentare studenti universitari, corsi frequentati ed esami svolti

Proviamo a fare un esercizio sulle tre tabelle di Figura 6; trovate i voti degli esami superati da Batini e i nomi dei corrispondenti corsi. Come "muoviamo gli occhi" per trovare risposta alla domanda? Beh, diciamo che prima cerchiamo nella tabella Studente il dato "Batini" e il corrispondente numero di matricola "13242"; a questo punto cerchiamo 13242 nella tabella Esami, dove lo troviamo una volta sola, accoppiato con il voto 25 e il codice del corso 27; a questo punto, infine, cerchiamo nella tabella Corsi il codice 27 e concludiamo che il nome del corso è Analisi. Come abbiamo fatto a trovare la risposta? Abbiamo dovuto navigare nelle tre tabelle, utilizzando per due volte dati di una tabella per muoverci verso un'altra tabella (la matricola 13242 per passare dalla prima alla seconda, e il codice corso 27 per passare dalla seconda alla terza).

Abbiamo scoperto che nel modello relazionale per poter collegare logicamente tabelle (e quindi navigare tra esse) lo strumento a nostra disposizione sono i *valori* dei dati, cioè le matricole e i codici dei corsi: rappresentandoli in due tabelle, possiamo navigare tra esse. Si dice anche che il modello relazionale e un *modello basato su valori*.

Un modello diverso è adottato in Figura 7; qui invece dei valori abbiamo usato frecce, anche dette puntatori, che visivamente indirizzano al dato corrispondente nelle tabelle Esame e Corso.

In particolare, le frecce che partono da Corso nella tabella Esame puntano ai corsi nella tabella Corso associati agli esami superati, le frecce che partono da Studente nella tabella Esame puntano al corrispondente studente nella tabella Studente. Nella memoria di un calcolatore, i puntatori possono essere rappresentati mediante gli indirizzi in memoria del dato "puntato".

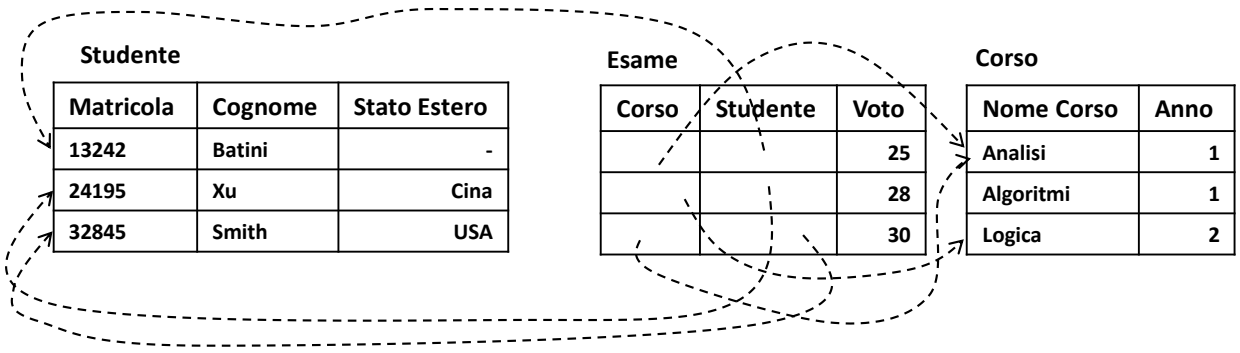


Figura 7 - Le tre tabelle in cui i legami sono rappresentati da puntatori invece che da valori

Un altro modo con cui possiamo rappresentare il contenuto informativo delle tre tabelle di Figura 6 è mediante l'unica tabella di Figura 8.

Studente-Esame-Corso

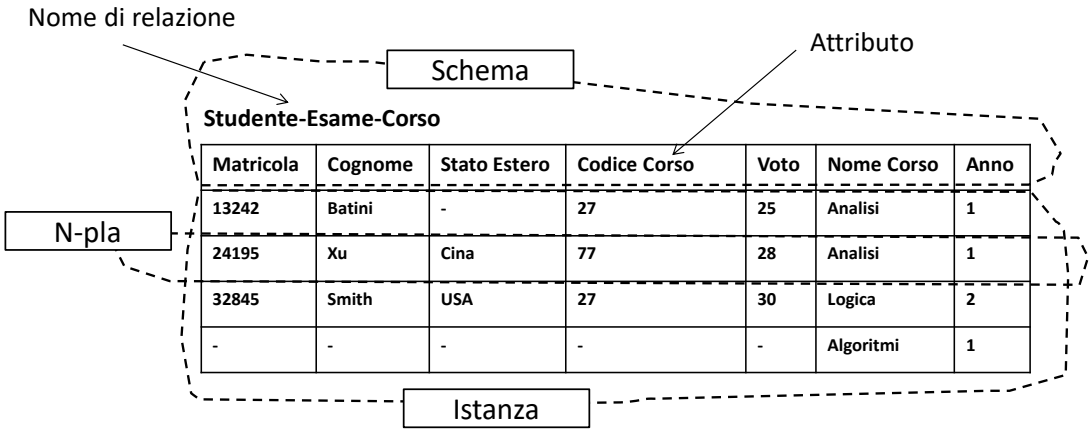
Matricola	Cognome	Stato Estero	Codice Corso	Voto	Nome Corso	Anno
13242	Batini	-	27	25	Analisi	1
24195	Xu	Cina	77	28	Analisi	1
32845	Smith	USA	27	30	Logica	2
-	-	-	-	-	Algoritmi	1

Figura 8 - I dati delle tre tabelle rappresentanti in un'unica tabella

Introduciamo ora alcune definizioni che esprimono i concetti più importanti per il modello relazionale. Facciamo riferimento alla Figura 9.

La tabella (o relazione) di Figura 9 è costituita da due parti distinte, lo schema e la istanza. Lo schema è descritto dal nome della relazione e da un insieme di attributi, che sono le proprietà che intendiamo rappresentare degli studenti, degli esami e dei corsi. L'istanza è costituita da righe tutte con la stessa struttura, che sono anche chiamate n-ple o record. Lo schema più la istanza (o in caso di più tabelle, l'insieme degli schemi + le istanze) è anche chiamato *base di dati*.

Lo schema è relativamente stabile nel tempo, mentre la istanza viene modificata ogni volta che c'è la necessità di aggiungere un nuovo studente o esame o corso. Possiamo dire che lo schema descrive il significato della tabella, mentre la istanza rappresenta i dati, descritti da valori.



Matricola	Cognome	Stato Estero	Codice Corso	Voto	Nome Corso	Anno
13242	Batini	-	27	25	Analisi	1
24195	Xu	Cina	77	28	Analisi	1
32845	Smith	USA	27	30	Logica	2
-	-	-	-	-	Algoritmi	1

Figura 9 – Concetti rilevanti nel modello relazionale

Vedi nella seguente Figura 10 lo schema delle tre tabelle di Figura 6, rappresentato in forma lineare con il nome di tabella seguito dai nomi degli attributi.

Studente (Matricola, Cognome, Stato Estero)
Esame (Matricola, Codice Corso, Voto)
Corso (Codice Corso, Nome Corso, Anno)

Figura 10 – Lo schema delle tre tabelle di Figura 6

Vi propongo a questo punto un esercizio.

Esercizio 1 - Provate a modificare lo schema precedente, aggiungendo le date degli esami, e inoltre i professori che insegnano corsi, con la matricola, il cognome, il nome, e i corsi che insegnano. Suggerimento: rappresentate i corsi più i professori che li insegnano con una tabella a parte. La soluzione in Appendice.

Torniamo ora alla base di dati di Figura 8. In questo caso, come si vede, le coppie di dati relative alle matricole e i codici corsi vengono rappresentati una sola volta; questo apparentemente dà luogo ad una rappresentazione più compatta. Potrebbe sembrare a prima vista che sia più immediatamente comprensibile la tabella di Figura 8 delle tre tabelle di Figura 6. In realtà la tabella di Figura 8 presenta diversi problemi:

1. Se un corso non è stato sostenuto come esame da nessuno studente (è il caso del corso di Algoritmi), e vogliamo rappresentare quel corso nella tabella, dobbiamo completare i dati del corso (nome e anno) con valori che hanno come significato “nessuna informazione”, anche detti valori nulli. Ciò sembra una sorta di forzatura innaturale.
2. Se più di uno studente ha sostenuto l’esame di Analisi, il fatto che analisi sia tenuto il primo anno è ripetuto più volte.

3. Se scorriamo le colonne della tabella ci accorgiamo che esse corrispondono a diversi concetti, Studente, Esame e Corso, esattamente quelli che avevamo utilizzato come nomi delle tre tabelle viste in precedenza.

Insomma, utilizzando una sola tabella noi mescoliamo diversi concetti. Questa è la ragione per cui si dice che lo schema a tre tabelle è *in forma normale*, o *normalizzato*, perché abbiamo usato ognuna delle tre tabelle per un concetto diverso, dando luogo a una rappresentazione meno confusa rispetto a quella costituita da una sola tabella, e perciò "normalizzata".

Quando c'è la esigenza di produrre una base dati, per esempio la base di dati degli studenti, dei corsi e degli esami di Figura 6, la sua produzione è formata da due attività distinte (vedi Figura 11):

- prima di tutto una attività di *progettazione*, che ha lo scopo di produrre lo schema;
- successivamente una attività di *alimentazione*, che ha lo scopo di inserire i dati nella istanza.

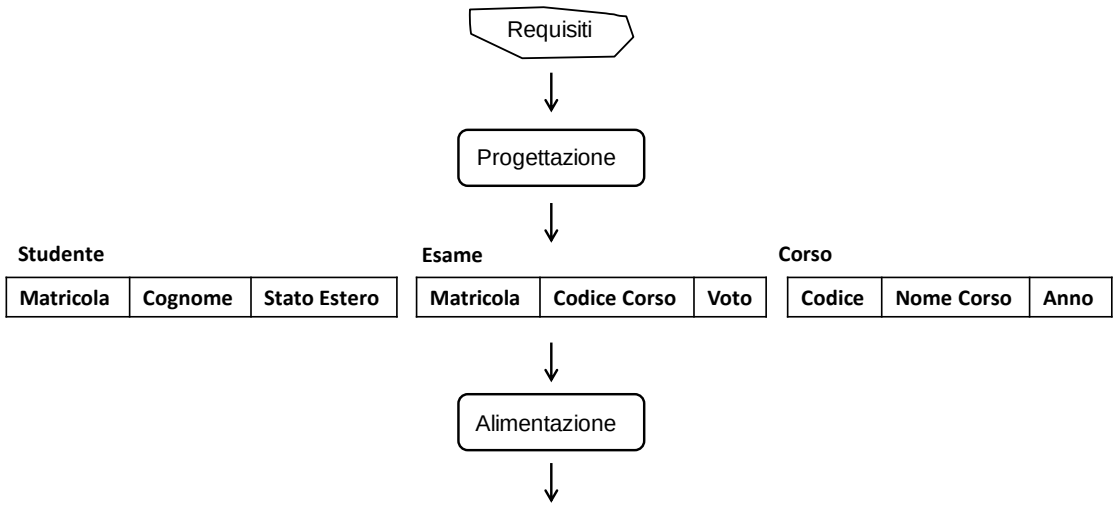


Figura 11 – Le due fasi del processo di produzione di una base di dati

Quindi: prima lo schema e poi la istanza. Se nella attività di progettazione produciamo uno schema non normalizzato, come lo schema di Figura 9, poi non sarà facile modificarlo in corso d'opera. Questo significa che è nel corso della progettazione di uno schema che dobbiamo avere l'obiettivo di produrlo *normalizzato*.

Lo strumento concettuale per produrre uno schema normalizzato da uno normalizzato sono le dipendenze funzionali, concetto che non approfondirò nel seguito (il lettore interessato può

approfondire sul testo [Batini 2016] liberamente scaricabile dal Web); piuttosto ne discutiamo una versione intuitiva.

Guardiamo la Figura 8, in cui lo schema è non normalizzato. e individuiamo nella tabella gli attributi associati agli studenti; è chiaro che essi sono Matricola, Cognome e Stato Estero. Possiamo rappresentare questi tre attributi con una tabella chiamata Studente. Possiamo operare nello stesso modo con Corso, dando luogo a una tabella con attributi Codice Corso, Nome e Anno. Infine con gli attributi residui (in pratica solo Voto) possiamo costruire una terza tabella relativa a Esame, in cui dobbiamo ricordarci di aggiungere i due attributi che permettono di collegare le tabelle, Matricola Studente e Codice Corso.

Concludiamo questa sezione, ricordando quali sono le due operazioni fondamentali che si effettuano su una tabella o un insieme di tabelle nelle basi di dati: le *interrogazioni*, che selezionano da una tabella o un insieme di tabelle tutti i record che rispettano una data proprietà, e le *transazioni*, che modificano la base di dati, aggiungendo n-ple, modificando n-ple o cancellando n-ple. I modelli dei dati sono molto utili perché permettono di esprimere le interrogazioni e le transazioni sulla base di dati mediante linguaggi semplici, come ad esempio per le interrogazioni nel modello relazionale il linguaggio SQL. Ad esempio. la interrogazione in linguaggio naturale

> Seleziona i corsi superati come esami dallo studente con Matricola = “13242” e il voto ottenuto

effettuata sulla tabella di Figura 9, è espressa nel linguaggio SQL con la seguente istruzione

```
SELECT NomeCorso, Voto
From Studente-Esame-Corso
Where Matricola = “13242”
```

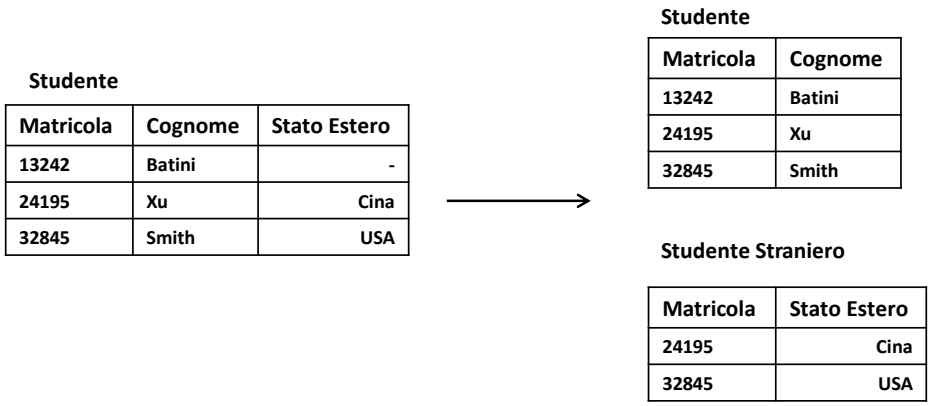
3.Il modello Entità Relazione

Il modello relazionale è adottato in molti sistemi di gestione di basi di dati, che costituiscono il software che permette ad una pluralità di utenti di accedere a una base di dati in modo efficiente e corretto, garantendo a tutti l’uso delle risorse elaborative condivise. Ora però, a ben vedere, il modello costituito da tabelle, se pur semplice, non fornisce agli utenti una rappresentazione intuitiva, ricca di significato, e facilmente comprensibile della realtà sensibile.

Pur garantendo quella che Codd chiamava una spartana semplicità, rappresentare il mondo con tabelle e attributi non è il massimo; tornando all’esempio di Figura 6, la tabella Studente rappresenta in realtà due diversi tipi di studenti, gli studenti italiani e gli studenti stranieri, e ciò è evidenziato dal fatto che la prima n-pla che fa riferimento a Batini ha un valore nullo per l’attributo Stato Estero.

Se noi vogliamo rappresentare gli stati degli studenti stranieri con una tabella a parte, possiamo trasformare la tabella nelle due tabelle di Figura 12. Tuttavia questa trasformazione non è risolutiva,

perché è compito nostro mantenere le due tabelle consistenti, e, ad esempio, rappresentare un nuovo studente straniero riportando la matricola e il cognome nella prima tabella e riportando matricola e stato nella seconda. Il fatto che gli Studenti stranieri siano un sottoinsieme degli Studenti non è rappresentato nello schema, ma deve essere gestito nel programma che effettua la transazione di aggiornamento.



Studente

Matricola	Cognome	Stato Estero
13242	Batini	-
24195	Xu	Cina
32845	Smith	USA

Studente

Matricola	Cognome
13242	Batini
24195	Xu
32845	Smith

Studente Straniero

Matricola	Stato Estero
24195	Cina
32845	USA

Figura 12 – Trasformazione dello schema relazionale per rappresentare lo stato degli studenti stranieri con una tabella a parte

Per questo ordine di ragioni, In fase di analisi, per rappresentare i requisiti degli utenti, si preferisce non utilizzare i modelli logici dei dati, in quanto troppo vicini al modo in cui i dati sono rappresentati nel sistema informatico, ed in particolare nel sistema di gestione della base di dati. I requisiti vengono piuttosto descritti per mezzo di un modello più astratto, formale, indipendente dal modello scelto per la realizzazione tecnologica, producendo in questo modo uno *schema concettuale* dei dati; successivamente, nella fase di progettazione, lo schema concettuale viene tradotto in uno *schema logico,* rappresentazione dei dati espressa nel modello relazionale.

3.1 Strutture del modello Entità Relazione e simbolismi grafici

In questa sezione esamineremo il il modello Entità Relazione (ER), che risponde alle esigenze emerse poco fa. Nel modello sono definite quattro strutture di rappresentazione: entità, relazioni, attributi, generalizzazioni a ciascuna delle quali è associata una rappresentazione grafica. Utilizzeremo i requisiti di Figura 13 per introdurre i vari concetti. Al contrario del modello relazionale, nel modello Entità Relazione è di interesse solo lo schema, non le istanze, perché il modello è utilizzato prevalentemente come modello di progettazione, e per esso non sono stati sviluppati sistemi di gestione di basi di dati.

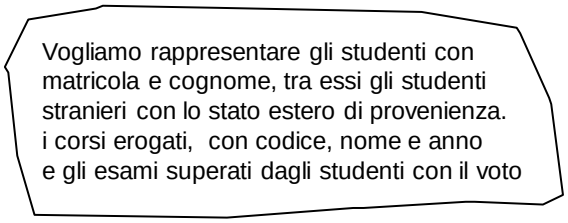


Figura 13 - I requisiti che usiamo per introdurre i concetti del modello Entità Relazione

Il grande filosofo Ludwig Wittgenstein afferma nel suo Tractatus Logico Philosophicus che "il mondo è la totalità dei fatti, non delle cose (o oggetti)". Nel modello Entità Relazione sono rappresentati sia gli oggetti della realtà che i fatti che legano gli oggetti.

Entità - Le *entità* corrispondono a classi di oggetti del mondo reale (oggetti che chiameremo in seguito *istanze* della entità) che hanno proprietà omogenee ai fini della applicazione. Le entità sono rappresentate nella rappresentazione grafica mediante rettangoli. Si veda per esempio la Figura 14. Nella parte inferiore sono rappresentati quattro studenti di una ipotetica Università; se noi vogliamo rappresentare l'insieme (o classe) degli studenti, possiamo farlo mediante una entità chiamata Studente.

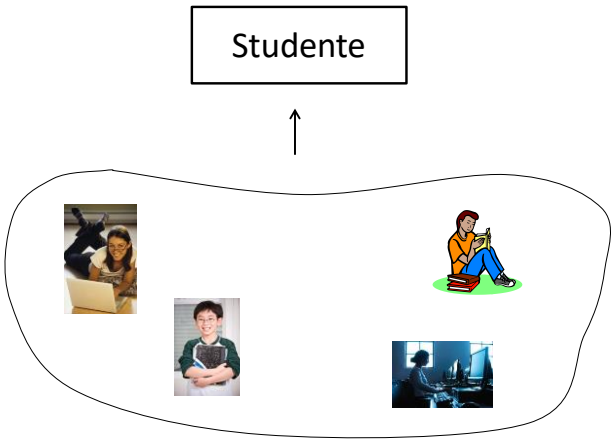


Figura 14 – L'entità Studente come classe di quattro studenti

Gli studenti di Figura 14 hanno proprietà o caratteristiche comuni, ad esempio la matricola, il nome, il cognome. Queste proprietà sono dette nel modello ER *attributi*. A un attributo è associato un insieme di valori, detto anche *dominio*, che rappresentano l'insieme dei valori elementari che l'attributo può assumere. Così, ad esempio, Matricola e Cognome sono proprietà che noi siamo interessati a rappresentare per ogni Studente, e quindi sono nel modello attributi della entità Studente. Vedi lo schema con le due entità Studente e Corso rappresentate in Figura 15; qui e nel seguito la parte dei

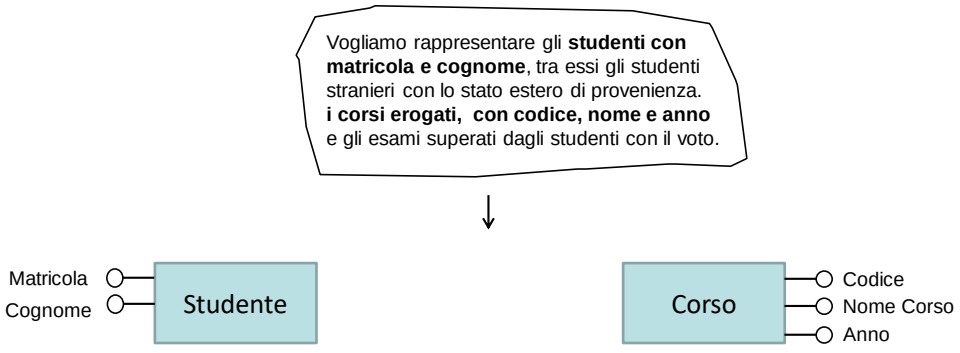


Figura 15 – Le entità Studente e Corso con i loro attributi

requisiti rappresentata nello schema è indicata in **neretto**.

Relationship o Relazioni - Mentre le entità corrispondono agli oggetti, nella terminologia di Wittgenstein le relazioni (da non confondere con le relazioni del modello relazionale, che sono tabelle) sono classi di fatti del mondo reale che sono significativi ai fini della applicazione; tali fatti mettono in relazione istanze di due entità. Ad esempio, in un censimento della popolazione possiamo essere interessati ad esprimere la relazione tra le persone e le città in cui sono nati, e chiamare tale relazione è-nato; nel nostro esempio, Esame è un concetto che esprime una relazione tra Studente e Corso. Le relazioni si rappresentano come rombi, che collegano le coppie di entità associate nella relazione, vedi Figura 16. Anche alle relazioni possono essere associati attributi, nel nostro caso Voto è attributo della relazione Esame.

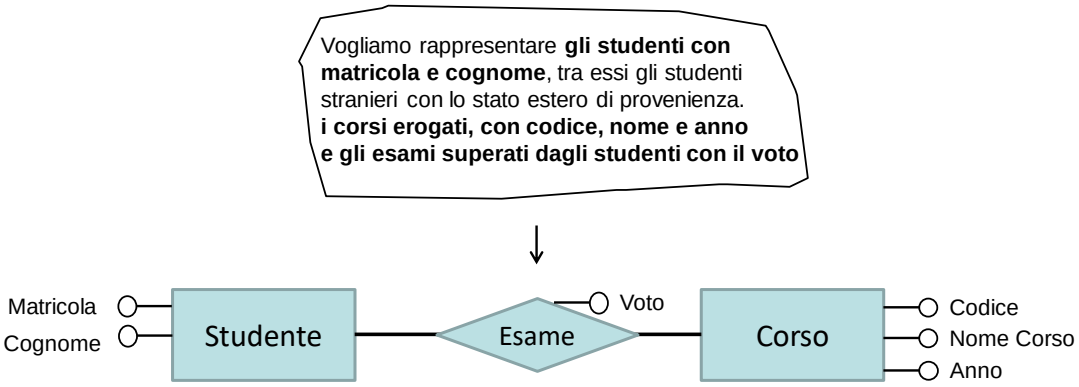


Figura 16 – La relazione Esame con attributo Voto

Generalizzazioni - Le Generalizzazioni mettono in relazione una entità EG (detta nel seguito entità genitore) con una o più entità EF1, EF2, .., EFn (dette entità figlie) che rispettano la seguente proprietà: le istanze di ciascuna delle entità EF1, EF2, .., EFn sono un sottoinsieme delle istanze di EG.

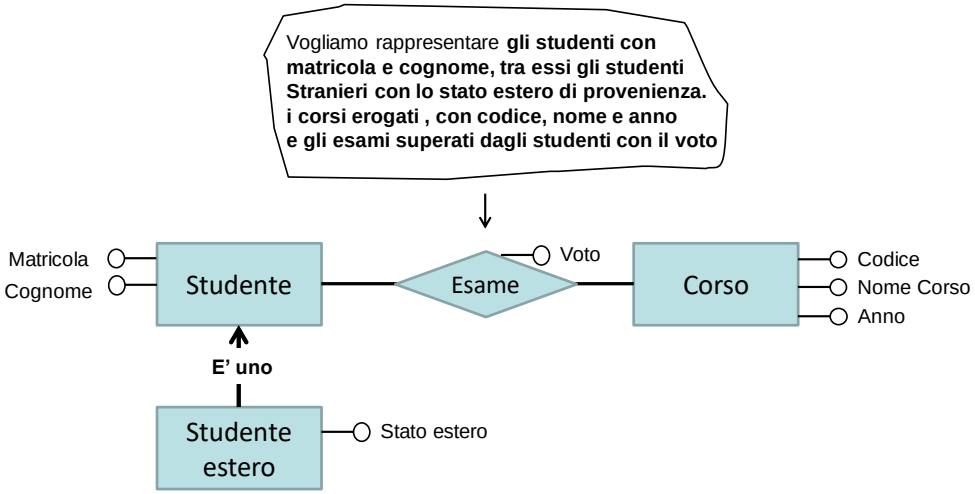


Figura 17 – La generalizzazione tra Studente e Studente Estero

Ad esempio, in una Università l'entità Studente è generalizzazione delle entità Studente italiano e Studente straniero, perche' tutti gli studenti italiani e gli studenti stranieri sono studenti. Come ulteriore esempio, la entità Luogo è generalizzazione di Comune e Stato Estero. Nei nostri requisiti riconosciamo un legame di generalizzazione tra l'entità Studente e una nuova entità Studente Estero, vedi Figura 17 della pagina precedente.

Una fondamentale proprietà delle generalizzazioni è la seguente: in una generalizzazione, ogni proprietà della entità genitore è anche proprietà delle entità figlie. Per proprietà intendiamo gli attributi, le relazioni e le generalizzazioni cui partecipa la entità. Nel nostro caso, siccome Studente è generalizzazione di Studente Straniero, gli attributi di Studente vengono ereditati da Studente Straniero, e così pure la relazione Esame con Corso. Ciò è come dire che se ogni studente ha una matricola, la avranno in particolare gli utenti stranieri. La proprietà è chiamata *proprietà di ereditarietà*.

Vi invito ora a svolgere un esercizio.

Esercizio 2 - Sulla base del significato dei concetti del modello Entità Relazione modifica lo schema Entità Relazione di Figura 17 con i seguenti requisiti:

- Rappresenta per i soli studenti italiani la città di nascita e la regione di nascita
- Rappresenta i professori con codice fiscale, cognome, nome, e i corsi che ogni professore insegna.

Hai quattro concetti a tua disposizione per rappresentare i nuovi concetti:

- l'entità, che è una classe di oggetti,
- la relazione, che è un legame tra entità,
- l'attributo, che è una proprietà elementare di entità e relazioni, e
- la relazione di generalizzazione E'-uno, definita tra entità.

Rileggi i requisiti e identifica prima le nuove entità, poi gli attributi delle entità, poi le relazioni e gli attributi delle relazioni, poi le eventuali generalizzazioni. Soluzione in Appendice.

3.2 Tipi di relazioni

In quante città di nascita è nata una persona? E' una domanda un po' strana, ciascuno di noi è nato in una sola città… Quanti film avete visto ciascuno di voi nell'ultimo mese? Mah, tanti, non mi ricordo esattamente il numero. In quante Università è iscritto uno studente universitario? Una sola! Quanti studenti sono iscritti in una Università? Tanti. Le precedenti domande, e le risposte, ci fanno capire che le relazioni nel modello Entità Relazione appartengono a varie tipologie, e che è importante distinguere tra queste.

Le relazioni nel modello ER sono di tre tipologie (le introduciamo facendo riferimento alla Figura 18):

- *uno a molti*, quando, come evidenziato in Figura 18, a ogni studente corrisponde una sola città di nascita, e, inversamente, in ogni città possono essere nati tanti studenti (e non ci interessa sapere quanti, sono certamente in genere più di uno). Si noti che dovremmo anche considerare il caso in

cui in una città non sia nato nessuno studente; questa è chiamata nel modello Entità Relazione cardinalità minima, non ci interessa approfondirla in questo testo.

- *molti a molti*, quando, ad esempio, uno studente possa aver suoperato diversi esami relativi a vari corsi, o l'esame associato a ogni corso sia stato superato da diversi studenti.

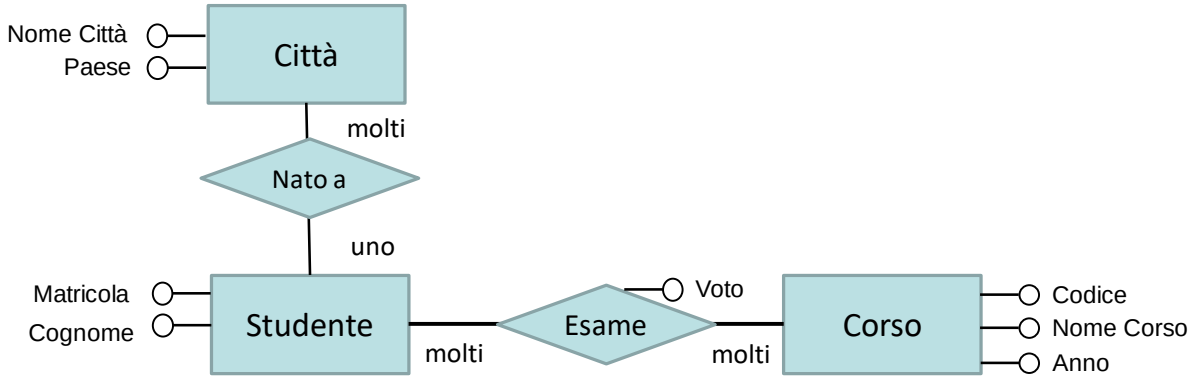


Figura 18 – Tipi di relazioni: molti a molti e uno a molti

Un terzo caso, non presente in Figura 18, riguarda le relazioni uno a uno. Concentriamoci sui matrimoni, istituti che, peraltro, nelle più recente legislazione italiana sono stati generalizzati con altri tipi di legami di vita, associati ad altre unioni di fatto. In Italia i matrimoni sono relazioni uno a uno, a ogni moglie corrisponde uno e un solo marito e a ogni marito una sola moglie. In altri paesi le relazioni di matrimonio sono uno a molti, a ogni moglie corrisponde un marito, ma a un marito possono corrispondere diverse mogli.

Anche in questi semplici esempi vediamo che rappresentare negli schemi Entità Relazione i tipi di relazioni porta ad esprimere proprietà semantiche importanti, che non possono essere espresse nativamente nel modello relazionale.

3.3. Metodologie di progettazione di basi di dati

Le basi di dati, utilizzate nei sistemi informativi in tutto il mondo, fanno uso del modello relazionale. Una base di dati è un insieme di tabelle; quindi, per progettare una applicazione software che usa una base di dati, è necessario che la applicazione informatica e il linguaggio adottato facciano riferimento al modello relazionale.

Allo stesso tempo, è più semplice da comprendere uno schema espresso nel modello Entità Relazione, anche per la sua intuitiva rappresentazione grafica. Come conseguenza, quando si vuole progettare uno schema di base di dati, si procede come in Figura 19: una prima fase di progettazione concettuale che produce uno schema Entità Relazione, e una seconda di progettazione logica che produce uno schema relazionale.

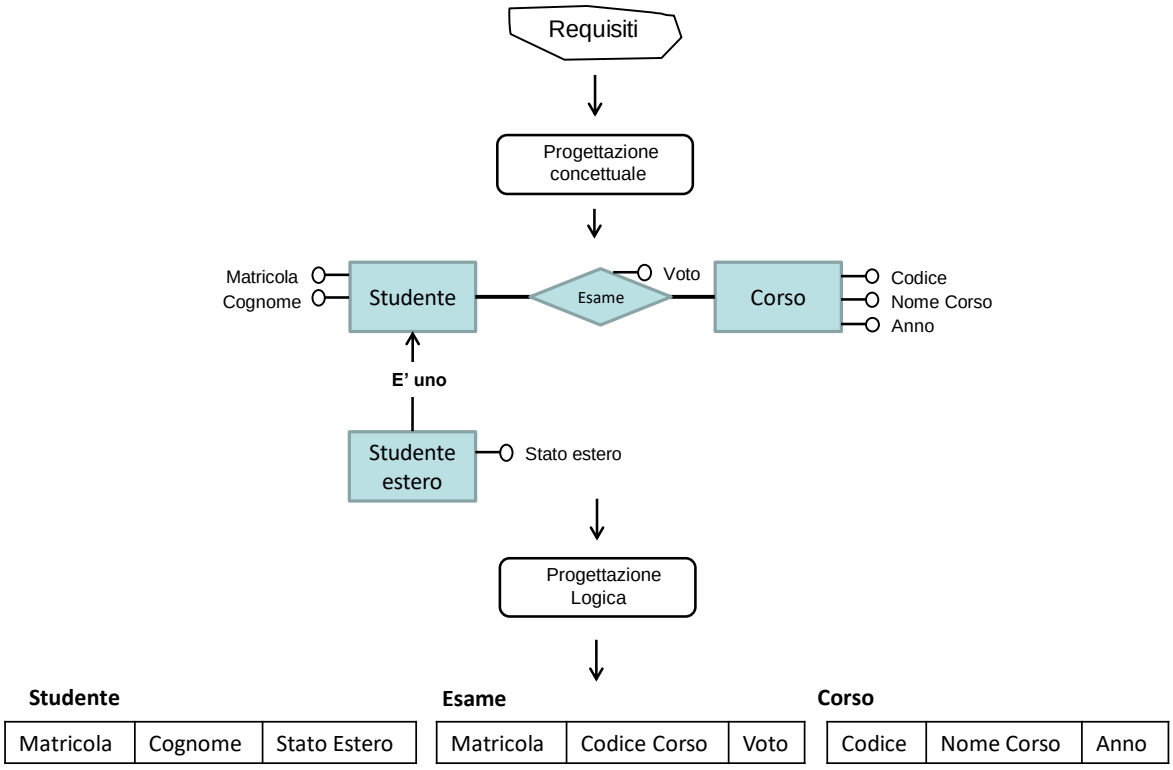


Figura 19 – La progettazione di basi di dati suddivisa
nelle due fasi di progettazione concettuale e logica

E' veramente utile la progettazione concettuale, o è una perdita di tempo? Occorre dire che una grande organizzazione ha nel suo sistema informativo centinaia di basi di dati, frutto di molte diversi attività di progettazione svolte da svariati analisti. Ciò porta a produrre basi di dati in cui sono rappresentati gli stessi oggetti del mondo reale, spesso con scelte modellistiche eterogenee. Per fare un solo esempio semplice, la data di nascita può essere rappresentata con un attributo in uno schema e tre attributi Giorno, Mese, Anno in un altro schema.

Perciò può essere utile comprendere quale sia il patrimonio informativo comune a due o più basi di dati, per capire:

- quale sia il patrimonio informativo complessivo rappresentato nelle basi di dati
- se sia opportuno unificare o fondere le due basi di dati, allo scopo di aumentare il numero di interrogazioni o elaborazioni che possono essere effettuate sull'insieme delle due basi di dati, e semplificarne la gestione e l'aggiornamento.
- quali interrogazioni o elaborazioni ulteriori potrebbero essere fatte anche mantenendo distinte le due basi di dati, collegando le due basi di dati in un sistema distribuito.
- se siano presenti ridondanze, intendendo per ridondanza la presenza delle stesse tipologie di informazioni nelle diverse basi di dati.

Avendo a disposizione gli schemi concettuali di partenza, il contenuto informativo comune può essere individuato producendo lo schema concettuale integrato; ciò può essere fatto analizzando le entità e le relazioni dei due o più schemi, individuando le eterogeneità e selezionando le entità e relazioni con gli stessi nomi, e quelle che pur non avendo gli stessi nomi risultano corrispondere allo stesso concetto del mondo reale (sinonimi). Analizzeremo più approfonditamente questi problemi nel Capitolo 7.

4. I modelli a grafo

Nel Web i dati possono essere illimitatamente condivisi, non ci sono gerarchie, tutti gli utenti sono potenzialmente parte di una comunità tra pari. Il fenomeno della condivisione dei dati sul Web porta necessariamente a adottare rappresentazioni dei dati che possano fare riferimento le une alle altre attraverso collegamenti o link, creando così una rete, un arcipelago di dati a partire da dataset isolati. Come sarebbe possibile infatti collegare logicamente due tabelle relazionali come quelle di Figura 6, una pubblicata sul Web a Milano e una a Melbourne? Nel Web i dati non possono più essere collegati attraverso i valori, come accade nel modello relazionale. Abbiamo bisogno di strutture di dati del tipo di quelle di Figura 7, in cui abbiamo utilizzato puntatori invece che valori.

Il primo ricercatore che si è posto il problema di stabilire regole e definire modelli per la condivisione dei dati sul Web è stato Tim Berners Lee, che nell'anno 2010 ha proposto cinque gradi di maturità per i dati pubblicati sul Web, corrispondenti a cinque livelli (chiamati anche *stelle*, *), essi sono:

1. Una * - Rendere disponibili i dati sul Web, in qualunque formato siano rappresentati, con una licenza open che li rende utilizzabili da tutti. Questa prima stella fa riferimento alla proprietà dei dati, che ai fini della condivisione non possono essere proprietari, ma aperti.
2. Due ** - Rendere disponibili i dati sul Web come dati strutturati, ad esempio in Excel invece che in formato scannerizzato. Questa seconda * fa riferimento alla possibilità di interrogare e elaborare i dati messi in condivisione, azione molto complessa o impossibile in un documento scannerizzato.
3. Tre *** - Rendere disponibili i dati sul Web in formato strutturato non proprietario (ad esempio Excel è un formato proprietario, mentre il formato CSV è aperto). In questo caso, l'enfasi è sulla possibilità di garantire l'accesso ai dati senza incertezza presente e futura riguardo ai diritti legali o le specifiche tecniche.
4. Quattro **** - Usare identificatori universali di risorsa per denotare gli oggetti descritti dai dati, così che la comunità degli utenti possa fare riferimento ai dati attraverso tali identificatori universali, collegandoli tra loro. Qui ci si riferisce al fatto che il dato abbia nel Web un "indirizzo" condiviso che permetta a tutti di accedere al dato.
5. Cinque ***** - Collegare i dati ad altri dati nel Web per condividerli ed integrarli nella comunità mondiale degli utenti. La condivisione può avvenire, come abbiamo detto, solo se esiste un meccanismo esplicito di collegamento tra i dati, detto link o puntatore.

Le cinque stelle di Berners Lee prefigurano modelli che permettano di *condividere* e *collegare* dati, modelli che quindi devono rispondere a un paradigma diverso e nuovo rispetto alla spartana semplicità di Codd. Tale paradigma richiede che siano rappresentati concetti e collegamenti tra concetti, quindi un modello a grafo.

Accanto alla esigenza della condivisione, fin dagli anni 90 del secolo scorso si sviluppò un'altra esigenza, quella di poter operare sui dati con funzionalità in grado di sfruttare il significato del dato, nel seguito *semantica*, per ragionare in modo automatico sul dato, inferire o dedurre nuove proprietà del dato. Se ritorniamo alla *proprietà di ereditarietà* nel modello ER tra due entità EG e EF, ricordiamo che applicandola possiamo attribuire tutte le proprietà della entità EG alla entità EF: questo è un esempio di inferenza. Possiamo sviluppare programmi software che automaticamente siano in grado di sfruttare la proprietà di ereditarietà e altre proprietà dei dati? Certamente saremo in grado, se il modello con cui sono rappresentati i dati esprime esplicitamente tali proprietà, cioè la semantica del dato. Ciò dà luogo ad una seconda esigenza insita nei modelli di dati utilizzati nel Web, essere caratterizzati da una *semantica più ricca* rispetto a quella dei modelli di basi di dati tradizionali.

Dunque, due esigenze: collegare i dati nel Web, attraverso modelli a grafo, e ragionare automaticamente sui dati, utilizzando modelli a grafo ricchi semanticamente. Secondo [Wikipedia 2019], un grafo semantico è "una rete che rappresenta relazioni semantiche tra concetti ... è un grafo diretto o simmetrico di vertici, che rappresentano concetti, e archi, che rappresentano relazioni semantiche tra concetti."

Molti modelli a grafo semantico sono stati sviluppati in anni recenti, nel seguito introduciamo il Property Graph mentre nel Capitolo 6 definiremo i Knowledge Graph e il Resource Description Framework (RDF). La caratteristica comune a tutti i modelli a grafo consiste nel rappresentare dati attraverso i concetti di nodo e arco che collega due nodi, da cui la denominazione "linked data", o dati collegati. Un property graph è formato da nodi e archi, che rispettano le seguenti regole:

- Ogni nodo ha un identificatore unico
- Ogni nodo ha un insieme di archi uscenti ed un insieme di archi entranti
- Ogni nodo ha un insieme di proprietà, espresse da una funzione chiave valore (es. Nome/Carlo)
- Ogni arco collega due nodi
- Ogni arco ha un insieme di proprietà, espresse da una funzione chiave/valore (es. Nome/Carlo)

Vi possono essere diversi tipi di archi e quindi di relazioni definibili tra i nodi. Le relazioni forniscono connessioni dirette e semanticamente rilevanti tra due nodi (ad es. Studente *Nato-in* Comune). Una relazione ha sempre una direzione, un tipo, un nodo iniziale e un nodo finale. Come i nodi, anche le relazioni possono avere proprietà; nella maggior parte dei casi, le proprietà delle relazioni sono quantitative, come pesi, costi, distanze, intervalli di tempo. Sebbene le relazioni siano rappresentate con una direzione specifica, possono sempre essere navigate nelle interrogazioni in entrambe le direzioni. Entriamo ora nel merito dei due modelli.

Vediamo in Figura 20 come si può rappresentare in un modello property graph una tabella costituita da una sola n-pla, che descrive distanza e costo della benzina tra due città, Roma e Milano.

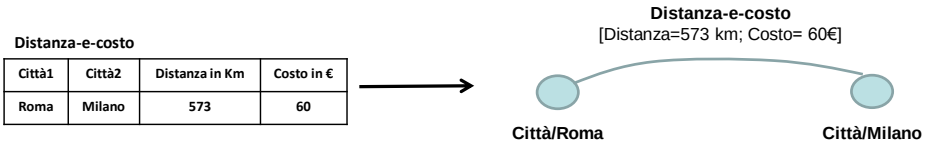


Figura 20 – Come si rappresenta una tabella di una n-pla con un property graph

Il grafo ha due nodi, caratterizzati dalle funzioni chiave/valore Cottà/Roma e Città/Milano, collegati da un arco che descrive una relazione Distanza-e-costo avente due proprietà, Distanza/573 km e Costo/60€

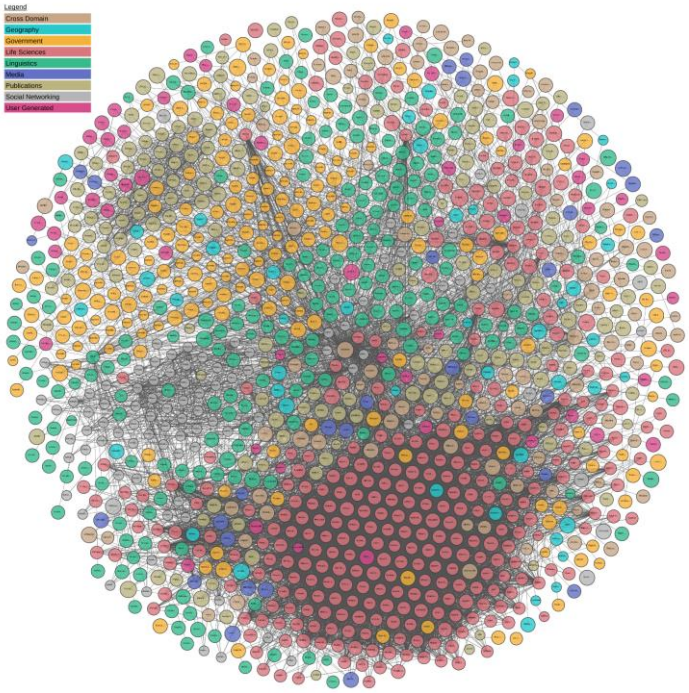


Figura 21 - Il linked open data cloud (tratta da https://lod-cloud.net/, 2019)

Concludiamo questa sezione mostrando in Figura 21 il repository di dati chiamato Linked open data cloud, creato a partire dal 2007 e che è in costante crescita; il Linked open data cloud, che è rappresentato nel modello RDF (vedi Capitolo 6), è costituito da dataset e archi tra dataset pubblicati da comunità in tutto il mondo seguendo il paradigma dei modelli a grafo. Nell'ottobre 2007 i dataset consistevano di oltre un miliardo di triple nodo-arco-nodo; nel settembre del 2011 le triple erano cresciute a 31 miliardi. Uno splendido esempio di condivisione, occorre dire, in un mondo che per altri aspetti sta continuando a erigere muri.

5. Le mappe

Da millenni gli esseri umani hanno avuto la necessità di spostarsi dai loro insediamenti per cercare cibo e migliori condizioni di vita, andare a lavorare o studiare, fare un viaggio, esplorare una città, e tante

altre occupazioni che hanno in comune la necessità di conoscere un territorio inizialmente ignoto. Questa necessità è da almeno due millenni soddisfatta dalle mappe. Io conservo tante mappe, del touring, del Club alpino italiano, e mi piace confrontarne le diverse edizioni, per confrontarne la capacità descrittiva e per seguire le evoluzioni del territorio rappresentato.

Una mappa può essere definita come una rappresentazione, usualmente su una superficie piatta, di una porzione della terra o del cielo, che ne descrive mediante simboli le forme, dimensioni e relazioni, in accordo a qualche convenzione, e la loro evoluzione nel tempo. Le mappe sono usate da millenni per un vasto insieme di attività umane come: navigare, guidare, camminare; a seconda delle azioni da intraprendere, l'utente richiede per la mappa differenti livelli di precisione e di astrazione nella rappresentazione del corrispondente territorio. Le proprietà rappresentate nelle mappe appartengono alle seguenti categorie: lo spazio, il tempo, e le caratteristiche o temi del mondo reale, considerate nella loro localizzazione spaziale e nella loro evoluzione temporale.

Una delle più antiche mappe pervenuteci è la mappa di Ecateo (vedi Figura 22); osservando la mappa troviamo diversi simboli per rappresentare i fiumi e le montagne.

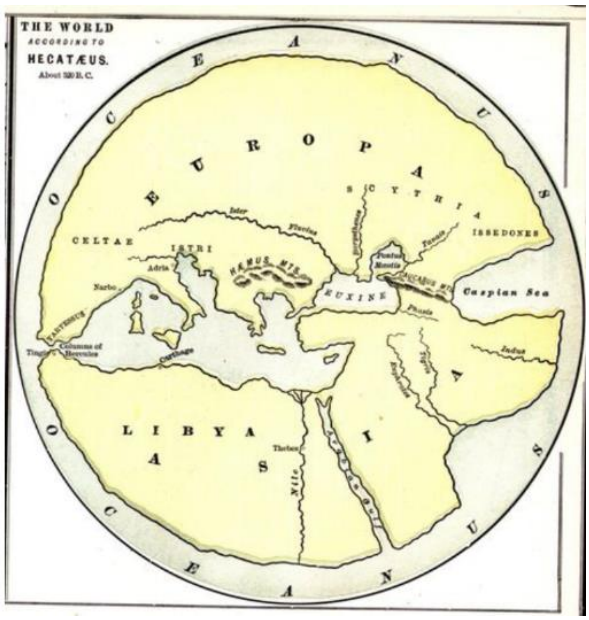


Figura 22 – La mappa di Ecateo (tratta da www.pinterest.it/)

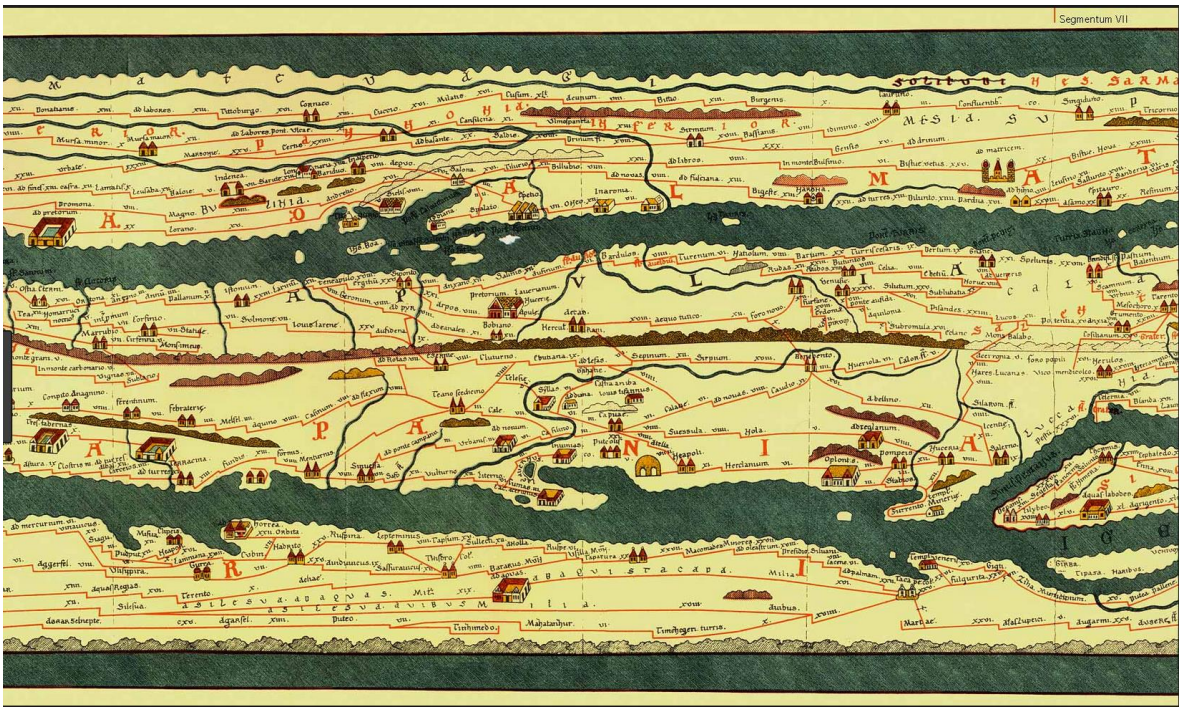


Figura 23 – La Tavola Pitungeriana (da www.wikipedia.org/)

La Tabula Peutingeriana è una copia del XII-XIII secolo di un'antica carta romana che mostra le vie militari dell'Impero romano. Nella Figura 23 si vede la riproduzione della mappa che compare in [Prontera 2003], mostrata in un suo frammento, l'intera mappa è lunga più di sei metri, il frammento circa 30 centimetri. Qui il simbolismo utilizzato è molto più vasto rispetto alla mappa di Ecateo, accanto alle montagne compaiono grandi edifici, gruppi di case che immaginiamo rappresentare città, case isolate di diversa forma, e strade.

Ciò che è comune alla maggior parte delle mappe è il fatto che sono rappresentazioni grafiche della realtà. In altre parole, vari simboli grafici sono usati per rappresentare caratteristiche geografiche o entità del territorio. L'annotazione o il testo è anche comunemente usato sulle mappe e facilita l'interpretazione della mappa. Inoltre, esistono regole implicite per cui, ad esempio, nella Tavola Pitungeriana non è possibile rappresentare case nel mare ma solo sul territorio, e due fiumi non si possono intersecare, ma semmai confluire in un unico fiume. In altre parole, anche le mappe rispondono al nostro concetto di modello, perché i simboli utilizzati devono rispettare determinate regole di composizione.

Le mappe utilizzano tre tipi di oggetti geometrici: il punto, la linea e il poligono o l'area. Un punto è definito da coordinate spaziali, una linea è definita da due punti e un poligono è definito da un minimo di tre punti. La definizione di un punto in uno spazio bidimensionale è analoga a una posizione definita da longitudine e latitudine. Inoltre, poiché linee e poligoni sono costituiti da punti, le informazioni sulla posizione sono intrinseche a punti, linee e poligoni.

Tutte le mappe bidimensionali possono essere create usando questi tre oggetti geometrici relativamente semplici. Inoltre, cambiando le caratteristiche grafiche di ciascun oggetto, emerge un numero infinito

di possibilità di rappresentazioni. Ad esempio, è possibile utilizzare punti di dimensioni diverse per riflettere le variazioni della dimensione della popolazione, il colore della linea o la dimensione della linea (cioè lo spessore) può essere utilizzato per denotare il volume o la quantità di interazione tra le posizioni, e possono essere utilizzati diversi colori e forme per riflettere diversi valori di interesse.

A completamento degli elementi grafici descritti in precedenza compaiono nelle mappe, come detto in precedenza, le annotazioni o testi. L'annotazione viene utilizzata per identificare particolari caratteristiche geografiche, come città, stati, corpi idrici o altri punti di interesse. Come gli elementi grafici, il testo può essere variato in base alle dimensioni, all'orientamento o al colore. Ci sono anche numerosi caratteri di testo e stili che sono incorporati nelle mappe. Ad esempio, i corpi idrici sono spesso etichettati in corsivo.

Un altro elemento delle mappe che va menzionato e che combina sia grafica che testo è la *legenda* della mappa. Una legenda fornisce al lettore informazioni sul modo in cui le informazioni geografiche sono rappresentate graficamente. Le legende consistono solitamente in un titolo che descrive la mappa, nonché i vari simboli, colori e motivi utilizzati sulla mappa. Tali informazioni sono spesso vitali per la corretta interpretazione di una mappa.

In Figura 24 e Figura 25 vediamo due diversi tipi di legenda. Si nota chiaramente che la legenda di Figura 24 è stata pensata per quella particolare mappa, mentre la legenda di Figura 25 ha uno scopo più generale.

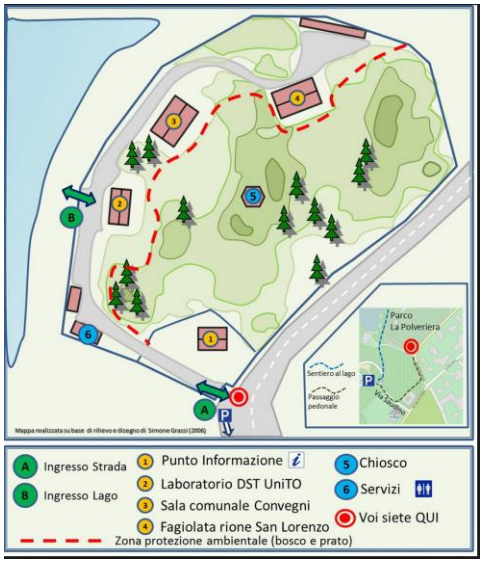


Figura 24 - Una mappa con legenda

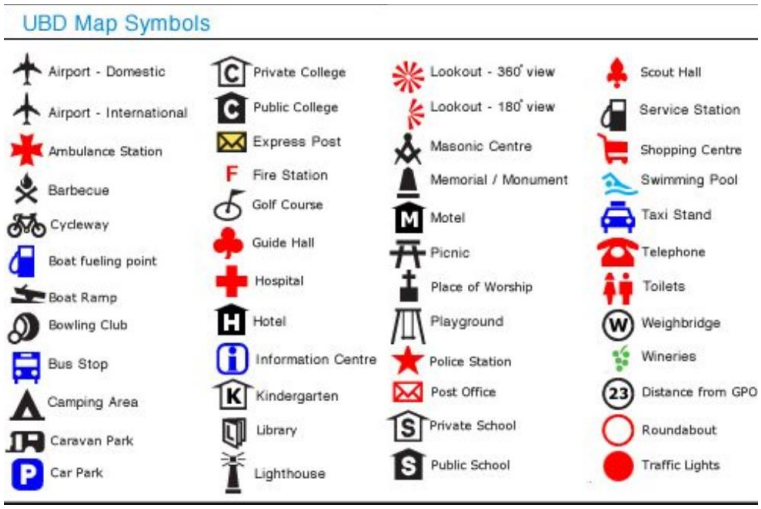


Figura 25 – Un esempio di legenda di natura più generale della precedente (da www.pinterest.it/)

Si notino in Figura 25 alcuni simboli, come i due tipi di aeroporti ovvero i due tipi di college, che appaiono come caso particolare di un elemento territoriale più generale, rispettivamente l'aeroporto e il college, rispetto ai quali sono caratterizzati da un legame concettuale di generalizzazione.

La forma e il contenuto delle mappe (o carte, nel seguito) variano in base allo scopo. Si distinguono infatti

- carte generali, per esempio, carte fisiche e politiche, e
- carte tematiche, che descrivono la distribuzione di un particolare fenomeno sul territorio

Tra i tanti tipi di carte tematiche si possono distinguere

- le carte geomorfologiche, rappresentano le forme del terreno e indicano i processi che le hanno originate.
- le carte pedologiche rappresentano i tipi di suolo.
- le carte climatiche forniscono informazioni sul clima di una data regione.
- un particolare tipo di carte climatiche sono le carte meteorologiche, che documentano la variazione del tempo.

Con l'avvento del Web si sono sviluppati nuovi tipi di mappe; si veda in Figura 26 un piccolo frammento di un quartiere della città di Roma, rappresentato con OpenStreetMap e con Google Earth.

OpenStreetMap è un progetto mondiale per la raccolta collaborativa di dati geografici da cui si possono derivare innumerevoli artefatti e servizi. Gli artefatti più popolari sono le mappe online, che però rappresentano solo la punta dell'iceberg di quel che si può ottenere da questi dati. La caratteristica fondamentale dei dati di Open Street Map è quella di possedere una licenza libera, e quindi di essere utilizzabile da chiunque. L'altra caratteristica molto importante di OpenStreetMap è che tutti possono contribuire arricchendo o correggendo i dati riportati nelle mappe; come in progetti simili (es. Wikipedia) la comunità è l'elemento fondamentale perché, oltre a inserire i dati e arricchire il progetto, ne controlla anche la qualità.

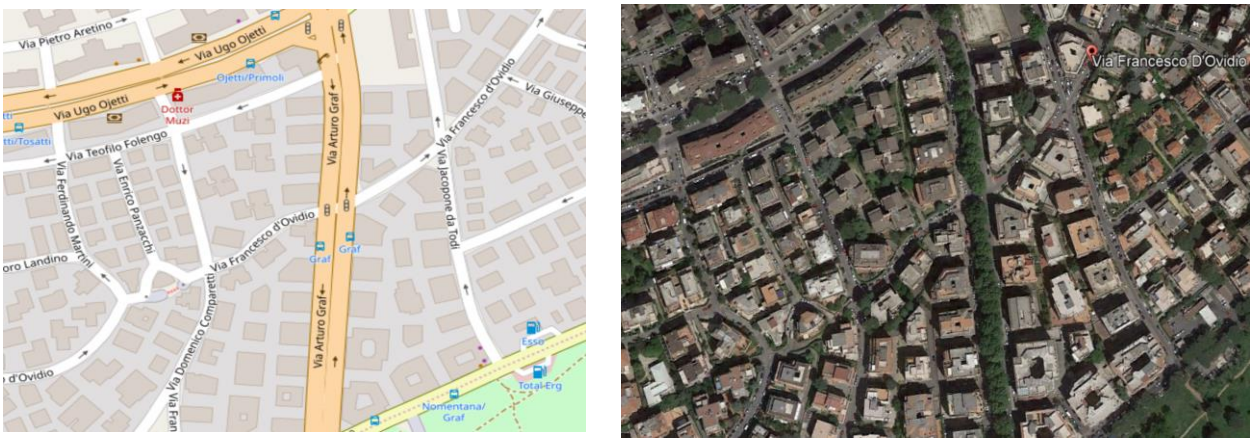


Figura 26 - Roma e Via D'Ovidio, dove ho vissuto a lungo,
viste con OpenStreetMap e con Google Earth

Le immagini di Google Earth [Wikipedia 2019] sono visualizzate su un globo digitale, mostrando la superficie del pianeta usando una singola immagine composita e vista da una distanza che può essere modificata con un cursore attraverso primitive di zoom in e zoom out.. Le fonti delle immagini sono satelliti o aerei. L'immagine finale è ottenuta combinando più serie di immagini prese dal satellite Landsat 7 con lo scopo di eliminare nuvole e spazi diagonali, creando una singola immagine "a mosaico". Recentemente (2019), Google ha iniziato a utilizzare il satellite Landsat 8 per fornire immagini di qualità superiore e aggiornate con maggiore frequenza. Le immagini di Google Earth sono sotto copyright.

Google Earth è ispirata da uno stile fotorealistico, che cioè riproduce il territorio nella maniera più fedele possibile. Tuttavia, a volte gli algortimi utilizzati, come accade in Figura 27, restituiscono rappresentazioni del territorio affette da errori che sono piuttosto clamorosi.

Figura 27 – Effetti distorsivi degli algoritmi usati in Google Earth
(tratta da Google Earth)

Riferimenti e approfondimenti

E.F. Codd - A Relational Model of Data for Large Shared Data Banks, Communications of the ACM, 1970.

C. Batini, S. Ceri, S.Navathe – Conceptual Database Design: An Entity-Relationship Approach, Addison-Wesley, 1991.

F. Prontera (a cura di) - Tavola Peutingeriana, Le antiche vie del mondo Ed. Olschki, 2003

Se vuoi approfondire il modello relazionale e modello Entità Relazione puoi

- Scaricare dal sito http://hdl.handle.net/10281/97114 le trascrizioni del corso su Data Base Modeling and Design in inglese di Carlo Batini.
- Accedere alle lezioni video dello stesso corso sul sito https://open.elearning.unimib.it/course/view.php?id=52
- Accedere alle lezioni in Power Point commentate a voce in italiano sul sito http://elearning.unimib.it/course/view.php?id=14058 (qui è anche trattato l'SQL e l'Algebra Relazionale)

Capitolo 4 – Le tecnologie per Big data

Andrea Maurino

4. Introduzione

All'inizio dell'Informatica, il calcolatore è stato pensato per eseguire operazioni matematiche, anche molto complesse, in poco tempo. Immediatamente si è però compreso come la capacità di memorizzare dati e di ritrovarli più facilmente di quanto possa fare un essere umano fosse un'altro importante utilizzo di questa nuova tecnologia. Sin dai suoi albori, l'Informatica ha sviluppato tecniche e strumenti per acquisire, memorizzare e restituire una quantità di dati sempre più grande; la data driven society dei nostri giorni è anche il risultato di questi continui investimenti e soluzioni tecnologiche.

Ogni volta che ci rechiamo presso uno sportello bancario per effettuare un bonifico, l'impiegato cerca il numero del nostro contro corrente ed esegue una operazione di trasferimento di fondi da un conto corrente ad un altro.

Quanto si diventa amici di un altro utente su un social network si crea una relazione fra due persone, ma anche quando si associa un libro all'acquirente si sta costruendo una rete di connessioni di diverso tipo, che possono essere usate poi per predire i comportamenti di acquisto futuro, ovvero per studiare le reti di influenza di un determinato utente in un social network.

Una cartella medica elettronica contiene diversi tipi di analisi, prescrizioni e risultati tutti facenti rifermento allo stesso paziente. Raccogliere insieme tutti i documenti relativi ad una persona in un unico contenitore è una attività abbastanza comune anche nella vita reale, si pensi alla cartella con i documenti sanitari dei figli, o la cartella delle tasse da pagare per un determinato appartamento o ancora la cartella che racchiude tutti i disegni fatti a scuola da un figlio.

In conclusione, tutte le volte che utilizziamo un dispositivo elettronico, un *computer*, *tablet*, o *smartphone*, questo produce ed elabora dati, a volte piccoli a volte molto grandi.

Come detto nel prologo, nel 1970 il ricercatore dell'IBM Ted Codd pubblicò su una rivista scientifica il primo articolo [Codd 1970] che introduceva il modello relazionale dei dati, per oltre 30 anni l'unico modello di rappresentazione dei dati strutturati in formato elettronico largamente utilizzato. Il modello relazionale, estensione della teoria degli insiemi, è stato uno dei casi di maggior rilevanza di sopravvivenza di una tecnologia digitale; lo abbiamo descritto in dettaglio nel Capitolo 3. Ancora oggi, una larga maggioranza di dati relativi al mondo bancario, sanitario, alla pubblica amministrazione oltre che nell'industria, è gestita tramite sistemi di gestione di basi di dati relazionali, i cui principi di base sono sostanzialmente quelli definiti da Codd nell'articolo citato.

A partire dai primi anni 2000, la tecnologia delle basi di dati relazionali è stata messa in discussione, per la sua incapacità di gestire i nuovi tipi di dati digitali e i loro volumi. Effettivamente, le nuove esigenze

applicative nate nel mondo di Internet hanno evidenziato i limiti della tecnologia relazionale; si è assistito alla nascita di una nuova serie di modelli per rappresentare i dati digitali e contestualmente sono stati definiti nuovi paradigmi di memorizzazione e analisi dei dati. Abbiamo già visto le motivazioni che hanno fatto evolvere i modelli verso i modelli a grafo nel Capitolo 3; intuitivamente, per descrivere una rete di amicizie su Facebook, è molto più adatto un modello a grafo, in cui i legami di amicizia sono rappresentati come relazioni, rispetto a un modello basato su relazioni.

Come abbiamo iniziato a osservare nel Capitolo 1, al giorno di oggi sempre più frequentemente fenomeni fisici del mondo reale (ad esempio la temperatura atmosferica, la velocità di un'auto), ma anche sentimenti umani come amicizia, amore, odio sono trasformati in dato elettronico. Il processo di "datification" della realtà ha portato a una crescita imponente dei dati disponibili in formato digitale anche grazie alle tecnologie di comunicazione. A titolo di esempio, si riportano in Figura 1, riferita all'anno 2020 i dati che in 60 secondi si possono generare sulla rete di comunicazione Internet.

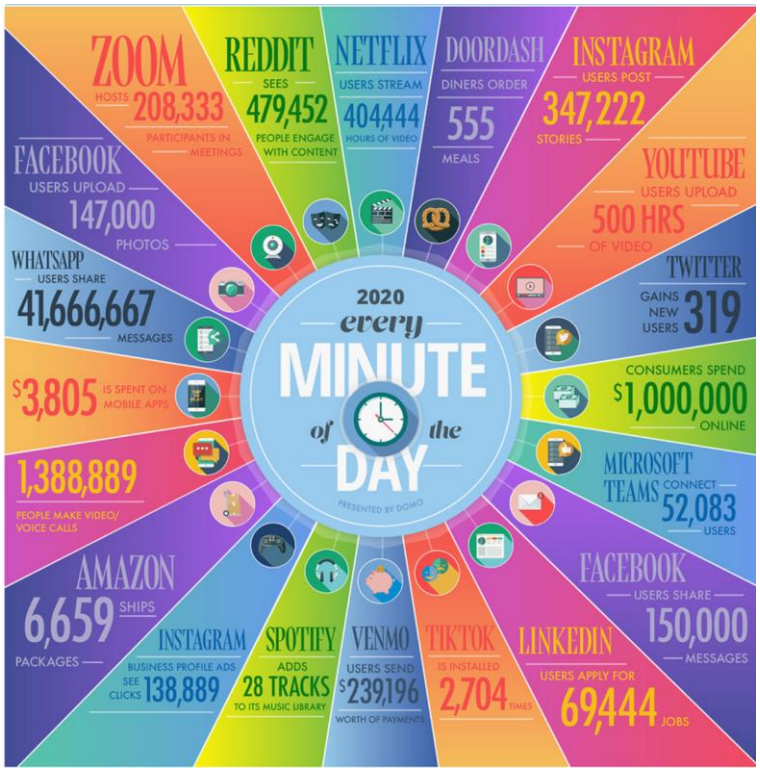


Figura 1 – I dati generati su Internet in 60 secondi nel 2020
(tratta da https://www.visualcapitalist.com/every-minute-internet-2020/)

Sono numeri certamente impressionanti, che richiedono profonde innovazioni nel modo con cui i dati sono codificati, memorizzati e recuperati. Il modello relazionale e le tecnologie dei sistemi di gestione di basi di dati ad esso collegate, avvertiamo subito, non è da considerare come una tecnologia obsoleta. È sempre più evidente, e lo vedremo nel capitolo, come al giorno di oggi non esista una soluzione di gestione dei dati che va bene per tutte le applicazioni, ed è importante conoscere di ogni modello dei dati e del relativo sistema software di gestione i punti di forza e di debolezza per poter sfruttare appieno le nuove tecnologie che si affiancano, ma non sostituiscono, il modello relazionale dei dati.

Il capitolo è organizzato come segue; nella Sezione 2 si presentano i nuovi modelli di rappresentazione dei dati digitali nell'epoca dei grandi volumi di dati. Va tenuto presente che spesso si tratta di modelli già noti nel passato, ma che oggi grazie alle tecnologie disponibili sono diventati competitivi con il modello relazionale, e utilizzabili in molti contesti applicativi. La Sezione 3 presenta le principali soluzioni tecnologiche per la rappresentazione di dati in una rete di basi di dati distribuita in un'area geografica. Al crescere del volume dei dati e per garantire la disponibilità degli stessi, è fondamentale distribuire i dati fra più elaboratori chiamati server. In base ai livelli di coordinamento fra i nodi della rete, esistono diverse soluzioni tecnologiche possibili. Nella Sezione 4 vengono presentati i principi architetturali della soluzione Hadoop, utilizzata per la gestione di dati caratterizzati da una grande varietà, velocità di aggiornamento e da elevato volume (Le tre V descrittive delle caratteristiche dei big data discusse nel Capitolo 1). Nella Sezione 5 formuliamo alcune considerazioni conclusive su quanto descritto in precedenza. Va sottolineato che ognuno degli argomenti citati precedentemente richiederebbe una trattazione molto articolata; per questo motivo, nello spirito del libro, si mostreranno i concetti fondamentali delle tecnologie per basi di dati, tralasciando i dettagli tecnici e teorici.

5. I nuovi modelli dei dati

Come evidenziato nell'introduzione, fino ai primi anni 2000 il modello relazionale dei dati era il principale e sostanzialmente unico modello per la descrizione e gestione di dati strutturati. E' bene ricordare come negli anni 70 e 80 del secolo scorso le tecnologie hardware e di rete non erano neanche lontanamente assimilabili a quelle attuali, e, di conseguenza, i requisiti applicativi erano profondamente diversi.

Basti ricordare che il termine "computer per uso personale" (personal computer o PC) è una invenzione dei primi anni 80 del secolo scorso, mentre all'epoca di Codd gli hard disk (inizialmente chiamati anche tamburi) erano come quelli mostrati in Figura 2, ed erano in grado di memorizzare solo 250 Mbytes (milioni di byte, un byte permette di memorizzare, ad esempio, un carattere alfabetico), ovvero l'equivalente di poco meno di 50 foto di un comune telefono cellulare, ad un costo di svariate decine di migliaia di dollari!

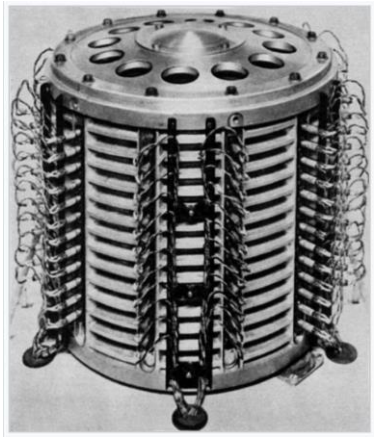


Figura 2 – Una memoria hard disk negli anni 80 del secolo scorso

(tratta da https://it.wikipedia.org/wiki/Memoria_a_tamburo)

È chiaro come di fronte a uno spazio di memorizzazione così costoso, uno degli elementi fondamentali di successo del modello relazionale è la unicità nella rappresentazione del dato, ovvero la proprietà consistente nel memorizzare una sola volta nella base di dati un determinato dato digitale. Inoltre, questa caratteristica consente una migliore gestione della qualità del dato; se devo aggiornare un dato, ad esempio la città di residenza di uno studente, lo devo fare una sola volta, mentre in presenza di numerose copie devo aggiornare tutte le copie dello stesso dato, e posso generare inconsistenze se mi scordo di aggiornane qualcuna. Tratteremo il tema della qualità dei dati nel prossimo Capitolo 5.

Per apprezzare a pieno la grande novità del modello relazionale nel momento in cui fu lanciato, occorre ricordare che negli anni 60 l'ingegneria del software, la disciplina che studia il ciclo di vita del software, dalla definizione dei requisiti alla progettazione e codifica di una applicazione software, era ai suoi albori, e il linguaggio di programmazione con cui si codificavano i programmi era il Cobol, linguaggio citato nel prologo di questo libro. Nel linguaggio Cobol i dati erano descritti da una Data Division, ed erano locali ai programmi; ogni programma aveva i suoi dati. Inoltre per esprimere una interrogazione, era necessario produrre un programma apposito, con istruzioni molto verbose e poco potenti. È importante tenere a mente queste caratteristiche, che hanno reso il modello relazionale e il linguaggio di interrogazione dei dati, l' SQL, di cui abbiamo mostrato un esempio nel Capitolo 3, molto popolari fino ai giorni nostri.

Le nuove applicazioni informatiche nate a seguito dell'avvento di Internet e del Web (social network, videogiochi on line, ecc.) male si adattano a essere realizzati usando il modello relazionale e le sue tecnologie software. Da un lato il modello ha delle rigidità modellistiche, in quanto tutti i dati devono essere rappresentati nello stesso modo, mediante tabelle, e dall'altro la tecnologia utilizzata, seppur evoluta nel corso degli anni, non ha mai abbandonato alcune caratteristiche fondamentali del linguaggio SQL. I moderni linguaggi di programmazione prevedono un paradigma di gestione di dati (detto a oggetti) diverso da quello del linguaggio SQL; di conseguenza per scrivere dei programmi con i moderni linguaggi di programmazione che interagiscono con una base di dati relazionale, è necessario prevedere la realizzazione di software chiamato middleware, in grado di far dialogare l'applicazione software con la base di dati, che in genere è molto complesso. La riduzione progressiva del tempo di vita sul mercato delle applicazioni, si pensi a videogiochi che vivono per meno di una stagione, ha imposto una forte spinta a ridurre il tempo di produzione del relativo software applicativo.

Questi due fattori (i vincoli modellistici e i limiti tecnologici delle basi di dati relazionali) hanno dato impulso alla ricerca di nuovi modelli di rappresentazione dei dati. Il termine NoSQL[1] (not only SQL) fa riferimento ai nuovi modelli di gestione dei dati che vanno oltre il modello relazionale e il suo linguaggio di interrogazione SQL.

[1] Il termine NoSQL è stato introdotto per la prima volta nel 1989 dall'italiano Carlo Strozzi come nome di un insieme di interfacce programmabili (API) per accedere a una base di dati relazionale non usando il linguaggio SQL. L'acronimo indicava inizialmente la volontà di NON usare il linguaggio SQL. Oggi si fa riferimento alla sigla NoSQL con il significato di "non solo SQL".

A causa del proliferare di questi nuovi modelli, si è scelto di presentarli secondo un ordine che parte dal modello più semplice di gestione dei dati, il modello Chiave-valore, per arrivare a modelli sempre più sofisticati per la gestione di dati complessi.

5.1. Modello Chiave-valore

Nel modello chiave-valore si associa ad una chiave, ovvero un valore univoco nell'insieme dei dati (dataset nel seguito), un valore. Ad esempio, se vogliamo rappresentare il cognome di un insieme di persone, possiamo usare come chiave il codice fiscale, e come valore il cognome. Il valore può essere non solo un tipo di dato semplice come una stringa di caratteri o un numero intero, ma anche un oggetto più complesso, come un video o una immagine o un frammento di una pagina Web. Questo modello è molto semplice, ed è utile nei contesti applicativi in cui è necessario accedere ad un valore nota la chiave.

Le basi di dati chiave-valore sono molto usate per le applicazioni che richiedono di gestire in memoria centrale una grande quantità di dati. Occorre ricordare che la memoria di un elaboratore è strutturata in diversi strati di memoria, tra cui possiamo distinguere una memoria centrale, "vicina" alla unità di calcolo dove vengono eseguiti i programmi, e una memoria secondaria, realizzata mediante dischi magnetici e, nelle versioni più recenti, mediante tecnologie a stato solido. Con il termine "vicina" intendiamo il fatto che per trasferire un dato da memoria centrale alla unità di calcolo occorrono 10^{-8} / 10^{-9} secondi; invece, per trasferire un dato da memoria secondaria alla unità di calcolo occorrono nel caso di memoria realizzata a dischi circa 10^{-2} secondi. La memoria centrale ha limiti di capacità, per cui in genere le basi di dati vengono memorizzate permanentemente in memoria secondaria.

Si pensi allora al carrello della spesa di una applicazione di e-commerce. Esso può essere modellato come una coppia chiave-valore dove la chiave è l'identificativo dell'utente e il valore è l'intero contenuto del carrello. L'accesso alla memoria centrale rispetto a quella secondaria consente un risparmio di tempo molto significativo. Per tale motivo, i grossi venditori on line come Amazon o le piattaforme di giochi on line usano massicciamente questo modello, perchè consente l'accesso a tutti i dati di un cliente in maniera molto veloce.

Essendo un modello molto semplice e utilizzato usualmente su memorie non persistenti, che cioè non garantiscono che a causa di un improvviso calo di tensione o guasto al sistema i dati non si perdano, il modello chiave-valore ha un raggio di applicazioni ben specifico; certamente, le capacità tecnologiche delle soluzioni disponibili sul mercato per il modello lo rendono ideale per gestire in modo efficace una memoria centrale dinamica.

5.2. Modello Wide Column

Come detto precedentemente, il modello chiave-valore è molto semplice ed efficace in diversi contesti applicativi. È possibile far evolvere il modello chiave-valore arricchendo la semantica del campo valore. Nel modello wide column è possibile ad esempio strutturare ogni singolo valore di attributo, nella terminologia del modello relazionale, in un insieme di campi che possono assumere significati e valori diversi.

il modello wide column si avvicina per alcuni aspetti al modello relazionale, in quanto riprende il concetto di relazione con attributi. Le più importanti differenze rispetto al modello relazionale sono

- la mancanza di uno schema rigido (si parla anche di modello schemaless, senza schema) e
- la presenza di una chiave identificativa di solito definita dal sistema e non definibile dall'utente.

Per quanto riguarda gli aspetti legati alla flessibilità dello schema, si vede nella Figura 3 come uno stesso dataset che descrive le informazioni gestite da un negozio virtuale di eCommerce può rappresentare nella prima riga un libro, con un identificatore (una coppia di attributi costituta da un codice prodotto e, nella prima riga, un identificatore del libro), il titolo, l'autore e l'anno di pubblicazione, mentre la riga successiva, che si riferisce a un album musicale, contiene oltre all'identificatore solo il titolo e l'autore. Successivamente, abbiamo una terza riga con il titolo di una traccia (si veda come anche l'identificatore cambia); ancora diversa è l'ultima riga, in cui è presente un film descritto dall'identificatore, il titolo, la tipologia e l'autore.

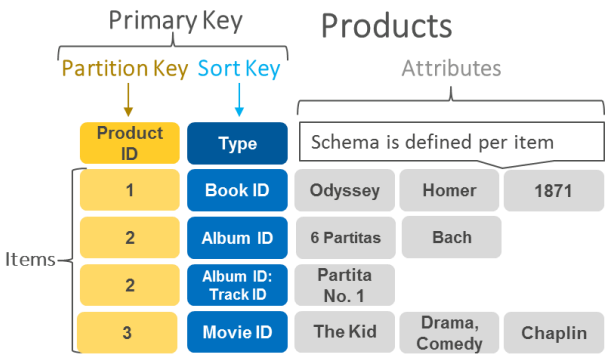


Figura 3 – Un dataset nel modello wide-column
(tratta da https://aws.amazon.com/it/nosql/key-value/)

È importante sottolineare come nel modello wide column la flessibilità dello schema è giustificata dalle necessità applicative dei moderni sistemi informativi, che devono gestire piattaforme tecnologiche a rete, complesse e che si modificano dinamicamente nel tempo; pensiamo al sistema informativo di Amazon, che gestisce una rete di utenti e fornitori distribuita a livello mondiale con cui intrattiene relazioni di business che si modificano nel tempo.

Questi sistemi informativi richiedono estrema flessibilità dei requisiti e, di conseguenza, nella progettazione o nella manutenzione evolutiva di applicazioni software. Lo sviluppo di applicazioni basate su basi di dati tradizionali prevede una lunga fase iniziale di strutturazione dello schema della base di dati, e solo successivamente lo sviluppo del codice. Nel caso in cui sia stato necessario durante lo sviluppo modificare lo schema della base di dati, è necessario fermare lo sviluppo e ricominciare dalla attività di progettazione della base di dati; questo processo di produzione, certamente efficace per applicazioni complesse e relativamente stabili nel tempo, è oggi troppo oneroso per i vincoli di rapidità e di flessibilità nello sviluppo software descritti in precedenza. Grazie alla flessibilità dello schema, è possibile durante lo sviluppo dell'applicativo, o successivamente al rilascio di esso, modificare lo schema

adattandolo ai nuovi dati da inserire. È evidente comunque che questa flessibilità debba essere ben governata per non generare problemi di qualità nei dati, derivanti dalle continue modifiche ai requisiti.

Il modello wide column è stato implementato da numerosi "giganti" delle tecnologie come Google, Amazon e Facebook con i loro prodotti denominati Big Table, Dynamo e Cassandra. Va notato che, in assenza di una standardizzazione, ognuno di loro ha sviluppato non solo diverse soluzioni per la scalabilità ma soprattutto diversi linguaggi di interrogazioni che comunque si ispirano in misura diversa al linguaggio SQL.

Nella Figura 4 troviamo un altro esempio che mostra la flessibilità del modello wide column rispetto al modello relazionale. Supponiamo di voler costruire una anagrafica dei clienti di una banca; è possibile che non tutti i clienti abbiano definiti gli stessi attributi, ovvero che per qualche attributo non sia noto il loro valore. Nel modello relazionale questo è risolvibile utilizzando il valore Null. Nel modello wide column è possibile invece semplicemente non assegnare i valori.

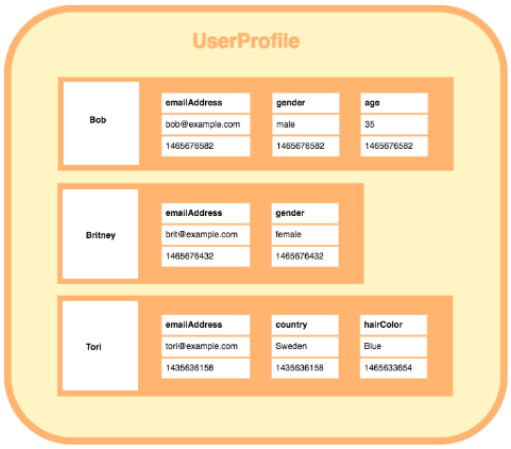


Figura 4 – Un frammento di una anagrafica dei clienti nel modello wide-column
(tratta da https://database.guide/what-is-a-column-store-database/

Dal punto di vista teorico, è possibile affermare che mentre il modello relazionale si basa sulla ipotesi di mondo chiuso (closed world assumption, tutte le informazioni note sono memorizzate nel dataset, e se non sono memorizzate non esistono), tutti i modelli NoSQL, e dunque anche il modello wide column, si basano sulla ipotesi di mondo aperto: su tutto ciò che non è rappresentato nella base dati non si può dire nulla, né che esista, né che non esista.

Va infine ricordato come il modello wide column non prevede nativamente di collegare logicamente dati fra di loro, come è possibile fare con il modello relazionale; si tratta di un elemento importante nella corretta scelta del modello. Facendo riferimento ai concetti del modello Entità Relazione, nel modello wide column le relazioni tra i dati sono solo di tipo uno a uno.

2.3 Modello documentale

È possibile arricchire il modello precedentemente mostrato, consentendo a ogni singolo attributo di essere non solo un attributo di tipo semplice avente come valori, ad esempio, stringhe di caratteri come in un cognome, ma anche un attributo di tipo composto caratterizzato da un insieme di valori, composti a loro volta da coppie chiave-valore, e così via. Il modello documentale, che descriviamo in questo paragrafo, può essere dunque visto come una struttura complessa nidificata, che descrive relazioni uno a molti.

Uno dei possibili linguaggi per rappresentare il modello documentale è Json. In Json un documento è formato da una chiave (rappresentata come una stringa) e un valore. Ogni valore può essere un tipo semplice ovvero un array di valori, cioè una colonna di n valori, associati alle posizioni 1, 2, .., ovvero un oggetto strutturato, cioè un insieme di coppie chiave-valore. Tutte queste possibilità modellistiche consentono un esteso utilizzo del modello documentale in molti ambiti applicativi.

Il modello documentale è dunque almeno dal punto di vista concettuale, più ricco di quello relazionale, tuttavia dal punto di vista tecnico la sua diffusione seppure vasta non è ancora paragonabile a quella del modello relazionale, per la mancanza di standard e per la presenza di applicazioni software e soluzioni tecnologiche cosiddette legacy, cioè ereditate dal passato, che difficilmente possono essere abbandonate dalle aziende.

Vediamo ora due diverse soluzioni modellistiche per rappresentare dati logicamente in relazione tra loro, chiamate referencing e embedding. Si assuma di voler modellare un sistema per la gestione dei biglietti da visita dei dipendenti di una azienda. La relazione che lega un dipendente ai suoi indirizzi, ad esempio l'indirizzo della sede della azienda e l'indirizzo di casa, ha una natura uno a molti, un dipendente può avere molti indirizzi associati. Nel modello relazionale e anche nel modello documentale si può modellare la relazione mediante due tabelle, una per il dipendente e una per gli indirizzi, come mostrato in Figura 5. Questa soluzione modellistica è chiamata referencing.

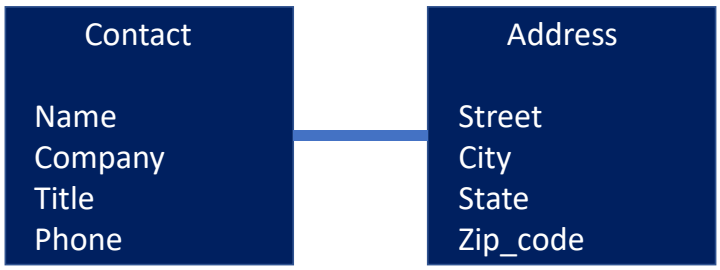


Figura 5 – Dipendenti e indirizzi mediante referencing

Figura 6 – Dipendenti e indirizzi mediante embedding

È possibile modellare lo stesso insieme di dati attraverso embedding. In questo caso, vedi Figura 6, è possibile inglobare le informazioni sugli indirizzi come una ulteriore specificazione delle informazioni del cliente. Va notato come sia possibile anche in questo caso che un cliente abbia diversi indirizzi a cui possa essere raggiunto; questo tipo di soluzione consente un più semplice inserimento dei dati, in quanto tutta l'informazione è riportata in un unico documento che a sua volta contiene uno o più documenti. Anche il ritrovamento dell'intera informazione è semplificato, in quanto, una volta individuato il contatto da cercare, tutta la informazione collegata logicamente si trova in un unico documento.

È peraltro evidente che la soluzione mediante embedding provoca una replicazione nelle informazioni dell'indirizzo, poiché l'indirizzo della azienda sarà replicato per tutti i dipendenti. Non va tuttavia dimenticato che i costi della memoria si sono ridotti significativamente negli ultimi 50 anni (un gigabyte "affittato" su una piattaforma cloud può costare qualche centesimo di euro l'anno). Resta invece ancora critico in questa soluzione dal punto di vista applicativo il problema della corretta gestione delle modifiche, ad esempio, dell'indirizzo della azienda, che dovranno essere effettuate su tutte le sue copie.

La scelta tra le due rappresentazioni dipende dal cosiddetto carico applicativo; se è più frequente cercare il contatto e poi, nel caso, l'indirizzo, allora la soluzione embedding sarà quella preferita. Se invece vi sono frequenti accessi direttamente agli indirizzi, sarà preferibile la soluzione referencing. In conclusione, le due soluzioni in Figura 5 e Figura 6 sono equivalenti come potere descrittivo, tuttavia possono dar luogo a prestazioni molto diverse in funzione del carico applicativo. I prodotti più popolari che utilizzano il modello documentale sono MongoDB e CouchDB.

2.4 Modelli a grafo

I modelli a grafo sono stati introdotti nel Capitolo 3, mettendo in enfasi in particolare la loro capacità di rappresentare nel Web insiemi di dati che possono essere condivisi e collegati; li riprendiamo qui per confrontarli con i modelli introdotti in precedenza nel capitolo. Un ulteriore approfondimento sarà fatto nel Capitolo 7. Dal punto di vista modellistico, un grafo è formato da nodi che rappresentano concetti del mondo reale e da relazioni che li collegano. Va notato come la teoria dei grafi sia molto antecedente l'apparizione dell'Informatica, e tutti i contributi teorici e metodologici relativi alla teoria dei grafi possono portare enormi interessi applicativi, e non solo nel campo delle basi di dati a grafo. Numerosi sono i possibili campi di applicazione delle basi di dati a grafo: i social network, i sistemi informativi geografici, i sistemi per la logistica, la bioinformatica sono solo alcuni degli esempi possibili di applicazioni.

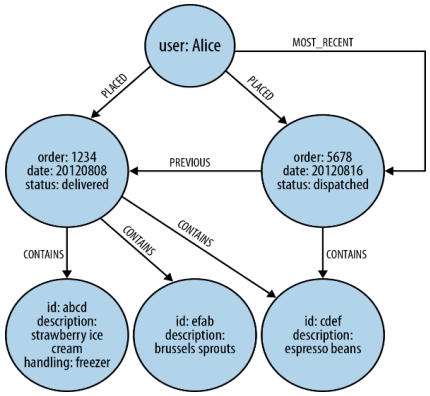


Figura 7 – Un esempio di grafo con tipi e proprietà

È possibile associare ai nodi e agli archi di un grafo dei tipi e proprietà come mostrato nel Capitolo 3 e ripreso nella Figura 7. Anche nei modelli a grafo, come nei precedenti modelli descritti in questo capitolo, ogni singolo nodo è indipendente dagli altri e può avere attributi diversi.

Di fatto nel modello a grafo, così come negli altri, scompare la tradizionale divisione fra descrizione intensionale, lo schema, ed estensionale, i valori, presente nel modello relazionale. Inoltre, in un grafo come quello riportato in Figura 7, il collegamento fra concetti è realizzato dagli archi, questa caratteristica rende il modello a grafo facilmente visualizzabile, anche se all'aumentare del numero di oggetto esistono significativi problemi di visualizzazione di grafi fortemente connessi. Affronteremo questi aspetti nel Capitolo 12 sulle astrazioni, in cui mostreremo come si possa rappresentare un grafo a diversi livelli di dettaglio.

Cosi come per gli altri modelli NoSQL, e come abbiamo già osservato nel Capitolo 3 quando abbiamo introdotto i property graph, non esiste per i grafi un unico modello di interrogazione. Ogni modello a grafo ha il suo proprio linguaggio di interrogazione, tuttavia negli ultimi anni stanno emergendo due linguaggi che possono diventare standard condivisi. Il primo è il linguaggio Cypher ideato dal Neo4J, produttore dell'omonimo modello a grafo; si tratta di un linguaggio concettualmente vicino al linguaggio SQL. Il linguaggio Gremlin proposto da Facebook ha un approccio diverso; il linguaggio esprime interrogazioni sul grafo per via esplorativa, immaginando di accedere a tutti i nodi, percorrendo per ciascuno i cammini costituiti da archi che soddisfano determinate condizioni, e così via fino al completamento dell'interrogazione.

2.5 Confronto fra modelli

Dopo aver presentato i vari modelli NoSQL, è necessario comprendere in quali contesti applicativi una soluzione sia preferibile ad un'altra. Come è naturale, non esiste una soluzione che vada bene sempre; per lo stesso problema applicativo, contesti diversi in termini, ad esempio, di volume (vedi Figura 8) di dati, possono richiedere soluzioni diverse.

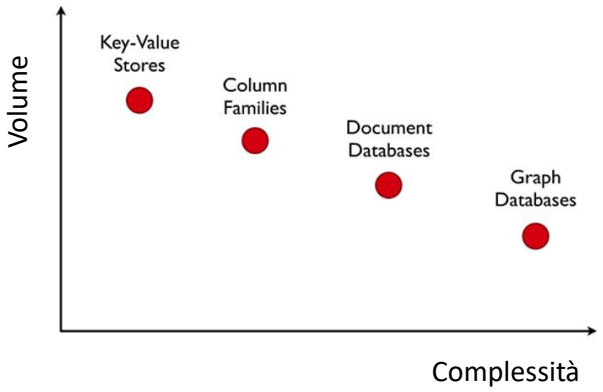


Figura 8 – I modello NoSQL confrontati in termini di volume dei dati e complessità

Nella Figura 8 sono riportati i modelli NoSql illustrati in questa sezione e il posizionamento fra:

- la dimensione dei dati che i sistemi di gestione basati sui modelli sono in grado di gestire e
- il livello di complessità del mondo reale che il modello può rappresentare.

Appare evidente come il modello chiave-valore sia il più semplice, e, allo stesso tempo, anche quello in grado di gestire una quantità di dati maggiore degli altri. Dalla parte opposta, abbiamo il modello a grafo, che è in grado di modellare realtà molto complesse ma, anche a causa di tale complessità non è in grado di gestire efficacemente quantità di dati paragonabili a quelle gestibili con tecnologie chiave-valore.

Al giorno di oggi non è infrequente usare diversi modelli dati per lo stesso contesto applicativo, in quanto le necessità applicative necessitano spesso di modelli diversi. Nella Figura 9 rappresentiamo lo stesso frammento di basi di dati su attori, registi e film nel modello relazionale e in un modello a grafo. Appare immediata la differenza nella leggibilità dei dati descritti, intesa come capacità dei dati di esprimere il significato, senza necessità di documentazione aggiuntiva, così come è semplice verificare quale sia il modello più adatto per capire in quali film l'attore Tom Hanks sia stato insieme regista e attore.

ID	Name	Surname	DateofBirth
1	Tom	Hanks	...

Movie	Actor
1	1
2	1
3	1

Id	Title	Director
1	The Da Vinci Code	2
2	The Green Mile	3
3	That thing you do	1
..		

```
{ "Name":"Tom",
  "Surname":"Hanks",
  "Works_on": [
        {"Title":"The Da Vinci Code",
          "role":"Actor"},
        {"Title": "That thing you do",
          "role":["Actor","Director"]}
        ]
}
```

Figura 9 – Diversa leggibilità di dati nel modello relazionale e nel modello a grafo
(tratta da https://database.guide/what-is-a-graph-database/)

6. Architetture di dati distribuite

Come detto più volte, nei primi anni 2000 si è assistito ad una crescita impressionante della quantità di dati disponibili in formato digitale. Da un lato le tecnologie della comunicazione hanno reso possibile il trasferimento di dati in maniera sempre più veloce, dall'altro le tecnologie hardware hanno reso possibile lo sviluppo di sensori in grado di acquisire dati in tempo reale descrittivi di svariati fenomeni naturali. Inoltre, i social network e i contenuti generati dagli utenti hanno consentito a chiunque di diventare un produttore di dati.

La necessità di rappresentare e gestire dati in una rete distribuita di sensori o nodi elaborativi era già presente nei sistemi relazionali e, di fatto, i sistemi NoSQL si sono avvalsi di questa esperienza pluridecennale per sviluppare architetture distribuite basate sui nuovi modelli. Esistono tuttavia alcune importanti differenze progettuali che rendono i sistemi NoSQL significativamente diversi da quelli relazionali.

I sistemi relazionali su architetture distribuite sono stati progettati per garantire rigidamente le proprietà cosiddette ACID. Con questo acronimo si fa riferimento ad un insieme di caratteristiche che un sistema software per la gestione dei dati relazionali deve garantire. Le proprietà impongono che un sistema di gestione di basi di dati relazionale nell'eseguire una transazione, cioè un insieme di operazioni di scrittura e lettura su una base di dati relazionale, debba garantire che:

- anche in presenza di guasti, la transazione sia eseguita tutta o per niente (Atomicità, pensate cosa succederebbe se di un bonifico fosse solo eseguito il prelievo e non l'accredito…)
- la transazione deve lasciare il sistema in uno stato consistente con quello di partenza (Consistenza, ad esempio in una operazione di bonifico la somma dei saldi dei due conti coinvolti deve rimanere la stessa)
- la transazione deve essere eseguita come se non ci fosse nessuna altra transazione eseguita in quello stesso momento sulle stesse risorse (Isolamento, se vengono eseguite due prenotazioni sull'ultimo posto libero in un treno, il posto va assegnato all'una o all'altra, senza intrecci tra le transazioni per cui il posto viene assegnato ad entrambe).
- i risultati della transazione devono essere memorizzati permanentemente nel sistema per tutta la vita dell'applicazione (Durabilità, non mi devo preoccupare che dopo aver prenotato un posto su un treno, quel posto venga in futuro assegnato ad altra transazione).

Il rispetto delle proprietà ACID ha reso i sistemi software per la gestione dei dati relazionali molto popolari e di riconosciuta affidabilità, e anche oggi sono estremamente utili in molti contesti. Tuttavia, il rispetto delle proprietà è andato a discapito dell'efficienza, per le tante attività che devono essere eseguite dal sistema software di gestione per garantirle; per fare solo un esempio, per garantire la atomicità in presenza di guasti, occorre ricordare tutta la sequenza di operazioni elementari eseguite, così che sia possibile quando si verifica il guasto annullarle o rieseguirle come un filo di Arianna.

Le nuove applicazioni nate con Internet e con il Web non sempre richiedono caratteristiche così stringenti. Si pensi ad un social network dove la garanzia che tutte le persone che seguono un determinato utente devono vedere subito ogni nuovo contenuto pubblicato è sicuramente eccessiva, mentre, come abbiamo visto nel caso di una operazione bancaria, l'atomicità è fondamentale perché il cliente si fidi della banca e tenga i suoi soldi in un conto corrente.

L'enorme volume di utenti e operazioni che ogni giorno si svolgono su Internet ha anche messo in evidenza che le soluzioni fino a quel momento utilizzate per gestire grandi carichi di lavoro non erano più sostenibili. Fino alla fine degli anni 90, l'idea principale per "scalare" nelle capacità di elaborazione, essere cioè in grado di mantenere il sistema efficiente in presenza di carichi sempre maggiori, era quella di usare server sempre più potenti. Tale misura non è sostenibile nell'era di Internet, perché da un lato il numero di utenti è in continua costante crescita e dall'altro le soluzioni non sono funzionali per carichi di lavoro ridotti. Si pensi ad esempio ad un sito di commercio elettronico che ha dei picchi stagionali significativi, mentre nel resto dell'anno ha flussi di accessi decisamente più bassi dei picchi stagionali. Alla scalabilità verticale della fine degli anni 90 si è sostituita l'idea di scalabilità orizzontale, ovvero l'uso di hardware poco costoso che si può dinamicamente aggiungere o rimuovere dal sistema distribuito quando serve.

La necessità di soddisfare un numero sempre maggiore di utenti e le diverse necessità applicative hanno portato a caratterizzare i sistemi NoSQL con proprietà diverse dalle proprietà ACID, rappresentate nell'acronimo BASE (Basic Available, Soft state, Eventual consistency), che afferma: i sistemi software NoSQL soddisfano una consistenza differita, garantendo che dopo qualche tempo, non specificato in maniera precisa, il sistema sarà consistente, mentre è garantito che sarà sempre disponibile.

Quasi contemporaneamente alla introduzione di questo nuovo insieme di caratteristiche, in occasione del Symposium on Principles of Distributed Computing tenuto nell'anno 2.000, Eric Brewer tenne un famoso keynote sulla sua esperienza nella realizzazione di database distribuiti. In quell'intervento Brewer enunciò quello che è comunemente chiamato CAP theorem. Il teorema afferma che in un sistema distribuito si possono garantire soltanto due delle seguenti tre caratteristiche:

- Consistenza dei dati,
- disponibilità dei dati (Availability) e
- tolleranza al partizionamento della rete Internet (cioè creazione di sottoreti non comunicanti tra loro) dovuto ad un guasto (Partition tolerance).

In base a questo teorema, ogni sistema di basi di dati distribuite deve "scegliere" due delle precedenti tre caratteristiche; ad esempio, i sistemi relazionali sono tutti AC, ovvero garantiscono la consistenza e la disponibilità, mentre non garantiscono il funzionamento del sistema nel caso di partizionamenti nella rete di server utilizzati. Definiti i termini generali che caratterizzano le architetture distribuite nei moderni sistemi distribuiti di basi di dati, nel seguito entriamo nel merito delle caratteristiche e funzionalità principali di tali architetture, mostrando successivamente alcuni esempi di soluzioni realizzate per i più popolari sistemi NoSQL distribuiti.

3.1 Architetture di distribuzione dei dati

È possibile classificare le architetture di distribuzione dei dati in base al livello di condivisione delle risorse hardware che i vari sistemi propongono.

Nella Figura 10 sono schematizzate le tre tipologie di condivisione possibili. La prima modalità denominata shared memory (memoria condivisa) o shared everything prevede che ogni applicazione software o processo di elaborazione (P) condivida con gli altri sia la memoria centrale che i dischi dove sono memorizzati i dati in modo persistente. Questo tipo di distribuzione prevede che i processi debbano essere eseguiti sullo stesso elaboratore per poter condividere l'area di memoria; tale modalità consente la distribuzione del carico di interrogazioni e transazioni che vengono eseguite dai processi. Questi sistemi sono tipici nelle architetture che prevedono una scalabilità verticale: all'aumentare del carico applicativo si attivano sulla stessa macchina altri processi di elaborazione, fintantochè l'infrastruttura hardware garantisce prestazioni adeguate.

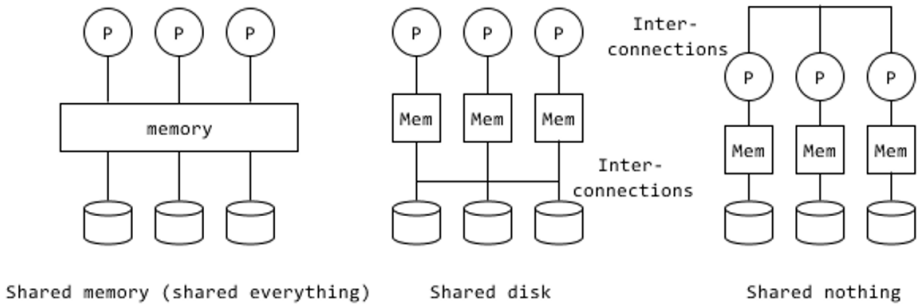


Figura 10 – Architetture di distribuzione dei dati
(tratta da https://pages.cs.wisc.edu/~zuyu/summaries/cs764/parallelDB)

La seconda modalità prevede di far condividere fra diversi elaboratori i dischi dove sono memorizzati i dati in maniera persistente. Questa architettura, implementata ad esempio da Oracle nella suite RAC (Real Application Cluster), porta ad una alta complessità nella gestione delle transazioni concorrenti appartenenti a più processi, per evitare la violazione della proprietà di isolamento delle transazioni. Inoltre, è necessaria una infrastruttura di rete ad alte prestazioni fra i dischi e i server di elaborazione, in quanto i tempi di trasmissione sulla rete Internet sono molto più lenti rispetto ai tempi di trasmissione su un canale che collega direttamente la memoria a dischi con la unità di calcolo.

L'ultima tipologia di condivisione denominata shared nothing (cioè, non è condivisa nessuna risorsa) è tipica delle soluzioni a scalabilità orizzontale. In questa architettura ogni server è indipendente dagli altri server e il coordinamento tra le transazioni non è gestito centralmente dal software di gestione, ma è demandato a livello della applicazione. Va notato che per i sistemi NoSQL, dove, come detto, i vincoli transazionali non sono stringenti come nelle soluzioni ACID, questa architettura è molto efficace, in quanto è possibile aumentare il numero di server coinvolti in maniera dinamica all'aumentare del carico applicativo.

3.2 Distribuzione dati nei sistemi NoSQL

Di seguito si riportano a titolo di esempio tre architetture dati di sistemi NoSQL largamente diffuse a testimonianza delle diverse soluzioni tecnologiche di gestione dei dati.

MongoDB, o come più comunemente noto, Mongo, è un database documentale orientato alla consistenza e alla gestione del partizionamento della rete; Mongo è uno dei database NoSQL più conosciuti e affermati. Il modello documentale ben si presta a numerosi ambiti applicativi, anche non in presenza di grandi volumi di dati. A partire dalla versione 4.0 di Mongo, rilasciata nell'estate del 2018, è presente anche il supporto alle transazioni ACID.

L'architettura di distribuzione dei dati di Mongo è basata su una architettura shared nothing, gestita con una soluzione cosiddetta master slave, che ora esemplifichiamo. Il dataset consistente in una collezione di documenti può essere distribuito in più frammenti, assegnati ciascuno a una diversa risorsa elaborativa, detta slave. Ogni risorsa slave gestisce il frammento come se fosse una collezione di documenti indipendente dalle altre.

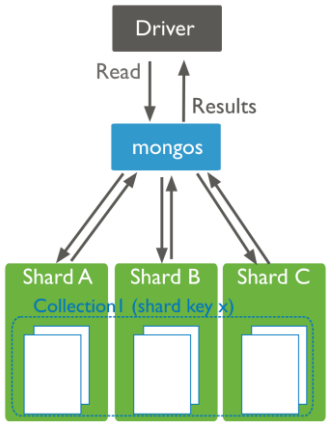


Figura 11 – Architettura dati ed esecuzione delle operazioni in Mongo
(tratta da https://docs.mongodb.com/manual/core/sharded-cluster-query-router/)

Nel caso di una interrogazione che coinvolge informazioni su più frammenti, il processo master, denominato MongoS, si occupa di ricevere le interrogazioni, distribuendo le stesse ai vari processi distribuiti, ognuno dei quali esegue l'interrogazione singolarmente, vedi Figura 11. I dataset locali di risposta alla interrogazione sono raccolte da MongoS che provvede a ricomporli e a produrli in output.

I frammenti di dati possono inoltre essere duplicati su più nodi; nel caso in cui si voglia interrogare un unico frammento locale, è possibile scegliere se interrogare il master, seguendo la precedente sequenza di azioni, oppure la copia del frammento più vicina nella rete al richiedente, per aumentare le prestazioni a scapito di eventuali problemi di consistenza nei dati (ricordiamo che Mongo supporta le proprietà BASE delle transazioni). L'inserimento e la modifica dei dati (cioè le operazioni di scrittura) effettuate dalle transazioni avvengono unicamente a partire dal processo master.

Il sistema HBase è l'evoluzione open source (cioè software libero e gratuito) del sistema wide column sviluppato da Google; entrambi garantiscono la consistenza e il supporto al partizionamento della rete. Anche HBase utilizza una architettura master slave per la distribuzione dei dati.

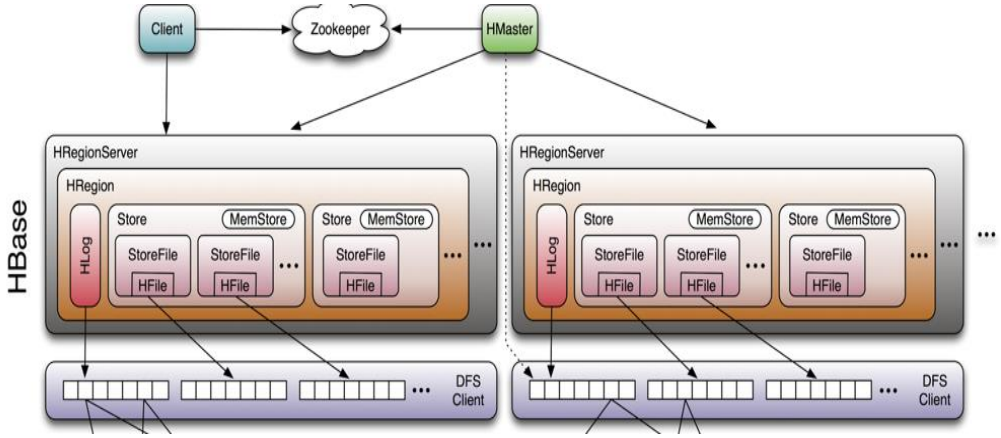


Figura 12 – Architettura dati di Hbase
(tratta da http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html(

Nella Figura 12 è mostrata l'architettura di Hbase. A parte una terminologia diversa, le funzionalità sono molto simili a quelle già presentate per MongoDB. Il client[2] (lo slave, nella terminologia precedente) interroga il master, che ha accesso ai dati memorizzati in componenti chiamati HRegionServer attraverso Zookeeper, un componente open source che si occupa di gestire architetture dati distribuite.

Ogni HRegionServer è un componente autonomo che contiene tutte le tipiche risorse e funzionalità di un sistema di gestione di basi di dati, cioè memoria centrale, file di log che memorizza tutte le operazioni elementari eseguite nella base di dati per garantire la atomicità, metodi di accesso ai dati su disco etc. In HBase i dati sono memorizzati su un file system distribuito denominato HDFS (Hadoop Distributed File Systems) che verrà illustrato nella prossima sezione.

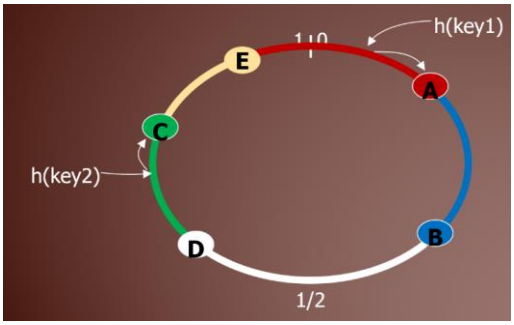


Figura 13 – Spazio di indirizzamento in Cassandra

[2] Assumiamo equivalenti le terminologie "client server" e "master slave", anche se realizzate nei prodotti dei fornitori con alcune differenze, che sono marginali negli argomenti svolti nel capitolo.

Una soluzione completamente diversa di distribuzione dei dati è proposta da Cassandra, sviluppato da Facebook. Pur essendo un database wide column come Hbase, Cassandra si caratterizza per garantire nella applicazione del CAP theorem la disponibilità e la tolleranza al partizionamento rispetto alla consistenza dei dati. Ciò non deve stupire, in quanto le applicazioni dei social network hanno l'esigenza di garantire un funzionamento continuo ("24x7") a scapito, come abbiamo già osservato, della consistenza dei dati. L'architettura di distribuzione dei dati prevede di distribuire lo spazio degli indirizzi delle chiavi sulle memorie di n nodi, formando topologicamente un anello.

Nell'esempio in Figura 13, lo spazio di indirizzamento delle chiavi viene gestito attraverso una opportuna funzione h(ash), così chiamata perché fa corrispondere a ogni valore key della chiave un indirizzo h(key) in memoria che viene calcolato efficientemente per mezzo della funzione hash; lo spazio di indirizzamento è diviso su cinque nodi, la chiave "key1" è assegnata al nodo A mentre la chiave "key2" è assegnata al nodo C, ogni volta che un valore viene scritto si procede alla scrittura anche nei nodi successivi, creando in tal modo una replicazione dei valori su più nodi per garantire la disponibilità.

4. Architettura Hadoop

Abbiamo visto che l'incremento del volume dei dati che possono essere generati in numerose applicazioni richiede un cambio radicale nel paradigma di gestione ed elaborazione dei dati. Fino agli inizi degli anni 2000, le applicazioni software assumevano che i dati risiedessero su memoria secondaria, e nel corso della elaborazione delle applicazioni fossero trasferiti dalla memoria secondaria ai nodi di elaborazione.

A causa dei volumi e della latenza della rete Internet, dove sono sempre più spesso eseguite applicazioni software, questo paradigma non è più funzionale alle nuove esigenze elaborative; di conseguenza, è stato sviluppato da Google, una delle aziende che tra le prime hanno sperimentato le criticità derivanti dal volume dei dati da gestire, una modalità di elaborazione dati completamente diversa.

La nuova modalità di elaborazione distribuita prevede non più il trasferimento dei dati verso i nodi di elaborazione, ma il viceversa, lo spostamento della capacità computazionale verso le risorse dove sono memorizzati i dati, ribaltando così il precedente paradigma. Per ottenere questo risultato sono disponibili almeno tre componenti architetturali:

- un file system, cioè un sistema di gestione dei file (o dataset) di dati in grado di distribuire i dati su più nodi di elaborazione.
- un motore di calcolo in grado di distribuire le elaborazioni fra nodi diversi.
- un sistema di gestione dell'intera infrastruttura.

Questi componenti architettural costituiscono la base della piattaforma Hadoop3, sviluppata originariamente da Google e poi rilasciata in formato open source alla comunità di sviluppatori. I tre componenti sopra menzionati sono rispettivamente chiamati, Hadoop Distributed File System (HDFS), Map-Reduce e YARN.

[3] Il nome è stato scelto da Doug Cutting responsabile del progetto di sviluppo usando il nome dell'elefante giallo di pezza di suo figlio.

4.1 Hadoop Distributed File System

HDFS è l'evoluzione di un progetto analogo sviluppato da Google denominato GDFS (Google Distributed File System), riguardante la costruzione di un file system distribuito; il sistema è stato pensato per funzionare con hardware poco costoso in grado di scalare orizzontalmente. Alla base del funzionamento di HDFS vi è l'idea di dividere i file, di solito di enorme dimensione, in frammenti (chunk) più piccoli e di memorizzare ogni chunk in un nodo. Per migliorare la disponibilità dei dati, ogni singolo chunk è replicato in almeno altri due nodi, seguendo così le buone pratiche di gestione dei dati.

Lo schema architetturale di HDFS è mostrato nella Figura 14. Ogni frammento di file è memorizzato in un nodo della rete di elaboratori, gestito da un componente software denominato HDFS datanode. Per gestire l'insieme dei datanode esiste un secondo componente denominato HDFS namenode, secondo una classica architettura client server. Il namenode contiene in memoria centrale la visione logica del file system compresa l'alberatura delle directory nei vari nodi, e per ogni file l'elenco dei datanode che gestiscono i frammenti del file.

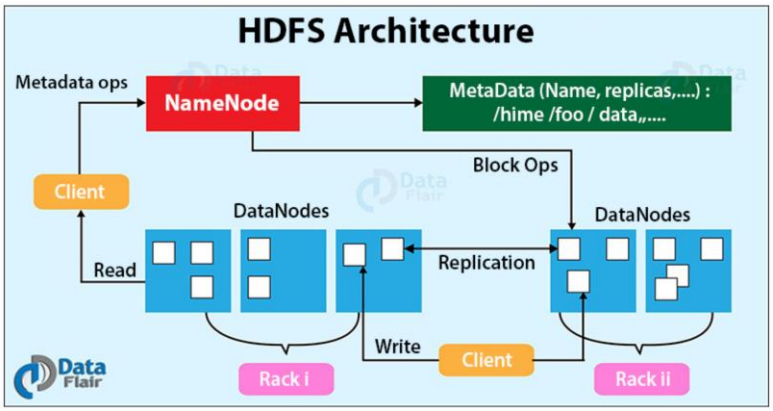


Figura 14 – Schema architetturale di HDFS
*tratto da https://data-flair.training/blogs/hadoop-hdfs-architecture/(

Ogni volta che un client vuole accedere a un file o a una sua porzione, effettua una richiesta al namemode che restituisce al client l'elenco dei datanode coinvolti. Il client può successivamente effettuare direttamente ai datanode le richieste di lettura o scrittura di dati. Come si può notare, l'architettura è relativamente semplice e, come tutte le architetture client server, il namenode è l'elemento più significativo dell'architettura.

Per evitare che una perdita del namenode renda il sistema inusabile, è prevista un namemode secondario che può intervenire se il namenode primario non è più disponibile. Il namenode esegue un insieme di funzionalità per garantire il corretto funzionamento dell'infrastruttura HDFS, come ad esempio la verifica periodica dello stato dei datanode, attraverso la ricezione di messaggi detti heartbeat, l'eventuale copia di frammenti (chunk) nel caso di malfunzionamento di un singolo datanode, e così via.

È importante segnalare che HDFS è stato pensato soprattutto per file di grandi dimensioni che crescono continuativamente, come ad esempio i file di log delle applicazioni, che abbiamo visto nel Paragrafo 3.2. Originariamente HDFS è stato pensato per applicazioni che prevedono prevalentemente operazioni in sola lettura, in cui le operazioni di scrittura consistono unicamente di inserimenti e non di modifiche, e in modalità batch, consistenti cioè di operazioni che possono essere effettuate in sequenza e che non sono caratterizzate da esigenze di tempo reale, cioè di esecuzione in tempi molto brevi. Nel corso degli anni, HDFS è stato utilizzato anche per file di dimensioni contenute e con un alto tasso di aggiornamenti anche di modifica, mettendo così in luce i limiti di questa architettura.

4.2 Map Reduce

Map Reduce è un motore di esecuzione di operazioni distribuite. L'idea alla base di questo paradigma è relativamente semplice ed è particolarmente efficace quando il problema consente di eseguire inizialmente in parallelo, cioè contemporaneamente, operazioni locali sui dati, e successivamente di aggregare i risultati della ricerca.

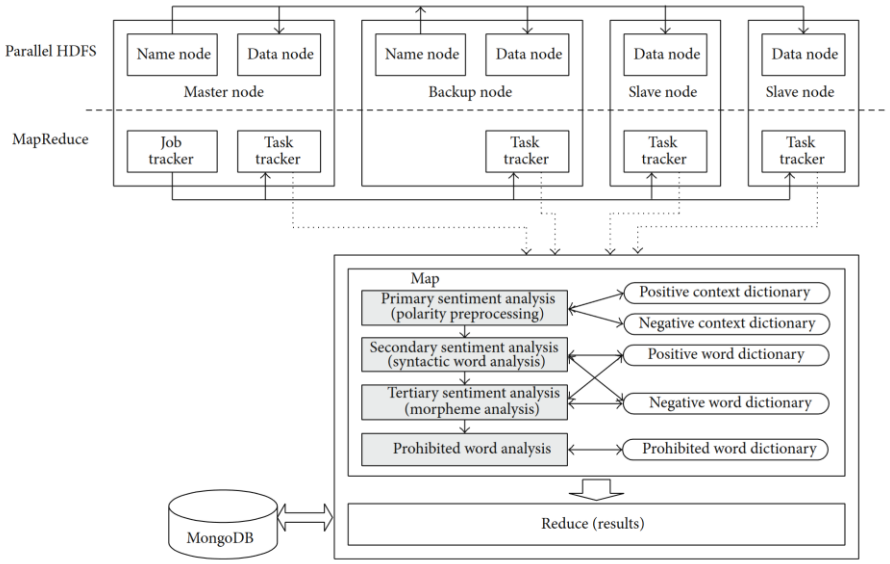


Figura 15 – Flusso delle attività in Map Reduce (tratta da Ha et al. 2015)

Si consideri ad esempio il problema consistente nell'effettuare una sentiment analysis di messaggi di un social network, in cui si vuole cioè determinare la percentuale di messaggi positivi, negativi o neutri rispetto ad un determinato tema o evento. L'analisi può essere suddivisa in due fasi. È necessario prima valutare per ogni messaggio se esso ha un sentimento positivo e neutro; ciò può essere fatto con tecniche anche raffinate, ciò che interessa ai fini del calcolo è che l'analisi riguarda i singoli messaggi, indipendentemente l'uno dall'altro. Una volta che tutti i messaggi sono stati così elaborati, è possibile contare i messaggi classificati come positivi, negativi o neutri. Map Reduce abilita questa analisi scomponendo il problema in due fasi (si veda la Figura 15).

Nella fase di Map (o Partiton) il problema del calcolo del sentiment viene partizionato in modo che ogni nodo locale possa elaborare un sottoinsieme dei messaggi da analizzare e restituire il numero dei messaggi positivi, negativi o neutri che ha elaborato. Questa attività può facilmente essere distribuita su più nodi (detti worker) ognuno dei quali può eseguire lo stesso algoritmo su porzioni di dati diversi. Una volta terminata questa fase si può procedere alla fase di Reduce (o Combine). In questa seconda fase i dati prodotti dai risultati della fase di Map sono aggregati (ad esempio ordinati, sommati etc.) e restituiti.

La libreria di programmi di gestione di Map Reduce, che è disponibile nella soluzione Hadoop per diversi ambienti e linguaggi di programmazione, si occupa di tutti i problemi infrastrutturali, quali ad esempio identificare i nodi liberi dove poter svolgere le elaborazioni, monitorare e gestire i processi di elaborazione, identificare eventuali nodi non più disponibili, è cosi via, in modo da lasciare al programmatore la scrittura delle sole funzioni di Map e Reduce. La libreria si occupa di spostare l'esecuzione delle funzioni sui nodi dove sono memorizzati i dati, realizzando così quel cambiamento di paradigma di calcolo indicato all'inizio della sezione. Map Reduce insieme con HDFS consentono dunque di eseguire calcoli anche complessi su un insieme molto grande di dati, soprattutto se il tipo di problema non prevede, come abbiamo discusso, una interdipendenza fra i dati.

Nonostante la potenza del motore Map Reduce, risolvere un problema individuando le fasi di Map e Reduce e la loro implementazione può essere molto complesso e a rischio di errori. Per tale motivi sono disponibili sistemi di analisi più semplici come ad esempio Pig e il suo linguaggio chiamato Pig Latin, che astraggono dalla difficoltà tecniche traducendo i programmi scritti nel linguaggo Pig Latin in una o più funzionalità da "mappare" su Map Reduce. Di fatto, al giorno di oggi i programmatori tendono a non scrivere direttamente in Map Reduce; usano piuttosto strumenti di più alto livello che successivamente traducono le proprie istruzioni in funzionalità Map Reduce. Ad esempio, è possibile interrogare una base di dati HBase con un linguaggio di interrogazione simile a SQL, la cui esecuzione non è altro che una serie di funzionalità Map Reduce.

4.3 Yet Another Resource Negotiation

Come visto per HDFS e per Map Reduce, la complessità dell'insieme di funzionalità software che gestiscono le varie componenti di una rete di nodi elaborativi è molto elevata. A queste complessità se ne aggiungono altre, relative alla topologia della rete e ai vari carichi applicativi che ogni nodo deve eseguire.

A partire dalla versione 2 di Hadoop è stato sviluppato un componente aggiuntivo in grado di gestire in maniera efficace le risorse disponibili, separando questo compito dalle funzionalità svolte dalle altre librerie per Map Reduce. Yet Another Resource Negotiator (YARN), come dice il suo nome, è un negoziatore di richieste in ambiente distribuito. Il suo compito fondamentale è decidere quali nodi allocare ai processi che richiedono di eseguire delle attività. Nella Figura 16 si mostrano i componenti fondamentali dell'architettura funzionale Yarn.

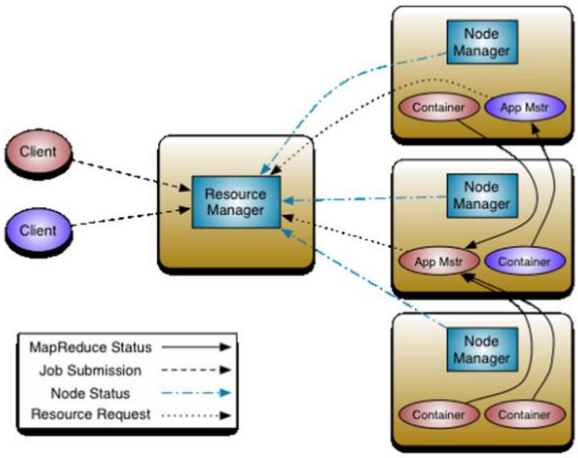


Figura 16 – Componenti della architettura funzionale Yarn
(tratto da https://www.slideshare.net/GabrieleLombari/hadoop-in-action)

Il cuore del sistema è il modulo di Resource manager (Gestore delle risorse), che dialoga costantemente con i Node manager installati sui nodi elaborativi slave dell'infrastruttura; in tal modo, il Resource manager ha sempre lo stato aggiornato dell'uso delle risorse. Quando un client, per esempio un gestore di attività map reduce, chiede risorse, effettua una richiesta al Resource manager, che controlla le disponibilità e autorizza l'installazione di un Application master (il gestore di una singola attività Map-Reduce) e dei Container (cioè, le porzioni di memoria assegnate alle singole attività Map Reduce).

5. Conclusioni

L'evoluzione delle tecnologie informatiche per memorie di dati consente di poter immagazzinare quantità sempre più grandi di dati anche di tipologie diverse. Le tecnologie relazionali che negli ultimi 40 anni erano state le protagoniste assolute nel settore della gestione dati stanno lasciando il passo a nuovi modelli di dati e nuove soluzioni tecnologiche. Tali cambiamenti sono legati alle nuove esigenze applicative, sia per applicazioni tradizionali sia per nuove applicazioni nate con l'avvento di Internet.

I nuovi modelli e i relativi sistemi di gestione riprendono quanto già investigato nella ricerca scientifica, con soluzioni tecnologiche moderne che consentono di rendere questi modelli utilizzabili anche per le sfide di oggi e future. Questo ampliarsi dello spazio delle possibili soluzioni tecnologiche per i problemi di gestione dati offre grande libertà di scelta ai progettisti di architetture dati; la loro semplicità d'uso rende l'utilizzo di queste nuove tecnologie accessibili a moltissime persone.

Va sottolineato come spesso le soluzioni NoSQL siano state promosse dai giganti del Web come Google, Facebook, Amazon etc. e che queste aziende non vendono tecnologie o infrastrutture hardware ma, in alcuni casi, offrono servizi basati su tali tecnologie. È importante sottolineare questo aspetto, in quanto se nel 1970 l'IBM sviluppava il modello relazionale per poi offrire prodotti proprietari sul mercato a titolo oneroso, al giorno di oggi le aziende rilasciano il software di gestione secondo il paradigma open

source, preoccupandosi di continuare a manutenerlo. Le motivazioni di questa scelta strategica, che può sembrare controintuitiva (come se i progettisti della scuderia Ferrari rendessero pubblici e usabili da chiunque i disegni dei motori delle monoposto di Formula 1) si può spiegare considerando due aspetti:

- il rilascio di software in formato open source ne abilita la diffusione e consente più velocemente a molti di individuare eventuali errori nel software, arrivando più rapidamente alla ingegnerizzazione del prodotto software.
- La spesa dei clienti e quindi i ricavi da parte delle aziende si riorientano verso i servizi di gestione e manutenzione, e verso lo sviluppo di nuove applicazioni.

Le tecnologie per la gestione di grandi quantità di dati con formati eterogenei e dei relativi modelli non sono ancora giunte ad un punto di maturazione, e anzi sono in continua evoluzione, questo per almeno due aspetti. Il primo è relativo alla continua ricerca e innovazione nel settore, che si integra sempre di più con le soluzioni applicative di analisi dati, il secondo è la continua evoluzione del linguaggio SQL e del software di gestione per basi di dati relazionali "tradizionali", per essere in grado di competere con i nuovi sistemi descritti in questo capitolo. Infine, sembra delinearsi per i nuovi prodotti lo sviluppo di sistemi di gestione ibridi, ovvero sistemi in grado di modellare dati secondo paradigmi diversi (per esempio documentale e a grafo).

Anche il settore industriale della gestione dati è in continua evoluzione; nel settembre del 2018 è stata annunciata la fusione delle due aziende Cloudera e Hortnworks che sviluppavano e gestivano piattaforme Hadoop, segno di una maturazione del mercato big data. Le soluzioni cloud dei principali fornitori di tecnologie dati si stanno orientando verso la costruzione di ecosistemi digitali per i quali è semplice rendere interoperabili, cioè far convivere, applicazioni diverse dello stesso provider ovvero integrare soluzioni di provider diversi.

A una grande disponibilità di soluzioni corrisponde una grande responsabilità da parte dei progettisti di sistemi informativi che utilizzano big data, in quanto è necessario scegliere con attenzione la migliore soluzione possibile rispetto al contesto applicativo in cui si agisce. Va sottolineato che le nuove esigenze di scalabilità, accennate in questo capitolo, così come le soluzioni cloud oggi esistenti e gli ecosistemi digitali che si stanno costituendo attorno ai grandi fornitori di tecnologie, impongono una attenta e non banale analisi su come risolvere un determinato problema applicativo che utilizzi grandi quantità di dati. E' anche cruciale la possibilità di utilizzare per lo stesso problema applicativo diversi paradigmi, passando da un modello ad un altro in maniera flessibile e veloce.

Riferimenti

E.F. Codd - A Relational Model of Data for Large Shared Data Banks.Communication of ACM 13,6 pp. 377-387, 1970

T. Brewer - Towards Robust Distributed System, PODC, 2000

I. Ha, B. Back, B. Ahn - Map Reduce Functions to Analyze Sentiment Informationfrom Social Big Data" Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2015.

G. Harrison / Next Generation Databases, Apress, 2015

A. Rezzani - Big Data Analytics, Apogeo 2017.

Capitolo 5 – La qualità dei dati e la grande sfera opaca

Carlo Batini

1. Introduzione

Quando mi capita di lavorare la domenica pomeriggio, per distrarmi guardo i risultati delle partite di calcio. Un pomeriggio del campionato 2012-13 vidi sul sito di Repubblica l'immagine riprodotta in Figura 1.

Figura 1 – Il risultato (meglio, i risultati) di Catania Inter nel campionato di calcio 2012-13 (dal sito di repubblica del 3 marzo 2013, ore 16 circa)

Nella stessa pagina sono riportati ben tre risultati diversi della partita Catania Inter! Come è possibile? Se guardiamo in alto a destra i marcatori, Palacio sembra aver segnato due volte nel giro di un minuto di recupero, al 47esimo e al 48esimo, e sembra che questo ultimo goal non sia conteggiato nella riga degli aggiornamenti corrispondente al "Finisce la partita! Vince l'Inter 3-2". Se dovessimo decidere noi sulla base delle informazioni presenti, poiché' il risultato 2-4 è coerente con i marcatori, propenderemmo per questo. Ma andando a guardare sul Web il risultato di quella partita, scopriamo che la partita è finita 2-3, non 2-4. Insomma, la concordanza dei diversi siti in cui è riportata quella informazione storica vince sulla informazione rappresentata nella pagina.

Ognuno di noi potrebbe portare tanti esempi di dati sbagliati, o, come diciamo in questo capitolo, dati di scarsa qualità. Nella Figura 2 mostro due ritagli di giornale presi a caso in una settimana del 2015

mentre preparavo un seminario sulla qualità dei dati per studenti delle scuole secondarie. Nel ritaglio a sinistra si vede come la scarsa qualità di un dato utilizzato in aeroporto possa essere dovuta a un valore alterato della umidità.

Inconveniente in aeroporto

Linate, l'umidità fa scattare l'allerta sicurezza

Il volo Milano-Napoli AZ1285 ha fermato il rullaggio in pista su ordine della torre di controllo. Ore 10 di ieri mattina, aeroporto di Linate. Un allarme sicurezza — quasi subito rientrato — ha preoccupato i passeggeri in partenza dallo scalo milanese. Dal velivolo sono partiti telefonate e cinguettii su Twitter. Sulla pista sono entrati i vigili del fuoco e i responsabili delle procedure di emergenza. È stato ispezionato l'aereo: tutto a posto, un falso allarme. Ma la causa del ritardo resta comunque anomala. L'umidità accumulata sulla pista ha «alterato» la sensibilità dei segnali di sicurezza. Il sistema informatico ha percepito — sbagliando — una presenza estranea.

Centinaia di dirigenti e funzionari dell'**Inps** sono al centro di un'indagine della Procura di Nocera Inferiore (Salerno) sull'erogazione di **premi** di produttività nel biennio 2012-2013.

Per raccogliere elementi utili all'inchiesta, carabinieri del reparto territoriale di Nocera Inferiore, con la collaborazione di quelli territorialmente competenti e del Ris di Roma, hanno eseguito a Roma un decreto di perquisizione presso la Direzione generale dell'Inps e in tutti i capoluoghi di regione un ordine di esibizione di documenti presso le direzioni regionali dello stesso istituto. L'Inps, in una nota, ha assicurato la "piena collaborazione con la magistratura e le forze dell'ordine", mentre il presidente dell'Anticorruzione, Raffaele Cantone, ha detto di ritenere che si tratti, come per altre vicende, di uno "spaccato inquietante dello stato della Pubblica amministrazione".

Secondo l'ipotesi investigativa, gli incentivi di produttività sarebbero stati concessi sulla base di ispezioni "gonfiate": nel database dell'Istituto sarebbero stati inseriti dati falsi, maggiorati rispetto a quelli reali, relativamente ai risultati delle ispezioni effettuate nelle aziende e ai rapporti di lavoro annullati. In questo modo sarebbero stati alterati i risultati delle "performance", cioè i parametri degli obiettivi indicati dalla direzione generale per la corresponsione degli incentivi. I dati attuali sono ancora parziali ma solo per gli anni 201[illegible]e

Figura 2 – Esempi di scarsa qualità dei dati presi dai giornali

Nella Figura 2 a destra la scarsa qualità dei dati è dovuta al fatto che, secondo una ipotesi investigativa, sono stati inseriti dei dati "gonfiati" nella base di dati, fornendo risultati distorti sulla produttività del personale.

Figura 3 – Due immagini intuitivamente di diversa qualità

I precedenti esempi sono una conferma che i dati digitali stanno rappresentando sempre nuovi tipi di informazioni utilizzate nel mondo reale, e perciò è fondamentale che, quando vengono usati, non ci si fidi acriticamente della loro qualità, che deve essere sempre controllata e, se inferiore ad una soglia ritenuta accettabile, migliorata.

Ma, quando parliamo di qualità possiamo stabilire delle misure assolute? Ed esiste un solo tipo di qualità? In Figura 3 della pagina precedente vengono mostrate due immagini di un fiore; guardando le due immagini non ci viene in mente una misura assoluta di qualità, ma certamente riusciamo ad affermare che l'immagine sulla destra è di maggiore qualità della immagine a sinistra, e questo perchè la figura a sinistra ci appare "sgranata", composta da tanti quadratini che producono una variazione discontinua dei livelli di grigio. Insomma, in questo caso riusciamo a dare una valutazione relativa della qualità, ma non assoluta.

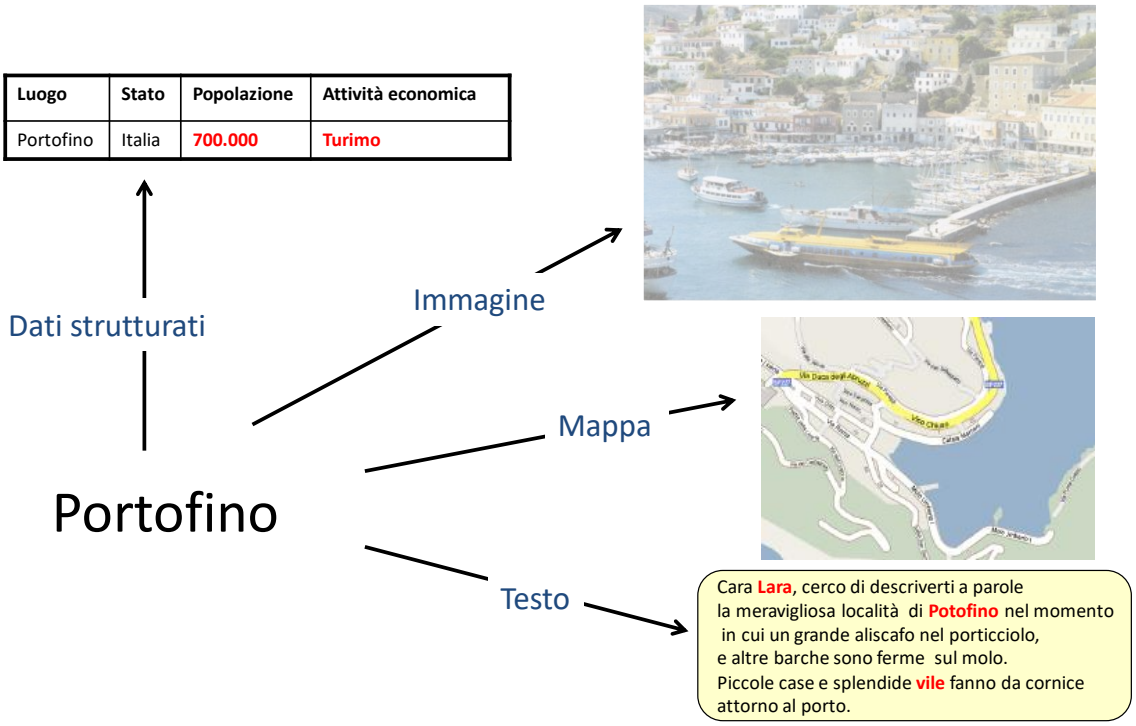


Figura 4 – Come possiamo scoprire gli errori nelle diverse rappresentazioni di Portofino?

E quando si parla di qualità, è un concetto le cui caratteristiche dipendono dal tipo di rappresentazione scelta per i dati, ad es. una tabella, un testo, una mappa, oppure è invariante con il cambiare della rappresentazione? In Figura 4 mostriamo le quattro rappresentazioni di Portofino che abbiamo visto nella Figura 1 del Capitolo 3. Guardando con attenzione, scopriamo che le quattro rappresentazioni relative alla tabella, alla immagine, alla mappa e al messaggio presentano alcune imperfezioni rispetto alle versioni di Figura 1 del Capitolo 3:

- la tabella presenta due palesi errori, nella popolazione (come fa Portofino ad avere 700.000 abitanti, è un piccolo paese di poche case…) e nella attività economica, non può essere "turimo", sarà sicuramente "turismo", è il nome più vicino a "turimo" tra le attività economiche. Volendo provare

a caratterizzare la qualità con una caratteristica che la esprima in maniera più precisa, potremmo parlare in questi casi di dati affetti da *inaccuratezza*.

- L'immagine ci appare opaca, poco contrastata nella scala di grigi, forse la foto è stata ripresa con una di quelle macchine fotografiche che non permettono di impostare il tempo dello scatto, forse la foto è stata troppo esposta, non ho mai visto Portofino avvolto dalla nebbia…). In questo caso possiamo parlare di scarsa *fedeltà* della immagine rispetto all'originale.
- La mappa non è recente, supponiamo di sapere che una mareggiata abbia costretto a chiudere temporaneamente la strada per Santa Margherita Ligure, non c'è traccia della interruzione nella mappa. In questo caso possiamo dire che la mappa non è *aggiornata* rispetto agli eventi più recenti.
- Il testo è stato probabilmente scritto in fretta o con una tastiera difettosa, la persona si chiamava Laura e ora è Lara, Portofino è scritto male, le "ville" sono diventate "vile", questi sono errori di *inaccuratezza* lessicale.

Accanto a questi errori, notate un errore un po' più difficile da rilevare? Pensateci un attimo……

Se osservate la frase "un grande aliscafo nel porticciolo", è evidente la mancanza di un verbo tra "aliscafo" e "nel porticciolo", il verbo "entra", che era presente nella versione originaria del messaggio; questo è un errore di inaccuratezza sintattica, perché in italiano deve esserci un verbo in una frase come quella che stiamo commentando.

Dunque: la qualità dipende dal tipo di rappresentazione, e, inoltre, presenta diverse caratteristiche, come, nel caso di Figura 1, l'accuratezza di valore nella tabella, la tempestività di aggiornamento nella mappa, l'accuratezza lessicale e sintattica nel testo. Queste caratteristiche di qualità dei dati, che chiameremo nel capitolo *dimensioni* di qualità, sembrano dipendere dal tipo di dato, ad esempio l'accuratezza del valore "Turimo" per l'attributo "Attività economica" consiste nel fatto che "turimo" non corrisponde a nessun valore nell'elenco delle attività economiche, mentre l'assenza di un verbo nel messaggio è di natura più complessa e afferisce alla sintassi dell'italiano.

Nelle basi di dati dei sistemi informativi tradizionali i dati sono rappresentati mediante modelli strutturati di tipo relazionale. I sistemi informativi che offrono servizi agli utenti, ad esempio un sistema di prenotazione, fanno uso di dati che sono *controllati*, nel senso che la fonte dei dati e il dominio di definizione è in genere noto, e la natura strutturata dei dati, insieme ai vincoli di integrità definibili, costituiscono un filtro alla immissione di dati scorretti. La prima parte del capitolo è dedicata alla qualità dei dati nei sistemi informativi tradizionali. La Sezione 2 si focalizza sulla qualità dei dati nelle basi di dati relazionali. Le Sezioni 3, 4 e 5 sono dedicate alla qualità dei dati nei testi, nelle mappe e nelle visualizzazioni. La Sezione 6 parla dei tradeoff che caratterizzano le dimensioni di qualità, per cui migliorandone una, talvolta se ne peggiora un'altra.

A questo punto il capitolo sposta l'osservazione della qualità sul Web, dove tutti possono in modo *non controllato* pubblicare un messaggio Twitter, fare un commento su Facebook, esporre una previsione del tempo, pubblicare una foto. Nel fare questo, spesso non vi è nessun controllo sulla qualità, e il concetto stesso di qualità cambia completamente prospettiva. Alla qualità dei dati nel Web sono dedicate le sezioni successive. In particolare la Sezione 7 è una introduzione generale, la Sezione 8 mostra come la qualità dei dati certe volte "vada di bolina", per cui aumentando i dati disponibili e

quindi la loro dimensione, varietà ed eterogeneità, come accade nel fenomeno dei big data, si possa certe volte migliorarne ugualmente la qualità. La Sezione 9 è dedicata al trust, cioè la fiducia, su cui spesso basiamo le nostre valutazioni di qualità, la Sezione 10 tratta il tema della credibilità e la Sezione 11 la questione così dibattuta e sentita ai nostri giorni della post-verità e delle fake news.

2. Le dimensioni della qualità nelle basi di dati

Osserviamo la tabella in Figura 5. Ogni riga della tabella, eccetto la prima, rappresenta un film del secolo scorso, nelle colonne compaiono nomi che forniscono il significato dei valori. Il valore NULL indica assenza di valore, o meglio, «non conosco il valore». Quando l'ho mostrata ai mei studenti, non avevano visto nessuno dei film citati, neanche nei cinema d'essai, o su You Tube. Questo rende l'esercizio che ora vi propongo un pò più interessante da risolvere. L'esercizio consiste in questo: provate a scoprire, magari consultando Wikipedia, gli errori di qualità presenti nella tabella, e provate a classificarli in termini di dimensioni, come abbiamo visto nell'esempio di Figura 4.

Id	Titolo	Regista	Anno	Numero di Remake	Anno Ultimo Remake
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Vacnze Romane	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Figura 5 – Una tabella con dati di scarsa qualità

Risposta - Nella tabella vi sono errori relativi a diverse dimensioni di qualità; le dimensioni coinvolte sono le seguenti:

- Accuratezza – *Vacnze Romane* è errato, "vacnze" non corrisponde a nessuna parla del vocabolario italiano. Grazie al meccanismo di autocorrezione utilizzato da Google e da altri motori di ricerca nel mostraci le pagine esito di una ricerca per parole chiave, scopriamo che il titolo di film più vicino è *Vacanze Romane*. Inoltre, *Wylder* è errato perché non esistono registi con quel nome, e il regista di Vacanze Romane è Billy *Wilder*. Questo errore è più difficile da rilevare e correggere, perchè una ricerca su Google non fornisce in questo caso nelle prime pagine il nome di un regista, e, piuttosto, dobbiamo cercare il nome del regista nel testo di Wikipedia associato a Vacanze Romane. Ciò è semplice per noi, ma non per una tecnica automatica.
- Completezza – I tre valori *NULL* non ci danno nessuna informazione, e quindi possono vedersi come errori di incompletezza.
- Consistenza – Riguardo al film *Casablanca,* non è possibile che l'anno del primo film con questo titolo sia *1942*, e l'anno dell'ultimo remake sia *1940*, c'è qualcosa che non va; così pure non è

possibile che il film *Sabrina* abbia 0 Remake, e l'ultimo Remake sia del *1985*; possiamo chiamare questi errori di inconsistenza.

Diamo a questo punto alcune definizioni. La *qualità* di un dato (o di un insieme di dati) è la caratteristica del dato che si basa sulla sua capacità di soddisfare necessità ed aspettative esplicite o implicite dei fruitori del dato; l'aspettativa più rilevante è che il dato sia una rappresentazione aderente alla realtà. Una *dimensione* di qualità è una specifica proprietà associabile alla qualità, usualmente non misurabile.

Le più importanti tra le dimensioni di qualità nelle basi di dati sono:

- *l'accuratezza*, intesa come aderenza del dato al fenomeno osservato,
- la *completezza*, cioè l'estensione con cui il dato rappresenta la realtà osservata,
- la *tempestività di aggiornamento*, intesa come rapidità con cui cambiamenti nel fenomeno osservato corrispondono ad aggiornamenti del dato digitale,
- la *consistenza*, cioè il rispetto di un insieme di regole logiche definite per rappresentare le proprietà del dato.

Nella tabella di Figura 6 mostriamo le dimensioni di qualità riscontrate nelle metodologie per la qualità dei dati proposte nella letteratura (vedi [Batini 2009]; mostro queste dimensioni senza nessuna intenzione di commentarle una per una, semplicemente per mostrare visivamente quanto sia esteso il tema della qualità dei dati nelle basi di dati e quante dimensioni diverse siano state proposte.

Acronym	Data Quality Dimension
TDQM	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of manipulation, Value added, Free of error, Interpretability, Objectivity, Relevance, Reputation, Security, Timeliness, Understandability
DWQ	Correctness, Completeness, Minimality, Traceability, Interpretability, Metadata Evolution, Accessibility (System, Transactional, Security), Usefulness (Interpretability, Timeliness (Currency, Volatility), Responsiveness, Completeness, Credibility, Accuracy, Consistency, Interpretability
TQDM	Inherent dimensions: Definition conformance (consistency), Completeness, Business rules conformance, Accuracy (to surrogate source), Accuracy (to reality), Precision, Nonduplication, Equivalence of redundant data, Concurrency of redundant data, Pragmatic dimensions: accessibility, timeliness, contextual clarity, Derivation integrity, Usability, Rightness (fact completeness), cost.
AIMQ	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of operation, Freedom from errors, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability
CIHI	Dimensions: Accuracy, Timeliness Comparability, Usability, Relevance Characteristics: Over-coverage, Under-coverage, Simple/correlated response variance, Reliability, Collection and capture, Unit/Item non response, Edit and imputation, Processing, Estimation, Timeliness, Comprehensiveness, Integration, Standardization, Equivalence, Linkage ability, Product/Historical comparability, Accessibility, Documentation, Interpretability, Adaptability, Value.
DQA	Accessibility, Appropriate amount of data, Believability, Completeness, Freedom from errors, Consistency, Concise Representation, Relevance, Ease of manipulation, Interpretability, Objectivity, Reputation, Security, Timelines, Understandability, Value added.
IQM	Accessibility, Consistency, Timeliness, Conciseness, Maintainability, Currency, Applicability, Convenience, Speed, Comprehensiveness Clarity, Accuracy, Traceability, Security, Correctness, Interactivity.
ISTAT	Accuracy, Completeness, Consistency
AMEQ	Consistent representation, Interpretability, Case of understanding, Concise representation, Timeliness, Completeness Value added, Relevance, Appropriateness, Meaningfulness, Lack of confusion, Arrangement, Readable, Reasonability, Precision, Reliability, freedom from bias, Data Deficiency, Design Deficiency, Operation, Deficiencies, Accuracy, Cost, Objectivity, Believability, Reputation, Accessibility, Correctness, Unambiguity, Consistency
COLDQ (Loshin)	Schema: Clarity of definition, Comprehensiveness, Flexibility, Robustness, Essentialness, Attribute granularity, Precision of domains, Homogeneity, Identifiability, Obtainability, Relevance, Simplicity/Complexity, Semantic consistency, Syntactic consistency. Data: Accuracy, Null Values, Completeness, Consistency,Currency, Timeliness, Agreement of Usage, Stewardship, Ubiquity, Presentation: Appropriateness, Correct Intepretation, Flexibility, Format precision, Portability, Consistency, Use of storage, Information policy: Accessiiblity, Metadata, Privacy, Security, Redundancy, Cost.
DaQuinCis	Accuracy, Completeness, Consistency, Currency
QAFD	Syntactic/Semantic accuracy, Internal/External consistency, Completeness, Currency, Uniqueness.
CDQ	Accuracy, Completeness, Consistency, Currency, Timeliness, Completability, Reputation, Accessibility, Cost.

Figura 6 – Quante solo le dimensioni di qualità dei dati?

Una *metrica* di qualità è una misurazione di una dimensione di qualità che, partendo dalla dimensione da misurare, associa ad essa un valore numerico o ordinale (es. alta) in un dominio di valori. Ad esempio, nella seconda riga della tabella di Figura 5 abbiamo cinque valori sul sei specificati, mentre il sesto ha valore nullo; possiamo associare alla completezza della riga il valore 5/6.

Più complesso è associare metriche ad altre dimensioni, come, ad esempio, la leggibilità di un testo, intesa come la capacità del testo di esprimere il significato, senza spiegazioni aggiuntive. Nel caso della *leggibilità*, sono state definite varie metriche, tra cui ad esempio, indici che misurano la percentuale di parole sul totale che non fanno parte di un elenco di parole considerate comprensibili a una persona che abbia raggiunto un determinato titolo di studio (ad esempio la scuola dell'obbligo). Se ricordate, abbiamo commentato questa dimensione all'inizio del prologo; torneremo sulla leggivilità tra poco, nella Sezione 3.

Supponiamo ora di voler rappresentare le informazioni contenute nella nuvoletta a sinistra della Figura 7, e che il risultato di questo processo di rappresentazione sia la tabella con una sola n-pla mostrata in basso. Assumiamo di ignorare i valori veri contenuti nella nuvoletta; ciò è usuale nelle basi di dati, i valori vengono inseriti una volta, e poi vengono interrogati da persone che non hanno partecipato all' inserimento iniziale del dato.

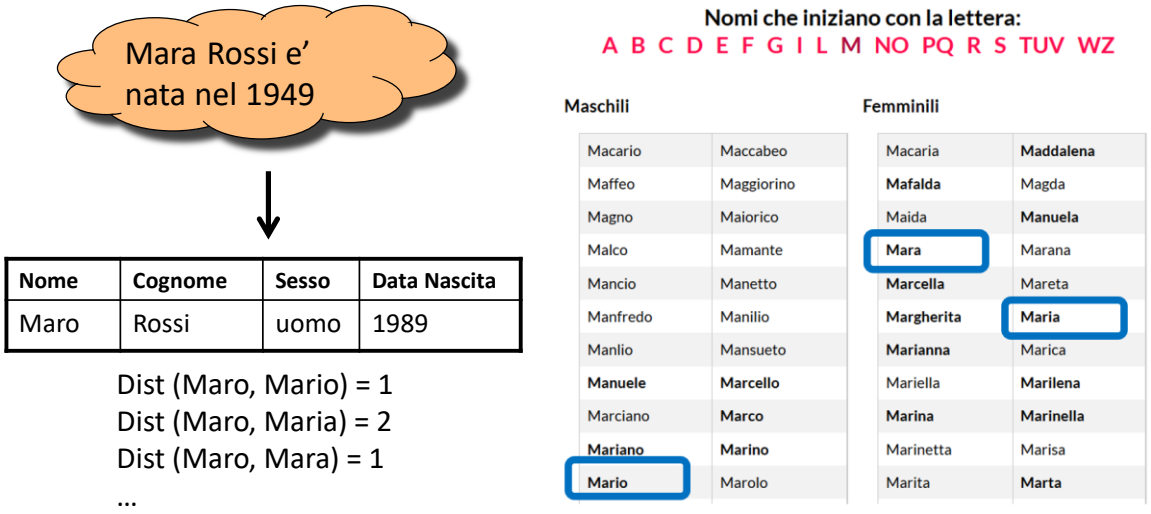


Figura 7 – Come troviamo il nome vero?

Osservando la tabella, ci rendiamo conto che il valore "Maro" è errato, perché non corrisponde a nessun valore noto dei nomi dati alle persone. Cosa possiamo fare per trovare il valore vero? Certamente il processo più affidabile consisterebbe nel cercare la persona e chiedergli come si chiama, sperando che la persona ci risponda in modo corretto, ma è chiaro che ciò è impossibile o molto costoso.

Un procedimento approssimato, e quindi soggetto ad errori, consiste prima di tutto nell'assumere che il nome sia un nome italiano, trovare un sito nel Web che elenchi tutti inomi italiani, e poi confrontare

“Maro” con questo elenco. In Figura 7 è mostrata una pagina con alcuni dei nomi di uomo e di donna che cominciano per M. Effettuare un confronto a questo punto porta a fare un’altra assunzione, che la prima lettera del nome sia corretta. In questo caso nel confrontare “Maro” con i nomi negli elenchi possiamo misurare la distanza tra “Maro” e i diversi nomi di uomo e di donna. In Figura sono evidenziati i tre nomi più vicini, ed è calcolata la distanza, misurata come numero dei caratteri che dobbiamo inserire, sostituire o cancellare per trasformare Maro nel nome (questa distanza è chiamata Edit distance, e non è l’unica che possiamo misurare). Ci sono due nomi a distanza 1, Mario e Mara, e uno a distanza 2, Maria. Come decidiamo tra Mario e Mara? Se facciamo l’assunzione che Sesso abbia valore corretto, allora possiamo dedurre che il nome corretto è Mario.

Il processo visto, pur nella sua semplicità, ha messo in evidenza un fatto importante. Per misurare la qualità dei dati, nel nostro caso la accuratezza dei nomi nella tabella, e per correggerli, dobbiamo cercare una conoscenza esterna alla tabella, che ci guidi nel processo. Nel nostro caso, la conoscenza è costituita dall’elenco dei nomi, e dal fatto che la persona in questione è affermata essere di sesso maschile. Inoltre abbiamo fatto diverse assunzioni, abbiamo assunto che:

- il primo carattere del nome sia corretto;
- il valore vero compaia nella lista trovata sul Web;
- il valore del sesso sia corretto;
- sia stato commesso un solo errore di digitazione nell’inserire nella tabella il nome.

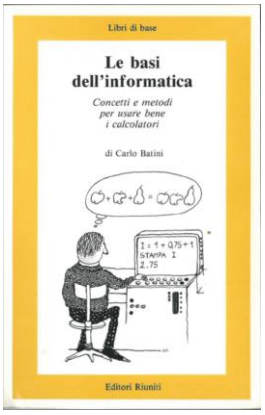
Le azioni di miglioramento della qualità dei dati nelle basi di dati si basano sul confronto tra i dati e una conoscenza di riferimento che può essere costituita da insiemi di dati certificati, vincoli logici tra dati, ovvero, ancora, può essere acquisita mediante ricerche o indagini.

Una trattazione esaustiva del tema della qualità dei dati nelle basi di dati, nelle immagini, nelle mappe geografiche e nei testi non strutturati, nonché la discussione di metodologie e tecniche per la valutazione e il miglioramento della qualità dei dati compaiono in [Batini, Scannapieco 2016].

3. La qualità nei testi

Quando dai dati strutturati passiamo ai testi non strutturati o debolmente strutturati (per esempio testi suddivisi in sezioni e paragrafi, le cose si complicano non poco. Mostriamo qui due esempi.

Abbiamo già visto nella Figura 1 del Capitolo 1 che in occasione della pubblicazione nel 1984 di un mio libro nella collana diretta da Tullio De Mauro, in sede di correzione delle bozze fu verificato che avevo usato nella prima versione vocaboli che non facevano parte delle 5.000 parole più usate del vocabolario, e note a chi aveva frequentato la scuola dell’obbligo che allora terminava a 13 anni. Questa misura corrisponde a ciò che abbiamo chiamato *metrica* nella sezione precedete. Possiamo chiamare *leggibilità* la corrispondente dimensione di qualità. L’importanza della conoscenza del significato delle parole era stata già affermata da Don Lorenzo Milani In una lettera al suo amico Gian Paolo Meucci: “Un operaio conosce 100 parole, il padrone 1.000. Per questo lui è il padrone”.



Il libro iniziava con ..

«**Quotidianamente ciascuno di noi** ha la necessità di usare e produrre un grande numero di informazioni.»

Il testo è stato modificato in

«Ciascuno di noi ogni giorno ha la necessità di usare e produrre un grande numero di informazioni.»

Figura 8 – Esempio di miglioramento della leggibilità

Per aderire alla regola delle 5.000 parole, dovetti sostituire le parole non incluse nelle 5.000 con altre che invece erano presenti nell'elenco. Come mostra la Figura 8, la sostituzione di "quotidianamente" con "ogni giorno" rispetta questa soglia di leggibilità, ma, allo stesso tempo, aumenta di una il numero delle parole della frase, peggiorando un'altra metrica di leggibilità, questa volta basata non sulle parole ma sulle frasi. Possiamo ovviare a questo problema spezzando le frasi in periodi più brevi; anche in questo caso si vede che per migliorare la qualità dobbiamo compiere un percorso talvolta complesso. Ricordo anche che tanto tempo fa scrissi delle dispense per un consorzio che produceva corsi per la Università a distanza, in cui lo strumento di apprendimento fondamentale erano appunto le dispense; il mio testo iniziale fu sottoposto a quattro valutazioni di leggibilità, ad ogni valutazione ero inviato a riscrivere il testo, e ad ogni nuova revisione la leggibilità peggiorava, la conclusione fu che tornammo al testo iniziale….

In Figura 9 ci occupiamo di una dimensione un po' diversa rispetto alla precedente, la comprensibilità. La comprensibilità è la capacità di un testo di essere capito, nelle sue singole frasi, nei nessi tra di esse, e nella sua globalità. I romanzi gialli tradizionali sono scritti, come noto, con lo scopo di descrivere uno o più delitti, introducendo personaggi e fornendo indizi nel testo su tali personaggi, fino ad arrivare a individuare l'assassino. Naturalmente ci sono molte varianti di questo canovaccio; Agatha Christie ha scritto molti romanzi gialli, e tra questi "The mysterious affair at Styles". All'inizio del romanzo, Agatha Christie ci fornisce l'elenco dei personaggi cui darà vita, con un breve profilo che serve al lettore per seguire il corso degli eventi, essendo in grado in ogni momento di contestualizzare il personaggio, facilitato in questo dalla consultazione del profilo.

Possiamo dire che la descrizione iniziale dei personaggi aumenta la comprensibilità, perché permette di ricordare in maniera più nitida, e tenendo conto del contesto, la storia complessiva descritta nel romanzo giallo.

Characters in "The Mysterious Affair at Styles"
Captain Hastings, the narrator, on sick leave from the Western Front.
Hercule Poirot, a famous Belgian detective exiled in England; Hastings' old friend
Chief Inspector Japp of Scotland Yard
Emily Inglethorp, mistress of Styles, a wealthy old woman
Alfred Inglethorp, her much younger new husband
John Cavendish, her elder stepson
Mary Cavendish, John's wife
Lawrence Cavendish, John's younger brother
Evelyn Howard, Mrs. Inglethorp's companion
Cynthia Murdoch, the beautiful, orphaned daughter of a friend of the family
Dr. Bauerstein, a suspicious toxicologist

Figura 9 – Esempio di miglioramento della comprensibilità

4. La qualità delle mappe

Abbiamo introdotto nel Capitolo 2 le mappe come esempio di dati digitali che utilizzano modelli di rappresentazione basati su punti, linee e superfici, una legenda che descrive i possibili simboli utilizzati nella mappa, e infine regole che disciplinano la relazione tra elementi geografici e simboli.

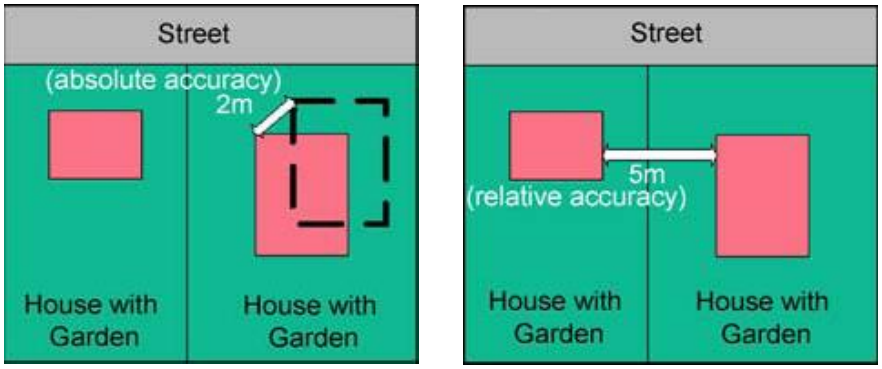
Figura 10 – Una mappa sbagliata

Gli abitanti di Lecco ed in particolare di via Marco d'Oggiono presente nella mappa rappresentata in Figura 10, in occasione della applicazione di nuove modalità di raccolta dei rifiuti, hanno osservato per diversi giorni un accumulo dei sacchi di rifiuti sui marciapiedi e fuori dai portoni. A individuare la ragione che ha generato questa situazione è stato un residente, che, tramite un controllo incrociato fra la mappa e uno stradario di Lecco, ha scoperto che nella mappa via Marco d'Oggiono risulta appartenere a una zona classificata come Zona 1, mentre nello stradario pubblicato sul sito del Comune risulta appartenere alla Zona 2, con giorni di raccolta completamente diversi. In questo caso l'errore è consistito in una scorretta attribuzione nella mappa di una via ad una zona; l'effetto dell'errore è stata l'esclusione della via dalla raccolta dei rifiuti e il conseguente accumularsi di rifiuti.

Nelle mappe, le dimensioni di qualità definite per le tabelle si modificano in funzione della diversa struttura spaziale delle mappe rispetto alle tabelle. Ad esempio, la accuratezza, che abbiamo discusso

per le tabelle nel caso dei nomi delle persone, diventa una proprietà più complessa e articolata, perché le mappe al contrario delle tabelle rappresentano una realtà con due (e talvolta, per le mappe in rilievo, tre) coordinate spaziali.

Per esempio, in Figura 11 sono mostrati due tipi di accuratezza caratteristici nelle mappe, che fanno riferimento alla posizione nel piano di due case. Nella accuratezza posizionale assoluta, siamo interessati a stabilire la distanza della casa a destra rispetto alla sua posizione reale, che può essere misurata con la latitudine e longitudine dei suoi quattro punti angolari; nel caso della accuratezza posizionale relativa, siamo invece interessati a misurare la variazione rispetto alla realtà della posizione relativa delle due case. E' possibile che entrambe le case abbiano una distanza di 10 metri rispetto alle coordinate reali, ma che la loro posizione relativa sulla mappa sia quella reale.



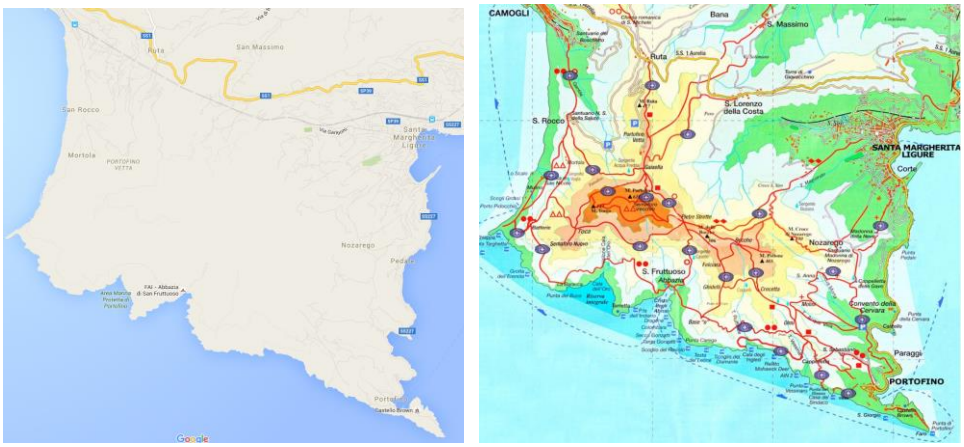
Accuratezza posizionale assoluta

Accuratezza posizionale relativa

Figura 11 – Accuratezza nelle mappe

Fino ad ora abbiamo discusso le dimensioni di qualità dei dati come proprietà assolute dei dati digitali, senza pensare all'uso che ne facciamo. La casa nella parte sinistra di Figura 11 ha una inaccuratezza posizionale assoluta di 10 metri, stop; in realtà, se approfondiamo un po' il problema, ci accorgiamo che la qualità dipende dall'uso e dagli scopi per cui utilizziamo i dati. Consideriamo i due esempi di Figura 12 e Figura 13. In Figura 12 sono rappresentate due mappe del monte di Portofino, dove sono andato diverse volte a fare passeggiate con mio figlio. Ci sono vari sentieri da Camogli a Portofino e a Santa Margherita Ligure, molto ben indicati con cartelli, ma è chiaro che ci torna molto comodo avere una visione di insieme dei sentieri, che soltanto una mappa ci fornisce. E' un pò come con i navigatori; per carità, utilissimi, ma quando li uso mi sento espropriato di una visione di insieme del territorio, mi sento un esecutore stupido di scelte fatte da un algoritmo.

Osserviamo le due mappe, qui rappresentate con la stessa scala. Se dobbiamo usarle per orientarci tra Camogli, San Fruttoso e Pertofino, non c'è materia del contendere, meglio la seconda, molto più completa rispetto alla prima nella descrizione dei sentieri. Una rappresentazione a scala più piccola della zona su Google Maps (la mappa a sinistra) introduce qualche sentiero, ma resta sempre molto incompleta. D'altra parte, se dobbiamo orientarci nel percorso di una strada statale o autostrada è chiaro che Google Maps è molto più adatta della mappa a destra, in cui il livello di dettaglio è eccessivo. Quindi, il livello di completezza desiderato di una mappa dipende dall'uso.



*https://**maps.google**.it/* *www.**portofino**trek.com/trek/6-mappa*

Figura 12 – La mappa a destra è migliore se intendiamo fare una passeggiata

Passando alla Figura 13, vediamo le mappe delle metropolitane di Londra e New York. Le due grandi metropoli hanno conformazioni diverse, più circolare Londra e più allungata Manhattan (la parte rappresentata in figura); certamente salta all'occhio che la disposizione delle linee e delle fermate nelle due mappe è ispirata a criteri diversi.

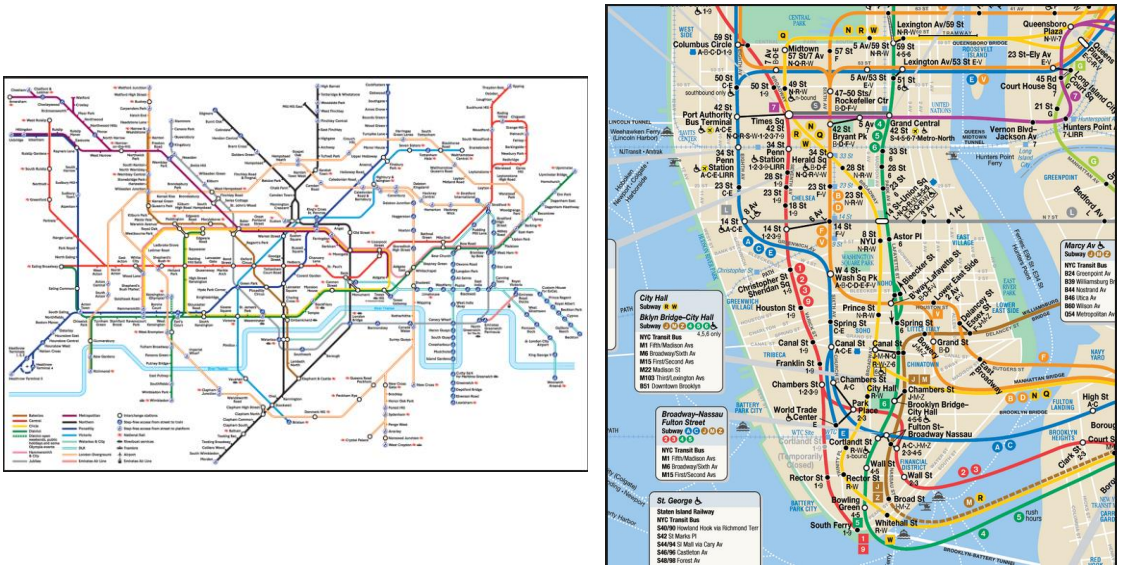


Figura 13 – Le mappe delle metropolitane di Londra e New York
(da www.bbc.co.uk e www.pinterest.com)

Nel caso di Londra il layout delle linee è stilizzato, non fa riferimento al territorio ed è ottimizzato nella disposizione sul piano, così da distanziare il più possibile linee e fermate ottimizzando la leggibilità, che verrebbe ridotta nel caso, ad esempio, di fermate così vicine da non essere chiaramente distinguibili. Nel caso di New York, invece le linee sono fatte corrispondere (approssimativamente) al territorio che percorrono, e quindi le fermate sono in corrispondenza con le strade dove esse si trovano.

Quale mappa è per voi di maggiore qualità? Per rispondere occorre prima mettersi d'accordo sulla dimensione di qualità osservata, e sull'uso che facciamo della mappa. Non posso entrare nella testa del lettore, ma se come dimensione pensiamo alla leggibilità, ebbene trovo la mappa di Londra di superiore qualità. Diverso è il discorso se invece pensiamo all'uso della mappa. Se ci troviamo a camminare in corrispondenza del cerchio rosso rappresentato per Londra e per New York in Figura 14, e vogliamo trovare la fermata più vicina, ebbene non c'è dubbio che ci è più di aiuto la mappa di New York, che ci fornisce un riferimento diretto alle strade che dobbiamo percorrere per raggiungere la più vicina linea di metropolitana.

Figura 14 – Dove è la fermata più vicina della metro?
(da www.bbc.co.uk e www.pinterest.com)

Le mappe di Figura 13 e Figura 14 ci confermano varie questioni: la qualità è un concetto con molte dimensioni, la percezione soggettiva di qualità può dar luogo a valutazioni diverse, la qualità dipende dall'uso.

5. La qualità delle visualizzazioni

Spostandoci dalle mappe alle visualizzazioni, la discussione sulla qualità dei dati digitali trova nuovi spunti. Se guardiamo la Figura 15, che riprende la Figura 14 del Capitolo 1, notiamo che la forma della strada e della sua larghezza deforma sensibilmente i dati numerici sul consumo di benzina riportati nella seconda colonna della tabella a sinistra.

Anno	Miglia per gallone
1978	18
1979	19
1980	20
1981	22
1982	24
1983	26
1984	27
1985	27,5

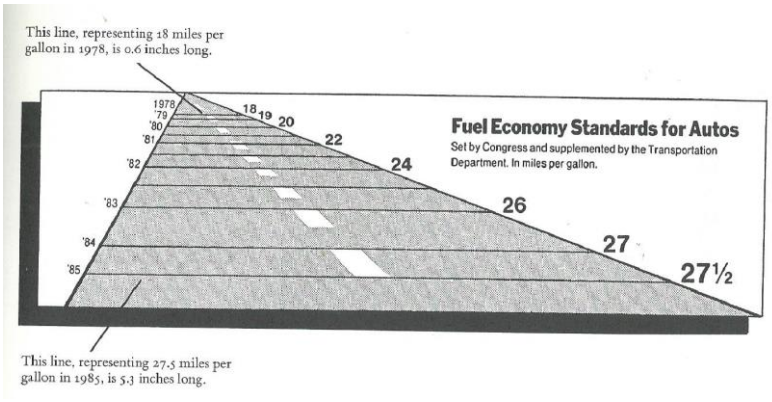


Figura 15 – La qualità come grandezza della bugia
(da E. Tufte – The visual display of quantitative information, 2001)

Nel bellissimo libro di Tufte [Tufte 1984] viene introdotto il concetto di *fattore di bugia*, che può essere visto come una metrica per misurare la accuratezza relativa tra i valori del consumo di benzina, espressi numericamente e tramite le larghezze della strada, e che possiamo esprimere come una frazione, vedi il box seguente (il livello di bugia mi ricorda tanto l'inizio di Anna Karenina, parafrasato in "tutte le verità sono uguali, ogni bugia è diversa una dall'altra").

Livello di «bugia» = rapporto tra lunghezze nella visualizzazione / rapporto tra valori numerici nel mondo reale = 15

L'esempio che mostro in Figura 16 mi ricorda quanto è avvenuto in Cina in due viaggi che ho fatto a trenta anni di distanza. Alla fine degli anni 80 fui invitato in Cina a Pechino da due professori della Beda University, una delle più antiche della Cina, con un finanziamento della World Bank. Allora mi interessavo tra gli altri argomenti di ricerca di disegno automatico dei diagrammi, e quindi decisi di includere questo tema tra quelli dei miei seminari; riparleremo di disegno automatico di diagrammi nel Capitolo 12. I seminari dovevano essere cinque, al primo seminario assistettero circa 16 tra professori, ricercatori e studenti. Nel primo seminario parlai del disegno di diagrammi e mostrai i due diagrammi Entità Relazione di Figura 16, diagrammi che utilizzano gli stessi simboli, sia pure con disposizioni diverse dei simboli sul piano. Anticipo subito cosa accadde nei seminari successivi; quell'epoca in Cina non comprendevano bene l'inglese, e, insomma, i partecipanti ai diversi seminari scesero a otto nel secondo, quattro nel terzo, due nel quarto, e uno nel quinto, nel sesto e nel settimo, una sorta di vittima sacrificale cui forse fu imposto di partecipare per educazione.

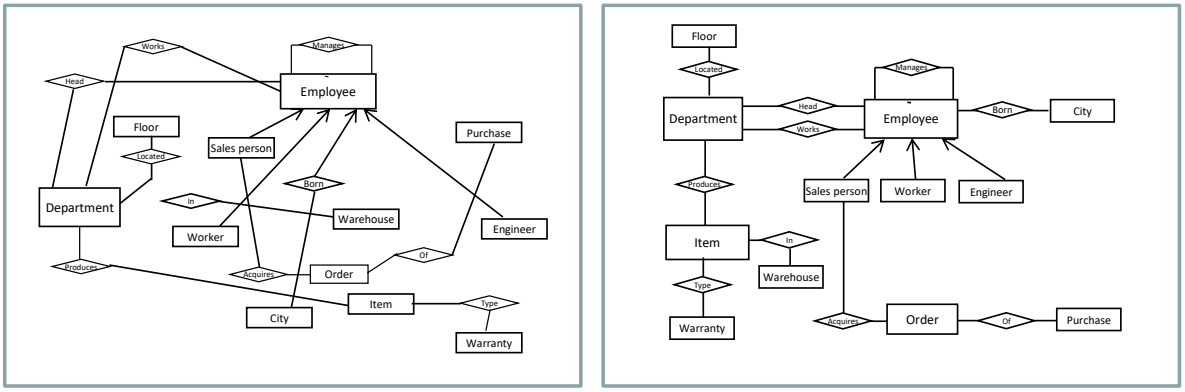


Figura 16 - Due diagrammi Entità Relazione che rappresentano lo stello schema concettuale

In occasione del primo seminario chiesi all'uditorio (ancora numeroso): which one of the two diagrams do you like more? E tutti alzarono la mano a favore del primo. Io rimasi stupefatto, e chiesi: why? E uno tra i partecipanti disse: perché è più mosso del secondo, dà più del secondo il senso del movimento, che a noi cinesi piace molto…

Trenta anni dopo feci un nuovo seminario, questa volta a Harbin, sulla qualità dei dati, e rifeci lo stesso esempio, con la stessa domanda. E tutti risposero: il secondo. Per dire, la qualità e la sua percezione sembrano proprio avere anche radici culturali.

Torniamo a noi. Se preciso meglio la domanda, e vi chiedo di decidere quale dei due diagrammi è più leggibile, intendendo con leggibile, ricordo, il fatto che lo si possa comprendere con basso sforzo cognitivo, cosa dite? Direi che possiamo concludere che è senza ombra di dubbio più leggibile il diagramma a destra. Ma perché? La ragione è che rispetta diversi criteri estetici che sono mostrati nella parte superiore di Figura 17, in forma di suggerimenti per migliorarne la leggibilità.

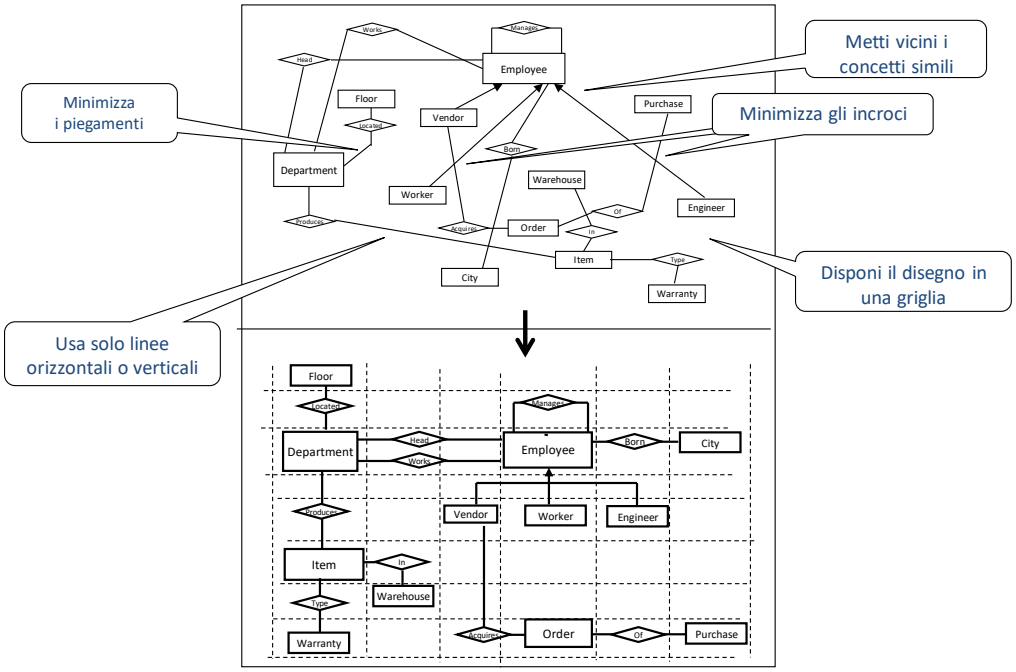


Figura 17 – Criteri estetici per migliorare la qualità

Nella parte bassa di Figura 17 vediamo il nuovo diagramma in cui sono state applicate tutte le regole di miglioramento. Questo esempio conferma quanto abbiamo discusso fino ad ora, aggiungendo altre due questioni importanti: il concetto di qualità è ampio, e si intreccia con la estetica e con la nostra cultura.

6. I tradeoff tra dimensioni di qualità

Finora abbiamo descritto e misurato la qualità dei dati digitali tramite un insieme di dimensioni e metriche che, volendo migliorare la qualità del dato, possono essere migliorate tutte insieme, senza conflitti. Confrontiamo ora le due immagini di Figura 18 che si riferiscono all'ingresso di un parcheggio. La prima è stata ottenuta in una giornata nebbiosa, e rappresenta fedelmente la scena in cui si è imbattuto l'automobilista che ha fatto la foto, mentre la seconda è stata ottenuta con un ritocco, in cui è stato aumentato il chiaroscuro. Dal punto di vista della leggibilità, è migliore la immagine a destra, ma dal punto di vista della fedeltà all'originale è migliore la immagine a sinistra. Vediamo dunque che tra diverse dimensioni di qualità vi possono essere tradeoff, nel senso che quando una migliora l'altra peggiora.

Spesso ci accade di incontrare situazioni come la precedente; ad esempio, se c'è stato un terremoto, i giornali on line cercano di essere i primi a pubblicare stime dei danni, ma se vogliono essere i più rapidi, e rinunciano a verificare la attendiblità della fonte, rischiano di pubblicare notizie che poi si rivelano inesatte. Analoga situazione accade quando si chiudono le urne in una votazione politica nazionale; gli exit poll danno informazioni quasi in tempo reale, che sempre più spesso si sono rivelate inesatte, per la tendenza di alcuni intervistati, nonostante l'anonimato, a nascondere il partito che hanno votato nell'urna; anche in questo caso, il dato più recente e più aggiornato non è il più accurato.

Figura 18 – Fedeltà vs Leggibilità

7. La qualità dei dati nel Web

7.1 Introduzione

La grande crescita dei dati prodotti nel Web modifica profondamente il concetto e le metodologie per la qualità dei dati rispetto a come li abbiamo considerati nelle sezioni precedenti; questo in virtù della grande varietà dei tipi di dati nel Web, e per la impossibilità, spesso, di confrontare il dato, come abbiamo fatto in precedenza, con una *conoscenza di riferimento*. Le differenze tra i due ambiti sono molteplici:

1. nel Web il costo di produzione e di trasmissione dei dati è praticamente nullo, a fronte di possibilità di diffusione sempre più ampie. E' praticamente gratis sia mandare una email a una persona che mandarla a un indirizzario di centinaia o migliaia di indirizzi; scrivere un messaggio su Twitter richiede uno sforzo cognitivo molto basso, e, allo stesso tempo, su Twitter ci possono essere milioni di followers, e i tweet si diffondono molto più velocemente dei dati riprodotti nelle fonti cartacee tradizionali; infine, leggere e comprendere un tweet richiede uno sforzo cognitivo enormemente inferiore rispetto a leggere e comprendere un articolo di giornale.
2. rispetto al carattere controllato dei flussi di dati nei sistemi informativi tradizionali e alla presenza, spesso, di un soggetto certificatore del dato che ne valida la qualità, nel Web si assiste ad una sostanziale disintermediazione tra la fonte e il ricevente. Ad esempio, sappiamo che è molto ampio il dibattito su quali siano i limiti non superabili delle notizie false nelle reti sociali, e chi debba

controllare e contrastare la veridicità delle notizie; torneremo su questi aspetti più avanti nel capitolo, ma è chiaro che la assenza di un soggetto certificatore del dato complica enormemente il tema della qualità.

3. nel Web abbiamo una distinzione tra *fonte* del dato, *mezzo* con cui è trasmesso e *messaggio* che viene trasmesso, e il tema della qualità riguarda tutti e tre i livelli, con intrecci rilevanti. Ad esempio, esiste una profonda differenza tra i messaggi inviati via email, via Facebook e via Twitter, derivanti dai vincoli sulla lunghezza del messaggio, il conseguente uso di vocabolari arricchiti con simboli, abbreviazioni e metafore, e il diverso livello di formalità che noi associamo alle tre tipologie di messaggi.
4. non ci sono (e non ci possono essere) standard universali per scambiare l'informazione nel Web; l'informazione può essere alterata o creata anonimamente sotto falsa identità o con la intenzione dell'inganno.
5. nei siti che nascono per crowdsourcing, cioè attraverso il contributo volontario di molti utenti, il dato digitale è il risultato di molteplici contributi e versioni e si perde la conoscenza sulla fonte e sul processo di provenienza.
6. nel Web mancano spesso fonti di conoscenza certe con cui confrontare il dato, anche perché spesso lo stesso significato del dato è conosciuto in modo vago e impreciso. Come conseguenza, spesso la qualità del dato viene di fatto stabilità sulla base della qualità della fonte (me lo ha scritto Giovanni, di lui mi fido…); e la qualità della fonte non è espressa in termini delle dimensioni viste in precedenza, ma, piuttosto, in termini di dimensioni ispirate da discipline che studiano gli esseri umani (Giovanni…), quali le scienze sociali, le scienze cognitive, la psicologia, la filosofia.
7. essendo la qualità del dato ricondotta alla qualità della fonte, acquista rilevanza nella analisi di qualità stabilire la provenienza del dato e il processo con cui si è formato.
8. nel processo di formazione del dato possono essere coinvolte diverse fonti, per cui può risultare molto difficile o impossibile accertare quale ruolo abbiano avuto le diverse fonti, e quindi la loro rilevanza rispetto alla qualità.
9. l'analisi, le classificazioni e le tecniche proposte per valutare la qualità dei dati, che nella letteratura scientifica sono presenti da pochi anni, si differenziano significativamente a seconda del tipo di dato, e si concentrano maggiormente nell'ambito della informazione linguistica, i messaggi, i testi in formato libero, gli articoli di giornale, i rumours (vedi [Zubiaga 2018]), le opinioni, le news, il microblogging, la informazione specialistica (es. in medicina), vedi [Viviani 2017].
10. Le dimensioni di qualità investigate per le varie tipologie di dati e le metodologie per la valutazione e il miglioramento della qualità, dovendo riguardare questa grande varietà di fenomeni, sono molto meno consolidate rispetto ai sistemi tradizionali, si veda per esempio [Batini 2015] e la relativa bibliografia sulla qualità nei linked open data.

Per esprimerci con una metafora, i dati si espandono nel Web come una sfera opaca (Figura 18), in cui accanto a dati di qualità, compaiono con sempre maggiore intensità dati imprecisi, sfocati, incompleti, eterogenei al loro interno, volutamente falsi, rendendo più arduo ricostruirne la validità.

L'immagine della sfera è anche utile per farci capire che, a seguito dei nostri limiti cognitivi, un ampliamento della conoscenza dei fenomeni può portarci a percepire tracce sempre più frammentate, eterogenee e imprecise di una realtà sempre più complessa, che vanno ricomposte in una conoscenza

comune. Anche se un po' alla lontana, vale il proverbio cinese: più invecchio, più conosco e meno capisco….

Nella grande sfera opaca il concetto “minimalista” di *qualità* del dato, adottato nei sistemi informativi tradizionali, si amplia a dismisura applicandosi ai vari tipi di informazioni prodotti dall’Internet delle cose, scambiati nel Web e nelle reti sociali, acquisiti e scambiati per mezzo dei telefoni cellulari, memorizzati nel Cloud, invadendo tutta la nostra vita di relazioni, e approdando in ultima analisi al concetto di *verità*, ampiamente studiato da secoli nella filosofia del linguaggio, nella filosofia morale, nella logica e nell’etica, e alle sue sistematizzazioni ed evoluzioni nei concetti di ipoverità, iperverità e post verità (vedi in seguito). La qualità dei dati digitali, da concetto limitato e intrinseco ai dati, diventa un concetto sempre più soggettivo e sempre più influenzato dal messaggio che esprime il dato, dalla cultura e condizione sociale del ricevente il dato e da aspetti emotivi e non razionali.

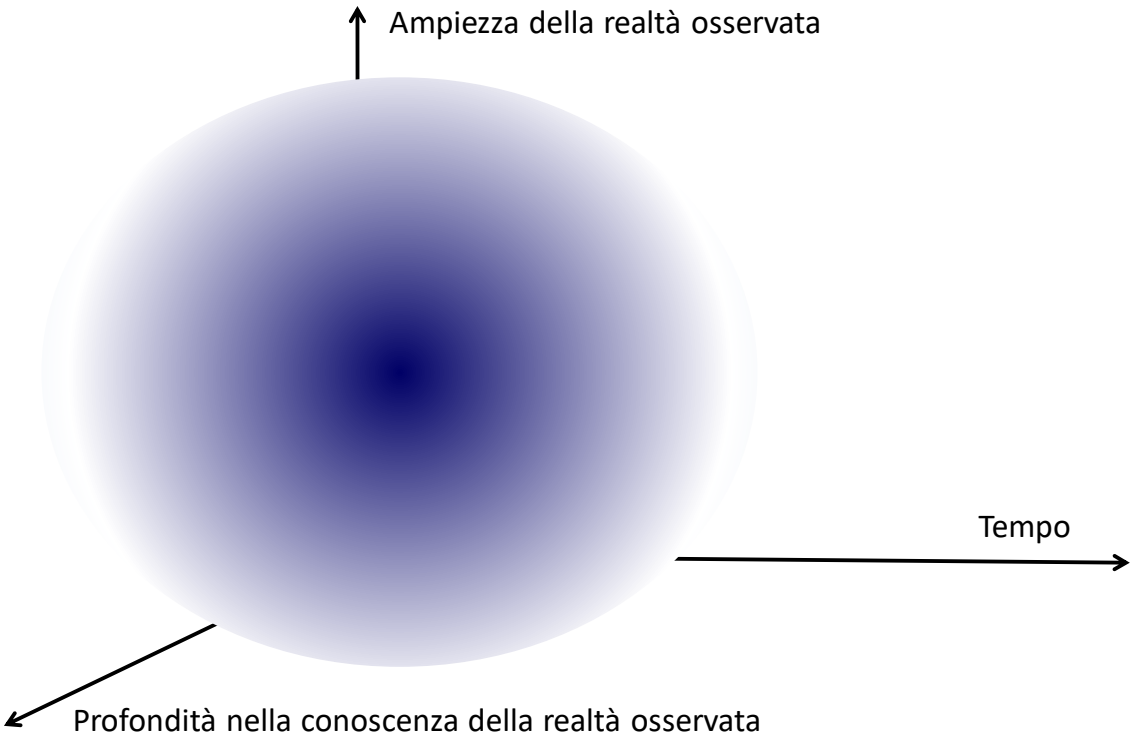


Figura 19 - La grande sfera opaca

Per fare un primo esempio, all’epoca della cerimonia di insediamento di Donald Trump come Presidente degli Stati Uniti, si diffusero foto che mostravano comparativamente le folle presenti all’insediamento alla Presidenza degli Stati Uniti di Obama e di Trump, vedi Figura 20. Apparentemente, la folla presente all’insediamento di Obama appariva essere di gran lunga superiore della folla presente all’insediamento di Trump, ma il portavoce di Trump Sean Spicer si espresse in senso totalmente opposto. E quando un giornalista chiese a una persona nello staff di Trump, Kellyanne Conway, come fosse possibile una così evidente alterazione della verità, disse che le affermazioni del portavoce erano da considerarsi “fatti alternativi”.

Figura 20 – I fatti alternativi (da www.wtop.com)

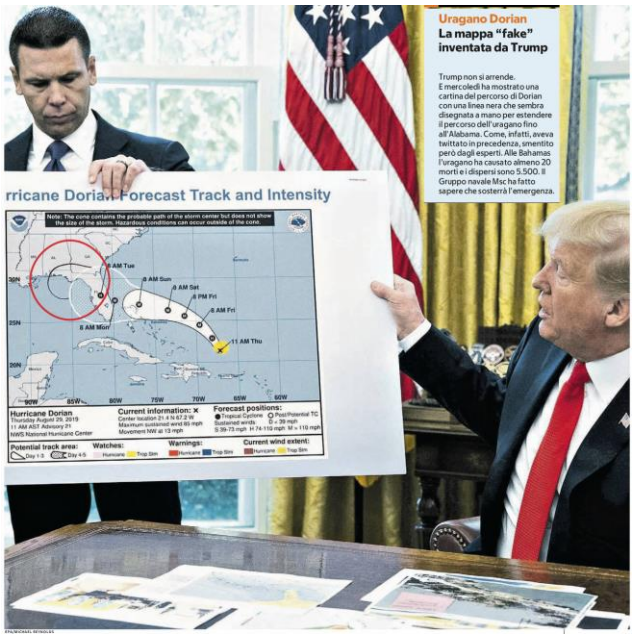


Figura 21 – L'area di azione dell'uragano Dorian nella mappa mostrata da Trump (www.theguardian.com)

Un secondo esempio di come ragionare sulla qualità dei dati diventi molto più complesso nel Web è la mappa che nel 2019 Trump ha mostrato per descrivere la traiettoria dell'uragano Dorian, in cui la linea curva cerchiata in Figura 21 sembra messa per confermare una previsione fatta da Trump sul fatto che Dorian avrebbe avuto un impatto sull'Alabama, previsione dichiarata subito infondata dalla agenzia federale che monitora gli uragani.

7.2 Le dimensioni di qualità nel Web, un'area ancora non assestata

Focalizziamo ora la attenzione sul tema delle dimensioni di qualità dei dati nel Web, vediamo quali sono le più investigate; adotteremo i termini inglesi, essendo imprudente utilizzare dizioni italiane e avvertendo che spesso nella letteratura sulla qualità nel Web non vengono fornite definizioni, piuttosto la dimensione trattata viene riferita ad altre dimensioni, creando nel complesso una rete di concetti che talvolta si richiamano l'uno all'altro, vedi per articoli generali [Lukoianova 2013], [Zubiaga 2018], [Rieh 2010].

La truthfulness è definita in termini di:

- honesty, un carattere morale di un essere umano che si riferisce alla sua abitudine di dire la verità
- accuracy, la propensione della informazione ad essere corretta
- credibility, a cui vengono fatti corrispondere diversi significati.

La credibility è stata studiata in molteplici aree scientifiche, dalla psicologia cognitiva al marketing alle discipline gestionali. In alcuni approcci ricomprende le componenti oggettive e soggettive della believability di una fonte o di un messaggio, accompagnata da componenti secondarie che includono il carisma; In altri approcci è definita come "people's assessment of whether information is trustworthy based on their own expertise and knowledge", e ha quindi due componenti fondanti: la trustworthiness e l'expertise, caratterizzate entrambe da elementi oggettivi e soggettivi. La trustworthiness è basata soprattutto su fattori soggettivi, ma può includere fattori oggettivi come la reliability [Nakamura 2007]. L'expertise può essere similmente percepita in modo soggettivo, ma include anche caratteristiche relativamente oggettive della fonte o del messaggio, come le credenziali o la certificazione.

La objectivity è un concetto più filosofico, ed esprime la proprietà di essere vero, indipendentemente da soggettività individuali causate da percezione, emozioni, immaginazione o fantasie.

La veracity è espressa da:

- accuracy, intesa come cioè conformità alla verità
- truthfullness, cioè la devozione e aderenza alla verità
- capacità di convogliare o percepire la verità.

La reliability ha due ruoli:

- se vista in relazione al metodo scientifico, è la consistenza o ripetibilità delle misure sperimentali;
- se vista in relazione ai modelli statistici, è la consistenza complessiva del processo di misurazione

Altri autori mettono in relazione l'expertise e la reliability, formulando un punto di vista molto lontano dai precedenti: la reliability si riferisce alla volontà di fornire informazioni corrette (intenzione), mentre l'expertise si riferisce alla capacità di fornire informazioni corrette (conoscenza).

Il trust (o fiducia) è stato studiato in molte discipline inclusa la sociologia, la psicologia, l'economia e la informatica. Ciascuna di queste discipline ha definito e considerato il trust da diverse prospettive. Il trust è una misura della fiducia che un'entità, persona, fonte informativa, programma informatico o altro artefatto, si comporterà in una modalità attesa, ovvero rispetterà delle qualità attese.

Nei prossimi due paragrafi ci concentriamo sul trust, e successivamente sui meccanismi cognitivi che applichiamo nel valutare la credibility.

7.3. Il Trust

Applicando la definizione appena data di trust ai dati digitali, il trust nei dati forniti da una parte emittente può essere interpretato come la fiducia da parte del ricevente sul fatto che i dati forniti siano corretti. Nel contesto dei social network, il trust è fortemente legato al capitale sociale della rete, che chiamiamo fiducia sociale. Il capitale sociale di un network è la ricchezza delle interazioni tra i suoi membri. La gestione del trust in una rete è vista in letteratura come:

- basata su credenziali o policy,
- basata su reputazione,
- basata su proprietà della rete.

L' idea di base dietro la gestione basata su credenziali e policy consiste nell'utilizzare credenziali per abilitare un accesso ai dati basato su criteri di controllo delle risorse. La gestione della fiducia basata sulla reputazione, al contrario, fornisce una valutazione della fiducia del proprietario della risorsa basata su valori di reputazione accumulati nel tempo. Il metodo basato su social network utilizza le relazioni sociali per valutare il trust dei singoli nodi nella rete sociale.

Nella scienza dei dati il trust può essere classificato secondo due punti di vista, dell'utente e del sistema. La nozione di trust "utente" è derivata da psicologia e sociologia con una definizione standard che può vedersi come specializzazione della precedente, come "soggettiva aspettativa che un'entità abbia sul comportamento futuro di un altro"; ciò implica che il trust sia una caratteristica essenzialmente personale. Nei sistemi online come eBay e Amazon, la fiducia si basa sul feedback sulle interazioni passate tra i membri; in questo caso, la fiducia è relazionale. In entrambi i casi, nel mentre due membri interagiscono l'uno con l'altro nel tempo, la loro relazione si rafforza o si indebolisce e la fiducia evolve in base alla esperienza di interazione. Nei sistemi online, la fiducia è considerata di due tipi: diretta e basata su raccomandazioni; a fiducia diretta si basa sull'esperienza diretta del membro con l'altra parte. La fiducia basata su raccomandazioni si basa su esperienze di altri membri nel social network; la fiducia delle raccomandazioni è basata sulla proprietà propagativa del trust.

Diversi tipi di trust possono essere individuati:

- quantitativo, quando il trust è il risultato di un calcolo, focalizzato a massimizzare l'interesse che il soggetto ha dalla interazione.
- relazionale, quando il trust è basato su una storia di interazioni ripetute tra i due soggetti coinvolti nel trust.
- emozionale, che definisce il livello di confidenza che si percepisce istintivamente nel fidarsi dell'altro.
- cognitivo, quando il trust è basato sul raziocinio e su comportamento razionale.
- Istituzionale, quando il trust è riposto in una istituzione che incoraggia la cooperazione tra i membri e scoraggia comportamenti malevoli.
- basato su una disposizione positiva, quando nel corso della vita si crea una aspettativa positiva e ottimistica verso gli altri.

Proprietà del trust sono:

- la dipendenza dal contesto, per cui possiamo fidarci di una persona quando ci parla di calcio ma, magari, non quando ci fa un'analisi politica.
- Il carattere dinamico, nel senso che il trust può aumentare o ridursi con nuove esperienze, interazioni o osservazioni.
- Il carattere propagativo, che non significa necessariamente transitivo; se A ha fiducia verso B e B ha fiducia verso C, A è in qualche modo influenzato nel decidere il suo trust verso C.
- Il carattere soggettivo, che porta sempre a una personalizzazione, basata su pesi di importanza soggettivi dei diversi elementi introdotti in precedenza.
- Il carattere componibile, per cui un membro di un social network consapevolmente o meno compone la sua valutazione su un nuovo soggetto tenendo conto delle catene di relazioni esistenti, e arrivando ad una sintesi anche in caso di valutazioni contrastanti relativi alle varie catene.
- l'amplificazione nel tempo della relazione di trust, sia in senso positivo che negativo
- la sensibilità agli eventi, per cui un singolo evento negativo può distruggere un lungo periodo di formazione del trust.

Per un approfondimento di tutta la precedente problematica il lettore può fare riferimento a [Sherchan 2013].

Prima di concludere in tema di trust, osserviamo che l'accessibilità pubblica delle reti sociali unita alla capacità di condividere opinioni, pensieri, informazioni e esperienza offre importanti prospettive alle imprese e alle pubbliche amministrazioni; oltre alle persone che utilizzano le reti per connettersi ai loro amici e famiglie, le imprese e le amministrazioni hanno iniziato a sfruttare queste piattaforme per fornire i loro servizi a cittadini e clienti. Tuttavia, il successo di tali tentativi si basa sul livello di fiducia che i membri hanno tra loro e con il fornitore di servizi; pertanto, la fiducia diventa un elemento essenziale e importante di un social network di successo.

Per bilanciare la natura aperta dei social network e salvaguardare le preoccupazioni sulla privacy degli utenti, è importante costruire comunità di fiducia, comunità cioè che creano un ambiente in cui i membri possono condividere i loro pensieri, opinioni ed esperienze in modo aperto e onesto, senza preoccupazioni sulla privacy e la paura di essere giudicati o che le informazioni fornite possano essere sfruttate per scopi malevoli.

7.4 Euristiche per la valutazione della Credibility

La ricerca si è soffermata recentemente [Metzger 2013] sulle euristiche utilizzate nella valutazione della credibility, cioè sui processi cognitivi che mettiamo in atto per valutare soggettivamente la credibility di una fonte. Le euristiche investigate sono:

- l'euristica basata sulla reputazione, che consiste nel privilegiare alternative riconoscibili rispetto a quelle meno familiari; tendiamo insomma a fidarci di più del noto che dell'ignoto.
- l'euristica basata sull'endorsement (sostegno, appoggio), basata sulle valutazioni espresse da altri, cui affidiamo la nostra; tendiamo a fidarci sulla base di quanto si fidano persone di cui ci fidiamo, ovvero che sosteniamo, o a cui ci affidiamo.

- l'euristica basata sulla consistenza, basata sul confronto tra fonti per evidenziarne le differenze; è la euristica che si forma se noi indaghiamo accedendo ad altre fonti per cercare conferme o contraddizioni, e troviamo solo, o prevalentemente informazioni, appunto, consistenti. Nel caso di inconsistenze, sono proposte varie tecniche per la scelta tra le alternative.
- l'euristica di auto-conferma, che misura la credibilità sulla base della conferma delle precedenti credenze; rientrano in questa tematica le analisi sulle *camere dell'eco*, documentate, ad esempio, in [Quattrociocchi 2016]. Viene presentata una analisi effettuata su circa mille agenzie di stampa e 400 milioni di utenti, in cui è stata esplorata *l'anatomia* del consumo di notizie su Facebook su scala globale. La conclusione che si può trarre è che gli utenti quando accedono al Web per fini informativi, tendono a focalizzare la loro attenzione su un numero limitato di pagine, andando a selezionare un gruppo ristretto di media da cui attingere informazioni e rafforzando così le proprie opinioni, senza mai metterle in discussione. Di fatto, si chiudono nella loro bolla.
- la euristica basata sulla violazione delle aspettative, assume una fonte non credibile se essa ha violato le aspettative in precedenti circostanze; è l'euristica basata sulla delusione, se una volta ci siamo sentiti ingannati, sarà molto difficile per la fonte recuperare in futuro.
- l'euristica basata sull'intento persuasivo tende a non considerare credibile il dato che si percepisce soffrire di una intenzionale distorsione; subentra quando ci sentiamo manipolati, e corrisponde ad un meccanismo di difesa; questa euristica è tipica della informazione commerciale.

Il lettore curioso potrebbe concepire un questionario da distribuire ai suoi amici o parenti, per capire quali sono tra le precedenti le euristiche più utilizzate nella vita di ogni giorno; io, confesso, nel corso della mia vita volta a volta le ho applicate un po' tutte….

Se la individuazione delle precedenti euristiche è un rilevante passo avanti nella comprensione della qualità dei dati sul web, la ricerca è ancora aperta nella individuazione della correlazione con il profilo utente e la influenza delle euristiche in contesti di comunicazione mediata dalla automazione.

8. L'irragionevole efficacia dei dati

Il titolo di questa sezione riprende alla lettera quello di un articolo del 2009 [Halevi, Norvig, Pereira 2009] in cui tre dei ricercatori più importanti nelle discipline che si occupano dei dati affrontano il problema della qualità dei dati in quella che abbiamo chiamato la grande sfera opaca. Nel seguito sintetizzo il loro articolo, ma consiglio senz'altro il lettore di leggere l'originale, da cui traspare una grande eleganza descrittiva unita a una ancor più grande competenza nella ricerca sperimentale.

L'articolo inizia con il ricordo di uno degli autori dell'articolo, che da studente universitario aveva accesso al Brown Corpus, un catalogo di termini considerato a suo tempo estremamente vasto, consistente di un milione di parole in lingua inglese. Recentemente, Google ha pubblicato un corpo di mille miliardi di parole, insieme al calcolo delle frequenze di tutte le sequenze di parole fino a cinque. Questo immenso corpo di parole (una delle possibili concretizzazioni della nostra sfera opaca) è di qualità di gran lunga inferiore al Brown Corpus, è pieno di frasi incomplete, errori di lessico e grammaticali, ecc, eppure ricomprende al suo interno una infinità di sfumature nella descrizione del comportamento umano, di gran lunga più efficace del Brown Corpus. Questo corpo di parole, sequenze di parole, immagini collegate, video ecc. può servire come base per la costruzione di un modello

completo del linguaggio naturale che risolva i tanti problemi legati al linguaggio, purchè siano disponibili tecniche in grado di comporre il mosaico che costituisce il modello.

I più grandi successi nel machine learning nell'ambito del linguaggio naturale sono stati finora nel riconoscimento del parlato e nella traduzione. Si noti che questi sono task difficili, rispetto ad esempio alla classificazione di documenti; l'aspetto vincente nella soluzione di questi task difficili, sta nel fatto di essere task che hanno a disposizione dati molto diffusi, e quindi portatori di corpus ampiamente disponibili nel Web; si pensi ad esempio alle voci di Wikipedia espresse da testi *più o meno simili* disponibili in diverse lingue.

Nel capitolo 2 quando abbiamo introdotto il Machine Learning supervisionato, abbiamo detto che esso opera su esempi noti (per esempio, coppie di frasi in cui una sia una traduzione di un'altra), così che la tecnica possa apprendere il modello generale a partire da essi. Nell'ambito dei corpus disponibili, i testi non sono annotati, e la annotazione da parte di esseri umani può essere molto onerosa, o, meglio, impossibile vista la loro estensione; quindi i dati di esempio prodotti da esseri umani non sono spesso di dimensioni significative. Ma questo non costituisce un problema, perché le relazioni semantiche tra entità descritte nei testi possono essere inferite automaticamente dalle statistiche sulle interrogazioni; ad esempio, in una base di dati definita sulle tre tabelle Studente, Esame e Corso della Figura 6 del Capitolo 3, se le interrogazioni collegano le tabelle Studente e Corso per il tramite della tabella Esame, possiamo inferire l'esistenza di una relazione semantica tra Studente e corso, che possiamo chiamare con il nome della tabella "ponte", e cioè Esame.

Un'altra opportunità sta nel combinare dati da più tabelle con dati da altre fonti, come Pagine Web non strutturate o query di ricerca Web. Per esempio, è possibile prima trovare classi come Company, e poi istanze di Company come Apple, e da una interrogazione che cerca lo Stock price di Apple, possiamo identificare Stock price come attributo di Company.

Similmente, il grande corpo di lavori di ricerca prodotto nello scorso secolo per realizzare programmi di traduzione di linguaggio naturale, basato su regole molto elaborate relative alle relazioni tra la sintassi e la semantica delle frasi, è stato rapidamente soppiantato da grandi tavole di corrispondenze tra frasi in un linguaggio A e frasi in un linguaggio B, costruite senza alcuna annotazione preliminare, e considerando regole di traduzione solo nei casi in cui sono effettivamente più efficienti, ad esempio per le date e i numeri. E si è visto che gli errori derivanti da dati inaccurati si ripercuotono sul modello prodotto dalla tecnica, ma non in percentuale tale da far degradare eccessivamente l'efficacia del modello.

E' pur vero che il numero di frasi ad es. in inglese è un numero immenso, ma in pratica esistono solo un numero finito di casi, con un ordine di grandezza del miliardo di istanze, trattabile dalle architetture descritte nel Capitolo 3. E per coloro che vedono ancora una lingua modellabile con un numero limitato di regole, bisogna ricordare che una lingua naturale è inerentemente complessa, con centinaia di migliaia di parole, che si arricchiscono dinamicamente di nuovi vocaboli, e un vasto insieme di regole

grammaticali e di eccezioni. E filtrare i casi rari è quasi sempre una cattiva idea, perché strutturalmente il Web consiste spesso di occorrenze che sono individualmente rare, ma collettivamente frequenti.

10.La post-verità

In [Lorusso 2017], cui ci ispiriamo in questa introduzione, viene richiamata la definizione di post-verità dell'Oxford Dictionary, in cui la post-verità viene vista come fare riferimento a circostanze in cui l'oggettività dei fatti è meno influente nel formare la pubblica opinione rispetto all'emotività e le credenze personali. Indica perciò una strategia retorico-persuasiva in cui è prevalente la componente soggettiva e passionale su quella referenziale.

Il problema della verità è investigato da oltre duemila anni in filosofia, e più recentemente nelle scienze cognitive, nelle scienze sociali, in psicologia e nella linguistica. Per Wikiedia, con il termine *verità* (in latino *veritas*, in greco ἀλήθεια) si indica il senso di accordo o di coerenza con un dato o una realtà oggettiva, o la proprietà di ciò che esiste in senso assoluto e non può essere falso. In un testo come questo, parlare di verità da parte mia fa un po' tremare le vene e i polsi.

Il mutato contesto in cui i dati sono rappresentati, comunicati e percepiti nel Web e la rilevanza della comunicazione attraverso le reti sociali, più "calda" di quella fornita, ad esempio, da una intervista su un giornale e quindi da un testo scritto, rende il tema della verità, della post-verità e delle fake news, termine per il quale non faccio neanche un tentativo di definizione tanto è abusato, estremamente arduo da trattare in questo libro. Per cui adotto un approccio "difensivo", indagando, nell'affrontare il problema della verità e della post-verità, quattro diversi punti di vista: l'approccio della informatica, l'approccio ontologico, l'approccio delle scienze cognitive, e l'approccio della filosofia del linguaggio. Esaminiamoli qui di seguito.

9.1 Approccio informatico

Devo ammettere che da quando faccio ricerca nel campo della qualità dei dati, sono stato tentato di dire che un dato è vero quando è aderente al frammento di realtà che rappresenta. Questo cortocircuito in tema di verità presenta due grandi debolezze concettuali di partenza, nel concetto di "aderente a" e nell'affermare che vi sia una diretta corrispondenza tra gli artefatti che noi chiamiamo dati e la realtà attorno a noi. Ma si sa, gli informatici, e io sono tra questi, non si pongono tanti problemi filosofici.

L' approccio informatico cerca di governare grandi problemi con operazioni di riduzione. Invece che ragionare sulla qualità dei dati in termini generali, circoscriviamo il problema, ad esempio, alle reti sociali; invece che ragionare sulla qualità in generale, circoscriviamo il problema a una dimensione di qualità, la credibility. Invece che ragionare su tutti i possibili tipi di dati, concentriamoci su particolari tipologie di dati, i testi scritti, e su particolari tipologie di testi scritti come le recensioni (reviews), l'informazione medico-scientifica, le voci o pettegolezzi (i rumors), i microblogging, pubblicazioni su Internet di piccoli contenuti (brevi messaggi di testo, immagini, video, Mp3, ecc.),

Le tipologie di testi scritti vengono ulteriormente classificate in sottoclassi, come le reviews, suddivise in untruthful, quando forniscono deliberatamente review positive o negative allo scopo di ingannare il fruitore del dato, *review sul brand*, più che sul prodotto o servizio, e *unreview*, che non contengono opinioni e perciò disorientano, ovvero messaggi diffusi attraverso il microblogging, suddivisi in conversation items, che riguardano l'utente e la sua cerchia di amici, e news items, che si riferiscono a informazione più generale, l'informazione medico-scientifica, rumors e altro.

Negli approcci più promettenti (vedi [Viviani 2017] per un survey esaustivo), la *credibility* è vista come una qualità percepita dal ricevente la informazione, non viene definita in modo preciso, ed è vista come composta di molteplici dimensioni, riguardando diverse caratteristiche riferite alla fonte/fonti, le relazioni tra gli utenti nella rete sociale, il messaggio, e il mezzo di trasmissione. La credibility come detto in precedenza, è collegata con *l'expertise* (la conoscenza, lo skill e l'esperienza percepita della fonte della informazione), e la *trustworthiness*, la percezione di quanto una informazione emessa dalla fonte è valida.

I modelli di base utilizzati per la valutazione della credibility sono *data-driven*, che utilizzano tecniche di machine learning per identificare frammenti del messaggio o fonti come credibile o non credibile, la informazione falsa, i modelli *model-driven*, che si basano su decisioni multicriterio e definiscono schemi di aggregazione per arrivare a calcolare una stima aggregata della credibility, e i modelli graph-based, che sfruttano la struttura delle entità connesse nel grafo della rete sociale.

Sia i modelli data driven che quelli *model-driven* si focalizzano su un certo numero di caratteristiche (features), che possono essere estratte sia dal messaggio che dalla fonte; le features possono essere sia linguistiche che facenti riferimento a meta-dati, cioè ulteriori dati associati al messaggio, ovvero, possono fare riferimento al comportamento del soggetto-fonte, alla struttura delle sue relazioni, alla natura del prodotto/servizio associato al messaggio. Tra le features linguistiche distinguiamo le features lessicali, quelle stilistiche e le inconsistenze semantiche. Tra le features comportamentali si distinguono i dati pubblici sulla fonte disponibili sui siti Web, i dati privati come ad esempio gli indirizzi IP, il tempo intercorso per postare un testo, la locazione fisica del fornitore dei contenuti, ecc. Una categorizzazione esaustiva degli approcci alla valutazione della credibility compare in [Viviani 2017].

Gli approcci, sia data-driven che model-driven, sono poi classificabili in content-based, quando sono basati esclusivamente su features lingustiche, e multiple features based, che rimuovono i limiti derivanti dal fatto che considerando solo caratteristiche linguistiche della informazione spesso non è possibile distinguere tra informazione veritiera e informazione falsa.

Gli approcci graph-based, infine, sfruttano la natura delle relazioni tra le entità valutate, estendendosi, tra le entità, alla struttura di connettività degli utenti, dei prodotti/servizi, dei messaggi.

9.2 Approccio ontologico

L'approccio ontologico qui descritto a partire da [Ferraris 2017] fornisce più che metodi e tecniche modelli classificatori e interpretativi. In [Ferraris 2017] l'ipoverità è ciò che è creduto tale da una comunità e che viene corroborato da un insieme di procedure. L'iperverità caratterizza, ad esempio, il contesto per cui la proposizione «la neve è bianca» sarebbe vera anche se non ci fosse mai stato un

essere umano sulla terra in grado di formularla. La *verità* è il risultato tecnologico del rapporto tra ontologia (ciò che esiste) e epistemologia (ciò che conosciamo). Le tre categorie possono essere esemplificate nei seguenti tre ambiti (si veda la Figura 22):

- Ontologico: in questo barattolo ci sono 12 fagioli. L'approccio ontologico fa riferimento a ciò che è, la realtà.
- Epistemologico, enuncio la frase «in questo barattolo ci sono 12 fagioli»; l'epistemologia fa riferimento a ciò che sappiamo, e che esprimiamo attraverso i concetti, il suo scopo è raggiungere la verità.
- Tecnologico: io conto 12 fagioli. L'approccio tecnologico si riferisce a ciò che possiamo fare, misurare, e riguarda perciò la interpretazione della realtà attraverso i fatti (è un fatto che ho contato 12 fagioli). Il metodo di misura che mi ha portato a contare 12 fagioli può risultare erroneo in qualche sua fase, attraverso la ricerca sperimentale posso concepirne un altro e dimostrarne la superiorità e maggiore precisione rispetto al precedete.

E ancora:

- Las frase "il barattolo ha un certo peso" fa riferimento alla ontologia,
- la frase "il barattolo pesa 100 grammi", così come la frase "il barattolo pesa tre once e mezza" fa riferimento alla epistemologia.
- La frase "metto su una bilancia il barattolo per misurarne il peso" fa riferimento alla tecnologia;

Disciplina	**Riguarda la**	**Osserva …**
Ontologia	Realtà	Oggetti
Epistemologia	Verità	Concetti
Tecnologia	Interpretazione	Fatti

Figura 22 - Ontologia, epistemologia, tecnologia

In sintesi, la verità "è relativa rispetto agli strumenti tecnici di verifica, ma assoluta rispetto alla sfera ontologica a cui fa riferimento e all'esigenza epistemologica a cui risponde".

9.3 Approccio cognitivo

Questo approccio parte dalla osservazione ([Metzger 2013]) che l'attività di interpretazione del dato richiede uno sforzo cognitivo, e tale sforzo cognitivo trova i suoi limiti nella idea della razionalità limitata [Simon 1955]. Sebbene il Web abbia ridotto alcuni costi connessi alla ricerca dei dati accrescendo la accessibilità ai dati, rimangono e sono relativamente incomprimibili costi significativi di interazione in virtù della vastissima area di dati disponibili, e in virtù della sua eterogeneità e opacità. Dalla psicologia cognitiva non arrivano dunque buone notizie. Riconoscere la cattiva informazione richiede processi cognitivi complessi. Un semplice mito è più attrattivo cognitivamente di una complicata correzione. Per coloro che sono fortemente convinti delle proprie idee, gli argomenti fortemente contrari possono rafforzare le loro convinzioni. Di conseguenza, non è tanto rilevante ciò che la gente pensa, ma *come* pensa.

La relazione tra psicologia cognitiva, post-verità e fake news è così rilevante che ho ritenuto di chiedere a un esperto del settore, Paolo Cherubini, un contributo monografico, contenuto nella seconda parte del Capitolo 17; ad esso rimando il lettore per un approfondimento su questa problematica.

9.4 Approccio della filosofia del linguaggio

Per [Lorusso 2017] "i media non rappresentano un reale già fatto, che sta da qualche parte nel mondo, i media costruiscono il reale, lo modellano. Gli spazi mediatici sono luoghi di costruzione del reale perché sono i luoghi in cui elaboriamo i modelli con cui poi classifichiamo il mondo e ci muoviamo in esso; da qui l'affermazione: è vero, o è reale solo ciò che passa dalla televisione, affermazione ormai datata, aggiornandola a partire dai nuovi media e reti sociali".

Un tempo c'erano (solo) i giornali; i Social media sono la generalizzazione (qualcuno dice la democratizzazione) delle agenzie di verità. Oggi chiunque sembra autorizzato a produrre non la sua versione del mondo, ma una versione del mondo che pretende di essere vera, che legittima solo la logica esclusiva del vero/falso. Non esiste piu' la coesistenza tollerante delle sfumature

Lorusso analizza criticamente quella che chiama la illusione del fact checking, cioè delle indagini che possono essere effettuate con il supporto di tecniche di ricerca e che hanno lo scopo di verificare la validità di affermazioni o previsioni diffuse via Web. In un mondo di verità moltiplicate la possibilità statistica dell'errore cresce a dismisura; se gli agenti della informazione si moltiplicano a dismisura, e possono essere ovunque, allora il controllo diventa infinitamente più difficile. Più che di verità assolutizzata, ha senso parlare di livelli di verità. In Politifact.com, ad esempio, esistono diversi livelli:

- true, notizie accurate e complete,
- mostly true, notizie che richiedono alcune integrazioni e chiarimenti,
- half true, notizie che trascurano dettagli importanti e decontestualizzano la informazione,
- mostly false, notizie che ignorano punti di vista corrisponde un'altra lettura dei fatti,
- false, notizie non accurate,
- pants on fire: notizie che sostengono cose ridicole.

L'intuizione più rilevante dell'approccio espresso da Lorusso riguarda la caratterizzazione della verità, che non è immagine della realtà, ma costruzione, è dunque un *processo*. Un itinerario verso una consapevolezza della verità insita nei fatti individua tre proprietà dei fatti:

- la *completezza*: nei fatti ci devono essere tutti gli elementi fondamentali
- la *contestualizzazione*, i fatti devono essere messi in relazione con gli elementi cui sono connessi
- la *tenuta*, i fatti devono aver tenuto in conto i punti di vista critici che potrebbero dare un'altra lettura

Si può affermare che problematizzare la verità, vederne la molteplicità, relativizzarne la natura, non significa che tutto va bene, che tutto è verità, ma significa recuperare il concetto di *prova*: non tutte le verità hanno *la stessa tenuta*. I discorsi hanno formazioni, forza, autorevolezza, raggi di applicazione. Guardare a questi elementi significa «provare» la tenuta dei discorsi.

La verità *non può essere raggiunta ma avvicinata*, mettendo alla prova le diverse verità espresse in ragione di interessi diversi, e via via *facendo la realtà*. Tenendo conto della intersoggettività, e in questo senso, dell'etica della interpretazione e della comunicazione.

Per Umberto Eco [Eco 1997]) non c'è dunque una verifica che basti da sola ad assicurare la veridicità di qualcosa; al massimo ci possono essere verifiche incrociate che possono escludere interpretazioni false o errate. Questa posizione è simile alla congettura di Dijkstra relativa alla prova della correttezza dei programmi informatici, per cui eseguendo un programma, non si può mai provare che è corretto, si possono solo trovare nuovi dati in ingresso al programma che dimostrano l'esistenza di errori.

Ricordiamo ancora di Umberto Eco l'affermazione che "I social media danno diritto di parola a legioni di imbecilli che prima parlavano solo al bar dopo un bicchiere di vino, senza danneggiare la collettività. Venivano subito messi a tacere, mentre ora hanno lo stesso diritto di parola di un Premio Nobel".

Nel mondo contemporaneo sta accadendo qualcosa di più rilevante di quanto trattato fino ad ora: sempre di più la costruzione del senso si dà per via *narrativa*. La notizia è sempre meno pensata come documento e sempre più come racconto. A prevalere non è quindi un criterio di attendibilità, ma di efficacia narrativa, chiamata in [Lorusso 2017] credibilità. C'è una profonda relazione tra fatti, favole, fole, bugie; la forza dei nuovi soggetti di informazione si misura più sulla capacità di riuso di ambiti narrativi consolidati che sulla attendibilità della informazione. Sembrerà esagerato, ma la dinamica è la stessa; quando leggo una favola, io ho delle esigenze che mi fanno apprezzare quella favola, e fanno sì che ci creda e mi appassioni.

L'impressione è che le verità si siano più che altro parcellizzate, e nel parcellizzarsi si siano moltiplicate, in un gioco di specchi e rimbalzi. Attraverso la condivisione sempre più estesa hanno acquisito credibilità. La logica di Facebook rafforza l'idea che le condivisioni siano indice di identificazione e adesione, dunque di credibilità.

In conclusione, oggi i soggetti della informazione siamo noi, persone comuni, dalle competenze comuni, il più delle volte con una nostra esperienza da condividere, con un nodo di emozioni da esprimere. In un mondo di verità moltiplicate, assolutizzate l'una rispetto all'altra, e isolate entro comunità chiuse, la cosa più seria che si perde *non è il vero ma il legame sociale*. Saper discriminare la verità significa condividere saperi; e condividere saperi significa essere parte della stessa comunità.

10. Conclusioni

La conclusione tentativa che possiamo tratte dalla precedente discussione è che il Web è una immensa prateria, in cui è difficile individuare elementi di riferimento per trovare la verità, che può essere semmai avvicinata, ma non raggiunta. Possiamo dire che sia un *dovere etico* cercare di avvicinare la verità, ma per farlo c'è bisogno di sforzo cognitivo e, soprattutto mente libera da convinzioni stratificate nel tempo e da ideologie; nel gennaio 2019 la neo deputata democratica Ocasio Cortez, in risposta a una osservazione di un giornalista del Washington Post che le contestava di aver fatto una affermazione imprecisa su un indicatore economico, per forzarlo verso la propria tesi, ha risposto: "I think that there's

a lot of people more concerned about being precisely, factually, and semantically correct than about being morally right". E' una osservazione libera da ideologie, questa?

Per avvicinare la verità, sono necessarie pazienza e perseveranza nella indagine; esattamente il contrario di quanto ci sollecita spesso la Rete, e il principale indizio di verità è la sua *tenuta* nel tempo [Eco 1990]. Possiamo dunque tornare alla immagine comparativa delle inaugurazioni a Washington della Presidenza di Obama e di Trump, per suffragare la policy che possiamo sintetizzare con l'espressione "i dati sono cocciuti". Nella visione che abbiamo sviluppato in precedenza, non potremo mai arrivare alla assoluta certezza che i partecipanti alla cerimonia di Obama fossero più dei partecipanti alla cerimonia di Trump, ma possiamo certamente contestare l'affermazione della collaboratrice di Trump su quelli che definì gli "alternative facts", acquisendo (vedi Figura 23) conoscenza comparativa ulteriore sulle due immagini, come ad esempio l'ora di ripresa delle foto, l'intervallo temporale tra l'ora di ripresa e l'inizio della cerimonia, il numero di biglietti validati sulla metropolitana di Washington in quei due giorni, il numero di auto parcheggiate nei parcheggi limitrofi all'area della cerimonia, avvicinando sempre più la verità fino, per così dire, a toccarla.

Ora della foto: 11.30

Ora della cerimonia: 12

Numero biglietti metropolitana nell'ora precedente: 80.000

Ora della foto: 11.30

Ora della cerimonia: 12

Numero biglietti metropolitana nell'ora precedente: 20.000

Figura 23 – I fatti sono cocciuti (da www.wtop.com)

Riferimenti

C. Batini e M. Scannapieco – Data quality: Concepts, Methodologies and Techniques - Springer, 2006.

C. Batini, A. Rula, M. Scannapieco, G. Viscusi - From Data Quality to Big Data Quality. J. Database Manag. 26(1): 60-82, 2015

C. Batini, C. Cappiello, C. Francalanci, A. Maurino - Methodologies for data quality assessment and improvement. ACM Computing Surveys 41(3), 2009

C. Batini e M. Scannapieco – Data and information quality: Dimensions, Principles and Techniques – Springer, 2016.

C. Dai, D. Lin, D., E. Bertino, E., M. Kantarcioglu - An Approach to Evaluate Data Trustworthiness Based on Data Provenance. Secure Data Management, 2008.

U. Eco – I limiti della interpretazione, Bompiani, Milano, 1997.

M. Ferraris – Post verità e altri enigmi, Il Mulino 2017.

L. Floridi – La quarta rivoluzione: come l'infosfera sta cambinado il mondo, Raffaello Cortina Editore, 2017.

A. Halevy, P. Norvig, e F. Pereira - The Unreasonable Effectiveness of Data, IEEE Intelligent System, 2009.

A.M. Lorusso – La post verità, Il Mulino, 2017.

T. Lukoianova, V. Rubin - Veracity Roadmap: Is Big Data Objective, Truthful and Credible? - Advances In Classification Research Online, 24(1), No. 4, 2013.

M. J. Metzger, A. J. Flanagin - Credibility and trust of information in online environments: The use of cognitive heuristics - Journal of Pragmatics 59, 2013

S. Nakamura, et al. - Trustworthiness Analysis of Web Search Results - ECDL, 2007.

W. Quattrociocchi - Misinformation, Franco Angeli, 2016

S. Y. Rieh, Yong-Mi Kim, Ji Yeon Yang, Beth St. Jean- A diary study of credibility assessment in everyday life information activities on the web: Preliminary findings. ASIST 2010.

J. Searle – Mente, linguaggio, società, Raffaello Cortina Editore, 2000.

W. Shercan, S. Nepal e C. Pari - A Survey of Trust in Social Networks, ACM Computing Surveys, Vol. 45, No. 4, 2013.

H. Simon - A Behavioural model of rational choice – Quart. J. Econ. 69, 1955

M. Viviani e G. Pasi - Credibility in Social Media: Opinions, News, and Health Information - A Survey. WIREs Data Mining and Knowledge Discovery, 2017

A. Zubiaga, A. Aker, K. Bonthceva, M. Liakata and R. Procter - Detection and Resolution of Rumours in Social Media: A Survey - ACM Computing Surveys, Vol. 51, No. 2, 2018.

Capitolo 6 – Integrazione

C. Batini

1. Introduzione

Dal 1993 al 2003 ho lavorato lontano dalla Università, in un ente chiamato Autorità per la informatica nella Pubblica Amministrazione, o AIPA. Aver lavorato per l'AIPA, per una Autorità, per un Ente la cui autorevolezza e terzietà derivava soprattutto dalla professionalità delle persone che ci lavoravano, professionalità che peraltro doveva essere percepita dagli enti da noi coordinati, Ministeri e Enti Pubblici, è stato per me un'onore e una meravigliosa occasione per confrontarmi con grandi progetti.

L'AIPA aveva un compito di indirizzo e controllo sulle Pubbliche Amministrazioni Centrali e sugli Enti Pubblici non Economici, come l'Inps e l'Inail, controllo che esercitava con diversi strumenti quali una analisi sullo stato della Informatica nella Pubblica Amministrazione, un piano annuale sulla evoluzione della Informatica nella PA, e infine i pareri di congruità tecnico economica, in cui valutavamo i singoli progetti che le amministrazioni presentavano, da un punto di vista di coerenza tecnica con la evoluzione delle tecnologie ICT, ed economica, per valutare se i costi esposti fossero ragionevoli o eccessivi. L' AIPA, creata nel 1993 dal Governo Amato, fu poi chiusa nel 2002 e trasformata in un Centro Tecnico dal Governo Berlusconi del tempo. Guido Rey fu per otto anni Presidente, seguito per un anno da Alberto Zuliani. Nell'ultimo anno di esistenza, in quanto componente più anziano in ruolo dell'organo collegiale, fui Presidente facente funzioni.

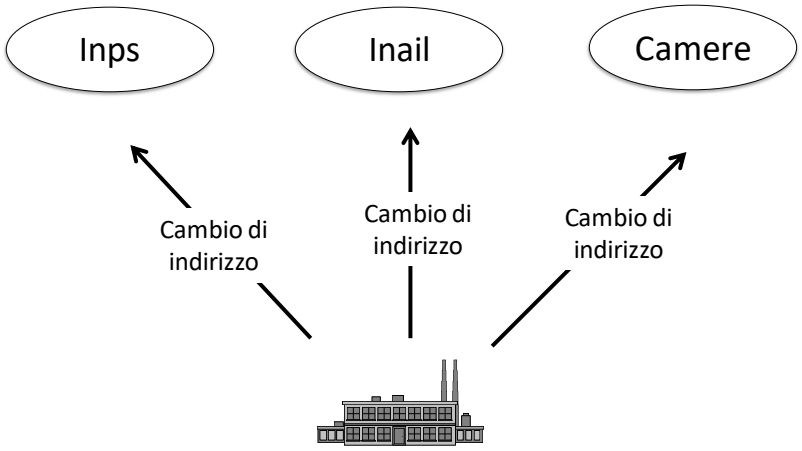
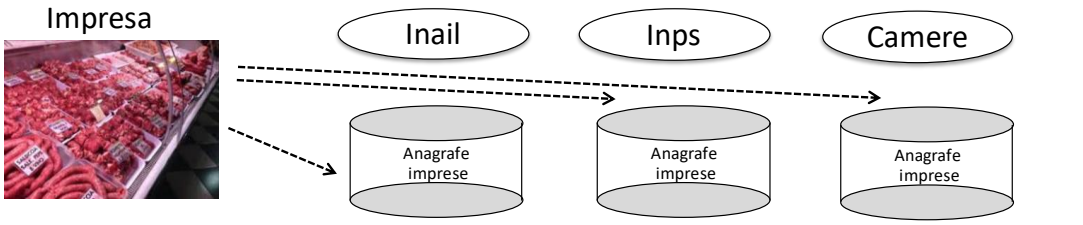


Figura 1 – L'impresa "pony express" deve fornire gli stessi dati di variazione di indirizzo a Inps, Inail e Camere di Commercio

L'AIPA ha assunto negli anni in cui ha operato un ruolo di impulso, proponendo progetti innovativi rispetto allo stato dell'arte nella PA. Uno di questi progetti, il progetto Servizi alle Imprese, lanciato nel 1999 e avente come capo progetto Sandro Osnaghi e me come referente all'organo collegiale, aveva lo scopo di superare le inefficienze di interazione tra imprese e Pubbliche Amministrazioni, che nascono

dalla cosiddetta sindrome della “Impresa pony express”, per cui le imprese si trovavano a fornire alle Pubbliche Amministrazioni, viste unitariamente, dati che esse già possedevano, dando luogo alla tipica interazione che ho esemplificato in Figura 1 per le imprese che interagivano con le Camere di commercio, l’Inps e l’Inail ogni volta che dovevano comunicare un cambio di indirizzo toponomastico. D’ora in poi invece che usare l’imperfetto (interagivano..) userò per non appesantire troppo il discorso il presente (interagiscono).

Questa modalità, per cui le imprese devono farsi carico di mandare un loro addetto presso i tre enti, o, nel caso che via via si sta affermando, interagire con i tre siti Web per comunicare la variazione, è una inutile vessazione per le imprese, e ha, inoltre, delle conseguenze molto dannose, come mostra l’esempio in Figura 2, esempio che fotografa un caso reale, in cui ho modificato il nome della impresa e altri dati per motivi di privacy.



Inail

Identificatore	Nome della Impresa	Tipo di attività	Indirizzo	Comune
CNCBTB765SDV	Produzione carni di Aldo Rossi	Vendita di carni bovine e ovine	Via Ostiglia 25	Roma

Inps

Identificatore	Nome della Impresa	Tipo di attività	Indirizzo	Comune	Numero dipendenti
0111232223	Produzione carni in scatola di Rossi Aldo	Drogheria e bevande	Via Roma 9	Marino	20

Camere

Identificatore	Nome della Impresa	Tipo di attività	Indirizzo	Comune	Fatturato
CND8TB76SSDV	Produzione Marino carni	Macellaio	Piazza Garibaldi 4	Marino	300.000

Figura 2 – Come la stessa impresa “appare” ai tre Enti con le attuali modalità di interazione (caso reale modificato per impedire la identificazione della impresa)

La figura mostra le tre basi di dati delle imprese detenute da Inail, Inps e Camere di Commercio, e le n-ple (record nel seguito) delle tre basi di dati che corrispondono ad una stessa impresa; questa impresa nella realtà svolge una attività di macelleria. I tre record ci raccontano una storia quasi inverosimile; i tre nomi della impresa, il tipo di attività, la sede sembrano fare riferimento a imprese diverse; il solo campo “nome della impresa” manifesta in qualche parola delle similitudini, la parola “produzione”, il nome del proprietario, che peraltro compare in due soli nomi di impresa, e il termine “carni”, anche

esso presente in due delle tre denominazioni. Eppure, vi posso assicurare si tratta della stessa impresa. Immaginatevi di NON sapere a priori che si tratta della stessa impresa; a quale conclusione arrivereste? Probabilmente le similitudini presenti nel nome non sarebbero assolutamente sufficienti a far arrivare alla conclusione che si tratti della stessa impresa.

C''e una ulteriore diversità tra i tre record, e sta nell'identificatore. In ogni tabella di una base di dati, che rappresenti persone, imprese, luoghi, eventi, prodotti, è sempre importante che vi sia un attributo o un insieme di attributi che indentificano univocamente l'oggetto del mondo reale rappresentato dal record, nel nostro caso una impresa. Per essere più precisi, ogni tabella di una base di dati, come abbiamo già osservato ed esemplificato nel capitolo sui modelli, descrive un frammento di mondo. Chiameremo nel seguito "osservabile" o "osservabili" tutti i fenomeni del mondo reale che vengono rappresentati in un record di una tabella. Ebbene, nei tre record, e nelle tabelle cui appartengono, gli osservabili sono le imprese, e gli identificatori associati sono di due tipi

- Codice fiscale/partita iva per l'Inail e le camere di commercio
- Indice numerico per l'Inps.

Inoltre, pur avendo lo stesso identificatore, il codice fiscale/partita iva, i valori che l'identificatore assume all'Inail e alle Camere di Commercio differiscono.

Perché la situazione di disallineamento tra i tre record che rappresentano la stessa impresa è in questo stato così critico, o, come diremmo dopo aver letto il Capitolo 5, di così scarsa qualità? Per saperlo, dovremmo sapere cosa è accaduto quando i dati sono stati inseriti nelle tre basi di dati, ma ricostruire ciò è impossibile. Possiamo fare delle ipotesi; ad esempio, possono essere stati commessi degli errori di inserimento per i due valori del Codice Fiscale, l'impresa potrebbe essersi scordata di comunicare una variazione di tipologia di attività, ovvero una variazione di indirizzo. E' difficile convincersi che possano essere stati commessi tanti errori, ma se ricordiamo che le imprese in Italia sono circa 5.000.000, allora statisticamente tutto può accadere. E' facile convincersi, in ogni caso, che il meccanismo utilizzato per comunicare e inserire i dati dei record ha qualcosa di profondamente sbagliato.

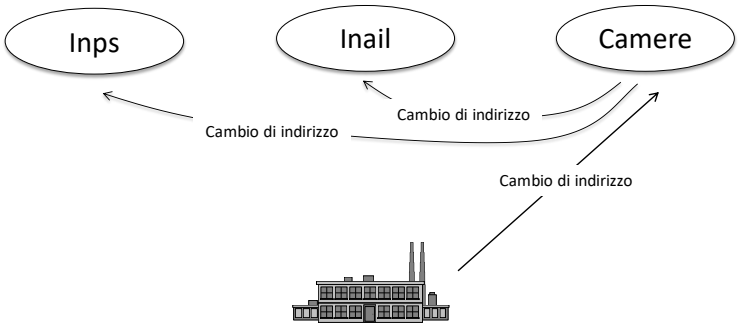


Figura 3 – La nuova modalità in cui la impresa fornisce una sola volta i dati di variazione, e sono le Camere di Commercio che comunicano i dati a Inps e Inail

Infatti, se invece di comunicare tre volte gli stessi dati, la impresa li comunicasse una sola volta, ad esempio alle Camere di Commercio, che, tra l'altro, per legge detengono e gestiscono il Registro delle Imprese, che costituisce in Italia il registro ufficiale sulle imprese, si potrebbe immaginare una

comunicazione tra gli enti coinvolti più razionale e affidabile, mostrata in Figura 3. In questo caso, l'ente deputato ad acquisire la variazione sull'indirizzo sono le Camere di Commercio, che, una volta memorizzata la variazione, la possono comunicare in forma affidabile all'Inps e all'Inail, eliminando alla radice i precedenti casi che generano errori; tutte le basi di dati sarebbero allineate, come dire, per definizione.

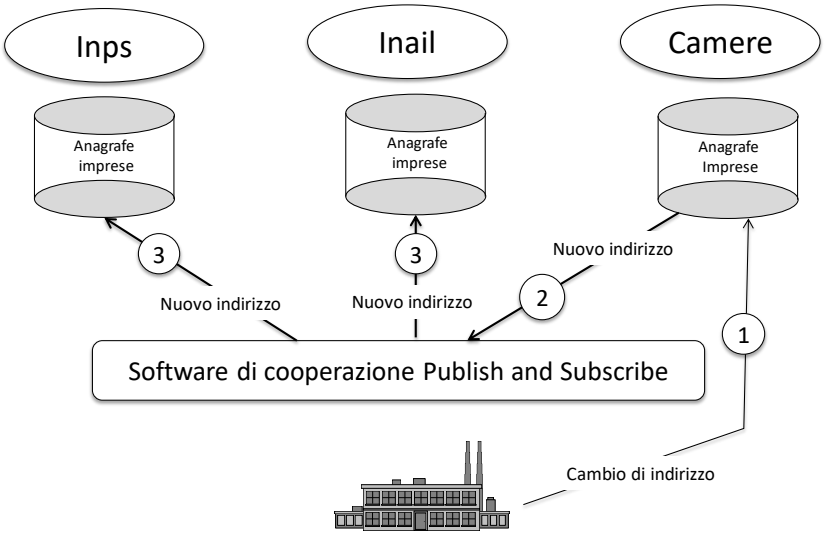


Figura 4 – Per permettere la comunicazione diretta occorre adottare un software di cooperazione di tipo publish and subscribe

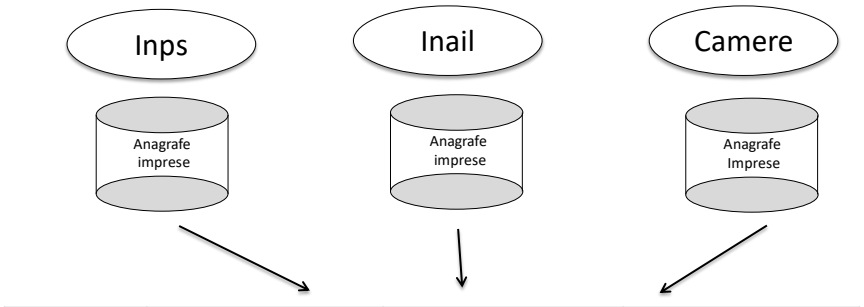
Restano due problemi da risolvere.

- Il primo riguarda la nuova modalità di interazione tra i tre enti. E' possibile risolvere questo aspetto utilizzando un software prodotto una volta per tutte, della tipologia chiamata middleware, perché si interpone tra le applicazione degli utenti (in questo caso le tre applicazioni che gestiscono le basi di dati) e la infrastruttura tecnologica sottostante: in particolare si può utilizzare un middleware di tipo Publish and subscribe, vedi Figura 4, in cui le Camere pubblicano la informazione di variazione e Inps e Inail sottoscrivono, come fosse una abbonamento, l'aggiornamento.
- Il secondo problema sta nel fatto che quando viene inviata la informazione di variazione sulla impresa, il sistema delle Camere di Commercio non conosce l'indirizzo fisico della stessa impresa nelle basi dati dell'Inail e dell'Inps, anzi, come mostra l'esempio di Figura 2, non hanno la minima idea di come l'impresa sia rappresentata nelle basi di dati, e, in particolare, quale valore assuma l'identificatore adottato nelle due basi di dati, sia esso un codice fiscale o un identificatore numerico.

Per poter risolvere il secondo problema, occorre capire per ogni impresa quali siano gli identificatori della *stessa* impresa e i relativi indirizzi di memoria nelle tre basi di dati, operazione che viene chiamata di record linkage. In Figura 5 vediamo quale struttura debba avere la nuova tabella di collegamento. Se riusciamo a creare questa tabella allora la procedura di aggiornamento dei tre indirizzi può procedere nel seguente modo:

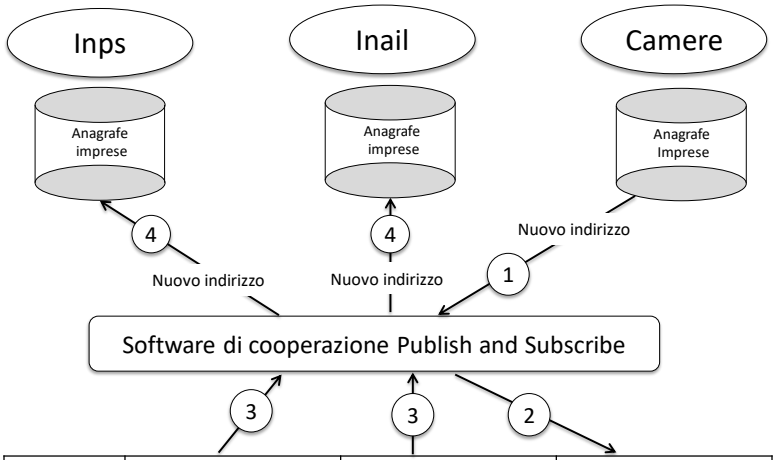
1. Dapprima l'impresa comunica il cambio di indirizzo toponomastico alle Camere di commercio che aggiornano la loro base di dati.

2. Le Camere di commercio inviano la informazione di variazione di indirizzo toponomastico dell'impresa al software di Publish and subscribe.
3. Il software di Publish and subscribe partendo dall'identificatore della impresa delle Camere di commercio cerca il relativo record, e trova i due identificatori (e relativi indirizzi in memoria) dell'Inps e dell'Inail.
4. L'informazione di variazione viaggia nella rete e aggiorna automaticamente negli indirizzi in memoria "giusti" i dati nelle due basi di dati di Inps e Inail.



Impresa	Base di dati Inps	Base di dati Inail	Base di dati Camere
Impresa 1	Indirizzo identificatore impresa 1	Indirizzo identificatore impresa 1	Indirizzo identificatore impresa 1
Impresa 2	Indirizzo identificatore impresa 2	Indirizzo identificatore impresa 2	Indirizzo identificatore impresa 2
……	……	…….	…….
Impresa 5.000.000	Indirizzo id. impresa 5.000.000	Indirizzo id. impresa 5..000.000	Indirizzo id. impresa 5.000.000

Figura 5 – Creazione di una tabella in cui "accoppiare" i tre valori degli identificatori per ogni impresa



Impresa	Base di dati Inps	Base di dati Inail	Base di dati Camere
Impresa 1	Indirizzo identificatore impresa 1	Indirizzo identificatore impresa 1	Indirizzo identificatore impresa 1
Impresa 2	Indirizzo identificatore impresa 2	Indirizzo identificatore impresa 2	Indirizzo identificatore impresa 2
……	……	…….	…….
Impresa 5.000.000	Indirizzo id. impresa 5.000.000	Indirizzo id. impresa 5..000.000	Indirizzo id. impresa 5.000.000

Figura 6 – Il messaggio inviato deve conoscere l'indirizzo fisico della impresa nelle basi dati dell'Inps e dell'Inail: è quindi necessario effettuare il record linkage tra le tre basi di dati per collegare gli indirizzi

Sembra l'uovo di Colombo! Eh, ma anche l'uovo va scoperto! Le architetture di Publish and Subscribe esistevano all'epoca del progetto, l'aspetto interessante sa nel fatto che siano state scelte tra le altre con lo scopo di modificare il meno possibile l'assetto tecnologico e normativo della Pubblica Amministrazione Centrale. Si sarebbe potuto consolidare le tre basi di dati in una unica, ma questo avrebbe leso i regolamenti che disciplinavano la gestione delle informazioni sulle imprese di responsabilità dei tre Enti.

Ho un aneddoto che testimonia la innovatività del progetto di cooperazione che veniva condiviso tra i enti; volendo verificare la efficacia di una prima realizzazione sperimentale, ci vedemmo con i tecnici di uno degli enti coinvolti, e chiesi come andavano le cose; mi fu risposto che c'era qualche problema. Quale problema? Beh, quando ci arriva una informazione di variazione, noi dobbiamo stamparla e poi reinserirla a mano nella base di dati. Ma perchè la ristampate! Che senso ha? Caro professore, abbiamo un regolamento che lo stabilisce, soltanto noi possiamo inserire dati che provengono dall'esterno….Dovemmo cambiare il regolamento.

Il problema che abbiamo visto, cioè la necessità di confrontare informazioni in diverse basi di dati che rappresentano lo stesso oggetto, si presenta spessissimo nelle basi di dati. In Figura 7 vengono mostrati i due casi che si presentano usualmente nelle basi di dati:

1. Un oggetto "osservabile" del mondo può essere rappresentato mediante due record e relativi schemi in due basi di dati diverse. Il processo di rappresentazione dà luogo a due oggetti, e occorre a posteriori ricostruire se essi provengano o mano da uno stesso osservabile
2. Due osservabili diversi sono rappresentati mediante lo stesso oggetto/record nella base di dati, ovvero con oggetti che differiscono in pochi particolari. In questo secondo caso occorre capire se i due oggetti e relativi schemi facciamo riferimento allo stesso oggetto, come induce a pensare la loro somiglianza, ovvero a oggetti diversi.

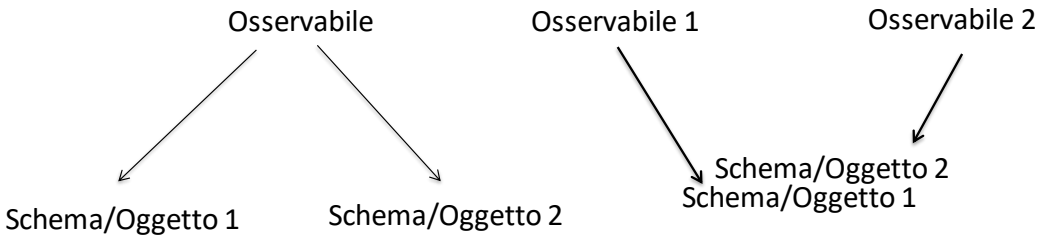


Figura 7 – Gli "osservabili" e la loro rappresentazione come dati digitali

Il problema della frammentazione della realtà in un gran numero di basi di dati diverse, in cui si possono presentare i due casi di Figura 7, è diffusissimo; ogni organizzazione può avere decine e spesso centinaia di basi di dati diverse, in cui gli stessi osservabili sono rappresentati più volte con schemi e record diversi. Ciò deriva dal fatto che le basi di dati vengono realizzate nel corso del tempo, non tutte insieme, e da diversi progettisti, ognuno dei quali adotta un punto di vista diverso. Al contrario, la visione che si dovrebbe avere del patrimonio informativo gestito dalla organizzazione dovrebbe essere unitaria. In fondo le basi di dati sono nate per questa esigenza organizzativa prima ancora che tecnologica, cioè rappresentare ogni osservabile una sola volta nelle basi di dati della organizzazione. Ma non è mai così.

Per queste ragioni la operazione di record linkage ha lo scopo di riconciliare le diverse rappresentazioni degli osservabili. Ad esempio, in Figura 8 nella parte superiore si vedono quattro basi di dati di una Università, in cui il personale è rappresentato con due basi di dati distinte, e a parte sono rappresentate l'offerta formativa e i prodotti della ricerca.

Con una "architettura dei dati" così concepita risulta difficile per la Università poter gestire unitariamente tutto il personale, per sapere ad esempio quanta parte degli stipendi sia destinata a personale docente e quale a personale tecnico amministrativo, quale sia la età media del personale ecc. Difficile ma non impossibile, purché si scoprano tutte le diversità di rappresentazione nelle basi di dati; ciò, tuttavia, è costoso, e ha il difetto di lasciare le cose come stanno, per cui ad ogni nuova esigenza si ripropone il problema della riconciliazione dei dati. L'unica soluzione che risolve il problema alla radice è riunificare le due basi di dati in una unica base di dati del personale.

Analogamente, nelle tre basi di dati della anagrafe docenti, della produzione della ricerca e della offerta formativa noi rappresentiamo *chi* sono i docenti e *cosa* fanno, nella ricerca e nella didattica. Ma anche qui tutto ciò è rappresentato in forma frammentata; per superare il problema, possiamo consolidare le tre basi di dati in una unica. Notate che nella nuova architettura i dati sui docenti sono rappresentati due volte, riproponendo il problema della doppia rappresentazione; esistono soluzioni che permettono di superare questi problemi, che vanno al di là degli scopi del presente capitolo.

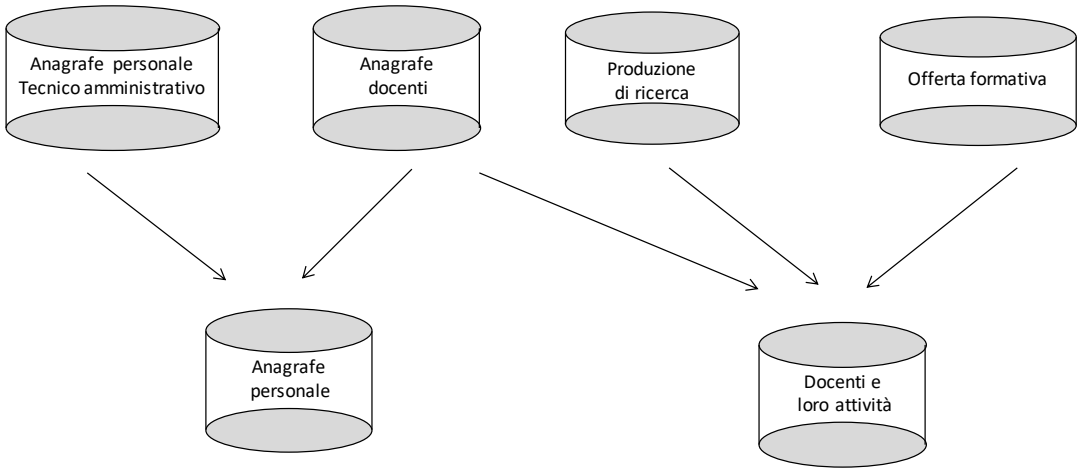


Figura 8 – Integrazione come riunificazione di basi di dati separate

Con questo esempio abbiamo concluso questa prima sezione introduttiva. Nelle prossime sezioni affronteremo in maggiore dettaglio le due fasi di cui si compone un processo di integrazione di due basi di dati, vedi la Figura 9:

1. Il *record linkage*, in cui si individuano gli osservabili comuni nelle due basi di dati, e i record con cui sono rappresentati. Si noti che uno stesso osservabile può essere rappresentato più di una volta in una stessa base di dati o tabella, in questo caso la individuazione dei record relativi ad uno stesso osservabile è chiamata con il termine di deduplicazione.

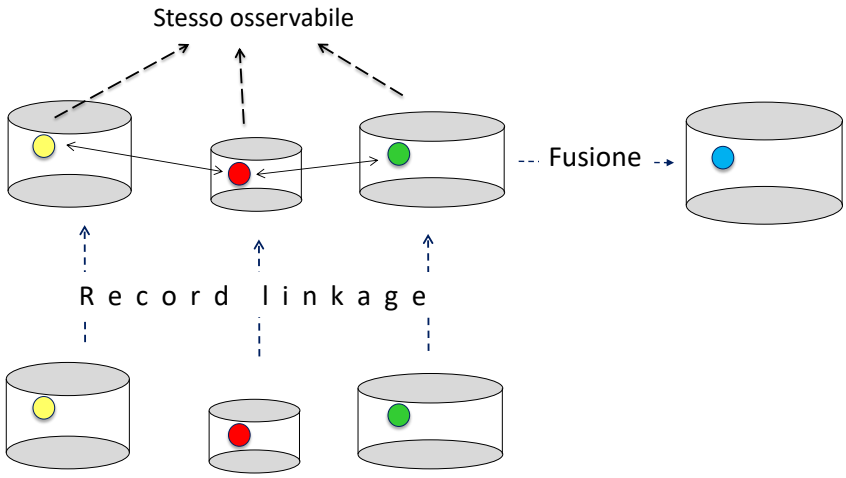


Figura 9 – Le due fasi della integrazione: il record linkage e la fusione

2. La *fusione*, in cui se ad un osservabile sono associati n record R1, R2, .., Rn, ciascuno definito in termini di uno stesso insieme di attributi A1, A2, …, Am, tali record sono sostituiti da un unico record Rint, e per tale record si scelgono volta a volta i valori di A1, A2, …, Am tra gli n record con apposite strategie di scelta che discuteremo più avanti nel capitolo.

Le due operazioni di record linkage (con la variante della deduplicazione) e di fusione operano sui record e corrispondno perciò alla integrazione dei dati, o data integration. Accanto alla integrazione dei dati possiamo essere interessati alla integrazione degli schemi, tema di cui non ci occupiamo in questo capitolo, il lettore interessato puà leggere [Batini et al. 1992].

Nelle prossime sezioni approfondiamo le problematiche che abbiamo evidenziato nello studio di caso Inps-Inail-Unioncamere. La Sezione 2 è dedicata al record linkage nei suoi aspetti generali, mentre la Sezione 3 si concentra sull'aspetto più rilevante del record linkage, la individuazione e misura delle funzioni di distanza tra dati candidati al linkage; scopriremo che queste funzioni di distanza devono adattarsi alla particolare struttura e dominio di definizione dei dati da collegare. La Sezione 4 si concentra su una particolare forma di integrazione, quella sui dati spaziali o territoriali; poiché questi dati rappresentano uno spazio a due dimensioni, si può immaginare che la integrazione presenti aspetti più complessi rispetto al caso dei dati, ad esempio, costituiti da stringhe di caratteri alfabetici. La Sezione 5 riguarda la fusione dei dati; una volta che ho deciso che due dati, ad esempio due tuple che descrivono il nome, il cognome, e la data di nascita, corrispondono allo stesso osservabile, come faccio a "metterli insieme", a fonderli in un unico record? La Sezione 6 applica i metodi visti fino a questo punto al caso di studio delle imprese con cui abbiamo iniziato il capitolo. La Sezione 7 è un po' particolare….applica i metodi di integrazione e fusione al contratto e alla azione di Governo della coalizione tra Lega e Movimento 5 Stelle, con qualche considerazione dedicata anche al Governo tra Movimento 5 Stelle e Partito Democratico che al momento di dare alle stampe questo libro sta emettendo i primi vagiti.

2. Il record linkage

Una trattazione esaustiva del record linkage e delle tecniche sviluppate negli anni recenti è fuori della portata di questo capitolo; pensiamo solo che nel libro [Batini Scannapieco 2016] al record linkage sono dedicate circa 100 pagine. Nella più ampia generalità, date due tabelle come in Figura 10, e due insiemi di osservabili O1 e O2 rappresentati mediante record nelle due tabelle, il problema del record linkage corrisponde a confrontare ogni coppia di record delle due tabelle, e stabilire se essi rappresentano uno stesso osservabile della parte comune ad O1 e O2, ovvero rappresentano due osservabili distinti di O1 e O2.

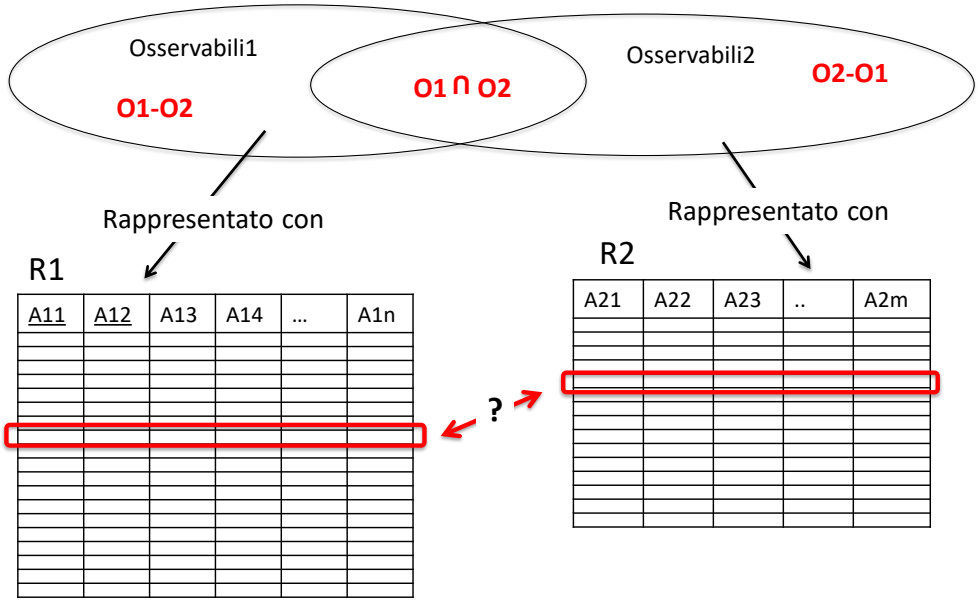


Figura 10 – Il problema da risolvere nel record linkage

Per poter stabilire se una coppia di record appartiene o non appartiene alla parte comune, e quindi corrisponde a un unico osservabile, dobbiamo prendere in considerazione i due record e sulla base della conoscenza disponibile stabilire il loro grado di *similitudine* o, all'opposto, di *diversità*. La similitudine e il suo complemento, la diversità, sono chiamati con il termine *distanza* nella letteratura informatica.

Come facciamo a misurare la distanza tra due record? Dobbiamo anzitutto esaminare la loro struttura. In Figura 11 mostriamo tre record dell'esempio Inps, Inail e Camere, questa volta con valori più simili rispetto ai precedenti. Chiaramente nel momento in cui confrontiamo i tre record per trovare le similitudini e le differenze, possiamo farlo solo sugli attributi comuni, rappresentati in grigio in Figura 11. In effetti, Inail e Camere di Commercio hanno in comune anche l'identificatore; si può immaginare che nella ricerca delle similitudini noi confrontiamo prima i record di Inail e Camere, che hanno più attributi comuni, e poi confrontiamo il record dell'Inps con gli altri due record.

Comunque, sembra ragionevole che il concetto di distanza sia ora applicato in modo diverso, ad esempio, al nome della impresa e al tipo di attività; nel primo caso non esistono restrizioni al nome,

mentre nel secondo caso probabilmente i valori seguono una classificazione, adottata dall'Istat nelle indagini sulle imprese.

Il problema di valutare la distanza è molto vasto ed è trattato ampiamente nella letteratura, per cui ad esso dedichiamo più avanti una sezione a parte. Qui supponiamo di essere riusciti ad esprime attraverso un valore numerico la distanza tra due record, arrivando alla rappresentazione di Figura 12.

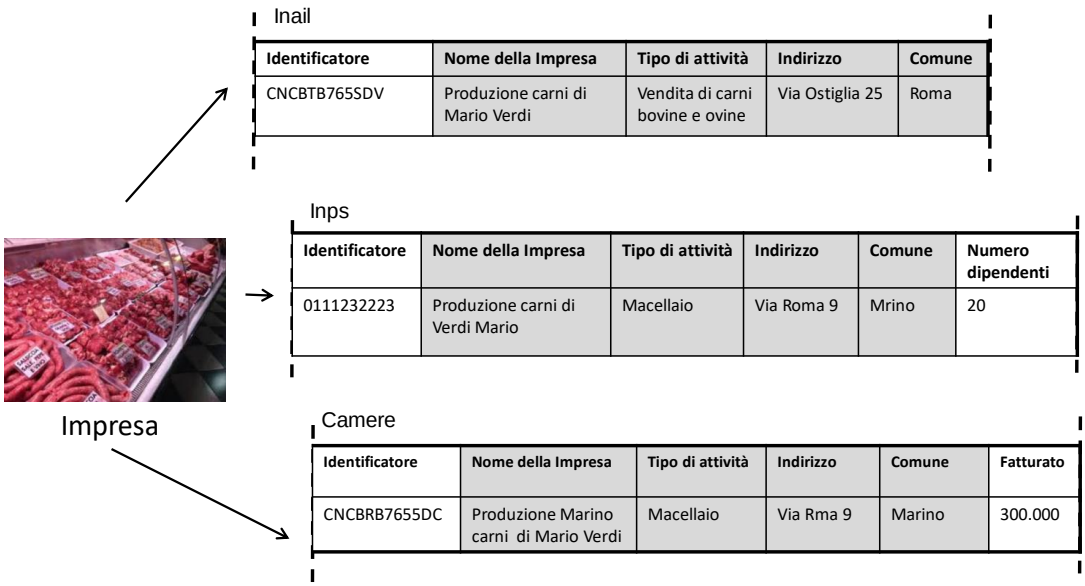


Figura 11 – I record che rappresentano le imprese nelle tre basi di dati vanno confrontati sugli attributi comuni

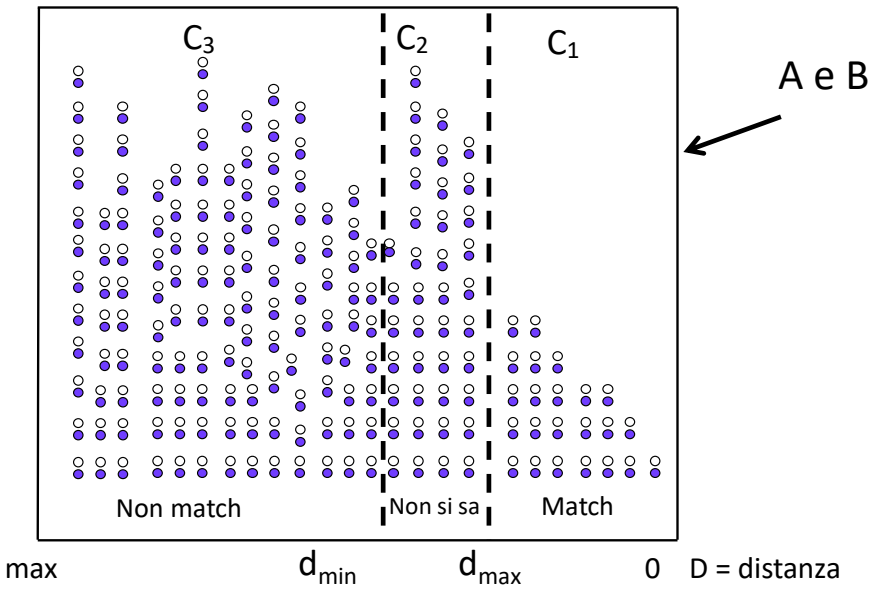


Figura 12 – La coppie di record rappresentate in accordo alla distanza

Nella Figura 12 sono rappresentate con pallini bianchi e neri coppie di record relativi a due tabelle A e B; le coppie sono collocate nel disegno secondo un asse verticale a seconda del loro valore di distanza, che va dal valore 0 ad un valore massimo. Il metodo di decisione consiste nello stabilire due soglie dmin e dmax, per cui:

- tutte le coppie aventi distanza minore di dmin sono assunte corrispondere allo stesso osservabile (si parla in questo caso di *match* tra le due coppie),
- tutte le coppie aventi distanza maggiore di un dato valore dmax sono assunte non corrispondere allo stesso osservabile (*non match*) e
- per quelle con distanza compresa tra i due valori in sostanza *non si decide*.

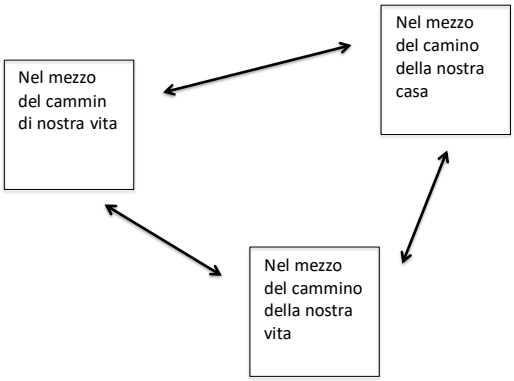
I metodi che si ispirano alla precedente strategia sono detti basati su errori, perché nella procedura di decisione, pur di arrivare a una qualche conclusione, si accetta un tasso di errore. Tale tasso di errore può essere calcolato producendo la distribuzione delle distanze per un insieme di coppie di record per cui sia nota la corrispondenza "match" e "non match", e valutando a questo punto il tasso di errore indotto dalla scelta di dmin e dmax; tale tasso di errore si riferisce ai *falsi positivi*, cioè alle coppie di record che non corrispondono alla stessa impresa, ma che sono state dichiarate "match" e ai *falsi negativi*, l'opposto. Quanto alle coppie all'interno dell'intervallo dmin e dmax, ciò che si può fare è acquisire nuova conoscenza attraverso navigazione nel Web, telefonate, approfondimenti, fino a creare nuovi match, e rinunciare agli accoppiamenti nei casi più complessi.

3. Il concetto di distanza

Nella vita siamo abituati a osservare il mondo e a cercare continuamente, a seconda della nostra attitudine mentale, similitudini o differenze. I temi della ricerca della distanza e della integrazione sono trattati oltre che dalla informatica, dalle scienze sociali, dalle scienze cognitive e dalla stessa filosofia. Il concetto di integrazione è spesso visto come un valore positivo (o osteggiato) nel grande problema dei migranti; in altri campi, nelle persone, nelle opere d'arte, negli stili letterari, noi continuamente cerchiamo similitudini o differenze (guarda come quel ragazzo assomiglia al padre…secondo me assomiglia più alla madre….).

In Figura 13 noi vediamo due casi in cui ci interessa trattare il problema di ciò che abbiamo chiamato distanza; nel caso 1 abbiamo tre testi che hanno parti del discorso simili: ma "quanto" simili?. Il caso 2 è quello che stiamo indagando. Le metriche che da più tempo sono utilizzate per esprimere il concetto di distanza sono quelle relative a dati espressi da stringhe di caratteri. Nel caso 2 di Figura 13 i record delle due tabelle hanno due campi in comune i cui valori corrispondono a stringhe di caratteri, i campi Nome e Cognome. Possiamo confrontare le stringhe di caratteri corrispondenti a Nome, a Cognome, e all'insieme Nome + Cognome, avendo la avvertenza di invertire le stringhe nella seconda tabella. In tutti i casi precedenti possiamo utilizzare per la distanza la edit distance, già introdotta informalmente nel Capitolo 5, Sezione 2. La edit distance (ED), corrisponde al numero di caratteri che dobbiamo cancellare, aggiungere, sostituire per trasformare la prima stringa nella seconda. Cosi' ad esempio

ED <Carlo, Crlo> = 1
ED <Ratti, Reti> = 2.



Caso 1 – Tre testi

CF	Cognome	Nome	Data Nascita
BTNCRL	Batini	Carlo	3-6-2000
RTTABC	Ratti	Gino	7-5-1997

⟷

Nome	Cognome	Città Nascita	Regione N.
Crlo	Batini	Pscara	Abruzzo
Gino	Retti	Roma	Lazio

Caso 2 – I record di due tabelle

Figura 13 – Cosa è la distanza? Cosa è la somiglianza?

Figura 14 – Inserimento dei caratteri che suggerisce una modifica della formula dell'edit distance

Si noti che esistono situazioni in cui possiamo modificare la formula della edit distance per tener conto di particolari tipi di errori. Ad esempio nel caso di Figura 14 può accadere che chi digita le lettere tenda a commettere continuamente l'errore di selezionare il tasto A al posto del tasto S; in questo caso possiamo modificare la formula dando un peso maggiore alla sostituzione S → A rispetto alle altre.

La Edit distance, come in fondo dimostra l'ultimo caso visto in precedenza, è troppo generale per essere in grado di trattare una vasta casistica. In particolare la edit distance mostra i suoi limiti quando vogliamo confrontare insiemi di parole, come accade nel caso di Figura 15.

Nomi di aziende 1

...	...	...
AT&T	...	...
IBM Corporation		
...	...	...

Nomi di aziende 2

...	...	...
AT&T Corporation	...	...
IBM Corporation		

Figura 15 – Un caso in cui la edit distance non va bene

Osservando le due tabelle, Intuitivamente noi facciamo corrispondere ad AT&T Corporation la stringa AT&T, e non IBM Corporation. Invece se calcoliamo la edit distance otteniamo:

ED <AT&T, AT&T Corporation> = 12
ED <IBM Corporation, AT&T Corporation> = 4

Considerati come stringhe di caratteri, sono dunque molto più simili IBM Corporation e AT&T Corporation rispetto a AT&T e AT&T Corporation. Quale è la ragione di questo risultato così contro intuitivo? La ragione risiede nel fatto che quando noi confrontiamo le due stringhe di caratteri non le vediamo come stringhe ma come sequenze di parole, e, viste come sequenze di parole, riconosciamo AT&T come abbreviazione di AT&T Corporation. Per cogliere questo aspetto, possiamo utilizzare la distanza di Jaccard, che calcola la distanza in due passi:

1. Prima si trasformano le stringhe in insiemi di parole:

AT&T Corporation → <AT&T, Corporation> (insieme P1)
IBM Corporation → <IBM, Corporation> (insieme P2)

2. A questo punto la metrica di distanza di Jaccard e' $Dist_{Jaccard}$ = 1 - numero parole (P1 ∩ P2)/numero parole (P1 ∪ P2)

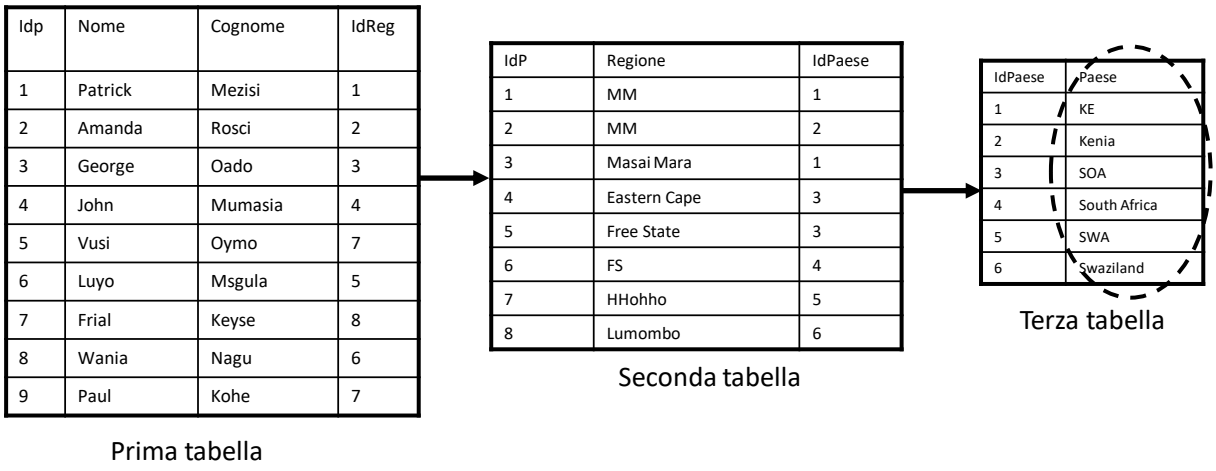
dove P1 ∩ P2 è l'insieme di parole comuni a P1 e P2 e P1 ∪ P2 è l'insieme delle parole di P1 a cui si aggiungono le parole di P2.

E' facile vedere che:
$Dist_{Jaccard}$ <AT&T, AT&T Corporation> = 1 - ½ = 0,5
$Dist_{Jaccard}$ <IBM Corporation, AT&T Corporation> = 1 - 1/3 = 0,66

Si noti che la distanza di Jaccard funziona molto bene anche nel caso delle inversioni, come ad esempio Distanza <Carlo Batini; Batini Carlo>. Anche in questo caso la Edit Disance assume valore molto alto, mentre invece $Dist_{Jaccard}$ <Carlo Batini; Batini Carlo> = 1 – 1 = 0, le due stringhe di caratteri, viste come insiemi di parole, coincidono, perché nel concetto di insieme non c'è una relazione di ordine tra elementi dell'insieme.

Il concetto di distanza cambia a seconda del modello con cui sono rappresentati I dati. Similmente, nella nostra esperienza quotidiana, il confronto tra oggetti e artefatti si modifica se li consideriamo in

isolamento, o nel contesto in cui essi operano o sono collocati. Consideriamo la struttura dati di Figura 16. I dati rappresentano persone che vivono in Africa (prima tabella), le regioni degli stati in cui risiedono (seconda tabella) e infine nella terza tabella gli stati in cui sono collocate le regioni. Le regioni nella seconda tabella e gli stati nella terza sono logicamente legati alle tabelle precedenti per mezzo di valori. In questo caso siamo interessati alla deduplicazione nella terza tabella, quella degli stati.



Idp	Nome	Cognome	IdReg
1	Patrick	Mezisi	1
2	Amanda	Rosci	2
3	George	Oado	3
4	John	Mumasia	4
5	Vusi	Oymo	7
6	Luyo	Msgula	5
7	Frial	Keyse	8
8	Wania	Nagu	6
9	Paul	Kohe	7

Prima tabella

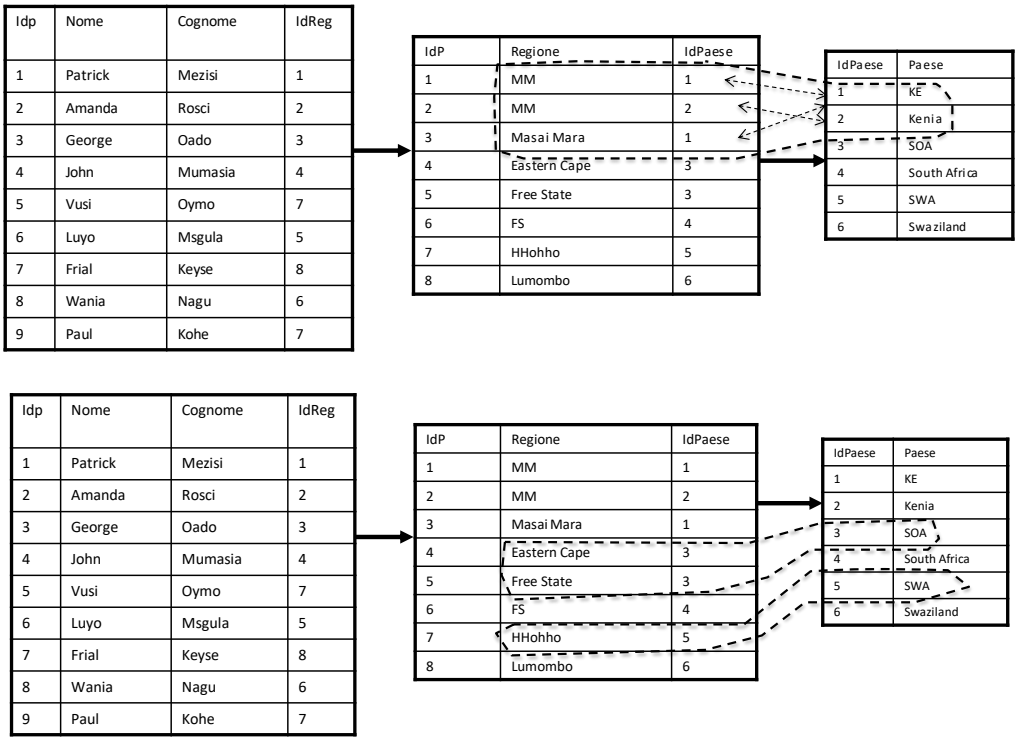
IdP	Regione	IdPaese
1	MM	1
2	MM	2
3	Masai Mara	1
4	Eastern Cape	3
5	Free State	3
6	FS	4
7	HHohho	5
8	Lumombo	6

Seconda tabella

IdPaese	Paese
1	KE
2	Kenia
3	SOA
4	South Africa
5	SWA
6	Swaziland

Terza tabella

Figura 16 – Quando cambia la struttura dati, cambia il concetto di distanza



Idp	Nome	Cognome	IdReg
1	Patrick	Mezisi	1
2	Amanda	Rosci	2
3	George	Oado	3
4	John	Mumasia	4
5	Vusi	Oymo	7
6	Luyo	Msgula	5
7	Frial	Keyse	8
8	Wania	Nagu	6
9	Paul	Kohe	7

IdP	Regione	IdPaese
1	MM	1
2	MM	2
3	Masai Mara	1
4	Eastern Cape	3
5	Free State	3
6	FS	4
7	HHohho	5
8	Lumombo	6

IdPaese	Paese
1	KE
2	Kenia
3	SOA
4	South Africa
5	SWA
6	Swaziland

Idp	Nome	Cognome	IdReg
1	Patrick	Mezisi	1
2	Amanda	Rosci	2
3	George	Oado	3
4	John	Mumasia	4
5	Vusi	Oymo	7
6	Luyo	Msgula	5
7	Frial	Keyse	8
8	Wania	Nagu	6
9	Paul	Kohe	7

IdP	Regione	IdPaese
1	MM	1
2	MM	2
3	Masai Mara	1
4	Eastern Cape	3
5	Free State	3
6	FS	4
7	HHohho	5
8	Lumombo	6

IdPaese	Paese
1	KE
2	Kenia
3	SOA
4	South Africa
5	SWA
6	Swaziland

Figura 17 – Dalla distanza tra record alla distanza tra co-occorrenze di record

Se osserviamo la tabella degli stati di Figura 16 e applichiamo la metrica della edit distance, chiaramente sono molto più simili SOA e SWA, e poco simili KE e Kenia. In effetti basta ragionare un attimo sulla

forma delle stringhe per arrivare alla conclusione che in alcuni casi le stringhe sono nomi di stati, in altri appaiono come acronimi o abbreviazioni. Questa osservazione è ben sostanziata dalla seguente idea: invece che considerare i nomi degli stati in isolamento, perché non li consideriamo nel contesto delle regioni che in essi sono collocate? Seguendo questo approccio, vediamo (Figura 17) che gli stati con nomi KE e Kenia hanno regioni in comune (vedi la parte superiore della Figura 17), mentre gli stati con nome SOA e SWA fanno riferimento a regioni diverse (parte inferiore della Figura 17).

Considerare il contesto è fondamentale nella comunicazione tra esseri umani come nella scoperta di similitudini e differenze, diverse tecniche semantiche che che verranno discusse nel Capitolo 8 applicano proprio questa filosofia). Come abbiamo anticipato nel Capitolo 1, se in due documenti compare un nome “Roma”, per capire se si fa riferimento alla capitale d’Italia o a un quartiere di Buenos Aires, potremo considerare le parole che contornano i due termini Roma, che chiamiamo parole di contesto, associarli alla loro posizione in lessici o in Wikipedia sul Web, e poi decidere sulla base dell’insieme delle distanze tra i due insiemi di parole di contesto.

Rientra nei metodi che estendono le distanze lessicografiche con regole semantiche anche il seguente esempio di deduplicazione. Stiamo considerando una tabella, vedi Figura 17, che riporta informazioni sui pensionati americani, sul loro reddito e sugli Stati in cui risiedono per una parte dell’anno. Negli USA è frequente che pensionati con alto reddito risiedano per una parte dell’anno, ad esempio d’estate, in stati del Nord, e l’altra parte dell’anno in stati del Sud, vedi Figura 18.

Numero	Nome	Cognome	Stato	Età	Reddito in $
1	Ann	Albright	Arizona	65	70.000
2	An	Olbrgth	Washington	65	70.000
3	Ann	Allbrit	Florida	85	15.000
4	Ann	Alson	Louisiana	72	70.000
5	Annie	Albight	Vermont	85	15.000
6	Annie	Allson	Florida	72	70.000
7	Georg	Allison	Vermont	71	66.000
8	George	Alsn	Florida	71	66.000

Figura 18 – Esempio dei pensionati americani

Figura 19 – Conoscenza aggiuntiva sui pensionati USA che può essere sfruttata nel record linkage (da www.commons.wikimedia.com)

Ci troviamo quindi di fronte a conoscenza aggiuntiva sui record della tabella, che possiamo sfruttare nella formula che esprime la distanza. Se non consideriamo tale conoscenza aggiuntiva, prendiamo in considerazione Nome e Cognome, e consideriamo la distanza = 3 come soglia della edit distance per decidere se accoppiare i record, allora risultano accoppiati (vedi in Figura 20 i diversi livelli di grigio) i record <1,2,5> , <4,6> e <7,8>.

Record #	First Name	Last Name	Country	Age	Income
1	Ann	Albright	Arizona	65	70.000
2	An	Olbrgth	Washington	65	70.000
3	Ann	Allbrit	Florida	85	15.000
4	Ann	Alson	Louisiana	72	70.000
5	Annie	Albight	Vermont	85	15.000
6	Annie	Allson	Florida	72	70.000
7	Georg	Allison	Vermont	71	66.000
8	George	Alsn	Florida	71	66.000

Figura 20 – Corrispondenze nel caso di applicazione della edit distance

Record #	First Name	Last Name	Country	Age	Reddito
1	Ann	Albright	Arizona	65	70.000
2	An	Olbrgth	Washington	65	70.000
3	Ann	Allbrit	Florida	25	15.000
4	Ann	Alson	Louisiana	72	70.000
5	Annie	Albight	Vermont	25	15.000
6	Annie	Allson	Florida	72	70.000
7	Georg	Allison	Vermont	71	66.000
8	George	Alsn	Florida	71	66.000

Figura 21 – Corrispondenze tenendo conto nella distanza della conoscenza aggiuntiva

Se invece prendiamo in considerazione gli Stati citati nei record, secondo la formula Nord/Sud, e accettiamo una maggiore distanza di edit pari a 5, allora gli accoppiamenti diventano <1,2>, <3,5> e <7,8>, vedi Figura 21.

4. L'integrazione di dati territoriali

La letteratura sulla integrazione di dati territoriali è molto ricca, e considera diversi casi; la difficoltà maggiore nella integrazione di dati territoriali si ha quando una o entrambe le fonti da integrare non sono georeferenziate, cioè non sono associate ad un sistema metrico di coordinate per il territorio. In questa sezione, consideriamo il caso di fonti da integrare costituite da:

- mappe vettoriali (cioè costituite da segmenti che uniti tra loro rappresentano le diverse caratteristiche del territorio) e
- immagini non georeferenziate.

Un approccio all'allineamento di mappe vettoriali e immagini non georeferenziate prevede le seguenti fasi (vedi anche Figura 21):

1. Rilevamento dei punti di controllo: i punti di controllo rappresentano la conoscenza estratta dalle mappe/immagini che si può assumere come punto di partenza per le successive attività di matching; gli incroci stradali sono buoni candidati per essere punti di controllo, perchè sono punti salienti per catturare la struttura della rete delle strade.
2. Filtraggio dei punti di controllo - A causa della complessità della scena naturale nelle immagini, la tecnica adottata in precedenza potrebbe portare ad errori; in questo passaggio viene utilizzato un filtro, per eliminare, nell'esempio degli incroci tra strade, incroci erroneamente identificati e mantenere solo quelli identificati in maniera inequivocabile.
3. Creazione della corrispondenza tra immagine e dati vettoriali (passo di conflation in Figura 22, sono anche riportati i riferimenti di alcuni algoritmi che possono essere utilizzati)): il sistema identifica un insieme di coppie di punti di controllo per cui si può presumere che la coppia di punti di controllo indichino la stessa posizione nel territorio (coppia matching). Per allineare tutti gli altri punti, le corrispndenze vengono calcolate a partire dalle coppie di punti di controllo precedenti.

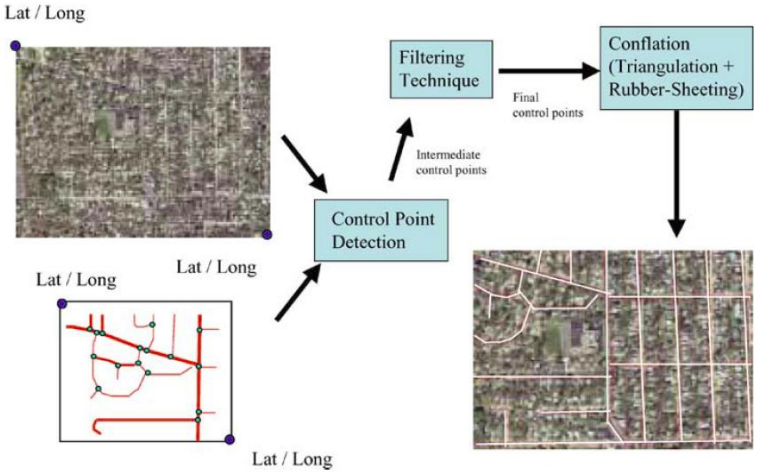


Figura 22 – L'integrazione dei dati spaziali (tratta da Chen C.C. et al. – Automatically and efficiently matching road networks with spatial attributes in unknown geometry systems, 2006)

5. La fusione dei dati

Ora che abbiamo accoppiato i record per effettuare il record linkage o la deduplicazione, ci troviamo di fronte a un dubbio amletico: a fronte di due record matching, quale record, o quali valori nei due record scegliamo per stabilire quella che viene chiamata “l’unica fonte della verità”?

Strategy	Classification	Short Description
PASS IT ON	ignoring	escalates conflicts to user or application
CONSIDER ALL POSSIBILITIES	ignoring	creates all possible value combinations
TAKE THE INFORMATION	avoiding, instance based	prefers values over null values
NO GOSSIPING	avoiding, instance based	returns only consistent tuples
TRUST YOUR FRIENDS	avoiding, metadata based	takes the value of a preferred source
CRY WITH THE WOLVES	resolution, instance based, deciding	takes the most often occurring value
ROLL THE DICE	resolution, instance based, deciding	takes a random value
MEET IN THE MIDDLE	resolution, instance based, mediating	takes an average value
KEEP UP TO DATE	resolution, metadata based, deciding	takes the most recent value

Figura 23 – Tecniche per la fusione

La letteratura recente si è occupata di questo problema, proponendo gli approcci che ora discutiamo.

Vi sono tre famiglie di strategie per operare la fusione (vedi Figura 23):

- ignorare le differenze (ignoring), in due modi possibili: a. lasciando all’utente o all’applicazione la decisione sui valori da scegliere, ovvero b. considerando tutti i possibili valori come accettabili. Insomma, riguardo alla strategia b, se un sito mi dice che domani la temperatura a Oslo sarà di 10 gradi sopra lo 0 e un secondo che sarà di 4 gradi sotto lo zero, io assumo per la temperatura di domani a Oslo contemporaneamente 10 gradi sopra e 4 gradi sotto lo zero.
- evitare le differenze (avoiding), con scelte che vanno da a. preferire i valori specificati a quelli che denotano assenza di valore, a b. considerare solo valori consistenti tra loro, ovvero c. considerare sempre i valori di una fonte considerata la più affidabile. Questo caso c. rientra nello studio di caso dei servizi alle imprese; poiché esiste una legge che affida la gestione del Registro delle imprese alle Camere di Commercio, quando si trattò di scegliere tra diversi valori, ad esempio per l’indirizzo toponomastico della impresa, la base dati delle Camere era considerata quella certificata di qualità.
- risolvere la differenza (resolution) con diverse metodologie: a. considerare il valore più frequente, b. scegliere un valore a caso, c. prendere la media (tornando al caso di Oslo, si assume una temperatura pari al varoe intermedio tra +10 e -4, cioè + 3 gradi sopra lo 0), d. considerare il valore aggiornato più recentemente (ciò si appura facilmente andando a guardare il log, il registro degli aggiornamenti alla base di dati, in cui sono riportate le operazioni effettuate, i valori e gli istanti temporali).

Strategy	Classification	Short Description
PASS IT ON	ignoring	escalates conflicts to user or application
CONSIDER ALL POSSIBILITIES	ignoring	creates all possible value combinations
TAKE THE INFORMATION	avoiding, instance based	prefers values over null values
NO GOSSIPING	avoiding, instance based	returns only consistent tuples
TRUST YOUR FRIENDS	avoiding, metadata based	takes the value of a preferred source
CRY WITH THE WOLVES	resolution, instance based, deciding	takes the most often occurring value
ROLL THE DICE	resolution, instance based, deciding	takes a random value
MEET IN THE MIDDLE	resolution, instance based, mediating	takes an average value
KEEP UP TO DATE	resolution, metadata based, deciding	takes the most recent value

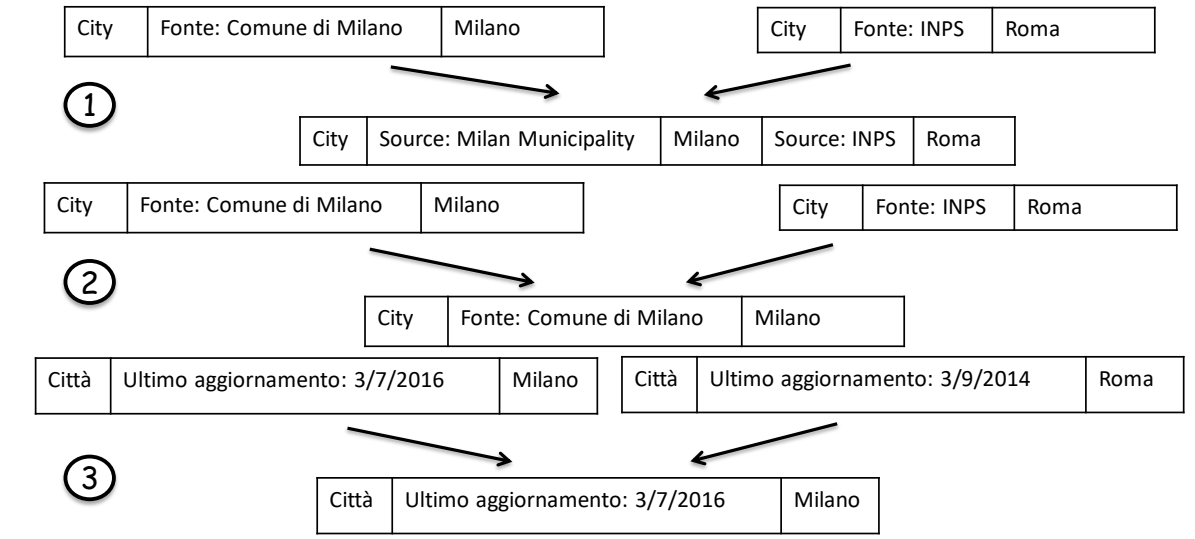


Figura 24 – Tre strategie per effettuare la fusione della stessa coppia di record

In Figura 24 vediamo l'applicazione di tre delle strategie nel caso di due record che rappresentano la città di residenza di una persona. Nel primo caso vengono considerate nella fusione entrambe le versioni della città di residenza, nel secondo caso viene presa in considerazione la fonte considerata più affidabile, nel terzo caso viene considerata la versione più recente.

6. Integrazione e fusione nello studio di caso delle imprese

Nello studio di caso delle imprese, l'attività di record linkage inizia scegliendo gli attributi da confrontare. Facendo riferimento all'esempio di Figura 24, possiamo confrontare gli attributi Nome della Impresa e Tipo di attività; non prendiamo in considerazione l'indirizzo perché dall'esempio di Figura 24 traspare chiaramente che si tratta di una informazione che può risultare non aggiornata, e di perciò nn c'è da fidarsi.

Focalizzandoci sulla terna di record di Figura 25, si tratta di capire se tutti e tre i record rappresentano la stessa impresa ovvero soltanto due tra essi, o nessuno. Decidiamo di applicare la distanza di Jaccard all'attributo Nome della Impresa. Applicando la formula di Jaccard otteniamo i valori di Figura 24 (dove 0,2 è calcolato dalla formula 1 – (# parole comuni = 5) / (# di tutte le parole = 6)).

Se decidiamo che tre record rappresentano la stessa impresa se la somma delle distanze di Jaccard dei campi “Nome di impresa” è minore di 0.5, allora la conclusione è positiva.

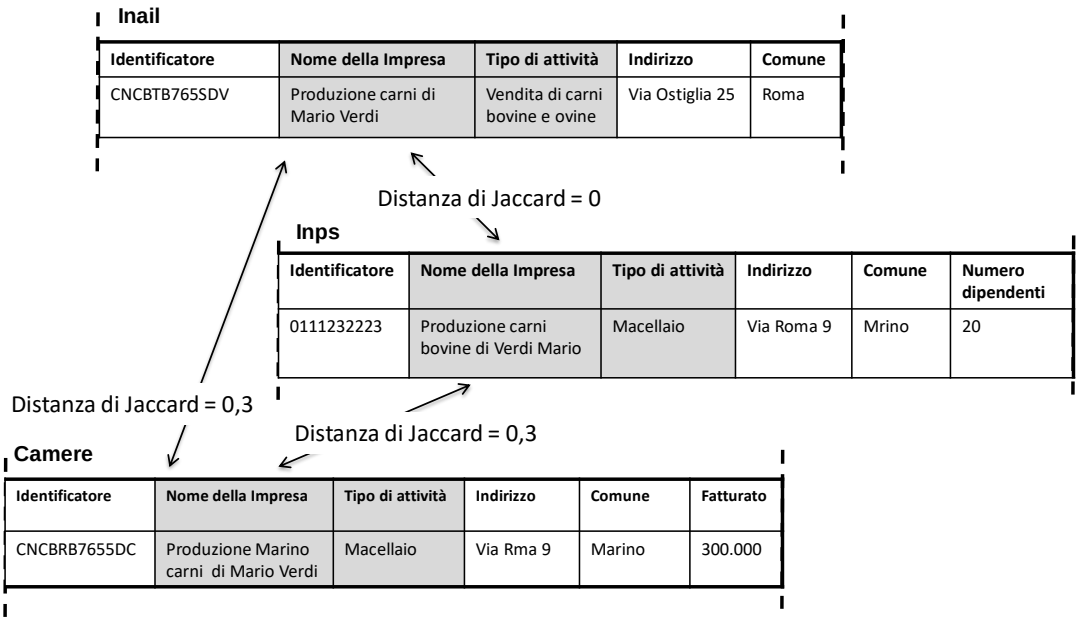


Figura 25 – Esempio di partenza nello studio di caso delle imprese

Riguardo al campo Tipo di attività, possiamo cercare qualche tassonomia delle attività delle imprese. Supponiamo di individuare una tassonomia a cui corrisponde il seguente frammento (semplifichiamo qui la tassonomia Istat).

……
Macellai
    Vendita di carni ovine
    Vendita di carni bovine
    Vendita di carni equine
    Vendita di carne di maiale
……

I tre valori sono dunque molto vicini nella tassonomia, e possiamo decidere che essi siano compatibili; il confronto rafforza la conclusione di identità dei tre record.

Dobbiamo a questo punto sottoporre a fusione i tre record, selezionando per ogni attributo il valore tra i tre che risulta dalla applicazione di una delle strategie di Figura 21. Si veda la figura 26; nel nostro caso, ad esempio, possiamo scegliere il Nome della Impresa dal record delle Camere di Commercio, che corrisponde alla base di dati certificata; analogamente per l’identificatore e il tipo di attività. Per l’indirizzo, il valore via Rma è certamente sbagliato; scegliamo il valore dell’Inps. Per il campo Comune scegliamo il valore associato a Via Roma, e infine aggiungiamo i campi Numero Dipendenti e Fatturato che vengono acquisiti rispettivamente da Inps e Camere.

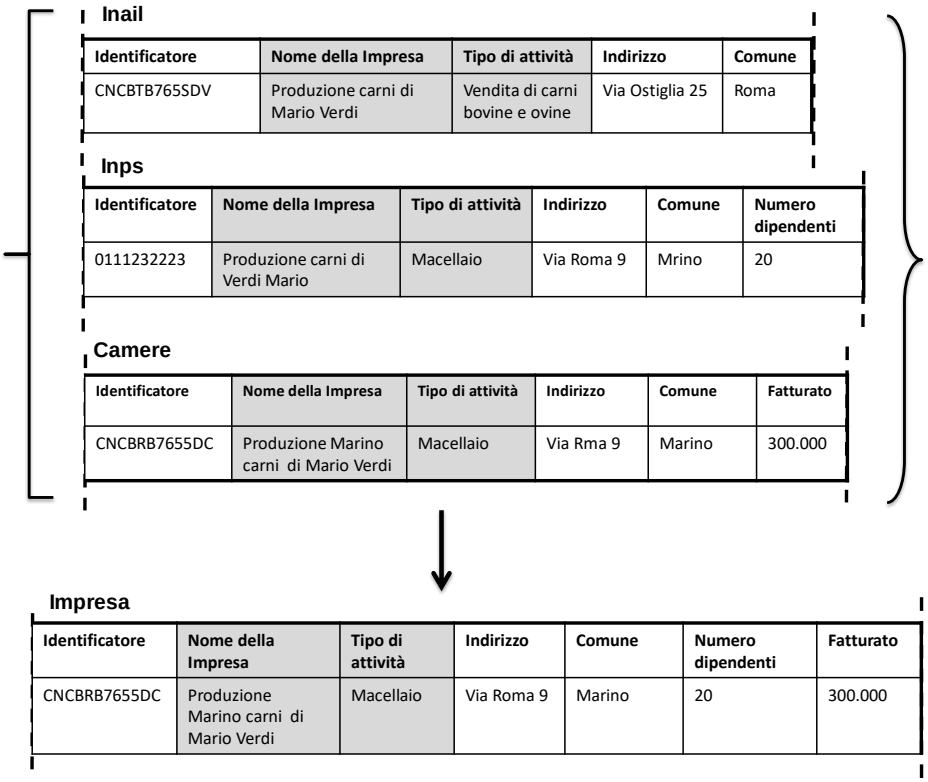


Figura 26 – Fusione dei tre record

Notiamo a conclusione della sezione che la integrazione, nel mettere insieme informazioni che in precedenza erano disperse in diverse basi di dati, porta valore; nel nuovo record risultato della fusione possiamo estrarre maggiori informazioni rispetto a prima, ed in particolare la correlazione tra numero dipendenti e fatturato. In una ipotetica analisi successiva con lo scopo di individuare, ad esempio, aree di evasione contributiva, questa correlazione ci può essere molto utile.

7. Integrazione e fusione nel contratto di governo tra Lega e Movimento 5 Stelle e nella successiva attuazione nella azione di Governo

Nel maggio del 2018, il Movimento 5 Stelle (M5S nel seguito) e la Lega, usciti come primo e terzo partito più votati dalle elezioni politiche del 4 marzo 2018, produssero e firmarono un contratto di Governo, sulla base del quale ebbero poi la fiducia nei due rami del Parlamento, varando così il Governo Conte.

Il Contratto di Governo fu prodotto come esito di intensi incontri tra i rappresentanti dei due partiti. In questa sezione intendo provare ad analizzare soprattutto la concreta attuazione del contratto nella azione di Governo come esito della integrazione e fusione delle visioni e delle strategie politiche dei soggetti coinvolti, quali i capi dei M5s e Lega e il Presidente del Consiglio. Nel condurre questa indagine, proverò ad applicare le strategie presentate per la fase di fusione della attività di integrazione dati nelle basi di dati. L'obiettivo è ambizioso e rischioso allo stesso tempo, soprattutto per la pretesa di adottare concetti, in fin dei conti, tecnologici, a prassi e comportamenti che certamente sono anni luce distanti

da tali concetti, essendo la politica un'arte non facilmente formalizzabile e modellabile. Il lettore non si aspetti una analisi esaustiva del problema, quanto piuttosto un esercizio volto ad analizzare il processo di creazione e le azioni di Governo del primo governo Conte alla luce delle metodologie di integrazione dati. Ciò che cercheremo di analizzare è il processo mostrato visivamente in Figura 27.

Figura 27 – Il processo di a. integrazione dei programmi elettorali di Movimento 5 Stelle e Lega nel Contratto di Governo e b. attuazione della azione di Governo attraverso provvedimenti ufficiali (da www.movimento5stelle.it e www.leganord.org, circa 2018)

Il processo è composto di due fasi:

- la predisposizione del Contratto di Governo, successivamente firmato dai due responsabili M5S e Lega, Luigi Di Maio e Matteo Salvini.
- La concreta attuazione del Contratto e le successive decisioni su materie anche non trattate nel contratto, che sono osservate nel seguito tramite una analisi delle scelte effettuate nella composizione del Governo e nei provvedimenti emanati.

Possiamo includere tra i provvedimenti di Governo i provvedimenti presi dal Consiglio dei Ministri, i decreti legge, le leggi proposte dal Governo, approvate dal Parlamento e firmate dal Presidente della Repubblica, i decreti ministeriali, le comunicazioni alla Unione Europea in risposta a richieste, rilievi o per dare attuazione a decisioni collegiali (ad es. il nome del rappresentante italiano come Commissario nella nuova Commissione Europea), e altri ancora. Un lavoro molto vasto, che ho cercato di semplificare focalizzandomi su un insieme di provvedimenti che sono rimasti nella memoria di tutti coloro che seguono con attenzione le vicende della politica italiana.

E' certamente rimasto nella memoria di tutti il fatto che nei primi mesi del governo "giallo-verde" ogniqualvolta uno dei due raggruppamenti politici spingeva per accelerare un determinato provvedimento affermava che "è nel contratto, quindi va attuato, anzi, abbiamo firmato il contratto proprio perché così non avremmo litigato successivamente"; nella analisi che sto effettuando, il

contratto è una sorta di base di dati integrata, in cui è stato trovato posto per ogni tema che i due raggruppamenti ritenevano importante. Il problema è che nel contratto c'era scritto il "cosa", non il "come", e sul "come" naturalmente i due raggruppamenti potevano avere idee differenti. Più avanti nella vita del Governo Lega M5S questo riferimento al contratto si è perso, probabilmente per due ragioni, la stanca ripetitività con cui si continuava a fare riferimento al contratto, che aveva perso il suo carattere rassicurante di mantra, e, forse soprattutto, il fatto che i nuovi temi dell'agenda politica non erano stati inclusi nel contratto, che evidentemente non poteva prevedere tutto quanto sarebbe potuto accadere.

In tutto questo vastissimo insieme di azioni, connesse alla scelta del struttura del Governo e ai provvedimenti emanati, ne ho scelte alcune. Esse sono:

1. Scelta del Presidente del Consiglio nella persona di Giuseppe Conte.
2. Scelta di nominare due Vicepresidenti del Consiglio nelle persone di Luigi Di Maio e Matteo Salvini.
3. Emanazione e approvazione dei due provvedimenti più rilevanti legati alla azione di governo nel primo anno: il reddito di cittadinanza e quota 100.
4. Decisione in merito alla prosecuzione della TAV Torino Lione.
5. Lettera alla Unione Europea in merito alla decisione sulla TAV.
6. I due decreti sicurezza.
7. Il meccanismo scelto nel decreto sicurezza bis per decidere in merito alle autorizzazioni di entrare nelle acque territoriali italiane delle navi Omg.
8. La non emissione del provvedimento che vieta l'ingresso nelle acque territoriali italiane della nave Open Arms, collegato al decreto di cui al punto precedente, perché, ricordo, i due ministri 5Stelle non lo firmarono.
9. La procedura di scelta del candidato Commissario europeo nella prima fase del processo decisionale (successivamente interrotta nel momento in cui la Lega ha presentato la mozione di sfiducia a Giuseppe Conte).

Per ciascuno di questi provvedimenti (nel caso 8 si tratta in effetti di un non-provvedimento) i due raggruppamenti presenti nella coalizione hanno dovuto compiere "azioni di integrazione"; la stesura in linguaggio naturale del provvedimento è stata fatta da esponenti dei due raggruppamenti, assistiti da consiglieri giuridici ed economici. L'esito della stesura è un testo che nasce dalla integrazione di due punti di vista, quello del M5S e quello della Lega. Nella integrazione dei due punti di vista, essi hanno dovuto fare operazioni di integrazione e di fusione, così da arrivare ad un testo condiviso; il mio intento è vedere se esiste qualche relazione logica tra le strategie di fusione descritte nella sezione precedente, e le strategie di fusione dei punti di vista adottate nella emanazione dei provvedimenti di cui alla lista precedente. Vediamo i vari punti.

1. <u>Scelta del Presidente del Consiglio</u>

Tutti ricordiamo che la scelta del nome del Presidente del Consiglio richiese alcuni giorni, e alla fine cadde sul Giuseppe Conte, persona poco nota al grande pubblico ma che aveva già ricoperto incarichi istituzionali. Si disse allora che Conte era più vicino al Movimento 5Stelle che alla Lega, ma considerato

sufficientemente al di sopra delle parti dalla Lega per poter svolgere un ruolo di mediazione tra Lega e M5S.

Strategia adottata (qui e nel seguito vedi le strategie in Figura 22): "meet in the middle", modificata pesando le due posizioni con la consistenza elettorale. I pesi sono valori la cui somma è uno, che si aggiungono alla formula che calcola la media delle posizioni così che essa diventi

Scelta = ((peso-5Stelle x scelta 5Stelle) + (peso-Lega x scelta Lega))/2

Naturalmente, siccome le persone non sono numeri, e tanto meno numeri con virgola, Conte corrisponde al "valore" che meglio approssima la formula. La strategia modificata è simile a quella proposta per la fusione in [Motro 2006].

2. Scelta di nominare due Vicepresidenti del Consiglio nelle persone di Luigi Di Maio e Matteo Salvini.

Qui secondo me la strategia adottata corrisponde alla **"**consider all possibilities", perché sono stati considerati tutti i capi/segretari politici dei raggruppamenti coinvolti nella coalizione.

3. I due provvedimenti più rilevanti delle decisioni di governo nel primo anno: il reddito di cittadinanza e quota 100

I due provvedimenti hanno occupato la agenda politica a lungo. Se li analizziamo, sappiamo che a entrambi i provvedimenti erano associate ingenti risorse economiche; il reddito di cittadinanza presenta *maggiori costi* per la macchina dello Stato, in quanto configura l'attribuzione di un reddito a una platea di soggetti che per definizione non lo percepiscono, e, inoltre, configura un investimento nella figura del navigator. Quota 100, poiché anticipa la andata in pensione di soggetti senza che automaticamente essi vengano sostituiti da altri lavoratori, configura da una parte *un maggior costo* nella erogazione di pensioni, e insieme una riduzione nella produzione di ricchezza che porta come effetto indotto una *riduzione delle entrate* da reddito di lavoro e dalla tassazione conseguente.

Insomma i due provvedimenti portano nel breve termine a un aumento della spesa oltre che a una riduzione delle entrate. Sono quindi in competizione, per cui la strategia di fusione corrisponde anche in questo caso a mio parere nella "consider all possibilities" cui possiamo associare anche, sia pure in forma più vaga, una influenza della strategia "keep up to date", nel senso che un leggero vantaggio di posizione derivava da chi faceva per primo la proposta di provvedimento, che poteva dire "i soldi ci sono", mentre il secondo poteva solo dire "se abbiamo trovato i soldi per il primo provvedimento, dobbiamo trovarli anche per il secondo".

4. Prosecuzione della TAV

Anche questo tema ha dato luogo a grandi discussioni e polemiche. Esso, dopo diversi rinvii, è sfociato nella votazione di una mozione del Partito Democratico a favore dell'opera, a cui il M5S ha dato voto negativo (e la Lega un voto positivo). Il Movimento ha deciso di non trarre le conclusioni da questo

esito; possiamo dunque arrivare alla conclusione che questa votazione, pur con il voto negativo del M5S, è da classificare tra i provvedimenti di integrazione, su cui cioè c'è stata alla fine una convergenza che non ha interrotto la esperienza di Governo, e quindi la strategia di fusione più vicina è la "trust your friends".

5. Lettera alla UE

Come noto la lettera alla Unione Europea in cui si confermava che l'opera sarebbe stata completata non è stata firmata dal Ministro competente ma da un Dirigente del Ministero delle Infrastrutture e dei Trasporti. In questo caso non riesco a trovare una strategia sensata tra quelle disponibili, mi sembra piuttosto una strategia che possiamo chiamare "fai come lo struzzo". Mi ricorda tanto un mio professore di italiano, il professor A., che una volta ci lesse un passo del Decamerone su Calandrino incinto, e quando arrivò a dover pronunciare questa parola, sospese la lettura, chiamò uno di noi, fece leggere quella parola a un mio compagno di classe, e poi proseguì tranquillamente la lettura.

6. I due decreti sicurezza

Qui propongo la strategia "cry with the wolves", non tanto perché sia stato preso "il valore più frequente" quanto perché ha vinto secondo me chi è stato più insistente, il Ministro Salvini, che ha fortemente voluto e accompagnato la preparazione e approvazione dei due decreti (lettralmente "cry with the wolves" significa urla con i lupi….).

7. Meccanismo scelto nel decreto sicurezza bis per decidere in merito alle autorizzazioni di entrare nelle acque territoriali italiane

Dopo che il Tar Lazio annullò un primo provvedimento volto a proibire la entrata nelle acque territoriali italiane della ONG Open Arms per la "situazione di eccezionale gravità ed urgenza" che si era creata a bordo della nave, il Ministro dell'Interno ha firmato un nuovo divieto di ingresso. Giustificato, a suo dire, dal fatto che alle ragioni citate nel provvedimento *sub judice* – quello annullato dal Tar – "se ne sono aggiunte altre. Per giorni, Open Arms si è infatti trattenuta in acque libiche e maltesi, ha anticipato altre operazioni di soccorso e ha fatto sistematica raccolta di persone con l'obiettivo politico di portarle in Italia», disse il Ministro.

Ma, al di là della veridicità dell'accusa, il nuovo divieto, come il vecchio, aveva bisogno della controfirma dei ministri della Difesa e delle Infrastrutture, i grillini Trenta e Toninelli. All'articolo 1, il decreto sicurezza bis stabilisce infatti che il Ministro dell'Interno "può limitare o vietare l'ingresso il transito o la sosta di navi nel mare territoriale" per ragioni di ordine e sicurezza, cioè quando si presuppone che sia stato violato il testo unico sull'immigrazione e in particolare si sia compiuto il reato di "favoreggiamento dell'immigrazione clandestina", quindi è necessario i concerto dei due ministeri citati.

In una prima versione del decreto, i Ministri delle Infrastrutture e dei Trasporti e della Difesa dovevano semplicemente essere informati dal Viminale dell'attuazione della interdizione all'ingresso nei porti. Nel testo definitivo, invece, il provvedimento doveva essere controfirmato dai titolari dei due Dicasteri

che erano alla data esponenti del Movimento 5 Stelle, rispettivamente Danilo Toninelli e Elisabetta Trenta.

In questo caso la presenza obbligatoria delle tre firme corrisponde alla strategia "no gossiping", in quanto, affinchè il provvedimento abbia validità, è *consistentemente* necessario il consenso dei tre ministri.

8. Non emissione del provvedimento di cui al punto precedente, perché i due ministri 5Stelle non lo firmarono.

La non emissione del provvedimento per assenza di due delle tre firme di ministri corrisponde a un esito diverso da quelli trattati fino a questo momento e cioè *nessuna integrazione*, esito che fa parte delle potenziali decisioni di governo, e che portò a una non-decisione e a nessun esito dal punto di vista di emanazione di provvedimenti ufficiali. E' anche interessante osservare che la ipotetica base dati dei provvedimenti adottati dal Governo Conte e dai Ministri rispetta quella che nelle basi di dati viene chiamata ipotesi di mondo chiuso (introdotta nel Capitolo 4): tutto ciò che non è registrato nella base dati, cioè che non ha dato luogo ad un provvedimento, non esiste, non è una informazione valida, come, appunto, il caso del provvedimento con una sola firma su tre.

9. Scelta del commissario europeo

Nella prima fase della istruttoria per individuare un nome da proporre a Bruxelles, si è imposto come argomento il fatto che avendo la Lega vinto le elezioni europee tra i partiti italiani, ad essa spettava la individuazione del nome. La strategia adottata appare dunque essere la "trust your friends".

Come sappiamo, l'esperienza di Governo della alleanza Lega – Movimento 5 Stelle si è interrotta bruscamente nell'agosto 2019. Ricordo al lettore che nella fase della discussione alle Camere suscitata dalla crisi di governo furono adottati due provvedimenti riferiti al precedente punto 8, in cui al contrario del primo caso, i due ministri 5 Stelle firmarono il decreto di interdizione. Tale decisione è difficilmente inquadrabile nelle problematiche affrontate in questa sezione.

8. Integrazione, fusione (e astrazione) nel secondo Governo Conte

I recenti avvenimenti dell' agosto 2019 che hanno portato Salvini a presentare una mozione di sfiducia al Governo Conte, e successivamente hanno portato il Primo Ministro a presentare le dimissioni, hanno fatto prendere agli avvenimenti successivi una piega imprevista, con le trattative tra Movimento 5 Stelle (M5S) e Partito Democratico (PD) come soggetti principali, che hanno portato ad un accordo di Governo e alla fiducia delle Camere, dando luogo al secondo Governo Conte.

Le trattative che hanno portato all'accordo, viste dal punto di vista delle metodologie di integrazione, hanno seguito un percorso molto diverso rispetto a quello che ha caratterizzato il precedente accordo, in particolare:

1. La trattativa è avvenuta attraverso una sequenza di fasi profondamente diverse dalla trattativa Lega – M5S, nel senso che mentre allora il Presidente incaricato fu scelto all'ultimo, in questo caso il Presidente è stato scelto all'inizio, e le trattative sul programma sono avvenute solo successivamente. Possiamo dunque dire, ricordando le metodologie di progetto di basi di dati, che nel caso Lega M5S è stata applicata una metodologia bottom-up (vedi Capitolo 3 sui modelli), in cui si è partiti dal basso, si sono integrati i due programmi e solo alla fine si è arrivati al nome del Primo Ministro, mentre nel caso M5S PD si è proceduto top-down, dall'alto verso il basso.
2. La scelta della persona del Primo Ministro, peraltro, è scaturita da un accordo (fusione) in cui il nome è stato inizialmente pesato in modo diverso dai due partiti/movimenti, nel senso che i 5 Stelle lo vedevano come espressione super partes di M5S e PD, mentre il PD lo vedeva come espressione del Movimento 5 Stelle.
3. Un altro aspetto, legato al precedente, riguarda il tema dei Vice Presidenti, su cui M5S e PD hanno oscillato tra due e uno, aspetto questo che riguarda nelle basi di dati lo schema, che sappiamo fare riferimento alle classi di osservabili, e non le istanze, cioè i valori. Quando si è passati dalle classi astratte ai valori (cioè i nomi e cognomi), l'accordo ha rischiato di saltare perché i M5S volevano Di Maio come uno dei due vicepresidenti, mentre gli esponenti PD affermavano che il M5s era già rappresentato da Conte, e quindi non poteva avere il suo massimo dirigente come Vice, arrivando alla proposta di Vice unico; il tutto si è risolto quando il ruolo di Vice è stato eliminato.
4. Il programma di Governo è stato definito essere un *accordo politico* e non più un contratto; inoltre, il testo dell'accordo politico è molto più *generale e generico* del testo del contratto. Vi è insomma una rilevante differenza di livello di astrazione tra il contratto di governo tra Lega e M5S e l'accordo politico tra M5S e PD. Partendo da un puro elemento quantitativo, il contratto è un documento di 38 pagine mentre l'accordo è un documento di sette pagine. Se poi leggiamo i due documenti, è facile arrivare alla conclusione che nell'accordo politico i punti sono espressi a un livello di astrazione decisamente maggiore rispetto al contratto. Tornando alle metodologie di progettazione di schemi concettuali, l'accordo politico è stato certamente ispirato ad un metodo bottom-up in cui M5s e PD hanno "messo insieme" i loro programmi, tra l'altro in versioni successivamente arricchite (ricordiamo i 10 punti di Di Maio che poi sono diventati 20); ma, sia per il poco tempo disponibile (il Presidente Mattarella premeva), sia probabilmente per raggiungere un livello di astrazione compatibile con le diversità e anche, probabilmente, la scia gli scontri e accuse scambiate tra M5S e PD nella precedente fase della legislatura, si è preferito astrarre, fino ad arrivare ad una versione il cui livello di astrazione fosse adeguato all'accordo. Non abbiamo in questo Capitolo gli strumenti per trattare il tema delle astrazioni in politica, che verranno discusse nel Capitolo 9, Sezione 4.5, a cui si rimanda il lettore.

Il fatto che l'accordo politico tra M5S e PD sia ad un elevato livello di astrazione naturalmente presenta de rischi, un po' come accade quando dopo una lunga discussione le due parti dicono per esaurimento: allora siamo d'accordo, è tutto risolto, salvo poi doversi ricredere alla prima discussione successiva. Ma accanto ai rischi, ha permesso di decidere che un governo ci sarà; è solo il futuro in questi casi che ci dirà come andranno le cose, se scendendo di livello di dettaglio prevarranno nella operatività gli elementi di convergenza ovvero gli elementi di divaricazione.

Riferimenti

C. Batini, Stefano Ceri, e Shamkant B. Navathe - Conceptual database design: an Entity-relationship approach - Vol. 116. Redwood City, CA: Benjamin/Cummings, 1992.

C. Batini, M. Scannapieco – Data and Information Quality, Springer Verlag, 2016.

A.Motro, P. Anokhin e A. C. Acar - Utility-based resolution of data inconsistencies. - Proceedings of the 2004 international workshop on Information quality in information systems. ACM, 2004.

A.Motro e P. Anokhin - Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information source - Information fusion 7.2., 2006

Capitolo 7 – Dati e Semantica

M. Palmonari

1. Introduzione

Quando si parla di semantica dei dati, intuitivamente, ci si riferisce al tentativo di considerare il *significato* dei dati ai fini di supportare la loro elaborazione. Il termine *semantica* viene in realtà usato in diversi ambiti dell'informatica e del sapere, con una accezione tecnica; ad esempio, si dice che si definisce la *semantica* di un linguaggio di programmazione o la *semantica* di un linguaggio logico formale. A partire dalla seconda metà degli anni novanta, però, il consolidarsi del World Wide Web ha favorito i processi di produzione e consumo dei dati attraverso la rete, rendendo possibile generare dati in quantità sempre maggiore e in formato diverso e richiedendo strumenti sofisticati per la loro elaborazione.

Uno dei padri fondatori del World Wide Web, Sir Tim Berners Lee, che abbiamo citato nel Capitolo 3, ha immediatamente riconosciuto la necessità che la grande mole di dati disponibili fosse elaborabile attraverso l'utilizzo di *informazioni semantiche*, di informazioni rappresentate in maniera tale da favorire l'elaborazione da parte di applicazioni software [Gandon 2018]. Lo stesso Tim Berners Lee ha definito una vera e propria roadmap per costruire quello che ha definito *web semantico* [Berners-Lee 1998]. Vennero gettate quindi le basi oltre vent'anni di ricerche e innovazioni tecnologiche legate al Web semantico, il cui prodotto è un insieme di best practices (ad esempio, i principi per pubblicare Linked Open Data citati nel Capitolo 3), linguaggi (ad esempio RDF[4]), tecniche (ad esempio *ontology matching*), e tecnologie (ad esempio i *triple store* – database per rappresentare dati in RDF) finalizzate alla costruzione di un web che possa essere esplorato, consumato, ed elaborato in maniera automatica da parte di applicazioni. Quando si parla di tecnologie semantiche ci si riferisce spesso a questi prodotti.

Tuttavia, l'utilizzo del termine "semantica" nell'ambito della Scienza dei dati fa riferimento a un insieme di modelli e processi più generali di quelli prodotti nell'ambito del Web semantico, nonostante questo filone di ricerca e innovazione tecnologica sia un esempio paradigmatico del tentativo di occuparsi in maniera nativa della semantica dei dati.

In questo capitolo cercheremo di introdurre, con un linguaggio il più semplice possibile rispetto a quello usato nei libri di logica, e corroborato da tanti esempi, alcuni dei principali obiettivi dell'applicazione della semantica alla Scienza dei dati. Utilizzeremo alcuni concetti base del Web semantico, ma cercheremo di inquadrarli in un ambito più ampio. Riteniamo che la semantica sia la disciplina che si occupa dell'*interpretazione dei dati* e che lo faccia da almeno due punti di vista: proponendo linguaggi e modelli che, qualora esplicitamente utilizzati, facilitino l'interpretazione dei dati; mettendo a punto tecniche per migliorare l'interpretazione dei dati rispetto a una data interpretazione di partenza (ad esempio, estraendo informazioni strutturate a partire da testi considerati come pure sequenze di parole).

[4] https://www.w3.org/RDF/

Tratteremo tre temi che riteniamo di particolare interesse nel momento in cui scriviamo: il rapporto tra interpretazione e inferenza; il rapporto tra interpretazione e similarità; il problema di interpretare dei testi in quanto sorgenti di informazioni fattuali, piuttosto che in quanto mere sequenze di parole.

Affrontando questi problemi di natura più teorica, introdurremo esempi concreti di strumenti introdotti per trattare la semantica dei dati quali: i grafi di conoscenza, le ontologie e i linguaggi proposti nel web semantico per rappresentarle, e alcune tecniche di base per l'estrazione di informazioni (come Named Entity Recognition). Dato lo spazio limitato, in questo capitolo ci poniamo soprattutto l'obiettivo di spiegare le relazioni che sussistono tra alcuni strumenti semantici particolarmente rilevanti oggi, evitando una trattazione esaustiva e rimandando, per questa, alla letteratura specializzata di riferimento.

Il capitolo è organizzato come segue. Prima di iniziare il nostro percorso, nella Sezione 2 introduciamo un piccolo esempio per discutere la relazione tra dati, significato e interpretazione, e per mostrare come il concetto apparentemente teorico di *interpretazione* abbia in realtà radici e implicazioni estremamente pratiche, legate, cioè, all'uso che vogliamo fare dei dati. Nella Sezione 3 traiamo alcune conclusioni che riguardano il rapporto tra semantica e interpretazione. Nella Sezione 4 discutiamo gli obiettivi della *data semantics* come disciplina. Nella Sezione 5 trattiamo la relazione tra semantica, rappresentazione della conoscenza e inferenza, introducendo concetti quali *grafo di conoscenza* e *ontologia* e linguaggi come RDF, RDFS e OWL per rappresentare e condividere grafi di conoscenza e ontologie sul web. Nella Sezione 6 discutiamo la relazione tra semantica e similarità, toccando alcuni temi legati all'integrazione dei dati e altre applicazioni in cui la similarità gioca un ruolo privilegiato; estendiamo in questo modo la trattazione effettuata nel Capitolo 6. Nella Sezione 7 discutiamo il rapporto tra semantica ed estrazione di informazioni da documenti non strutturati, introducendo brevemente tecniche ormai di uso comune come *Named Entity Recognition* e *Named Entity Linking*. Nella Sezione 8 traiamo alcune conclusioni e facciamo alcune note su alcune direzioni di ricerca recenti favorite dal successo delle tecniche di *deep learning*.

2. Dati, Significato e Interpretazione

Partiamo da un semplice esempio per prendere confidenza con il concetto di semantica. Ipotizziamo di avere dei dati costituiti dalle seguenti sequenze di caratteri: "753 a.C.", "44 a.C" "27 a.C" "14 d.C", "44 BC" "A.D 476". Dal punto di vista della sintassi, "753 a.C." è costituito dalla sequenza di caratteri <"7","5","3",spazio,"a",".","C","."> . Ora possiamo chiederci: cosa possiamo fare con queste sequenze di caratteri, ovvero, che tipo di elaborazioni possiamo fare?

Un informatico riconoscerebbe immediatamente in queste sequenze di caratteri delle *stringhe,* sequenze arbitrarie di caratteri definite a partire da un insieme di caratteri noti chiamato *alfabeto*. Interpretare queste sequenze di caratteri come stringhe ci permette, ad esempio, di ordinarle in base all'ordine lessicografico, un modello che estende a stringhe arbitrarie un ordinamento predefinito sull'alfabeto. Dato l'ordinamento sull'alfabeto (intuitivamente i numeri vengono prima delle lettere, e queste sono nell'ordine condiviso ad esempio della lingua inglese), possiamo ordinare queste stringhe

secondo l'ordine seguente: <"14 d.C", "27 a.C", "44 a.C", "44 BC", "753 a.C.", "A.D 476">. Sapendo che stiamo interpretando questi dati come stringhe, possiamo anche applicare altre operazioni definite su insiemi di stringhe arbitrarie. Ad esempio, possiamo individuare tutte quelle stringhe che contengono la sequenza di caratteri "44", ottenendo come risultato l'insieme costituito dalle stringhe distinte "44 a.C" e "44 BC".

Nell'informatica sono state sviluppate molte altre strutture dati di base, ovvero modelli interpretativi che permettono di elaborare i dati sfruttando la struttura del dominio su cui vengono interpretati i dati. Oltre alle stringhe, ci sono ad esempio i numeri interi (ad esempio, -753 e +14), i numeri binari, etc.. Ciascuna di queste strutture dati sfrutta il modello interpretativo per supportare alcune operazioni (ad esempio, dati due numeri interi possiamo sommarli). La semantica caratterizza innanzitutto le strutture dati e le operazioni che eseguiamo su di esse, e, da questo punto di vista, caratterizza ogni tipo di elaborazione. Tuttavia, quando oggi parliamo di semantica dei dati, ci riferiamo a qualcosa di leggermente diverso rispetto alla gestione di strutture dati come stringhe, interi, etc.. In realtà, vedremo che si tratta solo di generalizzare quanto discusso qui sopra, richiamando però modelli interpretativi più sofisticati.

Torniamo al nostro insieme di sequenze di caratteri. Una persona che abbia studiato su libri di storia in Italia potrebbe riscontrare alcune peculiarità nell'insieme di sequenze di caratteri individuati sopra. Ad esempio, sapendo che il sistema di datazione nella nostra cultura si basa su un anno zero e che per riferirsi agli anni precedenti e successivi si usano rispettivamente le espressioni "avanti Cristo" e "dopo Cristo", potrebbe interpretare alcune di queste sequenze come *anni*. Qualora la persona conosca anche la lingua inglese, potrebbe avere imparato che in Inglese si usa giustapporre la sigla "BC" per indicare un anno o una data *precedente* l'anno zero e anteporre "A.D." per indicare un anno o una data *successiva* all'anno zero. Sulla base di questa informazione, la persona potrebbe interpretare tutte le sequenze viste sopra come anni o come date.

Ma cosa significa interpretare tutte queste sequenze come anni o come date? Innanzitutto significa essere in grado di manipolare queste informazioni sfruttando il modello interpretativo di riferimento. Ad esempio, significa essere in grado di riconoscere quali sequenze si riferiscono al medesimo anno o alla medesima data ("44 a.C." = "44 BC") e ordinare tutti gli anni o le date in ordine cronologico (<"753 a.C.", "44 a.C", "27 a.C", "14 d.C", "A.D 476">). Si noti che non era possibile riprodurre l'ordine cronologico se avessimo interpretato questi dati come stringhe. Ma qual è il modello interpretativo che ci permette di raggiungere il nostro obiettivo? Ce ne sono molteplici. Il modello interpretativo più semplice è costituito dall'*insieme dei numeri interi*. L'anno zero corrisponderà al numero 0, l'anno "753 a.C" corrisponderà al numero intero -753, l'anno 14 d.C corrisponderà al numero intero +14, e così via.

Possiamo usare anche un altro modello interpretativo, quello cioè costituito dai *numeri reali*. Per gli anni elencati sopra, l'interpretazione apparentemente non cambia molto, anche se +753 e +14 saranno numeri reali. Entrambi i modelli raggiungono l'obiettivo pratico di supportare l'ordine cronologico dei dati, ma ci sono delle differenze nei due modelli interpretativi: la scelta di un modello interpretativo ha delle implicazioni. Ad esempio, se interpretiamo questi dati come numeri interi, stiamo rappresentando gli anni come punti su una linea temporale discreta. Non esiste un anno compreso tra +14 e +15. Se interpretiamo questo anni come numeri reali, stiamo rappresentando gli anni come punti su una linea

temporale continua. Sarà sempre possibile immaginare infiniti punti compresi tra +14 e +15, ad esempio per rappresentare date più specifiche o istanti temporali. Intuitivamente, il modello interpretativo dei numeri interi si applica meglio se implicitamente ci interessa rappresentare solo gli anni. Se invece vogliamo rappresentare date arbitrarie o istanti temporali a granularità estremamente fine (ore, minuti, secondi, millisecondi, etc.), ad esempio per gestire delle transazioni finanziarie, il modello interpretativo basato su numeri reali è più appropriato.

Tutti questi ragionamenti hanno a che fare con il dominio, ovvero l'insieme di oggetti, su cui interpretiamo i dati e sulla sua caratterizzazione. Il significato che attribuiamo a "753 a.C" è il risultato di un'operazione d'interpretazione, o - se vogliamo - una *funzione d'interpretazione*, che associa a ciascuna sequenza di caratteri, tra cui "753 a.C", un elemento di un insieme con proprietà note, in questo caso, ad esempio, un numero intero. Il significato è pertanto strettamente legato alla caratterizzazione dell'insieme di oggetti su cui interpretiamo la nostra sequenza di caratteri (ad esempio, i numeri interi) e ai rapporti interni tra di essi (ad esempio, -753 < +14 < +15).

Aggiungiamo ora un ultimo tassello prima di trarre alcune conclusioni. Uno studioso di storia romana antica potrebbe avere un bagaglio di conoscenze sufficiente per ricordare che tutti questi anni, o queste date, si riferiscono ad eventi che riguardano la storia di Roma Antica: la fondazione della città, la fondazione dell'impero, la sua caduta, etc.. In altri termini, possiamo interpretare questo insieme di anni in virtù del ruolo che giocano nella descrizione di Roma Antica, ad esempio come un sottoinsieme di quegli anni che caratterizzano la storia romana. L'aspetto relazionale è un altro aspetto caratterizzante del significato. Ovviamente tra queste interpretazioni esistono delle relazioni; l'insieme di anni elencati sopra è un sottoinsieme degli anni che caratterizzano la storia romana, che è a sua volta sottoinsieme dell'insieme di tutti gli anni, che può essere interpretato come un sottoinsieme dei numeri interi.

3. Interpretazione e Semantica

Cerchiamo di riassumere le principali conclusioni che possiamo trarre dal ragionamento fatto con il precedente semplice esempio e trarne le debite conclusioni:

- Il significato dipende dall'interpretazione dei dati all'interno di un modello, che possiamo chiamare *modello interpretativo*.
- Un dato può essere soggetto a diverse interpretazioni e quindi elaborato secondo modelli interpretativi diversi.
- Modelli interpretativi diversi hanno implicazioni diverse.
- Modelli interpretativi diversi possono essere maggiormente adeguati per scopi diversi (ad esempio, supportare l'ordine lessicografico rispetto all'ordine cronologico) ma anche egualmente adeguati per un medesimo scopo (ad esempio, numeri interi e numeri reali sono altrettanto adeguati per ordinare cronologicamente sequenze di caratteri che descrivono date indicate mediante gli anni di riferimento).
- La capacità di elaborare i dati all'interno di un modello interpretativo dipende dalla struttura associata al dominio di interpretazione, ovvero dall'insieme di vincoli che caratterizzano il modello (ad esempio, interpretazione del tempo come discreto verso continuo)

- Fissare un modello interpretativo può avere due principali effetti sulla nostra capacità di elaborare i dati:
    - *Interpretazione e validazione*: introduciamo dei vincoli che delimitano i dati effettivamente interpretabili, escludendo alcuni dati non compatibili con il modello e precludendo alcune elaborazioni (ad esempio, se interpretiamo le sequenze di caratteri con cui indichiamo gli anni come numeri interi, dovremo escludere in quanto non interpretabili tutte quelle sequenze di caratteri composte con numeri diversi dai numeri interi come "21:06:234" o "14,5 d.C"; se interpretiamo le sequenze di caratteri con cui indichiamo gli anni come stringhe, non possiamo sfruttare l'ordine cronologico). L'interpretazione e i vincoli associati a essa, in altri termini, possono permettere di individuare e riconoscere sequenze di caratteri non valide perché non interpretabili.
    - *Interpretazione e inferenza*. I vincoli introdotti ci permettono di elaborare i dati inferendo nuove informazioni, sfruttando la struttura del modello (ad esempio, se interpretiamo "44 a.C" e "44 BC" nel dominio dei numeri interi associando a entrambi il numero intero +44, e se fissiamo l'interpretazione del carattere "=" come l'uguaglianza tra numeri interi, possiamo derivare la sequenza di caratteri "44 a.C = 44 BC"). L'interpretazione e i vincoli associati a essa, in altri termini, ci consentono di derivare nuove sequenze di caratteri a partire da quelle note (ad esempio "44 a.C = 44 BC").
- L'assenza di un modello interpretativo fissato ci impedisce di elaborare i dati; per farlo dobbiamo effettuare un'operazione di interpretazione; tipicamente, diverse interpretazioni saranno possibili, ciascuna con le relative implicazioni in termini di validità e inferenza, e in generale, in termini di possibili elaborazioni che possiamo effettuare. Per questo motivo, l'interpretazione dipende, in ultima analisi, dalle esigenze di elaborazione dei dati. Ad esempio, data una sequenza arbitraria di caratteri, possiamo sempre interpretarla come una stringa, perché questa interpretazione è molto generale; però tale interpretazione ci preclude di sfruttare l'ordinamento cronologico; se il nostro scopo è ordinare queste sequenze in ordine cronologico, abbiamo bisogno di un'interpretazione più *specifica*, cioè di queste sequenze di caratteri come riferimenti temporali, ad esempio attraverso l'interpretazione nel dominio dei numeri interi.

4. Data Semantics: all'incrocio di diverse discipline

Le ultime due osservazioni della Sezione 3 ci permettono di definire gli obiettivi principali della disciplina che si occupa della semantica dei dati, che chiameremo *Data Semantics*:

- Fornire strumenti computazionali (modelli, tecniche, linguaggi, etc.) capaci di *esplicitare l'interpretazione dei dati*, con lo scopo di consentire elaborazioni basate sui modelli interpretativi scelti; esempi di questi strumenti che approfondiremo nel resto del capitolo sono i linguaggi e i modelli dei dati proposti nell'ambito del Web semantico per rendere esplicita l'interpretazione dei dati in processi di pubblicazione e condivisione di dati e conoscenze.
- Fornire strumenti computazionali (modelli, tecniche, algoritmi, etc.) capaci di *inquadrare in un modello interpretativo quei dati per cui tale modello non è fornito a priori*; esempi di questi strumenti sono gli algoritmi per l'elaborazione di testi volti a individuare nomi di entità descritte in una base di conoscenza di riferimento (ad esempio, riconoscendo che le sequenze di caratteri

"Barack Obama" "Italia" nella frase "Barack Obama è venuto in visita in Italia" sono menzioni di due entità del mondo reale descritte in una base di conoscenza come DBpedia[5].

Occorre però fare qui una precisazione. Poiché alcuni problemi relativi alla semantica dei dati sono stati studiati sin dalla nascita dei primi calcolatori, sistemi operativi e linguaggi di programmazione (ad esempio, modelli per la rappresentazione ed elaborazione di stringhe, numeri interi, date, etc.), quando parliamo di approcci semantici all'elaborazione di dati in genere ci riferiamo a quei casi in cui l'interpretazione desiderata non è riconducibile a domini - ormai noti - quali le stringhe, gli interi, le date, etc., ma richiama domini di complessità arbitraria; ad esempio, quando vogliamo interpretare sequenze di caratteri come nomi di persone, organizzazioni, luoghi, etc., considerando possibilmente le relazioni che intercorrono tra questi tipi di entità. In altri termini, quando usiamo espressioni come "trattare la semantica dei dati" ci riferiamo spesso a modelli interpretativi più vicini all'universo cognitivo delle persone, e che siano in grado di supportare elaborazioni simili a quelle che ci aspetteremmo da esseri umani.

Questa necessità emerge perché le elaborazioni supportate dall'interpretazione di un testo in linguaggio naturale come pura sequenza di stringhe, o di una tabella come una matrice composta di stringhe, numeri interi, numeri reali etc., sono limitate per un'ampia gamma di applicazioni. In applicazioni di nuova generazione abbiamo bisogno di modelli interpretativi che consentano elaborazioni sofisticate, quali: rispondere a domande formulate in linguaggio naturale, determinare quanto due parole usate in un testo sono simili, riconoscere quali parole in un testo si riferiscono a entità esistenti nel mondo reale, riconoscere quali stringhe in una tabella vanno interpretate come nomi di politici, dedurre quali politici menzionati in una tabella che riporti anche la loro città di nascita sono nati in Italia. Tutte queste elaborazioni richiedono modelli interpretativi che esulano dalla definizione di strutture dati per gestire stringhe, numeri, etc.. Per questo motivo. con il termine Data Semantics ci riferiamo spesso all'applicazione di modelli e tecniche elaborati in discipline legate all'Intelligenza artificiale (come la Rappresentazione della conoscenza, il Machine learning, e l'Information retrieval) in contesti riguardanti la gestione dei dati. La necessità, da un lato, di trattare dati di grandi dimensioni di natura fortemente eterogenea (che includono testi, immagini, video, etc.), e dall'altro, di supportare elaborazioni sofisticate di questi dati, rendono la Data semantics un crocevia tra Intelligenza artificiale e Gestione dei dati e uno strumento essenziale in una molteplicità di applicazioni della Scienza dei dati.

Come anticipato, non è ovviamente possibile approfondire l'ampio spettro di temi e risorse che caratterizzano la Data Semantics come disciplina. In questo capitolo ci limiteremo a discutere alcuni rapporti tra Semantica e altri concetti, che reputiamo centrali nel dibattito attuale relativo alla Data Semantics:

- Semantica, rappresentazione della conoscenza e inferenza;
- Semantica e similarità;
- Semantica ed estrazione di informazioni da documenti non strutturati.

[5] https://wiki.dbpedia.org/

5. Rappresentazione della conoscenza e inferenza

La Rappresentazione della conoscenza è una disciplina nata all'interno di quella branca dell'Intelligenza artificiale che viene spesso definita "simbolica" (in contrapposizione alla branca connessionista, o sub-simbolica, più legata a meccanismi di apprendimento e adattamento) [Russell&Norvig2016]. La disciplina si occupa di elaborare linguaggi e modelli computazionali che permettano di rappresentare conoscenze e supportare meccanismi di ragionamento automatico, ovvero di fare inferenze a partire da un insieme di conoscenze note. In questo capitolo sceglieremo una famiglia di linguaggi e modelli elaborati nell'ambito del Web semantico per esplicitare il significato dei dati condivisi via Web [Berners-Lee&al.2001, Hendler2001,Shadbolt&al.2006], che si sono affermati proprio in quell'intersezione tra Intelligenza artificiale e la Gestione dei dati che interessa la Data Semantics.

5.1 Grafi di conoscenza e RDF

Un *grafo di conoscenza* (in inglese *knowledge graph*), è un'astrazione che indica una base di conoscenza rappresentata mediante un grafo. Il termine è stato reso popolare da Google, che ha introdotto il proprio grafo di conoscenza[6] nel 2012. Intuitivamente, un grafo di conoscenza rappresenta un insieme di *entità* del mondo reale (film, cantanti, album, etc.), le reciproche relazioni (ad esempio, Elton John è l'autore del brano Sails), e altre proprietà (ad esempio, Elton John è un'entità *di tipo* musicista e Sails è un'entità di tipo canzone). Se nella comunità scientifica c'è un dibattito aperto su come definire in maniera più precisa un grafo di conoscenza [Ehrlinger&Wöß2016, Krötzsch2017], i grafi di conoscenza sono oggi molto usati nell'industria, anche in applicazioni in produzione, come strumento per migliorare l'integrazione e l'elaborazione dei dati. Alcuni grafi di conoscenza noti, in aggiunta a quello di Google, sono usati ad esempio da Facebook, Amazon, eBay, e molti altri [Noy&al2019].

Per sviluppare e gestire grafi di conoscenza si possono usare diversi modelli di rappresentazione dei dati e diverse tecnologie di gestione, soprattutto quando l'accesso al grafo è riservato alle organizzazioni che lo hanno creato o mediato da Application Program Interface (API, cioè librerie software di un linguaggio di programmazione). Per approfondire il concetto di grafo di conoscenza noi utilizzeremo il Resource Description Framework (RDF), un modello per la rappresentazione dei dati raccomandato dal W3C[7] e specificatamente pensato per supportare lo scambio di dati sul Web utilizzando un approccio semantico.

RDF è un modello a grafo per la rappresentazione dei dati che usa come primitive delle *triple*, ovvero, asserzioni dalla forma <soggetto, predicato, oggetto>[8]. Il soggetto può essere un *Unique Resource*

[6] https://www.google.com/intl/bn/search/about/

[7] https://www.w3.org/

[8] In realtà, RDF supporta anche l'uso di quadruple, con il quarto elemento che può identificare il grafo a cui appartengono gli altri tre elementi. Tuttavia, gli elementi rilevanti da un punto di vista logico sono quelli che formano la tripla, che può essere comunque considerata la primitiva essenziale del modello. In questo capitolo utilizziamo una notazione in cui le triple sono rappresentate tra parentesi angolari. RDF supporta diverse sintassi, tra cui la sintassi RDF/XML e quella basata su N-Triples; per dettagli sulla sintassi si rimanda ai diversi libri di testo oggi disponibili [Hitzler&al.2009, DiNoia&al.2013].

Identifier (URI) o un *blank node*, il predicato deve essere uno Uniform Resource Identifier (URI), mentre il soggetto può essere un URI o un *letterale*.

Gli URI sono identificativi, cioè nomi globali di oggetti arbitrari che chiameremo per comodità entità (persone, luoghi, documenti, tipi di cose, predicati, etc.); per coloro che hanno familiarità con la logica matematica, gli URI possono essere considerati costanti logiche e simboli di predicato. I letterali sono usati per rappresentare dati di tipi noti, ad esempio stringhe, numeri, date, etc. I blank node possono essere considerati dei nomi locali: mentre un URI utilizza uno standard riconosciuto per specificare un nome di una risorsa, un blank node denota un'entità solo all'interno di un grafo[9]. Utilizziamo come esempio l'insieme di triple in

Figura 1. Queste triple descrivono Elton John, specificando lo strumento musicale che suona, e asserendo che è autore di due prodotti musicali.

<dbr:Electric_Piano, rdfs:label, "Electric Piano"@en>
<dbr:Electric_Piano, rdfs:label, "Piano Elettrico"@it>
<dbr:Elton_John, dbo:instrument, dbr:Electric Piano>
<dbr:Empty_Sky, dbo:artist, dbr:Elton_John>
<dbr:Sails, dbo:musicalArtist, dbr:Elton_John>
<dbr:Sails, dbo:album, dbr:Empty_Sky>
<dbr:Elton_John, dbo:instrument, dbr:Electric_Piano>
<dbr:Elton_John, rdf:type, dbr:Musical_Artist>
<dbr:Sails, rdf:type, dbo:Single>
<dbr:Empty_Sky, rdf:type, dbo:Album>

dbr: for http://dbpedia.org/resource/
e.g.: http://dbpedia.org/resource/Elton_John
dbo: for http://dbpedia.org/ontology/
rdf for http://www.w3.org/1999/02/22-rdf-syntax-ns#

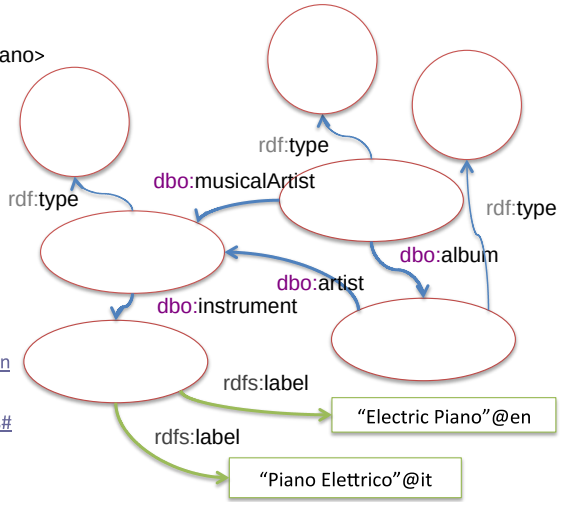


Figura 1 - Insieme di triple RDF che descrivono Elton John

Partendo dalla figura possiamo evidenziare alcune caratteristiche fondamentali di RDF:

- Una descrizione in RDF consiste di un insieme di *asserzioni atomiche* che possono essere combinate per costruire *asserzioni più complesse*; un'asserzione come <dbr:Elton_John, dbo:instrument, dbr:Electric Piano> asserisce, intuitivamente, che Elton John suona il piano elettrico come strumento; possiamo combinare questa asserzione con molteplici altre, ad esempio per specificare quali termini si usano in italiano e in inglese per riferirsi al Piano elettrico attraverso le asserzioni <dbr:Electric_Piano, rdfs:label, "Electric Piano"@en> e <dbr:Electric_Piano, rdfs:label, "Piano

[9] In questo capitolo non approfondiremo il ruolo dei blank node e il loro uso; dal punto di vista dell'interpretazione che discuteremo in questo capitolo si può pensare ai blank node come a delle costanti per riferirsi a entità di cui non sappiamo il nome.

Elettrico"@it>. In altri termini, un grafo RDF può essere visto come una *collezione di fatti* che descrivono ciò che è ritenuto vero in un certo dominio di interesse. Possiamo dunque *interrogare* un grafo RDF per recuperare informazioni sulla base delle asserzioni ritenute vere. SPARQL[10] è un linguaggio per l'interrogazione e la gestione di dati in RDF che può essere visto come l'equivalente per RDF di quello che SQL è per le basi di dati relazionali.

- Una descrizione RDF può essere vista come un multi-grafo etichettato, in cui i nodi sono URI, blank node o letterali e gli archi sono etichettati con URI che identificano i predicati; diciamo che è un multi-grafo, perché tra due nodi possiamo avere più archi, purché etichettati con predicati diversi; il punto di vista del *grafo* ci permette di ereditare tutto un insieme di operazioni che possono essere fatte sui grafi, quali l'attraversamento (ad esempio, "esiste un percorso che parte da Elton John e arriva a Michael Jackson?"), e l'analisi della connettività del grafo (ad esempio, individuazione di nodi principali).
- In entrambi questi due punti di vista, collezione di fatti e grafo, è importante rimarcare l'uso di URI come identificativi di entità e predicati (o connessioni, secondo il punto di vista a grafo).

Riassumendo, i dati rappresentati in RDF possono essere interpretati come un insieme di affermazioni che parlano di un dominio d'interesse ovvero come un grafo che rappresenta relazioni qualificate (annotate mediante predicati) tra i nodi. Se ci limitiamo a considerare RDF nella sua forma base, "la semantica" è legata fondamentalmente a queste due interpretazioni e alle operazioni che queste supportano. Da questo punto di vista, la semantica che possiamo attribuire ai dati in RDF non diverge molto dalla semantica che possiamo attribuire ai dati rappresentati utilizzando altri modelli, come, ad esempio, il modello relazionale. Rispetto a modelli tradizionali, RDF ha comunque alcune caratteristiche distintive.

In primo luogo, nelle basi di dati relazionali, abbiamo bisogno di uno *schema* prima di creare i dati veri e propri, e, inoltre, i dati inseriti devono rispettare i vincoli codificati in questo schema; in altri termini, lo schema pre-esiste ai dati e impone vincoli rigidi. RDF permette di definire uno schema (come vedremo tra poco), ma consente la massima flessibilità: non è necessario definire uno schema a priori e l'eventuale schema non impone vincoli rigidi sui dati che verranno inseriti [Batini&al. 2014]. Gli unici vincoli rigidi in RDF sono quelli definiti dalle sintassi definite per RDF (riprenderemo questa osservazione più avanti). In secondo luogo, è prassi consolidata usare molti termini del linguaggio naturale nelle descrizioni RDF, con il fine di renderle più facilmente interpretabili dagli esseri umani. Ad esempio, in

Figura 1 possiamo notare le triple che usano il predicato rdfs:label ma anche l'uso di URI comprensibili.

L'uso del linguaggio naturale nelle descrizioni è però utile anche per diversi tipi di elaborazione. Ipotizziamo di preparare una playlist da portarci in barca e cercare, in un insieme di descrizioni RDF di prodotti musicali, tutte le canzoni il cui titolo richiama la parola "boat". Ipotizziamo di avere una funzione di similarità applicabile a coppie arbitrarie di parole della lingua inglese, dove *sim*(*x*,*y*) rappresenta la similarità tra le parole *x* e *y*. Ad esempio, possiamo utilizzare Word2Vec [Mikolov&al.2013], un algoritmo che elabora grandi quantità di testo e costruisce funzioni di similarità che catturano il legame associativo tra le parole. Ipotizziamo ora di cercare tutte le canzoni, ovvero tutte le entità ?x per cui si afferma che <?x, rdf:type, dbo:Single> in un grafo RDF (ad esempio, con una

[10] https://www.w3.org/TR/rdf-sparql-query/

interrogazione SPARQL), ovvero tutti gli URI che identificano canzoni. Manteniamo di questi URI solo la parte che viene dopo il prefisso, anche chiamata *nome locale*. Applichiamo la funzione di similarità, e prendiamo tutte le entità tali che per il loro nome locale *x* vale *sim*("boat",*x*)>0.45[11]. Ad esempio, per la canzone dbr:sails otteniamo *sim*("boat","sails")=0.48 e la includeremo nella playlist. L'uso del linguaggio naturale nella descrizione di oggetti non è ovviamente peculiare del modello dei dati RDF (possiamo ottenere il medesimo risultato con una base di dati relazionale), ma è certamente soggetto a un'attenzione particolare quando si usa questo modello.

Per concludere, se pensiamo a RDF come un puro modello per la rappresentazione dei dati, avremo uno strumento per attribuire semantica ai dati con peculiari caratteristiche di flessibilità e ricchezza di descrizioni in linguaggio naturale, che è nativamente progettato per pubblicare e accedere a dati sul Web. Tuttavia, non ci stiamo discostando molto da altri modelli per la rappresentazione dei dati.

5.2 Grafi di conoscenza e semantica

Il vero motivo per cui RDF si pone come linguaggio di riferimento per la rappresentazione semantica dei dati è però un altro. Per scoprirlo dobbiamo considerare un secondo livello di rappresentazione, e, in particolare quel livello che ci consente di definire in qualche modo lo *schema* dei dati, e di utilizzare questo *schema* per supportare un tipo particolare di elaborazione. E' probabile che il lettore abbia una sufficiente conoscenza della lingua inglese per avere intuito il significato della descrizione riportata in

Figura 1, aiutato da un lato da nomi locali in linguaggio naturale (dbo:instrument, dbo:Musical_Artist, etc.), dall'altro dalla sue conoscenze di sfondo in campo musicale. E' probabile, infatti, che il lettore sappia che Elton John è un cantante, Eletric Piano è uno strumento, e che, quindi abbia intuito che dbo:instrument si usi per specificare lo strumento suonato dai musicisti.

Conoscenza linguistica e musicale pregresse permetteranno anche al lettore di intuire probabilmente che album e singoli sono entrambi prodotti musicali e che quindi possono avere un autore. Dobbiamo tenere presente però che tutte queste conoscenze di sfondo che possono fare parte del bagaglio culturale di molte persone, sono certamente inaccessibili alle macchine, o, se vogliamo, ad applicazioni che debbano interpretare questi dati. Il modello interpretativo fornito da RDF, se considerato da solo, supporta le tipiche operazioni di gestione dei dati (interrogazione, inserimento, etc.) ma è insufficiente per specificare il significato di nodi quali dbo:Musical_Artist o di predicati quali dbo:instrument.

L'utilizzo degli URI ci viene in aiuto, perché usando degli identificativi globali come gli URI, *possiamo* associare agli elementi del grafo un'interpretazione univoca, e recuperarla quando desiderato. Se non capisco il significato di dbo:artist (un nome di predicato ambiguo per un essere umano, potendo questo riferirsi alla relazione che collega musicisti ai loro prodotti musicali, o viceversa), l'URI mi dice che questo predicato è definito come un termine in uno spazio semantico che è dato dalla prima parte dell'URI, ovvero http://dbpedia.org/ontology/. Usando URI basati sul protocollo HTTP, mi aspetto di

[11] Il lettore può sperimentare i risultati ottenuti con una simile funzione di similarità ad esempio sul sito: http://bionlp-www.utu.fi/wv_demo/

poter localizzare la descrizione di questo predicato all'indirizzo http://dbpedia.org/ontology/artist. Una soluzione pragmatica potrebbe essere quella di puntare a descrizioni che chi pubblica dati è in grado di interpretare e possa seguire nella pubblicazione dei dati. Possiamo quindi indicare un insieme di termini consentiti, per esempio, tutti e soli i predicati che è lecito usare nelle descrizioni e dare di questi delle definizioni condivise in linguaggio naturale. Tuttavia, anche in questo caso, la semantica non sarebbe definita in maniera interpretabile da una macchina, ma solo da coloro che conoscono il linguaggio.

Cosa significa specificare il significato dei termini utilizzati in un grafo RDF? Abbiamo visto, nella prima sezione, che si tratta di utilizzare modelli interpretativi che, ponendo dei vincoli sulle possibili interpretazioni, permettano diversi tipi di elaborazione. Ciò che rende RDF *semantico* in maniera radicalmente diversa da altri modelli per la rappresentazione dei dati, è il supporto a meccanismi inferenziali. La capacità di fare *inferenze*, ovvero di trarre conclusioni da un insieme di premesse note, è una delle caratteristiche fondamentali dell'intelligenza che permette agli esseri umani di interpretare ed elaborare informazioni [Frixione1994]. Modelli interpretativi in grado di supportare inferenze sono stati elaborati nell'ambito della *logica matematica*, la disciplina che tradizionalmente si occupa dello studio del ragionamento corretto. La logica fornisce le basi teoriche per la maggior parte di quei linguaggi, modelli e applicazioni che oggi permettono di rappresentate conoscenze e fare inferenze su di esse. In questo capitolo non ci addentreremo nella definizione formale di specifici linguaggi, ma cercheremo di fare capire, usando alcuni linguaggi proposti nel web semantico, come modelli interpretativi di tipo logico contribuiscono all'elaborazione semantica dei dati.

5.3 Grafi di conoscenza e ontologie

Nelle sezioni seguenti, utilizzeremo il termine generico di *simbolo* per indicare URI, blank node e letterali utilizzati in RDF. Per poter specificare ulteriormente la semantica dei simboli usati nelle descrizioni RDF abbiamo bisogno di due ingredienti:

- Fissare l'interpretazione di alcuni simboli di base, ad esempio per predicati come rdf:type, rdfs:subClassOf, etc..
- Introdurre un insieme più ricco di asserzioni, che chiameremo *assiomi*, che usino tra gli altri quei simboli per cui abbiamo fissato l'interpretazione, cosi da codificare quei vincoli che permettono di specializzare l'interpretazione. Questo insieme di assiomi viene comunemente definito *ontologia*.

Partiamo da rdf:type (http://www.w3.org/1999/02/22-rdf-syntax-ns#type), un predicato speciale, il cui spazio semantico ci dice che è definito all'interno delle specifiche stesse di RDF. La definizione di questo predicato in RDF è la seguente:

```
rdf:type a rdf:Property ;
    rdfs:isDefinedBy <http://www.w3.org/1999/02/22-rdf-syntax-ns#> ;

    rdfs:label "type" ;
    rdfs:comment "The subject is an instance of a class." ;
    rdfs:range rdfs:Class ;
    rdfs:domain rdfs:Resource .
```

In questa definizione c'è un commento in linguaggio naturale che dice "The subject is an instance of a class.", e un insieme di altre asserzioni che richiamano un altro spazio semantico: RDFS. RDFS[12] è l'acronimo di RDF Schema, ed è un linguaggio che permette di definire schemi per dati in RDF.

Intuitivamente, la definizione ci dice che rdf:type è una rdf:Property e che il suo *dominio*, ovvero l'insieme di tutti i soggetti di triple che usano questo predicato, è costituito da generiche risorse (*risorsa* è il termine più generale usato in RDF per riferirsi a qualunque cosa possa avere un nome), e che il suo *codominio* (range, in Inglese), ovvero tutti gli oggetti di triple che usano questo predicato, è costituito da classi. Intuitivamente, stiamo dicendo che ogni volta che uso rdf:type sto specificando che l'oggetto della tripla è una *classe* e che il soggetto è un'*istanza* di questa classe.

Con RDFS abbiamo introdotto una fondamentale primitiva semantica, ovvero la distinzione tra simboli che denotano classi, ovvero insiemi di entità, e simboli che denotano *istanze*, ovvero entità individuali [Hitzler&al.2009]. Possiamo applicare questa distinzione ai letterali, dove le classi sono chiamate *tipi di dato,* in quanto denotano strutture per dati noti quali gli interi (xsd:integer), le stringhe (xsd:String), le date (xsd:DateTime) e così via. Per riferirsi genericamente a una classe o a un tipo di dato, possiamo usare anche il termine *tipo*. Con riferimento all'esempio in Figura 2, i nodi (simboli) dbo:Musical_Artist, dbo:Album e dbo:Single sono perciò nodi speciali, che individuano l'insieme degli artisti musicali, degli album e dei singoli (delle canzoni). E tuttavia, questa distinzione non aiuta ancora molto dal punto di vista della capacità che abbiamo di elaborare le informazioni in figura.

Di fatto, ciascuna asserzione è un assioma, ovvero una premessa che si ritiene essere vera. Tuttavia, gli assiomi rappresentati in Figura 2 ci permettono solo di rappresentare relazioni tra entità, descriverne alcune caratteristiche e definire la classe di cui le entità sono istanze. In altri termini, non abbiamo ancora aggiunto alcun assioma che ci permetta di interpretare meglio il significato di dbo:Musical_Artist, dbo:Album e dbo:Single e di definire dei rapporti generali tra tutte le entità che sono istanza di queste classi. Per farlo, possiamo estendere la rappresentazione in Figura 1, aggiungendo esplicitamente la distinzione tra istanze e classi.

Vediamo gli assiomi che abbiamo aggiunto. Abbiamo usato il predicato rdfs:subClassOf, definito nello spazio semantico RDFS. Questo predicato ci permette di dire che una classe è sottoclasse di un'altra classe. Ad esempio che dbo:Album e dbo:Single sono entrambi sottoclasse di dbo:Musical_Work. Abbiamo poi anche detto che dbo:Musical_Artist è sottoclasse di dbo:Artist, che, a sua volta, è sottoclasse di dbo:Person. Abbiamo, in pratica, introdotto una gerarchia tra le classi, con termini che ci permettono di riferirci a classi più generali, quali la classe dbo:Musical_Work per riferirsi sia ai brani (dbo:Single) sia agli album (dbo:Album).

[12] https://www.w3.org/TR/rdf-schema/

<dbr:Electric_Piano, rdfs:label, “Electric Piano”@en>
<dbr:Electric_Piano, rdfs:label, “Piano Elettrico”@it>
<dbr:Elton_John, dbo:instrument, dbr:Electric Piano>
<dbr:Empty_Sky, dbo:artist, dbr:Elton_John>
<dbr:Sails, dbo:musicalArtist, dbr:Elton_John>
<dbr:Sails, dbo:album, dbr:Empty_Sky>
<dbr:Elton_John, dbo:instrument, dbr:Electric_Piano>
<dbr:Elton_John, rdf:type, dbr:Musical_Artist>
<dbr:Sails, rdf:type, dbo:Single>
<dbr:Empty_Sky, rdf:type, dbo:Album>
<dbo:Musical_Artist, rdfs:subClassOf, dbo:Artist>
<dbo:Artist, rdfs:subClassOf, dbo:Person>
<dbo:Album, rdfs:subClassOf, dbo:Musical_Work>
<dbo:Single, rdfs:subClassOf, dbo:Musical_Work>

dbr: for http://dbpedia.org/resource/
e.g.: http://dbpedia.org/resource/Elton_John
dbo: for http://dbpedia.org/ontology/
rdf for http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs for https://www.w3.org/2000/01/rdf-schema#

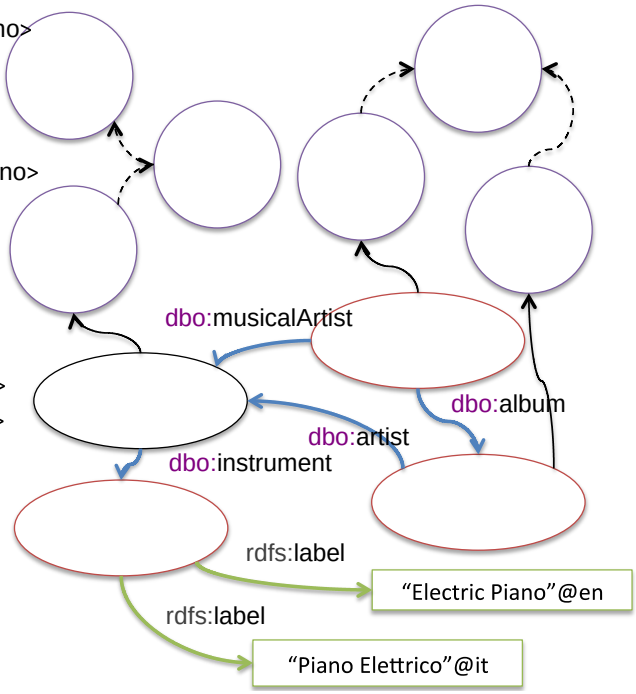


Figura 2 - Estensione del grafo RDF con assiomi per specificare il significato delle classi

Ovviamente possiamo usare assiomi più sofisticati. Ad esempio, dato un predicato come dbo:artist, come possiamo sapere utilizzarlo nelle triple? Di quali soggetti possiamo predicare dbo:artist e che tipo di oggetti compariranno nelle triple? Si noti bene che una possibile risposta è: qualsiasi cosa. Ma se diamo questa risposta, che differenza c’è tra dbo:artist e dbo:instrument? Se possiamo usare qualsiasi tipo di entità come soggetto e oggetto delle triple in cui compare un predicato, allora non stiamo specificando niente della sua semantica, ovvero del suo significato. Abbiamo detto che *il significato è il risultato di un’operazione di interpretazione*; come possiamo specificare in cosa diverge l’interpretazione di dbo:artist e dbo:instrument? Una possibile risposta consiste nello specificare appunto le classi a cui appartengono i soggetti e gli oggetti delle triple che contendono il predicato, ovvero specificare dei vincoli sul dominio e sul codominio dei predicati. Per esempio, possiamo introdurre i seguenti assiomi:

- <dbo:artist, rdfs:domain, dbo:Album>, che dice che tutti i soggetti delle triple in cui compare il predicato dbo:artist, appartengono alla classe dbo:Album;
- <dbo:artist, rdfs:range, dbo:Musical_Artist>, che dice che tutti gli oggetti delle triple in cui compare il predicato dbo:artist, appartengono alla classe dbo:Album;
- <dbo:instrument, rdfs:domain, dbo:Musical_Artist >, che dice che tutti gli oggetti delle triple in cui compare il predicato dbo:instrument, appartengono alla classe dbo:Musical_Artist

Si noti che ora abbiamo introdotto una distinzione tra dbo:artist e dbo:instrument: entità di tipo dbo:Musical_Artist e dbo:Album ricopriranno rispettivamente il ruolo di oggetti e soggetti nelle triple in cui appare dbo:artist. Possiamo fare di più, e usare un secondo predicato simile a dbo:artist, che chiameremo dbo: musicalArtist. A differenza del primo, il secondo lo usiamo per associare i brani

(dbo:Single) agli artisti musicali. Ma possiamo anche introdurre un predicato più generico di entrambi, ad esempio, il predicato dbo:relatedArtist. In questo modo avremmo introdotto una gerarchia tra i predicati.

Quello che stiamo facendo è costruire una semplice ontologia. Stiamo cioè specificando, attraverso degli assiomi, come usare i predicati (ad esempio dbo:artist) e le classi. Lo stiamo però facendo non attraverso linee guida, ma attraverso strumenti che ci consentono di elaborare i dati avendo fissato l'interpretazione di alcuni predicati (rdfs:subClassOf, rdfs:subPropertyOf, rdfs:domain, rdfs:range). Diciamo che nell'ontologia il significato è specificato in maniera formale e che è quindi elaborabile da una macchina. Cosa intendiamo esattamente con questa affermazione? Per spiegarlo, dobbiamo rispondere a due domande le cui risposte sono strettamente collegate:

1. Come funziona l'attribuzione di significato attraverso assiomi?
2. Cosa ci guadagniamo in termini di elaborazione delle informazioni?

5.4 Cosa significa definire la semantica dei termini di un ontologia?

La risposta alla prima domanda può essere data seguendo due approcci. Il primo approccio è quello ispirato alla teoria dei modelli e, in generale, alla tradizione della logica matematica. Cercheremo qui di dare solo qualche spunto utilizzando alcuni assiomi. Prendiamo le classi dbo:Musical_Artist e dbo:Artist e l'assioma che dice che la prima è sottoclasse della seconda. Ora, chiediamoci, come interpretiamo dbo:Musical_Artist? Intuitivamente lo interpretiamo come un insieme di individui in un dominio predefinito. A noi viene spontaneo, sulla base della conoscenza della lingua inglese, interpretarlo come l'insieme di tutti gli artisti musicali. Eppure, se non conosciamo la lingua, non abbiamo alcuno strumento per determinare che il simbolo dbo:Musical_Artist abbia proprio questa interpretazione. In altri termini, una macchina non avrà alcun modo di interpretare in questo modo il simbolo dbo:Musical_Artist e non quello, ad esempio, dbo:Artist.

Se consideriamo un simbolo in isolamento, è possibile attribuirgli significato mediante infinite interpretazioni che assegnano, al simbolo, un insieme di oggetti ciascuna. Allora possiamo fissare il significato del predicato di sottoclasse in maniera tale che l'espressione *A* è sottoclasse di *B* rappresenti un vincolo sulle possibili interpretazioni di *A* e di *B*. Tale vincolo stabilisce che tutti gli elementi che appartengono alla prima classe appartengono anche alla seconda, ovvero, che quale che sia l'insieme con cui interpretiamo *A* e quale che sia l'insieme con cui interpretiamo *B*, il primo insieme deve essere sottoinsieme del secondo. L'assioma <dbo:Musical_Artist, rdfs:subClassOf, dbo:Artist>, nel momento in cui lo prendiamo per vero, sta introducendo un vincolo sulle interpretazioni di dbo:Musical_Artist e di dbo:Artist, ovvero sta dicendo che quale che siano gli insiemi con cui interpretiamo dbo:Musical_Artist e di dbo:Artist, il primo insieme dovrà essere incluso nel secondo.

Secondo questo ragionamento, interpreteremo i simboli con cui identifichiamo specifiche entità (spesso chiamati costanti in logica) con degli elementi nell'insieme. Anche in questo caso, se prendiamo il simbolo dbr:Elton_John, possiamo interpretarlo con un infinito numero di elementi possibili presi da insiemi arbitrari. Però, nel momento in cui assumiamo per vero <dbr:Elton_John, rdf:type, dbr:Musical_Artist> stiamo introducendo un vincolo su tutte le possibili interpretazioni di

dbr:Elton_John e di dbr:Musical_Artist: stiamo dicendo che quale che sia l'individuo con cui interpretiamo dbr:Elton_John e quale che sia l'insieme con cui interpretiamo dbr:Musical_Artist, assumiamo che esista un vincolo che ci dice che l'individuo con cui interpretiamo dbr:Elton_John deve essere un elemento dell'insieme con cui interpretiamo dbr:Musical_Artist.

Possiamo ripetere questo ragionamento con assiomi in cui stabiliamo una relazione tra due entità. Innanzitutto assumiamo di interpretare un predicato, ad esempio, dbo:artist, come una relazione, ovvero, un insieme di coppie. <dbr:Empty_Sky, dbo:artist, dbr:Elton_John> ci dice che, quali che siano gli individui con cui interpretiamo dbr:Empty_Sky e dbr:Elton_John, e quale che sia la relazione con cui interpretiamo dbo:artist, la relazione con cui interpretiamo dbo:artist deve contenere una coppia costituita dai due individui con cui interpretiamo dbr:Empty_Sky e di dbr:Elton_John. Allo stesso modo, l'assunzione <dbo:artist, rdfs:subPropertyOf, dbo:relatedArtist> implica il vincolo per cui, quale che sia la relazione con cui interpretiamo dbo:artist, questa deve essere sottoinsieme di quella con cui interpretiamo dbo:relatedArtist.

L'insieme di tutti questi vincoli ontologici, nel complesso, determina il significato dei termini che definiamo nell'ontologia. In realtà, un insieme di simboli e un insieme di assiomi possono essere interpretati sempre in una molteplicità (tipicamente infinita) di modi; i vincoli che abbiamo introdotto hanno però eliminato una grande quantità di interpretazioni possibili, tenendo come buone solo quelle che soddisfano tutti i vincoli.

Ovviamente, può capitare che un insieme di assiomi comporti vincoli che non possono essere soddisfatti, nel loro complesso, da nessuna interpretazione. In questo caso, diciamo che l'ontologia è incoerente [Hitzler&al.2009]. I vincoli introdotti dagli assiomi, quando sono soddisfacibili, limitano le interpretazioni possibili dell'insieme di simboli in maniera sufficiente da consentirci di dedurre nuova conoscenza. Facciamo anche qui un esempio banale con i seguenti due assiomi: <dbr:Elton_John, rdf:type, dbr:Musical_Artist>, <dbo:Musical_Artist, rdfs:subClassOf, dbo:Artist>. Quale che sia l'interpretazione di dbr:Elton_John, di dbr:Musical_Artist e di dbo:Artist, sulla base dell'interpretazione che abbiamo dato a rdf:type e a rdfs:subClassOf, possiamo inferire con certezza che l'individuo con cui interpretiamo dbr:Elton_John sarà elemento anche dell'insieme con cui interpretiamo dbo:Artist, ovvero, se assumiamo come veri i due assiomi, saremo in grado di inferire l'asserzione <dbr:Elton_John, rdf:type, dbr:Artist>. In altri termini, potremo associare a dbr:Elton tutte le superclassi della classe che gli abbiamo attribuito negli assiomi. La semantica che abbiamo definito per i termini noti del linguaggio RDFS ci aiuta quindi a dedurre nuova conoscenza.

Il secondo approccio per attribuire la semantica è basato in effetti sulla definizione di un insieme di regole che specificano quali sono le conseguenze che possiamo trarre da un insieme di assiomi specificati nell'ontologia [Hitzler&al.2009]. Per il linguaggio RDFS, possiamo definire 13 regole, le cui più salienti, sono riportate in Figura 3.

If	**then**
<x, rdfs:subClassOf, y> <a, rdf:type, x>	<a, rdf:type, y>

<x, rdfs:subClassOf, y> <y, rdfs:subClassOf, z>	<x, rdfs:subClassOf, z>
<x, p, y> <p, rdfs:subPropertyOf, q>	<x, q, y>
<p, rdfs:subPropertyOf, q> <q, rdfs:subPropertyOf, r>	<p, rdfs:subPropertyOf, q>
<a, p, b> <p, rdfs:domain, x>	<a, rdf:type, x>
<a, p, b> <p, rdfs:range, x>	<b, rdf:type, z>

Figura 3 - Principali regole in RDFS

Se abbiamo discusso la semantica di diversi predicati utilizzando la semantica basata sulla teoria dei modelli, possiamo usare queste regole per analizzare più in dettaglio le restrizioni di dominio e codominio, ovvero quelle definite mediante i predicati rdfs:domain, rdfs:range. Come possiamo vedere, questi predicati inducono dei vincoli tali da farci inferire la classe a cui appartengono soggetti e oggetti delle triple. In altri termini, sebbene queste siano restrizioni, le usiamo per inferire nuove informazioni. Ipotizziamo, per esempio di avere una tripla <dbr:Empty_Sky, dbo:artist, dbr:Elton_John>, avendo definito <dbo:artist, rdfs:range dbo:Musical_Artist>. Ipotizziamo ora di aggiungere <dbr:help!, dbo:artist, dbr:The_Beatles>. Possiamo inferire che dbr:The_Beatles appartiene alla classe dbo:Musical_Artist. Eppure, se navighiamo lungo la gerarchia delle classi specificata in Figura 2 inferiremo anche che dbr:The_Beatles appartiene alla classe dbo:Person.

Quello che abbiamo fatto è avere ottenuto un'inferenza controintuitiva. Eppure non abbiamo ottenuto una contraddizione. Perché? Perché RDFS è un linguaggio molto semplice e non ha sufficienti primitive per consentirci di ottenere una contraddizione. Ci sono altri linguaggi come OWL 2[13] che estendono l'espressività di RDFS per permettere l'introduzione di vincoli più sofisticati [Hitzler&al.2009]. Ad esempio, OWL 2 introduce il predicato owl:disjointWith; se oltre agli assiomi precedenti, aggiungessimo <dbr:The_Beatles, rdf:type, dbo:Band>, allora inferiremmo che dbr:The_Beatles appartiene sia a dbo:Person sia a dbo:Band. Se infine nell'ontologia avessimo un assioma <dbo:Person, owl:disjointWith, dbo:Band>, ovvero un assioma che ci dice che le interpretazioni delle due classi non possono avere elementi in comune, otterremmo una vera e propria contraddizione. In OWL 2, infatti, possiamo avere ontologie contraddittorie e possiamo scoprire quando lo sono.

Possiamo ora caratterizzare l'approccio che abbiamo usato per definire la semantica dei dati, ovvero l'*approccio deduttivo*, impiegato nel Web semantico e a cui ci riferiamo quando parliamo di tecnologie semantiche. Tale approccio cerca di riprodurre alcune competenze cognitive proprie degli esseri umani, i quali sono in grado di trarre nuove conclusioni a partire da un insieme di premesse. Questo approccio è molto diverso dall'approccio tipico di molti sistemi di gestione dei dati in cui lo schema dei dati impone dei vincoli rigidi per assicurarsi che i dati rispettino lo schema [Batini&al.2014]. Quando definiamo un'ontologia, questa ontologia ci permette di dedurre nuove conoscenze mediante dei meccanismi

[13] https://www.w3.org/TR/owl2-overview/

inferenziali, che possono essere riprodotti meccanicamente da sistemi per il ragionamento automatico. In alcuni casi, il risultato è la scoperta di una contraddizione; in molti altri casi possiamo inferire asserzioni che non erano presenti nell'insieme delle conoscenze di partenza, ovvero nell'insieme degli assiomi che caratterizzano l'ontologia.

Si tenga presente che a titolo esemplificativo abbiamo considerato assiomi estremamente semplici. Linguaggi di rappresentazione della conoscenza come OWL 2 permettono di definire quali predicati sono transitivi, quali sono da interpretarsi l'uno come l'inverso dell'altro, quali si comportano come funzioni, etc. Altri linguaggi trattano problemi ancora più complessi, ovvero come effettuare inferenza sulla base di informazioni mancanti, o come trattare in maniera nativa l'incertezza associata ad alcune affermazioni.

5.5 A cosa serve definire formalmente la semantica dei termini usati in un grafo di conoscenza?

Piuttosto che approfondire i precedenti temi, ci preme rispondere ora alla seconda domanda: cosa ci guadagniamo in termini di elaborazione delle informazioni? In cosa può essere utile l'approccio deduttivo?

Una prima risposta discende intuitivamente dal ragionamento precedente: possiamo estendere l'insieme di asserzioni o, secondo la prospettiva del grafo, l'insieme di archi che connettono i nodi. Ma ci sono diverse ragioni per cui questo risulta particolarmente utile.

Pensiamo a una semplice domanda sul grafo in Figura 3. Ipotizziamo di volere cercare, nel nostro grafo, tutte le persone che suonano il piano. In SPARQL possiamo chiedere che i risultati siano di una certa classe, in questo caso dbo:Person, e possiamo usare il predicato dbo:instrument come filtro. Se non facciamo inferenza, non troveremo tra le risposte dbr:Elton_John, il quale non ha come classe dbo:Person ma una sua sottoclasse.

Aggiungendo al grafo le inferenze discusse nei precedenti paragrafi, siamo invece in grado di inferire che dbr:Elton_John è anche di tipo dbo:Person. Ma perché ciò dovrebbe esserci utile? Perché è vero che la classe principale di cui predichiamo dbo:instrument è quella dei Musical_Artist, però questo predicato è usato anche con altre classi, come, ad esempio, dbo:Music_Genre, dbo:Band, dbo:Classical_Music_Artist, dbo:Guitarist. Questo perché l'attribuzione della classe in grafi di conoscenza complessi avviene sulla base dell'individuazione di una classe principale. Ci sono chitarristi che suonano il piano e persino generi musicali il cui strumento principale è il piano: ma a noi interessano solo le persone, ovvero tutti quegli individui che appartengono a una sottoclasse di dbo:Person. Usare l'inferenza ci permette perciò di essere molto flessibili nell'organizzazione dei dati, una caratteristica che diventa fondamentale quando si gestiscono un gran numero di entità appartenenti a domini diversi.

Il risultato di quanto si è cercato di spiegare fin qui, è che le ontologie usate nel web semantico e in RDF hanno alcune differenze salienti rispetto all'uso degli schemi dei dati:

- Mentre nella maggior parte delle basi di dati, ad esempio quelle relazionali, si usa un approccio *prescrittivo* (i vincoli sono tutelati da regole che governano l'inserimento dei dati), l'approccio alla

definizione della semantica ispirato dalla logica è *deduttivo*: si estrapolano nuove conoscenze a partire dalla definizioni che costituiscono lo schema;

- L'uso delle ontologie permette un'organizzazione dei dati flessibile: l'ontologia non deve necessariamente pre-esistere ai dati (non è un caso se abbiamo potuto introdurre RDF prima di introdurre RDFS e OWL) e questi possono discostarsi dallo schema; ovviamente possono emergere contraddizioni e inconsistenze, che possono essere rilevate da un sistema per il ragionamento automatico.

Abbiamo cercato di approfondire la semantica legata all'uso di inferenze, in quando fondamentale tratto d'unione tra le nostre esperienze cognitive e le procedure di elaborazione delle informazioni dotate di utilità pratica. Abbiamo visto uno strumento per cercare di esplicitare la semantica dei simboli utilizzati attraverso assiomi che ci permettono di derivare nuove conoscenze. Oggi esiste una varietà di sistemi in grado di calcolare inferenze[14] a partire da assiomi definiti usando RDFS, OWL e altri linguaggi ancora (ad esempio RDFS++[15] e SWRL[16]).

6. Semantica e similarità

Un aspetto tipico della nostra capacità di manipolare il linguaggio sulla base di un'attribuzione di significato, è la capacità di inferire nuove conoscenze a partire da informazioni disponibili e conoscenze di sfondo (entrambe codificabili come assiomi). Un'altra capacità fondamentale è quella di riconoscere quando due cose (oggetti, entità, concetti) sono simili. Tale capacità è alla base di compiti cognitivi di base come la categorizzazione [Sloutsky&Fisher2004] e il ragionamento analogico, una forma di ragionamento che si sviluppa fin dai primissimi anni di età [Thibaut&al.2010,Vosniadou1989]. La valutazione della similarità è fondamentale per una molteplicità di compiti applicativi ed è uno degli aspetti costitutivi della semantica.

6.1 Similarità e integrazione di informazioni eterogenee

Uno dei requisiti per integrare dati provenienti da sorgenti differenti è l'*allineamento* tra gli elementi delle due sorgenti. In questo caso, ci ritroviamo con tutte le problematiche che riguardano l'integrazione dati che sono state discusse nel Capitolo 6. Nel campo dei grafi di conoscenza, per allineamento si intende la specifica delle *relazioni* che intercorrono tra le terminologie (tipi, predicati, entità) utilizzate in grafi distinti. L'allineamento è necessario per poter interrogare grafi di conoscenza distinti come se fossero un unico grafo. I linguaggi di rappresentazione della conoscenza introdotti nelle sezioni precedenti mettono a disposizione predicati, a cui è stata assegnata una semantica formale condivisa, per rappresentare esplicitamente le principali primitive da usare in un allineamento. Si prenda ad esempio la Figura 4.

[14] http://owl.cs.manchester.ac.uk/tools/list-of-reasoners/

[15] Un linguaggio che introduce alcune primitive di OWL 2 a RDFS, supportato, ad esempio dal sistema AllegroGraph https://franz.com/agraph/allegrograph/dynamic-materialization.lhtml

[16] https://www.w3.org/Submission/SWRL/

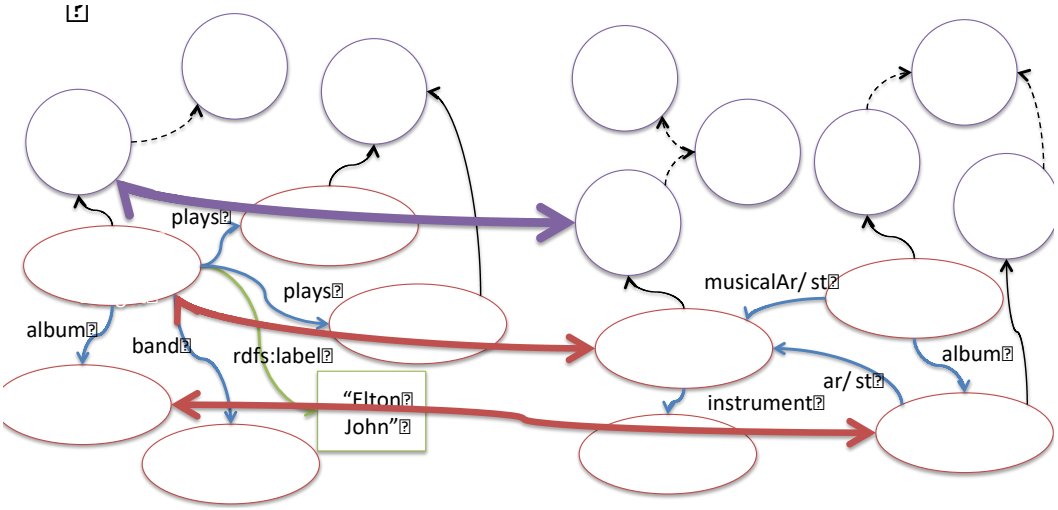


Figura 4 - Allineamento tra grafi di conoscenza

In questa figura mostriamo due descrizioni provenienti da grafi di conoscenza diversi di Elton John. Come si può vedere, la stessa entità è rappresentata da nomi locali differenti. Abbiamo stabilito delle frecce che ci dicono quali URI rappresentano la stessa entità del mondo reale e le stesse classi.

Usando linguaggi come RDFS e OWL possiamo definire formalmente questi mapping utilizzando un insieme di predicati condivisi. Ad esempio, ipotizzando che k1 e k2 denotino rispettivamente lo spazio semantico di due grafi distinti, possiamo definire le relazioni che definiscono l'allineamento tra i due grafi utilizzando le seguenti asserzioni:

- <k1:Reginald_Kenneth_Dwight, owl:sameAs, k2:Elton_John>
- <k1:Empty_Sky, owl:sameAs, k2:Empty_SkyA>
- <k1:plays, owl:equivalentPropertyOf, k2:instrument>
- <k1:album, owl:inverseOf, k2:artist, k2:artist>
- <k1:Musician, owl:equivalentClassOf, k2:MusicalArtist>
- <foaf:Person, owl:equivalentClassOf, k2:Person>

Come possiamo vedere, abbiamo introdotto i seguenti predicati: owl:sameAs, che rappresenta la relazione di identità tra identificativi di entità, owl:equivalentClassOf e owl:equivalentPropertyOf, che rappresentano rispettivamente l'equivalenza tra due classi e due predicati, owl:inverseOf, che stabilisce che il primo predicato rappresenta una relazione inversa rispetto a quella rappresentata dal secondo predicato. Asserzioni come queste possono essere utilizzate in due modi:

- per interrogare i due grafi di conoscenza sfruttando le inferenze supportate dall'interpretazione formale dei predicati; ad esempio, posso chiedere chi sono i musicisti (k1:Musician) autori (k2:artist) dell'album Empty Sky (k1:Empty_Sky), mescolando classi e predicati dei due grafi di partenza, e ottenendo in risposta i due identificativi di Elton John nei due grafi (k1:Reginald_Kenneth_Dwight e k2:Elton_John);

- per costruire un grafo unico, ad esempio quello rappresentato in Figura 5, una volta applicate politiche di fusione dei dati più sofisticate.

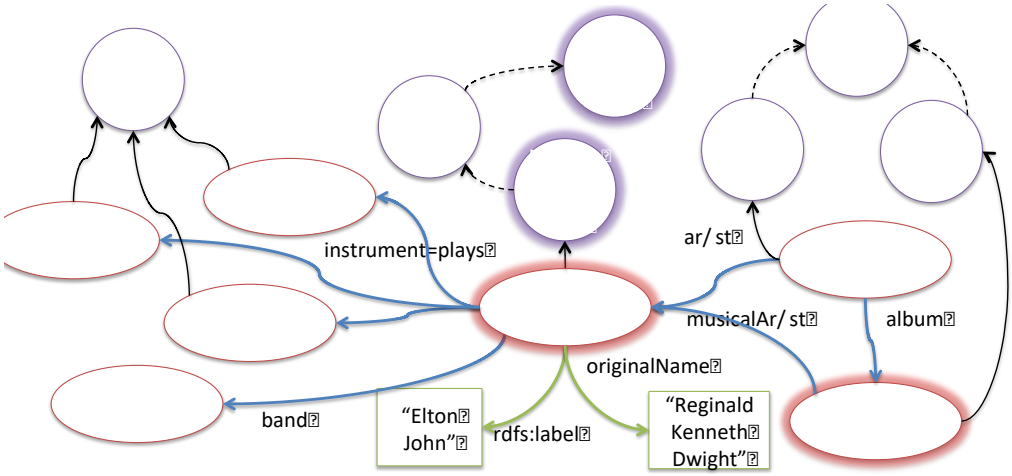


Figura 5 - Grafo integrato risultante dalla fusione dei grafi nella figura precedente

Definire un allineamento in maniera manuale è però un compito difficile e costoso, e può essere infattibile quando due grafi sono di grandi dimensioni (si pensi che l'ontologia usata da DBpedia contiene più di mille proprietà, circa cinquecento classi, e circa quattro milioni e mezzo di istanze[17]).

Il problema da affrontare è quindi *come derivare automaticamente* le relazioni che servono per allineare i due grafi di partenza [Cheatham&Pesquita2017]. Il problema non è molto diverso da quello visto con le basi di dati, anche se si usa una terminologia spesso leggermente differente. Quando dovremo trovare relazioni, o *mapping*, tra gli elementi di due ontologie usate in grafi di conoscenza differenti, dovremo occuparci del problema noto come *Ontology Matching* [Shvaiko&Euzenat2011, Ochieng&Kyanda2018].

Se ci concentriamo sulle entità, cioè sulle istanze delle classi, dovremo affrontare un problema noto come *Instance Matching* [Shvaiko&Euzenat2011]. Questi due problemi si riferiscono al tentativo di trovare classi e predicati che sono equivalenti in ontologie diverse (o sottoclassi e sottopredicati di altre classi [Cruz&al.2013]), e identificativi di entità che si riferiscono alla medesima entità del mondo reale. Possiamo però generalizzare questa interpretazione del problema e affrontare il problema noto come Link Discovery, ovvero il problema di scoprire delle relazioni tra i dati [Sherif&Ngonga-Ngomo2018,Achichi&al.2019]. In questo caso, le relazioni possono essere diverse da equivalenza, sottoclasse, sottopredicato e identità; ad esempio possiamo cercare delle relazioni di parte-tutto.

Sono state proposte in letteratura una grande quantità di tecniche e sistemi per cercare di fornire una soluzione automatica a questi problemi (si vedano i riferimenti indicati sopra). Tendenzialmente, queste

[17] https://wiki.dbpedia.org/about/facts-figures

tecniche prendono ispirazione da quelle sviluppate per le basi di dati relazionali, ma cercano di considerare e sfruttare alcune peculiarità di ontologie e grafi di conoscenza. Ad esempio, le ontologie includono profonde gerarchie di concetti; queste tecniche cercano di usare le relazioni tra le classi definite nella gerarchia per determinare la similarità dei concetti (si stima che sottoclassi di concetti equivalenti siano più simili fra loro). Per fare un altro esempio, la prassi che ci porta a usare in abbondanza nomi e descrizioni in linguaggio naturale nei grafi di conoscenza, ci consente di usare molti metodi provenienti dall'elaborazione del linguaggio naturale.

Queste tecniche fanno spesso ricorso a dizionari (ad esempio "musicista" è sinonimo di "musico" in Italiano) o a modelli basati sull'apprendimento del linguaggio naturale che ci permettono di definire la similarità tra parole (ad esempio un algoritmo come Word2Vec [Mikolov&al.2013] stimerà un legame di associazione piuttosto forte tra le parole "intrument" e "play"). Ovviamente, possiamo sfruttare altri elementi strutturali, ad esempio, predicati che hanno domini e codomini simili saranno più simili tra loro. Infine, possiamo usare la logica, per esempio per correggere alcune relazioni, scoperte magari usando le descrizioni in linguaggio naturale, ma che si rivelano contraddittorie, rispetto alle ontologie di partenza [Jiménez-Ruiz&Grau2011,Pesquita&al.2013].

In ogni caso, per problemi complessi come Ontology Matching, occorre pensare ai metodi automatici come a meccanismi per assistere gli utenti nello stabilire l'allineamento; a questi verrà poi chiesto di validare i risultati ottenuti mediante metodi automatici [Dragisic&al.2016], i quali, a loro volta, dovrebbero imparare dall'interazione con l'utente per raffinare i risultati ottenuti [Cruz&al.2016]. Esistono ormai diverse applicazioni che supportano compiti come Ontology Matching e Link Discovery; tra le più robuste segnaliamo soprattutto AgreementMakerLight [Faria&al.2013] e LogMap [Jiménez-Ruiz&Grau2011] per Ontology Matching e LIMES [Ngomo&Auer2011] e SILK [Isele&Bizer2012] per Instance Matching e Link Discovery.

6.2 - Similarità ed esplorazione della conoscenza: l'esempio dei sistemi di raccomandazione

La similarità è però uno strumento fondamentale non solo per integrare informazioni, ma anche per supportare l'esplorazione della conoscenza rappresentata. Un'applicazione fondamentale della similarità è quella a supporto dei sistemi di raccomandazione. Un sistema di raccomandazione cerca di raccomandare a un utente degli oggetti simili a quelli che sono interessati ad utenti a lui simili [Aggarwal&al.2016]; come si può intuire, implementare un sistema ispirato a questo principio euristico ci porta a trattare due problemi:

- definire quando possiamo stimare che un oggetto sia di interesse per un utente;
- definire delle funzioni per determinare la similarità tra utenti e oggetti.

L'approccio più tradizionale sviluppato nell'ambito dei sistemi di raccomandazione, chiamato *collaborative filtering*, ci permette di stimare similarità tra utenti e oggetti sulla base dell'interesse che gli utenti hanno espresso nei confronti degli oggetti. Tuttavia, questa tecnica di base, molto potente, comporta diversi problemi, come la raccomandazione di oggetti nuovi (non ancora di interesse per nessun utente) o per utenti nuovi (che non hanno ancora espresso interesse per nessun utente), la

scarsa disponibilità di dati sugli utenti e la sparsità della matrice che definisce le relazioni tra utenti e oggetti.
Per risolvere questi problemi, sempre più approcci oggi complementano questa tecnica con un analisi semantica dei contenuti [Lops&al.2011]. Se stiamo raccomandando brani, possiamo usare la similarità tra le onde di due brani diversi (in realtà non basta, ma ci rifacciamo a un tutorial online[18] per maggiori dettagli sullo stato attuale delle tecniche di raccomandazione in ambito musicale). Ma se vogliamo raccomandare ad un utente gli artisti più simili a quelli che gli sono piaciuti, come possiamo fare? Possiamo certamente usare anche i dati che descrivono questi artisti [Lops&al.2011, Musto&al.2017]. Ad esempio, due musicisti saranno tanto più simili quanto maggiore è il numero di strumenti che suonano in comune, e quanto più simili sono i generi musicali a cui sono associati.

Esistono una quantità di metodi più o meno complicati per determinare questa similarità. Possiamo usare misure di similarità sul grafo, ma possiamo anche re-interpretare il grafo in una maniera che renda più semplice, efficiente ed efficace la valutazione della similarità, ovvero utilizzando le informazioni nel grafo per generare questi vettori [Oramas&al.2017, Musto&al.2017]. Per ulteriori dettagli sull'utilizzo di grafi di conoscenza a supporto dei sistemi di raccomandazione, rimandiamo ai riferimenti bibliografici qui inclusi.

6.3 Similarità e interpretazione

Qui ci limitiamo a concludere osservando che la similarità è un elemento fondamentale per caratterizzare il significato delle entità, sia da un punto di vista cognitivo (la capacità di determinare la similarità tra gli oggetti è una competenza cognitiva fondamentale per gli esseri umani, che si sviluppa nei primi anni dello sviluppo) sia da un punto di vista pratico e particolarmente rilevante per Scienza dei Dati (la capacità di determinare la similarità tra entità e termini ontologici è fondamentale per supportare diversi tipi di applicazioni).

Diversi strumenti per analizzare la similarità saranno più o meno utili, a seconda dell'applicazione e dei dati che stiamo trattando. Ad esempio, se vogliamo usare una funzione di similarità per scoprire relazioni di identità tra entità, avremo bisogno che la funzione si comporti bene nel discriminare tra entità identiche ed entità distinte, mentre se vogliamo usare una funzione di similarità per raccomandare musicisti simili, avremo bisogno di una funzione di similarità in grado di scoprire pattern di similarità più sottili [Oramas&al.2017]. Se abbiamo una rappresentazione gerarchica dei generi musicali, per utilizzare i generi musicali nella valutazione della similarità tra musicisti, potranno essere utili le relazioni presenti nella gerarchia dei generi; se tale rappresentazione gerarchica non è disponibile nei dati, potremo solo valutare quali generi associati ai musicisti sono condivisi.

Una caratteristica comune alla diverse tecniche di similarità proposte in letteratura è che tali tecniche *interpretano* i dati a disposizione in strutture opportune, per poter determinare la similarità tra gli oggetti. Così facendo, queste tecniche estrapolano il significato dei dati elaborati secondo modelli interpretativi ottimizzati per risolvere un determinato compito, ad esempio, la scoperta di relazioni di identità tra entità, o la raccomandazione di entità di interesse per gli utenti.

[18] https://www.slideshare.net/FabienGouyon/music-recommendation-2018-116102609

Anche la valutazione della similarità è il risultato di un processo interpretativo. In questo processo si sviluppano algoritmi in grado di interpretare i dati disponibili in un modello, ad esempio il dominio delle stringhe, un grafo, o uno spazio vettoriale, che ci permetta di stabilire quanto due elementi sono vicini. Tipicamente, la similarità tra due oggetti può essere interpretata come una funzione inversamente correlata a una funzione di *distanza* definita nel modello (ad esempio, data una misura di distanza *dis*(x,y) con codominio nell'insieme [0,1] e due elementi *x* e *y,* è possibile definire una funzione di similarità come *sim*(x,y)=1-*dis*(x,y)). Per questo motivo, molti modelli interpretativi usati per valutare la similarità tra oggetti sono definiti in spazi metrici, ovvero in spazi in cui sia possibile calcolare la distanza tra gli oggetti.

Tipicamente le funzioni di similarità e di distanza sono vincolate da un limite inferiore (ad esempio, una similarità uguale a zero indica che due elementi sono completamente dissimili) e da un limite superiore (ad esempio, ci si aspetta che la similarità valutata tra due elementi identici sia uguale a uno). Esistono tuttavia diverse eccezioni, e diverse misure utili per calcolare la similarità che non rispettano alcune di queste proprietà formali (ad esempio, può essere conveniente utilizzare la funzione *Lucene Conceptual Scoring*[19] per valutare la similarità tra insiemi di parole, una funzione supportata da sistemi per la gestione di testi basati su Lucene[20]; eppure, per varie ragioni tra cui l'ottimizzazione delle prestazioni, il Lucene Conceptual Score non ha un limite superiore).

Nella valutazione della similarità, ci riferiamo alla semantica dal punto di vista di quelle tecniche volte ad estrapolare il significato dei dati, cercando il modello interpretativo migliore rispetto alla particolare applicazione della similarità che cerchiamo di supportare (ad esempio, Link Discovery vs. sistemi di raccomandazione).

7. Semantica ed estrazione di informazioni

Discutiamo infine un ultimo tipo di applicazione in cui trattiamo la semantica dei dati nel senso dell'estrapolazione del significato attraverso modelli interpretativi. Gran parte delle informazioni oggi prodotte da esseri umani e scambiate sono disponibili in linguaggio naturale [Chen&Zhang2014]. Abbiamo a disposizione certamente molte tecniche per interpretare il linguaggio naturale. Ad esempio, i sistemi tradizionali di Information Retrieval usano matrici termini-documenti, e quindi, vettori di termini associati a ciascun documento per stimare la similarità tra due testi, ovvero tra una interrogazione, rappresentata mediante un vettore di termini che la compongono, e i documenti compresi in una collezione (ad esempio utilizzando la funzione Lucine Conceptual Scoring). Questo tipo di modello interpretativo però non ci consente elaborazioni più sofisticate, come, ad esempio, rispondere a interrogazioni fattuali [Kolomiyets&al.2011] come: quali sono i membri della band Coldplay? Quali sono i documenti in cui si parla di almeno un membro della band Coldplay?

Per supportare applicazioni di questo tipo, occorre interpretare anche dei testi, all'interno di modelli interpretativi più sofisticati. Un approccio abbastanza consolidato consiste nell'usare modelli

[19] https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html
[20] https://lucene.apache.org/core/index.html

interpretativi *fattuali*, che supportino interrogazioni come quelle che possiamo eseguire su dati strutturati e grafi di conoscenza [Höffner&al.2017].

Secondo questo approccio, per supportare compiti di questo tipo abbiamo bisogno di estrarre dati strutturati a partire da documenti non strutturati, un problema tra quelli più caratterizzanti l'Intelligenza Artificiale [Etzioni&al.2006, Mitchell&al.2018]. Anche in questo caso intendiamo solo fornire alcune definizioni salienti e riferimenti bibliografici, lasciando al lettore l'approfondimento.

Facciamo uso di un esempio; in Figura 6 mostriamo il risultato dell'elaborazione mediante ClearForest Gnosis, una funzione del browser Firefox che effettua alcuni compiti base di analisi del linguaggio naturale che vanno nella direzione dell'estrazione di dati strutturati da testi[21].

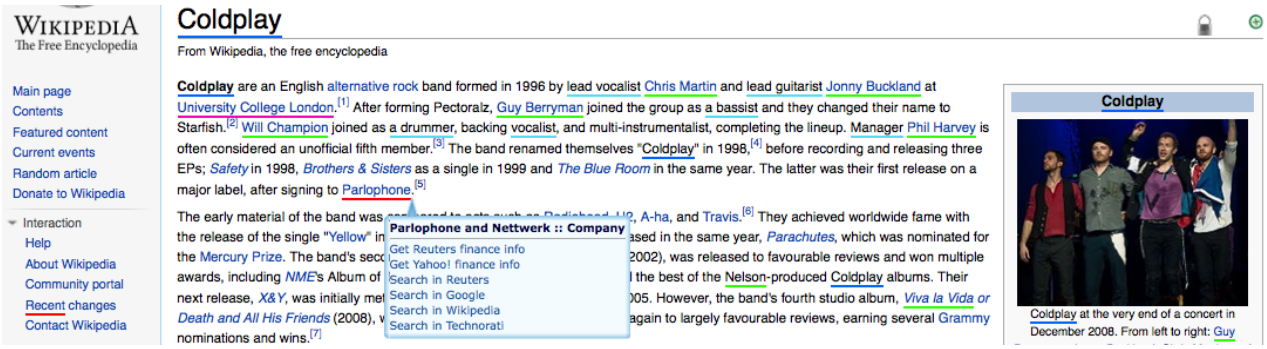


Figura 6 - Pagina di Wikipedia che descrive la band Coldplay
elaborata con lo strumento ClearForest Gnosis (da www.wikipedia.org)

Discutiamo qui due compiti tipici dell'elaborazione del linguaggio naturale legati all'uso di rappresentazioni strutturate come i grafi di conoscenza:

- **l'estrazione di entità** è il compito consistente nel riconoscere le sequenze di parole all'interno di un testo che rappresentano menzioni di entità del mondo reale, ad esempio, dbr:Coldplay; applicato un sistema che riconosce entità a un testo, otterremo un insieme di entità, ciascuna associata alle sequenze di parole in cui è stata *menzionata* nel testo;
- **l'estrazione di relazioni** è il compito di riconoscere relazioni tra entità, ad esempio <dbr:Coldplay, dbo:bandMember, dbr:Chris_Martin> che sono più o meno esplicitamente implicate nel testo; interpretando tali relazioni come asserzioni, ad esempio "i Coldplay hanno Chris Martin tra i loro membri", possiamo associare ai documenti un insieme di asserzioni provenienti dal testo, che potremo usare per popolare una base (o un grafo) di conoscenza, che possiamo poi interrogare.

E' abbastanza intuitivo riconoscere che il secondo problema dipende del primo e che richieda tecniche notevolmente più sofisticate, in grado di trattare la complessità del linguaggio naturale. Si noti, ad esempio, che nel testo non compare la parola "band member"; eppure, un essere umano che conosca il predicato che usiamo in un'ontologia per associare musicisti alle rispettive band (dbo:bandMember),

[21] Il plug-in non sembra più essere supportato; ci sono però diversi servizi che offrono funzionalità analoghe come ad esempio le API di Gate (https://gate.ac.uk/)

dovrebbe essere in grado di riconoscere che <dbr:Coldplay, dbo:bandMember, dbr:Chris_Martin> è implicata dall'espressione "Coldplay […] formed in 1996 by lead vocalist Chris Martin".

Si osserverà che in realtà questa interpretazione richiede certe conoscenze di sfondo; ad esempio se possiamo affermare con certezza che Chris Martin è stato membro della band perché chi fonda una band ne è membro, non sappiamo se Chris Martin sia ancora membro della band. In altri termini, l'interpretazione della conoscenza fattuale presente in un testo è un compito molto complicato e ancora ben lungi dall'essere risolto, nonostante gli incredibili progressi raggiunti nel campo dell'elaborazione del linguaggio naturale negli ultimi anni [Etzioni&al.2006,Mitchell&al.2018, Niklaus&al.2018].

Concentriamoci invece ora sull'estrazione di entità, un compito che, su alcune tipologie di testi semplici viene oggi svolto con livelli di precisione sufficiente a supportare molte applicazioni industriali. Quando parliamo di estrazione di entità di riferiamo tipicamente a due tipi di compiti:

- *Named Entity Recognition: il* compito di trovare menzioni di entità e determinare la classe a cui appartiene ciascuna entità tra un insieme di classi note. Questo è esattamente il compito svolto da Clear Forest Gnosis di cui vediamo il risultato in
-
- Figura . Lo strumento ha individuato che c'è una menzione dell'entità chiamata Coldplay, a cui associa una classe (identificata dal colore blu, che rappresenta le organizzazioni e i gruppi di persone), una menzione di Chris Martin (una persona), una menzione di University College London (un luogo), e così via.
- Named Entity Linking: se abbiamo una base di conoscenza di riferimento, ad esempio un grafo di conoscenza, è possibile che siamo particolarmente interessati a riconoscere menzioni delle entità presenti nel grafo di conoscenza all'interno di un testo. Allora non possiamo limitarci a riconoscere la menzione di un'entità e determinarne la classe, ma dovremo anche essere in grado di *disambiguare* la menzione, per capire a quale delle tante entità rappresentate nel grafo di conoscenza corrisponde l'entità trovata. Ad esempio, dovremo essere in grado di collegare la menzione dell'entità Coldplay di tipo Organizzazione, con l'entità dbr:Coldplay descritta in DBpedia. La disambiguazione è uno dei compiti più difficili e importanti che un sistema di Named Entity Linking deve risolvere, soprattutto se pensiamo a quanto comuni possono essere alcuni nomi di persone, organizzazioni e luoghi.

Anche in questo caso, il secondo compito è svolto successivamente al primo; in generale, un sistema che estrae entità dovrebbe essere in grado di riconoscere quando una menzione si riferisce a un'entità descritta nel grafo di conoscenza e stabilire il collegamento in maniera corretta, ma anche di riconoscere quando l'entità menzionata è nuova, ovvero non compare nel grafo di conoscenza, per eventualmente aggiornare il grafo di conoscenza. Per una dettagliata analisi di tecniche di Named Entity Linking si rimanda alla letteratura esistente [Shen&al.2015, Derczynski&al.2015].

Le tecniche di estrazione di entità sono oggi impiegate in una molteplicità di applicazioni anche di natura industriale, perché giocano un ruolo fondamentale nell'integrazione di dati non strutturati e di provenienti da sorgenti eterogenee (strutturati e non strutturati). Esempi di servizi che supportano la

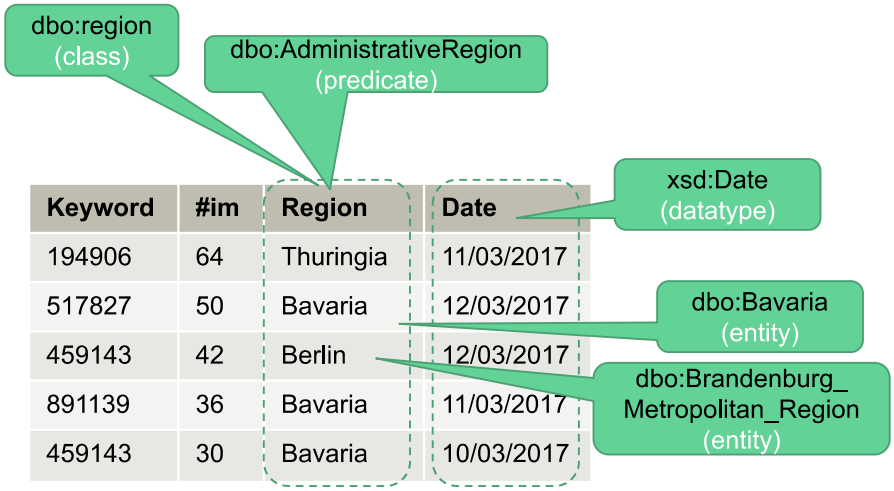
disambiguazione con DBpedia e Wikipedia sono DBpedia Spotlight[22], GATE APIs, TextRazor[23], Dandelion[24] (anche disponibile per la lingua Italiana), Wikifier[25] (multilingua) e Babelfy[26] (multilingua).

Questi servizi permettono di migliorare molte altre applicazioni, quali ad esempio gli agenti conversazionali. Ipotizziamo di voler contattare il centro assistenza di un negozio in cui abbiamo comprato il prodotto MacBook Pro. L'agente conversazionale che elaborerà la nostra richiesta dovrà riconoscere nel testo che scriviamo l'entità che rappresenta il prodotto da noi acquistato, recuperando i dati relativi ad esso per offrire un servizio migliore. Ad esempio, potrebbe riconoscere che abbiamo usato un'espressione ambigua e sulla base del confronto con il nome con la base di conoscenza ci farà una domanda per disambiguare la nostra espressione e capire se ci riferiamo a un MacBook Pro del 2016 o del 2019.

I grafi di conoscenza forniscono rappresentazioni particolarmente utili per supportare questi task di elaborazione del linguaggio naturale. Questo ruolo si basa sia su elementi strutturali (il grafo) che su elementi pragmatici (l'uso abbondante di annotazioni in linguaggio naturale). L'elaborazione del linguaggio naturale gioca a sua volta un ruolo fondamentale nelle applicazioni per l'integrazione di dati non strutturati e dati eterogenei [Kejriwal&al.2018]. Il ruolo fondamentale che giocano i grafi di conoscenza nel supporto all'elaborazione del linguaggio naturale, e di conseguenza, nelle architetture per integrare, organizzare e supportare l'accesso a sorgenti massive di dati, è probabilmente una della caratteristiche che ha favorito la loro diffusione più di qualunque altra.

Infine, segnaliamo che problemi simili a quelli riscontrati nell'interpretazione del linguaggio naturale, come l'estrazione di entità e di relazioni, si riscontrano anche quando vogliamo interpretare dati strutturati ma solo debolmente, e non associati con un'interpretazione esplicita del loro significato. Ad esempio, diversi metodi sono stati proposti per estrarre informazioni dalla grande quantità di tabelle presenti sul web, o dalla grande quantità di dati disponibili in formati tabellari come file CSV. L'interpretazione, in questo caso, consiste nell'associare classi e predicati di un'ontologia ed entità di una base di conoscenza a colonne, coppie di colonne e valori presenti nella tabella [Keshvari-Fini&al.2019,Vu&al.2019,]. Un esempio di queste associazioni, chiamate anche *annotazioni semantiche,* è riportato in Figura 7. Come nel caso dell'estrazione di informazioni da testi, le informazioni opportunamente interpretate possono essere usate per popolare basi e grafi di conoscenza, supportando elaborazioni più sofisticate [Paulheim2017].

[22] https://github.com/dbpedia-spotlight/
[23] https://www.textrazor.com/
[24] https://dandelion.eu/
[25] http://wikifier.org/
[26] http://babelfy.org/



Keyword	#im	Region	Date
194906	64	Thuringia	11/03/2017
517827	50	Bavaria	12/03/2017
459143	42	Berlin	12/03/2017
891139	36	Bavaria	11/03/2017
459143	30	Bavaria	10/03/2017

Figura 7 - Esempio di interpretazione semantica di una tabella che descrive statistiche su campagne di marketing digitale

8. Conclusioni

Abbiamo visto come l'oggetto principale della semantica applicata ai dati sia lo sviluppo di modelli interpretativi che permettano di elaborare i dati in maniera sofisticata, replicando alcune delle competenze cognitive che caratterizzano l'intelligenza umana. Questi modelli interpretativi sono spesso derivati da modelli propri dell'Intelligenza Artificiale, con applicazioni dirette alla gestione di dati eterogenei.

Tra le competenze cognitive che è utile replicare per favorire l'elaborazione dei dati ci sono in particolare:

- la capacità di effettuare inferenze, ovvero di derivare nuove conoscenze da conoscenze note,
- la capacità di riconoscere se oggetti rappresentati nei dati siano simili o addirittura identici, e
- la capacità di elaborare informazioni non strutturate o debolmente, come la sequenza di parole in un testo o tabelle, organizzando le informazioni in strutture su cui sia possibile applicare le altre due competenze discusse sopra.

Ovviamente, queste competenze non esauriscono lo spettro di competenze utili alla gestione dei dati, ma costituiscono certamente dei pilastri importanti. Abbiamo anche visto come un'astrazione come quella di grafo di conoscenza, può essere utile per supportare le tre capacità elencate. E in effetti, i grafi di conoscenza sono oggi utilizzati in sistemi informatici in produzione da molte aziende quali Google, eBay, Amazon, Facebook, per supportare uno spettro di applicazioni che usano, a diverso titolo, le competenze elencate [Noy&al.2019].

Per concludere questo capitolo, vorremmo almeno citare un tema diventato sempre più rilevante per la Data Semantics applicata alla Scienza dei dati: l'uso di tecniche di apprendimento automatico per costruire modelli interpretativi "data-driven", e cioè a partire da grandi collezioni di dati. I progressi

degli ultimi anni negli ambiti del Machine learning, di cui parleremo ampiamente nel Capitolo 10 e, in particolare, delle reti neurali profonde (deep learning) [LeCun&al.2015] (il lettore può valutare di leggere quel Capitolo prima di leggere queste conclusioni) hanno avuto - e avranno sempre di più - un impatto decisivo anche sulla semantica applicata alla Scienza dei dati.

Ci riferiamo qui soprattutto all'applicazione di reti neurali profonde all'elaborazione del linguaggio naturale [Li&Yang2018]. Algoritmi come word2vec [Mikolov&al.2013], ELMO [Peters&al.2018], o BERT [Devlin&al.2018], ad esempio, generano rappresentazioni di parole mediante reti neurali profonde, che vengono allenate cercando di compiere determinati task su grandi quantità di dati testuali (ad esempio, cercare di predire le parole intorno a una parola data in word2vec).

Alcuni di questi algoritmi si ispirano esplicitamente alla *semantica distribuzionale*, secondo cui parole con significato simile occorrono in contesti simili [Bruni&al.2014,Lenci2008]. Le rappresentazioni generate vengono chiamate *word embeddings* poiché codificano la semantica delle parole in vettori di dimensionalità fissata (ad esempio, vettori di 100 numeri), utilizzando un modello interpretativo geometrico basato su spazi vettoriali. Le rappresentazioni sono generate in maniera tale da ottimizzare il task con cui sono allenate, ma esibiscono alcune proprietà molto interessanti.

Generando rappresentazioni con word2vec possiamo valutare quanto sono simili due parole valutando la similarità del coseno tra i loro vettori [Mikolov&al.2013]; la valutazione della similarità può essere usata per derivare inferenze di tipo analogico (ad esempio, *Parigi sta alla Francia come Roma sta all'Italia*). Approcci simili, orientati cioè ad apprendere rappresentazioni a partire dai dati, sono stati applicati non solo a strutture linguistiche più complesse (frasi, paragrafi, documenti) [Le&Mikolov2014,Kiros&al.2015], ma anche a costrutti artificiali come i grafi di conoscenza e le ontologie.

In quel ramo di ricerca a innovazione tecnologica ormai noto come *knowledge graph embeddings*, si studiano modelli per apprendere rappresentazioni di entità, concetti ontologici, e relazioni a partire da diversi tipi di dati [Bordes&al.2013, Nickel&al.2016, Wang&al.2017, Dettmers&al.2018, Ristoski&al.2019, Bianchi&al.2019]. L'utilizzo di queste tecniche apre un'ampia gamma di nuove applicazioni per la Scienza dei dati, che possono variare dall'analisi testuale, ad esempio per scoprire distorsioni nei significati attribuiti alle parole in diverse collezioni testuali, alla predizione di relazioni, ad esempio per supportare sistemi di raccomandazione sofisticati.

Riferimenti

Achichi, M., Bellahsene, Z., Ellefi, M. B., & Todorov, K. (2019). Linking and disambiguating entities across heterogeneous RDF graphs. *Journal of Web Semantics*, *55*, 108-121.

Aggarwal, C. C. (2016). *Recommender systems* (pp. 1-28). Cham: Springer International Publishing.

Batini, C., Palmonari, M., & Viscusi, G. (2014). Opening the closed world: A survey of information quality research in the wild. In *The Philosophy of Information Quality* (pp. 43-73). Springer, Cham.

Berners-Lee T. (1998). Semantic web road map. Design Issues, Notes. Boston, USA, W3C. Retrieved from https://www.w3.org/DesignIssues/Semantic.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 28-37.

Bianchi, F., Palmonari, M., & Nozza, D. (2018, October). Towards encoding time in text-based entity embeddings. In *International Semantic Web Conference* (pp. 56-71). Springer, Cham.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems* (pp. 2787-2795).

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.

Cheatham, M., & Pesquita, C. (2017). Semantic Data Integration. In *Handbook of Big Data Technologies* (pp. 263-305). Springer, Cham.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.

Cruz, I.F., Palmonari, M., Caimi, F., & Stroe, C. (2013). Building linked ontologies with high precision using subclass mapping discovery. *Artificial Intelligence Review*, *40*(2), 127-145.

Cruz, I. F., Palmonari, M., Loprete, F., Stroe, C., & Taheri, A. (2016). Quality-based model for effective and robust multi-user pay-as-you-go ontology matching 1. *Semantic Web*, *7*(4), 463-479.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018, April). Convolutional 2d knowledge graph embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, *20*(1), 1-32.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

Di Noia, T., De Virgilio, R., Di Sciascio, E., & Donini, F. M. (2013). *Semantic Web: Tra ontologie e Open Data*. Apogeo Editore.

Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., Jiménez-Ruiz, E., & Pesquita, C. (2016, October). User validation in ontology alignment. In *International Semantic Web Conference* (pp. 200-217). Springer, Cham.

Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., ... & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, *51*(2), 32-49.

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, *48*.

Etzioni, O., Banko, M., & Cafarella, M. J. (2006, July). Machine Reading. In *AAAI* (Vol. 6, pp. 1517-1519).

Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. (2013, September). The agreementmakerlight ontology matching system. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 527-541). Springer, Berlin, Heidelberg.

Frixione, M. (1994). Logica, significato e intelligenza artificiale. Franco Angeli.

Fabien Gandon. A Survey of the First 20 Years of Research on Semantic Web and Linked Data. Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information, Lavoisier, 2018, ff10.3166/ISI.23.3-4.11-56ff. ffhal-01935898f

Hendler, J. (2001). Agents and the semantic web. IEEE Intelligent systems, 16(2), 30-37.

Hitzler, P., Krotzsch, M., & Rudolph, S. (2009). Foundations of semantic web technologies. Chapman and Hall/CRC.

Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., & Ngonga Ngomo, A. C. (2017). Survey on challenges of question answering in the semantic web. Semantic Web, 8(6), 895-920.

Isele, R., & Bizer, C. (2012). Learning expressive linkage rules using genetic programming. Proceedings of the VLDB Endowment, 5(11), 1638-1649.

Jiménez-Ruiz, E., & Grau, B. C. (2011, October). Logmap: Logic-based and scalable ontology matching. In International Semantic Web Conference (pp. 273-288). Springer, Berlin, Heidelberg.

Kejriwal, M., Szekely, P., & Knoblock, C. (2018). Investigative knowledge discovery for combating illicit activities. IEEE Intelligent Systems.

Keshvari-Fini, P., Janfada, B., & Minaei-Bidgoli, B. (2019, April). A Survey on Knowledge Extraction Techniques for Web Tables. In 2019 5th International Conference on Web Research (ICWR) (pp. 123-127). IEEE.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).

Kolomiyets, O., & Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24), 5412-5434.

Krötzsch, M. (2017, July). Ontologies for Knowledge Graphs?. In Description Logics.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: a survey. In Guide to Big Data Applications (pp. 83-104). Springer, Cham.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In Recommender systems handbook (pp. 73-105). Springer, Boston, MA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., ... & Krishnamurthy, J. (2018). Never-ending learning. Communications of the ACM, 61(5), 103-115.

Musto, C., Basile, P., Lops, P., de Gemmis, M., & Semeraro, G. (2017). Introducing linked open data in graph-based recommender systems. Information Processing & Management, 53(2), 405-435.

Ngomo, A. C. N., & Auer, S. (2011, June). LIMES—a time-efficient approach for large-scale link discovery on the web of data. In Twenty-Second International Joint Conference on Artificial Intelligence.

Nickel, M., Rosasco, L., & Poggio, T. (2016, March). Holographic embeddings of knowledge graphs. In Thirtieth Aaai conference on artificial intelligence.

Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2018). A survey on open information extraction. arXiv preprint arXiv:1806.05599.

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale Knowledge Graphs: Lessons and Challenges. Queue, 17(2), 20.

Ochieng, P., & Kyanda, S. (2018). Large-Scale Ontology Matching: State-of-the-Art Analysis. ACM Computing Surveys (CSUR), 51(4), 75.

Oramas, S., Ostuni, V. C., Noia, T. D., Serra, X., & Sciascio, E. D. (2017). Sound and music recommendation with knowledge graphs. ACM Transactions on Intelligent Systems and Technology (TIST), 8(2), 21.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8(3), 489-508.

Pesquita, C., Faria, D., Santos, E., & Couto, F. M. (2013, October). To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In Proc. 8th ISWC ontology matching workshop (OM), Sydney (AU), page this volume.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., & Paulheim, H. (2019). RDF2Vec: RDF graph embeddings and their applications. Semantic Web, 10(4), 721-752.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. IEEE intelligent systems, 21(3), 96-101.

Shen, W., Wang, J., & Han, J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering, 2(27), 443-460.

Ahmed Sherif, M., & Ngonga Ngomo, A. C. (2018). A systematic survey of point set distance measures for link discovery. Semantic Web, 9(5), 589-604.

Shvaiko, P., & Euzenat, J. (2011). Ontology matching: state of the art and future challenges. IEEE Transactions on knowledge and data engineering, 25(1), 158-176.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: a similarity-based model. Journal of Experimental Psychology: General, 133(2), 166.

Thibaut, J. P., French, R., & Vezneva, M. (2010). Cognitive load and semantic analogies: Searching semantic space. Psychonomic bulletin & review, 17(4), 569-574.

Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge acquisition: A developmental perspective. Similarity and analogical reasoning, 413-437.

Vu, B., Knoblock, C., & Pujara, J. (2019, May). Learning Semantic Models of Data Sources Using Probabilistic Graphical Models. In The World Wide Web Conference (pp. 1944-1953). ACM.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743.

Capitolo 8 – Trasformazione di modello e arricchimento semantico

C. Batini e A. Rula

1. Introduzione

In questo capitolo discutiamo del tema della trasformazione dei dati, focalizzandoci sulla trasformazione tra modello relazionale e modello RDF, e dell'arricchimento nel significato dei dati reso possibile dal fatto che i dati sono rappresentati in un modello semantico. Discuteremo di questi temi non in astratto, ma mostrando come la trasformazione di modello e l'arricchimento semantico costituiscano una valida attività preliminare alla integrazione dati, che abbiamo discusso nel Capitolo 6.

Attraverso uno studio di caso mostreremo come i due percorsi di Figura 1 diano luogo a risultati qualitativi diversi, e possa convenire portare avanti un percorso che includa prima la trasformazione e l'arricchimento semantico e poi la integrazione.

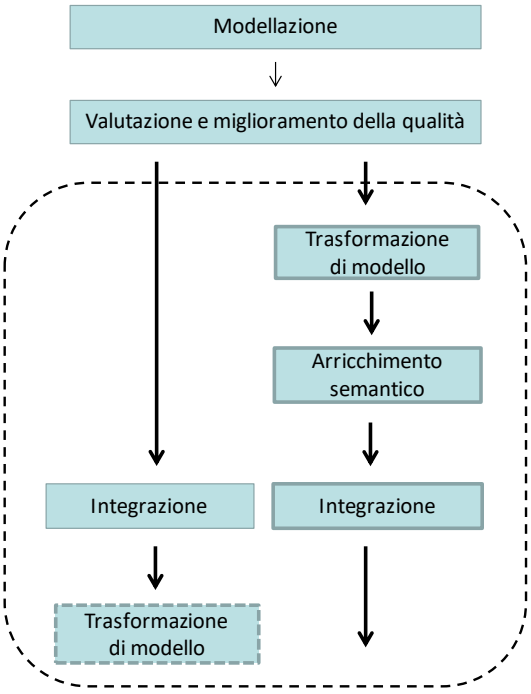


Figura 1 – I due cicli di vita che confrontiamo nel capitolo

La logica che presiede a questo nuovo percorso è che quanto più la attività di integrazione avviene in un contesto di conoscenza ricca sui dati da integrare, tanto più riusciremo a massimizzare i veri positivi e i veri negativi nel matching.

Lo studio di caso riguarda un insieme di posti di interesse turistico (vedi Figura 2), in cui siamo interessati a integrare due tabelle *Tourist Attraction* (chiamata nel seguito T1) e *Place* (chiamata nel seguito T2) che rappresentano tre luoghi di interesse, costituiti dagli osservabili relativi alla località di Torgiano, la Basilica di San Pietro, vista come Edificio religioso e alla Chiesa di San Pietro a Modica in Sicilia, vista come luogo di interesse in regione storica (Historic Region), vedi in figura le corrispondenze tra gli oggetti del mondo reale (gli osservabili) e i record delle due tabelle. Naturalmente, noi a priori nn conosciamo queste corrispondenze, e le dobbiamo ricostruire nel processo di integrazione.

T1 - Tourist attraction

Id	Denomination	Istat Code	Type
126759	Torgian	54053	Historic Place
125007	Saint Peter Abbey	246	Building

T2 - Place

Id	Label	Location	Type
1	Torgiano	Umbria	Populated Place
2	Saint Peter	Sicily	Historic Region
3	Basilica of Saint Peter	null	Religious Building

Figura 2 – L'esempio "Posti di interesse" considerato nel seguito

Il capitolo è organizzato come segue. Nella Sezione 2 approfondiamo il tema della trasformazione di dati da un modello "tradizionale" nel modello RDF. Nella Sezione 3 percorriamo il ramo di sinistra in Figura 1, applicando quanto visto nel Capitolo 6 sul record linkage e sulle funzioni di distanza. Nella Sezione 4 seguiamo il ramo a destra della Figura 1, e prima di affrontare il passo di integrazione effettuiamo una trasformazione da modello relazionale a modello RDF. Successivamente, sempre nella Sezione 4, si sfrutta l'arricchimento semantico arrivando a risultati di integrazione in termini di falsi positivi e falsi negativi (vedi Capitolo 6) decisamente migliori rispetto al record linkage "classico".

2. Il processo di trasformazione

Non esiste un modo unico per trasformare un dataset di partenza in un grafo *RDF* nel caso generale in cui il dataset contenga dati non strutturati (es. testo) o strutturati (es. tabella). Nel nostro esempio utilizziamo dati di partenza costituiti da una tabella. Esistono due modi per trasformare i dati strutturati in grafi RDF: mapping diretto e mapping personalizzato; in entrambi i casi la comunità del Semantic Web ha fornito degli standard approvati dalla World Wide Web Consortium[27] (W3C), l'ente che governa a livello mondiale i linguaggi, modelli e tecniche adottati nel Web.

Il primo metodo, mapping diretto, fornisce linee guida basate su un insieme di regole che riguardano:

[27] www.w3.org

- la specifica di generazione delle chiavi e degli attributi della tabella (la chiave, ricordiamo, è l'attributo o insieme di attributi i cui valori identificano univocamente tutti gli altri nei record; un esempio di chiave è Matricola nella tabella Studente della Figura 6 del Capitolo 3)
- la specifica degli identificatori in RDF per poter generare il grafo.

Il secondo metodo, mapping personalizzato, permette di personalizzare le trasformazioni (i mapping) in base al grafo finale che si vuole ottenere, tramite le due raccomandazioni del W3C denominate "Relational Database to RDF Mapping Language (R2RML)" e "Metadata Vocabulary for Tabular Data".

Nel caso del linguaggio R2RML, è l'utente a decidere quali tabelle e quali attributi della base di dati siano rappresentate nel grafo RDF e in base a quale schema. Dobbiamo tenere presente che le trasformazioni sono guidate dallo scopo, dall'uso che si farà dei dati e dal dominio e dalla qualità che sono associati allo scopo e all'uso. Nel nostro caso, possiamo assumere che lo scopo della trasformazione, oltre che portare a una più efficace integrazione, sia quello di preservare o migliorare la qualità dei dati; in particolare ci concentriamo sulla dimensione della *accuratezza*, poiché tanti problemi di qualità nella trasformazione e integrazione ricadono in questa dimensione di qualità.

Per il nostro esempio scegliamo un mapping minimale, che trasforma i dati della tabella tenuto conto dei soli legami di chiave, senza considerare altri aspetti più complessi dello schema relazionale, come ad esempio le chiavi esterne, che sono attributi di una tabella A che indentificano i record di una tabella B, da cui il nome di chiave esterna. Esempi di chiavi esterne sono nella Figura 6 del Capitolo 3 Matricola di Esame che è chiave esterna della tabella Studente e Codice Corso di Esame che è chiave esterna della tabella Corso. Per approfondimenti sulla trasformazione da modello relazionale a RDF si vedano [Lv 2008], [Thuy 2014] e [Zhou 2010].

Per poter decidere quale sia il più efficace tra i due percorsi di integrazione riprendiamo due misure già introdotte nel Capitolo 6. Ricordiamo che dati due record *r1* e *r2* che rappresentano due osservabili distinti nell'universo *U*, un processo di integrazione deve permetterci di decidere se:

- r1 e r2 fanno riferimento allo stesso osservabile, si parla in questo caso di match tra i due oggetti
- r1 e r2 non fanno riferimento allo stesso osservabile, non match.

Nel primo caso, se effettivamente r1 e r2 fanno riferimento allo stesso osservabile nella realtà parliamo di vero positivo, falso positivo altrimenti, nel secondo caso parliamo rispettivamente di vero negativo e falso negativo. La qualità della misura di distanza corrisponde a minimizzare i falsi positivi e i falsi negativi.

3. Integrazione con record linkage e funzioni di distanza

Mostriamo l'integrazione nell'esempio Punti di interesse in cui effettuiamo il record linkage applicando funzioni di distanza tra i record. Per le stringhe alfanumeriche utilizziamo la edit distance e per dati costituiti da liste di parole utilizziamo la Jaccard distance. Decidiamo di confrontare gli attributi Denomination della prima Tabella (T1) e Label della seconda (T2) e calcoliamo le seguenti distanze tra i record delle due tabelle:

- distanza tra la stringa del primo record R11 della T1 (“Torgian”) e la stringa del primo record R21 della T2 (“Torgiano”), vedi Figura 3. Utilizzando la edit distance si ottiene come valore 1 e quindi decidiamo per il match tra le due stringhe.

Tourist attraction

Id	Denomination	Istat Code	Type
126759	Torgian	54053	Historic Place
125007	Saint Peter Abbey	246	Building

Place

Id	Label	Location	Type
1	Torgiano	Torgiano	Populated Place
2	Saint Peter	Sicily	Historic Region
3	Basilica of Saint Peter	null	Religious Building

Primo record di T1 ↕ Primo record di T2

Figura 3 - Confronto di similarità tra "Torgiano" e "Torgian"

- distanza tra la stringa del secondo record R12 della T1 (“Saint Peter Abbey”) e secondo record R22 della T2 (“Saint Peter”). Si utilizza Jaccard e il risultato è uguale a 0,33, di conseguenza consideriamo questa corrispondenza come match, vedi Figura 4.

Tourist attraction

Id	Denomination	Istat Code	Type
126759	Torgian	54053	Historic Place
125007	Saint Peter Abbey	246	Building

Place

Id	Label	Location	Type
1	Torgiano	Umbria	Populated Place
2	Saint Peter	Sicily	Historic Region
3	Basilica of Saint Peter	null	Religious Building

Secondo record di T1 ↕ Secondo record di T2

Figura 4 - Confronto di similarità tra "Saint Peter Abbey" e "Saint Peter"

- distanza tra la stringa del secondo record R12 della T1 (“Saint Peter Abbey”) e terzo record R23 della T2 (“Basilica of Saint Peter”). Si applica la funzione di distanza di Jaccard tra i due insiemi di termini e il risultato è uguale a 0,60, di conseguenza consideriamo questa corrispondenza come non match, vedi Figura 5.

Tourist attraction

Id	Denomination	Istat Code	Type
126759	Torgian	54053	Historic Place
125007	Saint Peter Abbey	246	Building

Place

Id	Label	Location	Type
1	Torgiano	Umbria	Populated Place
2	Saint Peter	Sicily	Historic Region
3	Basilica of Saint Peter	null	Religious Building

Secondo record di T1 ↕ Terzo record di T2

Figura 5 - Confronto di similarità tra "Saint Peter Abbey" e "Basilica of Saint Peter"

Il risultato finale confrontato con le "vere" corrispondenze mostrate nella Figura 2 sono: un vero positivo (record R11 e R21), un falso positivo (record R12 e R22), un falso negativo (record R12 e record R23).

4. Integrazione preceduta da Trasformazione e Arricchimento semantico

Nel secondo percorso noi effettuiamo sulle due tabelle da integrare prima una trasformazione di modello e successivamente un arricchimento semantico, vedi Figura 6. Nei passi di trasformazione e arricchimento semantico i dataset vengono dunque trasformati in un modello più ricco semanticamente, il modello RDF, che, in virtù della sua struttura a grafo, può sfruttare risorse e tecniche semantiche disponibili nel Web. Ad esempio, possiamo collegare alcuni dei dati presenti nei due grafi a una tassonomia o ontologia, così che sia possibile confrontare i dati non solo mediante distanze misurate sui caratteri o sui termini, ma in modo più ricco attraverso il loro significato.

Mostriamo le attività nello studio di caso Punti di Interesse. Dapprima trasformiamo le tabelle dal modello relazionale nel modello RDF; nelle Figure 7 e 8 sono mostrati rispettivamente i grafi RDF relativi ai primi record delle due tabelle (R11 e R21) e poi quelli relativi al secondo record della tabella Tourist attraction (R21) e al secondo e terzo record della tabella Place (R22 e R23).

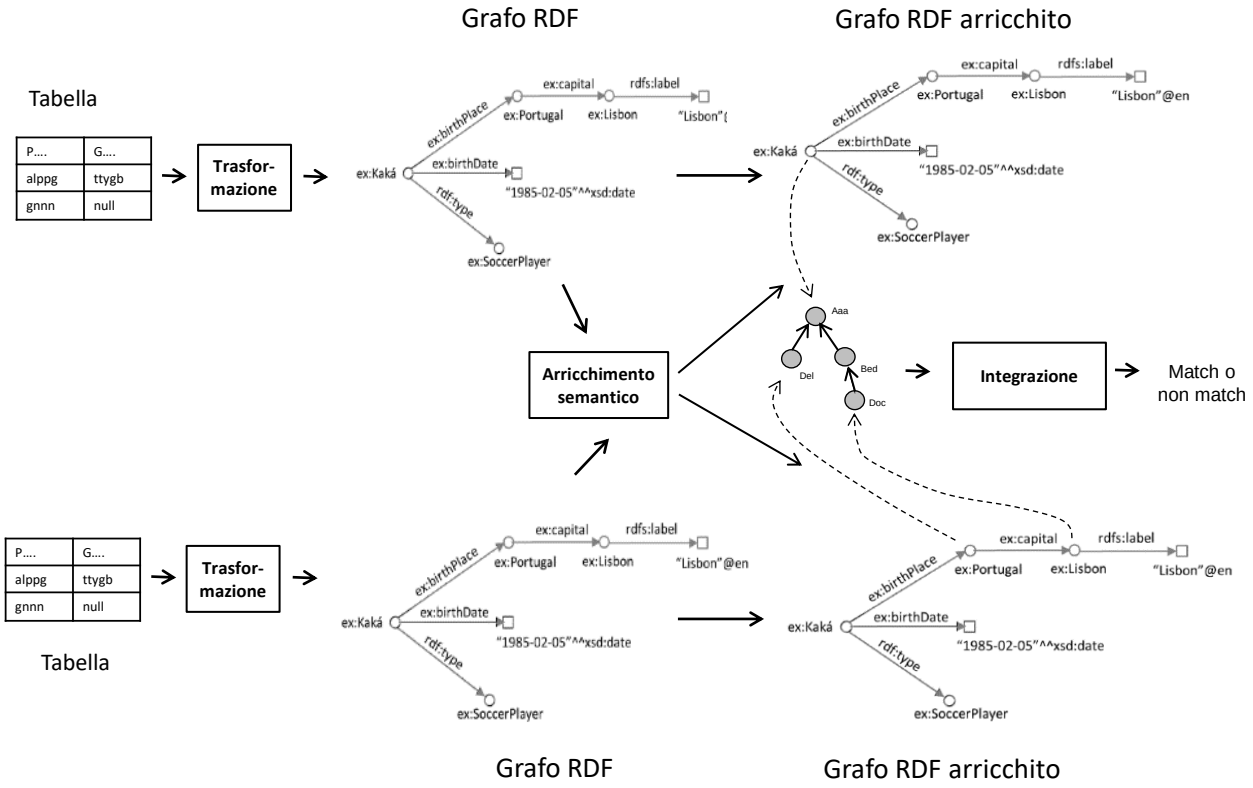
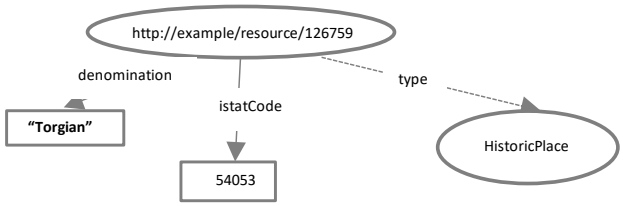


Figura 6 – La trasformazione e integrazione insieme

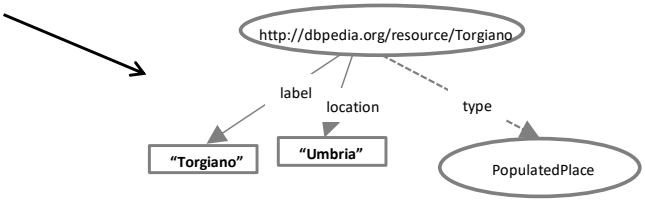
Tourist attraction

Id	Denomination	Istat Code	Type
126759	Torgian	54053	Historic Place
125007	Saint Peter Abbey	246	Building



Place

Id	Label	Location	Type
1	Torgiano	Umbria	Populated Place
2	Saint Peter	Sicily	Historic Region
3	Basilica of Saint Peter	null	Religious Building



7

Figura 7 - Trasformazione in RDF del record R11 e R21

Nel caso dei record R11 e R21, possiamo confrontare i nomi delle proprietà, cercando di capire se esistano casi di sinonimia, cioè di nomi diversi di proprietà che corrispondano allo stesso oggetto osservabile del mondo reale. Si noti che in questo caso non confrontiamo i valori, bensì le proprietà.

Con riferimento ai nomi delle proprietà, vediamo che entrambi i grafi hanno una proprietà con nome Id e una con nome Type, il secondo ha una proprietà Istat Code cui non corrisponde nessuna proprietà nel primo grafo, mentre, infine, il primo grafo ha una proprietà Denomination e la seconda una proprietà Label che, effettivamente, sembrano avere lo stesso significato con nomi diversi; siamo di fronte a un caso di sinonimia.

Possiamo risolvere la sinonimia cambiando il nome Denomination in Label nel grafo RDF (non mostriamo il nuovo grafo). A questo punto, se confrontiamo i valori associati nelle triple, cioè "Torgiano" e "Torgian", arriviamo alla conclusione, in virtù del valore della distanza pari a uno tra le stringhe di caratteri, che i due record rappresentano lo stesso oggetto osservabile. Eravamo arrivati alla stessa conclusione nel primo percorso.

Per quanto riguarda il record R12 e la coppia R22 e R23, cerchiamo nel Web una risorsa che arricchisca la nostra conoscenza sui dati rappresentati. Possiamo in questo caso sfruttare l'ontologia di DBpedia, disponibile nel Linked Open Data Cloud (https://wiki.dbpedia.org/).

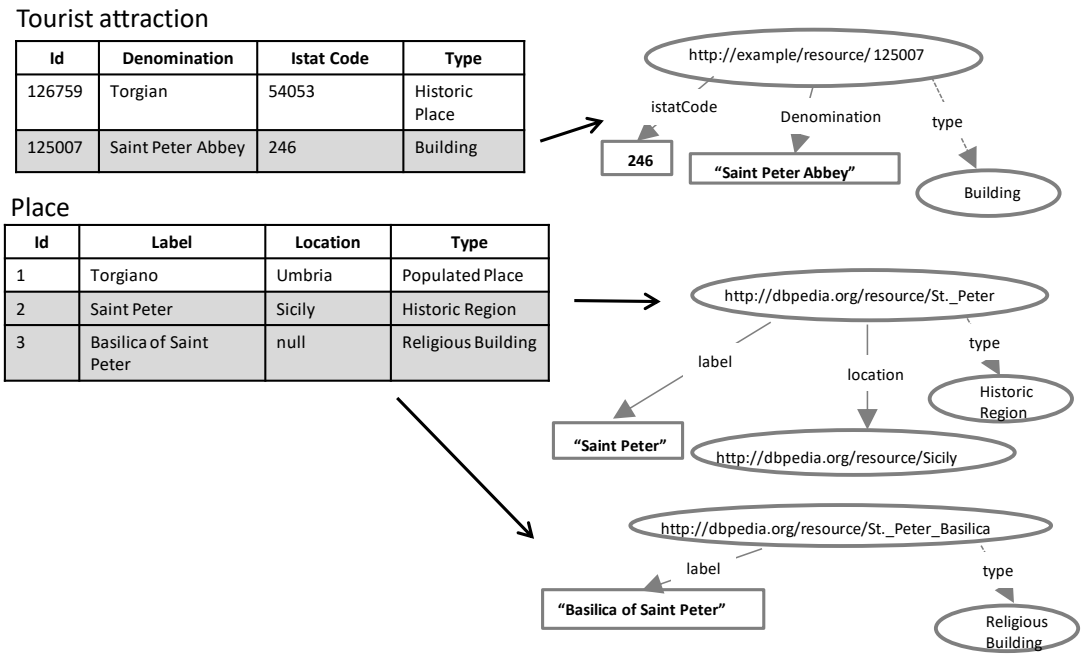


Figura 8 - Trasformazione in RDF dei record R12, R22 e R23

Navigando manualmente o con l'aiuto di una tecnica di ricerca automatica disponibile in DBpedia, possiamo anzitutto tramite arricchimento semantico individuare l'albero dei concetti di DBpedia che sono di interesse nel nostro esempio (vedi Figura 9) creando successivamente le corrispondenze tra i tipi (types) rappresentati nei grafi RDF e i concetti di DBpedia (sempre Figura 9).

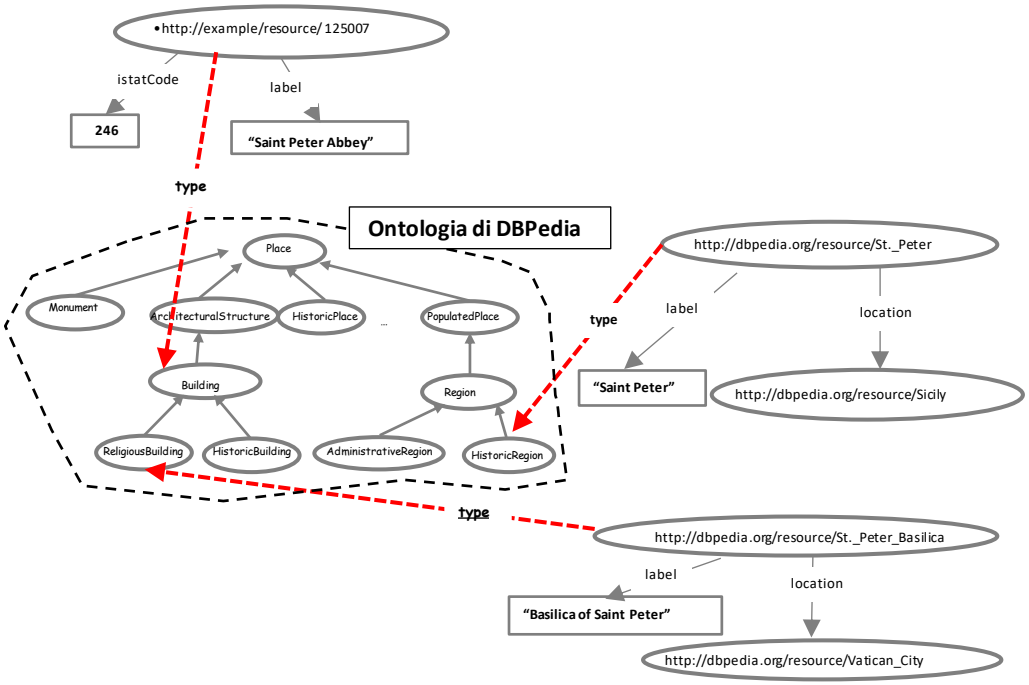


Figura 9 - Arricchimento semantico e corrispondenze tra i record R12, R21 e R22 con DBpedia

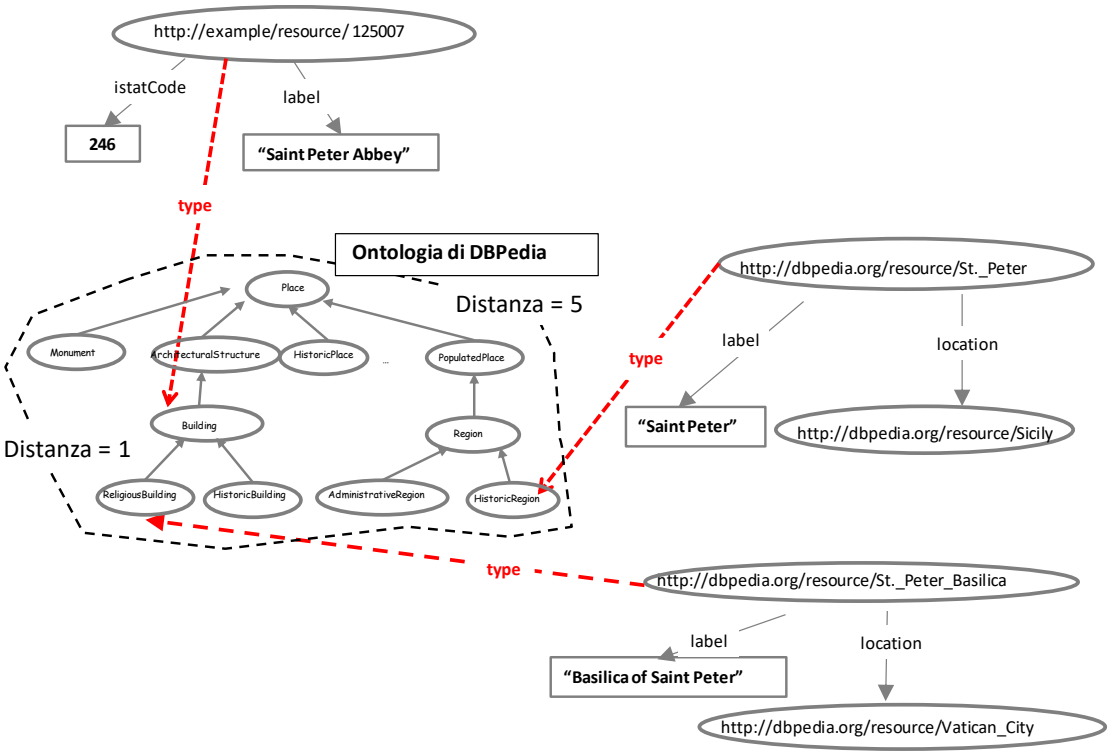


Figura 10 – Distanze nella ontologia di DBPedia

A questo punto (vedi Figura 10) possiamo confrontare i grafi RDF (confronto basato questa volta non sui caratteri o parole, come fatto in precedenza, ma basato sul significato riferito alla ontologia di DBpedia) misurando le distanze nell'albero di DBpedia tra type del record R12 e types dei record R22 e R23, arrivando a decidere che, in virtù della minore distanza, a "Saint Peter Abbey" in realtà corrisponde a "Basilica of Saint Peter" (Distanza nell'albero pari a 1), e non "Saint Peter" (Distanza pari a 5).

Seguendo il percorso a. trasformazione, b. arricchimento semantico, e c. integrazione, abbiamo due veri positivi, un vero negativo e zero falsi positivi e negativi (ti invitiamo a verificare per esercizio che concordi con questo bilancio).

Ora supponiamo di produrre un'altra trasformazione e un arricchimento semantico in cui siamo meno precisi; introduciamo in questo modo il problema dell'accuratezza semantica conosciuto come *annotazione o classificazione imprecisa*.

Supponiamo di classificare tutte le istanze di Type come "Place" invece di classificarle come abbiamo fatto in precedenza, cioè con tipi più specifici come "HistoricalPlace" oppure "Building". In questo modo, le trasformazioni risultanti sono quelle di Figura 11 e Figura 12.

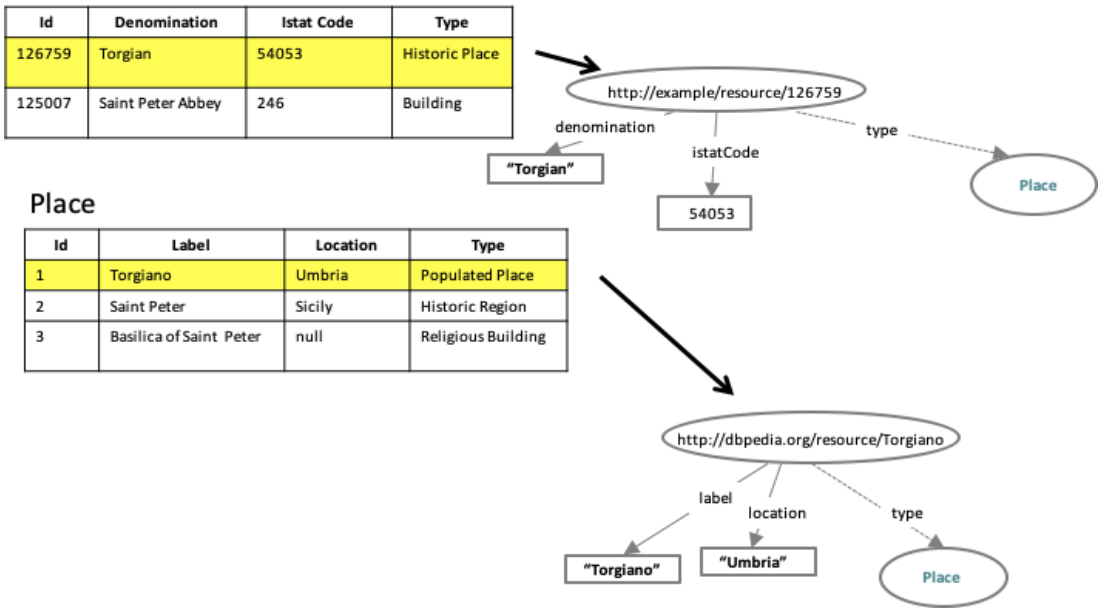


Figura 11 - Trasformazione in RDF del record R11 e R21 con la modifica sul tipo

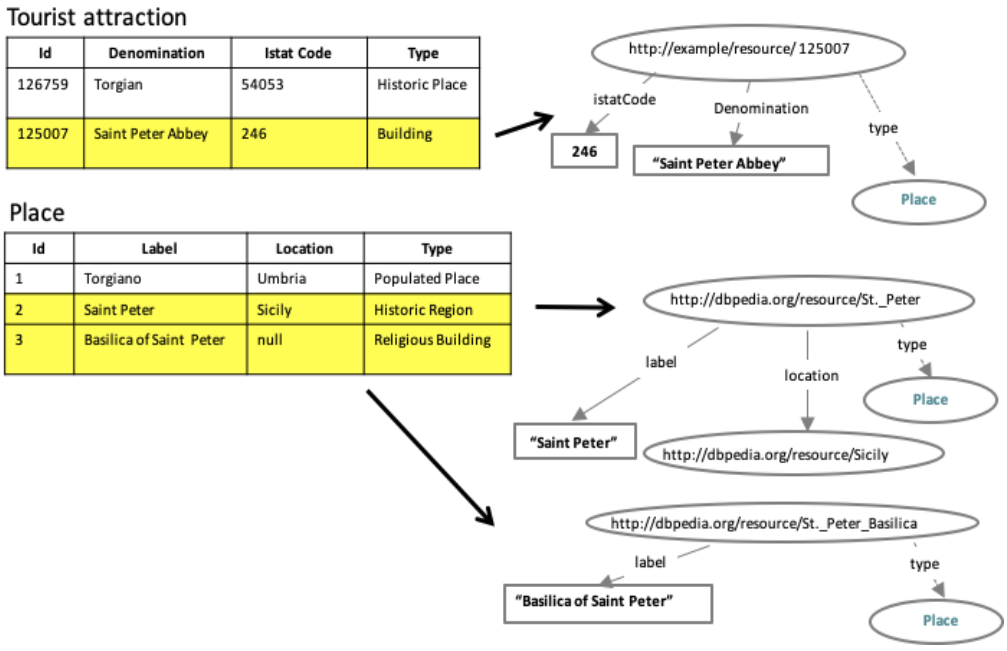


Figura 12 - Trasformazione in RDF del record R12, R22 e R23 con la modifica sul tipo

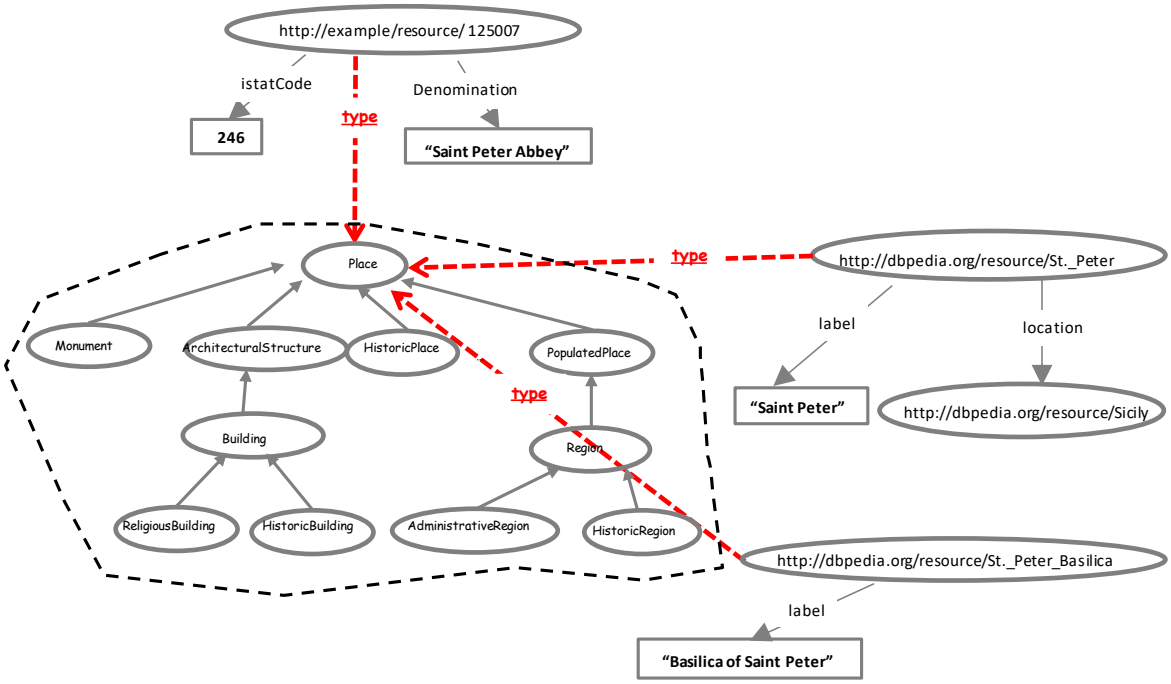


Figura 13 – Nuovo arricchimento semantico e nuove corrispondenze
tra i record R12, R21 e R22 con DBpedia

Se applichiamo il matching a quest'ultima trasformazione (vedi Figura 13) si ottengono distanze tutte pari a 0, producendo in questo modo un risultato qualitativo peggiore di prima, perché introduciamo un falso positivo. Ciò è dovuto al fatto che non viene sfruttata la classificazione dell'ontologia, e nella attribuzione del significato si rimane a un livello troppo astratto che non ci aiuta a sfruttare la classificazione ontologica. Questo esempio ha mostrato che nell'eseguire una trasformazione ci possono essere problemi di qualità, e si può incorrere in un "degrado" del significato dei concetti rappresentati. Quindi la qualità complessiva, *nel nostro caso della integrazione*, può migliorare se si sfrutta l'arricchimento semantico, ma allo stesso tempo, se non si è attenti nelle trasformazioni, i dati possono fare come il gambero, possono regredire peggiorando la precisione con cui rappresentano la realtà.

In conclusione, l'esempio iniziale ha dimostrato la superiorità del processo che effettua prima una trasformazione di modello da relazionale a RDF e poi un arricchimento semantico (creazione delle corrispondenze con DBpedia) e un confronto successivo.

Questi esempi dimostrano l'importanza della integrazione nel ciclo di vita dei dati, e la importanza insita nel saper valutare se una trasformazione di modello sia opportuna per essere in grado di sfruttare le risorse (nel nostro caso la risorsa DBpedia) e le tecniche disponibili del Web (il matching sulla base della distanza nell'albero). Abbiamo anche visto che, quando parliamo di integrazione o confronto semantico, possiamo fare riferimento ai concetti (ricordo il caso di Denomination verso Label, che sono stati riconosciuti come sinonimi) ovvero ai valori. Nelle metodologie di integrazione di basi di dati i due processi vengono suddivisi in *Integrazione di schemi* e *integrazione dei dati*.

Riferimenti

Y. Lv, e Z. M. Ma - Transformation of relational model to RDF model - IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2008.

P. Thuy et al. - RDB2RDF: completed transformation from relational database into RDF ontology. Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication. ACM, 2014.

S. Zhou - Exposing relational database as RDF - 2nd International Conference on Industrial and Information Systems. Vol. 2. IEEE, 2010.

Capitolo 9 – Io Statistica, le mie memorie

Anna Ferrari

Considerate la vostra semenza:
Fatti non foste a viver come bruti
Ma per seguir vitute e canoscenza
Dante

Mi trovavo in un posto dove niente era intorno a me e niente ero io; un posto senza spazio, senza tempo, senza un come e senza un perché. Niente di quello che si trovava qui era ancora. Altri come me aspettavano. Da quanto tempo eravamo in questo luogo? Quanto tempo dovevamo aspettare ancora? Altri, come noi, non erano mai stati qui, altri invece erano arrivati e andati via. Così aspettavamo quel momento in cui anche noi saremmo andati, se mai sarebbe successo...

"Respirare non è mai stata qua e Mangiare è andata subito via" ci ricordava spesso Storia. "Senza lasciare neanche un biglietto" aggiungeva Scrittura un po' triste e un po' compiaciuta. Anche loro aspettavano insieme a me, a Insiemistica, a Matematica, a Informatica, e tanti altri quel momento. Chissà chi di noi sarebbe andato via per primo, chissà se qualcuno sarebbe rimasto qui per sempre. Ma soprattutto, come faceva Storia a sapere tutte quelle cose? D'altronde anch'essa ancora non era. Oltretutto, quando si cercava di affrontare il discorso e farla ragionare, andava su tutte le furie "io so benissimo che anno è e cosa è successo finora! Siamo nel 6000 a.C. e ne manca ancora un bel po' sia per me che per voi". E così chiudeva il discorso e nessuno aveva il coraggio di contraddirla. Ogni tanto, per farsi grande, ci diceva in che anno fossimo. A Chiaroveggenza piaceva parlare di cose strane che spesso nessuno capiva. Sapeva davvero molto più di noi, e vorrei ben vedere! Sovente confermava le teorie di Storia che, infatti, aveva sempre ragione. Per noi comunque, conoscere l'anno voleva dire tutto e niente e sicuramente non ci dava indicazione su quando sarebbe arrivato il momento. E così aspettavamo speculando su chi tra noi sarebbe andato via per prima. Per Informatica e Matematica era chiara la precedenza, ma Matematica e Fisica litigavano sempre. Filosofia neanche entrava nel merito: a lei non importava altro che la lasciassimo in pace nei suoi ragionamenti personali.

E così, nel 5000 a.C., senza che nessuno se lo aspettasse, fu Insiemistica ad andare via per prima, lasciandoci questo biglietto, vedi Figura 1.

I 2.000 anni successivi furono molto turbolenti: moltissimi di noi se ne andarono, altri cambiarono forma e nuovi arrivarono. Subito dopo Insiemistica e Numeri andarono nel Regno di Babilonia e Scrittura tra i Sumeri, vedi Figura 2.

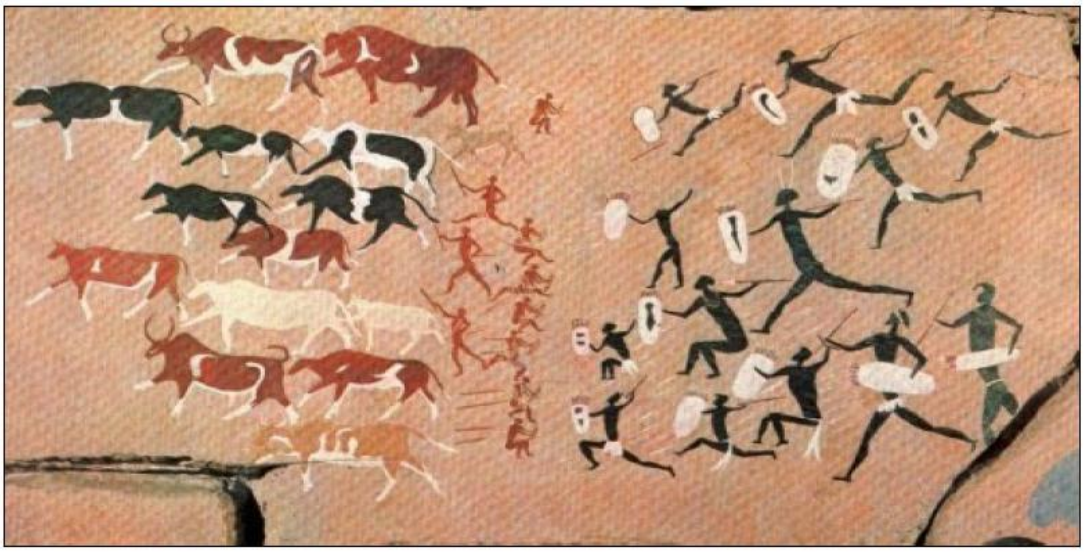


Figura 1 - L'idea di insiemistica - Si noti come fin dal 5000 a.C. l'uomo concepisce l'idea di Insiemistica: dipinti rupestri rappresentano una suddivisione di animali e persone in gruppi omogenei. La consapevolezza di gruppo e di insieme nasce prima della scrittura. Questa capacità fu cruciale per il miglioramento delle condizioni di vita dell'uomo rispetto alla caccia, all'agricoltura e alla difesa.

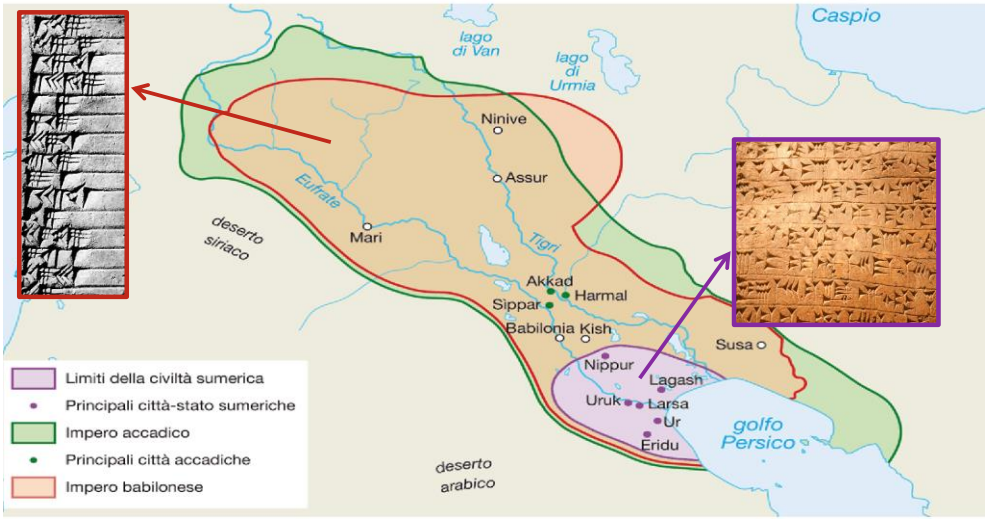


Figura 2 – L'invenzione dei Numeri e della Scrittura

Storia rimaneva lì, ma, mano a mano che passava il tempo, sbiadiva sempre di più. Somma, Sottrazione, Moltiplicazione e Divisione erano andate via con Base Sessaggesimale per poi tornare e andarsene di nuovo con Base Decimale. E così, un giorno, mentre facevo congetture su quando sarebbe arrivato il mio momento, sentii che una parte di me se ne stava andando via: la Statistica Descrittiva era stata concepita dall'essere umano.

1. Io Statistica, le mie memorie

Le mie prime memorie risalgono al 3000 a.C. Mi trovavo in moltissimi posti del mondo contemporaneamente: in Egitto, in Cina e in India. Ero la consigliera numero uno degli imperatori più importanti della terra e il mio compito

era quello di tenere traccia delle caratteristiche più salienti della popolazione. Ai tempi ero molto primitiva, e mi bastava essere enumerazione di persone, schiavi e bestiame. Seppur semplice, ero di vitale importanza per la spartizione delle risorse, per la riscossione delle tasse per il conteggio degli schiavi e dei soldati al servizio di faraoni e re.

E fu così che per molte centinaia di anni servii allo scopo di tenere traccia delle caratteristiche più importanti delle popolazioni. Grazie ai molti documenti che l'uomo decise di produrre e conservare oggi abbiamo una preziosissima raccolta di informazioni sulle popolazioni passate. Ad esempio, quale era la città più grande nel 50 a.C.? E Qual era la città più grande nel 1000 d.C.? Vedi Figura 3.

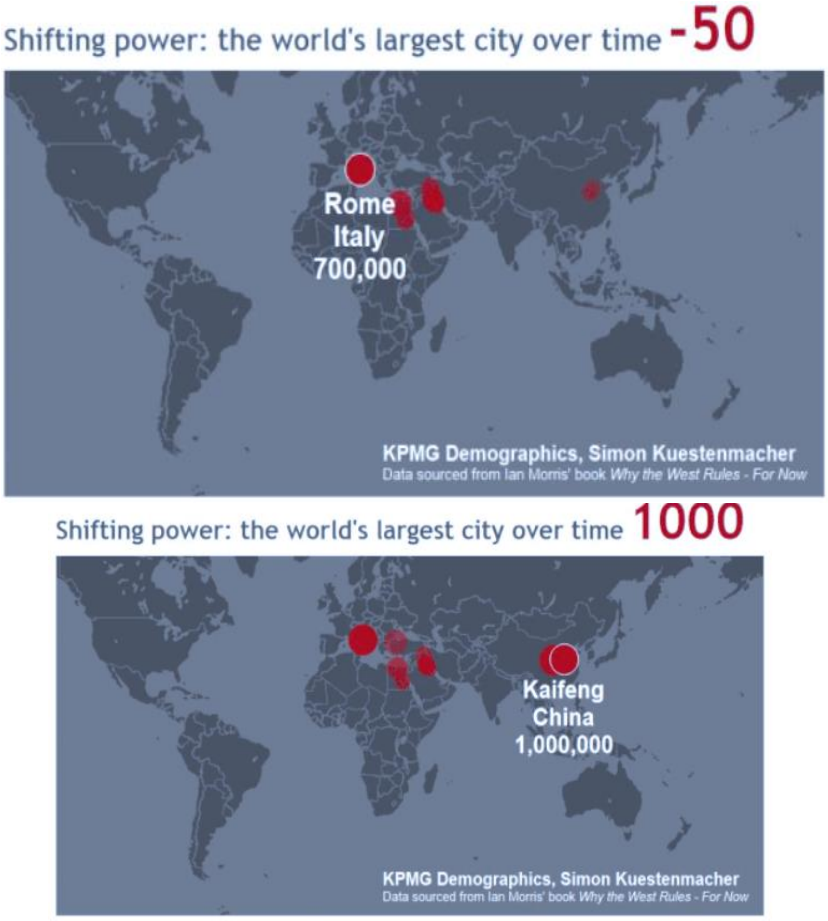


Figura 3 – Le città più grandi nel 50 a.C e nel 1000 d.C.

E quale *sarà* la città più grande nel 2050 d.C.? Un attimo! Non sono ancora così evoluta da poter fare previsioni. Per secoli il mio compito sarà esclusivamente la descrizione degli aspetti peculiari della popolazione, motivo per il quale vengo chiamata Statistica Descrittiva. Si usa attribuirmi la metafora della fotografia scattata; questa fotografia è statica, non dinamica, e rappresenta le caratteristiche principali in un preciso istante di una popolazione. Non posso sapere quello che è successo prima se non attraverso precedenti fotografie e non posso per nessun motivo dire cosa succederà in futuro. Questo non vuol dire che io sia antiquata: ancora oggi moltissime persone mi utilizzano e se non ci credete, guardate qua in Figura 4. In questa infografica realizzata dall'Istituto Nazionale di Statistica (ISTAT) sintetizzo le principali caratteristiche demografiche della popolazione residente in Italia nel 2017. Grazie a rilevazioni (fotografie) precedentemente realizzate è possibile fare dei confronti tra popolazione di ieri e popolazione di oggi. Ad esempio, nel 2017, rispetto al 2007, la popolazione

italiana è diminuita di 89.000 unità; il numero di anziani è aumentato da 11.7 a 12.5 milioni come gli over 90, e gli ultra centenari sono aumentati.

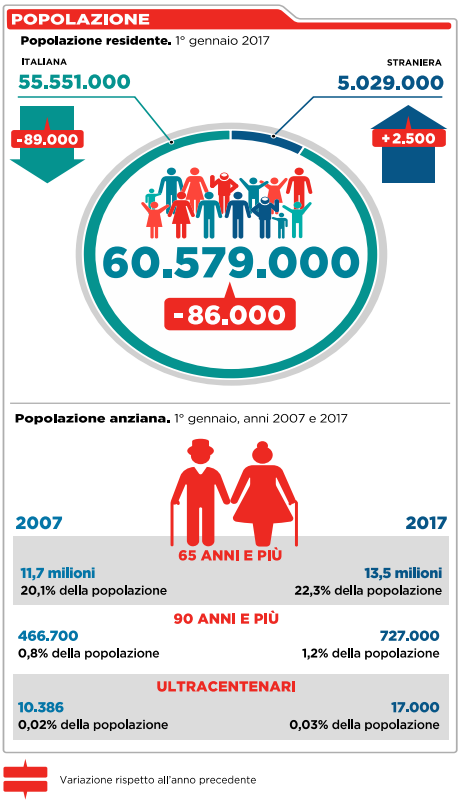


Figura 4 - Principali caratteristiche demografiche della popolazione residente in Italia nel 2017 e variazione rispetto al 2007 (Fonte: ISTAT)

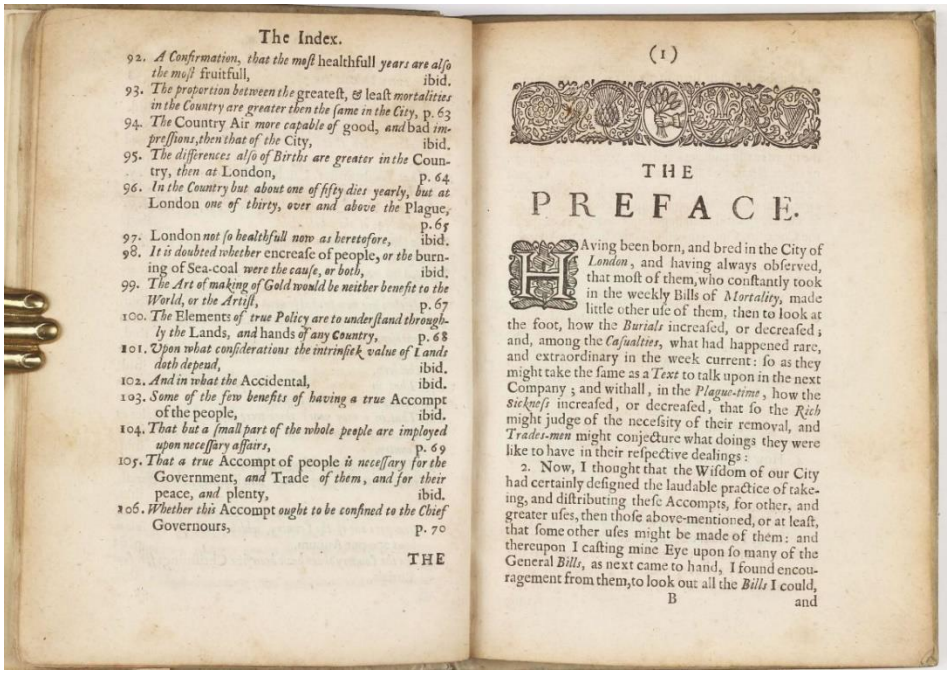
Questa per l'appunto è solo una descrizione e nessuna considerazione viene fatta sulle cause e sugli effetti di questa particolare configurazione della popolazione sulla popolazione di domani. Forse, però, istintivamente, qualcosa possiamo già concludere. Nessuno si è chiesto, ad esempio, se la popolazione italiana di anziani aumenterà nei prossimi anni?

Il Sig. Hermann Conring (1606-1681) iniziò a portarmi con sé al lavoro e a parlare di me davanti a moltissimi studenti. Sentivo dire che ero finalmente entrata in Università. Sembrava davvero un fatto prestigioso! Conring insieme al Sig. Gottfried Achenwall mi definirono come la "conoscenza dello Stato e della sua costituzione presente" e finalmente avevo un nome, anzi due: *Notitia Rerum Publicarum* o *Staatskunde*. Ero menzionata per l'analisi comparativa tra stati per cui ero espressa soprattutto attraverso concetti e spiegazioni verbali. Io, materia nobile, non potevo essere volgarizzata con l'impiego dei numeri. Se devo essere sincera, non mi sentivo proprio a mio agio espressa in questo modo: sentivo che molte delle mie potenzialità venivano perdute.

L'avvento e la diffusione sempre più ampia del metodo scientifico sperimentale, attribuito a Francis Bacon (1561 - 1626) e a Galileo Galilei (1564 – 1642), fu la goccia che fece traboccare il vaso. Essi sostenevano che la teoria che descrive un fenomeno è un insieme di ipotesi che devono essere verificate; la verifica avviene attraverso esperimenti che si servono di osservazioni empiriche e dai dati, che, a loro volta, permettono di confutare o

corroborare le ipotesi di partenza. La scuola aristotelica, assiomatica e deduttiva, stava facendo largo a un metodo prettamente induttivo; insomma, senza utilizzare i numeri non sarei andata da nessuna parte!

Così, quando re Carlo II istituì la Royal Society of London nel 1620, decisi anche io di prenderne parte insieme a molti studiosi che abbracciavano la Filosofia Sperimentale di Bacon. Lo scopo di questa associazione era di utilizzarmi per promuovere il benessere della società. Devo dire che presero questo scopo molto sul serio soprattutto John Graunt (1620 – 1674) che, ne suo libro *"Natural and Political observations mentioned in a following index, and made upon the bills of mortality",* analizzò moltissimi fenomeni basandosi su dati che aveva raccolto dalla lettura settimanale dei bollettini di mortalità per ricavarne delle regolarità scientifiche. Ad esempio, era in grado di prevedere lo svilupparsi di epidemie in città e proponeva contromisure come rifugiarsi nella campagna. Riporto in Figura 5 i titoli di alcuni capitoli del suo libro.



The Index.

92. *A Confirmation, that the moſt* healthfull *years are alſo the moſt* fruitfull, ibid.
93. *The proportion between the* greateſt, *&* leaſt *mortalities in the Country are greater then the ſame in the City,* p. 63
94. *The* Country Air *more capable of* good, *and* bad *impreſſions, then that of the* City, ibid.
95. *The differences alſo of Births are greater in the* Country, *then at* London, p. 64
96. *In the Country but about one of fifty dies yearly, but at* London *one of thirty, over and above the* Plague, p. 65
97. London *not ſo healthfull now as heretofore,* ibid.
98. *It is doubted whether* encreaſe of people, *or the* burning of Sea-coal *were the cauſe, or both,* ibid.
99. *The Art of making of Gold would be neither benefit to the World, or the Artiſt,* p. 67
100. *The* Elements *of true Policy are to underſtand throughly the* Lands, *and* hands *of any Country,* p. 68
101. *Upon what conſiderations the intrinſick value of Lands doth depend,* ibid.
102. *And in what the* Accidental, ibid.
103. *Some of the few benefits of having a true* Accompt of the people, ibid.
104. *That but a ſmall part of the whole people are imployed upon neceſſary affairs,* p. 69
105. *That a true* Accompt of people *is neceſſary for the* Government, *and* Trade *of them, and for their* peace, *and* plenty, ibid.
106. *Whether this* Accompt *ought to be confined to the Chief* Governours, p. 70

THE

(1)

THE

PREFACE.

HAving been born, and bred in the City of *London*, and having always obſerved, that moſt of them, who conſtantly took in the weekly Bills of *Mortality*, made little other uſe of them, then to look at the foot, how the *Burials* increaſed, or decreaſed; and, among the *Caſualties*, what had happened rare, and extraordinary in the week current: ſo as they might take the ſame as a *Text* to talk upon in the next Company; and withall, in the *Plague-time*, how the *Sickneſs* increaſed, or decreaſed, that ſo the *Rich* might judge of the neceſsity of their removal, and *Trades-men* might conjecture what doings they were like to have in their reſpective dealings:

2. Now, I thought that the Wiſdom of our City had certainly deſigned the laudable practice of taking, and diſtributing theſe Accompts, for other, and greater uſes, then thoſe above-mentioned, or at leaſt, that ſome other uſes might be made of them: and thereupon I caſting mine Eye upon ſo many of the General *Bills*, as next came to hand, I found encouragement from them, to look out all the *Bills* I could,

B and

Figura 5 – Prefazione dal libro di Graunt (Fonte: Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality, with reference to the Government, Religion, Trade, Growth, Air, Diseases, and the several Changes of the said City, the Second Edition).

Grazie agli studi di Graunt, da questo momento in poi, non sarò più solo concepita unicamente come enumerazione, ma servirò ad estrapolare informazioni più elaborate dai dati seguendo dei veri e propri metodi di analisi. Insieme Insieme a Sir Petty William, Graunt creò l'Aritmetica Politica, definita da C. Davenant come "l'arte di ragionare per mezzo di cifre sulle cose aventi attinenza con il governo". Influenzati dalla diffusione del metodo scientifico sperimentale, traevano dai dati l'esperienza che, successivamente, mutavano in metodi matematici e deduttivi. Le tavole di mortalità e di sopravvivenza diventarono gli strumenti più utilizzati dai cultori di questa corrente di pensiero, tra cui ricordiamo anche Edmund Halley, forse maggiormente conosciuto per la sua cometa.
In Figura 6 sono riportate diverse tavole e studi pubblicati nel libro di Graunt, presi dal libro "Natural and Political observations mentioned in a following index, and made upon the bills of mortality". Si tratta di un report dal valore inestimabile per la sua varietà e la meticolosità di raccolta e discussione dei dati.

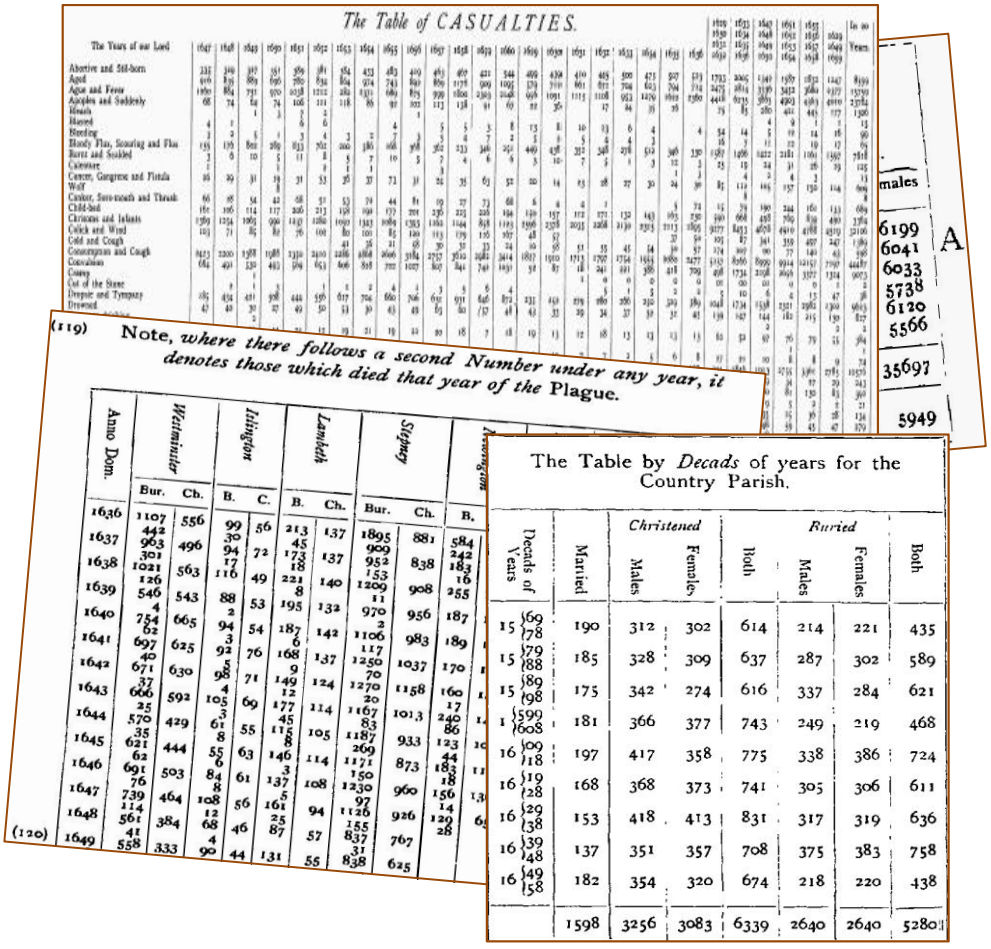


Figura 6 - Tavole e studi di Graunt

Chi sa leggere le tavole di mortalità pubblicate oggi dall'ISTAT? Impariamo a usarle, osservando la Figura 7. Per ogni intervallo di età sono riportati diversi valori: sopravvivenza, decessi, probabilità di morte, probabilità prospettiva di sopravvivenza e speranza di vita.

Ancora non mi capacitavo del fatto di essere rimasta in parte in questo posto, senza tempo, senza spazio, senza come e senza perché. Eppure Statistica Descrittiva era andata via da un bel pezzo. Mi chiedevo cosa mancasse affinché anche le altre parti di me potessero andare.

"E' arrivato l'illuminismo, non ti devi abbattere: questa sarà un'epoca di grandi invenzioni. Vedrai che ci sarà un posto anche per te" mi diceva Storia per tirarmi un po' su. Meccanica era andata via concepita dalla mente del grande Newton anch'egli parte della Royal Society di Londra, dove si stavano sviluppando teorie che avrebbero cambiato il mondo. Avevo visto andare via da poco Calcolo delle Probabilità, e non so perché, ma questo mi sollevava. Galileo Galilei ne usava le proprietà per calcolare delle probabilità al gioco dei dadi. Come lui anche Pacioli, Tartaglia, Pascal e Fermat si cimentavano con questi rompicapo.

Territorio	Italia				
Sesso	totale				
Seleziona periodo	2017				
Funzioni biometriche	sopravviventi - lx	decessi - dx	probabilità di morte (per 1.000) - qx	anni vissuti - Lx	probabilità prospettiva di sopravvivenza - Px
Età e classi di età					
fino a 4 anni	100000	345	3,44568	498435	0,999494
5-9 anni	99655	37	0,37575	498183	0,9996073
10-14 anni	99618	45	0,45552	497987	0,9992899
15-19 anni	99573	102	1,02938	497633	0,9987172
20-24 anni	99470	146	1,47119	496995	0,9984505
25-29 anni	99324	159	1,59971	496225	0,9982339
30-34 anni	99165	200	2,02166	495348	0,9975736
35-39 anni	98964	292	2,9461	494147	0,9962688
40-44 anni	98673	454	4,60459	492303	0,9940512
45-49 anni	98218	741	7,54762	489374	0,9904416
50-54 anni	97477	1155	11,84611	484697	0,9848401
55-59 anni	96322	1845	19,15281	477349	0,9753975
60-64 anni	94478	2915	30,85392	465605	0,9604903
65-69 anni	91563	4546	49,65126	447209	0,936938
70-74 anni	87016	7002	80,46646	419007	0,8936747
75-79 anni	80015	10857	135,68868	374456	0,8188747
80-84 anni	69157	16806	243,0107	306632	0,6762961
85-89 anni	52351	22180	423,67954	207374	0,4761596
90-94 anni	30171	19503	646,40642	98743	0,2839836
95-99 anni	10668	8544	800,82764	28041	0,1488559
100-104 anni	2125	1975	929,32624	4174	0,0494139
105-109 anni	150	148	985,10054	206	0,0107495
110-114 anni	2	2	998,05061	2	0,0015086
115-119 anni	0	0	999,838	0	0,0001356

Dati estratti il 19 lug 2019, 06h57 UTC (GMT) da I.Stat

Figura 7 - Tavole di mortalità pubblicate dall'Istituto Nazionale di Statistica (Fonte: ISTAT).

"La probabilità servirà all'uomo per gestire le situazioni di incertezza della vita quotidiana" disse un giorno Chiaroveggenza. Ad esempio, quando si esce di casa per andare a fare una passeggiata si guarda il cielo e si stima attraverso alcuni parametri come ad esempio la nuvolosità e il vento se sia meglio portare l'ombrello oppure no. Sostanzialmente ci si pone la domanda: "con quale probabilità pioverà?".

Chiaroveggenza rimase un po' in silenzio e poi esclamò "Incredibile!!" Si volse verso di noi e ci chiese "Sareste in grado di dirmi quante persone dovrebbero esserci in una stanza affinché la probabilità che due di loro festeggino il compleanno lo stesso giorno sia maggiore del 97%?". "Chiaroveggenza, Calcolo delle probabilità è andato via" risposi "ma se ce ne fossero 366 la probabilità di trovare due persone nate nello stesso giorno sarebbe 1" mi stupii di me stessa. Chiaroveggenza rispose: "Si tratta del *Paradosso del Compleanno*: ne bastano solo 50! Con 50 persone sono quasi sicura di trovarne due con lo stesso compleanno! Incredibile nevvero?"

Rimanemmo tutti a bocca aperta, 50 è davvero poco! "Non credete sia davvero così? Andate su Facebook nella sezione compleanni e guardate quanto è facile trovare due persone nate nello stesso giorno tra i vostri amici". Nessuno lì per lì capì cosa intendesse dire, vedi Figura 8. Io ho 291 amici su Facebook e il 20, 21 e il 28 Luglio e

anche il 3 e il 6 Agosto compiono gli anni due miei amici. Il 30 Agosto sono addirittura in tre a festeggiarlo. Non è raro trovare due persone nate nello stesso giorno, e sicuramente ne bastano molte meno di 366. Provare per credere!

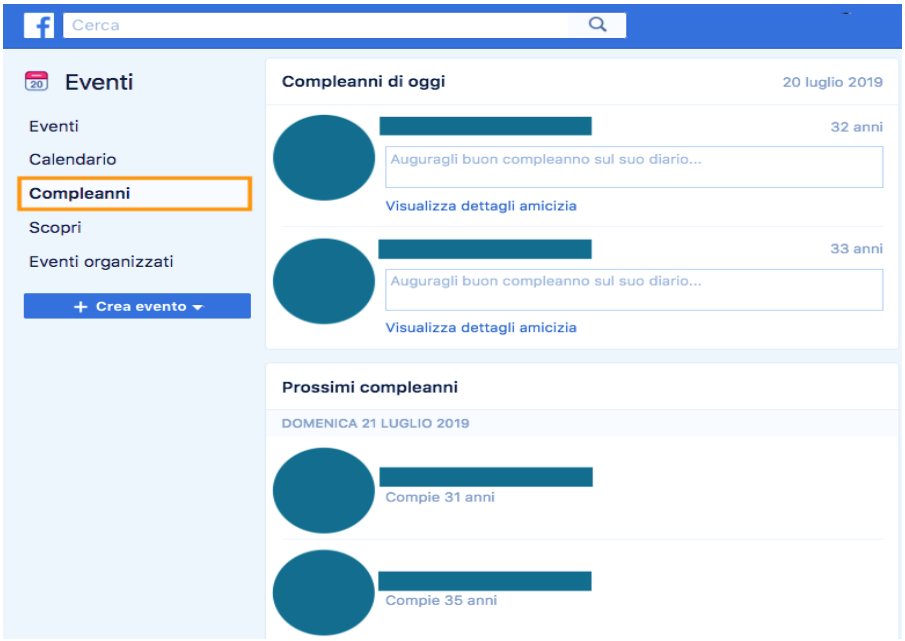


Figura 8 – La sezione compleanni su Facebook (Fonte: Facebook)

Ripetevo una frase mai sentita: "Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité."[28] Il calcolo delle probabilità era nella testa di Piere-Simon Laplace (1749-1827) e tra le menti più importanti del XVII secolo Friedrich Carl Gauss (1777-1855), Lambert-Adolphe-Jaques-Quételet (1796-1864), Francis Galton (1822-1911). Il calcolo delle probabilità si stava raffinando e presto si sarebbe potuto applicare anche a fenomeni naturali, legati alle scienze sociali, e così via.

Il punto cruciale fu scoprire che per molti fenomeni sembrava sottointesa una certa regolarità, detta anche distribuzione. La distribuzione di un fenomeno indica con quale probabilità si presentano determinati valori. Ad esempio, se immagino di estrarre dalla popolazione italiana una persona, sarà più probabile estrarne una con altezza 1.75m oppure di altezza 2.30m? Di peso 70kg o di peso 120kg?

Ogni fenomeno ha dei valori che si presentano più frequentemente e degli altri che si presentano più raramente. Gauss per primo aveva riscontrato questo comportamento sugli errori di misurazione; s era accorto che ripetendo un numero elevato di volte certe misurazioni, gli errori che si commettevano si distribuivano in un modo noto e, nella fattispecie, si distribuivano secondo una distribuzione detta *normale* (vedi tra poco). Ma non fu il solo: Quételet pubblicò uno studio, *Fisica sociale*, in cui studiò e misurò nei loro effetti e nelle loro mutue azioni le cause naturali e perturbatrici che agiscono sullo sviluppo dell'uomo e delle sue facoltà fisiche, morali e intellettuali (ref. Treccani, Luigi GALVANI). Così anche per lui molti fenomeni legati alle caratteristiche umane potevano essere ben rappresentati da una distribuzione normale.

[28] I fatti più importanti della nostra vita non sono altro che, nella maggior parte dei casi, dei problemi di probabilità, cit. Laplace.

Un'intuizione anticipata dagli scritti di Platone che qua riportiamo: "[…]*Credi forse che sia tanto facile trovare un uomo o un cane o un altro essere qualunque molto grande o molto piccolo o, che so io, uno molto veloce o molto lento o molto brutto o molto bello o tutto bianco o tutto nero? Non ti sei mai accorto che in tutte le cose gli estremi sono rari mentre gli aspetti intermedi sono frequenti, anzi numerosi?"* (Platone, Fedone, XXXIX)

Un fenomeno *normalmente distribuito* è quindi un fenomeno che presenta la maggior parte dei valori attorno alla media e presenta pochi valori estremi. Quindi, nell'esempio di Platone ci sono pochi persone (non distinguiamo qui tra i diversi generi) molto brutte e molto belle, molto alte e molto basse, molto lente e molto veloci, vedi Figura 9. In natura è più probabile osservare dei valori medi di tutte queste caratteristiche, quindi persone carine, persone di altezza normale (appunto!), persone né troppo lente né troppo veloci, normali insomma! Sei normale se i tuoi valori cadono nell'area verde di Figura 10, dove è rappresentata la distribuzione della altezza di un insieme di persone. Grazie alla distribuzione normale si calcolano i valori limite; all'interno di questi valori è compreso il 95% della popolazione, il restante 5% presenta valori maggiori o minori. Per cui una persona alta 1.63 m è nella norma perché compresa tra 154 e 193, i valori limite, vedi Figura 10.

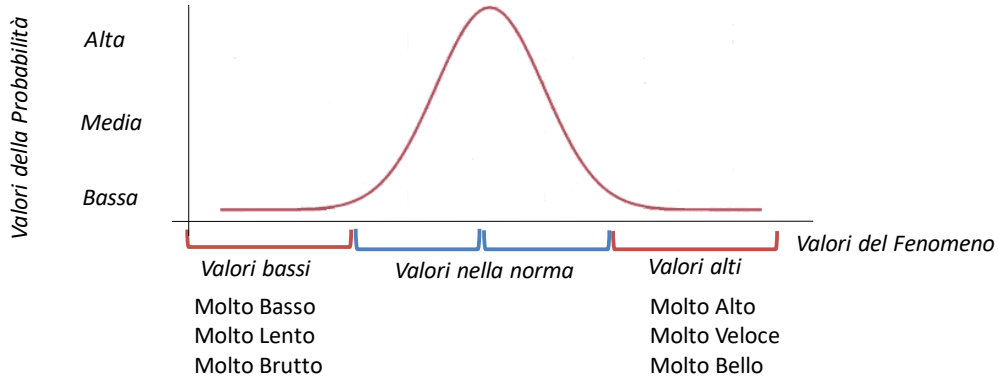


Figura 9 – Distribuzione dei valori di diversi fenomeni

.

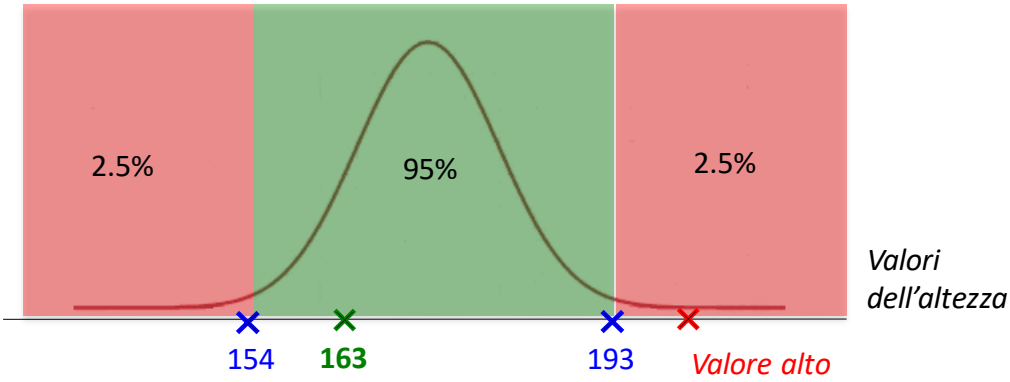


Figura 10 – Distribuzione dei valori della altezza

Anche i valori di riferimento del colesterolo negli esami del sangue sono ricavati dalla distribuzione normale: presa una popolazione sana si calcola quante persone hanno un certo livello di colesterolo. Così una nuova persona avrà colesterolo normale se cadrà all'interno di questi limiti ovvero avrà livello di colesterolo anormale se cadrà al di fuori, vedi Figura 11.

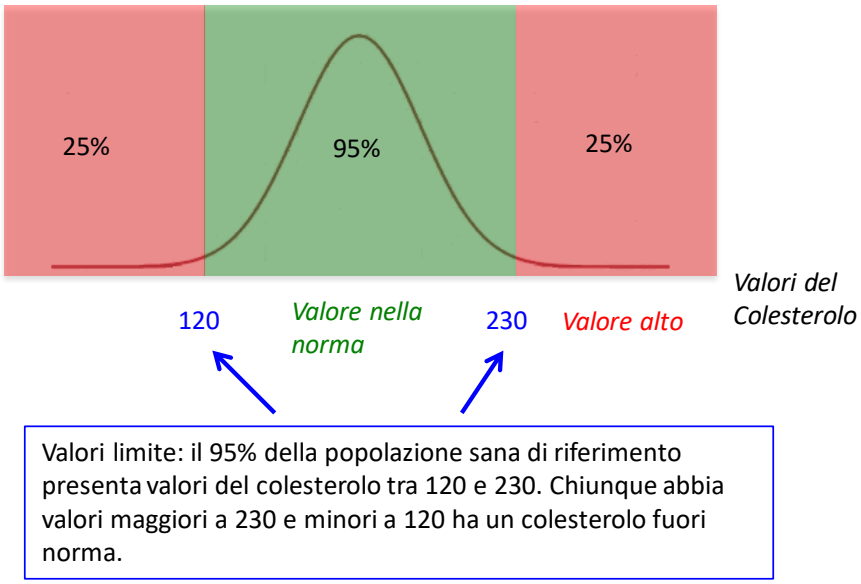


Figura 11 – Distribuzione dei valori del colesterolo

Mentre come Statistica Descrittiva continuavo ad occuparmi di aspetti prettamente demografici, un'altra me iniziava a diffondersi nella mente dell'uomo. Era sicuramente conseguenza della curva normale e di una serie di altre scoperte. Galton aveva creato questa macchina per fare un gioco, guardate la Figura 12.

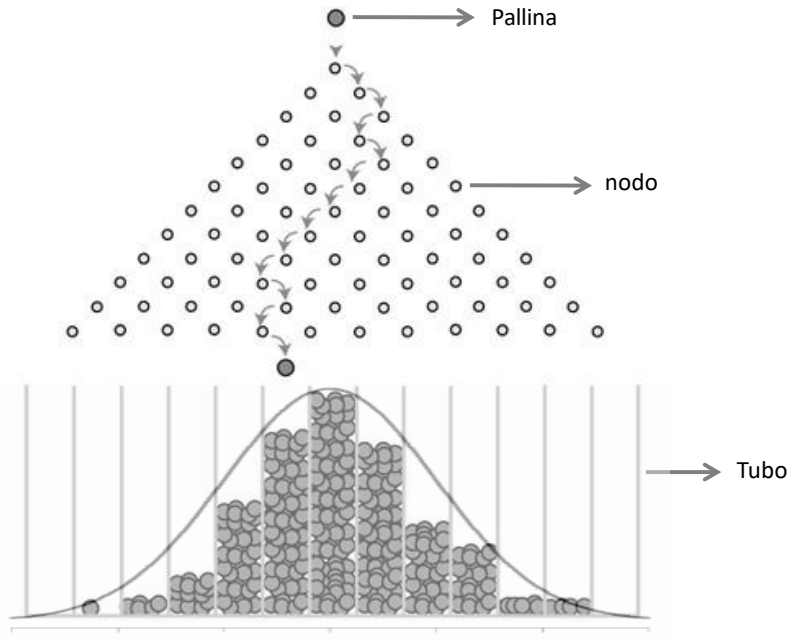


Figura 12 – Macchina di Galton (Fonte: https://pikabu.ru/tag/Galton%20Board/best)

Il gioco consiste nell'inserire una pallina in alto, come illustrato, e lasciarla cadere verso il basso. Ad ogni nodo, la pallina andrà a destra o a sinistra casualmente fino ad arrivare in uno dei tubi sottostanti. Ogni pallina farà un suo percorso e cadrà all'interno di un tubo. Lasciando cadere molte palline si nota che ci sono dei tubi che si riempiono maggiormente di altri. In particolare i tubi più esterni si riempiono meno dei tubi più interni. Se

pensiamo ai tubi esterni come a dei valori più alti e più bassi, e ai tubi interni come dei valori intermedi di un fenomeno, ci torna in mente proprio la distribuzione normale. Si scoprì anche che molti fenomeni che non hanno una distribuzione normale, se rilevati un gran numero di volte, tendono a distribuirsi normalmente.

E così anche Statistica Inferenziale se ne andò.

Come fossi collegata con la distribuzione normale non è semplice da spiegare. Vero è che gli illustri statistici del tempo mi usarono in maniera sconsiderata. Mi pareva di poter spiegare qualsiasi fenomeno attraverso questa distribuzione, ma forse in fondo non era poi così sbagliato.

In realtà la mia più grande espressione si sviluppò in concomitanza alla fisica dei gas. Immaginate di voler spiegare come si muove una molecola di un gas, e di voler prevedere i suoi movimenti. Ora prendete un po' di gas, quante molecole trovate? Miliardi! Era impossibile gestire e prevedere i movimenti di questa enorme quantità di molecole. Così si dovette selezionare una parte ristretta di molecole di gas, analizzarle e ricavare delle teorie su una parte limitata della popolazione da estendere alla popolazione totale. E così per molti altri casi in cui i dati erano limitati, per vari motivi, si cercava di trovare teorie generali sui fenomeni per poi estenderli, o meglio detto, inferirli sulla popolazione totale. Da qui il mio nome: Statistica Inferenziale.

Mi ricordo in particolare di William Gosset (1876-1937), che lavorava alla birreria Guinness. Allora come ora, i produttori della Guinnes volevano conoscere quali qualità di orzo e di luppolo e quali metodi di coltivazione, essiccamento, conservazione fossero migliori per la produzione della birra. Non era un compito facile: dovevamo accontentarci di piccole quantità di dati i cui risultati erano difficili da estendere correttamente su tutta la produzione di birra. I test statistici si sviluppano da questo momento in poi anche grazie a Wilhelm Lexis (1937-1914) e Karl Pearson (1857-1936). Brevemente cercherò di spiegarvi in cosa consistono questi test statistici affinché se ne possa apprezzare l'utilità: d'altronde chi meglio di me può spiegarvelo?

Rileviamo la quantità di grano prodotta da differenti tipologie di campo, rispettivamente morbido (indice 1) e duro (indice 2) e supponiamo che mediamente abbiamo prodotto una quantità di grano in quintali pari a

$$m_1 = 7.58 \text{ quintali} \quad m_2 = 8 \text{ quintali}$$

Vediamo bene che le due medie sono diverse e che la loro differenza è pari a 0.42 quintali. Possiamo dire che c'è una differenza tra la produzione di grano tra il campo morbido e il campo duro. Sbagliato! Seguite il mio ragionamento: noi abbiamo rilevato la quantità di grano da un campione di campi, quindi i dati ricavati riguardano una parte davvero minima rispetto alla produzione totale. Quello che vogliamo dimostrare è, però, se c'è una differenza generalizzabile a qualsiasi campo, ovvero che vi sia una differenza davvero importante, significativa che il campo morbido o duro evidenziano rispetto alla produzione di grano.
Evidentemente non è sufficiente dire che le due medie calcolate su un campione ristretto siano diverse per pensare che questo avvenga sempre. Per poterlo decidere si fa il seguente ragionamento: ipotizzo che la differenza di produzione di grano tra i due campi non esista quindi sia pari a zero. Se la differenza calcolata sul campione differisce *poco* da zero allora posso pensare che non ci sia differenza e che lo scarto sia dovuto al fatto di aver valutato un campione. Se invece differisce di *tanto*, è molto probabile che la mia ipotesi non sia giusta, ovvero i campi producono quantità di grano diverse. Quanto è *poco* e quanto è *tanto*? Per questo ci vengono in soccorso le distribuzioni statistiche che ci forniscono dei limiti (vedi Figura 13 le due linee verticali) a partire dai quali non è più possibile accettare l'ipotesi formulata.

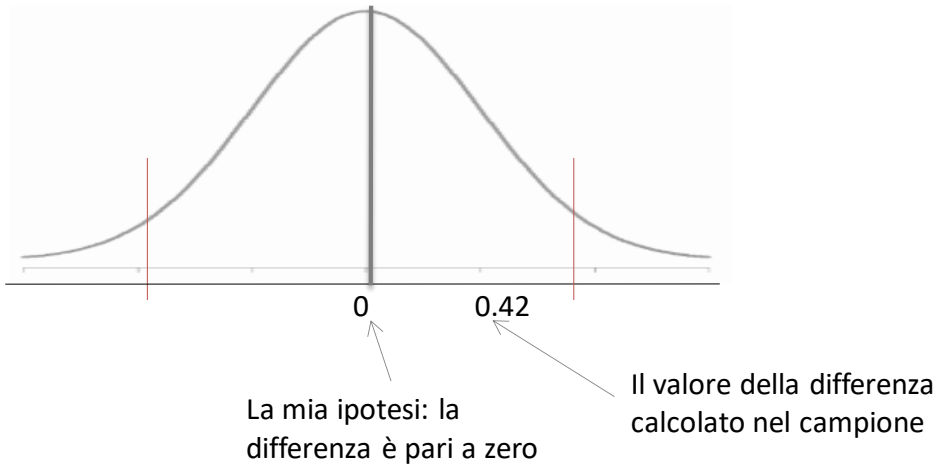


Figura 13 – Soglie per la accettazione delle ipotesi

Poiché 0.42 è compreso tra i limiti espressi dalle due linee verticali, è possibile accettare l'ipotesi che la differenza di produzione dei due tipi di campi non sia elevata, e inferire questo risultato a tutta la popolazione. Naturalmente questo è un risultato non assoluto, ma dipende da una probabilità; infatti, i limiti sono più o meno ampi a secondo della probabilità di "affidabilità" del risultato che si desidera ottenere. Maggiore sarà la probabilità di affidabilità richiesta, più ampi saranno i limiti e viceversa. Si veda schematicamente in Figura 14 il ciclo definito tra le tre fasi di ipotesi, raccolta dati e validazione.

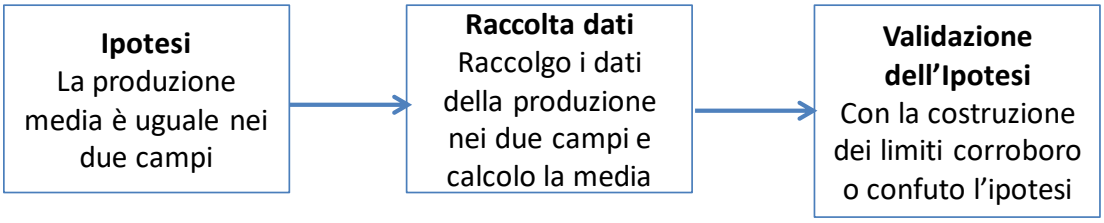


Figura 14 – Ciclo della ipotesi, raccolta dati e validazione

Ero diventata davvero sofisticata e avevo messo in crisi il metodo scientifico di Galileo Galilei, sostituendolo con il *metodo statistico.* La differenza essenziale era che ammettendo l'esistenza di eccezioni, il metodo scientifico non andava più bene perché non era possibile falsificare tutte le ipotesi. Ero diventata sofisticata, ma finalmente era possibile trarre delle conclusioni sufficiente-mente affidabili su campioni ridotti attraverso test statistici fare previsioni di fenomeni utilizzando dei modelli anch'essi calcolati a partire da una quantità di dati limitata.

Attraverso la formulazione di ipotesi riguardanti determinate caratteristiche di un fenomeno, permettevo agli statistici di capire quali fenomeni potessero essere la causa di certe osservazioni.

Nell'esempio dei campi ci si chiedeva se il tipo di suolo del campo potesse essere la causa di una produzione maggiore o minore di grano. Contemporaneamente, mi usavano per creare modelli che potessero spiegare le relazioni tra due o più fenomeni. Una volta trovata la causa di un certo fenomeno, diveniva interessante riuscire a capire in che modo e con quale intensità la causa influenzava il fenomeno. Ero diventata lo strumento di studio e conferma di analisi di una vastissima gamma di ambiti: economia, medicina, fisica, biologia, psicologia, e molti altri e le tecniche sviluppate cominciarono a moltiplicarsi, vedi Figura 15.

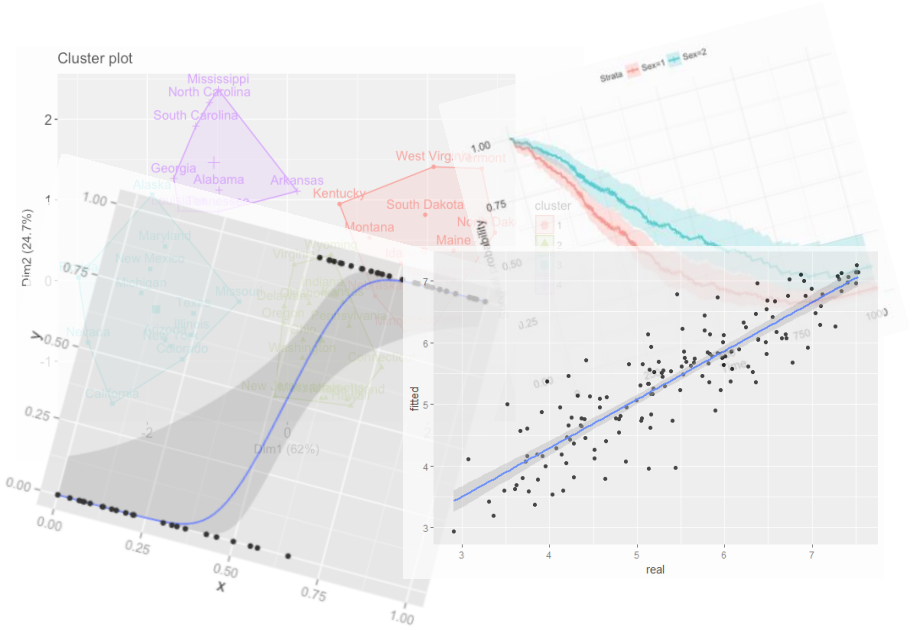


Figura 15 – Diverse visualizzazioni di tecniche statistiche di previsione

Tra le tecniche che mi videro applicata in questo periodo vorrei presentarne due che tutt'oggi sollevano un dibattito molto acceso tra gli Statistici. Galton e Pearson studiavano insieme a me le tecniche che potessero spiegare in modo approfondito le cause e le relazioni tra fenomeni. Così mi plasmarono e mi fecero diventare *Regressione Lineare* e *Coefficiente di Correlazione di Pearson*. Quest'ultimo è un coefficiente statistico che permette di calcolare se due fenomeni possono presentare un legame che porta a mutare insieme in senso crescente. I due legami possibili sono i seguenti:

- quando il primo fenomeno assume valori crescenti, cresce anche l'altro, ad esempio se la temperatura in un certo luogo aumenta, aumenta il consumo di acqua in quel luogo;
- all'aumentare dei valori di un fenomeno, diminuiscono i valori dell'altro, come ad esempio se lo stipendio diminuisce, aumenta il numero di ricerche di lavoro su Google.

La regressione lineare, rispetto al coefficiente di correlazione lineare, permette in più di stabilire la tipologia di relazione causa-effetto tra due fenomeni. Ad esempio, l'altezza e il peso sono due variabili correlate, se aumenta l'una, mi aspetto che aumenti anche l'altra. Ma chi tra i due fenomeni è la *causa* e chi l'*effetto*? In altre parole: è l'aumento del peso a causare l'aumento dell'altezza o, viceversa, è l'aumento di altezza che causa l'aumento di peso? Siamo sicuramente nel secondo caso: mentre una maggiore altezza causa un peso maggiore, al contrario non è detto che un peso maggiore comporti una maggiore altezza. La regressione lineare è uno dei tanti modelli che traduce questa relazione di causa effetto in linguaggio matematico.

La linearità della regressione stabilisce un certo tipo di legame tra i due fenomeni X e Y che è appunto lineare, $Y = a + b X$. Per stabilire questo legame, si misura il fenomeno su un campione, si raccolgono i dati e si determinano i coefficienti a e b secondo alcune tecniche statistiche, la più impiegata è il metodo dei minimi quadrati. Questo determina il modello, vedi esempio in Figura 16. Il modello retta è calcolato attraverso il metodo dei minimi quadrati impiegando i dati presi dal World Happiness Report. In questa analisi si voleva capire che cosa determinasse la felicità (effetto) in 155 paesi a partire da fattori (causa) come Libertà, Salute, Aspetti economici, altro (https://www.kaggle.com/unsdsn/world-happiness).

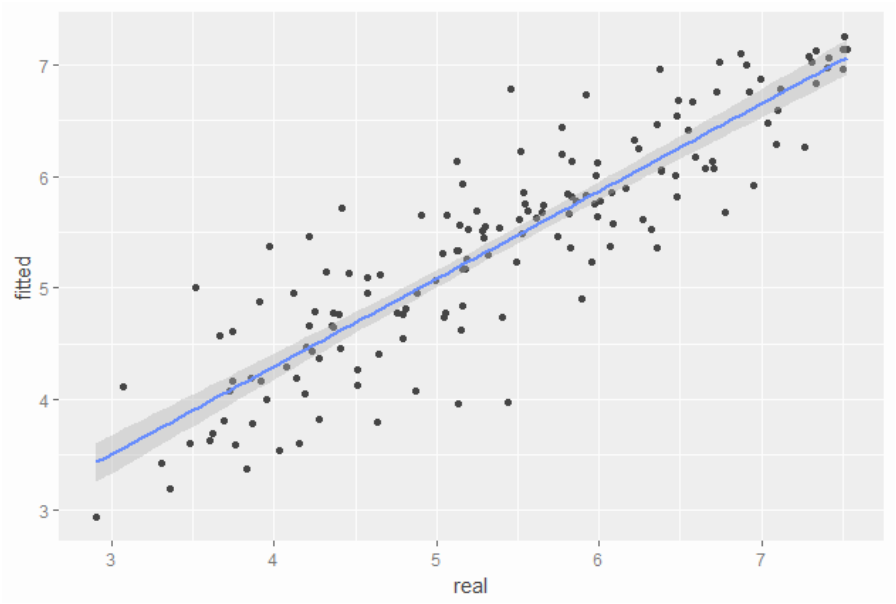


Figura 16 - Il modello retta calcolato attraverso il metodo dei minimi quadrati impiegando i dati acquisiti dal World Happiness Report

Il confronto tra correlazione e la regressione lineare, è molto importante: la corrispondenza tra fenomeni è sempre stata ed è tutt'oggi il mio punto di forza, ma anche di debolezza. Scegliere l'ordine di causalità (quale fenomeno è la causa e quale è l'effetto) non è semplice: richiede uno studio approfondito del fenomeno, richiede un'alta esperienza da parte di chi usa lo strumento di previsione e richiede anche una linea guida rispetto all'etica di utilizzo.

Confondere la causa con l'effetto è spesso molto facile a volte purtroppo molto comodo per tre motivi:

1) La correlazione tra due fenomeni non è indice di causalità: due fenomeni possono essere correlati, ma non essere in un rapporto di causa-effetto.
2) La correlazione tra due fenomeni può essere *spuria*: due fenomeni possono essere correlati perché sono correlati entrambi a un terzo fenomeno, spesso difficile da individuare, che però genera una "falsa correlazione".
3) Capire quale fenomeno è la causa e quale l'effetto non è semplice.

Caro lettore, tutti questi aspetti relativi alla relazione complessa tra i concetti di correlazione e causazione, o relazione causa effetto, saranno affrontati in maniera più approfondita nel Capitolo 17 sui limiti della Scienza dei dati. Mentre invece mi preme di approfondire qui gli aspetti relativi alla correlazione.


La Provincia

seleziona un comune

Como Cintura Lago Cantù Erba Olgiate Bassa Economia Sport BLOGAZZ

HOME / SALUTE / NOBEL E CIOCCOLATO UNA RELAZIONE C'È

Lunedì 15 Ottobre 2012

Nobel e cioccolato
Una relazione c'è

C'è una correlazione diretta tra il consumo di cioccolata di un Paese e il numero di premi Nobel che riesce a produrre.

Lo afferma uno studio pubblicato dal prestigioso «New England Journal of Medicine», secondo cui il fenomeno potrebbe essere correlato proprio agli

Figura 17 – "Nobel e cioccolato, una relazione c'è"(articolo da La Provincia di Como)

Ahimé, negli ultimi anni sono stata proprio strapazzata! Molti concetti possono essere mal interpretati e mal utilizzati proprio per via della mia natura. I dati, le interpretazioni e la selezione di essi, i concetti di regressione, correlazione e correlazione spuria possono essere oggetto di manipolazioni errate che dipendono dalla capacità dell'esperto che mi utilizza.

Cito lo studio dell'articolo "Chocolate Consumption, Cognitive Function, and Nobel Laureates", Franz H. Messerli, M.D. 2012 in cui viene dichiarato che il consumo di cioccolato è correlato con il numero di premi Nobel in uno stato, vedi anche in Figura 17 un articolo in propoito apparso sul giornale La Provincia di Como. La correlazione riportata nello studio, pari a $r = 0.791$ è molto alta e positiva, ovvero il consumo di cioccolato e il numero di premi Nobel hanno un rapporto crescente (aumentano simultaneamente e in maniera importante); il valore $p < 0.0001$ indica che questo valore è inferibile, valido, per tutta la popolazione. Inoltre, viene asserito che "The slope of the regression line allows us to estimate that it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1", ovvero consumando 0.4 kg di cioccolato pro capite per anno il numero di laureati che hanno vinto un premio Nobel di un dato paese aumenta di uno (ovviamente la Svizzera è al primo posto con maggior consumo di cioccolato e maggior numero di premi Nobel). In ultimo, nell'articolo scientifico, viene trovata una causa confirmatoria, ovvero che il consumo di cioccolato aumenti le capacità cognitive e, di conseguenza, la possibilità di eccellere nello studio. Sto assumendo che i dati siano rilevati in maniera consona all'analisi.

Non per fare la Statistica, ma l'articolo in Figura 17 e lo studio scientifico di riferimento non tengono in considerazione principalmente la problematica della correlazione spuria. L'analisi è incompleta, perché non include una ricerca di un'eventuale variabile che rende la correlazione spuria, e non indaga sull'eventualità che siano altre variabili a determinare l'alta correlazione di cui sopra e presentata in Figura 18. Vi spiego subito cosa intendo: nel grafico, in alto a destra, tra maggiori consumatori di cioccolato e i più virtuosi, io vedo paesi più freddi, in cui il consumo di cioccolato può essere effettivamente maggiore per via delle temperature. Considerando i paesi che sono stati scelti nella indagine, nei paesi più freddi, oltre al consumo di cioccolato, anche la produttività e il grado di istruzione sono maggiori, rispetto ai paesi in basso a sinistra, con climi più mediterranei (Italia, Portogallo, Spagna, Grecia) in cui grado di istruzione e produttività sono minori, vedi Figura 18.

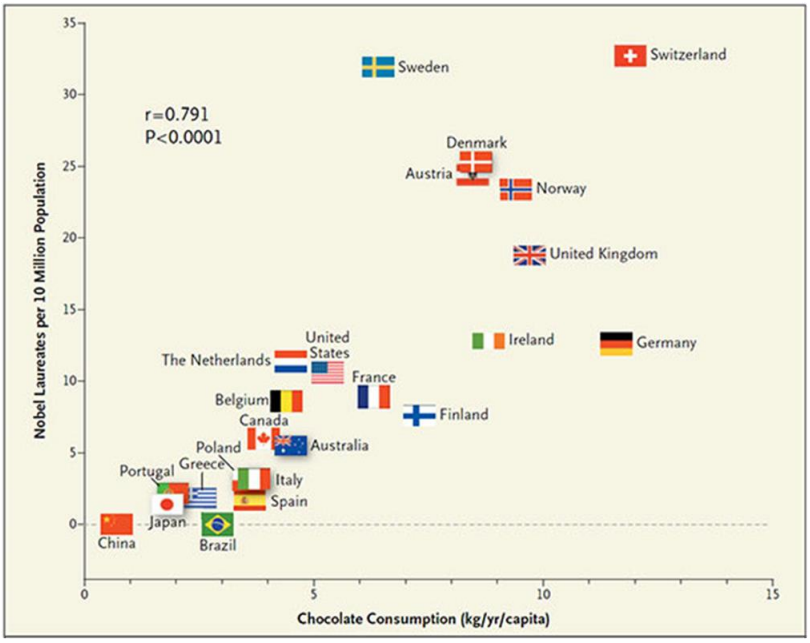


Figura 18 - La correlazione r = 0.791 è molto alta e positiva, ovvero il consumo di cioccolato e il numero di premi Nobel hanno un rapporto crescente (aumentano simultaneamente), p < 0.0001 indica che questo valore è inferibile, valido, per tutta la popolazione. (Fonte: Franz H. Messerli, M.D.)

Analizzare cause-effetto di un fenomeno non è semplice e anche coloro che utilizzano metodi molto sofisticati non possono prescindere dal dover integrare i propri studi con l'analisi e lo studio approfondito dei dati, prima di poter concepire un qualsiasi modello di previsione o una qualsiasi relazione tra fenomeni. Nel XX secolo, con l'aumentare della complessità e della quantità dei fenomeni che venivano studiati, lavoravamo insieme io, Statistica Inferenziale e io Statistica Multivariata. Anche informatica iniziò ad essere preponderante nel supporto alle nostre analisi; a quantità di dati e di calcoli da gestire diventavano sempre più impegnativi e richiedevano risorse di calcolo sempre maggiori.

Per riuscire a compiere studi approfonditi sui dati raccolti, Statistica Multivariata fu arricchita di svariate tecniche per l'identificazione di correlazioni, similarità, causalità, studio degli effetti dipendenti da differenti cause contemporaneamente, anche detti *multivariati*. Un esempio è *l'analisi dei gruppi*. Preso un certo numero di unità statistiche e rilevate su di esse alcune caratteristiche, come altezza, peso, età, stipendio, ecc.. si creano dei gruppi basati su queste caratteristiche, che saranno utili per capire se vi sono degli individui tra loro più simili. Trovate così delle caratteristiche discriminatorie, sarà più semplice svolgere ulteriori analisi. Nell'esempio seguente sono stati raccolti i dati di individui di specifiche città USA secondo diverse tipologie di arresto per reati. Applicando un'analisi dei gruppi si sono creati tre gruppi che vediamo nel grafico in Figura 19. Dal grafico si evince che, ad esempio, South e North Carolina siano simili in termini di tipologia di arresti.

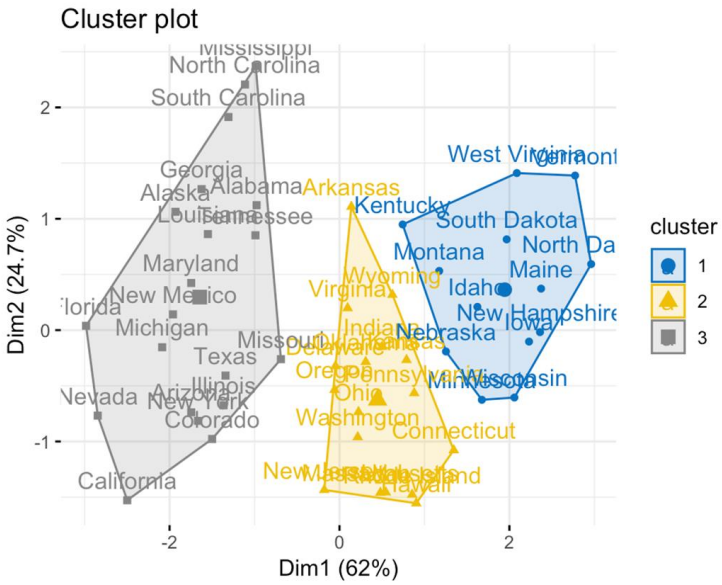


Figura 19 - Analisi dei gruppi basata sul metodo k-means di città USA su cui sono state rilevate differenti tipologie di arresti. Per maggiori approfondimenti consultare il sito https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/

“E’ l’era dei Big data” disse Presente, “finalmente i limiti computazionali sono stati sorpassati e Internet ha velocizzato lo scambio di informazioni”. “Le informazioni sono tante e ovunque. Basti pensare quante persone sono iscritte a Facebook” disse Machine Learning andandosene.

Machine Learning (che approfondiremo nel prossimo Capitolo 10) si trovava in moltissimi posti del mondo contemporaneamente: nei Social network, negli e-commerce, nelle case, negli smart-phone e negli smart watch delle persone, negli uffici, sugli aerei, nei taxi, insomma davvero dappertutto. C’era un sacco di lavoro per lui, Machine Learning, e infatti poteva fare mille cose: ascoltare discorsi, capirli, tradurre frasi in un sacco di lingue diverse, riconoscere volti, riconoscere persone, raggruppare le foto in album, dare consigli su come fare il selfie, riconoscere le attività fisiche svolte e calcolare le calorie, calcolare il numero di ore di sonno e decidere se fossero sufficienti, suggerire alle persone di fare movimento, suggerire una dieta personalizzata, sapere dove fosse un posto e portare le persone nel minor tempo possibile lì e contemporaneamente evitando caselli, autostrade, strade interrotte, strade senza uscita, strade disconnesse, e molte molte cose ancora.

Per ricompensare Machine Learning di questo enorme lavoro lo nutrivano; lo nutrivano di una quantità di informazione enorme. Io che ero dovuta divenire campionaria per poter riuscire a ricavare un briciolo di informazione dai dati, mi ritrovavo di fronte a questa nuova realtà, spiazzata. Per il riconoscimento facciale, ad esempio, riceveva in pasto milioni di immagini di volti, fotografie, tutte simili, un po’ modificate, ma anche diverse, con la nebbia, col sole, coi colori sfalsati, con dei visi simulati e chi più poteva dargli da mangiare meglio era, perché poteva lavorare meglio. Molti dicevano che diventava più accurato quando la quantità di dati aumentava e quindi Machine Learning veniva rimpinzato il più possibile di nuove informazioni, dati e molto altro.

Come faceva a funzionare così bene non lo sapeva nemmeno lui, ma funzionava e questo era un punto importante. Però, se da un lato il numero di coloro che lo utilizzavano e credevano nelle sue capacità cresceva in maniera rapidissima, dall’altro un significativo numero di persone cercavano di trovare una spiegazione al suo funzionamento e un modo quindi per poter gestire e controllare le sue capacità, io compresa (explanatory

machine learning). Non era semplice intraprendere uno studio del genere. La sua natura era molto diversa dalla mia. Io, Statistica, analizzavo i fenomeni in maniera completamente diversa.

La prima differenza in assoluto era l'approccio. Io ero basata sul metodo statistico, ora chiamato *model-driven approach*, dove lo studio del fenomeno e delle sue possibili cause è preponderante. Di conseguenza avveniva prima la stesura delle ipotesi, con la formulazione di un modello, poi la raccolta dei dati e infine la corroborazione o la confutazione dell'ipotesi, vedi ancora Figura 14. Ad esempio, se la relazione tra altezza e peso veniva ipotizzata lineare, una volta scelto un campione di dati e validata, la relazione tra altezza e peso sarebbe stata sempre lineare, fino a prova contraria.

L'approccio di Machine Learning, chiamato *data-driven,* era molto diverso dal mio. Non si facevano ipotesi a priori, ma ipotesi istantanee, dipendenti dai dati del momento. Così, la relazione tra altezza e peso non era definita da un modello specifico, ma data un'altezza, prevedeva direttamente in quel momento quale fosse il peso più realistico in base ai dati che aveva a disposizione.

Questa differenza è tutt'oggi dibattuta e discussa, la riprenderemo più diffusamente nel Capitolo 17; entrambi gli approcci presentano pro e conto. In generale l'approccio data-driven è un approccio molto libero poiché la relazione tra i fenomeni in gioco non è definita a priori con delle ipotesi specifiche, ma in grado di mutare insieme ai dati, e, nel momento in cui i risultati sono apprezzabili, non c'è motivo per cui non debba essere preferibile. La critica che mi sento di fare da Statistica, è quanto questa libertà possa portare a una crescita in termini di conoscenza in senso scientifico.

La scienza intesa come studio dei fenomeni, della loro interazione causa-effetto, non può limitarsi alla sola esplicazione momentanea dei fenomeni, ovvero non può ridursi al solo metodo induttivo. Il metodo induttivo può essere uno strumento iniziale per condurci a una spiegazione finale, come avviene con le scoperte che poi devono essere tramutate in realtà sistematiche e prevedibili. La Scienza ci insegna che le informazioni che siamo in grado di produrre o che osserviamo dall'ambiente che ci circonda sono spesso fallaci.

Pensiamo ad esempio al sole; secondo la nostra esperienza, il sole non è altro che un puntino nel cielo. E' solo grazie alla scienza e alle teorie confutate e corroborate che oggi sappiamo che il sole è un'enorme stella che ci riscalda. Dunque, una priorità è quella di arrivare a una comprensione intrinseca dei fenomeni per essere in grado di ricavare conoscenza dall'esperienza. Per utilizzare una metafora: molti di noi usano la macchina anche se pochi realmente sanno come funziona ciascun pezzo. Questo non ci impedisce di guidarla, ma, nel caso di malfunzionamento o di manutenzione, ci deve essere un meccanico o un esperto che sia in grado di ripararla, altrimenti saremmo costretti, ogni volta, a buttarla via e a costruirne una nuova.

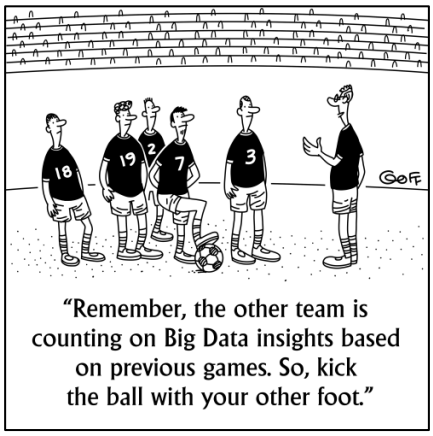


Figura 19 – I paradossi dell'apprendimento

La seconda differenza era rappresentata proprio dai dati. Io, Statistica, dedicavo sempre una larga parte dei miei studi all'esplorazione, allo studio, alla selezione e alla scelta dei dati. Machine Learning si nutriva di qualsiasi dato. Era nato coi Big data, era stato fatto apposta. Una sostanziale differenza, quindi, derivava proprio dalla sua "dieta". Come l'essere umano deve seguire una dieta che deve essere curata, equilibrata, monitorata per avere dei benefici a breve, medio, ma anche e soprattutto a lungo termine, anche Machine Learning doveva nutrirsi di dati corretti, equilibrati, mirati per lo scopo della mia analisi.

Questo processo di selezione non veniva e non viene praticamente mai tenuto in considerazione, ed è questa la maggior critica che mi sento di fare. In realtà c'è di più: come l'essere umano può costruire il proprio orto e il proprio allevamento e nutrirsi di quello, anche Machine Learning aveva imparato a costruire il proprio orto. Le informazioni che ricavava dai dati iniziali, le ritrasformava in "nutrizione", e quindi poteva riutilizzarle come nuova informazione, e così via. La scelta dell'infor-mazione da utilizzare è e deve essere una scelta ragionata, soprattutto nell'utilizzo ripetuto. Non essere in grado di *scegliere* le informazioni da utilizzare per la previsione di un certo fenomeno è sinonimo di non conoscenza del fenomeno. Tutto ciò può generare paradossi come quello evocato scherzosamente in Figura 19, che esprime un cortocircuito che l'allenatore di una squadra di calcio può creare tra le previsioni di una tecnica di Machine Learning basata sulle partite pregresse, e l'idea di cambiare le carte in tavola giocando con un piede diverso rispetto a quello usato nel passato.

Questa problematica, che si riallaccia fortemente alla differenza precedente, non solo non genera conoscenza globale, ma può portare a una perdita di conoscenza. Un'enorme quantità di informazione non controllata, può portare con sé delle contraddizioni, e le contraddizioni possono portare a una previsione sbagliata. Se da una parte è vero che a una realtà complessa deve corrispondere un'informazione altrettanto complessa, è anche necessario conoscere quale informazione deve essere utilizzata per ricavare dei risultati corretti. La mancanza di una struttura in questi termini proviene dalla mancanza di studio dei fenomeni. La troppa informazione diventa sovraccarico informativo e disinformazione, e le relazioni di causa-effetto tra fenomeni si perdono in questa moltitudine di informazioni. Non è un caso che, insieme ai Big data siano arrivate anche le (Big) Fake news.

Storia e Chiaroveggenza ci guardavano dall'alto. "Non sono servita a niente" diceva Storia, ancora pensando a tutte le fake news e alla grande confusione di questi momenti. "Non ti preoccupare" disse Chiaroveggenza "troveranno una soluzione. Questa non è la fine della teoria, è solo l'inizio di una teoria nuova".

Riferimenti per approfondimenti

M. Zenga - *Lezioni di Statistica Descrittiva* – Seconda Edizione*,* Giappichelli, 2004.

Lezioni di Statistica Descrittiva è un libro di introduzione e approfondimento della statistica descrittiva univariata e multivariata. E' un libro propedeutico a qualsiasi altra forma di statistica, poiché introduttivo alle nozioni di base della stessa e alle prime analisi. Approfondendo esclusivamente l'aspetto statistico descrittivo, si prefigge come obiettivo quello di analizzare i fenomeni statistici in maniera non probabilistica, ovvero considerando l'intera popolazione, non un campione. Partendo dalla mera rilevazione dei fenomeni propone un'eccellente raccolta di tecniche statistiche per l'analisi di uno o più fenomeni congiuntamente. Al suo interno si trova la spiegazione delle forme base di analisi come i grafici, le medie, gli indici di variabilità, ecc. Inoltre, vengono presentati i concetti di modello di regressione lineare e correlazione e di distribuzione normale semplicemente introdotti in questo capitolo.

S.M. Ross, *Calcolo delle Probabilità* - Terza Edizione, Apogeo Education, 2013.

E' uno dei libri maggiormente utilizzati per la didattica e per gli approfondimenti delle basi del calcolo delle probabilità. Tratta prevalentemente di argomenti di base, ma anche di argomenti molto avanzati soprattutto negli ultimi capitoli. Si presenta come un libro molto completo, spaziando dal calcolo combinatorio, dalle variabili causali e la loro applicazione, fino ad arrivare a modelli di simulazione e i processi stocastici, tra cui il processo di Poisson e le catene di Markov. E' una raccolta molto ricca di contenuti ed esempi e permette al lettore alle prime armi di poter approfondire la materia in maniera completa e rigorosa, oltre che permettere di arrivare ad un livello di conoscenza avanzato della modellizzazione probabilistica.

G.Landenna, D.Marasini, P.Ferrari - *La verifica di Ipotesi Statistiche* – Prima Edizione, Il Mulino, 1998

La verifica di Ipotesi Statistiche è un libro che introduce e approfondisce i concetti di statistica dei campioni passando dai concetti di stima e di stimatori. Conseguentemente, tratta dei test di significatività e della teoria di J. Neyman e di E. Pearson relativi ad essi. Oltre a questi, vengono presentati i modelli lineari e l'inferenza sui parametri del modello e sulla previsione. Ulteriori approfondimenti riguardano altri test statistici molto impiegati per il confronto tra più gruppi come l'analisi della varianza (ANOVA) e, in aggiunta, i test non parametrici. Completano i contenuti i test funzionali, di indipendenza e di associazione.

Capitolo 10 – Machine Learning

Fabio Stella

1. Introduzione

Come abbiamo cominciato a sottolineare alla fine del Capitolo 9, il *Machine Learning* è intorno a noi, in ogni minuto della nostra giornata, in ogni azione che compiamo, in ogni luogo nel quale ci troviamo, ci supporta e ci influenza in diversi aspetti più o meno rilevanti della nostra vita. Che lo si creda o meno, questo accade da anni, e la natura pervasiva del *Machine Learning* è destinata ad aumentare il suo impatto nella società, nella politica, nell'economia, nella sanità e in molti altri settori.

Quando camminiamo per la città, sembra che i pannelli pubblicitari sappiano chi siamo, vale a dire pare conoscano i nostri interessi e desideri. Il messaggio pubblicitario, ospitato dal pannello, varia in modo adattivo in base al mese dell'anno, al giorno della settimana, alla fascia oraria della giornata, e financo in base alla temperatura e alle condizioni metereologiche del momento. Cosa analoga accade quando, trovandoci in una località che non conosciamo bene o che non conosciamo affatto, tramite il nostro smartphone ci facciamo suggerire da *Yelp*, *TheFork* o *TripAdvisor*, quale sia il miglior posto dove pranzare o cenare. Il *Machine Learning* usa informazioni sulla nostra posizione e sfrutta il nostro profilo digitale per decidere non solo quali locali suggerirci ma anche l'ordine con il quale i locali alternativi debbano esserci mostrati.

Quando guidiamo un'automobile, tecnologia matura ma caratterizzata da livelli di complessità elevati, forse ignoriamo che tutte le sue componenti debbono cooperare per un fine ultimo, garantirci piacere e sicurezza della guida. Pertanto, è necessaria un'orchestrazione tra volante, freno, acceleratore, cambio, motore, sistema di trasmissione, impianto frenante, impianto di alimentazione e pneumatici, vale a dire un continuo adattamento dei parametri di funzionamento di queste componenti. Il *Machine Learning* è responsabile di apprendere il nostro stile di guida, consentendo di gestire in modo adattivo i parametri di funzionamento delle diverse componenti, assecondandoci e garantendoci una guida piacevole e sicura. Il nostro stile di guida interessa molto anche alle compagnie assicurative, che lo studiano per proporci un contratto assicurativo adeguato, vale a dire un contratto che preveda la corresponsione di un premio assicurativo equo per la nostra probabilità di sinistro. I dispositivi che un numero crescente di compagnie assicurative chiedono di installare sulle nostre vetture, per abbassare il premio assicurativo, hanno proprio lo scopo di raccogliere i dati che il *Machine Learning* utilizzerà per inferire il nostro stile di guida e quindi la nostra collocazione in quella che si chiama *griglia del rischio*.

Quando intraprendiamo un lungo viaggio in treno, e per ingannare il tempo decidiamo di assistere ad un film, come prima cosa ci colleghiamo tramite smartphone a Netflix o Infinity o ad una piattaforma analoga. Appena collegati, appaiono sullo schermo dello *smartphone* le locandine di alcuni film. Sia i film mostrati, che la posizione che essi occupano sullo schermo dello *smartphone* sono stabiliti da decisioni prese da un *Recommendation System*, un particolare tipo di algoritmo di Machine Learning che in base ai nostri comportamenti passati, in base a quali film abbiamo visto nel passato recente, ai nostri gusti e ai gusti delle persone che ci assomigliano, prende una decisione in merito a quali film

suggerirci per il nostro viaggio. Quando accediamo alla nostra casella di posta elettronica, la grande parte di messaggi pubblicitari ci viene risparmiata, questo accade in quanto un apposito algoritmo li ha indentificati come spazzatura (spam) e conseguentemente li ha filtrati o rimossi dalla nostra casella di posta elettronica.

In conclusione, tutte le volte che utilizziamo un dispositivo elettronico, *computer*, *tablet*, o *smartphone*, dobbiamo essere consapevoli del fatto che il *Machine Learning* è presente, persistente, pervasivo, ci supporta e ci influenza in diversi modi e a diversi livelli.

In passato, l'unico modo per far svolgere un compito ad un computer – a partire dal semplice calcolo della somma di due numeri sino al governare l'avionica per far volare un aeroplano – prevedeva la scrittura di un algoritmo che dettagliasse, istruzione per istruzione, come svolgere il compito medesimo. Al contrario, un algoritmo di Machine Learning, spesso indicato con il termine di *learner*, svolge il compito assegnatogli in modo molto differente, basandosi principalmente, se non esclusivamente, sulla disponibilità di dati, a partire dai quali è in grado di apprendere quali sono i modi possibili per raggiungere il proprio obiettivo, e quindi assolvere al compito che gli è stato assegnato. In generale, più dati sono disponibili sul compito che si desidera apprendere, migliori saranno le prestazioni ottenute dal *learner*, anche se non tutti i compiti possono essere appresi e, ancor peggio, in alcuni casi non è nemmeno possibile decidere se un determinato compito possa essere appreso o no a partire dai soli dati disponibili.

La realtà è che le aziende hanno raggiunto un livello di conoscenza dei propri clienti che solo pochi anni fa era inimmaginabile. Il politico che disponga del sistema di previsione del comportamento degli elettori più efficace è destinato a prevalere in una competizione elettorale. Veicoli di varia natura, ad esempio i droni, attraversano acqua, terra e aria viaggiando autonomamente o con minimo intervento da parte dell'uomo, e adattandosi alle condizioni del contesto nel quale si trovano ad operare. Nessuno ha esplicitamente codificato il tuo gusto o la tua preferenza per un determinato prodotto offerto da *Amazon.com*; è il sistema di raccomandazione di Amazon, tramite un algoritmo di Machine Learning, che ha provveduto ad analizzare i tuoi acquisti passati e gli acquisti passati di clienti a te simili, per giungere a decidere quando e quale prodotto suggerirti per l'acquisto.

Figura 1 - Large Hadron Collider. Copyright: Maximilen Brice/CERN (Fonte: https://home.cern/news/news/accelerators/cerns-large-hadron-collider-gears-run-2)

In questo scenario da *Blade Runner*, quello che appare chiaro è come il Machine Learning stia dettando i tempi del cambiamento di società, scienza, tecnologia, commercio, politica, salute, medicina e purtroppo anche dei conflitti armati. Infatti, satelliti, sequenziamento del DNA ed acceleratori di particelle, come il Large Hadron Collider (Figura 1 nella pagina precedente), indagano la natura che ci circonda a livelli di dettaglio finissimo, mentre gli algoritmi di Machine Learning cercano di trasformare questo enorme flusso di dati in nuova conoscenza il più delle volte a fini decisionali.

Questo capitolo è organizzato come segue; la Sezione 2 presenta una mappa dei tipi di problemi che il Machine Learning può formulare e risolvere. In particolare, la sezione in questione è finalizzato a fornire un mappa delle diverse tipologie dei problemi di Machine Learning. In particolare, per ogni tipologia di problema di Machine Learning, vengono descritte alcuni esempi di applicazioni reali. La Sezione 3 descrive alcuni dei principali algoritmi di Machine Learning, nello specifico viene presentato il modello dei Decision Tree, modello caratterizzato da un elevato livello di interpretabilità, insieme al modello delle Reti di Neuroni Artificiali, con particolare attenzione ai modelli di Deep Learning, che hanno assunto un'importanza notevole per la formulazione e per la risoluzione di complessi problemi di riconoscimento delle immagini, di estrazione automatica dell'informazione a partire da testo in linguaggio naturale e di riconoscimento del parlato. La Sezione 4 chiude il capitolo con alcune riflessioni e conclusioni sul Machine Learning ed il suo impatto a breve, medio e lungo termine.

2. Tipologie di Problemi

Il Machine Learning è definito come un insieme di metodi in grado di elaborare i dati a disposizione in modo da:
- Scoprire modelli nascosti che parlino del processo che li ha generati.
- Utilizzare modelli che sono stati scoperti, per prevedere dati futuri o altri esiti di nostro interesse.
- Prendere decisioni in condizioni di incertezza ed eventualmente indicare se e come collezionare dati aggiuntivi per migliorare l'efficacia di tali decisioni.

Uno degli approcci principali per formulare e trattare modelli di Machine Learning, sebbene non il solo, è offerto dal calcolo delle probabilità, che consente di descrivere, affrontare e risolvere problemi caratterizzati dalla presenza di incertezza. Nel Machine Learning, l'incertezza si presenta in forme differenti e queste sono alcune delle domande che ci poniamo:
- Qual è la previsione ottimale circa il futuro, alla luce dei dati disponibili sul passato?
- Qual è il miglior modello per spiegare un certo insieme di dati?
- Quali misurazioni è necessario effettuare per aumentare l'attendibilità dell'analisi?

Il Machine Learning risponde alle domande precedenti e ad altre domande simili, tramite differenti tipi di apprendimento. In particolare nel Machine Learning si distinguono:
- L'apprendimento supervisionato o predittivo,
- L'apprendimento non supervisionato o descrittivo,
- l'apprendimento per rinforzo.

2.1 Machine Learning supervisionato

Presentiamo Il primo tipo di Machine Learning con l'ausilio del seguente problema.

PROBLEMA 1: Il settore della telefonia mobile è competitivo e le società che vi operano lanciano campagne di marketing molto aggressive con l'obiettivo di strappare clienti ai concorrenti. La competizione causa migrazioni di clienti da una società all'altra, migrazioni motivate da vantaggi di costo o da una migliore qualità dei servizi offerti. In questo caso si parla di abbandono dei clienti (customer churn), Il fenomeno che porta uno o più clienti di una compagnia telefonica ad abbandonare la medesima per stipulare un nuovo contratto con una compagnia concorrente. Le compagnie del settore indicano questo fenomeno parlando del problema del cliente infedele.

Figura 2 - Metafora del cliente infedele. Copyright: Retentionscience.com (Fonte: https://www.retentionscience.com/blog/top-4-reasons-customers-churn-and-how-to-prevent-it/)

Nella narrazione della società di telefonia mobile, il cliente infedele, è assimilato ad un pesce rosso (Figura 2) che abbandona, tramite un guizzo, l'ampolla nella quale vive attualmente, la società di telefonia medesima, per tuffarsi in una nuova ampolla, una società di telefonia concorrente. Il cliente infedele mette in atto questa decisione in quanto ritiene più vantaggiosa o più adeguata alle sue esigenze di prezzo e/o di servizio l'offerta della compagnia di telefonia concorrente. Obiettivo della società di telefonia è prevedere quali clienti siano a rischio abbandono, vale a dire quali tra i loro clienti siano i clienti potenzialmente infedeli. Questo è uno dei problemi introdotti nel Capitolo 1.

La società di telefonia dispone di una base dati, dove per ogni UTENZA vengono misurate:
- PROVINCIA; provincia di residenza dell'intestatario del contratto di utenza,
- OPZIONE PLUS; se attivata assume valore 1, altrimenti ha valore 0,
- CHIAMATE GIORNO; numero di chiamate effettuate nella fascia oraria giorno,
- ADDEBITO GIORNO; addebito in valuta delle chiamate effettuate in fascia oraria giorno,
- CHIAMATE SERA; numero di chiamate effettuate nella fascia oraria sera,
- ADDEBITO SERA; addebito in valuta delle chiamate effettuate in fascia oraria sera,
- CHIAMATE NOTTE; numero di chiamate effettuate nella fascia oraria notte,
- ADDEBITO NOTTE; addebito in valuta delle chiamate effettuate in fascia oraria notte.

Inoltre, la società conosce quali utenze telefoniche in passato si siano comportate come clienti infedeli e quali invece si siano comportate fedelmente. Questo dato viene memorizzato nella base dati tramite

l'attributo denominato INFEDELE? Questo attributo assume il valore *si*, se il cliente ha abbandonato la compagnia telefonica, mentre assume il valore *no* in caso contrario.

Utenza	Provincia	Opzione Plus	Chiamate Giorno	Addebito Giorno	Chiamate Sera	Addebito Sera	Chiamate Notte	Addebito Notte	Infedele?
332-678921	VA	0	110	45.07	99	16.78	91	11.01	no
345-297735	MI	0	123	27.47	103	16.62	103	11.45	no
338-655112	PA	0	137	21.95	83	19.42	111	9.40	si
332-987324	PA	1	71	50.90	88	5.26	89	8.86	no
331-114854	NA	0	113	28.34	122	12.61	121	8.41	no
344-494765	GE	0	67	56.59	97	27.01	128	7.23	si
345-124356	FI	0	88	37.09	108	29.62	118	9.57	no
331-223224	PG	1	79	26.69	94	8.76	96	9.53	no
349-888634	MI	1	97	31.37	80	29.89	90	9.71	no
325-999456	CO	0	118	42.43	119	21.45	90	12.61	si

Figura 3 - Porzione della base dati della compagnia telefonica

Una porzione della base dati che mostra solo alcuni dei tanti attributi misurati in corrispondenza di ogni utenza e solo pochissime utenze tra tutte quelle che costituiscono il parco clienti della compagnia telefonica è riportata in Figura 3.

La società vuole prevedere per ogni utenza quale sia il valore dell'attributo INFEDELE? in modo tale da individuare quali siano i clienti potenzialmente infedeli. A tale fine pensa di utilizzare gli attributi; PROVINCIA, OPZIONE PLUS, CHIAMATE GIORNO, ADDEBITO GIORNO, CHIAMATE SERA, ADDEBITO SERA, CHIAMATE NOTTE, ADDEBITO NOTTE, insieme ad altri attributi memorizzati nella base di dati. La compagnia potrebbe pensare di seguire una strada molto lunga e complessa, che consiste nello scrivere un algoritmo che spieghi dettagliatamente come prevedere se un cliente sarà infedele o meno. Questo sarebbe particolarmente complesso e difficilmente porterebbe a risultati soddisfacenti.

L'alternativa è offerta dal Machine Learning che consente di apprendere un modello in grado di prevedere il valore dell'attributo denominato INFEDELE? Infatti, un tale modello le consentirebbe di rispondere ad un'esigenza primaria, vale a dire identificare i clienti potenzialmente infedeli. E' importante sottolineare che il modello di apprendimento avrà valore per la compagnia solo se in grado di prevedere quali tra i clienti attuali siano in procinto di abbandonare la compagnia telefonica, trasformandosi in clienti infedeli.

La società di telefonia può identificare i clienti potenzialmente infedeli formulando un problema di Machine Learning supervisionato. Infatti, questo tipo di problema assume siano disponibili diverse istanze della coppia ordinata $(\mathbf{X}, \mathbf{Y})$, che a fronte del valore misurato per un insieme di attributi $\mathbf{X}$, detti attributi di input, rende disponibile il valore di un insieme di attributi $\mathbf{Y}$, detti attributi di output.

Nel caso della società di telefonia, possiamo pensare ad $\mathbf{X}$ come agli attributi; PROVINCIA, OPZIONE PLUS, CHIAMATE GIORNO, ADDEBITO GIORNO, CHIAMATE SERA, ADDEBITO SERA, CHIAMATE NOTTE, ADDEBITO NOTTE, e ad $\mathbf{Y}$ come all'attributo INFEDELE? da prevedere.

In un problema di Machine Learning supervisionato si sfruttano le istanze (x, y) disponibili per la coppia ordinata $(\mathbf{X}, \mathbf{Y})$, dette esempi di apprendimento, per apprendere un modello che, data un record [29] $\mathbf{X} = \mathrm{x}$, vale a dire un particolare assegnamento x dei valori degli attributi di input $\mathbf{X}$, fornisca la previsione del valore y che verosimilmente verrà assunto dall'attributo di output Y (Figura 4)

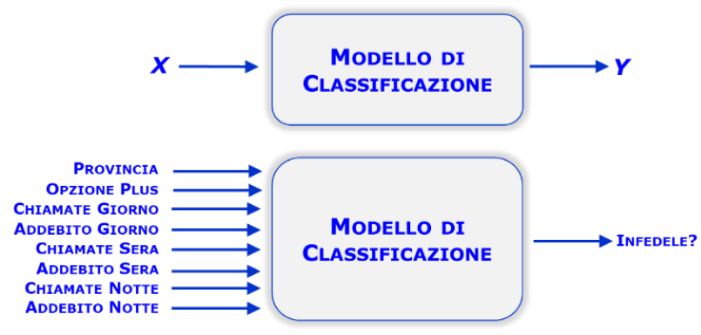


Figura 4 - Modello di classificazione.

Nel caso della compagnia di telefonia, il modello desiderato appartiene alla categoria dei modelli di classificazione, e nello specifico sarà un modello per la risoluzione di un *problema di classificazione binario*. Infatti, l'attributo di output INFEDELE?, che desideriamo prevedere, può assumere solo due valori mutuamente esclusivi, *si* o *no*.

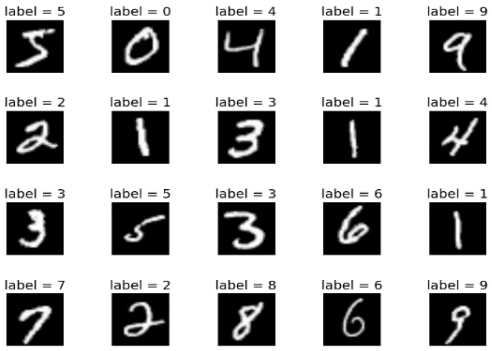


Figura 5 - Cifre da classificare. Copyright: Yann LeCun et al. (Fonte: https://medium.com/fenwicks/tutorial-1-mnist-the-hello-world-of-deep-learning-abd252c47709)

Al contrario, nel caso in cui l'attributo di output $\boldsymbol{Y}$ possa assumere più di due valori parleremo di *problema* o *modello di classificazione multi-classe*. Un'istanza significativa e frequentemente discussa è rappresentata dal problema di riconoscimento o classificazione delle cifre da 0 a 9 (Figura 5) a partire da un'immagine. In questo caso l'immagine disponibile costituisce l'attributo di input $\boldsymbol{X}$ a nostra disposizione, mentre l'attributo di output $\boldsymbol{Y}$, che abbiamo il compito di prevedere utilizzando l'attributo di input $\boldsymbol{X}$, può assumere dieci differenti valori, vale a dire tutti gli interi da 0 a 9.

[29] Con questo termine, di significato analogo a come è stato usato nei capitoli precedenti, intenderemo l'assegnazione di un valore specifico ad ogni attributo di input, anche nel caso in cui gli attributi di input siano di tipi differenti, numeri reali, interi, binari, immagini, testo, o altro ancora.

Infine, nel caso in cui siano presenti più attributi di output, ognuno dei quali possa assumere due valori, parleremo di problema o modello di classificazione multi-etichetta. Un problema multi-etichetta si ha quando è necessario decidere di quali argomenti, per esempio POLITICA, SPORT, SPETTACOLO, ECONOMIA, FINANZA, e CULTURA, tratti un testo in linguaggio naturale. Infatti, un testo in linguaggio naturale, per esempio un articolo di stampa, una pagina Web o un tweet, può trattare di nessuno, uno o contemporaneamente più argomenti tra quelli sopra menzionati (Figura 6).

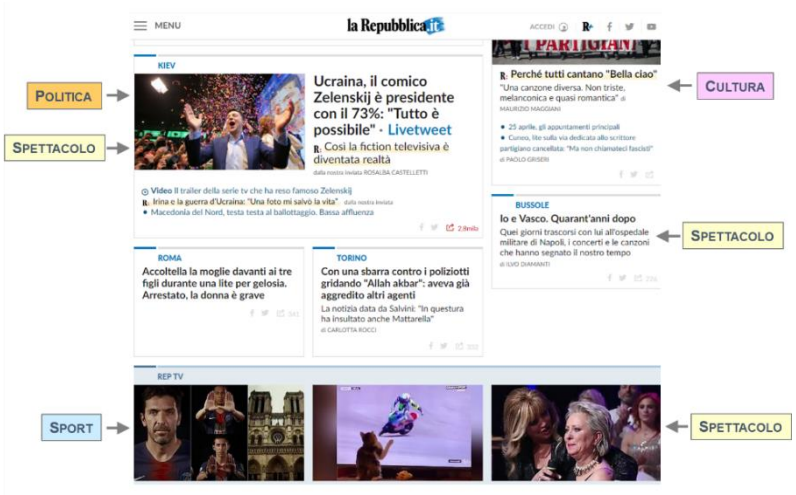


Figura 6 - Classificazione multi-etichetta di una pagina Web

Il problema di classificazione non è il solo problema di Machine Learning supervisionato. Infatti, nel caso in cui l'attributo di output $\mathbf{Y}$ assuma valori continui, si parla di problema o modello di regressione, caso che presentiamo sfruttando il seguente problema.

ID Abitazione	Mq calpestabili	Stanze da letto	Piani	Anno di costruzione	Prezzo (Euro)
1027	1,180	3	1	1955	221,900
2062	2,570	3	2	1951	538,000
1349	770	2	1	1933	180,000
834	1,960	4	1	1965	604,000
13	1,680	3	1	1987	510,000
1239	1,715	4	1	2001	257,500
349	1,060	3	2	1995	291,850
374	1,780	3	1	1963	229,500
28	1,890	3	1	1960	323,000
2945	3,560	3	2	2003	662,500

Figura 7 - Un frammento della base di dati della società immobiliare

PROBLEMA 2: una società immobiliare dispone di una base dati commerciale che contiene dati sulle abitazioni per le quali ha dai proprietari un mandato di vendita. Nello specifico, per ogni abitazione, identificata tramite il campo denominato ID ABITAZIONE, la base di dati riporta diverse caratteristiche strutturali e di costruzione dell'abitazione medesima quali; MQ CALPESTABILI, STANZE DA LETTO, PIANI ed ANNO

DI COSTRUZIONE, insieme alla stima del valore di mercato, PREZZO (EURO), fornita da un agente immobiliare. Una porzione delle base dati è mostrata in Figura 7.

In una tale situazione, la società immobiliare sarà interessata a conoscere il valore commerciale di altre abitazioni, abitazioni per le quali non ha mandato di vendita dai legittimi proprietari. Infatti, la società immobiliare potrebbe desiderare di conoscere il valore di un'abitazione per sfruttare l'informazione a livello commerciale o strategico. La società immobiliare potrebbe certo incaricare il proprio agente di procedere ad una stima del valore di mercato di tutte le abitazioni di diverse zone di una città. Questo però costerebbe all'agente immobiliare del tempo che si ritiene meglio impiegato nell'azione di visita, convincimento e vendita delle abitazioni ai potenziali acquirenti.

La società immobiliare sarà allora interessata a prevedere il valore di mercato di quelle abitazioni per le quali non dispone di una valutazione economica del proprio agente immobiliare. A tale fine decide di utilizzare alcuni o tutti gli attributi $\mathbf{X}$ che caratterizzano un'abitazione, per prevederne, tramite un modello di regressione, il valore di mercato espresso dall'attributo PREZZO (EURO) che in questo caso sarà l'attributo di output $\boldsymbol{Y}$.

Il *modello di regressione* è assimilabile al *modello di classificazione*, in termini concettuali e relativamente al **PROBLEMA 2** esso è rappresentabile come mostrato in Figura 8.

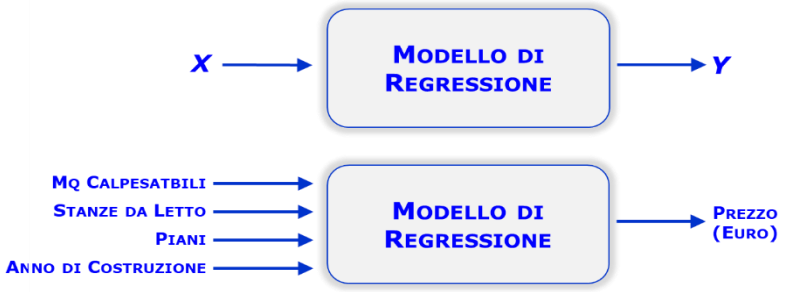


Figura 8 - Modello di regressione

2.2. Machine Learning non supervisionato

Il secondo tipo di Machine Learning, quello non supervisionato o descrittivo, viene spesso riferito con il termine di *knowledge discovery*, ed è un compito di apprendimento definito in modo molto meno preciso di quanto accade per il Machine Learning supervisionato. Infatti, mentre nel caso del Machine Learning supervisionato, disponendo di un attributo di output $\mathbf{Y}$ è possibile apprendere un modello che massimizzi una determinata misura di prestazione, questo non accade nel caso del Machine Learning non supervisionato, dove non disponiamo di istanze ordinate $(\mathbf{X}, \mathbf{Y})$ ma solo di istanze relative agli attributi $\mathbf{X}$.

Infatti, la bontà di un modello di classificazione viene usualmente valutata tramite la sua *accuratezza*, una misura di prestazione definita come la probabilità che il modello, se interrogato su un record $\mathbf{X}$, fornisca una previsione corretta del valore dell'attributo di output $\mathbf{Y}$. L'accuratezza, che abbiamo già

incontrato nel Capitolo 6 e nel Capitolo 8 sulla integrazione dati, per misurare l'efficacia di una tecnica di record linkage, viene spesso stimata tramite la frazione di istanze che il modello di classificazione prevede correttamente, rispetto al numero totale di istanze per le quali il modello ha fornito una previsione. In base a questa definizione è chiaro come maggiore sarà l'accuratezza del modello di classificazione, maggiori saranno il suo valore e la sua utilità. Analogamente, nel caso di un modello di regressione, la bontà è valutata come la propensione che il modello ha, quando interrogato su un record $\mathbf{X}$, a fornire una previsione $\widehat{\mathbf{Y}}$ prossima al valore effettivamente assunto dall'attributo di output $\mathbf{Y}$.

Nel Machine Learning non supervisionato non viene indicato quale debba essere il valore dell'attributo di output $\mathbf{Y}$ in corrispondenza di un determinato record $\mathbf{X}$ degli attributi di input. In tale situazione, il nostro compito è scoprire strutture e schemi nascosti nei dati, in modo tale da aumentare la conoscenza in nostro possesso e prendere di conseguenza decisioni più efficaci.

Il Machine Learning non supervisionato è indiscutibilmente una forma di apprendimento tipica degli esseri umani e degli animali. Inoltre, è una forma di apprendimento applicabile in modo più esteso di quanto non possa essere fatto con il Machine Learning supervisionato. Infatti, il Machine Learning non supervisionato non richiede che un esperto etichetti manualmente i dati a nostra disposizione, come invece accade nel caso del Machine Learning supervisionato.

Un'istanza tipica *di Machine Learning non supervisionato* è fornita dal PROBLEMA 3.

PROBLEMA 3: una società che fornisce energia elettrica ai privati di una grande città, decide di studiare il profilo di utilizzo energetico dei propri clienti al fine di progettare e proporre ai medesimi una tariffa personalizzata che al tempo stesso le garantisca un certo guadagno. La società dispone di una base di dati alimentata con frequenza oraria grazie a dispositivi smart meter installati presso le abitazioni cui fanno riferimento le utenze. Pertanto, per ogni utenza identificata tramite il campo ID UTENZA viene memorizzata la curva oraria di carico giornaliero, vale a dire la rilevazione del consumo orario di elettricità dell'utenza nell'arco delle 24 ore. La base di dati è alimentata con rilevazioni di consumo energetico orario per tutte le utenze della società di fornitura e per ogni giorno dell'anno. Una porzione della base dati viene illustrata in Figura 9, dove nello specifico vengono mostrati i consumi orari per le sole prime dieci ore della giornata.

In questa situazione la società che fornisce energia elettrica può utilizzare il Machine Learning non supervisionato, nello specifico può utilizzare la cluster analysis, per formare gruppi di utenze simili dal punto di vista del consumo energetico o per meglio dire gruppi simili dal punto di vista della curva oraria di carico giornaliero.

ID Utenza	Giorno	00:00-01:00	00:01-02:00	00:02-03:00	00:03-04:00	00:04-05:00	00:05-06:00	00:06-07:00	00:07-08:00	00:08-09:00	00:09-10:00
1345	27-Maggio-2019	0.15	0.05	0.00	0.00	0.00	0.25	0.75	1.40	1.50	0.90
4533	27-Maggio-2019	0.00	0.00	0.00	0.00	0.00	0.00	0.55	1.20	1.80	2.00
2938	27-Maggio-2019	0.35	0.30	0.10	0.00	0.00	0.00	0.00	0.00	0.90	1.60
9228	27-Maggio-2019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.20	1.70	2.10
4773	27-Maggio-2019	0.20	0.00	0.00	0.00	0.00	0.40	1.30	1.70	1.60	1.90

Figura 9 - Curve orarie di carico giornaliero, porzione della base dati dalle 00:00 alle 10:00

Altri esempi di cluster analysis si trovano nell'ambito delle reti sociali o biologiche (Figura 10), dove l'interesse risiede nell'identificare comunità, vale a dire insiemi di nodi che hanno somiglianza elevata e/o tendono ad interagire in modo sistematico tra loro.

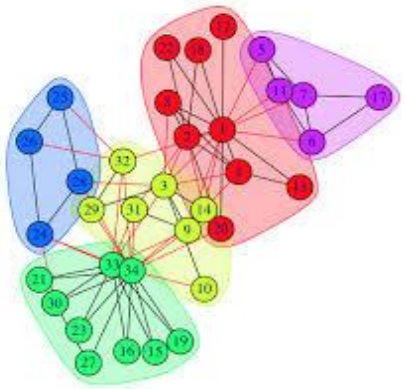


Figura 10 - Comunità di una rete. Copyright: stackoverflow.com (Fonte: https://stackoverflow.com/questions/24513339/how-to-find-measures-after-community-detection-in-igraph-r)

Un'altra istanza di Machine Learning non supervisionato è offerta dal PROBLEMA 4.

PROBLEMA 4: una catena di supermercati dispone di una base dati popolata tramite gli acquisti che giorno dopo giorno vengono registrati alle casse dei diversi punti vendita. La base di dati (Figura 11) contiene tutte le transazioni registrate in ogni GIORNO di esercizio. Una transazione, identificata in modo univoco dal codice TID, consiste nell'insieme di codici prodotto o della loro descrizione sintetica, insieme al dato di molteplicità per ogni cestello di acquisito che si presenti ad una delle casse della catena di supermercati. La prima transazione, univocamente identificata dal valore <TID=23312>, è stata registrata il <GIORNO=11-Apr-2019>, ed il cestello di acquisto con il quale il cliente si è presentato alla cassa conteneva; una confezione di <mayonese calvè>, una confezione di <pomodoro pachino>, 2 <baguette>, una confezione di <grana padano 200g.>, una confezione di <fusilli barilla 250g.> ed infine 1 confezione di biscotti <oro saiwa>.

TID	Giorno	prodotto 1	prodotto 2	prodotto 3	prodotto 4	prodotto 5	prodotto 6	prodotto 7
23312	11-Apr-19	mayonese calvè	pomodoro pachino	baguette	baguette	grana padano 200g.	fusilli barilla 250g.	oro saiwa
23313	11-Apr-19	coca cola 33cl.	coca cola 33cl.	coca cola 33cl.	popcorn san carlo	patatine san carlo		
23314	11-Apr-19	bonarda giunti	pannolini pampers	dentifricio colgate	spazz. dental plus	colluttorio listerine	conf.mele trent. 6x	
23315	11-Apr-19	acqua giuzza x 6	birra moretti x 4	grissini torino	doccia schiuma			
23316	11-Apr-19	spazz. dental plus	colluttorio mentadent	cotone conf. maxi	limoni rete da 10	coca cola 33cl.		
23317	11-Apr-19	banana 6x	mirtillo conf. grande	panna montata				
23318	11-Apr-19	coppa samm	bastoncini findus	bastoncini findus	barbera piacenza	parm. Regg. 250g.		
23319	11-Apr-19	zuppa del casale	asparagi conf 24x	mix rucoloso	birra moretti x 4			

Figura 11 - Transazioni di acquisto alla cassa di un supermercato.

La catena di supermercati sa che analizzando la base dati delle transazioni di acquisto, vale a dire analizzando il contenuto del cestello di acquisto dei propri clienti, può imparare a conoscerli meglio e quindi può aumentare i propri guadagni. Infatti, conoscere il comportamento di acquisto dei clienti

consente alla catena di supermercati di progettare e lanciare campagne pubblicitarie o operazioni di marketing basate su sconti. Inoltre, conoscere i propri clienti, dal punto di vista delle associazioni di prodotti fornisce dati importanti per gestire scorte, ordini e logistica della catena di supermercati. In questo caso, la catena di supermercati è interessata ad analizzare il contenuto del cestello di acquisto, e quindi ad applicare un particolare tipo di Machine Learning descrittivo che va sotto il nome di *market basket analysis* o di analisi delle associazioni, il cui obiettivo principale è scoprire regole associative o gruppi di prodotti che tendono a co-occorrere nel cestello di acquisto.

L'analisi delle associazioni è applicabile anche ad altri domini e problemi, come ad esempio quello biologico, quello della diagnosi medica, quello della formazione digitale, quello delle reti di comunicazione, per l'identificazione di software e applicazioni malware, ed infine per analizzare le valutazioni di gradimento di prodotti o servizi da parte di utenti e clienti.

Il Machine Learning descrittivo è popolato da ulteriori tipologie come ad esempio scoprire la struttura di reti regolatorie di geni, reti di comunicazione, reti sociali, reti di distribuzione, analisi delle immagini, compressione, e filtraggio collaborativo con specifico riferimento a problemi di raccomandazione.

2.3. Machine Learning per rinforzo

La terza ed ultima tipologia di Machine Learning è quella *per rinforzo*, anche se più propriamente nota con il termine di *Reinforcement Learning*. Si tratta di una tecnica nella quale un agente apprende come comportarsi all'interno di un ambiente a lui sconosciuto, eseguendo azioni e osservandone il relativo esito, tipicamente espresso in termini di un premio o di una punizione.

Per parecchi anni il Reinforcement Learning è stata la tipologia di Machine Learning meno studiata ed impiegata da ricercatori e addetti ai lavori. Tuttavia, negli ultimissimi anni è balzata agli onori della cronaca in forza di alcuni strabilianti risultati raggiunti grazie al suo impiego nell'ambito dei giochi. Un esempio su tutti è rappresentato dall'algoritmo AlphaGoZero, un algoritmo di Reinforcement Learning che ha dimostrato di aver acquisito capacità sovraumane nel gioco del GO, un gioco estremamente complesso che prevede 10^{761} possibili mosse (vedi in Figura 12 la copertina di un numero della rivista Nature su GO) Per comprenderne la complessità del gioco del GO, basti pensare che il gioco degli scacchi prevede 10^{123} mosse e che il numero totale di atomi nell'universo è ad oggi stimato esser pari a 10^{81}.

Figura 12 - Un numero della rivista Nature in cui si parla di GO. Copyright: Nature. (Fonte:https://www.nature.com/nature/volumes/529/issues/7587)

L'idea base di questo approccio è probabilmente la *legge dell'effetto* enunciata da E. Thorndike, uno psicologo statunitense, che nel 1911, in un esperimento svolto sugli animali, osservò come ogni comportamento seguito da una conseguenza positiva veniva più facilmente ripetuto di quanto non accedesse quando la conseguenza era negativa. Lo studio prevedeva di inserire un gatto in una scatola, all'interno della quale avrebbe dovuto risolvere un puzzle: in caso di successo il gatto avrebbe ottenuto la libertà e del cibo (premio), in caso contrario il gatto sarebbe rimasto all'interno della scatola (punizione). Thorndike notò come il gatto diventasse sempre più veloce nel compiere quelle azioni che portavano alla risoluzione del puzzle, così da poter uscire dalla gabbia nel minor tempo possibile e di conseguenza ricevere la propria ricompensa, il cibo.

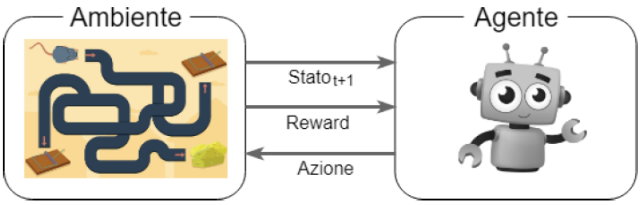


Figura 13 - Interazione tra Agente ed Ambiente

L'approccio del Reinforcement Learning (Figura 13) prevede sempre l'utilizzo di due componenti fondamentali: un agente ed un ambiente. Agente ed ambiente interagiscono determinando una sequenza di stati dell'ambiente, azioni attuate dall'agente e reward (premi o punizioni) che l'agente riceve come reazione alle azioni che applica all'ambiente quando questo si trova in un determinato stato.

Il Reinforcement Learning ha recentissimamente conosciuto successi anche in robotica, nel processamento automatico del linguaggio naturale (come la machine translation, generazione automatica di testi, e i sistemi di dialogo), così come nella computer vision, ed ancora in ambiti rilevanti quali la finanza, la sanità, l'educazione e l'Industria 4.0.

3. Modelli e Algoritmi

Il tipo di problema di Machine Learning che si desidera risolvere, insieme alla natura degli attributi di input $\mathbf{X}$ e di output $\mathbf{Y}$ disponibili, determinano quali modelli possano essere impiegati. In questo paragrafo, per ragioni di sintesi, presenteremo due soli modelli per la risoluzione di problemi di Machine Learning supervisionato, uno adatto alla risoluzione di un problema di classificazione, nello specifico il modello degli alberi di decisione, ed uno in grado di risolvere sia problemi di classificazione che problemi di regressione, vale a dire il modello delle reti di neuroni artificiali. Per quanto riguarda il Machine Learning non supervisionato presenteremo il *clustering partizionale*, con specifico riferimento all'algoritmo delle K-medie, ed il *clustering gerarchico agglomerativo* nelle sue principali formulazioni. Infine, data la complessità matematica delle diverse formulazioni di Reinforcement Learning, accenneremo solamente ai principali modelli ed algoritmi impiegati in questo ambito.

3.1. Machine Learning supervisionato

Consideriamo il PROBLEMA 1 nel quale la società di telefonia mobile desidera prevedere quali clienti siano potenzialmente infedeli. La società di telefonia decide di utilizzare i seguenti attributi; PROVINCIA, OPZIONE PLUS, CHIAMATE GIORNO, ADDEBITO GIORNO, CHIAMATE SERA, ADDEBITO SERA, CHIAMATE NOTTE, ADDEBITO NOTTE come attributi di input $\boldsymbol{X}$ per prevedere il valore dell'attributo INFEDELE?, che svolge il ruolo di attributo di output $\boldsymbol{Y}$.

ALBERI DI DECISIONE: un albero di decisione (Figura 14) risolve il PROBLEMA 1 conducendo una sequenza di test. Ogni test è associato ad un nodo dell'albero, e di norma ha ad oggetto, sebbene non sempre, uno degli attributi di input $\mathbf{X}$. L'esito di ogni test determina quale sarà il prossimo test al quale sottoporre il record $\mathbf{X}$ per prevedere il valore dell'attributo $\mathbf{Y}$. La sequenza di test origina dal nodo radice (CHIAMATE GIORNO), un nodo che ha solo archi uscenti, e prosegue con un nodo intermedio (ADDEBITO NOTTE, OPZIONE PLUS, CHIAMATE SERA), nodo che ha un arco entrante e più archi uscenti, o con un nodo foglia (SI, NO), che non ha archi uscenti, e nel quale avviene la previsione del valore dell'attributo di output $\mathbf{Y}$.

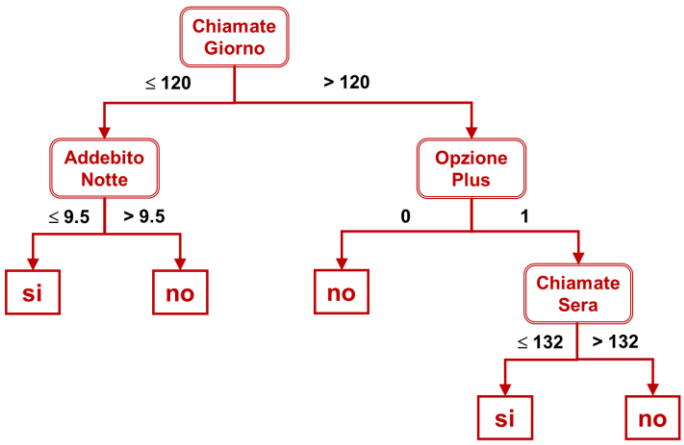


Figura 14 - Albero di decisione

L'albero di decisione mostrato in Figura 14 classifica un record $\mathbf{X}$ a partire dalla condizione di test associata al nodo radice (CHIAMATE GIORNO). Nello specifico viene effettuata la seguente condizione di test <Chiamate Giorno $\leq$ 120 ?>. Se la risposta al test è affermativa, allora la prossima condizione di test sarà quella associata al nodo intermedio ADDEBITO NOTTE, in caso contrario sarà quella associata al nodo intermedio OPZIONE PLUS. In questo secondo caso il test condotto sarà il seguente <Opzione Plus = 0 ?>, che, se riceverà risposta affermativa, porterà ad un nodo foglia associato al valore no, con conseguente previsione dell'attributo di output INFEDELE? = no.

Al contrario, nel caso in cui la risposta al test risultasse negativa, si proseguirebbe con un nuovo test associato al nodo intermedio CHIAMATE SERA. Il test da condurre sarebbe allora il seguente <Chiamate Sera $\leq$ 132 ?>, che porterà a due nodi foglia, uno associato al valore si ed uno associato al valore no, determinando le conseguenti previsioni per l'attributo di output $\mathbf{Y}$. In ogni caso la sequenza di test condotti terminerà non appena raggiunto un nodo foglia che fornirà la previsione del valore dell'attributo di output $\mathbf{Y}$, in questo caso l'attributo denominato INFEDELE?.

Un albero di decisione viene appreso dai dati disponibili tramite opportuni algoritmi. Infatti, esiste un numero esponenziale di alberi di decisione che possono essere costruiti a partire da un determinato insieme di attributi di input. Alcuni tra questi alberi saranno più accurati di altri, ma determinare quale tra tutti gli alberi possibili sia l'albero ottimale è computazionalmente proibitivo. D'altra parte, nel corso degli anni sono stati sviluppati algoritmi che, sebbene basati su una strategia miopica, che non riesce cioè a prevedere nel lungo termine, hanno mostrato di essere in grado di apprendere alberi con buoni livelli di accuratezza, impiegando quantità di tempo ragionevoli. Questa classe di algoritmi fa crescere l'albero prendendo decisioni localmente ottimali circa quale attributo di input utilizzare per l'operazione di partizionamento dei dati. Uno di questi algoritmi, l'algoritmo di Hunt, è la base di diversi algoritmi di apprendimento di alberi di decisione come ID3, C4.5, e CART.

L'algoritmo di induzione di Hunt fa crescere un albero di decisione in modo ricorsivo, partizionando i dati disponibili in modo da aumentarne via via il grado di purezza, vale a dire aumentando l'omogeneità delle istanze appartenenti alla stessa partizione dal punto di vista dell'attributo di output. Nello specifico, ogni nodo dell'albero ha associata una partizione dei dati. Se tale partizione contiene istanze appartenenti tutte alla stessa classe, per esempio tutte istanze tali che INFEDELE? = si, allora il nodo sarà un nodo foglia e verrà etichettato con la classe INFEDELE? = si. Al contrario, se la partizione contiene istanze che appartengono a più classi, viene formulata una condizione di test che partizionerà le istanze aumentando il grado di purezza complessivo. Per ogni possibile esito della condizione di test viene creato un nodo figlio del nodo corrente. Ogni nodo figlio si vedrà assegnare esattamente una delle partizioni di dati come determinato dall'applicazione della condizione di test. L'algoritmo viene applicato ricorsivamente ai nodi figlio, fino a che non sia più possibile formulare condizioni di test o fino a che qualche condizione di arresto dell'algoritmo di apprendimento non venga soddisfatta.

Ogni algoritmo di apprendimento di un albero di decisione deve affrontare le seguenti questioni:

– Come partizionare i dati? Ad ogni passo ricorsivo della crescita dell'albero di decisione è necessario scegliere una condizione di test, condizione che riguarda di norma, anche se non obbligatoriamente, un singolo attributo di input, e che partiziona i dati in sottoinsiemi più piccoli. L'implementazione di questo passo richiede di specificare una condizione di test nel caso di attributi di input di diversa

natura, e al tempo stesso di specificare una funzione di merito che consenta di comparare il livello di bontà di ogni condizione di test.

– Come arrestare la procedura di partizionamento dei dati? Una condizione di arresto è necessaria per terminare il processo di crescita dell'albero di decisione. Una strategia adottabile consiste nel continuare ad espandere un nodo fino a che accada uno dei seguenti eventi; i) tutti i dati appartengono alla stessa classe, ii) tutte i dati hanno gli stessi valori degli attributi di input $\mathbf{X}$, vale a dire sono indistinguibili dal punto di vista degli attributi di input $\mathbf{X}$. Sebbene entrambe le condizioni siano sufficienti per arrestare la fase di apprendimento di un albero di decisione, altri criteri possono essere impiegati per consentire alla procedura di apprendimento di terminare anticipatamente.

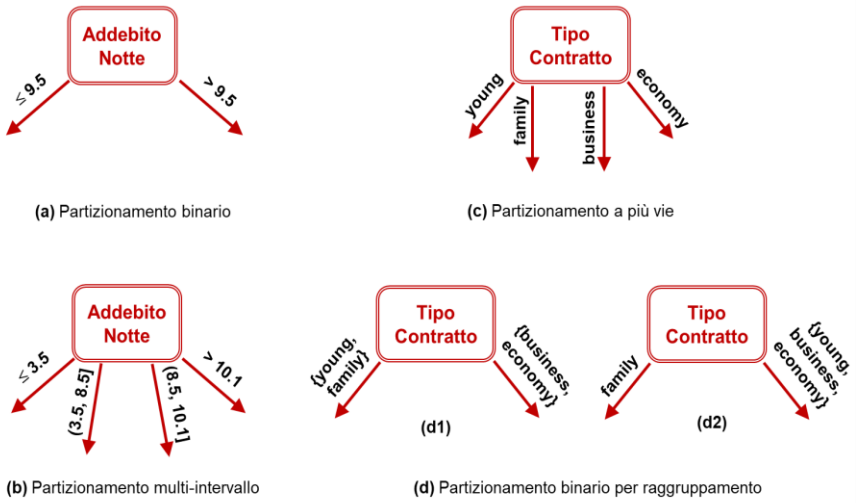


Figura 15 - Partizionamento per differenti tipi di attributo

L'albero di decisione in Figura 15, in corrispondenza del nodo associato all'attributo ADDEBITO NOTTE, effettua un partizionamento binario che viene riportato in Figura 15 (a). Tale partizionamento non è il solo possibile per attributi di tipo continuo come l'attributo ADDEBITO NOTTE. Un partizionamento alternativo è mostrato in Figura 15 (b), si tratta di un *partizionamento multi-intervallo*, caratterizzato da intervalli disgiunti, vale a dire senza sovrapposizioni, e tale da rappresentare l'intero dominio dei valori dell'attributo ADDEBITO NOTTE.

Il partizionamento dipende dal tipo di attributo considerato nella condizione di test. In Figura 15 (c) viene mostrato un partizionamento a più vie per l'attributo TIPO CONTRATTO che è di tipo nominale, vale a dire tale da non ammettere relazioni d'ordine tra i valori <young, family, business, economy>. Anche in questo caso sarebbero stati possibili altri partizionamenti, per esempio il partizionamento binario per raggruppamento come mostrato in Figura 15 (d1 e d2).

Ad ogni passo dell'algoritmo di apprendimento viene generato un insieme di condizioni di test ottenute a partire dagli attributi di input $\mathbf{X}$. Ogni attributo di input può essere coinvolto in più condizioni di test, dipendentemente dal proprio tipo e dallo specifico tipo di albero di decisione che si desidera apprendere. Inoltre, ogni condizione di test è relativa ad una partizione dei dati, e quando efficace ne determina un'ulteriore partizionamento. Ad ogni elemento dell'insieme di condizioni di test viene

associato il valore che una determinata funzione di merito assume in corrispondenza del partizionamento dati indotto dalla specifica condizione di test.

L'algoritmo di apprendimento utilizza i valori della funzione di merito per selezionare la condizione di test ottimale, vale a dire quell'elemento dell'insieme delle condizioni di test che ottiene il miglior valore della funzione di merito. La condizione di test ottimale determina il prossimo partizionamento dei dati. L'algoritmo di apprendimento procede in questo modo sino a che una qualche condizione di arresto non risulti soddisfatta. Tra le scelte possibili per la funzione di merito menzioniamo l'Entropia, il Gain Ratio, e l'indice di Gini.

Una seconda questione estremamente importante, che ogni algoritmo di apprendimento di un albero di decisione deve affrontare, è la condizione di arresto della procedura ricorsiva di partizionamento dei dati. La rilevanza del criterio di arresto è imputabile all'esistenza di un fenomeno particolarmente pericoloso noto come fenomeno di overfitting del modello di Machine Learning. Questo fenomeno, che presenteremo e discuteremo tra poco, consiste nell'adattamento esasperato dell'albero di decisione al fine di modellare ogni più piccola variazione dei dati di input, ignorando che un certo grado di variazione dei dati possa essere imputabile alla presenza di rumore e non alla presenza di un effettivo segnale.

Per controllare il fenomeno dell'overfitting vengono tipicamente impiegati criteri di *Pre-pruning* o di *Post-pruning*. Un criterio di Pre-pruning evita di generare un albero completo, imponendo una condizione restrittiva circa il livello minimo di miglioramento della purezza da ottenere per implementare un nuovo partizionamento. Al contrario, se si utilizza un criterio di Post-pruning, l'algoritmo apprende un albero completo che successivamente viene potato in base a diverse strategie.

Infine, è importante menzionare che gli alberi di decisione sono interpretabili e consentono di spiegare la natura dell'inferenza fatta su un record. Per tale ragione gli alberi di decisione rappresentano spesso la prima opzione per motivare all'esperto di dominio la natura della classificazione effettuata su uno specifico record. Riprenderemo questo importante aspetto degli algoritmi di Machine learning nel Capitolo 15.

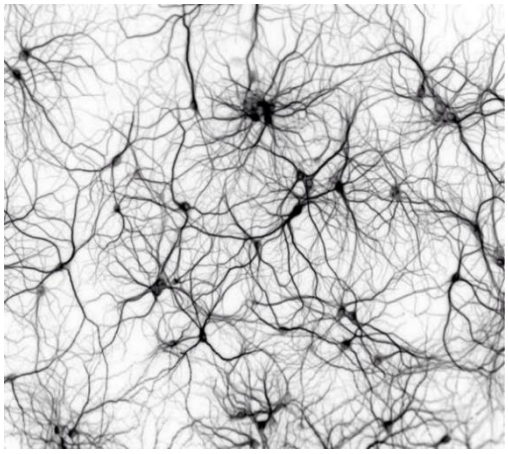


Figura 16 - Neuroni biologici. Copyright: winnerinstitute.eu. (Fonte: http://www.winnerinstitute.eu/2019/07/01/le-motivazioni-biologiche-delle-reti-neurali-artificiali/)

RETI DI NEURONI ARTIFICIALI: appartengono alla classe dei *modelli connessionisti*, facenti capo allo psicologo Canadese Donald Hebb, e si ispirano al cervello umano che consiste di circa 10^{10} a 10^{11} cellule nervose chiamate neuroni (Figura 16). I neuroni sono tra loro interconnessi per mezzo di fibre chiamate assoni che trasmettono impulsi nervosi da un neurone all'altro in presenza di stimoli. Il punto di contatto tra una dendrite, che consente ad un neurone di ricevere impulsi nevosi, ed un *assone* è detto *sinapsi*. Gli elementi principali del *neurone biologico* sono; il *corpo*, che ne implementa tutte le funzioni logiche, l'*assone* che trasmette impulsi nervosi ad altri neuroni, la *dendrite* che consente al neurone di ricevere impulsi nervosi da altri neuroni, la *sinapse* che costituisce il ramo terminale dell'*assone* e trasmette impulsi nervosi provenienti dal neurone alle dendriti di altri neuroni.

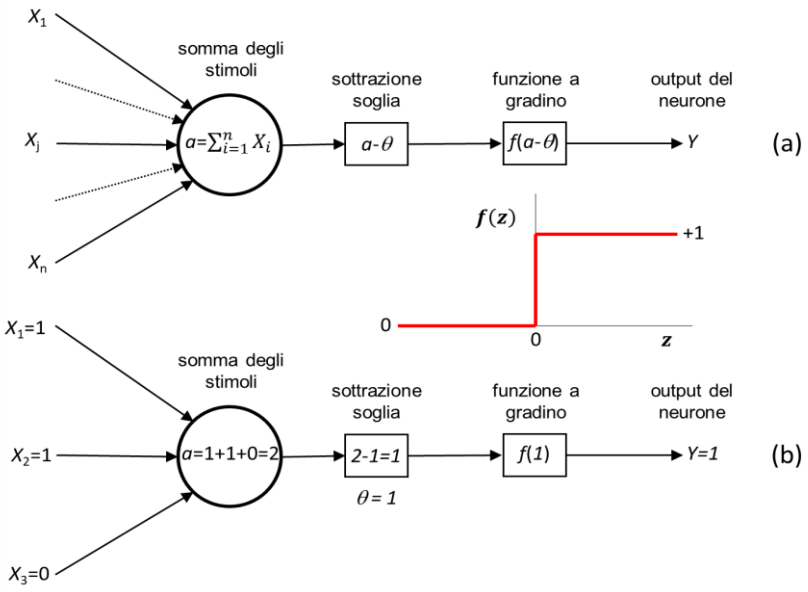


Figura 17 - Modello del neurone artificiale di McCulloch e Pitts

Il primo *modello formale di neurone artificiale* fu proposto da Warren McCulloch e Walter Pitts nel 1943 (Figura 17 (a)). Si trattava di un modello elementare nel quale i due ricercatori assumevano che il *neurone artificiale* venisse stimolato tramite *n* attributi binari di input $\boldsymbol{X} = (X_1, \dots, X_n)$. Il neurone computava il valore $\boldsymbol{a}$, vale a dire la *somma degli stimoli* che esso riceveva in input dagli *n* attributi binari $\boldsymbol{X}$. A tale somma veniva successivamente sottratto il valore di un parametro θ detto *soglia* del neurone artificiale. Il risultato di tale differenza, ($\boldsymbol{a}$-θ) veniva fornito come argomento alla *funzione di attivazione* $f(\bullet)$ che restituiva il valore dell'attributo di output $\boldsymbol{Y}$ anch'esso assunto binario.

In Figura 17 (b) viene mostrato un esempio di computazione del neurone artificiale di McCulloch e Pitts. In particolare, si considera il caso di n=3, nel caso in cui gli attributi binari in input $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ assumono rispettivamente valore 1, 1 e 0. Il valore del parametro soglia θ del neurone viene posto pari ad 1. Il neurone calcola la somma degli stimoli X_1, X_2, X_3 che risulta pari a 2, alla quale viene sottratto il valore 1 della soglia, portando a fornire in ingresso alla funzione di attivazione il valore 1. La funzione di attivazione restituisce il valore 1 come previsione dell'attributo di output $\boldsymbol{Y}$.

Il modello di McCulloch e Pitts ricevette enorme attenzione, immaginando che di li a pochi anni sarebbero stati disponibili modelli di neuroni artificiali in grado di apprendere tutti i compiti svolti dal cervello umano. Questo entusiasmo scemò non appena si realizzò che il neurone artificiale, in grado di modificare il proprio comportamento solo tramite la variazione del valore della soglia θ, offriva limitatissime capacità di adattamento ed apprendimento, che non gli consentivano di apprendere funzioni non linearmente separabili, al contrario di quanto accadeva al cervello umano. Infine, il neurone artificiale di McCulloch e Pitts tratta solo attributi di input binari, mentre in diversi problemi reali abbiamo a che fare con attributi continui.

Le critiche mosse al neurone artificiale di McCulloch e Pitts portano nel 1958 a proporre il neurone artificiale di Rosenblatt (Figura 18). Questo modello differiva dal modello di McCulloch e *Pitts* in quanto prevedeva che gli attributi di input fossero variabili reali, e che la loro combinazione avvenisse assegnando ad ogni X_j un diverso livello di importanza rappresentato dal parametro w_j detto *peso della connessione* tra l'attributo X_j ed il neurone artificiale.

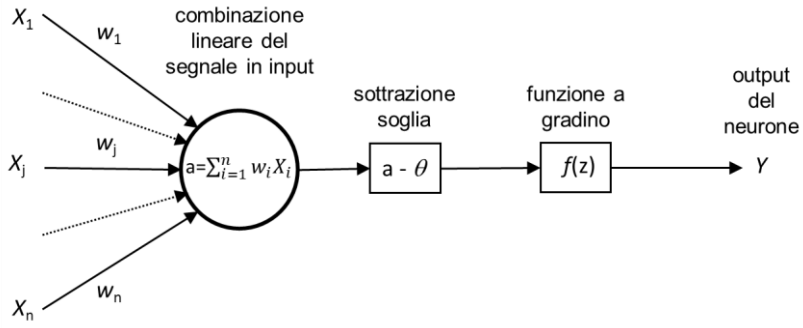


Figura 18 - Modello del neurone artificiale di Rosenblatt

Il neurone artificiale di Rosenblatt, a differenza di quello di McCulloch e Pitts, disponeva di un algoritmo di apprendimento a partire dai dati. Tuttavia, esso non superava il limite del neurone artificiale di McCulloch e Pitts in merito all'apprendimento di funzioni non linearmente separabili, come evidenziato da Papert e Minsky. Questo limite portò inevitabilmente al primo inverno dei modelli connessionisti, inverno che durò per circa 20 anni.

Infatti, nel 1986 Rumelhart, Hinton e Williams, avviarono la seconda primavera del connessionismo, formulando il modello del Percettrone Multi-Strato (Multi-Layer Perceptron) e rendendo disponibile il relativo algoritmo di apprendimento noto come *algoritmo di retro-propagazione* (back-propagation algorithm). Il Percettrone Multi-Strato era in grado di apprendere funzioni non linearmente separabili a partire dai dati, e l'algoritmo di retro-propagazione ne automatizzava tale apprendimento.

Se la società immobiliare del **PROBLEMA 2** decidesse di utilizzare solo l'attributo MQ CALPESTABILI ($\boldsymbol{X}$) per prevedere il valore del PREZZO (EURO) ($\boldsymbol{Y}$), allora potrebbe utilizzare un Percettrone Multi-Strato con uno strato di neuroni nascosti e due strati di pesi (Figura 19). Il percettrone consiste di uno strato di cinque *neuroni nascosti*, indicizzati con h_1, h_2, h_3, h_4, h_5. L'input $\boldsymbol{X}$ è collegato con i neuroni dello strato nascosto tramite un arco orientato al quale è associato un peso (pesi strato nascosto 1). Tutti i neuroni

dello strato nascosto sono collegati con l'output $\boldsymbol{Y}$ per mezzo di archi orientati, ognuno dei quali ha associato un peso (pesi strato nascosto 2).

La previsione del valore dell'attributo PREZZO (EURO) ($\boldsymbol{Y}$) viene ottenuta come segue; il valore dell'attributo MQ CALPESTABILI ($\boldsymbol{X}$) viene trasmesso ai neuroni dello strato nascosto h_1, h_2, h_3, h_4, h_5 i quali utilizzando i pesi dello strato nascosto 1, computano valori di attivazione e conseguenti valori delle funzioni di trasferimento. Questi valori vengono utilizzati, insieme ai pesi dello strato nascosto 2, per prevedere il valore dell'output $\boldsymbol{Y}$ (PREZZO (EURO)).

Il Percettrone Multi-Strato può consistere di due (Figura 19) o più strati di pesi nascosti. Inoltre, è possibile scegliere diverse funzioni di attivazione per i neuroni nascosti. Infatti, nel Percettrone Multi-Strato la funzione di attivazione associata ai neuroni nascosti non è la funzione a gradino ma tipicamente è la funzione tangente iperbolica o la funzione logistica. Il neurone di output può avere associata una delle due funzioni di attivazione menzionate o una funzione di attivazione lineare; il numero di strati di pesi nascosti, di neuroni per ogni stato nascosto e la scelta delle funzioni di attivazione definiscono l'architettura del Percettrone Multi-Strato.

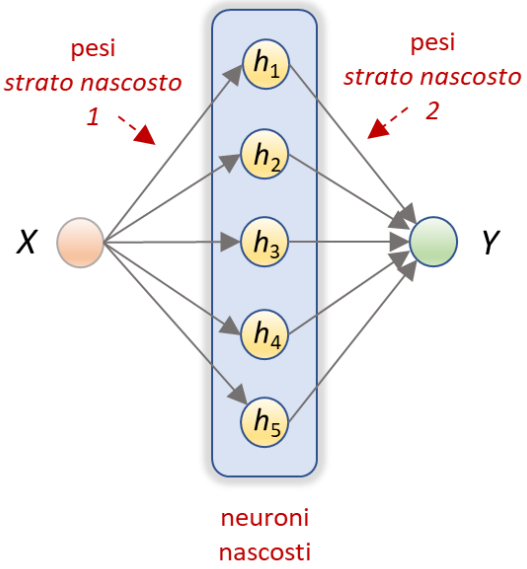


Figura 19 - Percettrone Multi-Strato

La rete è costituita da 64 attributi di input, 4 neuroni nel primo strato, 6 neuroni nel secondo e nel terzo strato, 3 neuroni nel quarto strato, 4 neuroni nel quinto strato e 10 neuroni di output, ognuno associato ad una delle dieci cifre 0, …, 9.

Negli ultimi dieci anni, l'enorme disponibilità di dati, il basso costo di unità di calcolo come le Central processing unit (CPU), le Graphics Processing Unit (GPU) e le Tensor Processing unit (TPU) e la disponibilità di software gratuito che implementa le reti di neuroni artificiali rendendone la definizione dell'architettura e l'apprendimento estremamente accessibili, hanno portato ad un incremento esponenziale del numero di applicazioni basate su questi modelli connessionisti che sono oggi meglio conosciuti come modelli di *Deep Learning*.

I fattori appena menzionati hanno consentito di addestrare istanze di Percettrone Multi-Strato estremamente flessibili, vale a dire con un'architettura con molti strati nascosti. In Figura 20 viene mostrata l'architettura di un Percettrone Multi-Strato per risolvere il problema del riconoscimento delle cifre da 0 a 9 (Figura 5), noto anche come problema MNIST. La cifra da riconoscere è descritta tramite un'immagine di 8x8 pixel, ognuno dei quali può assumere il valore 0 (bianco) o 1 (nero). I dieci neuroni di output sono associati alle dieci cifre da riconoscere, per cui ogni attributo Y_i (i=0, 1, …, 9) assume valori nell'intervallo continuo [0,1], modellando la probabilità che l'immagine fornita in input $X_1, X_2, \ldots, X_{64}$ sia relativa alla cifra i. E' utile osservare che mentre il numero di strati di neuroni nascosti è pari a cinque, il numero di strati nascosti di pesi è pari a 6.

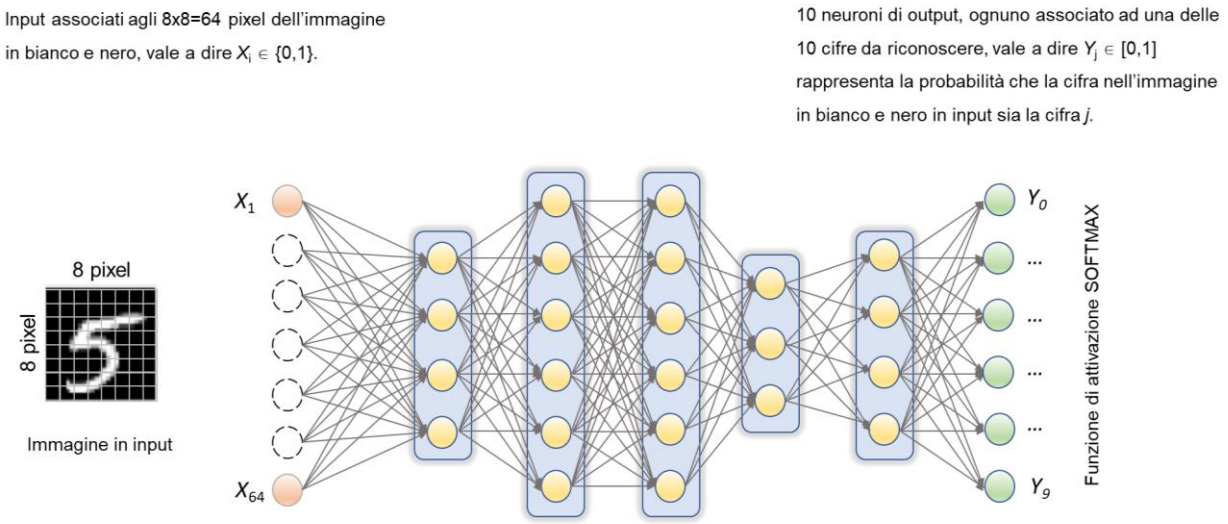


Figura 20 - Percettrone Multi-Strato per il riconoscimento delle cifre del problema MNIST.

Nel caso specifico della elaborazione delle immagini, il Deep Learning ha concepito modelli molto più efficaci ed efficienti del Percettrone Multi-Strato, come ad esempio le *reti neurali convoluzionali*, note come *Convolutional Neural Networks* (*CNNs)*, le *Generative Adversarial Networks* (*GANs*) e le *Capsule Neural Networks* (*CapsNet*).

Una rete convoluzionale (Figura 21) è basata sull'operazione di convoluzione, nello specifico la convoluzione degli elementi di una matrice o, nel caso di più di due dimensioni, di un *tensore*, una matrice a più di due dimensioni.

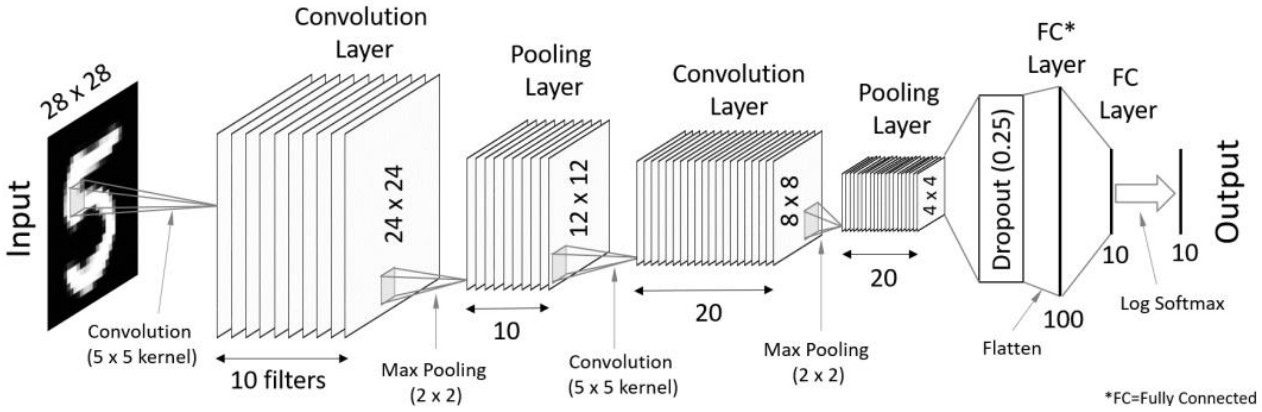


Figura 21 - Rete Neurale Convoluzionale per il riconoscimento delle cifre del problema MNIST. Copyright: codetolight.wordpress.com. (Fonte:https://codetolight.wordpress.com/2017/11/29/getting-started-with-pytorch-for-deep-learning-part-3-neural-network-basics/)

Non è scopo di questa trattazione fornire una definizione formale di convoluzione; è utile però sottolineare che l'operazione di convoluzione combina tra loro elementi di un tensore e i *pesi* che collegano un *filtro*, detto in termini tecnici *kernel*, con gli elementi di un altro tensore. Caratteristica fondamentale di tale operatore è la sua località, che consente di ridurre sensibilmente il numero di pesi da addestrare, riducendo in generale il tempo di addestramento e contribuendo a limitare il rischio di *overfitting* che caratterizza le reti di neuroni artificiali e in generale tutti i modelli di Machine Learning.

In questo caso l'immagine in input è costituita da 28x28 pixel. Dall'immagine in input allo strato di output, dove avviene il riconoscimento delle cifre, si alternano strati convoluzionali e strati di pooling che sono responsabili di calcolo del massimo del valore degli elementi dei vari tensori di cui si compone la rete neurale. In particolare, il primo strato convoluzionale è costituito da 10 filtri, ognuno di dimensione pari a 24x24. Agli elementi di questo strato convoluzionale viene applicata l'operazione di Max Pooling che determina il valore degli elementi dello strato di Pooling (Pooling Layer). Spostandoci più a destra, troviamo un ulteriore strato convoluzionale, un successivo strato di pooling ed infine uno strato densamente connesso che termina nei dieci neuroni di output, ognuno associato, come nel caso del Percettrone Multi-Strato, alle dieci cifre da riconoscere.

Il Deep Learning non si esaurisce con i modelli che abbiamo presentato, è popolato da molti altri modelli con supervisione particolarmente interessanti e potenti come i modelli delle *reti ricorrenti*, delle *reti ricorsive* e delle *reti Long Short Term Memory*. Questi modelli sono stati sviluppati per trattare problemi sequenziali come ad esempio: il riconoscimento del parlato, l'analisi sintattica del testo in linguaggio naturale, la previsione dell'andamento dei corsi azionari, la previsione dell'evoluzione di sistemi dinamici ed altri problemi complessi nei quali la sequenzialità dei dati ricopre un ruolo importante.

Un'importante corrente di pensiero, che va affermandosi negli ultimissimi anni sotto il nome di *explainable artificial intelligence*, critica aspramente i modelli di Deep Learning a causa della loro mancanza di interpretabilità, contrariamente a quanto accade invece nel caso degli alberi decisionali; parleremo di interpretabilità, come già detto, nel capitolo 15 sull'etica dei dati. Le critiche non hanno

tuttavia rallentato la diffusione dello studio e soprattutto dell'applicazione del Deep Learning, campo dove si stanno ottenendo prestazioni allo stato dell'arte in molti domini applicativi.

Alberi di decisione e reti di neuroni artificiali non sono i soli modelli di Machine Learning supervisionato disponibili. Infatti, nel corso degli ultimi 30 anni diversi modelli sono stati studiati ed applicati. Gli anni novanta furono un periodo di grande interesse per la classe dei modelli noti come *Support Vector Machines* (*SVM)*; questo accadde grazie all'importante impianto teorico reso disponibile da V. Vapnik, matematico e statistico sovietico che fu l'ideatore delle SVM, impianto teorico che forniva un limite superiore probabilistico al livello di errore che un modello, addestrato su un insieme di dati, avrebbe sofferto quando chiamato a fornire previsioni.

Gli anni duemila hanno visto l'ascesa dei modelli Bayesiani, con particolare riferimento al modello che va sotto il nome di *naive Bayes* (NB). Il successo di questo modello è probabilmente da imputare a due aspetti rilevanti, vale a dire la capacità del modello di apprendere da grandi quantità di dati ed in presenza di un elevatissimo numero di attributi di input, e gli eccellenti livelli di prestazione raggiunti in diversi domini applicativi. Altri modelli che hanno ricevuto l'attenzione della comunità scientifica sono il modello di *regressione logistica*, per la risoluzione di problemi di classificazione binaria, i modelli denominati *Random Forests* che sfruttano il concetto di *apprendimento d'insieme* (ensemble learning). Recentemente ha ricevuto attenzione anche il modello denominato gradient boosting.

3.2 Machine Learning non supervisionato

Nel **PROBLEMA 3** una società che fornisce energia elettrica desidera progettare e proporre tariffe personalizzate ai propri clienti. La personalizzazione è basata sulla curva oraria di carico giornaliero che consente di distinguere la modalità con la quale il cliente utilizza l'energia elettrica messa a sua disposizione. L'azienda decide di formare diversi gruppi o cluster di curve orarie di carico giornaliero; inoltre, decide di formare gruppi disgiunti, in modo che ogni curva oraria di carico giornaliero appartenga ad uno solo dei gruppi formati. L'azienda ha l'obiettivo di formare questo insieme di gruppi in modo tale che curve orarie di carico giornaliero simili vengano assegnate allo stesso gruppo, e curve orarie di carico giornaliero tra loro differenti vengano assegnate a gruppi differenti (Figura 22).

Figura 22 - Gruppi di curve orarie di carico

I tre gruppi, denominati Gruppo 1, Gruppo 2 e Gruppo 3, sono costituiti ciascuno da quattro istanze di curve orarie di carico giornaliero; ogni gruppo contiene istanze di curva oraria di carico giornaliero tra loro molto simili, le istanze di curva oraria di carico giornaliero tra loro differenti sono assegnate a gruppi differenti. In altri termini, le istanze del Gruppo 1 hanno un andamento estremamente simile tra loro, la stessa cosa accade per le istanze associate al Gruppo 2 e le istanze associate al Gruppo 3. Al tempo stesso, le istanze assegnate al Gruppo 1 sono differenti dalle istanze assegnate al Gruppo 2 e dalle istanze assegnate al Gruppo 3. La stessa considerazione può essere fatta comparando le istanze associate al Gruppo 2 alle istanze assegnate al Gruppo 3. In sintesi, istanze tra loro simili devono appartenere allo stesso gruppo ed istanze tra loro differenti devono appartenere a differenti gruppi.

ALGORITMO DELLE K-MEDIE: è probabilmente l'algoritmo più elementare di clustering e certamente quello maggiormente applicato. L'algoritmo si basa sul concetto di *centroide di un gruppo* (*cluster*), vale a dire un elemento prototipo deputato a rappresentare o sintetizzare *tutte le istanze* che vengono assegnate a quel cluster, e che mai o quasi mai corrisponde a un record della base di dati a nostra disposizione.

L'algoritmo delle *K-medie* richiede di specificare il numero *K* di gruppi che desideriamo formare. La fase di apprendimento procede iterativamente, ad ogni passo l'algoritmo assegna ogni record della base di dati a esattamente uno dei *K gruppi*, nello specifico ogni record viene assegnato al gruppo con *centroide*

ad essa più *prossimo*. Una volta che tutte le istanze siano state assegnate, l'algoritmo computa il *centroide* di ogni gruppo. L'algoritmo procede iterativamente, assegnando istanze ai gruppi e ricomputandone i *centroidi*, sino a che ogni record della base dati sia stato assegnato allo stesso gruppo per due iterazioni consecutive. La soluzione è costituita da *K gruppi* di istanze, ed ogni gruppo può contenere un numero diverso di istanze della base di dati. Un gruppo, e quindi le istanze a lui assegnate, viene sintetizzato tramite il relativo *centroide*.

L'algoritmo delle K-medie si basa sul concetto di prossimità tra istanze e centroidi dei gruppi. La letteratura rende disponibili diverse misure di prossimità, per esempio la misura di distanza Euclidea ben si adatterebbe al PROBLEMA 3, dato che gli attributi assumono valori reali, anche se ulteriori considerazioni potrebbero essere fatte per il caso in questione che interessa una serie temporale. Altre misure di prossimità sono proposte ed utilizzate nella letteratura specializzata. La coseno-similarità è particolarmente adatta quando si desiderino raggruppare documenti rappresentati tramite conteggio dell'occorrenza delle parole. Altre misure tipicamente utilizzate dall'algoritmo delle K-medie sono la distanza Manhattan e la misura di Jaccard; non approfondiremo nel seguito il tema delle misure.

L'applicazione dell'algoritmo delle *K-medie* presenta diverse criticità;
- la *scelta dei centroidi iniziali*, scelta che può influenzare anche pesantemente la velocità di convergenza dell'algoritmo e la qualità della soluzione restituita,
- la gestione dei *gruppi vuoti*, vale a dire gruppi ai quali non è stata assegnata alcun record della base di dati
- la gestione delle *osservazioni anomale*, osservazioni derivanti da errori di registrazione del record, da errori di imputazione, o da altri tipi di errori.

L'algoritmo delle *K-medie* offre vantaggi e soffre svantaggi. Il vantaggio principali sono rappresentati dalla sua semplicità e dal fatto che esso è applicabile ad una vasta tipologia di dati ed attributi; inoltre, è necessario sottolineare che l'algoritmo delle *K-medie* è particolarmente efficiente, sebbene di norma sia necessario effettuarne più esecuzioni.

Tra gli svantaggi troviamo il fatto che non è applicabile ad ogni tipo di dato e che fatica a indentificare correttamente gruppi non sferici. Inoltre, tratta in modo non soddisfacente gruppi con numerosità differente, vale a dire ai quali corrispondono numeri molto differenti di istanze, o gruppi con *densità* differente, vale a dire gruppi dove il livello di prossimità tra istanze dello stesso gruppo varia molto da gruppo a gruppo. Nello specifico, l'algoritmo potrebbe frammentare in più gruppi istanze che appartengono "naturalmente" allo stesso gruppo, oppure potrebbe assegnare allo stesso gruppo istanze che appartengono naturalmente a gruppi differenti. È importante che queste limitazioni vengano prese nella dovuta considerazione ogni volta che si decida di utilizzare il risultato dell'algoritmo delle *K-medie* per trarre conclusioni o per prendere decisioni.

ALGORITMO GERARCHICO AGGLOMERATIVO: si tratta dell'algoritmo gerarchico maggiormente diffuso la cui esecuzione porta ad ottenere un *raggruppamento gerarchico agglomerativo* che di norma viene rappresentato graficamente tramite un albero binario detto *dendrogramma*. Ogni record della base dati è associato ad uno ed esattamente uno dei *nodi foglia* dell'albero. Riferendoci al **PROBLEMA 3** e

considerando per semplicità espositiva solo le otto istanze con ID UTENZA uguale a 23318, 43982, 33318, 87362, 11212, 45721, 22671 e 98364, otteniamo il dendrogramma in Figura 23.

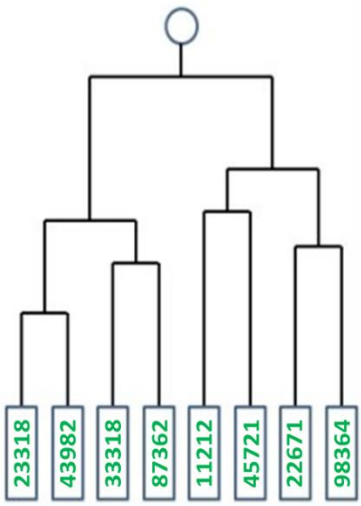


Figura 23 – Dendrogramma

Il *dendrogramma* descrive come avviene l'aggregazione delle istanze da raggruppare. Al primo passo, l'*algoritmo gerarchico agglomerativo*, computa la distanza esistente tra ogni coppia di istanze che è possibile formare a partire dalle otto istanze. La coppia di istanze cui compete il valore minimo della distanza viene raggruppata a formare un unico *cluster*. Nel caso della Figura 23, le prime due istanze ad essere raggruppate hanno ID UTENZA uguale a 23318 e a 43982. Il fatto che le due istanze siano quelle cui corrisponde il minimo valore di distanza lo si evince dal fatto che i *nodi foglia* corrispondenti a tali istanze sono i primi a congiungersi tramite un segmento orizzontale come si può notare partendo dai *nodi foglia* e risalendo verso il *nodo radice* rappresentato da un cerchio. Al secondo passo l'algoritmo calcola la distanza tra i sette elementi esistenti dopo il primo passo, vale a dire le sei istanze ed il cluster formato al primo passo. Anche in questo caso l'algoritmo decide quale raggruppamento effettuare secondo il criterio della distanza minima. Nel nostro caso il nuovo raggruppamento consiste delle istanze con ID UTENZA uguale a 33318 e a 87362.

L'*algoritmo gerarchico agglomerativo* procede fino a che tutte le istanze disponibili non siano raggruppate in un unico cluster, che viene rappresentato dal nodo radice del dendrogramma. La lettura del dedrogramma consente di decidere quale sia il livello di aggregazione ottimale, vale a dire in quanti raggruppamenti sia opportuno partizionare le istanze della base di dati. Diversi criteri sono stati presentati, discussi, e comparati nella letteratura specializzata per determinare il numero ottimale di gruppi da formare utilizzando il dendrogramma. Tuttavia, scegliere il numero ottimale di gruppi continua ad essere un problema complesso.

Un'importante differenza tra l'*algoritmo delle K-medie* e l'*algoritmo gerarchico agglomerativo* è rappresentata dal fatto che:

- nel primo caso la scelta del numero di gruppi da formare deve essere effettuata prima che l'algoritmo venga eseguito, mentre

- nel secondo caso l'algoritmo viene eseguito indipendentemente dal numero di gruppi che si intendono formare, numero che viene determinato sfruttando il *dendrogramma* e un criterio di ottimalità scelto tra quelli disponibili nella letteratura specializzata.

Anche l'*algoritmo gerarchico agglomerativo*, come l'*algoritmo delle K-medie* e tutti gli *algoritmi di clustering*, si basa su una misura di distanza/prossimità per decidere come formare gruppi di istanze. Pertanto, le misure di distanza/prossimità utilizzabili dall'*algoritmo delle K-medie* sono utilizzabili anche dall'*algoritmo gerarchico agglomerativo*.

A proposito di distanza/prossimità, nell'esposizione del funzionamento dell'*algoritmo gerarchico agglomerativo* potrebbe essere sfuggito un aspetto fondamentale, vale a dire il fatto che mentre è chiaro come calcolare la distanza/prossimità tra due istanze della base di dati, non è altrettanto chiaro come calcolare la distanza/prossimità tra due gruppi di istanze. Scopriamo l'importanza del concetto di distanza, che abbiamo introdotto la prima volta nel Caitolo 5, anche nel campo del Machine learning. Nel contesto degli algoritmi che stiamo discutendo, sono di norma possibili tre opzioni; la *distanza/prossimità single linkage*, la *distanza/prossimità complete linkage* e la *distanza/prossimità average linkage*.

Se consideriamo istanze dove vengono considerati, solo due attributi continui, PESO ed ALTEZZA, avremo quanto segue. La distanza/prossimità single linkage tra due gruppi di istanze è ottenuta calcolando la distanza/prossimità tra tutte le coppie di istanze appartenenti a gruppi differenti e selezionandone il valore minimo (Figura 24 (a)). La distanza/prossimità complete linkage tra due gruppi di istanze è ottenuta calcolando la distanza/prossimità tra tutte le coppie di istanze appartenenti a gruppi differenti e selezionandone il valore massimo (Figura 24 (b)). Infine, la distanza/prossimità average linkage tra due gruppi di istanze è ottenuta come valore medio della distanza/prossimità tra tutte le coppie di istanze appartenenti a gruppi differenti (Figura 24 (c)).

L'*algoritmo gerarchico agglomerativo* viene tipicamente impiegato nelle situazioni in cui la produzione di una gerarchia delle istanze risulta di particolare interesse, per esempio nei casi in cui sia utile ottenere una tassonomia delle istanze costituite da record di una base di dati. Alcuni studi hanno mostrato come questo algoritmo tenda ad ottenere soluzioni di migliore qualità rispetto ad altri algoritmi, incluso l'algoritmo delle K-medie. Purtroppo, l'algoritmo è costoso in termini di memoria e tempi di computazione; infine, l'algoritmo soffre nel caso di istanze caratterizzate da alti livelli di errore o istanze descritte tramite molti attributi, come accade nel caso del testo in linguaggio naturale.

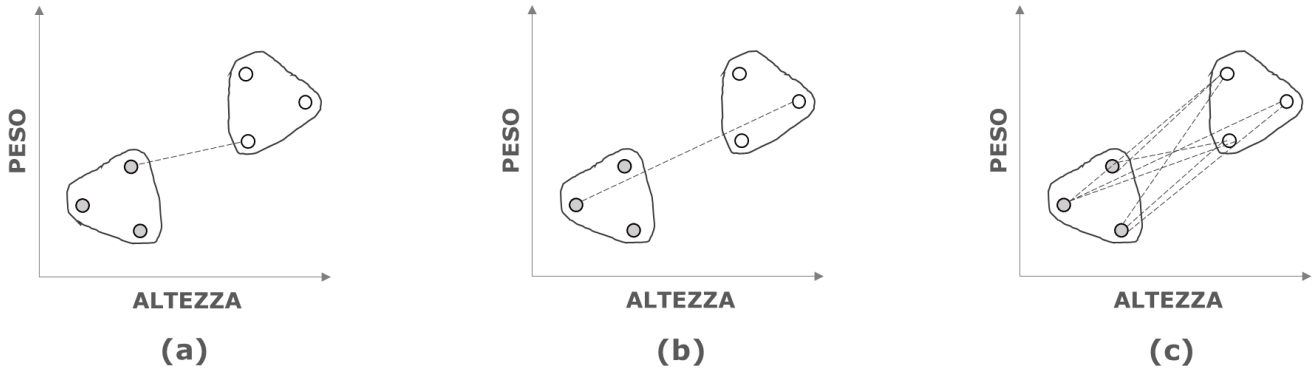


Figura 24 - Distanza/Prossimità single linkage (a), complete linkage (b) e average linkage (c).

ALGORITMO APRIORI: fu il primo algoritmo sviluppato per l'apprendimento di *regole di associazione* in modo tale da limitarne l'esplosione combinatoria. Una *regola di associazione* è un'implicazione logica del tipo $A \rightarrow B$, dove A e B sono insiemi disgiunti, vale a dire A e B non condividono elementi. A e B sono di norma riferiti rispettivamente con i termini di *antecedente* e *conseguente* della regola di associazione. Se facciamo riferimento al PROBLEMA 4, A e B sono insiemi di prodotti, per esempio $\mathrm{A} = \{\mathrm{pop\ corn}\}$ e $\mathrm{B} = \{\mathrm{coca\ cola\ 33cl.}\}$, e la regola di associazione è $\{\mathrm{pop\ corn}\} \rightarrow \{\mathrm{coca\ cola\ 33cl.}\}$, qualora appresa a partire dalla base di dati a nostra disposizione circa i contenuti del cestello di acquisto, suggerirebbe una relazione *forte* (vedi tra poco) tra la vendita di <pop corn> e la vendita di <coca cola 33cl.>. In altre parole, la regola di associazione in questione, indicherebbe che molti clienti che acquistano <pop corn> acquistano anche <coca cola 33cl.>.

La *forza di una regola di associazione* è misurata tramite il suo *supporto* e la sua *confidenza*. Il *supporto* indica quanto frequentemente una regola di associazione sia applicabile ad una determinata base dati, mentre la *confidenza* determina quanto frequentemente occorra il *conseguente* B in quelle istanze della base di dati che contengono l'*antecedente* A. Considerando il PROBLEMA 4, avremo che il *supporto della regola di associazione* $\{pop\ corn\} \rightarrow \{coca\ cola\ 33cl.\}$ sarà il rapporto tra il numero di *cestelli di acquisto* contenenti sia $\{pop\ corn\}$ che $\{coca\ cola\ 33cl.\}$ ed il numero totale di *cestelli di acquisto* che si sono presentati alle casse. La *confidenza della regola di associazione* sarà invece data dal rapporto tra il numero *cestelli di acquisto* contenenti sia $\{pop\ corn\}$ che $\{coca\ cola\ 33cl.\}$ ed il numero di *cestelli di acquisto* nei quali è contenuto solo il prodotto $\{pop\ corn\}$.

L'*algoritmo apriori* risolve il problema della *market basket analysis* o dell'*analisi delle associazioni*, vale a dire, fornita una base di dati in cui sono memorizzati i contenuti di ogni cestello di acquisto, trova tutte le regole associative che contemporaneamente hanno valore del supporto *maggiore o uguale* ad un supporto minimo e confidenza *maggiore o uguale* ad una confidenza minima. L'algoritmo agisce in due passi, il primo consiste nella generazione di insiemi frequenti, vale a dire nella generazione di insiemi la cui frequenza sia maggiore o uguale alla soglia imposta dal supporto minimo, il secondo passo consiste nella generazione di regole associative forti, vale a dire regole associative ad elevato valore di confidenza.

La *generazione di insiemi frequenti* viene effettuata tramite una struttura a grafo che consente di enumerare gli insiemi generabili a partire da un certo numero di elementi. Se nel PROBLEMA 4 esistessero solo quattro prodotti distinti, a, b, c, d, allora i differenti *cestelli di acquisto* che potrebbero presentarsi alla cassa, ipotizzando per semplicità che nessun cestello contenga più di una singola unità di prodotto, sono rappresentati in Figura 25.

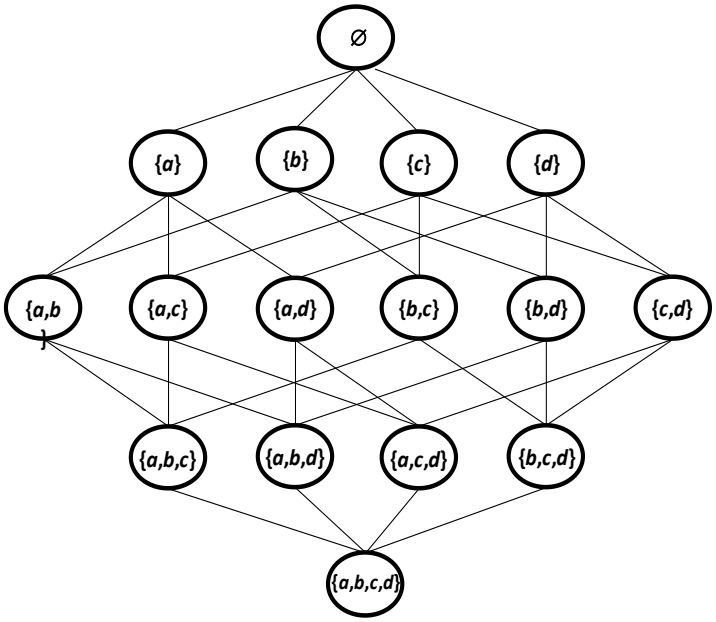


Figura 25 – Reticolo di insiemi

Percorrendo il reticolo dall'alto in basso, ogni nodo di livello "*i*" è ottenuto per estensione, vale a dire per aggiunta di un prodotto, di uno o più nodi del livello precedente "*i-1*". Per esempio, Il nodo {*a*,*b*,*d*} del livello quattro è ottenibile dai nodi di livello tre {*a*,*b*}, {*a*,*d*} e {*b*,*d*} aggiungendo loro rispettivamente il prodotto {*d*}, il prodotto {*b*} e il prodotto {*a*}.

L'*algoritmo apriori* sfrutta il *principio apriori* secondo il quale se un insieme non è frequente allora non lo saranno tutti gli insiemi ottenuti tramite sua estensione, vale a dire tramite l'aggiunta di un altro elemento; in questo modo è possibile potare il reticolo di insiemi, eliminando tutti i nodi discendenti di un nodo associato ad un insieme non frequente, e quindi viene sensibilmente ridotto il numero di insiemi per i quali è necessario calcolare il valore del supporto e della confidenza.

Prendiamo ora come riferimento la Figura 26. Se l'insieme {*a*,*c*} in Figura 8.26 (a) risultasse non frequente, vale a dire se esso avesse un valore del *supporto* minore al valore del *supporto minimo* da noi scelto, allora in base al *principio apriori* sarebbe possibile evitare la computazione del supporto degli insiemi {*a*,*b*,*c*} e {*a*,*c*,*d*} ottenuti per sua estensione; in cascata si potrebbe evitare la computazione di supporto e confidenza per le estensioni di tali due insiemi, vale a dire l'insieme {*a*,*b*,*c*,*d*} che certamente, come {*a*,*b*,*c*} e {*a*,*c*,*d*}, risulterà un insieme non frequente.

Allo stesso modo, se l'insieme {*d*} in Figura 26 (b) risultasse non frequente, sarebbe possibile potare il reticolo di insiemi eliminando dapprima i nodi associati alle sue estensioni, vale a dire gli insiemi {*a*,*d*},

{*b*,*d*} e {*c*,*d*}, e successivamente eliminando le estensioni di ognuno dei tre insiemi, vale a dire eliminando gli insiemi {*a*,*b*,*d*}, {*a*,*c*,*d*} e {*b*,*c*,*d*}. Infine, si procederebbe ad eliminare l'insieme {*a*,*b*,*c*,*d*} in quanto estensione degli insiemi {*a*,*b*,*d*}, {*a*,*c*,*d*} e {*b*,*c*,*d*}.

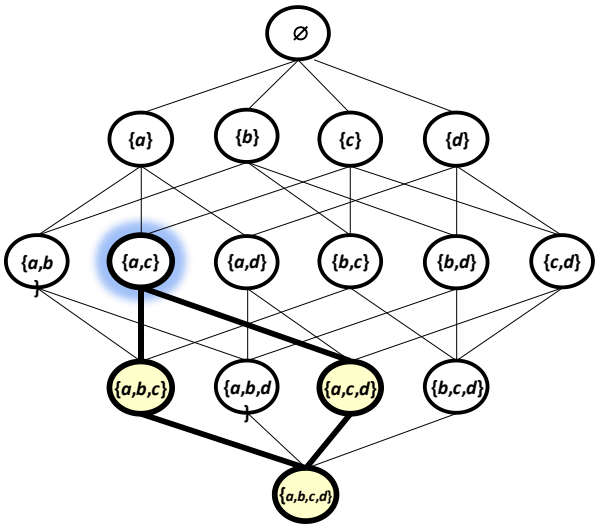


Figura 26 (a)

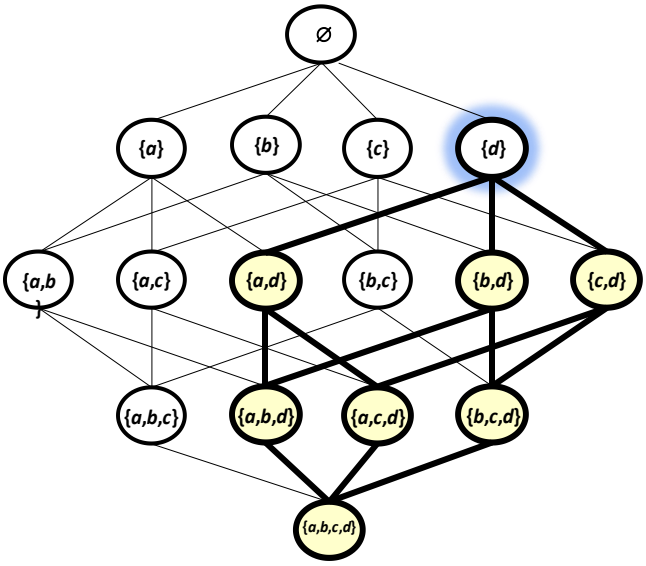


Figura 26 (b)

Figura 26 - Utilizzo del principio apriori per effettuare la potatura del reticolo degli insiemi, eliminando gli insiemi non frequenti.

Tramite l'operazione di potatura del reticolo degli insiemi, il principio apriori consente all'algoritmo apriori di ridurre sensibilmente il tempo dedicato alla computazione del supporto degli insiemi generabili a partire da un determinato numero di elementi/prodotti. La seconda fase dell'algoritmo apriori, quella dedicata alla computazione della confidenza di una regola associativa, è di norma molto

meno onerosa in termini computazionali. Comunque, anche in questo caso l'algoritmo apriori utilizza considerazioni circa il livello di *confidenza* delle regole per applicare, quando possibile, una riduzione del tempo di computazione per apprendere le *regole di associazione*.

3.3 Machine Learning per rinforzo

Il Reinforcement Learning studia l'interazione tra due attori, un *agente* e un *ambiente*. L'obiettivo è verificare se e come sia possibile dotare l'agente di un'intelligenza che gli consenta di apprendere come comportarsi in un ambiente a lui ignoto. Secondo questo paradigma l'apprendimento avviene tramite prove successive, nelle quali l'agente esegue determinate azioni sull'ambiente e ne riceva un premio o una punizione.

La caratteristica principale del Reinforcement Learning è il compromesso tra esplorazione e sfruttamento, vale a dire decidere quando l'agente debba continuare ad esplorare l'ambiente che costituisce un territorio a lui sconosciuto, e quando debba invece iniziare a sfruttare la conoscenza sull'ambiente che ha acquisito sino a quel momento tramite le esplorazioni precedenti.

Molti sono gli approcci di Reinforcement Learning proposti negli anni, e la loro trattazione è per ragioni di spazio fuori dallo scopo di questo capitolo. Tuttavia, al fine di comprendere l'essenza del Reinforcement Learning, ritengo utile descrivere brevemente il *Multi-Armed Bandit Problem* che è stato molto studiato in Statistica, Ricerca operativa, Informatica ed Economia.

Figura 27 - Slot machine del casinò. Copyright: medium.com. (Fonte: https://medium.com/gradientcrescent/fundamentals-of-reinforcement-learning-the-k-bandit-problem-illustrat-940eea430296)

Il *Multi-Armed Bandit Problem* o *problema del bandito multi-braccia* origina da una narrazione colorita nella quale uno scommettitore entra in un casinò nel quale sono presenti diverse slot machines (Figura 27). Lo scommettitore inizia a giocare, e ad ogni turno di gioco può scegliere liberamente la slot machine con la quale tentare la fortuna.

Ogni slot machine, quando se ne tiri il braccio, fornisce un ritorno che può essere positivo o negativo, vale a dire può premiare lo scommettitore con un certo ammontare di denaro o lo può punire, trattenendo tutto il denaro che egli ha scommesso in quella puntata. Tra tutte le slot machine ve ne è una ottimale, vale a dire una slot che tende a ripagare lo scommettitore in misura maggiore di quanto non facciano le altre slot machine. Purtroppo, lo scommettitore non conosce quale sia questa slot machine, vale a dire quale sia la slot machine da sfruttare, e cerca di apprenderlo tramite giocate

successive, dette *esplorazioni*. In questo caso lo scommettitore gioca il ruolo dell'agente, mentre le slot machine, con i relativi ritorni, giocano il ruolo dell'ambiente.
La domanda di ricerca alla quale la comunità scientifica cerca di fornire risposta è se e come sia possibile per lo scommettitore sviluppare una strategia sequenziale di attivazione del braccio delle diverse slot machine, in modo tale che venga bilanciato il compromesso tra esplorazione, la scelta cioè di giocare con diverse slot machine, e sfruttamento, la scelta di giocare con una slot machine specifica, possibilmente quella che massimizza il guadagno dello scommettitore.

Il problema del bandito multi-braccia, che conosce diverse formulazioni in dipendenza delle ipotesi circa l'ambiente considerato ed il livello di conoscenza accessibile all'agente, è stato applicato in molti domini che vanno dai piani di sperimentazione clinica di un farmaco o di un trattamento medico in senso lato, alla gestione dei banner pubblicitari da esporre in una pagina Web dipendentemente dal tipo di visitatore, fino alla risoluzione di problemi di cammino minimo formulato in condizioni di incertezza.

In questo ultimo caso una persona (l'agente) deve spostarsi dalla propria abitazione al posto di lavoro ogni mattina; è normale immaginare che lui/lei desideri effettuare tale spostamento seguendo il cammino più conveniente, vale a dire il cammino che in media richiede il minimo valore del tempo di viaggio. Il suo problema è però che egli ignora quale sia il tempo di viaggio richiesto dai cammini alternativi che può seguire per raggiungere il posto di lavoro partendo dalla propria abitazione; in tali ambiti sono state formulate e risolte con successo diverse istanze del problema del bandito multi-braccio; il problema negli ultimi anni ha riscosso un discreto successo anche nell'ambito dei sistemi di raccomandazione o recommendation systems.

4. Conclusioni

Le ricerche che effettuiamo tramite un motore di ricerca, i post che pubblichiamo tramite *Twitter*, le foto che esponiamo nella vetrina di *Instagram*, le coordinate geografiche della posizione dove ci troviamo, le recensioni dei ristoranti che redigiamo su *TripAdvisor*, gli acquisiti di beni e servizi tramite piattaforme di e-commerce, sono alcuni esempi dell'enorme quantità di dati che quotidianamente generiamo e che "doniamo" a Google, Facebook, Amazon, Twitter, Whatsapp, Instagram, e molte altra piattaforme digitali, molto spesso senza consapevolezza dell'uso che ne verrà fatto.

L'enorme disponibilità di dati in formato digitale, insieme alle elevate capacità di memorizzazione e bassi costi dei dispositivi di computazione, hanno innescato la rivoluzione tecnologica alla quale stiamo assistendo. Ogni volta che accediamo ad uno dei nostri dispositivi digitali, computer, tablet o smart phone, dobbiamo esser certi che esso ci conosce estremamente bene, conosce le nostre abitudini, i nostri gusti, le nostre amicizie, dove siamo stati oggi, cosa abbiamo fatto, quali pagine Web abbiamo consultato, quanto tempo abbiamo speso a comunicare con la nostra rete sociale, quale è il nostro orientamento politico, se abbiamo problemi di salute, ed molto altro ancora. Questa rivoluzione, resa possibile dal Machine Learning, per anni è stata tale solo per ricercatori e addetti ai lavori; ai giorni nostri, ed ancor più negli anni a venire, il Machine Learning interesserà ogni fase ed ogni aspetto della nostra vita.

Il Machine Learning offre enormi opportunità al genere umano, ed al tempo stesso lo interroga su cosa si sia effettivamente ottenuto sino ad oggi, su cosa sia ragionevole ipotizzare che venga raggiunto in un futuro non troppo distante, sulla capacità di comprendere e criticare i suggerimenti che ci vengono forniti dallo specifico algoritmo di Machine Learning che utilizziamo, ed ancora su aspetti etici come ad esempio la discriminazione di genere o etnia, di provenienza geografica e infine sulla responsabilità morale e giuridica delle decisioni che vengono basate su suggerimenti forniti da modelli ed algoritmi di Machine Learning. Riprenderemo questi aspetti nel Capitolo 15 dedicato all'etica dei dati digitali.

In questo scenario il Deep Learning sta, senza dubbio, giocando il ruolo del leone. Veicoli a guida autonoma, sistemi di riconoscimento del volto, algoritmi per l'annotazione automatica di in una scena (identificare gli elementi presenti nella scena, fornirne una classificazione di tipo ed eventualmente inferirne le relazioni tra loro), sistemi conversazionali basati su attori virtuali che interagiscono con noi, inducendo empatia tramite opportune manipolazioni delle espressioni facciali e del tono della voce. Questi sono solo alcuni esempi dell'incisività che il Deep Learning, non a caso il Turing Award del 2018, omologo del premio Nobel per l'informatica, è stato assegnato insieme a tre illustri ricercatori del settore, Geoffrey Hinton, Yan Lecun e Joshua Bengio.

Il Machine Learning è estremamente potente, se basato sulla disponibilità di grandi quantità di dati, su una potenza computazionale crescente, accessibile a costi molto contenuti, condizione sine qua non, sulla capacità dell'essere umano di interrogarsi, criticare, e soprattutto sulla capacità di immaginare. D'altra parte la narrazione che viene fatta del Machine Learning da parte di diverse testate giornalistiche, carta stampata, Web, emittenti televisive e reti sociali, va valutate con grande attenzione. L'iperbole narrativa alla "Blade Runner", di cosa si sia già ottenuto e di cosa sia a portata di mano, già in passato, ha portato a grandi entusiasmi, irrazionali ed ingiustificati, che puntualmente si sono trasformati in enormi delusioni, culminate nel cosiddetto inverno dell'Intelligenza Artificiale.

In conclusione, ritengo necessario consigliare di avere equilibrio, informarsi, ascoltare il parere di più esperti del settore, aumentare la nostra comprensione su cosa sia la rivoluzione del Machine Learning, sui suoi effetti palesi e su quelli che verosimilmente si realizzeranno in un futuro non troppo distante, con la consapevolezza che il Machine Learning svolgerà un ruolo fondamentale nel "disegnare" la società nella quale vivremo domani. Non lasciamo il Machine Learning solo agli addetti ai lavori.

Riferimenti

P. Domingos, P. - The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books. ISBN: 978-0-465-06570-7, 2015.

I. Goodfellow, Bengio, Y., Courville, A. - Deep Learning. MIT Press, 2016.

K.P. Murphy - Machine Learning: A Probabilistic Perspective. The MIT Press, 2012

P.N. Tan, Steinbach, M., Kumar, V. - Introduction to Data Mining. Addison Wesley. ISBN: 0321321367, 2005.

I.H. Witten, Frank, E., Hall, M. A. - Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufmann. ISBN: 978-0-12-374856-0, 2011.

Capitolo 11 – Introduzione alla Visualizzazione dei Dati

Federico Cabitza

In questo capitolo parliamo degli elementi essenziali della visualizzazione dei dati (o data visualization) e del contributo di questa pratica accademica e professionale per la Scienza di dati. Ci soffermeremo su tre aspetti fondamentali della data visualization: il cosa, il perché e il come. Ponendo la questione sotto forma di domanda in altrettante sezioni dedicate, cercheremo di fornire al lettore le nozioni, gli argomenti ed alcuni esempi utili affinché questi trovi, nello spazio concettuale aperto da queste domande essenziali, le risorse utili a capire la natura delle pratiche che producono visualizzazioni di qualità, e la natura delle competenze necessarie per interpretare tali oggetti visuali in maniera non distorta in qualsiasi attività in cui i dati possono aiutarci a capire il mondo che ci circonda.

> *La maggior parte di noi ha bisogno di ascoltare la musica per capire quanto sia bella. Ma spesso è così che presentiamo le statistiche: mostriamo solo le note, non suoniamo la musica.*
> **Hans Rosling**

5. Cosa è la data visualization?

L'insegnamento di *Data Visualization* è stato il primo di un corso di Laurea Magistrale che mi fu affidato appena divenuto professore associato: per questo motivo fu per me una grande emozione ricevere questo incarico da Carlo Batini in persona, che era allora il coordinatore del corso in Data Science che sarebbe iniziato nell'autunno successivo. Come spesso mi capita, però, pochi giorni dopo trovai comunque modo di sentirmi un po' insoddisfatto: per la definizione del regolamento didattico, avevo proposto un nome diverso da Data Visualization per quelle 21 ore di insegnamento, un nome che mi sembrava più adatto al profilo altamente innovativo che Carlo voleva dare a quel corso di laurea e, soprattutto, che rispecchiava maggiormente l'approccio empirico e sperimentale che avrei voluto adottare. Ma Carlo non reputava quel nome né attraente né attrattivo nei confronti degli studenti, e neppure rappresentativo di come la maggior parte degli studiosi di riferimento intendessero quel campo di studi: l'insegnamento si sarebbe quindi chiamato Data Visualization (e si chiama ancora così!).

Anche se Carlo aveva probabilmente ragione, io da quel primo anno non posso fare a meno, all'inizio del corso, di raccontare agli studenti questo aneddoto. E se ne affollano davvero tanti sui banchi della prima ora: il corso è infatti uno dei primi dell'intero corso di laurea e le prime ore sono sempre molto popolari tra gli studenti, convinti come sono di poter recuperare in quell'occasione indicazioni utilissime sulle modalità d'esame (e, penso, anche su che tipo sia il docente che dovrà giudicarli alla fine del corso). Quindi condivido con l'intera classe i miei scrupoli vagamente nominalistici (convinto come sono che

“ogni nome è presagio”, come dicevano i Latini), e spiego loro che durante l’anno vorrò sensibilizzarli verso alcuni temi, tra tutti quelli possibili nell’ampio panorama della visualizzazione dei dati, che avrei voluto rendere chiari ed espliciti, quasi programmatici, a partire fin dall’indicazione che sarebbe apparsa sul loro libretto.

Ai lettori di questo capitolo dirò un poco più avanti quale nome avessi proposto al posto di Data Visualization per un insegnamento che nell’ambito degli studi della Scienza dei dati si occupasse di come visualizzare i dati in maniera ricca, suggestiva e spesso anche interattiva attraverso sistemi dedicati e, sempre più spesso, il Web. Prima di farlo, però, introdurrò comunque l’oggetto del discorso riferendomi ad esso con il nome più comune e tradizionale.

Appunto, che cosa dovremmo chiamare *Data Visualization*, nei termini che usiamo per denotare una specifica materia di studio e approfondimento nella complessa articolazione del concetto di Scienza dei dati? Questa domanda è molto affine al chiedersi: cosa significa visualizzare i dati? Certamente non possiamo limitarci a pensare che si tratti di “rendere visibili” i dati, cioè semplicemente renderli disponibili alla nostra vista. Anche stampare dei dati in forma testuale e simbolica tra gli stretti margini di una tabella equivarrebbe a renderli visibili; ma farlo come proposta di una “visualizzazione dei dati” sembrerebbe più una provocazione che una esemplificazione corretta (o quanto meno, nell’ambito del nostro discorso, rappresenterebbe una specie di occasione persa). In realtà, occuparsi di visualizzazione dei dati significa concepire, progettare, realizzare e validare modalità opportune e adeguate per rendere i dati *visuali*, o meglio, *più visuali*.

5.1. Una scena d’altri tempi (molto lontani)

Ora, sono consapevole che una qualche precisazione si impone, per non essere presi per ciarlatani. Anche la dimensione della testualità, cioè quella abitata dai testi scritti, ha a che fare con un linguaggio visuale e con qualcosa che si offre alla vista e alla comprensione di chi “sa leggere” (lo sa bene chi si occupa di tipografia e composizione di testi e ama scegliere con cura una font tra mille tipologie e sottospecie).

Eppure, io penso che non dobbiamo essere troppo diversi, noi che viviamo alla fine del secondo decennio del ventunesimo secolo, da quegli spettatori che, più o meno 2437 anni fa, si erano accalcati, un po’ come faremmo ora per la prima di un film di Cameron, per assistere all’ultima rappresentazione teatrale del più grande autore del loro tempo, Euripide, un’opera di cui non ci rimangono che pochissimi frammenti. Quante risate (ce lo dicono le cronache del tempo miracolosamente giunte a noi) scatenò la seguente scena: un incolto pescatore, incalzato dal protagonista, un comandante che aveva avuto l’importante incarico di ritrovare Teseo, il mitologico re di Atene dato per disperso in mare nel viaggio di ritorno dalla missione contro Minosse, descrive come può i diversi segni che la mattina precedente aveva scorto sulla chiglia di una nave arenata sulla battigia: “e il terzo segno è come una ciocca di capelli ricci”. Con una certa fatica, cercando di capire cosa il pescatore analfabeta avesse visto sulla chiglia della nave, il protagonista capì che quel segno doveva essere una Sigma, il terzo dei simboli raffigurati in Figura 1; e così, un segno alla volta, ricostruì che il pescatore dovesse aver visto la scritta ΘΗΣΕΥΣ (cioè Theseus in Greco) e che quindi quella arenata la notte prima fosse effettivamente la nave di Teseo (una buona notizia, perché almeno non era affondata!).

Figura 1 - Una inscrizione in Greco antico su un marmo bianco del Partenone (particolare). Si noti il terzo simbolo da sinistra della prima riga in alto, e il primo simbolo della seconda riga: una sigma per greci e grecisti, ma più probabilmente per noi una M rovesciata su un fianco, oppure una W rovesciata sull'altro, o, perché no, una ciocca di capelli ricci. Copyright: Karel Miragaya. Fonte: https://bit.ly/2mhsEBE

Le risate suscitate da quella scena ci dicono qualcosa del livello di istruzione del pubblico Ateniese del tempo (difficilmente degli analfabeti avrebbero riso guardandosi allo specchio della propria ignoranza); ma anche di quale trovata originale dovette sembrare allora quella di Euripide, di ricondurre i segni delle lettere che gli spettatori avevano imparato a *leggere* a quello che in realtà erano, null'altro che segni da *vedere* su un pezzo di legno: arbitrari, convenzionali, semplici "simboli".

È proprio la cultura greca, quella di Aristotele e di Platone (che aveva infatti un rapporto ambivalente nei confronti del potere della scrittura), a cominciare a separare il mondo della scrittura da quello delle figure, a tal punto che allora si potesse effettivamente ridere del fatto che qualcuno potesse impacciatamente confondere i due piani distinti della "trascrizione del linguaggio verbale" [Falcinelli, 2014] (p. 145), cioè la testualità, e del "ricalco del visibile", la figuratività.

Allora cosa significa rendere i dati "più visuali"? Possiamo provare a tradurre in termini semplici e poi complicare un poco la faccenda. Per ora basti dire: significa dare loro una forza comunicativa che "parla" anche per elementi grafici, immagini, forme, colori, e per una qualche somiglianza e vicinanza alla "cosa reale", quella che in altri capitoli del libro, notate, abbiamo chiamato *osservabile*, o alle grandezze a cui i dati vogliono riferirsi o di cui sono misura. Visualizzare i dati significa quindi *mostrarli*, illustrarli, esporli, in contrapposizione al parlarne, descriverli a parole, scriverli e trascriverli.

Arrischierò più avanti in questo capitolo il suggerimento di alcuni motivi per cui penso sia importante visualizzare i dati; in questa sede non posso che solo accennare al cosiddetto "visual turn", o la *svolta visuale*, termine con cui si denota un nuovo e relativamente recente grande interesse da parte di un numero crescente di studiosi nelle scienze umane e sociali per tutte le forme visuali (e non scritte) che abitano e caratterizzano la nostra cultura contemporanea (così come molte altre culture, incluse molte di quelle cosiddette "tradizionali").

In questi studi che, dagli inizi del secolo ventunesimo, si concentrano sul ruolo degli artefatti e prodotti visuali e sull'importanza delle corrispondenti pratiche di produzione e consumo nella comprensione delle culture e società che ne sono caratterizzate, si nota come la tecnologia della seconda metà del ventesimo secolo abbia portato nella nostra vita una presenza sempre maggiore e sempre più attraente (per non dire, talvolta, invadente) di forme visuali, che hanno lo scopo di rappresentare i fatti, ciò che sappiamo del mondo che ci circonda, e raccontare nuove storie, in modo più persuasivo e potente, sul mondo che abitiamo.

Si pensi alla "riproducibilità tecnica" degli oggetti artistici; al cinema e alla televisione ovviamente; ma anche ad innovazioni solo apparentemente meno influenti, quali la pubblicità in televisione; la stampa offset nella stampa di periodici e giornali; i videogiochi; e infine il personal computing e, all'interno di questo paradigma, la transizione dai terminali a carattere alle interfacce visuali, delle "finestre", delle scrivanie ("desktop") e delle icone che indicano tanto oggetti metaforici (ad esempio cartelle e documenti) che operazioni su di essi (quali "salva" e "taglia").

In questa luce, rendere "più visuali" i dati significa tradurli, nel senso di trasportarli, da un ambito principalmente astratto e in cui è preponderante (se non egemonico) l'elemento scritto "simbolico", in cui qualcuno ci parla di fatti, misure, e codifiche che vengono da lontano, ad un ambito in cui il messaggio si offre silenzioso ma più vicino, vibrante, palpitante, particolare e vivo, tale da richiedere un atto di interpretazione più attivo e coinvolgente da parte del singolo fruitore, in cui ciascuno di noi guardi un pezzo di mondo e si racconti una *sua* storia su di esso, libero di vedere nello stesso segno un ricciolo di capelli, o un serpente o il profilo del becco di una gallina.

5.2. Per un approccio semiotico alla data visualization

Non posso più evitare, a questo punto, di accennare a qualche elemento di quella complessa e frastagliata disciplina, la semiotica, che studia le modalità con cui i segni abilitino, e anzi siano parte indissolubile di ogni pratica di comunicazione e collaborazione umana. Nel farlo, ritengo che quello che l'informatico chiama "dato" non sia che un insieme di segni che è appunto "dato" ad un processore macchinico (e non diciamo meccanico a ragion veduta), ma potremmo anche dire un interprete, e quindi sia adatto ad innescare in questo un processo di interpretazione (che nelle macchine è computazionale, si chiama elaborazione, e che è volto a trasformare ciò che è dato in ingresso, come input in un nuovo insieme di dati, un output).

Il dato, esattamente come ogni segno, ha bisogno di due elementi per diventare informazione: un *modo* di rappresentazione, ad esempio tracce di inchiostro su un supporto cartaceo; fori a distanza regolare in un supporto cartonato; la polarità del campo magnetico di anelli in un materiale ferromagnetico; e

un insieme di convenzioni e prassi produttive (diremmo un *codice*), che rendano possibile un processo di interpretazione. È in questo processo che può instaurarsi quella *relazione*, vibrante e viva, tra il significante (la rappresentazione), uno dei suoi possibili significati, e l'oggetto di riferimento. La magia dei dati e dei segni è tutta qua: per qualcuno, cioè per noi, *stanno per* qualcos'altro, "*sotto certi aspetti e possibilità*", come diceva Peirce, uno dei padri fondatori della semiotica moderna.

Ebbene, quando Peirce considerava i diversi modi in cui un segno può darsi all'esperienza umana, proponeva di distinguere tra tre modalità: *simboli*, *indici* e *icone*. Ci dava così un modo semplice per riconoscere quello che è comune nelle innumerevoli specifiche espressioni in cui i segni innescano il nostro riconoscere e capire. Dopo aver accennato di seguito a questo semplice paradigma, la mia espressione precedente, "rendere più visuale il dato", sarà più chiara.

Allora: il *simbolo* è quel segno che rappresenta l'oggetto in virtù di una convenzione, di un significato arbitrario ma preciso che gli è assegnato all'interno di una comunità da chi vi riconosce quel significato preciso e che lo usa per astrarre sinteticamente molte, se non tutte, le caratteristiche fisiche e percepibili dell'oggetto riferito. Come imparano presto i bambini a cui sono svelati i primi rudimenti dell'algebra, simboli particolari indicano: quantità precise (numeri); oppure quantità ignote ma di cui si conosce la costanza o continua variabilità (lettere); e operazioni da fare su quelle quantità, quali la somma o la sottrazione.

Nulla di un insieme di simboli è veramente universale quanto vorrebbero quanti vi affidano verità logiche, istruzioni algoritmiche e relazioni aritmetiche, a meno che i membri di una comunità di interpreti non siano stati istruiti a riconoscerli o non sia fornita loro una qualche "stele di Rosetta" che permetta la traduzione di tali simboli in un altro sistema di simboli (un altro codice). Il simbolo deve essere visibile, ovviamente, nel senso di percepibile; ma *non è visuale*, nel senso che abbiamo voluto dare alla parola in questo capitolo. Non è visuale proprio perché è distante da come appare l'oggetto a cui si riferisce e, piuttosto, vuole avvicinarsi al suono della parola che pronuncia il nome o la descrizione dell'oggetto.

I segni entrano nel mondo visuale quando, nella terminologia di Peirce, o sono *indici* (o, anche nel seguito, *segni indicali*) o sono *icone (segni iconici)*. Per il segno iconico è facile: l'icona, in quanto tale, ricorda l'oggetto per affinità visuale diretta: il segno assomiglia all'oggetto per il quale sta e il senso comune di questa parola, per una volta, non si contrappone alla interpretazione tecnica, che qui abbiamo ricordato. Per i segni indicali invece la cosa è più sottile, ma non meno rilevante nell'ambito del nostro discorso. L'*indice* ricorda l'oggetto per sostituzione, lo evoca senza rappresentarlo direttamente, ma in virtù di una correlazione tanto forte quanto implicita nel suo innescare una deduzione del tipo "se c'è questo, allora significa quello". Gli esempi di indici che si fanno spesso a riguardo comprendono: il fumo che esce da un camino; esso *rappresenta* il fuoco che deve scaldare gli ambienti della casa sottostante (e può indicare tante altre cose, ad esempio che quella casa è abitata e che vi si sta cucinando qualcosa). Oppure le orme ancora visibili sulla sabbia di una spiaggia, che indicano il passaggio di qualcuno, avvenuto prima che la marea ne potesse cancellare le tracce.

Questa breve disamina ci ricorda che i segni molto raramente si presentano unitari e univoci: spesso icone, simboli e indici coesistono in una stessa visualizzazione, o perfino in un suo singolo tratto, e si

sovrappongono agli occhi di un interprete, con tutti i significati che questi è in grado di scorgervi. Ad esempio, se, persi in una fitta foresta straniera, finalmente scorgessimo alla sommità di un palo un vecchio cartello di legno a forma di freccia, che reca su di esso una scritta incomprensibile in un alfabeto sconosciuto, è l'*indacalità* del segno, e cioè la forma appuntita del cartello su un solo lato verticale, ad indicarci la direzione verso cui dirigersi per la meta dal nome sconosciuto. Se poi il cartello recasse alcuni tratti neri che ricordano nelle forme stilizzate un essere umano e qualcos'altro, come in Figura 2, una via o un tetto sopra la sua testa, sarebbe sempre l'iconicità del segno a rassicurarci che un rifugio è vicino, anche se non sapremo mai il suo nome o quanta strada dovremo ancora fare per raggiungerlo.

In ogni caso, e qui torniamo alla domanda di questa prima sezione, gli indici e le icone innescano in noi un riconoscimento più immediato, anche se forse una interpretazione meno ricca di dettagli e più vaga, della parola scritta e del simbolo. *Rendere più visuali* i dati significa quindi *rappresentarli come segni*. Sia chiaro: difficilmente un segno può perdere la sua natura simbolica, cioè la capacitò di *significare* qualcosa in base a qualche convenzione nota o alla familiarità del suo codice; ciò nonostante, visualizzare i dati significa arricchire le loro rappresentazioni con caratteristiche di indicalità e iconicità, in virtù dei diversi vantaggi che questa operazione comporta, e di cui parleremo nella sezione dedicata al "perché" dovremmo visualizzare i dati, subito dopo la prossima.

Figura 2 - Cartelli segnaletici in un bosco. Il segno per andar su esprime un simbolo, un'icona e un indice, in questo ordine. Copyright: tripiteasy.it (Fonte: https://bit.ly/2kALaVe)

1.3 Un esempio propedeutico alla definizione

La digressione semiotica della sezione precedente ci permette di tornare a quanto ci preme definire in questo nostro capitolo: di cosa parliamo quando parliamo di visualizzazione dei dati? Come abbiamo

anticipato sopra, visualizzare i dati significa rendere "visuale" quello che sappiamo di una certa porzione della realtà di nostro interesse, cioè in grado di attivare un processo di interpretazione e comprensione che vada oltre (per integrarlo) quello basato sul significato convenzionale e arbitrario dei simboli, quello del testo scritto e dei numeri. Più precisamente, in termini semiotici: rendere visuali i dati significa costruirne rappresentazioni grafico/visuali che operino meno sulla base di una competenza tecnica profonda, e più come indici e icone dei concetti e delle informazioni a cui quei dati si riferiscono.

Prima di fornire la definizione che mi sento di offrire ai lettori di questo capitolo, reputo però doveroso fare alcuni esempi, così che parlare di icone e indici non risulti un *latinorum* senza alcuna utilità pratica (mentre la mia opinione è tutto l'opposto, pensare semioticamente è essenziale per fare della buona *data visualization*): ne farò uno preso dall'ambito medico, che è quello in cui ho operato maggiormente negli ultimi 14 anni.

Una azienda britannica, la Bodyscan Limited, nel 2017 ha pubblicato[30] i dati relativi a più di 2.700 mineralometrie ossee computerizzate (di tipo DEXA) fatte nei due anni precedenti a soggetti di sesso maschile, a Londra. La scansione DEXA è in genere utilizzata per diagnosticare e valutare l'evoluzione dell'osteoporosi, ma in questo caso immagineremo di aver estratto da tali esami il dato relativo al Fat Mass Index (o FMI, indice di massa grassa). Questo è un indicatore importante per capire se il soggetto è veramente obeso (e più informativo a riguardo del suo peso), e soprattutto a rischio di malattie cardio-vascolari e metaboliche.

Il dato di cui disponiamo ha due dimensioni, entrambe numeriche e quantitative: l'indicatore FMI, appunto; e l'età del soggetto esaminato. Una tabella che visualizzasse questi dati non sarebbe troppo diversa da quella riportata in Figura 3.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Età	28	33	44	51	26	38	42	56	61	57	39	51
2	FMI	5.76	7.23	3.92	4.72	5.14	5.20	6.51	5.06	4.08	8.08	6.92	8.10
3													
4													

Figura 3 - Vista di una tabella che riporta valori di età e FMI per un certo numero di pazienti (noi ne vediamo solo un piccolo estratto).

A colpo d'occhio questa visualizzazione è piena di simboli: numeri, caratteri alfanumerici, anche le scritte "età" e "FMI" sono simboli di tre lettere (dal significato non scontato); ma a ben guardare anche le linee orizzontali e verticali che disegnano la matrice e dividono i numeri gli uni dagli altri sono segni a tutti gli effetti che, nel loro dipanarsi per lo spazio a loro disposizione, definiscono colonne, e quindi pazienti; e righe, le loro caratteristiche di interesse. Ma questo avviene solo in virtù di una qualche convenzione[31]: appunto si tratta di segni simbolici.

[30] https://www.bodyscanuk.com/bodyscan-data.html

[31] Non sarà sfuggito al lettore che per riportare in questa pagina una figura che avesse uno sviluppo orizzontale ho adottato una convenzione trasposta rispetto a quella, molto più comune, per cui le righe di una tabella indicano le istanze, in questo caso i pazienti, e le colonne gli attributi di interesse.

Questi simboli, ad un lettore competente che sappia far di conto, conosca l'Italiano, e abbia familiarità con alcuni degli acronimi usati in ambito medico, "mostrano" i dati con grande fedeltà: i dati originali sono infatti misure quantitative. Eppure, nonostante la loro fedeltà, è abbastanza pacifico sostenere che sarebbe difficile rispondere a qualsiasi domanda di ricerca di interesse epidemiologico e clinico limitandosi a guardare quell'insieme di dati, soprattutto se pensate al fatto che la tabella di Figura 3 dovrebbe avere più di 2700 colonne! Domande interessanti di questo tipo possono essere: è vero che i soggetti più giovani, che presumibilmente fanno più attività fisica, hanno un FMI più basso dei soggetti più anziani? E se, come è presumibile, le persone sopra i 50 anni hanno un FMI più alto di quelli che hanno meno di 50 anni, per quanti di loro, in proporzione, tale FMI è superiore ad una certa soglia e quindi li espone ad un rischio non trascurabile di salute?

Affidiamo a delle visualizzazioni, anche dalla struttura molto semplice, il compito di aiutarci nell'indirizzare queste domande di ricerca. In Figura 4 possiamo vedere un semplice diagramma a barre; le due barre che si vedono indicano ciascuna la media del FMI (indice di massa grassa) in due sottogruppi: quello di soggetti con età inferiore a 50 anni, e il gruppo dei pazienti che hanno 50 o più anni.

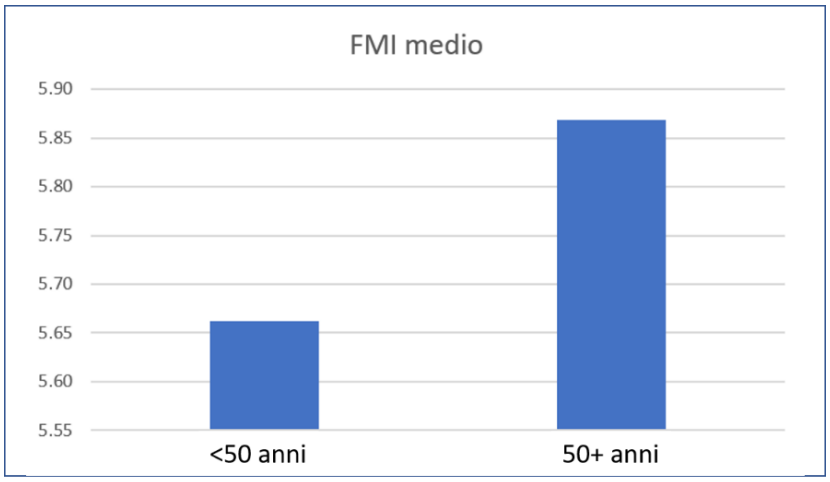


Figura 4 - Un grafico a barre relativo a due gruppi di pazienti, e il loro FMI medio.
Questa visualizzazione ha un problema, ma non vi svelo ancora quale.

E' abbastanza immediato, per chi ha una certa familiarità con i diagrammi, capire che l'estensione verticale di ogni barra rappresenta il valore di FMI del gruppo la cui denominazione campeggia alla base della barra, e che il valore preciso della media può essere letto sulla scala graduata riportata nell'asse verticale in corrispondenza del "tetto" della barra.

Abbastanza chiaramente si intende che più una barra è alta, maggiore è il corrispondente valore di FMI (medio). Ancora convenzioni: quindi la barra è un simbolo. Corretto: ma quello che ci interessa ora *non* è l'estensione delle barre; bensì la *differenza di estensione* tra le due barre, che balza all'occhio per semplice vicinanza di forme. In questo caso, lo scarto evidente tra la prima barra a sinistra, più bassa, e la seconda, a destra, *indica* che c'è differenza tra il FMI medio dei pazienti con meno di 50 anni, e il

medesimo indicatore per i pazienti più anziani. Una volta creato il segno simbolico "barra" che sta (per i lettori competenti) per una caratteristica di un certo gruppo, ecco che, avvicinando un altro segno barra-gruppo, sorge alla percezione un terzo segno indicale: c'è una differenza tra le due barre, e *quindi* tra i due gruppi. E lo vediamo bene!

I grafici a barre non mi hanno mai entusiasmato, a fronte della loro popolarità (seconda sola, forse, ad un grafico ancor peggiore, che è il grafico a torta, o aerogramma): secondo un modo di vedere reso popolare da Edward Tufte, uno dei padri fondatori della disciplina, una barra rappresenta un ingiustificato "spreco di inchiostro" e di superfice poiché, di tutto il suo corpo ed estensione, siamo interessati evidentemente solo alla sua sommità.

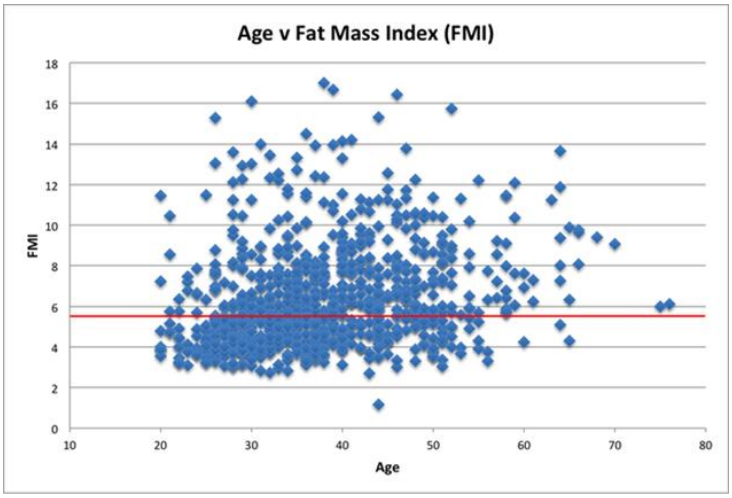


Figura 5 - Diagramma a dispersione che mostra i dati di FMI ed età relativi a tutti i pazienti considerati.

Passiamo quindi ad un'altra visualizzazione, per cui invece non nascondo di provare una certa simpatia: un diagramma a dispersione (o scatter plot, in Figura 5). Tale tipo di diagramma è, per il data scientist, quello che il fumo che si scorge in lontananza tra gli alberi è per una guardia forestale o l'orma nel fango di una radura per un cacciatore esperto: una visualizzazione che ha l'ambizione di mostrare *tutti* i dati, come insieme di punti nello spazio di *tutti i possibili valori* (nelle due dimensioni di nostro interesse). Nel nostro caso, quindi, ogni punto è un punto-dato nello spazio di tutte le possibili coppie di età e FMI; e quindi, ogni punto in Figura 5 rappresenta un soggetto maschile.

Guardando le nuvole o sciami di punti che i grafici a dispersione presentano alla vista in caso di insiemi di dati (data set) numerosi come il nostro, la persona esperta di data visualization "vede" più segni, proprio come farebbero la guardia forestale o il cacciatore in un bosco: la densità complessiva della nuvola di punti-dato gli "parla" della numerosità del campione coinvolto, mentre le regioni più dense gli parlano delle *mode* del campione, cioè delle configurazioni età-FMI più frequenti, *del valore medio* e della *variabilità* intorno a questi punti particolari. L'occhio esperto riesce anche a intravedere come si distribuiscono i valori relativi a ciascuna delle due dimensioni; la semplice forma della nuvola, più o meno oblunga, parla della intensità della correlazione tra le due variabili (abbiamo visto la correlazione nel Capitolo 9): nel caso specifico, quanto al crescere (o diminuire) dell'età dei soggetti coinvolti cresca anche (o diminuisca) il loro FMI.

L’esperto, anziché affidarsi ad un calcolo complesso che produce un numero reale tra -1 e +1, capirà che non c’è correlazione se la nuvola assomiglia ad un cerchio regolare o una ellisse parallela ad un asse; e che essa invece è presente quando i punti sembrano disporsi stretti e ammassati lungo una retta immaginaria che taglia l’asse orizzontale con un certo angolo, tanto più forte quanto più stretta e vicina alla retta che taglia esattamente a metà il piano. L’inclinazione, poi, gli parla della “direzione” della correlazione, cioè se al crescere di una variabile, anche l’altra cresca (correlazione positiva) o diminuisca (correlazione negativa).

Tutti questi sono *segni indicali* per il lettore esperto, che si formano nella testa dell’osservatore insieme alla presentazione della visualizzazione e che possono competere per la sua attenzione per imporsi in una certa interpretazione dei dati. L’esperto sarà anche in grado di vedere più visualizzazioni, dove il non esperto ne vedrà solo una, un po’ come il giocatore di scacchi principiante vede solo i 32 pezzi sulla scacchiera e le prime mosse che possono fare, mentre l’esperto vede diverse possibili configurazioni di mosse e l’esito della maggior parte di esse.

Ad esempio, sempre con riferimento alla Figura 5, la linea retta parallela all’asse orizzontale indica il valore medio del FMI maschile, che è di 5,78 kg/m^2. Come scrive l’azienda che ha pubblicato il grafico, “c'è un'evidente tendenza nei dati dal basso a sinistra verso l'alto a destra, suggerendo che i pazienti mettono su grasso con il crescere dell’età. Ma la tendenza è leggera e sembra essere assente fino alla metà di 50 anni; dopo i 60 anni, quasi tutti i pazienti hanno un FMI superiore, e anche molto superiore, al valore mediano”. Per l’esperto quindi c’è fumo e fumo: il” fumo” che proviene da una massa di sterpaglie ammassate in mezzo ad una radura di conifere e il fumo che proviene da un sottobosco di latifoglie; solo il secondo segnala un possibile pericolo e richiede un suo intervento. Allo stesso modo, l’occhio esperto osserva il grafico in Figura 5, e riconosce che un fenomeno interessante, di natura correlativa, emerge solo per un sottogruppo di soggetti, i pazienti che hanno più di 50-55 anni (si veda Figura 6).

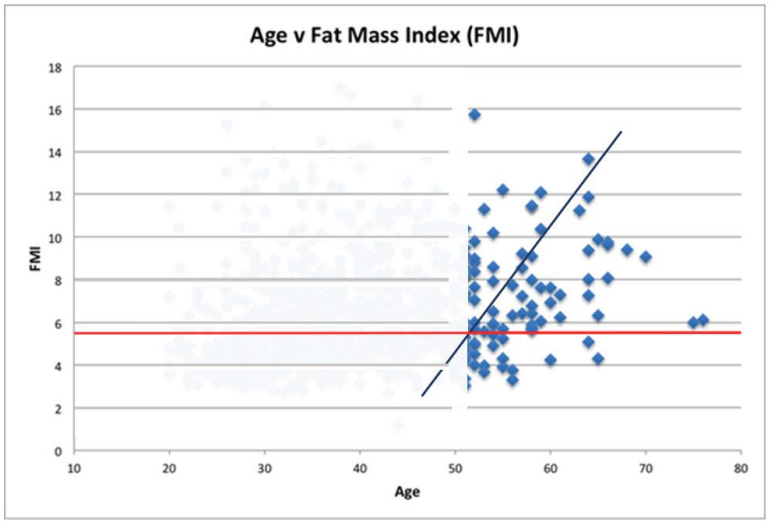


Figura 6 - Il medesimo scatter plot di Figura 5, ma visto da un esperto di data visualization.

La Figura 7 riporta invece un grafico dalla comprensione probabilmente più immediata, dove non operano segni indicali, bensì iconici. Esso è un tipo di visualizzazione che si sta diffondendo da quando si è osservato che ragionare per "frequenze naturali", cioè in termini di numeri da 1 a 100, sia più naturale e facile che in termini percentuali, soprattutto in ragionamenti di tipo Bayesiano che richiedono di manipolare probabilità condizionate [Gigerenzer 2003]. Ciò nonostante, questo diagramma non ha ancora un nome consolidato in letteratura: alcuni lo chiamano *pictorial fraction chart* e faremo così anche noi, evitando di proporre una traduzione estemporanea.

Questo grafico aiuta a rispondere alla seconda domanda di ricerca che ci siamo posti: quale proporzione di persone con più di 50 anni sono a rischio di sindrome metabolica, cioè hanno un FMI più alto di 7 kg/m^2 [Liu et al. 2013]? Qui sono la composizione della figura e il livello di grigio delle icone a forma di uomo stilizzato a suggerire quanti, in un determinato campione della popolazione di riferimento, siano attualmente a rischio di questa sindrome, e quanti no. Il simbolo di uomo è qui il segno iconico che suggerisce il concetto di probabilità nei termini più elementari di numerosità, e il livello di grigio diviene l'indice che crea una distinzione e segnala un rischio significativo.

In un lavoro (non pubblicato) che ho condotto con un gruppo di cardiologi abbiamo toccato con mano come uno strumento simile a quello rappresentato in Figura 7[32] può aiutare a migliorare la comunicazione tra medico e paziente, quando il primo deve presentare diverse opzioni terapeutiche mostrando al secondo diverse grafici di quel tipo, uno per ogni trattamento, e la visualizzazione funge da supporto alla decisione condivisa (shared decision aid). Lo fa in virtù del fatto che è in grado di "mostrare" concretamente la probabilità che il paziente ha di finire nel gruppo dei pazienti rappresentati in grigio più chiaro piuttosto che rimanere nel gruppo rappresentato in grigio più scuro.

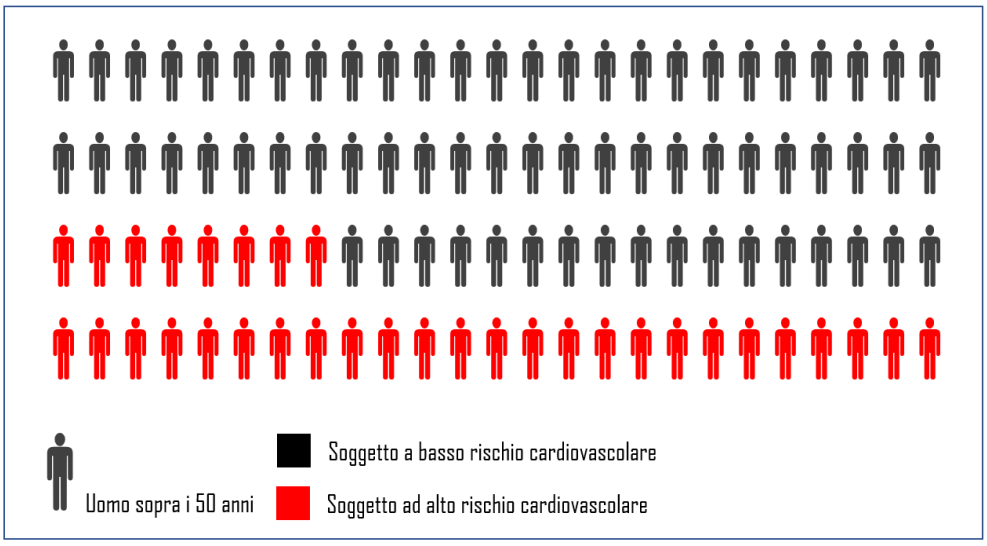


Figura 7 - Un grafico che mostra la proporzione di persone a rischio di sindrome metabolica tra quelle esaminate.

[32] Il prototipo è utilizzabile, accedendo al seguente indirizzo: http://diabetesprevention.altervista.org/.

1.4 Finalmente, cosa è data visualization?

Dopo questi esempi, possiamo rispondere alla domanda posta ansiosamente nel titolo di questa sezione: cos'è la data visualization? Quando ho abbozzato una prima risposta a questa domanda ho adottato la prospettiva per cui questa espressione la si deve associare non a qualcosa di statico ed assoluto, bensì ad un insieme di tecniche, metodi e pratiche: insomma ad un "fare", orientato ad uno scopo. È poi ovvio che con "data visualization" possiamo anche intendere il risultato finale di tali pratiche e quindi l'oggetto visuale creato, un prodotto che intende soddisfare una esigenza o rispondere ad un requisito: ma prima viene il fare, incluse le convenzioni, e le "prassi produttive e regole del gusto" [Falcinelli, 2014].

I motivi per proporvi questa prospettiva sono molteplici. Prima di tutto, nel campo della data visualization, si assiste ad una grande eterogeneità di prodotti finali: il risultato dei metodi e pratiche suddette potrebbe essere una rete, un grafo, un diagramma, un grafico, o una mappa, per citarne solo le tipologie più comuni. In altri casi, niente affatto infrequenti, lo stesso oggetto visuale può essere chiamato da uno *data visualization* (o, più affettuosamente, *dataviz*, come faremo sovente anche noi nel proseguo del capitolo), da un altro *information visualization*, e ancora da altri *infografica*.

A proposito, poiché questo ultimo termine è molto comune e questa distinzione terminologica confonde spesso il soggetto che si avvicina alla data visualization, non di rado qualche studente all'inizio del corso mi avvicina chiedendomi se gli insegnerò a fare data visualization o infografiche. In quei casi io cerco di nascondere l'imbarazzo di dover alimentare una distinzione che a mio parere non è cruciale, anche perché le differenze che si sono sedimentate irregolarmente nel tempo, anche grazie ad una certa superficialità terminologica, riguardano aspetti secondari rispetto al processo di produzione e di interpretazione: in entrambi i casi abbiamo:

- un materiale, i dati;
- un processo di produzione di un artefatto con elementi visuali rilevanti (rispetto al testo scritto e all'uso di simboli numerici); e
- una interazione tra artefatto e fruitore che è volta a creare o diffondere della conoscenza.

Ciò a cui noi diamo nomi diversi, così da creare enti dove ci sono solo processi di produzione e fruizione mediata da artefatti grafico/visuali, è solo un mezzo, che può assumere molte forme: come una rosa è una rosa, anche se ci sono diverse varietà - la rosa Alba, la rosa Bourbon, e la Centifolia - una visualizzazione di dati è una visualizzazione di dati.

Dopo essermi espresso in questi termini, se vedo ancora una certa delusione negli occhi dello studente che nutre questo dubbio, di solito mi arrendo e proseguo: "OK, se proprio ci tiene le posso dare qualche criterio per riconoscere una rosa "Damascena" quando la vede". Ad esempio, una infografica può contenere più dataviz, mentre il contrario non si dà (a quel punto mi permetto di dare criteri anche troppo netti ma semplificatori). In una infografica, come dice il nome, l'elemento grafico, estetico, illustrativo ed iconico è spesso preponderante; e questo anche e soprattutto per la sue finalità principali:

- rivolgersi ad un pubblico a cui le visualizzazioni di dati più tecniche non sono necessariamente familiari, o che può non aver voglia o tempo di interagire con esse e studiarle;

- raccontare loro una storia dotata di una sua narrativa articolata anche se in maniera molto sintetica, appunto per immagini, con l'immediatezza di un fumetto e non disdegnando numerose, sebbene sintetiche, parti scritte e simboliche (ad esempio, numeri e percentuali).

Una dataviz, invece, è meno immediata; richiede un maggior coinvolgimento del fruitore, e la comprensione da parte sua di qualche caratteristica precisa dei dati, attraverso l'interpretazione di specifici elementi visuali, quali ad esempio gli assi orientati e il piano di un diagramma cartesiano. Questi elementi rendono gli elementi indicali (quali ad esempio dei semplici punti) *significanti* dotati di un senso preciso (ad esempio "qui c'è un dato che vale tot").

E incalzo: "In una dataviz l'uso di indici è frequente, mentre nelle infografiche è padrona l'icona". In particolare, in una infografica la comunicazione avviene per analogia grafica con l'oggetto a cui si riferisce, mentre in una dataviz questa analogia opera molto più frequentemente per proporzione di un elemento grafico (ad esempio, la lunghezza di linee rette, l'angolo tra di esse, il grado di saturazione del loro colore) con delle grandezze espresse nei dati.

1.5 Dagli enti ai processi, e quindi all'interazione

Ma insomma, come diceva Ockham, non dovremmo riempire il mondo di enti e cose per il solo piacere del loro riconoscimento tassonomico. Ebbene, allora, cosa tiene insieme artefatti grafico-visuali che variano grandemente per tipo, nome e specifica finalità pur essendo tutti basati su quello che sappiamo del mondo, cioè su dei dati? Io propongo di considerare come *elemento unificante e motivazionale* il processo generale di produzione e fruizione, e quindi il suo scopo di più alto livello, l'obiettivo finale generale. E qual è allora l'obiettivo generale della data visualization?

Per un famoso teorico della disciplina, Colin Ware [Ware 2012], la data visualization, vista come "l'uso di rappresentazioni visuali interattive di dati astratti" (p. vii), deve "*amplificare la cognizione*", e prima di lui Arnheim [Arnheim 1969] era dello stesso avviso. Alberto Cairo, autore di libri molto citati sull'argomento, spinge ancora più in là l'obiettivo di ogni singola data visualization (intesa suggestivamente da lui come un *display* di dati, cioè una loro manifestazione visibile, una loro espressione ed esposizione): "*illuminare le persone*" e informarle [Cairo, 2016]. Per altri, forse meno ambiziosamente, l'obiettivo della data visualization (sia della disciplina che dei suoi risultati) è presentare informazioni, permettere la loro esplorazione e scoperta; Unwin e colleghi [Unwin 2016] aggiungono anche un elemento che mi è caro: "rivelare la inerente variabilità e incertezza" dei dati.

E Cairo coglie un altro elemento per me fondamentale: "Le visualizzazioni dei dati non sono destinate principalmente a trasmettere messaggi predefiniti dai loro progettisti. Invece, sono spesso concepite come strumenti che permettono alle persone di *trarre le proprie conclusioni* dai dati." (mia enfasi) Questo passaggio evidenzia uno degli elementi più importanti della data visualization nella prospettiva semiotica: questa, più che concentrarsi sulle *possibili forme* dei prodotti dell'attività, si focalizza sul *processo di riconoscimento, interpretazione e comprensione* del fruitore del prodotto, visto come intreccio di segni che agiscono su più livelli; e quindi come *mezzo* di comunicazione, non *messaggio*; e neppure come fine in sé, bensì come strumento grazie al quale un utente raggiunge i suoi obiettivi più facilmente.

Finalmente possiamo riprendere la definizione di Ware, per estenderla dando la *nostra* definizione di data visualization. Con tale espressione indichiamo *"un insieme di metodi e pratiche in cui sono concepite, progettate, realizzate e usate rappresentazioni visuali, anche interattive, di dati volte a supportare il loro fruitore nel raggiungimento di determinati obiettivi in vari contesti d'uso, con efficienza, efficacia e soddisfazione maggiore che in loro assenza"*[33].

Questa definizione riflette i miei interessi di ricerca e l'essenza del mio lavoro, sia come docente che come ricercatore, e strizza evidentemente l'occhio alla definizione di usabilità che è data dalla normativa ISO 9241, lo standard della International Organization for Standardization che riguarda l'Ergonomia e l'Interazione uomo-macchina.

La definizione che propongo in questo capitolo vuole essere un invito alla comunità di chi si occupa di data visualization ad uscire da quell'equivoco secondo cui, quasi come se fosse una sineddoche, dovremmo preoccuparci più di una parte che dell'intero, e quindi guardare alla visualization più come uno *scopo* che come *mezzo*. È anche un invito a tutti coloro che si occupano di data visualization a cominciare a guardarla come parte di una disciplina empirica più ampia, che mette l'essere umano al centro, sia come produttore che consumatore, e pone le esigenze situate di questo (cioè dipendenti dalle specifiche situazioni in cui si trova) come precisi requisiti per valutare la qualità di quanto fatto.

Per questo motivo io interpreto la data visualization come una articolazione della disciplina che si occupa di studiare e progettare modalità innovative in cui le persone interagiscono con le tecnologie della informazione e della comunicazione per raggiungere i loro scopi, tipicamente comunicare, informarsi, e prendere decisioni. Ora sì che posso svelare quale nome nuovo proposi a Carlo per l'insegnamento che si occupasse di data visualization: *interazione uomo-dato*!

Questa piccola "rivelazione" chiude adeguatamente la lunga sezione sul "*cosa*", per la quale reputavo fondamentale circoscrivere la questione per spire concentriche, concedendomi anche il lusso di parlare di semiotica; e ci permette di passare ad altre domande molto importanti per mettere a fuoco gli elementi principali della data visualization per la Scienza dei dati: il *perché* e il *come*.

6. Perché dovremmo fare data visualization?

Perché quindi dovremmo visualizzare, o meglio, rendere visuali, i dati? I *perché* nascondono sempre una duplice natura e vocazione: esprimono tanto *i motivi, le ragioni*, quanto anche *le intenzioni e le finalità*; quindi sia la causa che l'effetto di una pratica e di un fare. Aver svelato che per me la data visualization è una parte di quello che caratterizza la nostra interazione con i dati ci permette di prendere una scorciatoia: se lavoriamo con i dati, o vogliamo far lavorare gli altri con i dati, è importante la data visualization *nella misura in cui* essa ci aiuta a rendere i dati *usabili*, soprattutto quando questi sono raccolti in grande quantità, e quando la loro sintesi statistica (attraverso vari indicatori, quali quelli

[33] Ammettiamo che mettere insieme metodi, cioè conoscenza, e pratiche è per certi versi pleonastico, perché ogni pratica si svolge nell'alveo di una conoscenza che, sia essa tacita o esplita, gode anche della importante proprietà di essere condivisa tra le persone che collaborano e instaurano rapporti di interdipendenza in vista degli obiettivi comuni e locali.

di tendenza, variabilità, correlazione e differenza) non sempre può bastare o è reputata inaffidabile (tornerò su questo punto quando accennerò al dinosauro di Cairo).

Eppure, evocare l'usabilità come finalità non può esimerci dal provare ad essere più specifici, pur rimanendo su un piano di voluta semplificazione e generalizzazione: possiamo dire che le finalità della pratica della data visualization sono principalmente due, e rivolte a tipologie di interpreti non sempre coincidenti: la *esplorazione* e la *comunicazione*. Ovviamente, tali scopi non esauriscono le esigenze di chi sente necessario rendere i dati visuali, ma sono un primo livello di requisiti.

In particolare, l'esplorazione è tipicamente volta alla *comprensione* (forse un fine più realistico dell'illuminazione a cui fa riferimento Cairo); mentre la comunicazione è tipicamente rivolta alla *persuasione*, nel senso più ampio del termine, che riguarda anche il semplice convincersi, da parte del singolo fruitore della visualizzazione, che ciò che è indicato dai dati e visualizzato sia vero e attendibile. Vediamo queste due finalità in maggior dettaglio.

La esplorazione dei molti fatti che i dati possono esprimere è volta a facilitare la comprensione dei fatti semplici, e misurati direttamente, quali "sta piovendo da sei ore sul bacino del fiume Lambro"; e di facilitare la derivazione per inferenza di altri fatti più articolati e indiretti, quali ad esempio "e quindi è necessario attivare un piano anti-allagamento per la zona nord-est di Milano". D'altro canto, la comunicazione si occupa di esprimere sia i primi che i secondi, e gli innumerevoli altri fatti di qualche interesse, come "ha piovuto in sei ore la quantità di pioggia che di solito si accumula in due mesi", o che si pongono come possibile spiegazione dei primi, quali quelli inerenti il nesso tra piogge consistenti, la mancanza di invasi artificiali, e i disagi dell'allagamento in vaste zone di una città a forte cementificazione; o quelli relativi ad associazioni ancora più deboli e discutibili, quale il nesso tra le emissioni di anidride carbonica e altri gas serra da parte delle attività umane, il riscaldamento globale, e l'intensificarsi dei fenomeni meteorologici in territori dal clima tradizionalmente temperato umido.

Se tanto basta per identificare le finalità generiche, perché un dato visuale dovrebbe permettere l'esplorazione e la comunicazione più efficacemente di un dato non visuale? Appellarsi al famoso detto che "una immagine vale più di mille parole" non sarebbe sufficiente (anche perché non è vero: ma guardare una immagine è molto più veloce che leggere un testo così lungo!). Ora la mia generalizzazione delle motivazioni rischia di essere un poco più arbitraria che nel caso delle finalità: rispetto al codice testuale/simbolico, quello grafico/visuale gode di due caratteristiche interessanti: una maggiore *universalità*, mediante la quale il suo utilizzo permette di *semplificare* il messaggio e, al contempo, renderlo più *accessibile* anche ad un pubblico meno specialista; e una maggiore *vaghezza* o *indeterminatezza* che, al di là di essere considerata un difetto in molti contesti scientifici, è anche ciò che in una visualizzazione permette, tanto allo specialista che al profano, di "trovare più di quanto stia cercando", cioè di capire o identificare associazioni, correlazioni, relazioni (ad esempio di uguaglianza o superiorità) che non si aspetta, e a cui egli non avrebbe pensato soffermandosi solo al livello simbolico e testuale dei dati. Se ci pensiamo, queste constatazioni rivalutano quella che nel Capitolo 5 abbiamo chiamato la grande sfera opaca!

È evidente che la "sottospecifica" intrinseca nei codici visuali la proponiamo come caratteristica principale che facilita l'esplorabilità dei dati, mentre la loro presunta universalità è proposta come ciò

che facilita il compito della comunicazione più ampia ed efficace, rispetto al testo scritto. Ho indugiato sull'aggettivo "presunto", non perché non sia convinto di quanto una icona che rappresenta una toilette possa infondere un senso di sollievo a chi ne sta cercando una più facilmente di quanto possa fare la scritta туалет impressa sulla porta di un aeroporto a Nursultan in Kazakistan; ma perché sia chiaro che l'immediatezza della fruizione di ogni oggetto prevalentemente visuale, sia esso un dipinto, la scena di un film, una visualizzazione di dati o una indicazione per il bagno, è in buona misura un mito: si veda ad esempio Figura 8, per capire come il segno più essenziale non è sempre quello più immediato o universale.

Figura 8 - Alcuni cartelli che indicano la presenza di un bagno in luoghi pubblici (reperibili sul Web, fonti diverse). Non ci sono scritte e i segni sono semplici, ma quali sono veramente immediati e quali comprensibili solo alla luce di convenzioni culturali e, perfino, conoscenze specialistiche?

Come dice Falcinelli [2014] "il nostro guardare è sempre un guardare esperto" (p. 15) e ogni rappresentazione visuale non è mai completamente neutra, né più naturale di un passaggio scritto, ma possiede anch'essa "un linguaggio, un funzionamento, un'ideologia" (ibidem) e viene interpretata più o meno accuratamente sulla base di fattori esterni ad essa: talora grandi e pervasivi come la cultura materiale di una società in cui essa è prodotta; altre volte piccoli e locali come la cultura (nel senso di grado di istruzione) di chi la osserva; e financo assolutamente contingenti, come quello a cui si sta pensando in un certo momento. In poche parole, ogni segno e rappresentazione, anche il più iconico, supporta un processo di comprensione nella misura in cui il suo fruitore è *competente riguardo al codice* in cui è inscritta.

A questo punto, ai miei studenti faccio spesso riferimento al concetto di *graphicacy,* indicandola come un loro preciso obiettivo formativo: questo termine Inglese segue evidentemente il calco di altri termini più comuni e frequenti, quali *literacy* e *numeracy*, che indicano rispettivamente l'alfabetismo e la dimestichezza con numeri e formule matematiche: quindi graphicacy indica una *competenza*.
Il Dizionario Italiano-Inglese Sansoni non traduce direttamente questo termine ma lo spiega così: "capacità di interpretare rappresentazioni grafiche." Questa definizione, sebbene apparentemente generica, è però limitante in un senso importante, e cioè si limita ad una capacità del *fruitore*.

Per tale motivo, in questa sede preferiamo la definizione che ne danno Aldrich & Sheppard [2000], che considera sia produzione che fruizione: la graphicacy è "la capacità di *comprendere* e *presentare* le

informazioni sotto forma di schizzi, immagini, diagrammi, mappe, grafici e altri formati non testuali e bidimensionali" (mia enfasi).

In questa sede io non considererei entrambe le competenze "comprendere e presentare" necessarie, ma considererei la graphicacy come una competenza da modulare in base al profilo della persona coinvolta:

- se si tratta di un fruitore, ad esempio un normale cittadino o un dirigente d'azienda, o chiunque debba prendere una decisione sulla base di dati disponibili e della loro rappresentazione grafico/visuale, allora la graphicacy ha il suo nucleo caratterizzante nella *comprensione*.
- se si tratta di un data scientist, o comunque di qualcuno che ha l'esigenza di produrre visualizzazioni dati per sé stesso o per gli altri (ad esempio i fruitori di cui sopra), allora la capacità di *presentare* i dati adeguatamente è parimenti importante e, insieme alla capacità di comprendere (che non va data per scontata neppure in chi fa data visualization di mestiere), sono entrambe una parte essenziale della sua preparazione in ambito visuale.

Se poi, al contrario, una persona, sia essa un produttore o un fruitore, difettasse di graphicacy, il rischio sarebbe duplice: in un caso, quello di creare visualizzazioni inadeguate o, peggio, che distorcono i fatti rappresentati da dati. Nell'altro quello di non saper leggere e interpretare un messaggio o, peggio, di farsi fuorviare e indurre in interpretazioni erronee da chi crea visualizzazioni difettose o fraudolente.

Ora posso ammettere che una parte importante del mio corso è dedicata ad un argomento che da sempre intriga lo statistico (almeno a partire dal grande classico [Huff 1993] e, più di recente, con riferimento all'esperto di visualizzazione dati[34]): come si può ingannare la gente con le visualizzazioni. In realtà, la cosa è più semplice di quanto possa sembrare, anche (e spero di poter dire "soprattutto") inavvertitamente e inconsapevolmente: tanto che Rogowitz e colleghi [Rogowitz 1996] preferiscono parlare di come "non ingannare" con le visualizzazioni!

Si può sostenere che se la graphicacy, in termini di comprensione, fosse elevata sia tra chi produce visualizzazioni sia tra chi le legge, distorsioni involontarie o fraudolente sarebbero molto rare (anche perché i lettori le noterebbero subito, evitando di farsene fuorviare o condizionare). Purtroppo, attualmente è vero il contrario. A lezione mostro decine di casi, presi sia dal Web che dalla stampa generalista (e a volte anche dalla letteratura specialistica) in cui ricorrono una varietà notevole di modi di *deformare l'informazione*; un elenco ben fatto di questi modi possibili è curato da Claire Bradley tra le risorse rese disponibili nel sito "From Data to Viz"[35]; si veda Figura 9.

A titolo puramente esemplificativo, mostro in Figura 10 un errore ricorrente, riguardo il quale è possibile escludere malafede nella quasi totalità dei casi: quando si confrontano delle grandezze, e quindi si intende visualizzare una differenza, si dovrebbe sempre evitare di rappresentare le differenze come superfici (ad esempio aree circolari), perché l'occhio umano non è sensibile a tali differenze, almeno non tanto quanto lo è invece nel confrontare estensioni lineari.

[34] Qui cito il breve contributo di Yau [2017] e il libro di Cairo [2019], totalmente dedicato all'argomento, in uscita il prossimo Novembre.

[35] https://www.data-to-viz.com/caveats.html. Acceduto il 13 Settembre 2019. Archiviato su: http://archive.is/vConN

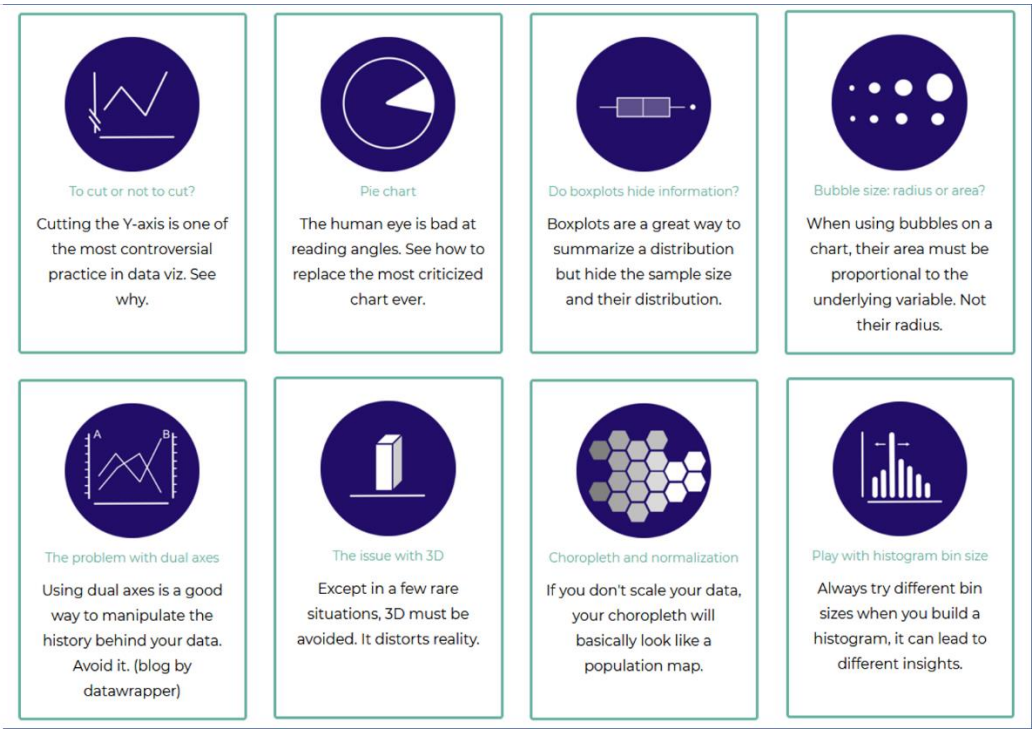


Figura 9 - Otto tra gli aspetti che si dovrebbero tenere presenti per evitare errori o distorsioni nelle dataviz - Da https://www.data-to-viz.com/caveats.html

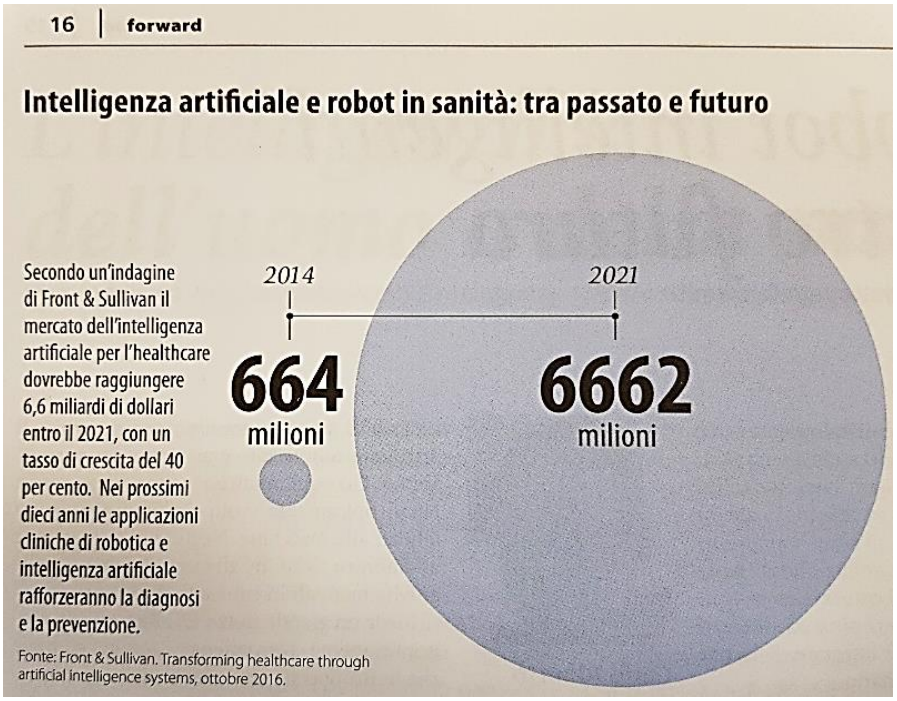


Figura 10 - Una figura che illustra il problema dei grafici ad aree proporzionali

Ma questo, del resto, è un peccato veniale: il problema vero è che, se anche si desiderasse impiegare dei cerchi per evidenziare grandi differenze, si dovrebbe rappresentare le grandezze da confrontare in termini di superfici dei cerchi, e non dei loro raggi o diametri (!), come invece spesso, del tutto inopinatamente, avviene. Nel caso specifico, in Figura 10 la differenza rappresentata simbolicamente è quasi di un ordine di grandezza (un aumento di 10 volte), mentre quella "suggerita" visualmente è di più di due ordini di grandezza (un aumento di circa 120 volte).

Di seguito (in Figura 11) mostro altri quattro possibili modi di distorcere i dati, presi da Chiqui [2015]. Ciascuno di essi rappresenta un modo in cui si può indurre il lettore in interpretazioni errate, non potendo sempre escludere una certa malizia se non proprio una volontà falsificatoria. I primi due casi (1 e 2 in Figura 11) riguardano la distorsione che può essere introdotta "troncando" un asse della visualizzazione, cioè visualizzando i dati così da focalizzare l'attenzione del lettore solo su un particolare intervallo dei possibili valori. In Figura 4 io avevo applicato coscientemente questa distorsione, allo scopo di amplificare (surrettiziamente) il divario rispetto al FMI tra maschi sotto i 50 anni di età e quelli con più di 50 anni.

La mia dataviz ha un *fattore menzogna* (*lie factor*) molto alto: 18. Il *lie factor*, che abbiamo introdotto nel Capitolo 1, è una misura che sempre Edward Tufte [Tufte 2001] ha proposto per rendere esplicito il rapporto tra la dimensione del fenomeno, come questo è *mostrato nella visualizzazione*, e la dimensione del fenomeno come questo è *rappresentato dai dati*: una delle sue raccomandazioni per evitare effetti distorcenti (inconsapevoli o meno) è che tale rapporto sia circa uno.

Questi tre esempi (compreso il mio) riguardano il caso più frequente, in cui ad essere troncato è l'asse verticale; anche il troncamento dell'asse orizzontale può indurre fraintendimenti importanti, soprattutto nel caso delle serie temporali e quindi per la scelta arbitraria di un certo intervallo temporale; a tal riguardo si veda l'esempio di Figura 12, relativo alla efficacia dell'introduzione del vaccino nel caso dell'incidenza di casi di poliomielite negli Stati Uniti d'America.

Il terzo caso in Figura 11 riguarda invece un doppio errore: prima di tutto l'uso della terza dimensione (inutile e spesso giustificata menzionando una gradevolezza estetica che è tutta da dimostrare), che nella rappresentazione a sinistra ingrandisce impropriamente uno dei settori dell'aerogramma (o grafico a torta) poi, l'aerogramma stesso che, nonostante la sua popolarità, non è indicato per indicare proporzioni simili o numerose, in quanto l'occhio umano non distingue bene differenze minute né in termini di angoli né, come si è già detto, di superfici.

Il quarto caso, infine, è un caso a cui è possibile applicare la nota massima "la correlazione non è causalità"[36] (cioè, il fatto che due grandezze appaiono correlate non implica che una sia causa dell'altra), che gli statistici ricordano spesso (anche a sé stessi), quando non è immediato trovare un nesso causale che spieghi perché due grandezze sembrano avere lo stesso andamento (ad esempio, se aumenta una aumenta anche l'altra). In questo caso, per chi produce una visualizzazione del genere è difficile evitare che il lettore non ravvisi una qualche relazione tra gli andamenti visualizzati, perché di fatto è lui stesso a suggerirgliela, con un uso indicale della giustapposizione o vicinanza delle linee.

[36] Introdotta nel Capitolo 9 nel commento alla Figura 16 e che verrà approfondita nel Capitolo 16

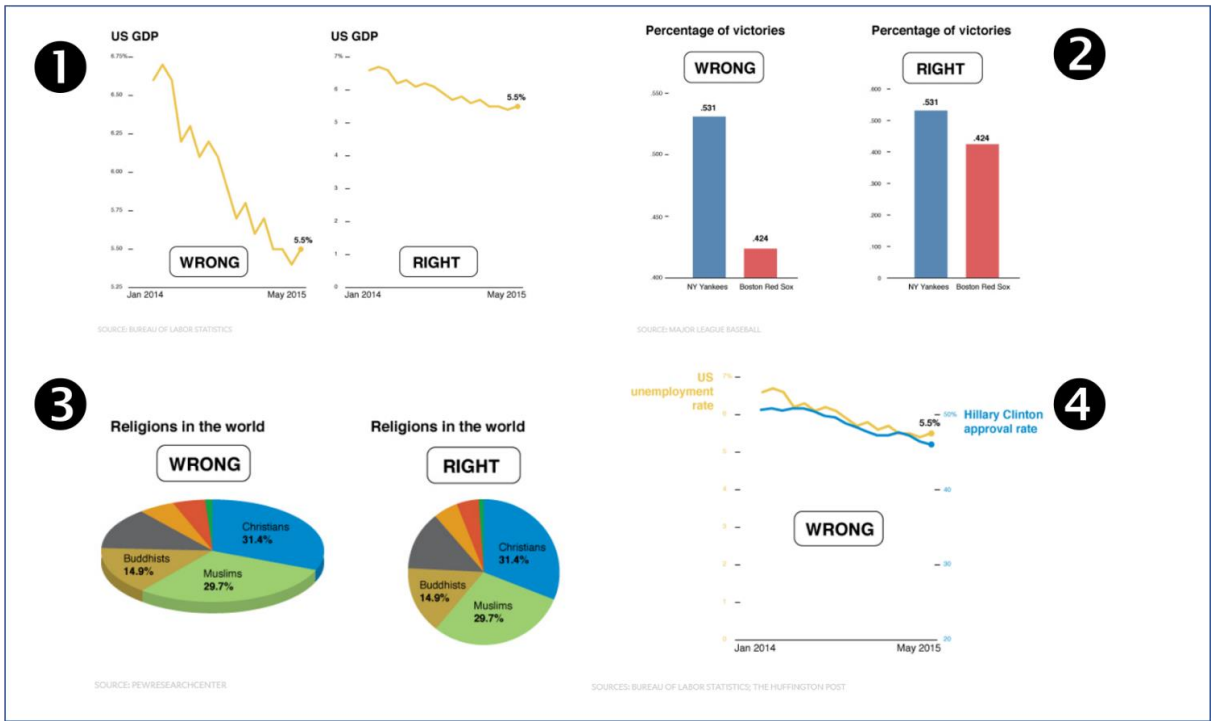


Figura 11 - Quattro modi in cui è possibile ingannare (anche involontariamente) con le visualizzazioni

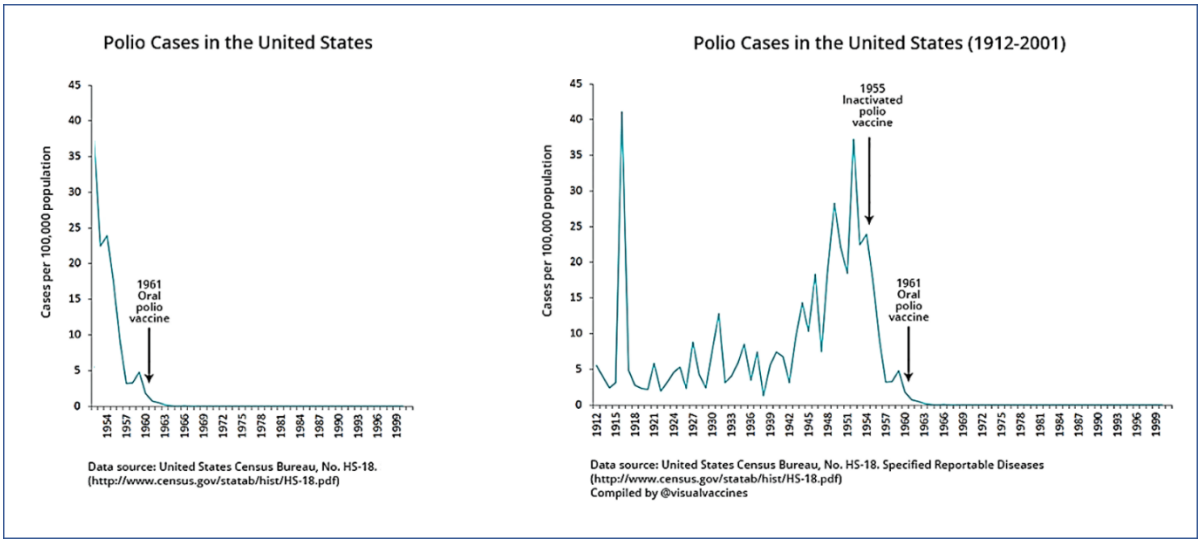


Figura 12 - Due linee temporali che visualizzano gli stessi dati, anche se su intervalli temporali, probabilmente con finalità comunicative e persuasive diverse. A sinistra c'è una versione modificata opportunamente da me; a destra la versione originale di VisualVaccines. (Fonte: https://bit.ly/1T2nv63)

Per tornare al discorso sulla graphicacy, la cui scarsità genera i mostri di cui agli esempi precedenti, di seguito mi concentrerò sugli elementi di questa competenza che riguardano i data scientist più da

vicino. Questi, per dimostrare capacità di graphicacy, sono tenuti a padroneggiare una serie di tecniche e strumenti, che però possono cambiare a seconda della disponibilità e della opportunità: ancora più importante, quindi, è comprendere il quadro complessivo, che passiamo ad illustrare nella prossima sezione.

7. Come dovremmo fare data visualization?

Riprendiamo la definizione che abbiamo dato nella Sezione 5 su cosa sia la data visualization: *"un insieme di metodi e pratiche in cui sono concepite, progettate, realizzate e usate rappresentazioni visuali, anche interattive, di dati volte a supportare il loro fruitore nel raggiungimento di determinati obiettivi in vari contesti d'uso, con efficienza, efficacia e soddisfazione maggiore che in loro assenza"*

Questa definizione ha il merito di far menzione di diversi elementi: della *conoscenza* (si tratta di *metodi*); dell'*attività* (si tratta di *pratiche*); del *risultato* della loro combinazione virtuosa, e cioè le *rappresentazioni* visuali come *strumenti*; e del loro *obiettivo* più nobile: portare valore al fruitore di tali strumenti, che è a sua volta attore e decisore nelle proprie pratiche.

Ma nella definizione c'è anche un altro elemento più implicito: essa accenna a una sequenza logica di attività, a un processo sistematico e strutturato, caratterizzato da una articolazione predefinita e replicabile, assimilabile ad una attività di progettazione: questa sequenza parte dalla raccolta e formalizzazione di una esigenza ("rendere visuali i dati per un certo scopo dell'utente"); passa attraverso un progetto (si potrebbe dire un "disegno" senza perdere troppo il senso originario della parola "design") e la sua realizzazione concreta, con la quale l'utente può interagire; e si conclude con una fase di validazione, cioè di valutazione della adeguatezza di quanto fatto rispetto alle richieste iniziali, le esigenze comunque riconosciute o gli standard di riferimento. Quest' ultimo passaggio di retroazione e verifica è necessario per capire se il fine nobile dell'intero processo, la *creazione di valore* per un attore o decisore, è raggiunto; oppure se si deve rivedere e correggere qualche elemento per migliorare la comunicazione o i processi interpretativi del fruitore.

È con questo atteggiamento, penso, che chi si occupa di data visualization può avere l'ambizione di contribuire adeguatamente ad un discorso più ampio, e a processi di utilizzo e valorizzazione dei dati condotti con approccio o finalità scientifiche, appunto per contribuire alla Scienza dei dati.

3.1 Chi sa, fa

Quando a lezione avverto che la lezione del giorno e quelle successive verteranno su *come fare* una data visualization di qualità, ottengo subito l'attenzione degli studenti. Forse li anima la convinzione che, piuttosto che le definizioni e gli inquadramenti generali, quello che rimarrà loro degli anni passati all'Università, nonostante le loro migliori intenzioni, saranno le competenze procedurali, la conoscenza delle tecniche più efficaci e degli strumenti più efficienti, e un "saper fare" che li equipaggerà adeguatamente nel mondo del mercato professionale. Io sono di avviso quasi opposto: mentre le nozioni e le tecniche potranno essere dimenticate presto, o rese obsolete dalla stessa evoluzione del mercato. Il miglior risultato del periodo universitario è quello di ottenere per sé la salda convinzione di poter affrontare compiti difficili e problemi complessi, pur nella consapevolezza di non sapere molto,

anzi praticamente nulla; aver quindi sviluppato una fiducia nei propri mezzi intellettuali, ad esempio ottenuta superando prove difficili e apparentemente insensate (come il ricordare nozioni di scarsa utilità per un esame), e nella propria capacità di potersela cavare, anche se destinati a vivere e lavorare nell'incertezza e, spesso, nell'ignoranza.

In quest'ottica, inquadrare una tematica, gettar luce sui concetti centrali ad essa e, soprattutto, fornire delle salde motivazioni per capire l'importanza della stessa in un contesto più ampio e incentrato sulle capacità di migliorare la condizione dell'essere umano (o anche solo alcuni ambiti circoscritti nella vita di una sola persona) è molto più importante e proficuo che investire tempo nell'apprendimento di tecniche che potranno presto risultare superate, per non parlare della padronanza di uno strumento software che in pochi anni potrebbe non essere neppure più disponibile.

Per questo motivo, una volta che ho attirato l'attenzione degli studenti con il proposito di occuparmi di come è possibile fare dataviz di qualità, induco sempre una certa delusione, che leggo nei loro volti quando dico loro che chiunque si presenti a loro dicendo che può insegnare loro a fare delle buone visualizzazioni, o non sa cosa dice o è in malafede (e che io non ricado in nessuna di queste due categorie): "non si può insegnare a fare delle visualizzazioni di qualità!". Di solito attendo qualche secondo in silenzio, mentre probabilmente non riesco a nascondere uno sguardo di divertita provocazione.

"Ho detto che non si può insegnare, non che non si può imparare!" Avverto a quel punto che il sollievo è soltanto parziale, ma almeno ho ottenuto di nuovo una attenzione massima. Fare buone visualizzazioni è come fare delle buone torte. Si possono dare consigli su quali strumenti siano i più adeguati ad una certa preparazione; mettere in guardia sugli errori più frequenti; anche dare quelle che possono sembrare delle vere e proprie ricette, alla luce degli ingredienti a disposizione (fuori di metafora: dei dati che si sono raccolti). Ma l'abilità necessaria a fare delle buone torte si sviluppa necessariamente nel prepararne parecchie; buttarne via alcune; bruciarne altre; servirne altre ancora dal gusto terribile; e imparare dalla propria esperienza, così come dalle opere dei maestri che le condividono nella comunità degli appassionati e dei professionisti del settore (avendo però la sensibilità di distinguere un maestro da chi non lo è).

Non esistono quindi, che io sappia, metodologie per fare buone visualizzazioni di dati, o meglio una procedura che garantisca che il risultato finale sia sempre una buona visualizzazione; per questo motivo sono sempre molto scettico nei confronti degli applicativi di analisi numerica dei dati che forniscono la funzionalità per la generazione automatica di un grafico che sia adatto ai dati a disposizione e adeguato al compito di visualizzazione. Pur divenendo sempre più popolari, suggerisco tali sistemi per la loro funzionalità di suggerire come *non* si deve fare una visualizzazione; e suggerisco agli studenti di affidarsi piuttosto allo studio e all'esperienza: all'imitazione di ciò che è piaciuto, che si è ammirato e, possibilmente, invidiato (se si coltiva la passione delle visualizzazioni efficaci); al ricordo delle proprie soluzioni che sono piaciute di più; allo studio di una serie di linee guida e raccomandazioni da innumerevoli fonti, libri, articoli, blog che un insegnante esperto può aiutare a indicare e sintetizzare.

Mi preme però precisare che il rifiuto a ragionare proceduralmente, e il ricorso alla creatività, non rende l'esame più facile; bensì molto più difficile. Piuttosto che affidarsi ad una autorità esterna (ad esempio

il professore), gli studenti devono mettere mano allo strumento che preferiscono; immaginarsi una domanda che un insieme complesso di dati potrebbe aiutare a rispondere; o inventarsi una storia che la visualizzazione possa permettere di raccontare. E devono cercare di stupirmi con la loro visualizzazione.

La mia citazione di una scena del film di animazione Ratatouille che vede protagonista il personaggio di un noto critico di cucina, Egò, è semplicemente il mio modo di sottolineare che non c'è altro modo per provare che il budino, come dicono gli Inglesi, sia buono: *bisogna assaggiarlo*; è cioè necessario che qualcuno, non coinvolto con la sua preparazione e che possibilmente non abbia frequentato la cucina, lo assaggi. Fuori di metafora, è necessario che un utente, che non abbia una conoscenza approfondita dei dati di partenza, *provi la visualizzazione*, interagisca con essa, la esplori, e provi a farsi raccontare una storia, senza ovviamente che questa gli venga presentata in forma scritta ed esplicita.

L'unica metodologia che offro, sebbene chiamarla così pare velleitario o iperbolico, è quella raffigurata in Figura 13. Si tratta di una rielaborazione della ricetta cibernetica di Deming per il miglioramento continuo della qualità di qualsiasi oggetto di design (e noi consideriamo ogni visualizzazione il risultato di una attività di visual design): Plan, Do, Check, Act, l'originale. Progetta, Realizza, Valuta, Migliora, quello che propongo io per la data visualization. Vediamo queste fasi nel dettaglio.

Figura 13 - Variazione sul tema del ciclo di Deming per la data visualization

3.2 Progettazione

La prima attività, la progettazione, ha un nome ambizioso, un perimetro molto ampio, e tale da essere spesso la nota "parte per il tutto", almeno in molti contesti sia accademici che professionali. E' chiaro che progettazione e sviluppo non sono attività da intendersi distinte in modo netto e da svolgersi in una stretta sequenza: a meno che queste attività non siano in capo a team diversi e che interagiscono solo attraverso canali e flussi di comunicazione molto strutturati, è auspicabile che le persone coinvolte nella realizzazione di una dataviz "ciclino iterativamente", come si dice, tra una fase e l'altra, cioè eseguano più volte l'una e l'altra attività, avvicendandosi in una sequenza di tentativi ed errori, o ipotesi ed esperimenti, da cui far uscire un artefatto che si reputa "buono" abbastanza da essere fatto vedere all'utente (o al committente, o comunque ad un rappresentante degli "altri" che non fanno parte del team di progettazione e sviluppo della dataviz).

Come diceva Herbert Simon, uno dei fondatori delle cosiddette *scienze della progettazione* [Simon, 1996], "progettare è ideare linee d'azione volte a trasformare le situazioni esistenti in situazioni che si preferiscono". Nella mia esperienza, il buon design nasce dalle pratiche di persone che mettono in discussione ciò che li circonda (che, probabilmente, è a sua volta frutto di un design che li ha preceduti) e si pongono delle domande per conto dell'utente che interagirà con le loro opere, senza avere la pretesa che quello che costruiranno dovrà necessariamente rispondere a quelle domande o esigenze, quanto piuttosto aiutare a porne di nuove e diverse.

Per questo motivo, come strumento alla progettazione sono solito mostrare ai miei studenti qualcosa che assomiglia ad un diagramma di flusso come quello riprodotto in Figura 14, ma in realtà è solo un modo per cominciare a porsi alcune domande che li aiutino a inquadrare il problema della progettazione di dataviz in modo equilibrato e corretto.

3.3. La metodologia "Socrate"

Chiamo così il mio approccio, nel ricordo scolastico di quello che amava fare il grande filosofo greco Socrate con i suoi allievi e interlocutori, il *metodo maieutico*, per farli arrivare alla comprensione di una verità razionale attraverso l'incalzare di domande sempre più precise e implacabili. Come il diagramma riportato in Figura 14 può far intuire, la progettazione di una dataviz riguarda molto "sudore" da dedica-

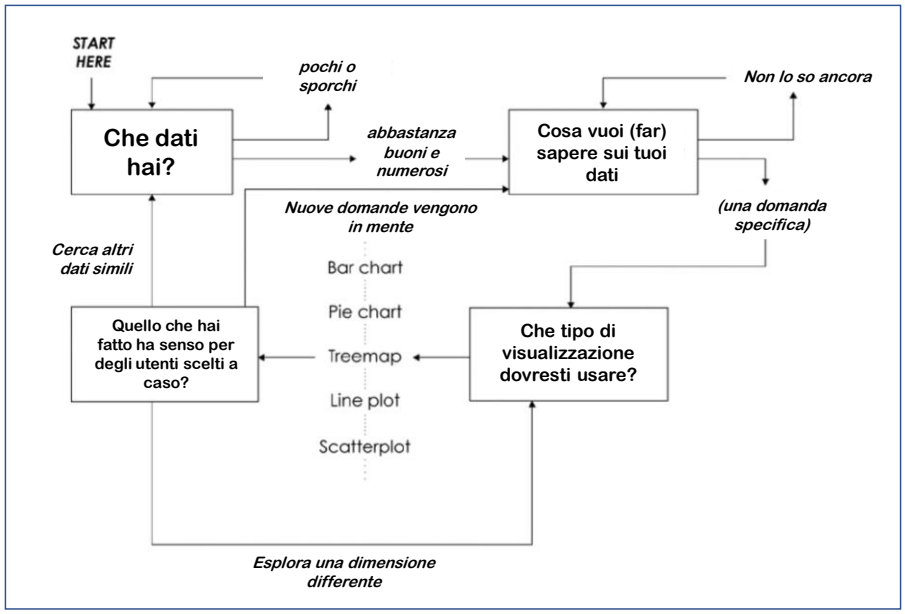


Figura 14 - Un diagramma di flusso di domande, rielaborato a partire da [Yau 2013], p. 137

re soprattutto allo studio dei dati a disposizione, e un po' di ispirazione, soprattutto riguardo alla parte grafico/visuale. Consideriamo le varie domande nella Figura 14.

<u>Che dati hai?</u>

Dalla prospettiva dei dati, si tratta di rispondere in modo onesto e preciso alla domanda “che dati hai?”. Per provare a farlo è necessario esplorarli, e per questa attività ci si può affidare a semplici grafici e diagrammi, come quelli che gli strumenti più diffusi sono in grado di generare semi-automaticamente, cioè fissando se non pochi e determinati parametri. A tal scopo, i grafici a barre, ad esempio, sono utili per capire se tutte le categorie presenti nei dati sono adeguatamente rappresentate; i grafici di densità, quelli a dispersione e gli istogrammi sono invece utili per “vedere i dati” e capire se la loro distribuzione dei dati presenta dei “buchi” (il che farebbe sospettare un problema nella loro raccolta nella misura in cui “natura non facit saltus”) o si discosta molto da una curva normale, cioè da una forma a campana; in quel caso, così come in tutti quei casi in cui i dati non seguono una qualche curva caratteristica, molte assunzioni per le quali è possibile applicare dei test statistici verrebbero a cadere, e il raggio delle azioni lecite di elaborazione dei dati per la loro analisi ed elaborazione si restringerebbe molto.

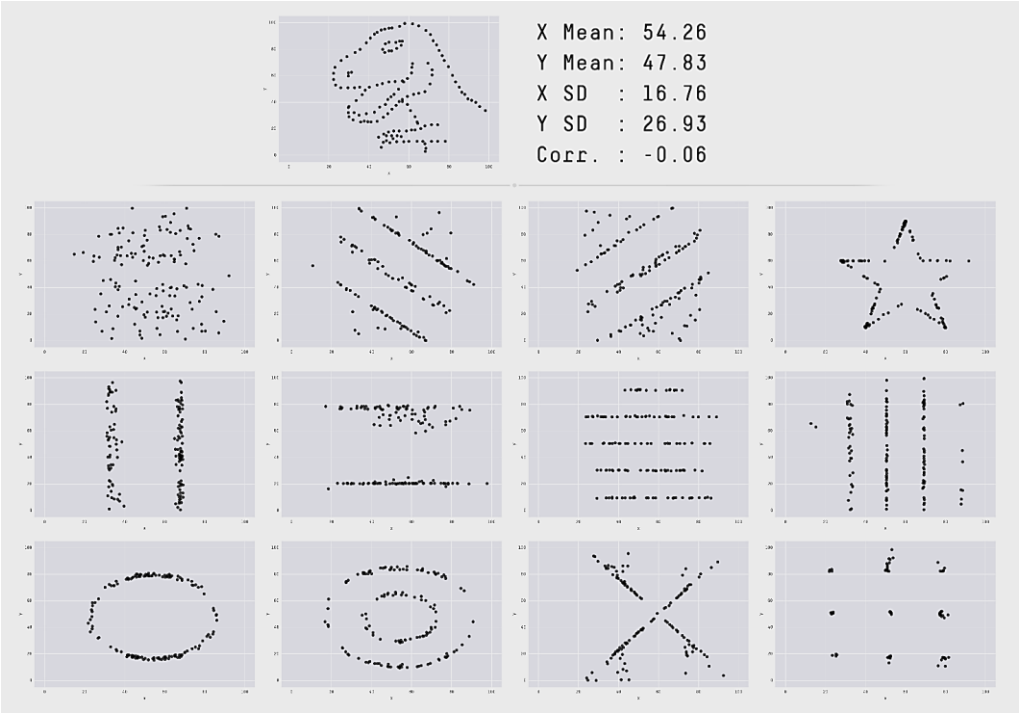


Figura 15 - La dozzina del Datasaurus. (da [Matejka & Fitzmaurice 2017])

Per questo motivo, ad un qualsiasi data scientist, la raccomandazione di “guardare i dati” il prima possibile in qualsiasi attività di interpretazione ed elaborazione dei dati non può essere fatta più convintamente: fin dagli anni Settanta infatti, gli statistici si sono resi conto che affidarsi solo a numeri e indicatori di tendenza centrale o dispersione (tra cui l’onnipresente media e la deviazione standard) per capire come siano fatti i dati a disposizione può portare a spiacevoli sorprese. Il quartetto di Anscombe (costruito appositamente dall’omonimo statistico nel 1973) sono quattro insiemi di dati che, pur essendo caratterizzati dai medesimi indicatori statistici, differiscono molto per forma, come solo un diagramma a dispersione potrebbe mostrare.

A tale riguardo a me piace citare i 12 data set costruiti al medesimo scopo dimostrativo da Alberto Cairo, tra i quali spicca il suo famoso “datasaurus”: il motivo per cui quell’insieme di dati si chiama così

lo lascio all'intuizione visiva del lettore, guardando Figura 15 nella pagina precedente. Si noti che i data set, anche se di aspetto molto diverso hanno le stesse statistiche riassuntive (media, deviazione standard e correlazione di Pearson, riportati in alto a destra)

Riprendiamo il filo del discorso. Nella fase di esplorazione è importante capire sostanzialmente due aspetti dei dati che si hanno a disposizione: primo, capire di quale tipo siano; e secondo, determinare, anche approssimativamente, il loro livello di qualità, secondo le consuete dimensioni della accuratezza, completezza e tempestività di aggiornamento. Abbiamo parlato di queste dimensioni di qualità nel Capitolo 5.

Capire di che tipo siano i dati significa determinare quali sottoinsiemi siano:

- *categorici* (o nominali), cioè categorie scelte per ciascun soggetto/istanza da un insieme predefinito (detto tassonomia o schema di classificazione); oppure
- *ordinali*, cioè categorie tra le quali è possibile definire una relazione d'ordine (come ad esempio tra le parole: sufficiente, discreto, buono, molto buono, ottimo); oppure
- *scalari*, cioè numeri e quantità che godono di certe proprietà (ad esempio quella di essere sommate tra loro, o divise per un certo numero).

Questa valutazione è fondamentale per non partire con il piede sbagliato e pensare se su certi dati sia sensato calcolare la *media* (lo è solo se i dati sono scalari e continui), o la *mediana* (non è lecito farlo sui dati categorici, sui quali invece è sensato calcolare la *moda*, cioè identificare la categoria più frequente).

<u>Cosa vuoi (far) sapere sui tuoi dati?</u>

Questa domanda, come si intuisce, è duplice: cosa il progettista vuole sapere sui suoi dati è solo il primo passo per capire cosa può *far* sapere a chi vedrà e consulterà la sua creazione, per progettare qualcosa che indirizzi questa esigenza informativa. Sebbene entrambe le domande siano concepite in forma volutamente aperta (il metodo Socrate prevede che il progettista si ponga le domande, ma non suggerisce risposte, perché è il processo di ricerca delle risposte ad alimentare una buona progettazione), in via del tutto generale è possibile delimitare il perimetro delle risposte più ricorrenti: "voglio fornire evidenza visibile di …..una qualche differenza tra cose o gruppi di individui"; "…dell' andamento nel tempo di una certa grandezza"; "…di una correlazione o associazione tra due o più variabili". Queste risposte sono generali, ma possono essere applicate a qualsiasi ambito di interesse. Ad esempio, Aaron Penne, un famoso data scientist e abile progettista di dataviz, nel 2018 si pose questa domanda e si rispose: "voglio far vedere la differenza che esiste tra come noi percepiamo una certa causa di morte e il suo reale contributo al numero totale di decessi". Il risultato è la dataviz in Figura 16. I problemi cardiovascolari sembrano essere sottovalutati rispetto al cancro. Si noti poi come due importanti giornali diano più visibilità a omicidi e terrorismo di quanto in effetti questi fattori incidano sulla vita (e morte) delle persone.

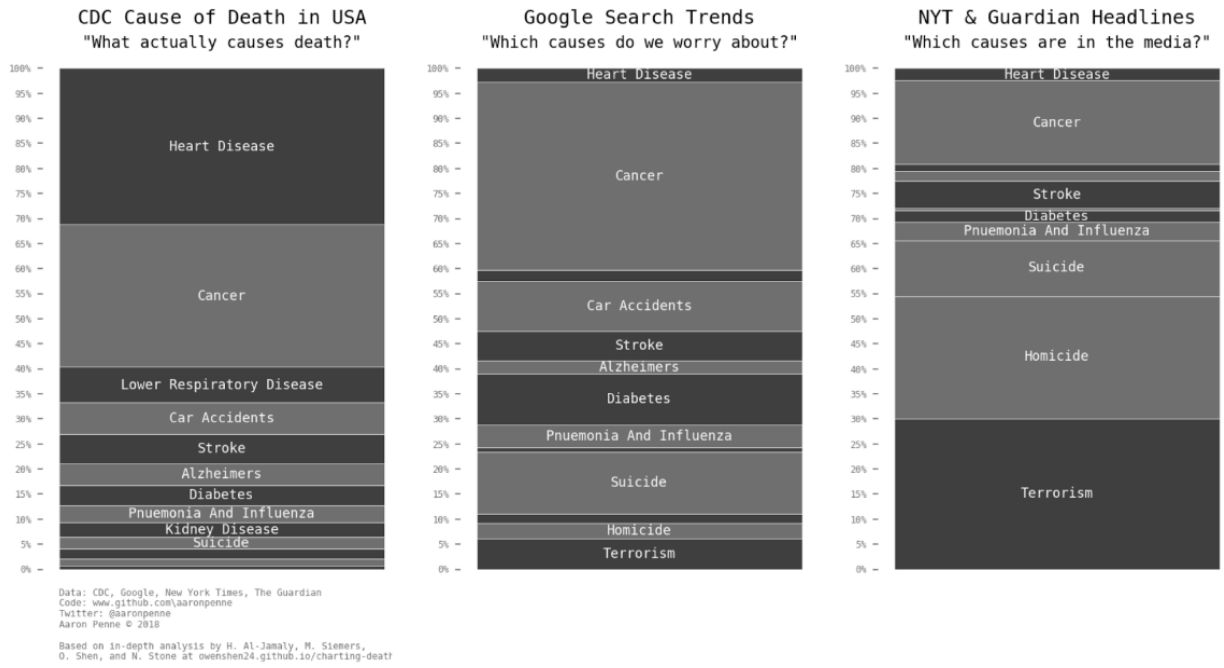


Figura 16 - Dataviz che mostra la differenza di percezione delle cause di morte.
Copyright: Aaron Penne, 2018. (Fonte: https://bit.ly/2HxItNt)

Che tipo di visualizzazione dovrei usare?

Una volta che si è capito di quali dati si dispone, che la loro qualità è buona o accettabile, e quali informazioni rappresentate possono essere messe in evidenza con una visualizzazione dati, inizia la parte delle domande più imbarazzanti, volte a capire una cosa in sé e per sé molto semplice: "che tipo di visualizzazione dovrei usare?". È difficile per un progettista non gettarsi subito nel compito del "disegno", che nel nostro caso significa concepire e dedicarsi all'elemento visuale. Ciò nonostante, l'esperienza dei maestri insegna che per farlo adeguatamente si deve essere in grado di avere risposte pronte a domande di questo tipo:

- che cosa si vuole ottenere con la dataviz? Informare, comunicare o convincere?
- che storia o narrativa c'è dietro i dati, che li ha generati, per la quale sono stati raccolti, o che aiutano a scoprire?
- E soprattutto: a chi è rivolta la dataviz? e dove sarà pubblicata? È per dei cittadini da informare nella sala d'aspetto di un ambulatorio medico tramite un opuscolo cartaceo; oppure per l'uomo da affari che legge velocemente un editoriale sul sito Web del suo giornale preferito; o ancora per il collega ricercatore sulle pagine (anche elettroniche) di un articolo scientifico?

Ogni risposta suggerisce una certa tipologia di visualizzazione e non altre.

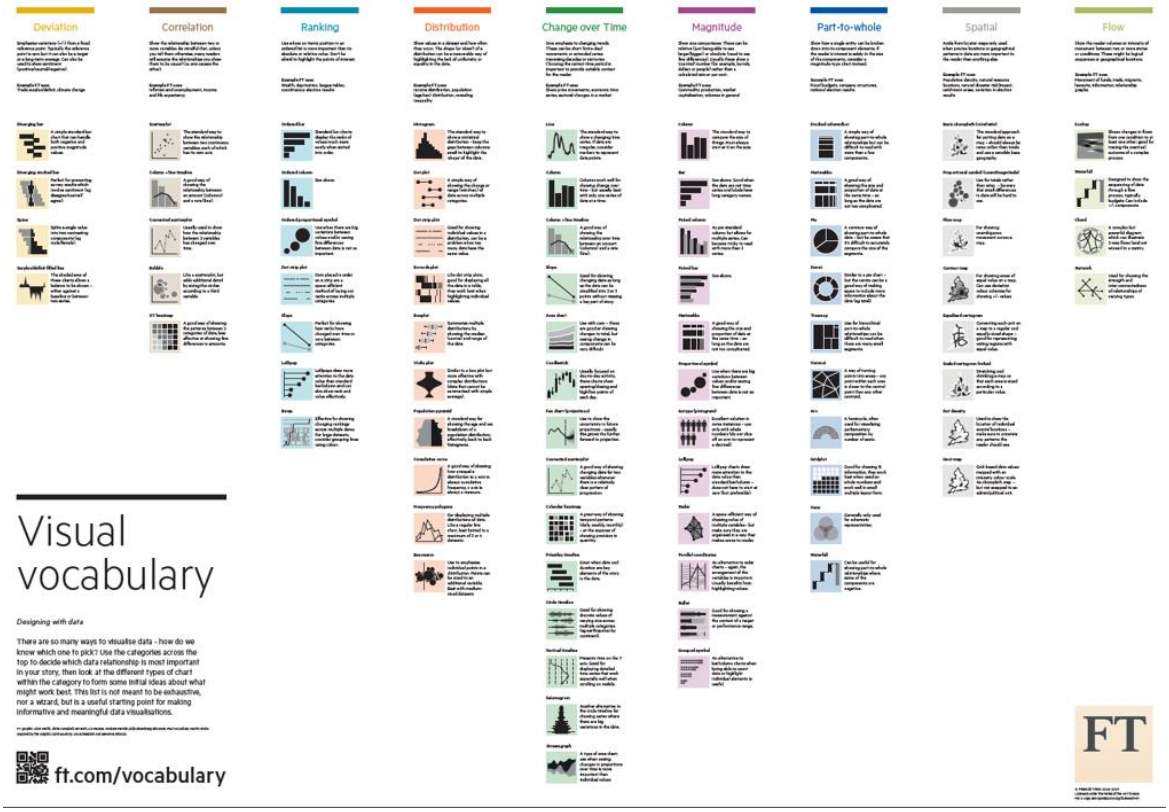


Figura 17 - "Il catalogo è questo" o almeno uno di quelli disponibili per scegliere la visualizzazione più adatta.

A supporto della scelta del tipo di visualizzazione, negli ultimi tempi la comunità accademica e professionale si è interrogata sull'utilità delle cosiddette "tassonomie di grafici", cioè cataloghi disponibili sul Web che riportano le tipologie di visualizzazione più note, descrivendole nei loro elementi principali e suggerendo gli impieghi più comuni o indicati: buoni esempi di queste risorse elettroniche sono il Dataviz project[37], il Visual Vocabulary[38], il Dataviz Catalogue[39] (che riproduciamo integralmente in Figura 17), e la risorsa "From Data to Viz"[40], solo per citare quelli che apprezzo di più.

Iniziative come queste possono essere molto utili per i principianti (o per chi come me deve spesso dare un "nome", più o meno consolidato, ad una certa tipologia di visualizzazione dati), ma sono anche accusate [Makulec, 2019] di banalizzare la scelta della visualizzazione più idonea, favorendo l'ignoranza dei principi fondamentali; riducendo il processo di progettazione ad una attività analoga alla scelta di un vestito preconfezionato; appiattendo così le modalità di rappresentazione a poche modalità tradizionali e generali; riducendo l'opportunità per l'esplorazione per modi più originali (e per questo anche di maggiore impatto grafico, come in Figura 18); o peggio ancora, favorendo il rischio di

[37] https://datavizproject.com/
[38] https://github.com/ft-interactive/chart-doctor/blob/master/visual-vocabulary/Visual-vocabulary.pdf
[39] https://datavizcatalogue.com/
[40] https://www.data-to-viz.com/caveats.html

trascurare il più ampio ecosistema di informazioni in cui una certa visualizzazione potrebbe risultare formalmente corretta ma incapace di “passare” e quindi di comunicare qualcosa o qualcuno.

Nella visualizzazione di Figura 18 Ogni cerchio indica il numero di casi di influenza (proporzionale al raggio), in diverse stagioni (a seconda della intensità d igrigio) e in diversi anni e mesi (a seconda della posizione); la dimensione temporale si dipana per spire concentriche, dalla più interna alla più esterna.

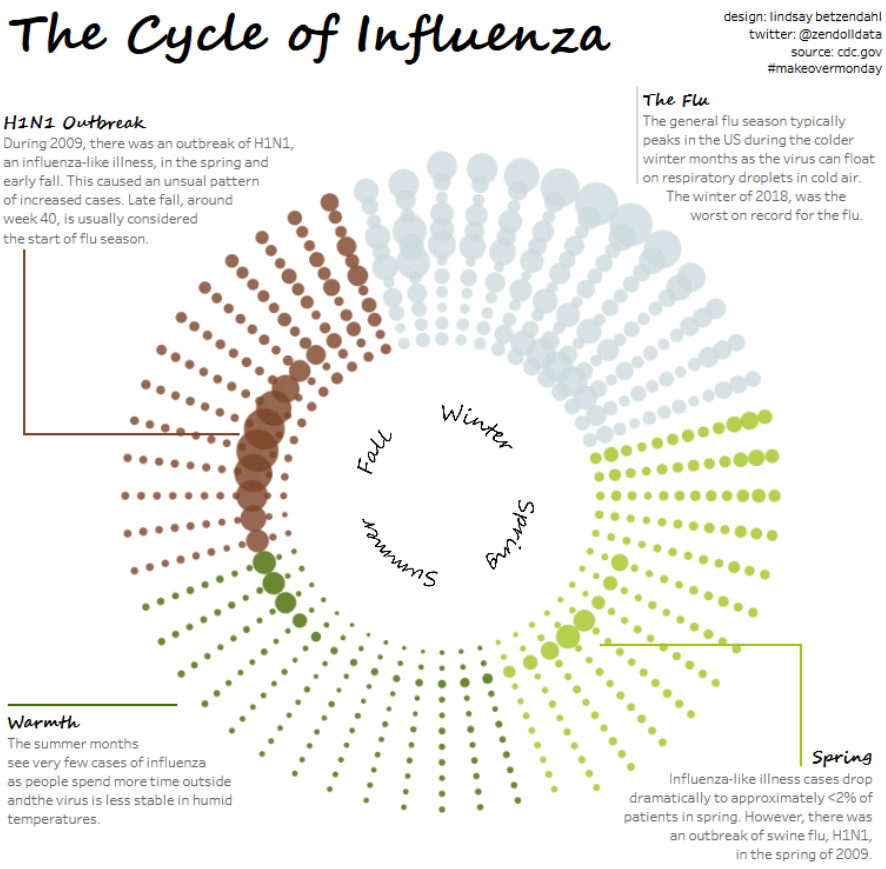


Figura 18 - Una visualizzazione originale, il cui progettista ha resistito alla tentazione di rappresentare una serie temporale con un tradizionale line chart – (Fonte: https://tabsoft.co/2mhj55H)

Sebbene anche io apprezzi e segnali agli studenti certe risorse, soprattutto per dare loro una idea della ricchezza della “tavolozza” a disposizione del progettista per fare qualcosa di meno convenzionale e più adeguato dei soliti grafici a barre o aerogrammi (che sono invece i grafici suggeriti più frequentemente dalle applicazioni foglio di calcolo quali MS Excel o Google Sheets), personalmente trovo più utile il metodo suggerito da Yau [2013]: cioè, continuare a farsi delle domande relative alle esigenze informative che motivano la raccolta dei dati o la loro fruizione.

In questo caso non andrebbe enfatizzata l’associazione tra la domanda e le possibili visualizzazioni: questa associazione deriva dall’applicazione di un numero limitato di euristiche e linee guida. Ad esempio: se si hanno serie temporali di dati scalari continui, allora è bene usare delle line chart; se invece si hanno dati categorici sono indicati i grafici a barre; i grafici a dispersione devono avere due assi continui; se una variabile è categorica allora si devono fare dei dot-plot e le due cose non vanno

confuse. Eppure, anche queste, come i cataloghi di visualizzazioni, hanno un valore limitato nella pratica professionale, quello che hanno dei semplici galleggianti per chi inizia ad imparare a nuotare: permettono solo di galleggiare.

Ciò che invece trovo utile nell'approccio a domande, il cosiddetto "metodo Socrate", è che invita il progettista a porsi nei panni di un fruitore esterno (che non sa nulla dei dati che lui ha a disposizione) e quindi in una condizione di difetto: "cosa non so e vorrei sapere?". E rispondersi ponendosi domande sempre più specifiche e vicine alle possibili esigenze o curiosità del fruitore: "cosa vorrei vedere per poter rispondere a questa domanda?"

Una volta che il progettista ha un'idea almeno vaga su quale tipo di visualizzazione dovrebbe impiegare, è molto importante "metter giù" il prima possibile le idee che si hanno, e quindi "vedere la visualizzazione", per provare direttamente l'effetto che fa.

3.4 Realizzazione

La seconda attività ha volutamente un nome generico, "realizzazione", in quanto essa può comprendere un misto di attività molto specifiche ma anche molto diverse tra loro, che hanno in comune la caratteristica di rendere "reale", o meglio visibile, l'oggetto della progettazione: queste tecniche possono quindi variare dall'uso di lavagna e pennarelli colorati (come in Figura 19, dove nella metà superiore è possibile vedere gli schizzi preparatori, e in quella inferiore le dataviz finali), all'uso di

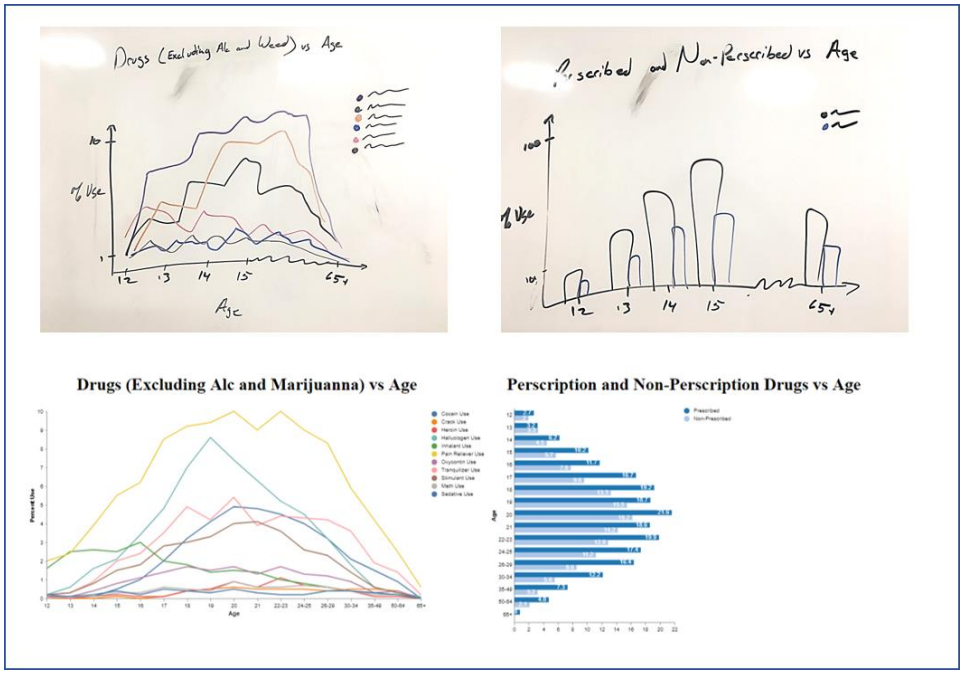


Figura 19 - Evoluzione di due visualizzazioni durante la fase di realizzazione. (da Nguyen [2017])

uno strumento professionale di elaborazione grafica vettoriale; e può comprende anche: lo sviluppo di specifici script (cioè pezzi di codice software) per creare visualizzazioni interattive molto complesse (ad

esempio utilizzando librerie software molto potenti quali matplotlib[41] e seaborn[42] per il linguaggio Python); oppure la combinazione di diversi moduli configurati attraverso uno delle molte piattaforme software che permettono la creazione e pubblicazione di visualizzazioni complesse, quali Tableau™[43], Plotly™[44] o Microsoft™ Power BI™[45], solo per citarne alcuni[46]; oppure tutte queste tecniche insieme.

La creazione di una visualizzazione interattiva è il risultato di un codice software che, anche se non deve essere sviluppato da zero, richiede comunque molto lavoro. Ciò nonostante, numerose fonti sostengono che, se si considera 10 il tempo necessario ad un data scientist per confezionare una dataviz di qualità accettabile, la maggior parte del tempo non viene investito nella scrittura del codice vera e propria, bensì nella preparazione dei dati, fino a 7-8 parti dell'intera durata, affinché essi siano adatti alla visualizzazione, o meglio affinché il codice software che genera la visualizzazione li possa "leggere", elaborare e infine visualizzare. Infatti, la fase di progettazione potrebbe avere identificato problemi a livello di dati, quali la presenza di numerosi "buchi" o errori evidenti, o l'impiego di formati diversi e incompatibili (ad esempio, relativi alle date o valute), o la presenza di misure fatte con granularità temporali e su intervalli di valori (range) diversi. In quel caso, il data scientist deve anche pulire, curare (come si dice in gergo) e "ingegnerizzare" l'insieme dei dati per renderlo più adatto ad una visualizzazione fedele e non distorta.

3.5 Valutazione (nell'uso)

Con riferimento alla Figura 14, il lettore attento avrà notato che non ho ancora parlato dell'ultima domanda, forse la più maliziosa: "Quello che hai fatto ha senso per degli utenti scelti a caso?" Per rispondere a questa domanda è necessario raccogliere, in maniera quanto più sistematica e attendibile, il giudizio dell'utente; e valutare la bontà della visualizzazione dalla sua prospettiva, cioè nell'uso: questa è l'essenza di ciò che segue la fase di realizzazione e che io ho chiamato "Valutazione", sebbene altri termini siano possibili, tra cui validazione, test e controllo qualità.

Ovviamente la migliore prova si dovrebbe fare "su strada", cioè in condizioni reali e, come si dice, ecologiche (cioè in ambienti sociali e lavorativi reali). Questo non è sempre possibile, e comunque richiederebbe molto tempo. Per ovviare a questo e non rinunciare ad una valutazione, ci si accontenta spesso di condizioni realistiche, non reali, e di simulazioni in ambienti sperimentali controllati.

Questo approccio si può avvalere di metodi e tecniche agili e veloci sviluppati all'interno delle comunità scientifiche e professionali della "interazione uomo-macchina" ed "ergonomia cognitiva". In particolare, io propongo agli studenti di immaginare dei semplici compiti che richiedono a delle persone non esperte di data visualization di usare il loro artefatto: ad esempio, stabilire quale fosse l'anno in cui

[41] https://matplotlib.org/
[42] https://seaborn.pydata.org/
[43] www.tableau.com/
[44] https://plot.ly
[45] https://powerbi.microsoft.com
[46] Penso che citare questi strumenti renda quello che scrivo estremamente sensibile ai moti del tempo, in quanto non so prevedere se tra cinque o anche dieci anni esisteranno ancora o come si saranno evoluti in piattaforme ancora più usabili e versatili.

la NASA ha assorbito il budget maggiore per le sue missioni; o la stagione in cui Messi ha segnato più gol non considerando i gol da calcio piazzato; e così via: cose che non siano immediatamente mostrate dalla dataviz, ma che richiedano all'utente di utilizzare la visualizzazione in modi non banali né superficiali, come ad esempio filtrare alcuni dati o usare altri controlli di interfaccia per guardare certe parti dei dati con maggiore dettaglio.

La sessione di valutazione è un momento delicato e che di solito appassiona molto gli studenti, anche se sulle prime se ne sentono atterriti: il loro compito è coinvolgere delle persone in carne ed ossa, dei "tester", solitamente almeno una dozzina (ma più sono meglio è) e possibilmente fuori dalla cerchia degli amici più prossimi (e quindi più simili a loro). Devono spiegare a questi "estranei" in modo breve e chiaro la finalità del test, sottolineando in quella occasione il fatto che durante il test non saranno sotto esame loro, bensì la visualizzazione; descrivere i compiti in cui le persone dovranno cimentarsi; e infine, far eseguire i compiti, prendendo appunti e registrando l'intera sessione sperimentale per un suo esame più approfondito a posteriori.

In un esperimento tipo, il tester è dapprima lasciato libero di esplorare la visualizzazione senza interferenze, ma è pregato di "pensare a voce alta" (questa tecnica di analisi si chiama *think-aloud protocol*) in modo tale che eventuali incomprensioni o perplessità siano facilmente associabili da parte dei valutatori (cioè dagli studenti) a momenti precisi di interazione o parti della interfaccia; in un secondo momento al tester viene spiegato cosa fare e l'esecuzione del compito viene cronometrato e registrato come successo/insuccesso a seconda che il tester sia stato in grado di fornire una risposta corretta o abbia comunque eseguito tutte le operazioni richieste.

Sia il tempo di esecuzione che il tasso di errore sono "misure" analizzate quantitativamente per valutare la *efficacia e la efficienza* e, rispettivamente, delle dataviz rispetto ai compiti assegnati, cioè la loro capacità di informare adeguatamente (efficacia) e più velocemente del testo scritto (efficienza[47]). In particolare, tali misure sono usate su base aggregata, per capire se e quanto le medie dei tempi di esecuzione, o i tassi di errore esibiti dal campione coinvolto, differiscano da delle soglie di ottimalità convenzionali[48]; oppure dalla media e tassi registrate per versioni leggermente diverse della medesima dataviz. Quest'ultimo tipo di confronto è molto informativo, perché permette di scegliere e concentrarsi su soluzioni che sono oggettivamente migliori delle altre, sebbene richieda ovviamente un maggior sforzo di sviluppo ai progettisti, e sessioni di test più lunghe ai tester.

Anche la prima fase del test, quella della raccolta delle impressioni a caldo, è molto informativa e può tradursi in una esperienza che mette a dura prova gli studenti: è spesso un bagno di umiltà e realtà, in cui soluzioni a cui potevano essere affezionati come progettisti, per la loro originalità o complessità, o su cui hanno passato lunghe ore per farle funzionare come sviluppatori, risultano, alla prova dei fatti, totalmente incomprensibili o inutilizzabili agli utenti.

[47] In particolare, in rete circola la notizia che "le visualizzazioni sono elaborate 60.000 volte più velocemente del testo." [iScribbers, 2019], che è di difficile verifica. Se fosse vero, potremmo riformulare il famoso detto in "una immagine vale più di 60.0000 parole", almeno dal punto di vista della efficienza.

[48] La soglia dell'efficacia, cioè il tasso di errore ottimale può essere fissato al 1% o al 5%, a seconda della complessità della dataviz e del compito; l'efficienza ottimale è invece solitamente fissata empiricamente sulla base di quanto velocemente riesce a svolgere i loro compiti chi conosce bene la dataviz, ad esempio i progettisti stessi, in diversi tentativi.

Infine, le registrazioni, ovviamente raccolte con il consenso dei tester, permettono ai valutatori di prendere appunti e identificare i cosiddetti "problemi di usabilità", termine a cui si dovrebbe dare l'accezione più ampia: se un utente non ha capito che significa una banda colorata intorno a certi pallini, e l'ha confessato senza remore durante la prova su strada della dataviz, questo è un campanello d'allarme che richiede più di una riflessione: si è sbagliato colore? Manca un commento chiarificatore? Si è usato un termine troppo tecnico?

Da ultimo, gli utenti sono invitati a compilare un breve questionario psicometrico, a risposte chiuse, in cui valutare la loro esperienza d'uso della dataviz. A tal riguardo, da un po' di anni propongo loro di utilizzare un questionario molto breve che abbiamo sviluppato (e poi parzialmente ripreso in [Locoro et al. 2017]) nel novero delle attività scientifiche condotte in un progetto di cui ero il responsabile e che è stato finanziato dalla Regione Lombardia, incentrato sul "Valore sociale della visualizzazione degli open data di Regione Lombardia".

Il progetto ha prodotto un sito di dataviz interattive che possono essere ancora consultate e provate[49]. Il questionario che abbiamo sviluppato consta di sole 6 domande a risposta chiusa, in cui l'utente è invitato a riferire il grado di Utilità, Intuitività, Chiarezza, Informatività e Bellezza che ha percepito durante l'uso della dataviz, in una scala a differenziale semantico di sei livelli, dal primo, "pochissimo" all'estremo opposto, "moltissimo"; e infine indicare anche il valore complessivo della dataviz sulla medesima scala.

È interessante notare come queste dimensioni di qualità contribuiscano diversamente al valore complessivo delle dataviz fatte analizzare a campioni di utenti molto assortiti ed eterogenei: in genere una buona dataviz è quella che ottiene buoni risultati lungo tutte queste dimensioni, e quale sia il contributo di ciascuna di esse al valore complessivo dipende dalla singola dataviz e da logiche che non possono essere generalizzate facilmente.

Ciò detto, in uno studio che abbiamo condotto nel progetto succitato abbiamo mostrato ad un campione molto eterogeneo di 618 persone una coppia di dataviz scelta tra tre coppie. A ciascuna coppia di dataviz era associato un diverso compito informativo in ambito sanitario in scenari di diversa urgenza e serietà, quale quello di dover trovare l'ospedale migliore dove farsi ricoverare per risolvere un problema di lombalgia di lunga data; oppure quello dove portare la propria figlia di un anno di età soggetta ad un episodio di febbre molto alta. In ciascuna coppia di dataviz mostrata, una era stata progettata per essere innovativa e utilizzare codici linguistici relativamente nuovi; e l'altra era di tipo più tradizionale (da catalogo, per così dire); in un caso si trattava addirittura di una semplice tabella.

La eterogeneità dei compiti informativi, delle visualizzazioni mostrate, e del campione di persone coinvolte (molte delle quali erano passanti in una piazza molto affollata di Milano) era ovviamente voluta, ed un aspetto molto delicato del disegno sperimentale, che era volto a massimizzare la generalità delle evidenze trovate, al di là della semplice significatività statistica.

Il risultato che abbiamo estratto da questo studio è che il maggior contributo al valore complessivo lo dà la *Chiarezza*, seguita dalla *Utilità* e dalla *Bellezza*. Queste dimensioni, soprattutto le prime due, sono

[49] Il progetto è disponibile al seguente indirizzo Web: http://calib.ro/hdil-polls/#!/home

solitamente molto correlate tra loro, con un indice di correlazione tra il 75% e l'85% (il 100% è ovviamente il massimo e una correlazione del 40% è già considerata molto alta in molti contesti psicometrici). L'informatività sembra essere stata la dimensione meno importante nel giudicare il valore di una infografica, e questo risultato appare ragionevole se si pensa che i compiti e i bisogni informativi erano sì stati concepiti per essere di esperienza comune, ma erano pur sempre indotti sperimentalmente, e quindi non realmente sentiti dai soggetti coinvolti.

Queste dimensioni, che nel questionario sono espresse anche con una serie di sinonimi e locuzioni descrittive per non ingenerare fraintendimenti legati ai termini usati, tutte insieme "spiegano" più di tre quarti della variabilità della dimensione del valore. Questo è un risultato notevole, se si pensa che è abbastanza plausibile che molto del valore dipenda *anche* da elementi contestuali e legati alle esigenze situate delle persone coinvolte (in Figura 20 è possibile vedere diversi modi in cui visualizzare i risultati del nostro questionario, in termini di distribuzioni e correlazione tra le dimensioni).

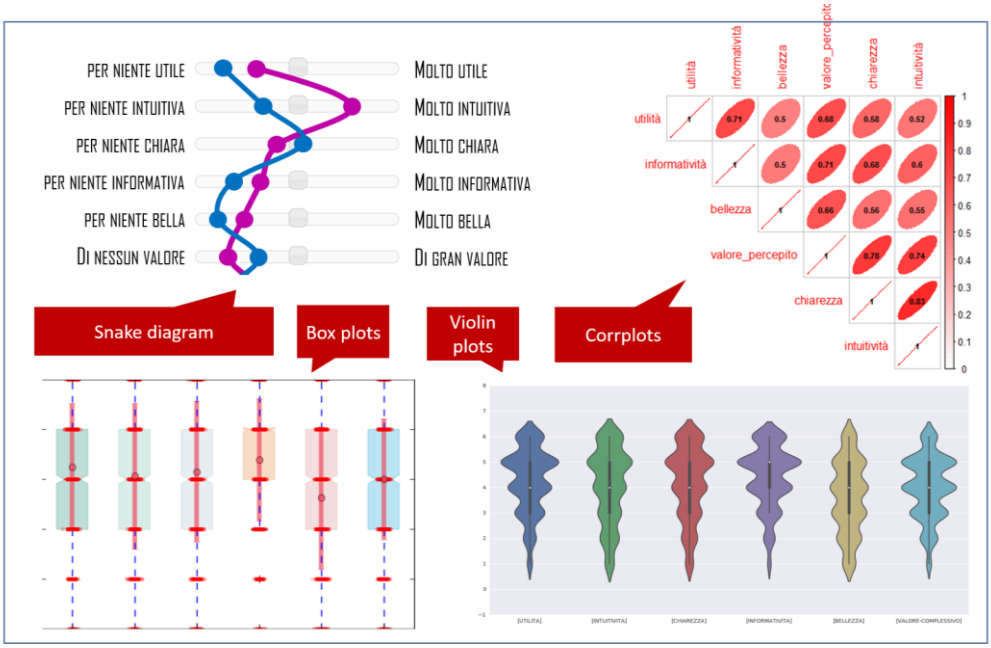


Figura 20 - Diverse modalità di visualizzazione per mostrare l'apporto di diverse dimensioni di qualità al valore totale percepito.

La ricerca appena commentata verrà ripresa nel Capitolo 14, in cui l'attenzione si sposterà sul valore sociale inteso come utilità del dato per risolvere le sue esigenze per un problema di salute.

3.6 Miglioramento

In questa fase la dataviz dovrebbe essere modificata sulla base di come sono andati gli esperimenti della fase precedente. Idealmente tutto il processo andrebbe ripetuto qualche volta, o finché gli utenti dimostrino concretamente di saper usare proficuamente la visualizzazione per informarsi e apprendere qualcosa di nuovo. Porre l'accento sull'aspetto semiotico e interazionale delle visualizzazioni porta su

un secondo piano gli aspetti più legati alla specifica implementazione, e quindi ad eventuali difetti sul piano delle prestazioni e delle caratteristiche funzionali del prodotto.

In un mondo perfetto tutti gli aspetti sono importanti, ma in mancanza di risorse (e quindi spesso di tempo) il mio messaggio agli studenti è quello di previlegiare gli aspetti relativi all'*uso*: cerco di convincerli che sia meglio richiedere un po' di pazienza o comprensione all'utente sugli aspetti precipuamente tecnici, piuttosto che incorrere in fraintendimenti sul senso della visualizzazione, o lasciare inespresse alcune delle sue potenzialità, solo perché certi aspetti, apparentemente marginali dal punto di vista del progetto, non sono adeguatamente presentati nel "linguaggio dell'utente".

La sfida più difficile, in questo senso, è produrre un oggetto di design visuale che, da una parte, risulti nuovo, intrigante, in grado di sollecitare la curiosità e spingere l'utente alla esplorazione; dall'altra, un oggetto che non adotti soluzioni troppo innovative o che rompano convenzioni di lettura e interazione; ciò perché certe novità potrebbero avere l'effetto contrario alle intenzioni sulla qualità dell'esperienza d'uso, e quindi ingenerare nell'utente incomprensione o sensazioni di frustrazione. Alla luce di un approccio semiotico alla data visualization, la cosa più importante è riuscire a comunicare qualcosa che può riguardare da vicino le persone coinvolte, siano queste degli scienziati interessati ai risultati di un esperimento, o dei cittadini interessati a capire, ad esempio, se le retribuzioni sono cresciute di pari passo con la produttività oppure no (si veda Figura 21).

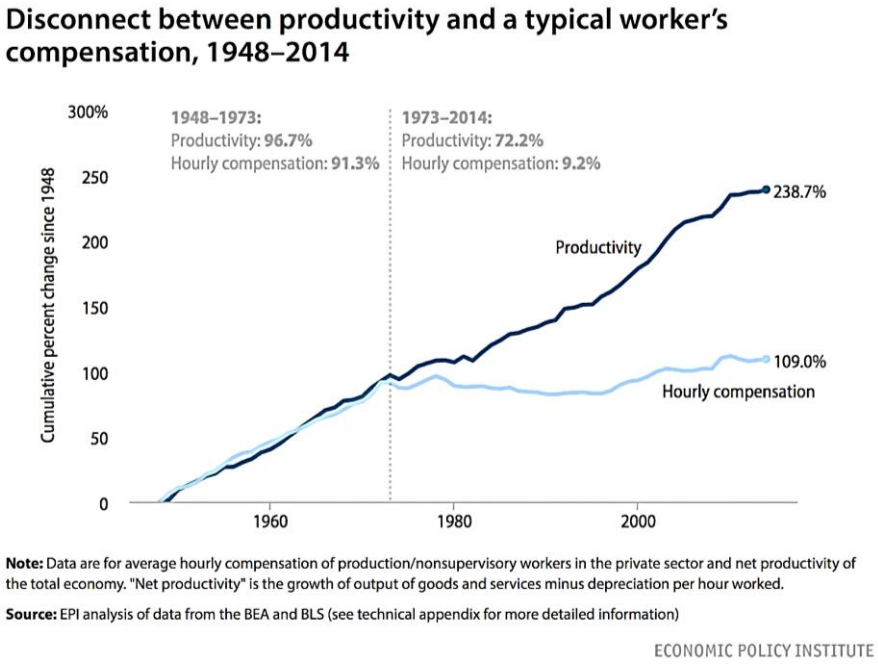


Figura 21 - Divaricazione tra produttività e salari negli ultimi sessant'anni.

8. Conclusioni

Per concludere questo capitolo sulle basi della Sata visualization nell'ambito più ampio della Scienza dei dati, vorrei riprendere il messaggio principale e infine riportare un aneddoto personale in cui questo

messaggio mi è divenuto molto chiaro. La visualizzazione dei dati, cioè rendere i dati più visuali, è una pratica su cui incidono diverse competenze: un “saper fare” più tecnico, legato a metodi e strumenti; e una metodologia così scarna e sotto-specificata da fungere più come promemoria che come procedura predefinita e ripetibile in maniera sistematica.La competenza più importante richiede di pensare *semioticamente*, cioè considerando l’artefatto grafico/visuale prodotto come un mezzo per innescare e facilitare processi di interpretazione, comprensione e comunicazione, processi che ovviamente includono aspetti linguistici (il codice visuale), ma anche il contesto di fruizione e le capacità interpretative del fruitore.

Questo richiede di vedere (e perseguire!) la pratica della visualizzazione dei dati come un fare progettuale che interroga il mondo e le persone in modo analogo a come la Scienza dei dati interroga il mondo dei fenomeni per registrare informazioni codificabili: con *rigore e precisione*, certo, ma anche *onestà e apertura verso ciò che non si conosce e non ci si aspetta*. Come ha scritto il già citato Herbert Simon “lo studio corretto della umanità coincide con la scienza della progettazione”.

Durante il progetto per la Regione Lombardia menzionato poc’anzi, mi ero prodigato molto per la realizzazione di una dataviz (la numero 1 in Figura 22) intesa a supportare la decisione di una persona che, nello scenario proposto, dovesse scegliere in quale pronto soccorso portare un proprio caro che stesse soffrendo di dolorose coliche renali ma insistesse (con la ragionevolezza tipica di una situazione del genere) per andare nell’ospedale dove operavano specialisti riconosciuti per quel tipo di disturbi.

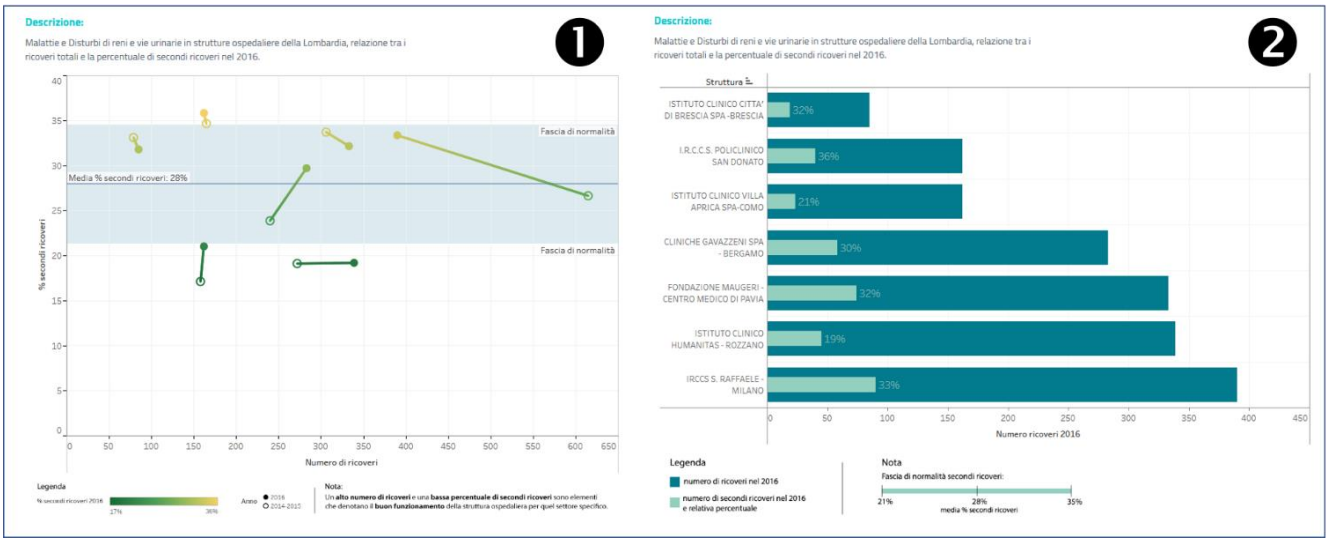


Figura 22 - Le due visualizzazioni in competizione tra loro per informare il cittadino su quali siano gli ospedali migliori per farsi trattare una colica.

La dataviz 1 rappresenta ogni ospedale con un piccolo cerchio pieno, posto in un piano dove la dimensione verticale indica la percentuale di secondi ricoveri, cioè di ricoveri non pianificati che si rendono necessari poco dopo il primo ricovero, solitamente per qualche complicanza occorsa durante il primo; e la dimensione orizzontale indicava il numero di ricoveri.

Entrambe le dimensioni sono considerate indicatori indiretti ma fedeli della qualità della cura erogata in una struttura ospedaliera (assumendo che la tipologia di paziente medio che le strutture considerate trattano non vari molto tra un centro e l'altro): meno secondi ricoveri (in percentuale) occorrono, più risolutivi sono i primi ricoveri per un determinato problema; più casi di un certo tipo sono trattati in un centro, e maggiore è l'esperienza dei medici che vi lavorano. Insomma, quel piano rappresentava uno "spazio" della qualità dei centri ospedalieri per i disturbi renali e quindi: più un cerchio vi era rappresentato in basso e a destra, e più il centro corrispondente poteva essere considerato raccomandabile per il trattamento di quei disturbi.

Non pago della ricchezza di informazioni che anche solo questa indicazione potesse dare, mi ero inoltre immaginato che ogni centro fosse in realtà rappresentato da due cerchi uniti da un segmento: un cerchio pieno, ad indicare la rilevazione più recente del numero di ricoveri e secondi ricoveri; e uno vuoto, ad indicare lo stesso dato due anni prima. In questo modo, i segmenti che unissero questi due cerchi e fossero inclinati verso il basso indicavano che i centri corrispondenti fossero anche migliorati negli ultimi anni; e lo fossero tanto di più quanto più fossero stati lunghi, a rappresentare il fatto che avessero attuato politiche che erano risultate efficaci nel miglioramento della cura. Questo è ciò che viene in effetti rappresentato in Figura 22.

Inutile dire che fossi molto soddisfatto della mia dataviz, convinto com'ero che fosse in grado di sintetizzare elegantemente molte informazioni in un ambito molto delicato come quello della valutazione della qualità delle cure. Ero molto curioso di vedere come sarebbe stata considerata, non solo in termini assoluti ma anche rispetto ad una dataviz più tradizionale, come quella denotata con il numero 2 in Figura 22, che, a mio parere, era più confusionaria, sebbene (o forse proprio perché) si limitasse ad impiegare semplici barre colorate.

Devo dire che mi fu difficile nascondere la mia delusione quando vidi i risultati dello studio sperimentale: rifeci anche i conti tre volte, per sincerarmi che non avessi fatto qualche errore nell'elaborazione dei dati. Meno di un terzo dei rispondenti trovò la mia dataviz intuitiva, e ancor meno la trovò chiara, a fronte di più di 8 persone su 10 nel caso del grafico a barre.

Pur ammettendo che la mia visualizzazione potesse richiedere un qualche sforzo interpretativo, soprattutto a chi non fosse aduso ai diagrammi a dispersione (di cui quella era una variazione elegante), ero però certo che lo sforzo avrebbe ripagato in termini informativi, e che la bellezza della dataviz non sarebbe passata inosservata: ebbene, anche riguardo alla bellezza, appena un terzo dei rispondenti considerò la mia dataviz di una bellezza accettabile, contro più del doppio di persone che trovarono bella l'altra. La maggioranza assoluta dei rispondenti considerò infine la mia dataviz di basso valore, mentre il grafico a barre piacque (anche nel senso di fu "trovata utile") in maniera quasi plebiscitaria, da quasi nove soggetti su dieci (si veda Figura 23 per i risultati).

Che morale trassi da questa esperienza? Avevo creato una dataviz che trovavo bella, elegante e gratificante, pur a fronte del piccolo sforzo che avrebbe richiesto al lettore per decifrarne il codice visivo, e che ero convinto sarebbe risultata molto chiara a chi, come me, avesse una buona familiarità con grafici e diagrammi cartesiani: ero convinto che un mio collega, o comunque un esperto di

visualizzazioni o dati sanitari, avrebbe potuto apprezzare il mio sforzo di comunicare qualcosa in modo visuale e accattivante, rispetto alla piattezza di un grafico a barre.

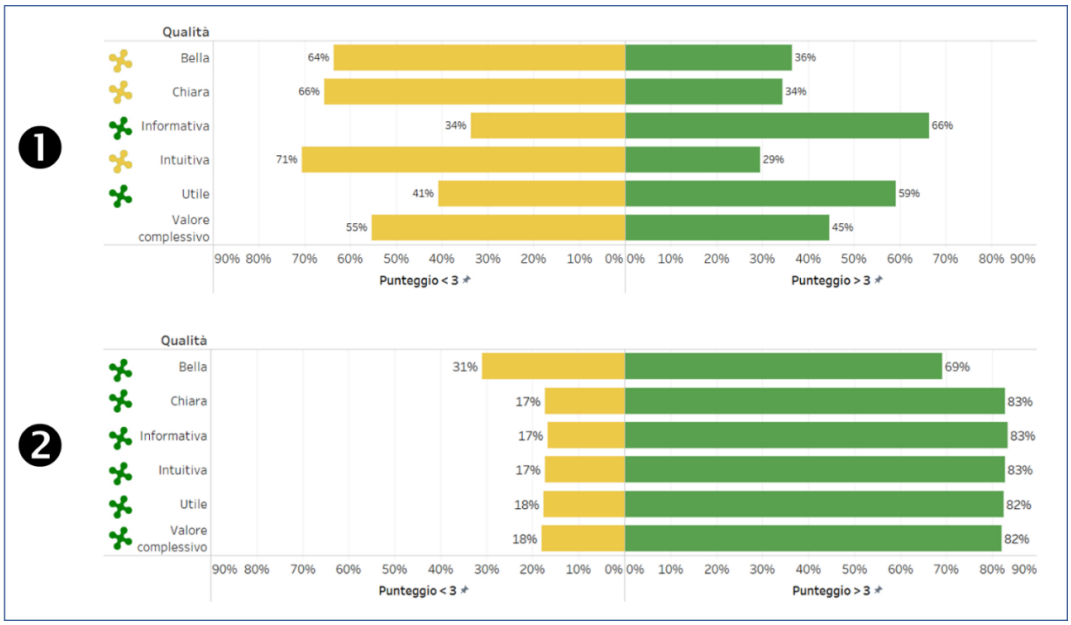


Figura 23 - I risultati delle valutazioni di valore delle due dataviz rappresentate nella figura precedente, rigorosamente rappresentati sotto forma di barre, a dimostrazione di aver imparato la lezione. (Fonte: Dessì, 2018)

Ciò nonostante, non mi ero posto due delle domande più importanti che ho citato in questo capitolo: *perché voglio visualizzare questi dati, e per chi?* Intendevo davvero solo gratificare il senso estetico di uno scienziato dei dati? E davvero volevo aiutarlo nel momento in cui avesse sentito l'esigenza di cercare quella dataviz, cioè quando lui o un suo caro avesse provato un dolore lancinante alla schiena e avesse avuto bisogno di cercare il migliore ospedale dove farsi curare?

Le visualizzazioni di dati sono il modo in cui i dati, che sono come semplici note che si accumulano negli spartiti delle nostre macchine e dei nostri libri, possono comunicare con noi, attraverso i segni e i codici che possiamo capire più velocemente e in maniera meno mediata da convenzioni e nozioni specifiche: sono per l'occhio, quello che la musica è per le nostre orecchie: una esperienza che noi, e soltanto noi, possiamo riempire di senso e che, se l'ascoltiamo attentamente, può aprirci la mente a pensieri stupendi e modi nuovi di guardare il mondo.

Riferimenti

F. Aldrich, & Sheppard, L. Graphicacy; The fourth 'R'? Primary Science Review, 64, 8–11, 2000

R. Arnheim - Visual Thinking. University of California Press, Berkeley, CA, 1969

A.Cairo - The truthful art: data, charts, and maps for communication. New Riders, 2016

A.Cairo - How Charts Lie: Getting Smarter About Visual Information. W W Norton & Co Inc, 2019.

E.A. Chiqui - Quick Guide to Spotting Graphics That Lie. National Geographic. 2015. URL: https://www.nationalgeographic.com/news/2015/06/150619-data-points-five-ways-to-lie-with-charts/. Acceduto il 13 Settembre 2019. Archiviato su: http://archive.is/hhKhu

L. Dessì - Indagine sul valore sociale degli Open Data: il caso della Regione Lombardia. Tesi di Laurea Magistrale in Teoria e Tecnologia della Comunicazione. A.A. 2017-2018. Università degli Studi di Milano-Bicocca.

R. Falcinelli - Critica portatile al visual design. *Torino: Einaudi*, 2014

G. Gigerenzer - Quando di numeri ingannano: Imparare a vivere con l'incertezza. Raffaello Cortina, 2003.

D. Huff - How to lie with statistics. WW Norton & Company, 1993.

P. Liu et al. - The utility of fat mass index vs. body mass index and percentage of body fat in the screening of metabolic syndrome. BMC public health, 13.1: 629, 2013

A.Locoro, A., Cabitza, F., Actis Grosso, R. & Batini, C. - Static and interactive infographics in daily tasks: A value-in-use and quality of interaction user study. Computers in Human Behavior, 2017, 71: 240-257.

A.Makulec - Pros and Cons of Chart Taxonomies. Medium. URL: https://medium.com/nightingale/the-pros-and-cons-of-chart-taxonomies-5c8e094578c4. Acceduto il 14 Settembre 2019. Archiviato su: http://archive.is/ujtVo, 2019.

J.F. Matejka, Fitzmaurice, G. - Same Stats, Different Graphs: Generating Datasets With Varied Appearance And Identical Statistics Through Simulated Annealing. In: Proceedings Of The 2017 Chi Conference On Human Factors In Computing Systems. Acm, 2017

K. Nguyen - Two Approaches to Data Visualization — Understanding and Persuasion. Medium. 2017 URL: https://medium.com/bucknell-hci/a-detail-implementation-of-two-approaches- to-data-visualization-design-for-understanding-and-f2ae307b360. Acceduto il 13 Settembre 2019.

B. E. Rogowitz, Treinish, L. A., e Bryson, S. - How not to lie with visualization. Computers in Physics, 10(3), 268-273. 1996

H. Simon, H. - The Sciences of the Artificial, MIT Press, 1996.

E.R: Tufte - The visual display of quantitative information. Cheshire, CT: Graphics press, 2001.

T. Unwin, T. Martin e H. Hofmann - Graphics Of Large Datasets: Visualizing A Million. Springer-Verlag New York, 2016.

C. Ware - Information visualization: perception for design. Elsevier, 2012

N. Yau - Data points: visualization that means something. John Wiley & Sons, 2013.

N. Yau - How to Spot Visualization Lies. FlowingData. 2017. URL: https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/ Acceduto il 13 Settembre 2019. Archiviato su: http://archive.is/HdbEA

Capitolo 12 – Le Astrazioni

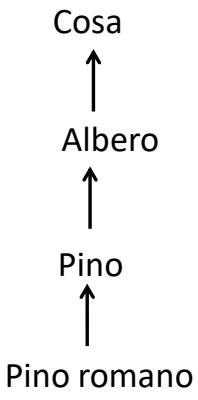
Carlo Batini

1. Introduzione

Immaginate di essere a Roma con un amico e supponete di visitare un'area archeologica (vedi Figura 1): se volete indicare il soggetto principale nella foto nella Figura 1.a, quale espressione scegliereste tra:
- Guarda quella cosa;
- Guarda quell'albero;
- Guarda quel pino;
- Guarda quel pino romano?

Provate a rispondere alla domanda…..

a. Una foto di un'area archelogica a Roma



b. Differenti generalizzazioni dello stesso oggetto

Figura 1 – Un'astrazione nella nostra vita di ogni giorno (da http://dryades.units.it/Roma/)

I quattro termini sono tra di loro in una relazione di generalizzazione; *Cosa* esprime un concetto intuitivamente più generale e astratto di Albero, Albero più astratto di Pino e Pino di Pino Marittimo; nel passare da un termine all'altro, ogni volta circoscriviamo sempre di più gli oggetti del mondo reale che corrispondono al termine usato.

Tipicamente, per indicare un'oggetto, noi scegliamo un termine, nel caso specifico tra i quattro in Figura 1.b, a seconda della nostra cultura generale e del lessico a noi noto. Probabilmente, non sceglieremo il termine *Cosa*, a meno che il nostro vocabolario non sia limitato, mentre scegliamo tra Albero, Pino e

Pino romano a seconda della nostra cultura in botanica. Possiamo affermare che nel denotare un oggetto della realtà noi, spesso inconsapevolmente, effettuiamo una operazione di *astrazione*.

Un secondo esempio di astrazione è il concetto di prezzo. Nel suo libro [Mayer-Shonberger 2018], Victor Mayer-Shonberger discute il ruolo del prezzo come formidabile fattore di sviluppo dei mercati monetari, che costituiscono il nucleo dell'attività economica in tutto il mondo. Possiamo dire che il prezzo di un oggetto o di un servizio è un'astrazione di una grande quantità di caratteristiche dell'oggetto. La sua conoscenza semplifica enormemente le transazioni tra un acquirente e un venditore, rispetto alle precedenti forme di scambio come il baratto, dove i due giocatori hanno la necessità di concordare ogni volta le unità o le quantità da scambiare di beni diversi.

Un terzo esempio, più ampio, da cui partire per iniziare a indagare il concetto di astrazione riguarda un progetto che ho portato avanti quando ho lavorato all'Aipa, Autorità per la Informatica nella Pubblica Amministrazione. Nel 1995 abbiamo organizzato una rilevazione dello stato dei sistemi informativi nelle Pubbliche Amministrazioni Centrali che ha riguardato le strutture organizzative amministrative, i processi amministrativi, le applicazioni software, le basi di dati e le relazioni tra tutti i precedenti temi.

Riguardo alle basi di dati, censimmo circa 400 basi di dati, e con l'aiuto di cinque bravissimi borsisti fummo in grado di produrre, a partire dagli schemi logici relazionali (vedi Capitolo 3) i corrispondenti schemi concettuali descritti nel modello Entità Relazione. Ricordo ancora quando posammo sul pavimento di una grande stanza della sede Aipa dell'Eur i 400 schemi rappresentati mediante diagrammi, senza un ordine particolare. Chiaramente il nostro lavoro non poteva finire lì, cosa ce ne facevamo di un volume di 400 pagine in ognuna delle quali era rappresentato uno schema concettuale?

Mi venne inizialmente in mente che dovevamo cercare di integrare gli schemi, producendo un unico grande schema concettuale di tutte le tipologie di dati trattati dalla Pubblica Amministrazione Centrale, vedi Figura 2; solo così saremmo stati in grado di avere una visione unitaria dell'intero patrimonio informativo della PA centrale, trovare entità rappresentate in più schemi, trovare gerarchie Is-a tra entità in diversi schemi (per esempio Soggetto Fiscale Is-a Cittadino), ecc. Ma quanto darebbe stato grande questo schema? Facemmo un rapido calcolo e assumendo che in media i singoli schemi contenessero circa 15 entità e altrettante relazioni, il numero di entità e di relazioni dello schema integrato era di circa 6.000! Avevamo bisogno di un foglio grande come la intera stanza, ma, soprattutto, era impossibile pensare di disegnare uno schema così grande.

Qui mi venne una seconda idea; perchè non provavamo a suddividere gli schemi in gruppi secondo aree tematiche omogenee, come ad esempio le aree Personale, Istruzione, Fisco, ecc., cercando di produrre gruppi non tanto grandi? Costruimmo una trentina di gruppi di schemi, ognuno con circa 14/15 schemi. A questo punto provammo a costruire per ogni gruppo di schemi il corrispondente schema concettuale. Ogni schema, però, aveva circa 200 entità e 200 relazioni, "eravamo da capo a dodici", come si diceva una volta.

Mi venne una terza idea: partendo da un generico schema concettuale, dovevamo avere la possibilità di produrne uno più compatto, in modo tale da poterlo poi integrare con altri; chiamammo questa operazione con il nome di astrazione. Nel produrre uno schema astratto, dovevamo cercare di individuare i concetti più importanti dello schema di partenza, e rappresentare solo questi nello schema

astratto. A questo punto, dovevamo capire come applicare congiuntamente le operazioni di integrazione e astrazione. Vediamo a questo proposito la Figura 3, in cui ho volutamente scelto degli esempi “giocattolo”.

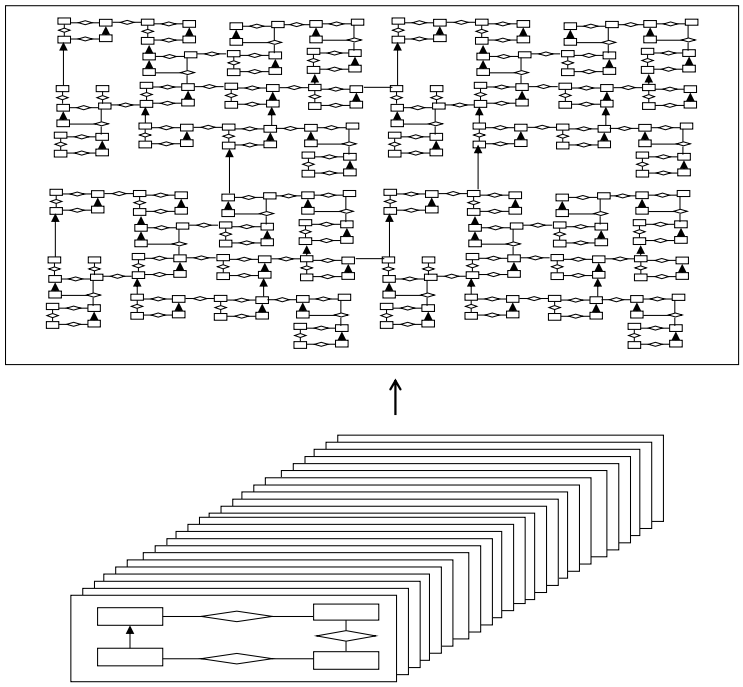


Figura 2 - Problemi nel rappresentare il risultato della integrazione di un grande numero di schemi

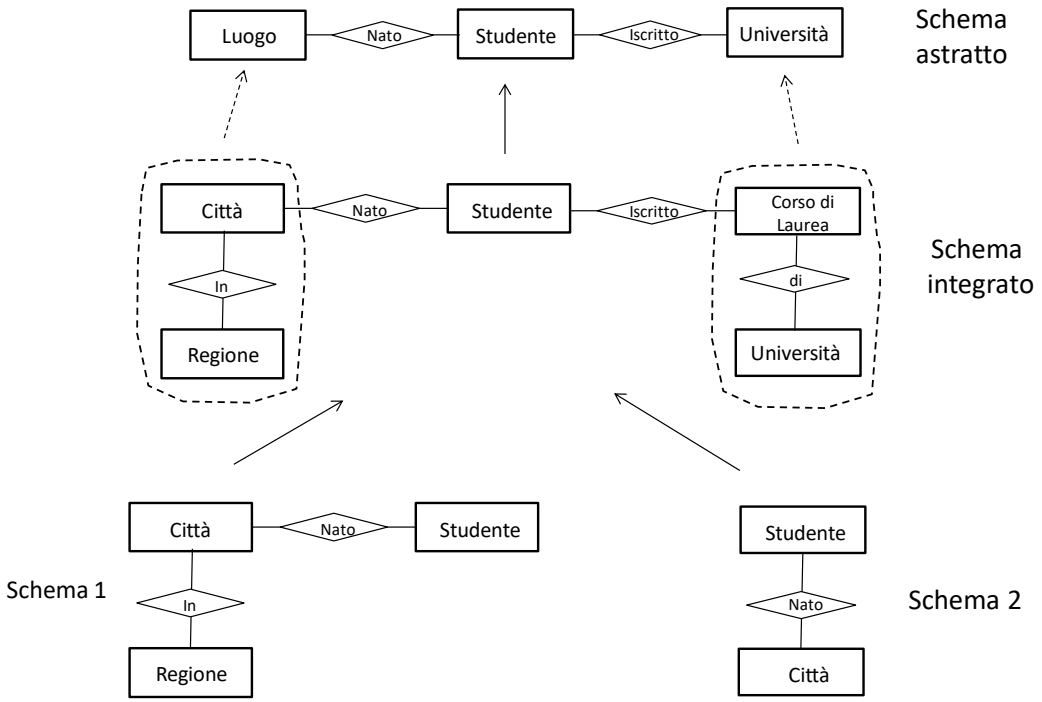


Figura 3 – Operazione di integrazione-astrazione

Partiamo dai due schemi chiamati Schema 1 e Schema; pemezzella perazioe di integrazione possiamo generare uno schema integrato (parte centrale della figura); dopodichè per produrre uno schema astratto dobbiamo scegliere quali sottoschemi sostituire con un unico concetto; i sottoschemi rappresentati in modo più compatto sono racchiusi da superfici con linea tratteggiata, e pure i concetti che li sostituiscono sono indicati da linee tratteggiate. Le due operazioni di integrazione e astrazione possono essere applicate nell'ordine di Figura 3, ovvero in ordine inverso: posso prima generare schemi più astratti e poi integrarli. Possiamo chiamare le due operazioni insieme come una unica operazione di integrazione astrazione.

Ho iniziato a interessarmi in modo sistematico di astrazioni nella preparazione di un tutorial che presentai all'Entity Relationship Conference tenutasi nel 2016 a Gifu, in Giappone, il lettore interessato può trovare le presentazioni in [Batini 2016]. Spesso ho tenuto tutorial su diverse tematiche per costringermi a investigare un tema scientifico in tutti i suoi aspetti, così da comprenderlo a fondo. Se dovevo spiegarlo ad altri, dovevo prima averlo chiaro io nella mia testa.

Nel preparare il tutorial, ho ragionato su una frase tratta da [Mylopoulos 1998], che diceva:

Per definizione, un'astrazione porta a una soppressione di dettagli considerati irrilevanti…

Effettivamente nelle astrazioni delle Figure 1 e 3 noi tendiamo ad eliminare i dettagli; la definizione pur così generica sembra individuare un paradigma generale. L'astrazione è in effetti una categoria utilizzata in tutte le discipline scientifiche e tecnologiche; possiamo iniziare a comprendere il concetto, introducendo le etimologie dell'astrazione che compaiono in [Dizionario Etimologia Online 2018]:

1. c. 1400, "un ritiro dagli affari mondani, dall'ascetismo", dall'antica astrazione francese
2. (14c.) Dal latino abstractionem (nominativo abstractio), nome dell'azione del participio passato di abstrahere "trascina via, allontana, devia"
3. Il significato di "idea di qualcosa che non ha esistenza" è del 1640.

Tra i tre riferimenti, il primo e il terzo hanno un sapore di assenza, ritiro, distacco. Solo il secondo, "trascinare via, allontanare, deviare" è vicino alla osservazione di Mylopoulos, e porta una potenziale utilità per il nostro discorso; peraltro, quando "trasciniamo via", manteniamo, simmetricamente, la parte più importante.

Nelle due ore del Tutorial ho cercato di sintetizzare i risultati della mia indagine, estesa a molti campi scientifici, che mi ha portato alla scoperta di oltre 400 astrazioni riportatein appendice. Prima d ientrare neerito delle caratteristiche principali delle astrazioni, conviene continuaa ragionare su altri esempi.

Come esempi nei domini della Fisica e della Matematica, vediamo in Figura 4 le equazioni di Maxwell e una formula matematica nella Figura 5.

Le equazioni di Maxwell (si veda [Wikipedia]) sono un insieme di equazioni alle derivate parziali che costituiscono il fondamento dell'elettromagnetismo classico, dell'ottica classica e dei circuiti elettrici.

La conoscenza astratta delle equazioni di Maxwell può trovare applicazione in uno spettro estremamente ampio di fenomeni fisici; le equazioni sopprimono i dettagli degli specifici fenomeni fisici, fornendone quindi una astrazioni nel senso della definizione di Mylopoulos.

Legge di Maxwell	**Significato**	**Equazione differenziale**
Legge di Gauss	il flusso del campo elettrico attraverso una superficie chiusa è proporzionale alla carica interna alla superficie	$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$
Legge di Gauss per il magnetismo	Non ci sono cariche magnetiche analoghe alle cariche elettriche. Il flusso del campo magnetico attraverso una superficie chiusa e' pari a zero.	$\nabla \cdot \mathbf{B} = 0$
Equazione di Maxwell–Faraday	il lavoro per unità di carica necessario a spostare una carica intorno a una spira chiusa è pari al tasso di diminuzione del flusso magnetico attraverso la superficie racchiusa	$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
Legge di Ampere	il campo magnetico indotto intorno a un circuito chiuso è proprorzionale alla corrente elettrica più la corrente di spostamento (proporzionale al tasso di cambiamento del flusso) attraverso la superficie chiusa	$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right)$

Figura 4 - Equazioni di Maxwell (da Wikipedia)

La formula matematica di Figura 5 può essere applicata a qualsiasi numero intero di due e mette in relazione il quadrato della somma dei due numeri con la somma dei quadrati dei due numeri più il prodotto dei due numeri moltiplicato per due. La formula è quindi anche essa una astrazione nel senso della definizione di Mylopoulos, perchè elimina il dettaglio relativo alla specifica coppia di numeri ed esprime una legge generale. Si noti che la stessa nozione di numero intero è un'astrazione di un ampio insieme di fenomeni fisici o virtuali (es. tre pere, tre dinosauri, ecc.)

$$(a+b)^2 = a^2 + b^2 + 2ab$$

Figura 5 – Un'astrazione espressa in una formula matematica

Consideriamo la definizione di "astrazione in matematica" fornita da Wikipedia: "L'astrazione in matematica è il processo per estrarre l'essenza sottostante di un concetto matematico, rimuovendo ogni dipendenza da oggetti del mondo reale con i quali potrebbe originariamente essere stato collegato, e generalizzandolo in modo che abbia applicazioni più ampie o corrispondenze tra altre descrizioni astratte di fenomeni equivalenti." Questa definizione è anche essa coerente con la definizione di Mylopoulos; inoltre, sottolinea che gli obiettivi della soppressione dei dettagli sono duplici: a. rendere il concetto indipendente dagli oggetti del mondo reale collegati, e b. raggiungere una più ampia applicazione in fenomeni tra di loro equivalenti.

Passiamo ora a un dominio a noi noto perché lo abbiamo affrontato nel Capitolo 3, il modello relazionale dei dati. Nel suo libro sul modello relazionale scritto venti anni dopo il lavoro originario [Codd 1990], Ted

Codd mentre motiva il suo sforzo negli anni per costruire un modello di dati completo che mirasse ad essere una rottura rispetto ai precedenti modelli a rete e gerarchici, afferma che nel concepire il modello relazionale ha cercato di seguire il consiglio di Einstein: "Rendi il più semplice possibile, ma non più semplice." Inoltre, nel lavoro originario in [Codd 1970], Codd afferma che il suo scopo era quello di definire un modello in grado di rimuovere tutti i dettagli nei modelli precedenti che fanno riferimento ad aspetti di implementazione fisica del modello in un calcolatore, dettagli che non interessano all'utente che deve comprendere il significato dei dati nella base di dati, ovvero che debba concepire una interrogazione uo una transazione.

Come abbiamo visto nel Capitolo 3, il modello relazionale è basato sulla struttura di relazione. Vediamo nuovamente in Figura 6 un esempio di un insieme di oggetti osservabili descritti in linguaggio naturale e la corrispondente rappresentazione nel modello relazionale.

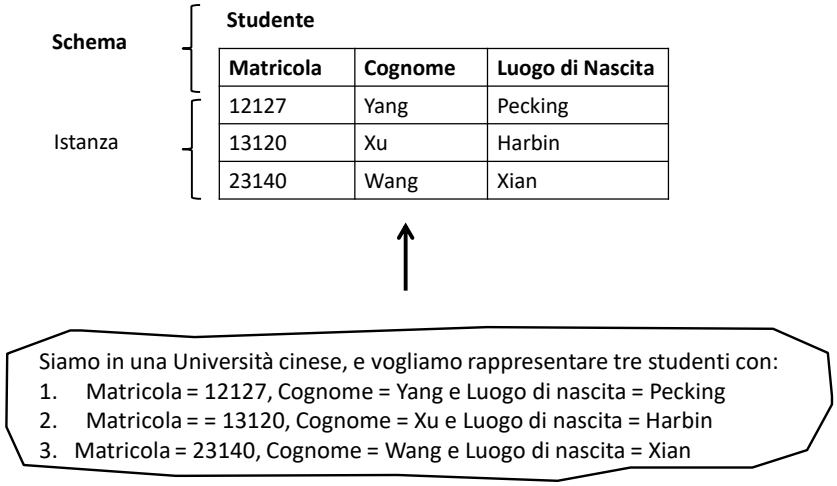


Figura 6 – Astrazione nel modello relazionale

Ogni osservabile (nel nostro caso ogni studente) è rappresentato da un record nella relazione ed esiste una netta separazione nella relazione tra il cosiddetto schema, composto dal nome della relazione e dai nomi degli attributi, e l'istanza, composta da record composti a sua volta di valori atomici che appartengono agli attributi Matricola, Cognome e Luogo di nascita; gli attributi sono astrazioni delle classi di valori che compaiono nella rispettiva colonna, e la relazione Studente può considerarsi astrazione dei valori che possono assumere i record.

Come ultima astrazione di questa introduzione, riconsideriamo le metodologie di progettazione di basi di dati che sono state proposte alcuni anni dopo l'introduzione del modello relazionale; apparentemente, il passo chiamato *logico* del progetto di basi di dati, in cui i requisiti che descrivono gli osservabili del mondo reale sono rappresentati nel modello relazionale, dovrebbe essere sufficiente, dal momento che il modello relazionale è più astratto dei precedenti modelli fisici.

I lavori di John Miles Smith e Diane Smith sulle astrazioni di generalizzazione e aggregazione [Smith 1977a], [Smith 1977b] e di Peter Chen sul modello Entità Relazione (ER) [Chen 1976] chiariscono che il modello relazionale è adeguato come modello di basi di dati, ma è troppo povero nell'esprimere le

astrazioni utili per poter procedere nel processo di progettazione di basi di dati. Abbiamo visto questi aspetti nella sezione 3 del Capitolo 3 sui modelli di dati, li riprendiamo brevemente; per convincerci della maggiore attitudine del modello ER a rappresentare astrazioni, vediamo in Figura 7 come viene rappresentata la generalizzazione Is-a nel modello relazionale e nel modello ER[50].

Studente

Matricola	Cognome	Italiano?	Stato
12127	Rossi	si	Italy
13120	Wang	no	Cina
23140	Macron	no	Francia

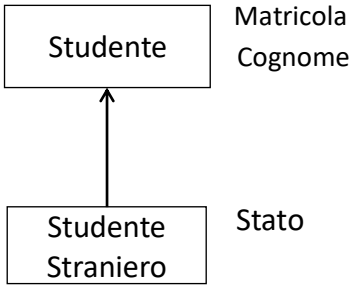


Figura 7 – L'astrazione Is-a nel modello Relazionale e nel modello Entità Relazione

L'astrazione Is-a, che esprime la proprietà per cui la classe Studente straniero è un sottoinsieme della classe Studente, deve essere rappresentata mediante un attributo binario nel modello relazionale, cioè, in certo senso, con un artificio linguistico, mentre è rappresentata in modo nativo nel modello Entità Relazione. La proprietà di *ereditarietà,* introdotta nel Capitolo 3 è un potente meccanismo di progettazione, soprattutto se adottiamo una strategia top-down nella progettazione dello schema, in cui dapprima modelliamo gli oggetti più generali, nel nostro caso Studente, e poi gli oggetti più specifici come Studente straniero.

Analogamente, nella progettazione di basi di dati viene adottato un processo di progettazione mostrato nella Sezione 3 del Capitolo 3, in cui le due fasi, progettazione concettuale e progettazione logica sono chiaramente separate insieme ai modelli adottati nelle due fasi.

Il Capitolo è organizzato come segue. La Sezione 2 discute due diverse accezioni secondo cui possiamo vedere le astrazioni: astrazioni come strumento di *rappresentazione* della realtà, e astrazioni come *processo*. La Sezione 3 si concentra sul tema della qualità delle astrazioni: scopriremo che ci sono astrazioni, sia come rappresentazione che come processo, di maggiore o minore qualità. Nella Sezione 4 vediamo che il concetto di astrazione è stato usato, oltre che nella Informatica e nelle basi di dati, in molte discipline e scienze, diventando perciò, a mio parere, uno dei concetti di più ampia applicazione nella scienza moderna. La Sezione 5 si focalizza sui big data, e mostra come lo strumento della astrazione sia uno strumento concettuale indispensabile per rappresentare grandi quantità di dati; sono presi in considerazione in particolare le mappe e i grafi.

2. Astrazioni come rappresentazione e astrazioni come processi

[50] In questo capitolo per semplicità rappresentiamo gli attributi invece che con un pallino riportando il nome accanto alla entità o relazione.

Le astrazioni vengono viste nella letteratura scientifica secondo due diversi punti di vista, che corrispondono a:
• astrazione vista come rappresentazione, e
• l'astrazione vista come un processo

Prendiamoli ora in considerazione distintamente.

3.1 Astrazione come rappresentazione

Un esempio di astrazione come rappresentazione appare nella Figura 8. Cosa ci dice lo schema? Quale realtà ci descrive? Osservando questo schema, e anche a causa della mia conoscenza contestuale sul mondo universitario, posso affermare che lo schema rappresenta:
• studenti, che hanno una matricola,
• gli studenti sono di due tipi, vale a dire
- studenti italiani, di cui ci interessa la data di nascita
- studenti stranieri, di cui ci interessa il paese di provenienza.

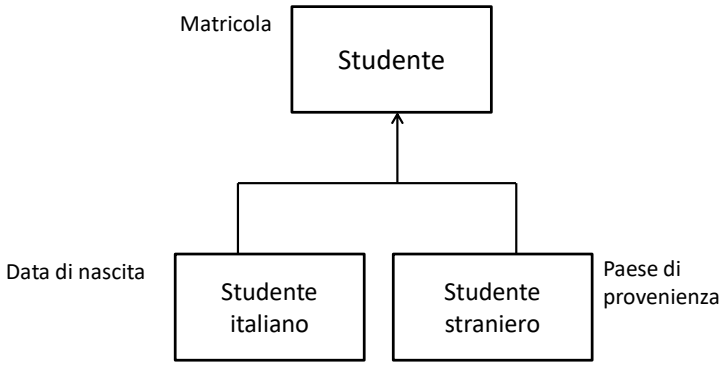


Figura 8 - Astrazione come rappresentazione

Tutte le astrazioni presentate fino ad ora nella introduzione sono astrazioni di rappresentazione. Noi utilizziamo anche inconsapevolmente le astrazioni di rappresentazione per comprendere in maniera chiara un aspetto della realtà e per comunicare con gli altri. Per “dominare” la complessità della realtà noi creiamo astrazioni in tutti i campi del sapere.

Consideriamo ora una astrazione in una vecchia mappa inglese in cui sono mostrati i simboli usati per diversi tipi di vegetazione, vedi Figura 9. Se non vogliamo fare nessuna distinzione, possiamo usare il simbolo in alto a destra in cui compaiono delle linee orizzontali generiche; dentro il rettangolo è anche scritto che volendo si può lasciare l’area bianca, ad indicare l’assenza di vegetazione significativa. I simboli successivi sono ottenuti dal simbolo generale fornendo una informazione addizionale rilevante sui tipi di vegetazione. Fornire informazione addizionale è la concettualizzazione complementare alla eliminazione dei dettagli di cui alla prima definizione che abbiamo introdotto.

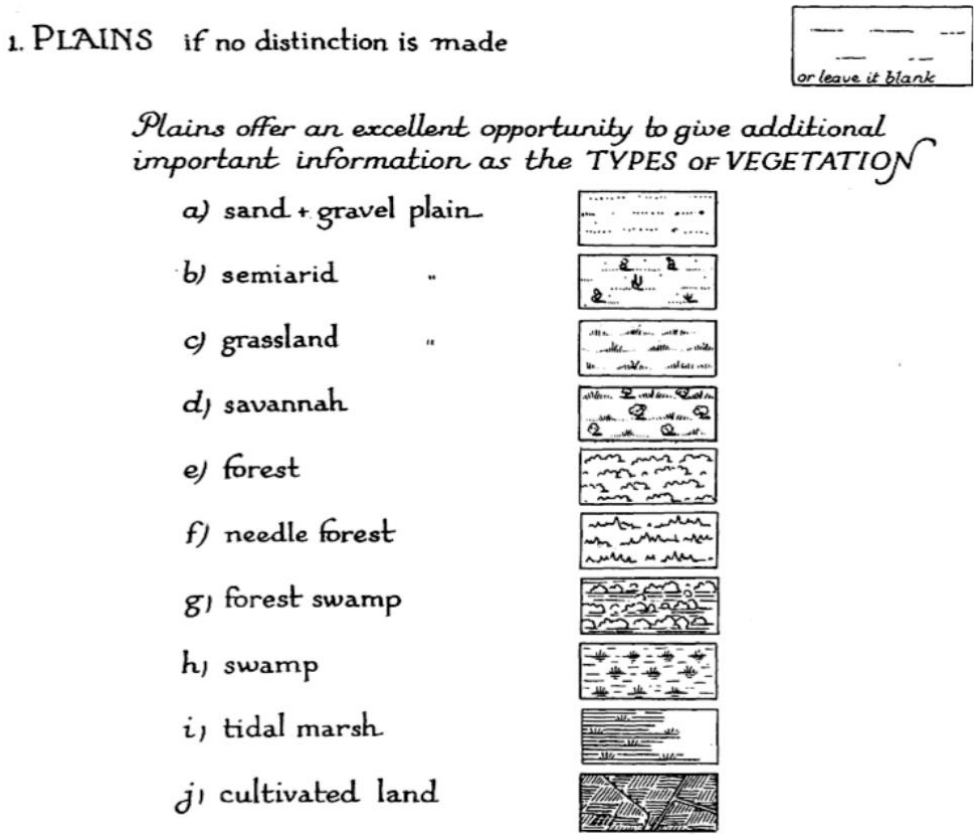


Figura 9 – Tipi di vegetazione in una antica mappa inglese (tratta da www.pinterest.it)

Le tre successive astrazioni, classificazione, generalizzazione e aggregazione, sono quelle utilizzate nel modello Entità Relazione, per cui nella loro descrizione faremo riferimento a figure già presenti nel Capitolo 3.

Classificazione

La Figura 14 del Capitolo 3 è un esempio di astrazione di classificazione: l'entità Studente è una classe le cui istanze sono gli studenti; come astrazione, rimuove le differenze tra le istanze e mantiene le proprietà comuni alle istanze, che possono essere la Matricola, il Cognome, o una relationship Nato-a con una entità Città.

Generalizzazione

Nella Figura 17 del Capitolo 3 compare una astrazione di generalizzazione Is-a tra le entità Studente straniero e Studente. La generalizzazione è un potente strumento di rappresentazione perché attraverso la proprietà di ereditarietà ci permette di rappresentare in modo compatto e quindi più facilmente comprensibile realtà complesse, mettendo in evidenza le proprietà (es. gli attributi) in corrispondenza al giusto livello di astrazione nelle generalizzazioni, e sfruttando la ereditarietà per far migrare le proprietà ai livelli più bassi della gerarchia di generalizzazione.

Aggregazione

Focalizziamo l'attenzione nella Figura 16 del Capitolo 3 sulla relationship Esame. Esame può essere visto come una astrazione rispetto a Studente e Corso diversa dalla generalizzazione, nel senso che Studente e Corso possono essere viste come parti di Esame. Allo stesso modo se abbiamo tre attributi Giorno, Mese e Anno essi possono essere visti come parte di un concetto più astratto che chiamiamo Data. Esame e Data sono esempi di astrazione di aggregazione, nel senso che aggregano concetti più elementari, Studente e Corso nel caso di Esame e Giorno, Mese e Anno nel caso di Data.

Introduciamo infine due ulteriori astrazioni, l'astrazione di raggruppamento e l'astrazione "che dimentica", abstraction by forgetting.

Raggruppamento

Pensiamo al legame concettuale tra un insieme di giocatrici di pallavolo, rappresentato come Entità Giocatrice, e le loro squadre, insieme anche esso rappresentato mediante una entità Squadra in Figura 10.

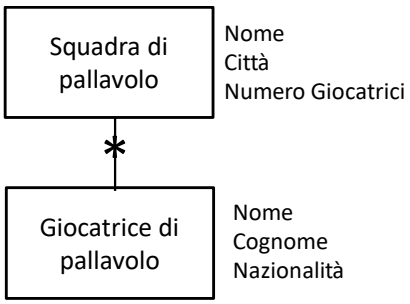


Figura 10 – L'astrazione di aggregazione

Forse vi state chiedendo: ma perché sta introducendo una nuova astrazione? tra Giocatrice e Squadra è definita una astrazione di generalizzazione. Se vi state facendo questa domanda, devo contraddirvi; sarebbe definita una generalizzazione Is-a se ogni istanza di Giocatrice fosse una istanza di Squadra, ma non è così! Una istanza di Squadra corrisponde a un insieme di giocatrici, che è una istanza di squadra. A riprova di ciò, a ogni gruppo di giocatrici corrispondente a una squadra corrisponde un valore dell'attributo Numero giocatrici di Squadra.

Perciò ci troviamo di fronte a una nuova astrazione che chiamiamo di raggruppamento, che rappresentiamo con un simbolo di asterisco nella Figura 10.

Astrazione che dimentica (Abstraction by forgetting) – Questa astrazione, introdotta da Palmonari nel lavoro [Palmonari & Batini 2009] è definita su grafi; fissato un nodo di un grafo, rappresentato in Figura 11 al centro del grafo di sinistra con etichetta Israel, possiamo trasformare il grafo cancellando, e quindi scordandoci, tutti i cammini di lunghezza maggiore di tre che partono dal nodo Israel.

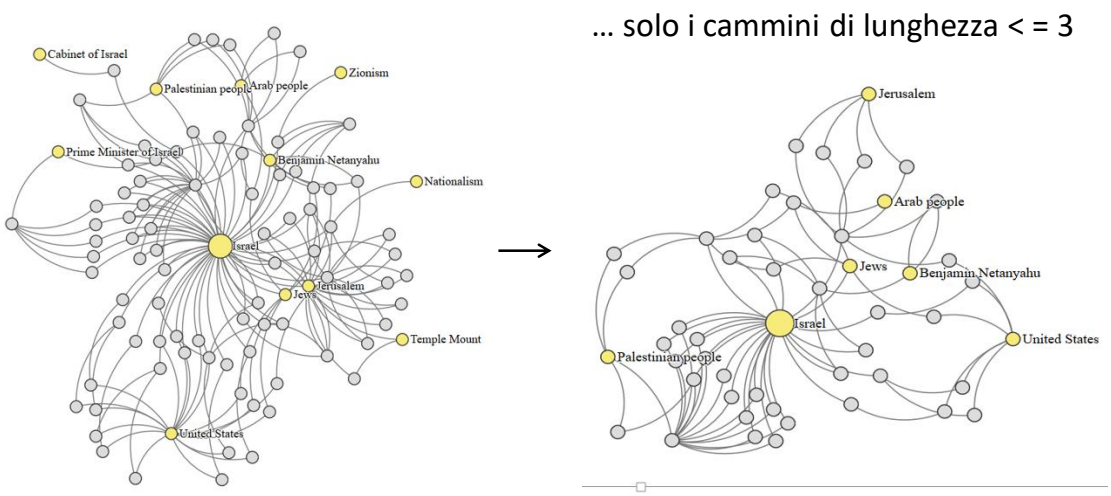


Figura 11 – Abstraction by forgetting

Le astrazioni come rappresentazione sono state intensamente studiate in molti campi del sapere; per una trattazione più approfondita nella modellazione concettuale e nella modellazione ontologica, il lettore interessato può approfondire consultando [Batini 2016].

3.2 Astrazione come processo: le trasformazioni

Oltre che nelle rappresentazioni, le astrazioni possono essere utilizzate nei processi di trasformazione. In molte discipline ci poniamo un obiettivo di progettazione, consistente nella produzione di un artefatto conforme a determinati requisiti. Questo è tipico, per menzionare solo alcune aree, nella progettazione di basi di dati, la ingegneria del software, gli artefatti nella ingegneria civile come edifici, ponti, infrastrutture o nella produzione di artefatti letterari come libri di narrativa o saggistica. Quando produciamo artefatti, di solito concepiamo per prima cosa una bozza dell'artefatto, che viene iterativamente raffinata, modificata, arricchita, adattata, attraverso un processo in cui vengono eseguiti diversi tipi di trasformazioni.

Una delle strategie ben note che vengono proposte per disciplinare tale processo, come abbiano iniziato a discutere nel Capitolo 3 sui modelli, è la strategia *top-down*, che adotta trasformazioni fatte di raffinamenti, in cui una parte semplice dell'artefatto viene trasformata in una complessa, che rappresenta lo stesso insieme di requisiti a più alto livello di dettaglio. I raffinamenti top-down possono essere visti come trasformazioni inverse o complementari delle astrazioni, che trasformano strutture complesse in strutture semplici.

La trasformazione della Figura 12 è un esempio di raffinamento, in questo caso lo schema semplice è formato da una sola entità, che il raffinamento trasforma in uno schema più grande, costituito da una generalizzazione tra Studente e le due entità Studente italiano e Studente straniero.

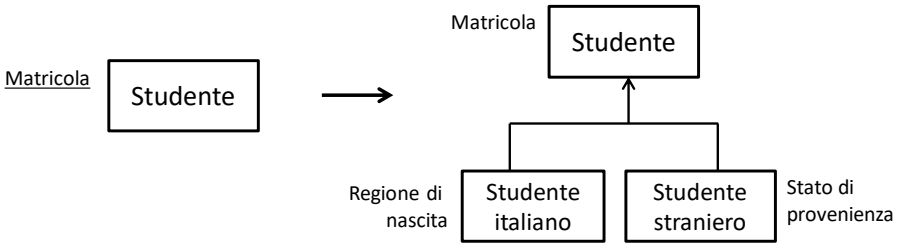
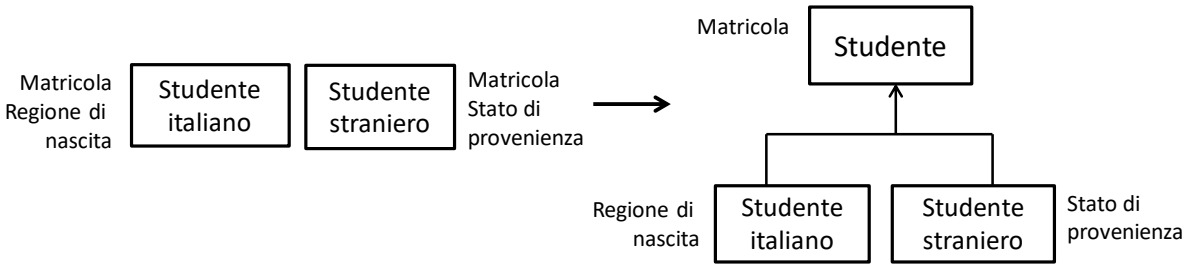
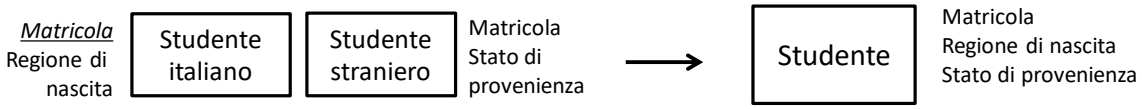


Figura 12 – Esempio di raffinamento

Nella Figura 13 vediamo due trasformazioni con caratteristiche diverse dalla precedente. In entrambe le trasformazioni lo schema nella parte sinistra è composto dalle due entità Studente italiano e Studente straniero; nella trasformazione (a) noi aggiungiamo l'entità genitore di una generazione, nella trasformazione (b) facciamo altrettanto, ma cancelliamo le entità di partenza. Queste trasformazioni ci appaiono piuttosto come trasformazioni di astrazione, perché creano una generalizzazione che nel primo caso viene conservata nello schema e nel secondo caso produce la sola entita' genitore.



Transformazione (a)



Transformazione (b)

Figura 13 – Astrazione come processo nel modello relazionale

Nella Figura 14 applichiamo la trasformazione di Figura 12 (riprodotta nella Figura 14 (b)) allo schema iniziale (a), ottenendo lo schema finale (c). Vediamo un po' più in dettaglio come si applica la trasformazione (b). Dobbiamo prima eliminare dallo schema (a) la entità Studente e poi "attaccare" allo schema (a) la parte destra della trasformazione (b); in questa operazione dobbiamo anche decidere a quale entità dello schema (b) associare i legami che la entità Studente aveva con il resto dello schema (a). Decidiamo di associare la relationship Vive-in che lega Studente a Stato alla entità Studente Straniero, e di mantenere associata a Studente la relationship con Città; con questo abbiamo completato l'applicazione della trasformazione.

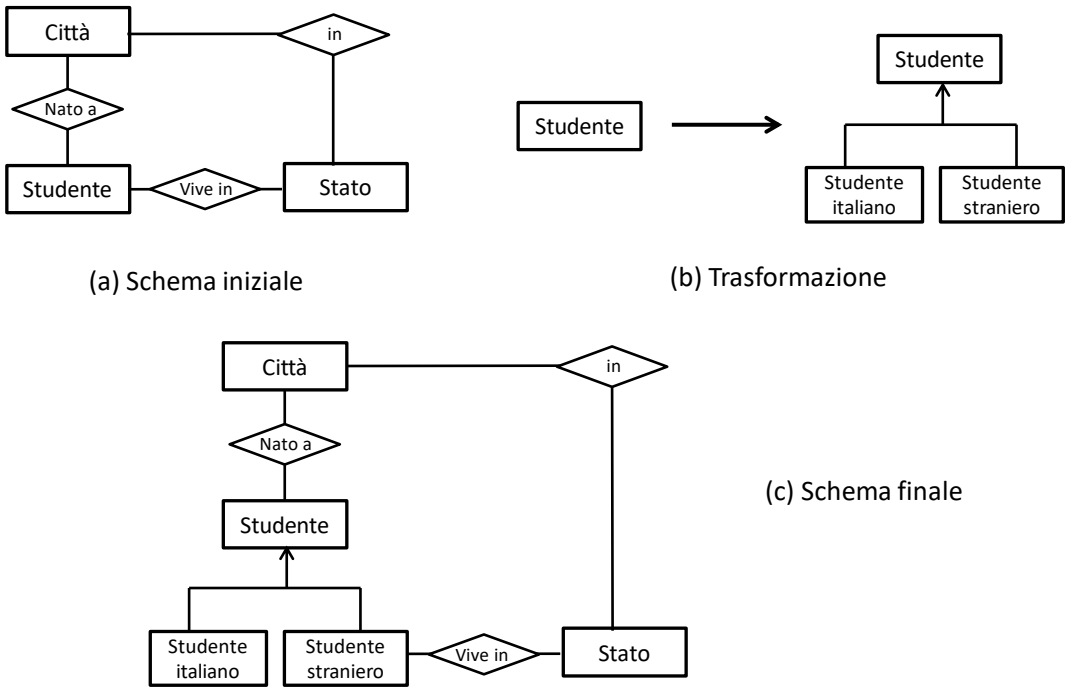


Figura 14 – Come si applica una trasformazione

Tornando ai 400 schemi concettuali della Pubblica Amministrazione Centrale, riconsideriamo la Figura 3 in cui sono mostrate trasformazioni di integrazione e astrazione. Pensate di applicare queste trasformazioni a 400 schemi suddivisi in gruppi omogenei; ciò che otteniamo, applicando via via le trasformazioni di integrazione astrazione, è una piramide di schemi via via più integrati e più astratti; chiamiamo questo insieme di schemi Repository di schemi concettuali.

Non abbiamo la possibilità in questo testo di rappresentare l'intero Repository degli schemi della Pubblica Amministrazione centrale; nella Figura 15 vediamo la parte del Repository che rappresenta le basi di dati del Sistema della fiscalità. La rappresentazione utilizza alcuni simbolismi metaforici al posto della usuale rappresentazione grafica del modello Entità Relazione, per impossibilità pratica di rappresentare la versione originale.

Le ellissi rappresentano singoli schemi concettuali. Quando ad una ellissi di un grigio tenue corrisponde più in basso una ellissi di colore grigio più forte, allora ci troviamo di fronte a un raffinamento, in senso inverso abbiamo una astrazione; quando ad un insieme di ellissi corrisponde immediatamente più in alto una sola ellissi, ci troviamo di fronte ad una *integrazione* di schemi (in senso inverso, una *segmentazione*). I "coni" cui sono associati gruppi di schemi a crescente intensità di grigio sono insiemi di raffinamenti, seguiti da una integrazione/segmentazione; agli schemi foglia, per ragioni di spazio, non abbiamo associato un nome né un acronimo. Con le quattro operazioni, complementari a due a due, di astrazione/raffinamento e segmentazione/integrazione abbiamo rappresentato una realtà complessa che a sua volta può essere vista come parte del più grande Repository degli schemi concettuali della Pubblica Amministrazione Centrale.

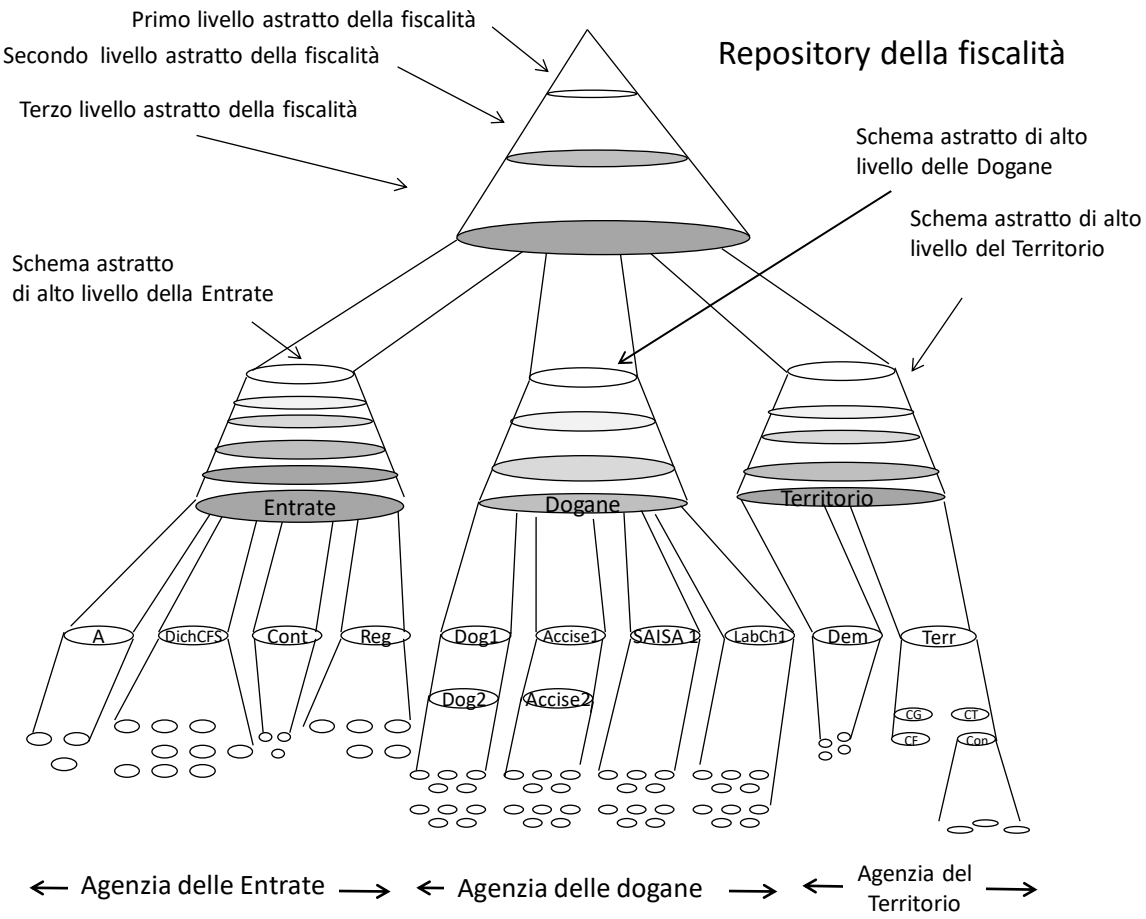


Figura 15 – Il Repository di schemi concettuali della Fiscalità

Il lettore interessato ad esaminare in modo più approfondito i temi delle trasformazioni nel modello Entità Relazione può dare un'occhiata a [Batini 1993], mentre per una discussione sulle possibili utilizzazioni del Repository si veda [Batini et al 2005].

Le trasformazioni sono strumenti concettuali che possono essere utili anche per molti altri modelli, e in molte altre discipline. Vediamo ora esempi di trasformazioni nelle Reti di Petri, formalismo adottato per descrivere processi, cioè sequenze di attività tra loro legate da vari tipi di relazioni, come ad esempio la relazione di precedenza.

Una rete di Petri cosiddetta Eventi/Condizioni consiste di rappresentare:
1. elementi attivi, che corrispondono ad azioni, chiamati *eventi*, vedi Figura 16, dove gli eventi sono rappresentati con riquadri rettangolari;
2. elementi passivi, che non corrispondono ad azioni, e che determinano l'evoluzione del processo, ad esempio: se sta piovendo (condizione) esci con l'ombrello.
3. relazioni causali tra eventi e condizioni, che nella Figura 16 sono rappresentate come frecce tra condizioni ed eventi.

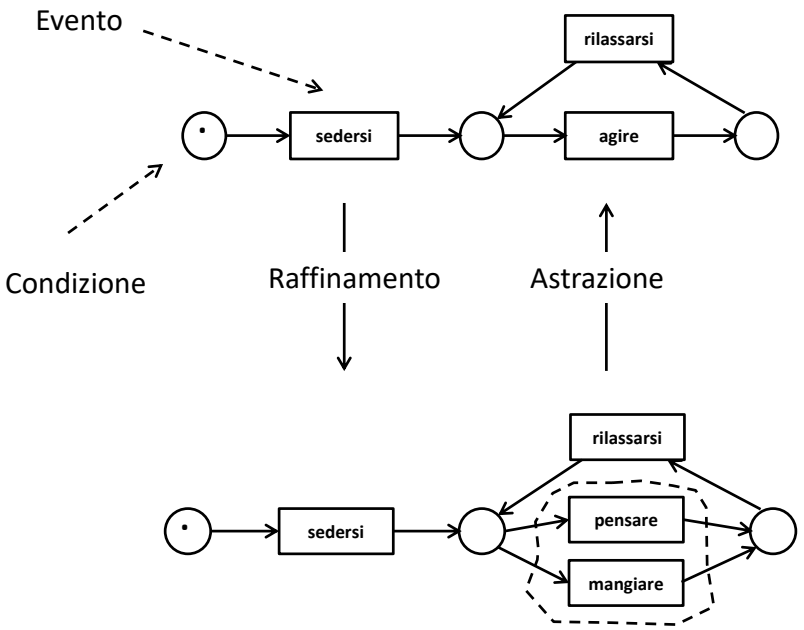


Figura 16 – Esempi di astrazione come processo nelle Reti di Petri

La Figura 16, tratta da [Brauer 1989], rappresenta nella sua parte superiore (con un po' di ironia) la vita di un filosofo, per come gli scienziati informatici la vedono. Il filosofo si siede a un tavolo e ripetutamente fa qualcosa e poi si rilassa. La Figura 16 nella sua parte inferiore si può vedere come un raffinamento della precedente rete di Petri, in cui l'azione "agire" viene dettagliata in termini delle azioni consistenti in "mangiare" o "pensare". Anche in questo caso l' applicazione della trasformazione sostituisce l'evento "agire" con i due eventi precedenti, collegati in parallelo con le condizioni che precedono e seguono gli eventi.

Nella Figura 17 vediamo un altro esempio di trasformazione espressa sulle reti di Petri Eventi/Condizioni. La trasformazione di raffinamento (b) specifica che la trasformazione applicata, ad esempio, all'evento t_r della rete (a) porta alla sostituzione con la rete specificata nella parte destra di (b). Anche in questo caso, nella applicazione della trasformazione occorre specificare quali eventi in (b) ereditano i legami che l'evento t_r ha nella rete (a).

Per altri tipi di reti di Petri dette Posti/Transizioni e per una trattazione completa delle trasformazioni di raffinamento nelle Reti di Petri si veda [Suzuki and Murata 1983], e il già citato [Brauer 1989].

Le trasformazioni precedenti erano definite all'interno di un modello, cioè si applicavano a uno schema dati o a una rete di Petri e fornivano come risultato un nuovo schema o rete, le chiameremo perciò intramodello.

Nella Figura 18 vediamo un esempio di trasformazione inter-modello, in cui l'artefatto che viene prodotto è descritto in termini di un modello diverso rispetto all'artefatto oggetto della trasformazione. Abbiamo visto esempi di questi tipi di trasformazioni nel Capitolo 8, quando abbiamo trasformato u tabella relazionale in una rappresentazione Linked Open Data.

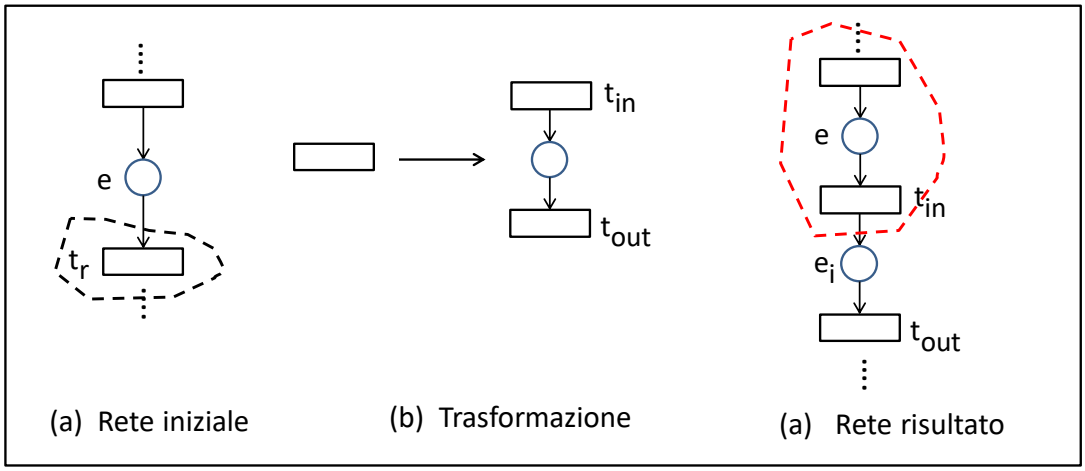


Figura 17 – Una trasformazione di raffinamento nelle Reti di Petri

Nella Figura 18 trasformiamo uno schema relazionale in uno Schema Entità Relazione; la trasformazione non riguarda la istanza. Questa trasformazione inter modello è una trasformazione di astrazione, perché il modello di arrivo, il modello Entità Relazione, è a più alto livello di astrazione del modello di partenza; la trasformazione inversa, dal modello Entità Relazione al modello relazionale corrisponde alla fase di progettazione logica di una base di dati.

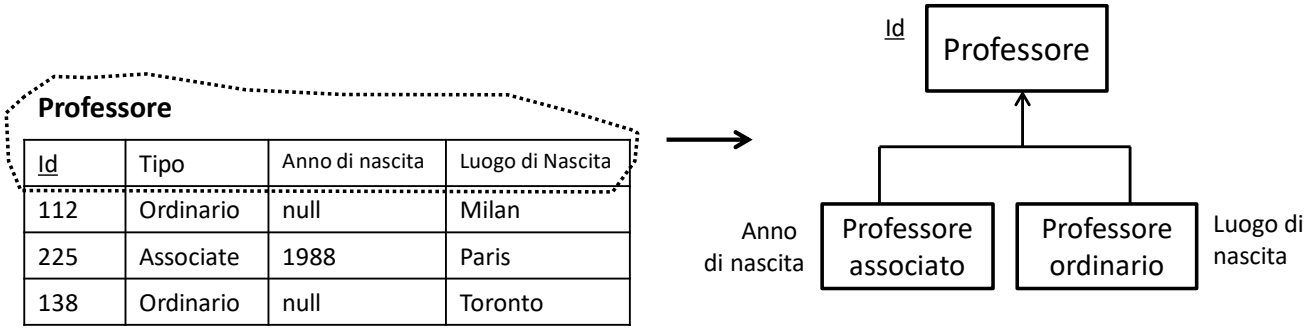


Figura 18 – Trasformazione di astrazione intermodello
tra modello relazionale e modello Entità Relazione

Per esempi di trasformazioni inter-modello nelle Reti di Petri, si vedano i lavori citati in precedenza.

3. Astrazioni e qualità

Le astrazioni come rappresentazione viste nella sezione precedente descrivono artefatti, evidenziando le proprietà rilevanti di tali artefatti e nascondendo dettagli non necessari. Le astrazioni (e i raffinamenti) viste come processo, operano su artefatti trasformandoli in versioni più sintetiche (astrazioni) o più dettagliate (raffinamenti). Sia nella loro natura di rappresentazione che nella loro natura di processo, possiamo associare alle astrazioni diversi livelli di qualità, concetto che con riferimento ai dati digitali abbiamo investigato nel Capitolo 5.

Un esempio di qualità di un'astrazione come rappresentazione è mostrato nella Figura 19. In cui vediamo due diversi schemi Entità Relazione che rappresentano lo stesso insieme di requisiti nella nuvoletta a sinistra. Lo schema in Figura 19.a è incomprensibile, sono utilizzati nomi che non ci dicono nulla sul significato dei concetti rappresentati e i simboli grafici sono disposti in modo causale sul piano. Un disastro dal punto di vista della comprensibilità, la caratteristica di qualità di uno schema che ne favorisce la comprensione anche in assenza di ulteriori informazioni disponibili.

Al contrario, lo schema di Figura 19.b è caratterizzato da alta comprensibilità, dovuta al fatto che abbiamo utilizzato:
1. nomi significativi per entità e attributi
2. L'entità genitore nella generalizzazione (Studente) è posta in alto e le entità figlie in basso.
3. La regola di ereditarietà delle generalizzazioni viene applicata rappresentando ciascun attributo una sola volta nello schema e ottenendo in questo modo una rappresentazione più compatta e comprensibile.

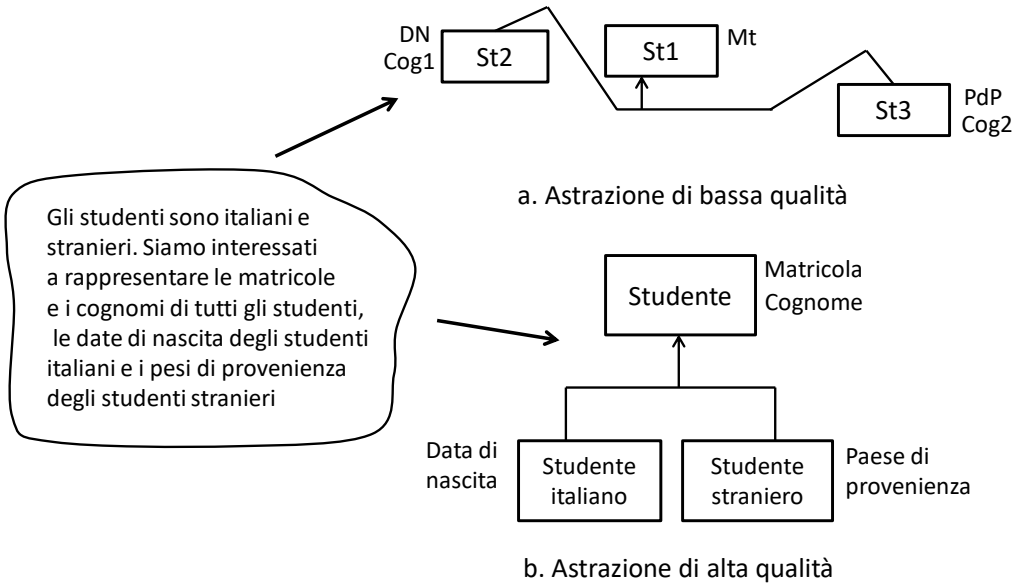


Figura 19 – Astrazioni di bassa e alta qualità nella rappresentazione diagrammatica del modello Entità Relazione

Analizziamo ora un esempio di qualità delle astrazioni viste come processi, si veda Figura 20. In questo caso vengono rappresentati tre diversi processi di astrazione dello stesso schema iniziale. Per ogni passo di astrazione, mostriamo nello schema inferiore la parte dello schema coinvolta dalla trasformazione. L'esito finale delle trasformazioni è lo stesso, ma notiamo che:

- nel processo (di astrazione) 1 si compattano prima due entità in una e successivamente quattro entità in una.
- nel processo 2 si fa l'inverso, prima quattro entità e poi due entità
- nel processo 3 si compattano per due volte tre entità.

Il più equilibrato processo dal punto di vista delle "intensità" con cui lo schema è astratto è il processo 3, in cui in entrambi passi vengono compresse tre entità.

Possiamo chiamare questa caratteristica di qualità "bilanciamento delle astrazioni". Questa caratteristica ricorda un pò certi discorsi o certi documenti in cui si passa talvolta da temi molto astratti a particolari insignificanti, o, viceversa ci si concentra a lungo in particolari dettagliatissimi, per poi improvvisamente salire di livello di astrazione parlando dei massimi sistemi.

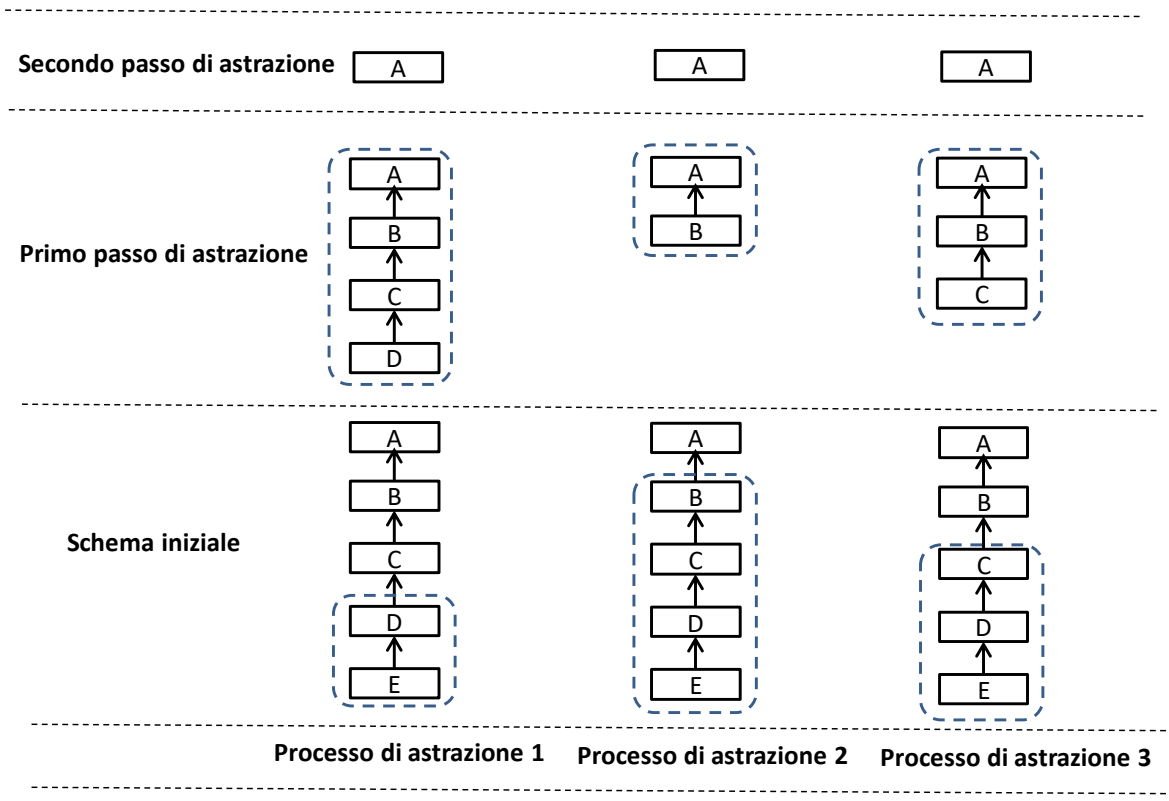


Figura 20 - Processi di astrazione con differenti qualità

Si noti che nel precedente esempio abbiamo assunto implicitamente che tutte le entità nelle relazioni Is-a abbiano la stessa importanza; in caso di diversi gradi di rilevanza, possiamo evidenziare tale rilevanza *rallentando la velocità* delle astrazioni. Chi fosse interessato ad approfondire il tema della qualità delle astrazioni viste come processi, può consultare [Batini 2012].

4. Dalle astrazioni nelle basi di dati alle astrazioni in altre discipline

4.1 Le discipline investigate nel Tutorial del 2016

In assenza di una scienza completa e matura delle astrazioni, possiamo costruire i primi mattoni di tale disciplina investigando la letteratura scientifica che utilizza in modo esplicito o implicito il paradigma dell'astrazione. Questo è l'approccio seguito durante la preparazione del Tutorial 2016, in cui i domini scientifici e i sottodomini che sono stati studiati sono mostrati nella Figura 21, insieme ai riferimenti

principali per i lettori che vogliono approfondire. L'insieme completo delle presentazioni del tutorial e i riferimenti bibliografici sono disponibili su [Batini 2016].

Dominio	Sottodominio	Riferimenti
Scienze cognitive e lingustica		[Tenenbaum et al. 2011]
Informatica		[Colburn et al. 2007]
Modellazione concettuale		[Batini et al. 1989]
Modelli dei dati		
	Relazionale	[Codd 1979]
	Linked open data	[Ferrara 2005]
Economia		[Boisot 2005]
Matematica		[Ferrari 2013]
Filosofia		[Floridi 2008]
Reti di Petri		[Suzuki 1983], [Brauer 1989]
Robotica		[Kuipers 2000], [Zender 2008]
Scienza dei Servizi		[Comerio 2015]
Ragionamento, Logica		[Saitta & Zucker, 1998]
Ontologie		[Sowa 2000], [Keet 2008]
Visualizzazione		
	Artefatti	[Ravish 2009]
	Diagrammi	[Batini 1984], [Batini et al. 1993]
	Fisheye	[Furnas 1992], [Gansner 2003]
	Grafi	[Quigley 2000]
	Immagini	[Lin 2011]
	Mappe	[Shashi 2007]

Figura 21 – Domini e sottodomini investigati in [Batini 2016]

Il titolo del tutorial:"Abstractions in conceptual modeling and surroundings" ci dice che il punto di partenza dell'indagine è l'astrazione nella modellazione concettuale di basi di dati; ambiti vicini alla modellazione concettuale sono i modelli di dati in generale e le ontologie. Allargando la portata dell'indagine, le aree più ampie "vicine" alla modellazione concettuale sono l'informatica, le scienze cognitive, la matematica, mentre generalizzando i diagrammi concettuali incontriamo altre forme di visualizzazione, come grafici, viste fisheye, artefatti e mappe e immagini in generale. Il tutorial ha infine preso in considerazione anche altre discipline, come la economia nelle sue relazioni con la rappresentazione della conoscenza, in cui è stato investigato il contributo di Boisot alle astrazioni in economia. Oltre a ciò, è stata indagata la modellazione dei processi, i sistemi formali basati sulla logica, e la robotica la scienza dei servizi,.

Infine, poiché uno dei due autori del Tutorial era laureato in Filosofia, decidemmo di indagare anche in questa disciplina, con il risultato di trovare una cinquantina di diverse definizioni di astrazione proposte nella storia della Filosofia. Abbiamo deciso di non trattare le astrazioni in altre aree della conoscenza, come l'Arte e la Musica, non perché le astrazioni siano assenti in tali discipline, ma a causa della assoluta incapacità di applicare un metodo rigoroso in ambiti così vasti, in conseguenza della nostra ignoranza sui concetti di base in queste aree.

Per sostanziare alcuni risultati nelle precedenti discipline, vediamo nel seguito dapprima due importanti applicazioni delle astrazioni nell'ambito della prova di teoremi e nel layout automatico dei diagrammi, e successivamente un confronto tra astrazioni in matematica e astrazioni in informatica, concludendo con una analisi comparativa delle astrazioni utilizzate in diverse discipline.

4.2. Astrazioni nelle prove di teoremi

Provare un teorema corrisponde a trovare un insieme di inferenze che nel loro complesso permettano di dimostrare che dalle ipotesi del teorema è possibile generare la tesi espressa dal teorema. Ora, se il dominio del problema su cui "predicano" le ipotesi e la tesi del teorema è molto complesso, il modello di rappresentazione del problema su cui occorre trovare le inferenze può essere troppo complesso per farci essere in grado di concepire la prova.

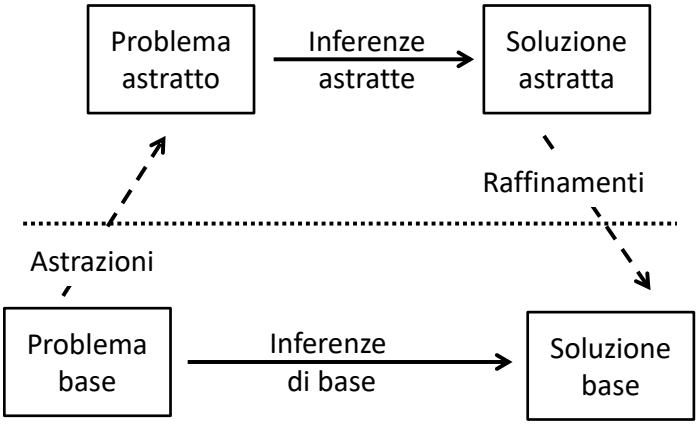


Figura 22 - Semplificazione della prova dei teoremi attraverso le astrazioni

Ciò che possiamo fare, consiste nel formulare una versione astratta del problema (vedi Figura 22) che, sotto opportune ipotesi, possiamo considerare come punto di partenza della prova. Se riusciamo a concepire un insieme di inferenze che trasformano le ipotesi, ora astratte e quindi "trattabili", nella tesi, anche essa astratta rispetto alla tesi iniziale, ciò che rimane da fare è verificare se sia possibile raffinare la tesi, introducendo tutti i dettagli del dominio di partenza, nella tesi della soluzione base.

Il tema delle astrazioni nei sistemi formali è molto vasto, il suo approfondimento richiede uno studio impegnativo; il riferimento per approfondire il tema precedente è [Saitta 2015]; una teoria delle astrazioni è proposta in [Giunchiglia 1992], mentre [Keet 2006] approfondisce il concetto di granularità, studiato nei sistemi formali e nell'ambito dei fuzzy set.

4.3. Astrazioni nel layout automatico di diagrammi

Passiamo ora al tema del layout automatico di diagrammi, cioè del loro disegno sul piano effettuato per mezzo di algoritmi. Quando tanto tempo fa ho cominciato a ragionare sul problema del layout automatico di diagrammi, ho cercato anzitutto di capire quali criteri estetici tendiamo ad applicare quando disegniamo un diagramma. Abbiamo trattato questo problema nella discussione sulle qualità nel Capitolo 5, si veda in particolare la Figura 17. I principali criteri che mi sono venuti in mente riguardavano:

- la minimizzazione degli incroci tra linee: la presenza di incroci tende a complicare la comprensione del significato del diagramma.

- la minimizzazione dei piegamenti delle linee: una linea diritta crea meno problemi di comprensione della relazione logica che esprime, rispetto a una linea con tanti piegamenti.
- la minimizzazione dell'area del diagramma; quanto più grande è l'area tanto più dobbiamo spostarci con i nostri occhi per comprendere, nel vero senso della parola, prendere insieme, catturare tutto in un colpo, il diagramma.

Mentre ragionavo senza costrutto sull'algoritmo che mi avrebbe permesso di disegnare diagrammi che rispettavano i precedenti criteri, venne da me Roberto Tamassia, studente di ingegneria, per sapere quali tesi gli potessi proporre. Io gli parlai del precedente problema. Non è stata l'unica volta che ho proposto a qualcuno un problema molto difficile; ricordo che una volta proposi un testo d'esame per il corso di Programmazione dei Calcolatori Elettronici in cui lo studente doveva affrontare un problema che era di difficoltà paragonabile al precedente, da risolvere in un'ora, scrivendo il programma espressione dell'algoritmo risolutivo. Per fortuna che poi mi accorgo della eccessività della mia azione; in quel caso, ovviamente, annullai il compito, e da allora mi sono sempre posto questo, forse ovvio, criterio: quando concepisci una domanda d'esame, prova prima a risolverla tu, e se ci riesci e se vedi che non è eccessivamente complicato come quesito, conferma quella domanda, altrimenti trova una domanda più semplice.

Roberto ci pensò un secondo, e mi disse: le chiedo una settimana per ragionarci con calma. Dopo una settimana, Roberto entrò nella mia stanza con un fogliettino, dicendo che aveva trovato un algoritmo, che ora descriverò nei suoi aspetti essenziali, senza entrare in tecnicismi.

Prima di tutto dobbiamo tradurre il diagramma Entità Relazione in un grafo, rappresentato come una relazione matematica, in termini di un insieme di variabili (nodi del grafo) e relazioni binarie tra variabili (rami del grafo, vedi Figura 23 in alto). E' bene far notare al lettore che nel diagramma è definita una relazione ternaria, cioè tra tre entità, una estensione del concetto di relazione che ho definito nel Capitolo 3. Questa relazione ternaria è trasformata, come anche le entità, in un *nodo* del grafo, mentre le relazioni binarie sono trasformate in *rami* del grafo.

A questo punto, dobbiamo vedere, modellare il grafo secondo una astrazione *topologica*, in cui *l'unico aspetto che conta riguarda le superfici* che il grafo creerebbe se lo rappresentassimo su un piano. In questa fase, dobbiamo cercare un algoritmo che permette di modificare le superfici in modo tale che la rappresentazione finale delle superfici corrisponda alla *assenza di incroci*. Questo è ottenuto con operazioni del tipo di Figura 24, in cui abbiamo identificato le superfici nei due grafi disegnati sul piano mediante numeri interi; nel grafo a destra c'è una superficie in meno. In questo modo abbiamo tenuto conto e realizzato l'obiettivo della *minimizzazione degli incroci tra linee*.

Passiamo ora, scendendo di livello di astrazione, a considerare la *forma*. Il grafo ottenuto dal passo precedente va ora visto considerando le linee non più come separatrici di superfici, ma nella loro disposizione in una griglia, quindi *composte di segmenti orizzontali e verticali*; quando si passa da un segmento orizzontale a uno verticale, c'è un *piegamento* di 90 gradi. Questa è una astrazione *morfologica*, perché prende in considerazione la forma, ma non ancora, ad esempio, la *distanza* tra nodi. Ora dobbiamo generare un algoritmo che tra le tante rappresentazioni trovi quella o quelle con il minor numero di piegamenti.

Il gioco è quasi fatto. Ora consideriamo anche la *distanza*, e la *misura dei segmenti* che formano le linee. Ho concepito, diceva Roberto a conclusione della sua proposta, un algoritmo che tra le tante rappresentazioni sul piano cartesiano, piano cioè in cui consideriamo anche una metrica per la distanza, trova quella o quelle contenute nell'area di *dimensione più piccola*. Quando dico che sono un uomo fortunato ad aver incontrato nella mia vita professionale persone come Roberto Tamassia, e non solo lui, non scherzo!

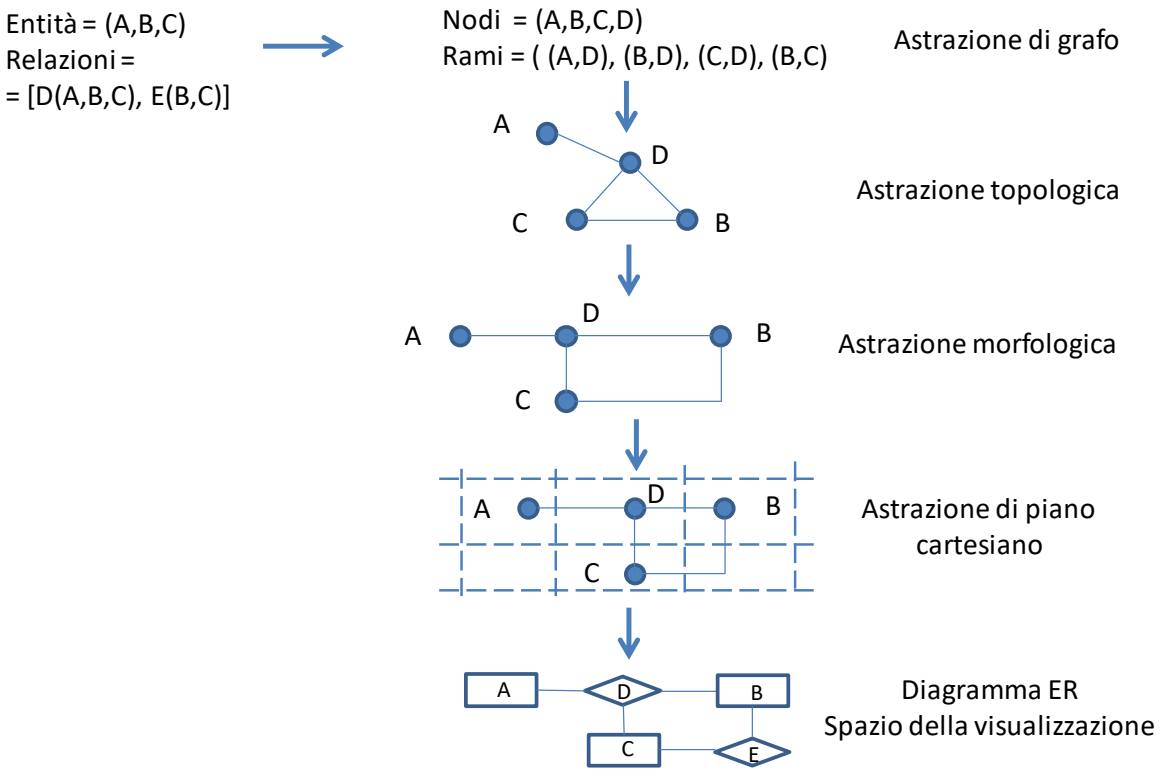


Figura 23 - Livelli di astrazioni utilizzati nel layout automatico dei diagrammi

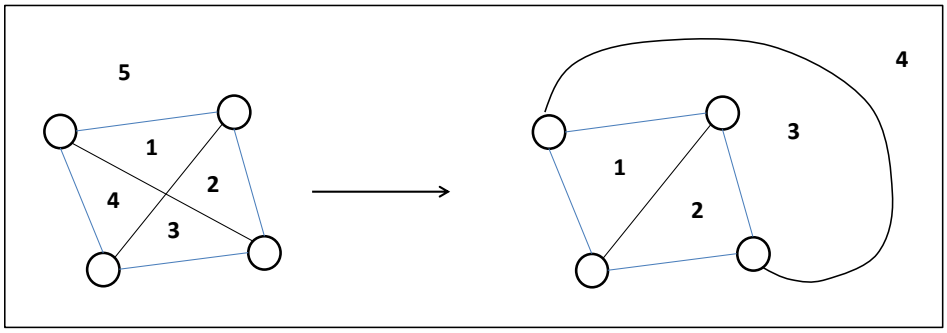


Figura 24 – Come modificare le superfici create dal grafo sul piano, con il fine di produrre un disegno senza incroci tra le linee

Per ottimizzare il diagramma rispetto ai tre criteri enunciati in precedenza, Roberto ha adottato una strategia "divide et impera", che trasforma il grafo astratto in termini di tre successive rappresentazioni, in cui vengono adottate astrazioni via via di livello più concreto, con più dettagli, come l'astrazione *topologica, morfologica e cartesiana*. Vedi anche [Tamassia 1988] per una descrizione più formale dell'algoritmo di disegno automatico.

4.4 Astrazioni nella Matematica e in Informatica

Ci concentriamo ora su un confronto tra le astrazioni in Matematica e Informatica. Tratteremo dapprima le astrazioni nella loro accezione *generale* come strumenti per rimuovere aspetti considerati non rilevanti, focalizzandoci successivamente sulla *astrazione di relazione* in Matematica e nelle Basi di dati.

Le scienze empiriche tradizionali si occupano sia di modelli concreti, sotto forma di apparati sperimentali, sia di modelli astratti trattati in Matematica. L'Informatica si distingue dalle scienze empiriche in quanto i suoi "prodotti" sono in prevalenza immateriali - essi sono costituiti da infrastrutture hardware, programmi software e dati, tutti artefatti la cui forma costitutiva li rende più vicini ai servizi immateriali che ai beni materiali, come vedremo per i dati nel prossimo Capitolo 13 sulla economia digitale. L'informatica ha bisogno di astrazioni per modellare i suoi artefatti, come abbiamo visto in maniera inequivocabile per i dati nel Capitolo 3, in cui abbiamo discusso vari modelli, e abbiamo individuato tra i modelli vari livelli di astrazione.

Poiché anche la Matematica si distingue dalle scienze empiriche in quanto i tipi di modelli che studia sono astrazioni, si può essere tentati di arrivare alla conclusione che esiste una stretta relazione tra Matematica e Informatica. Colburn in [Colburn 2007] confronta la natura delle astrazioni in Matematica e in Informatica, comparando la natura degli artefatti primari da esse trattate.

Per Colburn gli artefatti principali della Matematica sono le *strutture di inferenza*, mentre gli artefatti dell'Informatica sono i *modelli di interazione*, cioè i modelli che descrivono la relazione tra utente e macchina. Questa è una differenza cruciale, e determina la natura le astrazioni usate nelle due discipline:

• La Matematica, essendo interessata principalmente allo sviluppo di strutture di inferenza, usa le astrazioni con l'obiettivo di "dimenticare" (la già discussa abstraction by forgetting definita in [Palmonari & Batini 2009]); in una formula matematica non c'è più nessuna traccia del fenomeno che i numeri rappresentano, aspetto cui corrisponde l'obiettivo dell'astrazione.
• L'informatica, essendo interessata principalmente allo sviluppo di modelli di interazione, ha l'obiettivo di creare astrazioni che nascondano i dettagli fisici della infrastruttura di calcolo, ma *non li annullino*, perché deve sempre essere possibile reintrodurre tali dettagli (questa astrazione è chiamata in [Palmonari & Batini 2009] abstraction by collapse).

Facciamo due esempi di questi differenti processi di astrazione.

Il primo esempio riguarda il progetto di basi di dati che abbiamo descritto nella Figura 19 del Capitolo 3; la prima fase di progettazione concettuale permette al progettista di utilizzare un modello, il modello

Entità Relazione, vicino al suo modo di pensare e lontano dal modello che i dati presentano nella memoria fisica; accanto ad essa esiste una seconda fase in cui lo schema concettuale è tradotto in tabelle nel modello relazionale, quello in cui vengono formulate le interrogazioni e le transazioni di accesso.

Il secondo esempio riguarda i cosiddetti linguaggi programmativi di alto livello che permettono a un programmatore di scrivere un programma senza conoscere le caratteristiche fisiche della memoria e della unità di calcolo; anche in questo caso, il programma nel linguaggio di alto livello è tradotto nel linguaggio eseguibile dalla macchina da un programma software chiamato compilatore.

Insomma, la complessità dei moderni dispositivi di calcolo informatici rende impossibile programmarli senza strumenti di astrazione che *nascondano, ma non cancellino*, i dettagli realizzativi che sono essenziali in un contesto di elaborazione, ma non rilevanti in un contesto di progettazione e programmazione del software.

4.5 Astrazioni in politica ed economia

Concludiamo questo capitolo con una discussione sull'uso delle astrazioni in politica ed in economia. Abbiamo iniziato a parlare delle astrazioni in politica nella Sezione 7 del Capitolo 6, a proposito dell'accordo di governo tra Movimento 5 Stelle e Partito Democratico. La figura seguente mostra un esempio dei diversi livelli di astrazione che possono essere adottate nelle policy di Governo. Come sappiamo, gli uomini politici sono usualmente radicati in un particolare territorio, dove hanno sempre vissuto o dove è collocato il loro collegio elettorale. Hanno inoltre una formazione e una professionalità diversificata, e diverse visioni politiche, quindi sono più sensibili a determinate materie piuttosto che ad altre. In Figura 25 mostriamo un insieme di politiche che possono essere adottate, per esempio, nella pianificazione degli investimenti a livello locale o statale.

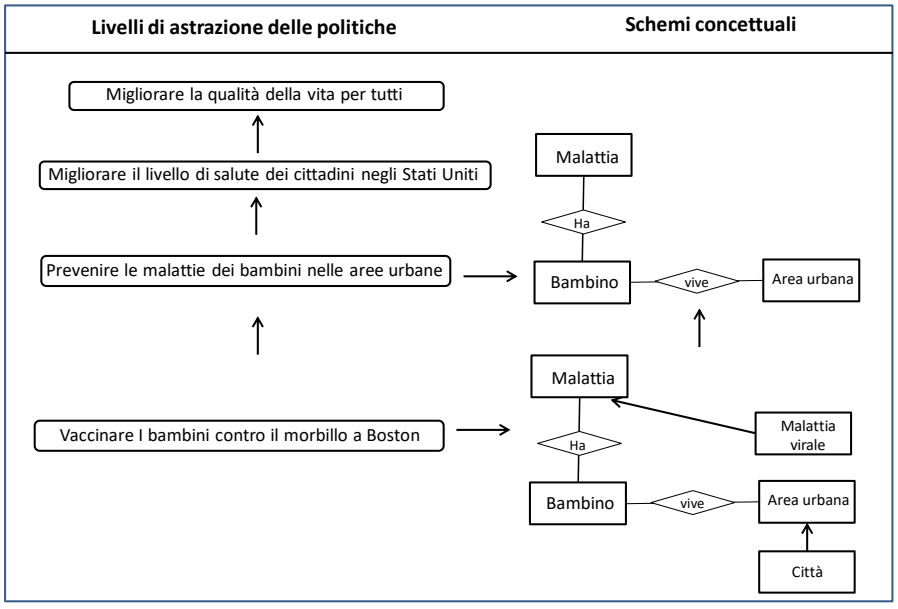


Figura 25 - Livelli di astrazioni nelle politiche

Le politiche sono rappresentate a diverso livello di astrazione, dal livello più specifico che riguarda una particolare comunità di una particolare città degli Stati Uniti, al livello più astratto, che riguarda gli interessi generali della popolazione degli Stati Uniti. Il diverso livello di astrazione che può essere adottato nei discorsi dei politici dipende da tanti elementi legati alla strategia di comunicazione e ai media che i politici utilizzano; concludiamo qui il ragionamento.

Per le astrazioni in Economia rimandiamo al Capitolo 13 sulla Economia digitale.

4.6 Tipi di astrazioni comuni a diverse discipline

In questo capitolo, ragionando sul tema delle astrazioni in diverse discipline, ho cercato di indagare su quali siano i concetti fondanti di una "Scienza" delle astrazioni, consapevole peraltro che siamo ai primi vagiti di tale scienza. La Figura 26 è un primo tentativo di individuare tipologie di astrazioni comuni a diverse discipline. Le tipologie sono classificate in livelli, secondo un processo simile a quello che abbiamo discusso per il layout di diagrammi nel Paragrafo 4.3. I livelli di astrazione sono:

- (il livello) formale, che riguarda modelli basati sulla logica o su sistemi formali, che si possono considerare al massimo livello di astrazione
- orientato ai processi, che fa riferimento alla descrizione dei processi di una organizzazione che producono i beni o servizi oggetto della missione della organizzazione. La specifica dei processi è l'attività iniziale della progettazione di sistemi informativi, si può quindi considerare a un elevato livello di astrazione.
- concettuale, quando l'artefatto è descritto con un modello o linguaggio vicino al punto di vista del progettista e lontano dalla implementazione fisica.
- topologico e morfologico, come li abbiamo intesi quando abbiamo discusso gli algoritmi per il layout automatico di diagrammi, facenti dunque riferimento alla natura topologica dell'artefatto e alla sua forma in termini di superfici iscrivibili in linee sul piano
- orientato alla interazione tra l'artefatto e l'ambiente o persona con cui l'artefatto interagisce; quando l'artefatto, ad esempio, è un robot, l'interazione riguarda la pianificazione della navigazione nell'ambiente in cui il robot opera.
- logico, inteso come livello che descrive in maniera "nativa" la struttura dell'artefatto, astraendo ancora dagli aspetti fisici; ad esempio, nelle basi di dati, il livello logico rappresenta la base di dati mediante relazioni, che sono la rappresentazione utilizzata dal software che gestisce la base di dati, ma non rappresenta gli aspetti fisici legati alla rappresentazione dei dati in memoria. Nelle mappe sul piano, la generalizzazione di modello, collocata a livello logico, utilizza punti, linee e superfici come primitive di base per descrivere una porzione del territorio.
- fisico, che descrive tutti gli aspetti fisici dell'artefatto (ad es. nelle basi di dati, la struttura fisica della memoria); questo è chiaramente il livello più basso di astrazione.

Livello	Intelligenza Artificiale	Disegno automatico dei diagrammi	Pianificazione in robotica	Progetto di basi di dati	Processo cartografico	Reti di Petri	Progetto di Sistemi informativi
Formale	Teoretico						
Orientato ai processi							Business
Concettuale	Linguaggio		Concettuale	Concettuale	Generalizzazione cartografica	Sistemi P/T Sistemi C/E	Concettuale
Topologico		Topologico	Topologico				
Morfologico		Morfologico	Morfologico / Spaziale				
Orientato alla interazione	Percettivo		Navigazionale	Interfaccia			Presentazione
Logico	Logico	Metrico	Metrico	Logico	Generalizzazione di modello		
Fisico				Fisico			Implementa-zione

Figura 26 - Confronto tra livelli di astrazione in domini e sottodomini analizzati in [Batini 2016]

Concludiamo qui la trattazione delle astrazioni nelle diverse discipline oggetto del Tutorial documentato in [Batini 2016].

5. Astrazioni e big data

Gli esempi precedenti di processi di astrazione ci hanno mostrato che noi possiamo migliorare la comprensione della struttura di un artefatto utilizzando un meccanismo generativo basato sui concetti di astrazione, e il suo complemento, il raffinamento. Questo paradigma, consistente nel rappresentare un artefatto a diversi livelli di dettaglio, è un pò, se permettete la espressione, come “sparare alla mosca con un cannone” nel caso di artefatti di piccole dimensioni (pensiamo a uno schema concettuale con dieci/venti tra entità e relazioni) ma diventa una strada obbligata quando l’artefatto sia composto da un numero di parti che supera una certa soglia, diciamo venti/cinquanta oggetti.

Nella Figura 27 vediamo tre diversi grafi, che rappresentano il primo un diagramma Entità Relazione, il secondo e il terzo un grafo composto di nodi e rami.

Il diagramma Entità Relazione è composto da 22 entità e 23 relazioni; per rappresentarlo in modo comprensibile abbiamo seguito i criteri estetici di cui abbiamo parlato nella Sezione 4.3: abbiamo collocato i simboli grafici in una griglia, abbiamo minimizzato gli incroci tra linee e abbiamo usato solo segmenti orizzontali e verticali. Inoltre, abbiamo rappresentato l’entità Genitore della generalizzazione al di sopra e simmetrico rispetto alle entità Figlie. Il tutto rende comprensibile il diagramma, e per identificare concetti e relazioni possiamo esplorare il diagramma secondo diversi movimenti degli occhi con un ridotto sforzo cognitivo.

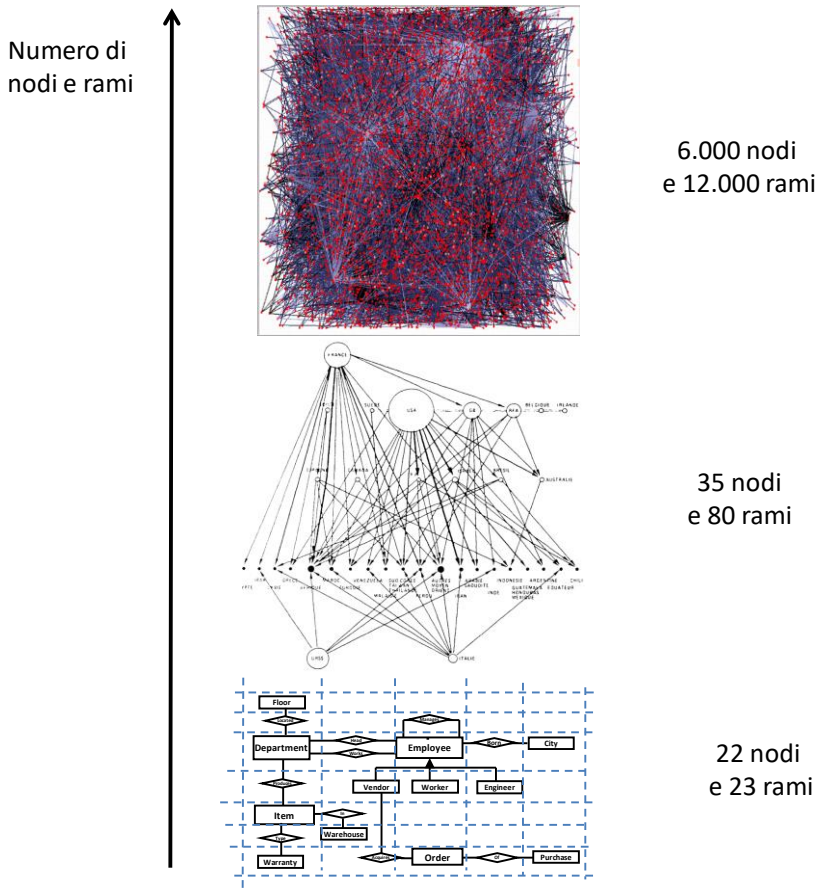
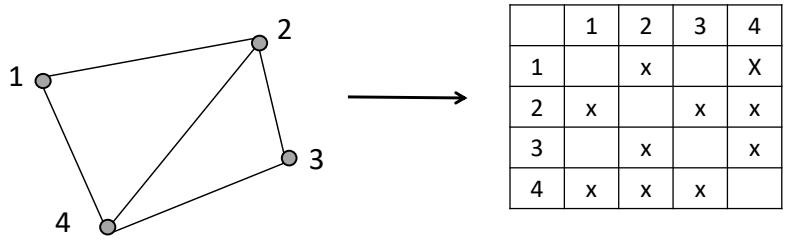


Figura 27 – Problemi nel rappresentare grafi con numero di nodi e rami molto grande

Il secondo grafo, al centro della figura, è composto da 35 nodi e 80 rami. La crescita del numero di rami da 23 a 80 rende impossibile rispettare la minimizzazione del numero degli incroci tra linee; la congettura che formulo (e che non provo in maniera formale, affidandomi piuttosto alla vostra intuizione) è che finchè sia possibile azzerare gli incroci tra linee ovvero ridurli a pochi, noi tendiamo a perseguire questo obiettivo; quando il grafo superi un certo livello di complessità, noi rilassiamo questo criterio, non ce ne preoccupiamo più, cercando di ottimizzare altri criteri. In questo caso, per dare un ordine e un minimo di struttura al grafo abbiamo collocato i nodi in corrispondenza a fasce verticali, in ogni fascia i nodi sono tutti allo stesso livello; infine per permettere una collocazione dei rami che non abbia sovrapposizioni eccessive, i nodi con più rami entranti sono rappresentati con cerchi più grandi.

Il grafico nella parte superiore della figura è composto da migliaia di nodi e rami, il suo layout sul piano produce una rappresentazione caotica e assolutamente incomprensibile.

Come possiamo utilizzare le astrazioni per riuscire a rappresentare strutture dati, e in particolare grafi, molto grandi? Anzitutto vediamo in Figura 28 le due più usate rappresentazioni dei grafi, quella con *nodi e rami* che abbiamo utilizzato finora, e la corrispondete rappresentazione con *matrice di adiacenza* in cui un grafo di n nodi è rappresentato mediante una matrice con n righe e n colonne, e in cui c'è una x in una cella (i,j) della matrice se i nodi i e j sono collegati da un ramo.



	1	2	3	4
1		x		X
2	x		x	x
3		x		x
4	x	x	x	

Nodi e rami

Matrice di adiacenza

Figura 28 – Rappresentazioni di grafi di tipo nodi e rami e di tipo matrice di adiacenza

La rappresentazione a matrice richiede meno spazio nel rappresentare grafi molto grandi, e, soprattutto, può essere utilizzata come struttura dati elaborabile da un programma; di per sé, però, non risolve il problema della compattezza, una matrice per rappresentare il grafo con 6.000 nodi di Figura 28 richiede una matrice di 6.000 righe e 6.000 colonne.

Una rappresentazione che cerca di conciliare l'efficacia visiva della rappresentazione di tipo nodi e rami e la compattezza della rappresentazione a matrice è mostrata in Figura 29.

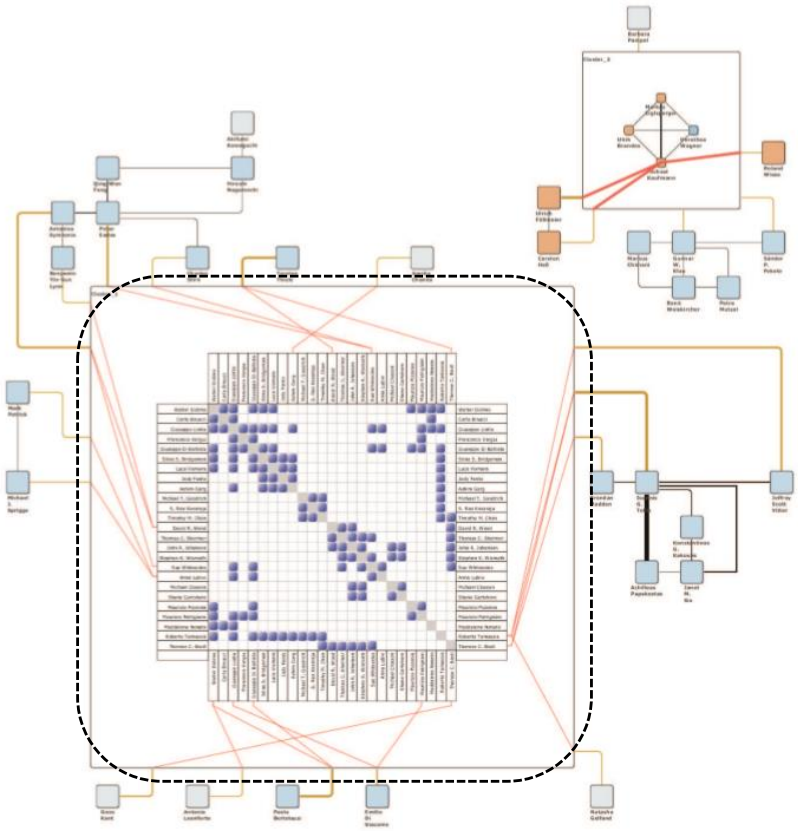


Figura 29 – Rappresentazione ibrida "nodi e rami" + "matrice di adiacenza" (tratta da [Batagelj 2010])

Vediamo rappresentata nell'area tratteggiata in Figura 29 una matrice di adiacenza, matrici, collegata con il resto del grafo mediante rami, che a lor ovolta collegano i nodi identificati dalle posizioni nella matrice con altri nodi esterni ad essa.

Anche la rappresentazione ibrida di Figura 29 non risolve il problema dei grafi molto grandi. E' necessario adottare un meccanismo analogo quello che abbiamo visto per il Reposotory di schemi concettuali, adattando le trasformazioni alla struttura di dati matrice di adiacenza.

In Figura 30 tratta da [Elmqvist 2008] mostriamo una rappresentazione piramidale di una matrice di adiacenza, in cui nel livello più basso è rappresentata una matrice con n righe e n colonne, che rappresenta dunque un grafo di n nodi. Al livello immediatamente più alto è rappresentata una matrice con n/2 nodi, e così via per i livelli superiori; a un quadrato di dimensione 2 x 2 della matrice al livello più basso corrisponde una sola cella nel livello immediatamente più alto. La matrice ha log2(n) livelli.

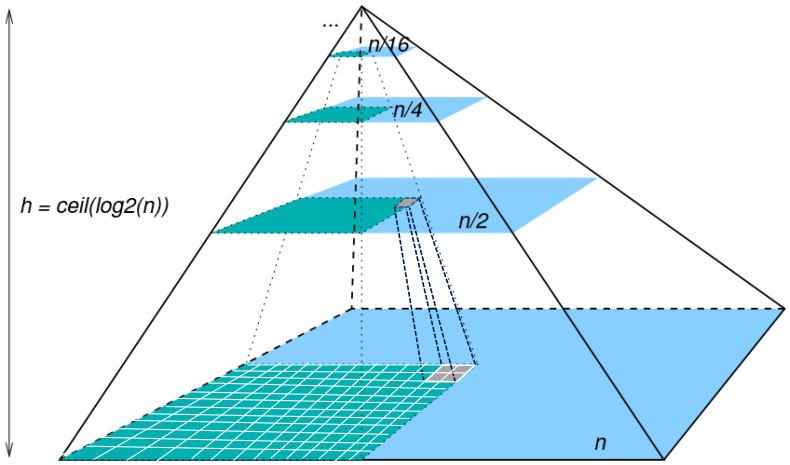


Figura 30 – Rappresentazione di un grafo mediante matrici di adiacenza
a successivi livelli di astrazione

Se passiamo dai grafi alle mappe, troviamo problemi di analogo "affollamento" di simboli anche per piccole mappe. In Figura 31 mostro una mappa di Lione, che secondo me è assolutamente illeggibile; troppi simboli, troppo vicini, troppo piccoli.

I cartografi da molto tempo sono abituati a trattare il problema delle mappe, rappresentandole a scale diverse; la scala di una mappa è il rapporto tra, ad esempio, una linea lunga un centimetro nella mappa e la corrispondente lunghezza della linea nel territorio rappresentato.

Permettete una breve divagazione sulle mappe. So bene che orami sono diffusi sui telefoni cellulari i navigatori, e che le mappe non si uso quasi più quando uno guida. E so bene che le mappe per coloro che camminano in montagna possono essere scaricate sul cellulare, e quindi le mappe cartacee tradizionali vivono di questi tempi una vita grama; tuttavia, a me consultare le mappe piace moltissimo, sia quando guido che quando vado in montagna, perché non fanno soltanto vedere il punto in cui siamo,

ma fanno vedere il contesto, il territorio più ampio in cui siamo immersi, e non mortificano, nel secondo caso, la grandiosità della montagna.

Per le mappe stradali viene usualmente fornita una mappa a scala 200.000 (un centimetro corrisponde a due chilometri), mentre per mappe di sentieri montani si usano mappe al 25.000 (un centimetro corrisponde a 250 metri).

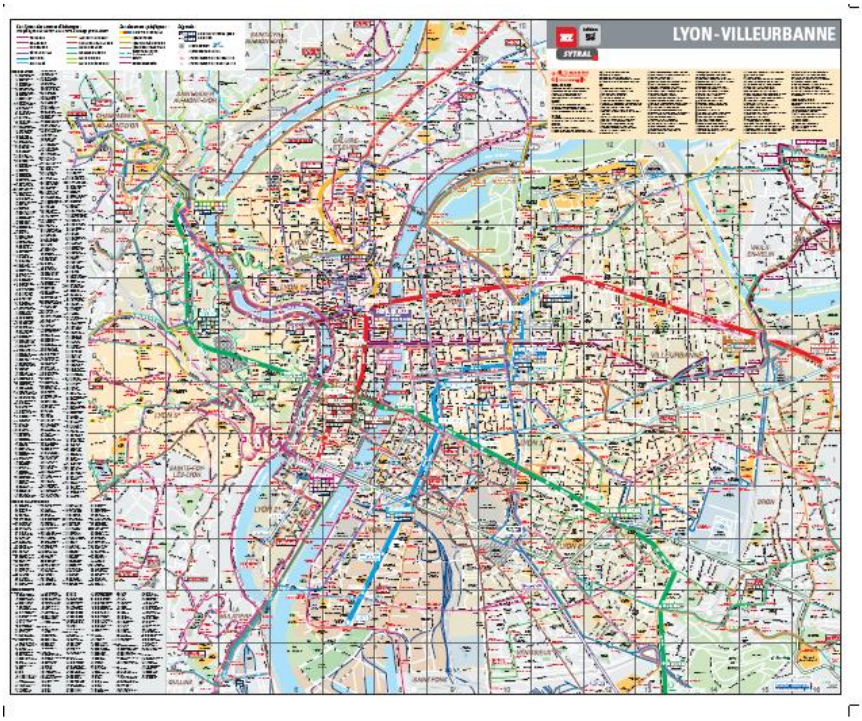


Figura 31 – E' difficile leggere le mappe con una elevata densità di simboli (das www.mapsa.net)

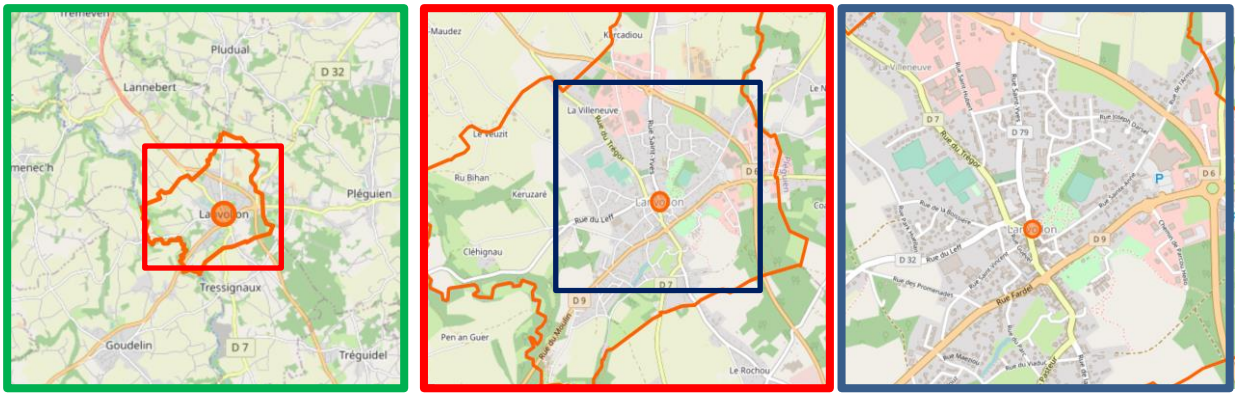


Figura 32 – Una mappa stradale rappresentata con tre scale diverse
(Lanvollon, Francia, da Open street map)

Dunque, nelle mappe vengono usati diversi livelli di dettaglio corrispondenti a diverse scale. In Figura 32 mostro un esempio di carta stradale a tre successivi livelli di dettaglio. Nel passare da una mappa ad una determinata scala a una a scala più ridotta, come vedete, vengono adottate delle trasformazioni, il

paese di Lanvollon che nella prima mappa è rappresentato con le sole strade principali, successivamente è rappresentato mediante una rete di strade più ampia, e nella terza mappa evidenziando gruppi di edifici. Quindi, anche per le mappe vengono usate tante trasformazioni, che i cartografi adottano da molto tempo. Siamo lontani dalla mappa uno a uno del regno di Babilonia, citata in uno scritto di Borges "Storia universale dell'infamia", Il Saggiatore, 1961, traduttore Mario Pasi, che i geografi babilonesi produssero per il loro Imperatore, rappresentando l'impero con una mappa che lo riproduceva esattamente!

Un insieme non esaustivo di trasformazioni adottate per passare da una scala ad una più piccola è mostrato in Figura 33. Vediamo che la prima trasformazione (smoothing) opera una riduzione degli angoli nei segmenti che rappresentano una linea; ciò corrisponde a una astrazione in cui i dettagli che vengono eliminati sono gli angoli tra segmenti, sostituiti da linee curve. L'ultima trasformazione (symbolization) sostituisce una rappresentazione veridica di una stazione ferroviaria con un simbolo di locomotiva (un pò di vecchia concezione…), quindi in questo caso l'astrazione sostituisce una metafora (la locomotiva) a un insieme di dettagli costituiti da un fascio di binari.

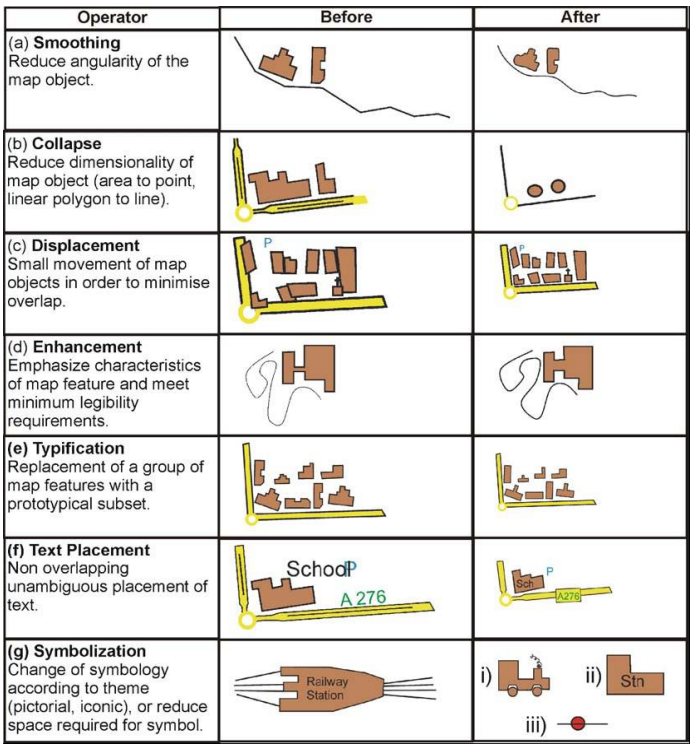


Figura 33 - Alcune trasformazioni adottate nelle mappe per passare da una scala a una inferiore (da Encyclopedia of Geographic Information Systems, Springer Verlag)

Insomma, le trasformazioni sono tipiche di molte rappresentazioni di artefatti, e la loro natura, i simboli su cui operano, le regole per la loro applicazione dipendono dal modello adottato nella rappresentazione e dal significato dei simboli o concetti che costituiscono il modello. Per concludere questa sezione, confrontiamo in Figura 34 artefatti costituiti da schemi concettuali, mappe e grafi e trasformazioni che possono essere adottate per costruire rappresentazioni compatte a partire da rappresentazioni caratterizzate da maggiori dettagli.

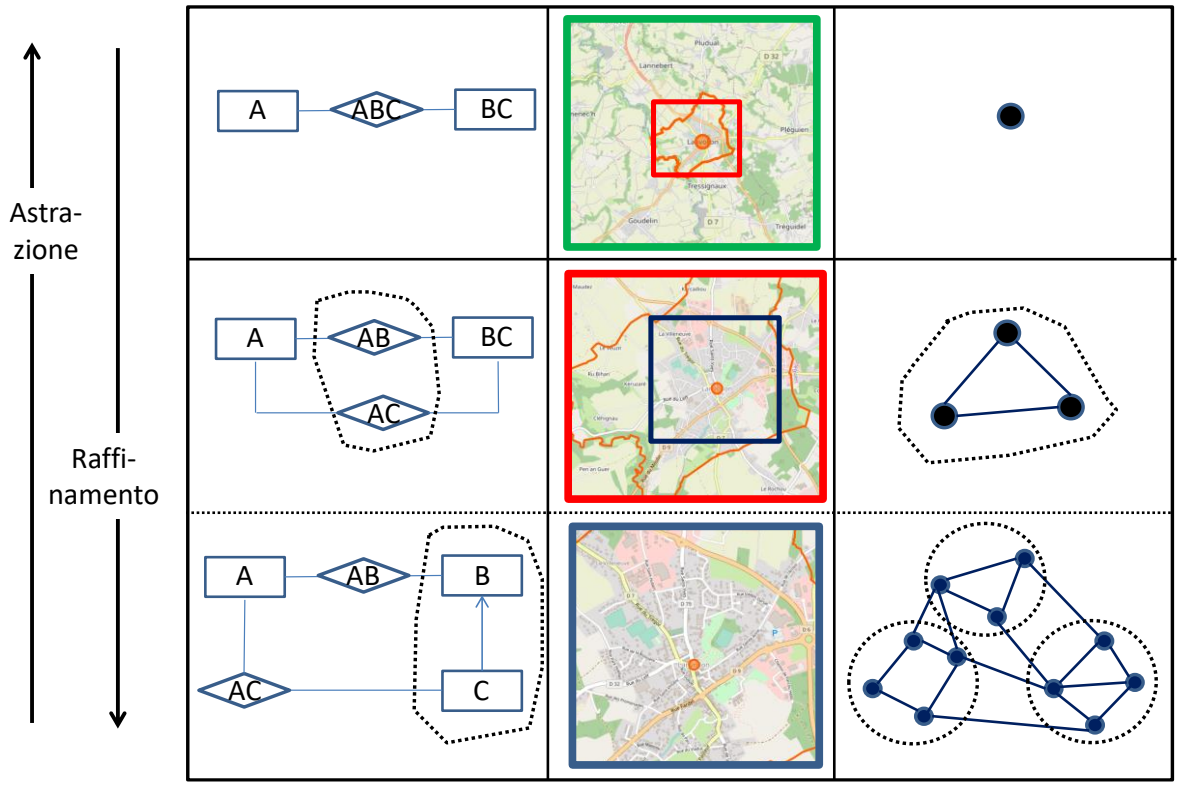


Figura 34 – Processi di astrazione e trasformazioni a confronto tra schemi concettuali, mappe e grafi

Come si vede, tre tipi di rappresentazioni molto diverse tra loro, gli schemi concettuali, le mappe, e i grafi condividono nei processi di astrazione un insieme di concetti che ho descritto nel capitolo: astrazioni e raffinamenti, livelli di astrazione, tipi di trasformazioni e trasformazioni elementari necessarie per passare da un livello di astrazione all'altro.

Riferimenti

V. Batagelj et al. - Visual analysis of large graphs using (x, y)-clustering and hybrid visualizations." IEEE Transactions on visualization and computer graphics 17.11 - 2010.

C. Batini, S. Ceri, e S. Navathe - *Entity Relationship Approach* - Elsevier Science Publishers, North Holland, 1989.

C. Batini, G. Di Battista, G. Santucci - Structuring primitives for a dictionary of entity relationship data schemas - *IEEE Transactions on Software Engineering* - 19.4, 1993.

C. Batini, M. Comerio, G. Viscusi - Managing Quality of Large Set of Conceptual Schemas in Public Administration: Methods and Experiences. MEDI, 2012.

C. Batini – Presentazioni power point su Tutorial su Abstractions in Computer Science and surroundings, Entity Relationship Conference, Gifu (Japan), 2016, accessibile su https://boa.unimib.it/preview-item/187839?queryId=mysubmissions&

C. Batini, M. Talamo, R. Tamassia: Computer aided layout of entity relationship diagrams. Journal of Systems and Software 4(2-3), 1984.

M. Boisot - Information space: a framework for learning in organizations, institutions, and culture. London ; New York: Routledge, 1995.

W. Brauer, R. Gold, e W. Vogler - A survey of behaviour and equivalence preserving refinements of Petri nets. In International Conference on Application and Theory of Petri Nets (pp. 1-46). Springer Berlin Heidelberg, 1989.

L. Campbell e T. A. Halpin, and Henderik Alex Proper. "Conceptual schemas with abstractions making flat conceptual schemas more comprehensible." *Data & Knowledge Engineering* 20.1, 1996.

E. R. Gansner, Yehuda Koren e Stephen North - Topological Fisheye Views for Visualizing Large Graphs – IEEE Transactions on Visualization and graphics. 2003.

N. Elmqvist et al. - ZAME: Interactive large-scale graph visualization - IEEE Pacific Visualization Symposium., 2008.

P. L. Ferrari - Abstraction in mathematics - The Royal Society 2013.

F. Giunchiglia & Walsh, T. A theory of abstraction. Artificial intelligence, 57(2-3), 1992.

M. Keet - A taxonomy of types of granularity - GrC. 2006.

B. Kuipers - The spatial semantic hierarchy - Artificial intelligence 119.1-2, 2000.

S. Ichiro e T. Murata - A method for stepwise refinement and abstraction of Petri nets - Journal of computer and system sciences 27.1., 1983.

M. Monmonier - How to Lie with Maps. University of Chicago Press, Chicago, 1991.

D. Moody e Andrew R. Flitman - A Decomposition Method for Entity Relationship Models: A Systems Theoretic Approach - *ICSTM* 2000.

D. Moody - The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. Ieee Trsansactions in Software Engineering, 2009.

J. Mylopoulos - Information Modeling in the Time of the Revolution - Information systems 23.3, 1998.

Online Ethimology dictionary, https://www.etymonline.com/word/abstraction, acceduto il 10 Agosto 2018.

M. Palmonari, e C. Batini - Abstract ERIA: a web language for conceptual metadata integration and abstraction in the large. MEDES, 2009.

M. Palmonari e C. Batini. "Design, redesign and publication of linked schema repositories in the large." Proceedings of the International Conference on Management of Emergent Digital EcoSystems. ACM, 2010.

L. Saitta e J.D. Zucker - Abstraction in Artificial Intelligence and Complex Systems, Springer 2015.

S. Shekhar e H. Xiong, eds. - Encyclopedia of GIS - Springer Science & Business Media, 2007.

J. M. Smith e D. C.P. Smith - Database abstractions: aggregation and generalization - ACM Transactions on Database Systems, 2.2, 1977.

J. M. Smith e D. C.P. Smith - Database abstractions: aggregation - Communications of the ACM 20.6, 1977

J.F. Sowa - Knowledge representation: logical, philosophical, and computational foundations. China Machine Press. 2000.

N. Sturtevant, e M. Buro - Partial pathfinding using map abstraction and refinement - AAAI. Vol. 5. 2005.

I. Suzuki e T. Murata. "A method for stepwise refinement and abstraction of Petri nets." Journal of computer and system sciences 27.1 (1983): 51-76.

R. Tamassia, G. Di Battista e C. Batini - Automatic graph drawing and readability of diagrams. IEEE Trans. Systems, Man, and Cybernetics 18(1), 1988.

J. B. Tenenbaum e T. L. Griffiths - Generalization, similarity, and Bayesian inference - Behavioral and Brain Sciences, 2001

S. Timpf, et al. - A conceptual model of wayfinding using multiple levels of abstraction - Theories and methods of spatio-temporal reasoning in geographic space. Springer Berlin Heidelberg, 1992.

Appendice 1 – Le 400 astrazioni del Tutorial 2016

In questa pagina e nelle successive mostro le 400 astrazioni che ho individuato durante la preparazione del Tutorial che feci nel 2016 [Batini 2016]. Le astrazioni compaiono classificate in diverso modo, e la classificazione cui corrispondono è riportata in colonna 1.

- **Abstraction**
  - atomic
    - term
    - formula
    - symbol
    - arity
  - type of
    - hiding
    - equating
    - hierarchy building
    - combining
    - approximating
  - operational
- *Abstraction (by forgetting)*
  - Viewpoint (View)
  - Abstraction from restriction (by forgetting)
  - Abstraction from functions (by forgetting)
  - Deletion
  - Delete-less relevant-concepts a. in ontologies
  - Deleting argument
- Precondition-elimination A.
- (Abstraction) by clustering
- (Abstraction) by summarization
- Abstraction by hiding
- Elimination
- Breadth abstraction
- Depth abstraction
- Cartographic Abstraction
- Schema abstraction
- Abstracting away
- Abstracting out
- Physical abstr.
- Mathematical abstr.
- Metaphysical abstr.
- Semi-topology a.
- Simple node a.
- Full Mesh a.
- Source Star a.
- Star a.
- Transition abstraction (refinement)
- State abstraction (refinement)

- Organizational abstraction
- Spatial a. (reasoning)
- Visualization A.
- Analytical A.
- **Generalization (is-a, Inclusion)**
  - On the fly generalization
  - Dynamic
  - Generic/generic (Brachmann)
  - Generic/individual
  - Lenght
    - sic
  - Angle
  - Shape
- Conceptual generalization
- Generalization in Vis. - GIS and maps
- Map generalization
- Cartographic Generalization
- Model generalization (Geodatabase a.)
  - Selection
  - Re-classification
  - Aggregation
  - Area collapse
- Streaming generalization
- Subsumption
- **Aggregation (Part-of)**
  - Mereological
    - Structural part-of
      - Functional part-of
    - Spatial part-of
      - Contained-in
      - Located-in
    - Involved-in
  - Meronymic
    - Member-of
    - Constitued-of
    - Sub-quantity-of
    - Particicipates-in
- State aggregation
- Linked data aggregation
- Function aggregation

- **Classification (instance-of)**
  - Recursive
  - Constrained
  - Unconstrained
  - Dynamic
- **Granularity**
  - Scale dependent
    - with relation to granularity size
    - with some form of aggregation
      - carving up ad different levels
      - using aggr. of the top type
  - Non scale dependent
    - Levels adopt one single type of relation
    - Folding
    - With some form of aggregation
      - aggr. is attributed to a collective ent.
      - ..that can be partitioned
  - Sowa granularities
    - Firstness
    - Secondness
    - Thirdness
- **Module A. in ontologies**
  - Functional
    - Design pattern
    - Subject domain
    - Isolation branch
    - Locality
    - Privacy
  - Structural
    - Domain coverage
    - Ontology matching
    - Optimal reasoning
  - Abstraction hiding
    - Axiom abstraction
    - Entity type
    - High level abstraction
    - Weighted
  - Expressiveness
    - Sublanguage
    - Feature

- **Other abstractions in ontologies (Keet)**
  - **Relation-to-function a.**
    - Basic - sub-supertype (class) abstraction
    - B - part into its whole
    - B - part-process into the whole-process
    - B -Abstract an endurant into a perdurant
    - B - Abstract a type into 'nothing'
  - **Folding a.**
    - Comp. - Abstract two end. into anoth. end.
    - C - Abstract two perdurants into a perd.
    - C - Abstract an end. and a perd. into an end
    - C - Abstract an end. and a perd. into a perd.
  - Simplifying a.
  - Semantic deletion a.
    - Remove an attribute
- **Sampling**
  - 1. remove noise nodes
  - 1. remove unimportant nodes
  - 2. structure based
  - 2. content based
- **Hierachization A.**
  - k-level abstraction hierarchy
  - Hierarchization
  - Hierarchical structuring
  - Pyramid
  - Repository of conceptual schemas
  - Hierarchical simplification
  - Hierachy A. in GIS
  - (Hierarchical) Embedding
  - Hierarchical Bayesian models
  - Abstraction hierarchy
  - Repository of ontologies
  - Hierarchical video summarization
- **Grouping**
  - Multiple
  - Grouping (in visualization)
  - G. by logical horizon
  - Subject g.
  - Category

- **Filtering**
  - Property-based
  - Local connectivity
  - Global connectivity
  - Complex global connectivity
  - Sophisticated connectivity
  - Formal ontological
  - External criteria f.
  - Importance based
  - Filter (in feature selection)
  - Closeness between entitiy types based
  - Bilateral (in images)
  - Region boundary simplific. by shock filt. (images)
  - Filtering by patterns
  - Interest based
- **Smoothing**
  - Surface
  - Curve
  - Region Smoothing in visualization
- **Contextualization**
  - Workspace
  - Version
  - Configuration
  - View
  - Scope
  - Perspective
- **Clustering**
  - Still image a. by Clustering
  - Moving image a.
- **Summarization**
  - Summary
  - Video skimming by summarization
- **Selective a.**
  - **Fisheye**
    - Distorsion
    - Topology driven fisheye view
    - Variabile zooming
    - Semantic fisheye
  - Enhancig Region Perceptibility in visualiz.

- **A. in Visualization - images**
  - Categorization
  - 3D Abstr.
  - Segmentation
    - Color based
    - Mean shift
    - triangle-based
    - Mean curvatureflow
  - Multi-scale based on gradient reconstruction
  - Importance-adaptive abst.
  - Regularization by Mean curvature flow
  - R. by constrained mean curvature flow
  - Non linear diffusion
  - *Evolution of abstractions in images*
    - Symmetry based, volumetric shape abstractions
      - Superquadratic ellipsoid
      - Geon
      - Shock model
    - Shape abstr.
      - Functional object description
    - Structure abstraction
      - Parametrizing
    - Lowest common a. of a set of graphs
- **Abstr. In Visualizations - photos**
  - Stylistc abstr.
  - Non-photorealistic rendering
- **A. in Visualization - video**
  - Still image abstr. (video summary)
    - Sampling-based Keyframe Extraction
    - Shot-based keyframe Extraction
    - Color-based
    - Motion-based
    - Mosaic-based
    - Segment-based keyframe
  - Moving image abstr. (video skimming)
    - Summary sequence
    - Highlight
      - by Dynamic sampling
- **A. in Vis. - Automatic layout of diagrams**
  - Topological A.
  - Shape A.
  - Metric A.

- **(Other) Abstraction in mathematics**
  - Decontextualization
  - Reification
  - Equivalence relation
  - Lexicalization
  - Nominalization
  - Register
  - Generic example
  - Elicitation of form over content
- **Mapping Abstractions**
  - Predicate mapping
  - Mapping btw formal syst (semantic)
  - Mapping of languages (syntactic)
  - Axiomatic mapping
  - Object mapping
  - Syntactical mapping
  - Semantic mapping
  - Perceptual mapping
  - Theorem..
    - decreasing
    - increasing
    - constant
- **Pattern**
  - Creational
  - Structural
- Abstract knowledge pattern
- Filtering by patterns
- Design pattern
  - Behavioural
    - using delegation
    - using aggregation
    - using consultation
- **Bayesian A.**
  - Bayesian Inference
  - Hierarchical Bayesian model
- Approximation
  - Function
  - Probability function approximation

- **A. in Machine Learning**
  - Feature selection
  - Instance selection
  - Feature discretization
  - Feature construction
  - Predicate invention
  - Term abstraction
  - Propositionalization
- **Abstractions in technologies, architectures, systems**
- *Abstractions in software engineering*
- A. by parametrization
- A. by specification
- Realization abstraction
- Procedural a.
- Polymorphic abstr.
- Control a.
- Iteration a.
- Abstract data type
- Procedural data type
- Interface abstr.
- Subroutine
- Function
- Procedure
- Pointer
- Metamodel
- Parts of an abstraction in se.
  - Hidden
  - Variable
  - Fixed
- *Abstraction in databases*
  - Integrated schema
  - Global schema in Distributed data bases
  - Fragmentation schema
  - Replication schema
  - Conceptual schema
    - Filtered
  - Logical Schema
  - Physical Schema
  - Normal form (in relational theory)
  - Key abstraction

- *Abstractions in Information Systems*
  - Corporate high level model
  - Corporate subject area
  - Organizational view area
  - Information Area
- *Zachman model for Information Systems*
  - Scope model (Contextual)
  - Business model (Conceptual)
  - System model (logical)
  - Technology model (Physical)
- Separation of concern
- *Abstr. In Model Driven Architecture*
  - Business (or domain) model
  - Logical system
  - Implementation
- *Abstr. In Meta Object Facility*
  - M3 - level Meta-metamodel
  - M2 - level Meta-model
  - M1 - level User-defined model
  - M0 - level Object Diagram
- *Abstr. In information system design*
  - Strategic
  - Business
  - Conceptual
  - Presentation
  - Implementation
- **Abstractions in processes**
  - Process model
  - Process abstraction
  - Macroprocess
  - Process
  - Activity
  - Subactivity
- **Abstractions in Feature selection**
  - Wrapper
  - Embedded
- Feature discretization
- Instance selection
- Feature construction

Abstractions in Event/time		
	Explicit cl.	
	Multilevel (hierarchical) cl.	
	Cluster adjusting	
	Temporal a. (reasoning)	
A. In roads networks		
	Planning	
	Instructional	
	Driver	
Other non classified types of abstractions))		
Conceptualization		
Materialization		
Ownership		
Generation		
Versioning		
Realization		
Normalization (normal case first)		
Integration		
Parametrization		
Deep instatiation		
Multi level objects		
Multi level relationship		
Powertype		
Ownership		
Role		
Modularization		
Well-definedness		
Entity expansion		
Semantic		
Geometric		
Geographic data reduction		

Vocabulary of map symbols enrichement	
Geometry enhancing	
Schematizing	
Caricaturing	
Change of geometry type	
Simple dilatation	
Simplification	
Squaring	
Enlargement	
Thematic abstr.	
Perceptual a.	
Perceptual constancy	
Exoskeleton	
Lossy compression	
Regularization	
Recursion (Recursive cognition)	
Information a.	
Graphic	
Layout	
Degree of abstraction (in graphs)	
Contraction (telecommunications)	
Clique	
Behavioural	
Numerical	
Data a.	
Formal	
Descriptive	
Conceptual	

Generic		
Ground		
Domain		
Predicate invention		
Term abstraction		
Propositionalization		
State feature selection		
Task decomposition		
Function approximation		
Categorization		
Naming		
Vocabulary of terms enrichement		
Navigation A.		
Conceptual A.		
Effect A.		
Task A.		
Sensor annotated hierarchical a.		
Dataflow a.		
Iteration		
Building the abstract		
Empirical		
Reflective		
Codification		
Distilling the essence		

Capitolo 13 – L'Economia Digitale

Roberto Masiero

1. Introduzione

Le tecnologie dell'informazione, con la loro rapida evoluzione, contribuiscono in modo determinante alla trasformazione della nostra economia, delle nostre imprese e dei nostri stessi modi di vita. Tra le altre, le cinque considerate più rilevanti (vedi [Grefen 2016]) come abbiamo detto nel Capitolo 1, sono il cloud computing, l'internet delle cose, i social media, il mobile computing, che tutti insieme contribuiscono alla esplosione dei dati digitali.

Come risultato, l'Economia digitale o Digital Economy sta crescendo rapidamente. Sotto questo nome vengono inclusi diversi fenomeni, per cui il significato e le metriche della Digital Economy sono molto vari. Citiamo innanzitutto la definizione che ne dà Deloitte, che ci pare particolarmente puntuale: *"La Digital Economy è l'attività economica che risulta da miliardi di connessioni online che avvengono ogni giorno fra persone, business, dispositivi, dati e processi. Il cuore della Digital Economy è l'"Iperconnettività", che significa la crescente interconnessione fra persone, organizzazioni e macchine fra Internet, le tecnologie mobile e l'internet of Things" (IoT)"* [Deloitte 2017].

Per quanto riguarda un approfondimento delle definizioni, la concettualizzazione e la misurazione della Economia digitale facciamo riferimento al lavoro [Bukht 2017]. Secondo questo studio il concetto di Economia digitale si articola in tre diverse dimensioni. Il "cuore" è il settore digitale, ovvero il settore IT/ICT che produce le tecnologie e i servizi di base. La "vera" Economia Digitale – definita come quella parte dell'output economico che deriva soltanto o primariamente dalle tecnologie digitali con un business model basato su beni o servizi digitali – consiste del settore digitale più i servizi digitali emergenti e l'economia delle piattaforme (torneremo su tutti questi concetti). Gli autori definiscono infine l'Economia Digitale nella sua nozione più ampia – da essi definita Economia digitalizzata – come l'insieme degli utilizzi dell'ICT in tutti i settori economici (vedi Figura 1).

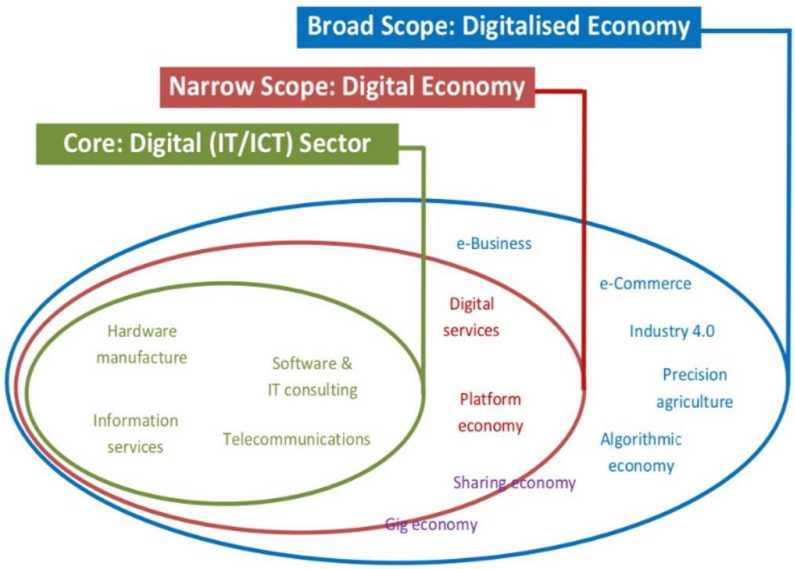


Figura 1 - Le tre dimensioni della Economia Digitale (tratta da [Bukht 2017])

Definito così il perimetro del nostro campo di analisi, si tratta ora di chiarire i principi e le regole che governano l'Economia digitale. Le domande cui intendiamo rispondere in questo capitolo sono: l'informazione e la conoscenza vanno considerati come asset strategici? Quali modelli economici sono stati proposti per determinarne il valore? E perché questo non è ancora adeguatamente registrato nei bilanci di esercizio delle aziende? Che opportunità e che problemi pone l'ascesa impetuosa dell'economia degli artefatti intangibili? Come rispondere alle sfide dell'economia dell'accesso, della società a costo marginale zero, dei mercati "data rich" rispetto a quelli "capital rich"? E infine quali nuovi parametri governano l'economia delle piattaforme, quali cambiamenti intervengono quando si passa dai modelli "pipeline", che seguono cioè un approccio classico della catena del valore, a modelli a piattaforma, e in che modo, con quali modelli di business i nuovi protagonisti che stanno facendo irruzione nella Economia digitale si candidano a dominare l'economia mondiale?

A queste e ad altre domande ci proponiamo di rispondere nelle pagine seguenti. Nella Sezione 2 discutiamo su come la informazione possa diventare un asset (cioè un patrimonio) delle imprese, e delle difficoltà connesse alla misurazione del loro valore economico. La Sezione 3 espone la visione di Max Boisot sui concetti di dato, informazione e conoscenza, cui abbiamo accennato nel primo capitolo, e mostra come il valore della informazione cresca in uno spazio a tre dimensioni, che consistono nella scarsità, concetto da molto tempo proprio della economia, la codifica e la astrazione, concetto quest'ultimo discusso nel capitolo precedente. La Sezione 4 affronta il problema di quali nuove regole economiche sono rispettate dai dati digitali visti come bene economico. La Sezione 5 affronta il tema del rapporto tra i dati e il prezzo nello scambio dei beni, delineando una evoluzione da mercato "capital rich" verso un mercato "data rich", esprimendo un punto di vista nuovo e originale. La Sezione 6 si concentra sulla economia degli intangibili, cui appartengono gli artefatti costituiti dai servizi e dai dati, individuando le nuove forme che caratte-rizzano il mercato degli intangibili. La Sezione 7, infine, discute la grande rivoluzione che nella Economia digitale è svolta dalle piattaforme digitali planetarie, ad es. Uber e Amazon.

2. Gestire le informazioni come asset strategico per creare valore economico

2.1 L'informazione come asset strategico della impresa

Il primo punto da chiarire quando si affronta il tema generale dell'economia digitale è: possiamo legittimamente considerare l'informazione (termine che useremo in questa sezione al posto dei dati per rispettare la terminologia dell'autore) come un asset strategico dell'impresa, intendendo per asset il patrimonio della impresa, e dunque la risorsa utilizzabile nei processi di una organizzazione per creare valore economico? In altre parole, l'informazione soddisfa le caratteristiche essenziali che contraddistinguono un asset aziendale?

Secondo Moody in [Moody 1995] la risposta è positiva per i seguenti motivi:

1. Un asset è caratterizzato dal fatto di poter fornire servizi o benefici economici futuri, attraverso il suo utilizzo o la vendita dello stesso. L'informazione soddisfa questo requisito, poiché essa soddisfa la capacità di fornire servizi e di formulare decisioni efficaci.
2. Un asset è controllato dall'organizzazione a cui appartiene, nel senso che solo questa può beneficiarne e negarne o regolarne l'accesso ad altri. L'informazione soddisfa anche questo requisito, in quanto solo la organizzazione ha la possibilità di accedere alle informazioni che essa detiene, salvo che decida di venderla o di concederne accesso a terzi.
3. Un *asset* è il risultato di transazioni precedenti, ovvero il *controllo sull'asset* deve essere il risultato di transazioni precedenti come acquisti, sviluppi interni o scoperte. L'informazione soddisfa anche questo terzo requisito, in quanto essa può essere raccolta come il by-product di precedenti transazioni (sviluppo interno), o esser stata acquistata (una mailing list), ovvero essere l'esito di una scoperta (attraverso l'analisi di dati).

L'informazione è quindi generalmente riconosciuta come una risorsa economica e un asset strategico per le imprese, anche se non adeguatamente valorizzato. Mentre infatti le imprese consumano quantità di risorse organizzative sempre maggiori per acquisirla, memorizzarla ed elaborarla, essa invece non viene adeguatamente valorizzata all'interno del bilancio d'esercizio delle imprese stesse. Mentre l'hardware e – in alcuni casi – anche il software vengono capitalizzati all'interno dello stato patrimoniale, altrettanto non accade per quanto riguarda la valorizzazione delle informazioni, benché queste rappresentino spesso un vantaggio competitivo essenziale per l'impresa che la controlla. Chiunque provi ad andare in banca ad offrire come garanzia per un mutuo una base di dati che fornisce una descrizione analitica di un mercato, che tanto lavoro e tante risorse ha assorbito, capirà immediatamente di cosa stiamo parlando.

L'informazione peraltro offre la capacità di fornire servizi, assumere migliori decisioni, migliorare i risultati, acquisire un vantaggio competitivo e può anche essere commercializzata direttamente come prodotto. Usando un'analogia col settore manifatturiero:

- I dati sono la materia prima
- Il software e l'hardware sono gli impianti e gli strumenti.

L'informazione può essere sia il semilavorato che il prodotto finale offerto al cliente, vedi Figura 2. Nella figura l'informazione è vista come l'esito delle trasformazioni e arricchimenti prodotti dalle tecnologie utilizzate nella organizzazione, quindi in modo coerente con il significato che abbiamo attribuito al termine nel Capitolo 1. In questo capitolo continueremo perciò ad utilizzare il termine informazione, anche in coerenza, come già detto, con il testo [Moody 1995] cui ci stiamo ispirando.

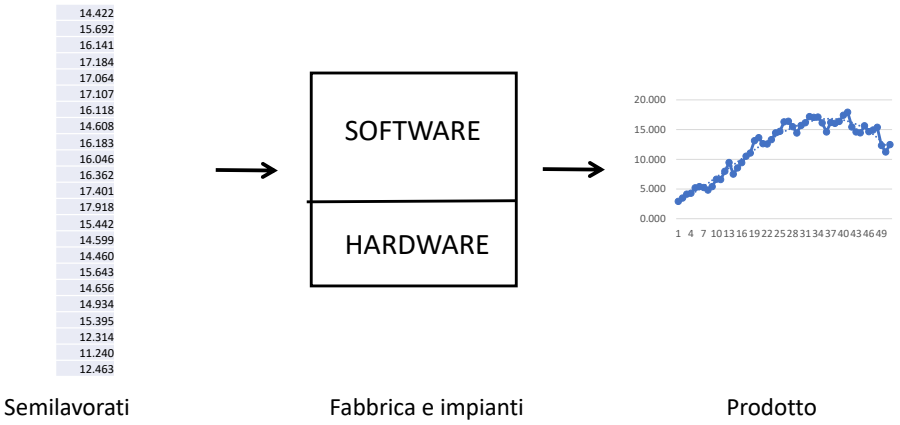


Figura 2 - L'Informazione come asset importante, ma non abbastanza valorizzato

Possiamo quindi concludere che l'informazione è un asset intangibile. Inoltre:

- come ogni altro asset organizzativo, l'informazione ha un costo (il costo per acquisirla, memorizzarla e mantenerla) e un valore (il suo valore per l'organizzazione).
- l'informazione non obbedisce alle stesse leggi economiche che caratterizzano gli altri asset – ma ha alcune proprietà peculiari che devono essere chiaramente comprese per poterne misurare il valore.

Possiamo quindi definire la natura dell'informazione in quanto asset intangibile identificando alcuni principi generali – o leggi, nel seguito – che ne governano il comportamento come bene economico. La sezione successiva approfondisce queste leggi.

2.2. Le sette leggi di Moody e Walsh che governano il comportamento dell'informazione come bene economico

Il testo cui facciamo riferimento in questa sezione è [Moody 1995], in cui vengono proposte sette leggi sulla informazione come bene economico. Le leggi, come vedremo tra poco, sono espresse in modo qualitativo, senza esprimere, in genere, una precisa relazione matematica tra la variabile descrittiva della informazione (ad esempio nella Figura 3 il numero di utenti) e il suo valore economico. Tuttavia, esprimono un punto di vista interessante, se vogliamo, proprio per la loro semplicità. Vediamole.

<u>Prima Legge: l'informazione è infinitamente condivisibile.</u>

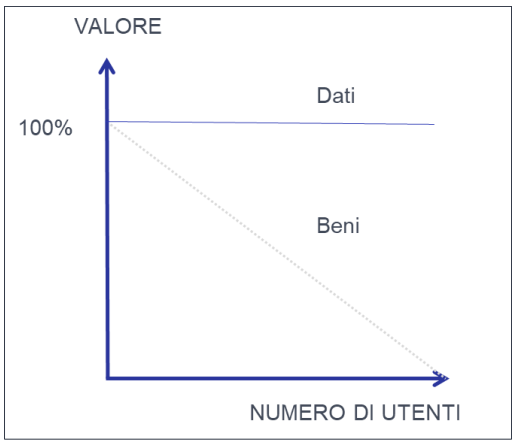


Figura 3 - L'Informazione è infinitamente condivisibile

La caratteristica pressoché unica dell'informazione come asset è che essa può venire condivisa fra un numero qualsivoglia di persone, aree di business ed organizzazioni, senza perdita di valore per alcuno, vedi Figura 3. Un esempio tipico è l'enciclopedia libera Wikipedia: essa è consultabile da chiunque disponga di una connessione a Internet, senza limiti di accesso.

Ne consegue che:

- Mentre la maggior parte degli asset sono appropriabili (o li possiedi o non li possiedi), l'informazione può essere condivisa fra molteplici aree di business di una impresa, mantenendo lo stesso valore per ognuna di esse: il valore è cumulativo invece che venire suddiviso tra i vari soggetti.
- In generale, condividere l'informazione tende piuttosto a moltiplicare il suo valore; cioè, il valore aumenta con il numero di persone che la utilizzano.
- L'accaparramento e la concentrazione delle informazioni rappresenta quindi una perdita di opportunità di business.
- L'informazione può esser replicata indefinitamente.

Seconda Legge: Il valore dell'informazione cresce con l'utilizzo della stessa

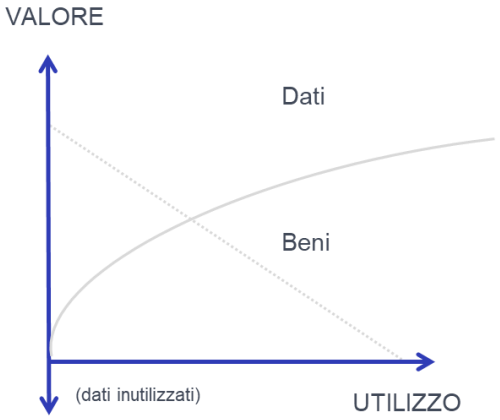


Figura 4 - Il valore dell'informazione cresce con l'utilizzo della stessa

l'Informazione, a differenza degli altri asset, mostra ritorni crescenti con l'utilizzo, vedi Figura 4. Possiamo considerare ad esempio i numerosissimi dataset disponibili online, ad es. sulla piattaforma Kaggle, www.kaggle.com, che collega aziende che pubblicano dataset alla ricerca di analisti che forniscono modelli previsionali o decisionali. Più questi dati vengono utilizzati dagli analisti, più informazioni si possono generare; inoltre, il dataset originario in questo modo acquista valore, e tende ad attirare nuovi utenti.

Dalla legge consegue che:

- L'informazione non ha valore di per sé, ma acquisisce valore quando viene utilizzata.
- L'informazione non utilizzata è di fatto una passività, poiché da essa non viene estratto alcun valore. In molte organizzazioni vi sono grandi quantità di informazioni raccolte e memorizzate con costi elevati, ma che non vengono mai utilizzate: si tratta di fatto di uno spreco.
- Il massimo potenziale di generazione di valore dall'informazione si ha quando ognuno nell'organizzazione sa dove essa si trova, ha accesso ad essa e sa come utilizzarla.

Terza Legge: L'informazione è deperibile

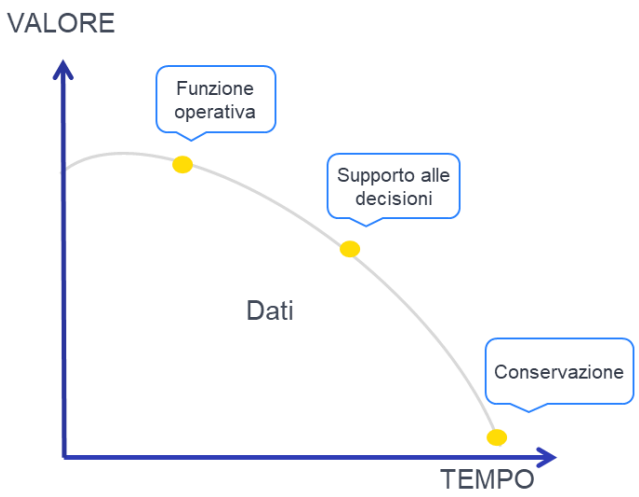


Figura 5 - L'informazione è deperibile

Consideriamo l'informazione relativa ai tempi di arrivo dei mezzi pubblici in una città: ovviamente, dati real-time hanno per i passeggeri un valore molto elevato; quando il mezzo pubblico arriva, il dato avrà solo, insieme a tanti altri, valore statistico.

Come molti altri asset, il valore dell'Informazione tende a deprezzarsi col tempo, a un ritmo variabile con il tipo di informazione, vedi Figura 5, in cui all'inizio l'informazione ha rilevanza per i processi operativi o decisionali, nel tempo ha valore solo a fini statistici e infine, dopo un determinato periodo in cui deve essere disponibile a fini giuridici (ad esempio i biglietti aerei), termina il suo ciclo di vita e non serve più.

Distinguendo due tipi di processi nelle organizzazioni, e cioè processi operativi che forniscono servizi finali agli utenti, e processi decisionali per prendere decisioni tattiche o strategiche, le informazioni utilizzate dai processi operativi hanno un tempo di vita relativamente breve a livello operativo, un tempo molto più lungo per il supporto alle decisioni, e ancora superiore per utilizzi per scopi legali.

Quarta Legge: Il valore dell'Informazione cresce con la sua accuratezza

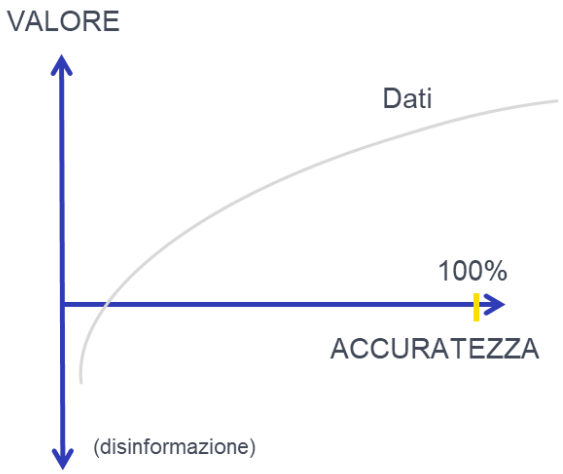


Figura 6 - Il valore dell'Informazione cresce con la sua accuratezza

Abbiamo trattato il tema della qualità dei dati nel Capitolo 5, dove, in particolare, abbiamo discusso la dimensione della *accuratezza*. In generale, quanto più accurata è un'informazione, tanto più essa è utile e quindi assume valore; al contrario, le informazioni inaccurate possono risultare estremamente costose per le organizzazioni sia in termini di errori operativi che di decisioni errate, vedi Figura 6.

Possiamo ad esempio considerare un'azienda che svolge una Sentiment analysis ad esempio su messaggi Twitter. La sentiment analysis è un tipo di analisi sui dati che cerca di comprendere il tipo di percezione che gli utenti hanno di una persona, un prodotto, un evento. E' chiaro che più i tweet sono accurati (soprattutto, riguardo alla accuratezza sintattica), più facilmente l'azienda riuscirà ad estrarre valore (si pensi al confronto tra un tweet scritto correttamente, contro un altro pieno di abbreviazioni, errori lessicali, emoticon ecc…) .

Vi è peraltro un punto di ritorni marginali decrescenti in cui un aumento della accuratezza fornisce scarsi benefici addizionali, particolarmente in considerazione dell'aumento più che proporzionale dei costi relativi, in quanto produrre dati accurati costa. D'altra parte, quando l'accuratezza dell'informazione scende al di sotto di un livello minimo accettabile, essa diventa una passività piuttosto che un asset, e le persone smettono di utilizzare l'informazione inaccurata.

Quinta Legge: Il valore dell'Informazione cresce quando questa è combinata con altre informazioni

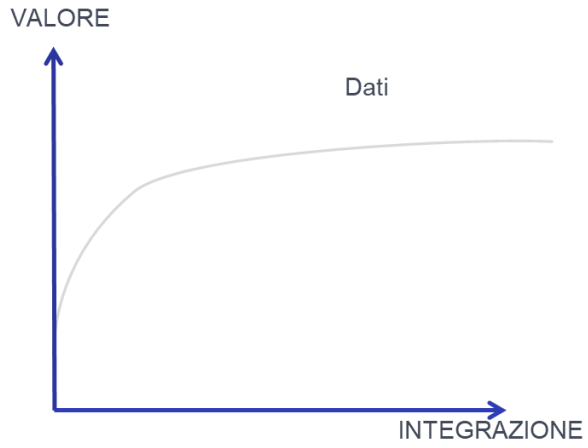


Figura 7 - Il valore dell'Informazione cresce quando questa è combinata con altre informazioni

Abbiamo di fatto introdotto questa legge mostrata in Figura 7 nel Capitolo 7 dedicato alla integrazione di basi di dati. Ad esempio, I dati di vendita dei prodotti e le informazioni sui clienti sono già di per sè informazioni di valore. Tuttavia, essere in grado di correlare questi due dataset rende le informazioni che se ne possono estrarre assai più preziose. La capacità di correlare le caratteristiche dei clienti con i profili di acquisto consente di orientare le iniziative di marketing, in modo da promuovere al momento giusto i prodotti più adatti ai clienti più disponibili alla spesa. Inoltre, secondo la Legge di Pareto (o regola dell'80/20), integrare il 20% dei dati in generale porta all'80% dei benefici; al di là di questa soglia si possono avere ritorni decrescenti, il che può risultare economicamente non più convenente.

Sesta Legge: "di più" non è necessariamente meglio, ovvero il fenomeno dell'information overload

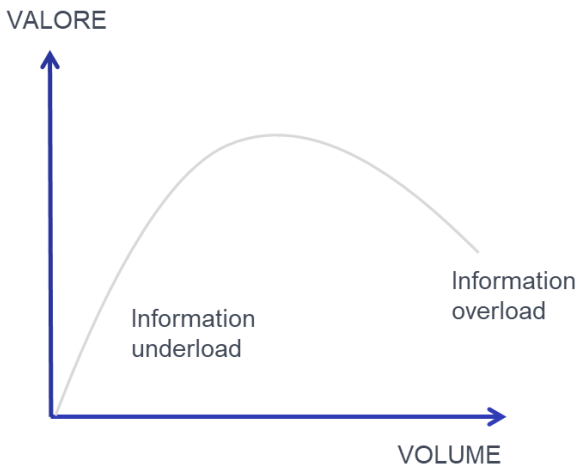


Figura 8 - Di più non è necessariamente meglio

In molte organizzazioni oggi il problema non è la mancanza di informazioni, ma il contrario, l'enorme sovrabbondanza di informazioni disponibili, fenomeno anche chiamato dell'information overload o sovraccarico della informazione, vedi Figura 8. Vi sono esperienze evidenti che confermano che la

capacità degli esseri umani di apprendere, memorizzare e elaborare informazioni è limitata e che, quando l'ammontare di informazioni eccede questo limite, subentra il sovraccarico di informazioni, la capacità di comprensione diminuisce rapidamente, così come l'efficienza nell'assunzione delle decisioni; vedi Figura 9, in cui viene fatto l'esempio della utilità portata dalla assunzione di acqua per il corpo umano.

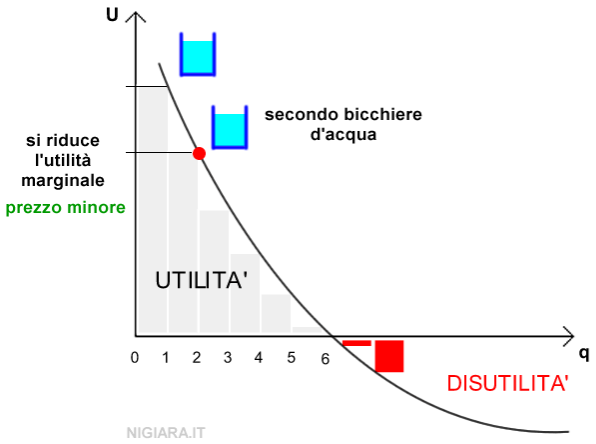


Figura 9 - Utilità e disutilità marginale di un bene

In questo ambito di considerazioni vogliamo citare un altro risultato non intuitivo della ricerca e delle indagini portate avanti in questi anni in tema di valore della informazione e information overload: mentre la legge di Pareto sull'utilità marginale decrescente afferma che all'aumentare del consumo di un bene, l'utilità marginale di quel bene diminuisce, studi empirici mostrano che il valore percepito dell'informazione da parte dei decisori continua a crescere oltre il punto di overload. Questo significa che i decisori umani tendono a ricercare più informazione di quanto possa essere processata in modo ottimale, nello sforzo di evitare errori e di ridurre l'incertezza. Ciò suggerisce che le persone ritengano che "più informazione è comunque meglio", e non siano coscienti dei propri limiti nella capacità di elaborare efficientemente l'informazione stessa.

Settima Legge: L'Informazione non è esauribile

La maggior parte delle risorse sono esauribili – più le usiamo, meno ne abbiamo disponibili. Al contrario, l'informazione ha la capacità di auto-generarsi – più la utilizziamo e più ne abbiamo. Questo perché nuove informazioni sono spesso create come risultato della combinazione di diverse fonti di informazione fra di loro; l'informazione originale rimane utilizzabile e le informazioni derivate si aggiungono alla base di informazioni esistente.

2.3 Modelli alternativi per misurare il valore della informazione

Finora ci siamo occupati di leggi generali relative al valore della informazione, in questa sezione discutiamo modelli che ci permettono di misurare il valore; discuteremo anche dei vantaggi e degli svantaggi che l'applicazione del modello comporta.

1. Modello del costo storico: L'asset informazione viene valutato sulla base del costo pagato in origine per acquistarlo o per generarlo.

- Vantaggi: è il metodo più facile da applicare, e probabilmente il più obiettivo.
- Svantaggi: può non riflettere affatto il valore corrente dell'asset, che potrebbe essere stato influenzato da contingenze o eventi favorevoli o sfavorevoli, ad esempio il fatto che procurarsi la informazione sia diventato più immediato e meno costoso di un tempo.

2. Modello del valore corrente di mercato: La valutazione dell'asset informazione è basata su quanto altre persone o organizzazioni sono disposte a pagarla al prezzo di mercato corrente.

- Vantaggi: Questo metodo fornisce una buona indicazione del valore corrente dell'asset.
- Svantaggi: Il punto debole di questo metodo consiste nel fatto che il suo calcolo può portare ad un impegno di risorse umane o economiche maggiore che non la misurazione del semplice costo storico, perché effettuare indagini costa.

3. Modello della utilità (o del Present Value): L'asset informazione è valutato sulla base del valore attuale dei futuri benefici economici attesi.

- Vantaggi: Concettualmente questo metodo rappresenta la migliore approssimazione al vero valore economico dell'asset informazione, si veda [Godfrey et al, 1997].
- Svantaggi: La principale criticità di questo metodo consiste nella difficoltà di determinare gli specifici flussi di cassa e in generale i benefici economici che un asset (ad es., un dataset, un data base, una mailing list) può generare in futuro.

Le precedenti considerazioni fanno capire che non si è ancora giunti ad un modello univoco per il calcolo del valore dell'asset informazione; ad esse occorre aggiungere altri motivi per cui il valore della informazione non è ancora adeguatamente riconosciuto nei bilanci di esercizio:

1. Come abbiamo in parte visto, al momento non c'è una approccio generalmente accettato a misurare il valore dell'informazione, anche perché la natura dell'informazione digitale come vero e proprio asset non è adeguatamente e generalmente compresa. Il risultato tuttavia è che, finché il valore dell'informazione non comparirà nello stato patrimoniale, esso tenderà fatalmente ad essere sottovalutato rispetto a quello degli altri asset.
2. Vi sono significative barriere culturali rispetto alla misura del valore dell'informazione, per cui molte organizzazioni preferiscono classificare i costi dell'informazione come spesa corrente piuttosto che capitalizzarla nel corso della sua vita utile. Un altro problema è che esistono significative barriere nella pratica misura della informazione, e appare più vantaggioso per le organizzazioni classificarne il valore come spesa nel periodo di bilancio corrente, piuttosto che capitalizzarla nel suo periodo di validità.
3. Solo recentemente, di fronte allo sviluppo di un enorme volume di investimenti intangibili in varie forme di beni e servizi consistenti essenzialmente in informazioni digitali, si è diffuso un interesse su questa tematica che ha portato a riconoscere anche dal punto di vista legale la possibilità di registrare nello stato patrimoniale alcuni degli asset intangibili. Svilupperemo questa importante tematica in una sezione successiva sull' ascesa dell'economia intangibile.

3. Dati, informazione e conoscenza. Max Boisot e lo spazio del valore economico

La riflessione di Boisot spazia tra i concetti di dato, informazione, conoscenza, vista quest'ultima come asset e come artefatto che può essere oggetto di scambio economico, Sezione 3.1. Successivamente, l'attenzione si sposta sulle relazioni tra i concetti di conoscenza, codifica e astrazione e la categoria della complessità, Sezione 3.2. Infine, il focus si sposta sulla economia dello scambio di prodotti fisici, e sulle relazioni tra utilità, valore, codifica, astrazione e scarsità, Sezione 3.3.

3.1. Caratteristiche della conoscenza come asset

Finora ci siamo occupati della informazione come asset strategico. Nel Capitolo 1 abbiamo visto quale relazione concettuale leghi la informazione alla conoscenza; elaborare informazione porta alla scoperta di nuove informazioni che, se accumulate con la conoscenza pregressa, portano a generarne di nuova. L'esempio dei geografi babilonesi, che osservando le eclissi furono in grado di indurre la legge temporale di ripetizione ciclica delle eclissi è un esempio di conoscenza generata dal mettere insieme tante osservazioni e relative informazioni.

Facciamo ora un passo avanti, e passiamo a osservare la *conoscenza* come asset strategico, attraverso l'analisi delle relazioni tra dati, informazioni e conoscenza. Per questo utilizzeremo l'analisi di Max Boisot [Boisot 1998], riassumibile nello schema riportato in Figura 10, che riprende ed estende le relazioni tra dato, informazione e conoscenza introdotte nel Capitolo 1.

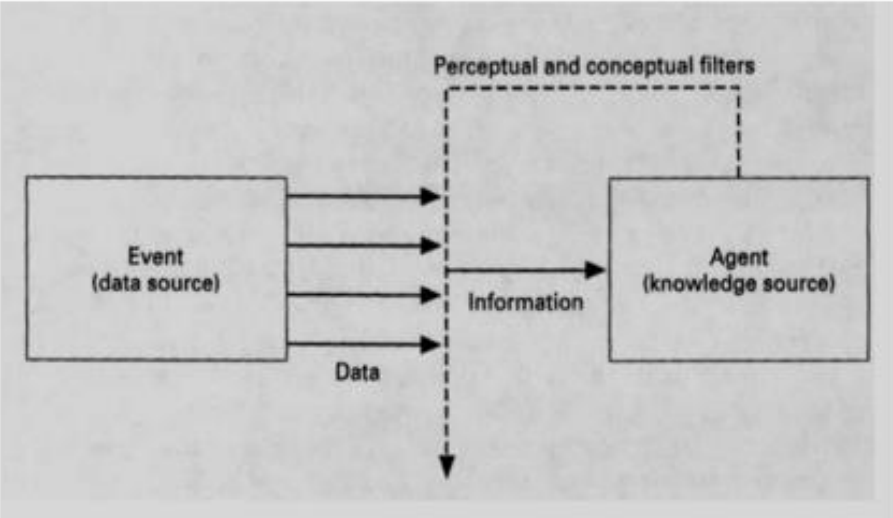


Figura 10 - Relazioni fra Dati, informazioni e conoscenza (tratta da [Boisot 1995])

I *dati* sono una distinzione fra stati fisici delle cose – nero, bianco, pesante, leggero, ecc. – che può o meno portare *informazioni* a un agente, in funzione del precedente *stock, o magazzino, di conoscenza* dell'agente stesso.

Mentre i dati possono essere caratterizzati come una proprietà delle cose, la conoscenza è una proprietà degli *agenti* che li predispone ad agire in certe circostanze; la conoscenza è quindi un asset

costruito sulle informazioni estratte dai dati. L'informazione è quel sottoinsieme di dati residente nelle cose che attiva un agente, ed è filtrata dai dati attraverso l'apparato percettivo o concettuale dell'agente. L'informazione quindi stabilisce *una relazione fra l'agente e le cose*.

La conoscenza può essere concettualizzata come una serie di distribuzioni di probabilità detenuta da un agente e tale da orientare le sue azioni ("devo andare a Merano, dove domani c'è il 60% di probabilità che piova, quindi mi conviene portare l'ombrello"). A differenza delle informazioni, la conoscenza non può esser direttamente osservata: la sua esistenza si può soltanto inferire dalle azioni degli agenti.

Il patrimonio di conoscenze risultante da questo processo non può essere osservato direttamente, ma deve essere inferito indirettamente. E, tuttavia, è il patrimonio di conoscenze accumulato dalla scienza, dalla tecnica e dalla nostra esperienza quotidiana al centro dello straordinario sviluppo economico caratterizzato dal progresso tecnologico e dalla rivoluzione dell'informazione degli ultimi decenni, che si può spiegare soltanto con la continua e sistematica applicazione di conoscenze in continua estensione a una serie sempre crescente di processi fisici. In molte aree infatti *la conoscenza è diventata più importante* del mondo fisico a cui essa si riferisce.

Per questo insieme di ragioni si è finito col vedere la conoscenza come un "asset di per sé", e non semplicemente come un attributo o come una valorizzazione di altri tipi di asset; a maggior ragione in quanto, alla fine del XX secolo, i patrimoni di conoscenze vengono a costituire la base principale delle economie postindustriali.

Occorre peraltro dire che i *patrimoni di conoscenze* non si comportano come gli asset fisici, e implicano un sostanziale cambio di paradigma. Ricordando al lettore che un asset è una risorsa economica che ci aspettiamo fornisca benefici, e da cui possano fluire servizi per un certo periodo di tempo, i patrimoni di conoscenze sono quegli asset costituiti da risorse di conoscenza dalle quali ci si aspettano benefici e servizi per un periodo di tempo che, come abbiamo visto con le considerazioni della Sezione 2, è difficile da specificare a priori; anzi, a differenza degli asset fisici, un patrimonio di conoscenze può in teoria durare per sempre.

Ciò implica che non vi è corrispondenza tra lo sforzo richiesto per creare conoscenza e il valore dei servizi che esso genera. La conoscenza accumulata, a differenza dei puri asset fisici – come ad esempio le materie prime – è non-lineare nel processo di accumulazione rispetto agli effetti che essa produce; e se è incorporata all'interno di asset fisici – come ad es. un sistema manifatturiero - possono diventare una maggiore fonte di discontinuità.

Boisot distingue tra *utilità* e *valore*: mentre infatti il condividere la conoscenza non riduce l'utilità di questa per il possessore originale, che può continuare a ricevere utili servizi anche dopo averla condivisa, ne riduce invece il valore. La condivisione della conoscenza ne riduce infatti la scarsità, categoria economica che ha un ruolo rilevante nel determinare il valore, si pensi ad esempio alla legge della domanda e della offerta. Torneremo su questi concetti nella Sezione 3.3.

Un altro concetto chiave della visione di Boisot è che la conoscenza si diffonde naturalmente e rapidamente in certe circostanze, a differenza di altre, in cui si comporta maggiormente come gli oggetti fisici: per salvaguardare il valore del patrimonio di conoscenze, si tende in questi casi a porvi intorno delle barriere protettive, ad es. brevetti, proprietà intellettuale.

Comprendere quando la conoscenza fluisca liberamente o meno è essenziale per orientare la conoscenza come maggior fonte di ricchezza. In quali condizioni dunque la conoscenza fluisce liberamente e in quali tende a diventare "vischiosa"? In breve:

- la *conoscenza fluida* è quella ben codificata ed astratta, liberata da tutti i dati inessenziali; si noti la forte consonanza nel concetto di astrazione con quanto abbiamo discusso nel Capitolo 12.
- la *conoscenza vischiosa,* al contrario, è ricca di dati, qualitativa, "ambigua", e tende, appunto, a fluire molto lentamente.

Ad esempio, si pensi alla differenza tra il trasmettere un codice identificativo e descrivere un quadro di Rembrandt nel corso di una telefonata. Per poter utilizzare la conoscenza, ad un certo punto essa deve perdere la sua vischiosità, perché tale vischiosità rende problematico l'estrarne e realizzarne il valore.

Ci imbattiamo qui in un paradosso: più utile diventa un patrimonio di conoscenza, più diventa difficile controllarlo. Per sfruttare con successo un patrimonio di conoscenze sono necessarie strategie efficaci per affrontare e risolvere questo paradosso.

Il livello in cui la conoscenza è strutturata e condivisa definisce inoltre una *cultura*. Ad esempio, una cultura burocratica è caratterizzata per Boisot da un patrimonio di conoscenze ben codificato, astratto e privo di ambiguità (in realtà sappiamo che spesso le norme e le leggi possono essere interpretate, e in tale interpretazione si annida spesso il potere del burocrate); la cultura burocratica non è portata a condividere la conoscenza, e tende invece a bloccarne la diffusione. Una cultura di mercato preferisce anch'essa forme di conoscenza ben codificate ed astratte, ma è molto più orientata alla condivisione.

3.2 Codifica, astrazione e riduzione di complessità

In questa sezione siamo interessati a discutere la relazione che sussiste tra i concetti di *codifica*, della *astrazione* e la categoria della *complessità*, investigate da Boisot nel modello da lui proposto.

Partiamo da questa considerazione: ogni attività in una organizzazione, dal livello operativo a quello del controllo ai ruoli strategici, implica un aumento della complessità, intesa qui come:

- il numero di elementi che interagiscono fra di loro nella attività e
- il numero dei differenti stati a cui queste interazioni danno origine.

In che modo l'alta direzione può gestire la complessità che sorge dall'interno e dall'esterno dell'Impresa? La risposta più semplice è: sviluppando modelli astratti che aiutino a mettere sotto controllo la complessità riducendola a proporzioni gestibili. Vediamo un po' più da vicino i legami tra astrazione e codifica con il concetto di complessità.

Quanto ad astrazione e complessità, possiamo considerare l'astrazione come un atto di semplificazione cognitiva che ci libera dalla necessità di aver a che fare con la complessità che ci circonda.

Intuitivamente possiamo vedere come astrazione e complessità siano antitetiche fra di loro, in quanto le astrazioni ci permettono di organizzare il patrimonio di conoscenza per produrre un flusso di servizi utili nel tempo, economizzando l'utilizzo di risorse fisiche nel tempo, e così riducendone il livello di entropia (disordine) e complessità.

Riguardo alla codifica e alla complessità, gli artefatti che sono il risultato della produzione di massa incorporano una conoscenza che, per essere trasmessa e riutilizzata, deve essere sistematicamente formalizzata e codificata. La conoscenza trasmessa in modo discorsivo, al contrario, permette abitualmente un maggior grado di informalità; anche questa conoscenza discorsiva tuttavia, per essere trasmissibile, deve essere codificabile in qualche misura.

In conclusione, astrazione e codifica rappresentano due concetti fondamentali per l'analisi e la trasmissione di un patrimonio delle conoscenze; inoltre, esse rappresentano due modalità diverse, anche se interrelate, per ridurre il costo necessario per convertire conoscenza potenzialmente utilizzabile in patrimonio di conoscenza, da riutilizzare nel tempo. I principi astratti, in questo ambito, hanno una copertura più ampia e trovano applicazione più generale che quelli concreti, poiché abilitano un'efficace realizzazione concreta dei principi tecnici coinvolti.

3.3 Lo spazio del valore economico (I-Space)

Approfondiamo infine il concetto che sta alla base del modello proposto da Boisot, l'*utilità* dei beni economici scambiati nel mercato. Secondo la teoria Neoclassica del valore, l'utilità è un fatto personale e misura ciò che un agente economico individuale (una persona o un'impresa) ottiene dal consumo di una data quantità di bene economico. Ad esempio, possiamo pagare lo stesso prezzo per andare a vedere un film, ma ciascuno ne ottiene qualcosa di diverso. Secondo l'approccio di Boisot, l'utilità che può essere estratta da un bene economico è funzione di tre coordinate, la scarsità, la codifica e la astrazione. Scarsità, codifica e astrazione tutte insieme formano secondo Boisot le tre coordinate dello spazio del valore economico (i-space, vedi Figura 11) dei dati digitali, vedi [Boisot 1995] e [Boisot 1998].

Nell'approccio di Boisot, l'utilità economica che è possibile tratte da un prodotto venduto in un mercato è anzitutto funzione della *scarsità*, in quanto più un bene è scarso, più nella relazione domanda offerta tende ad aumentare il valore. Questa prima coordinata è peraltro ben nota in economia.

Una seconda coordinata riguarda la *codifica*. Pensiamo al modello relazionale descritto nel Capitolo 3: la possibilità di assegnare codici univoci ai prodotti, di descriverne le caratteristiche attraverso valori di domini numerici o alfabetici è fondamentale per far interoperare mercati diversi. La codifica riduce la complessità e aumenta il valore in quanto identificare un prodotto con un codice abilita le transazioni economiche perché crea una tassonomia comune.

E infine la *astrazione*; quanto più un prodotto può essere descritto in forma astratta, tanto più esso potrà flessibilmente essere usato in diverse applicazioni. Pensiamo alle leggi di Maxwell, la loro generalità permette di applicarle in una grandissima varietà di fenomeni fisici.

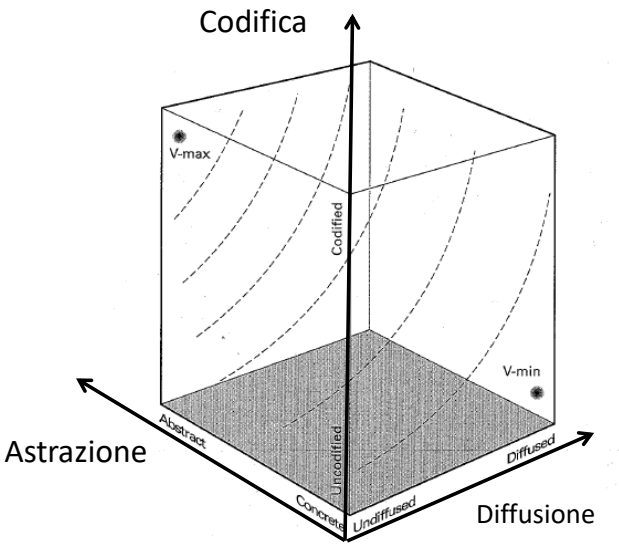


Figura 11 – Le tre dimensioni del dato come bene economico per Boisot (tratta da [Boisot 1998])

Alcune conclusioni finali sull'approccio di Boisot:

- esiste una fondamentale differenza tra il comportamento dei beni fisici e della conoscenza.
- Le aziende il cui business è fondato sul proprio patrimonio di conoscenze hanno bisogno di una prospettiva teorica diversa da quelle che lavorano prioritariamente sulla base dei beni fisici.

Lo schema teorico dell'I-Space consente tra l'altro di:

- Analizzare le condizioni per la sostituzione reciproca di beni fisici e patrimoni di conoscenza.
- Analizzare i flussi della conoscenza fra e all'interno dei gruppi sociali (vedi Figura 12).
- Sviluppare due diverse strategie di gestione dei patrimoni di conoscenza ("appropriazione" rispetto a "condivisione") per far fronte al paradosso per cui**:** "Più utili diventano, più difficile diventa controllarli".
- Studiare l'integrazione dei patrimoni di conoscenza a livello di prodotti, tecnologie e organizzazione
- Comprendere l'impatto dell'Information technology sui flussi della conoscenza e sulla gestione dei patrimoni di conoscenza.

La nuova prospettiva è dunque destinata a influenzare profondamente le modalità con cui le imprese basate sulla conoscenza concepiscono le loro strategie, le tecnologie che adottano, i loro processi di gestione delle risorse umane e il loro ambiente.

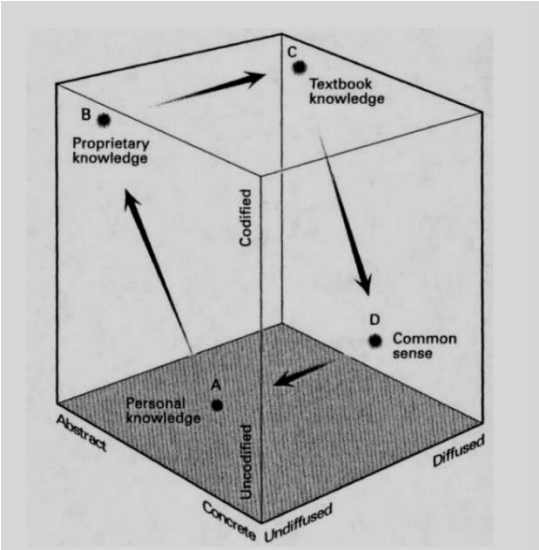


FIG. 3.2 The Movement of Knowledge in the I-Space

Figura 12 - Il movimento della conoscenza nell' I-Space (tratta da [Boisot 1998])

4. Le nuove regole dell'informazione nell'era del digitale secondo Shapiro e Varian

Affrontiamo a questo punto il tema seguente: nel mondo dominato dallo sviluppo impetuoso delle tecnologie digitali, quali nuove regole e principi economici governano l'informazione? Come evolvono i meccanismi di creazione del valore? E le leggi di base dell'economia sono ancora valide nel nuovo contesto? Per rispondere a queste domande ci rifaremo all'elaborazione ormai classica discussa in [Shapiro e Varian 1999].

Il primo aspetto da osservare è che in un mondo che negli ultimi 30 anni ha sperimentato un rapidissimo processo di innovazione tecnologica e in cui i vecchi modelli di business hanno subito una profonda trasformazione, le leggi di base dell'economia si sono comunque confermate valide: tanto che coloro che hanno continuato a tenerne conto sono sopravvissuti - e molti di loro hanno prosperato – a differenza di chi le ha ignorate ed è fallito.

Secondo Shapiro e Varian, le tecnologie cambiano. le leggi economiche, no. E alcuni concetti economici di base consentono di spiegare efficacemente anche l'evoluzione delle imprese di oggi, sotto la spinta del digitale.

Essenzialmente, ogni cosa o fenomeno che possa esser digitalizzato è informazione. Il valore dell'informazione, come in parte abbiamo già visto, varia a seconda dei diversi consumatori: alcune informazioni hanno valore per il tempo libero, altre per il business; a prescindere dalla fonte specifica, le persone sono pronte a pagare per l'informazione, tanto che molte strategie dei fornitori sono basate sul diverso valore che le varie categorie di consumatori attribuiscono al bene informazione.

Due diverse strutture caratterizzano il mercato dell'informazione.

- Il modello dell'azienda dominante (es. Microsoft) che raggiunge la leadership rispetto ai costi attraverso economie di scala, riuso e rivendita (Reuters).
- Il modello di mercato basato sulla differenziazione dei prodotti; molte aziende producono lo stesso tipo di informazione, ma in molte diverse varietà, ad es. publishing, film, televisione, alcuni mercati del software; queste aziende differenziano il prodotto aggiungendo valore all'informazione base.

Il processo di creazione e assemblaggio della informazione ha un costo, come è evidente dalla complessità e ricchezza del suo ciclo di vita cui abbiamo dedicato alcuni capitoli. Partiamo dall'analisi della struttura di questi costi, poiché essa è assai particolare e determina la natura della competizione sui mercati dell'informazione.

Il costo di produzione della informazione

L'informazione è costosa da produrre, ma il costo di riproduzione è basso: in altre parole si può dire che la produzione di un bene basato sulla informazione digitale (si pensi ad esempio a un e-book) richiede costi fissi elevati, ma bassi costi marginali. Il costo per produrre una copia addizionale in generale non aumenta, anzi tende a diminuire ulteriormente se viene prodotto un elevatissimo numero di copie. Inoltre, la maggior parte dei costi fissi sono "sunk costs", ovvero costi che non possono essere recuperati se la produzione viene interrotta. In queste condizioni decidere il prezzo della informazione basandosi sul costo di produzione non funziona, e ai prodotti basati sulla infor-mazione deve essere attribuito un prezzo basato sul valore per i consumatori.

Tuttavia, i consumatori valutano l'informazione in modi diversi. Si possono quindi distinguere strategie che vanno dal pricing basato sul valore (massimizzando il valore della proprietà intellettuale) al prezzo differenziale.

Possiamo distinguere tre forme principali di *prezzo differenziale*:
1. *Prezzo personalizzato* (basato sulla profilazione): consiste nel formulare per ogni cliente un prezzo diverso, basato su un design di prodotti ottimizzati secondo le caratteristiche di ogni cliente.
2. *Versioning*, consistente nell'offrire una linea di prodotti in diverse versioni e permettere ai clienti di scegliere la versione del prodotto più adatta a loro.
3. *Pricing di gruppo*, consistente nel fissare prezzi diversi per gli stessi prodotti per diversi gruppi di clienti, come ad esempio sconti per studenti.

L'informazione come bene esperienziale

Secondo gli economisti, un bene è definito come *esperienziale* se i clienti devono provarlo per essere in grado di percepirne il valore. Virtualmente ogni nuovo prodotto è un bene esperienziale, e gli operatori di marketing hanno sviluppato strategie come ad es. campioni gratuiti, prezzi promo-zionali e testimonials per introdurre i clienti ai nuovi prodotti. Ma l'informazione è un bene esperienziale *ogni volta* che viene consumata; come faccio a sapere se il giornale di oggi vale effettivamente € 1,50 finché non l'ho letto? Si sono quindi sviluppate forme specifiche di promozione, come navigazione limitata nel tempo, previews, freemium, ecc.

Inoltre, sommersi come siamo da un enorme sovraccarico di informazioni, possiamo ben apprezzare l'affermazione del Premio Nobel Herbert Simon *"Una ricchezza di informazioni crea una povertà di attenzione".* Oggi, ai fini della vendita di informazione e quindi del suo prezzo, il problema non è l'accesso all'informazione, ma l'overload di informazioni; il valore reale fornito da un information provider viene dalla sua capacità di individuare, filtrare e comunicare ciò che è utile al consumatore, identificato attraverso una efficace profilazione. Di qui il ruolo centrale che i motori di ricerca hanno assunto nell'indirizzare il problema dell'economia dell'attenzione.

Lock in e switching costs

Quando un individuo o un'impresa fanno significativi investimenti durevoli in beni tecnologici complementari specifici a un brand, un sistema o un ambiente, spesso investono in beni con diversi tempi di vita utile. In questo modo non è facile sincronizzare i tempi per muovere a nuovi sistemi, tecnologicamente più avanzati ma incompatibili coi precedenti. Come risultato essi vengono a trovarsi in una situazione in cui, per migrare ad altri sistemi, devono affrontare i cosiddetti "switching costs"; quando i costi per migrare da una tecnologia ad un'altra sono sostanziali, gli utenti si vengono a trovare in una situazione cosiddetta di lock-in.

Switching costs e lock-in non sono limitati all'aspetto puramente tecnologico. Ad esempio la scelta in fase di produzione di certi ambienti tecnologici può implicare sostanziali costi di formazione del personale o profonde trasformazioni organizzative, con impatti importanti sui modelli di business. Le strategie per garantirsi situazioni di lock-in a proprio vantaggio, o di neutralizzare quelle di altri, possono essere fondamentali per assicurare ai fornitori di tecnologie informatiche un vantaggio competitivo.

I beni basati sull'informazione presentano effetti di rete o economie di scala demand-side; un bene presenta effetti di rete se la domanda del bene stesso dipende da quante altre persone lo acquistano. I feedback positivi fanno sì che le reti diventino sempre più grandi; quando la base installata di utenti cresce, sempre più utenti sono portati ad aderirvi.

Gli effetti di rete possono distinguersi in:

- *effetti di rete diretti*: un aumento dell'utilizzo porta a un aumento diretto del valore della rete per altri utenti
- *effetti di rete indiretti*: aumenti dell'utilizzo di un prodotto o della copertura di una rete accrescono il valore di prodotti o di reti complementari, il che produce a sua volta un incremento di valore nel prodotto o nella rete originali.

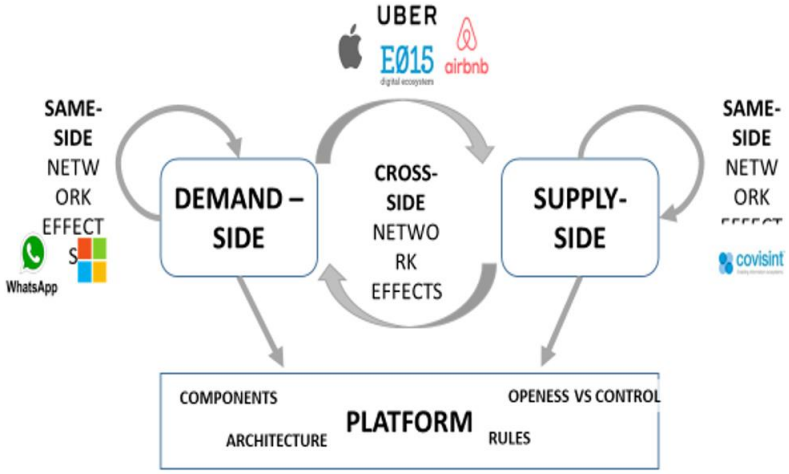


Figura 13 - Effetti di rete (tratta da [Shapiro e Varian 1999])

Infine, possiamo distinguere effetti di rete (vedi Figura 13):

- lato domanda: il valore del network generato dal prodotto aumenta grazie all'aumento degli utenti che vi aderiscono (è il caso di Facebook e What's App)
- lato offerta: il valore del network aumenta col numero dei fornitori che vi aderiscono (es. Tripadvisor)
- effetto combinato: Il valore del network aumenta grazie alla combinazione sinergica dell'aumento degli utenti e dei fornitori, che si rafforzano reciprocamente (esempi sono Uber, Airbnb).

5. **Jeremy Rifkin, l'economia dell'accesso e la società a costo marginale zero.**

Desideriamo ricordare qui due principali contributi di Jeremy Rifkin alla comprensione della evoluzione dell'economia digitale: il concetto di "Economia dell'accesso" e quello di "Società a costo marginale zero".

Economia dell'accesso

Il libro [Rifkin, 2000] è del 2000, e rappresenta l'opera più lucidamente profetica di Jeremy Rifkin. Nella nuova era del Digitale, stiamo muovendo rapidamente dai mercati alle reti e dalla proprietà all'accesso; per imprese e consumatori comincia ad assumere minore importanza lo scambio di proprietà fra venditori e acquirenti sul mercato, mentre si diffonde sempre più l'accesso a breve termine fra fornitori e clienti che operano in una relazione di rete; i mercati rimangono, ma giocano un ruolo sempre più ridotto negli affari umani.

Mentre il mercato metteva al centro nel passato venditori e acquirenti, ora i protagonisti sono *fornitori e utenti, in una relazione di rete*. E la forza dominate nella nuova era è ora il *capitale intellettuale*; il principale valore nella nuova economia sono concetti, idee, immagini, e non oggetti: alcuni business scelgono anzi di offrire gratuitamente i loro prodotti nella speranza di entrare in una relazione di lungo termine con i loro clienti. E si comincia ad assistere a un passaggio di lungo termine dal predominio dalla produzione industriale a quella culturale.

Rifkin osserva come anche i consumatori tendono a spostarsi dalla proprietà *all'accesso*: mentre i beni meno costosi e di uso quotidiano continueranno ad essere comprati sul mercato, articoli più costosi come le case e le automobili saranno di proprietà dei fornitori e i consumatori vi avranno accesso tramite leasing di breve termine, affitti, e altri tipi di accordi (è da notare come Rifkin nel 2000 prefiguri già la Sharing Economy, l'economia della condivisione: Airbnb nasce nel 2008, Uber nel 2009.

Si vedono anche i primi segni del passaggio ad *un'economia esperienziale* (vedi quanto detto nella sezione su Shapiro e Varian), un mondo in cui la vita personale diventa un mercato commerciale: il valore di vita del cliente è la misura teorica del valore del consumatore se ogni momento della sua vita venisse mercificato in qualche forma nella sfera commerciale. Nella nuova era, le persone "acquistano" la loro esistenza accedendo a beni e servizi in piccoli segmenti commerciali.

L'abilitatore fondamentale del business in questa nuova era è la *connettività*, che consente di gestire in modo agile le relazioni tra fornitori e utenti in quella che prenderà forma successivamente come "economia delle piattaforme", vedi più avanti la Sezione 7. Le reti, a loro volta, sono molto adatte alla natura volatile della economia globale, favorendo la collaborazione, il lavoro in team e risposte rapide alle mutevoli condizioni del mondo esterno.

Desideriamo infine su questo punto citare due ultime grandi intuizioni di Rifkin, che verranno poi sviluppate sia nella sua opera successiva che da altri studiosi:

- L'economia della smaterializzazione - Rifkin rileva come, in conseguenza dello sviluppo della economia dell'accesso, lo spazio dell'economia fisica si stia contraendo. Mentre l'era industriale era caratterizzata dall'accumulazione di capitale fisico e di proprietà, la nuova era premia forme intangibili di potere basate sull'informazione e sul capitale intellettuale. E i prodotti fisici, che erano stati a lungo la misura della ricchezza nell'era industriale, stanno dematerializzandosi.
- Gli asset intangibili - La nuova era dell'accesso favorisce le imprese "leggere", il cui valore è basato sulle idee piuttosto che sugli asset fisici. Comincia quindi a manifestarsi lo spostamento dal predominio dagli asset tangibili a quelli intangibili, che si esprime nella divaricazione tra valore di mercato e valore contabile. Oggi molte delle imprese che presentano le migliori prestazioni hanno profitti estremamente elevati, come conseguenza del rilievo crescente assunto dagli asset intangibili nella valutazione del valore di mercato, benchè spesso questi non vengano adeguatamente registrati nella parte patrimoniale del bilancio di esercizio delle imprese.

Svilupperemo nella prossima parte di questo lavoro gli effetti di questo fenomeno sull'economia mondiale nei nostri giorni.

<u>La società a costo marginale zero</u>

Già Shapiro e Varian avevano osservato come la produzione di un artefatto basato sull'informazione richiede costi fissi elevati, ma bassi costi marginali. Nel 2014 Rifkin porta alle estreme conseguenze questa affermazione rilevando come, nel mondo digitale, i costi marginali di riproduzione tendano a zero. Egli rileva come molti autori di libri tendano già a disintermediare gli editori, rendendo disponibili i propri libri su Internet a un costo molto basso, o addirittura gratuitamente. Questo fenomeno è

l'avanguardia di una tendenza molto più ampia, per cui più di un terzo dell'umanità genera le proprie informazioni tramite cellulari o computer relativamente economici e le mette in condivisione tramite video, audio o testi in un mondo interconnesso e collaborativo a costo marginale "quasi zero".

Dunque, una parte sempre maggiore dei beni e servizi che costituiscono la vita economica della società muove verso il quasi azzeramento dei costi marginali e diventa quasi gratuita; da questa osservazione, Rifkin trae la previsione che il mercato capitalistico sia destinato a ritrarsi in nicchie sempre più ristrette, dove le imprese a scopo di lucro sopravvivranno ai margini dell'economia, contando su una base di consumatori sempre più limitata e rivolta a prodotti e servizi altamente specializzati. E la rivoluzione dell'Internet delle cose contribuirebbe a sua volta a spingere la produttività al punto in cui il costo marginale di molti prodotti sarebbe sempre più vicino a zero.

Secondo Rifkin, una società caratterizzata da costi marginali prossimi allo zero rappresenterebbe il contesto a massima efficienza in cui promuovere il benessere generale e, nel contempo, costituirebbe il definitivo trionfo del capitalismo. Tale trionfo però segnerebbe anche "l'inevitabile uscita del capitalismo dalla scena mondiale".

La previsione di Rifkin va tuttavia collocata in una visione generale più problematica: se in un mercato perfetto i consumatori pagassero veramente solo il costo marginale e questo tendesse a zero, le imprese non sarebbero in grado di garantirsi un ritorno sui propri investimenti, per cui i leader di mercato cercheranno di acquisire posizioni di monopolio per imporre prezzi superiori al costo marginale. Lo sviluppo dell'economia delle piattaforme e l'affermarsi di modelli di business diversificati e altamente profittevoli come quelli di Google e di Amazon, si incaricheranno presto di dimostrare la creatività con cui i nuovi protagonisti dell'economia digitale sapranno affrontare questa sfida, come vedremo tra poco.

5. Mercati "data rich" vs "capital rich"

L'economia digitale cambia i mercati e il concetto stesso di denaro, come argomentato in [Mayer-Shonberger 2018]. La sempre maggiore ricchezza dei dati disponibili sui prodotti porta a superare il prezzo del prodotto visto come astrazione o sintesi di un insieme di caratteristiche, spesso usato da solo o con poche altre informazioni, per limiti cognitivi o comunicativi, al fine di caratterizzare il bene o servizio da acquistare (ad es. quella orata è pescata e costa 25 euro al chilo, quell'altra è allevata e costa 10 euro al chilo, la prima è quasi sicuramente migliore della seconda).

Nei mercati chiamati "data-rich" (in contrapposizione a "capital-rich") in [Mayer-Shonberger 2018], la tecnologia dei dati offre nuove opportunità per i paesi in via di sviluppo. Ad esempio, per centinaia di anni, i pescatori del Kerala, in assenza di informazioni sul mercato, sono stati obbligati a fare stime rischiose nel momento in cui dovevano decidere dove far approdare le loro barche per vendere il pesce pescato. Quando si diffusero nel 1997 i telefoni mobili e le reti di comunicazione, le transazioni poterono iniziare nel momento stesso in cui il pesce era pescato; i telefoni mobili diventarono un mezzo di comunicazione che resero il mercato più efficiente (occorre dire che questa posizione è criticata in [Steyn 2016], i cui vengono opposte diverse osservazioni metodologiche e logiche).

Un altro aspetto affrontato in [Mayer-Shonberger 2018] riguarda la previsione per cui i dati sostituiranno sempre di più il denaro nelle transazioni economiche. Questo in certo senso già accade nelle nostre relazioni con Facebook, Amazon e gli altri operatori, che leggono i nostri comportamenti trasformandoli in profili economici da utilizzare per favorire nuove transazioni. In un crescente numero di casi, le aziende forniscono dati ad altre aziende che li analizzano, e pagano questi servizi con dati, autorizzando le aziende di analisi a riutilizzarli per altri scopi.

I dati hanno un effetto rilevante anche sulla concentrazione dei mercati. Accanto agli effetti di scala e di rete, già discussi nella Sezione 4, stanno assumendo grande rilevanza gli effetti legati ai *dati di feedback*, cioè dati che fotografano l'esito di transazioni, che si manifestano ogniqualvolta gli algoritmi usano dati di feedback per apprendere. I servizi basati su apprendimento, alimentati dai dati di feedback, "comprano" innovazione diminuendo i costi man mano che si allarga la platea degli utenti.

Ciò ha grande rilevanza per la competizione, perché le start up non possono competere con le grandi compagnie, in quanto mancano ad esse i dati di feedback che possono guidare lo sviluppo dei prodotti. La trasparenza degli algoritmi è utile, ma non basta. I regolatori dovrebbero esigere la condivisione dei dati, così che il valore derivante dai dati si diffonda.

Un'altra proposta in [Mayer-Shonberger 2018] riguarda la necessità che gli algoritmi di apprendimento si diffondano il più possibile, così che la progettazione e lo sviluppo permettano un cross-checking e un continuo miglioramento.

6. L'ascesa dell'Economia intangibile

In questa parte raccoglieremo molti degli spunti emersi nei paragrafi precedenti per delineare le caratteristiche, le implicazioni e le prospettive dell'economia degli Intangibili, che svolge ormai un ruolo determinante nel mondo attuale. Per questo ci riferiremo a [Haskel 2018], testo che analizza a fondo questo tema.

Investimenti tangibili e Intangibili

Gli investimenti svolgono un ruolo centrale nella realtà economica. Ricordiamo innanzitutto che il Prodotto interno lordo di una nazione o entità amministrativa (ad esempio, una regione) è dato da

PIL = Consumi +Investimenti + spesa pubblica + esportazioni al netto delle importazioni

Fra queste variabili gli investimenti sono spesso il driver di espansioni e recessioni economiche, poiché essi tendono a crescere e a contrarsi in risposta alla politica monetaria e alle condizioni di business. Fino ai tempi recenti gli investimenti misurati dagli Enti nazionali di statistica erano solo in beni (assets) tangibili (edifici, impianti, macchinari, ecc.); per secoli infatti i sistemi contabili hanno misurato soltanto i beni fisici, come abbiamo già visto nella Sezione 2.2. Ma l'economia non è basata solo su investimenti in beni tangibili. Ad esempio il valore di un aeroporto non è basato solo su quello delle piste, dei terminali e degli impianti, ma anche su know-how, sul software, gli accordi con linee aeree, cioè su idee, conoscenze e relazioni sociali: in una parola, su asset intangibili.

La tendenza degli investimenti in intangibili ad assumere un ruolo sempre maggiore nell'ambito di un economia sempre più basata sulla conoscenza si manifesta nel divario crescente tra valore di mercato e valore contabile nelle società IT. Ad esempio nel 2006 il valore di mercato di Microsoft era pari a $250 miliardi, ma gli asset registrati nello stato patrimoniale era di soli $70 miliardi, di cui $60 in cash e asset finanziari, e solo $3 in impianti e macchinari. La differenza tra valore di mercato e valore contabile era quindi dovuta a beni intangibili quali brevetti, licenze, know how, software, accordi, competenze specifiche, ecc.

Come abbiamo visto nella prima parte di questo capitolo, vedi ancora la Sezione 2.2, molte convenzioni contabili ignorano tuttavia gli investimenti intangibili, ritenendoli difficilmente misurabili; inoltre, un'economia basata su un'elevata intensità di intangibili tende a differenziarsi a causa delle proprietà generali di questi investimenti, come vedremo fra poco. Cominciamo quindi a precisare in modo più approfondito cosa intendiamo per Investimenti tangibili e intangibili, osservando la Figura 14 (R&D è la ricerca e sviluppo).

Tangible investments	Intangible investments
Buildings	Software
ICT equipment (e.g., computer hardware, communications equipment)	Databases
Noncomputer machinery and equipment	R&D
Vehicles	Mineral exploration
	Creating entertainment, literary or artistic originals
	Design
	Training
	Market research and branding
	Business process reengineering

Figura 14 - Investimenti tangibili e Intangibili (tratta da [Haskel 2018])

Alcuni esempi di imprese intangible-Intensive sono:

- Il design e la catena di fornitura di Apple
- Le reti di autisti e di clienti di Uber e Airbnb
- Il know-how di Tesla nel manufacturing.

Gli investimenti intangibili si sono sviluppati rapidamente, e a metà degli anni '90 hanno superato quelli tangibili negli USA, vedi Figura 15. Lo stesso è accaduto all'inizio degli anni 2000 in UK, e poi gradualmente in molti altri Paesi europei, vedi Figura 16.

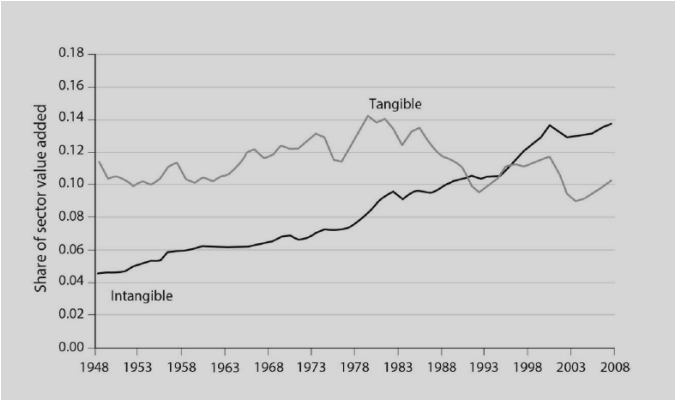


Figura 15 – Investimenti in tangibili e intangibili, US (tratta da [Haskel 2018])

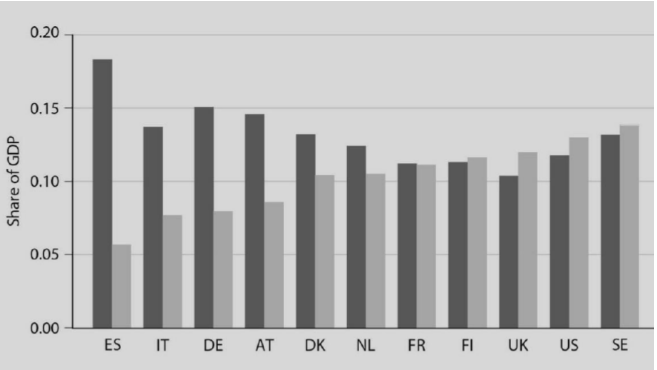


Figura 16 - Investimenti in tangibili e intangibili come percentuale del PIL in Europa (1999-2013) (tratta da [Haskel 2018])

Molto si è discusso sulle ragioni di questa crescita impetuosa degli Investimenti Intangibili. Queste sono alcune delle motivazioni che appaiono più convincenti:

- Una prima motivazione è data dal fatto che i servizi con elevata intensità di capitale umano diventano più costosi rispetto ai prodotti industriali, e ciò mentre gli investimenti intangibili richiedono un'elevata intensità di lavoro.
- L'Information technology e le reti sociali hanno aumentato il ritorno sugli investimenti sociali
- Una maggiore deregulation favorisce gli investimenti intangibili
- Si rileva una certa correlazione tra gli investimenti in intangibili e la spesa in R&D dei Governi
- Un fattore importante è la dimensione dei mercati: mercati più piccoli sarebbero meno attrattivi per investimenti intangibili per brand (come Starbuck o Facebook) che possono essere scalate indefinitamente. Per questo, ad esempio, la Brexit potrebbe ridurre l'incentivo a fare investimenti intangibili in UK.

<u>Caratteristiche degli Investimenti Intangibili: le 4S</u>

Approfondiamo l'analisi sugli investimenti intangibili; [Federico 2018] riassume così le quattro caratteristiche che differenziano profondamente gli investimenti intangibili da quelli tangibili:

“In cosa si distinguono le attività intangibili da quelle tangibili? In quattro proprietà (le quattro S: scalability, sunkenness, synergies e spillovers).

- La scalability indica le fortissime economie di scala: a differenza di un macchinario, che può produrre soltanto una determinata quantità in un determinato luogo, un brevetto può essere utilizzato contemporaneamente in molteplici unità produttive; le economie di scala sono spesso amplificate dall’effetto rete che caratterizzano alcune attività intangibili.
- La sunkenness è legata al fatto che alcune attività intangibili sono molto specifiche all’impresa che le ha prodotte e, a differenza di un macchinario o di un edificio, hanno pertanto un valore di rivendita più basso.
- Le synergies sottolineano i grandi vantaggi che spesso possono essere utilizzati con la combinazione di diverse attività intangibili.
- Gli “spillovers” riflettono la facilità con cui le attività intangibili possono essere utilizzate da soggetti diversi dal proprietario (è più semplice copiare un software rispetto a un macchinario complesso).

Queste proprietà delle attività intangibili hanno effetti pervasivi sulle imprese, sull’economia e più in generale sulla società attuale. Un punto di forza consiste nel cogliere tali effetti nella loro ampiezza, utilizzando prospettive di volta in volta differenti (dalla macroeconomia all’economia industriale, dall’economia regionale alle teorie sull’organizzazione delle imprese).

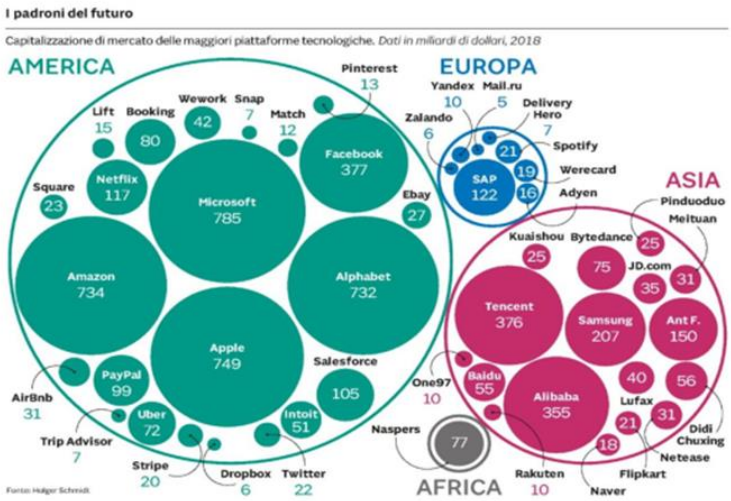
Una prima trasformazione indotta è quella della struttura di mercato, che mostra in diversi settori una tendenza verso la concentrazione, al fine di beneficiare di economie di scala e sinergie. Aumenta anche l’incertezza e la rischiosità degli investimenti, sia perché le economie di scala danno vita a effetti “winner-take-all” sia per il rischio di spillover a favore dei concorrenti. Tali sviluppi hanno conseguenze macro, contribuendo a spiegare la debolezza degli investimenti nell’ultimo decennio; Si modifica l’organizzazione dell’impresa, favorendo da un lato strutture aziendali più flessibili e creative ma dall’altro anche, grazie alle nuove tecnologie, un monitoraggio più intenso dei lavoratori e un’azione di coordinamento più efficace dai vertici manageriali. Cambia anche la distribuzione geografica delle attività economiche, con il ritorno delle agglomerazioni urbane come centri in cui si conduce innovazione e più ampi sono i possibili spillover. La crescita degli investimenti intangibili comporta infine una sfida per il sistema finanziario, in quanto la loro maggiore opacità e incertezza li rende più difficili da finanziarie con il solo credito bancario, mentre il modello del venture capital non sembra applicabile su larga scala.

7. L’ Economia Digitale e la rivoluzione delle Piattaforme

Gli ultimi anni hanno visto lo sviluppo impetuoso delle piattaforme digitali, che hanno assunto un ruolo chiave nell’economia globale nell’era del digitale [Federico 2018].

Si tratta di un nuovo business model che utilizza la tecnologia digitale per connettere persone, organizzazioni e risorse, in un ecosistema interattivo nel quale vengono prodotte e scambiate immense quantità di valore. Amazon, Airbnb, Uber, Alibaba e Facebook sono solo alcuni esempi: ognuna di esse è focalizzata su una industria e un mercato distintivi, e ognuna ha sfruttato il potere della piattaforma

per trasformare profondamente settori dell'economia globale, tanto da permettere ad imprese nate per la maggior parte negli ultimi dieci anni di raggiungere i primi posti nella lista delle aziende mondiali, vedi Figura 17.



Fonte: Il Sole 24 Ore, 08-09-2019

Figura 17 – I padroni del futuro (tratta dal Sole 24 Ore, 8 settembre 2019)

Dal punto di vista della distribuzione geografica (vedi Figura 18), si tratta di business dominati ancora dagli USA, in cui tuttavia le imprese cinesi stanno assumendo un ruolo rapidamente crescente, mentre l'Europa soffre della frammentazione dei suoi mercati, vedi ancora la Figura 17 e la Figura 18.

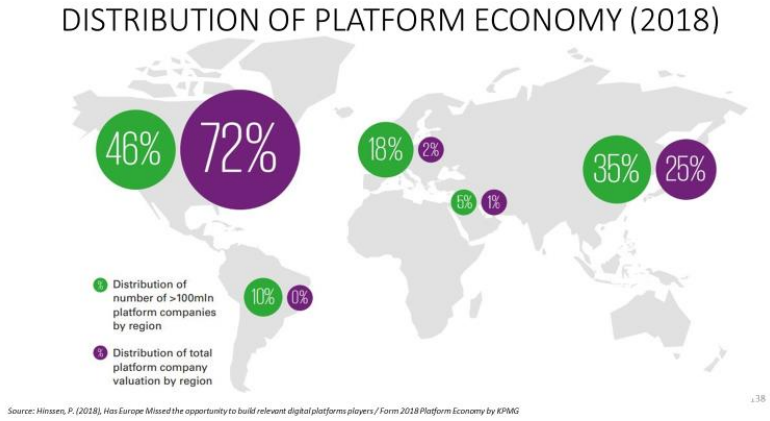


Figura 18 - Distribuzione geografica della Platform Economy (tratta da [Hinnsen 2018])

Definiamo dunque cosa intendiamo per "*piattaforma*": una piattaforma è un business che consente di abilitare interazioni e comunicazione che producono valore fra produttori esterni e consumatori. La piattaforma fornisce un'infrastruttura aperta e partecipativa per queste interazioni e ne garantisce la

governance. L'obiettivo principale della piattaforma è quello di mettere in contat-to gli utenti e di facilitare lo scambio di beni, servizi o di valore sociale, favorendo la creazione di valore per tutti i partecipanti. Approfondiamo ora diversi aspetti che caratterizzano le piattaforme rispetto a modelli di business più tradizionali.

Modelli di business pipeline e modelli a piattaforma

La modalità tradizionale in cui la maggior parte dei processi di business è organizzata è quella detta *a pipeline*. A differenza del modello a piattaforma, in questo caso il valore del business viene prodotto e trasferito passo a passo, dal produttore al consumatore, passando attraverso diverse fasi, dalla progettazione alla produzione, alla promozione e alla vendita: possiamo descrivere questa struttura come "*catena del valore lineare*" [Parker 2016], vedi Figura 19.

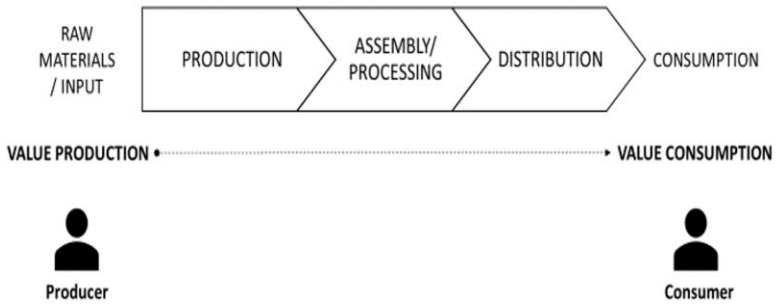


Figura 19 - Struttura di business tradizionale a "Pipeline" (tratta da [Parker, Platfom Revolution 2016])

Negli ultimi anni molte aziende stanno muovendo da una struttura a pipeline ad una a piattaforma; la struttura lineare caratteristica della pipeline si trasforma in una relazione complessa tra produttori e consumatori, rispetto a cui la piattaforma stessa assume una serie di diverse relazioni, vedi Figura 20. Molteplici tipi di utenti (produttori, consumatori, "prosumers", cioè i produttori e consumatori allo stesso tempo) entrano in collegamento e sviluppano interazioni fra di loro utilizzando le risorse a cui possono accedere tramite la piattaforma. La piattaforma "batte" il modello a pipeline perchè la piattaforma consente di scalare più efficientemente attraverso l'eliminazione degli intermediari, vedi mostrato in Figura 21 il caso di Airbnb.

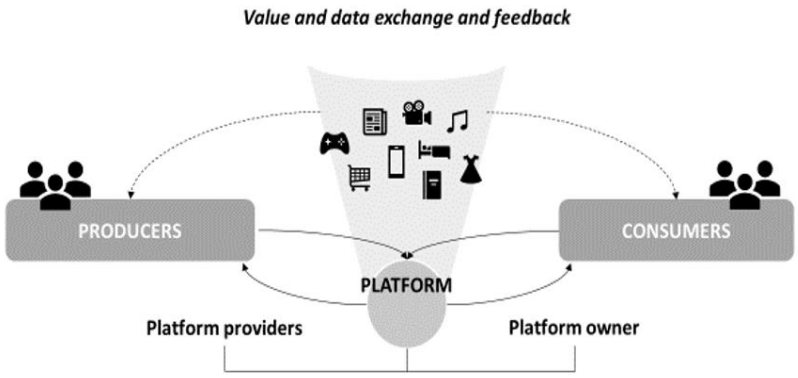


Figura 20 - Struttura del business a piattaforma (tratta da [Parker, Platfom Revolution 2016])

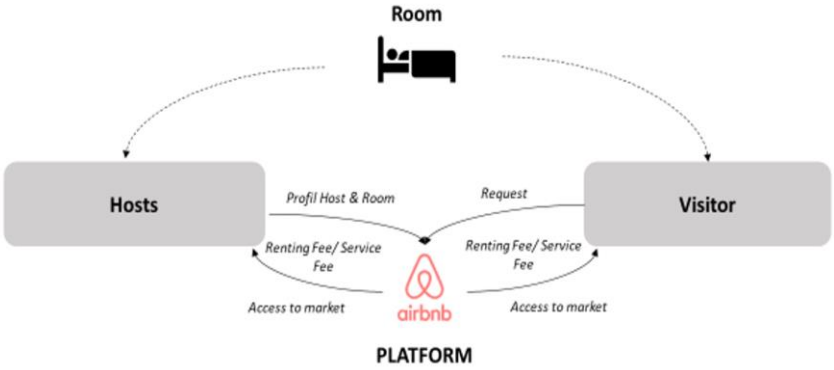


Figura 21 - Il business model di Airbnb, basato sulla piattaforma digitale
(tratta da [Parker, Platfom Revolution 2016])

La piattaforma:

- non possiede direttamente le risorse (il valore deriva dagli utenti e dalle comunità che essa serve) è aperta (almeno parzialmente) e permette una partecipazione organizzata.
- dipende solo dalle sue regole e dalla sua architettura.
- consente di scalare molto più velocemente e fa leva sugli effetti di rete (vedi Sezione 4).

La Figura 22 seguente mostra, attraverso il confronto fra le strategie di Microsoft e di Amazon, le profonde differenze fra obiettivi e visioni strategiche di due tra i maggiori protagonisti di questa nuova economia digitale.

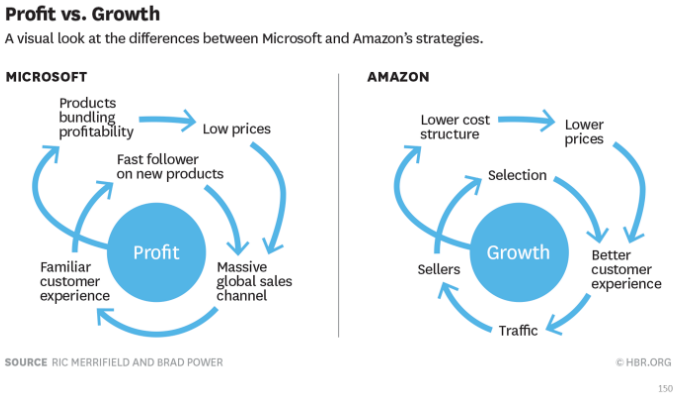


Figura 22 - Le strategie di piattaforma di Microsoft e Amazon
(tratta da [Parker, Platfom Revolution 2016])

I Protagonisti dell'Ecosistema Piattaforma

Una piattaforma fornisce dunque l'infrastruttura e le regole per un marketplace che mette in collegamento fornitori e consumatori. I players dell'ecosistema coprono quattro ruoli principali, ma possono assumere rapidamente ruoli diversi. Comprendere le relazioni all'interno e all'esterno dell'ecosistema è quindi essenziale per identificare la strategia della piattaforma; i ruoli sono (vedi anche Figura 23):

- *Produttori*, creatori dei beni e servizi che vengono offerti sulla piattaforma
- *Consumatori*, acquirenti dei beni e servizi offerti
- *Fornitori*, che mettono a disposizioni i dispositivi e le interfacce per operare sulla piattaforma
- Il *Proprietario*, che controlla la piattaforma e definisce chi può partecipare, in che modo e secondo quali regole.

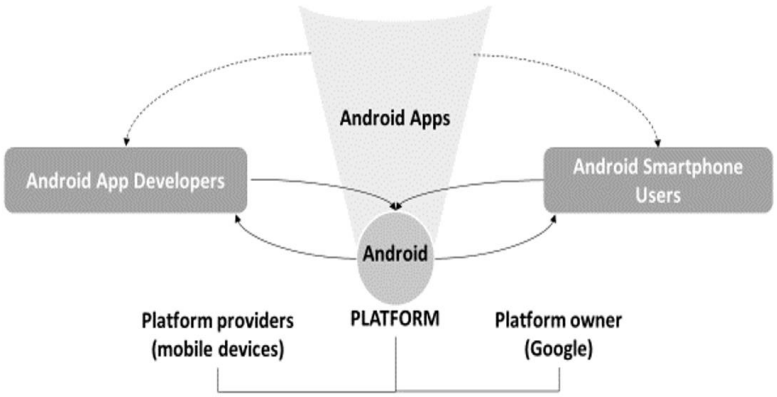


Figura 23 - I *player* in un Ecosistema Piattaforma: il caso Android
(tratta da [Parker, Platfom Revolution 2016])

La più importante forma di attività che ha luogo sulla piattaforma è la "*core interaction*", ovvero lo specifico scambio di valore che attrae in primo luogo la maggior parte degli utenti della piattaforma. La core interaction si articola in tre componenti-chiave, che rappresentano la struttura -base del modello piattaforma:

- I partecipanti, ovvero il Produttore e i Consumatori: chi crea e chi consuma valore
- L'unità di valore; ogni core interaction comincia con uno scambio di informazioni che ha valore per i partecipanti (d es. prodotti o servizi basati su informazioni, video, tweet, profili di utenti, ecc.)
- Il filtro, un tool software basato su un algoritmo utilizzato dalla piattaforma per abilitare lo scambio di unità di valore fra gli utenti.

Per favorire lo sviluppo di un elevato volume di interazioni di valore, la piattaforma deve essere in grado di eseguire tre funzioni-chiave:

- Pull: la piattaforma deve innanzitutto essere in grado di attirare produttori e consumatori di un certo settore.
- Facilitate: la piattaforma deve facilitare le loro interazioni fornendo I tools e definendo le regole che consentono loro di connettersi e di realizzare scambi di valore apprezzabile.
- Match: la piattaforma deve mettere efficacemente in relazione produttori e consumatori, utilizzando informazioni su ciascuno di essi così da connetterli in un modo apprezzabile da ciascuna delle parti.

Nel tempo, le piattaforme di successo tendono a crescere, aggiungendo nuovi tipi di transazioni alla "core interaction" di base. Ad esempio, possono cambiare le unità di valore scambiata fra gli utenti esistenti, o introdurne di nuove; introdurre nuove categorie di produttori o di consumatori, ecc.

<u>La rivoluzione delle Piattaforme sta rapidamente trasformando l'economia mondiale.</u>

In realtà l'economia delle piattaforme esiste da sempre, dai primi mercati all'aria aperta in cui agricoltori e artigiani vendevano i loro prodotti ai consumatori, alle Borse e alla grande distribuzione.
La principale differenza con queste piattaforme di business tradizionali è data dalle tecnologie digitali, che hanno enormemente ampliato la copertura, la velocità, la capacità di connessione e la convenienza delle piattaforme. Internet e le tecnologie associate hanno aperto una enorme capacità di trasformazione delle industrie tradizionali, spesso in modo assolutamente imprevedibile.

INDUSTRY	EXAMPLES
Agriculture	John Deere, Intuit, Fasal
Communication and Networking	Linkedin, Facebook, Twitter, Tinder, Instagram, Snapchat, WeChat
Consumer Goods	Philips, McCormick Foods FlavorPrint
Education	Udemy, Skillshare, Coursera, edX, Duolingo
Energy and Heavy Industry	Nest, Tesla Powerwall, General Electric, EnerNOC
Finance	Bitcoin, Lending Club, Kickstarter
Healthcare	Cohealo, SimplyInsured, Kaiser Permanente
Gaming	Xbox, Nintendo, Playstation
Labor and Professional Services	Upwork, Fiverr,99designs, Sittercity, LegalZoom
Local Services	Yelp, Foursquare, Groupon, Angie's List
Logistics and Delivery	Munchery, Foodpanda, Haier Group
Media	Medium, Viki, YouTube, Wikipedia, Huffington Post, Kindle Publishing
Operating Systems	iOS, Android, MacOS, Microsoft Windows
Retail	Amazon, Alibaba, Walgreens, Burberry, Shopkick
Transportation	Uber, Waze, BlaBlaCar, GrabTaxi, Ola Cabs
Travel	Airbnb, Trip Advisor

Figura 24 - La penetrazione delle piattaforme digitali nei maggiori settori dell'economia mondiale (tratta da [Parker, Platfom Revolution 2016])

La Figura 24 mostra la grande penetrazione che le piattaforme hanno rapidamente raggiunto, fino a dominare spesso ampi settori dell'economia globale.

Riferimenti

M. Boisot - Information space: a framework for learning in organizations, institutions, and culture. Routledge, 1995

M. Boisot - Knowledge Assets - Securing Competitive Advantagein teh Information Economy, Oxford University Press, 1998.

R. Bukht e R. Heeks - Defining, Conceptualising and Measuring the Digital Economy - Center for developing Informatics, University of Manchester, 2017

Deloitte - What is digital economy? Unicorns, transformation and the internet of things", 2017,"https://www2.deloitte.com/mt/en/pages/technology/articles/mt-what-is-digital-economy.html

S. Federico - Capitalism without Capital - Doppiozero, 27 dicembre 2018

P. Grefen - Beyond E-Business: Towards networked structures London, 2015

J. Haskel e S. Westlake - Capitalism without Capital: The rise of the Intangible Economy, Princeton University Press, 2018.

P. Hinnsen - Has Europe missed the opportunity to build relevant digital platforms players?, Post, 30 gennaio 2018.

V. Mayer-Schonberger e Thomas Range, " Reinventing Capitalism in the Age of Big Data, London, 2018

D. Moody e D. Walsh - Measuring The Value Of Information: An Asset Valuation Approach, 1995

G. G. Parker, M. W. Van Alstine e S. Choudary - Platfom Revolution, New York, 2016.

G. G. Parker, M. W. Van Alstine e S. Choudary - Pipelines, Platforms and the new Rules of Strategy, Harvard Business Review, 2016.

J. Rifkin - The Age of Access, New York, 2000

J. Rifkin - The Zero Marginal Cost Society, New York, 2014

C. Shapiro e H. R. Varian - Information Rules: a Strategic Guide to the Network Economy, 1998

Capitolo 14 – Dati digitali e società

Carlo Batini

1. Introduzione

La guerra in Siria ha prodotto grandi atrocità, morti, bombardamenti, migrazioni, e al momento in cui questo libro viene distribuito non si vede all'orizzonte una soluzione umanitaria. Attivisti e organizzazioni umanitarie denunciano l'aviazione di Assad che continua a colpire con bombardamenti indiscriminati la popolazione civile, che spesso è intrappolata e non riesce a porsi in salvo in luoghi più sicuri. Nel numero del Foglio di martedì 6 agosto 2019 si racconta che la società Hala Systems ha sviluppato nel 2016 un modello predittivo che attraverso tecniche di machine learning basati su tracce di voli aerei e elementi atmosferici fa una analisi dei possibili obiettivi di bombardamenti e invia messaggi agli utenti tramite i social media. Questa tecnologia salva potenziali vittime innocenti e viene naturale definirla una tecnologia *buona*, caratterizzata da un importante obiettivo sociale: preservare la vita di esseri umani.

In questo capitolo prendiamo in considerazione le problematiche che nascono con riferimento ai dati digitali nel loro utilizzo nelle società. Una società è definita in Wikipedia come un insieme di individui dotati di diversi livelli di autonomia, relazione ed organizzazione che, variamente aggregandosi, interagiscono al fine di perseguire uno o più obiettivi comuni. Questa definizione può applicarsi nelle società a vari livelli, che corrispondono ad ambiti territoriali (es. gli abitanti di Roma), politici (i simpatizzanti di un certo partito o movimento), lavorativi (i lavoratori di una certa azienda), comunicativi (gli utenti di Facebook), culturali (i lettori di libri), merceologici (gli acquirenti di gelati), e altro. Si usa spesso anche la espressione di stakeholder o portatori di interessi per indicare i gruppi di utenti che ricevono, nel nostro caso da elaborazioni su dati digitali, una specifica utilità.

Esaminare tutte le possibili relazioni tra Scienza dei dati e Società è al di fuori della portata di questo testo e anche della mia cultura personale. Parleremo solo in maniera circoscritta di open data, che nella più ampia generalità richiederebbero un capitolo a parte, non parleremo dell'uso dei big data per analisi sociali, non parleremo del tema delle reti sociali e delle dinamiche e degli effetti che esse suscitano nella società. Mi concentro qui su alcuni aspetti, per arrivare alla fine di questa introduzione a chiarire quali saranno gli approfondimenti nelle sezioni successive.

I dati digitali possono essere utilizzati per migliorare tanti aspetti della nostra vita personale e collettiva. Da diversi anni, uso la dichiarazione dei redditi precompilata, risparmio tempo e non impazzisco a tenere le carte in ordine per un anno, perché queste carte sono già state inviate alla Agenzia delle entrate dai soggetti con cui ho avuto rapporti economici che influiscono sulla dichiarazione (es le farmacie, i cui scontrini possono essere utilizzati per le deduzioni dovute a spese mediche). E' chiaro però che a monte della mia interazione con la Agenzia delle entrate vi deve essere una reciproca fiducia, altrimenti la affronterò in modo sospettoso, e alla fine perderemo sia io che lo Stato.

La Scienza dei dati può essere utilizzata per un miglior rapporto con le amministrazioni pubbliche. Utilizzando le tecnologie di integrazione tra basi di dati e di collaborazione tra amministrazioni, si può evitare la sindrome da cittadino pony express, per cui in ogni nuova interazione devo comunicare alla amministrazione B ciò che nel passato ho comunicato alla amministrazione A. Il sistema di cooperazione tra Inps, Inail e Camere di Commercio di cui abbiamo parlato nel Capitolo 2 aveva proprio questo scopo.

D'altra parte, realizzare il sistema ha comportato la realizzazione di un collegamento delle posizioni delle imprese nelle basi di dati dei tre enti, e quindi ha portato come effetto collaterale la possibilità di contrasto alla evasione contributiva, incrociando i dati dei tre enti sulle imprese. L'integrazione dei dati, di cui abbiamo parlato nel Capitolo 6 può essere utilizzata per tanti altri scopi, per pubblicare dati comparativi sui prezzi dei beni, o segnali di miglioramento/decadimento sulla qualità dei servizi erogati, e una infinità di altre informazioni utili. Allo stesso tempo, è proprio la integrazione sui dati che permette ai grandi vendors come ad es. Amazon di sapere tutto su di noi.

Per riprendere un tema molto dibattuto quando nel 1967 ho iniziato la Università, le tecnologie digitali non sono mai neutrali, e possono essere utilizzate per tanti fini; una prima provvisoria conclusione che possiamo trarre è che sta a noi come cittadini e come elementi di comunità, sviluppare la consapevolezza critica sugli usi attuali e sugli usi potenziali dei dati digitali. Alla fonte di questa consapevolezza è la trasparenza sul funzionamento delle tecnologie che operano su dati digitali, tema questo che affronteremo nel prossimo Capitolo 15 dedicato all'etica.

Un aspetto che mi interessa particolarmente riguarda il tema dei big data in ambito statistico. Le statistiche sociali sono tra le più importanti tra quelle prodotte dagli enti statistici pubblici, e sono utilizzate dai governi per pianificare le politiche sociali. Gli enti nazionali di statistica hanno maturato nei decenni una posizione di terzietà nelle istituzioni dello stato, per cui le statistiche che vengono prodotte sono assunte corrette, obiettive e esenti da distorsioni (o bias). I big data possono migliorare le statistiche ufficiali prodotte dagli enti nazionali di statistica, non solo dal punto di vista della efficienza, perché possono (parzialmente) sostituire i costosi censimenti nazionali con fonti alternative di dati, ma anche perché possono estendere le statistiche sociali con ulteriori informazioni di provenienza dal Web, e possono permettere indagini su fenomeni a scala territoriale variabile, ad esempio in un territorio circoscritto come una periferia di una città per individuare segnali di disagio; ovvero possono monitorare fenomeni di inquinamento o diffusione di malattie contagiose, in modo ancor più efficace di quanto fece Snow durante la epidemia di colera a Londra nel 1854.

La precedente formulazione del tema dei dati digitali per la società, è ispirata ad una visione positiva e tendenzialmente "progressista" sull'uso dei dati digitali. Altre analisi esprimono un punto di vista più problematico. [Desouza 2014] nota che i problemi sociali vengono spesso chiamati problemi "complicati". I problemi sociali non solo sono più "caotici" dei problemi scientifici, come abbiamo detto fin dal Capitolo 1, ma sono anche più dinamici e complessi a causa del numero di stakeholders (portatori di interessi) coinvolti e a seguito delle molteplici interrelazioni e feedback che esistono tra i relativi componenti. Numerose agenzie governative e organizzazioni no profit sono potenzialmente coinvolte nell'affrontare questi problemi, con una cooperazione e condivisione dei dati tra loro in genere difficile e limitata. La maggior parte di queste organizzazioni ha risorse informatiche inadeguate rispetto alle

loro controparti nelle scienze "dure" o nelle applicazioni gestionali che riguardano problemi di business, le quali, al contrario, hanno ampio accesso a dati finanziari, sui prodotti e sui clienti.

Oltre agli ostacoli infrastrutturali che si presentano agli analisti di problemi sociali, gli stessi dati possono essere un problema. Spesso i dati sono mancanti e incompleti o archiviati in basi di dati "verticali" o in formati proprietari, e quindi non elaborabili facilmente. Altre rigidità derivano da norme e regolamenti che rendono complesse le collaborazioni tra le diverse parti interessate ad affrontare lo stesso tipo di problema. [Desouza 2014] elenca un insieme di barriere nel creare e usare big data nel settore pubblico, in sintesi:

1. I dati sono "sepolti" nei sistemi informativi amministrativi. Sulla base della mia esperienza all'Aipa, il termine è quello giusto; non esiste spesso documentazione, le tabelle e gli attributi vengono chiamati con nomi simili a quelli usati nella Figura 19 del Capitolo 12, il significato dei dati è perso ed è molto difficile e costoso ricostruirlo.
2. Sono carenti gli standard di rappresentazione, ad esempio la data 3 gennaio 2018 è rappresentata come *030118* in un sistema e come *20180103* in un altro, rendendo difficile e costosa la navigazione tra dati con formati diversi.
3. I dati sono spesso inaffidabili e di scarsa qualità.
4. I dati possono causare conseguenze inattese; ad es. la raccolta e pubblicazione di dati aperti sui possessori di fucili nel Connecticut dette un potenziale vantaggio informativo ai ladri nel concentrarsi sui soggetti privi di fucile.

In [Desouza 2014] sono anche indicati possibili azioni per incrementare l'uso dei big data nel sociale, quali:

- costruire archivi digitali globali sui temi critici per la umanità
- coinvolgere attivamente i cittadini
- sviluppare una scienza dei cittadini (Citizen science)
- creare un gruppo mondiale di scienziati dei dati, curatori (curators) dei dati.

Serge Abiteboul, uno tra i più importanti ricercatori nel campo delle basi di dati nella fase della maturità ha investigato in [Abiteboul 2016] problemi di ricerca sulla relazione tra dati digitali e società, ragionando in particolare sul tema della *responsabilità*, con particolare riferimento agli aspetti di responsabilità legati alle politiche pubbliche.

Per Abiteboul gli enti governativi hanno la responsabilità di regolare, nell'ambito del ciclo di vita che abbiamo trattato nel Capitolo 2, le attività di acquisizione e, soprattutto, di diffusione delle e accesso alle tipologie di dati che abbiamo classificato nella Appendice 1 del Capitolo 1, e tra essi quelli pubblici, quelli essenziali, e quelli personali. Questa responsabilità è stata recentemente estesa alla qualità dei dati e alla problematica delle fake news e della post verità.

Peraltro, una regolamentazione dall'alto verso il basso è difficile da realizzare e mantenere, per diversi motivi. Innanzitutto, il problema è internazionale: i dati vengono generati e gestiti in diversi paesi con leggi diverse mentre serve una base di regole internazionali (da questo punto di vista, l'Europa è in una posizione privilegiata vista la sua natura sovranazionale); secondo, in una fase in cui le tecniche e le tecnologie per big data si stanno sviluppando rapidamente è difficile "inseguire" i cambiamenti

tecnologici. Infine, ogni regolazione sui dati è difficile da implementare, ad esempio potrebbe non essere possibile determinare chi ha prodotto un determinato dataset e in quali circostanze, perché spesso mancano o sono volutamente occultati i metadati corrispondenti. Su questa stessa problematica si è soffermato recentemente Franco Debenedetti il 13 settembre del 2019 in un articolo sul Foglio, accessibile dal Web; è impossibile sintetizzare qui le tesi di Di Benedetti, riporto la tesi di fondo che compare nel sottotitolo: una regolamentazione privata dei contenuti è preferibile a una pubblica consiglio la lettura dell'articolo nella sua versione integrale.

Tornando al ragionamento di Abiteboul, anche se la regolamentazione è sicuramente una questione complessa e potrebbe non essere fattibile oggi, c'è una grande necessità di di consapevolezza e coinvolgimento da parte dei governi. I governi potrebbero usare congiuntamente una combinazione di regolamentazione e incentivi; ad esempio, i governi possono fornire incentivi alle organizzazioni che aderiscono a regole di condivisione e scambio di dati e di tecniche e programmi applicativi, così da permettere verifiche aperte sulla adozione di pratiche "data-responsible".

Infine, è compito dei governi definire norme sulla integrazione dei dati provenienti da applicazioni e organizzazione diverse, così da impedire le forme di integrazione che permettono la identificazione dei soggetti, violando in tal modo la loro privacy, e allo stesso tempo abilitare le integrazioni che aumentano il valore aggiunto dei dati, permettendo nuove analisi e correlazioni.

Il Capitolo approfondisce alcune delle tematiche introdotte in precedenza, e ne propone altre. La Sezione 2 tratta diversi problemi di giustizia sociale che sorgono nel ciclo di vita del dato aperto. La Sezione 3 affronta il problema del *data divide* come evoluzione del digital divide, la condizione cioè per cui l'accesso ai dati digitali è diversificato e iniquo per i diversi ceti sociali nei singoli paesi, e tra i diversi paesi. La Sezione 4 approfondisce la questione del ruolo delle statistiche pubbliche, sia dal punto di vista della loro effettiva applicabilità nei paesi in via di sviluppo, sia dal punto di vista delle novità che i metodi per la produzione di statistiche pubbliche presentano nell'era dei big data. La Sezione 5 presenta un insieme di risultati derivanti dalla mia ricerca sul valore sociale dei dati. La Sezione 6 tratta il problema cruciale nelle società democratiche del declino dei giornali.

2. Il divario sociale nel ciclo di vita del dato aperto

Abbiamo visto nel Capitolo 2 come il dato segua un ciclo di vita, strutturato secondo un insieme di fasi e passi. Nell'ambito di tale ciclo di vita, i *dati aperti* (o open data) sono dati generati dalle amministrazioni e istituzioni pubbliche, che vengono messi a disposizione e resi accessibili in formato elaborabile a tutti tramite il Web. Il movimento dei dati aperti è in gran parte un movimento che riflette l'etica libertaria presente nel settore delle tecnologie della informazione e in particolare la cultura open source.

Come affermato in [Open Government Working Group 2010], Internet e il Web sono lo spazio pubblico del mondo moderno; i governi del mondo attraverso i dati aperti hanno l'opportunità di comprendere meglio le esigenze dei loro cittadini e i cittadini possono partecipare più pienamente ai processi decisionali del loro governo.

Secondo il precedente punto di vista l'accessibilità al dato e il suo uso diventano l'elemento fondante dei dati digitali nella società, per cui recentemente si usa il termine di data democratization a indicare il fenomeno della diffusione del dato per tutti. Accanto al fenomeno sociale dell'uso, i dati sono anche e forse soprattutto, visto il diffondersi delle reti sociali, una forma di comunicazione tra attori che incorpora in ciò che viene comunicato le ipotesi e la visione del mondo degli attori partecipanti nello scambio comunicativo. I dati sono, come tutte le tecnologie, un artefatto, costituito da concetti espressi da attori, che interpretano il mondo fisico attraverso strutture intellettuali per mezzo delle quali capiscono il mondo; come tali, i dati sono incorporati in un insieme di pratiche sociali attraverso cui i dati stessi vengono creati, interpretati e utilizzati.

I dati non devono essere considerati e analizzati "in isolamento", ma in relazione agli altri dati che possono emergere dalla realtà sociale. Pensiamo, per esempio, ai dati utili per costruire un osservatorio sulla recidiva nei reati, fenomeno sul quale supponiamo di voler produrre statistiche descrittive del fenomeno e interpretative dei fenomeni sociali. Un primo insieme di dati può riguardare le informazioni anagrafiche dei detenuti ed ex detenuti e informazioni sulle restrizioni alla libertà cui sono stati condannati. Con questi dati si può fare ben poco, se non correlare dati anagrafici con dati sulle pene e i reati connessi. Possiamo estendere i dati a rappresentare la storia delle detenzioni, le misure alternative, gli interventi applicati per reinserire il detenuto nella società; potremmo aggiungere dati che descrivono l'ambiente sociale in cui il soggetto ha vissuto, la sua formazione, il contesto familiare. In questo modo il soggetto analizzato si trasforma progressivamente da una matricola ad un essere sociale.

La natura sociale dei dati presenta aspetti intimamente legati al concetto di *giustizia sociale*. Tale concetto ha due diverse accezioni; nella prima, o *giustizia distributiva*, si è interessati agli effetti provocati da una contesa tra due soggetti sociali provocata da obiettivi opposti, e al modo con cui arbitrare tra aspettative in contrasto tra di loro. Accanto ad essa, occorre considerare anche un concetto di *giustizia strutturale*, intesa come il grado con cui una società considera e supporta le condizioni istituzionali necessarie per la realizzazione dei valori associati alla qualità della vita.

Riguardo a ciò, la natura dei dati insita nel codificare fenomeni sociali rende possibili l'instaurarsi di situazioni di privilegio nella loro formazione e nel loro utilizzo, [Johnson 2013]. Quando i valori e i privilegi che i dati creano sono ingiusti, nel duplice senso riferito al concetto di giustizia sociale introdotto poco fa, nessun fenomeno di dati aperti e nessuna elaborazione statistica può annullare le imprecisioni o le incompletezze sui dati originali. 'Garbage in, garbage out' 'è un concetto centrale nell'uso sociale dei dati. Alla qualità dei dati abbiamo dedicato il Capitolo 5.

Un esempio molto citato a tale proposito è la sottostima di alcuni gruppi sociali quali africani e spagnoli nel Censimento della popolazione negli Stati Uniti. Le famiglie o membri delle famiglie non sono rappresentati nel censimento per un bias verso gli africani e gli spagnoli; tali etnie non sono rappresentate perché la base di dati del censimento contiene errori e incompletezze, nei casi in cui:

- il nucleo familiare sia composto da persone non in relazione di parentela ovvero con legami sociali labili,
- i membri siano raramente a casa,

- quando vi sia un senso debole di responsabilità civica, e forse un sfiducia attiva nei confronti del governo
- quando la permanenza nelle case è volatile,
- quando l'inglese non è parlato bene.
- dove i legami di comunità non sono forti.

[Johnson 2013] osserva che a seguito della miriade di modalità in cui il privilegio sociale può essere "immerso" nei dati, i dati aperti non hanno spesso in sé la caratteristica intrinseca di promuovere la giustizia, e possono facilmente marginalizzare i gruppi non rappresentati nei dati; gruppi che la mancanza di potere nella società esclude dai tipi di interazioni che producono i dati, rendendo perciò le loro esigenze invisibili nel nuovo sistema digitalizzato. Rendere i dati *aperti* non fa che propagare le ingiustizie che permeano i dati nel momento in cui sono raccolti; qualunque tentativo si faccia per promuovere l'equità nell'uso dei dati è per intrinsecamente portato al fallimento, e i dati danno luogo a un processo per cui "ingiustizia in input provoca ingiustizia in output".

Un secondo esempio molto studiato è la digitalizzazione dei dati sul possesso di terreni nello stato di Karnataka [Johnson 2016]. Nello stato sono stati progressivamente digitalizzati il catasto terreno, il catasto dei beni immobili e la base di dati delle proprietà dei terreni e beni immobili. Ad essi sono stati aggiunti dati sulla età, la casta, e la religione dei proprietari, insieme ad altri dati territoriali. I programmi di digitalizzazione sono stati finanziati dal governo indiano, un partenariato pubblico privato ha reso le informazioni accessibili tramite chioschi internet distribuiti in tutto lo stato.

I primi effetti del programma di digitalizzazione sono stati la esclusione della casta Dalit (spesso indicati come gli 'Intoccabili'), i cui beni spesso non erano documentati nei precedenti archivi, ma erano descritti in altre fonti cartacee o frutto di una tradizione orale. Aspetti problematici del programma sono stati la selezione solo di alcune tipologie di documenti per la inclusione nella nuova base di dati, insieme alla decisione di archiviare i dati risultanti in una base di dati relazionale (vedi Capitolo 3), orientata verso rappresentazioni strutturate dei dati che non permettono di descrivere la ricchezza delle informazioni provenienti da documenti testuali.

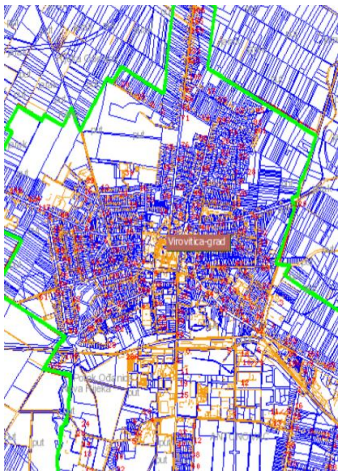
Le modalità di progettazione esposte in precedenza hanno quindi impedito di rappresentare conoscenza informale e storica, agendo come una accetta su tutto quanto faceva parte di accordi e pratiche sociali non codificate, caratteristica tipica delle rivendicazioni di appartenenti alla casta dei Dalit. Mancando la informazione, quelle rivendicazioni diventavano basate sul nulla. Torneremo su questo punto nel prossimo capitolo inquadrandolo nel tema della critica al concetto di trasparenza.

La scelta di digitalizzare i tre archivi menzionati, insieme alla aggiunta di informazioni demografiche, riflette anche una tipica mentalità e deformazione burocratica: i progettisti del nuovo sistema vedevano le informazioni imprecise e la presenza di informazione inconsistente o ridondante (ad esempio molteplici possessori dello stesso terreno) non come una spinta a ricostruire e disambiguare, ma, piuttosto, come una manifestazione di inefficienza e di opacità, che doveva essere sanata identificando un singolo proprietario, quello attestato dagli archivi ufficiali presi come singola fonte della verità. Il lettore attento ritroverà qui alcune tecniche che abbiamo introdotto nella Sezione 6 del Capitolo 6 dedicata alla fusione di dati. Questo atteggiamento riflette l'intrinseca natura burocratica che viene

attribuita alla efficienza e alla consistenza della informazione, quando il mondo attorno a noi è spesso incompleto e contraddittorio. L'amministrazione pubblica assume di poter semplificare la società rendendola falsamente decifrabile e non neutrale in quella che abbiamo chiamato la *giustizia distributiva.*

a. La «realtà»

b. Le vecchie mappe



b. Le nuove mappe digitali

Figura 1 - Realtà, vecchie mappe, nuove mappe e percezione distorta dei piccoli proprietari

Riguardo alle conseguenze sui processi di compravendita di terreni, come documentato in [Johnson 2016], quando agli agricoltori indiani che erano piccoli proprietari terrieri, furono mostrate (vedi Figura 1) le nuove mappe catastali digitali, che sostituivano la simbologia delle vecchie mappe analogiche con una nuova simbologia, non più comprensibile, con lo scopo di rendere più trasparenti e meno soggette a corruzione le attività amministrative di vendita e acquisto di terreni, l'effetto fu l'esatto opposto; il mutamento dei linguaggi simbolici, l'aumento della complessità e la necessità della intermediazione amministrativa ridussero la trasparenza e aumentarono gli eventi corruttivi.

Molti agricoltori non erano infatti alfabetizzati, e quindi in grado di negoziare gli affari attraverso i processi amministrativi tradizionali; con le nuove procedure automatizzate, sulle quali non avevano nessun controllo, e delle quali non riuscivano ad avere nessuna padronanza, diventarono totalmente dipendenti dagli operatori di front office che operavano nei villaggi.

La corruzione non fu debellata negli uffici tecnici e amministrativi, questo per la mancanza di formazione e orientamento verso coloro che attuarono il progetto, che hanno visto la iniziativa come un modo per ottenere maggiori vantaggi economici rispetto al passato. Analogamente, la trasparenza sulle attività amministrative, invece che aumentare si tramutò in una maggiore opacità, leva tipica del fenomeno corruttivo. In sostanza, mentre in teoria la iniziativa aveva lo scopo di democratizzare l'accesso ai dati, il risultato finale fu di aumentare il potere di chi già lo aveva.

3. Il data divide e la data democratization

Il tema dei big data si coniuga con nuove forme di "divide" nell'uso delle tecnologie della informazione e della comunicazione. Tradizionalmente, con il termine digital divide si intende (vedi Wikipedia) "il divario esistente tra chi ha accesso effettivo alle tecnologie dell'informazione (in particolare personal computer e Internet) e chi ne è escluso, in modo parziale o totale. I motivi di esclusione comprendono diversi aspetti: condizioni economiche, livello d'istruzione, qualità delle infrastrutture, differenze di età o di sesso, appartenenza a diversi gruppi etnici, provenienza geografica.

Con il diffondersi dei dati digitali, si intende con *data divide* la relazione asimmetrica tra coloro che raccolgono, archiviano e analizzano grandi quantità di dati e la vastissima comunità di persone che vorrebbe poter utilizzare in maniera estesa i dati nella propria vita personale, ma non vi riesce, per mancanza anzitutto dell'accesso ai dati, e secondariamente dei modelli con cui poter comprendere il significato dei dati e dei linguaggi per accedere ai dati e analizzarli. La forma più critica di "big data divide" è vista in questo senso nella asimmetria tra I dati disponibili ai grandi operatori mondiali, come Facebook e Google, e coloro che forniscono i dati, cioè tutti noi, che in questo scambio non riceviamo nessun vantaggio (gli operatori affermano che in realtà, al contrario, riceviamo nuovi servizi, ad esempio di georeferenziazione).

Il data divide, al di là della asimmetrica possibilità di accedere ai dati, ha una ulteriore dimensione nella asimmetrica modalità con cui i dati possono essere utilizzati per decisioni, come dimostrano diversi modelli sviluppati per predire la recidiva nei comportamenti di gruppi sociali o etnici, fenomeno diffuso nelle tecniche di predictive policing [Gangadharan 2014], [Lyon 2003].

I big data, in virtù della loro ineguale diffusione nel mondo, tendono a creare nuove forme di "digital divide". Abbiamo detto che la disponibilità di big data permette di raccogliere dati dall'intero universo di persone o eventi che si intende analizzare; tuttavia, per il modo con cui i dati vengono acquisiti, questo universo può essere rappresentativo solo di una parte del fenomeno; si pensi ad esempio ancora a Google, che non è accessibile in diversi paesi del mondo. Osservate anche la Figura 2, in cui sono mostrate le immagini di Google Earth che rappresentano Times Square a New York e un frammento dello slum Kibera a Nairobi, uno dei più grandi slum del mondo. Il dettaglio disponibile per Times Square per le strade, gli edifici, le sedi delle banche, degli esercizi commerciali, dei ristoranti, ecc. è infinitamente maggiore rispetto alla piatta e non informativa rappresentazione di Kibera. Questo porta a un grande disequilibrio nelle possibilità di espansione economica, turistica, commerciale tra i due territori. Questo disequilibrio, peraltro, vale non solo per rappresentazioni proprietarie come Google Earth, ma anche per rappresentazioni open, quali Open Street Map.

Un progetto che, cito tra i tanti, cerca di recuperare questo divide è il progetto "Map Kibera", organizzato nell'ambito della iniziativa *GroundTruth*, messo in atto proprio per contrastare la mancanza di dati cartografici e tematici e di altra informazione pubblica su uno dei più grandi slum del mondo, lo slum Kibera a Nairobi. Il progetto ha tra gli altri i seguenti obiettivi:

- la identificazione delle risorse interne al campo per i donatori di fondi.

- la identificazione dei servizi che sono in corso di progressivo miglioramento a seguito dei progetti finanziati (per esempio fontane dispensatrici di acqua),
- il contrasto alla informazione falsa e preconcetto su Kibera
- la diffusione di informazione certificata su eventi rilevanti, come ad esempio le elezioni.

Il disequilibrio discusso negli esempi precedenti riguarda anche la informazione statistica, sia economica che sociale. Ad esempio, il prodotto interno lordo e l'indice di povertà, tipici indicatori statistici raccolti nei paesi in tutto il mondo, sono indicatori di grande importanza nella determinazione degli aiuti ai paesi in via di sviluppo; alcune indagini delle Nazioni Unite e dell'OCSE dimostrano che tali indicatori in diversi paesi sono calcolati, per la carenza di dati, in modo notevolmente sottostimato.

Il data divide riguarda anche, in misura rilevante, la ricerca sperimentale, che basa sulla disponibilità di dati la sua stessa possibilità di esistenza. Essendo molti big data accessibili solo a pagamento, la loro disponibilità è garantita in modo ineguale, e i gruppi più ricchi si trovano ad avere un vantaggio competitivo che aumenta il divario rispetto ai gruppi di paesi caratterizzati da minore ricchezza. Mentre dunque alcune forme di accesso dipendono dai mezzi tecnici utilizzati per raccogliere, archiviare e gestire i dati, altre rivestono aspetti di carattere economico, tra coloro che hanno la possibilità di pagare per accedere alla conoscenza che i dati creano, e coloro che non se lo possono permettere. Ad esempio, Twitter fornisce accesso limitato ai propri dati al pubblico e pieno accesso a un numero limitato di aziende e ricercatori.

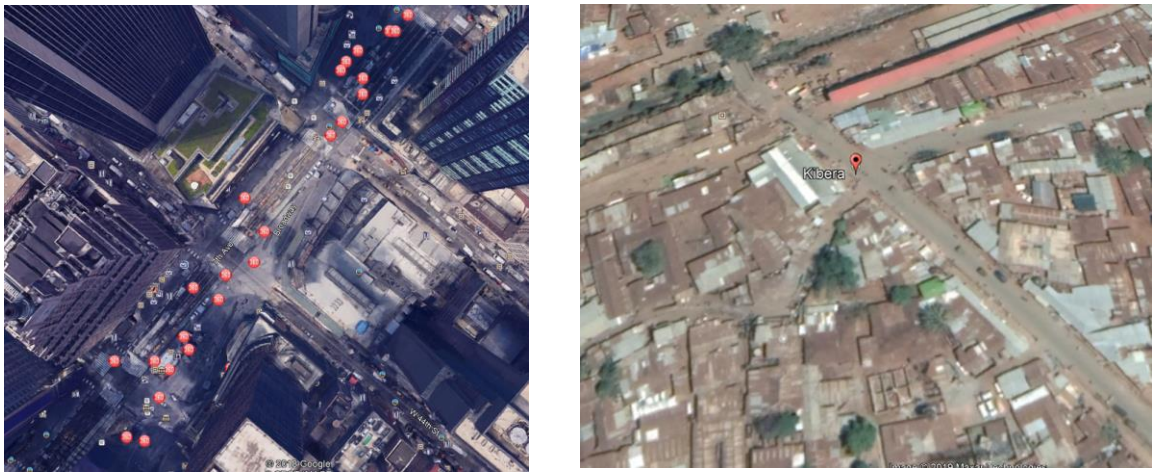


Figura 2 – Times Square a New York e lo slum Kibera a Nairobi (tratte da Google Earth)

Il libro di Cathy O'Neill [O'Neil 2016] "Weapons of Math Destruction", in italiano Armi di distruzione matematica, o Armi matematiche di distruzione, discute un impressionante serie di casi in cui l'autrice individua una discriminazione nell'uso dei dati. I casi discussi includono, tra gli altri:

- i modelli utilizzati per valutare le prestazioni degli insegnanti,
- gli abusi associati ai i modelli di recidiva utilizzati nelle decisioni giudiziarie;
- le ingiustizie perpetrate dall'uso dei test di personalità nelle decisioni di assunzione;
- i modelli algoritmici per massimizzare l'efficienza sul lavoro in cambio di bassi salari;

- le asimmetrie nei modelli che valutano Il rischio di credito;
- il potenziale vulnus al processo democratico nell'uso di big data in campagne politiche.

Altre forme di divide riguardano la “data analysis divide”, che evidenzia una relazione asimmetrica tra coloro che raccolgono, archiviano e analizzano grandi quantità di dati, e coloro che non dispongono degli strumenti e della cultura per analizzare i dati. Altri ancora evidenziano il divide nella disponibilità da parte di alcuni delle risorse per accesso al cloud, tecnologia che è chiaramente disponibile per chi ha la possibilità di pagarla.

Il data divide acquista una rilevanza particolare nell’ambito dell’accesso e elaborazione di dati individuali. E’ evidente che sia potenzialmente molto più efficace analizzare modelli di comportamento individuali nel contesto di modelli sociali più ampi piuttosto che affidarsi esclusivamente ai dati storici relativi a un particolare individuo. Si potrebbe immaginare che un avvio alla soluzione del problema consista nell’obbligo da parte dei grandi data providers (es. Amazon, Google, ecc.) di consentire agli utenti di accedere ai propri dati.

Ciò, tuttavia, non è risolutivo dal punto di vista dell’equo uso dei dati individuali: anche se gli utenti avessero accesso ai propri dati, non avrebbero la possibilità di inferire sui dati e sviluppare le capacità predittive rispetto a coloro che possono estrarre i dati aggregati e i modelli che su di essi si basano. In questo caso possiamo parlare di un divide nella ampiezza dell’universo dei dati disponibili e aggregabili. Inoltre, anche se agli individui fossero forniti tutti i dati (una condizione puramente ipotetica), essi mancherebbero della capacità di memorizzazione e della potenza di elaborazione per dare un senso ai dati e metterli in uso. Ne consegue che il divario strutturale associato all'avvento di nuove forme di creazione di senso basata sui dati sarà sempre più evidente in futuro, poche’ la dimensione dei dati, come abbiamo visto nel primo capitolo, cresce continuamente nel tempo.

Nonostante tutte queste difficoltà, occorre anche osservare che negli ultimi cinquant'anni molti stati e i loro governi hanno formalmente riconosciuto il diritto di accesso alle informazioni, attraverso leggi chiamate Freedom of Information Act, e alcune, ad esempio il Sudafrica, hanno incardinato il diritto di accesso nelle loro Costituzioni. Queste iniziative vanno inquadrate in un fenomeno chiamato recentemente Data democratization, che riguarda l’utilizzo dei dati nelle società per far crescere il livello e la qualità dei principi democratici, in cui rientrano il diritto all’accesso, il diritto ad essere informati sulle leggi e norme, il diritto ad essere informati sui processi decisionali, il diritto ad avere disponibili i dati utili per il miglioramento delle condizioni di vita.

L'accesso ai dati è essenziale affinché le persone realizzino il loro diritto fondamentale di partecipare al governo del loro paese e vivere sotto un sistema basato sul consenso informato della cittadinanza. In qualsiasi stato, e in particolare in quelli in cui le capacità di analisi delle politiche della società civile sono scarsamente sviluppate, i diritti di partecipazione politica non possono essere esercitati efficacemente senza accesso alle informazioni del governo.

Gli strumenti per garantire la data democratization citati in letteratura sono spessi in sintonia con le tematiche che abbiamo affrontato nei capitoli precedenti. Riguardano in sintesi:

1. Facilitare l'accesso ai dati attraverso interfacce e strumenti di ricerca usabili, comprensibili, efficaci.
2. Fornire dati di qualità, non affetti da errori, approssimazioni, incompletezze, obsolescenze.
3. Accrescere sempre più le fonti disponibili, fornendo anche filtri progressivi che evitino il fenomeno del sovraccarico di informazioni.
4. Mettere a disposizione di tutti le tecniche utilizzate dagli analisti per la produzione di modelli a partire da big data
5. Dare la possibilità di condividere conoscenza, permettendo a tutti di utilizzare modelli che facilitino la condivisione, come i linked open data.

Per un approfondimento sugli ultimi due punti, si veda [Espinosa 2014]. Per altri contributi di discussione sulla data democratization si veda [Alexander 2016], [Fahey 2014], [Treuhaft 2016], e la Chief Data Officer Guide to Data Democratization citata nei riferimenti.

4. Ruolo delle statistiche pubbliche nell'era dei big data

La disponibilità di dati statistici di natura economica o sociale è un elemento fondamentale per le moderne società; vale il motto "conoscere per agire". In una società si può agire solo se si conoscono a fondo non solo i fenomeni particolari e focalizzati, ma anche rappresentazioni astratte e generali descrittive della intera comunità che opera in una nazione. Questa visione si è imposta fin dal diciottesimo/diciannovesimo secolo, e l'episodio di Snow descritto nel primo capitolo è un esempio plastico di questo punto di vista innovativo. Naturalmente, le statistiche non possono essere piegate ad un punto di vista o ideologia specifica, ma devono quanto possibile esprimere una visione e analisi obiettiva della società; per questa ragione, (quasi) ogni paese nel mondo è dotato di un Ente statistico nazionale, cui è riconosciuta dalle leggi e regolamenti ampia autonomia di giudizio e indipendenza dalla politica e dalle altre istituzioni nella produzione delle statistiche ufficiali sulla società e sulla economia.

Vediamo ora di approfondire le precedenti problematiche, focalizzandoci dapprima sul tema della rilevanza delle statistiche ufficiali nella società, e discutendo successivamente tracce recenti di declino che alcuni ricercatori individuano nel ruolo della statistica; descriveremo infine come le metodologie statistiche sono influenzate e arricchite dal fenomeno dei Big data.

Importanza delle statistiche ufficiali e ruolo della statistica nella produzione di conoscenza

Come nota Giovannini in [Giovannini 2010], il termine "statistica" viene da "scienza dello Stato". Nella nostra epoca, in cui siamo bombardati da messaggi rivolti talvolta alla nostra "pancia" o alla nostra percezione, che come sappiamo è altamente soggettiva e influenzabile, la statistica diventa rilevante [Giovannini 2012] quando si tratta di sviluppare opinioni su fenomeni di cui non abbiamo una conoscenza diretta. Questa osservazione porta anche a fare una prima distinzione tra Statistica e scienza dei dati. Storicamente, la statistica è lo studio della collezione, analisi, interpretazione, presentazione e organizzazione dei dati. La scienza dei dati è lo studio della estrazione generalizzabile di conoscenza dai dati; essa incorpora elementi e tecniche e teorie da molti altri campi del sapere. Un

aspetto chiave che ha enormemente incrementato la applicabilità della Scienza dei dati è lo sviluppo del machine learning, che è utilizzato per scoprire pattern dai dati e sviluppare modelli predittivi e prescrittivi. Naturalmente, statistica e scienza dei dati si complimentano, come vedremo tra poco.

Dove entra in gioco la Statistica, per Giovannini? È ben nota la tendenza della stragrande maggioranza delle persone ad acquistare un quotidiano, o a seguire programmi televisivi, più in linea con le proprie convinzioni, e lo stesso accade per la consultazione dei siti Internet; abbiamo già parlato del fenomeno delle camere dell'eco nella Sezione 10 del Capitolo 5. E' qui che la statistica dovrebbe entrare in gioco. Infatti, "*la statistica*[51] *è stata sviluppata per andare al di là di quello che, come singoli, possiamo conoscere sulla base delle nostre esperienze: non deve quindi stupire che, nella Società dell'informazione siano sempre più spesso i dati statistici a formare l'immagine collettiva dello Stato di un sistema socio-economico, giocando un ruolo decisivo nel determinare l'insieme informativo che la nostra mente archivia e richiama quando necessario*". La Statistica stimola quindi ciascuno di noi a responsabilizzarci nell'assumere una visione collettiva, che tenga conto di tutto l'ambito sociale in cui viviamo, e non soltanto del nostro particolare.

Ma per fare questo, serve in una Società una Istituzione che abbia la capacità (ancora [Giovannini 2012] a proposito dell'Istat) "*di servire la collettività, sviluppando un'approfondita conoscenza della realtà sociale, economica e ambientale dell'Italia ai diversi livelli territoriali e favorendo i processi decisionali di tutti i soggetti (cittadini, amministratori, ecc.), attraverso la produzione e la comunicazione di informazioni sta tistiche e analisi di elevata qualità, realizzate adottando rigorosi principi etico-professionali e i più avanzati standard scientifici."*

"*Per realizzare il proprio compito nel mondo «digitale», gli statistici ufficiali devono fare il percorso inverso, cioè evitare di essere concentrati unicamente sui numeri, ma tenere conto anche della «proiezione mentale» che questi generano nel cervello degli individui e nella società. Solo così, infatti, le statistiche cesseranno di essere considerati codici incomprensibili, ma descriveranno persone, imprese, situazioni e comportamenti concreti, cioè ologrammi immediatamente intellegibili dall'osservatore. È possibile fare ciò senza perdere la propria anima di servitori «neutrali» della società? Fermo restando che nessun dato statistico è del tutto neutrale, in quanto elaborato a partire da un certo modello di misurazione (cioè da una certa visione della realtà)…."* la risposta di Giovannini è positiva. Lo statistico deve trasformarsi da produttore di informazione in generatore di conoscenza, utilizzando anche, aggiungo io, gli strumenti che la Scienza dei dati le può mettere a disposizione nella modellazione della realtà osservabile.

Si può misurare il *valore*, per la Società, della produzione di dati statistici? Ancora Giovannini afferma che "*se, dunque, l'attività statistica è un servizio, possiamo allora domandarci come si misuri la produzione, e quindi il valore aggiunto, di una tale attività. Il Sistema dei conti nazionali ci dice che il valore di un servizio deriva dal cambiamento (fisico o mentale) che la fruizione del servizio produce nel consumatore. Naturalmente, per un servizio di mercato il prezzo a cui si effettua la transazione è una misura della willingness to pay del consumatore, cioè dell'utilità marginale che egli attribuisce alla fruizione del servizio. Ma per le attività non di mercato? Il "Rapporto Atkinson" elaborato alcuni anni fa nel Regno Unito sottolinea come l'output di un'attività non di mercato debba essere misurato in*

[51] Qui e nel seguito in corsivo i testi tratti da Giovannini e altri autori

funzione del contributo che essa fornisce al risultato finale ricercato dall'utente del servizio. Possiamo allora domandarci: quale cambiamento dovrebbe essere prodotto dalla fruizione della statistica, cioè dalla lettura di una tavola o di un grafo contenente dati statistici? La risposta è: la conoscenza che genera. La lettura dei dati su un certo fenomeno (i prezzi, la produzione, l'occupazione eccetera) dovrebbe, cioè, accrescere la conoscenza di esso nell'utente finale".

Questa conoscenza generata dipende da vari elementi: la quantità di statistiche ufficiali prodotte; il ruolo che i media svolgono nella loro diffusione; la loro rilevanza per ciascun consumatore; la fiducia che quest'ultimo ha in chi produce i dati e la sua capacità di trasformare le statistiche in conoscenza (*literacy statistica* o numeracy). Basta che uno solo di tali fattori sia mancante, perché il valore della produzione statistica sia pari a zero, mentre più cresce il numero di persone che allargano la propria conoscenza grazie alle statistiche, più aumenta il valore totale della produzione statistica.

"*Mettere al centro della valutazione del servizio l'utente e il processo cognitivo che compie per trasformare dati in informazioni comprensibili e poi in conoscenza cambia radicalmente la prospettiva. In quella che viene chiamata "Statistica 2.0" la catena di produzione dei dati non si interrompe al momento della diffusione dell'informazione, ma prosegue curandosi di come quest'ultima sia portata all'utente finale dai media, così da soddisfare i bisogni del massimo numero possibile di individui (e non solo del governo, della pubblica amministrazione o di una élite economica o culturale), di quanto gli utenti si fidino di quelle informazioni*".

La crisi delle statistiche ufficiali

In [Letouze 2015] si sottolinea un tema che abbiamo già discusso nel capitolo sulla qualità dei dati: la crescita dei dati disponibili nel Web non favorisce spesso la loro qualità. Nell'ambito della informazione statistica si parla della "disillusione" statistica, definita come la disponibilità di dati inaccurati o incompleti, che tende ad essere maggiore nei paesi in via di sviluppo. Ad esempio, il Prodotto interno lordo (PIL) del Ghana è cresciuto di oltre il 60% da un giorno all'altro dopo aver intrapreso un esercizio di ricalcolo nel 2010; il PIL della Nigeria è cresciuto di oltre il 75% dopo un esercizio analogo nell'aprile 2014.

Un gruppo di paesi - non tutti poveri, ma tutti con una storia di labilità della democrazia - non ha condotto il censimento della popolazione in decenni. Le popolazioni e qualsiasi dato pro-capite possono quindi essere solo stime, anche le cifre "ufficiali", che sono presumibilmente precise, sono molto probabilmente inesatte. Un paese come il Kenya non ha prodotto dati sulla povertà in quasi un decennio. La scarsa qualità dei dati sulla povertà, in particolare, trasforma il monitoraggio e le previsioni in stime molto approssimate, dove le medie sub-continentali devono essere utilizzate come proxy grezzi per i dati a livello di paese.

Ma ci sono altri segni sul momento problematico che vivono le statistiche ufficiali nella nostra epoca. [Davies 2019] osserva che "*i dati statistici hanno l'utile effetto di circoscrivere l'arena del conflitto democratico, descrivendo economia e società in termini oggettivi e consentendo a cittadini e politici di avere concordanza di vedute almeno rispetto alla realtà in cui tutti si trovano. Se una statistica afferma che in un anno vi è stato un declino della natalità rispetto all'anno precedente del 15%, ovvero che l'età*

media è cresciuta in un anno di due mesi e in dieci anni di un anno e mezzo, e quindi è in corso un processo di invecchiamento della popolazione, questo è un fatto oggettivo, sulle cui cause possiamo dissentire, ma non sul dato in quanto tale".

Ci sono tuttavia segnali crescenti, si afferma in [Davies 2019], del fatto che Statistica ed Economia stiano perdendo la loro capacità di porre fine o comporre le discussioni (si pensi a tutto il dibattito sul rapporto deficit/PIL, e sulla sua rilevanza, o scarsa rilevanza, nello sviluppo economico e nel patto di stabilità della Unione Europea). Negli Stati Uniti, la fiducia nella veridicità delle statistiche si riflette pesantemente nelle divisioni politiche: l'86% degli elettori di Hillary Clinton ha espresso fiducia nei dati economici prodotti dal Governo federale, rispetto al 13 % degli elettori di Trump. Nella generale crisi delle elite, e con l'affermarsi di strumenti di comunicazione rapidi, sintetici, specifici ed emotivi, ma concreti e reali, come possono essere i messaggi di Twitter, anche le statistiche ufficiali, così astratte e oggettive, così fredde e algide, perdono valore e diventano anzi spesso controproducenti. Si pensi alla differenza di impatto emotivo tra i dati aggregati sulla diminuzione dei reati, e la percezione soggettiva di paura che si ha passeggiando la sera tardi in un quartiere periferico poco illuminato.

Come i big data modificano i processi di produzione delle statistiche ufficiali

I Big data sono generati da eventi non pianificati a fini statistici; in Figura 3 sono riportate le nuove fonti di dati utilizzabili nelle indagini statistiche.

Tipo di fonte	**Dominio di applicazione**
Dati di ricerche online	Statistiche su forze di lavoro
Dati estratti da siti	Statistiche sui prezzi
Dati da telefoni mobili	Statistiche su mobilità e turismo.
Social media	Statistiche sociali
Ortoimmagini	Statistiche su Agricoltura

Figura 3 – Nuovi tipi di fonti e applicazioni alle statistiche ufficiali

Una metodologia adeguata per un uso statistico dei big data deve perciò essere in grado di [Baldacci 2015]:

- collegare eventi di incertezza nota o stimata cui fanno riferimento i big data alle unità di rilevazione della popolazione di interesse per le statistiche ufficiali (individui, famiglie, imprese o istituzioni);
- elaborare i dati raccolti allo scopo di renderli coerenti con il quadro statistico desiderato (concetti, definizioni, classificazioni);
- dare pesi (con incertezza nota o stimata) ai dati, in modo da garantire rappresentatività rispetto alla popolazione target;
- stimare gli aggregati di interesse e accompagnarli con una misura della loro qualità, sulla base delle misure di incertezza nelle fasi precedenti.

Nella Figura 4, adattata da [Baldacci 2015] sono confrontate le fonti di dati coinvolte nel processo di produzione delle statistiche, con vantaggi e svantaggi legati alla loro adozione. Le indagini (campionarie o censimenti) sono le fonti di dati più adatte per raggiungere l'intera popolazione target, sebbene la costruzione e il mantenimento di un frame della popolazione possa essere molto costoso. Per quanto

riguarda la raccolta dei dati, questa fonte consente di ottenere i dati desiderati attraverso domande finalizzate che soddisfano le esigenze dell'indagine stessa con un insieme chiaramente definito di classificazioni, concetti e definizioni. In questa fase, soprattutto nel caso di indagini molto complesse che richiedono enormi quantità di informazioni dettagliate, il carico degli intervistati aumenta e il rischio di mancata risposta totale e parziale diventa molto elevato. D'altra parte, nelle fasi dell'elaborazione dei dati, le indagini possono fare affidamento su metodologie valide e condivise e non sono interessate da problemi tecnologici, dal momento che la quantità di dati raccolti è ridotta. Per quanto riguarda la stima, questo tipo di fonte utilizza metodologie consolidate, ma alti tassi di non risposta possono influire sulla precisione dei risultati.

Per quanto riguarda le basi di dati amministrative, l'identificazione della popolazione target potrebbe non essere immediata, a causa sia della difficoltà di collegare le informazioni di interesse alle persone nella popolazione target sia dei problemi di sotto-copertura. Nella fase di raccolta dei dati, i costi delle informazioni fornite sono molto limitati e vi è l'opportunità di rendere le definizioni e le classificazioni coerenti con quelle utilizzate in contesti statistici; nondimeno, la mancanza di corrispondenza tra informazioni desiderate e disponibili è un problema concreto. Riguardo alla elaborazione dei dati, i dati amministrativi utilizzano metodologie ben definite e non mostrano problemi tecnologici evidenti.

Elementi di confronto		**Fonti dei dati**		
		Censimenti e Indagini	**Basi di dati amministrative**	**Big Data**
Popolazione di riferimento	A favore	Se disponibile uno schema di identificazione della popola-zione, vi è quasi identità con la popolazione di riferimento	E' possibile effettuare record linkage con la popolazione di riferimento	
	Contro	La costruzione e la manutenzione sono costose	Possibile incertezza nel collegamento con popolazione target Possibile copertura solo parzi-ale	Il collegamento con la popolazione di riferimento può essere complesso per mancanza di metadati adeguati Possibile copertura solo parziale
Acquisizio-ne dati	A favore	Concetti, classificazioni e defini-zione chiaramente espresse nei questionari e funzionali agli scopi delle rilevazioni	Costo ridotto di fornitura dei dati	Costo ridotto di fornitura dei dati
	Contro	Suscitano rifiuto da parte del rispondente con conseguenti risposte parziali	Problemi legati alla debole corrispondenza tra le classificazioni e definizioni desiderate e quelle disponibili	Problemi legati alla debole corrispon-denza tra le classificazioni e definizioni desiderate e quelle disponibili
Elaborazio-ne	A favore	Esistono metodologie mature Non esistono problemi tecnologici dovuti alla dimensione dei dati	Esistono metodologie mature Limitati problemi tecnologici dovuti alla dimensione dei dati	-
	Contro	-	-	Esistono metodologie solo per casi specifici Richiesto un passo di estrazione della conoscenza Possono insorgere problemi tecnologici dovuti alla dimensione dei dati
Stima	A favore	Esistono metodologie mature	Esistono metodologie mature Maggiore accuratezza se i dati amministrativi sono usati congiuntamente con i dati sulle rilevazioni	Esistono metodologie mature Maggio-re accuratezza se i dati amministrativi sono usati congiuntamente con i dati sulle rilevazioni Possibilità di coprire nuove esigenze informative
	Contro	Problemi di accuratezza nel caso di alti tassi di non risposta	Possibili distorsioni ed errori nelle stime	Esistono metodologie solo per casi specifici Possono insorgere problemi tecnologici dovuti alla dimensione dei dati

Figura 4 – Fonti di dati coinvolte nel processo di produzione delle statistiche, vantaggi e svantaggi. (da [Baldacci, 2015])

In [Baldacci 2015] si propone un approccio alla integrazione dei big data nel processo di produzione delle statistiche ufficiali, in cui si afferma che considerando una determinata popolazione target (famiglie, individui, imprese, istituzioni, ecc.), diverse fonti di dati sono originate dalla (o possono essere correlate alla) popolazione target (vedi ancora Figura 4):

- dati statistici, raccolti mediante indagini tradizionali; tutte le informazioni identificative e strutturali relative alla popolazione target sono organizzate in un contesto, la base per i censimenti o le indagini campionarie; i dati vengono raccolti contattando direttamente le unità selezionate nella popolazione e trattati per produrre stime che vengono diffuse agli utenti;
- dati amministrativi, derivanti dall'esecuzione di procedure amministrative (tasse, sicurezza sociale, sanità, istruzione, carte d'identità, dati contabili interni, ecc.). Questi dati possono essere collegati ad altre fonti statistiche con certezza o con una procedura di record linkage probabilistico e possono essere usati come variabili ausiliarie o possono coincidere con variabili di interesse;
- big data, originati dall'uso di dispositivi digitali come già visto nella Figura 4.

Mentre l'obiettivo finale di produrre informazioni statistiche affidabili rimane lo stesso, diversi scenari possono essere delineati sulla base del grado di utilizzo dei big data nel processo di produzione:

- Uso di dispositivi digitali (in particolare di quelli collegati via Internet) come media per la raccolta di dati: il quadro statistico complessivo rimane invariato;
- uso di big data *in combinazione* con dati statistici, come informazioni aggiuntive che possono essere utilizzate al fine di migliorare la qualità dei dati statistici (ad esempio nella fase di modifica e acquisizione) o per migliorare l'affidabilità delle stime utilizzando i big data; allo stesso modo in cui dati amministrativi o i dati del censimento possono essere usati insieme ai dati di campionamento (ad esempio negli stimatori compositi nella stima di piccole aree);
- uso di Big Data *in sostituzione* di dati statistici. Quando si verificano determinate condizioni, al fine di ridurre il costo e il carico di risposta, è possibile sostituire completamente il processo classico di produzione basato su indagini statistiche e adottare processi radicalmente diversi basati sull'uso integrale dei Big Data.

In Figura 5 mostriamo infine il flusso di attività che integra nuovi dati dal Web e Internet nelle statistiche ufficiali.

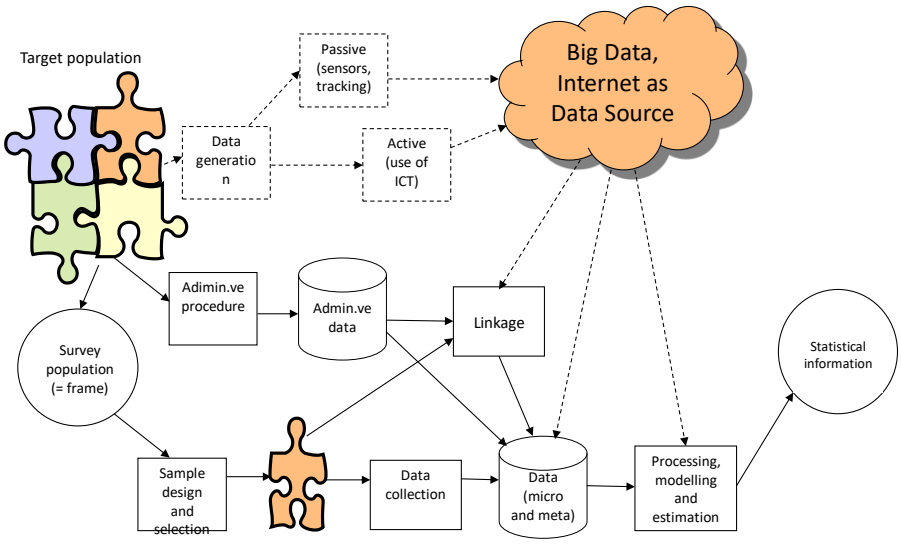


Figura 5 - Flusso di attività che integra nuovi dati dal Web e Internet nelle statistiche ufficiali (da [Baldacci, 2015])

5. Il valore sociale dei dati

Osserviamo la Figura 6, tratta da un numero dell Economist del 2011; il testo parla di una ricerca svolta dalla Università di Stoccolma, in cui si fornisce evidenza del fatto che in Uganda la disponibilità da parte delle famiglie di dati sulla qualità della cura negli ospedali ha permesso di ridurre di un terzo le morti dei bambini da 0 a 5 anni. Dunque i dati possono migliorare la qualità della vita delle persone.


The Economist

The Open Government Partnership

The parting of the red tape

Is it just another global talking-shop—or a fresh approach to shaking out government secrecy?

Oct 8th 2011 | NEW YORK AND TALLINN | from the print edition

UGANDA is not best known as a testbed for new ideas in governance. But research there by Jakob Svensson at the University of Stockholm and colleagues suggested that giving people health-care performance data and helping them organise to submit complaints cut the death rate in under-fives by a third. Publishing data on school budgets reduced the misuse of funds and increased enrolment.

Figura 6 - La disponibilità di dati sulla qualità della cura negli ospedali in Uganda (dall'Economist, 8 ottobre 2011)

Il valore sociale dei dati può essere definito come la capacità che hanno i dati di fornire una risposta alle esigenze delle comunità in termini di qualità della vita. L'organizzazione per la cooperazione e sviluppo economico (OCSE) ha sviluppato da tempo un quadro concettuale per la misurazione del

benessere e del progresso di una nazione o di un gruppo sociale. Esso propone una distinzione tra benessere presente e benessere futuro; il benessere presente è costituito da due domini principali: le condizioni materiali di vita e la qualità della vita. Più specificamente riguardo alla qualità della vita gli ambiti del quadro concettuale riguardano:

- Stato di salute (in cui ricade il caso dell'Uganda)
- Relazione vita lavoro
- Formazione e competenze
- Relazioni sociali
- Coinvolgimento civico e partecipazione alle decisioni pubbliche
- Qualità dell'ambiente
- Sicurezza personale
- Sensazione soggettiva di benessere

e con riferimento alle condizioni materiali riguardano:

- Reddito e ricchezza
- Impiego e condizioni abitative

Come secondo esempio per ragionare sul valore sociale dei dati consideriamo l'iniziativa della polizia inglese che pubblica sul Web (www.police.uk) per ogni città nel territorio di sua competenza e per ogni quartiere della città una mappa dei reati compiuti in un determinato arco temporale, con numerosità e tipo di reato, vedi Figura 7. Il servizio ha richiesto lo sviluppo di un sistema che acquisisce a partire dalle denunce le informazioni sui reati, che vengono poi georeferenziate e aggregate su una mappa della Gran Bretagna.

Se ad esempio un italiano va a vivere in Gran Bretagna, può scegliere il luogo dove vivere tenendo anche conto della rischiosità dei reati nella via o nel quartiere. D'altra parte, se le informazioni sui reati acquistano una valenza sociale perché forniscono una risposta sia pur parziale alle esigenze di sicurezza, le stesse informazioni hanno anche una valenza economica per gli affittuari e per coloro che intendono vendere le loro case, perché esse possono essere deprezzate dalla conoscenza del fatto che il relativo quartiere è a forte rischio reati. Questo caso ricade nella categoria OCSE della *sicurezza*.

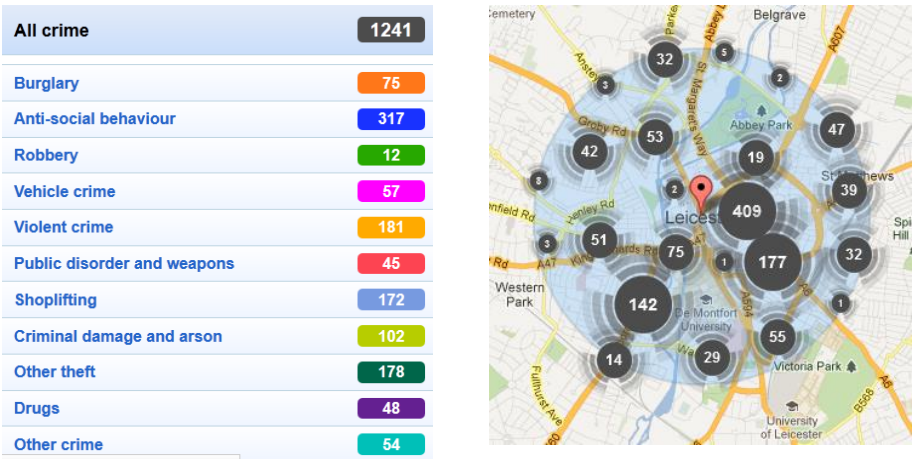


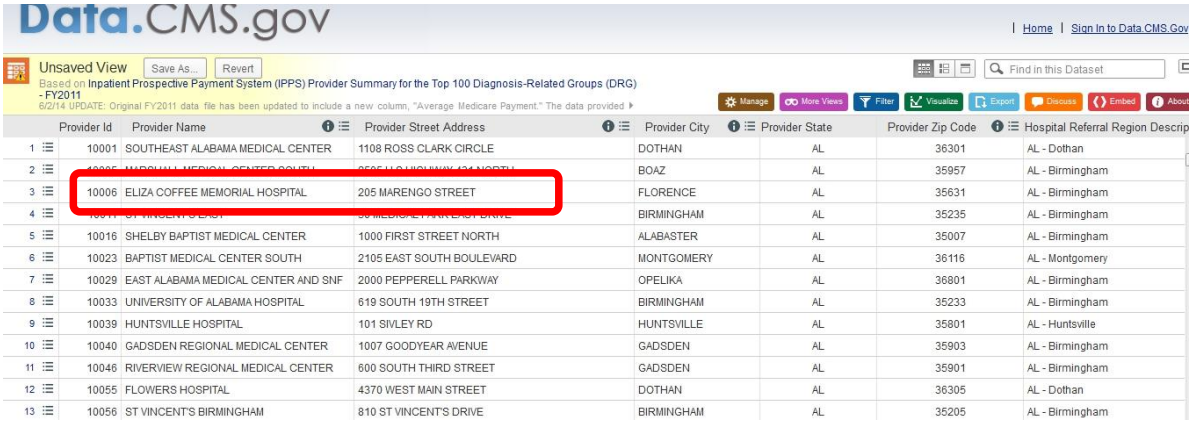
Figura 7 – Sicurezza nelle città inglesi (tratta da https://data.police.uk/)

In entrambi i casi precedenti la disponibilità di dati permette di risolvere (almeno parzialmente) un problema, nel primo caso dove portare a curarsi un bambino che soffre di una patologia aumentando la probabilità che possa uscire guarito, nel secondo caso dove prendere in affitto una casa riducendo la probabilità che io possa subire un reato.

Anche l'economista premio Nobel Amartya Sen ha sviluppato una teoria economica orientata al valore sociale, distinguendo, nel determinare l'impatto degli investimenti economici, tra "funzionamenti" (cioè stati degli esseri umani e delle azioni o attività che una persona può intraprendere) e la effettiva possibilità e libertà di perseguire il benessere ("capacità"). Volendo applicare l'approccio si Sen al valore sociale dei dati, si tratterebbe di misurare *quanto essi incrementano le capacità e i funzionamenti*, rispetto al caso in cui non fossero disponibili.

I precedenti esempi e approcci ci dicono che il valore sociale dei dati è soggettivo; sapere che un quartiere nella periferia di una grande città è rischioso per viverci mi può servire per evitarlo se ho possibilità economiche, mi dà una informazione meno utile se posso andare in affitto soltanto lì con il mio reddito.

Anche le tre tabelle successive di Figura 8 mostrano esempi che possono essere riferiti al valore sociale dei dati. In questo caso si tratta di open data messi a disposizione nell'ambito della iniziativa Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS). Come si vede la Tabella 1 fornjisce informazioni sugli indirizzi degli ospedali, la Tabella 2 i decessi osservati, attesi e normalizzati rispetto al rischio per tipo di trattamento cardiaco e per chirurgo, e la Tabella 3 la percezione soggettiva della esperienza vissuta in un ricovero da parte dei pazienti.



Data.CMS.gov

Home | Sign In to Data.CMS.Gov

Unsaved View Save As... Revert

Based on Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG) - FY2011

6/2/14 UPDATE: Original FY2011 data file has been updated to include a new column, "Average Medicare Payment." The data provided

Find in this Dataset

Manage | More Views | Filter | Visualize | Export | Discuss | Embed | About

	Provider Id	Provider Name	Provider Street Address	Provider City	Provider State	Provider Zip Code	Hospital Referral Region Descrip
1	10001	SOUTHEAST ALABAMA MEDICAL CENTER	1108 ROSS CLARK CIRCLE	DOTHAN	AL	36301	AL - Dothan
2				BOAZ	AL	35957	AL - Birmingham
3	10006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	FLORENCE	AL	35631	AL - Birmingham
4				BIRMINGHAM	AL	35235	AL - Birmingham
5	10016	SHELBY BAPTIST MEDICAL CENTER	1000 FIRST STREET NORTH	ALABASTER	AL	35007	AL - Birmingham
6	10023	BAPTIST MEDICAL CENTER SOUTH	2105 EAST SOUTH BOULEVARD	MONTGOMERY	AL	36116	AL - Montgomery
7	10029	EAST ALABAMA MEDICAL CENTER AND SNF	2000 PEPPERELL PARKWAY	OPELIKA	AL	36801	AL - Birmingham
8	10033	UNIVERSITY OF ALABAMA HOSPITAL	619 SOUTH 19TH STREET	BIRMINGHAM	AL	35233	AL - Birmingham
9	10039	HUNTSVILLE HOSPITAL	101 SIVLEY RD	HUNTSVILLE	AL	35801	AL - Huntsville
10	10040	GADSDEN REGIONAL MEDICAL CENTER	1007 GOODYEAR AVENUE	GADSDEN	AL	35903	AL - Birmingham
11	10046	RIVERVIEW REGIONAL MEDICAL CENTER	600 SOUTH THIRD STREET	GADSDEN	AL	35901	AL - Birmingham
12	10055	FLOWERS HOSPITAL	4370 WEST MAIN STREET	DOTHAN	AL	36305	AL - Dothan
13	10056	ST VINCENT'S BIRMINGHAM	810 ST VINCENT'S DRIVE	BIRMINGHAM	AL	35205	AL - Birmingham

Tabella 1 - Indirizzi degli ospedali (da https://www.medicare.gov/hospitalcompare/data/overview.html)

Tabella 2 - Percentuale di decessi osservati, attesi e normalizzati rispetto al rischio per trattamento cardiaco per chirurgo.
(da https://www.medicare.gov/hospitalcompare/data/overview.html)

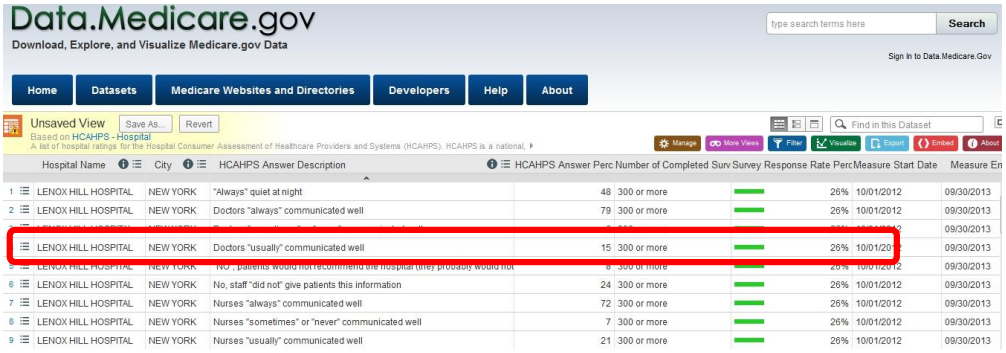


Tabella 3 - Percezione della esperienza vissuta in un ricovero da parte dei pazienti
(da https://www.medicare.gov/hospitalcompare/data/overview.html)

Figura 8 – Varie misurazioni che possono ricondursi al concetto di valore sociale dei dati

Osservando le tre tabelle ci vengono in mente le seguenti considerazioni:

- Certamente i dati della Tabella 2 e 3 sono intuitivamente di maggior valore dei dati della Tabella 1, ma non è immediatamente chiaro scegliere tra la 2 e la 3.
- La Tabella 2 fa riferimento ad un fenomeno doloroso ma oggettivo, la Tabella 3 fornisce dati che basandosi sulla percezione hanno il vantaggio di effettuare rilevazioni sui fruitori del servizio di cura, p pazienti, e allo stesso tempo lo svantaggio per cui potrebbero anche essere influenzati da fattori esteriori e non di sostanza (es. il fatto che il medico sia simpatico e si intrattenga con il paziente).

L'intento della parte restante della sezione è quello di esporre alcuni metodi di misurazione delle qualità sociale dei dati che ho concepito nei miei lavori di ricerca, svolti in collaborazione con Gigi Viscusi, Federico Cabitza, e Angela Locoro, vedi [Viscusi 2014], [Cabitza 2015] e [Viscusi 2016]. Il valore sociale è incentrato su informazioni sulla salute.

Nella successiva Figura 9 vengono mostrate due metodologie, la prima oggettiva e basata sugli schemi delle basi di dati di cui misurare il valore sociale e e la seconda basata sulla percezione e focalizzata sui valori dei dati nella base di dati (vedi Capitolo 3 sui modelli).

Misura obiettiva
1. Individua le fonti di dati da confrontare
2. Per ogni fonte, costruisci lo schema concettuale
3. Produci lo schema concettuale integrato
4. Produci un insieme di domande e relative decisioni che influenzano in specifici eventi la qualità della vita con riferimento alle condizioni di salute.
5. Misura il grado di copertura degli schemi associati alle fonti per rispondere alle domande

Misura percepita
1. Produci un insieme di domande e relative decisioni che influenzano in specifici eventi la qualità della vita con riferimento alle condizioni di salute.
2. Chiedi a un campione di utenti la rilevanza percepita delle domande e pesa i rispettivi risultati.

Figura 9 - Metodologie nel caso di valutazione obiettiva e di valutazione basata sulla percezione

Metodo basato sugli schemi

Assumiamo di confrontare quattro indagini svolte nel passato sulla qualità della cura negli ospedali, quella degli ospedali negli Stati Uniti che abbiamo appena visto, l'analoga effettuata per gli ospedali Canadesi, la indagine svolta dalla rivista Wired qualche anno fa, e infine l'indagine Dove Salute, una iniziativa del Ministero della Salute italiano.

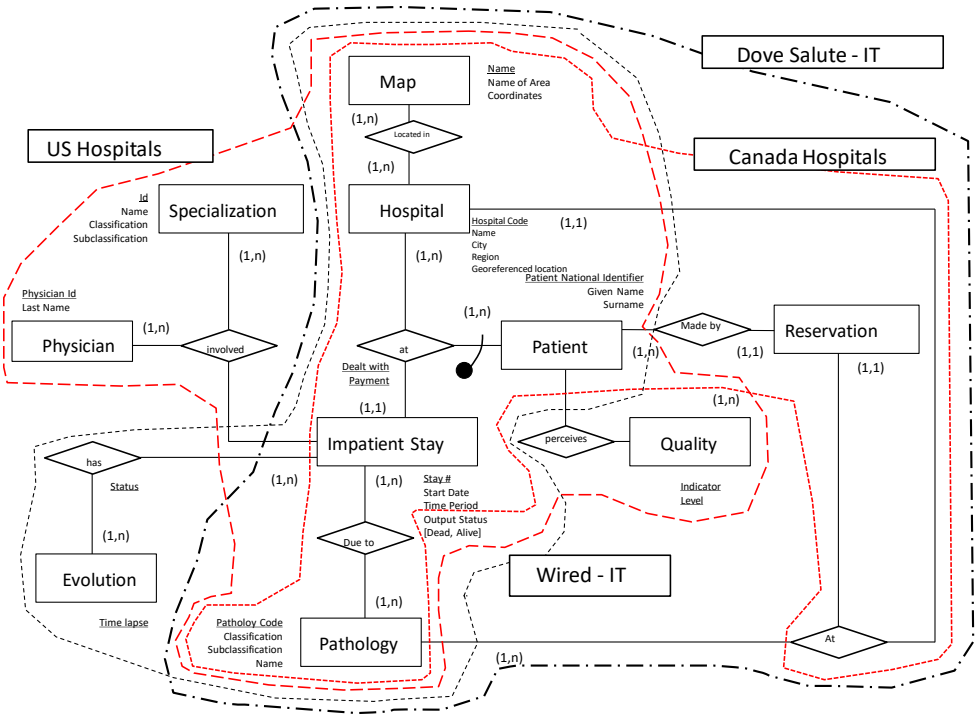


Figura 10 – Schemi locali relativi alle quattro fonti di dati e schema globale

Per ciascuna delle quattro indagini è stata effettuata una ricerca che permettesse di costruire lo schema concettuale sottostante l'indagine. Possiamo per esempio osservare tutte le tabelle del tipo di quelle di Figura 9, creare per ogni tabella lo schema concettuale della tabella e poi fondere gli schemi. Oppure possiamo partire dai questionari di rilevazione associati alla indagine, e domanda pe domanda procedere alla concettualizzazione fino a costruire lo schema finale. I quattro schemi di dati sono visti integrati in uno schema finale in Figura 10 nella pagina precedente.

A questo punto possiamo metterci nei panni di una persona che per le sue condizioni di salute abbia la necessità di accedere ai dati che lo schema rappresenta. Potete fare un esperimento, e formulare un insieme di domande o di statistiche che vi farebbe comodo avere. Nel lavoro di ricerca da cui è tratto questo esempio sono emerse le domande di Figura 11.

Domanda
Q1. Ospedale più vicino
Q2. Ospedali che curano la patologia
Q3. Percentuale di dimissioni dovute a complicazioni
Q4. Percentuale di secondi ricoveri entro 6 mesi
Q5. Percentuale di dimissioni dovute a complicazioni per medico curante
Q6. Percentuale di secondi ricoveri per medico curante
Q7. Percentuale di soddisfatti per patologia per paziente
Q8. Percentuale di soddisfatti per patologia per i parenti
Q9. Tempo medio prima del ricovero per patologia

Figura 11 - Domande nel caso di valutazione sugli schemi

A questo punto, osservando lo schema integrato, ma con un occhio anche agli schemi locali, dobbiamo verificare per ogni schema quali entità siano ricomprese nello schema e per ogni interrogazione o statistica quali entità tra quelle visitate dalla interrogazione o statistica appartengano allo schema, vedi Figura 12.

Osservando la mappatura di Figura 12 appare chiaro che lo schema in grado di rispondere a più domande è quello della rilevazione effettuata negli US. Si noti anche che la Domanda 8 riguarda anche una entità non presente in nessuno degli schemi, i parenti, questa informazione potrebbe portare in futuro ad arricchire le indagini misurando anche la percezione dei parenti.

	Map	Hospital	Specialization	Patient	Physician	Reservation	Inpatient Stay	Quality	Pathology	Evolution	Relative
US	X	X	X	X	X		X	X	X		
Canada	X	X		X		X	X		X		
Wired - It	X	X		X			X		X	X	
Dove Salute - IT	X	X		X		X	X	X	X		
Q1	o	o									
Q2		O							O		
Q3		O		O			O		O		
Q4		O		O			O		O	O	
Q5		O	O	O	O		O		O		
Q6		O	O	O	O		O		O	O	
Q7		O		O			O	O	O		
Q8		O		O			O	O	O		O
Q9		O		O		O	O		O		

Figura 12 – Copertura delle fonti e delle domande

Metodo basato sulla percezione

Quanto alla valutazione basata sulla percezione, in un'altra indagine abbiamo preso in considerazione i tre casi di Figura 13, che sono a diverso livello di gravità, e che coinvolgono rispettivamente un parente, la persona oggetto della rilevazione, e un figlio.

- *Ictus di un proprio caro* - Uno dei tuoi cari sta mostrando i sintomi di un ictus improvviso. Le proponi di chiamare un'ambulanza, ma lei risponde che non si sente così male e ti chiede se potessi portarla tu personalmente al miglior ospedale.
- *Il tuo mal di schiena* - Hai sofferto di dolore di schiena per anni. Soffri ancora per il fatto che fare certi movimenti è estremamente difficile per te; sulla base di un esame medico, la diagnosi del tuo ortopedico è ernia del disco, e la cura è una operazione, da prenotare e fissare rapidamente.
- *Febbre del bambino* - Tua figlia di un anno è malata e ha la febbre alta. Dato che è tarda notte non la puoi portare dal pediatra, quindi decidi di andare in uno degli ospedali della città

Figura 13 – I tre casi dei questionari

Le domande estendono le precedenti dettagliando il ruolo dei parenti, come emerso dalla discussione precedente.

Q1. Ospedale più vicino
Q2. Ospedali che curano la patologia
Q3. Percentuale di dimissioni dovute a complicazioni
Q4. Percentuale di secondi ricoveri entro 6 mesi
Q5. Percentuale di dimissioni dovute a complicazioni per medico curante
Q6. Percentuale di secondi ricoveri per medico curante

Q7a. Percentuale di soddisfatti per patologia
Q7b. Percentuale di soddisfatti su tutti i pazienti
Q8a. Percentuale di soddisfatti tra i parenti dei pazienti per patologia
Q8b. Percentuale di soddisfatti su tutti i parenti dei pazienti
Q9. Tempo medio prima del ricovero per patologia

Figura 14 - Domande nel caso della valutazione della percezione

A questo punto, possiamo cercare caratteristiche che permettano di associare pesi differenti di rilevanza ai tre casi; esse fanno riferimento a:

- La *criticità*, che riguarda la misura in cui una situazione è grave per quanto riguarda la necessità di ricevere cure ospedaliere, la condizione correlata o il problema di salute.
- La *probabilità*, stimata, ad esempio, su base della frequenza da dati epidemiologici, del presentarsi della patologia o dell'emergere del bisogno correlato.
- Il *coinvolgimento,* che si riferisce al grado di coinvolgimento emotivo che il rispondente che legge lo scenario potrebbe sentire rispetto alle condizioni di salute: "me stesso" e la prole sono i più alti livelli di coinvolgimento, conoscenti e sconosciuti il più basso, parenti e amici il livello intermedio
- *L'urgenza* si riferisce alla necessità di un intervento rapido e quindi di un rapido recupero delle informazioni per eseguire l'intervento terapeutico.

Vedi in Figura 15 i pesi da 1 a 4 associati a priori per le quattro caratteristiche associate alle tre patologie.

Caratteristica/Patologia	Ictus	Mal di schiena	Febbre
Criticità	4	2	2
Probabilità	1	3	1
Coinvolgimento	3	1	3
Urgenza	2	4	4

Figura 15 – Pesi associati alle caratteristiche nelle tre patologie

In Figura 16 compaiono I risultati ottenuti. La figura riporta per tutte le domande il livello di valore sociale percepito della informazione connessa alla domanda, calcolato in accordo ai pesi della Figura 16. I risultati ottenuti indicano che, indipendentemente dal campione di rispondenti coinvolti, pochi aspetti sono in maniera chiara considerati socialmente di valore, forse non sorprendentemente; le risultanze dell'indagine affermano che, al fine di individuare l'ospedale dove ricevere la cura (mediato sulle diverse patologie), le persone considerano rilevante sapere:

- quanto è lontano l'ospedale dalla loro residenza
- se l'ospedale è specializzato per la patologia di interesse
- quanto devono aspettare per essere ricoverati.

Domanda	Livello di valore sociale
Q1. Ospedale più vicino	Molto alto
Q2. Ospedali che curano la patologia	Molto alto
Q3. Percentuale di dimissioni dovute a complicazioni	Molto basso
Q4. Percentuale di secondi ricoveri entro 6 mesi	Molto basso
Q5. Percentuale di dimissioni dovute a complicazioni per medico curante	Molto basso
Q6. Percentuale di secondi ricoveri per medico curante	Molto basso
Q7a. Percentuale di soddisfatti per patologia	Probabilmente basso
Q7b. Percentuale di soddisfatti su tutti i pazienti	Molto basso
Q8a. Percentuale di soddisfatti tra i parenti dei pazienti per patologia	Molto basso
Q8b. Percentuale di soddisfatti su tutti i parenti dei pazienti	Molto basso
Q9. Tempo medio prima del ricovero per patologia	Molto alto

Figura 16 – Valore sociale associato alle domande

E' interessante il risultato relativo alla domanda Q4. La percentuale di secondo ricoveri entro 6 mesi è utilizzata in molte valutazioni oggettive sulla qualità degli ospedali, insieme al dato assoluto sui pazienti curati per la patologia. Le valutazioni oggettive differiscono da quelle basate sulla percezione perché si basano sulla misurazione di fenomeni fisici, e quindi, definito il metodo di misura, arrivano a risultati che prescindono dalle percezioni degli utenti.

Per esempio quando arriviamo in un aeroporto e ci rechiamo alla sala per il recupero bagagli, talvolta troviamo dei display che dicono: in questo aeroporto il tempo medio di consegna dei bagagli è di 15 minuti. Questa è una valutazione oggettiva, basata sulla misurazione di un intervallo temporale. Naturalmente, le valutazioni oggettive, per essere efficaci, devono confrontarsi con un valore di riferimento, ma non sempre questo accade. Ebbene, mentre quell'indicatore è uno tra i più rilevanti per misurare la qualità della cura di un ospedale, risulta di scarso valore nella percezione degli utenti.

6. Dati digitali e declino dei giornali

L'uso delle reti sociali come forma di comunicazione oramai prevalente tra umani sta provocando un progressivo declino delle forme di comunicazione tradizionali, quali i giornali quotidiani [Rusbridger 2018] ricorda quasi con nostalgia i 19 passi che erano necessari nei giornali cartacei per poter trasformare un insieme di testi prodotti dai giornalisti in una edizione cartacea di giornale. E mentre descrive questi 19 passi a un uditorio di giovani millenials, osserva preoccupato e stupito le loro facce perplesse e un po' annoiate e i loro gesti rapidi per scegliere le app sui loro telefoni smart.

Costruire una notizia nei giornali tradizionali era e rimane un processo costoso, che nei giornali seri ha sempre previsto una fase di verifica sulla veridicità della fonte originaria, verifica che richiede un lavoro sul campo, talvolta rischioso; come noto, sono centinaia i giornalisti uccisi in diversi paesi del mondo

per le notizie, inchieste e opinioni da essi espresse in teatri di guerra, in territori controllati da mafie, cartelli della droga, regimi politici illiberali.

Certo ci sono giornali e giornali, ma nelle grandi testate giornalistiche la cura per fornire notizie precise è sempre stata molto attenta. E tuttavia, per il Guardian [Jack 2018], stabilire la verità costa; i giornalisti che hanno appreso il loro mestiere nell'epoca pre-digitale, non sapevano molto di profitti e perdite, né applicavano ciò che viene chiamato il *business model*. In questo ambito, le notizie grezze e non verificate sono gratuite e le notizie verificate molto costose. Questa situazione è aggravata dal fatto che le tecnologie digitali attraverso il fenomeno dei big data e delle reti sociali estendono enormemente per loro natura la capacità di diffondere informazione falsa.

Ciò che sta accadendo è spiegato in modo straordinariamente chiaro da Stefano Quintarelli in un articolo sul Foglio del 13 gennaio 2019. Diamo la parola a lui (il testo si trova sul Web).

"La diffusione di contenuti che esacerbano gli animi, come per esempio le fake news, ha un motivo economico razionale, radicato nelle caratteristiche proprie dei beni immateriali e, in particolare, nel fatto che essi hanno dei costi variabili sostanzialmente nulli. Considerando il lato costi, possiamo infatti fare le seguenti considerazioni generali.

1. *L'attività degli intermediari online, come ad esempio i social network, è caratterizzata da costi variabili sostanzialmente nulli: una volta fatto il sistema e collegato alla rete, non vi sono costi variabili per servire un utente. Inoltre, operando su scala globale, ripagano gli investimenti dei loro sistemi informatici con i proventi da una base di utenza globale.*
2. *Gli editori tradizionali, che svolgono l'importante ruolo di portare informazioni e notizie alla società, hanno un'attività con costi infrastrutturali (per i loro sistemi informatici) che deve essere ripagata con i proventi da una base di utenza solo locale e quindi assai ristretta ed hanno costi variabili non trascurabili derivanti dall'attività giornalistica di produzione di contenuti.*

Venendo al lato ricavi, osserviamo che, per l'editoria tradizionale, i costi di vendita e di pubblicità dovevano coprire, oltre agli ammortamenti, i costi variabili industriali (carta, rotative, distribuzione, ecc). Il pallino della determinazione del prezzo della pubblicità era in mano all'offerta, cioè agli editori che conoscevano il proprio conto economico. Dato che tutti gli editori usavano tecnologie simili, i costi variabili erano confrontabili e, soprattutto, non nulli.

Oggi competono per l'attenzione del cliente non soltanto gli editori tradizionali ma anche tutti quei soggetti a matrice tecnologica che monetizzano l'attenzione del cliente (social network, sistemi di condivisione, ecc.). Questi nuovi "editori", forti di costi variabili pari a zero, possono accettare qualunque prezzo maggiore di zero per la loro pubblicità. I costi variabili nulli consentono cioè all'editore di chiedere all'inserzionista "quanto vuoi darmi?", adottando un meccanismo basato su aste. Nella dimensione immateriale il pallino della formazione del prezzo della pubblicità passa al lato della domanda.

Grazie a tecniche di riconoscimento dell'utente assai accurate, è possibile per un intermediario tecnologico "pedinare" l'utente su pressoché tutti i mezzi online che egli visita. Anche questo pedinamento digitale, come praticamente tutte le attività immateriali, ha un costo variabile nullo. Un inserzionista accorto ottiene il massimo beneficio sfruttando il pedinamento digitale per agganciare un

utente su un mezzo pubblicitario ad alta qualità e ad alto costo, come può essere un "editore tradizionale", e seguendolo con le proprie inserzioni sui mezzi a costo inferiore che costui visiterà. Questo spiega perché se leggiamo articoli sulle Azzorre su una testata blasonata, poi troveremo pubblicità relative alle Azzorre anche sui social network o siti di chat.

Questi intermediari del pedinamento digitale incassano una fetta del budget pubblicitario incidendo sui ricavi disponibili per i mezzi tradizionali: il costo variabile nullo della tracciabilità degli utenti e delle aste in tempo reale per inserzioni sposta quote crescenti di ricavi dagli editori agli intermediari tecnologici. Secondo dati diffusi in un recente forum dello Iab, ***gli intermediari tecnologici (social network, motori di ricerca, pedinatori digitali, ecc.) intercettano oggi il 75 per cento dei ricavi pubblicitari digitali****.*

Un'altra fonte di ricavi importanti per gli editori tradizionali era la vendita in edicola, oggi sotto pressione perché il pubblico dei lettori cartacei si assottiglia, in parte per ragioni anagrafiche e comunque per una aumentata offerta gratuita di informazione online. I contenuti disponibili online infatti proliferano molti ordini di grandezza in più di quanto potesse avvenire quando erano fissati su supporti materiali. Il fatto è che, essendo immateriali, il costo di distribuzione dell'informazione è nullo e l'informazione è prodotta anche da attori non sono interessati a una remunerazione diretta. Think tank, blogger, aziende, centri di ricerca, appassionati, associazioni, sono tutti soggetti che producono contenuti che competono per l'attenzione dell'utente. Un utente che trova una cornucopia di contenuti accessibili gratuitamente avrà una disponibilità a pagare erosa rispetto a quanto non fosse in precedenza. Fatto accentuato, ovviamente, nel pubblico più giovane e più avvezzo all'uso di strumenti digitali.

I ricavi degli editori sono quindi stretti in una morsa: quelli da pubblicità sono sottoposti alla crescente pressione degli intermediari tecnologici e quelli da digital circulation dall'erosione della disponibilità a pagare. *Nel giro di pochi anni i ricavi dell'industria dei produttori professionali di informazione si sono ridotti a una frazione di quelli che, in precedenza, sostenevano le loro strutture di costo. Cosa fanno quindi gli "editori" per monetizzare al massimo l'attenzione degli utenti? Cercano di massimizzare i ricavi pubblicitari aumentando il numero di utenti e loro permanenza sui loro servizi. Quanto maggiore sarà il numero complessivo degli utenti, la loro permanenza e frequenza di visita, tanto maggiore sarà la disponibilità di un "inventario" di esposizioni pubblicitarie da mettere all'asta per cercare di mitigare il calo complessivo dei ricavi.*

La tecnologia aiuta a perseguire questi obiettivi di massimizzazione dell'esposizione: grazie al tracciamento e alla capacità di elaborazione, consente una personalizzazione individuale dei contenuti, fatta su una massa di lettori, con un costo variabile pressoché nullo.

Il Mass media, per come lo abbiamo conosciuto, lascia il posto al Personal Media di massa. Per massimizzare attrazione e permanenza degli utenti si sfruttano meccanismi psicologici noti. E' così che si spiega l'aumento del "click baiting" acchiappa lettori (titoli quali "non immaginerai mai …", "10 cose da fare…"), di sfruttamento della gratificazione istantanea dell'utente con la conferma delle sue idee (con il c.d. "confirmation bias" abilitato dal Personal Media di massa che crea delle filter bubbles) e dei contenuti esacerbanti per stimolare reazioni immediate. Ogni reazione infatti è una occasione per rimettere in circolo i contenuti, esponendoli ad altri utenti, e generando altro inventario pubblicitario.

Fake news, filter bubbles e conversazione pubblica aspra, in una certa misura, sono sempre esistite, ma l'attuale pressione sui ricavi portata dalla dematerializzazione e i conseguenti comportamenti di editori e intermediari, di per sé economicamente ragionevoli, ne determinano una amplificazione senza precedenti.

Nessuno ha ancora trovato un modo che possa contribuire a rimuovere o mitigare gli incentivi economici che determinano questi comportamenti, ma anche considerando gli effetti sulla conversazione pubblica è importante che la ricerca continui."

I giornali stanno tentando di porre freno a questa decadenza tramite le edizioni on line, ma anche per esse valgono molte delle considerazioni di Quintarelli. La cornucopia di cui lui parla sta trovando la sua massima espressione nella comunicazione politica nei messaggi Twitter e nei video su You Tube. Le differenze tra un articolo di giornale, sia cartaceo che on line, e un messaggio Twitter sono mostrate in Figura 17.

Caratteristica	**Articolo su giornale**	**Messaggio Twitter**
Tempo richiesto per la produzione	Alto	Basso
Tempo richiesto per la percezione	Alto	Basso
Numero di utenti potenziale max.	Milioni	Miliardi
Livello di qualità	Alto	Basso
Costo della qualità	Alto	Basso
Livello di qualità	Alto	Basso
Professionalità richiesta	Alta	Bassa
Possibilità di ridiffusione	Bassa	Alta
Livello di emotività	Basso	Alto

Figura 17 – Differenze tra articolo di giornale e messaggio Twitter

Le grandi novità dei messaggi Twitter e in generale delle tecnologie di comunicazione che si sono sviluppate negli ultimi venti anni sta nei seguenti aspetti:

- il cambio di ordine di grandezza (1 a 100, 1 a 1.000, 1 a 10.000) nella quantità di informazione che viene diffusa,
- la velocità con cui questa informazione si diffonde,
- la concentrazione in poche mani di questa informazione,
- il suo carattere digitale e quindi elaborabile, e quindi ri-diffondibile,
- la crescita della diffusione dei soggetti che possono generarla,
- la diffusione dei soggetti che possono acquisirla,
- la riduzione della sua qualità e delle risorse cognitive investite nel generarla,
- la frammentazione e la contestualizzazione con cui viene generata e fruita,
- l'assenza di costi economici nel produrla e nel diffonderla,
- la difficoltà di ricostruirne la storia e capire la sequenza degli eventi e dei soggetti che hanno contribuito a farla arrivare a te.

Tutti questi elementi insieme, aggiunti alla nostra limitata razionalità e capacità di investire risorse cognitive, porta al trionfo della superficialità e di chi la spara più grossa per coprire al meglio il mondo mediatico offerto gratuitamente dalla rete. Chi la spara più grossa ha diversi vantaggi: ha più impatto emotivo, più potenzialità di polarizzazione, ha più difficoltà ad essere contraddetto, perchè ci vuole tempo e argomenti per contraddire, l'esatto opposto del messaggio di pancia, si imprime meglio nella mente del lettore, riesce a guadagnare quel tempo, e quindi ad avere un crescente vantaggio di posizione, che l'altro perde nell'argomentare.

Ha ragione il Presidente americano Trump quando dice che lui arriva con i suoi tweet a 100 milioni di persone mentre il New York Times lo leggono in due milioni. Qualunque cosa scriva, lui intercetta una comunità 50 volte più grande dei lettori del New York Times; non solo, per leggere i suoi tweet ci mettono 50 volte meno tempo che leggere un noioso articolo del New York Times. Questo è quanto. Qualcuno ricorda lo Stefano Satta Flores di “C'eravamo tanto amati”? che perde tempo a rispondere a Mike Bongiorno, approfondisce, precisa, e perde.

Al giornalismo resta il fondamentale compito di fornire al lettore una visione realistica, ragionata e completa di tutti gli elementi di giudizio dei fatti ed eventi che influenzano la vita dei lettori. Questa è però una battaglia molto difficile.

Riferimenti

S. Abiteboul, Miklau, G., Stoyanovich, J. e Weikum, G. - Data, responsibly (dagstuhl seminar 16291). In *Dagstuhl Reports* (Vol. 6, No. 7). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

Agid - Linee guida nazionali per la valorizzazione del patrimonio Informativo Pubblico, 2014.

J. C. Alexander, MD, and Girish P. Joshi Smartphone Application-based Medical Devices: Twenty-first Century Data Democratization or Anarchy? The Open Mind, 2016.

P. Arora - The Bottom of the Data Pyramid: Big Data and the Global South – International Journal of Communication, 2016.

M. Andrejevic - The Big Data Divide. International Journal of Communication N. 8, 2014.

S. Badiee, J. Jütting, Deirdre Appel, Thilo Klein *and* Eric Swanson The role of national statistical systems in the data revolution - Development Co-operation Report 2017 Data for Development - OECD 2017.

E. Baldacci, M. Scannapieco eds. – Istat's Roadmap for the Adoption of Big Data for Official Statistics, 2015.

G. Barcaroli et al. - Using Big Data for Statistical Purposes – conference on Big data, social mining & social indicators, Pisa 2013.

C. Batini et al. - A User Study to Assess the Situated Social Value of Open Data in Healthcare –- HCIST 2015 & Procedia.

C. Batini et al - Assessing Social Value in Open Data Initiatives: A Framework - Future internet, 2014

C. Batini - Conceptual Schemas as a Means to Compare and Measure Open Data Social Value Putting Open Data to the Test of Life: Conceptual Schemas as a Means to Compare and Measure Social Value. SEBD 2015: 12- Springer-Verlag Berlin Heidelberg 2011

F. Beltram, F. Giannotti, D. Pedreschi – Joint Statement on New Economic growth: the role fo science, technology, innovation and infrastructure, Position Paer on Data Science, 2017.

Data and discrimination – Collected Essays, Open Technology institute, 2014.

F. Cabitza, A. Locoro, and C. Batini. "A user study to assess the situated social value of open data in healthcare." Procedia Computer Science 64, 2015

W. Davies - Stati nervosi, come la emotività ha conquistato il mondo, Einaudi, 2019.

K. Desouza, K. Smith – Big Data for social innovation, Stanford Social Innovation Review, 2014

R. Espinosa - Linked Open Data Mining for Democratization of Big Data 2014 IEEE International Conference on Big Data.

S. Fahey - The Democratization of Big Data, 2014.

S. Gangadharan (edito da) - Data and discrimination: collected essays - Open technology institute, 2014

E. Giovannini – Statistics 2.0 – From the Data Revolution to the Next Level of Official Statistics, 2010

E. Giovannini – Undertanding Economic Statistics – An OECD Perspective.

E. Giovannini – Conoscere per Capire, La Lettura del Mulino, 2013

E. Giovannini – Towards a comprehensive global policy agenda: what deos it mean for statistics? – Global conference on a Transformative Agenda, New York 2015.

E. Giovannini - Data for Policy: a Myth or a Must?, 2016.

M. Giugale - Fix Africa's statistics, The World Post, www.huffingtonpost.com/ marcelogiugale/, 2012.

S. Grimmelikhuijsen - A good man but a bad wizard. About the limits and future of transparency of democratic governments - Information Polity 17.3, 2012.

M. Gurdstein - Open Data - Empowering the Empowered? First Monday, 2011.

I. Jack - Breaking News by Alan Rusbridger review – the remaking of journalism and why it matters now, Book of the day, The Guardian, 1 settembre 2018.

S. Jäppinen, Tuuli Toivonen, and Maria Salonen. "Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach." Applied Geography 43 (2013): 13-24.

J. Johnson - From open data to information justice - Ethics and Information Technology 16.4, 2014

J. Johnson - The question of information justice, Communications of the ACM, *59*(3), 2016.

W. Hartzog - Evan Selinger, Big Data in Small Hands, 66 Stan. L. Rev. Online 81 (2013-2014).

E. Letouze, J. Jutting – Official Statistics, Big Data and Human Development, Data-Pop Alliance White Paper Series – March 2015

D. Lyon – Surveillance as social sorting – Routledge 2003.

E. Kalampokis - Combining Social and Government Open Data for Participatory Decision-Making

G. Kelly, G. Mulgan and S. Muers - Creating Public Value: An analytical framework for public service reform.

J. Metcalf e K. Crawford Where are human subjects in Big Data research? The emerging ethics divide – Big Data and Society, 2016.

M. Michael - Toward a manifesto for the 'public understanding of big data'Public Understanding of Science, 2016

M. McCarthy - The big data divide and its consequences, Sociology Compass 10.12, 2016.

K. O'Hara - Transparent government, not transparent citizens: a report on privacy and transparency for the Cabinet Office, 2011.

C. O'Neill - Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 2016.

S. Rolando - La statistica come risorsa. Colloquio con Enrico Giovannini (Istat). Riv.it.Com Pubblica n.41/2010 Rivista italiana di comunicazione pubblica (Franco Angeli editore) n.41/2010

L. Rosenbaum - Bridging the Data-Sharing Divide — Seeing the Devil in the Details, Not the Other Camp, 2017.

A. Rusbridger - Breaking News: The Remaking of Journalism and Why It Matters Now, Canongate, 2018.

A. L. Schmidt, F. Zollo, M. Del Vicario, Alessandro Bessi, Antonio Scala,, Guido Caldarelli, H. Eugene Stanleyd, and Walter Quattrociocchi - Anatomy of news consumption on Facebook – 2016.

B. Solomon, R Bhuvaneswari, P. Rajan, Manjunatha Bhoomi - E-Governance - Or, An Anti- Politics Machine Necessary to Globalize Bangalore? Casum Working Paper, 2007.

The Chief Data officer Guide to Data Democtratization, www.happiestminds.com, 2017

S. Treuhaft - The democratization of data: How the internet is shaping the work of data intermediaries, EconStor, 2016.

G.Viscusi e C. Batini - Information production and social value for public policy: a conceptual modeling perspective - Policy & Internet 8.3, 2016.

G. Viscusi, Marco Castelli, e Carlo Batini - Assessing social value in open data initiatives: a framework - Future Internet 6.3, 2014

M. Wigan, R. Clarcke – big data's big unintended consequences – omputer 2013.

Y. Zheng - Different Spaces for e-Development: What Can We Learn from the Capability Approach?

Capitolo 15 - Etica e Big data

Carlo Batini

1. Introduzione

I dati digitali sono un artefatto che è al tempo stesso tecnologia, servizio, risorsa, rappresentazione del mondo. Le tecnologie dei telefoni mobili, dell'internet delle cose e delle reti sociali nascono e si diffondono con la promessa di rappresentare potenzialmente ogni aspetto del mondo. In tal modo, assistiamo ad una progressiva commistione tra i due mondi dell'analogico e del digitale, che porta a rendere più complessa la definizione della responsabilità etiche e del libero arbitrio della persona e l'influenza che su di esse esercitano la macchina e l'algoritmo.

In tale contesto di pervasività dei dati digitali nella vita delle comunità e degli individui, i ricercatori di diverse discipline hanno iniziato a interrogarsi sulla relazione esistente tra i dati digitali e l'etica.
Per [Wikipedia 2019] l'etica è una branca della filosofia che studia i fondamenti razionali che permettono di assegnare ai comportamenti umani uno status deontologico, ovvero distinguerli in

- buoni, giusti, leciti, rispetto ai comportamenti ritenuti
- ingiusti, illeciti, sconvenienti o cattivi

secondo un ideale modello comportamentale (ad esempio una data morale).

Parlare di etica è per me molto molto rischioso, per la mia immaturità di pensiero in questo campo, eppure penso che questo rischio vada corso; in fondo, se vediamo la nostra azione e il nostro desiderio di organizzare il pensiero anche proiettati al di là della nostra esperienza di vita, cosa resterà della nostra speculazione se non l'impatto, in questo caso, della Scienza dei dati sulla nostra vita, sui nostri comportamenti, su come le future generazioni sapranno fare un uso giusto dei dati digitali?

Il capitolo è organizzato come segue. Nella Sezione 2 riprendiamo e commentiamo le categorie generali che in Wikipedia vengono associate all'etica dei dati digitali. Nella Sezione 3 ci interessiamo dell'etica dei dati come evoluzione dell'etica delle tecnologie digitali, in accordo alla prospettiva filosofica tracciata da Luciano Floridi.

La ricerca sulla relazione tra etica e dati digitali è molto vasta. Eppure, quando si parla di etica nella letteratura sui dati digitali, raramente si parte da una discussione focalizzata direttamente sul concetto di etica, piuttosto si fa riferimento ad altri concetti, che chiamiamo nel seguito *determinanti*, che consistono in proprietà o caratteristiche dell'etica. Nella Sezione 4 affrontiamo il tema dei determinanti dell'etica nell'ambito dei dati digitali.

Ci concentriamo a questo punto su tre dei principali determinanti, la trasparenza, la equità e la interpretabilità, scoprendo che per tutti e tre i temi la ricerca, pur avendo prodotto prime sistematizzazioni, è da considerarsi ancora come un libro aperto, in cui diversi aspetti sono caratterizzati da visioni talvolta anche in grande contrasto tra di loro.

Nella Sezione 5 approfondiamo il tema della trasparenza dei dati digitali, tema peraltro controverso, per cui la Sezione 6 successiva indaga sui rischi connessi alla trasparenza. Nella Sezione 7 discutiamo del concetto di equità del modello derivante dalla analisi, nella Sezione 8 trattiamo il tema della trasparenza del modello (da non confondersi con il concetto di trasparenza dei dati oggetto della Sezione 5, vedi tra poco un chiarimento sui due concetti), che chiameremo con il termine di interpretabilità. La Sezione 9 tratta l'importante novità in ambito Europeo in tema di protezione della privacy, costituita dal Regolamento Europeo chiamato con l'acronimo GDPR. La Sezione 10 affronta il tema di come si possa impostare un progetto che riguardi i dati digitali che rispetti i principi etici, il cosiddetto "ethics by design".

2. L'etica dei dati digitali: categorie generali tratte da Wikipedia

In [Wikipedia 2019] l'etica dei dati (digitali) si riferisce alla sistematizzazione e orientamento dei concetti di giusto e sbagliato in relazione ai dati, e specificamente ai *dati personali*; altre sistematizzazioni, come vedremo nelle sezioni successive, coprono ambiti molto più ampi.

L'etica dei dati è di crescente rilevanza al crescere della quantità dei dati e della vastità dei fenomeni sociali su cui i dati digitali hanno impatto. Essa è vista in [Wikipedia 2019] riguardare i seguenti principi e fenomeni:

1. il possesso dei dati (ownership), per cui *gli individui sono proprietari dei dati che li riguardano*. La protezione dei diritti morali di un individuo è basata sulla visione per cui i dati personali sono una diretta espressione della personalità dell'individuo; i diritti morali sono perciò specifici e riferiti a quell'individuo, e non possono essere trasferiti ad altra persona ad eccezione che tramite testa-mento, quando l'individuo muore. I diritti morali includono il diritto (dell'individuo) ad essere identificato come la fonte dei dati e il diritto di opporsi ad ogni distorsione o manipolazione dei dati che potrebbe essere pregiudizievole alla sua reputazione. Tali diritti morali associati ai dati personali, sono perpetui.
2. la trasparenza di elaborazione (transaction transparency), sulla base della quale se vengono utilizzati dati personali in una elaborazione, algoritmo o tecnica, deve essere garantito l'accesso trasparente all'algoritmo o tecnica utilizzata sui dati.
3. Il consenso, se un individuo o persona giuridica intende utilizzare dei dati personali, è necessario ottenere il consenso informato ed esplicitamente espresso dell'owner dei dati su quali dati personali vengono utilizzati, a chi vengono forniti, e a quale scopo e quando vengono utilizzati.
4. La privacy: se avvengono transazioni sui dati, deve essere compiuto ogni possibile sforzo per preservare la riservatezza.
5. L' apertura: i dati aggregati, non facenti riferimento quindi a dati personali, devono essere liberamente accessibili.
6. Le persone dovrebbero essere messe a conoscenza delle transazioni finanziarie e del loro ammontare in cui è coinvolto l'uso dei loro dati personali.

Secondo una diversa prospettiva possiamo vedere l'emergere di temi etici a partire dal ciclo di vita dei dati di cui abbiamo parlato nel Capitolo 1. La Figura 1, tratta dal Responsible Data Science Program[52], mostra gli elementi critici che nascono da temi etici, con riferimento alla trasparenza (come rendere gli algoritmi adottati e i processi adottati comprensibili a tutti gli utenti coinvolti), la confidenzialità (come proteggere la privacy degli utenti coinvolti), la accuratezza (come produrre modelli aderenti alla realtà, discussa nel Capitolo 5), l'equità (come evitare trattamenti o decisioni discriminatorie), la confidenzialità.

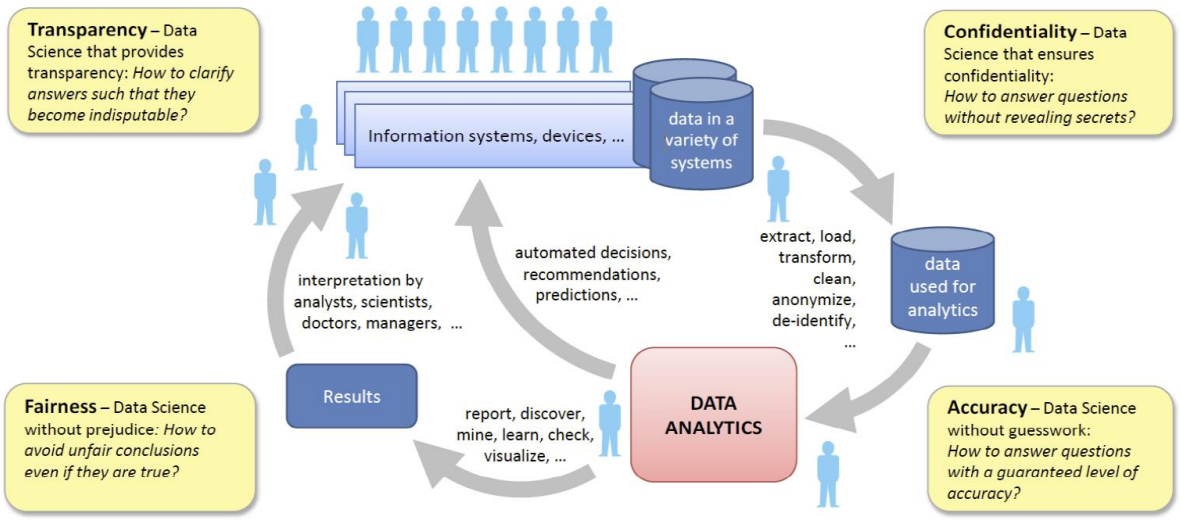


Figura 1 – Il ciclo di vita dei dati e le sfide etiche

3. Etica dei dati e filosofia, l'approccio di Luciano Floridi

Riassumiamo qui uno dei più recenti contributi di Luciano Floridi, dedicato all'etica dei dati. L'etica dei dati è vista nella speculazione di Floridi [Floridi 2016] come una nuova branca dell'etica che studia e valuta i problemi morali legati:

- ai dati nel loro ciclo di vita, dalla acquisizione all'utilizzo da parte dell'utente finale,
- agli algoritmi (includendo l'Intelligenza artificiale, il machine learning, la robotica) che fanno uso dei dati
- alle relative pratiche (includendo l'innovazione responsabile, la pratica dell'hacking, intesa come l'insieme dei metodi, delle tecniche e delle operazioni volte a conoscere, accedere e modificare un sistema informatico hardware o software, i codici professionali)

in modo da formulare soluzioni e comportamenti moralmente validi. Questo significa che le sfide etiche poste dalla Scienza dei dati possono essere messe in corrispondenza con i tre percorsi di ricerca corrispondenti all'etica dei dati, all'etica degli algoritmi, e sull'etica delle pratiche. Prima di esplorare in dettaglio la classificazione espressa nella definizione precedente (dati nel loro ciclo di vita, algoritmi, pratiche) vediamo altri due percorsi della riflessione di Floridi sempre riferiti all'etica dei dati.

[52] https://redasci.org/

Etica delle tecnologie informatiche, della informazione e dei dati

L'etica dei dati si fonda sull'etica delle tecnologie di calcolo (Computer Ethics), ma allo stesso tempo raffina e amplia l'approccio di tale campo di ricerca, spostando l'attenzione delle domande etiche, da quelle che Floridi chiama *centrate sulla informazione* (information centric) a quelle *centrate sui dati* (data-centric). La distinzione è rimarcata da Floridi, nel senso di ampliare la riflessione etica alle dimensioni morali di *tutti i tipi di dati*, anche quando *il dato non completa* il processo di trasformazione in informazione semanticamente "matura" (perché nella pratica quotidiana noi spesso usiamo dati grezzi, opachi, incompleti), e può essere usata per supportare azioni e generare comportamenti.

Ciò mette in evidenza la necessità per le analisi etiche di concentrarsi sul contenuto e la natura delle operazioni computazionali sui dati – le interazioni tra hardware, software e dati – piuttosto che sulla varietà delle tecnologie digitali che le abilitano; ed enfatizza la complessità delle sfide etiche poste dalla Scienza dei dati.

Livelli di astrazione nell'etica in informatica

L'evoluzione intervenuta tra etica delle tecnologie informatiche, della informazione e dei dati viene vista da Floridi anche secondo una diversa luce, e cioè come un passaggio tra livelli di astrazione, tema questo di cui in una prospettiva piuttosto ampia abbiamo parlato nel Capitolo 12. Vediamo i livelli individuati da Floridi.

La ricerca sull'etica nell'informatica ha dapprima adottato un livello di astrazione centrato sugli *esseri umani*, affrontando i problemi posti dalla disseminazione degli elaboratori in termini di responsabilità professionali, sia dei progettisti che degli utenti. A questo punto c'è stata una evoluzione verso un livello di astrazione centrato sugli *elaboratori*, che ha messo in luce la natura degli elaboratori (e del software) come strumenti universali e malleabili, facilitando la comprensione dell'impatto degli elaboratori sulle dinamiche sociali e sul progetto dell'ambiente attorno a noi.

L'adozione dei livelli di astrazione si spostò a questo punto *dalla tecnologia al contenuto*, che nella terminologia di Floridi è *l'informazione*, che può essere creata, memorizzata, scambiata. Furono analizzate le diverse dimensioni morali della informazione, come fonte, risultato, obiettivo delle azioni morali, e ciò portò allo sviluppo di un approccio macro etico sull'intero ciclo di vita della informazione, dalla creazione, alla protezione, all'oblio e distruzione.

La fase finale di questa rivoluzione ha portato in pochi decenni a comprendere che non bisogna più concentrarsi sulle tecnologie, siano esse i tablet, gli smart phone, il cloud computing, le piattaforme on line, ma sui *dati, ciò che ogni tecnologia rappresenta e usa*. Lo spostamento dall'etica della informazione all'etica dei dati è più legato al significato che concettuale, e mette in luce la necessità di concentrarsi su ciò che rappresenta il vero invariante della nostra speculazione etica; non l'hardware, non il software, ma *ciò che le tecnologie producono operando sui dati*.

Torniamo ora alla classificazione insita nella definizione di etica dei dati fornita all'inizio della sezione.

Etica dei dati, degli algoritmi e delle pratiche

L'etica dei dati si focalizza sui problemi etici posti dalla raccolta e analisi di grandi dataset e sui temi legati all'uso dei big data nella ricerca biomedica e nelle scienze sociali, così come negli open data. In quest'ambito le tematiche più rilevanti riguardano la identificazione degli individui tramite tecniche di estrazione di conoscenza dai dati, il collegamento e la integrazione di dataset, così come il rischio di violare accanto alla privacy individuale la "privacy dei gruppi o comunità", portando a rischi di discriminazione di gruppi sulla base ad es. di etnie o abitudini sessuali. Il trust e la trasparenza sono pure temi cruciali nell'etica dei dati. Per esempio la trasparenza è spesso invocata come aspetto che abilita il trust, ma non è ancora chiaro quale informazione dovrebbe essere resa trapsarente e a chi dovrebbe essere divulgata.

L'etica degli algoritmi indaga le questioni poste daslla crescente complessità e autonomia degli algoritmi intesi nel senso più ampio, e quindi comprendenti gli algoritmi del machine learning. In quest'ultimo caso, alcuni sfide cruciali includono la responsabilità morale e la verificabilità (accountability) sia da parte dei progettisti dedgli algortimi rispetto alle conseguenze impreviste e indesiderabili (es. discriminazioni o promozione di contenuto antisociale).

L'etica delle pratiche (includendo l'etica professionale e la deontologia) riguarda le domande pressanti riguardanti le responsabilità delle persone e delle organizzazioni coinvolte nelle attività, strategie e policy riguardanti i dati, con l'obiettivo di definire un quadro di riferimento etico avente lo scopo di definire codici professionali per una innovazione responsabile, che possa assicurare pratiche etiche che abilitino sia il progresso della scienza dei dati che la protezione dei diritti degli individui e dei gruppi. Tre sono i temi centrali in questa linea: la privacy personale, il consenso, e gli utilizzi dei dati successivi all' utilizzo primario (secondary use).

L'etica dei dati, degli algoritmi e delle pratiche sono ovviamente interdipendenti, e questa è la ragione per cui può essere preferibile definire uno spazio concettuale all'interno del quale i problemi etici sono come punti identificati da tre valori sugli assi. Per esempio, le analisi che si focalizzino sulla privacy devono anche fare riferimento a temi riguardanti il consenso e le responsabilità professionali.

Come conclusione, l'etica dei dati deve fare riferimento all'intero spazio concettuale definito in precedenza, e quindi a tutti e tre gli assi, pur se con differenti priorità e focus.

Opportunità e sfide della Scienza dei dati

Per Floridi la Scienza dei dati offre enormi opportunità per migliorare la qualità della vita privata e pubblica, nonché l'ambiente in cui viviamo. Sfortunatamente, tali opportunità sono anche forte-mente in relazione con significative sfide etiche. Le grandi tendenze riguardano:

- L'uso estensivo nella nostra vita dei dati digitali che descrivono frammenti di mondo sempre più estesi, dati che rappresentano sia aspetti personali che collettivi, territoriali, temporali, per decisioni e scelte sempre più importanti nella nostra vita
- L'uso estensivo di algoritmi e tecniche di apprendimento basati sui dati digitali

- La graduale riduzione del coinvolgimento umano a favore di algoritmi basati sull'appren-dimento,

e pongono domande urgenti sui temi della equità di trattamento, della responsabilità personale e collettiva, sul rispetto dei diritti umani.

Non affrontare o sottovalutare i problemi etici posti dalla diffusione dei dati digitali può portare ad un rifiuto sociale sulle grandi questioni. L'accettabilità sociale (Floridi usa anche il termine social preferability, difficilmente traducibile, ma con significato intuitivamente chiaro) deve essere il criterio guida per ogni progetto, (intervento legislativo di regolazione o abilitazione, metodo, modello, parte tra parentesi aggiunta da me) sviluppati nell'ambito della Scienza dei dati che abbia anche un lontano impatto sulla vita umana, per assicurare che le opportunità insite nel progetto non siano perse. Allo stesso tempo, sopravvalutare la protezione dei diritti individuali in contesti sbagliati (wrong contexts) può portare a regolazioni troppo rigide, vanificando le potenzialità insite nel valore sociale della Scienza dei dati; ancora una volta, anche e più ancora nella presente epoca dominata dai dati digitali, il problema più rilevante delle società sta nel conciliare i diritti individuali e i diritti collettivi.

In virtù di tale insieme di complessità, nella visione di Floridi l'etica dei dati dovrebbe essere sviluppata fin dal suo nascere come Macroetica, cioè come un framework complessivo che eviti approcci ad hoc e affronta l'impatto etico e le implicazioni della Scienza dei dati all'interno, appunto, di un framework consistente, olistico e inclusivo.

Ho volutamente descritto la visione espressa in [Floridi 2016] quasi parola per parola per offrire al lettore un punto di vista, quello del filosofo, eccentrico rispetto ai punti di vista espressi nel resto del libro e anche, a mio parere, per arrivare alla conclusione che anche la filosofia, su queste tematiche, e ai suoi primi passi.

4. Determinanti dell'etica

Considerando la letteratura sull'etica dei dati, è facile arrivare alla conclusione che le riflessioni tendono a concentrarsi non su temi generali, ma su aspetti specifici che sono elementi costituenti l'etica dei dati, che chiameremo *determinanti*. Nel seguito propongo una lista di determinanti, accompagnandoli con una breve definizione; occorre osservare che in taluni casi le differenze di significato tra i vari determinanti sono davvero minime, ho tuttavia voluto riportarle, consapevole che siamo solo all'inizio di un cammino che non sarà breve.

1. *Trasparenza dei dati digitali*, intesa come capacità dei dati digitali di descrivere in modo comprensibile il fenomeno (il processo, l'evento, l'organizzazione, una legge) che essi rappresentano.
2. *Interpretabilità o proprietà di esistenza di una spiegazione o trasparenza del modello predittivo, descrittivo, interpretativo, prescrittivo* costruito a partire dai dati digitali, intesa come capacità del modello di far comprendere il percorso computazionale che il modello esprime, in sintesi, *come* il modello opera, così che tutti possano comprendere il processo che ha portato a risolvere in un certo modo quel problema o a prendere quella decisione.

3. *Accessibilità*, esprime la capacità degli utenti di poter fruire dei dati digitali di loro interesse, sia in termini di tecnologie di accesso (ad es. la rete Internet) che in termini di comprensione del loro significato.
4. *Accountability*, ovvero capacità di rispondere dei propri atti, esprime la esistenza e la messa a disposizione di strumenti conoscitivi per identificare chi, a partire dai dati, abbia preso una decisione o abbia effettuato una azione cui è connessa una responsabilità, e le ragioni di tale decisione.
5. *Assunzione di responsabilità*, la accettazione dei costi potenziali e dei doveri connessi alla decisione presa.
6. *Attribuzione di responsabilità*, un processo di comprensione basato sui dati, a partire dal quale sia possibile associare la responsabilità morale a una persona per aver prodotto un comportamento o effetto censurabile. E' molto vicina all'accountability.
7. *Verificabilità*, capacità di esaminare e valutare in profondità comportamenti o azioni riferiti a dati digitali.
8. *Data divide*, ineguaglianza di natura economica, sociale o culturale che dà luogo ad una rappresentazione della realtà, ad un utilizzo, ad uno scambio di dati digitali caratterizzato da asimmetrie riguardo al loro accesso, al loro uso o impatto o valore per l'utente.
9. *Fairness* (equità, Imparzialità, correttezza), lo stato dell'essere vero e del comportarsi, nell'utilizzo e nella comprensione del dato e dei problemi che ha permesso di risolvere o le decisioni prese, *come parte terza*, anche al di là delle propensioni, interpretazioni e sensibilità individuali. Anche: lo stato, condizione o qualità di essere liberi da preconcetti o ingiustizie. Anche: correttezza di una tecnica di learning; cioè, i classificatori o modelli prodotti dalle tecniche di machine learning dovrebbero nelle loro elaborazioni o applicazioni di tecniche essere indipendenti da aspetti sensibili, quali, ad esempio, gli elementi riferibili al sesso, alla etnia, alle convinzioni religiose. Un'aspetto della fairness è la discriminazione statistica, diseguaglianza tra gruppi demografici o etnici basata su stereotipi.
10. *Parità di opportunità (ovvero non discriminazione, egualitarismo),* I dati digitali utilizzati a fini decisionali o nella formulazione di classifiche e selezioni dovrebbero permettere a tutti di competere su base egualitaria.
11. *Generalizzazione vs Personalizzazione*, individuazione di un equilibrio tra la messa a disposizione di dati per tutti, ovvero la personalizzazione e l'adattamento dei dati verso specifiche comunità o individui.
12. *Qualità della informazione*, proprietà delle informazioni di essere corrette, complete, aggiornate, essendo in tal modo aderenti alla realtà. Abbiamo affrontato questo aspetto nel Capitolo 5.
13. *Privacy*, la caratteristica dei dati personali (es. codice fiscale) e sensibili (es. genere, religione) di essere accessibili solo a soggetti autorizzati (es. pubbliche amministrazioni) e per particolari usi.
14. *Condivisione*, che abilita la possibilità di avere in comune i dati digitali e a non considerarli come un bene privato.

E' impossibile in un breve scritto entrare nel dettaglio o anche fornire esempi di tutti i precedenti temi. Il lettore curioso può fare rifermento alla mia presentazione in formato Power Point [Batini 2018] e alla bibliografia al termine di questo testo. Come detto nella introduzione, nel seguito ci concentriamo sulla trasparenza, sulla equità, e sulla interpretabilità.

5. Trasparenza dei dati

Abbiamo poco fa definito la trasparenza dei dati digitali come la capacità dei dati digitali di descrivere in modo comprensibile il fenomeno (il processo, l'evento, l'organizzazione, una legge) che essi rappresentano. Tratteremo dapprima la trasparenza dei dati digitali in generale, traendo spunto da [Turilli e Floridi, 2009], per focalizzare l'attenzione verso la fine della sezione sulla trasparenza dei dati utilizzati negli algoritmi di machine learning.

Tendiamo ad usare in questa prima parte della sezione più che il termine *dato* il termine *informazione*, assicurando il lettore che pur essendo fedeli al termine usato dagli autori, rimaniamo coerenti con i contenuti complessivi della sezione e non facciamo forzature di significato.

Vista dalla prospettiva di coloro che accedono alle informazioni, la trasparenza dipende da fattori come la disponibilità della informazione, le condizioni tecnologiche e normative per la accessibilità, e da come l'informazione che è stata resa trasparente può pragmaticamente o in relazione al corpus della conoscenza scientifica fornire supporto al processo decisionale dell'utente.

Gli information providers (aziende, organizzazioni o istituzioni pubbliche) trattano tali aspetti scegliendo quali informazioni volta in volta potrebbero o dovrebbero essere divulgate, anche in accordo alla legislazione vigente, a decidendo in che forma le informazioni possano essere rese disponibili. Tali scelte sono legate alla valutazione di temi giuridici e vincoli e implicazioni etiche, oltre che, come accade in prevalenza, ad aspetti di business e di mercato.

Ora, la trasparenza della informazione, intesa in termini di informazione divulgata, non implica necessariamente conseguenze etiche, dal momento che la informazione divulgata può essere eticamente neutrale. Dunque, la trasparenza della informazione può avere soltanto effetti non in relazione con l'etica, o anche non avere nessun effetto. Ad esempio, l'interfaccia utente di un sistema operativo (Windows, Linux,ecc.) spesso descrive i processi computazionali sottostanti senza portare a conseguenze etiche. Divulgare tali informazioni è una scelta progettuale, fondamentale per le interazioni funzionali uomo-computer, ma che non si qualifica come una scelta etica.

La trasparenza della informazione può diventare un fattore abilitante o di impedimento, rappresentando così una condizione "proetica", quando la informazione rivelata ha un impatto su principi etici; tale impatto dipende da almeno due tipi di relazioni tra informazioni rese trasparenti e principi etici. La prima è la *dipendenza*: per soddisfare principi etici, è necessario che il processo di trasparenza riguardi qualche aspetto della informazione, ad esempio in riferimento ai temi della sicurezza, al benessere, al consenso informato. L'altro è la *regolazione*: i principi etici regolano i flussi informativi limitando l'accesso e la disseminazione della informazione, ad esempio, quando vi siano esigenze di privacy, o copyright.

La trasparenza della informazione può minare alla base i precedenti principi etici quando dettagli informativi falsi, fuorvianti, parziali o inappropriati sono resi pubblici; ad esempio, informazioni parziali possono minare la privacy e l'anonimità, e possono anche generare false credenze che possono minare a loro volta l'accountability o giustificare un falso senso di sicurezza o benessere.

La trasparenza ha poi un carattere ambivalente, può essere a seconda dei casi eticamente abilitante o controproducente, ponendo un problema generale rispetto al modo migliore per decidere quali informazioni dovrebbero essere rivelate quando sono presi in considerazione aspetti etici.

Sfortunatamente, non esiste nessun modo semplice per assicurarsi che la trasparenza della informazione sia garantita massimizzando le sue caratteristiche "ethically enabling"; quale informazione diffondere deve essere deciso con attenzione, valutando le conseguenze etiche caso per caso; In conclusione, per [Turilli e Floridi, 2009] non abbiamo nessuna certezza e nessuna verità assoluta, dobbiamo volta a volta cercarla, affinando nel contempo gli strumenti a noi disponibili.

Passando brevemente alla trasparenza dei dati utilizzati negli algoritmi di machine learning, che cos'è in questo caso la trasparenza dei dati e come possiamo raggiungerla? Un'interpretazione nel contesto dell'analisi predittiva include "rendere i dataset di training e di validazione disponibili pubblicamente". Però, mentre è ragionevole che i dati siano resi aperti ogni volta che sia possibile, i dati rappresentano spesso proprietà sensibili che non possono essere condivise. Alla luce di ciò, oltre a rilasciare i dataset di training e di validazione ogni volta che sia possibile, i fornitori dovrebbero rendere disponibili aggregazioni statistiche sui dati che possono aiutare a interpretare le decisioni prese utilizzando i dati, applicando metodi allo stato dell'arte per preservare la privacy, come, ad esempio, la privacy differenziale [Dwork and Roth 2014]. Non approfondiamo ulteriormente questa problematica.

6. Problemi con la trasparenza

La trasparenza è uno di quei concetti che sembrano universali; chi può dichiararsi contro la trasparenza? Nella letteratura sono, al contrario, molti i contributi critici.

Un primo contributo che mi sembra utile approfondire riguarda il progetto citato nel Capitolo 14 relativo alla iniziativa nello stato indiano del Karnatka sulla compravendita di terreni, vedi per approfondimenti il testo [Raman 2012], il cui titolo "The rhetoric of transparency and its reality: Transparent territories, opaque power and empowerment" fa comprendere bene le tesi discusse nell'articolo.

[Raman 2012] analizza due iniziative intraprese dallo Stato indiano nell'ambito del programma di Open Government, il cui scopo è stato quello di rendere disponibili i dati posseduto dallo Stato in formato aperto e elaborabile: la digitalizzazione delle informazioni spaziali, e il Right of Information Act, mediante il quale ogni cittadino indiano può chiedere l'accesso ai dati non di natura sensibile posseduti dallo Stato.

Riguardo alla prima iniziativa, lo Stato indiano ha investito molti fondi per produrre una base dati centralizzata per i dati spaziali e per le proprietà terriere; tuttavia, mentre la trasparenza delle informazioni spaziali viene utilizzata come giustificazione dallo Stato per progetti come quello discusso nel precedente capitolo, che come abbiamo visto hanno presentato gravi criticità, ci sono anche altre logiche politiche ed economiche alla base della loro introduzione.

Piuttosto che assumere acriticamente gli effetti benefici della trasparenza, è fondamentale esaminare non solo i risultati dell'apertura di diversi tipi di dati, ma anche porsi domande su come e perché sono stati aperti i dati governativi; la tesi espressa da [Raman 2012] è che la celebrazione indiscussa dell'open Government e il paradigma della trasparenza possono servire da strumento per ridurre le rivendicazioni politiche a questioni tecno-gestionali, e quindi realizzare un'agenda "anti-politica" per contrastare le contestazioni sui territori urbani.

Il processo che ha portato alla realizzazione della base di dati centralizzata di dati spaziali sulla proprietà dei terreni, non ha affrontato il problema fondamentale della costruzione di una base dei dati fondiari accurato, unificato e coerente, e rischia, al contrario, di cancellare tutta la rete di accordi informali, storie, sfumature, ricordi che sono analogicamente rappresentati nella memoria delle persone e nelle comunità, e non trovano spazio nella "fredda" e strutturata rappresentazione fornita dalla base dati. I dati rappresentano bene stati di cose, ma non le storie, le contese, i patti tra persone e tra gruppi sociali. La difficoltà nella costruzione di una base di dati deriva dunque dalle complessità implicate nell'accurata acquisizione e rappresentazione di documenti informali, non strutturati, deteriorati, e dalle difficoltà insite nel ricostruire la memoria storica sulle forme di possesso della terra o sul funzionamento della città, e nel rappresentare gli aspetti informali.

Inoltre, la trasparenza ha potenziali conseguenze impreviste e non intenzionali su vasta scala che possono influire sulla società e sull'economia. In particolare, i dati resi disponibili invocando fini di trasparenza (pensiamo per esempio alle intercettazioni telefoniche rese pubbliche nonostante il segreto sulle indagini, anche per le parti che non sono direttamente attinenti alle indagini), sono spesso utilizzati in modo distorto per dare sostanza di prova a presunti comportamenti non leciti, con il rischio che la trasparenza incentrata sulla attribuzione di una colpa possa comportare il rischio di una evoluzione dei media e di gruppi organizzati verso una "democrazia inquisitrice".

Un altro rischio è legato al rilascio involontario di informazioni. Quando WikiLeaks ha rilasciato informazioni che contenevano i nomi di soldati e informatori afgani, ha messo seriamente a rischio la vita delle persone; ciò, come sappiamo, suscita importanti problemi di sicurezza e privacy. Il rilascio involontario può riguardare il processo decisionale, e coloro tra funzionari e politici che sono coinvolti nel processo decisionale e che hanno il sospetto di essere intercettati; in tale condizione, tali soggetti possono decidere di agire in modo difensivo e censurarsi, dando luogo ad un processo decisionale di scarsa qualità.

Un episodio che vale la pena ricordare per confermare le precedenti tesi è la famosa diretta streaming tra Bersani e gli esponenti del movimento 5Stelle nell'anno 2013, per verificare la possibilità di un governo tra il Partito democratico e il Movimento 5stelle. Indipendentemente dalle posizioni politiche di ciascuno di noi, penso di poter affermare che certamente l'esporre pubblicamente in diretta le proprie opinioni e le offerte di collaborazione, ha censurato la discussione certamente nel caso di Bersani, che ha esposto in termini molto generali e astratti la propria visione della possibile collaborazione, in tal modo accentuando probabilmente nella visione degli esponenti del Movimento la presunta vaghezza delle proposte lanciate. Insomma, la trasparenza in quel caso ha significativamente favorito il fallimento della trattativa, rispetto al metodo usuale di incontri riservati

che permettono un linguaggio più espressivo delle reali intenzioni di collaborazione e una maggiore capacità di arrivare a posizioni di equilibrio e accordi.

La trasparenza involontaria può non solo causare, come ho argomentato, una qualità inferiore del processo decisionale, ma può anche portare a ulteriore erosione dell'accettazione delle decisioni. Una persona con cui ho lavorato all'Aipa e che ora non c'è più osservò una volta che nelle decisioni si sbaglia due volte su tre. Se le persone possono vedere tutti gli errori dietro le quinte del governo, al di là di quelli già resi visibili dalla azione politica palese (i decreti e le leggi), possono progressiva-mente e ulteriormente disilludersi, erodendo così la fiducia e la legittimità dei governi. Alla fine questo potrebbe diventare un processo a spirale in cui un processo decisionale sempre più conformista e difensivo porta a minore dissenso, con una qualità complessiva della azione politica, lato governo e lato opinione pubblica, sempre più ridotta.

Esiste poi nel fenomeno dell'Open Government un altro rischio di spostamento della attenzione su aspetti marginali della azione di governo, a scapito dei problemi sostanziali. Questo accade quando le amministrazioni arricchiscono i propri siti di dati di "facciata", informazioni di descrizione della normativa, ovvero di cattiva informazione, che nasce da intenzionale distorsione dei dati nel loro significato originario, oppure da filtro su quali dati pubblicare e quali no.

[Lessig 2002] si concentra su casi negli Stati Uniti relativi al tema dei finanziamenti delle corporazioni ai membri del Congresso, sostenendo che "tutti i dati nel mondo non ci diranno se un determinato contributo ha prodotto un risultato, assicurando un voto o un atto che altrimenti non si sarebbe verificato. Il massimo che è possibile dire è che laddove un membro del Congresso agisca in modo incompatibile con i suoi principi o i suoi elettori, ma coerente con un contributo significativo, tale atto solleva almeno una domanda sull'integrità della decisione. Ma al di là di una domanda, i dati dicono poco altro". Ritroviamo qui un tema che ho trattato nel Capitolo 5 a proposito delle fake news; entre su quel tema arrivavo alla conclusione che la verità non è mai raggiungibile nella sua interezza, ma che dobbiamo fare ogni sforzo per indagare e per costruire se non la verità almeno frammenti di verità, qui [Lessig 2002] arriva a una conclusione pessimistica, del tipo "sì, possiamo farci una opinione ma nulla di più, possiamo arrivare a conclusioni generali, ma non alla certezza in casi specifici".

[Lessig 2002] riprende anche l'argomento che ho discusso nel Capitolo 5 in merito ai nostri limiti cognitivi nello scoprire dati manipolati o artefatti. Capire qualcosa - un saggio, una discussione, una prova di innocenza - richiede una certa attenzione. Ma su molte questioni, la quantità media o persino razionale di attenzione che noi decidiamo di investire nella comprensione delle implicazioni diffamatorie è sempre inferiore alla quantità di tempo richiesta. Il risultato finale è quello della creazione di una zona grigia, in cui il problema diventa riuscire a capire in quali casi si potrà arrivare a una comprensione completa di fatti ed eventi, e in quali casi no; questa è una delle chiavi per capire cosa ci sia di sbagliato nella "tirannia della trasparenza".

Riguardo ai sottili legami tra trasparenza e privacy, su cui si potrebbe scrivere un intero capitolo di questo libro, e su cui sono stati scritti interi libri, riporterò semplicemente un passo di [Lessig 2002] tratto da un articolo di Peter Lewis in un numero del New York Times del 1998: "Le telecamere di sorveglianza hanno seguito l'attraente giovane donna bionda attraverso l'atrio dell'hotel nel centro di

Manhattan, l'hanno tenuta d'occhio mentre saliva sull'ascensore fino al 23° piano e scrutava discretamente il corridoio mentre bussava alla porta della mia stanza. Non ho visto le videocassette, ma posso immaginare la lettura digitale sovrapposta alle scene, notando l'ora esatta dell'incontro. Sarebbe stato utile se qualcuno in seguito si fosse chiesto perché questa donna, che non era mia moglie, stava visitando la mia camera d'albergo durante un recente viaggio d'affari. Le telecamere in seguito ci hanno visto dirigerci a cena e a teatro - un uomo di mezza età, sposato proveniente dal Texas, con un braccio intorno a una graziosa donna dell'East Village abbastanza giovane da essere sua figlia…. È un dato di fatto, era mia figlia".

7. Equità (Fairness)

La Figura 2 mostra un insieme di classificazioni prodotte da Google Photos, in cui due persone africane vengono classificati come gorilla. Questo può essere visto come un errore da parte dell'algoritmo, che non è riuscito a discriminare esseri umani rispetto a scimmie, ovvero un errore nei dati di training, tra i quali potrebbero esserci un numero insufficiente di esempi che esprimono questa distinzione. In ogni caso, sia nel caso di deliberata scelta nella classificazione, sia nel caso che derivi da omissioni progettuali o nei dati, certamente possiamo dire che il modello di classificazione non è equo nei confronti di persone proveniente dall'Africa.

Molteplici sono gli esempi di algoritmi iniqui. [Guidotti 2018] riporta una previsione effettuata nel 2017 da Gartner per cui nel 2018 almeno metà delle violazioni etiche nelle attività delle aziende sarebbero dovute alle analisi effettuate su big data. Un altro esempio riguarda i modelli di predictive policing, attività che viene definita in [Ensing 2017] come la attività il cui scopo consiste nella allocazione ottima degli agenti incaricati di prevenire i reati.

Questi modelli, sulla base di dati storici sui reati, su chi li ha compiuti, sui luoghi e sull'ora del giorno in cui sono stati perpetrati, forniscono una previsione sul luogo, sul momento della giornata e talvolta su chi in futuro saranno soggetto di un reato nel territorio di una città; questi modelli sono in uso in molte città americane.

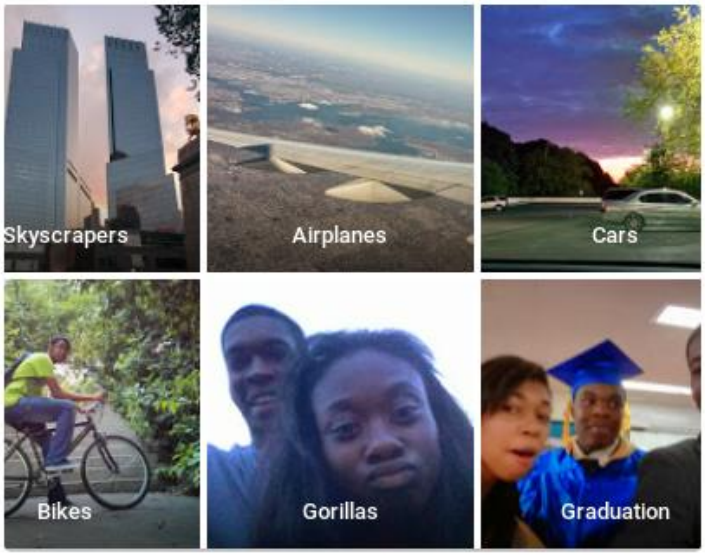


Figura 2 – Diverse classificazioni prodotte la Goolge Photos
(tratte da https://www.google.com/photos/about/)

ProPublica (https://www.propublica.org/), un sito di giornalismo investigativo che sulla base di una analisi di un programma di predictive policing utilizzato negli Stati Uniti, vedi Figura 3, fornisce un confronto tra previsioni di rischio di commettere reati per bianchi e afro americani e effettiva successiva recidiva, che mostra un chiaro sbilanciamento a favore dei soggetti bianchi.

Risultato del Modello predittivo	**Bianchi**	**Afro Americani**
Valutato come rischio alto, ma senza una successiva recidiva	23,5	44,9
Valutato come rischio basso, ma con successiva recidiva	47,7	28,0

Figura 3 – Differenti previsioni e effettiva recidiva per bianchi e afro americani
per una tecnica di predictive policing utilizzata negli Stati Uniti.

I precedenti esempi chiariscono che tutti siamo d'accordo che nelle decisioni o previsioni, siano esse effettuate da un essere umano o da un algoritmo, sia necessaria imparzialità o equità (in seguito equità, in inglese fairness), ma è poi difficile mettersi d'accordo su una definizione non ambigua di equità.

Definizioni quantitative di ciò che è equo e ciò che è iniquo o discriminante sono state introdotte in diversi processi decisionali e discipline per oltre 50 anni, tra esse nella formazione e i processi di selezione del personale (processi decisionali) e nell'apprendimento automatico (disciplina). [Hutchinson 2018] esplora la storia delle definizioni quantitative del concetto di equità, indagando sulla cultura e il contesto sociale in cui sono state prodotte. In alcuni casi, le precedenti definizioni di equità sono simili o identiche alle definizioni di accuratezza nel Machine learning, in altri casi, le definizioni su cosa significa equità e come misurarla sono cadute nell'oblio.

Approfondendo il concetto di discriminazione, l'opposto della equità, il termine ha una etimologia dal latino, che fa riferimento al concetto del *distinguere*. Tuttavia, mentre distinguere non è di per sè una

azione negativa, il concetto di discriminazione applicato alle persone presenta intrin-secamente una connotazione negativa, facendo riferimento a un trattamento oppositivo e ostile verso singole persone o gruppi di persone, basati su caratteristiche diverse rispetto al merito personale.

La legge sui diritti civili degli Stati Uniti del 1964 mise al bando le discriminazioni sulla base della razza, del colore, la religione, il sesso o origine nazionale di un individuo. La legge conteneva due importanti disposizioni che esprimevano la comprensione della comunità dei cittadini su cosa significava essere ingiusto: il titolo VI, che impediva alle agenzie governative (comprese le università) di ricevere fondi federali che discriminavano in base alla razza, colore o origine nazionale; e il titolo VII, che impediva ai datori di lavoro con 15 o più dipendenti di discriminare nei rapporti di lavoro in base alla razza, colore, religione, sesso o origine nazionale.

L'attenzione delle società civili verso la prevenzione della discriminazione è significativamente cresciuta negli ultimi anni, vedi [Zliobaite 2015]; le legislazioni nazionali e internazionali anti-discriminazione si stanno progressivamente estendendo riguardo all'ambito in cui contrastare la discriminazione.

Il machine learning "discrimination-aware" (consapevole riguardo alla discriminazione) è una disciplina emergente che studia come prevenire la discriminazione potenziale derivante dagli algoritmi. Assume a priori che le leggi nazionali e internazionali contro la discriminazione prescrivano quali caratteristiche personali siano considerate sensibili (useremo anche il termine protette) ovvero quali gruppi debbano essere protetti dalla discriminazione. Lo scopo della ricerca in questo campo è quello di tradurre le leggi anti discriminazione in vincoli non discriminatori, e quello consistente nello sviluppare modelli predittivi che tengano in conto tali vincoli e allo stesso tempo risultino essere il più accurati possibile.

Cosa significa che un algoritmo è equo o non discriminatorio? Diversi lavori usano nozioni diverse di equità algoritmica coerenti internamente, ma che appaiono tra di loro incompatibili. La letteratura sulla equità è molto vasta, e da alcuni anni viene svolta annualmente una conferenza denominata Conference on Fariness, Accountability and Transparency (FAT*).

Partendo da queste premesse, e facendo riferimento, per non rimanere troppo astratti, invece che a dati in generale ai dati sulle persone la non discriminazione di un modello di machine learning può essere definita in termini generali nel seguente modo. Definiamo anzitutto il concetto di caratteristica protetta, una caratteristica fondamentale che identifica una comunità di persone, come ad esempio la razza, la religione, l'etnia, la lingua, la disabilità, il genere, l'orientamento sessuale etc. La non discriminazione afferma che

1. persone che siano classificate simili in termini di caratteristiche non protette dovrebbero essere oggetto di classificazioni simili, e
2. le differenze nelle predizioni tra gruppi di persone possono essere caratterizzate da una ampiezza della differenza, ma questa ampiezza deve essere giustificata da caratteristiche *non protette*.

La prima condizione fa riferimento alla discriminazione diretta, e può essere esemplificata dal cosiddetto test dei gemelli; se il genere è l'attribuito protetto e i due gemelli hanno le stesse caratteristiche, eccetto il genere, dovrebbero avere classificazioni simili.

La seconda condizione assicura che non vi sia *discriminazione indiretta*, che ad es. nel caso delle banche è chiamata in inglese *redlining*, o rifiuto di un prestito o fornitura di servizi a prezzi più alti; ad esempio, si è visto che negli US le banche hanno usato il redlining verso clienti che abitavano in quartieri popolati per la maggioranza da popolazione non-bianca. Anche se abitanti "gemelli" nello stesso quartiere sono trattati allo stesso modo, vi è discriminazione per abitanti gemelli di quartieri diversi. La differenza di trattamento può avere una ampiezza che sia giustificata esclusivamente da caratteristiche non protette, e questo è ciò che dice la seconda regola.

Ma cosa significa "classificazioni simili"? Per poter definire il termine in maniera più precisa abbiamo bisogno di introdurre un po' di terminologia, basata sul concetto di qualità predittiva di un modello. Ricordiamo dal Capitolo 10 che un *modello* predittivo di learning supervisionato è organizzato in termini di:

- un insieme di variabili di input
- una variabile di output, che fornisce attraverso i suoi valori la classificazione predittiva
- un insieme di dati su cui è nota la relazione tra dati di input e dati di output (detti dati di training)
- una tecnica (nel seguito algoritmo) che apprende dai dati di training una legge generale che lega nell'intero universo osservato i dati di input a quelli di output.

Assumiamo per semplicità che lo scopo del modello sia di fornire una classificazione binaria, positiva o negativa, ad esempio una persona lascia una compagnia telefonica ovvero rimane cliente (uno dei problemi di classificazione affrontati nei capitoli precedenti). Definiamo un dato classificato come:

- vero positivo, quando il dato è classificato positivo sia nella predizione che nella realtà
- vero negativo, quando il dato è classificato negativo sia nella predizione che nella realtà
- falso positivo, quando il dato è classificato positivo nella predizione e negativo nella realtà
- falso negativo, quando il dato è classificato negativo nella predizione e positivo nella realtà.

Un modello "perfetto" non ha falsi positivi e falsi negativi; sono definite misure di *qualità* di modelli di classificazione, chiamate di *precision* e *recall* che misurano gli scostamenti in termini di veri posi-tivi e veri negativi rispetto a quello che abbiamo chiamato modello perfetto. Come conseguenza, le diverse definizioni di equità si basano su diverse applicazioni delle metriche di qualità, e sui diversi modi in cui possono essere prese in considerazione le caratteristiche protette e non protette. Quindi la equità o non equità della classificazione può dipendere da:

- la scelta dei dati di training
- la scelta dell'algoritmo.

Riguardo a tali diverse definizioni, [Friedler 2018] presenta un formalismo matematico in cui sono confrontate venti (!) differenti definizioni di equità. Oltre a caratterizzare lo spazio degli input (lo spazio "osservato") e degli output (lo spazio "delle decisioni"), [Friedler 2018] introduce la nozione di *spazio costruttivo*: uno spazio che permette di catturare variabili non osservabili, ma significative per la previsione. Inoltre, mostra che per provare proprietà desiderabili del processo decisionale, le differenti definizioni di equità richiedono differenti assunzioni sulla natura delle corrispondenze tra lo spazio costruttivo allo spazio decisionale.

Nell'approccio di [Zlobiaite 2017] la ricerca sulla equità deve essere necessariamente interdisciplinare (vedi Figura 4), e deve riguardare insieme le scienze giuridiche, le scienze sociali, e la informatica. Le scienze giuridiche aiutano a definire il perimetro dei requisiti anti discriminazione, mentre il ruolo delle scienze sociali è quello di definire una giusta allocazione delle risorse negli interventi anti discriminazione; infine l'informatica deve sviluppare le tecniche per la analisi dei modelli di machine learning. Le linee continue mostrano interazioni interdisciplinari, mentre le linee tratteggiate mostrano gli obiettivi possibili.

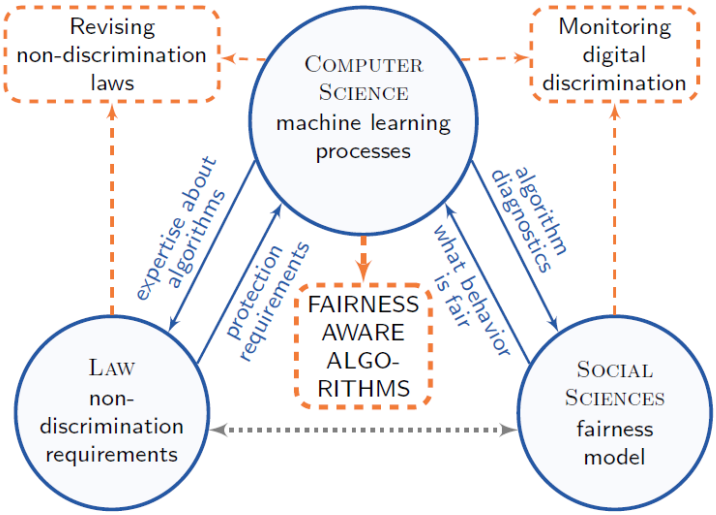


Figura 4 – Interazioni tra scienze giuridiche, sociali e informatiche
in tema di discriminazione/equità (da [Zlobiaite 2017])

Infine, [Abiteboul 2018] arriva alla conclusione che le ricerche svolte negli anni recenti dimostrano che definizioni differenti di equità non possono essere garantite contemporaneamente, e che in genere devono essere definiti dei trade-off; ciò è in certo senso simile a quanto abbiamo visto nel Capitolo 6 sulla fusione; questo non è di per sè un risultato negativo. L'equità è un concetto che racchiude in sé aspetti soggettivi (equo è ciò che conviene a me; diciamoci la verità, noi professori universitari, quando in un concorso vince un nostro allievo, pensiamo che la commissione ha lavorato in modo equo, quando il nostro allievo non vince, tendiamo a pensare l'opposto), dipendente dal contesto e dallo scopo, ed è influenzato da posizioni ideologiche, politiche e utilitaristiche. Un consenso globale su ciò che è equo e ciò che non lo è non è all'orizzonte della ricerca. La proposta di [Abiteboul 2018] è quella di sviluppare tecniche che investigano le differenti definizioni di fairness, e i trade-off esistenti tra di esse.

8. Proprietà di esistenza di una spiegazione, o interpretabilità

"L'ha detto il computer"; diverse volte mi è capitato di sentirmi rispondere in questo modo in un ufficio quando chiedevo spiegazione del perché di un certo risultato di una elaborazione o notifica, ad esempio quella volta che da una nota azienda fornitrice di elettricità mi arrivò una bolletta di circa 850 euro per due mesi di utenza; occorre peraltro dire che negli ultimi anni c'è stato un significativo miglioramento,

ad esempio, nelle spiegazioni che vengono fornite sulle bollette. In questa sezione indaghiamo il tema della interpretabilità nelle tecniche di machine learning, intesa come esistenza di una spiegazione del modello fornito dalla tecnica e dei risultati della classificazione. Cerchiamo di capire il concetto di interpretabilità partendo da un esempio semplice.

Anche se non sapete nulla di linguaggi programmativi, provate a cercare di capire "cosa" calcolano i seguenti due programmi

Programma 1

```
SOMMA = 100  * (100  +  1) / 2
TERMINA
```

Programma 2

```
SOMMA = 0
PER TUTTI I VALORI DI I DAL VALORE 1 ALM VALORE 100 ESEGUI
SOMMA = SOMMA + I
I = I + 1
TERMINA
```

Ebbene, il Programma 1 calcola con una sola istruzione il valore che corrisponde alla somma dei primi 100 numero interi. Se non siete convinti, provate a verificare la validità della formula per i primi 5 numeri: SOMMA = 1 + 2 + 3 + 4 + 5 + 6 = 21 = 30 = (6 x 7) / 2

Il secondo programma calcola lo stesso numero, ma in modo più complicato, calcolando prima 0 + 1, poi incrementando la variabile I di 1, aggiungendo a SOMMA il valore 2, e così via, fino a 100. Certamente, invece che dover capire da soli il calcolo effettuato dal programma, sarebbe utile che all'inizio dei due programmi comparisse una frase, detta *commento*, così concepita:

COMMENTO – IL PROGRAMMA CALCOLA LA SOMMA DEI PRIMI CENTO NUMERI
Il precedente commento chiarisce *cosa* fa il programma, non *come* lo fa. Certe volte, può essere utile conoscere, oltre al cosa, il come, ad esempio per poter valutare la efficienza dell'algoritmo, misurata dal numero di istruzioni eseguite. Possiamo perciò aggiungere al Programma 1 il com-mento:

LA SOMMA E' CALCOLATA IN BASE AD UNA SEMPLICE FORMULA MATEMATICA CHE LEGA IL NUMERO N ALLA SOMMA DEI PRIMI N NUMERI INTERI

e al programma 2 il commento:
LA SOMMA E' CALCOLATA SOMMANDO AL VALORE 0 SUCCESSIVAMENTE I VALORI 1, 2, …, FINO A 100.

E’ chiaro che il secondo programma esegue più istruzioni del primo; il primo programma ha anche la caratteristica di essere *scalabile*, intendendo che il numero di istruzioni eseguite per sommare i primi 1.000, 10.000, ecc. numeri è sempre lo stesso, mentre invece il numero di istruzioni cresce linearmente nel secondo programma.

Il problema di rendere gli algoritmi comprensibili agli esseri umani, già presente nella informatica da quando esiste il software, diventa ancor più critico nell’epoca del Machine learning, perché ora l’algoritmo che prima veniva concepito e prodotto da esseri umani, ora è autonomamente prodotto dalla stessa tecnica di apprendimento, come già osservato nella sezione precedente con riferimento alla equità.

[Guidotti 2018] osserva che facendo affidamento su sofisticati modelli di Machine learning automatico grazie a infrastrutture scalabili e ad alte prestazioni, rischiamo di creare e usare sistemi decisionali che non comprendiamo pienamente. Questo non ha impatto solo sull'etica, ma anche sulla sicurezza e sulla responsabilità delle aziende. Il grande rischio che corriamo è quello che [Pasquale 2015] chiama la “black box society”, una società in cui le decisioni sono prese basandosi su analisi e modelli previsionali i cui meccanismi di funzionamento sono noti solo a pochi, e certe volte neanche ad essi, ma solo a chi li ha concepiti, soggetti che per ragioni di concorrenza o sicurezza, non intendono o non vogliono condividerli.

Prima di entrare nel merito sulle tecniche che abilitano la interpretabilità, occorre chiedersi anzitutto, seguendo [Guidotti 2018], cosa si intenda con interpretabilità e con i collegati concetti di esistenza di una spiegazione (o spiegabilità) e comprensibilità.

[Burrel 2016] introduce tre tipi di *opacità*, l’opposto della intepretabilità. La prima forma di opacità è la deliberata forma di auto-protezione da parte del proprietario del dato con il fine di proteggere segreti commerciali per vantaggio competitivo. Ben note alternative ai modelli cosiddetti proprietari sono il software open source, e i dati open. Non ci occuperemo nel seguito di questa forma di opacità.

La seconda forma di opacità è quella introdotta in precedenza con gli esempi dei programmi; scrivere programmi è una competenza caratteristica di programmatori, che hanno competenze specializzate. La comprensione dei programmi resta inaccessibile alla maggioranza degli utenti. Le metodologie della Ingegneria del software (software engineering) enfatizzano l’importanza della scrittura di programmi chiari, eleganti e comprensibili. Per contrastare questa forma di opacità, citiamo il movimento che a livello mondiale porta avanti l’idea di insegnare il pensiero computazionale a tutti i livelli della formazione: pensiamo solo alla importanza del pensiero computazionale nella professione del giornalista, per il quale le tradizionali tecniche di indagine sono sempre più spesso arricchite o sostituite da un lavoro di ricerca di dati descrittivi di fatti sul Web, dati che, come abbiamo notato nel Capitolo 5, sono spesso per loro natura non verificati e opachi.

La terza forma di opacità è quella che ci interessa più da vicino, e riguarda i modelli abilitati dal machine learning. In questo caso la opacità non riguarda solo la comprensibilità del programma o del collegato algoritmo, ma, piuttosto, l’essere capaci di comprendere l’algoritmo in azione, mentre apprende dai dati il modello di learning. Anche se si possono concepire tecniche di machine learning facilmente

comprensibili, è difficile che la comprensibilità si concili con la utilità; modelli costruiti con tecniche di apprendimento che siano effettivamente utili, ed in particolare, accurati, nel senso introdotto in precedenza) posseggono un grado di inevitabile complessità.

L'interpretabilità si occupa di rendere esplicite le interazioni tra la tecnica di apprendimento e i dati su cui essa opera; essa è rilevante sia quando un modello decisionale è investigato per scoprire discriminazioni sistematiche, sia quando si vuole spiegare una decisione che riguarda un singolo individuo. Supponiamo per esempio che un modello decisionale produca una graduatoria per accedere a un servizio. Se un individuo inserisce i suoi dati e riceve come risultato un punteggio, questo numero da solo non fornisce alcuna informazione sul perché sia stato assegnato tale punteggio e sul perché della posizione comparativa rispetto agli altri partecipanti.

Interpretare significa dare o fornire il significato o spiegare e presentare in termini comprensibili dei concetti. Pertanto, nel nel machine learning l'interpretabilità è la capacità di spiegare o fornire significato in termini comprensibili per un essere umano; in sostanza, una spiegazione è una specie di contratto tra esseri umani e un decisore, spiegazione che è allo stesso tempo una appros-simazione sia accurata dell'azione svolta dal decisore che comprensibile agli umani.

Una importante caratteristica della interpretabilità, come già ho avuto modo di dire, è la *comprensibilità*, che possiamo definire come lo sforzo cognitivo necessario all'essere umano per interpretare il modello di apprendimento. Potrebbe accadere che siamo riusciti a fornire una spiegazione che rende il modello di apprendimento interpretabile, ma che questa spiegazione sia troppo complessa per essere compresa dall'umano.

Una ulteriore caratterizzazione che dobbiamo fare sulla interpretabilità è la distinzione tra interpretabilità globale e locale; un modello è *globalmente interpretabile* se siamo in grado di comprendere la logica complessiva del modello e seguire l'intero ragionamento che porta ai differenti possibili risultati della classificazione, è *localmente interpretabile* se siamo in grado di comprendere le motivazioni per una specifica decisione/predizione.

Allo stato dell'arte, i modelli considerati interpretabili sono tre: gli *alberi di decisione*, i *sistemi a regole*, i *modelli lineari*. Consideriamo le prime due. Gli alberi di decisione sono stati introdotti nel Capitolo 10, ad esso rimandiamo per le definizioni e gli esempi; anche in virtù della rappresentazione grafica (un grafico spiega più di mille parole…) possiamo convenire che essi sono una tecnica comprensibile anche senza molte conoscenze di Machine learning. I sistemi a regole esprimono il procedimento decisionale/classificatorio per mezzo di formule logiche del tipo:

se la febbre è superiore a 38 *and* la gola è rossa *and* il paziente starnutisce frequentemente allora il paziente ha una influenza

Quando applichiamo un sistema a regole, dobbiamo applicare l'antecedente della regola ai dati in input al sistema, e se essi rispettano la formula logica, allora possiamo associare a tali dati il valore di classificazione che compare a destra della regola (nel nostro caso, "ha una influenza"). Anche se i sistemi a regole non hanno naturalmente associata una rappresentazione grafica, essi esprimono un tipo di ragionamento logico che ci è piuttosto usuale.

Individuati negli alberi di decisione e nei sistemi a regole i modelli che per la loro natura comprensibile sono candidati a esprimere le spiegazioni connesse alla interpretabilità, vediamo ora di precisare un po' meglio il perimetro della interpretabilità, distinguendo due aspetti in cui è coinvolta una spiegazione; essa può riguardare:

- il modello nel suo complesso; chiameremo il modello "a scatola nera" perché non ci è immediatamente evidente la spiegazione del suo funzionamento
- uno specifico risultato del modello, relativo ad uno specifico valore di input.

dando luogo ai seguenti due problemi.

Problema 1 - Trovare una spiegazione per il modello a scatola nera

Possiamo definire il problema in questo modo: dato un modello M1 a scatola nera che risolve un problema di classificazione, trovare una spiegazione per il modello M1 consiste nel costruire un nuovo modello M2 tra quelli considerati nativamente interpretabili (albero o regole), che imita il comportamento del modello non nativamente interpretabile e che è anche globalmente interpretabile (cioè si è in grado di interpretare tutto il modello e non solo casi particolari). M2 deve inoltre auspicabilmente essere *accurato*, cioè fornire risultati di classificazione con qualità simile a quella di M1. Per quanto riguarda quest'ultimo punto, è chiaro che non basta spiegare, la spiegazione non deve divergere rispetto al modello originario fornendo risultati eccentrici rispetto ai risultati forniti da M1. Nella Figura 5 tratta da [Guidotti 2018] si veda un esempio in cui il modello di spiegazione adottato è un sistema a regole.

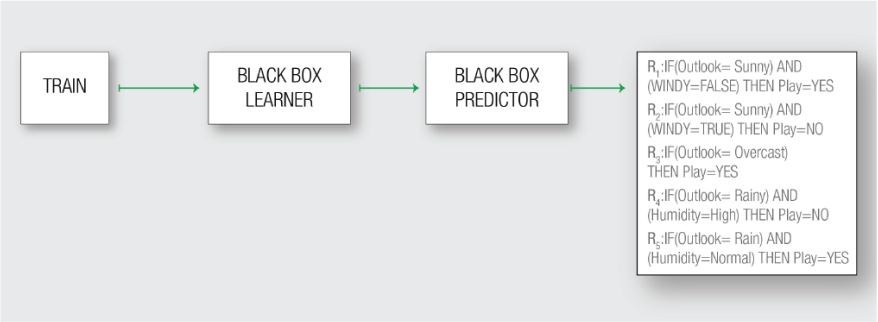


Figura 5 – Problema 1: trovare una spiegazione globale per un modello

Problema 2 - Trovare una spiegazione per uno specifico risultato prodotto dal modello

In questo caso il problema consiste nel fornire un risultato interpretabile; in altre parole, il modello interpretabile deve fornire un modello predittivo, insieme alle ragioni di tale predizione per un particolare valore di input. Chiaramente in questo caso la predizione è solo localmente interpretabile, e on si ritiene necessario spiegare l'intera logica del modello. Vedi Figura 6 tratta ancora da [Guidotti 2018].

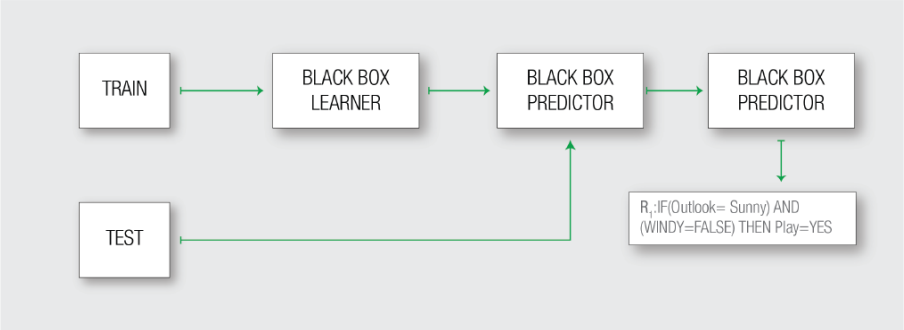


Figura 6 – Problema 2: trovare una spiegazione per un particolare risultato.

9. Il Regolamento generale sulla protezione dei dati (GDPR)

Il Regolamento generale sulla protezione dei dati (GDPR), la nuova legge sulla privacy per l'Unione Europea (UE), è entrato in vigore il 25 maggio 2018. I principi su cui si basa il GDPR fanno riferimento alle categorie introdotte all'inizio del capitolo e in particolare al principio generale per cui così come ognuno di noi è proprietario del proprio corpo, lo è anche dei dati che lo descrivono. Gli aspetti più rilevanti riguardano:

- Il principio del consenso informato: vengono stabilite regole rigide per ottenere il consenso come base legale per l'elaborazione
- La portabilità dei dati, che corrisponde al diritto di trasferire i dati personali da un fornitore di servizi a un altro
- La trasparenza, intesa come opacità dei dati personali, cioè il diritto di cancellare le informazioni su quali dati vengono raccolti e su come vengono elaborati
- La qualità dei dati, il diritto di correggere dati personali inesatti
- La cancellazione, il diritto in alcuni casi alla cancellazione dei dati personali
- Detto in maniera informale, non vale più l'affermazione: "l'ha detto il calcolatore", ogni utente ha il diritto a non essere passivamente soggetto a una decisione basata su modelli basati su tecniche automatiche.
- Una nuova e più ampia classificazione dei dati personali e sensibili rispetto alla precedente legislazione, che include gli identificatori on line, e i dati genetici e biometrici.

.

Le organizzazioni pubbliche o private devono rispettare i seguenti principi:

- La responsabilità, devono cioè dimostrare la conformità alle regole attraverso una registrazione di tutte le attività di elaborazione dati.
- L'analisi di impatto della protezione dei dati, obbligatoria se le attività di elaborazione dei dati possono dar luogo ad elevato rischio per i diritti degli individui.
- La sicurezza dei dati, che consiste nel conservare i dati in modo tale da garantire la sicurezza, attraverso appropriate misure tecniche e organizzative.

- Il trasferimento dei dati personali fuori dalla UE può avvenire solo se sono garantite adeguate salvaguardie.
- L'obbligatorietà di istituire un data protection officer, se l'organizzazione:
  1. è pubblica;
  2. ha relazioni con utenti su vasta scala;
  3. elabora dati sensibili.

L'ambito territoriale innova profondamente rispetto all'esistente ed è costituito da:
- organizzazioni con basi nella Unione Europea che raccolgono o elaborano dati personali di residenti nella UE;
- organizzazioni esterne alla UE che monitorano il comportante o vendono beni e servizi a cittadini residenti nella UE;
- service providers che elaborano dati personali come servizio ad altra organizzazione

Le sanzioni prevedono:
- multe fino a 20 milioni di euro o 4% del fatturato complessivo;
- rimborsi per i danni subiti

E infine il GDPR afferma il controverso ma fondamentale problema del diritto alla spiegazione da parte dei modelli di machine learning discusso nella Sezione precedente, che stabilisce che coloro che sono direttamente coinvolti dai modelli predittivi (es. una persona sta per compiere un delitto) e interpretativi (è lui il responsabile del delitto) hanno diritto a una spiegazione su come il modello è stato costruito. Occorre dire che vi sono opinioni ed analisi molto contrastanti su questo punto, e la ricerca è ancora in fase immatura; alcuni hanno parlato in modo veementemente contrario anche solo sulla possibilità che un tale diritto esista, altri sostengono che il diritto è auto evidente.

Non mancano le critiche al GDPR. [Chivot 2019] afferma che un anno dopo la entrata in vigore, ci sono prove crescenti che la legge non ha prodotto i risultati previsti; inoltre, le conseguenze indesiderate sono gravi e diffuse, in particolare il GDPR:

1. Influisce negativamente sull'economia e sulle imprese dell'Unione Europea
2. Riduce le risorse dell'azienda
3. Ha effetti di freno allo sviluppo delle startup tecnologiche europee
4. Riduce la concorrenza nella pubblicità digitale
5. È troppo complicato da implementare per le aziende
6. Non riesce ad aumentare la fiducia tra gli utenti
7. Ha Impatto negativo sull'accesso online degli utenti
8. È troppo complicato per essere compreso dai consumatori
9. Non è attuato in modo coerente in tutti gli Stati membri
10. Distorce l'azione e le risorse dei regolatori.

10. Ethics by design (l'Etica tramite regole di progettazione)

Nelle sezioni precedenti, soprattutto nell'ambito della equità e della esplicabilità, abbiamo visto metodi che permettono di valutare o di incrementare il livello di estensione che questi due determinanti dell'etica hanno in un modello o sistema esistente. Nella letteratura e tra gli enti e le conferenze che discutono aspetti etici della Scienza dei dati sono diffusi insiemi di regole che un progettista dovrebbe applicare per produrre nel corso stesso del ciclo di vita del dato digitale sistemi che rispettino principi etici, promuovendo quindi l'etica tramite regole di progettazione.

Ad esempio in [Mulvenna 2017] vengono proposti i seguenti principi generali validi per tutte le applicazioni informatiche utilizzate in prodotti o servizi.

- Progettare l'applicazione avendo come principale fine supportare le persone che useranno il prodotto o il servizio, generando empatia per gli utenti.
- Fornire agli utenti informazioni sufficienti per consentire lro di prendere decisioni informate in ogni fase del progetto sulla opportunità, su quando e come utilizzare il prodotto o il servizio.
- Rispettare il diritto delle persone di scegliere il modo in cui interagiscono con il prodotto o il servizio; offrire alternative o personalizzazioni.
- Equilibrare le esigenze di privacy e sicurezza con un accesso al servizio da quanti più sistemi e persone possibile.
- Non limitarsi a definire per la applicazione una strategia di corto periodo, ma inquadrarla in una policy d lungo periodo.
- Individuare proattivamente possibili pregiudizi e potenziali valori che potrebbero riflettersi nel prodotto o servizio.
- Tenere conto di esigenze, capacità, punti di vista e principi morali diversi.
- Tenere conto di processi decisionali e feedback diversi.
- Puntare a progetti economicamente, ecologicamente e socialmente sostenibili.
- Integrare nelle specifiche della applicazione le modalità con cui gestire i comportamenti erronei, inclusi trasparenza e reportistica.
- Essere realistici su ciò che è possibile e necessario.
- Supportare il prodotto o servizio per tutto il suo ciclo di vita.

Nel corso della Conferenza di Asilomar del 2017, sono stati sviluppati i corrispondenti *principi di Asilomar*, che forniscono indicazioni generali focalizzate sui sistemi di Intelligenza Artificiale (IA nel seguito), disciplina con molti punti di contatto con la Scienza dei dati [53]. Vediamo quelli che fanno più diretto riferimento all'etica (vedi anche [Estropico 2017].

- Sicurezza: i sistemi basati sulla IA devono essere sicuri e protetti per tutta la loro vita operativa – tali proprietà devono esse verificabili, quando necessario e fattibile.
- Mancata trasparenza: se un sistema di IA provoca danni, dovrebbe essere possibile accertarne il motivo.
- Trasparenza giuridica: in qualsiasi coinvolgimento di un sistema di IA in una decisione giuridica dovrebbe essere possibile fornire una spiegazione soddisfacentemente verificabile da un'autorità umana competente.

[53] Il tema delle relazioni concettuali e delle aree di sovrapposizione tra le discipline della Scienza dei dati e della Intelligenza Artificiale non è ancora giunto nella letteratura a un livello di maturità adeguato per poter essere trattato in questo libro.

Responsabilità: i progettisti e costruttori di sistemi avanzati di IA sono parti interessate nelle implicazioni morali del loro uso, abuso ed azioni, con la responsabilità di influire su tali implicazioni.

- Allineamento dei valori: i sistemi di IA dotati di autonomia devono essere progettati in modo che i loro obiettivi e comportamenti possano essere affidabilmente allineati con i valori umani in tutte le attività in cui sono coinvolti.
- Valori Umani: i sistemi di IA devono essere progettati e gestiti in modo di essere compatibili con gli ideali di dignità umana, dei diritti, della libertà e della diversità culturale.
- Privacy personale: le persone devono avere il diritto di accedere, gestire e controllare i dati da essi stessi generati, in tutti i casi in cui i sistemi di IA hanno il potere di analizzare e utilizzare tali dati.
- Libertà e Privacy: l'applicazione di IA ai dati personali non deve irragionevolmente limitare la libertà delle persone (reale o percepita).
- Condivisione dei benefici: le tecnologie di IA dovrebbero beneficiare e potenziare il maggior numero di persone possibile.
- Prosperità condivisa: la prosperità economica creata dall'IA deve essere condivisa a beneficio del più ampio insieme di comunità e nazioni possibile.
- Controllo umano: gli esseri umani dovrebbero scegliere come e se delegare le decisioni ai sistemi di IA, per raggiungere gli obiettivi che dovrebbero rimanere loro propri.
- Non-eversione: il potere conferito dal controllo dei sistemi di IA altamente avanzati dovrebbe rispettare e migliorare, piuttosto che sovvertire, i processi sociali e civili da cui dipende la qualità della vita della società.
- Corsa agli armamenti: la corsa agli armamenti in armi letali autonome dovrebbe essere bandita.

Altri principi in tema di Ethics by design si trovano in [Dennis 2018], [Dignum 2018]. [Jansen] e [Stoyanovich 2017]. Quest'ultimo lavoro è particolarmente interessante perché descrive un sistema, chiamato *Fides*, che definisce un insieme di tecniche che, oltre che garantire la accuratezza dei risultati, "forzano" i principi etici della equità e della trasparenza, e, come vedremo tra poco, della privacy, in tutte le fasi del ciclo di vita, che in Fides è definito in termini delle seguenti fasi:

- Acquisizione e cura dei dati, che raccoglie tutte le fasi del ciclo di vita definito nel Capitolo 2 che precedono la analisi.
- Ricerca esplorativa, in cui possono essere tollerati falsi positivi e tecniche discriminatorie per permettere un più ampio insieme di direzioni nella esplorazione
- Analisi confermativa, in cui l'attività sperimentale deve essere rigorosa e riproducibile
- Diffusione dei risultati, in cui i risultati della analisi sono resi disponibili agli utenti e sono usati per prendere decisioni che influiscono sul "mondo".

Come si vede, il ciclo di vita differisce da quello mostrato e discusso nel libro, nella direzione di dare maggiore rilevanza alle attività connesse con la analisi e la diffusione. Inoltre, le sotto fasi mostrano una tipologia di analisi più orientata ad utilizzare metodi e modelli statistici rispetto a quello visto nel libro; in questo modo il lettore ha la possibilità di vedere approcci diversi.

Fase	Sottofase	Proprietà considerata
Acquisizione e cura dei dati	Acquisizione di conoscenza di dominio	Equità
	Annotazione automatica di dati sensibili	Privacy
	Scoperta di relazioni tra i dataset	Accuratezza
Ricerca esplorativa	Testing sulla base di ipotesi multiple	-
	Spiegazioni esogene	-
	Verifiche statistiche	Accuratezza
Analisi confermativa	Gestione delle ipotesi	Equità/Trasparenza
	Tracciatura degli esperimenti	Equità
Diffusione dei risultati	Provenienza semantica	Interpretabilità
	Verifica e spiegazione	Equità/Interpretabilità

Figura 7 – Fasi e sottofasi nel ciclo di vita in [Stoyanovich 2017] e proprietà considerate

In Figura 7 mostriamo per ogni sottofase la proprietà etica considerata. Commentiamo le varie sottofasi.

- Acquisizione di conoscenza di dominio – Per permettere al sistema di verificare automaticamente situazioni di discriminazione nei risultati, occorre costruire un modello della popolazione da cui il dataset in input al ciclo di vita è tratto. Fides permette ai data owners di asserire conoscenza di dominio che può essere usata per quantificare e correggere le distorsioni.
- Annotazione automatica di dati sensibili – E' necessario definire politiche di controllo di accesso ai dati, e annotare i dati sensibili. E' impensabile che ciò venga fatto manualmente; Fides fornisce tecniche di machine learning per effettuare la annotazione tramite apprendimento.
- Scoperta di relazioni tra dataset – Trovare le relazioni nascoste tra data set aiuta a produrre risultati accurati e che tengono conto di tutta le fonti disponibili.
- Testing sulla base di ipotesi multiple – E' un testing tipico in statistica in cui diverse ipotesi sui dati possono essere verificate contemporaneamente.
- Spiegazioni esogene – Il sistema può esplorare diverse spiegazioni, anche mediante applicazione di tecniche di corrispondenza tra i dataset disponibili e altri dataset disponibili nel Web. Ad esempio in una analisi sull'utilizzo di biciclette in una città, è possibile correlare il dataset con altri in cui il tempo è stato inclemente, per trovare una correlazione tra i due fenomeni (biciclette/tempo atmosferico).
- Verifiche statistiche – I modelli di regressione lineare adottati in Statistica assumono che i dati abbiano una distribuzione cosiddetta normale, le verifiche permettono di verificare questa proprietà evitando risultati spuri e non corretti.
- Gestione delle ipotesi – E' fondamentale essere in grado di verificare in questa fase le questioni poste dagli utenti e i loro dubbi sulla assenza di distorsioni, garantendo equità e trasparenza, ciò viene fatto con tecniche di testing sulla base di ipotesi multiple.
- Tracciatura degli esperimenti – Questa attività permette di investigare distorsioni dovute a selezioni di dati non rappresentativi dell'universo, collezionando nuove fonti e tracciando la sequenza degli esperimenti.
- Provenienza semantica - La provenienza dei dati è di grande rilevanza per la fase di analisi; essa può essere *sintattica* quando semplicemente traccia i dati che hanno partecipato nella formazione di un

risultato; *semantica* quando entra nel merito del significato dei dati permettendo una analisi più ricca e una maggiore usabilità e interpretabilità.

- Verifica e spiegazione – Siamo all'atto finale, il più delicato in cui i risultati vanno condivisi con gli utenti. La verifica riguarda l'equità e assenza di distorsioni; la spiegazione fa riferimento alla interpretabilità.

11. Conclusioni

Le precedenti considerazioni sono un primissimo contributo al tema nascente del rapporto tra etica e dati. Via via che i dati descriveranno un crescente numero di aspetti della nostra vita, crescerà la tentazione di considerare i dati e non la realtà fisica, mentale e emozionale quale rappresentazione (talvolta falsamente) obiettiva ed esclusiva della nostra vita. Investigare l'etica nella sua relazione con i dati digitali è un modo per salvaguardare la nostra dignità di esseri umani, e mantenere le nostre responsabilità nelle relazioni con gli altri e con l'infosfera dei dati digitali.in continua espansione.

Riferimenti

S. Abiteboul e Stoyanovich, J. - Transparency, Fairness, Data Protection, Neutrality: Data Management Challenges in the Face of New Regulation. Journal of Data and Information Quality 11(3), 2018.

AGID, Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino, AGID, Marzo 2018.

Asilomar AI principles, https://futureoflife.org/ai-principles/ (verificato 10 agosto 2019)

R. Binns, Fairness in Machine Learning: Lessons from Political Philosophy, Proceedings of Machine Learning Research 81:1–11, 2018

S. Barocas and D. Boyd - Computing Ethics Engaging the Ethics of Data Science in Practice, Comm. of the ACM, Novembre 2017.

C. Batini - Etica e Big Data - L'etica, gli archivi e la cultura, Associazione nazionale archivistica italiana, Sezione Trentino Alto Adige Südtirol, 19 e 20 aprile 2018, Trento

R. Binns, Fairness in Machine Learning: Lessons from Political Philosophy, Proceedings of Machine Learning Research 81:1–11, 2018.

D. Boyd and K. Crawford – Critical questions for big data, Information, communication and society, 2017

J. Burrel - How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society January–June 2016: 1–12, Conference on Fairness, Accountability, and Transparency: Preface, Proceedings of Machine Learning Research 81:1–2, 2018

E. Chivot e Daniel Castro - What the Evidence Shows About the Impact of the GDPR After One Year, Information Technology and Innovation Foundation, 2019

A. Croll - Big data is our generation's civil rights issue, and we don't know it – O'Reilly Radar, 2012.

L. Dennis e M. Fisher - Practical challenges in explicit ethical machine reasoning. - arXiv preprint arXiv:1801.01422 (2018).

V. Dignum, et al. - Ethics by Design: necessity or curse? - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 2018.

K. Donovan - Seeing Like a Slum towards Open, Deliberative Development, Science and Technology, 2012

D. Ensign, Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. - Runaway feedback loops in predictive policing. arXiv preprint arXiv:1706.09847, 2017.

Estropico Blog: I principi di Asilomar per l'Intelligenza Artificiale - http://estropico.blogspot.com /2017/02/i-principi-di-asilomar-per.html#ixzz5wD6gyyuO, 2017

B. Fish, Kun, J., & Lelkes, Á.D. - A Confidence-Based Approach for Balancing Fairness and Accuracy. SDM, 2016.

L. Floridi - Information ethics: On the philosophical foundation of computer ethics, Ethics and Information Technology 1: 37–56, 1999.

R. Guidotti, A. Monreale, F. Turini, Dino Pedreschi, Fosca Giannotti - A Survey Of Methods For Explaining Black Box Models. CoRR abs/1802.01933, 2018.

M. Herschel, M., Diestelkämper, R., & Ben Lahmar, H. - A survey on provenance: What for? What form? What from? The VLDB Journal—The International Journal on Very Large Data Bases, 26(6), 881-906, 2017.

K. Holstein, Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. - Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 600). ACM, 2019.

B. Hutchinson e M. Mitchell - 50 Years of Test (Un)fairness: Lessons for Machine Learning - ACM Conference on Fairness, Accountability, and Transparency, 2018.

S. Grimmelikhuijsen - A good man but a bad wizard. About the limits and future of transparency of democratic governments." Information Polity 17.3, 4, 2012.

F. Jansen - Engineering for social justice instead of fairness: a perspective from the context of policing."

J. Johnson, From open data to information justice - Ethics Inf Technol (2014) 16:263–274

J. Johnson, The question of Information Justice, Communications of the ACM, March 2016

Kebira Project - Making all voices count – a grand challenge for development, Research report, October 2017

S. Hajian, F. Bonchi, Carlos Castillo - Algorithmic Bias from discrimination discovery to fairness-aware data mining, KDD 2016 Tutorial, August 13th, 2016. San Francisco, US

J. Leidner e V. Plachouras, Ethical by Design: Ethics Best Practices for Natural Language Processing, Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 30–40, Valencia, Spain, April 4th, 2017

L. Lessig – Against transparency – New Republic digital edition, 2009.

E. Hagen - Open mapping from the ground up: learning from Map Kibera, Ground Truth Initiative, 2017

J. King - Computing Ethics Humans in Computing: Growing Responsibilities for Researchers, Communications of the ACM, 2015.

K. Kirkpatrick – It's not the algorthim, it'ìs the data – Communications of the ACM, February 2017.

J. A. Johnson, From open data to information justice, Ethics Inf Technol (2014) 16:263–274

J. Metcalf et al.- Where are human subjects in Big Data research? The emerging ethics divide, Big Data & Society January–June 2016: 1–14, March 2015.

M. Mulvenna et al. - Ethical by Design - A Manifesto, Ecce, ACM, 2017.

L. Noren Course on Ethics of Data Science, NYU, 2017.
Principles for Algorithmic Transparency and Accountability: A Provenance Perspective, unpublished blog, 2017.

The council for Big Data, Ethics and Society, Perspectives on Big Data, Ethics, and Society, 2017.

European General Data Protection Regulation (GDPR), European Union, 2016.

S. Friedler e Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236.

Propublica, COMPAS Recidivism Risk Score Data and Analysis, 2018

B. Raman - The rhetoric of transparency and its reality: Transparent territories, opaque power and empowerment - The Journal of Community Informatics 8.2, 2012.

S. Ruggieri, Pedreschi D., e Turini F. - Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4.2: 9.

S. Ruggieri e Turini F. "A KDD process for discrimination discovery." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2016.

B. Selbst, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68). ACM.

Shroff R. – Combatting Police discrimination in the age of Big data, in https://speakerdeck.com/fatml/algorithmic-accountability-and-transparency-in-journalism-nick-diakopoulos

J. Stoyanovich, Bill Howe, Serge Abiteboul, Gerome Miklau, Arnaud Sahuguet, et al.. Fides: Towards a Platform for Responsible Data Science. SSDBM'17 - 29th International Conference on Scientific and Statistical Database Management, Jun 2017, Chicago, United States.

J. Stoyanovich et al. "Fides: Towards a platform for responsible data science." Proceedings of the 29th International Conference on Scientific and Statistical Database Management. ACM, 2017.

J. Stoyanovich et al - Follow the data! The alg. Transparency, The Ethical Machine 2018

M. Swan - Philosophy of Big Data: Expanding the Human-Data Relation with Big Data Science Services, 2015.

M. Turilli e Floridi L. - The ethics of information transparency, Ethics Inf Technol (2009) 11:105–112

S. Verma, & Rubin, J. (2018, May). Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (pp. 1-7). IEEE.

W. Xiaolin and Zhang X. - Automated Inference on Criminality using Face Images, CoRR abs/1611.04135, 2016.

I.Zliobaite - A survey on measuring indirect discrimination in machine learning." arXiv preprint arXiv:1511.00148 (2015).

I.Zliobaite - Fairness-aware machine learning: a perspective - arXiv preprint arXiv:1708.00754 (2017).

Do No Harm et al.: Ethical Guidelines for Applying Predictive Tools Within Human Services, Data Science and Human Services, 2018

Capitolo 16 – I limiti della Scienza dei dati

Carlo Batini e Fabio Stella, con contributi di Anna Ferrari

1. Introduzione

Perché il Problema 2 del Capitolo 1, trovare il giorno in cui conviene acquistare il biglietto aereo più economico per un dato viaggio, è stato risolto solo recentemente? Ne abbiamo parlato nel Capitolo 1: un modo per risolvere il problema poteva consistere nel cercare di individuare il modello di pricing adottato dalle compagnie aeree per attribuire un prezzo ai biglietti. Una volta formulata una ipotesi sul modello di attribuzione del prezzo ai biglietti (pricing) si sarebbe potuto verificare la accuratezza del modello sulle serie storiche dei prezzi; qualora la accuratezza fosse stata troppo bassa, si sarebbe potuto modificare il modello fino, auspicabilmente, a convergere verso un modello affidabile, che poi si sarebbe potuto applicare ai biglietti da acquistare in futuro. La scienza sperimentale opera quindi con la seguente sequenza di passi che qui riassumiamo:

1. Formuliamo dapprima una ipotesi scientifica o modello che riteniamo approssimare bene il fenomeno di interesse
2. Validiamo o meno il modello sulla base di uno o più criteri quantitativi adatti per il modello desiderato.

Ora, il modello di pricing è troppo complicato per poter essere espresso attraverso un insieme di formule matematiche ovvero di regole logiche, ciò che abbiamo chiamato un modello. Inoltre, nel tempo, la legge del prezzo può essere stata modificata per un cambiamento del modello di business della compagnia, ovvero per il mutamento della domanda, ad esempio, per ragioni geopolitiche.

La Scienza dei dati innova rispetto al metodo sperimentale. Per capire come, dobbiamo chiederci quali sono le sue caratteristiche salienti; queste caratteristiche nascono dalla grande trasformazione tecnologica e sociale in atto per cui:

- La rappresentazione del mondo è costituita sempre più da dati digitali (vedi Figura 1)
- E' perciò possibile descrivere fenomeni sempre più complessi mediante dati digitali
- Possiamo analizzare fenomeni complessi e risolvere problemi legati a tali fenomeni mediante algoritmi automatici, anche quando la complessità della rappresentazione è tale che sia impossibile costruire un modello che leghi i dati in ingresso ai dati in uscita
- Mediante tecniche statistiche o di apprendimento automatico, gli algoritmi si costruiscono da soli.
- Gli algoritmi permettono di costruire modelli descrittivi, interpretativi, predittivi, prescrittivi.
- Gli algoritmi possono essere utilizzati in linea di principio in qualunque dominio applicativo, e in particolare nelle scienze sociali e scienze della vita, in cui in virtù della complessità del dominio spesso è difficile individuare modelli.
- Esistono tecnologie che possono operare su grandi quantità di dati senza perdere in efficienza.

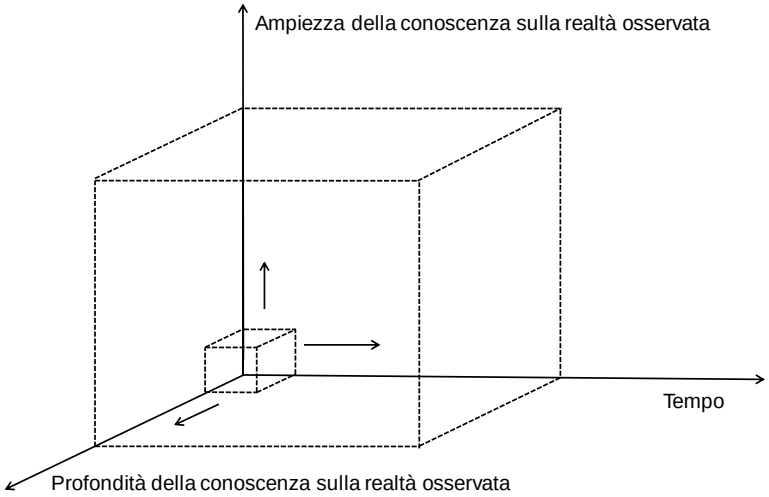


Figura 1 – L'espansione dei dati digitali nella rappresentazione del mondo – dai piccoli dati ai grandi dati

Chris Anderson, Editor nel 2009 della rivista Wired, scrive in [Anderson 2009]: "La nuova disponibilità di enormi quantità di dati, insieme agli strumenti statistici per operare su questi numeri, offre un modo completamente nuovo di comprendere il mondo. La correlazione sostituisce la causalità e la scienza può avanzare anche senza modelli coerenti, teorie unificate o una qualsiasi spiegazione meccanicistica".

Leo Breiman, un famoso statistico che ha prima lavorato come consulente di grandi aziende e poi ha lavorato alla Università, è sferzante con i suoi colleghi ricercatori in Statistica in questo articolo [Breiman 2001] di cui qui riassumiamo e adattiamo alcuni passi dell'abstract: "ci sono due culture nell'uso della modellistica statistica per trarre conclusioni dai dati. Una assume che i dati siano troppo complessi per poter sperare di individuare un modello che ne descriva le proprietà, e tratta il meccanismo insito nei dati come sconosciuto. La comunità statistica si è impegnata nell'uso quasi esclusivo di modelli di dati. Questo impegno ha portato a una teoria irrilevante, a conclusioni discutibili e ha di fatto impedito agli statistici di lavorare su una vasta gamma di interessanti problemi attuali".

Il grande fenomeno in atto, il tentativo che la Scienza dei dati ha in corso, è di aggiungere all'approccio storico della Statistica e della scienza sperimentale basato su modelli, o *model driven*, un approccio in cui *sono i dati che suggeriscono il modello*, che non è quindi dato a priori; questo nuovo approccio è chiamato *data driven*.

La precedente grande questione è anche riassunta da diversi ricercatori affermando che nella nascente Scienza dei dati si è passati dal *perché*, dalla ricerca in altre parole di modelli generativi dei dati, al *cosa*, alla ricerca di correlazioni e legami logici tra i dati senza porsi a priori il problema del modello. E' vera questa affermazione? Ne parleremo più diffusamente nel seguito del capitolo.

Accanto al tema del perché e del cosa, molti ricercatori si sono resi conto di un particolare che finora ci era sempre sfuggito in questo libro, e a cui avevo accennato nel capitolo introduttivo. Le tecniche statistiche o di apprendimento di cui abbiamo parlato operano su *dati*; questi dati descrivono fenomeni

passati, e non potrebbe essere altrimenti… Ma se tutto ciò di cui abbiamo parlato, se l'enorme produzione di modelli predittivi e interpretativi/diagnostici di questi anni, nasce da tecniche che operando su dati, e hanno solo la capacità di osservare il passato, come potranno esse sostituire la "vecchia" scienza sperimentale? Ciò che permette la scienza sperimentale è di passare dalla osservazione passiva dei dati, all'intervento sul mondo. Ebbene, il fatto che la nuova scienza che si sta formando, la Scienza dei dati, operi sul passato, non porta ad una limitazione intrinseca attraverso la generazione di modelli inerentemente limitati?

Affrontiamo un altro aspetto del problema del cosa e del perché; ricercatori apprendono presto nel loro percorso di formazione che "correlation is not causation", che abbiamo già introdotto nel Capitolo 9: il concetto di correlazione è profondamente diverso dal concetto di causa-effetto, e nessuna conclusione di relazione causale dovrebbe essere tratta a partire dalla esistenza di una correlazione tra i dati coinvolti nella correlazione.

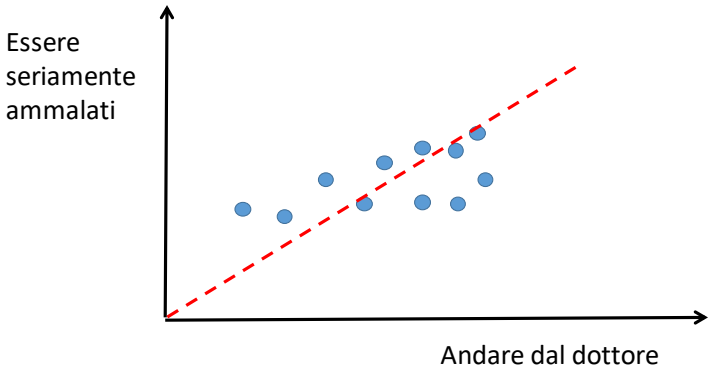


Figura 2 – Correlazione misurata su un insieme di pazienti
tra l'andare dal dottore e l'essere seriamente ammalati

Ad esempio, in Figura 2 mostriamo dei dati che esprimono, per un insieme di pazienti, la frequenza delle visite dal dottore e lo stato di gravità della malattia da essi sofferta. In questo caso, è evidente, non c'è una relazione causale tra andare dal dottore e la gravità della malattia, tant'è che, se tale relazione ci fosse, basterebbe andare *meno* dal dottore per sperare di non avere malattie serie.

Ancora Chris Anderson: "Confrontato con i big data, l'approccio della scienza sperimentale – formula una ipotesi, costruisci un modello, verifica il modello sulla realtà - sta diventando obsoleto. I big data ci permettono di arrivare alla conclusione che la correlazione è tutto ciò che ci serve. Non si devono più cercare modelli; si possono analizzare i dati senza formulare alcuna ipotesi su cosa potrebbero evidenziare. Possiamo gettare i dati in un cluster di calcolatori più potenti, e fare in modo che gli algoritmi trovino da soli i pattern che la Scienza non è in grado di trovare".

"Questo è un mondo in cui grandi quantità di dati elaborati usando metodi e modelli della matematica applicata sostituiscono ogni altro strumento cui si potrebbe attingere. Non servono più le teorie del comportamento umano, dalla linguistica alla sociologia; dimentichiamo le tassonomie, le ontologie, la psicologia. Che importanza ha chiedersi perché le persone fanno quello che fanno? Il punto è che lo

fanno, e noi possiamo tracciare e misurare le azioni umane con una accuratezza senza precedenti. Avendo a disposizione abbastanza dati, i dati parlano da soli."

In questo capitolo cercheremo di approfondire tutti i precedenti problemi. Il capitolo è organizzato come segue. Nella Sezione 2 diamo la parola a Leo Breiman, che, come abbiamo detto, dopo un periodo speso come consulente per aziende entrò all'Università portando con sé tutto il bagaglio di tecniche di analisi applicate nella sua attività di consulenza, e sviluppò nell'articolo che discutiamo una critica al metodo statistico classico. La Sezione 2 discute la distinzione tra approcci alla analisi model-driven e data-driven.

La Sezione 3 è dedicata alla vicenda di Google, che lanciò una tecnica per prevedere l'insorgere e il diffondersi di patologie come l'influenza, chiamata Google Flu, ma fu successivamente costretta a chiudere questa esperienza per manifesta inadeguatezza delle tecniche utilizzate. La Sezione 4 è dedicata al problema della relazione tra i concetti di correlazione e causazione, intesa quest'ultima come relazione causa effetto. La Sezione 5 osserva più da vicino di quanto fatto finora il tema del passaggio nelle tecniche di analisi dai piccoli dati ai grandi dati. La Sezione 6 affronta un insieme di considerazioni critiche e punti di debolezza espressi in letteratura sugli approcci della Scienza dei dati, che arrivano a prefigurare addirittura una prossima decadenza della scienza dei dati nel panorama scientifico delle scienze. Le ultime due Sezioni 7 e 8 affrontano da vicino il tema cruciale della relazione tra Scienza dei dati e Intelligenza Artificiale, e tra l'intelligenza raggiunta dalle tecniche di analisi e l'intelligenza degli esseri umani.

2. La critica al metodo statistico nella visione di Leo Breiman

In questa sezione presentiamo la critica espressa in [Breiman 2001] riferita al confronto tra la cultura classica della analisi statistica, incentrata su ciò che abbiamo chiamato nella introduzione il metodo model driven, e la nuova cultura data driven.

L'approccio model driven considera la relazione tra dati di input e dati di output come espressa da un modello che è individuato a priori, ed è basato, ad esempio, sul legame di correlazione. La definizione del modello non è semplice, ed è spesso necessario ricorrere a tecniche preliminari di selezione tra diversi modelli candidati e di studio dei fenomeni in oggetto che solitamente devono rispondere a delle ipotesi precise, tra cui la ipotesi di distribuzione Gaussiana dei dati (vedi quanto detto nel Capitolo 9). Una volta definito il modello, viene validata la sua capacità esplicativa, utilizzando diversi indici chiamati *indici di bontà di adattamento*, e il *tasso di errore*, che valuta in quale misura il modello rappresenta la reale relazione tra dati di input e di output. Se questi indici soddisfano determinati requisiti, il modello potrà essere impiegato come *modello predittivo*. Inoltre, proprio in quanto definito da un'equazione precisa, la relazione che sussiste tra input e output è nota ed espressa dal modello.

L'approccio data driven non definisce a priori una relazione tra input e output, ma cerca di dedurne un'approssimazione direttamente dai dati attraverso l'impiego di un algoritmo, in genere basato su tecniche di machine learning, di cui abbiamo parlato nel Capitolo 10. In questo caso non è possibile

stabilire esattamente la natura della relazione tra variabile di input e di output, non essendo questa definita espressamente.

La critica di Breiman agli approcci model driven si concentra soprattutto sulla scarsa efficacia degli approcci, secondo tre aspetti principali:

1) *La non unicità dei modelli plausibili:* La selezione del modello migliore è sicuramente una fase molto delicata per l'approccio model driven. Diverse tecniche possono essere impiegate per selezionare le variabili da tenere in considerazione a partire dai dati di input, ma è senz'altro vero che si possono determinare classi di modelli che, pur selezionando variabili differenti, presentano un error rate simile. Questo può essere indice di instabilità e l'esperto deve tenerlo in considerazione.
2) *La semplificazione indotta nella ricerca dei modelli, a fronte di fenomeni complessi:* l'approccio model driven predilige la selezione di modelli semplici. Breiman evidenzia che l'accuratezza dei modelli molto semplici come quelli privilegiati nell'approccio model driven è usualmente inferiore rispetto a quella raggiunta dall'approccio data driven. Peraltro, ricordiamo, lo svantaggio dell'approccio data driven è nella ridotta, e in molti casi la totale, assenza di interpretabilità; il concetto di interpretabilità è stato discusso nella Sezione 8 del Capitolo 15. Si genera quindi un dilemma tra complessità e interpretabilità che potrebbe essere risolto in base al contesto in cui l'analista si trova; ad esempio, per quanto riguarda le analisi mediche, l'interpretabilità del modello è molto importante perché può determinare l'uso o meno di procedure invasive per i pazienti.
3) La *problematica della "dannazione della dimensionalità":* l'approccio model driven privilegia un numero di variabili basso per non incorrere nella *"dannazione della dimensionalità"*. In generale, l'aggiunta di una variabile porta informazioni aggiuntive e può facilitare l'individuazione del modello predittivo; vi è però un limite, in cui l'aggiunta di ulteriori variabili non porta benefici al modello, anzi, può causarne la perdita di efficienza. Per questo motivo, si preferisce selezionare le variabili più informative, in modo tale da non aumentare la dimensionalità. L'approccio data driven non limita in alcun modo a priori la complessità del modello ma mette in contrapposizione la complessità del modello con le prestazioni che il modello ottiene su un insieme di dati di test, vale a dire un insieme di dati non usato nell'addestramento del modello.

Per Breiman, non c'è da stupirsi se molta parte della ricerca e delle aziende si siano spostati sull'approccio data-driven: non solo questo approccio funziona, ma funziona anche molto bene. Il punto è capire se questa nuova forma di analisi è effettivamente il punto di snodo tra il paradigma del metodo statistico e un nuovo paradigma basato sui dati. Oggigiorno, siamo ancora in un momento di transizione che, per essere effettuato con successo, deve essere in grado di motivare l'integrazione o, perfino, la sostituzione totale dei paradigmi che la scienza sta offrendo. Per arrivare a ciò, due sono le domande che devono trovare una risposta.

1. *Gli approcci model driven e data driven hanno lo stesso obbiettivo? Se sì, Sono sostituibili e in quale modo? Se no, in quale misura possono coesistere per dare un contributo costruttivo alla conoscenza?*

La risposta di Breiman è no, i due approcci non hanno lo stesso obbiettivo, almeno per ora. Mentre l'approccio model driven risponde alla domanda "è vero?", l'approccio data-driven risponde alla domanda "è utile?". In un mondo ideale in cui la realtà è descrivibile senza errore, non ci sarebbe scelta:

attraverso la verità si arriva all'utilità, mentre il viceversa non è detto. Il problema è che, come discusso poco fa, descrivere la realtà implica una serie di semplificazioni che, alla fine dei conti, non ci permette di descrivere una *realtà reale*, ma una *realtà ideale, astratta*.

Se la realtà non è dunque possibile da descrivere, non dovrebbero esserci dubbi sullo scegliere l'approccio data-driven. In realtà non è così; la risposta alla domanda "è utile?" è certo molto allettante e, tra l'altro, i metodi di machine learning sono molto accurati e forniscono risultati molto interessanti. Il problema si pone sul lungo periodo. Se da un lato, infatti, l'approccio data-driven porta a un vantaggio in termini di adattabilità rispetto a una realtà in continuo cambiamento, dall'altro esso è intrinsecamente un'ammissione di ignoranza rispetto alla realtà stessa.

Se ammettiamo di non conoscere come determinati algoritmi riescano a arrivare a un certo risultato, e non riusciamo a spiegare perché producano quel risultato, allora significa che non stiamo spiegando la realtà in un senso più profondo. Il fatto che i risultati siano ottimi perde significato di fronte alla impossibilità di dimostrare perché li abbiamo raggiunti. Per usare una metafora: sarebbe come tirare a caso e fare canestro. E' vero che tirando posso raggiungere il canestro mille volte, ma ciò che fornisce un valore aggiunto è aver capito come scegliere la traiettoria e capire questa da quali altre variabili dipende; solo così, noi esseri umani possiamo incrementare la nostra conoscenza. Questo non significa che l'approccio data-driven non lo faccia, ma bisogna capire come; alla mancanza di una generalizzazione metodica corrisponde un'insicurezza, sia sull'affidabilità sia sulla possibilità di intervenire a lungo andare sugli algoritmi in maniera efficace. In parole semplici, possedere una macchina, ma non avere nessuno che possa aggiustarla, significherebbe doverla cambiare ogni volta che non funziona più, e questo è tutt'altro che utile.

Sulla base delle precedenti considerazioni, seppure l'approccio model driven presti il fianco a un insieme di limiti concettuali, siamo in una fase (cito sempre [Breiman 2001]) in cui l'approccio data-driven è tutt'altro che pronto per sostituire l'approccio model driven. Con gli strumenti che ha a disposizione, la comunità statistica dovrebbe impegnarsi a cercare di proporre una nuova formalizzazione, per poter sfruttare i metodi data-driven di analisi di dati. E' impensabile e anacronistico non predisporsi a una trasformazione in seguito a questo fenomeno; nella sua storia, la statistica ha attraversato molte mutazioni negli scopi e di conseguenza nei metodi e oggi è tempo di compiere una nuova trasformazione.

2. *Come si devono selezionare i dati nell'approccio data-driven e quindi come si può validare un risultato prodotto?*

Per quanto riguarda la selezione delle fonti e la acquisizione dei dati digitali, tra le prime fasi del ciclo di vita definito nel Capitolo 2, la comunità scientifica è divisa. Se da un lato si sostiene che alla complessità della realtà debba corrispondere analoga complessità nei dati che sono utilizzati per analizzarla, dall'altro si ritiene che la troppa informazione non sia sempre un valore aggiunto, ma possa generare disinformazione. Come facciamo a scegliere di quanta e quale informazione abbiamo bisogno? Pensiamo a quando utilizziamo Google per informarci su un fatto di cronaca e formarci una nostra opinione; abbiamo a disposizione moltissimi dati, ma tra tutti quelli disponibili quali sono effettivamente quelli che ci servono e di quali possiamo effettivamente fidarci? Se non facessimo una

precisa selezione dei dati, non riusciremmo ad arrivare al dato che stiamo cercando, e spesso o non ci si arriva oppure c'è un alto rischio di acquisire dati scorretti o contraddittori. Sorge spontanea la domanda: è meglio avere "tanta" informazione e scegliere quale utilizzare, o è meglio averne poca e già selezionata da qualcuno?

Il punto fondamentale è che chi sostiene l'approccio data-driven è portato a preferire che non sia stata fatta una selezione preliminare dei dati. Dalla "dannazione della dimensionalità" siamo passati alla "benedizione della dimensionalità"; il modello data-driven, nonostante sia effettivamente più elastico rispetto alla complessità della realtà, nella pretesa di voler utilizzare tutti i dati disponibili tende paradossalmente a creare una descrizione ad hoc della realtà che quindi è difficilmente generalizzabile.

Le soluzioni ad hoc non possono essere soluzioni durevoli, e devono essere sostituite con soluzioni derivanti da un'astrazione tale da permettere di comprendere gli aspetti peculiari della realtà. Una soluzione potrebbe essere impiegare entrambi i metodi, model driven e data-driven, ma in quale ordine e con quale ruolo? Brieman fa una proposta di metodo: si potrebbe utilizzare l'approccio data-driven per una prima esplorazione dei dati, al fine di dare una direzione all'analisi più formale o, viceversa, usare l'approccio model-driven per avere un'indicazione di causa-effetto al fine di effettuare la scelta dei dati, sui quali può essere poi utilizzato un algoritmo di machine learning. Il dibattito è aperto.

3. I dati NON parlano da soli – La parabola di Google Flu Trends e l' Hubris dei dati

La visione di Anderson per cui i dati *parlano da soli*, ha portato in diversi casi ad utilizzare i big data con una certezza fideistica sul fatto che tutto ciò che producevano in termini di modelli fosse di migliore qualità rispetto ai metodi tradizionali. Ciò ha portato a generare quella che è stata chiamata la *Data Hubris*, una arroganza dei dati come strumento di affermazione della verità. L'esempio che più di ogni altro è stato studiato come fenomeno in questa direzione è Google Flu Trends, vedi [Butler 2013] e [Lazer 2014].

In molti paesi esistono da tempo strutture sanitarie che hanno tra i loro compiti quello di monitorare l'andamento nel tempo della percentuale di persone che presentano patologie epidemiche, il caso più usuale è quello della influenza. Questo permette di pianificare gli interventi sia per la prevenzione che per la cura. Negli Stati Uniti, in particolare, esiste un sistema di Centri per la prevenzione e controllo delle malattie (centri PCM nel seguito), che nel 2013 comprendeva [Butler 2013] circa 2.700 strutture sanitarie che registravano circa 30 milioni di visite di pazienti ogni anno.

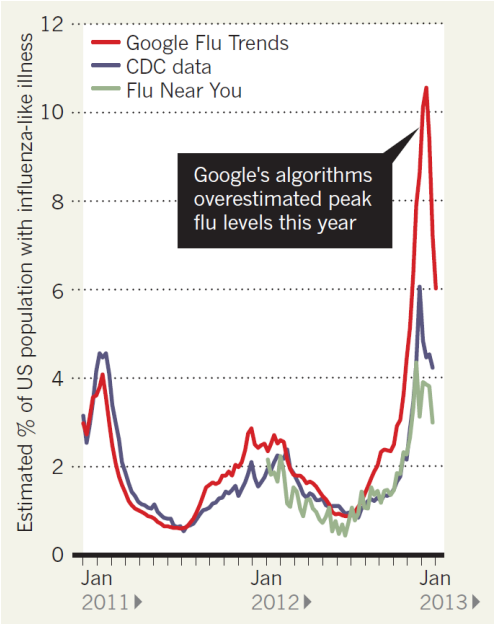


Figura 3 - Confronto tra tre metodi di stima della percentuale della popolazione degli Stati Uniti con patologia influenzale (tratta da https://www.nature.com/news/when-google-got-flu-wrong-1.12413)

In quegli anni, le reti sociali sono state investigate per capire come il diffondersi di epidemie potesse essere valutato con metodi più precisi e più rapidi delle rilevazioni effettuata dai Centri PCM. Google nel 2008 lanciò il sistema Google Trends, che si basava sulla analisi testuale delle parole chiave sottoposte al motore di ricerca di Google, combinata con algoritmi di modellazione predittiva.

Nei primi tempi, le stime di Google trends erano in stretta corrispondenza con le misurazioni risultanti dal sistema dei CPCM. Il sistema venne distribuito in 29 paesi in tutto il mondo, e fu esteso per includere il monitoraggio di una seconda malattia, la dengue.

A partire dal 2013 il sistema ha cominciato a produrre risultati che si discostavano significativamente da quelli del sistema dei centri PCM (vedi Figura 3). La stima di Google Flu Trends per il picco di Natale 2013 è stata almeno doppia di quella del sistema dei centri PCM, e in alcuni stati americani le discrepanze sono state anche maggiori.

In effetti, Google Flu Trends ha sovrastimato la diffusione della influenza anche in periodi temporali precedenti il 2013. Gli errori, peraltro, non sono distribuiti casualmente; per esempio, gli errori della settimana n sono predittori degli errori nella settimana n+1, e la direzione e grandezza degli errori varia durante l'anno, introducendo una stagionalità. Questi pattern dimostrano che l'algoritmo trascurava una quantità significativa di dati che poteva essere estratta mediante i metodi statistici tradizionali.

Riguardo al livello di trasparenza offerto da Google sulla possibilità di analizzare gli algoritmi utilizzati, i dati messi a disposizione da Google (vedi [Lazer 2014]) non erano adeguati agli standard stabiliti dalla comunità dei ricercatori. Inoltre, Google non ha mai fornito i 45 termini su cui basava la ricerca. Occorre osservare che era peraltro impossibile per Google fornire un corpus di informazioni più ampio, né sarebbe stato eticamente accettabile, dati i problemi di privacy. I vincoli alla diffusione, naturalmente,

non riguardavano i dati aggregati. Anche se uno avesse avuto accesso a tutti i dati di Google, sarebbe stato impossibile replicare le analisi prodotte da Google.

Google Flu Trends non è più attivo da tempo, resta nel sito la possibilità di accedere a serie storiche e utilizzare dati a scopi di ricerca. Goolge Flu Trends è talvolta visto come esempio di "Big data hubris", di arroganza dei Big data, tutte le volte in cui si assume che i big data sono un sostituto, piuttosto che un complemento, alle tradizionali fonti di raccolta e analisi dei dati.

4. Correlazione e causazione

[Lehikoinen 2014] e [Borrowman 2014] tra gli altri analizzano a fondo il senso della affermazione "correlation is not causation", che tradurrò nel seguito con la frase "la correlazione non implica la causazione". Per causazione intendiamo ne seguito Il complesso di fattori che fondano la relazione tra causa ed effetto. Il seguente testo è liberamente ispirato ai precedenti due riferimenti.

Per [Lehikoinen 2014] la perdita di una prospettiva orientata alla *ricerca delle cause* nei fenomeni del mondo reale porta a porsi domande su varie tematiche di grande rilievo:

- La responsabilità – Chi è responsabile delle decisioni e delle loro conseguenze che si fondano unicamente su modelli predittivi basati su big data?
- L'apprendimento – Quando non conosciamo le cause di un fenomeno, non apprendiamo; non c'è alcun modo di sviluppare un sistema di riferimento cognitivo basato su esperienze pregresse, se ogni volta che abbiamo bisogno di prendere una decisione, deleghiamo la decisione ai dati e ai modelli costruiti per mezzo di essi.
- La fiducia – E' molto più difficile avere fiducia nei risultati di un modello basato sui dati se non si può comprendere in maniera trasparente la natura del modello. Noi per nostra natura non siamo portati ad assegnare a un algoritmo un carattere di autorevolezza; peggio, se la predizione appare scorretta (e' il caso di falso positivo) può essere impossibile intraprendere in futuro azioni correttive. Per esempio, se per qualche ragione Facebook arriva alla conclusione che a me piace il pesce anche se io sono letalmente allergico al pesce, riceverò per sempre raccomandazioni di mangiare il pesce, indipendentemente da quel poco che è mio potere fare.

L'analisi di [Borrowman 2014] è estremamente ampia. la causazione è stata a lungo un mistero, tormentando i filosofi e i ricercatori nei secoli. Le domande fondamentali sono: cosa è la causazione? Come può essere misurata la "forza" della causazione? Che segnali ci dà una correlazione su un possibile legame causale tra fenomeni? In che modo fattori multipli influenzano congiuntamente un fenomeno? E, una delle domande fondamentali dei filosofi: La relazione causa-effetto esiste "nel mondo", o è soltanto una nostra costruzione mentale, una connessione che noi tratteggiamo tra due eventi che abbiamo osservato in successione tante volte? Non è possibile riassumere in poche pagine la ricchezza della elaborazione concettuale e degli esempi che compaiono in [Borrowman 2014], discutiamo qui per punti alcune tesi tra le più rilevanti, suddividendole in argomenti. Allacciate le cinture!

Introduzione al legame tra correlazione e causazione

- Quando parliamo di causazione noi diamo per scontata una qualche nozione su che cosa sia, e un qualche metodo che possiamo usare per valutarne la forza. Alcuni legami causali ci appaiono forti, quando diciamo che una cura ha guarito una malattia, altri vengono percepiti in forma più debole, come quando diciamo che la detenzione di un leader dell'opposizione ha influito su una valutazione negativa della opinione pubblica internazionale.

- In tempi recenti, è diventato diffusamente accettato il fatto che in diversi campi le decisioni dovrebbero essere basate sulla evidenza (evidence based), nel senso che la conoscenza dei risultati della decisione, basata su studi scientifici e altre fonti empiriche, dovrebbe orientare le nostre scelte; inoltre ci aspettiamo che tali scelte *causino* i risultati finali. Come conseguenza, la statistica è diventata progressivamente sempre più importante; peraltro, basta leggere con attenzione un giornale, una intervista a un politico, ovvero un suo messaggio Twitter, per rendersi conto di quanto la statistica venga distorta, nel senso di sovrastimare i legami causali (es. il deficit è tutta colpa del governo precedente) ovvero non indagare su sfumature più sottili di tali legami.

- Il fenomeno dei big data sta radicalizzando questa visione (vedi ancora [Anderson 2009], portandoci a pensare che con la potenza di calcolo disponibile potremo scoprire tutte le correlazioni rilevanti, e queste saranno così forti e decisive nel produrre modelli predittivi che potremo fare a meno della causazione. Al contrario, è fondamentale indagare ancora di più sulla relazione tra correlazione e causazione, entrando anche nel merito dei diversi fattori che influenzano positivamente ovvero minano alle radici i due concetti.

Problemi legati alla causazione

- Per esempio, dire che il fumo causa il tumore al polmone, se ci pensiamo bene non esprime né una condizione sufficiente, né una condizione necessaria tra i due aspetti. Molti ricercatori hanno investigato sull'effetto congiunto di diverse cause coinvolte nella causazione, ed è diventato naturale esprimere la causazione per mezzo di probabilità. Avere un tumore al polmone può dipendere da diverse cause, come la esposizione a componenti chimici dannosi, la predisposizione genetica e l'età. Alcuni fattori possono essere addirittura sconosciuti, altri ancora non compresi nella loro interezza, o non misurabili. Pensare la causazione in termini di probabilità ci permette di semplificare il problema considerando a parte alcuni di tali fattori, almeno tentativamente.

- Il dibattito scientifico su correlazione e causazione nasce a metà del diciannovesimo secolo quando il filosofo Hume osservò che noi non possiamo mai osservare direttamente la causazione, ma piuttosto soltanto "la congiunzione costante di due oggetti". La causazione fu a lungo trascurata come concetto da indagare, e anche K. Pearson, l'inventore nel 1911 del coefficiente che misura la forza della correlazione tra due variabili, svalutò la causazione come "un altro feticcio tra i concetti arcani e imperscrutabili della scienza moderna". Ma già dagli anni 20 del secolo scorso si cominciò a sviluppare un insieme di studi che indagarono inizialmente i problemi e le "trappole" nascoste nel concetto.

- Una fonte di confusione riguardo alla causazione è che essa esprime in forma sintetica un concetto decisamente più complesso da comunicare, e perciò è molto adatto ad essere utilizzato nelle moderne forme di comunicazione che non lasciano spazio alla riflessione, come abbiamo discusso in precedenza. Confrontiamo ad esempio le due frasi seguenti: "le cinture di sicurezza salvano vite" e "l'uso di cinture di sicurezza è associato ad una minore mortalità"; la prima esprime in modo decisamente più diretto il messaggio, la seconda è certamente più precisa e veritiera della prima, ma appare al contrario elusiva e volutamente oscura. Ciò non toglie che la causazione non possa sostituirsi alla correlazione, salvo forzare in modo non logicamente corretto un assunto. Negli anni 90' del secolo scorso J. Gottman studiò il legame tra atteggiamenti in una coppia, come ad esempio stare sulla difensiva, creare incomunicabilità ecc. e i divorzi, riuscendo a prevedere quali coppie avrebbero divorziato con una accuratezza del 94%. Alcuni giornali suggerirono ai lettori che avrebbero potuto ridurre significativamente il rischio di divorzio cambiando il modo in cui comunicavano con il coniuge. Questo messaggio non corrispondeva ai risultati del lavoro di Gottman, perché la correlazione non implicava che il divorzio era dovuto a quei fattori, né che cambiare tali atteggiamenti avrebbe cambiato il risultato finale.

- Come è possibile che un insieme di fattori sia correlato e non collegato da legami causali? Una ragione è il puro caso, si parla in questo caso di *correlazioni spurie*, contingenza che, come abbiamo detto, diventa *più* probabile, e *non meno* probabile, incontrare nei big data. Gli statistici hanno sviluppato teorie e strumenti per trattate il problema, chiamato anche della *significatività statistica*. Il concetto è talvolta usato in modo tale da generare idee errate, come la credenza che la correlazione non implica la causazione *a meno che la correlazione* non sia statisticamente significativa. L'errore in questa credenza è facilmente individuabile nei big data, in cui una correlazione tra dati è virtualmente garantita come statisticamente significativa. Il grande volume dei dati non dà nessuna garanzia in merito alla esistenza di una relazione causale.

- Un'altra ragione ben nota per cui due fattori possono essere correlati senza avere nessun legame causale sta nel fatto che essi possono avere una causa comune derivante dal legame con un terzo fattore, si parla in questo caso di *fattore confondente*, che va scoperto, a meno di non ricadere in false conclusioni, come negli effetti di cure e medicine.

- Altre correlazioni ingannevoli possono nascere dalla distorsione nella selezione (selection bias) del gruppo selezionato in uno studio, che riguardi, per esempio, la relazione nelle donne tra impianti per il seno e malattie del tessuto connettivo; se i due fattori caratterizzano in modo distorto la scelta del campione la forza della correlazione può essere differente dal vero. Questo tema, ben noto in statistica, diventa di estrema rilevanza nel mondo dei big data, in cui spesso si selezionano fonti che forniscono dati raccolti con procedure non note e non impostate in modo tale da evitare a priori le distorsioni.

- Ancora, non è neanche vero ciò che potrebbe sembrare evidente, e cioè che la causazione implichi la esistenza di una correlazione; si pensi a un impianto di riscaldamento che alimenta il calore in una abitazione in cui un termostato mantiene la temperatura costante; pur essendoci una causazione tra energia termica e temperatura, non c'è correlazione, rimanendo la temperatura costante al

variare della energia termica sviluppata. Questo è un esempio di sistema con retroazione, in cui i legami tra causazione e correlazione si perdono.

- Anche in assenza di un legame di causazione, le correlazioni possono essere estremamente utili. I sintomi di una malattia sono vitali per arrivare a una diagnosi, e certi indicatori economici possono far presagire una recessione; gli statistici chiamano queste informazioni *segnali*. Le compagnie di assicurazione sono interessate a correlazioni tra fattori di rischio, indipendentemente dal rapporto di causazione. Per esempio, se un certo modello di automobile è a più alto rischio di incidente, la compagnia deciderà per un premio più alto, e si disinteresserà dei motivi del più alto rischio. Secondo una prospettiva di sicurezza pubblica, al contrario, conoscere i motivi degli incidenti favorisce la pianificazione di interventi per mitigare il rischio.

- Un altro aspetto cruciale della relazione di causazione, è la *direzione* della relazione, quale fattore sia causa e quale sia effetto. Quando due eventi siano correlati, e un evento A accade prima e un secondo evento B accade dopo, siamo tentati di concludere che il primo evento causa il secondo, ma tutta la discussione precedente dimostra che la relazione può essere dovuta al caso, o a una delle distorsioni di cui abbiamo discusso. Tutto ciò ha grande importanza nei processi civili o penali. In questi casi la proposizione "la correlazione non implica la causazione" può diventare uno strumento argomentativo generale per indebolire una accusa.

I controfattuali

- E arriviamo alle affermazioni di Anderson con cui abbiamo iniziato il capitolo. La precedente discussione ci dice che è vero che le correlazioni possono essere utili, soprattutto per costruire predizioni, ammesso, naturalmente, che non siano dovute al caso o a distorsioni. Ciò che le correlazioni non possono fare è dirci cosa accadrebbe se noi volessimo intervenire sulla realtà, cambiandone qualche aspetto. Per questo, dobbiamo conoscere se esiste una relazione causale. Supponiamo, ad esempio, che uno studio affermi che coloro che bevono il caffè vivono di più. Se una persona che non beve il caffè ragiona in questo senso: bene, mi metto a bere il caffè, così vivrò di più, il suo ragionamento è fondato? Non è fondato, per diversi motivi.

- Il primo motivo sta nel fatto che i bevitori di caffè nello studio avevano probabilmente caratteristiche diverse rispetto a coloro che non lo bevevano (es. con riferimento alle diete alimentari, l'esercizio fisico, ecc.), alcune di queste erano legate al bere il caffè, altre no; tutto ciò non rende automaticamente la persona appartenente al secondo gruppo.

- Il secondo motivo è ben più profondo e importante. Cosa significa "vivono più a lungo?", più a lungo rispetto a cosa?. L'interpretazione più sensata è: più a lungo rispetto a cosa sarebbe accaduto se essi non avessero bevuto il caffè. Per cui se una persona inizia a bere il caffè, bisognerebbe sapere cosa sarebbe accaduto se non avesse iniziato. Questo concetto è anche chiamato *controfattuale*, perché richiede di prendere in considerazione un *mondo parallelo*, il mondo nel quale l'individuo avrebbe vissuto se non avesse bevuto il caffè, rispetto a ciò che effettivamente è accaduto. I controfattuali giocano un ruolo centrale nelle moderne teorie sulla causazione, come vedremo meglio nell'ultima sezione del capitolo.

- I controfattuali portano al cuore di ciò che rende la relazione causa effetto così problematica. Noi possiamo osservare ciò che è accaduto, non *ciò che sarebbe potuto accadere* se non avessimo effettuato la decisione che abbiamo di fatto effettuato. La valutazione di un effetto causale non è possibile senza fare assunzioni o incorporare informazioni esterne al contesto in cui stiamo operando. E mentre noi non possiamo mai osservare direttamente l'effetto causale che sospettiamo essere responsabile della associazione che stiamo osservando, noi siamo in effetti in grado di osservare la *associazione*. Peraltro, la associazione può essere fuorviante in presenza di effetti dovuti al caso o alle distorsioni che abbiamo visto. Per rispondere a una domanda su una relazione causale, il ragionamento controfattuale, che si pone la domanda "cosa accadrebbe se?" è indispensabile. Nessun insieme di big data e nessun potente elaboratore può sostituire questo *passaggio del pensiero.*

Esperimenti e osservazioni

- Le minacce delle distorsioni come ad esempio quella vista in precedenza di *legame confondente*, e le complessità insite nel ragionamento causale appaiono come ostacoli formidabili alla scienza. Gli scienziati hanno concepito strumenti potenti per aggirare queste difficoltà, e cioè *l'esperimento*. In un esperimento, gli scienziati manipolano le condizioni, tenendo alcuni fattori costanti e variando i fattori di interesse nel corso di molte ripetizioni, e misurano i risultati. Quando è possibile fare ciò (e anticipiamo che non sempre è possibile) si possono ottenere inferenze valide nella relazione tra causa ed effetto. Ciò, peraltro, è vero anche nel caso di dati osservazionali, ovviamente sotto opportune ipotesi.

- Nel momento in cui le tecniche scientifiche sono state estese alle scienze sociali nel diciannovesimo secolo, si è dovuto condurre gli esperimenti in contesti così complessi che è stato spesso impossibile controllare tutti i fattori rilevanti dell'esperimento. E' intervenuto a questo punto un secondo elemento decisivo, la *randomizzazione*, cioè la scelta casuale dei soggetti o fenomeni o eventi coinvolti dall'esperimento.

- L'esperimento randomizzato e controllato (randomized control trial) permette di superare i problemi legati alle distorsioni, e, inoltre, fornisce una risposta decisiva a ciò che fino ad ora ci è sembrato un ostacolo insormontabile, il controfattuale. Basta infatti dividere casualmente i soggetti dell'esperimento in due insiemi, e, per simulare il controfattuale, ad esempio in un esperimento su una medicina, basta prescrivere la medicina al primo gruppo, e non prescriverla al secondo, realizzando in tal modo il "cosa accadrebbe se non" di cui alle discussioni precedenti.

- Purtroppo gli esperimenti non sono sempre possibili, per ragioni di privacy o per ragioni etiche. Per esempio, non possiamo certo organizzare esperimenti che incidono negativamente sulla salute di pazienti. Si effettuano in questi casi, invece che esperimenti, *osservazioni*. Si effettuano osservazioni in tutti gli studi sulle cause delle malattie. Naturalmente, gli studi basati su osservazioni presentano problematiche legate alle distorsioni. Ad esempio, nel misurare la forza dei legami nelle reti sociali, si adotta come parametro la frequenza dei contatti, spesso utilizzato nella analisi di configurazioni tipiche (pattern) di sequenze di chiamate telefoniche; in questo caso può subentrare un *fattore di confusione interpretativa*, come per esempio accade quando abbiamo contatti frequenti con

soggetti con cui abbiamo legami deboli, ad es. nelle interazioni di routine, ma superficiali, tipiche delle chiamate a una compagnia di taxi.

- Per contrastare i fattori di confusione interpretativa, negli studi basati su osservazioni, sono state prodotte tecniche opportune. Per esempio nello studio della incidenza dell'ordine di nascita (primo figlio, secondo figlio, ecc.) sulla sindrome di down, l'età della madre è un fattore di confusione, perché l'età cresce con l'ordine di nascita. Si può effettuare allora una analisi stratificata, che riguarda separatamente i diversi gruppi di figli a seconda dell'ordine. Anche in questo caso, resta la possibilità che esistano ulteriori fattori di confusione, e ciò può essere visto come un limite intrinseco dei dati raccolti mediante osservazione, limite che può dar luogo a controversie legali e riduzione di efficacia nei dibattimenti simili a quelli visto in precedenza.

La relazione di causazione ai tempi nostri

- Fino ad ora abbiamo "girato attorno" al tema della relazione causa effetto. La domanda a questo punto è: sono stati sviluppati metodi per inferire le cause di un certo fenomeno su cui sono state individuate correlazioni?

- Certamente, fin dagli anni 20 del secolo scorso sono stati proposti metodi. Ad esempio, S. Wright introdusse la "path analysis" (o analisi dei cammini), che consiste nel costruire grafi che rappresentino insieme tutte le potenziali relazioni di causa effetto; partendo da questi grafi e dalle correlazioni osservate, vengono costruiti sistemi di equazioni, la cui soluzione porta a individuare coefficienti che rappresentano i diretti effetti dei fattori coinvolti in ogni relazione causa effetto.

- Negli anni 80 del secolo scorso sono stati introdotti i grafi diretti aciclici, su cui rimandiamo il lettore alla bibliografia che appare in [Borrowman 2014]. Concludiamo con questa nota la sezione, non senza aver informato il lettore interessato che troverà in [Borrowman 2014] altri spunti interessanti che non hanno trovato qui spazio.

5. Dai piccoli dati ai grandi dati: è tutto oro quel che luccica?

[Markus 2014] offre una interessante lista di contesti in cui i big data (nel seguito grandi dati) e le tecniche e modelli su essi formulabili presentano limitazioni e distorsioni rispetto agli obiettivi per cui vengono usati. Vediamoli.

Mentre usando i grandi dati rispetto ai piccoli dati il numero di correlazioni che le tecniche sono in grado di scoprire cresce, le tecniche non sono in grado di individuare quali correlazioni siano significative. Inoltre, la percentuale di correlazioni veramente significative decresce, per la "casualità" con cui le correlazioni vengono scoperte; quindi, in assenza di un controllo umano, cui in genere si rinuncia proprio per la mole esponenziale di lavoro, si rischia di generare molto più rumore che sostanza.

Una questione collegata alla precedente porta a percepire i grandi dati come fornitori di soluzioni che appaiono valide scientificamente, pur essendo i problemi formulati in modo impreciso. Sono stati

effettuati studi per fornire graduatorie di letterati sulla base della importanza storica ovvero dei contributi culturali rilevanti. In uno di questi, Francis Scott Key, noto per aver scritto i versi dell'inno degli Stati Uniti, è collocato al diciannovesimo posto tra tutti i più importanti poeti della storia.

I grandi dati aiutano efficacemente nelle analisi, ma solo se utilizzati come uno strumento in più rispetto alla ricerca sperimentale tradizionale. Ad esempio, con riferimento al problema della derivazione della struttura tridimensionale delle proteine dalla sequenza del DNA, nessuno scienziato penserebbe di risolvere questo problema semplicemente "lavorando con" i dati, per quanto possa essere potente la analisi statistica e gli algoritmi di machine learning; c'è sempre bisogno di fare riferimento alle leggi della fisica e della biochimica, quindi delle ipotesi e dei modelli di cui Anderson aveva profetizzato la fine.

Quando si abbia una qualche idea della tecnica adottata nella analisi dei grandi dati, è facile ingannarla. La valutazione di saggi letterari di studenti in discipline umanistiche è fatta da modelli che si basano sulla lunghezza dei periodi e su parole valutate come di raro utilizzo, e, in quanto tali, classificate come indicatore di lessico sofisticato; questo, per la alta correlazione con i punteggi assegnati da valutatori umani. Ma una volta che uno studente abbia scoperto come opera l'algoritmo, può ingannarlo scrivendo lunghi periodi, magari di significato incerto, e usando parole oscure, piuttosto che apprendere come formulare un testo in modo chiaro e coerente.

Anche per le tecniche di apprendimento esiste un rischio di effetto "camera dell'eco", di cui abbiamo parlato nel Capitolo 5. Ogniqualvolta la fonte dei dati è essa stessa un prodotto generato da analisi su grandi dati, abbondano possibilità di circoli viziosi. Per esempio Google Translate in tempi recenti manifesta un tasso di imprecisione e di errore decisamente più ridotto rispetto al passato nelle traduzioni da lingua a lingua (ricordo una traduzione automatica dei curricula dei Ministri di un Governo di inizio millennio, in cui un Ministro si era laurea alla "University of Mouthful", invece che alla "Bocconi University"); questo viene ottenuto cercando con una tecnica di learning lo stesso pattern linguistico in due differenti versioni di una voce di Wikipedia. La strategia è ragionevole, ma se si comincia a produrre le voci di Wikipedia in un linguaggio poco comune, mediante traduzione utilizzando Google Translate, un qualunque errore iniziale "infetta" Wikipedia, provocando un effetto rinforzo sulle voci successive.

I grandi dati si comportano bene quando il dominio su cui operano è di uso comune, mentre mostrano forti limiti quando il dominio è specifico e poco frequente. Provate a tradurre successivamente dall'inglese all'italiano e viceversa una frase comune e una frase con parole rare, l'esito sarà di una certa stabilità di traduzione nel primo caso, di una divergenza nel secondo.

Infine, alcuni ricercatori afferma che i grandi dati siano all'apice di popolarità di una curva, cui sta per seguire una lenta o rapida decadenza. Se pensiamo ai risultati ottenuti finora con i big data, non ci vengono in mente scoperte paragonabili alle grandi innovazioni degli ultimi due secoli, come gli antibiotici, l'automobile, e l'aeroplano.

Un'altra caratteristica legata intrinsecamente ai grandi dati è la imprecisione. L'approccio dei grandi dati implica l'utilizzo di più fonti di dati eterogenei o frammentati in "piccoli dati", In molti casi, i dati sono stati originariamente raccolti per scopi completamente diversi e le diverse fonti hanno formati diversi e diversa qualità (accuratezza, completezza, livello di aggiornamento). Abbiamo visto nei precedenti

capitoli i metodi per migliorare la qualità e per integrare dati eterogenei, ma questi metodi sono costosi e richiedono sforzo professionale spesso troppo elevato per i budget disponibili; la conclusione è che, Inevitabilmente, queste diverse fonti di dati portano imprecisioni. Quindi spesso occorre decidere in accordo a un qualche compromesso (tradeoff) tra quantità/qualità che peraltro è difficile da valutare e dipende dalle circostanze.

Un altro aspetto è quello insito nella coordinata temporale nello "spazio" di Figura 1 del capitolo. Il mondo cambia, e quindi, a fronte di modifiche significative dell'ambiente rappresentato dai dati, l'analisi deve essere rifatta. C'è una figura molto bella nel libro [Pearl 2017], riprodotta in Figura 4. Il robot non ha appreso dagli algoritmi di learning che non è opportuno usare l'aspirapolvere la mattina presto, e la persona che si sveglia lo rimprovera addirittura con una argomentazione che nella Sezione 8 chiameremo ragionamento controfattuale (il concetto di controfattuale è già stato introdotto nella Sezione 4).

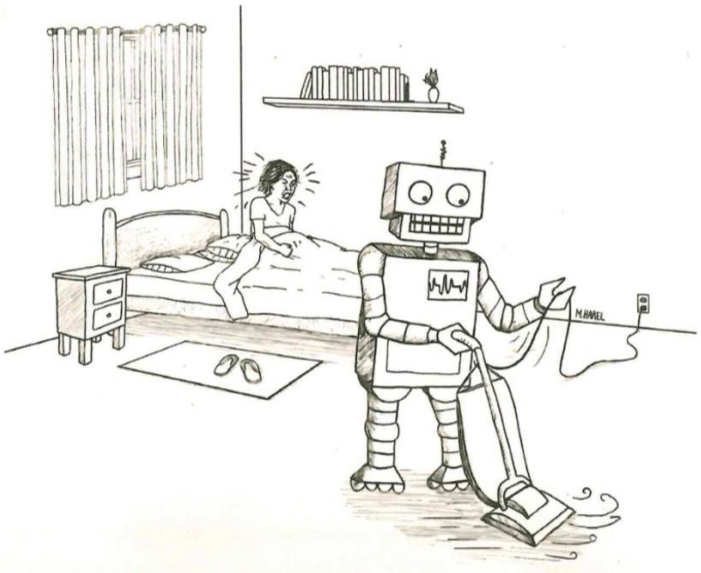


Figura 4 – "Non avresti dovuto svegliarmi!" (tratta da [Pearl 2018])

6. Con la crescente attenzione ai grandi dati, siamo alla fine del metodo scientifico?

Il lavoro [Succi 2018] forse più di ogni altro esprime un insieme di argomenti in cui il tema dei big data e dell'apparato di modelli, tecniche, algoritmi in corso di sviluppo nel mondo viene visto secondo una prospettiva critica e con una previsione di interesse declinante; ad esso ci ispiriamo nel seguito. Il lavoro discute quattro tematiche:

1. i sistemi complessi sono descrivibili mediante dati digitali che, in virtù della complessità del sistema, sono fortemente correlati tra loro e quindi non rispettano la distribuzione normale o Gaussiana (introdotta nel Capitolo 9).
2. il legame tra correlazione e causa si indebolisce al crescere della dimensione dei dati.

3. in un mondo di risorse finite (come quello in cui viviamo), avere tanti dati digitali disponibili diventa equivalente a non averne nessuno.
4. i dati non sono mai sufficienti, per quanto grandi essi siano, nella descrizione di sistemi che siano significativamente sensibili alla accuratezza nei dati.

Non discuteremo il punto 4, il lettore interessato può consultare [Succi 2018].

I sistemi complessi sono descrivibili solo mediante dati che non rispettano la distribuzione Gaussiana

Nelle assunzioni che nella Scienza dei dati si fanno in genere sui dati descrittivi di fenomeni, si ipotizza che i dati siano tra di loro non correlati. Supponiamo di provocare successivi schizzi di acqua per il tergicristallo della nostra automobile; ogni schizzo modifica sia pure impercettibilmente il grado di umidità e di polveri sottili dell'ambiente circostante, per cui lo schizzo successivo opera su un ambiente un po' diverso da quello su cui ha operato lo schizzo precedente; dunque, se descriviamo gli schizzi s1 e s2 mediante dati digitali d(s1) e d(s2), che descrivono ad esempio la posizione dello schizzo, l'istante di emissione e la portata, i dati d(s1) e d(s2) non descrivono fenomeni completamente indipendenti.

Diversi "pilastri" della statistica, come

- la legge dei grandi numeri, che stabilisce che in presenza di incertezza associata ad un insieme di n valori (o dati descritti da numeri), all'aumentare di n l'incertezza, ad esempio, sul valor medio diminuisce.
- la convergenza verso il valore vero m* di parametri statistici come il valor medio m associati ad un insieme di n numeri (o dati descritti da numeri
- la distribuzione Gaussiana o normale "a campana"

hanno validità solo se a. i dati sono tra di loro non correlati, cioè descrivono fenomeni indipendenti e b. la varianza o grado di diversità dei valori dei dati è finita (cioè i dati sono "simili tra di loro" e sono percentualmente pochi i dati con valori anomali.

La distribuzione Gaussiana rispetta molte importanti proprietà, tra cui la seguente: i dati anomali (outlier) sono molto rari nella distribuzione, e siccome sono essi i responsabili del livello di incertezza esistente nei dati, l'incertezza è anch'essa minimizzata e "soppressa" al crescere dei valori che essa rappresenta. Questa sembrerebbe una ottima notizia per i grandi dati; al crescere del volume dei dati, l'incertezza diminuisce, l'esatto contrario di quanto abbiamo ipotizzato nel capitolo dedicato alla qualità. Purtroppo, i grandi numeri non rispettano le due ipotesi a. e b. descritte in precedenza, come dimostra l'esempio degli schizzi d'acqua prodotti nel tempo proposto in precedenza. E quanto più i dati crescono in numerosità, tanto più aumentano i valori anomali, cioè le diversità tra dati, e ci si allontana dalla distribuzione Gaussiana e dalla sua tendenza a rappresentare un mondo molto stabile, conforme e normale al suo interno.

Aumentando il volume dei dati nelle tre dimensioni di Figura 1, aumenta il livello di dettaglio con cui possiamo rappresentare la realtà, ma, aumentando il livello di dettaglio, aumentano le proprietà (o attributi, nella terminologia del modello relazionale discusso nel Capitolo 3) che non sono tra loro indipendenti; insomma, i grandi dati ricadono nei grandi principi della matematica e della logica: quanto più cresce la loro capacità di rappresentare il mondo, tanto più manifestano al loro interno una

complessità che li allontana dalla possibilità di essere modellati in termini di proprietà statistiche "trattabili". E' una illusione quella insita nella possibilità che i numeri parlino da soli; man mano che crescono, passatemi la metafora, i numeri cominciano a balbettare e a emettere suoni inarticolati….

Prima di concludere questo punto, ricordo, peraltro, che un altro lavoro [Halevi, Norvig, Pereira 2009] discusso nel Capitolo 5 Sezione 8 si arrivava a una conclusione opposta, i grandi dati manifestano al loro interno quella che abbiamo chiamato una "irragionevole efficacia".

Il legame tra correlazione e causa si indebolisce al crescere della dimensione dei dati

Abbiamo a lungo discusso nella Sezione 4 che anche se due caratteristiche di un fenomeno di interesse sono caratterizzate da un'alta correlazione, questo non necessariamente implica che sono collegate da un nesso causale. Vi possono essere correlazioni false ovvero correlazioni vere, queste ultime soltanto esprimono una correlazione che segnala una connessione causale.

Il problema generale di distinguere tra correlazioni false e vere, e correlazioni spurie, non è semplice da affrontare: distinguere tra correlazioni false e vere rimane un'arte, e il problema si colloca quindi tra quelli importanti e allo stesso tempo difficili da risolvere nella Scienza dei dati. Il fatto non piacevole è che le correlazioni false crescono molto più rapidamente delle correlazioni vere; come provato in [Calude 2017] il rapporto tra correlazioni false e vere è una funzione che cresce molto rapidamente con la dimensione dei dati.

Occorre peraltro ancora ricordare che i dati digitali rappresentano sempre fenomeni della realtà, in questa rappresentazione il fatto che siano grandi dati non garantisce necessariamente che il fenomeno sia rappresentato nella sua completezza; un altro risultato che esprime un limite della Scienza dei dati è discusso in [Meng 2014], dove si dimostra che per poter esprimere delle inferenze statisticamente affidabili attraverso un algoritmo di machine learning, è necessario poter accedere a una percentuale sostanziale (superiore al 50%) dei dati che rappresentano il fenomeno di interesse.

Resta una questione aperta quale sia, e da quali leggi sia disciplinata, la relazione tra la dimensione dei dati e l'affidabilità dei modelli predittivi basati su Machine learning. Ciò che è richiesto nell'ambito della Scienza dei dati e del Machine learning sono molti più studi, teoremi, condizioni, che specifichino in maniera chiara il dominio di validità dei metodi e la dimensione dei dati necessari per produrre modelli statisticamente affidabili.

In un mondo di risorse finite avere tanti dati disponibili può diventare equivalente a non averne nessuno

Vi è una ultima questione legata ai grandi dati, che afferisce alla epistemologia e alle scienze sociali, perché ha a che fare con una caratteristica tra le più apprezzabili di un essere umano, la intuizione o saggezza (wisdom), che corrisponde alla capacità di prendere la decisione giusta. La intuizione è spesso rappresentata al vertice di una piramide costituita da quattro livelli (vedi Figura 5), che corrispondono ai dati, la informazione, la conoscenza e la saggezza. E' la capacità di trasformare dati in informazioni, queste in conoscenza, e infine la capacità di utilizzare la conoscenza facendo affidamento alla saggezza, che ci porta a prendere decisioni ben informate ed efficaci.

Figura 5 – La piramide dati-informazioni-conoscenza-intuizione/saggezza

La piramide dà una immagine ingannevole della relazione tra dati e saggezza, come se a più dati corrisponda più saggezza. Piuttosto, tutti i sistemi con capacità finita rispondono a un fenomeno di saturazione non lineare, un concetto ben noto nei sistemi complessi, in cui una maggiore disponibilità di dati porta a saturazione e talvolta a perdita di informazione. Discutiamo questo fenomeno.

Un ben noto esempio di saturazione non lineare è la funzione logistica di crescita della popolazione. Se x è il numero di individui di una specie che si riproducono a un tasso t > 0, in un mondo a risorse illimitate la crescita della popolazione è esponenziale; in un ambiente finito a risorse limitate, con spazio e cibo finito, la crescita non può durare all'infinito, e la competizione per le risorse porta all'esaurimento della espansione, vedi Figura 6.

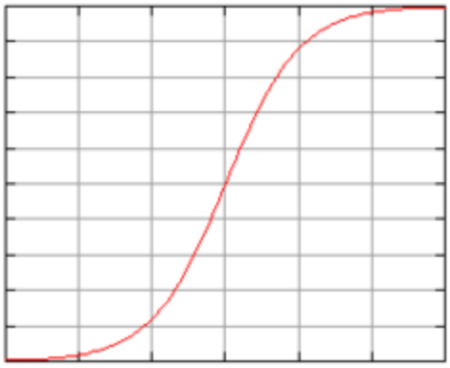


Figura 6 – La funzione logistica di crescita della popolazione

La precedente legge si riferisce alla crescita della popolazione, e non alla crescita della informazione e della conoscenza in funzione dei dati disponibili. Tuttavia, come mostrato qualitativamente in Figura 7, anche nel mondo dei dati digitali è nostra esperienza comune che oltre una certa soglia, avere a disposizione più dati non porta a più informazione, conoscenza, saggezza, ma, piuttosto, subentra un fenomeno di saturazione, detto *data overload*, per cui oltre un certo limite non siamo più in grado di creare valore dai dati, e nuovi dati creano confusione e degrado della informazione. Abbiamo iniziato a

discutere questo fenomeno nel Capitolo 13 sulla economia digitale, riproduciamo in Figura 7 l'analoga figura che compare in quel capitolo.

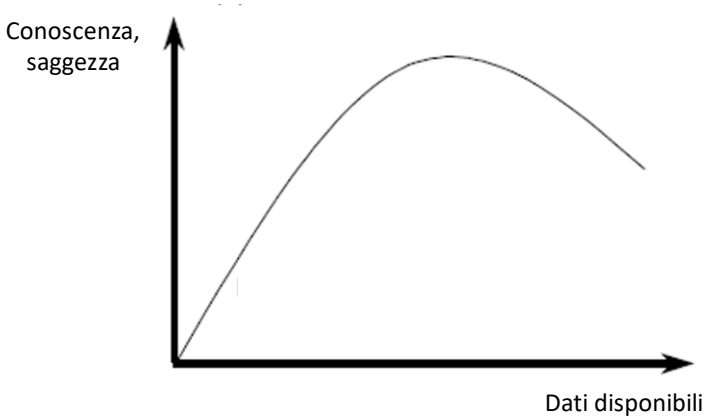


Figura 7 – Il fenomeno del data overflow

In molti dei problemi di cui abbiamo parlato in questo libro, si pensi per esempio al Problema 2 del Capitolo 1 sul giorno più opportuno in cui acquistare un biglietto aereo, la crescita dei dati porta a un miglioramento della predizione. Qui si vuole piuttosto esprimere il fatto che nella accezione più generale, la crescita incontrollata dei dati nello spazio di Figura 1 di questo capitolo porta a dati che possono essere ridondanti, eterogenei, contraddittori, caratterizzati da qualità deteriorata, per cui la nostra razionalità limitata e la immaturità degli algoritmi di gestione nel ciclo di vita del dato, portano a dei limiti intrinseci nella nostra capacità, e nella capacità degli algoritmi, di generare valore dai dati, inteso come intensità e ricchezza della conoscenza e della saggezza.

Come commento personale, e deviando dal tema conduttore della sezione, quando mi sono occupato di astrazioni, tema che abbiamo visto nel Capitolo 12, ho immaginato che la curva di Figura 7 non sia ineluttabile, e che il raggiungimento del valore massimo e il decadimento successivo possa essere ritardato se noi effettuiamo delle operazioni di astrazione, il cui esito, vedi Figura 8, è quello di innalzare il livello di conoscenza complessiva creata a partire dagli stessi dati, e "ritardare" il punto di saturazione oltre il quale vi è decadimento.

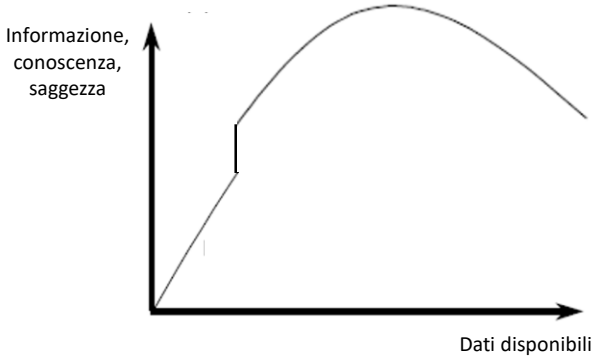


Figura 8 – Aumento di conoscenza mediante astrazione

7. Modelli e funzioni - Il punto di vista di Adnan Darwiche

In questa sezione presentiamo il punto di vista di *Adnan Darwiche*, così come lo abbiamo compreso ed elaborato leggendo il suo lavoro intitolato "*Human-Level Intelligence or Animal-Like Abilities?*", pubblicato nel 2017.

I tanti successi riscossi recentemente tramite le *reti di neuroni artificiali*, rinominate con il termine di *deep learning*, in domini applicativi come quello del riconoscimento del parlato, dell'analisi delle immagini, e della navigazione autonoma hanno portato grande eccitazione nella comunità scientifica dell'*Intelligenza Artificiale*, negli addetti ai lavori e nella cosiddetta società civile. In tempi incredibilmente brevi, se visti alla luce dell'orologio del progresso scientifico, si è stati in grado di automatizzare compiti che per anni ci avevano visto sconfitti applicando le diverse tecniche prodotte dai migliori ricercatori del settore.

I risultati ottenuti e la loro enfatizzazione stanno veicolando l'idea che l'automazione di questi compiti abbia raggiunto un livello di intelligenza paragonabile a quello umano. Questa percezione, emersa dall'Accademia, ha riscosso crescente credito e ha portato ad alcune implicazioni. Ad esempio, sta emergendo una linea di pensiero in base alla quale la ricerca in ambito *Machine Learning* sia destinata ad essere dominata dalle *reti di neuroni artificiali* o *deep learning* che dir si voglia. Questa linea di tendenza ha portato alcuni a chiedersi se valesse la pena di continuare ad investire in altri approcci di *Machine Learning*, o in aree di grande importanza nell'*Intelligenza Artificiale* come la *rappresentazione della conoscenza*, il *ragionamento simbolico* e la *pianificazione*.

Alcune pubblicazioni ed iniziative a livello divulgativo, lasciate a figure di rilievo, hanno mescolato la comprensibile eccitazione per i risultati ottenuti con un senso di timore e financo di paura dell'impatto che l'Intelligenza Artificiale potrebbe avere sulla vita di tutti noi. *Stephen Hawking*, illustre fisico ebbe a dire "*Un pieno sviluppo dell'intelligenza artificiale potrebbe significare la fine della razza umana*". Un altro monito è stato recentemente lanciato da *Elon Musk* il quale ha avuto ad affermare "*L'intelligenza Artificiale è potenzialmente più dannosa del nucleare*".

La comunità dell'Intelligenza Artificiale si trova oggi di fronte ad un dilemma. Da una parte, non si può che essere impressionati e godere di quanto è stato ottenuto utilizzando le reti di neuroni artificiali. Dall'altra parte, permane il comprensibile scetticismo nell'accettare che un metodo che non richiede alcuna modellazione esplicita o ragionamento sofisticato sia in grado di riprodurre un livello di intelligenza paragonabile a quella umana. Questo dilemma è ulteriormente amplificato dal fatto che i recenti sviluppi non sono culminati in una chiara e profonda scoperta scientifica, come una nuova teoria della mente in grado di richiedere un aggiornamento massivo dei principi e dei contenuti dell'Intelligenza Artificiale. Gli studiosi di aree di ricerca differenti dall'Intelligenza Artificiale e dall'Informatica percepiscono crescentemente questo dilemma e si lamentano del fatto che non sono fornite loro risposte intellettualmente soddisfacenti alla seguente domanda "Cosa è appena accaduto di importante nell'Intelligenza Artificiale?"

La possibilità di fornire una risposta al dilemma è legata alla capacità di affrontare in modo attento la valutazione di quanto siamo stati in grado di raggiungere tramite il *deep learning*, e nell'identificare ed apprezzare gli esiti chiave dei recenti sviluppi di questa area di ricerca. Sfortunatamente questo è mancato per parecchi anni, ed inizia solo oggi ad essere argomento di discussione e confronto tra gli addetti ai lavori.

Discutere di un aspetto così articolato impone di porne innanzitutto le basi. Come abbiamo già commentato nella Sezione 2 sul pensiero di Leo Breiman, è fondamentale riconoscere che esistono due approcci distinti per affrontare problemi di Machine Learning, Intelligenza Artificiale e Data Science; l'approccio che abbiamo chiamato model-driven, e quello data-driven, che Darwiche chiama function-based o function-driven, terminologia che per rispetto dell'autore utilizzeremo in questa sezione. Riprendiamo la discussione iniziata nella Sezione 2, approfondendo qui il pensiero di Darwiche.

Consideriamo per esempio il compito di riconoscere se in un'immagine sia presente un cane o un gatto. Svolgere questo compito tramite l'approccio model-driven richiede di rappresentare la conoscenza relativa a cani e gatti e non solo, e coinvolge il ragionamento in un contesto siffatto. In questo caso gli strumenti principali saranno la logica ed il calcolo delle probabilità, in termini generali avremo necessità di modellazione matematica che possiamo pensare essere alla base dell'approccio concepito dai fondatori dell'Intelligenza Artificiale. L'approccio function-driven d'altra parte risolverebbe questo compito come un problema di approssimazione di una funzione dove gli input vengono forniti direttamente dai pixel dell'immagine e con un output corrispondente al riconoscimento o meno degli elementi di nostro interesse; lo strumento principale di questo approccio è costituito dalle reti di neuroni artificiali.

Negli ultimi anni è divenuto palese come l'approccio *function-driven* sia molto più efficace dell'approccio *model-driven* nello svolgere alcuni compiti di Intelligenza Artificiale; questa evidenza ha colto di sorpresa non solo i ricercatori nell'ambito dell'Intelligenza Artificiale ma anche i ricercatori dell'ambito del Machine Learning, che di norma studiano ed utilizzano entrambe gli approcci. Questo accadimento ha avuto molte implicazioni, alcune delle quali molto positive mentre altre hanno sollevato questioni particolarmente importanti e delicate.

Il lato positivo è legato al numero crescente di compiti ed applicazioni che siamo in grado di svolgere, utilizzando uno strumento che è accessibile da persone con cognizioni ingegneristiche di base. L'aspetto che solleva preoccupazione è dovuto allo sbilanciamento tra sfruttamento/beneficio di questo strumento e lo sforzo intellettuale dedicato a riflettere su potenzialità e rischi conseguenti al suo impiego. La carenza di attitudine a riflettere è colpevole delle errate interpretazioni circa il livello raggiunto/raggiungibile dall'Intelligenza Artificiale e su dove essa possa portarci in un futuro prossimo o venturo.

Se si analizzano i tanti successi ottenuti dai modelli basati su reti di neuroni artificiali o del deep Learning, come questo ambito di ricerca è stato rinominato, si nota che essi dicono molto sullo specifico problema affrontato e molto meno sulla *natura* delle reti di neuroni artificiali. Questi modelli calcolano funzioni degli attributi di input utilizzando un'enorme flessibilità messa a disposizione da parametri i cui valori vengono appresi; questo fatto è noto da parecchi decenni e descritto egregiamente in moltissimi libri

sul tema. Allora è naturale chiedersi cosa abbia causato lo scoccare della scintilla che ha innescato la rivoluzione alla quale stiamo assistendo nell'ambito del Machine Learning e dell'Intelligenza Artificiale.

Rispondere a questa domanda passa dal constatare che alcune abilità complesse tipicamente associate alla cognizione, possono essere colte e riprodotte a livelli accettabili semplicemente utilizzando i dati disponibili per approssimare una funzione, senza la necessità di modellazioni esplicite del contesto considerato e del ragionamento simbolico da condurre su di esso. Questo è indubitabilmente un risultato importante, che evidenzia problemi e soglie più dell'aspetto tecnologico. Infatti, ogni comportamento, intelligente o meno, può essere colto utilizzando una funzione che mette in relazione l'input e l'output. In questo contesto le domande chiave sono forse le seguenti:

- Queste funzioni sono sufficientemente semplici da ammettere una rappresentazione compatta tale da consentire di mappare efficientemente input e output come accade nelle reti di neuroni artificiali?
- Se la risposta alla precedente domanda fosse affermativa, disponiamo dell'abilità per stimare queste funzioni a partire da coppie di dati input-output?

Quanto recentemente accaduto nel Machine learning e nell'Intelligenza artificiale è dovuto a tre sviluppi.

Il primo è l'aumentata abilità nel trattare problemi di approssimazione di funzioni a partire dai dati, che è stata determinata da *i*) disponibilità di enormi quantità di dati; *ii*) disponibilità di un'aumentata capacità computazionale; e *iii*) sviluppo di tecniche sofisticate per l'approssimazione di funzioni.

Il secondo sviluppo è dovuto al fatto che abbiamo identificato una classe di applicazioni che oggi sappiamo essere sufficientemente semplici da essere rappresentabili in forma compatta tale da essere calcolabile efficientemente, e la cui stima è fattibile dati i limiti sulla disponibilità di dati, la velocità di computazione e le tecniche di stima a nostra disposizione. Tra queste applicazioni stanno l'identificazione e la localizzazione di diversi tipi di oggetti all'interno di un'immagine e alcuni compiti di riconoscimento del linguaggio naturale e del parlato.

Il terzo sviluppo, che sembra essere passato sotto silenzio, è dovuto al fatto che abbiamo drasticamente cambiato il nostro modo di misurare il successo raggiunto da un modello, riducendo sensibilmente la complessità tecnica della sfida che affrontiamo; in altri termini, siamo divenuti meno esigenti nel valutare la qualità del risultato ottenuto tramite un modello.

È interessante osservare che nessuno degli sviluppi menzionati costituisca di per sè un'innovazione tecnica o una pietra miliare, come per esempio è accaduto nell'adottare il calcolo delle probabilità come fondamento del *buon senso* (common sense) verso la fine degli anni 80, o come accadde per l'introduzione delle *reti di neuroni artificiali* più di 50 anni or sono. Nonostante questo, resta un fatto che la combinazione dei tre fattori di sviluppo che abbiamo menzionato ha avuto un impatto notevole sulle applicazioni reali del Machine learning e dell'Intelligenza artificiale.

Supponiamo che durante un viaggio in aereo siate seduti accanto ad un individuo che, presentandosi, sostiene di essere laureato e quindi di disporre di un buon livello di formazione. A questo punto vi presentate come ricercatore di Machine learning. Il vostro compagno di viaggio, incuriosito da quanto avete detto e da quanto ha potuto leggere ed ascoltare nei media generalisti, vi porge la seguente

domanda “Quali sono gli sviluppi che hanno abilitato i recenti progressi nel Machine learning?” Voi rispondete con la narrazione dell’approccio function-driven ed elencate i tre sviluppi abilitanti. Probabilmente il vostro interlocutore rimarrà impressionato ed intellettualmente soddisfatto. Comunque, se alla sua domanda aveste risposto “È stata appena scoperta una nuova teoria della mente”, non vi sorprendereste affatto se l’interlocutore manifestasse grande preoccupazione in merito ad entità latenti nell’imminenza di rovinare le nostre vite.

La percezione e consapevolezza che la società civile ha in merito ai progressi di Machine learning ed Intelligenza artificiale sono molto importanti. Comunicare cosa ad oggi siamo effettivamente in grado di ottenere, e quello che possiamo verosimilmente immaginare di ottenere in un futuro non troppo lontano, è di enorme rilevanza ed urgenza. A questo scopo è necessario comunicare in modo elementare ed intellettualmente onesto non solo cosa siamo stati in grado di ottenere sino ad oggi sfruttando il Machine learning e l’Intelligenza artificiale, ma anche quale potrebbe essere l’impatto sociale, economico, etico su noi umani. Questi temi sono stati affrontati nei capitoli precedenti del presente libro.

Darwiche ritiene che affrontare temi sociali, economici, legislativi ed etici richieda di porre grande attenzione a considerazioni politiche e normative in grado di guidare e controllare l’impatto potenziale dei nuovi livelli di automazione nei quali la nostra società si sta imbarcando. I risultati impressionanti che sono stati raggiunti nell’ambito del riconoscimento del linguaggio naturale e della traduzione di testi da una lingua ad un’altra, o quelli strabilianti ottenuti nel riconoscimento e della localizzazione delle componenti di un’immagine o ancora quelli ottenuti tramite il *Reinforcement learning* nel gioco del GO, potrebbero farci formulare domande come le seguenti:

- Il livello di successo raggiunto da prodotti commerciali nello svolgere questi compiti giustifica preoccupazioni in merito ad un imminente “giorno del giudizio universale”?
- I successi ottenuti autorizzano a sostenere che abbiamo a disposizione un nuovo modo di formalizzare il buon senso?
- Possiamo affermare di essere in grado di comprendere il linguaggio naturale, testo o parlato, o di disporre di un sistema di visione a livello di quello degli esseri umani?

Concordiamo con quanto sostenuto da *Adnan Darwiche*, e riteniamo che le risposte alle tre domande formulate siano altrettanti “no”. In sostanza, quanto accaduto non è una svolta tale da doversi far preoccupare di scenari da giorno del giudizio universale. Quanto accaduto è l’applicazione di tecnologie sofisticate a classi di applicazioni commerciali che ne hanno tratto enorme beneficio; tuttavia, questi successi estremamente rilevanti in termini commerciali non hanno, a mio modesto parere, raggiunto le ambizioni dell’Intelligenza artificiale.

Sebbene il livello corrente di Machine learning ed Intelligenza artificiale sia abbastanza limitato, il loro impatto sull’automazione e, pertanto, sulla società, può essere notevole (ad esempio nel lavoro e nella sicurezza). Questo richiede di porre grande attenzione alla normativa che regolamenta l’impiego e la diffusione di tecnologie così invasive e pervasive.

Personalmente concordo con quanto dichiarato da Adnan Darwiche, vale a dire ritengo che attribuire un livello di intelligenza umano ai modelli di *Deep Learning* sia discutibile, sebbene questi abbiamo ottenuto risultati strabilianti nell'affrontare e risolvere diversi compiti estremamente complessi; tuttavia, questi compiti possono essere svolti egregiamente anche da molte specie animali. Una riflessione importante penso possa essere fatta a partire dalla seguente frase di *Judea Pearl*:

> *Il sistema di visione di un'aquila e di un serpente dal punto di vista delle prestazioni surclassano ogni modello che siamo ad oggi in grado di sviluppare in laboratorio, tuttavia nè serpenti nè aquile possono costruire occhiali, telescopi o microscopi.*

I gatti dispongono di un'abilità di navigazione di gran lunga superiore a quella disponibile dai sistemi di automazione più sofisticati, inclusi i veicoli autonomi. I cani sono in grado di riconoscere e reagire alla voce umana ed ancora i pappagalli grigi Africani sono in grado di emettere suoni che mimano in modo strabiliante la voce umana; resta il fatto che nessuno di questi animali possiede abilità cognitive e livelli di intelligenza tipicamente attribuiti agli esseri umani.

Un'altra ragione di cautela, in un clima di grande entusiasmo come quello che stiamo vivendo, è costituita dal fatto che si sono già verificati in passato fenomeni simili, grandi aspettative, iperboli narrative, scenari da "*Blade Runner*". Infatti, negli anni 80 vissero un periodo di splendore i *sistemi esperti basati su regole*; questi sistemi promettevano di raggiungere prestazioni superiori a quelle umane in alcuni casi, in particolare in ambito di diagnostica medica. In quei giorni si potevano leggere frasi come "la conoscenza è potere" così come oggi leggiamo di meraviglie ed iperboli imminenti legate al Deep learning. Il periodo che seguì questa ubriacatura di entusiasmi è noto come l'inverno dell'Intelligenza artificiale, e per anni l'ambito di ricerca visse solo grazie agli sforzi ed al lavoro dei ricercatori del settore.

Abbiamo già trattato nel Capitolo 15 il tema della interpretabilità delle funzioni e modelli di learning. Negli ultimissimi anni sta crescendo la consapevolezza, almeno negli addetti ai lavori, circa l'importanza della interpretabilità delle funzioni che vengono apprese tramite modelli di Deep Learning, modelli che di norma non consentono di spiegare i risultati che ci forniscono. Ad esempio, se un sistema diagnostico medico basato sul Deep learning suggerisce la necessità di sottoporre un certo paziente ad un intervento chirurgico, il paziente ed il medico di fiducia vorrebbero conoscere il perché di tale suggerimento. Se un'autovettura a guida autonoma uccide un essere umano, dovremmo essere in grado di conoscerne il perché. Alla stessa stregua, se un sistema di riconoscimento facciale induce ingiustamente ad arrestare un individuo, anche in questo caso dovremmo essere resi partecipi del perché ciò sia stato suggerito. Rispondere alla domanda *Perché?* è centrale per attribuire giudizi, responsabilità ed eventuali colpe che sono alla base dei sistemi giuridici. Certamente gli approcci model-driven possono essere utilizzati per rispondere a tali domande, ma gli approcci function-driven, allo stato attuale della ricerca, certamente no.

8. Il punto di vista di Judea Pearl e la scala della causalità

Il testo intitolato "*The Book of Why*", scritto da Judea Pearl e Dana McKenzie, tratta in modo egregio la differenza tra un modello causale ed un approccio data-driven, sebbene nel testo non venga impiegato

esplicitamente il termine data-driven. Il testo pone in evidenza quello che un modello è in grado di fare, e che invece un approccio data-driven non è in grado di ottenere.

È interessante osservare come il primo capitolo del libro inizi con una discussione in merito all'evoluzione a partire da 40.000 anni fa, sostenendo la tesi che l'abilità di rispondere a domande del tipo "*Cosa succede se?*" e "*Perché?*" sia l'elemento distintivo degli esseri umani. Il testo pone in evidenza tre ostacoli che ci separano tra lo stato attuale della ricerca a quello in cui si potrà forse affermare che l'Intelligenza Artificiale ha raggiunto livelli paragonabili a quelli dell'essere umano.

Il primo ostacolo è rappresentato dall'*adattabilità* o *robustezza* di un modello. I ricercatori hanno incontrato e riconosciuto che i sistemi attuali di *Machine learning* non sono in grado di riconoscere o reagire adeguatamente a nuove circostanze per le quali non erano stati esplicitamente programmati o sottoposti ad apprendimento. Testimonianza di tale difficoltà è l'intenso sforzo teorico e sperimentale profuso dalla comunità scientifica del Machine Learning sui temi del trasferimento dell'apprendimento, dell'adattamento al dominio e dell'apprendimento continuo, sforzi che allo stato attuale non hanno fornito risposte efficaci.

Il secondo ostacolo è rappresentato dalla caratteristica che abbiamo già discusso, della *interpretabilità*; infatti, allo stato attuale delle cose i modelli di *Machine Learning* rimangono incapaci di spiegare e motivare le loro previsioni o raccomandazioni, erodendo di fatto la quota di fiducia che possono ragionevolmente chiedere ai loro utilizzatori, ed impedendo ai ricercatori di effettuare una diagnosi che consenta di riparare agli errori commessi dai modelli addestrati.

Il terzo ed ultimo ostacolo è dovuto alla mancata comprensione della connessione tra *causa ed effetto*. Questa caratteristica della cognizione umana è, secondo Judea Pearl, vincitore del premio Turing nel 2011, un ingrediente necessario, sebbene non sufficiente, per raggiungere un livello di intelligenza paragonabile a quello umano. Questo ingrediente dovrebbe consentire a un elaboratore di dati digitali, a un robot, a una applicazione su uno smart phone, di disporre di una rappresentazione coreografica parsimonio-sa e modulare dell'ambiente nel quale opera, di interrogare tale rappresentazione, eventualmente di distorcerla tramite l'immaginazione, per consentire di rispondere a domande del tipo "Cosa succederebbe se?". In tale direzione potrebbero essere formulate domande di natura diversa, ad es. domande di natura *manipolativa* come "Cosa accadrebbe se inducessimo un determinato evento?", e di natura *retrospettiva* o *esplicativa* basate su controfattuali come "Cosa sarebbe accaduto se avessi agito in modo differente *da come feci?*" o ancora "*Cosa sarebbe successo se il mio volo non fosse stato in ritardo?*"

Un esempio particolarmente efficace per controbilanciare la comprensibile eccitazione ed il senso di onnipotenza che i recenti risultati di *Machine learning* hanno diffuso è costituito dal *Paradosso di Simpson*, probabilmente il paradosso più noto e che illustreremo utilizzando il seguente esempio. Consideriamo uno studio clinico che per un gruppo di individui misuri le ore di esercizio fisico settimanale (Exercise) ed il livello di colesterolo (Cholesterol), rappresentando le due quantità tramite il grafico in Figura 9.

Nella figura, per ogni gruppo ad età fissata (Age; che varia da 10 a 50), si osserva una linea che individua una tendenza discendente o correlazione negativa, quasi a suggerire che l'incremento del livello di esercizio (Exercise) porta ad una riduzione del livello di colesterolo (Cholesterol). D'altra parte, se non segreghiamo gli individui in base alla loro età (Age), rendendoli indistinti come accade nella figura in basso, notiamo una linea che individua una tendenza ascendente, o correlazione positiva, quasi a suggerire che più un individuo svolge esercizio fisico (Exercise) più aumenta il suo livello di colesterolo (Cholesterol).

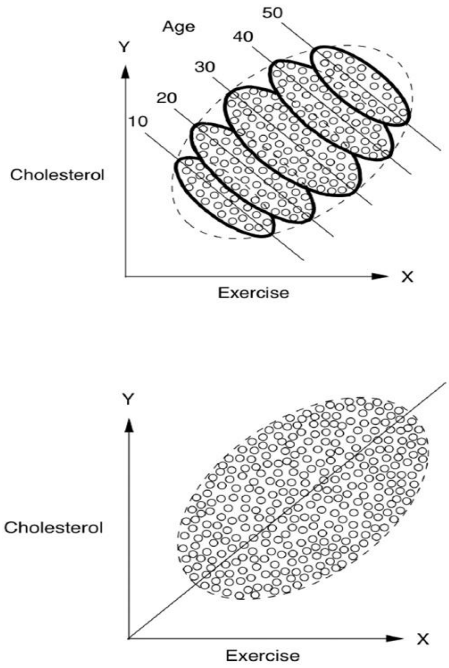


Figura 9 – Il paradosso di Simpson (tratta da [Pearl 2016])

Alla luce di quanto esposto saremmo portati a concludere che se si ignora l'età (Age) dell'individuo, allora l'esercizio fisico è dannoso, infatti aumentando le ore di esercizio settimanale pare aumenti il livello di colesterolo, mentre se si conosce l'età dell'individuo, indipendentemente da quale essa sia (10, 20, 30, 40 o 50), fare esercizio fisico sembra ridurre il livello di colesterolo. In altri termini, fare esercizio fisico sembra essere salutare per ogni gruppo di età (Age) ma al tempo stesso pare essere dannoso per gli individui se visti indistintamente.

Questo è chiaramente un paradosso, appunto un'istanza del *paradosso di Simpson*, che ad oggi non può essere risolto anche nel caso in cui si disponga di infiniti dati. In questo caso, per decidere se l'esercizio fisico sia salutare o dannoso, abbiamo bisogno di interrogare il processo che ha generato i dati e di far riferimento al concetto di *causalità*. I dati mostrano che gli individui più vecchi della nostra popolazione svolgono maggiore esercizio fisico. Ora, apparendo più verosimile che sia l'età (Age) a causare il livello di esercizio dell'individuo e non il vice versa, possiamo concludere che l'età (Age) *può avere un effetto causale* sul livello di colesterolo, e pertanto l'età può essere un *fattore confondente* per il livello di esercizio ed il livello di colesterolo.

Questo significa che per scoprire quale è l'effetto dell'esercizio fisico sul livello di colesterolo dovremo far riferimento ai dati segregati per età, concludendo in via definitiva che l'esercizio fisico è salutare, indipendentemente dall'età (Age) dell'individuo.

In base alla *scala della causalità* (Figura 10), proposta da *Judea Pearl*, la *gerarchia causale* è costituita da tre strati, ognuno dei quali è associato ad un piolo della scala medesima. In questo caso si parla di gerarchia in quanto ogni domanda associata ad un determinato livello può ottenere risposta solo se si dispone dell'informazione del livello medesimo o dei livelli gerarchicamente superiori, i pioli che stiano più in alto nella *scala della causalità*.

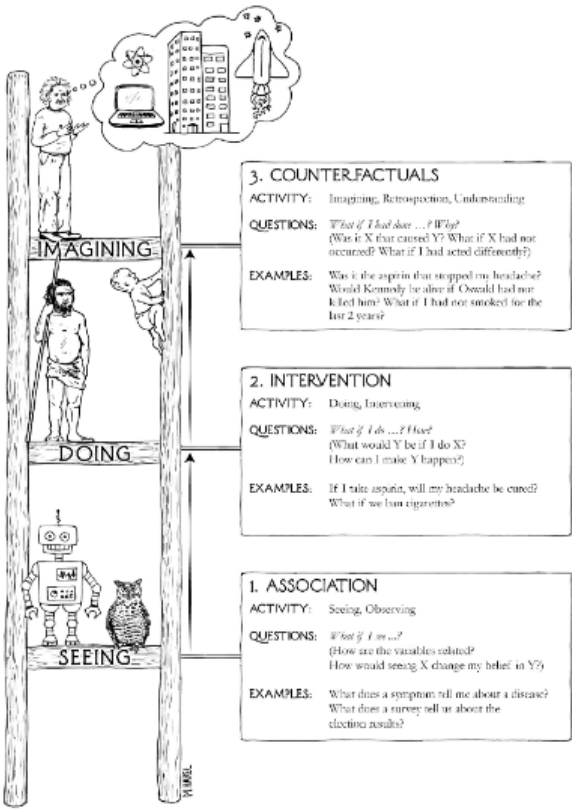


Figura 10 – La scala della causalità di Pearl (tratta da [Pearl 2018])

I tre livelli della gerarchia causale sono denominati:

1. associazione,
2. manipolazione,
3. controfattuale,

con l'obiettivo di enfatizzarne l'impiego.

Il primo livello è denominato *associazione* in quanto è puramente relativo a relazioni statistiche, definite solo a partire dai dati disponibili. Ad esempio, sapere che un cliente di un supermercato acquista un sacchetto di patatine rende più probabile che acquisti anche una bibita; questo tipo di relazione può essere inferita direttamente data la sola disponibilità dei dati. Tutte le domande di questo strato, in quanto non richiedono alcuna informazione di natura causale, stanno nel livello più basso della

gerarchia, associate al primo gradino della *scala della causalità*. Questo è il livello della quasi totalità degli attuali algoritmi di Machine learning e Deep learning.

Il secondo livello, denominato *manipolazione*, richiede non solo di osservare il valore di determinate quantità, ma la possibilità di manipolare (modificare) il valore di una o più delle quantità che possiamo osservare; una domanda tipica di questo livello gerarchico è "*Cosa accadrebbe se raddoppiassimo il prezzo?*". Non è possibile rispondere a domande di questo tipo disponendo solo dei dati di vendita. Infatti, questo tipo di domanda fa riferimento ad un'azione da noi compiuta che determinerebbe un cambiamento del comportamento del cliente all'aumento del prezzo. In questa condizione, a meno di non replicare esattamente le stesse condizioni di mercato che in passato avevano portato ad un incremento dei prezzi, il cliente si comporterà verosimilmente in modo diverso da quanto fatto in passato.

Il livello più alto della gerarchia causale è rappresentato del controfattuale. Una domanda che appartiene certamente a questo livello è la seguente "Cosa sarebbe accaduto se avessi agito o deciso in modo differente?". Una domanda di questo tipo richiede di ragionare retrospettivamente. Il controfattuale è l'ultimo piolo della scala della causalità in quanto include associazione e manipolazione. Se disponiamo di un modello in grado di rispondere a domande controfattuali, allora esso risponderà a domande di manipolazione e di associazione.

9. Conclusioni

Quando una nuova tecnologia si affaccia con piglio deciso e si afferma in modo così pervasivo come sta facendo il *Deep Learning,* che indubitabilmente sta cambiando le regole del gioco in modo drastico e repentino, non è per nulla saggio lasciare che questa tecnologia rimanga una scatola nera, vale a dire che resti un elemento imperscrutabile da accogliere ed accettare come un oracolo. Infatti, l'opacità di un algoritmo, vale a dire l'ignoranza delle ipotesi sulle quali si basa e il non considerarne adeguatamente i limiti di applicabilità, portano ad un suo impiego errato o ad interpretare le previsioni fornite dall'algoritmo in termini non sempre scientificamente fondati, e porta pertanto a prendere decisioni prive delle necessarie motivazioni e spiegazioni.

Noi tutti dobbiamo avere contezza di cosa sia la rivoluzione che va sotto il nome di *Machine learning*, dei suoi vantaggi, delle enormi opportunità ma anche dei suoi limiti e dei conseguenti rischi. Questa necessità non vale solo per chi come uno degli scriventi (Fabio Stella) abbia dedicato la propria attività di ricerca a studiare, analizzare, sviluppare ed applicare algoritmi di Machine learning in domini applicativi quali la finanza, la produzione industriale, il commercio elettronico, la biologia e la medicina, ma anche a tutti noi intesi come esseri umani e cittadini che stiamo assistendo ad una rivoluzione della quale dobbiamo essere consapevoli ed alla quale dobbiamo prendere parte attivamente per garantire che l'essere umano sia sempre, comunque e saldamente al primo posto, prima di ogni Intelligenza artificiale.

Da questo punto di vista, riconoscendo che questa rivoluzione offre opportunità ma contiene anche tanti rischi, pensiamo comunque che ci si debba guardare molto più da noi stessi che non dal rischio che

le macchine ci controllino, se questo accadrà sarà solo colpa e responsabilità nostra, per non avere opportunamente agito ed interagito, analizzato e criticato quanto sta accadendo e quanto accadrà prossimamente.

Riferimenti

C. Anderson - The end of theory: The data deluge makes the scientific method obsolete." Wired magazine 16.7, 2008.

N. Barrowman - Correlation, causation, and confusion, The New Atlantis, 2014.

L. Breiman - Statistical modeling: The two cultures (with comments and a rejoinder by the author) Statistical science 16.3, pp. 199-231, 2001.

D. Butler - When Google got flu wrong - Nature News 494.7436, 2013.

K. Byeong Soo, et al. "Data modeling versus simulation modeling in the big data era: Case study of a greenhouse control system." Simulation 93.7, 579-594, 2017.

P. Coveney, E. R. Dougherty, e R. R. Highfield. "Big data need big theory too." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016.

A.Darwiche, Adnan. "Human-level intelligence or animal-like abilities?." arXiv preprint arXiv: 1707. 04327, 2017.

C. Calude e G. Longo. "The deluge of spurious correlations in big data." Foundations of science 22.3 (2017): 595-612.

D. Lazer - The parable of Google Flu: traps in big data analysis - Science 343.6176, pp. 1203-1205, 2014.

J. Lehikoinen e Ville Koistinen - In big data we trust? - Interactions 21.5, 2014.

G. Markus e E. Davis – Eight (no, nine!) problems with big data, New York Times, 2014.

X. Meng - A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it) 2014

J. Pearl, M. Glymour, N.P. Jewell, Causal inference in Statistics: A Primer, Wiley, 2016.

J. Pearl e D. Mackenzie – The Book of Why - Penguin; 2 edizione, 2019.

S. Succi e P. V. Coveney - Big data: the end of the scientific method? - Philosophical Transactions of the Royal Society A 377.2142, 2019.

Capitolo 17 - Big data e psicologia: luci e ombre

Paolo Cherubini

L'avvento di internet, e soprattutto dei social media, ha aperto prospettive di ricerca entusiasmanti e senza precedenti per la scienza del comportamento [Adjerid & Kelley, 2018[[Jones 2016]. Parallelamente, al di fuori del mondo della ricerca, la capacità di tutti gli utenti della rete di comunicare liberamente con tutti gli altri sta mostrando aspetti inattesi e interessanti, che però sembrano – quasi paradossalmente – mettere a repentaglio alcuni tradizionali assunti alla base della democrazia [Madsen, Bailey, & Pilditch, 2018]. Questi aspetti meritano di essere posti al centro dell'attenzione nella ricerca psicologica.

1. Big data e ricerca in psicologia sperimentale

In 150 anni di esistenza il principale ostacolo alla rapida crescita della psicologia sperimentale è stato la scarsa disponibilità di dati. Negli esperimenti la platea di soggetti sperimentali si compone di esseri umani (o altri animali, ma solo i livelli più fondamentali dei processi psicologici possono essere studiati su animali non umani con qualche pretesa di generalizzabilità agli esseri umani). Il ricorso a soggetti sperimentali umani è intriso di difficoltà. Alcune sono di livello etico, come la necessità di informare i soggetti sui dettagli della ricerca e i suoi scopi, di ottenere il loro consenso informato, di non danneggiarli in alcun modo, di non pagarli per partecipare alla ricerca, ecc. I requisiti etici si sono fatti lodevolmente sempre più stringenti via via che passavano i decenni, rendendo sempre più lunghi e onerosi gli aspetti pratici da affrontare prima della raccolta di dati su un qualsivoglia campione umano. Altre difficoltà sono pratiche: la gente ha molte altre cose da fare, oltre a prestarsi per partecipare a esperimenti di psicologia, e trovare "soggetti" volontari è un'impresa sempre più ardua; la disponibilità dei soggetti deve essere compatibile con la disponibilità dei laboratori; e se gli studi richiedono più sedute in giorni, mesi, o settimane diversi, queste difficoltà diventano particolarmente gravi. Per gli studi longitudinali su più anni, le difficoltà diventano quasi insormontabili, tanto da renderli molto rari nella storia della psicologia.

La scarsa reperibilità di soggetti, insieme alla complessità di alcuni dei fenomeni studiati, ha contribuito a dare forma alla ricerca psicologica tradizionale: è un tipo di ricerca empirica che si focalizza sugli effetti di poche variabili esplicative su una singola variabile dipendente. Quando le variabili sono misurate o controllate più volte nel tempo, questo avviene in situazioni altamente strutturate, specificate a priori: per esempio, con incontri in laboratorio a cadenze prefissate. Quasi tutte le ricerche nella letteratura psicologica si basano su poche variabili "fotografate" trasversalmente, o – al più – in poche occasioni altamente strutturate disseminate in un lasso di tempo relativamente breve. Oltre a poche variabili e pochi punti di registrazione nel tempo, le ricerche tradizionali usano campioni relativamente piccoli. I ricercatori calcolano la dimensione minima del campione necessaria per avere almeno, per esempio, l'80% di possibilità di rilevare un effetto di dimensioni medie o grandi: si attengono poi strettamente a

quella numerosità nella raccolta dati, a causa dei già menzionati problemi nel reclutare campioni di ampie dimensioni.

La combinazione tra l'uso di un numero limitato di variabili e campioni di piccole dimensioni è stato il "marchio di fabbrica" della ricerca empirica in psicologia, nell'arco dell'intera esistenza di questa scienza. Le domande che si possono affrontare con questo approccio sono necessariamente di portata limitata. Metodologicamente, ci si concentra sul partizionamento della varianza relativa alla variabile dipendente, e sulla stima degli effetti di e tra le variabili: questo ha consentito un grande sviluppo di alcune tecniche statistiche ben note, come i test *t*, l'analisi della varianza, la regressione multipla, e le misure del goodness of fit basate sul chi-quadrato. D'altra parte, a poco a poco, sono emersi alcuni importanti limiti di questo approccio alla ricerca. Sono passati più di 50 anni da quando Cohen [Cohen 1962] indicò che i ricercatori usavano troppo spesso campioni troppo piccoli. Successivamente è stato ben illustrato il problema dei "campioni di convenienza", cioè selezionati perché disponibili e non perché rappresentativi della popolazione a cui i ricercatori speravano di generalizzare le scoperte [Henrich, Heine, & Norenzayan, 2010].

I campioni più comuni in gran parte delle ricerche sono tratti dagli studenti dei corsi di psicologia, o comunque dagli studenti universitari: il materiale umano più disponibile, per la loro vicinanza fisica ai laboratori, per la maggiore disponibilità di tempo rispetto ad altre categorie, e per la maggiore motivazione a contribuire alle ricerche; ma proprio per queste stesse ragioni, non necessariamente sono campioni rappresentativi di popolazioni più variegate.

È poi stato sottolineato che la misurazione di alcuni costrutti per loro natura altamente dinamici e soggetti a rapide fluttuazioni nel tempo, come l'umore o le emozioni, è spesso stata troppo grossolana, con una o poche misurazioni "statiche" [Csikszentmihalyi & Larson, 2014]. Infine, sono state individuate alcune importanti frodi sui dati [Simonsohn 2013]; le vere e proprie frodi sono auspicabilmente poche, ma molto più frequenti sono diversi tipi di *malpractices* metodologiche [John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011], che indeboliscono notevolmente la robustezza dei risultati di diverse ricerche. Questi problemi hanno contribuito in modo significativo alla cosiddetta "crisi della replicabilità dei risultati" in psicologia [Maxwell, Lau, & Howard, 2015].

L' era dei Big data sta cambiando la natura dei dati disponibili per almeno tre aspetti. Già Cattell [Cattel 1946], [Cattel 1966] classificava i metodi di ricerca basandosi sulla struttura dei dati disponibili, e organizzandoli secondo tre dimensioni: persone disponibili per lo studio (ampiezza del campione, *n*), varietà delle variabili considerate (*v*), e capacità di osservare cambiamenti in queste variabili in diverse occasioni nel tempo (*t*). È un approccio classico, che tuttavia trova una chiara eco in descrizioni recenti dei big data, come quella di Borgman (2015), che, come già detto nel Capitolo 1, li scompone nelle "3 V" di *volume*, *varietà*, e *velocità*. I big data generati da grandi piattaforme digitali non sostituiranno le ricerche di laboratorio tradizionali, ma affiancandosi ad esse in modo complementare stanno consentendo, e consentiranno sempre di più nel prossimo decennio, di superare molti dei problemi prima descritti.

Già si intravedono cambiamenti metodologici migliorativi su tutte le tre dimensioni *n*, *v* e *t*. Per quanto riguarda *n*, cioè l'ampiezza dei campioni, da poche decine o centinaia di soggetti reclutati per una

ricerca, può diventare prassi comune osservare campioni di decine o centinaia di migliaia, o addirittura di milioni di persone. Questo risolve due problemi metodologici: da una parte, i grandi campioni consentono di individuare con grande precisione effetti specifici anche di modesta entità; d'altra parte, la platea di persone che interagiscono con le piattaforme digitali è molto vasta, e consente di superare il problema dei "campioni di convenienza", migliorando la rappresentatività dei campioni verso la popolazione generale (si veda ad es. [Shannon, Andrew, & Duggan, 2016]). Certamente non tutte le categorie di persone hanno eguale facilità, possibilità, o volontà di accesso a diverse piattaforme digitali; quindi i ricercatori dovranno prestare molta attenzione ai gruppi non sufficientemente rappresentati nelle utenze di quelle piattaforme.

Ma è fuori di ogni dubbio che vastissimi campioni tratti da popolazioni di utenti del Web permetteranno di rappresentare la popolazione generale più adeguatamente di quanto lo facciano piccoli campioni tratti da popolazioni di convenienza, come gli "studenti unversitari". Per quanto riguarda *v*, cioè la varietà delle variabili considerate nello studio, osserviamo che le piattaforme digitali stanno diventando sempre più interattive. Gli utenti possono esibire in esse diversi comportamenti: acquisti, vendite, accessi a prodotti o informazioni, tempi di esplorazione degli stessi, interazioni comunicative verbali (commenti e quesiti), interazioni comunicative non verbali (come i "like"), caricamenti di fotografie, persino "comportamenti negativi" come "non cliccare" su un qualche link; e molti altri. Questa varietà di interazioni possibili consente di operazionalizzare diverse variabili, da esplorare in connessione a diversi tipi di comportamento. Qui è necessario un *caveat*, per quanto ovvio: i ricercatori dovranno prestare molta attenzione al fatto che stanno misurando comportamenti certamente reali e concreti, ma "nuovi", perché esibiti in un ambiente "virtuale" che solo vent'anni fa nemmeno esisteva.

Non è garantito che quanto riscontrato valido, per esempio, per un comportamento d'acquisto su un sito di vendite sia altrettanto valido per comportamenti d'acquisto quando si interagisce direttamente con un commesso in un negozio. Ma, al netto di queste naturali precauzioni, il ventaglio di variabili comportamentali operazionalizzabili su dati provenienti da piattaforme digitali è sempre più vasto, consentendo studi ad altissimo livello di granularità. Infine, per loro stessa natura le grandi piattaforme digitali sono "sempre in ascolto": quindi la capacità di misurare e controllare le variabili più volte nel tempo aumenta in maniera pressoché illimitata, rispetto a quanto possibile in un laboratorio tradizionale, e ad un costo infinitamente più basso.

L'osservazione dei comportamenti sulle piattaforme può diventare, se non continua, quasi continua per prolungati periodi di tempo. Questo permette ai ricercatori di individuare con precisione quando avviene un evento o un comportamento di interesse (nelle ricerche di laboratorio, se il comportamento non avviene nei brevi periodi di permanenza presso il laboratorio, non viene individuato), e di tenerne traccia a basso costo per un lungo periodo di tempo, e registrarne le variazioni fini. La capacità di catturare con facilità flussi comportamentali quasi continui consentirà sia un miglior focus sugli aspetti "di processo" dei comportamenti studiati, sia una maggiore attenzione al comportamento di singoli individui, in un certo senso ricucendo un tradizionale "iato" tra la ricerca quantitativa e quella qualitativa in psicologia.

2. Big data e progresso sociale

Nel precedente paragrafo abbiamo indicato sinteticamente le entusiasmanti prospettive di ricerca aperte alla psicologia sperimentale dall'avvento delle piattaforme digitali, che provocano, ospitano e registrano una varietà di comportamenti umani, rendendo disponibili e abbastanza facilmente accessibili data set di big data su quei comportamenti.

Abbiamo anche brevemente ricordato come quelle piattaforme, in un certo senso, "modifichino" quelli che erano "analoghi" comportamenti umani pre-Web: acquistare un prodotto sul Web non è la stessa cosa che acquistarlo in un negozio, fare una comunicazione pubblica "veemente" sul Web non è la stessa cosa che proclamarla a voce stentorea sulla pubblica piazza, e così via. L'emersione di nuovi comportamenti, fotografabili e studiabili tramite i big data, sta provocando effetti sorprendenti nella società, imprevedibili a priori, e che stanno diventando – e auspichiamo che diventeranno sempre di più – oggetti di studio di fondamentale importanza.

Il più famoso tra tutti questi effetti sono le "fake news". La psicologia ha studiato per decenni – tanto al livello quantitativo quanto a quello qualitativo – i meccanismi di pensiero e di comunicazione alla base sia del "credere vere" alcune false notizie, teorie, e storie (per esempio, le diverse forme di confirmation bias, si veda [Nickerson, 1998]), sia al loro diffondersi nella società [Heath, Bell, & Sternberg, 2001]. Non è certo una novità che le persone tendano a credere in massa alle fandonie più implausibili, totalmente prive di fondamenti empirici: basti pensare al diffondersi millenario delle religioni, delle credenze esoteriche ed occultiste, delle pseudoscienze e parascienze – con relative ideologie politiche al seguito.

Queste false credenze hanno sempre avuto effetti molto concreti, sia sulla società sia sugli individui. Ma l'impressione – solo di impressione si può parlare perché le ricerche in merito, pur numerose negli ultimi sei-sette anni, ad es. [Cook & Lewandowsky, 2016; Ngampruetikorn & Stephens, 2015; Olsson, 2011], non sono ancora tali da darci alcuna sicurezza – è che nelle "fake news ai tempi dei social media" ci sia qualcosa di diverso rispetto alle fandonie di una volta: la diversità potrebbe essere quantitativa (per esempio, maggiore rapidità di generazione, diffusione, e resistenza all'estinzione), qualitativa (per esempio, diversa visibilità, o diversità nei meccanismi motivazionali o cognitivi sottostanti), o entrambe le cose.

È forte l'impressione che, nella loro interazione con alcuni meccanismi decisionali sociali, quali le elezioni, le attuali fake news abbiano effetti molto rapidi, e potenzialmente devastanti [Flaxman, Goel, & Rao, 2016]: alcune *fake news* possono aver influenzato in maniera significativa l'elettorato in situazioni critiche, come nel voto sulla Brexit, o il voto nelle ultime elezioni presidenziali americane, al punto da mettere a repentaglio alcune scelte politiche forse determinanti per il futuro dell'umanità [Lewandowsky et al. 2019].

Come si diceva all'inizio di questo capitolo, assistiamo a un effetto che è quasi un paradosso storico, di fronte al quale l'unico approccio sensato sembra essere studiare e ricercare, finché non lo si sarà capito meglio. Il paradosso è questo: siamo sempre stati abituati a pensare che ogni blocco alla libera

circolazione del pensiero e delle idee fosse un blocco al progresso civile e sociale. L'umanità è molto progredita nei secoli, sia materialmente sia a livello conoscitivo, eliminando – via via – diversi blocchi alla circolazione della conoscenza. Si pensi agli effetti dell'invenzione della stampa, della traduzione e circolazione delle "sacre scritture" in lingue correnti, della diffusione di idee e teorie scientifiche "vietate" da alcuni sistemi ideologici (come il sistema Copernicano, o l'idea di evoluzione naturale della specie umana); o agli effetti dell'invenzione e diffusione dei quotidiani, che hanno portato il dibattito politico dalle regge alle piazze e ai caffè; all'alfabetizzazione, che ha consentito a sempre più persone di leggere quei quotidiani, attraverso l'introduzione prima e l'estensione poi di obblighi scolastici sovvenzionati dagli stati; agli effetti della diffusione di massa della radio e della televisione sulle capacità linguistiche e sulla maturità civile della popolazione; agli effetti dell'apertura dell'istruzione universitaria alle masse.

Ogni volta che, nella storia, abbiamo visto cadere una censura alla proliferazione delle idee, e abbiamo visto allargarsi la comunicazione, abbiamo anche assistito ad un progresso; simme-tricamente, ogni volta che il dibattito politico è diventato più inclusivo, abbiamo assistito ad un progresso (si pensi alla progressiva estensione del voto nelle democrazie – nell'ordine – agli uomini di classi sociali medie e basse, alle donne, e alle minoranze etniche, per arrivare ai moderni sistemi a suffragio universale). Quindi, tutto faceva pensare che attraverso un "Web 2.0" che consentisse la comunicazione globale di tutti gli utenti con tutti gli utenti e la libera circolazione di ogni opinione senza il minimo filtro censorio, si sarebbero dovuti sviluppare meccanismi compensatori tali da contrastare e far cadere in minoranza ogni teoria falsa e non suffragata da fatti, facendo emergere un "consenso maggioritario e democratico" sempre più orientato verso opinioni sostenute fattualmente, rendendo sempre più realistica l'idea di implementare una "democrazia diretta" su vaste masse di popolazione (Grimes, 2016). Invece stiamo assistendo alla proliferazione di gruppi e sottogruppi, talvolta così vasti da influenzare gli esiti di elezioni nazionali o continentali, disposti ad accettare, diffondere ed autoalimentare credenze obiettivamente smentite da fatti a disposizione di tutti; i "consensi condivisi" che vengono raggiunti, possono in quei casi rivelarsi assai deteriori per il progresso dell'umanità.

Questa sarebbe la prima volta nella storia in cui un ampliamento della libertà di idee, di parola, e di comunicazione porta a un arretramento, invece che ad un avanzamento, della società, e del benessere collettivo medio (materiale, ma non solo). Esistono preoccupanti analisi [Madsen, Bailey, & Pilditch, 2018] che mostrano come il formarsi e il mantenersi di "echo chambers" (o camere dell'eco, si veda il Capitolo 5) in grado di riverberare e autoalimentare false credenze sia una caratteristica emergente della semplice *dimensione* della rete, e del suo alto grado di *interconnessione* interna: anche agenti perfettamente bayesiani, se inseriti in una rete sufficientemente vasta, tutti con eguale accesso alle stesse informazioni di partenza, e senza alcun bias cognitivo *tranne* l'aspetto più generale e meno evitabile tra quelli che compongono il *confirmation bias* – e cioè la tendenza spontanea a preferire di comunicare con persone che hanno idee più vicine alle nostre, piuttosto che con quelle che hanno idee più lontane dalle nostre – generano echo chambers che li portano a diventare progressivamente sempre più "certi" di nozioni false.

La ricerca su questi aspetti è appena agli inizi: ma è un importante obiettivo approfondirla e giungere a qualche nozione scientificamente fondata sul come limitare o evitare questi pericoli, senza limitare la libertà di comunicazione sulla rete; se la scienza non saprà dare una risposta a queste preoccupazioni,

rischiamo di vedere sempre più persone auspicare o predicare il reinserimento di “censure”, limitazioni all’accesso, o altri vincoli allo scambio delle informazioni in rete.

Riferimenti

I. Adjerid, & Kelley, K. - Big data in psychology: A framework for research advancement. American Psychologist, 73(7), 899-917, 2018.

C. L. Borgman - Big data, little data, no data: Scholarship in the networked world. Cambridge, MA: MIT Press, 2015.

R. Cattell - Personality structure and measurement; the operational determination of trait unities. British Journal of Psychology, 36, 88–103, 1946.

R. Cattell - The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology (pp. 67–128). Chicago, IL: Rand-McNally, 1966.

J. Cohen - The statistical power of abnormal-social psychological research: A review. The Journal of Abnormal and Social Psychology, 65, 145–153, 1962.

J. Cook & Lewandowsky, S. - Rational irrationality: Modeling climate change belief polarization using Bayesian networks. Topics in Cognitive Sciences 8, 160–179, 2016.

M. Csikszentmihalyi & Larson, R. - Validity and reliability of the experience-sampling method. In Flow and the foundations of positive psychology (pp. 35–54). Rotterdam, the Netherlands: Springer, 2014.

S. Flaxman, S., Goel, S. & Rao, J. M. - Filter bubbles, echo chambers, and online news consumption, Public Opinion Quarterly 80 (special issue), 298–320, 2016.

D.R. Grimes - On the Viability of Conspiratorial Beliefs. PLoS One 11(1), e0147905, 2016.

C. Heath Bell e Sternberg, E. - Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology, 81*(6), 1028-1041, 2001.

J. Henrich Heine, S. J., e Norenzayan, A. - The weirdest people in the world? Behavioral and Brain Sciences, 33, 61–83.

L. John, Loewenstein, G., e Prelec, D. - Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23, 524–532, 2012.

M.N. Jones (Ed.) - Big data in cognitive science. Psychology Press, 2016.

S. Lewandowsky, Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. - Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. Cognition, 188, 124-139, 2019.

J.K. Madsen, Bailey, R. M., e Pilditch, T. D. - Large networks of rational agents form persistent echo chambers. Scientific Reports, 8, 12391, 2018.

V. Ngampruetikorn & Stephens, G. J. - Bias, Belief, and Consensus: Collective opinion formation on fluctuating networks, arXiv1512.09074v1, 2015.

R.S. Nickerson - Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220, 1998.

E.J. Olsson - A simulation approach to veritistic social epistemology. Episteme 8(2), 127–143, 2011.

G. Shannon, Andrew, P., & Duggan, M. (2016). Social media update 2016. Washington, DC: Pew Research Center, 2016.

J.P. Simmons, Nelson, L. D., & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366, 2011.

U. Simonsohn - Just post it: The lesson from two cases of fabricated data detected by statistics alone. Psychological Science, 24, 1875–1888, 2013.

Capitolo 18 – La Datacy

Carlo Batini

1. Introduzione

La cultura di una comunità è tradizionalmente misurata da due indicatori, la literacy e la numeracy. Per *literay* si intende la capacità di usare il linguaggio scritto sia passivamente che attivamente, attraverso le attività di lettura o scrittura di un testo, e ascolto o espressione verbale di un discorso. Uno standard adottato spesso per valutare la literacy è la lettura di un giornale. La *numeracy* complementa la literacy, è definita come la capacità di fare calcoli con i numeri e i simbolismi matematici, applicandoli in un insieme di contesti per risolvere problemi.

Tutto quanto abbiamo discusso in questo libro ci fa arrivare alla conclusione che una nuova scienza, la Scienza dei dati, si sta formando, basata su altre scienze quali la Informatica, la Matematica, la Statistica. Poiché i dati digitali stanno pervadendo tutti gli aspetti della nostra vita, appare importante introdurre accanto alla literacy e alla numeracy una nuova categoria riferita ai dati digitali, la datacy. Possiamo definire la *datacy* come l'insieme dei modelli, metodologie, linguaggi, tecniche, e delle loro applicazioni, che permettono di elaborare, analizzare, ragionare su un vasto insieme di tipologie di dati digitali, come i dati tabellari, i testi, le immagini, essendo in grado di:

- ricostruirne e modellarne il significato attraverso linguaggi concettuali,
- applicare tecniche per la valutazione del livello di veridicità e qualità,
- integrare dati eterogenei riconciliandone le differenze,
- interrogare e analizzare i dati mediante linguaggi e ambienti per estrarne conoscenza
- applicare tecniche e modelli statistici e basati su apprendimento per costruire modelli decisionali, interpretativi, predittivi e prescrittivi
- visualizzare i dati (digitali) per comprendere meglio la natura dei dati e i risultati delle analisi
- risolvere problemi con il supporto dei dati e prendere decisioni complesse.
- comprendere l'impatto sulla economia e sulla società del fenomeno dei dati
- analizzare i corpi giuridici sviluppati dalle istituzioni pubbliche in tema di dati
- comprendere i principi delle scienze cognitive che presiedono all'uso consapevole dei dati nella nostra vita.
- affrontare i temi etici che nascono dall'uso dei dati.

E' naturale che il concetto così vasto di datacy abbia portato le diverse istituzioni e Università che si occupano di formazione a vedere la Scienza dei dati, alla base della datacy, secondo varie angolazioni culturali, questo anche in funzione delle competenze esistenti presso le Università. Esaminando varie

esperienze esistenti, di cui tra poco discuteremo, si giunge alla conclusione che le principali aree culturali che forniscono un contributo e allo stesso tempo sono talvolta profondamente influenzate dalla Scienza dei dati, siano quelle mostrate in Figura 1; accanto alla Informatica e alla Statistica, sono l'Economia, le Scienze filosofiche ed etiche, le Scienze sociali, le Scienze cognitive.

Nelle sezioni che seguono esploreremo i contenuti che emergono nei corsi di Data Science nel mondo riferiti alle precedenti aree culturali. Inizieremo dalla Università degli Studi di Milano Bicocca, dove a partire dall'anno accademico 2017-18 è stato istituito un corso di laurea magistrale in Data Science.

Parleremo anche del materiale didattico complementare che è stato prodotto e che viene erogato su tematiche aggiuntive a quelle insegnate nel corso di laurea, con lo scopo di costruire nel tempo un corpus di conoscenze che ricadono in quella che mi piace definire la Cultura del dato digitale.

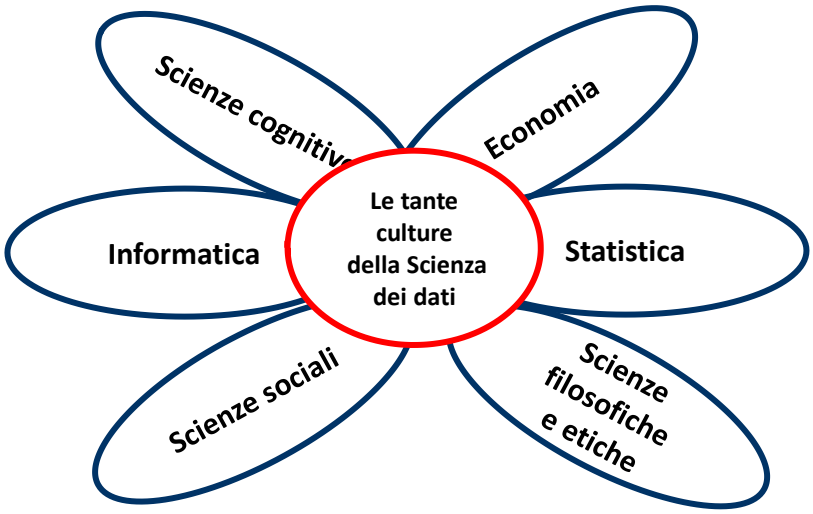


Figura 1 – Le diverse discipline che contribuiscono alla scienza dei dati

Successivamente, esploreremo le diverse discipline di Figura 1, mostrando essenzialmente esempi selezionati di corsi che approfondiscono il legame tra Scienza dei dati e tali discipline. Per ogni corso, sono riportati il titolo, l'Università che lo eroga, il sito dove si possono trovare approfondimenti. In molti casi è stato conservato il linguaggio inglese; n alcuni casi, per i quali non è stato possibile individuare un intero corso di studi, sono riportati contenuti di seminari.

2. La Scienza dei dati nel corso di Laurea Magistrale della Università di Milano-Bicocca

In Figura 2 mostriamo l'organizzazione dei corsi erogati nel primo e secondo anno del corso di laurea in Data Science.

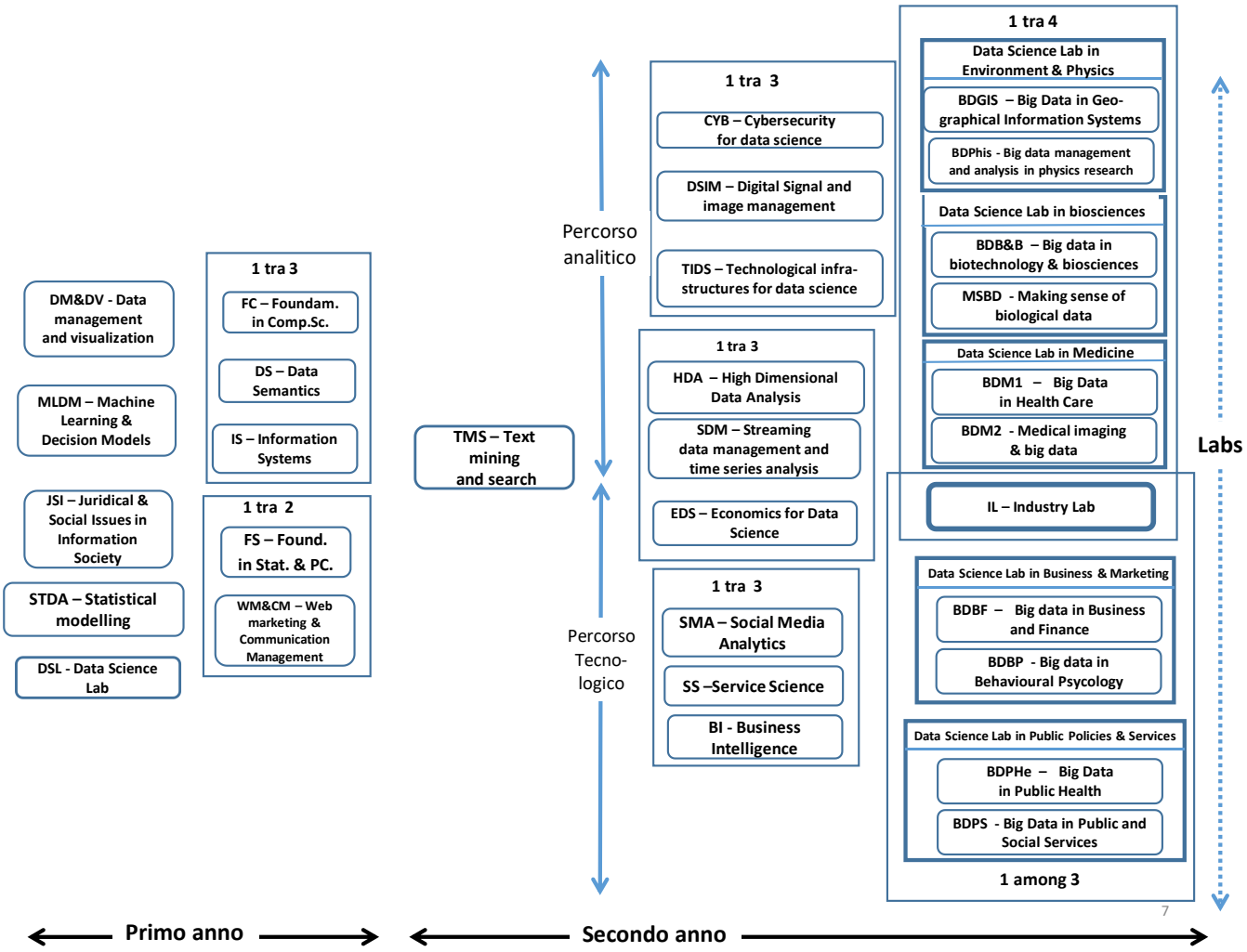


Figura 2 – Il Curriculum del corso di laurea magistrale in Scienza dei Dati della Università di Milano Bicocca (Anno Accademico 2019-20)

Nel seguito forniamo per tematiche i titoli dei singoli insegnamenti e i contenuti erogati

<u>Computer science – Basi, tecnologie, modelli, metodi</u>

Fondamenti di Computer Science – Come organizzare I dati in una base di dati, il modello relazionale, il linguaggio di interrogazione SQL, Introduzione alla programmazione in Python: array, liste, cicli, funzioni e programmi. Programmazione esplorativa, introduzione all'ambiente Pandas. Il test e la correzione dei programmi.

Data Management – Il ciclo di vita dei dati, dalla acquisizione alla integrazione. Modelli per Big data, modelli NoSQL, Modelli basati su documenti, architetture dati distribuite, la qualita dei dati, l'ambiente MapReduce.

Data Visualization – Introduzione alla visualizzazione, la percezione grafica, le codifiche di simboli visuali, l'uso dei colori, la trasformazione dei dati in rappresentazioni visuali di conoscenza, come produrre una narrativa per il data journalism.

Data Semantics – Modelli per rappresentare conoscenza: knowledge Graphs, RDF. Il linguaggio SPARQL. Vocabolari e ontologie, RDFS, OWL e regole. La data integration e il matching semantico. Estraizone della informazione e entity linking, l'esplorazione della conoscenza.

Sistemi informativi – Cosa è un sistema informativo, il suo ruolo nelle organizzazioni. Come si progettano I sistemi informativi. Modelli per rappresentare e collegare organizzazione, processi e dati. Il modello Entità Relazione per I dati e il linguaggio BPMN per i processi. Efficienza ed efficacia dei sistemi informativi e dee processi. L'evoluzione dei sistemi informativi verso I sistemi networked.

Cybersecurity for Data Science – Attori coinvolti nella cybersecurity. I fini della cybersecurity: confidenzialità, integrità, disponibilità. Incidenti, vulnerabilità, attacchi: errori nel software, sniffing, spoofing, social engineering. Denial of service. Difese, crittografia, security nei big data. Strumenti.

Digital Signal and Image Management – Conversione analogico–digitale, classificazione e riconoscimento di segnale, classificazione delle immagini e video, metodi per indexing e retrieval.

Streaming data management e time series analysis – Natura delle serie temporali, serie storiche e dati streaming, la gestione delle serie temporali, le funzionalità per serie temporali, similarità, clustering, regression, forecasting. La predizione statistica. Processi stocastici a tempo discreto, funzione di autocovarianza, stazionarietà, funzione di cross-varianza, filtro di Kalman, Stime di massima similarità, Approcci non parametrici, Reti neurali artificiali.

Text Mining and Search – L'information Retrieval, la text summarization, la classificazione di testi, l'estrazione di informazione strutturata da testi. La rappresentazione dei testi, le search engines. Il Web Crowling. Gli strumenti.

Technological infrastructures for Data science – L'Internet delle cose, tipi di sensori, qualità dei dati da sensori, reti di sensori, piattaforme per gestione di sensori. Le architetture per la gestione dei dati, architetture parallele, modelli di comunicazione, piattaforme distribuite, Data Center, Cloud Computing.

<u>Statistica</u>

Fondamenti di probabilità e statistica – Introduzione alla gestione dei dati con SAS e R, la statistica descrittiva, il concetto di inferenza statistica, la regressione logistica.

Modelli statistici per la analisi dei dati - Modelli lineari avanzati confrontati con I modelli lineari classici. Modelli lineari generalizzati, modelli non lineari, trattamento degli outlier. Modelli multivariate e multilivello

High Dimensional data analysis – Gli spazi dimensionali di grandi dimensioni e la "maledizione" della dimensionalità. Il testing multiplo, l'inferenza in spazi di grandi dimensioni, modelli grafici.

Scienze giuridiche

Juridical Issues in Information society – Introduzione alla costituzione; la protezione dei diritti, il diritto di informare e di criticare, il lavoro del giornalista, il diritto alla privacy, la comunicazione politica, la legge di internet, doveri dei provider, General data protection regulation.

Scienze sociali

Social Issues in the information society – L'economia digitale, perchè I dati sono il nuovo petrolio, la società a costi marginali, le relazioni economiche a rete, automazione e lavori, il valore economico dei dati, dati digitali e sviluppo sostenibile, il digital divide.

Social media analytics – Social Web, Social networks, Social media, Raccogliere dati da social media, crowling nel social Web, rappresentazioni basate su grafi, analisi di dati sociali, problemi aperti, sentiment analysis, emotions detection, Named entity recognition and Linking, Social media tagging e summarization.

Tecniche e modelli per la Data Science

Machine Learning – Introduzione al data mining. Classificazione supervisionata e non supervisionata, stima delle prestazioni, selezione delle caratteristiche, misure di validità, analisi associative.

Prendiamo ad esempio il corso di Machine Learning per dire che in questo e in altri corsi vengono effettuate sfide lanciate da aziende e pubblicate sul sito Kaggle (www.kaggle.com), nelle quali gli studenti sono invitati a partecipare alla concezione di una tecnica basata su Machine Learning per risolvere un problema decisionale o predittivo. Le migliori tesine sono premiate con un reward economico (occorre dire che fino ad ora sono stati pochi i casi di aziende che hanno scelto questo reward) ovvero un reward accademico, un voto alto all'esame di Machine Learning. Il lettore interessato (e arrivato fin qui nella lettura….) a esaminare tesine prodotte nel corso di Machine Learning può accedere all'indirizzo

https://www.dropbox.com/sh/l7m6589htim26o6/AACiHbcsjHAeKEMYAI5e0EDsa?dl=0

Decision models – Analisi decisionale, tipi di decisioni, decisione e incertezza, modelli di ottimizzaizone, concetti di simulazione, il valore della informazione.

Economia, marketing e uso dei dati a fini decisionali

Economics for Data science – Big data: nuove frontiere per l'analisi economica, esperimenti randomizzati e naturali, matching estimator, previsioni e simulazioni, modelli strutturali, valutazione delle policy ex ante, applicazioni empiriche. L'economia della innovazione, le regole della economia dei

dati digitali, la società a costo marginale, l'impresa digitale a rete, l'economia della condivisione, la nuova rivoluzione industriale.

Web Marketing – La gestione della relazione con il cliente, la vision unificata del cliente. Strategie di contatto. L'analisi dei social media, la gestione delle campagne di marketing. Identificazione di gruppi target.

Business intelligence – Scopi della Business Intelligence (BI), Architetture di BI, I key performace indicators, la knowledge discovery nelle basi di dati, modelli e metodi per la qualità dei dati e per la analisi dei dati, la visualizzazione nella BI.

Scienza dei servizi

Service Science - Il concetto di servizio, il valore dei servizi, strategie di business dei service provider, la progettazione dei servizi basata su conoscenza, open data e servizi pubblici, I servizi e I big data.

Domini applicativi della Data Science

- Laboratorio del primo anno – Programmare in R. Gestire i dati in R. Tecniche di learning, Random forest e support vector machine. Associazione e raccomandazione, Naïve Bayes. Studi di caso su vari domini della vita di ogni giorno

Laboratori del secondo anno

- Data Science Lab in Environment & Physics Big Data in Geographic Information Systems
- Data science lab in biosciences Big data in biotechnology & biosciences
- Data science lab in biosciences Making sense of biological data
- Data Science Lab in Medicine Big Data in Health Care
- Data Science Lab in Medicine Medical imaging & big data
- Data Science Lab in Business and Marketing - Big data in Business, Economics and Society
- Data Science Lab in Business and Marketing - Big data in Behavioural Psycology
- Data Science Lab in Public Policies and Services - Big Data in Public Health
- Data Science Lab in Public Policies and Services - Big Data in Public and Social Services
- Industry Lab - Big Data in Application Domains

L'industry Lab prevede una presenza attiva delle aziende che propongono agli studenti I loro problemi e li seguono nel corso dei progetti; i temi delle aziende riguardano una specifica tematica applicativa, per esempio nel primo anno di erogazione del corso la tematica è stata la manutenzione predittiva.

Altri contenuti liberamente accessibili

In questa sezione sono descritti alcuni corsi prodotti su piattaforme di eLearning ad accesso libero. Tali corsi sono inquadrati nel programma Bbetween della Università di Milano Bicocca, che prevede sia corsi faccia a faccia che corsi della tipologia MOOC (Massively Open OnLine Course), questi ultimi in collaborazione con il Consorzio Federica Web Learning della Università Federico II di Napoli. Per i corsi

sono previste due certificazioni Open Badge, la certificazione di frequenza e la certificazione di attività. Il programma dei corsi verrà arricchito nel tempo

Le basi della scienza dei dati - Carlo Batini

- Introduzione alla Scienza dei dati
- Il ciclo di vita dei dati – Gestione
- Il ciclo di vita dei dati – Analisi
- Strutture di dati nei linguaggi
- Strutture di controllo

Accessibile alla pagina https://www.federica.eu:443/c/le_basi_della_scienza_dei_dati

Linguaggi per la Data Science:Linguaggio Python – Gianluca Della Vedova

- Concetti base propedeutici alla conoscenza del linguaggio Python: variabile, istruzione di assegnazione, espressione logica, tabella, matrice, vettore.
- Organizzazione di un programma in Python.
- La struttura del linguaggio, i tipi di dati, le istruzioni
- I sottoprogrammi sviluppati nella comunità dei ricercatori e professionisti che realizzano un vasto insieme di tecniche di analisi e di visualizzazione
- Come si scelgono i package per risolvere un problema di analisi dati.

Accessibile alla pagina https://www.federica.eu/c/introduzione_a_python_per_la_data_science

Linguaggi per la Data Science: Linguaggio R - Matteo Pelagatti

- Caratteristiche essenziali dell'ambiente R,
- Le strutture dati e le principali strutture di controllo.
- Funzioni utili per la gestione dei dati,
- Realizzare semplici funzioni e creare grafici di base.
- Funzionalità dell'IDE R-Studio e dei pacchetti dplyr e ggplot2
- Tipi di file editabili in R-Studio (script, markdown, ecc.) e panel disponibili nell'interfaccia (source, console, environment, ecc.).

Accessibile alla pagina https://www.federica.eu/c/introduzione_a_r

Database Modeling and Design – Carlo Batini

- Introduzione alle basi di dati
- Il modello Entità Relazione
- Il modello relazionale
- Progettazione concettuale
- Progettazione logica

Accessibile alla pagina http://elearning.unimib.it/course/view.php?id=17573

Le prossime sezioni riportano materiali e contenuti di corsi erogati nel mondo nell'ambito di diverse discipline su temi riconducibili alla Scienza dei Dati; il materiale, tratto dai siti Web dei corsi, è frammentario e mancante di una sintesi adeguata, ma ho preferito includerlo in questo libro perché in questo modo ho la speranza di contribuire a lanciare un lavoro interdisciplinare per una fondazione più matura della Scienza dei Dati.

3. Scienze giuridiche

Corso su "Legal issues in Data Science" della Università di Pisa
https://esami.unipi.it/esami2/programma.php?c=38542

– The Algorithmic Society: the Classifying Society
– Background and Overview, Surveillance Society
– Big Other, Networks of Control, Predicting Behavior
– People Analytics, Behavioural "Nudging
– New Emerging Human Rights in the age of Behavioral Data Science and Neurotechnologies
– Towards "Mental Privacy" and "Decision Integrity"
– Legal and ethical implication of computational capacity.Building Legally-Compliant Algorithms:
– Legal Pitfalls of Algorithms, The Problems of Personalization, Data Handling & Sharing,
– Deploying Algorithms for Human Rights: Complications & Challenges
– Classification of Algorithms in the Information Society
– Legal Implications and Business Applications, Exploitation of Public Sector Data
– Competition Law in the Age of Algorithms, Transparency
– Accountability and traceability of algorithm based decision-making
– Accountability in the Machine Learning Context
– Technical and Legal Options to Enhance Transparency & Accountability
– Legal Liability for Algorithm Autocomplete (ISP Liability)
– Open Data Governance, Data Ethics.General principles of privacy law: The American approach, The European approach.The General Data Protection Regulation:
– Notions and principles, GDPR global reach and compliance
– Google Spain Decision
– Invalidation of Data Retention Directive (US Safe Harbour Decision)/Schrems.Privacy in operation– Privacy-by-Design, GDPR Solutions: The Right to an Explanation, etc.
– Notions of Privacy in the Algorithmic Age, Privacy from the Government
– Surveillance Capitalism, Governance by Proxy, Privacy from Private Entities
– Privacy from Platforms, Privacy from Employers, Privacy from our Devices (IoT).Comparative Perspectives & Crossborder Issues:– Comparative Privacy and security Regimes: GDPR vs. USA
– Comparative Privacy and security Regimes: GDPR vs. China.

4. Economia e management

<u>Temi di economia</u>

Corso di Digital Economy, University of Athens,
http://en.econ.uoa.gr/

Technological developments become relevant to the extent that they are economically and socially meaningful. This theoretical and methodological stance has inspired a rather inter-disciplinary approach to the study of the digital economy. There is an ongoing vivid discussion on themes such as:

- the emergence of new markets;
- network technologies and new forms of industrial and business organization;
- ICTs and productivity;
- the transformation of management practices; work and employment;
- social networks, trust and social capital in the digital economy; cultural and motivational aspects;
- ICTs and globalisation;
- ICTs and the political economy of inequality etc.

Understanding the uses and the impact of ICTs on economy and society presupposes thus a focus on the overall dynamics in which they are embedded. The main theoretical prerequisites for this kind of inquiry on shaping and impact of the digital economy can be found in three distinct but interrelated disciplinary fields: in economics, in information systems studies and in economic sociology.

The first part of lectures will be dedicated to the interface between economic and technological aspects of the emergence and the consequences of the digital economy. The main topics in this part are:
a) The macroeconomic perspective,
b) Market structure, Competition, and the Role of Small Business,
c) Employment and Human Resource Development,
d) Organisational Change and Economic Transformation.
In the second part the focus will be on principles of the new economic sociology and its contribution to the understanding of the digital economy.
The third part will discusses the relevance of central facets of information and network economics for the study of digital economy (as well as on their interface with approaches based on new economic sociology), and a concluding overview lecture.

Corso di Digital Economy, Graduate School of Economics, Barcelona
https://www.barcelonagse.eu/summer-forum/workshop-digital-economy

Platform Competition
-Competition & Switching Costs
-Network Externalities..
-Pricing in Two Sides Markets. Information Economics in Digital Markets. -Introduction to auctions. -"Digital" auctions Intellectual Property Rights- Patents, copyrights, and trademarks - R&D races - Empirical tools
- Innovation and advertising
- Litigation cases Competition Policy in Digital Markets
- Abuse of dominant position in digital markets (related competition cases: Google, Microsoft, qualcomm, etc...)

- Vertical restraints in online markets.
- Big data and competition
- Mergers in digital market

Corso su Business and Professions in the Digital Economy, Western Norway University of Applied Sciences

https://www.hvl.no/en/studies-at-hvl/study-programmes/course/mo%C3%B8235

- The economic and technological factors driving the digital revolution and how digital changes transform businesses and professions;
- The clash between existing business models and new digitally enhanced business models;
- Central technologies and concepts in the context of digital businesses and professions;
- Specific characteristics of digital business models and sources of competitive advantage in the digital economy;
- The opportunities and challenges for international and Norwegian companies in taking part in the digital competition;
- Ethical issues of the digital economy and organizations' use of data.

Temi di Management (titoli di corsi)

- Principles of Data Science - Gain an understanding of the foundations of data science and its applications.
- Data Science Processes, Impact, and Functions - Evaluate how the data science process can be used to address business problems.
- Modern Applications of Data Science - Realize the potential of data science in your organization through modern applications.
- Building a Data Science Team - Learn how to recruit, retain and develop data science talent for your organization.
- Measuring the Success of Data Science Projects
- Measure key performance indicators to recognize the drivers of value from data science impacts.
- Digital Transformation Strategies - Develop your planning and strategy skills by implementing a data science project.

5. Scienze sociali

Corso su Politics of Data, University of Michigan
http://pne.people.si.umich.edu/PDF/poldatasyll.pdf

Introduction: Learning from Big Data
- Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are
- Big Data: Are We Making a Big Mistake?
- Seeing the Sort: The Aesthetic and Industrial Defense of ‘the Algorithm

- How to Read a Book

Concepts: algorithmic culture, big data, information waste
- How Statistics Lost Their Power – And Why We Should Fear What Comes Next
- Data Doxa: The Affective Consequences of Data Practices
- Heteromation and Its (Dis)contents: The Invisible Division of Labor Between Humans and Machines
- Notes on Platform Capitalism
- Big Data is the Answer... But What is the Question?

Algorithmic culture
- Data Politics
- Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook.

Concepts and methods
- Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms
- Algorithmic Accountability: Journalistic Investigation of Computational Power Structures
- The Scored Society: Due Process for Automated Predictions
- Governance By Algorithms: Reality Construction By Algorithmic Selection on the Internet
- First I “like” it, Then I Hide it
- We’re building a dystopia just to make people click on ads
- How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms

Big Data, Machine Learning, and the Social Sciences
- How Big Data is Unfair: Understanding Unintended Sources of Unfairness in Data Driven Decision Making
- A Few Useful Things to Know About Machine Learning
- The Production of Prediction: What Does Machine Learning Want?
- Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
- Information Flow Experiments: Determining Information Usage from the Outside

Fake news and biased algorithms
- Social Bots Distort the 2016 US Presidential Election Online Discussion
- Polarization, Partisanship and Junk News Consumption Over Social Media in the US.
- Facebook overhauls News Feed in favor of 'meaningful social interactions’
- The Rise of the Weaponized AI Propaganda Machine
- When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software
- Trade Secrets and Algorithms as Barriers to Social Justice
- The Networked Nature of Algorithmic Discrimination.

6. Filosofia

Poiché nei capitoli precedenti non ho affrontato temi filosofici, in questa sezione adotto un approccio alla descrizione di corsi più introduttivo e discorsivo, basandomi su [Swan 2016].

Diversi filosofi considerano il campo nel suo insieme e le tre principali categorie filosofiche di ontologia (esistenza), epistemologia (conoscenza, logica e ragionamento) ed etica ed estetica. Altri considerano la storia dello studio del concetto di informazione, e la possibilità di unificare diverse teorie dell'informazione, vedi Figura 3.

Tipo di informazione	**Paradigma base**	**Tipo di teoria**
Informazione di Shannon	Probabilità	Quantitativa
Informazione di Fisher	Probabilità	Quantitativa
Complessità di Kolmogorov	Calcolo	Quantitativa
Informazione quantistica	Meccanica quantistica	Quantitativa
Semantica della informazione	Verità, Accuratezza	Qualitativa
Informazione come stato di un agente	Credenze (proposizioni che non sono necessaria-mente vere, ma che sono credute vere)	Qualitativa

Figura 3 – Le diverse teorie della informazione

Come per le altre filosofie della scienza, ci sono due livelli principali di interesse nella filosofia dei dati digitali, che possono essere considerati a diversi livelli di astrazione. Il primo livello è interno all'ambito di osservazione: riguarda un'articolazione dei concetti, teoria e sistemi che emergono per descrivere la natura generale della scienza all'interno dell'ambito. Il secondo è esterno all'ambito, e riguarda l'impatto più ampio su individui, società e mondo. Riguardo al livello esterno, vien investigata la collocazione nella società dei concetti di dato, informazione e conoscenza, valutati nel contesto delle questioni sociali, giuridiche ed etiche. Le tre categorie di interesse della filosofia (ontologia, epistemologia ed etica) si applicano ugualmente al livello pratico e al livello sociale della filosofia della informazione.

Primo livello interno di astrazione

- Domande epistemiche sulla produzione di conoscenza, ciò che conta come conoscenza, come conosciamo, qual è il metodo di prova e come è correlato con la sperimentazione empirica.
- Domande etiche ed estetiche sulla produzione di dati, come e quali dati vengono osservati, inclusi, omessi e valorizzati, e come vengono utilizzati nella pratica.

Secondo livello di astrazione: impatto sociale

- Domande sulla definizione ontologica del campo, quali sono le sue dimensioni, i confini e le traiettorie di crescita, cosa significa per gli umani e come il suo utilizzo pratico può cambiare il modo in cui pensiamo a noi stessi e al mondo.
- Casi di studio: Il cambiamento nel modo in cui gli esseri umani nell'accedere ai propri dati sulla composizione del genoma personale ha spostato la sensibilità su ciò che significa essere umani e ha abilitato nuove responsabilità e modi di agire nel mondo.

Altri temi

- Ruolo delle ipotesi a fronte di un numero potenzialmente illimitato di dati disponibili ed esperimenti su di essi, che può essere condotto senza che la domanda di ricerca debba essere formulata a priori.
- Influenza nella scienza dei dati del dinamismo intrinseco insito nella tecnologia e nelle fonti, in contrasto con la concettualizzazione tradizionale della Scienza come statica osservazione e analisi della realtà.
- Evoluzione del concetto di causalità rispetto a forme di interrelazione che non sono relazioni causali nel senso tradizionale, come correlazioni episodiche che innescano situazioni, che non presentano la struttura tradizionale di causa/effetto.

Concetti, metodi e strumenti della Scienza dei dati - Un insieme di concetti primari e delle domande filosofiche da essi suscitate è mostrato nella Figura 4.

Concetto	**Domanda filosofica**
Causalità	Come dobbiamo investigare le cause nell'era dei big data? Dobbiamo sviluppare una nuova visione della causalità conforme alle nuove pratiche?
Qualità	Come assicurarci che i dati sono di sufficiente qualità per gli scopi per cui sono utilizzati? Come affrontare il tema dell'open access? Quali nuove tecnologie sono necessarie?
Sicurezza	Come possiamo garantire la sicurezza dei dati, rendendoli tuttavia accessibili a coloro che hanno necessità di utilizzarli?
Big data	Quali novità introducono i big data nel metodo scientifico? Quali sono le sfide legate ai big data?
Incertezza	Possono i big data governare la incertezza, ovvero ne generano di nuova? Quali tecniche e tecnologie sono necessarie per ridurre la incertezza nelle scienze "data driven"

Figura 4 – Concetti della Scienza dei dati e domande filosofiche che suscitano

Rappresentazioni

I dati, come abbiamo visto in diversi capitoli e in particolare quello dedicato ai modelli (Capitolo 3) esistono in una varietà di "forme fisiche", e la loro percezione da parte degli esseri umani avviene sempre attraverso approssimazioni, astrazioni, o rappresentazioni. Ciò accade perché i dati non hanno nessuna forma olistica, intendendo con olismo il paradigma filosofico secondo cui le proprietà di un dato sistema non possono essere determinate dalla somma delle sue componenti, bensì è il sistema in generale che determina il comportamento delle parti, ovvero forme intrinseche che siano accessibili a essi umani. In altre parole, è impossibile comprendere il significato dei dati attraverso una esperienza diretta. Ciò suscita numerose questioni su quanto accurata sia una particolare rappresentazione nel

catturare il fenomeno sottostante che i dati ci rivelano. Nasce quindi la necessità di indagare il concetto di autenticità della rappresentazione, che include sia la questione ontologica riguardante il grado con cui una rappresentazione corrisponde al rappresentato, sia la questione epistemica che riguarda il come si possa conoscere che la rappresentazione sia accurata, e quale sia il metodo di prova adottato.

In conclusione, la filosofia dei dati digitali si occupa di concetti di base, metodi e implicazioni insite nell'uso dei dati digitali; e delle definizioni, le concettualizzazioni, la possibilità di conoscere, i livelli di verità e le pratiche, in situazioni che coinvolgono insiemi di dati caratterizzati da elevati volumi, alta velocità e alta varietà.

7. Etica

Corso su "Behind the Data: Humans and Values
Berkeley School of Information
https://www.ischool.berkeley.edu/courses/datasci/231

- Introduction to the legal, policy, and ethical implications of data, including privacy, surveillance, security, classification, discrimination, decisional-autonomy, and duties to warn or act.
- Examines legal, policy, and ethical issues throughout the full data-science life cycle collection, storage, processing, analysis, and use with case studies from criminal justice, national security, health, marketing, politics, education, employment, athletics, and development.
- Includes legal and policy constraints and considerations for specific domains and data-types, collection methods, and institutions
- Technical, legal, and market approaches to mitigating and managing discrete and compound sets of concerns strengths and benefits of competing and complementary approaches.

Course on Big Data, Big Responsibilities: The Law and Ethics of Business Analytics – Wharton Business School - University of Pennsylvania
https://mba.wharton.upenn.edu/

- The promise and the peril - How might data science change the relationships among firms, customers, employees, other firms, and governments? What are some of the legal or ethical concerns that may arise?
- "It's Just Math" - Algorithms rely on human decisions about how data are collected, analyzed, and used. Failure to appreciate this can lead to problems.
- Limits of Algorithms - The first step to responsible use of analytics is to appreciate its limitations and known statistical issues.
- Privacy in a Big Data World - Are there limits on how data should be collected, used, and shared?
- Privacy Law - Privacy is the subject of many legal and regulatory regimes in the U.S. and elsewhere. How well do those rules apply to big data and business analytics?
- Perils of Prediction - The ability to make predictions about individual behavior based on models means that "private" information, including very sensitive facts, can be inferred without ever having access to personal data. Can we even talk about privacy if personally identifiable information isn't being shared?

- Influencing Users - To what extent does analysis itself influence behavior? And what are the limits on using analytics not merely to understand and predict customer actions, but to shape them?
- Algorithmic Market Power - What issues arise when analytics are used for dynamic pricing? Should we be concerned about algorithmic monopolies or other anti-competitive practices?
- Algorithmic discrimination - When is a differential effect a neutral reflection of the state of the world, and when is it tantamount to illegitimate discrimination? The use of analytics has the potential both to counteract and to reinforce systematic biases against women, people of color, and other groups.
- Addressing disparate impact - The primary legal theory for unintentional discrimination is known as "disparate impact" in the United States. Can it apply to cases of algorithmic discrimination? Should it? Are there technical alternatives?
- Algorithmic inequality - As information and insight from data become more evaluable, should be concerned that their benefits are not universally available? Will analytics worsen the gap between haves and have-nots?
- Moving forward - How to balance the various interests involved to develop responsible approaches to business analytics. Can industry be part of the solution, and not just the problem?

Royal Society, London
https://royalsocietypublishing.org/journal/rsta

Phlosophical transactions of the Royal Society: Mathematical, Physical and Engineering Sciences.
Pubblicati molti lavori di ricerca che si riferiscono etica e società per la Data science, ad esempio:

Paper on "Opportunities for data science in government, the need of an Ethical Framework"
Approach to ethics by design

1. Start with clear user need and public benefit.
2. Use data and tools that have the minimum intrusion necessary.
3. Create robust data science models.
4. Be alert to public perceptions.
5. Be as open and accountable as possible.
6. Keep data secure.

Corso su "Human context and ethics of Data"
University of California, Berkeley
https://data.berkeley.edu/degrees/human-contexts-and-ethics

- How does data science transform how people live and how societies function?
- How do cultures and values inform how data-driven tools are developed and deployed?
- What assumptions do data-enabled algorithms and tools carry with them?
- What projections does AI make on the future?
- How can we shape the outcomes we want to see?
- Doing ethical data science amid shifting definitions of human subjects, consent, and privacy;
- The changing relationship between data, democracy, and law;

- The role of data analytics in how corporations and governments provide public goods such as health and security to citizens;
- Sensors, machine learning, and artificial intelligence and changing landscapes of labor, industry, and city life;
- The implications of data for how publics and varied scientific disciplines know the world.

Corso su Ethics for Data Science
Purdue Engineering, Purdue University
https://engineering.purdue.edu/datamind/datascience/19spring/phil-syllabus.pdf

- Identify ethical issues associated with applications of data science in a variety of professional settings by reading assigned texts and viewing/listening to assigned media content;
- Assess and critique the actions of individuals, corporations, governments and other organizations as ethical or unethical by participating in classroom and group discussions;
- Apply general ethical principles to the specific, concrete actions of individuals, corporations, governments and other organizations by completing written projects related to case study analysis;
- Formulate sound, well-reasoned arguments, and communicate them clearly, by writing reports implementing the case study procedure developed in class;
- Generate a case study of their own, by submitting a final case study report implementing the case study procedure developed in class.

8. Scienze cognitive

Corso su Cognitive modeling
Universitat Osnabruck
http://computational-cognition.eu/files/ikw_info.pdf

Cognitive modeling is the theoretical branch of cognitive psycho-logy concerned with simulating mental processes, typically with computer programs. Psychological experiments test human per-formance on cognitive tasks. Typical tasks include detecting faint sounds, tracking several objects on a screen, remembering lists of words, recognizing objects, learning new concepts, and solving puzzles. Cognitive models describe the processes and mechanisms that underlie human cognitive performance. For example, memory models describe in detail the differences between short-term and long-term memory and how the processes of encoding and retrieving information work. These models are specified formally in mathematical equations. Usually the models are implemented as computer simulations. Since the aim of cognitive modeling is to develop computer programs that behave intelligently in various cognitive tasks, it has close connections to artificial intelligence. However, contrary to programs in artificial intelligence, cognitive models should also have the same limitations as humans, e.g., in working memory or attention span. Hence, cognitive models not only explain how people master cognitive tasks, but also when and why they fail.

- How do we recognize objects?

- Seeing is effortless. You open your eyes and you see what is where. However, your eye only registers the light intensity on different parts of the retina. There aren't any objects in the image that is captured by the eye. The objects are detected in the image by your visual system. Edges and object boundaries have to be extracted from the image in order to separate the objects from the back-ground before objects can be recognized.
- Which representations underlie our capacity for recognizing objects?
- Which processes transform image-based representations into object-based rep-resentations?
- How are these processes implemented in the brain?
- How does memory work?
- Broadly, memory can be divided into short-term and long-term memory. Short-term memory can only hold small amounts of information for a short time, whereas long-term memory can store vast amounts of information for a much longer time. Like working memory in a computer, short-term memory only holds information that is immediately relevant to the task at hand.
- But how is information transferred between these two buffers? How is information encoded, retrieved, and forgotten? How does the structure of memory limit memory performance and what can be done to improve memory? What is the architecture of the mind?
- Just like a modern computer, the mind is hypothesized to consist of separate functional modules that interact with each other. Key components are different memory buffers, perceptual systems, the motor system, and a central controller. A cognitive architecture is a unified computer model of the mind that explains how these mod-ules interact. Can such a cognitive architecture explain intelligent behavior?

Corso su su Neurobiopsycology

- Can humans learn a new sensory modality?
- Using a sensory augmentation device, the feelSpace belt, subjects who train in the natural environment receive information on magnetic north via vibrating elements onto the waist. This changes how space is perceived, while increasing trust in navigational abilities.
- How does the level of visual information change the sampling behavior of an object? We investigate the infl uence of eye movements on the percep-tion and recognition of ambiguous objects. We demonstrate that action precedes perception, i.e., that there is a reverse sensorimotor coupling.
- How does visual, vestibular, and kinesthetic information interact while moving in space? In a fully immersive virtual reality setup, subjects perform a triangle completion task actively walking and turning. To study integration of multiple sensory modalities, we investigate brain activity using mobile EEG.
- Can partners in a joint visual search task use tactile and auditory cues to exchange gaze information?
- In a psychophysics experiment, we examine whether partners im-prove their performance through multisensory cueing, translating close links of their individual perception and action to a teamwork context.

Corso su Neuroinformatics

- How do neurons learn?

- We investigate how the brain can self-organize, adapt to new tasks, and compensate damage in neuronal systems. To this end, we use mathematical and simulation-based analyses of larger populations of neurons and define principles of how neurons can optimize information processing induced by neuronal plasticity.
- Can we build an artificial brain?
- We study and develop models for neuro-inspired computing devices composed of silicon neurons or even built by optical laser elements. The size and complexity of these artificial brains will equal the size of the human brain already in the coming years. We investigate how such systems can be used for computation and made adaptive in order to learn cognitive processes and behavior in a brain-like way. To this end, we implement cutting-edge machine learning principles, such as deep learning or reservoir computing, based on neuro-inspired artificial systems.
- Can we read brain activity?
- The brain processes information and produces electrical signals. We develop tools and technologies to read these signals and to decode the information they contain. This information is then used in brain-computer-interfaces that allow us to control cognitive processes and also enable us to identify critical states that can be used as early markers of diseased neuronal activity.

Corso su Psyco and Neurolingusitics

- How do we learn the rules of language?
- Language enables us to create an infinite number of sentences out of a limited number of words. This makes sentences sometimes very complex. When and how do we learn the relation between words in sentences such as “The cat the dog chased fought back”
- Are humans the only animals able to learn complex linguistic rules?
- How do we acquire the meaning of words?
- Many different indicators support the learning of word meanings. We are interested in finding out how meanings related to our different senses, such as the shape of an object, the ringing of a telephone, or the softness of a touch, are learned as aspects of word meaning by young infants. How are language and memory related?
- Language supports our memory. We easily remember linguistic information and often even predict upcoming sentence parts. How are memory and language related and which neural processes support both processes?
- What is special when we learn a second language?
- Production and comprehension in a second language is often more effortful compared to our native language. What are the neural differences between first and second language processing? How do the mechanisms of language learning change across development?

Corso su Computational linguistics

- How can we analyze the language used in spoken or text-based communication?
- Our research focus is the ubiquitous phenomenon of context dependence. We develop robust analyses of linguistic structure that can cope with degenerate input, e.g., the fact that both human dialogues and large corpora of text from the internet provide us with noisy, fragmentary, and ill-formed data. We design formal systems involving semantic/pragmatic representations that can

model, e.g., the interaction of lexical and compositional meaning with discourse and background knowledge. Such representations can be integrated into automated reasoning and human-computer interaction systems, providing more naturalistic user interfaces.

- How does linguistic information interact with non-linguistic experience?
- Using eye tracking methods, we measure at which point in time people focus on which objects in a visually presented scene while listening to a brief discourse.
- How can we make computers creative? We develop cognitively inspired models for enabling computers to compose music, to write poetry, or invent new mathematical concepts.
- How can we measure creativity?
- What are economically interesting domains for computational creativity?
- How can we base models on cognitive mechanisms?
- What are adequate computational models for analogy, conceptual blending, heuristic reasoning, or similarity?
- What are good formal methods in order to represent such mechanisms? To which extent do such mechanisms facilitate general intelligence?
- What is needed for artificial general intelligence?
- How can we construct generally intelligent systems that achieve human-level performance in a wide variety of tasks?
- Are there models for integrating low-level streams of input with high-level structured knowledge?
- Is the cognitive computing paradigm a way to achieve general intelligence?

Le seguenti tematiche appartengono al campo delle scienze cognitive, ma su esse non ho trovato corsi dedicati. Sono però a mio parere così importanti che tentiamo nel seguito di organizzare una bozza di contenuti.

Euristiche

- basata sulla reputazione, che consiste nel privilegiare alternative riconoscibili rispetto a quelle meno familiari.
- basata sull'endorsement (sostegno, appoggio), basata sulle valutazioni espresse da altri, cui affidiamo la nostra.
- basata sulla consistenza, fa riferimento sul confronto tra fonti per evidenziarne le differenze. Nel caso di inconsistenze, sono proposte varie tecniche per la scelta tra le alternative.
- di auto-conferma, che misura la credibilità sulla base della conferma delle precedenti credenze;
- basata sulla violazione delle aspettative, assume una fonte non credibile se essa ha violato le aspettative in precedenti circostanze.
- basata sull'intento persuasivo, tende a non considerare credibile il dato che si percepisce può soffrire di un pregiudizio

Pregiudizi (o bias), basato su [Wikipedia]

Il bias come lo intendiamo in questo testo è una forma di distorsione della valutazione sul significato o sulla qualità di un dato o insieme di dati. La mappa mentale di una persona presenta bias laddove è condizionata da concetti preesistenti non necessariamente connessi tra loro da legami logici e validi. Il

bias, contribuendo alla formazione del giudizio, può quindi influenzare un'ideologia, un'opinione e un comportamento. È generato in prevalenza dalle componenti più ancestrali e istintive del cervello.
Dato il funzionamento del sistema cognitivo umano, il bias è in genere considerato non eliminabile ma si può tenerne conto "a posteriori" (per esempio in statistica e nell'analisi sperimentale) o correggendo la percezione per diminuirne gli effetti distorsivi.

Un gran numero di bias (o distorsioni) è citato nella letteratura, ne vediamo di seguito alcuni, tratti dalla voce di Wikipedia (si veda anche in Appendice 1 il Cognitive Bias code, accessibile anche in versione interattiva all'indirizzo
https://upload.wikimedia.org/wikipedia/commons/6/65/Cognitive_bias_codex_en.svg):

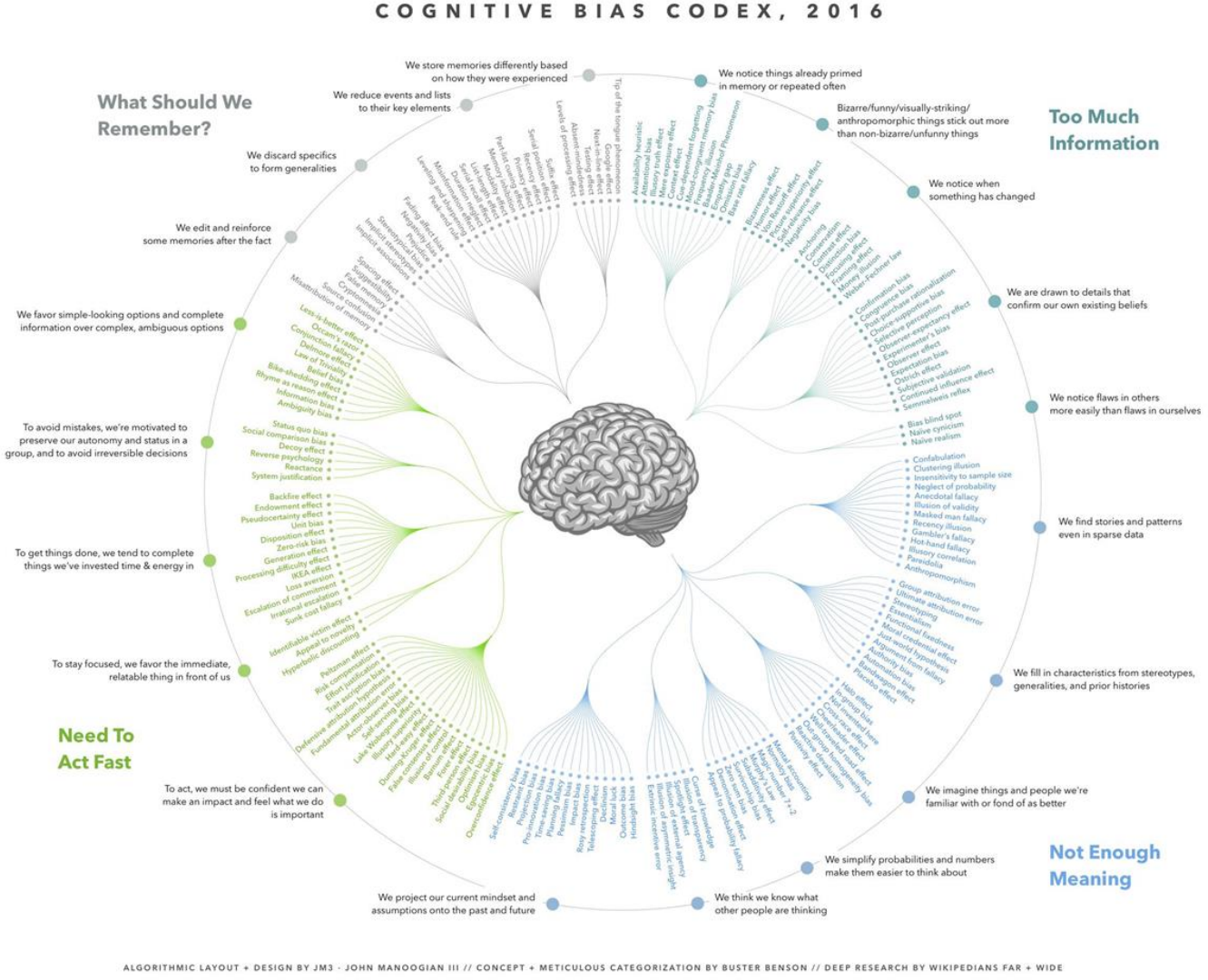
- Basata sull'incertezza - La tendenza ad evitare opzioni per le quali non è nota la probabilità del verificarsi di un risultato favorevole.
- Ancoraggio - La tendenza a fare troppo affidamento su un dato specifico quando si prendono decisioni (di solito il primo dato acquisito sull'argomento).
- Ragionamento antropocentrico - La tendenza a usare analogie umane come base per ragionare su altri fenomeni biologici meno familiari.
- Antroporfismo - La tendenza a caratterizzare animali, oggetti e concetti astratti come aventi tratti, emozioni e intenzioni simili all'uomo
- Bias basato sull'attenzione - La tendenza della percezione a essere influenzata da pensieri ricorrenti
- Basata sulla automazione - La tendenza a dipendere eccessivamente da applicazioni informatiche o basate sull'apprendimento automatico che possono portare a valutazioni errate che prevalgono sulle decisioni corrette
- Basata sulla memoria e sulle emozioni - La tendenza a sopravvalutare la probabilità di eventi che può essere influenzata da quanto recenti sono i ricordi o da quanto possono essere insoliti o emotivamente ricchi.
- Basata sull'effetto bandwagon - La tendenza a interpretare i dati sulla base del fatto che molte altre persone li interpretano allo stesso modo.
- Basata sul paradosso di Berkson – La tendenza a intepretare in modo erroneo esperimenti statistici basati sul proprietà condizionali.

9. Linguistica e Semiotica

Seminario su "Come non farsi ingannare dai dati – la post verità" [Lorusso 2019]
- Prospettiva semiotica
- Il ribaltamento semiotico
- Semantiche dei dati
- Dati, fatti, realtà
- L'illusione dei big data
- Dalla logica della corrispondenza alla logica della pertinenza
- I regimi di credenza
- L'illusione del fact checking
- L'infosfera in cui viviamo

- I rischi dell’epoca della post verità
- Surrogati di esperienza
- La semiotica e la sua utilità

Appendice 1 – Il Cognitive Bias Codex, 2016



Riferimenti

Acca – Ethics and Professional Skills Syllabus, 2019
https://www.accaglobal.com/gb/en/student/ethics.html

Ethics for Data Science - Purdue Engineering University of Purdue,
https://engineering.purdue.edu/datamind/datascience/19spring/phil-syllabus.pdf

A. M. Lorusso – Come farsi ingannare dai dati, la Post Verità – Seminario su Il cittadino consapevole verso la scienza dei dati, i big data, il machine learning, Digital Week, Cariplo Factory Lab 16 marzo 2019

Swan, M. (2015, March). Philosophy of big data: Expanding the human-data relation with big data science services. In *2015 IEEE First International Conference on Big Data Computing Service and Applications* (pp. 468-477). IEEE.

Drew C. Data science ethics in government. *Phil. Trans. R. Soc. A* **374**: 2016.01.19.
http://dx.doi.org/10.1098/rsta.2016.0119

L. Hunter, Ethics in Data Science – University of Colorado, School of Medicine
http://compbio.ucdenver.edu/Hunter

University of Pennsylvania, Warthon Business School - Big Data, Big Responsibilities: The Law and Ethics of Business Analytics, 2016.

J. Metcalf, Kate Crawford, and Emily Keller Pedagogical Approaches to Data Ethics, Data & Society Research Institute, Council for Big Data, Ethics, and Society, April 21, 2011

University of California, Berkeley Human Contexts and Ethics of Data - www.hce-sts.org

World Economic Forum - The Future of Jobs Report, 2018

Epilogo

La Scienza dei dati e la categoria della Datacy sono ancora molto giovani, e solo agli albori di sviluppi che permetteranno di sistematizzare il corpo di conoscenze che ad esse fanno riferimento. Tuttavia, il loro ruolo nei nuovi lavori e nella società si sta già imponendo in modo rilevante. La Figura 1 tratta da [The Future of Jobs, World Economic Forum 2020] sui lavori in crescita e in decadimento secondo i manager delle principali aziende mondiali, indica chiaramente che le tematiche della Data Science e della Intelligenza Artificiale sono le più indicate per creare nuovi lavori.

↗ Increasing demand		↘ Decreasing demand	
1	Data Analysts and Scientists	1	Data Entry Clerks
2	AI and Machine Learning Specialists	2	Administrative and Executive Secretaries
3	Big Data Specialists	3	Accounting, Bookkeeping and Payroll Clerks
4	Digital Marketing and Strategy Specialists	4	Accountants and Auditors
5	Process Automation Specialists	5	Assembly and Factory Workers
6	Business Development Professionals	6	Business Services and Administration Managers
7	Digital Transformation Specialists	7	Client Information and Customer Service Workers
8	Information Security Analysts	8	General and Operations Managers
9	Software and Applications Developers	9	Mechanics and Machinery Repairers
10	Internet of Things Specialists	10	Material-Recording and Stock-Keeping Clerks
11	Project Managers	11	Financial Analysts
12	Business Services and Administration Managers	12	Postal Service Clerks
13	Database and Network Professionals	13	Sales Rep., Wholesale and Manuf., Tech. and Sci.Products
14	Robotics Engineers	14	Relationship Managers
15	Strategic Advisors	15	Bank Tellers and Related Clerks
16	Management and Organization Analysts	16	Door-To-Door Sales, News and Street Vendors
17	FinTech Engineers	17	Electronics and Telecoms Installers and Repairers
18	Mechanics and Machinery Repairers	18	Human Resources Specialists
19	Organizational Development Specialists	19	Training and Development Specialists
20	Risk Management Specialists	20	Construction Laborers

Source

Figura 1 – Nuovi lavori, lavori in decadimento

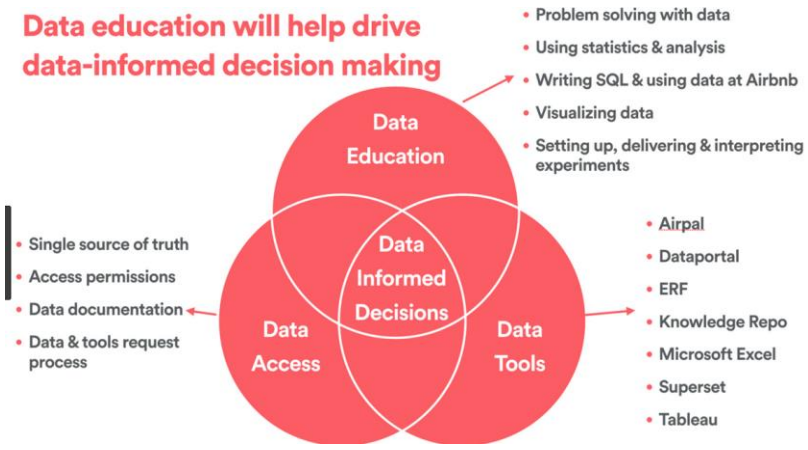


Figura 2 – La Scienza dei dati per decisioni informate

Accanto ai nuovi lavori, la Scienza dei dati, come ci dice la Figura 2, tratta da Je. Feng, E. Coffman & E. Grewal "How Airbnb Democratizes Data Science With Data University", è sempre più importante per prendere decisioni informate e consapevoli, aspetto questo che coinvolge tutti i livelli della società; decisioni su come votare alle elezioni, su come curarsi, su quale scuola scegliere per i propri figli, su come comportarsi nelle reti sociali, su come vivere la propria vita e su come vivere con gli altri, essendo esseri liberi che pensano con la propria testa e che si realizzano nel mondo vedendo sempre le tecnologie come un mezzo e non come un fine.

In conclusione, questo libro, che è largamente imperfetto, ha cercato di sondare tutto quanto possa essere sondato, di dissodare quanto possa essere dissodato, di collocare quanto possa essere collocato al suo posto, per costruire su basi solide la nuova Scienza dei dati.

Caro lettore, grazie per essere arrivato fino a qui.

Milano, Dicembre 2020

Carlo Batini
Federico Cabitza
Paolo Cherubini
Anna Ferrari
Andrea Maurino
Roberto Masiero
Matteo Palmonari
Fabio Stella

Per approfondire

Carlo Batini

In questa sezione sono citati diversi testi che possono essere letti per approfondire gli argomenti trattati.

A. Baricco – The game, Einaudi Stile Libero, 2018.

Da una delle presentazioni del libro “Quella che stiamo vivendo non è solo una rivoluzione tecnologica fatta di nuovi oggetti, ma il risultato di un'insurrezione mentale. Chi l'ha innescata - dai pionieri di Internet all'inventore dell'iPhone - non aveva in mente un progetto preciso se non questo, affascinante e selvaggio: rendere impossibile la ripetizione di una tragedia come quella del Novecento. Niente più confini, niente più élite, niente più caste sacerdotali, politiche, intellettuali. Uno dei concetti più cari all'uomo analogico, la verità, diventa improvvisamente sfocato, mobile, instabile. I roblemi sono tradotti in partite da vincere in un gioco per adulti-bambini.”

C. Batini, S. Ceri, S. Navathe – Conceptual Database Design – Benjamin and Cummings, 1992

Il libro più importante che ho scritto (prima di questo…) e anche quello che ha avuto maggiore diffusione nel mondo. Sistematizza tutto ciò che c’era da dire sulla progettazione di basi di dati, concentrandosi di più sulla parte concettuale. Vi sono tanti esempi ed esercizi. La qualità del libro è in parte scaturita da un severissimo assessment che fu fatto sulla prima versione, e che lo migliorò moltissimo.

C. Batini, M: Scannapieco – Data and Information Quality: Dimensions, Principles and Techniques, Springer Verlag 2016.

Questo libro attesta anche attraverso il numero di riferimenti (oltre 700) il grande sforzo fatto con Monica Scannapieco per sistematizzare il tema vastissimo della qualità dei dati, per i dati costituiti da relazioni, e per altre tipologie di dati quali ad esempio immagini, mappe, linked open data, testi, leggi.

C. Batini – Data Base Modeling and Design, Creative Commons licence, 2016

Libro e sui modelli per basi di dati e sulla progettazione di basi di dati, consiste nelle trascrizioni delle lezioni del corrispondente corso on line accessibile alla pagina https://open.elearning.unimib.it/enrol/index.php?id=52, accedere con provider Google (2020)

Il libro è liberamente scaricabile dalla pagina http://hdl.handle.net/10281/97114

Batini, C., Castelli, M., Comerio, M., Cremaschi, M., Iaquinta, L., Torsello, A., et al. (2015) - The Smart Methodology for the Life Cycle of Services. Creative commons Licence.

Rivisitazione del ciclo di vita dei servizi rispetto al libro Springer del 2010, liberamente scaricabile dalla pagina http://hdl.handle.net/10281/98632

Batini, C. Le Basi dell'Informatica: concetti e metodi per usare bene i calcolatori. Roma, Editori Riuniti 1984. Fuori catalogo.

E' il libro citato nel prologo; scritto 35 anni fa, guarda ai concetti dell'Informatica, non alle tecnologie, per cui è ancora attuale. Scaricabile all'indirizzo http://hdl.handle.net/10281/ 97703

C. Batini, G. Viscusi, M. Mecella – Information Systems for eGovernment, a Quality of Service perspective - Springer, 2010.

E' il libro in cui ho cercato di capitalizzare la mia decennale attività presso l'Autorità per la Informatica nella Pubblica Amministrazione, ente che fu creato per indirizzare le Amministrazioni Centrali in Italia ad un più efficace ruolo di servizio verso cittadini e imprese, utilizzando al meglio le tecnologie informatiche. Centrato quindi sul concetto di servizio, fornisce un metodo per progettare servizi digitali di qualità utilizzando le tecnologie ICT

C. Batini – Abstractions in Computer Science and surroundings – Tutorial alla Entity Relationship Conference, Haiku, Japan, 2016, accessibili alla pagina https://boa.unimib.it /preview-item/187839?queryId=mysubmissions&

Chi fosse interessato a immergersi in circa 400 astrazioni che ho scovato nelle discipline più disparate, dal conceptual modeling, alle visualizzazioni, i sistemi formali, i modelli per processi, le mappe, i grafi, senza naufragare, è invitato a percorrere un viaggio che ho trovato straordinario; "ho visto cose, che voi… no, cosa state pensando, che vorrei condividere…".

C. Batini - Corso online su "Le basi della scienza dei dati" Federica Web Learning, 2019
Liberamente Accessibile alla pagina
https://www.federica.eu:443/c/le_basi_della_scienza_dei_dati

Chi ha letto questo libro, non trova nel corso niente di nuovo. Chi non vuole leggere questo libro, ma vuole avere una panoramica generale sulla scienza dei dati, può essere interessato a frequentare questo corso on line.

E. Bencivegna – La scomparsa del pensiero, perché non possiamo rinunciare a ragionare con la nostra testa, Feltrinelli 2017

Dalla presentazione: “un saggio schietto e tagliente che ci mette in guardia rispetto alle insidie di una mutazione antropologica che sottrae alla nostra specie la sua risorsa più preziosa: il ragionamento.”

J. Bertin - Semilogy of Graphics – Diagrams, Networks, Maps- Esri Press 1993.

Una vertiginosa sequenza di simbolismi usati per rappresentare la Terra e, in particolare, la Francia. La tesi di fondo del libro è: l’elaborazione dei dati comporta lamenibilità, lacomunicazione richiede semplificazione

M. Boisot – Information Space, a Framework for learning in organizations, institution and culture, Rutledge 1995

Libro difficile, molto più difficile e direi dispersivo rispetto ad altri lavori in cui Boisot mi ha affascinato con il suo I-Space, lo spazio della informazione, le cui coordinate sono l’astrazione, la codifica, la scarsità, dimensioni che definiscono lo scambio di artefatti tra organizzazioni e quindi le basi della economia di mercato. La intuizione che la astrazione possa essere una fondamentale dimensione economica, con la sua capacità di far comunicare efficientemente gli umani negli scambi economici mi ha profondamente colpito.

P. Domingos – The Master Algorithm, how the quest for the ultimate learning machine will remake our world - Basic Books, 2018.

Un libro ricchissimo di esempi di tecniche di machine learning (“if the data-ism is today’s philosophy, this book will be its bible”). Il libro compie un viaggio partendo dal principio di induzione, commentando la regola di Bayes, e discutendo la rivoluzione insita negli algortimi che si costruiscono da soli. Avendo come obiettivo ultimo il Master Algorithm, capace di scoprire ogni forma di conoscenza dai dati, e di scoprire dai dati ogni cosa ancor prima che noi la chiediamo.

G. da Empoli – Gli ingegneri del caos, teoria e tecnica dell’internazionale populista, Marsilio Nodi, 2019

Analizza il nuovo populismo alla luce dei cambiamenti che l’analisi dei Big Data ha provocato nel messaggio politico, e il ruolo che gli ingegneri e gli spin doctors stanno svolgendo per spingere questo cambamento.

W. Davies – Stati Nervosi – Come l’emotività ha conquistato il mondo, Einaudi, 2019

Libro molto interessante, con una a traduzione approssimativa che talvolta infastidisce, sugli strumenti culturali e il pensiero che nella storia hanno influito sulla pace e sulla guerra, sulla relazione tra elite e popolo, tra esperienza e ignoranza, tra razionalità ed emotività, con alcune idee che ho trovato nuove sul declino dei numeri e della statistica a favore degli elementi soggettivi ed emotivi nella formazione del consenso.

B. Duffy – I rischi della percezione: perché ci sbagliamo su quasi tutto, Einaudi 2019.

Dalla presentazione: “Bobby Duffy, basandosi su una ricerca esclusiva condotta da Ipsos su quaranta Paesi, ci spiega perché non conosciamo i fatti fondamentali relativi al mondo che ci circonda. Il ventaglio di temi trattati è eccezionalmente ampio e restituisce una panoramica unica sulle aspettative che la popolazione del mondo ha della realtà statistica. L'esito, lampante e sconvolgente, è che a prescindere da età, livello di istruzione e ceto sociale sviluppiamo tutti una visione distorta di pressoché ogni aspetto della realtà. Duffy si chiede perché e, svincolando il concetto di «ignoranza» dalla sua accezione negativa, esamina come pensiamo e cosa ci viene detto per produrre questi risultati fittizi.”

U. Eco – Sulle spalle dei giganti – La nave di Teseo, 2017. Sono le conferenze che Eco teneva alla Milanesiana, segnalato qui per il saggio su “Dire il falso, mentire, falsificare” che ricorda la dedica a Eco di A.M. Lorusso nel libro La postverità: Per Umberto, che mi ha insegnato a ragionare sul falso molto più che a credere il vero.

V. Eubanks – Automating inequality – How High Tech tools profile, police, and punish the poor - St Martin’s Press 2017.

Virginia Eubanks si basa su studi di caso, nei quali riscopre la tesi di fondo alla base del libro: l’automazione aumenta le differenze tra ricchi e poveri, e lo fa in modo opaco e quindi molto più difficilmente opponibile.

M. Ferraris – Postverità e altri enigmi, Il Mulino, 2017

Dalla presentazione “L’ideologia che animpstverità è l’atomismo di milioni di persone convinte di avere ragione non insieme (come credevano, ,sbagliando, le chierse ideologiche del secolo scorso) ma da sole.

L. Floridi - La quarta rivoluzione. Come l'infosfera sta trasformando il mondo, Raffaello Cortina, 2017.

Dalla presentazione del libro “Luciano Floridi sostiene che gli sviluppi nel campo delle tecnologie dell'informazione e della comunicazione stiano modificando le risposte a domande così fondamentali. I confini tra la vita online e quella offline tendono a sparire e siamo ormai connessi gli uni con gli altri senza soluzione di continuità, diventando progressivamente parte integrante di un' "infosfera" globale. In ogni campo della vita, le tecnologie della comunica-zione sono diventate forze che strutturano l'ambiente in cui viviamo, creando e trasformando la realtà. Saremo in grado di raccoglierne i frutti? Quali, invece, i rischi impliciti? Floridi suggerisce che dovremmo sviluppare un approccio in grado di rendere conto sia delle realtà naturali sia di quelle artificiali, in modo da affrontare con successo le sfide poste dalle tecnologie correnti e dalle attuali società dell'informazione.”

C. Guzzanti – Aborigeno, vedi https://www.youtube.com/watch?v=l-qD_3o_obg

Guzzanti sintetizza problematiche complesse in tema di dati digitali con un monologo di 15 secondi; un genio.

D. Hand – Il tradimento dei numeri: i dark data e l'arte di nascondere la verità, Rizzoli 2019

Libro originale nell'occuparsi non dei dati che abbiamo, ma di quelli che non abbiamo, ovvero che crediamo erroneamente di possedere, suddividendoli nelle diverse tipologie e cercando di capire quali problemi creino e come si possa fare per imparare a gstire i problemi che essi provocano.

Y. N. Harari – 21 Lezioni per il XXI secolo – Saggi Bompiani 2018

Libro che definirei tecno ottimista, in cui si affrontano in maniera eccentrica e originale e con scrittura molto scorrevole un vasto insieme di questioni, dai migranti, alle fake news, ai cambiamenti climatici, al terrorismo, al ruolo dei modelli predittivi, ecc. Nella seconda di copertina, e all'inizio del libro si afferma che in un mondo inondato da informazioni irrilevanti, la lucidità è potere. E conclude: se questo libro servirà ad aggiungere al dibattito sul futuro della nostra specie anche solo un numero ristretto di persone, allora avrà raggiunto il suo scopo.

Peccato che un sito russo, the Insider, abbia scoperto che nell'edizione russa di 21 lezioni per il XXI secolo, alcune parti siano state cambiate per sostituire o rimuovere i passaggi critici verso il governo del presidente Vladimir Putin. In un passaggio della versione inglese e internazionale, Harari parla di post-verità facendo riferimento alla propaganda russa sull'Ucraina, e scrivendo che quando Putin smentiva che ci fossero truppe russe in Ucraina, definendole formazioni filorusse autonome, «sapeva perfettamente di mentire». Nella versione russa, per parlare del tema Harari cita invece il presidente statunitense Donald Trump, citando un conteggio delle sue bugie fatto dal *Washington Post.* I rappresentanti di Harari hanno risposto con una lettera a *Newsweek* in cui hanno spiegato che la preoccupazione principale dell'autore è «permettere che le idee centrali del libro, sulle minacce rappresentate dalle dittature, dagli estremismi e dall'intolleranza, raggiungano pubblici diversi». "Alla faccia" del numero ristretto di persone.

J. Haskel e S. Westlake – Capitalismo senza capitale, l'ascesa dell'economia intangibile, Franco Angeli 2018.

Libro dedicato alla economia intangibile, quella dei servizi basati sui dati digitali, sulle sue leggi originali rispetto alla economia classica, e sulle possibili conseguenze in termini di distribuzione della ricchezza e diseguaglianze sociali.

M. Karutani – La Morte della Verità, la menzogna nell'era di Trump, Solferino 2018

Riflessioni sul concetto di verità e post-verità della critica letteraria del New York Times. Con un percorso di analisi storica che include il relativismo, il decostruzionismo, il culto narcisistico del sé, la riscrittura narcisistica della storia, il discredito della scienza.

J. Posetti and Alice Matthews - A short guide to the history of 'fake news' and disinformation, International Center for Journalism, https://www.icfj.org/news/short-guide-history-fake-news-and-disinformation-new-icfj-learning-module, 2018.

J. Lanier – Ten Arguments for Deleting your Social Media Accounts Right Now, The Bodely Head, London 2018.

La tesi del libro è che se vogliamo una vita più felice, un mondo più pacifico e meno nervoso o semplicemente la possibilità di ragionare da soli senza essere influenzati dalle corporazioni più ricche della storia umana, allora la cosa migliore che possiamo fare è cancellare il nostro account dalle reti sociali.

A.M. Lorusso – Postverità – Edizioni Laterza, 2018

Il libro e' una analisi nell'ambito della filosofia del linguaggio del concetto di postverità. Oggi "realtà e finzione si sono intrecciate in una logica culturale che premia le emozioni e le identificazioni, non la messa alla prova e le competenze. "

V. Mayer-Schonberger e K. Cukier – Big Data, a Revolution that will transform how we Live, Work and Think – John Murray 2013.

E' il libro più bello che io abbia mai letto sui Big Data e sulle conseguenze del proliferare dei dati digitali sulle scienze, sulle tecnologie e sulla nostra vita: intelligente e informativo.

V. Mayer-Schonberger e T. Ramge – Reinventing capitalism in the Age of Big Data – Basic Books, 2018.

E' il manifesto del data capitalism. Mentre il capitalismo classico usa i prezzi come espressione del valore d'uso di beni e servizi, i dati che noi generiamo e che sono utilizzati dai fornitori per capire e condizionare le nostre intenzioni di acquisto permettono di collegare acquirenti e venditori in modo più efficiente, riducono i costi di ingresso nella creazione della impresa, e allo stesso tempo aumentano i rischi di monopolio, in ogni caso cambiano radicalmente la economia.

M. Monmonnier – How to Lie with Maps, The Uiversity of Chicago Press – 1991

Questo saggio piacevole da leggere sull'uso e l'abuso delle mappe ci insegna come consi-derare criticamente le mappe e promuove un sano scetticismo su questi modelli della realtà così facili da manipolare.

K. Murphy – Machiine Learning: a probabilistic perspective – The MIT Press, 2012.

Con le sue oltre mille pagine e I suoi circa milleduecento riferimenti è la Bibbia del Machine learning. Libro difficile, per esperti.

T. Munzner – Visualization Analysis & Design, CRC Press, 2015
Il libro più completo e chiaro su tutti I temi della visualizzazione.

D. Norman – La Caffettiera del Masochista - il Design degli oggetti quotidiani – Giunti 2014.

L'interfaccia uomo macchina è componente fondamentale della rivoluzione digitale, perché per poter usare le macchine dobbiamo essere in grado di comunicare con loro. Quante volte ci arrendiamo difronte a oggetti che non siamo in grado di maneggiare, oppure di fronte a un messaggio inviatoci da una app? La visione di Norman è che spesso ci rroviamo di fronte a interfacce mal progettate.

T. Nichols – The Death of Expertise, the Capaign against Esta blished Knowledge and Why it Matters, Oxford University Press 2017.

Grazie alla innovazione tecnologica e allo sviluppo del Web e dei motori di ricerca, oggi ognuno di noi conosce, o crede di conoscere, ogni aspetto della realtà e della scienza. Tutte le opinioni, anche le più ridicole, hanno pari dignità sul Web, e ogni tentativo di riflessione è bollato come elitismo antidemocratico. Tom Nichols analizza il fenomeno della morte della competenza.

C. O'Neill - Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy

Libro provocatorio e ripetitivo, ma con punti di vista molto originali sul tema della diseguaglianza indotta dal fenomeno dei big data digitali. Una per tutte, l'affermazione per cui in un mondo massificato dalle procedure automatizzate, gli unici che continueranno a interagire con quella risorsa preziosa che sono i formatori umani saranno i figli dei ricchi.

J. Pearl e D. Mc Kenzie – The book of Why, the new Scence of Cause and Effect, Allen Lane 2018.

Libro fondamentale, per comprendere il grande salto di paradigma tra le vecchie tecniche per l'analisi dei dati e l'innovazione introdotta dal machine learning, per comprendere i limiti intrinseci che riguardano l'osservare i dati, cioè il passato, rispetto alle tecniche che realmente possono portare le macchine a diventare autonome, e il grande salto che interviene nel passare dal cosa al perché. L'ultimo capitolo e' quanto di più profondo ci sia sul libero arbitrio, sull'etica e sulla posibilità per le macchie di imitare il pensiero umano.

N. Polson and J. Scott – AIQ: How Artificial Intelligence works and how we can harness its power for a better world. Penguin Random House, 2018.

Bellissimo libro, che parte con una premessa: se vogliamo comprendere il mondo moderno e dove ci sta portando, dobbiamo conoscere un po' del linguaggio matematico "parlato" dalle macchine intelligenti. Questo linguaggio è raccontato nel libro attraverso storie piuttosto che attraverso

formalismi. Da Isacco Newton a Florence Nightingale queste storie parlano di dati, probabilità e migliori strumenti per ragionare, che nella nostra epoca vengono ora rivisitati e applicati ai big data e alle macchine intelligenti. Il lettore è portato per mano, avvertito e incoraggiato nei momenti impervi, premiato quando si raggiunge la meta.

S. Quintarelli – Capitalismo immateriale, le tecnologie digitali e il nuovo conflitto sociale, Bollati Boringhieri, 2019.

Siamo pronti a gestire le conseguenze dell'immane scossone che la Rivoluzione Digitale sta provocando? Il libro aiuta a capire il nuovo ambente digitale (quello che Baricco chiama l'oltremondo e Floridi la infosfera) in cui stiamo vivendo, secondo la nuova prospettiva economica dei beni immateriali, occupata ormai sena confini e limiti dai dati digitali.

J. Rifkin – The zero marginal cost society, Palgrave MacMillan 2014.

Rifkin analizza la grande trasformazione che le tecnologie digitali e l'Internet delle cose stanno apportando alla società e alla economia, riducendo i costi di produzione e di distribuzione che per i dati digitali e i servizi ad essi collegati sono "zero-marginal cost". L'economia si sta trasformando in un mercato ibrido che vede convivere il mercato capitalistico con una nuova economia dei beni comuni.

H. Rosling – Factfullness, the Reasons we're wrong sbout the world – snf Why Things sre better than you think.

Il progresso umano è osservato con ottimismo da Rosling, e i suoi occhiali per vedere e analizzare in maniera chiara la realtà è la statistica descrittiva, usando la quale combatte molte idee preconcette e pessimistiche sul mondo.

A. Rusbridger - Breaking News: The Remaking of Journalism and Why It Matters Now, Casnongate 2018.

La storia del Guardian raccontata da un suo Direttore, dai tempi in cui per fare una edizione, il giornale doveva passare per 30 mani, all'epoca di Twitter e delle fake news, che sono molto più economiche da pubblicare di una notizia verificata, con alcune proposte per non far morire i giornali. Un libro splendido.

D. Rushkoff – Team human, W. Norton & Company, 2019

Il libro analizza con enfasi e talvolta moralismo il ruolo degli esseri umani come creature sociali, in contrasto alla atomizzazione e radicalizzazione che il Web provoca nelle relazioni sociali

E. Sadin – La siliconizzazione del mondo, l'irresistibile espasione del liberismo digitale, Einaudi 2018.

L. Saitta e J.D Zucker - Abstraction in Artificial Intelligence and Complex Systems, Springer Verlag 2015

Dalla presentazione: “Abstraction is a fundamental mechanism underlying both human and artificial perception, representation of knowledge, reasoning and learning. This mechanism plays a crucial role in many disciplines, notably Computer Programming, Natural and Artificial Vision, Complex Systems, Artificial Intelligence and Machine Learning, Art, and Cognitive Sciences. This book first provides the reader with an overview of the notions of abstraction proposed in various disciplines by comparing both commonalities and differences. After discussing the characterizing properties of abstraction, a formal model, the *KRA* model, is presented to capture them.”
C. Shapiro e H. Varian – Information Rules, a Strategic Guide to Network Economy, Harvard Business review Press, 1999.

Dalla presentazione: “la Sylicon Valley è oggi il luogo di una frenesia innovatrice che intende ridefinire ogn iaspetto della nostra esistenza per fini privti, diciharando tutta via di agire per il bene della umanità. Ma la Sylicon Valley non è solo un territorio; è oggi soprattutto una mentalità, che sta muovendosi per colonizzare il mondo.”

Una guida alle leggi della economia digitale, scritta da due professori che hanno fatto ricerca e consulenza sull’argomento da tempo; tratta su come investire negli information asset nell’epoca delle reti.

R. Staglianò – Lavoretti, così la Sharing Economy ci rende tutti più poveri, Einaudi 2018

Scritto con stile giornalistico, si legge bene; con una miriade di esempi fa vedere come l’economia della condivisione creata dalla diffusione di reti di dati digitali e servizi tenda a far scomparire molti lavori, creandone altri che spesso sono però molto più parcellizzati e sottopagati, concentrando immensi capitali nelle mani di pochi come mai si era visto nel passato.

M. Thompson – What’s gone wrong with the language of politics? - St. Martin’s Press, 2016.

Il linguaggio politico è profondamente cambiato, e non solo per twitter e you tube. Il messaggio svalutativo di Sarah Palin sul “Death Panel” che sarebbe stato contenuto nella riforma della sanità di Obama, fu diffuso attraverso la televisione. La tesi dell’autore è che la crisi della politica è soprattutto una crisi d linguaggio.

E. Tufte – The Visual Display of Quantitative information, Graphic Press, 2001.

Uno dei libri più belli che abbia mai letto, straordinaria sequenza di visualizzazioni efficaci e banali, veriteire e mentitrici, con una teoria dei grafici sui dati che parte dall’inchiostro e arriva alla estetica nel disegno dei grafi.

S. Vaidhyanathan – Anti-social media, How Facebook Disconnects Us and Undermines Democracy, Oxdord University Press, 2018.

Critica feroce a Facebook, analizza come la rete sociale è evoluta da un innocente sito sociale creato da student di Harvard in una froza che, se può rendere la vita personale più piacevole, rende la salute della democrazia più incerta e sfidante da mantenere.

J. Wajcman – Pressed for Time, the acceleration of time in digital capitalism, the University of Chicago Press, 2015.

E' una approfondita analisi sul ruolo del tempo, della accelerazione del tempo che stiamo vivendo nella era digitale, del paradosso per cui più usiamo efficientemente il tempo, e più il tempo ci manca.

M. Wolfe – Reader, Come Home – Harper Collins, 2018

Il libro parla di scienze cognitive applicate alla lettura dei libri, nell'epoca del libro virtuale e dell' eBook. Nell'epoca della nostra immersione totale nel Web e del nostro affidarsi orami totale agli strumenti digitali, il nostro modo di elaborare il linguaggio scritto sta cambiando radicalmente. Il libro cerca di rispondere ad alcune domande cruciali: riusciremo in futuro a sviluppare processi cognitivi sul pensiero critico, la riflessione, l'empatia? La domanda di at-tenzione continua insita negli apparati digitali, e l'accesso a immense quantità di informazione avrà impatto sulla nostra capacità di fare analogie, costruire inferenze, arrivare a valutazioni indipendenti? Diventeremo più sensibili e autonomi nell'affrontare la minaccia delle fake news e della demagogia, e, in questo senso, che ne sarà della democrazia?