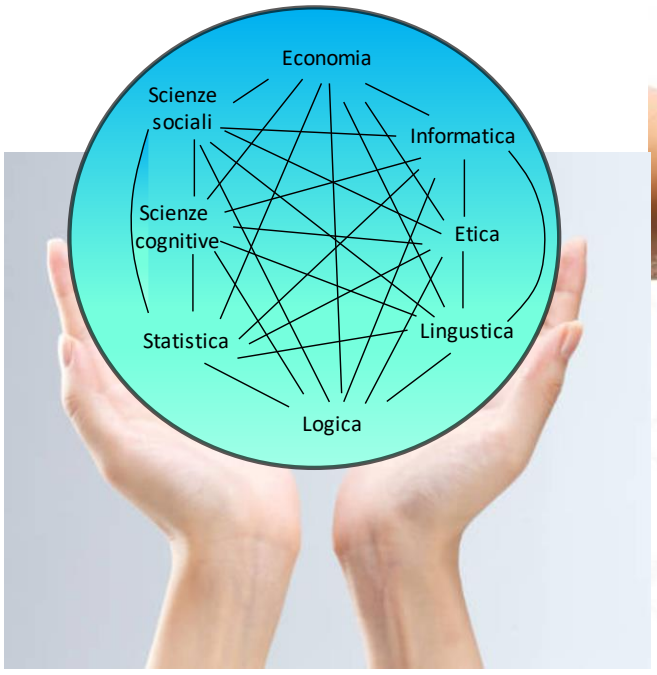
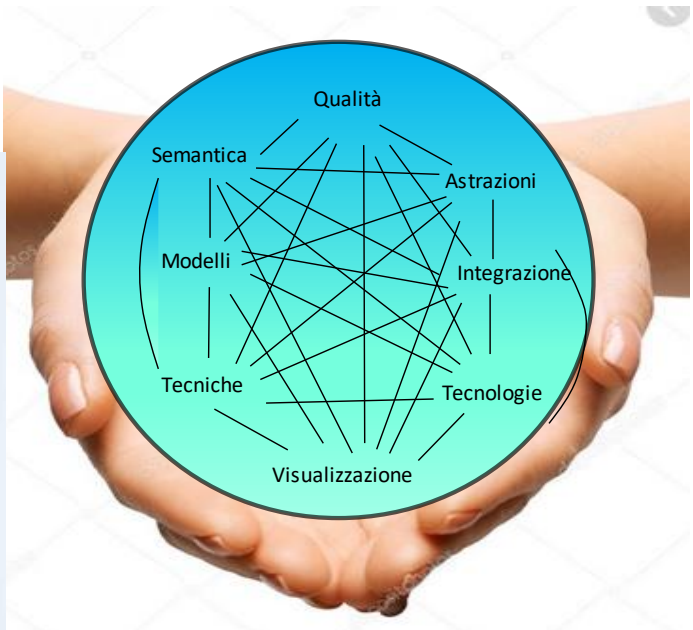
La Scienza dei Dati

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





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.

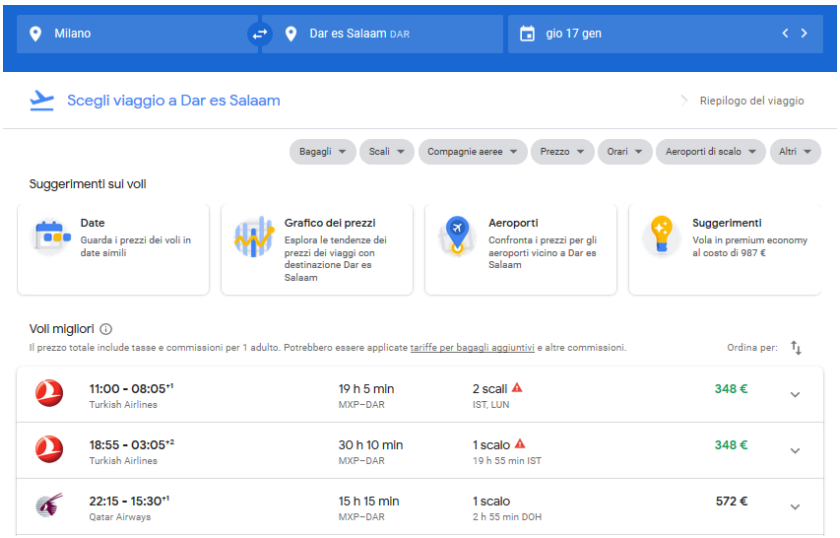


Figura 3 – Prenotare e acquistare un volo (dal sito eDreams)

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.

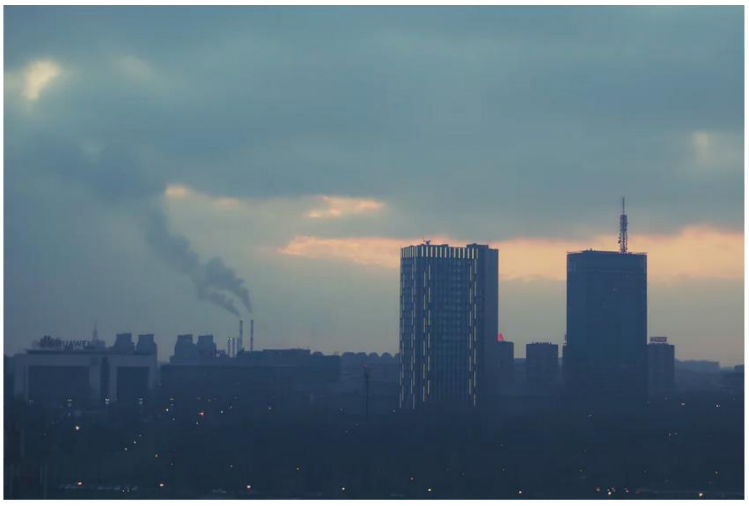


Figura 4 – Dove andare a correre senza farsi avvelenare dallo smog?

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.

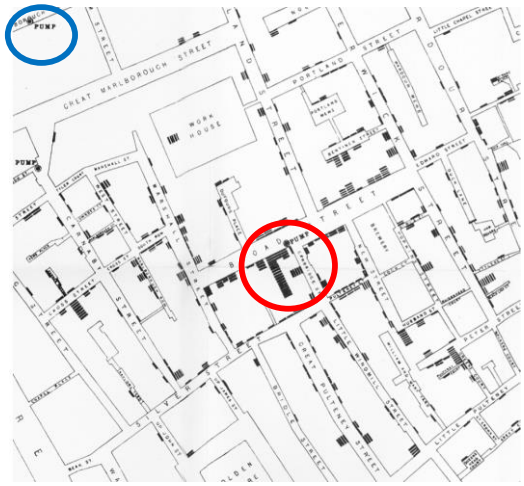


Figura 7 – Osservazioni sul colera a Londra

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.

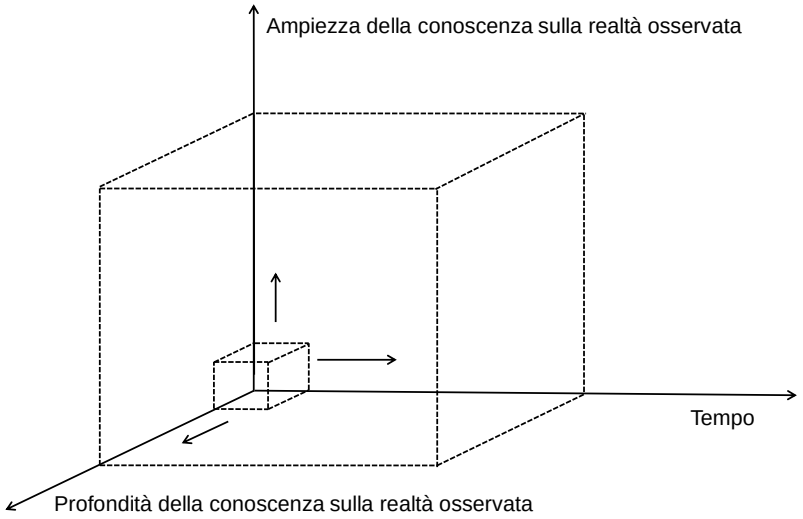


Figura 8 – Lo spazio dei big data

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.

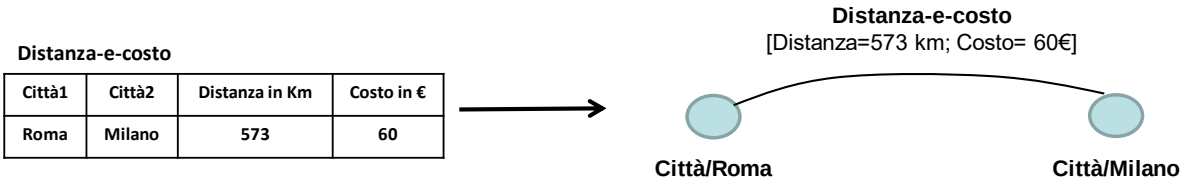


Figura 11 – Dati rappresentati mediante grafi

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.

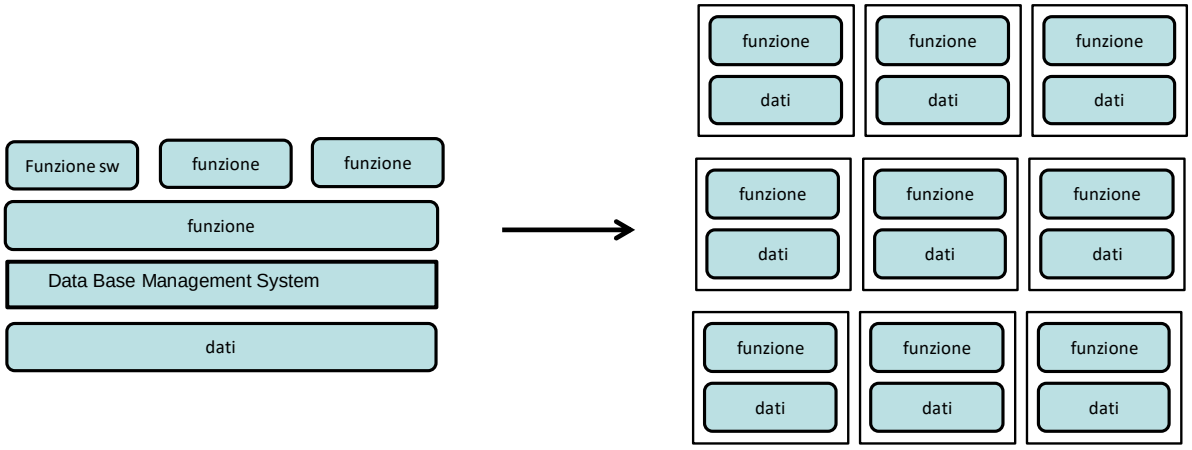


Figura 12 - Evoluzione delle architetture

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.

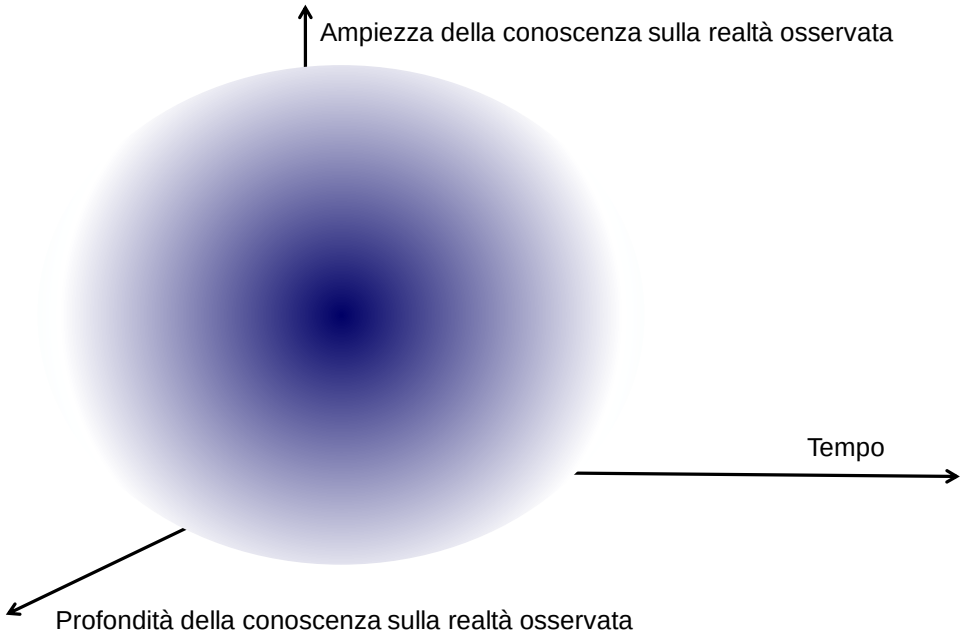


Figura 13 - La grande sfera opaca

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

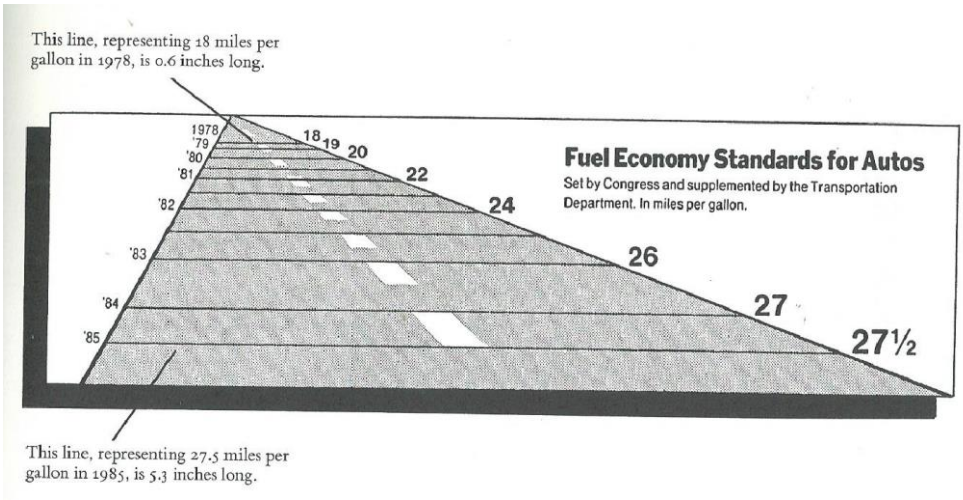


Figura 15 - Visualizzazione dei dati (tratta da E. Tufte – The Visual Display of Quantitative Information)

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.

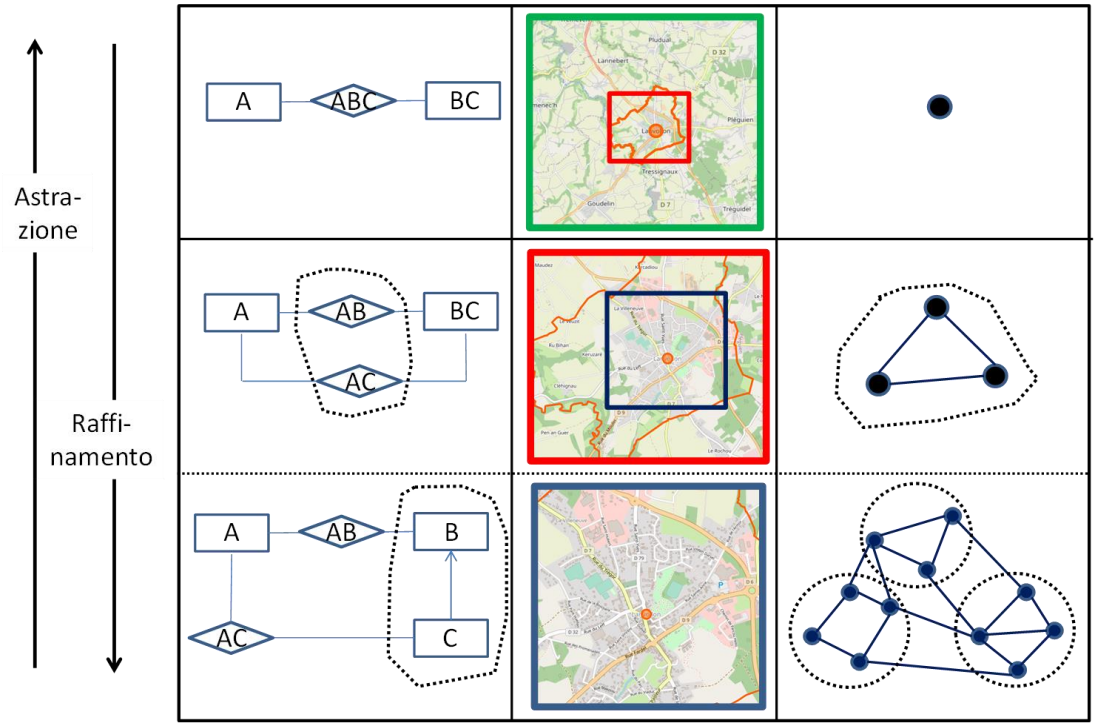


Figura 16 - Livelli di astrazione in un diagramma concettuale, una mappa, un grafo

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.

<u>Il valore dei dati</u>

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.

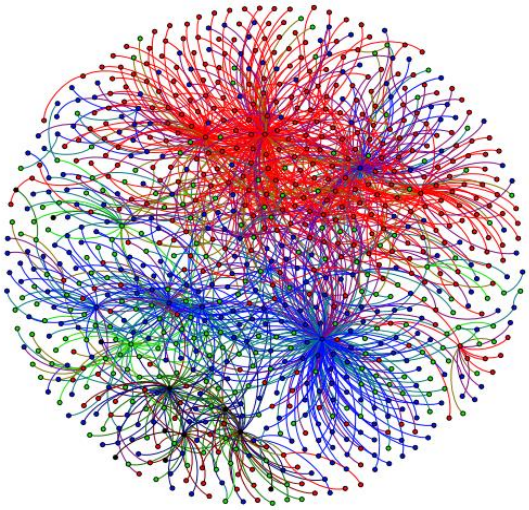


Figura 18 – I sentimenti nelle reti sociali

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.

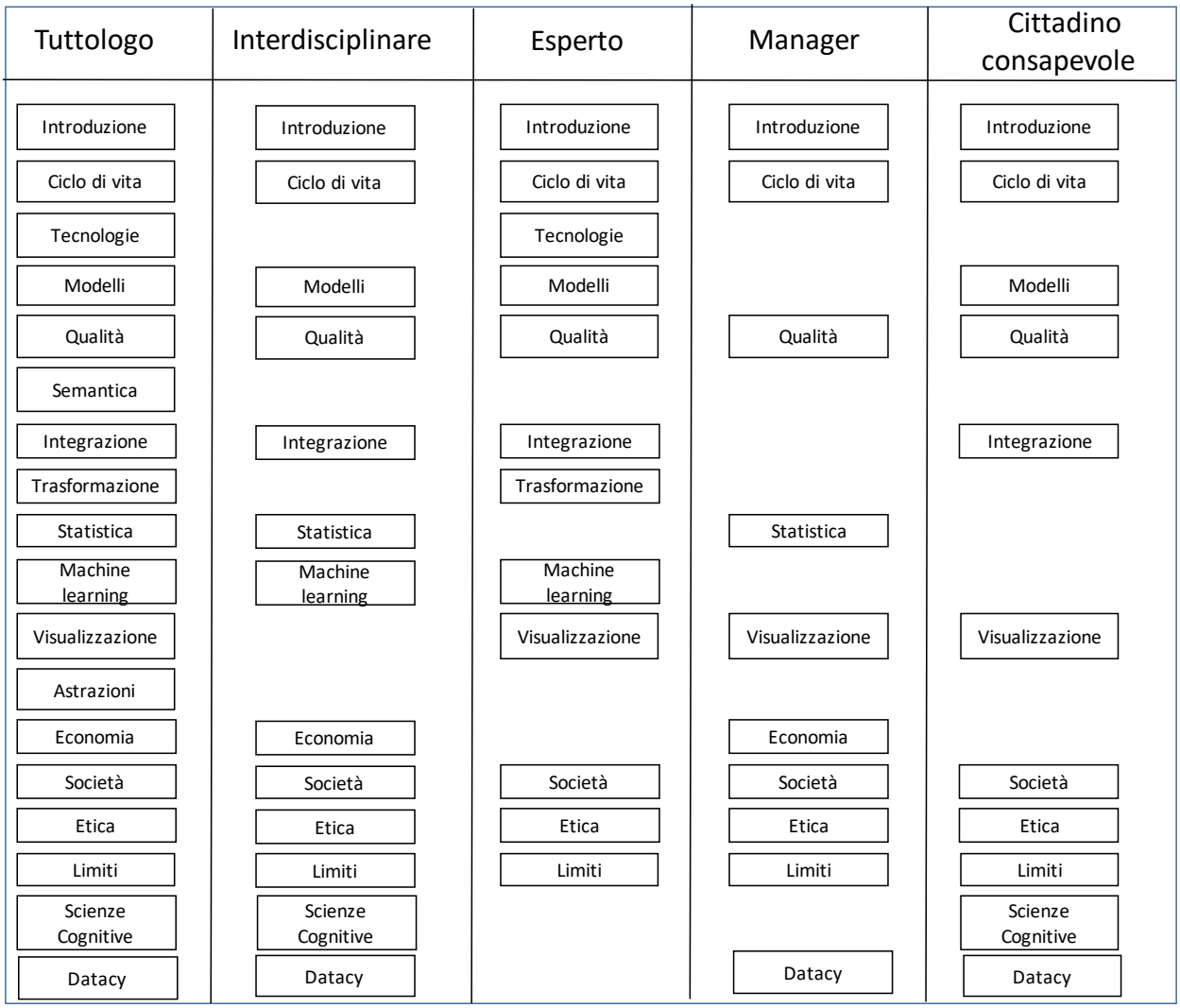


Figura 19 – Tipi di lettori e percorsi di lettura

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.