Enciclopedia dei dati digitali

Carlo Batini

Libro Primo

I dati sono una finestra sul mondo

Versione 1

2 febbraio 2021

Indice

Prologo

Ma insomma, cosa sono questi dati digitali?

I dati digitali sono una finestra sul mondo….

Adesso mi spiego meglio….

Capitolo 1

Dalla alfabetizzazione linguistica alla alfabetizzazione dei dati digitali

Nella seconda metà del secolo scorso vi fu in Italia una vasta iniziativa sociale, che aveva lo scopo di promuovere l' alfabetizzazione linguistica, cioè l'insegnamento a tutta la popolazione italiana del linguaggio italiano scritto e orale. Il processo di alfabetizzazione era iniziato fin dalla unificazione dell'Italia nel 1861, epoca in cui secondo Tullio de Mauro solo il 2.5% della popolazione si esprimeva con la lingua italiana.

Figura 1 – Il maestro Manzi

Favorita dall'avvento prima della radio e poi della televisione, l'opera di alfabetizzazione visse una esperienza fondamentale nella trasmissione televisiva della RAI *Non è mai troppo tardi*, messa in onda fra il 1960 e il 1968.

il maestro Alberto Manzi era il protagonista delle trasmissioni: usando semplici strumenti di comunicazione come il gesso, il pennarello, la lavagna di ardesia, pannelli di carta e disegni, insegnò dapprima a scrivere le lettere dell'alfabeto, e sviluppò successivamente un insieme di contenuti formativi. Questi contenuti riguardarono il lessico, cioè le parole del vocabolario, la *sintassi* delle frasi, cioè le regole da seguire nella loro costruzione, e la *semantica*, cioè il loro significato; vediamo una immagine da una puntata di *Non è mai troppo tardi* in Figura 1.

Successivamente, due figure tra le tante posero l'apprendimento della lingua italiana scritta e orale come base per una emancipazione delle classi sociali meno abbienti. La prima figura è Don Lorenzo Milani, che basò gran parte della sua opera pedagogica sull'insegnamento della lingua e in particolare sulla produzione di testi scritti; il momento più significativo di quella

esperienza fu il libro *Lettera a una Professoressa* del 1967, da cui riportiamo in Figura 2 alcune frasi rimaste celebri.

> La mamma è di quelle che si intimidiscono davanti a un modulo di telegramma. Il babbo osserva e ascolta, ma non parla.
>
> La timidezza dei poveri è un mistero più antico.
>
> Chiamo uomo chi è padrone della sua lingua.
>
> L'operaio conosce 100 parole, il padrone 1.000, per questo è lui il padrone.

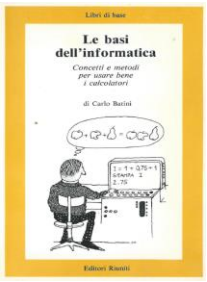
Figura 2 – Alcune frasi da *Lettera a una Professoressa*, di Don Lorenzo Milani

La seconda figura è Tullio De Mauro, professore di linguistica generale alla Università "La Sapienza" di Roma, Ministro della Pubblica Istruzione dal 2000 al 2001, e, soprattutto, grande saggista e autore di testi sulla linguistica e sulla lingua italiana. De Mauro diresse la collana dei *Libri di base*, composta da 139 volumi di divulgazione scientifica, pubblicati tra il 1979 e il 1989 dagli Editori Riuniti.

De Mauro, nel primo libro della collana, *Guida all'uso delle parole*, espose i risultati di una indagine svolta su un campione di cittadini che avevano portato a termine la scuola dell'obbligo, che allora terminava con la terza media. L'indagine condotta da De Mauro portò a individuare un elenco di circa 5.000 parole, tutte quelle che un diplomato di terza media era in grado a quei tempi di comprendere nel linguaggio scritto e usare nel linguaggio verbale.

Ho avuto il privilegio di scrivere il testo della collana Libri di base su *Le basi dell'informatica*, in virtù del rapporto professionale che mi legava a Mimmo Cioffi, a sua volta autore del testo *Che cos'è il calcolatore*. Il testo fu pubblicato nel 1984; una sua prima versione fu sottoposta a una severa correzione di bozze, che esemplifico in Figura 3, mostrando la prima pagina del libro, insieme a una bella foto di De Mauro che compare su Wikipedia e la copertina del libro.

Nel contratto stipulato, mi impegnavo ad usare solo le 5.000 parole dell'elenco stilato da De Mauro; se volevo introdurre una nuova parola, ad esempio *algoritmo*, dovevo spiegarla con le parole contenute nell'elenco delle 5.000. La parola *quotidianamente*, quarta parola del primo periodo della prima pagina, fu cancellata con una doppia linea e mi fu chiesto di sostituirla con parole nel vocabolario base. Io scelsi come sostituzione: *quasi ogni giorno* (vedi sempre Figura 3). Le 164 pagine del testo furono trattate allo stesso modo, e ancora ricordo il cimitero di cancellazioni che osservai con sconcerto quando sfogliai il pacco di carta che mi arrivò con la revisione delle bozze…..



Come si cerca un numero di telefono – Come si consulta un orario – Come si fa un orario – Come si calcola (in alcuni casi) quanto dobbiamo aspettare un autobus – Conclusioni

Ciascuno di noi ~~quotidianamente~~ **quasi ogni giorno** ha la necessità di usare e produrre un grande numero di informazioni. Informazione è quella che acquisiamo dall'orologio quando leggiamo che ora è, quella che forniamo a un automobilista straniero che ci chiede come fare per andare a San Pietro, quella che scopriamo sulla faccia del medico quando gli chiediamo come sono andate le analisi.

Figura 3 – Userai solo 5.000 parole!

L'alfabetizzazione linguistica del secondo novecento è stato un processo straordinario, e fondamentale in un momento storico in cui la diffusione della cultura avveniva attraverso la lingua scritta e orale. *A me sembra che ai giorni nostri si debba cominciare a pensare, con tanta umiltà di pensiero ma anche una certa urgenza, a un'altra alfabetizzazione, quella dei dati digitali.*

Da tanto tempo, almeno quarantacinque anni, il mio lavoro di ricerca, didattico e di divulgazione è centrato sui dati digitali. Tuttavia, è stato agli inizi del mese di novembre del 2020 che ho iniziato a pensare convintamente che il processo di alfabetizzazione dei dati fosse oramai necessario e urgente.

Nella prima settimana di novembre del 2020 vi furono due eventi concomitanti. Il 3 novembre si sono svolte le elezioni presidenziali negli Stati Uniti; in quella stessa settimana fu emesso un Decreto del Presidente del Consiglio dei Ministri, in cui si stabilivano le modalità con cui le Regioni italiane venivano suddivise in tre diverse fasce di rischio epidemico, associate informalmente ai tre colori giallo, arancione e rosso, vedi Figura 4.

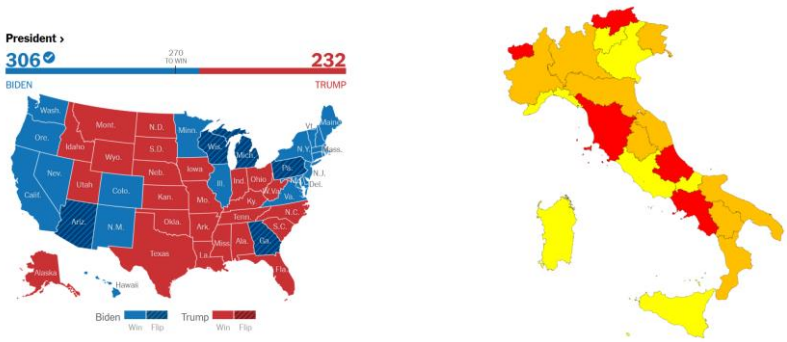


Figura 4 – Una cartina dal New York Times con i risultati delle elezioni americane a un certo punto dello spoglio, e una cartina dell'Italia dal Corriere della Sera con le regioni gialle, rosse e arancione

Gli avvenimenti di quei giorni, legati all'esito delle elezioni e alle restrizioni previste nel decreto e alla loro durata, suscitarono grande impatto emotivo e grande incertezza sul futuro. Negli Stati Uniti terminava il quadriennio della presidenza Trump, e mai come allora, anche in virtù delle politiche divisive di Trump, la società americana si presentava polarizzata ed estremizzata. L'Italia, uscita esausta dalla prima ondata, iniziava a essere consapevole che la pandemia si stava diffondendo per la seconda volta in maniera molto seria tra la popolazione.

Le cartine come quelle di Figura 4 servirono nel corso dello spoglio dei voti negli Stati Uniti e nella evoluzione della pandemia in Italia per far comprendere a tutti in maniera chiara e sintetica l'evoluzione delle due vicende: nel caso delle elezioni americane, mostrando attraverso l'uso dei colori rosso (repubblicano) e blu (democratico) quali stati potessero considerarsi aggiudicati all'uno e all'altro candidato o fossero ancora in bilico (nella figura questi stati sono segnati con righe diagonali); nel caso della suddivisione delle regioni in aree di rischio, in quale area fosse collocata la propria regione di residenza abituale e quali norme e restrizioni venivano applicate.

Le due cartine erano la sintesi finale di un complesso processo di raccolta ed elaborazione di dati: nel caso delle elezioni americane, i dati esprimevano il voto degli elettori, aggregati per contea e stato, nel caso delle regioni italiane, l'attribuzione dei colori alle regioni era il risultato di una complessa procedura di raccolta di dati da parte dagli enti territoriali (ad es. gli ospedali), e di calcolo dei colori basato sui valori assunti da 21 indicatori e dall'indice Rt di trasmissione del contagio.

La raccolta dei dati sulle espressione di voto negli Stati Uniti

Nel caso degli Stati Uniti, i dati raccolti erano tutti dello stesso tipo, nel senso che riguardavano le preferenze di ogni singolo elettore americano tra Biden, Trump o un candidato indipendente; per capire chi aveva vinto le elezioni in ogni stato, conquistando automaticamente tutti i grandi elettori della votazione finale, si doveva compiere una operazione relativamente semplice di conteggio (ma vedremo tra poco che anche le operazioni di conteggio, ed in particolare delle schede per le espressioni di voto può presentare seri problemi..).

Dai diversi stati i dati cominciarono ad arrivare con il contagocce; il loro spoglio e la prevalenza di uno o dell'altro candidato erano influenzati anche dal fatto che per la prima volta, a seguito della epidemia Covid, era molto alto il numero di voti inviati per posta.

Siccome i voti per posta nelle precedenti elezioni erano stati in prevalenza a favore dei democratici, a seconda dell'ordine con cui i voti venivano scrutinati tra quelli in presenza e quelli arrivati per posta, si verificarono casi di alternanza su chi fosse in testa nei voti scrutinati; un esempio è quello dell'andamento dei voti scrutinati in Pennsylvania, vedi le due linee continue raffigurate in Figura 5.

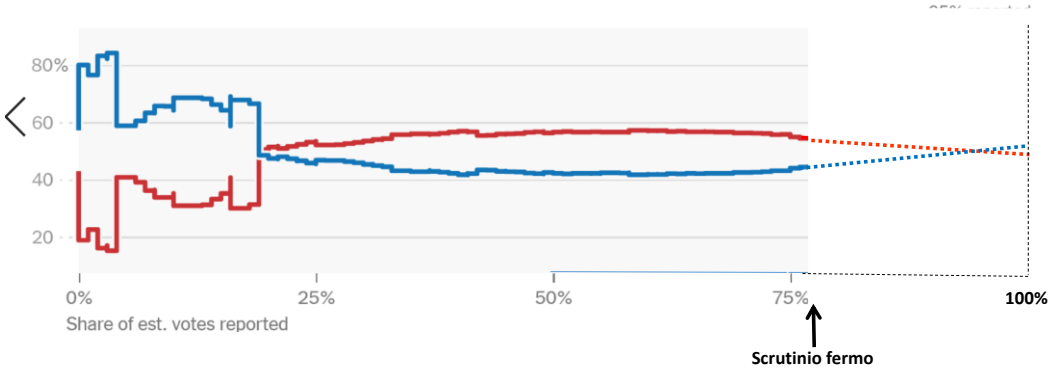


Figura 5 – Andamento dei voti ottenuti da Biden e Trump in Pennsylvania per il primo 80% dei voti scrutinati, in blu Biden, in rosso Trump.

La Figura 5 mette in evidenza il fatto che a un certo punto lo scrutinio si fermò, probabilmente perchè iniziarono ad essere esaminati i ricorsi di Trump; per lunghe ore guardai quella curva, che non andava avanti. Non sapevo cosa pensare, e i giornali on line non fornivano molte informazioni su cosa stava succedendo.

Ho pensato allora di fare una semplice operazione grafica, consistente nel prolungare le curve con linee rette fino alla posizione corrispondente al 100% dei voti scrutinati, per vedere, in modo approssimato, come sarebbe andata finire. E ho visto che con ogni probabilità Biden avrebbe vinto in quello Stato.

La stessa operazione di approssimazione devono aver fatto i media americani più importanti, che, per tradizione, proclamano il Presidente eletto, assegnando uno Stato a uno dei due contendenti quando il vantaggio è ormai incolmabile, cioè la percentuale dei voti da scrutinare diventa inferiore al vantaggio che il candidato in testa ha in quello stato.

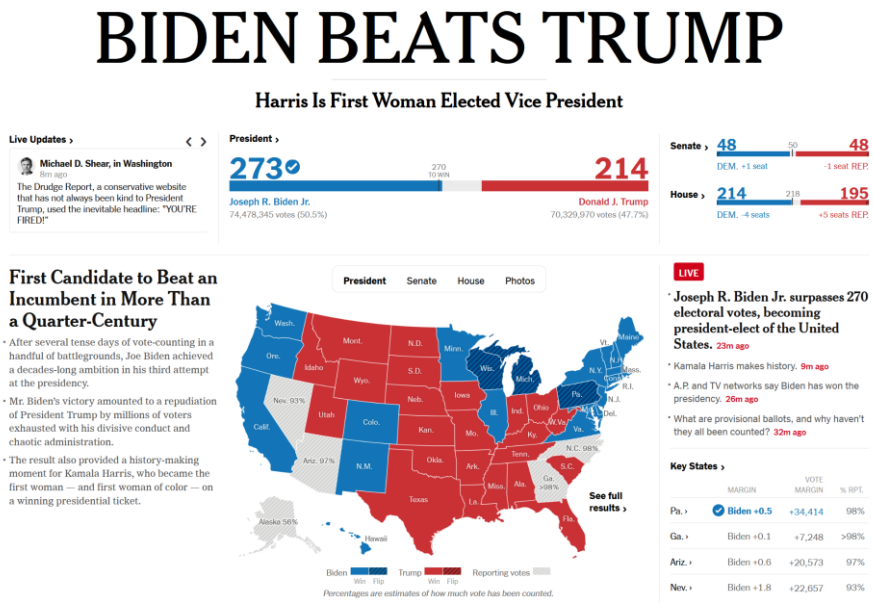


Figura 6 – La proclamazione del Presidente eletto

Siccome gli scrutini erano fermi in diversi stati, si rischiava una situazione di instabilità sociale e politica pericolosa, e quindi i media, sulla base di operazioni di approssimazione simili a quella fatta da me, decisero come sarebbe andata a finire, vedi Figura 6: Biden batte Trump!

Il calcolo del colore delle Regioni e le statistiche pubblicate sui giornali

In Italia il Ministero della Salute e l'Istituto Superiore di Sanità pubblicarono documenti, tabelle e "domande frequenti" sul complesso processo di raccolta di dati ed elaborazione di indicatori che permettevano di calcolare il colore da attribuire alle Regioni.

I dati coinvolti nel *calcolo del colore* erano costituiti da 21 indicatori che rappresentavano fenomeni diversi tra loro, come la diffusione della malattia e l'impatto dei malati bisognosi di cure specialistiche sulle strutture sanitarie. In particolare assumeva rilevante importanza per il calcolo del colore delle regioni l'indice Rt, cioè il numero medio di infezioni trasmesse da ogni individuo infetto, calcolato nel corso della evoluzione della epidemia.

Cari lettori, da questo momento cercherò di rendere il testo più interattivo di quanto accade usualmente nei testi divulgativi, rivolgendomi alcune volte o proponendo quesiti alla comunità dei lettori usando il ***voi.***

Altre volte dialogherò con un interlocutore immaginario, che mi pone quesiti o formula osservazioni, insomma un lettore che svolge un ruolo attivo nell'interagire con me. Questo lettore talvolta mi obbliga a essere più chiaro, ovvero a scoprire nuovi aspetti sui temi affrontati. A questo interlocutore darò del ***tu***.

Altri dati che furono pubblicati da molti siti e giornali sulla epidemia Covid in Italia riguardavano sintesi statistiche, rappresentate mediante vari tipi di visualizzazioni. Per andare con ordine, ci concentriamo prima sulle statistiche e successivamente sull'indice Rt.

Guardate dunque le statistiche di Figura 7, comparse sul quotidiano Corriere della Sera nelle due edizioni del 21/12/2020 e del 16/1/2021. Le cornici di colore verde, rosso, azzurro, blu contengono vari tipi di dati, che chiameremo anche con il termine *valori, valori numerici o numeri*. Nelle due cornici verdi vi sono *valori assoluti*, cioè valori che nascono da un conteggio su una popolazione; il conteggio può riguardare i casi totali, le persone attualmente positive, i guariti, i deceduti.

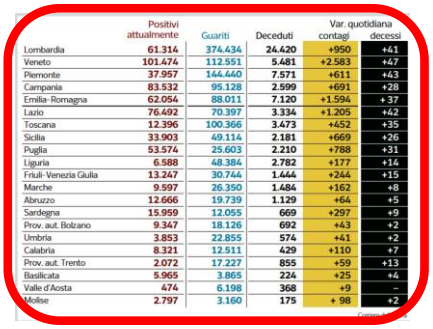
Questi valori assoluti conteggiano i diversi casi *a partire dall'inizio del contagio*, e sono forniti in forma aggregata per l'intera Italia *nelle cornici verdi*, e per le singole regioni e province autonome di Bolzano e Trento *nelle cornici rosse*.

Altri valori assoluti sono di natura diversa, nel senso che, come gli altri, *contano* persone, ma questa volta i conteggi riguardano solo la giornata precedente, sono dunque *variazioni*. Ad esempio, la variazione dei contagiati per regione contribuisce al calcolo di uno dei 21 indicatori di cui abbiamo parlato in precedenza.

Nelle *cornici azzurre* compaiono *grafici* e mappe, che forniscono una rappresentazione visuale che intende essere più immediata e espressiva dei numeri.

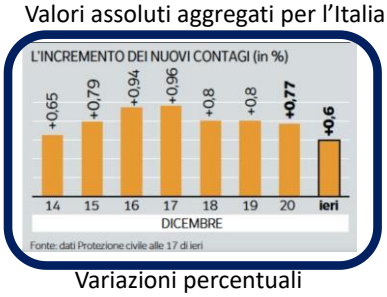
Nella edizione del 21/12/2020 il grafico delle due funzioni nella cornice azzurra riguarda il conteggio dei decessi nella prima e seconda ondata della pandemia, che viene rappresentato per entrambe le funzioni a partire dall'inizio della rispettiva ondata. Questi dati sono *comparativi*, nel senso che confrontano due fenomeni; e nel confrontarli ci danno tante informazioni e ci permettono di fare previsioni come ad esempio il fatto che la seconda ondata fosse destinata purtroppo a fare più morti, perché non poteva accadere che la curva si appiattisse improvvisamente; e in effetti ciò accadde, i morti della seconda ondata superarono quelli della prima.

Valori assoluti aggregati per l'Italia

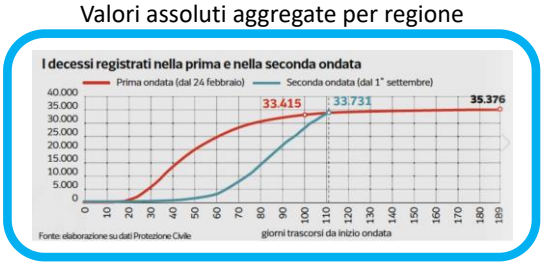


	Positivi attualmente	Guariti	Deceduti	Var. quotidiana contagi	Var. quotidiana decessi
Lombardia	61.314	374.434	24.420	+950	+41
Veneto	101.474	112.551	5.481	+2.583	+47
Piemonte	37.957	144.440	7.571	+611	+43
Campania	83.532	95.128	2.599	+691	+28
Emilia-Romagna	62.054	88.011	7.120	+1.594	+37
Lazio	76.492	70.397	3.334	+1.205	+42
Toscana	12.396	100.366	3.473	+452	+35
Sicilia	33.903	49.114	2.181	+669	+26
Puglia	53.574	25.603	2.210	+788	+31
Liguria	6.588	48.384	2.782	+177	+14
Friuli-Venezia Giulia	13.247	30.744	1.444	+244	+15
Marche	9.597	26.350	1.484	+162	+8
Abruzzo	12.666	19.739	1.129	+64	+5
Sardegna	15.959	12.055	669	+297	+9
Prov. aut. Bolzano	9.347	18.126	692	+43	+2
Umbria	3.853	22.855	574	+41	+2
Calabria	8.321	12.511	429	+110	+7
Prov. aut. Trento	2.072	17.227	855	+59	+13
Basilicata	5.965	3.865	224	+25	+4
Valle d'Aosta	474	6.198	368	+9	-
Molise	2.797	3.160	175	+ 98	+2

Valori assoluti aggregate per regione

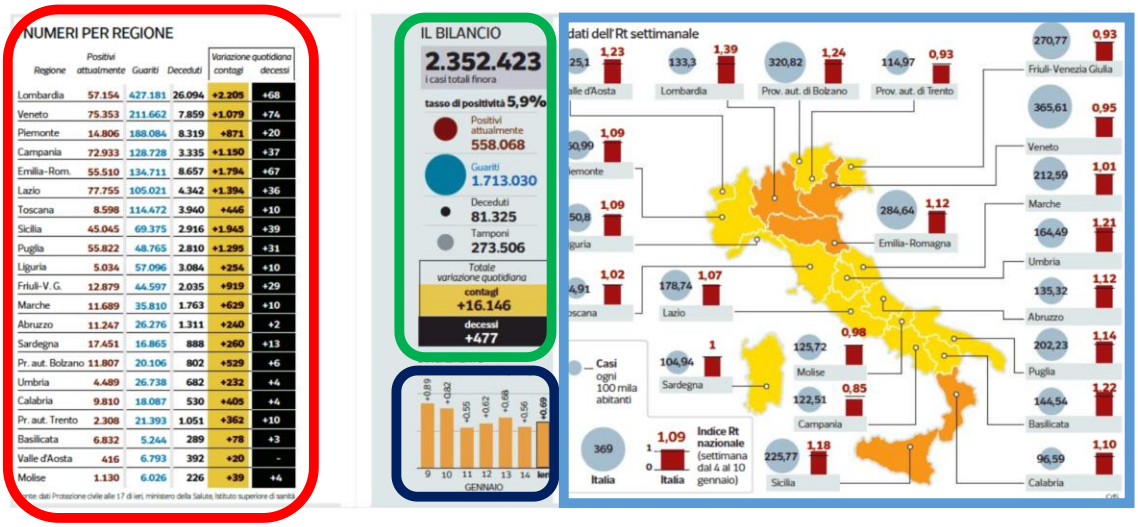


Variazioni percentuali



Confronto tra prima e seconda ondata

Dal numero del Corriere della sera del 21/12/2020



Dal numero del Corriere della sera del 16/1/2021

Figura 7 – Statistiche pubblicate sui giornali

Il grafico con cornice azzurra della edizione del 16/1/2011 mostra una cartina dell'Italia in cui sono evidenziati regione per regione l'indice Rt (che riprendiamo tra poco) e il numero di casi

giornalieri per 100.000 abitanti. Oltre all'indice Rt, anche i casi per 100.000 abitanti hanno assunto nel corso della evoluzione della pandemia un ruolo rilevante; possiamo dunque dire, confrontando comparativamente le due cartine che la seconda, relativa al 16/1/2011 è decisamente più informativa della prima, nel senso che mostra una maggiore quantità di dati rilevanti per noi lettori.

Nei due *riquadri blu scuro* troviamo altri grafici molto interessanti, seppur per un motivo completamente diverso. Vediamo infatti nei due riquadri un diagramma a barre, che fornisce i valori *percentuali* dei nuovi contagiati in diversi giorni, rispettivamente, dei mesi di dicembre 2020 e gennaio 2021.

Ora, una percentuale è un rapporto tra un numeratore che descrive un *numero assoluto*, nel nostro caso le nuove persone contagiate, e un valore a denominatore che rappresenta l'insieme delle persone in un *universo di riferimento*. Nella figura non viene detto quale sia questo universo di riferimento, ma si raggiunge facilmente la conclusione che sia *il totale dei contagiati a partire dall'inizio della epidemia fino al giorno in cui sono misurati i dati*. Quindi il valore è calcolato come:

(a) Nuovi contagiati nella giornata di ieri / Totale dei contagiati dall'inizio dell'epidemia

Se ci pensiamo un attimo, questo dato è veramente strano, insomma, *poco significativo*: che senso ha misurare una variazione che ha a denominatore un valore che cresce continuamente, perché ogni giorno ci sono nuovi contagiati che si aggiungono a quelli del periodo precedente? La scarsa significatività dell'indicatore è ancor più evidente se guardiamo la Figura 8, in cui vediamo due serie di numeri di persone positive al Covid, in cui, nella prima, ad un incremento costante dei positivi (pari a 20) corrisponde una diminuzione del valore dell'indicatore (a), e, nella seconda, la diminuzione della percentuale avviene anche nel caso di incremento crescente (20, 21, 22) dei positivi.

totale positivi	nuovi	percentuale
200		
220	20	0.100
240	20	0.091
260	20	0.083

Aumenti uguali percentuali decrescenti

totale positivi	nuovi	percentuale
200		
220	20	0.100
241	21	0.095
263	22	0.091

Aumenti crescenti percentuali decrescenti

Figura 8 - Due serie di valori che evidenziano la scarsa utilità dell'indicatore (a)

Se ci pensiamo un attimo, ha più senso usare una percentuale in cui a denominatore ci sono, ad esempio, i contagiati del giorno precedente, così che abbiamo un valore superiore a uno se i contagiati sono aumentati, e inferiore a uno se sono diminuiti:

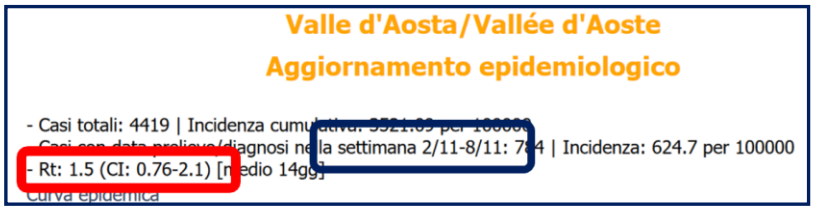
(b) Nuovi contagiati nella giornata di ieri / Nuovi contagiati nella giornata dell'altro ieri

Abbiamo visto che anche semplici tabelle o semplici grafici possono differire di molto quanto a significatività dei valori numerici che esprimono.

L'indice Rt

Riguardo all'indice Rt, sul sito del Ministero della Salute (http://www.salute.gov.it/portale/nuovocoronavirus/ verificato nel gennaio 2021) sono comparse informative molto chiare sulla formula di calcolo, che qui non approfondiamo. Vediamo invece tre estratti dai rapporti di monitoraggio che ogni settimana venivano prodotti dal Ministero.

Nell grafico di Figura 7, riferito alla settimana precedente il 16/1/2021, è riportato per la Val d'Aosta un valore di Rt pari a 1.23. Se ora osserviamo i dati nelle cornici rosse della Figura 9, che fanno riferimento ai tre periodi evidenziati nelle cornici blu, notiamo che, accanto al valore Rt, compare un altro dato, chiamato CI.



Regione.PA	Incidenza per 100.000 ab 14gg	Nuovi casi segnalati nella settimana	Trend settimanale COVID-19 Casi (Fonte ISS)	Trend settimanale COVID-19 Focolai	Stima di Rt-puntuale	Dichiarata trasmissione non gestibile in modo efficace con misure locali (zone rosse)	Valutazione della probabilità	Valutazione di impatto	Allerte relative alla resilienza dei servizi sanitari territoriali	Compatibilità Rt sintomi puntuale con gli scenari di trasmissione**	Classificazione complessiva di rischio	Classificazione Alta e/o equiparata ad Alta per 3 o più settimane consecutive
Molise	494.60	775	↑	↑	1.17 (CI: 0.88-1.5)	No	**Moderata**	Bassa	0 allerte segnalate	1	**Moderata con probabilità alta di progressione a rischio Alto**	**No**
Piemonte	1115.73	21401	↓	↓	0.89 (CI: 0.88-0.9)	No	Bassa	**Alta**	0 allerte segnalate. Ind. 2.2 non costituisce allerta in quanto 2.3 risulta sotto soglia	1	**Moderata con probabilità alta di progressione a rischio Alto**	**No**
PA Bolzano/Bozen	1423.85	3461	↓	↓	1 (CI: 0.96-1.04)	No	Bassa	**Alta**	2 allerte segnalate. Ind 2.1 in aumento. Ind 2.2 e 2.3 sopra-soglia	1	**Alta**	**Si**
PA Trento	537.09	1293	↓	↑	0.81 (CI: 0.75-0.88)	No	Bassa	**Alta**	0 allerte segnalate	1	**Moderata con probabilità alta di progressione a rischio Alto**	**No**
Puglia	443.85	8745	↓	↓	0.99 (CI: 0.96-1.02)	No	Bassa	**Alta**	2 allerte segnalate. Ind 2.1 in aumento. Ind 2.2 e 2.3 sopra-soglia	1	**Alta**	**Si**
Sardegna	157.93	1064	↓	↑	0.71 (CI: 0.65-0.77)	No	**Moderata**	**Alta**	2 allerte segnalate. Ind 2.1 in aumento. Ind 2.6 sotto 75%. Ind. 2.2 non costituisce allerta in quanto 2.3 risulta sotto soglia	1	**Alta**	**Si**
Sicilia	370.16	7559	↓	↑	1.04 (CI: 1.01-1.07)	No	**Moderata**	Bassa	1 allerta segnalata. Ind 2.6 in diminuzione (già segnalato la settimana precedente)	2	**Moderata**	**No**
Toscana	716.62	11156	↓	↑	1.2 (CI: 1.17-1.22)	No	**Moderata**	**Alta**	0 allerte segnalate	2	**Alta**	**Si**
Umbria	718.63	2571	↓	↑	0.74 (CI: 0.71-0.77)	No	Bassa	**Alta**	0 allerte segnalate	1	**Moderata**	**No**
V.d'Aosta/V.d'Aoste	1322.70	651	↓	↓	0.99 (CI: 0.92-1.07)	No	Bassa	**Moderata**	0 allerte segnalate. Ind 2.6 sotto 90% però in aumento	1	**Moderata con probabilità alta di progressione a rischio Alto**	**No**
Veneto	848.10	20743	↓	↑	1.2 (CI: 1.17-1.22)	No	**Moderata**	Bassa	1 allerta segnalata. Ind 2.1 in aumento	2	**Moderata con probabilità alta di progressione a rischio Alto**	**No**

Rt periodo 4 – 17 novembre Valle d'Aosta

Figura 9 – Il valor medio di Rt e la varianza in Val d'Aosta in tre periodi temporali riferiti al mese di novembre 2020

Cosa rappresenta CI? Il dato CI fornisce *l'intervallo di confidenza*, cioè l'intervallo di valori in cui si collocano la gran parte dei valori calcolati di Rt attorno al valor medio; non bisogna infatti dimenticare che Rt è una *stima* dotata di una sua *intrinseca variabilità*, espressa appunto dall'intervallo di confidenza. Ebbene, questo intervallo è molto ampio all'inizio del

periodo di osservazione, e un intervallo di confidenza alto significa che il valor medio perde di significatività. Insomma, diventa un pòcome tirare ai dadi…..Ci troviamo qui di fronte a un dato che è noto solo in modo approssimato, espresso dalla ampiezza dell'intervallo di confidenza.

I dati che vediamo sono la punta di un iceberg

Gli esempi precedenti e la descrizione che ho fatto sul processo di raccolta e elaborazione dei dati utilizzati nelle elezioni americane e nella scelta del colore delle regioni, ci lanciano *un messaggio importante*: i dati che vediamo nei siti Web, i dati che sono citati nelle reti sociali, sui giornali, nei dibattiti televisivi, e che ambiscono a descrivere fenomeni per noi rilevanti, sono spesso soltanto la punta di un iceberg, vedi Figura 10.

Noi non vediamo mai la parte sommersa dell'iceberg, quella parte del processo di raccolta e elaborazione dei dati che porta alla costruzione del dato finale (chi alla fine abbia vinto tra Biden e Trump, o che colore abbia una regione).

Figura 10 – I dati che vediamo sono la punta di un iceberg
(da http://www.cim-fema.it/web/blog/)

Tutto ciò che in quegli eventi veniva chiamato con il termine generico *dati,* rappresentava in realtà una miriade di fenomeni diversi, e lo rappresentava nello stesso modo che lega la punta emergente, la parte che noi percepiamo (nel caso della epidemia Covid, la cartina dell'Italia con i colori delle regioni) alla parte sommersa dell'iceberg (i dati raccolti negli ospedali e nei centri che facevano tamponi e individuavano le cause dei decessi)[1]. Nel corso di questa serie

[1] Se avrete pazienza, discuteremo più a fondo questo problema nel Capitolo 8, I dati che non abbiamo, con riferimento a un caso che nel gennaio 2021 occupò tutti i giornali, la grande polemica tra Regione Lombardia e Istituto Superiore di Sanità sulla attribuzione del colore rosso alla Lombardia.

di fasi, i dati vivono una vita complessa, fatta di approssimazioni, di inciampi, di trasformazioni, di errate interpretazioni, fino ad arrivare a noi in forma di numero o di grafico.

Da quei giorni del novembre 2020, la parola *dati*, fino ad allora confinata prevalentemente in un linguaggio tecnico, fu pronunciata una quantità innumerevole di volte nei telegiornali, scritta nei giornali on line, diffusa nelle reti sociali, usata nelle trasmissioni di approfondimento.

E gli aggettivi con cui veniva coniugata furono spesso (ma non sempre…) polemici e svalutativi: dati inaccurati, dati vecchi, dati incompleti, dati inaffidabili, dati sporchi, dati mancanti, vedi alcuni esempi in Figura 11. Talvolta non si usava la parola dati ma, piuttosto, la parola *numeri*, cioè dati appartenenti a un dominio matematico. Talvolta si usava una parola ancora diversa, la parola *fatti*.

Nella parte sinistra di Figura 11 compare un tweet di Rudolph Giuliani, per diverso tempo il principale avvocato a cui Donald Trump si rivolse per la sua titanica battaglia volta a dimostrare che nel voto erano state commesse frodi: Giuliani dice che *in diverse contee e stati, ad esempio il Michigan, c'erano stati più voti che elettori, in rapporti che arrivavano al 300%!* Il governatore della Campania De Luca mise in dubbio che per l'indicatore delle terapie intensive alcune Regioni inviassero dati corretti. Ma ci furono anche affermazioni positive: "Numbers don't lie, i numeri non dicono bugie", a indicare la sensazione di una forza intrinseca dei dati rispetto a tutte le manipolazioni che venivano effettuate su di essi.

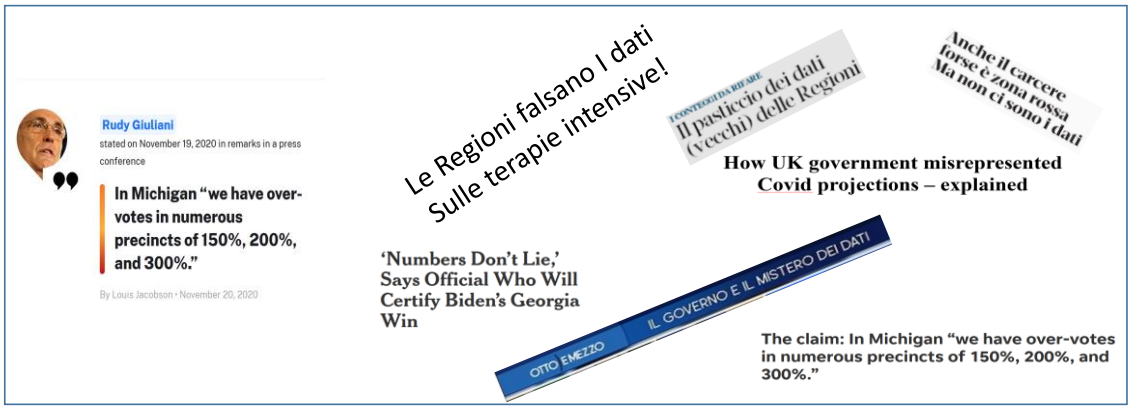


Figura 11 - Le polemiche di quei giorni

Occupandomi dei dati da tanto tempo, ho cercato di capire perché improvvisamente i dati fossero diventati così importanti. Ho cercato di capire se avesse un senso compiuto usare questa parola, i *dati*, in maniera così generica, o, piuttosto, dovessi provare a indagare sui tanti aspetti della nostra vita che sono influenzati, condizionati, migliorati, peggiorati, modificati, trasformati dalla grande disponibilità di dati digitali.

Provare a indagare su quanto dobbiamo ai dati riguardo al miglioramento della qualità della nostra vita, e quanto è necessario stare attenti e esercitare consapevolezza critica, ogni volta che un dato ci apre una finestra sul mondo, ci rappresenta il mondo.

Ho cercato di capire quale fosse *l'alfabeto dei dati*, quale fosse *il linguaggio* attraverso cui essi ci parlano della realtà, quali *strumenti* potessero essere condivisi con i miei lettori per diventare tutti persone consapevoli.

Ed è venuto fuori il viaggio attraverso i dati che ci apprestiamo a compiere. Viaggio che inizio con questo libro, e che, per essere concluso, richiederà uno sforzo ben maggiore, richiederà di scrivere una Enciclopedia….[2]

Riassumendo

Il secondo novecento è stata l'epoca della ***alfabetizzazione linguistica***. Oggi, in virtù della grande importanza che i dati stanno assumendo nella nostra vita, nasce la esigenza di una **alfabetizzazione in tema di dati digitali**, fornendo a tutti gli strumenti per acquisire una consapevolezza critica sul loro uso. I dati vengono prodotti e mostrati a noi attraverso un itinerario spesso complesso, simile a un **iceberg** di cui noi vediamo la parte emergente.

[2] Questa mia affermazione sulla Enciclopedia non prendetela come una *minaccia*, questo libro può essere letto tranquillamente da solo….

Capitolo 2

Breve storia del passaggio dalla carta ai dati digitali
Come rappresentare il mondo con zero e uno

Comincio a interagire con te con unma curiosità: perché, pur parlando tanto di dati, fino ad ora non ne hai fornito una definizione? In una Enciclopedia non si dovrebbe usare un concetto senza definirlo.

Certamente. Possiamo inizialmente definire un dato come una discriminazione tra stati fisici delle cose, ad esempio il colore di una maglietta è "rosso", oppure "bianco", sono nato nel 1949, oppure sei nato nel 1990, le parole del testo che corrispondono a dati sono sottolineate.

La definizione di dato che ho appena proposto si applica a un vastissimo insieme di fenomeni. Per concentrarci su un fenomeno specifico, osserviamo la storia dei censimenti della popolazione degli Stati Uniti.

Questa storia è raccontata in un meraviglioso sito accessibile all'indirizzo https://www.census.gov/history/ (verificato nel gennaio 2021). In Figura 12 vediamo tre diverse modalità con cui dall'anno del primo censimento della popolazione, il 1790, sono stati rappresentati i dati raccolti nell'immenso territorio degli Stati Uniti.

I primi censimenti utilizzavano grandi fogli cartacei riempiti a mano, come quello a sinistra in Figura 12; questo modo di rappresentare i dati dilatava molto i tempi di elaborazione delle statistiche che venivano effettuate con il censimento (ad esempio quante donne appartengono alle diverse fasce di età); alla fine dell'800 erano richiesti circa 8 anni per completare queste statistiche; dunque, a quell'epoca i dati diventavano vecchi, non aggiornati, ben prima di diventare disponibili nelle statistiche.

Il grande cambiamento è iniziato da quando, per il censimento del 1890 e successivi fu adottata la scheda Hollerith, vedi ancora Figura 12, in cui i dati erano rappresentati per mezzo di codifiche basate su piccoli buchi rettangolari nella scheda; in questo modo la scheda poteva essere letta e i dati rappresentati potevano essere elaborati dalle macchine elettromeccaniche progenitrici dei moderni computer.

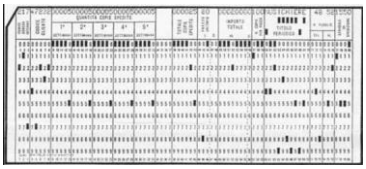
Con l'avvento dei primi computer[3], dopo la seconda guerra mondiale, per memorizzare i dati del censimento si usarono vari tipi di tecnologie di memorizzazione digitale, tra cui in Figura

[3] Uso qui la dizione inglese *computer*, essendo la dizione italiana *calcolatore* ormai in disuso.

12 mostriamo i *CD Rom*, una tecnologia con lettura a laser ottico che ebbe grande successo negli ultimi due decenni del 900.

Le tecnologie di memorizzazione di cui il CD Rom è un esempio adottano per i dati una codifica binaria, convenzionalmente indicata con i numeri “0” e “1”. Questa codifica è utilizzata nei moderni computer per rappresentare tutte le possibili tipologie di dati, dai numeri ai testi, le immagini, le registrazioni audio, i video.

Moduli cartacei utilizzati nei primi censimenti

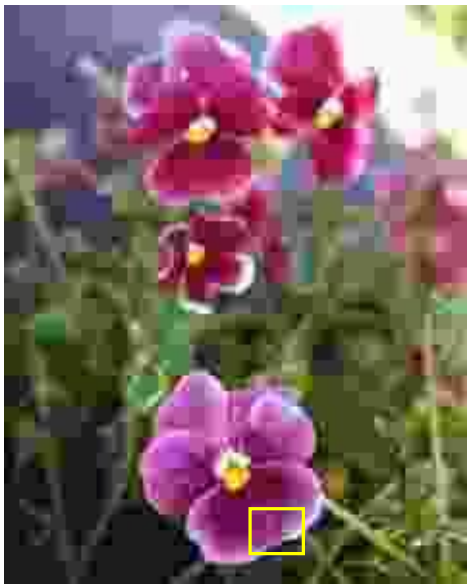


Schede Hollerith

CD-Rom

Figura 12 – I censimenti della popolazione negli Stati Uniti: dalla carta alla memoria digitale

Per comprendere come si possa rappresentare con le cifre zero e uno qualunque tipo di dato, guardate la Figura 13, in cui vediamo una viola del pensiero fotografata con macchine fotografiche digitali o telefoni smart phone a differente sensibilità.



Bassa sensibilità

Alta sensibilità

Figura 13 - Una viola del pensiero fotografata con telecamere di diversa sensibilità

Concentriamoci sul quadratino giallo che nelle due foto rappresenta lo stesso frammento di petalo, ed in particolare sul quadratino della foto a bassa sensibilità. Poniamoci l’obiettivo di

rappresentare il frammento di petalo, sezionandolo in 12 porzioni diverse, delimitate dalla griglia di segmenti gialli orizzontali e verticali in Figura 14.

Anzitutto, possiamo decidere di rappresentare tutti i colori (ad esempio rosso, verde, nero, ecc.) mediante le cifre da 0 a 7, e la intensità del colore ancora mediante cifre da 0 a 7 (0 corrisponde alla intensità più bassa, 7 corrisponde alla intensità più alta). Ad esempio (vedi le cornici blu in Figura 14) il primo quadratino della prima riga ha come colore base il *rosso*, che rappresentiamo con la cifra 3, ed è ad *alta* intensità, che rappresentiamo con la cifra decimale 6.

Con questa codifica decimale abbiamo bisogno di due cifre per ogni frammento di viola, e quindi in tutto di 24 cifre, in quattro righe e sei colonne.

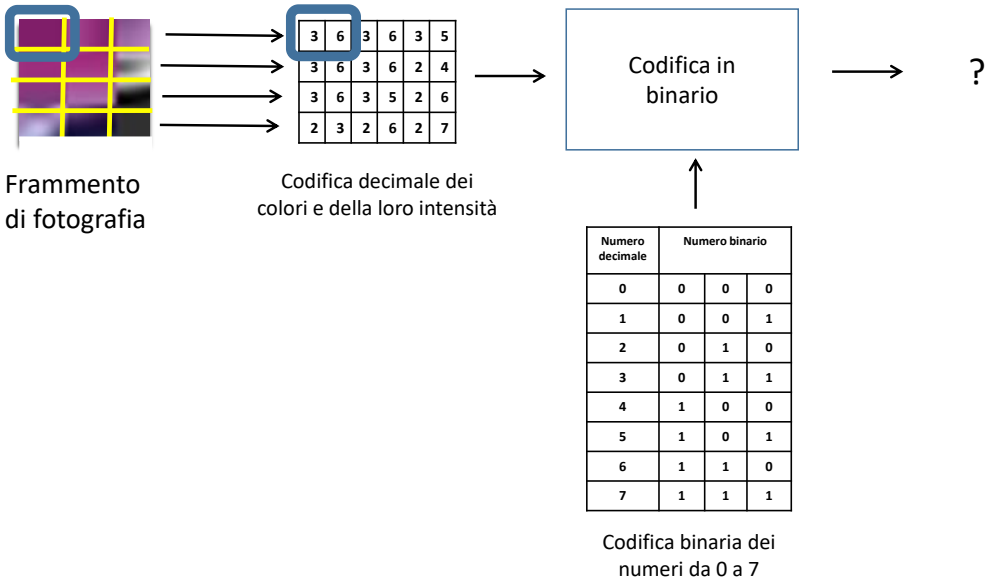


Figura 14 - Rappresentare una viola del pensiero con le cifre zero e uno

Adesso trasformiamo i numeri decimali della matrice in numeri binari, parte destra di Figura 14. Se dovessimo rappresentare le sole cifre da 0 a 3, potremmo usare *due* cifre binarie per ogni cifra, facendo corrispondere alla cifra 0 in decimale la coppia di cifre 00 in binario, a 1 → 01, a 2 → 10, a 3 → 11.

Ma noi dobbiamo rappresentare le cifre decimali da 0 a 7, e quindi abbiamo bisogno di *tre* cifre binarie. A questo punto vi propongo una prima domanda; guardate nuovamente la figura 14; secondo voi come possiamo rappresentare in binario la matrice di numeri decimali composta da quattro righe e sei colonne? Quante righe ha questa matrice? Quante colonne ha? Quali valori binari sono rappresentati nella matrice?

La risposta nella prossima pagina.

In Figura 15 vediamo la risposta: abbiamo bisogno di una matrice di quattro righe e 18 colonne, tre cifre binarie per ogni cifra decimale. Nelle cornici e frecce blu e rosse vediamo evidenziate due corrispondenze tra cifre decimali e gruppi di cifre binarie.

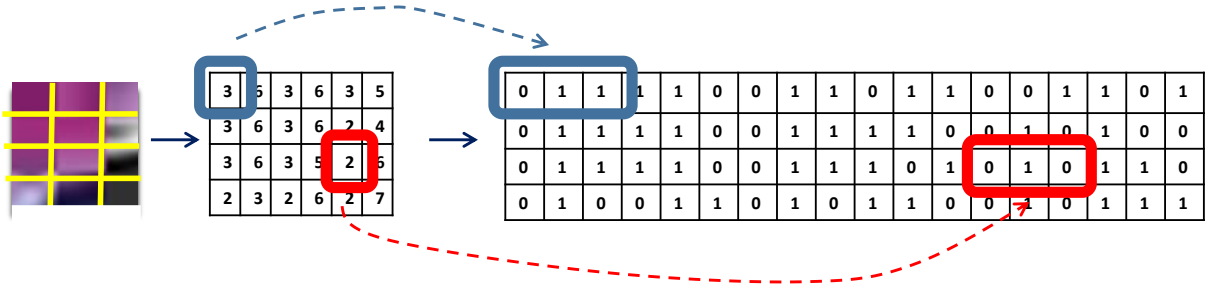


Figura 15 – Soluzione alla domanda

I *dati digitali*, cioè i dati rappresentati nella memoria di un computer, sono dunque sequenze di zero e uno che possono rappresentare in principio qualunque aspetto del mondo attorno a noi. Quindi i testi, le fotografie su Istagram, i video su You Tube, i messaggi in una rete sociale, sono tutti dati digitali rappresentati mediante sequenze di cifre zero e uno.

Anche una frase, ad esempio "oggi mi sento stanco" può diventare un dato digitale?

Certamente sì, una volta che tu abbia definito una corrispondenza tra le lettere dell'alfabeto e gli spazi bianchi tra le lettere, e le sequenze di 0 e 1, come abbiamo fatto poco fa con i numeri decimali da 0 a 7. Attenzione! Posso certamente rappresentare la frase nella memoria di un computer, ma questo non significa che il computer la *capisca*. Perchè un computer capisca il significato della frase, è necessario sviluppare tecniche che ricadono nella elaborazione del linguaggio naturale, come vedremo nel terzo Libro della Enciclopedia dedicato alle forme e al significato dei dati.

I precedenti esempi ci fanno capire perché la diffusione dei dati digitali stia crescendo negli ultimi anni a ritmi sempre più intensi.

Fino all'avvento di Internet e del *World Wide Web*[4], o *Web* per brevità, i dati venivano scambiati prevalentemente tra le organizzazioni, cioè le pubbliche amministrazioni, le banche, le compagnie aeree, ecc. per fornire servizi di varia natura, come i certificati, i bonifici bancari, la prenotazione di viaggi aerei, ecc. La Figura 16 mostra un grafico disegnato alla lavagna dal grande probabilista Bruno De Finetti; questa figura, che personalmente trovo bellissima, mostra i flussi di dati scambiati nel 1962 tra le pubbliche amministrazioni italiane centrali (i Ministeri e le loro organizzazioni periferiche, ad esempio il Catasto) e locali (i Comuni), l'Istat (Istituto Nazionale di Statistica), le aziende concessionarie di servizi, come ad esempio le aziende per l'elettricità, le banche e le assicurazioni. Nel riquadro sulla sinistra si riconosce il simbolo utilizzato a suo tempo per indicare la memoria secondaria.

[4] Il Wordl Wide Web è nato al CERN di Ginevra, dove ancora oggi vi è affissa una targa (vedi http://web-marketing.netd.it/il-luogo-in-cui-e-nato-il-web-where-the-web-was-born/)

Da allora è passato poco più di mezzo secolo, ma tutto è cambiato, in particolare con l'avvento del Web, la immensa *biblioteca* di dati digitali diffusa nel mondo. Per fornire un solo indicatore, i dati digitali 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 i dati e la conoscenza nella storia della umanità.

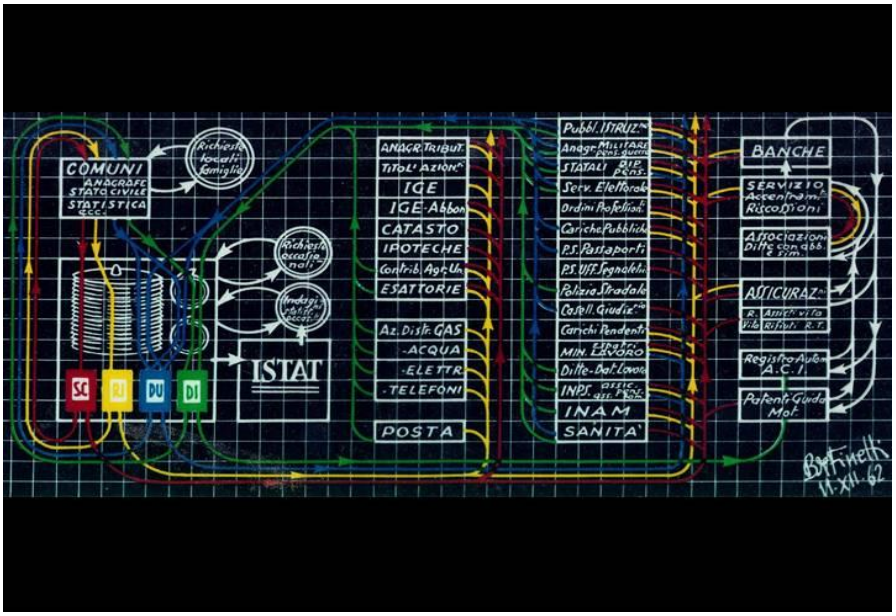


Figura 16 – I flussi di dati tra pubbliche amministrazioni italiane visti da Bruno De Finetti

I big data, i grandi dati

Per questa ragione i dati digitali sono chiamati oggi *big data*, grandi dati. Ciò porta 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 dei dati digitali nella nostra vita, portatrice di grandi innovazioni e di servizi come le app sui nostri telefoni mobili che rispondono a tante nostre esigenze, e portatrice di grandi scoperte scientifiche. Una presenza, come vedremo meglio nel seguito del libro, che può essere potenzialmente intrusiva nella vita della singole persone e delle comunità.

Cosa ha reso possibile l'improvviso irrompere dei dati digitali in tanti aspetti della nostra vita? Tutto ciò è dovuto al moltiplicarsi di tecnologie che fanno uso di dati digitali, ed in particolare allo sviluppo e diffusione delle cinque "grandi" tecnologie nate negli ultimi due decenni mostrate in Figura 17. Esse sono:

1. I *telefoni mobili*, che ci permettono di essere connessi "sempre e ovunque", effettuando telefonate, scambiando dati di varia natura (messaggi, foto, ecc.) e usando applicazioni che acquisiscono, elaborano, inviano dati sul Web.
2. le *reti sociali*, che permettono a chiunque con pochi e semplici comandi di comunicare pensieri, opinioni, emozioni.
3. *L'internet delle cose* (*internet of things o IoT* in inglese), che permette attraverso la installazione nel mondo degli oggetti fisici di una miriade di sensori, la integrazione del mondo

fisico sensibile e del mondo virtuale dei dati digitali (torneremo sulla relazione tra questi due mondi nel prossimo Capitolo).
4. il *cloud* (o nuvola) che fornisce risorse di calcolo e di memoria, rendendo facilmente accessibili e condivisibili l'elaborazione e la memorizzazione dei dati digitali.
5. i *big data*, i grandi volumi di dati che, come vedremo diffusamente nel libro, sono sempre più utilizzati nel mondo della ricerca e dei servizi digitali.

Accanto a queste tecnologie ci sono *Internet*, la rete di comunicazione mondiale che permette di inviare e ricevere dati digitali tra ogni luogo del pianeta, e il *Web*, luogo virtuale dove sono accessibili miliardi di siti e di pagine di dati digitali; sono state scoperte tante altre tecnologie digitali, come la posta elettronica, la stampa 3D, i sistemi di gestione di basi di dati, e molte altre ancora, che qui è impossibile riassumere.

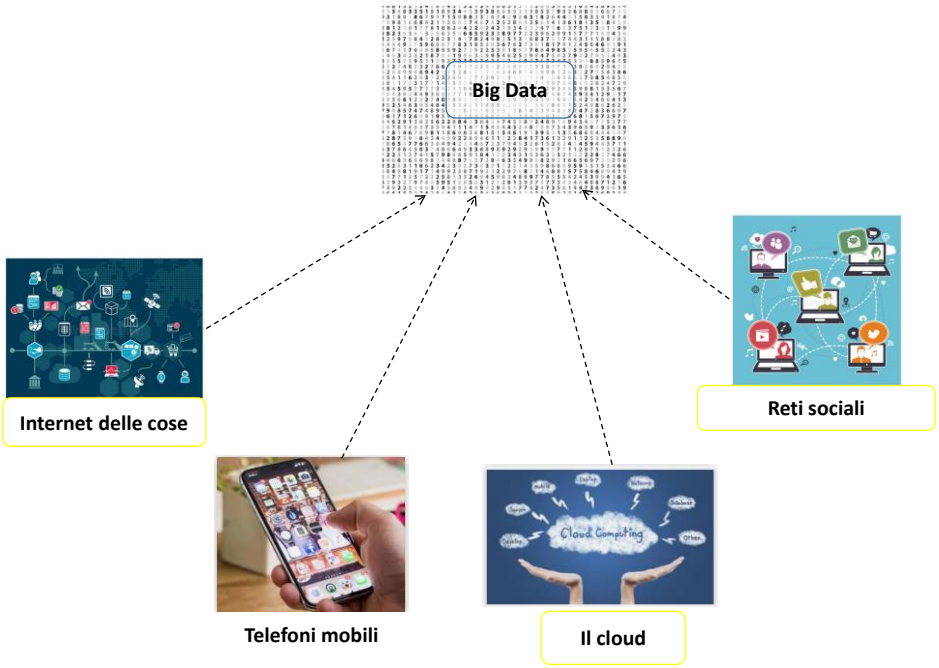


Figura 17 – Le cinque grandi tecnologie digitali

Non riesco a capire una questione di base: perché è così importante che i dati siano digitali?

Perché così possono essere elaborati digitalmente! Ma per elaborarli digitalmente, devo introdurre la tecnologia più importante di tutte: il *computer*, la cui concezione iniziale risale addirittura al diciannovesimo secolo, vedi in Figura 18 la macchina di Babbage, progettata nei primi decenni dell'1800.

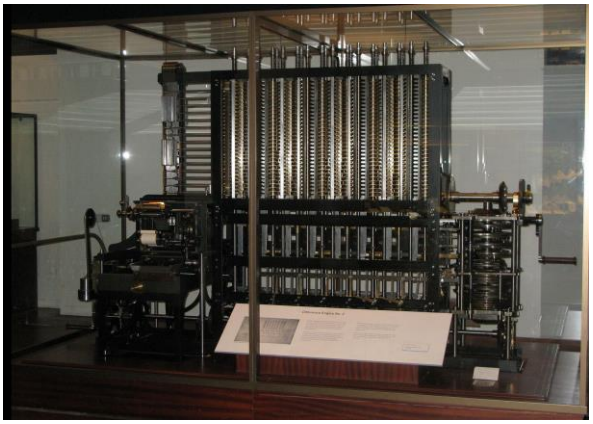


Figura 18 – La macchina di Babbage, tratta dal sito https://en.wikipedia.org/wiki/

Come fanno i computer a elaborare tanti diversi tipi di dati? Come fa uno smart phone a correggere una foto cambiando la luminosità dei colori, come fa il programma Alpha Go a sconfiggere il più grande giocatore di Go del mondo?

Ottima domanda; a questo punto dobbiamo capire …

Come è fatto un computer, in poche parole[5]

Guardate la Figura 19. Un computer è una tecnologia che svolge le seguenti funzioni:

1. *acquisisce dati* dal mondo esterno, tramite *unità di ingresso* (nella figura, una tastiera e un termometro digitale);
2. *elabora dati* tramite *l'unità di elaborazione;*
3. *memorizza dati* temporaneamente (*memoria principale*) o permanentemente (*memoria secondaria*);
4. *trasferisce dati* al mondo esterno tramite *le unità di uscita* (nella figura uno schermo e una stampante).

Lo scopo delle *unità di ingresso* è quello di codificare in alfabeto binario i dati del mondo reale che rappresentano fenomeni fisici: ad esempio, la temperatura corporea, nel caso di un termometro digitale, ovvero un sensore fotoelettrico che segnala il passaggio di una automobile dal casello di una autostrada.

Lo scopo delle *unità di uscita* è complementare, rappresentare cioè i datI digitali risultato della elaborazione in un formato *leggibile* dal mondo esterno. Lo schermo di un tablet rappresenta i dati in modo che siano leggibili con i nostri occhi, la stampante rappresenta i dati su fogli di carta così che li possiamo leggere o conservare.

L'unità di elaborazione è l'elemento del computer che elabora i dati ed esegue *programmi, composti da istruzioni (*vedi tra poco). Fisicamente, è composta da un microprocessore costituito da un monocristallo di silicio, materiale che è in grado di condurre elettricità.

[5] I lettori non particolarmente interessati alla tecnologia del computer e ai programmi software possono saltare il resto del capitolo

Non mi è ancora chiaro, però, perché la memoria sia di due tipi, memoria centrale e memoria secondaria.

La funzione della memoria principale l'abbiamo vista, memorizzare i dati utilizzati o prodotti dai programmi che vengono eseguiti nella unità di elaborazione. I *dati su cui opera un programma* si trovano nello stesso componente di silicio in cui si trova la unità di elaborazione, questo con lo scopo di ridurre il più possibile il tempo necessario per eseguire le istruzioni che operano sui dati (per trasferire un dato dalla memoria principale alla unità di elaborazione ci vogliono 10^{-8}/10^{-9} secondi).

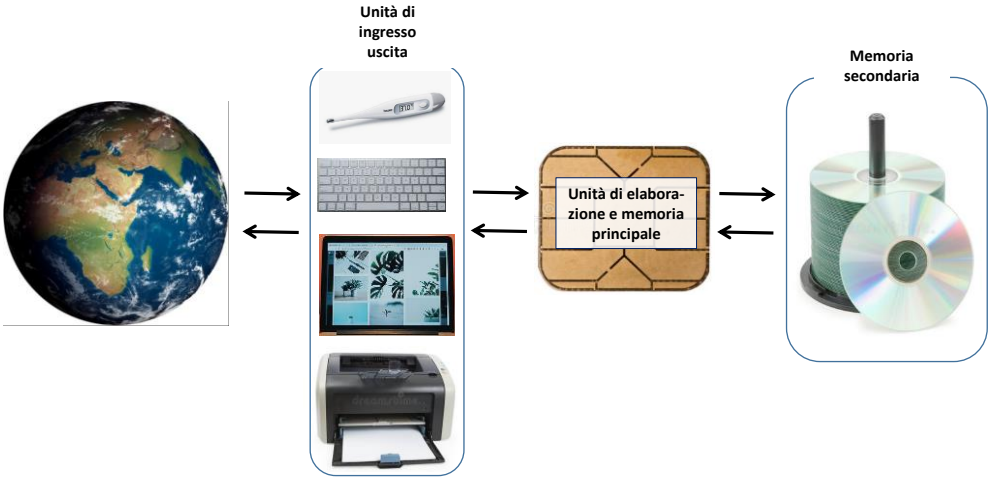


Figura 19 - Come è fatto un computer

Se noi vogliamo memorizzare *permanentemente* i dati prodotti da un programma, così che in futuro possiamo recuperarli (pensa al registro digitale di una professoressa, con i nomi e cognomi degli studenti e i voti assegnati nelle interrogazioni), allora dobbiamo utilizzare un diverso tipo di memoria, la *memoria secondaria*, che è molto più capiente della memoria principale, meno costosa, e che ha però tempi di trasferimento dalla/alla unità centrale molto più elevati (ordine dei millisecondi).

Tutti i computer del mondo sono fatti in questo modo?

Lo schema è sempre questo, ma ci sono molte varianti. Per esempio, per poter elaborare grandi quantità di dati, è possibile che invece di una unità di elaborazione ce ne siano tante, collegate tra di loro; un'altra recente evoluzione è nell'uso del cloud, per cui la memoria secondaria non è più o non è solo direttamente collegata al computer, ma risiede sul cloud.

Questa è la soluzione adottata in Dropbox e Google Drive, che sono utilizzati ormai da molte persone per memorizzare le proprie foto, i documenti o che si ha interesse ad avere a portata di mano o di interesse storico, ma la cui dimensione eccede lo spazio di memoria del proprio computer.

Come fa l'unità di elaborazione a elaborare dati? Cosa intendi per "elaborare"?

Siamo appena all'inizio e ti poni domande (*cosa intendi per elaborare?*) che hanno occupato la mente di matematici e filosofi per tanto tempo! Cercherò di rispondere alla prima domanda, mentre per la seconda domanda, più teorica, puoi leggere il mio testo *Le basi dell'informatica* liberamente scaricabile dal sito http://hdl.handle.net/10281/97703.

I programmi software

Una unità di elaborazione esegue *istruzioni*. Le istruzioni, come i dati, sono rappresentate in *formato binario*; un esempio di istruzione (scritta in italiano….) è "leggi il dato dal termometro e trasferiscilo all'indirizzo 100 della memoria principale", oppure "leggi il dato memorizzato nell'indirizzo 1100 della memoria principale e trasferiscilo all'indirizzo 10100 della memoria secondaria (per memorizzarlo permanentemente)". Il linguaggio che esprime le istruzioni in fomato binario è detto linguaggio macchina.

Un *programma* o *programma software* è un insieme di istruzioni che leggono dati dalle unità di ingresso, scrivono dati sulle unità di uscita, eseguono calcoli (per esempio moltiplica due numeri), prendono decisioni (per esempio ora vai a eseguire la istruzione 10).

Attenzione! I programmi non vengono scritti dai programmatori in linguaggio macchina, altrimenti …. uscirebbero di senno a dover imparare le codifiche binarie delle diverse istruzioni, e a ricordarsi tutti gli indirizzi di memoria utilizzati! Piuttosto, i programmi sono scritti in un linguaggio simbolico, più facile da ricordare e da usare. Un esempio semplicissimo è il programma che somma due numeri composti sulla tastiera di un telefono mobile.

Se i programmi si rappresentano anch'essi in un alfabeto binario, come fa l'unità di elaborazione a "capire" ed eseguire queste istruzioni?

Sono stati scritti una volta per tutte altri programmi, chiamati *traduttori*, che traducono i precedenti programmi in linguaggio macchina, così che le istruzioni siano intepretabili ed eseguibili dalla unità di elaborazione.

Algortimi e programmi

Ora cerchiamo di capire la differenza tra i concetti di problema, algoritmo e programma. Anche se non sapete nulla di linguaggi programmativi, provate a cercare di capire "cosa" calcolano i seguenti due programmi, sulla base delle indicazioni che darò. Per descrivere i due programmi usiamo un semplice linguaggio programmativo adattato da un linguaggio reale.

Il *Programma 1* consiste di due istruzioni:

1. SOMMA = 10 * (10 + 1) / 2
2. TERMINA

la prima istruzione calcola il valore a destra del simbolo di =, e lo assegna alla variabile SOMMA.

Cosa è una variabile?

Hai ragione! Prendi un foglio di carta, e chiamalo con il nome SOMMA. Questo foglio di carta, e il suo nome, SOMMA, ci serve per ricordare i risultati delle istruzioni del programma. La seconda, e ultima, istruzione del programma dice che è terminato. Ora, quale è il numero che scrivi sul foglio di carta?

110/2, è 55! Corretto?

Corretto!

Ho capito. Scusa, non servirebbe anche una istruzione che mi comunica il risultato? Se no, il computer se la tiene per sé la somma!

Hai ragione! Sono stato troppo sintetico. Modifichiamo il programma, aggiungendo una istruzione di stampa, vedi il box di Figura 20.

```
SOMMA = 10 /2  * (10  +  1)
STAMPA SOMMA
TERMINA
```

Figura 20 - Programma 1 modificato

Cerchiamo di capire ora quale è il *problema* che è risolto dal programma. I primi dieci numeri interi sono:

1 2 3 4 5 6 7 8 9 10

Se li sommo a coppie, cioè sommo 1 e 10, 2 e 9, 3 e 8,. 4 e 7, 5 e 6, per ogni coppia ottengo il valore 11. Le coppie da sommare sono 10/2 uguale a 5, quindi il risultato è 10/2 * 11 uguale a 55.

Quindi il problema risolto dal Programma1 è

Somma i primi 10 numeri interi

Infatti per 5 volte (10/2) somma una coppia di numeri (1+10, 2 + 9, …) che ha come risultato 11, cioè 10 + 1. Se non sei convinto ti prego di riguardare tutto il ragionamento numerico.

Ora osserviamo il Programma 2 in figura 21.

```
SOMMA = 0
PER I CHE VA DA 1 A 10 ESEGUI
SOMMA = SOMMA + I
STAMPA SOMMA
TERMINA
```

Figura 21 - Programma 2

La prima istruzione dice di scrivere il valore 0 sul foglio di carta con nome SOMMA.

La seconda istruzione dice che per dieci volte (PER I CHE VA DA 1 A 10 ESEGUI) devi eseguire l'istruzione seguente; occorre eseguirla la prima volta con I uguale a 1, la seconda con I = 2, ecc. la decima volta con I uguale a 10; I valori calcolati devono essere memorizzati ogni volta sul foglio di carta che ha nome SOMMA.

Ti faccio anzitutto una domanda: che valore scrivi sul foglio SOMMA la prima volta che esegui la seconda istruzione?

Sul foglio la prima volta scrivo il valore 0, poi sommo 0 a 1 e mi viene 1, e a questo punto scrivo 1 sul foglio. E' un po' strano, un po' arzigogolato…

E' vero, c'è scritto 1! Non è complicato, basta che mi segui un attimo. Il foglio SOMMA contiene la prima volta 0 + 1, la seconda 0 + 1 + 2, e così via, è chiaro?

Sì, è chiaro, ma dove vuoi andare a parare?

Voglio capire insieme a te quale numero compare alla fine della esecuzione del programma?.

Dunque, 0 + 1 + 2 + 3…, fino a 10….

E' così, e quanto fa?

…..55. lo stesso di prima!

Esatto! i due programmi *calcolano lo stesso valore, la somma dei primi 10 numeri interi*! Il primo programma effettua questo calcolo sulla base di una semplice proprietà matematica, per cui la somma dei primi N numeri interi è pari a N x (N + 1) /2. Il secondo programma invece li somma a uno a uno.

Ricapitolando, guardate la Fiigura 22.

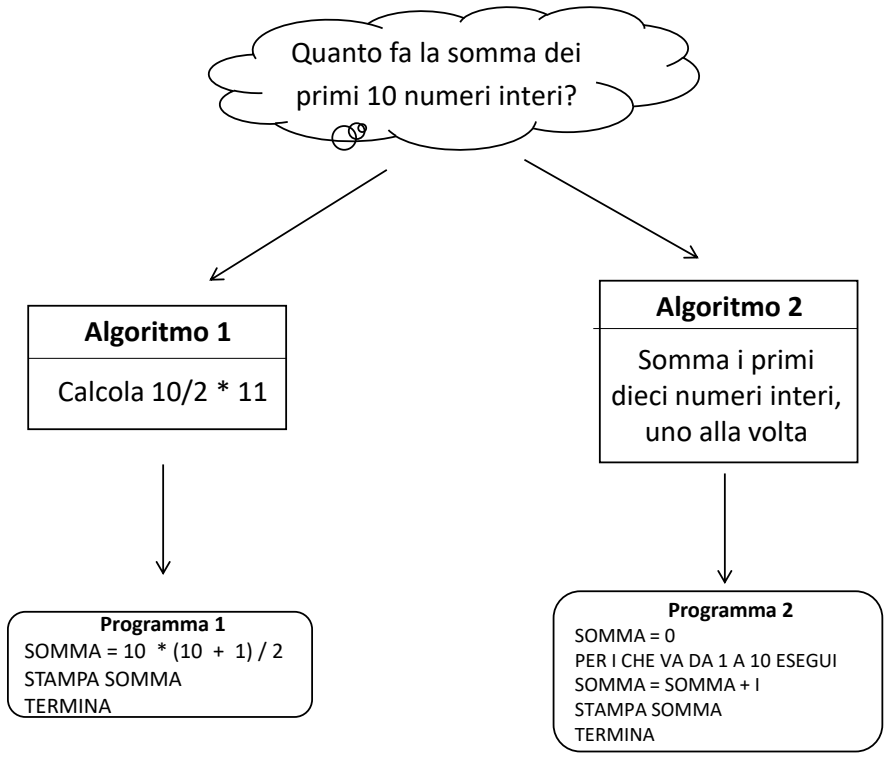


Figura 22 – Problema da risolvere, algoritmi risolutivi, programmi scritti in un linguaggio intepretabile da un computer

Dato un problema, nel seguito useremo il termine *algoritmo* quando intendiamo parlare di un metodo riusolutivo del problema descritto in forma astratta, useremo il termine *programma* o *programma software* quando pensiamo l'algoritmo scritto in un linguaggio programmativo, un linguaggio cioè comprensibile ed eseguibile da un computer.

Se non hai saltato nulla di questo capitolo, e sei arrivato fin qui, prometto che nel seguito non parlerò più di tecnologie digitali…

Riassumendo

Il grande uso che facciamo dei dati digitali deriva anzitutto dalla recente comparsa delle ***tecnologie*** del ***telefono mobile***, le **reti sociali**, ***l'internet delle cose***, il ***cloud***, che producono e fanno uso di una quantità crescente di ***grandi quantità di dati*** detti anche ***big data***.
Le altre tecnologie fondamentali legate ai dati digitali sono la ***rete Internet***, che permette di trasmettere dati in tutto il mondo a costi molto bassi, e il ***computer***, che permette di elaborare dati digitali producendo altri dati.
Il computer è costituito da una ***unità di elaborazione***, da ***organi di ingresso e di uscita***, che fanno interagire i computer con il mondo esterno, e da diverse memorie, la ***memoria centrale***, dove risiedono i dati che sono elaborati da programmi, e la **memoria secondaria** dove risiedono i dati che sono memorizzati permanentemente.
I dati digitali sono rappresentati con due cifre, ***0 e 1***. I ***programmi software*** sono definiti in un ***linguaggio programmativo***, che può essere il ***linguaggio macchina***, direttamente eseguibile dalla unità di elaborazione, ovvero ***in linguaggio ad alto livello***, comprensibile da esseri umani. Un ***algoritmo*** è un programma descritto in maniera più astratta.

Capitolo 3

I dati sono la nostra finestra sul mondo
Mondo analogico e mondo digitale

L'introduzione dei computer e delle grandi tecnologie descritte in Figura 17 sta enormemente espandendo il mondo dei dati digitali, che chiameremo nel seguito *mondo digitale*, e che filosofi come Floridi chiamano *infosfera*, e scrittori come Baricco *l'oltre-mondo*.

Ben prima che nascesse il mondo dei dati digitali, a partire dall'homo sapiens, gli esseri umani sono abituati da tempo immemorabile a interagire nella propria vita con il *mondo sensibile*, dal quale continuamente, con i nostri sensi, percepiamo fenomeni ed eventi; siamo abituati a codificare questi fenomeni ed eventi mediante dati che hanno una natura diversa dai dati digitali, e che siamo abituati a memorizzare ed elaborare nella nostra mente. Chiameremo questi dati *analogici*.

Il Semaforo Vecchio sul monte di Portofino

Antichi documenti citano il Semaforo Vecchio che si trova sul monte di Portofino come sede di segnalazioni effettuate con fuochi e fumate per i naviganti, e successivamente come stazione di Telegrafo di epoca napoleonica. Queste due funzioni sono ben chiarite nel cartello che riproduco in Figura 19.

Figura 23 – I segni utilizzati nel telegrafo del semaforo vecchio

Il semaforo (notate la etimologia: semaforo in greco antico significa *portatore di segni*) trasmetteva un alfabeto di segni ben definito e codificato: a ogni segno corrispondeva un

significato, e i segni venivano comunicati originariamente tramite segnali col fuoco. Il telegrafo (etimologia: scrivo lontano) permetteva di trasmettere messaggi più complessi, corrispondenti a frasi del linguaggio italiano, esprimibili come sequenze di segni elementari mostrati in Figura 24.

Tutti i naviganti a cui venivano trasmessi dal Semaforo e dal Telegrafo i dati analogici costituiti dai simboli o dalle frasi del telegrafo dovevano conoscere il loro *significato*, altrimenti non potevano condividere le azioni conseguenti. Ma conoscere il significato dei simboli del Semaforo era più semplice che comprendere il significato delle frasi costituite dalle sequenze di parole inviate con il Telegrafo; questo accade anche nella nostra vita: capire un simbolo, ad esempio un segnale stradale, è molto più semplice e immediato che capire una sequenza di parole in linguaggio italiano.

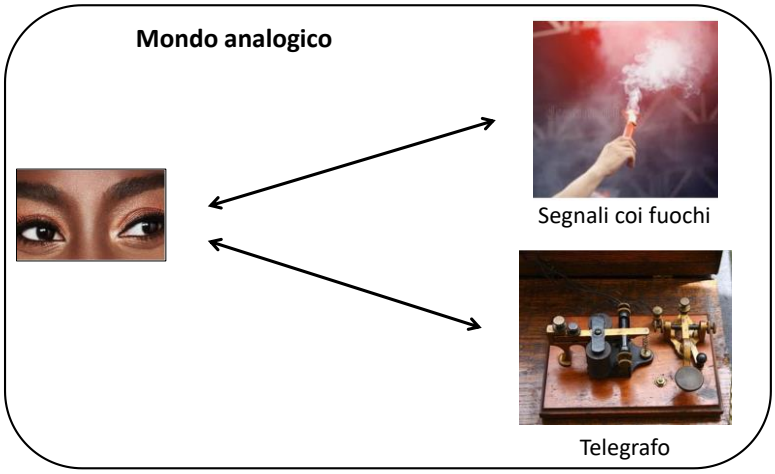


Figura 24 – Comunicazione dei dati nel mondo analogico del Semaforo vecchio del monte di Portofino

Il mondo del Monte di Portofino era un mondo esclusivamente analogico. In quel mondo, i naviganti dovevano conoscere bene il significato dei dati, a quel tempo tutti dati analogici. E nel nostro mondo, cosa accade?

Mondo analogico e mondo digitale nelle elezioni americane

In occasione delle elezioni presidenziali americane del 2000, si scontrarono per il partito repubblicano George W. Bush e per il partito democratico Al Gore; per settimane le elezioni non poterono essere decise perché non si era in grado di contare con precisione i voti assegnati all'uno e all'altro nel decisivo stato della Florida.

Ciò derivava dal fatto che a quell'epoca (e ancora oggi) venivano usate diverse modalità di espressione del voto, in Figura 25 vediamo le modalità adottate nelle elezioni successive del 2004.

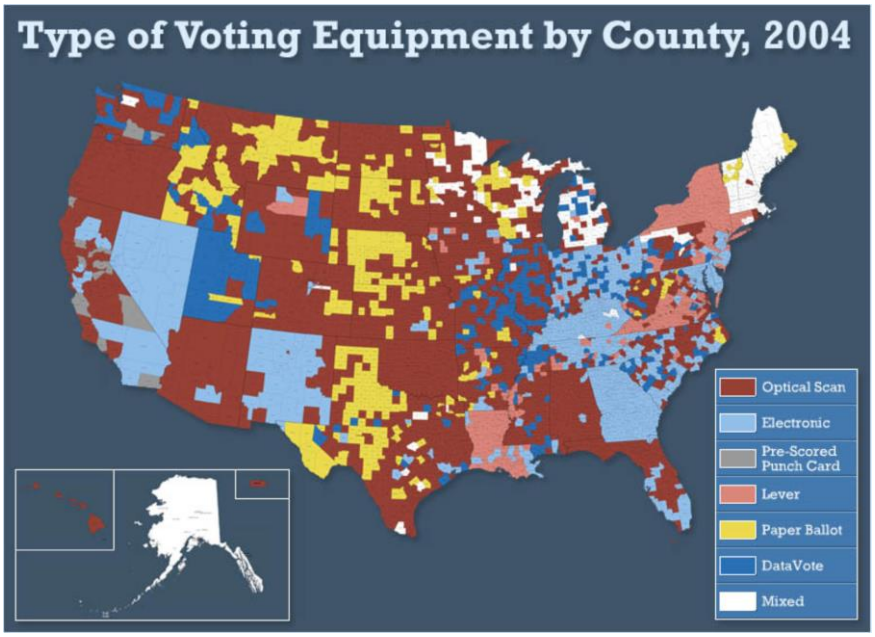


Figura 25 – I diversi modi di raccogliere le espressioni di voto

Una modalità usata in molti seggi elettorali consisteva nel perforare schede Hollerith, che abbiamo già incontrato nella Figura 12. Un buco in una certa posizione della scheda corrispondeva a un voto a Bush, un buco in una posizione vicina corrispondeva a un voto a Gore. Ora, siccome le perforatrici, fatte di componenti meccaniche, tendevano a perdere l'allineamento tra le colonne della scheda e i buchi riferiti a Bush e a Gore, non ci si metteva d'accordo sul … significato dei buchi. Le due persone mostrate in Figura 26 non sono due cittadini curiosi che cercano di capire a cosa corrisponde un buchetto, sono due giudici del tempo, che esaminavano le schede una a una per capire a quale dei due candidati associarle!

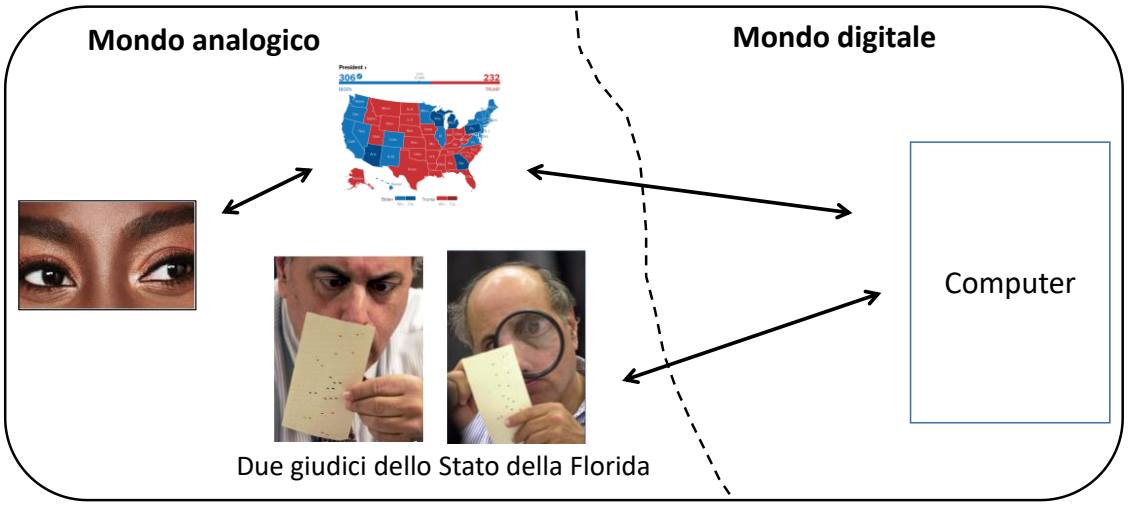


Figura 26 – Due giudici della Florida che cercano di "leggere" due schede perforate

La preferenza di un elettore americano a favore di Bush o Gore è un esempio di dato; questo dato ha due possibili valori, o significati: Bush o Gore. Con le diverse modalità di esprimere la

preferenza mostrate in Figura 25, il valore, corrispondente alla preferenza dell'elettore, può essere rappresentato fisicamente:

1. un buchetto su una scheda Hollerith associato a Bush o Gore.
2. la scrittura del cognome di Bush o Gore su una scheda, ovvero
3. una croce messa a mano sul cognome Bush o sul cognome Gore in due riquadri, come accade nelle elezioni comunali in Italia

Il mondo di Figura 26 è un *mondo ibrido*, in parte analogico e in parte digitale. I voti decisi dai giudici sono dati del mondo analogico, che vengono poi comunicati alla applicazione software che trasforma i dati analogici in dati digitali e conta i voti dei due contendenti, producendo la visualizzazione a colori degli Stati Uniti che vediamo riprodotta in Figura 26.

Cominciamo a vedere con l'esempio delle elezioni americane una stretta connessione tra i due mondi; la produzione della cartina degli Stati Uniti è frutto di dati che, in parte, ma solo in parte, sono stati *decisi* da esseri umani. Il programma software che li conteggia, tuttavia, non distingue tra dati selezionati da una tastiera di un computer e poi inviati al computer centrale, e dati decisi dai giudici, a meno di non tener traccia, voto per voto, della *loro provenienza*. Possiamo concludere dall'esempio quanto scritto nella cornice.

Tener traccia della provenienza e dell'origine del dato digitale ci fa capire la sua storia, e ci fa conoscere meglio la sua natura e il suo significato. Il significato dei dati digitali comprende la loro storia.

Quando la Regione Lombardia diventò zona rossa senza meritarlo

Nel gennaio 2021 il Ministero della Salute assegnò alla Regione Lombardia il colore rosso, suscitando le vibranti proteste della Regione, che si sentì penalizzata da questa decisione, e affermò che sulla base dei suoi calcoli il colore non doveva essere rosso.

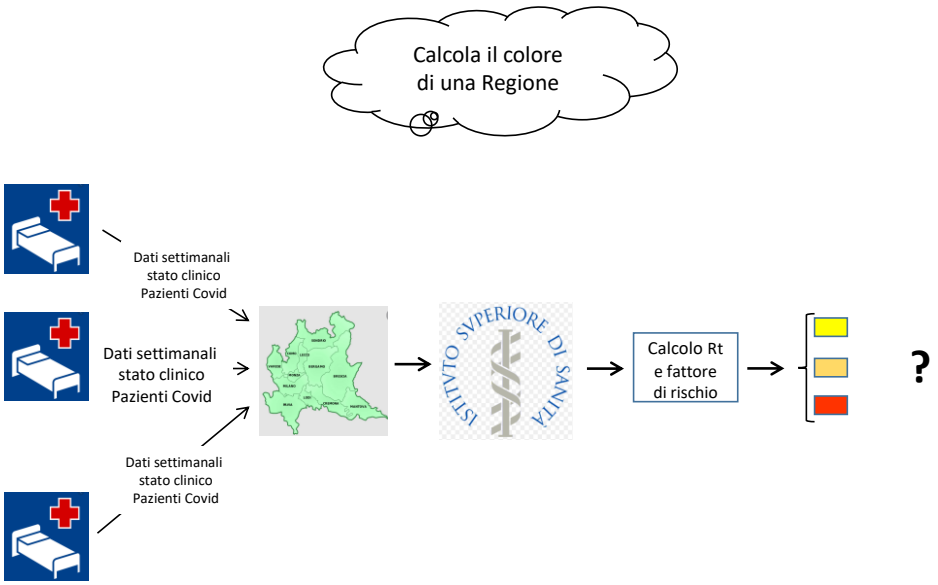


Figura 27 – I dati raccolti e scambiati e gli enti coinvolti nel calcolo dell'indicatore Rt

In Figura 27 vediamo rappresentanti i soggetti coinvolti nella decisione:

- gli enti territoriali (ad esempio gli Ospedali) che inserivano in un programma software i dati richiesti per il calcolo dei 21 indicatori e dell'indicatore Rt,
- la Regione Lombardia che raccoglieva i dati dai diversi enti territoriali e li inviava all'Istituto Superiore di Sanità,
- e infine quest'ultimo che sottoponeva i dati pervenuti dalla Regione al programma di calcolo dell'Rt.

Chi aveva ragione tra la Regione Lombardia e l'Istituto Superiore di Sanità? Nel Capitolo 8 discuteremo a fondo quanto accaduto, ora, piuttosto, vorrei invitarvi a ragionare nell'ambito dei mondo rappresentato nella Figura 27 sulle seguenti questioni:

1. Quale parte del mondo rappresentato in Figura 26 corrisponde al mondo analogico e quale parte corrisponde al mondo digitale?
2. Cosa c'è al posto del punto interrogativo? Per spiegarmi meglio: chi subisce gli effetti della decisione legata alla assegnazione del valore giallo, arancione, rosso?

Le risposte nella prossima pagina.

In Figura 28 vediamo rappresentate le parti del mondo che corrispondono an mondo analogico e al mondo digitale. Riguardo alla seconda domanda, le conseguenze della assegnazione del colore ricaddero sulle persone, su tutti noi, e sugli esercizi commerciali, perché il colore rosso portava a limitazioni ai nostri movimenti e alla attività degli esercizi commerciali. I dati inseriti dagli enti territoriali entravano nella "scatola grigia" del mondo digitale, e ne uscivano solo al termine del calcolo dell'indicatore Rt.

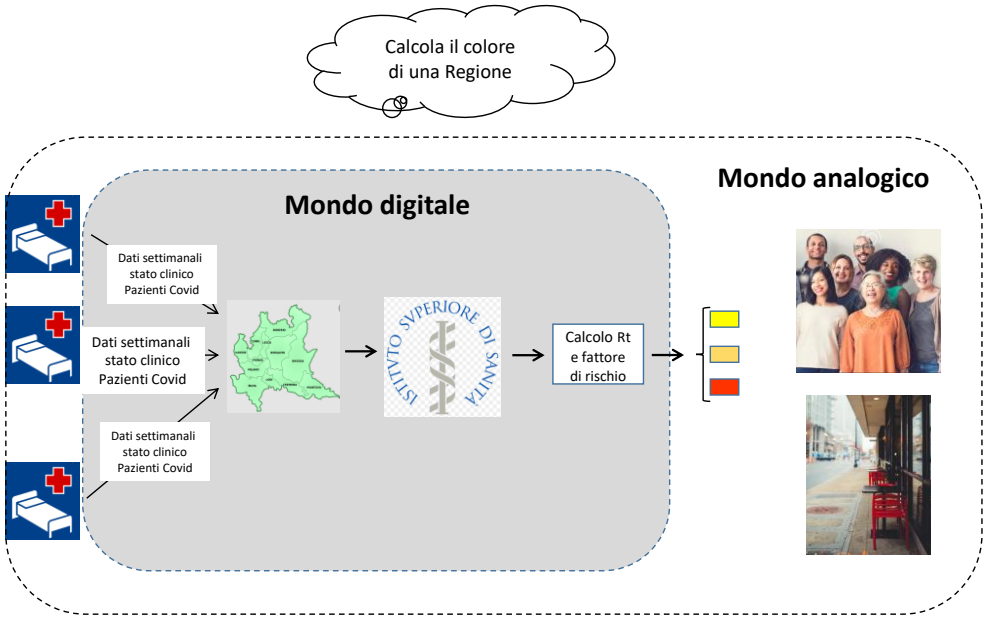


Figura 28 – Mondo analogico e mondo digitale nel calcolo dei colori

Ancora sul significato dei dati - Sento la fronte calda: ho la febbre?
Dati, informazioni, conoscenza

Sentiamo un leggero calore sulla nostra fronte, e misuriamo la temperatura corporea per capire se abbiamo la febbre; cerchiamo di capire cosa accade nei due casi mostrati in Figura 29, in cui nella parte superiore compiamo tutte le attività da soli e nella parte inferiore con l'aiuto di un computer.

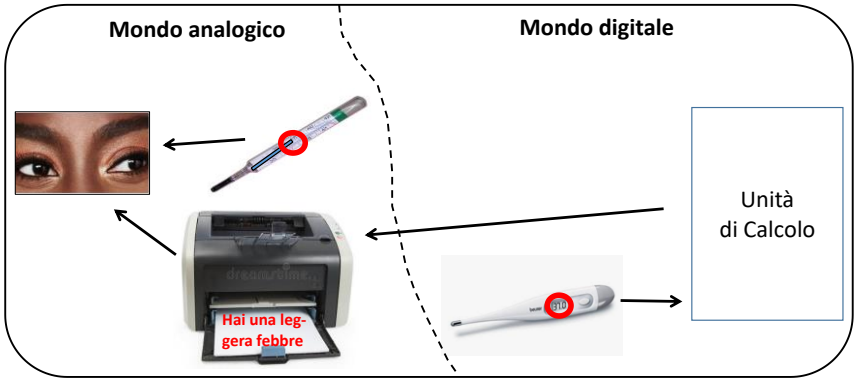


Figura 29 – Mondo analogico e mondo digitale nella misura della temperatura

Se rimaniamo all'interno del mondo analogico, dobbiamo munirci di un termometro; se il termometro è di quelli tradizionali con una colonnina di mercurio che si espande in modo proporzionale alla temperatura misurata, dopo qualche minuto dobbiamo estrarlo e traguardare il punto della colonnina dove il mercurio si è fermato rispetto alla scala di valori sottostante. In Figura 29, in corrispondenza del cerchietto rosso noi leggiamo il numero 37.0.

Attenzione, e scusate la pignoleria: *siamo in grado di leggerlo se conosciamo i numeri con virgola*, argomento di matematica che si studia nella scuola elementare.

Se poi, oltre che saper leggere i numeri con virgola, noi siamo consapevoli del *significato di quel numero*, allora possiamo affermare che

37.0 è la mia temperatura corporea in questo momento

Anche in questo caso, in qualche momento del passato abbiamo dovuto apprendere cosa è una temperatura corporea. Sei convinto della cosa, o trovi qualcosa che manca?

Ci sono diversi modi di misurare la temperatura corporea, e in Italia adottiamo la scala Celsius.
.
Hai ragione, la nostra conoscenza in realtà è

37.0 è la mia temperatura corporea in gradi Celsius in questo momento

I due passaggi che abbiamo compiuto non sono per nulla banali. E' la nostra *conoscenza* pregressa del fenomeno che ci permette di associare un *significato*, cioè *temperatura corporea in gradi Celsius*, al numero letto.

Si suole anche dire che il dato iniziale, 37.0, si è ora trasformato in *informazione*, cioè un dato *dotato di significato*, e che la trasformazione è stata possibile perché nella nostra mente noi abbiamo conoscenza dei concetti di

- termometro,
- temperatura e
- temperatura misurata con la scala Celsius.

Dopodichè, noi probabilmente sappiamo che la *soglia della febbre* sono 37 gradi, e quindi arriviamo alla conclusione che siamo alla soglia tra la temperatura normale e la febbre. Per arrivare a questa conclusione abbiamo fatto ricorso a una risorsa preziosa, la nostra conoscenza pregressa sul nostro corpo e sulla sua temperatura in condizioni normali.

Seguiamo ora il caso in cui non vogliamo sprecare neanche una goccia di fatica mentale per capire se abbiamo la febbre. In questo caso dobbiamo usare uno di quei termometri che codificano la temperatura per mezzo di una rappresentazione costituita da un dato digitale; se il termometro non è dotato di un suo computer interno, può trasmettere questo dato digitale a un telefono mobile o a un computer, che deve svolgere le elaborazioni che noi abbiamo fatto a mente sul dato analogico.

Ma, attenzione! per fare questo, è necessario trasferire al computer *non solo* il dato "37.0" rappresentato mediante le cifre zero e uno, ma *anche il suo significato di temperatura corporea in gradi Celsius*. Questo perchè il computer è come un bambino, all'inizio della sua vita non sa niente, tutto gli deve essere insegnato!

In questo caso risolvere il problema non è difficile. Basta scrivere un programma che, quando il computer legge il dato digitale, *attribuisca il significato* di *temperatura corporea* al dato che legge; successivamente il programma, con opportune istruzioni del tipo

SE TEMPERATURA CORPOREA HA VALORE MAGGIORE DI 36.9 E MINORE DI 37.4 ALLORA STAMPA "Hai una leggera febbre"

può confrontare "37.0" con i vari intervalli di valori, e stampare un messaggio come in Figura 28.

Se devo misurare la temperatura corporea del mio cane, oppure ho un termometro americano con i gradi in scala Fahrenheit, come faccio? Non si può usare lo stesso programma…

Hai ragione, devi scrivere ogni volta un nuovo programma! Da tempo, peraltro, sono stati sviluppati sistemi software che permettono di rappresentare permanentemente il significato dei dati in un certo contesto, per esempio *la temperatura corporea degli esseri umani*. Questi sistemi, chiamati *sistemi di gestione di basi di dati*, rappresentano i *valori* dei dati e il loro *significato* per mezzo di *tabelle*. Un esempio di tabella è riportato in Figura 30, dove nella prima riga trovi il significato dei valori e nelle righe successive i vari intervalli di temperatura e i relativi livelli di temperatura diagnosticati.

Intervallo di temperatura corporea		**Livello temperatura**
35.0	36.9	Normale
37.0	37.5	Leggera febbre
37.6	38.0	Febbre bassa
38.1	39.5	Febbre alta
39.6	40.5	Febbre molto alta
40.6	43	Febbre altissima

Figura 30 – Tabella per i diversi tipi di febbre

Ma anche i sistemi di gestione di basi di dati non bastano. Tutto lo sforzo che è in corso in questi anni, e che descriverò nel terzo libro della Enciclopedia, ha lo scopo di includere nei computer la conoscenza del mondo, ad esempio il fatto che i termometri misurano la temperatura corporea, che la soglia della febbre in Italia è assunta pari a 37.0, ecc.

In questo modo le risposte alle domande come quella all'inizio del capitolo possono essere fornite direttamente da un computer; in questo modo il computer è dotato non solo della capacità di *elaborare dati* digitali, ma anche di *utilizzare ed elaborare conoscenza autonomamente*, sia che si tratti di misurare una temperatura, sia che si tratti di giocare al gioco del Go.

D’altra parte, ancora siamo alla infanzia di questo grande sforzo! Se scriviamo sulla tastiera di un computer la frase

37.0 è la mia temperatura corporea in gradi Celsius in questo momento

oppure la pronunciamo con la nostra voce, ci sarà bisogno di un programma che comprenda il linguaggio naturale scritto o parlato, nel senso che dovrà;

1. riconoscere le songole parole della frase, poi
2. ricostruire il significato complessivo,
3. riconoscere che il valore all’inizio, 37.0, è il valore da associare al concetto che compare a metà della frase, cioè “temperatura corporea in gradi Celsius”,
4. confrontare questo valore con la conoscenza disponibile.

Insomma, una lunga strada prima di poter dare una risposta alla nostra domanda! Gli esempi che abbiamo visto, il Semaforo vecchio, le schede Hollerith nelle elezioni americane, il calcolo del valore Rt per la Regione Lombardia, la risposta alla domanda “Ho la febbre?”, ci fanno capire la grande importanza che ha vedere nitidamente *separati e allo stesso tempo interagenti* i due mondi in cui viviamo e con cui interagiamo, analogico e digitale.

Quando trasformiamo un dato analogico in un dato digitale, a questo punto nel mondo digitale viene rappresentato insieme al suo valore (nell’esempio della temperatura, 37.0) anche il suo significato (temperatura corporea). E noi, da questo momento, percepiamo il mondo analogico attraverso il mondo digitale; i dati digitali diventano una finestra sul mondo, vedi Figura 30, finestra che ci fa vedere i dati digitali e che nella nostra vita si sostituisce al mondo sensibile.

Ora che sempre più spesso i dati analogici sono sostituiti dai dati digitali, e il mondo digitale si sta espandendo progressivamente a scapito del mondo analogico, il significato dei dati digitali va cercato in ***quel*** *mondo. Ed è in* ***quel*** *mondo, non più nel nostro mondo tradizionale, che dobbiamo capire come metterci d’accordo sul significato delle cose che stanno al di là della finestra.* Se volete approfondire questo punto, vi do appuntamento al Capitolo 8 e all’esempio del calcolo dell’Rt della Regione Lombardia nel gennaio 2021.

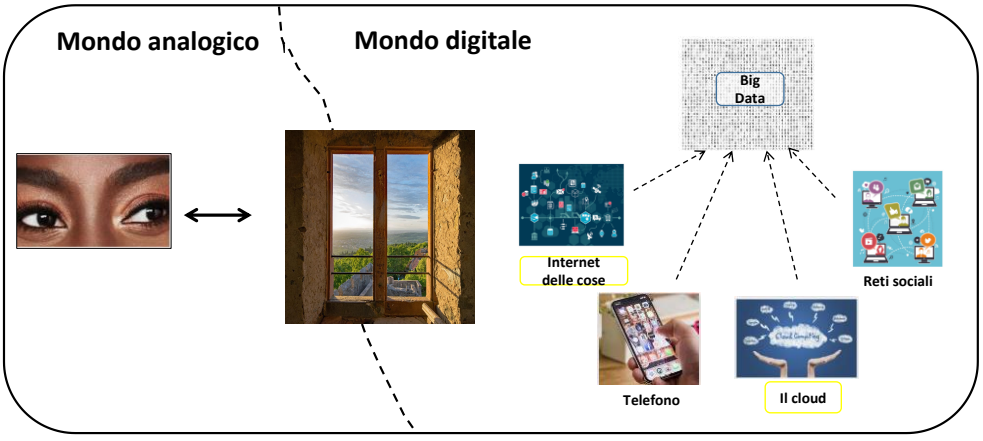


Figura 31 – I dati digitali sono una finestra sul mondo

Accade talvolta che gli occhiali digitali con cui osserviamo il mondo provochino *distorsioni* nella nostra vista sul mondo, analoghe a quelle del normale processo di invecchiamento degli occhi, come la *miopia*, il vedere male da lontano, o la *presbiopia*, il vedere male da vicino, vedi Figura 31.

Presbiopia

Miopia

Figura 32 - Gli occhiali sul mondo possono portare distorsioni

Cerchiamo di capire come devono essere fatti questi occhiali, così da acquistare consapevolezza e vedere correttamente il mondo.

Riassumendo

Gli esseri umani hanno sempre convissuto con i ***dati analogici***, i dati che ci forniscono informazioni sul mondo. Un dato analogico è una codifica di fenomeno del mondo, una temperatura letta da un termometro o il semaforo rosso a un incrocio.
Con il diffondersi dei dati digitali, accanto al ***mondo*** dei dati analogici o ***analogico*** si sta sempre più espandendo il ***mondo digitale***. Sia nel mondo analogico che nel mondo digitale i dati hanno un ***significato*** che noi possiamo comprendere se abbiamo conoscenza del fenomeno rappresentato.
Quando un dato analogico viene trasformato in un dato digitale, il significato risiede insieme al dato nel mondo digitale. Il significato è influenzato e modificato dal percorso che il dato segue e dalle ***trasformazioni*** che il dato subisce.
La comprensione dei **dati digitali** e del loro **significato** sono ***entrambi*** rilevanti per utilizzare in modo ***consapevole*** e ***corretto*** i dati digitali.

Capitolo 4

Dai piccoli dati ai grandi dati

Da lungo tempo l'umanità ha usato i dati analogici per le proprie attività; in molte circostanze, i dati disponibili erano tutto ciò che serviva per agire e per decidere, insomma, perdonate la espressione colloquiale, *le persone se li facevano bastare*. Gli agricoltori sapevano che il sole sorge e tramonta in diversi momenti della giornata a seconda delle stagioni, e per sapere l'ora del giorno consultavano una meridiana posta sul muro di una chiesa o di un edificio del paese, vedi Figura 33.

Figura 33 – Una meridiana

Per molto tempo, i dati disponibili per affrontare un problema erano *pochi*. Pensiamo alla quantità enorme di dati che nel 2020 e nel periodo successivo sono stati raccolti e analizzati per comprendere e combattere il virus Covid-19 e alla ricerca che si è potuta sviluppare da allora; quando all'inizio del ventesimo secolo l'umanità fu colpita dalla epidemia nota come *spagnola*, si riuscì solo a produrre poche tavole riassuntive del fenomeno, perché le tecnologie del tempo, come il telefono o il telegrafo, non permettevano di raccogliere e trasmettere dati con adeguato livello di estensione e aggiornamento.

Nonostante la penuria di dati, però, si ottennero risultati scientifici con scoperte straordinarie, e in diverse circostanze si riuscì a risolvere problemi anche con i pochi dati che si potevano rilevare.

Nel 1854 Londra fu colpita da una epidemia di colera veramente devastante. Un medico, John Snow, si mise alla ricerca delle cause dell'epidemia, e a tal fine elaborò una piantina di Londra con la diffusione dei casi nei diversi periodi. Diamo adesso una versione delle scoperte di Snow ispirata a quanto compare nella voce di Wikipedia dedicata a Snow e sul sito *https://www.ph.ucla.edu/epi/snow/snowcricketarticle.html*.

Inizialmente Snow concentrò la sua attenzione su una strada dove c'erano state molte morti di colera, Broad Street. In Figura 34, che rappresenta una parte di Broad Street, ogni lineetta nera rappresenta un morto, ed è posta in corrispondenza della casa dove abitava la persona deceduta. Il cerchietto rosso rappresenta una pompa dell'acqua. Notate niente di particolare?

Figura 34 – Le prime osservazioni di John Snow sulla epidemia di colera del 1854

Non so voi, ma Snow notò che un alto numero di morti si verificava *in vicinanza di una pompa dell'acqua*. Naturalmente un solo dato non crea una relazione di causa ed effetto, ma permise di formulare una ipotesi, e cioè che ci fosse un legame tra la diffusione della epidemia e l'erogazione dell'acqua nel quartiere.

Snow si procurò mappe della rete delle strade del quartiere di Soho, dove era situata Broad Steeet. Si procurò i dati sulle persone decedute nella intera area, e i dati sulla collocazione delle pompe dell'acqua. E costruì una nuova mappa come quella mostrata nella parte sinistra della Figura 35.

Nine Surrey Districts of London.

Houses supplied by	Estimated Population, 1851.*	Deaths from Cholera. 8th July to 5th Aug.	6th Aug. to 26th Aug.	27th Aug. to end of Year.	Total.
(a) Lambeth Company	155,057	14	84	513	611
(b) Southwark Company	240,326	236	977	2,213	3,476
(c) Wells and other sources	106,122	84	119	1,283	1,436
(d) Total	511,435	334	1,180	4,009	5,523

	Deaths from Cholera to 10,000 living. First Stage.	Second Stage.	Third Stage.	All Stages of Epidemic.
Southwark water, drawn from Battersea and containing London sewage	115	392	888	1,395
Wells and other sources	33	112	1,189	1,333
Purer Lambeth water, drawn from Thames beyond sewage range	9	54	330	393

Figura 35 – Con pochi dati Snow capì la causa della epidemia di colera nel 1854 a Londra

Nella mappa è rappresentata con un cerchietto rosso la pompa che abbiamo già visto nella Figura 34, mentre è cerchiata in blu una pompa dell'acqua erogata da un'altra compagnia. Snow contò in modo sistematico le morti verificatesi nel tempo in edifici vicini alle pompe delle diverse compagnie, ottenendo statistiche mostrate in Figura 35, in cui i morti per 100.000 abitanti che abitavano in case vicine alle pompe della compagnia Southwark erano decisamente superiori a quelli associabili alle altre compagnie. Era l'acqua erogata da Southwark a portare il colera.

Per lungo tempo, dunque, i dati a disposizione per agire, per diagnosticare, per decidere, per prevedere, sono stati pochi. Ai nostri giorni, con l'avvento delle grandi tecnologie mostrate in Figura 17, i dati digitalizzati a noi disponibili stanno crescendo ad un ritmo tumultuoso. Possiamo inquadrare questa espansione continua dei dati secondo uno spazio a tra dimensioni, mostrato in Figura 36, in cui abbiamo collocato due degli esempi descritti in precedenza. Le dimensioni riguardano:

1. *L'ampiezza* nella rappresentazione della realtà osservata, intesa come la vastità spaziale dei fenomeni osservati; nel nostro caso il giudice sta interpretando la scheda elettorale di *un elettore* residente in Florida, e la cartina fotografa la situazione dei decessi in *una strada di Londra*.
2. La *profondità* nella rappresentazione della realtà osservata, intesa come vastità delle *caratteristiche* osservate dei fenomeni; per caratteristica intendiamo i tipi di dati con cui descriviamo i fenomeni. Nel nostro caso, oltre a osservare il *voto* (prima caratteristica delle persone votanti) e registrare lo *stato* (seconda caratteristica) in cui è stato espresso il voto, potremmo conservare anche la *contea* (terza caratteristica) e magari correlare i voti con le caratteristiche sociali della popolazione nelle varie contee. Riguardo a Snow, oltre che i decessi e le pompe di benzina, fu decisivo che Snow acquisisse anche le informazioni sulle compagnie che erogavano acqua dalle diverse pompe.

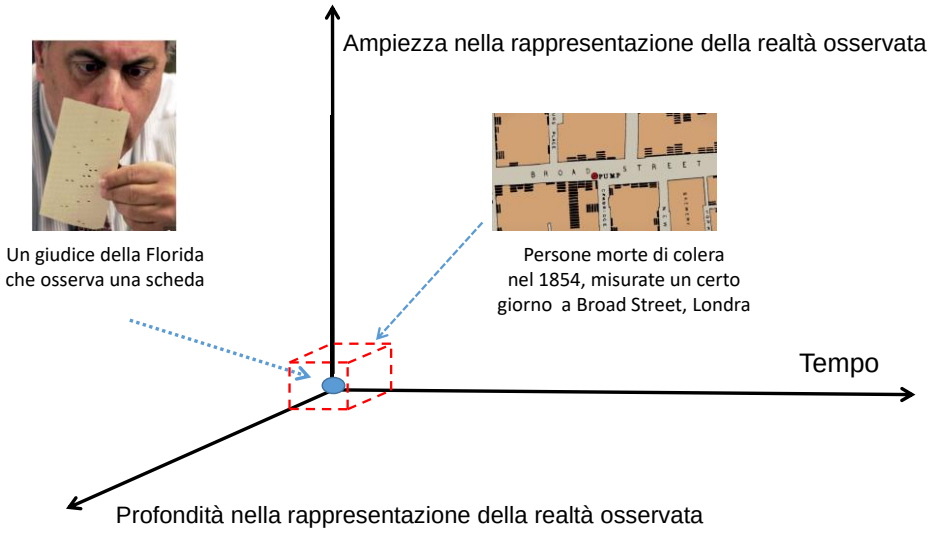


Figura 36 – Lo spazio dei dati

3. il tempo in cui i dati vengono generati; nel nostro caso, i voti scrutinati arrivarono nel corso di diversi giorni, e Snow estese i suoi conteggi per tutto il periodo della epidemia.

Ampiezza della rappresentazione, *profondità* della rappresentazione e *tempo* sono le tre coordinate che possono essere associate ai dati digitali, e secondo cui si espandono i dati digitali. Queste tre coordinate sono così importanti che per capirle bene vorrei fare altri esempi.

Riguardo alla *ampiezza della rappresentazione*, una delle più antiche mappe del mondo (vedi Figura 37) è quella di Ecateo ed è databile al 520 avanti Cristo. Il centro della mappa è il Mediterraneo, e dei paesi affacciati al grande mare vengono rappresentati pochi caratteri territoriali, la forma della costa, alcuni fiumi, alcune catene montuose.

Ai giorni nostri, i satelliti Landsat permettono di rappresentare aree della superficie terrestre di 30 metri per 30 metri, fornendo quindi rappresentazioni molto più grandi in *ampiezza*.

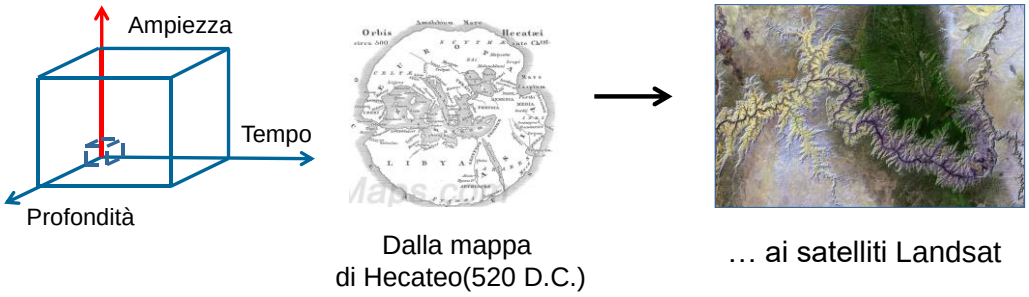


Figura 37 – Le mappe del mondo

Questo permette di effettuare il monitoraggio ambientale del territorio, elaborare modelli metereologici molto più precisi che nel passato, monitorare tornadi e inondazioni, pianificare l'uso dei terreni per le colture.

Riguardo alla *profondità della rappresentazione*, la tecnologia dell'Internet delle cose sta portando alla installazione di miliardi di sensori su dispositivi, apparecchiature, impianti e sistemi, materiali e prodotti tangibili, infrastrutture e beni, macchine e attrezzature. Gli oggetti fisici possono essere, a titolo di esempio, le catene produttive, gli elettrodomestici, le automobili, i pneumatici (vedi Figura 38), le scarpe da corsa, i palloni nel gioco del calcio e le racchette nel gioco del tennis, e così via. Le funzioni rese possibili sono la identificazione, la connessione, la localizzazione, in generale, la capacità di interagire con l'ambiente esterno.

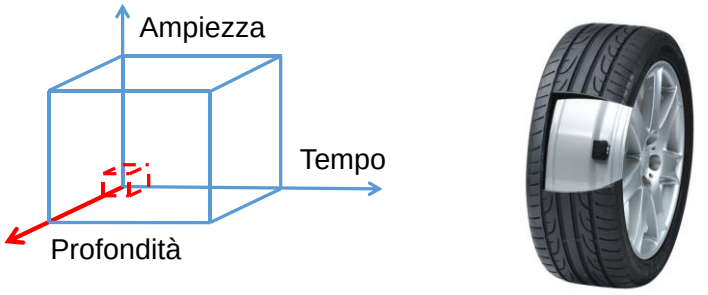


Figura 38 – I pneumatici intelligenti

Quanto alla *evoluzione nel tempo*, i dati prodotti da tutte le precedenti tecnologie possono essere seguiti nel tempo per fornire informazioni sui cambiamenti nella vita e stato di salute delle persone, nel territorio, negli oggetti, nei materiali. Se un sensore posto in un pneumatico fornisce un dato istantaneo sulla usura e sulla pressione e aderenza al terreno dei penumatici, il dato nel tempo permette di identificare il momento in cui conviene cambiare pneumatico o montare i penumatici da neve.

Nella Figura 39 compaiono due esempi di dati che cambiano nel tempo relativi al territorio e al traffico aereo. Nel primo, basato su Google Earth, vediamo che è possibile seguire nel tempo il processo di urbanizzazione di Dubai, effettuando una foto del territorio, ad esempio, una volta l'anno; nel secondo, basato su Flight Radar[6], possiamo seguire la evoluzione della

[6] https://www.flightradar24.com/45.47,9.19/8

posizione degli aerei nelle piste di decollo e di atterraggio di un aeroporto, facilitando il controllo delle distanze di sicurezza da parte dei controllori di volo.

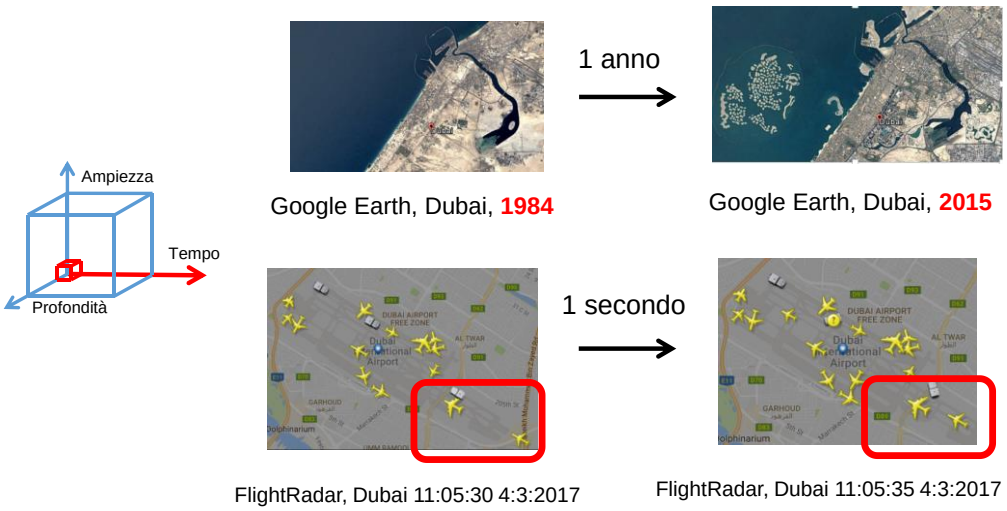


Figura 39 – Evoluzione nel tempo

L'esempio di Figura 40 riassume le tre coordinate. L'esempio fa riferimento al genoma umano; fIno a pochi anni fa le informazioni sulla salute e le patologie degli esseri umani si riferivano a esami di laboratorio o esami specialistici come la glicemia, le radiografie, le risonanze magnetiche.

E' del 2003 l'annuncio del sequenziamento del genoma umano; da allora, i costi connessi con il sequenziamento sono scesi di diversi ordini di grandezza (cioè, potenze di 10), permettendo così di estendere il sequenziamento a genomi di centinaia di migliaia di persone (percorriamo qui l'asse relativo alla *ampiezza della realtà osservata*).

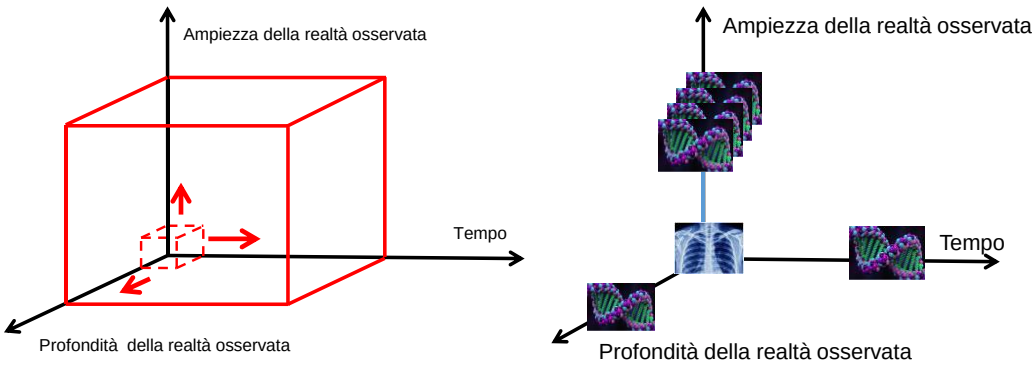


Figura 40 – Dalle radiografie al genoma umano

Ciò estende in maniera fondamentale rispetto agli esami tradizionali i dati disponibili e le scoperte scientifiche sulla natura del corpo umano e sul suo funzionamento (*profondità della realtà osservata*), e, inoltre, permette di conoscere e analizzare nel tempo le *mutazioni*, tematica che tanta importanza ha nel governo della epidemia Covid-19.

La disponibilità di grandi quantità di dati, l'aumento della potenza di calcolo dei computer e il perfezionamento delle tecniche di apprendimento automatico (tema che sarà discusso nel volume nel Capitolo 10) hanno reso recentemente (dicembre 2020) possibile da parte della azienda Alphamind nuovi fondamentali risultati in tema di ripiegamento delle proteine.

Il *ripiegamento delle proteine* è il processo fisico mediante il quale una catena proteica costruisce in modo rapido e riproducibile la sua struttura tridimensionale nativa, una conformazione che è biologicamente funzionale, cioè risponde a uno scopo. È quindi il processo fisico mediante il quale una catena di aminoacidi si ripiega nella sua caratteristica e funzionale struttura tridimensionale da una bobina casuale. Questo apre grandi prospettive per la cosiddetta *medicina di precisione*, che studia la posibilità di immaginare cure specifiche per le singole persone.

Vediamo in Figura 41 i miglioramenti che sono intervenuti nella accuratezza del processo di ripiegamento prodotto dalle tecniche sviluppate negli anni, intendendo per *accuratezza* la identificazione esatta della forma del ripiegamento.

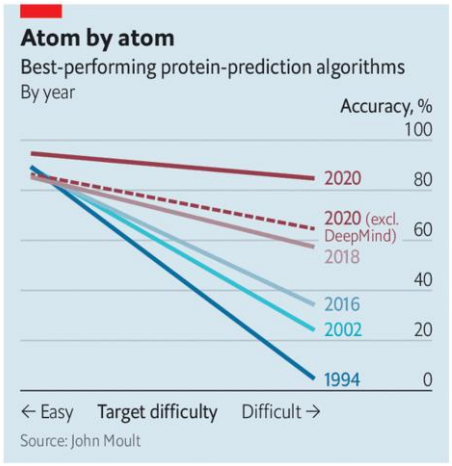


Figura 41 – La tecnica di Alphafold, da
https://www.economist.com/science-and-technology/2020/11/30/how-do-proteins-fold

Infine, è bene ricordare che i *grandi dati* fanno riferimento a tutti gli oggetti dell'universo, dall'atomo all'universo delle galassie, vedi Figura 42.

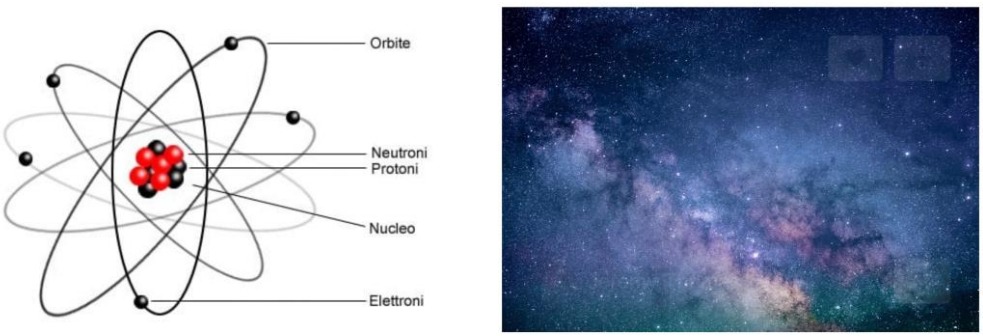


Figura 42 - Dall'Atomo all'Universo
https://www.biopills.net/la-struttura-atomo/

Una particella subatomica è stimata avere la dimensione di 10^{-35} metri, mentre una galassia ha la dimensione di 10^{24} metri, con una differenza di circa 60 ordini di grandezza (lo ricordo, potenze di 10); comprendiamo dunque che con i big data non potrà mai essere raggiunto il sogno dei geografi babilonesi di costruire una mappa in scala 1 a 1 dell'Impero babilonese, sogno descritto nel testo di Jorge Luis Borges, *Storia universale dell'infamia* (Il Saggiatore, 1961 traduzione di Mario Pasi)!

Riassumendo

L'umanità ha sempre dovuto risolvere ***problemi*** per ***prendere decisioni***, per ***prevedere il futuro***, per **comprendere le cause dei fenomeni naturali e sociali**, usando ***dati***, prima ***analogici*** e poi ***anche digitali***.
Molti problemi possono essere affrontati solo avendo a disposizione tanti dati digitali, perché è necessaria una conoscenza che riguardi tanti fenomeni (***ampiezza***), tante proprietà dei fenomeni (***profondità***) e l'evoluzione dei fenomeni nel **tempo**, le tre **dimensioni** dei big data. I fenomeni che si possono rappresentare con i dati digitali vanno dalle ***particelle subatomiche*** all' ***universo***.

Capitolo 5

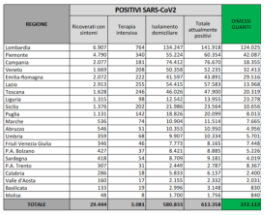
Le diverse forme che assumono i dati digitali: Le tabelle, i testi, le immagini, gli odori……

Dagli esempi che hai fatto, i dati possono assumere tante forme diverse, tabelle, mappe, immagini. Perché non mi parli di queste forme?

Hai ragione, è proprio il momento di capire in questo testo la natura profonda dei dati, le diverse *forme* che i dati possono assumere.

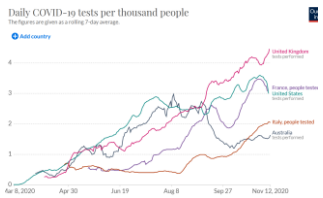
Guarda la Figura 43, in cui ho rappresentato diversi insiemi di dati usati per fornire informazioni utili sulla evoluzione della epidemia Covid.

Abbiamo già visto in precedenza le tabelle; in questo caso (Figura 43) la tabella fornisce dati giornalieri per le diverse regioni sui nuovi positivi e su altre tipologie di malati Covid. La tabella, è forse la forma di dati più usata, rappresenta i dati per mezzo di una struttura formata da righe e colonne; in conseguenza di questa sua struttura, la tabella facilita le ricerche delle informazioni a cui siamo interessati (per esempio, vogliamo individuare la regione con il maggior numero di nuovi positivi, ovvero le regioni che hanno un numero di pazienti ricoverati superiore a una data soglia).



Tabella

Mappa



Grafico

Ibrido testo/tabella

Dieci Regioni/PA sono ancora classificate a rischio alto o a esso equiparate, di queste, 9 sono state classificate a rischio Alto e/o equiparate a rischio Alto per 3 o più settimane consecutive.

Questo andamento non deve portare a un rilassamento prematuro delle misure o a un abbassamento dell'attenzione nei comportamenti.

Si conferma la necessità di mantenere la drastica riduzione delle interazioni fisiche tra le persone. È fondamentale che la popolazione eviti tutte le occasioni di contatto con persone al di fuori del proprio nucleo abitativo che non siano strettamente necessarie e di rimanere a casa il più possibile. Rimane essenziale evitare gli eventi aggregativi che, se effettuati, porteranno a un rapido aumento nel numero di nuovi casi.

Testo

Video

Figura 43 - In quanti modi si possono comunicare dati utili per la pandemia Covid

La tabella presenta la importante proprietà di esprimere esplicitamente il significato dei valori rappresentati, perché a ciascuno di essi associa, in questo caso, un *nome nella colonna* in cui compare il dato e un *nome nella riga*. Altre tabelle, che vedremo più approfonditamente nel Capitolo 6, hanno solo nomi sulle colonne, e non hanno nomi sulle righe.

Le *mappe* forniscono una descrizione di una porzione di territorio, e usano vari simbolismi per esprimere proprietà delle diverse parti del territorio rappresentate, in questo caso la mappa rappresenta l'Italia con vari colori per le regioni. I *grafici* rappresentano mediante linee, superfici e simboli di varia natura funzioni matematiche o statistiche. Mappe e grafici verranno discussi più approfonditamente nel Capitolo 10 dedicato alle visualizzazioni.

I *testi* sono composti da frasi in una lingua naturale, nel nostro caso l'italiano. Sono dotati di una struttura, ad esempio i punti separano le frasi e gli "a capo" caratterizzano periodi, ciascuno con un senso compiuto. La struttura dei testi definita dalle regole sintattiche e dai simboli di interpunzione è molto più *labile* di quella delle tabelle, dove, come abbiamo detto, ogni dato è in una ben precisa posizione, con un significato espresso esplicitamente dal nome della riga e della colonna associata. Torneremo tra poco su questo aspetto.

I *video* sono sequenze di immagini; i *dati ibridi* nascono dalla composizione di dati di diverse tra le precedenti tipologie, nel nostro caso testo e tabella.

Le possibili forme dei dati possono essere classificate a seconda del *senso* che usiamo per percepirli, come si vede in Figura 44. Questa classificazione, basata sui sensi, riguarda evidentemente i dati analogici, ma possiamo adottarla anche per i dati digitali. Essa estende le forme viste nella Figura 43.

Dati visti
- Tabelle
- Testi in linguaggio naturale
- Documenti dotati di una struttura (es. leggi, tesine, ecc.)
- Documenti esito di una scannerizzazione da documenti cartacei
- Segnali (ad esempio, la temperatura misurata con un termometro)
- Grafici
- Diagrammi
- Mappe
- Immagini fisse (es. Fotografie)
- Video

Dati ascoltati
- Audio

Dati olfattivi
- Odori

Dati tattili
- Testi tattili (es. In alfabeto Braille)
- Mappe tattili

- Dati sul Web
  - Dati collegati
- Dati ibridi

Figura 44 – Le diverse forme dei dati

E’ interessante osservare la Figura 44. Tra i *dati visti,* vediamo la distinzione tra diversi tipi di testi; i *testi in linguaggio naturale* sono testi in senso generico, fatti di lettere, parole, frasi, periodi, segni di interpunzione. Usano un lessico, rispettano una sintassi per la composizione delle frasi, e hanno un significato che però, come abbiamo già osservato, non è immediatamente percepibile come nelle tabelle.

I *documenti*, come ad esempio le leggi, sono dati organizzati in forma chiamata *semistrutturata*, in quanto prevedono, rimanendo nell’esempio delle leggi, un preambolo o premessa, gli articoli, i commi; i documenti sottoposti a scannerizzazione sono un po' un ibrido tra documenti cartacei e documenti digitali, perché, sottoponendoli ad applicazioni di riconoscimento dei caratteri e delle frasi, non sempre queste applicazioni sono in grado di ricostruire il testo originario.

Ancora tra i *dati visti*, i segnali sono tutto quanto misuriamo del mondo fisico, abbiamo visto l’esempio della temperatura corporea. Grafici, diagrammi e mappe saranno discussi nel Capitolo 10 sulle visualizzazioni. Le *immagini* e i *video* sono parte della nostra esperienza quotidiana; come e più dei testi, il significato delle immagini e dei video è complesso da descrivere anche da noi esseri umani, perché sono usati per catturare momenti particolari, raccontare storie ed esprimere emozioni. E’ peraltro importante poterli analizzare automaticamente, ad esempio, in un sistema di sorveglianza.

I *dati ascoltati* sono usati nei telefoni mobili o altri strumenti di riproduzione. Riguardo ai *dati olfattivi*, forse vi siete meravigliati quando li avete visti! Ma non c’è da meravigliarsi; così come i colori, che si possono ottenere mediante combinazioni di colori base, anche gli odori possono essere scomposti in odori di base, che, opportunamente codificati, trasformano gli odori in dati digitali.

I *dati tattili* sono utili per ipovedenti e non vedenti, e sono usati a questo scopo per rappresentare e rendere accessibili altre tipologie di dati (ad esempio, per mezzo dell’alfabeto Braille posso rappresentare testi).

I *dati collegati (o linked data)* sono un tipo di dati relativamente recente, e sono usati nel Web per collegare dati prodotti da diverse persone o organizzazioni; infine i *dati ibridi* nascono dall’uso congiunto delle precedenti forme.

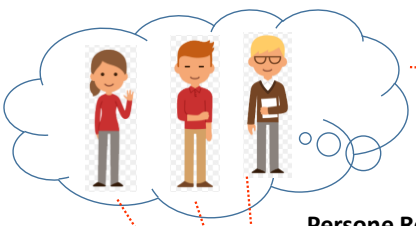
Dati ibridi? Mi stai dicendo, quindi che potrei usare per un qualche scopo sia il profumo dei fiori che la loro immagine?

Certamente! E’ quello che noi esseri umani facciamo spesso quando usiamo sensi diversi per avere una conoscenza complessiva del mondo attorno a noi…

Le tabelle e i testi

Confrontiamo ora più da vicino le tabelle e i testi riguardo al modo con cui rappresentano un frammento di mondo, e su come possiamo usarli per le nostre esigenze.

Guardate la Figura 45. Nella figura sono rappresentate mediante un testo e una tabella alcune proprietà di tre persone residenti a Milano: il cognome, il nome, ecc. Nel caso del testo, queste proprietà sono descritte da parole inserite in frasi; ad esempio la città di nascita di Carlo Bini è descritta dalla frase "è nato a Pescara". Nel caso della tabella, queste proprietà, come abbiamo visto, sono descritte nelle celle, e il loro significato compare espresso all'inizio della relativa colonna, nel nostro caso *Comune di nascita.*



Vogliamo rappresentare dati riguardanti tre persone residenti a Milano. La prima è Carlo Bini, nato a Pescara il 7 giugno del 1949, Residente in via Rossini 18, Codice Avviamento Postale (CAP nel seguito) 20122. La seconda è Giulio Verdi, nato a Roma il 9 luglio 1999, e residente anche lui in via Rossini 18, CAP 20122, la terza è Laura Rossi, nata a Milano il 20 dicembre 1996 e residente in via Verdi 31, CAP 20121.

Persone Residenti a Milano

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 45 – Un testo e una tabella

Prova ora a fare un esperimento. Munisciti di un cronometro o di un orologio con i secondi, e misura il tempo che ci metti a rispondere alle seguenti domande osservando il testo e la tabella:

1. Chi abita in Via Rossini 18?
2. Quante persone sono nate dopo l'anno 1995?

Per non favorire nessuna delle due rappresentazioni tra testo e tabella, inizia dal testo per la prima domanda e dalla tabella per la seconda, e ….non barare!

Credo che possiamo convenire sul fatto che la forma strutturata che hanno i dati nella tabella porta ad un minor tempo necessario per rispondere alle due domande. Sei d'accordo?

Sono d'accordo!

Ora conduciamo un secondo esperimento. Mentre in precedenza testo e tabella contenevano la stessa informazione, ma in formati diversi, in Figura 46 testo e tabella rappresentano due realtà che hanno una parte in comune, ma soltanto una parte; per rispondere alla domanda di Figura 46 dobbiamo *mettere insieme* dati contenuti sia nel testo che nella tabella. Prova a rispondere alla domanda e osserva bene le operazioni mentali che effettui.

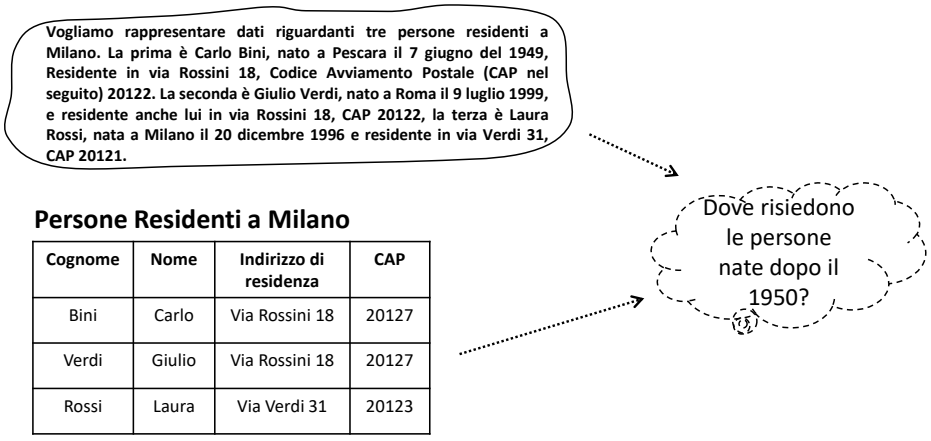


Figura 46 – Come rispondere a una domanda avendo a disposizione un testo e una tabella

Ho pensato che prima devo capire leggendo il testo chi è nato dopo il 1950, e poi guardare nella tabella le righe associate a queste persone, arrivando infine all'indirizzo. E' giusto? La risposta giusta è: Via Rossini 18 e Via Verdi 21?

Certamente, è tutto giusto! Cerchiamo ora di analizzare la tua strategia al rallentatore; hai dovuto:

1. prima capire chi erano le persone,
2. poi quali erano le loro date di nascita, *estraendo questi dati* dal testo.
3. poi hai dovuto selezionare quelli nati prima del 1950,
4. infine hai dovuto *collegare* i nomi e cognomi nel testo con quelli nella tabella, arrivando alla fine a trovare il loro indirizzo.

Perciò in sintesi: hai dovuto *riconoscere* ed *estrarre* dal testo le persone, *selezionare* quelle che rispondono alla condizione sull'anno, *collegare* questi dati con le persone nella tabella, *selezionare* gli indirizzi. In sintesi:

Riconoscere + estrarre + selezionare nel testo + collegare + selezionare nella tabella

L'operazione di *riconoscere + estrarre*, che noi umani facciamo in modo quasi naturale, è molto difficile per i computer, ed è oggetto di un'area di ricerca in grande sviluppo, quella della *Elaborazione del linguaggio naturale*, o *Natural language processing*, costituita da un insieme di tecniche che cercano di estrarre significato dai testi, facenti parte della disciplina chiamata *Intelligenza artificiale*.

L'elaborazione del linguaggio naturale mi sembra un'area molto importante per il futuro, spero di ritrovarla nella enciclopedia. L'elaborazione del linguaggio naturale si usa negli assistenti vocali che rispondono talvolta alle nostre telefonate?

Certo, in quel caso è necessaria una elaborazione sul linguaggio parlato, che è ancora più complicato da interpretare del linguaggio scritto, ma il principio è simile.

L'elaborazione del linguaggio naturale si sta sviluppando secondo due direttrici di ricerca che vorrei ora accennare. Supponiamo di voler tradurre dall'italiona in inglese.

Una *prima direzione* riconosce prima le parole, scompone la frase da tradurre nelle parti costituenti, assegna a ogni parola la sua forma sintattica (verbo, aggettivo, nome, ecc.) e poi, confrontando le parole con i loro diversi significati, cerca di individuare il significato giusto nel contesto della frase, per tradurre correttamente la frase nella corrispondente frase inglese.

Si sta sviluppando una *seconda direzione*, nell'ambito dell'apprendimento automatico o machine learning. Se provi a usare Google Translator[7], scoprirai che riesce a tradurre un buon numero di frasi da una lingua naturale a un'altra per un vasto insieme di lingue naturali, per esempio dall'italiano all'inglese, e viceversa.

Ebbene Google Translator e diversi altri traduttori si basano su un meccanismo analogo a quello che permise di interpretare il linguaggio dei geroglifici usando la stele di Rosetta, mostrata in Figura 47.

Nella stele di Rosetta lo stesso testo è scritto in tre scritture: il geroglifico egiziano, il greco e l'egizio. Poiché si tratta sostanzialmente dello stesso testo in tre lingue diverse, gli archeologi e i linguisti hanno potuto individuare le corrispondenze tra simboli, ad esempio tra geroglifici e greco, arrivando alla comprensione dell'egizio.

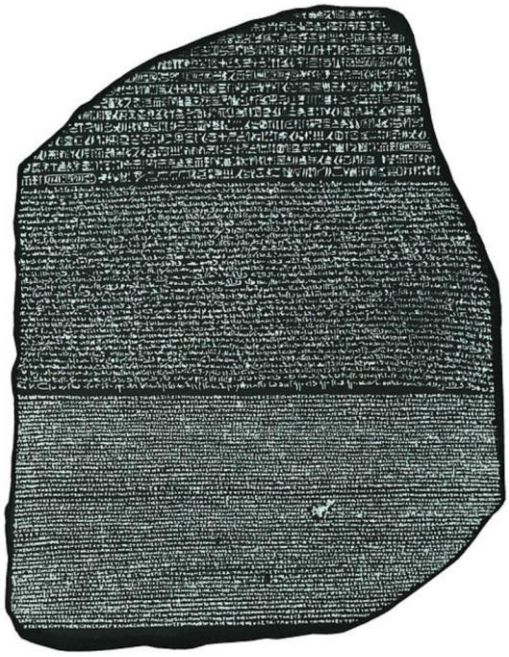


Figura 47 – La stele di Rosetta

Più o meno lo stesso meccanismo è adottato da Google Translator, con la diversità che ora a imparare a tradurre è un algoritmo di apprendimento automatico, che viene istruito a far questo sottoponendogli testi, ad esempio, in italiano e in inglese, tratti dal grandissimo patrimonio di testi tradotti dagli utenti; pensate, per esempio, ai documenti della Unione Europea che devono essere tradotti, in alcuni casi, nelle lingue di tutti i paesi della Unione.

[7] https://translate.google.it/

Naturalmente, l'algoritmo di Google può sbagliare, così come sbagliano o non capiscono gli assistenti vocali, ma questo accade per il fatto che siamo ancora nella infanzia della ricerca su questi temi. Di Machine Learning (apprendimento automatico) parleremo diffusamente nel Capitolo 11.

Riassumendo

I dati digitali, come anche i dati analogici, per cercare di rappresentare il mondo possono assumere tante **forme** diverse, dalle ***tabelle*** ai ***grafici*** ai ***diagrammi***, le ***mappe***, le ***immagini***, i ***video***. Queste diverse forme servono per rappresentare mediante dati digitali ciò che percepiamo con i sensi della ***vista***, ***dell'udito***, ***l'olfatto***, il ***tatto***. Vi sono anche altri tipi di dati nati con l'avvento del Web, che permettono di collegare tra di loro dati su siti diversi.
Le tabelle e i testi sono i dati più utilizzati nelle organizzazioni, le immagini, i video e i testi sono utilizzati nei telefoni mobili e nelle reti sociali.

Le ***tabelle*** sono composte di ***righe*** e ***colonne*** dove vengono rappresentati i ***valori*** dei dati, e per questa caratteristica sono chiamati **dati strutturati**, i ***testi*** sono costituito da ***frasi in una lingua naturale***, e in genere hanno una struttura meno strutturata rispetto alle tabelle, chiamata anche ***debolmente strutturata***, costituita da capitoli, sezioni, articoli, commi, a seconda della natura del testo, ad esempio una legge o un articolo di ricerca o un documento amministrativo o un rapporto.

Capitolo 6

I modelli dei dati sono gli occhiali con cui possiamo dare un significato al mondo [8]

Se avete letto un romanzo russo del diciannovesimo secolo (vedi in Figura 48 alcuni romanzi russi conservati in uno scaffale di libreria), sapete che una delle maggiori difficoltà nella lettura consiste nel ricordare i nomi dei personaggi. Infatti, in Russia il nome completo di una persona è composto dal nome di battesimo, dal patronimico che deriva dal nome del padre e dal nome della famiglia. Ad esempio., in *Guerra e Pace* di Tolstoj il nome completo di uno dei personaggi principali è Conte Pëtr *Pierre* Kirillovič Bezuchov.

A complicare le cose, nei romanzi russi i personaggi sono di volta in volta citati con il solo nome di battesimo, con il nome e il patronimico, con il nome della famiglia o con il diminutivo.

Figura 48 - Romanzi russi in una libreria

Confermo. Ricordo quando lessi Anna Karenina di Tolstoj: Il fratello di Anna è il Principe Stepan Arkadyevich Oblonsky, a volte viene chiamato Stepan, altre Oblonsky, altre ancora Stiva, che è il diminutivo in russo di Stepan.

Infatti. Noi non abbiamo questa abitudine, e la distanza tra la lingua russa e l'italiano rende ancora più difficile ricordare il nome completo dei personaggi, a volte molto numerosi. Sarebbe utile, quando leggiamo un romanzo russo, poter consultare facilmente un inventario dei personaggi, soprattutto per quanto riguarda quelli secondari, che compaiono occasionalmente durante il racconto.

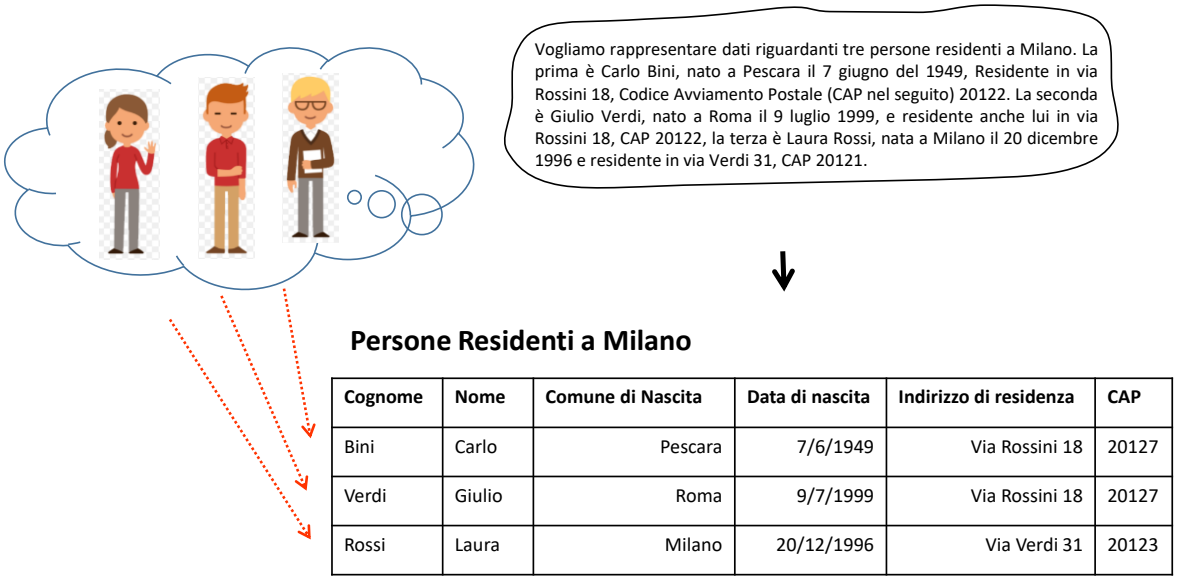
[8] Capitolo scritto con Gaetano Santucci

Credo che sarebbe utile anche disporre di un repertorio dei legami tra i personaggi: parentele, amicizie, ecc.

Certamente. Tutto ciò di può fare con un modello dei dati, insieme ad una sua rappresentazione grafica.

Cosa è un modello dei dati?

Guarda la Figura 49, che riproduce la Figura 45. Viene mostrata una tabella con un nome (*Persone residenti a Milano*), un insieme di sei colonne, che rappresentano proprietà delle persone e possiamo chiamare *attributi* della tabella (es. Cognome e Nome) e poi tre righe con i valori che assumono gli attributi (ad esempio il Cognome *Bini* e il Nome *Carlo*). Ebbene, tabelle e attributi sono le due *strutture di rappresentazione* usate dal *modello relazionale* dei dati, il modello più usato dai programmi informatici.



Persone Residenti a Milano

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 49 – Una tabella che rappresenta tre cittadini milanesi

Avrei bisogno, prima di continuare, di capire se il concetto ti è chiaro. Se nel caso della tabella di Figura 49 noi vogliamo rappresentare anche la città in cui risiedono le tre persone, cioè Milano, e poi vogliamo rappresentare *una quarta persona* chiamata Gerli Antonio, nato a Venezia il 7/4/1990, residente a Padova in Via Dolomiti 23, Cap 35123, come facciamo? La risposta nella prossima pagina.

La mia soluzione in Figura 50.

Persone Residenti in Italia

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	Comune	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	Milano	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	Milano	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	Milano	20123
Gerli	Antonio	Venezia	7/4/1990	Via Dolomiti 23	Padova	35123

Figura 50 – Risposta alla domanda

Ero arrivato alla conclusione che bisogna aggiungere una colonna, però l'avevo chiamata Città, non Comune; sulla nuova riga sono d'accordo; sul nome della tabella ero incerto, perché l'hai chiamata Persone Residenti in Italia? Quelle quattro non sono tutte le persone residenti in Italia!

Rispondo con ordine.

Nuova colonna – *Città* è un nome certamente lecito, io ho preferito Comune perché più preciso dal punto di vista del ruolo amministrativo che il Comune ha, e che non ha Città: infatti, il comune è una pubblica amministrazione, con Sindaco e Assessori, il comune detiene l'anagrafe dei residenti nel territorio del comune, Città è un po' generico…..

Nome della tabella – Certamente il nome va cambiato. Anche prima in fondo il nome della tabella era un po' strano, non è possibile che a Milano risiedano solo tre persone…. ma era l'unico nome possibile con la conoscenza a disposizione. Perciò questa volta o scegliamo il nome *Persone residenti a Milano e Padova*, che però non mi suona bene, oppure facciamo come prima, diamo un nome generico, che può essere *Persone residenti*, oppure *Persone residenti in Italia* ad indicare, come prima, *alcune persone residenti in Italia*.

Possiamo a questo punto condividere il concetto di modello di dati. Un *modello di dati* è un insieme di strutture di rappresentazione con cui è descritto un frammento di mondo. ad esempio, il modello relazionale è costituto da due strutture, le tabelle e gli attributi.

Fai attenzione; fissato un frammento di mondo e scelto un modello dei dati, finora conosciamo solo il modello relazionale, quel frammento si può descrivere in tanti modi diversi con le strutture del modello. Guarda per esempio la Figura 51; la tabella descrive lo stesso frammento di mondo della tabella nella Figura 50, ma ora, converrai, non si capisce niente! Ebbene, spesso nelle applicazioni informatiche le tabelle utilizzate sono descritte in modo più simile alla Figura 51 che alla Figura 50….

Tabella 1

Att1	Att2	Att3	Att4	IRes	Att6	CAP
Bini	Carlo	Pescara	7-6-1949	Via Rossini 18	Milano	20127
Verdi	Giulio	Roma	09071999	Via Rossini 18	Milano	20127
Rossi	Laura	Milano	201296	Via Verdi 31	Milano	20123
Gerli	Antonio	Venezia	7-4-990	Via Dolomiti 23	Padova	35123

Figura 51 – Un' altra tabella

Davvero le tabelle sono progettate in modo così sciatto?

Certo. E questo perché coloro che progettano le tabelle non sempre sono consapevoli del fatto che le tabelle e gli attributi devono esprimere in maniera chiara il *significato dei dati*, e quindi devono essere *comprensibili per tutti*, e non solo per loro. Mi raccomando,usa bene i modelli dei dati.

Bene, a questo punto torniamo ai romanzi russi. In questo caso, non usiamo il modello relazionale, useremo un nuovo modello, il *grafo semantico*, scopriremo presto il perché.

Ci ispiriamo ad semplice esempio tratto da *I Demoni* di Dostoevskij, un romanzo di quasi novecento pagine. Utilizziamo una brevissima descrizione dei principali personaggi e dei loro legami, vedi Figura 52.

Nella Russia del secondo '800 vive la nobile Varvara Petrovna Stavrogina, che è legata da profonda e platonica amicizia e mantiene economicamente lo scrittore e poeta incompreso Stepan Trofimovič Verchovenskij. Il figlio di Stepan, Petr Stepanovic Verchovenskiji, cresce lontano dal padre che è il tutore del figlio di Varvara, Nikolaj Vsevolodovic Stavrogin.

Entrambi i figli, cresciuti e dopo una lunga assenza all'estero, tornano a casa, ove tramano per compiere attentati. Nikolaj, detto anche Nicolas, è l'ideologo ispiratore di Petr, il quale coordina e comanda una "cinquina" di cospiratori composta da Virginsky, Sigalef, Liputin, Tolkacenko e Ljamsin.

Figura 52 – Descrizione sintetica dei principali personaggi dei Demoni, e dei legami tra essi

Anche se ho reso la più semplice possibile la precedente descrizione dei legami tra i personaggi, certamente a una prima lettura, si capisce ben poco…

Per non complicarci la vita con il primo esempio di grafo semantico, esaminiamo solo il primo paragrafo della descrizione, e procediamo per gradi. Identifichiamo subito i personaggi: Varvara, Stepan, Petr e Nikolaj. Sono soggetti rilevanti per il racconto e quindi sono oggetti del grafo semantico che rappresentiamo come *entità*; una entità è tutto ciò che ha esistenza autonoma, ad esempio una persona, un libro, un albero…. Rappresentiamo le entità per mezzo di nodi a forma di ellisse; qui i personaggi da rappresentare come entità sono quattro, e quindi abbiamo bisogno di quattro nodi, vedi Figura 53.

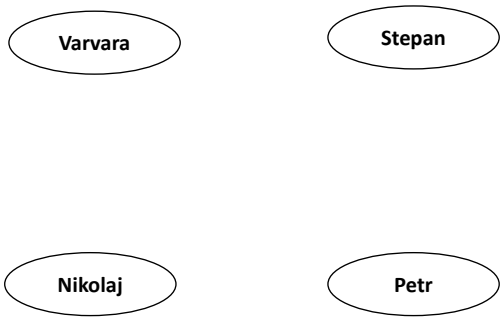


Figura 53 – Il grafo semantico dei Demoni: i quattro personaggi

Per capire quale parte del testo nella Figura 53 è stata fino ad ora rappresentata nel grafo semantico, rappresentiamo in Figura 54 le parti del testo corrispondenti in ***corsivo/neretto***.

*Nella Russia del secondo '800 vive la nobile **Varvara** Petrovna Stavrogina, che è legata da profonda e platonica amicizia e mantiene economicamente lo scrittore e poeta incompreso **Stepan** Trofimovič Verchovenskij. Il figlio di Stepan, **Petr** Stepanovic Verchovenskiji, cresce lontano dal padre che è il tutore del figlio di Varvara, **Nikolaj** Vsevolodovic Stavrogin.*

Entrambi i figli, cresciuti e dopo una lunga assenza all'estero, tornano a casa, ove tramano per compiere attentati. Nikolaj, detto anche Nicolas, è l'ideologo ispiratore di Petr, il quale coordina e comanda una "cinquina" di cospiratori composta da Virginsky, Sigalef, Liputin, Tolkacenko e Ljamsin.

Figura 54 – Descrizione sintetica con le parti del testo
rappresentate nel grafo in ***neretto/corsivo***

Consideriamo ora i *legami* tra i personaggi, intendo, ad esempio, che un personaggio è *figlio* di un altro personaggio; questi legami sono *relazioni* tra entità, che rappresentiamo mediante *archi* del grafo. Vedi in Figura 55 il nuovo grafo semantico e in Figura 56 il nuovo testo in neretto.

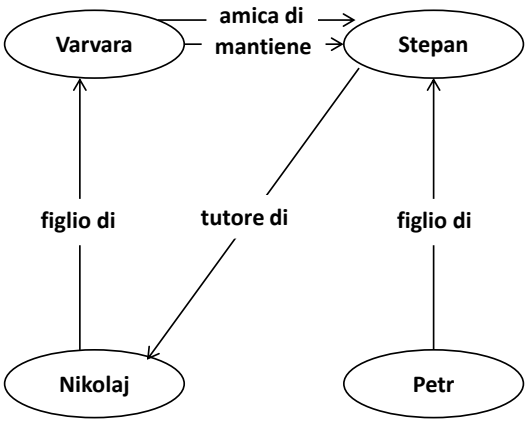


Figura 55 - La rete semantica dei Demoni: le relazioni tra i quattro personaggi

*Nella Russia del secondo '800 vive la nobile **Varvara** Petrovna Stavrogina, che è* **legata** ***da*** *profonda e platonica **amicizia** e **mantiene economicamente** lo scrittore e poeta incompreso **Stepan** Trofimovič Verchovenskij. Il **figlio di** Stepan, **Petr** Stepanovic Verchovenskiji, cresce lontano dal padre che è il **tutore** del **figlio di** Varvara, **Nikolaj** Vsevolodovic Stavrogin.*

Entrambi i figli, cresciuti e dopo una lunga assenza all'estero, tornano a casa, ove tramano per compiere attentati. Nikolaj, detto anche Nicolas, è l'ideologo ispiratore di Petr, il quale coordina e comanda una "cinquina" di cospiratori composta da Virginsky, Sigalef, Liputin, Tolkacenko e Ljamsin.

Figura 56 – Nuova descrizione con le parti del testo rappresentate nel grafo in ***neretto/corsivo***

Consideriamo infine le *proprietà* dei personaggi, quelle che nel modello relazionale sono gli attributi (es. Nome, Cognome); come abbiamo detto poco fa, il nome completo in russo è costituito dal nome di battesimo, dal patronimico e dal nome della famiglia (in effetti in russo anche il nome della famiglia si declina al maschile e al femminile). In un grafo semantico, anche le proprietà sono rappresentate con nodi del grafo, collegati con archi alle relative entità, come mostra la Figura 57.

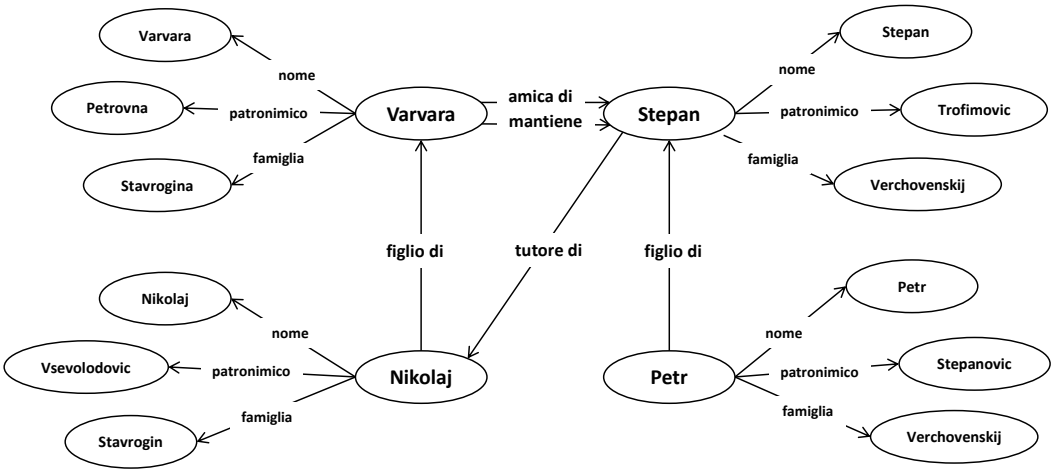


Figura 57 - La rete semantica dei Demoni: gli attributi dei quattro personaggi

Al solito, mostriamo le parti del testo che abbiamo rappresentato nel grafo semantico nel nuovo testo in Figura 58.

*Nella Russia del secondo '800 vive la nobile **Varvara Petrovna Stavrogina**, che è* **legata** ***da*** *profonda e platonica **amicizia** e **mantiene economicamente** lo scrittore e poeta incompreso **Stepan Trofimovič Verchovenskij**. Il **figlio di** Stepan, **Petr Stepanovic Verchovenskiji**, cresce lontano dal padre che è il tutore del **figlio di** Varvara, **Nikolaj** Vsevolodovic Stavrogin.*

Entrambi i figli, cresciuti e dopo una lunga assenza all'estero, tornano a casa, ove tramano per compiere attentati. Nikolaj, detto anche Nicolas, è l'ideologo ispiratore di Petr, il quale coordina e comanda una "cinquina" di cospiratori composta da Virginsky, Sigalef, Liputin, Tolkacenko e Ljamsin.

Figura 58 - Nuova descrizione con proprietà rappresentate in ***neretto/corsivo***

Molto chiaro e facile da leggere. Aiuta molto la rappresentazione grafica con nodi e archi. Come si può rappresentare la stessa informazione mediante tabelle?

Ottima domanda, che ci permette di fare un esercizio molto istruttivo, rappresentare lo stesso frammento di mondo con due modelli diversi. Procediamo un passo alla volta; prova prima tu a progettare la tabella dei Personaggi. Guarda prima il grafo semantico di Figura 57 e concentrati sulle tre proprietà di ogni Personaggio, quali sono?

Nome, Patronimico e Famiglia, è giusto?

Sì, è giusto! Adesso guarda la tabella di Figura 59, è composta di una prima riga in cui inserisco ho rappresentato le tre proprietà, e quattro colonne in cui devo inserire i dati relativi ai quattro personaggi. Ebbene, prova a riempire la tabella, la soluzione nella prossima pagina.

Personaggio

Nome	Patronimico	Famiglia

Figura 59 – Tabella dei Personaggi

Ecco la soluzione, non era difficile. A ogni personaggio è associata una riga; in ogni riga devo inserire i dati associati al personaggio, prima il nome, poi il patronimico e infine la famiglia.

Personaggio

Nome	Patronimico	Famiglia
Varvara	Petrovna	Stavrogina
Stepan	Trofimovič	Verchovenskij
Petr	Strpanovich	Verchovenskij
Nokolaj	Vsevolodovic	Verchovenskij

Figura 60 – La tabella dei Personaggi con i valori

Adesso procediamo con le relazioni tra i personaggi. Ti chiedo: quante e quali relazioni ci sono tra i personaggi?

Dunque, vediamo…le relazioni sono quattro: *amica di, figlio di, mantiene e tutore di*.

Giusto! Le quattro relazioni le possiamo rappresentare in due modi:

1. con quattro tabelle, ognuna associata a una relazione, oppure
2. con un'unica tabella in cui rappresento le cinque relazioni tra le coppie di personaggi, e per ciascuna coppia la relazione esistente tra i due. Guarda la Figura 61.

Amica di

Personaggio1	Personaggio2
Varvara	Stepan

Figlio di

Personaggio1	Personaggio2
Petr	Stepan
Nikolaj	Varvara

Mantiene

Personaggio1	Personaggio2
Varvara	Stepan

Tutore di

Personaggio1	Personaggio2
Stepan	Nikolaj

Soluzione 1

Relazioni tra personaggi

Personaggio 1	Relazione	Personaggio2
Varvara	Amica di	Stepan
Petr	Figlio di	Stepan
Nokolaj	Figlio di	Varvara
Varvara	Mantiene	Stepan
Stepan	Tutore di	Nokolaj

Soluzione 2

Figura 61 – Le due soluzioni per modellare le relazioni tra personaggi

Sono colpito dal fatto che la stessa frase, lo stesso pezzetto di mondo descritto dalla frase da cui siamo partiti, possa essere rappresentato in modi così diversi, nel modello a grafo semantico e nel modello relazionale con tabelle. E poi nello stesso modello, il modello relazionale, con due soluzioni così diverse…..

Hai ragione. Ti rimando al secondo libro della Enciclopedia sui modelli per continuare a ragionare su queste questioni, ma fin da ora abbiamo scoperto che i modelli dei dati sono occhiali che ci permettono di capire il mondo. Occhiali che possono essere molto diversi tra loro….

Mi cominciano a piacere, i dati….

Riassumendo

Per dare un ordine ai dati, e poterli usare, è opportuno rappresentarli per mezzo di un insieme di ***strutture***, che tutte insieme costituiscono un ***modello dei dati***, ad esempio la ***tabella*** e ***l'attributo*** per il ***modello relazionale*** dei dati. Un modello dei dati può essere visto come un **paio di occhiali** attraverso cui poter osservare **il mondo dei dati digitali**, che a sua volta permette di rappresentare il **mondo analogico**, sia pure solo in parte.
Anche i ***grafi semantici*** sono un modello dei dati, che, rispetto al modello relazionale, rappresenta il mondo per mezzo di ***entità*** e ***relazioni tra entità***; al contrario del modello relazionale, i grafi semantici hanno una ***rappresentazione grafica*** o ***mediante diagrammi,*** sia per le entità, i **nodi** del grafo semantico, sia per le **relazioni tra entità**, gli **archi** del grafo semantico.
Esistono dunque ***modelli di dati*** e loro ***rappresentazioni grafiche***, i due concetti vanno tenuti distinti.

Capitolo 7

I dati vanno rispettati
Occorre prendersi cura della qualità dei dati

Non capisco nessuna delle due parti del titolo, vengono usate per i dati espressioni ("i dati vanno rispettati", "prendersi cura…") che tipicamente si usano per esseri umani….

E' vero, il titolo di questo capitolo all'inizio appare un po' oscuro. Siamo arrivati a questo punto del libro, e ci siamo resi conto che l'espressione *dati* raccoglie in sé tante sfumature diverse…

Con la espressione *prendersi cura dei dati* intendo dire che i dati sono diventati così importanti nella nostra vita, per cui, proprio in virtù di questa importanza, li dobbiamo trattare "con i guanti", dobbiamo curarne le patologie, gli errori, tutto ciò che in questo capitolo chiameremo la loro *qualità*. Per cominciare a chiarire l'aspetto della qualità, inizio con un aneddoto.

Quando mi capita di lavorare, qualche volta, la domenica pomeriggio, per distrarmi guardo i risultati delle partite di calcio. Un pomeriggio del campionato 2012-13 vidi sul sito di Repubblica questa immagine riprodotta in Figura 62.

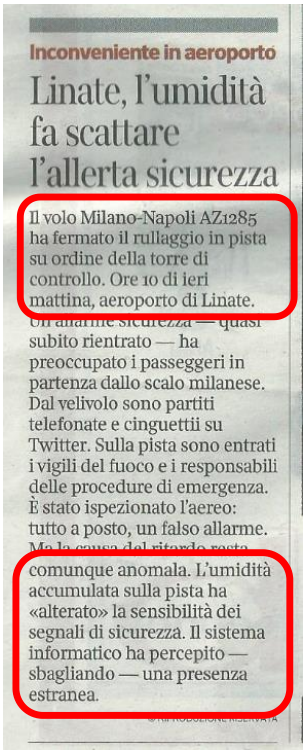
Figura 62 – Il risultato (meglio, i risultati) della partita Catania Inter nel campionato di calcio 2012-13

Nella stessa pagina sono riportati ben *tre* risultati diversi della partita! 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 minuto, e sembra che questo ultimo goal non sia conteggiato nella riga degli aggiornamenti corrispondente all'annuncio "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, i dati che rappresentano il risultato di Catania Inter nella pagina sono tre (2-2, 2-3. 2-4), sono tutti diversi uno dall'altro, e, naturalmente, solo uno è quello giusto!

Ci capita spessissimo di imbatterci in dati sbagliati. Per esempio, quando devo dire il mio cognome al telefono, e dico "Batini", spesso sento ripetere "Badini", "Battini", "Barini", e devo ripetere lettera per lettera, citandole come iniziali di nomi di città, Bologna, Ancona, ecc.

Nella Figura 63 mostro un ritaglio dal Corriere della Sera del 14 novembre del 2015, in cui si vede come basti un segnale rilevato in un aeroporto sensibile all'umidità, e che quindi fornisce un valore inaccurato, per far scattare una falsa allerta di sicurezza.



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.

Figura 63 – Esempi di scarsa qualità dei dati presi dai giornali

In ogni momento della nostra giornata rischiamo di incorrere in dati che non rappresentano esattamente il mondo attorno a noi (il risultato di una partita, un sensore di sicurezza, ecc.), e che manifestano degli errori. E' così importante questo aspetto della nostra relazione con i dati, che un importante immunologo, Alberto Mantovani, nel suo libro dedicato agli aspiranti

scienziati include tra le dieci norme del decalogo proposto, vedi Figura 64: rispetta i dati; a lui mi sono ispirato per il titolo di questo capitolo.

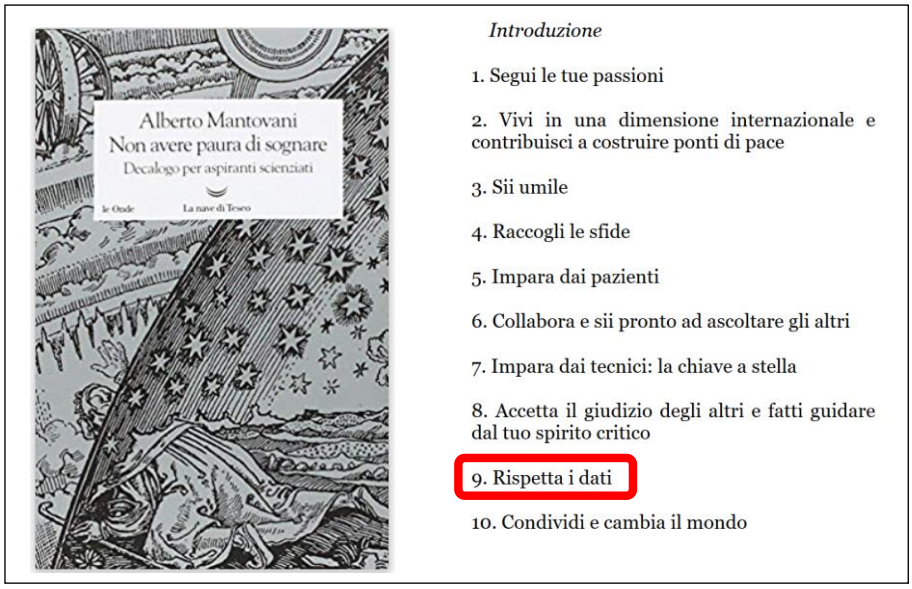


Figura 64 – Il decalogo di Alberto Mantovani

Se avete sentito Mantovani in qualche intervista, sapete che è persona molto competente nel suo campo, e allo stesso tempo molto umile nel riconoscere i limiti delle sue conoscenze.

Cominciamo a entrare nel merito del tema della qualità dei dati. La parola *qualità* è molto utilizzata nel linguaggio comune, nel marketing, nella erogazione di servizi. Quel prodotto per lavapiatti costa un po' di più, ma è di grande qualità; quella scuola è un pò distante da casa, ma ha professori di grande qualità; e così via. Ma cosa intendiamo per qualità? Riprendiamo la figura sulle viole del pensiero che abbiamo già utilizzato per descrivere come si possono codificare i dati digitali per mezzo delle cifre zero e uno, questa volta mostrando tre foto diverse, vedi Figura 65. Quale foto e relativa viola vi piace di più, e perché? Pensateci un attimo, le possibili risposte nella prossima pagina.

Viola 1 **Viola 2** **Viola 3**

Figura 65 – Tre foto di viole del pensiero

Riguardo alla prima domanda, *quale foto e viola vi piace di più*? penso che non ci siano dubbi, la foto di migliore qualità è Viola2. Certo, se un lettore o lettrice fosse un pittore divisionista, di quella scuola che tende a rappresentare le cose e i paesaggi con colori frammentati e discontinui, probabilmente sceglierebbe Viola1.

Pur essendo i pittori divisionisti molto pochi, anche loro hanno diritto di esprimere la propria opinione!

Perfetto! Dietro la tua osservazione si nasconde un mondo di questioni! *Anche i pittori divisionisti hanno diritto di esprimersi* significa che in certa misura la qualità è un fatto culturale, soggettivo. E certamente di questa soggettività dobbiamo tenere conto anche nel mondo dei dati. Tuttavia, questo non significa che possiamo fare riferimento a un senso comune, e rispetto al senso comune spero tu convenga che Viola2 è la foto migliore….

Sì certo….

Perché è la migliore? Proviamo a procedere per sottrazione, cercando di capire perché Viola1 e Viola3 non lo sono. Viola1 è frammentata, e quindi non è *fedele* ad una viola del pensiero reale. Qui vale la pena riprendere la questione dei pittori divisionisti; la pittura non deve riprodurre fedelmente la realtà, spesso è una *interpretazione* della realtà.

Viola3 è fedele nel senso che non ha discontinuità nella rappresentazione dei colori della viola, ma è troppo opaca, usa una gamma di intensità di colori ridotta rispetto a Viola2. Noi sappiamo che una viola in un giardino è più simile a Viola2 che a Viola3; e lo sappiamo perché esiste nella nostra mente un esempio, diciamo un ideale di viola, che abbiamo appreso nel tempo, simile a Viola2.

Finora ho capito, ma osservo che tutto rimane un po' vago, generico. In particolare, non stai valutando in termini assoluti la qualità, ma, piuttosto, stai confrontando immagini tra loro usando termini comparativi, "più di", "meno di". Non si potrebbero definire dei tipi di qualità, delle caratteristiche di qualità dei dati digitali, e misurarle?

Diciamo in generale che non tutto può essere espresso per mezzo di numeri; però, avendo tanto studiato il tema della qualità dei dati, ci provo.

La qualità può essere valutata attraverso un certo numero di *caratteristiche*, cioè in termini di diverse proprietà che esprimono un particolare aspetto di cosa intuitivamente intendiamo per qualità, e nel caso delle immagini di viola, qualità delle immagini.

La prima caratteristica che abbiamo visto è la *fedeltà all'originale*. Per poter definire e misurare la fedeltà all'originale, possiamo procurarci una immagine di riferimento della "viola perfetta" e calcolare pixel per pixel la distanza tra viola perfetta e ciascuna delle tre viole di Figura 65. Le cose non sono così semplici, ma mi interessa qui introdurre il metodo, non i suoi dettagli.

La fedeltà, però, non ci dice niente su quella che abbiamo chiamato rappresentazione frammentata di Viola1. Certamente, sulla base della definizione che ho dato, Viola1 è poco fedele all'originale, ma la fedeltà non rende bene un altro aspetto della qualità, che possiamo definire, in negativo, il *grado di frammentazione* della fotografia: dico *in negativo* perché foto altamente frammentate sono da considerarsi di scarsa qualità. La frammentazione può essere misurata sommando tutte le distanze tra pixel *adiacenti*, dove la distanza riguarda il colore e la sua intensità, e assegnando un voto maggiore alle fotografie in cui la somma delle distanze è minore, cioè, appunto, c'è meno discontinuità, meno frammentazione.

Ultima caratteristica per discriminare tra Viola2 e Viola3 è la *estensione del colore*, intesa come gamma di intensità di colori utilizzata. La *estensione del colore* ha come opposto la *opacità*. L'*estensione del colore* è misurabile come somma di tutte le intensità di colore dei singoli pixel rispetto ad una intenità base. Ad esempio, la estensione del colore di Viola3 è bassa, perché le intensità di colore dei pixel differiscono di poco, rispetto a Viola1 e Viola2.

Adesso cerchiamo di approfondire il tema delle caratteristiche di qualità, non solo per le immagini, ma anche per le altre forme dei dati. Ricorda le forme che possono assumere i dati, viste nel Capitolo 5; consideriamo nel seguito le tabelle, i testi, le mappe, le immagini e il Web.

Qualità delle tabelle

Osserviamo la tabella in Figura 66. Ogni riga della tabella rappresenta un film del secolo scorso; hai mai visto uno di questi film?

Veramente…no!

Va bene, in fondo è meglio così. Di ogni film la tabella rappresenta il titolo originale, il titolo italiano, il regista, l'anno in cui è uscito e la durata in minuti. Nella tabella vi sono diversi valori che non rispettano una delle due principali caratteristiche di qualità delle tabelle, la *accuratezza dei valori* e la *completezza dei valori*. Per *accuratezza* intendiamo la vicinanza al suo valore vero, per *completezza* intendiamo che tutti i dati sono specificati, e non ci sono "non lo so".

In Fiigura 66 sono evidenziati due errori, uno di *accuratezza* e l'altro di *completezza*; in particolare, è inaccurato il valore "Vacanze Romae", perché in italiano non esiste la parola "Romae", ed è incompleto il dato relativo al regista di Sabrina perché non c'è scritto niente.

Film del ventesimo secolo

Titolo originale del film	Titolo in italiano	Regista	Anno	Durata in minuti
Casablanca		Curtz	1983	102
Dead Poets Society	L'attimo fuggente	Weir	1989	199
Roman Holidays	Vacanze Romae	Wylder	1953	
Sabrina	Sabrina		1954	113

Inaccurato

Incompleto

Figura 66 – Una tabella con dati di scarsa qualità

Come possiamo correggere i due errori?

Per il primo, possiamo cercare un sito Web che elenchi i nomi italiani dei film, sperando trovare i nomi in ordine alfabetico. Questo sito esiste, ed ha come indirizzo https://www.film.it/film/film-a-z; magari ce ne sono altri, ma a noi basta questo.

Se consideriamo che la parola *Vacanze* è probabilmente corretta e supponiamo che "romae" sia il risultato di un errore di inserimento del titolo nella tabella, possiamo considerare i soli titoli nell'elenco che cominciamo con la parola *Vacanze*, vedi figura 67. Tra essi scegliamo quello decisamente più simile a *Vacanze Romae* che è *Vacanze Romane*.

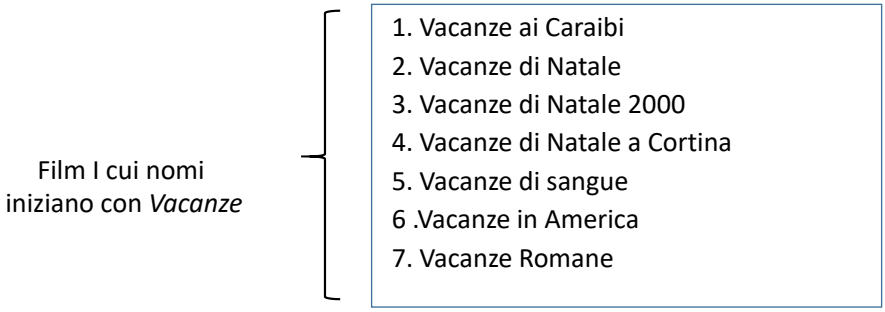


Figura 67 – Film i cui titoli iniziano con vacanze

Invece il valore incompleto può essere corretto accedendo alla voce Wikipeda del film, e cercando il regista. Un tempo di cercava nella Enciclopedia cartacea che avevamo in famiglia, ma difficilmente si poteva trovare una informazione così specifica, bisognava andare in blblioteca e sperare di trovare una enciclopedia specializzata; adesso basta un click!

Adesso è importante che tu svolga un ruolo attivo, le cose si capiscono solo se le si fa. Prova a trovare sul Web siti e informazioni che ti permettano di identificare tutti i restanti dati di scarsa qualità nella tabella di Figura 66, dove le caratteristiche di qualità da considerare sono la *accuratezza* e la *completezza*. Mi raccomando, per ora identifica gli errori ma non correggerli, procediamo un passo alla volta. La tabella con i dati errati e il tipo di errore è nella prossima pagina.

Vedi la tabella nella Figura 68, dove sono evidenziati in *grigio scuro* gli errori di accuratezza e in *grigio chiaro* gli errori di completezza.

Film del ventesimo secolo

Titolo originale del film	Titolo in italiano	Regista	Anno	Durata in minuti
Casablanca		Curtz	1983	102
Dead Poets Society	L'attimo fuggente	Weir	1989	199
Roman Holidays	Vacanze Romae	Wylder	1953	
Sabrina	Sabrina		1954	113

Inaccurato

Incompleto

Figura 68 – Tutti gli errori di accuratezza e completezza

Evidenziare gli errori di completezza è banale, basta non trovare un valore; per gli errori di accuratezza, dobbiamo verificare dato per dato la sua validità, cercando, ancora una volta, in Wikipedia.

Adesso correggiamo gli errori, cioè, a partire dai siti dove abbiamo trovato i dati di confronto, sostituiamo i dati di confronto ai dati inaccurati o incompleti nella tabella. Attenzione! Per la durata in minuti del film *L'attimo fuggente*, ti propongo di accedere ai due siti (prova fatta a inizio 2021)

- https://www.comingsoon.it/film/l-attimo-fuggente
- https://it.wikipedia.org/wiki/L%27attimo_fuggente

Cosa trovi?

Trovo che nei due siti la durata in minuti è diversa, rispettivamente 128 e 129 minuti, vedi Figura 69. Ma come è possibile, delle due l'una, o il film dura 128 o dura 129 minuti?!

L'attimo fuggente è un film di genere drammatico del 1989, diretto da Peter Weir, con Robin Williams e Robert Sean Leonard. Durata 129 minuti. Distribuito da Warner Bros Italia (1989) - Touchstone Home Video.

https://www.comingsoon.it/film/l-attimo-fuggente/

Titolo originale	*Dead Poets Society*
Paese di produzione	Stati Uniti d'America
Anno	1989
Durata	128 min

https://it.wikipedia.org/wiki/L%27attimo_fuggente

Figura 69 – Due valori diversi per la durata del film "L'attimo fuggente"

E' possibile! Cominci ad accorgerti che spesso è difficile, se non impossibile, trovare ….la *verità*. In questo caso, riguardo ai due valori diversi della durata di L'attimo fuggente, 128 e

129 minuti, dobbiamo capire cosa fare. Questa attività viene chiamata *fusione*, questo è il termine che si usa per unire valori diversi in un unico valore; ci sono almeno tre possibilità:

1. considerare la media, oppure
2. considerare il valore riportato nel sito di cui ci fidiamo di più, oppure ….
3. possiamo prenderli tutti e due.

Facciamo così: come si usa dire, *salomonicamente* li prendiamo tutti e due, dando luogo alla tabella *corretta* di Figura 70.

Film del ventesimo secolo

Titolo originale del film	Titolo in italiano	Regista	Anno	Durata in minuti
Casablanca	Casablanca	Curtiz	1943	102
Dead Poets Society	L'attimo fuggente	Weir	1989	128-129
Roman Holidays	Vacanze Romane	Wyler	1953	118
Sabrina	Sabrina	Wilder	1954	113

Figura 70 – La tabella corretta

Insisto: non è possibile che il film Dead Poets Society abbia due durate in minuti!

Si, è possibile, e bada bene, accade tutti i giorni di dover approssimare la misura di un fenomeno. Prova a guardare sul Web che tempo farà domani a Milano, troverai temeprature minime, temperature massime, ecc. diverse tra di loro…

Questo esempio non mi va bene, il tempo che farà domani riguarda il futuro, la durata di un film riguarda il passato…

Bene, allora cerca sul Web la durata delle registrazioni dell'Arte della Fiuga di Bach. Io sono appassionato di questo meravigliosa musica di Bach, sento tante registrazioni diverse, e hanno tutte durate diverse…

….Mi hai convinto…

Qualità dei dati nei testi

Quando dalle tabelle, passiamo ai testi, la questione della qualità diventa più complessa, per cui non ce la possiamo cavare con poche pagine, è necessario trattarla con calma nel libro della Enciclopedia che parlerà della qualità. Un assaggio però lo possiamo fare, concentrandoci su una importante caratteristica di qualità che possiamo chiamare la *comprensibilità* del testo, con due varianti, la *comprensibilità del lessico* e *la comprensibilità dei contenuti*.

Della comprensibilità del lessico ne abbiamo già parlato all'inizio di questo libro, quando ho discusso il contributo storico di Tullio de Mauro sulla comprensibilità della lingua italiana.

Una volta stabilito un elenco di parole che si assume noto ai lettori, in particolare le 5.000 parole note a coloro che hanno ottenuto il diploma di terza media, l'accessibilità del lessico si può misurare contando le parole utilizzate nel testo che non rientrano nelle 5.000, magari calcolando la percentuale sul totale delle parole usate.

Riguardo alla comprensibilità dei contenuti, la prenderemo in considerazione per i *romanzi*, in cui, come abbiamo visto nel Capitolo 6 sui modelli dei dati, spesso l'intreccio si snoda introducendo vari personaggi, che vengono successivamente citati in particolari storie o contesti.

Il grafo semantico di Figura 57 intende essere una rappresentazioni che facilita la comprensione dei nomi di persone e dei loro legami di relazione affettiva o di altra natura nel romanzo I Demoni. Alcuni libri gialli, come ad esempio il libro di Agatha Christie The Mysterious Affair at Styles in Figura 71, riportano all'inizio del libro i nomi dei personaggi principali, e brevi descrizioni dei loro ruoli, senza peraltro dare indizi che portino a sospettare di qualche personaggio, perché in un libro giallo il colpevole si deve sempre scoprire all'ultima pagina……

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

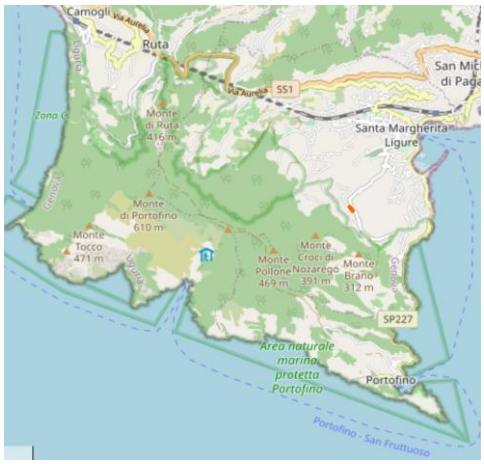
Figura 71 – I personaggi di un giallo di Agatha Christie

Qualità dei dati nelle mappe

Fino ad ora abbiamo discusso le caratteristiche di qualità dei dati senza pensare all'uso che ne facciamo, allo scopo che abbiamo nel farne uso.

Nella mia famiglia siamo appassionati di passeggiate, e abbiamo individuato nel parco di Portofino un luogo ideale. Prima di iniziare a camminare nei sentieri, abbiamo cercato delle cartine che ci illustrassero la rete dei sentieri; abbiamo trovato le due cartine mostrate in Figura 72, una da OpenStreetMap e l'altra da un sito specializzato.

Non c'è dubbio, fin da un primo sommario esame, che la mappa a destra ci è molto più *utile* della mappa a sinistra, quando il nostro scopo è fare passeggiate, mentre se noi dobbiamo usare la mappa per muoverci in treno o in automobile da Camogli a Santa Margherita Ligure, allora la maggiore evidenza che hanno le strade e la via ferrata nella mappa di OpenStreetMap ci aiuta certamente di più.



https://www.openstreetmap.org www.portofinotrek.com/trek/6-mappa

Figura 72 – Se intendiamo fare una passeggiata, la mappa a destra è migliore

Vorrei osservare che le mappe ormai non le usa più nessuno, ci sono i navigatori per le automobili, le app per i sentieri…

Non è proprio vero. La carta ti fornisce una visione panoramica del territorio in cui si svolge la passeggiata molto più ampia dello schermo del telefono mobile, Tu hai mai fatto una passeggiata in alta montagna, vuoi mettere l'ampiezza del campo visivo coperto dalla mappa, rispetto al piccolo schermo del telefono? e quando non hai campo come fai?

Qualità delle immagini e tradeoff tra caratteristiche di qualità

Vorrei introdurre nell'ambito delle immagini un altro aspetto importante della qualità dei dati. Guardate le due immagini di Figura 73, in entrambe le immagini viene mostrato l'ingresso di un parcheggio in una giornata nebbiosa. Non c'è dubbio che la immagine sulla sinistra sia più *fedele* alla scena originale, riproduce molto bene la nebbia fitta della zona. Ma decisamente è la immagine a destra a essere più *utile* per la nostra esigenza di sapere quanto si paga e quali sono le altre regole del parcheggio….

Figura 73 – Tradeoff tra fedeltà e utilità

C'è insomma in questo caso un *tradeoff*, un punto di equilibrio tra due caratteristiche di qualità, in questo caso la *fedeltà* e la *utilità per uno scopo;* questo si verifica spessissimo nella nostra vita, sia per i dati che utilizziamo, sia per un acquisto che dobbiamo fare, in cui dobbiamo decidere tra diverse alternative, basate su un certo numero di caratteristiche del bene o del servizio che vogliamo acquisire. Pensate all'acquisto di un telefono mobile, o di una bicicletta, oppure alla scelta delle condizioni contrattuali tra diverse aziende che forniscono servizi di connettività per il nostro telefono.

Qualità dei dati nella epidemia Covid

Il ricordo di quanto accaduto nel 2020 e nel periodo successivo non è facilmente cancellabile. Diciamo subito che il fenomeno del Covid è stato all'inizio un fenomeno caratterizzato da pochi dati, molto sparsi nello spazio che abbiamo definito in Figura 36. Con l'evolvere della epidemia in Italia, il Ministero della Salute, l'Istituto Superiore della Sanità e varie altri enti hanno iniziato a elaborare serie storiche e indicatori diffusi attraverso visualizzazioni del tipo di quelle mostrate in Figura 7. A livello mondiale, il più autorevole sito per la raccolta e la analisi comparativa delle serie storiche è stato il John Hopkins Institute, il cui sito ha come link https://coronavirus.jhu.edu/map.html.

E' apparso rapidamente chiaro che i dati forniti e le elaborazioni relative soprattutto a serie storiche di indicatori erano affette da significativi problemi di qualità.

1. Per quanto riguarda i dati sulla mortalità, è doloroso dirlo, ma sembrerebbe che sia il dato più certo su cui non si possono avere dubbi. E invece occorre ricordare che ad ogni decesso va associata una causa di morte. Sempre facendo riferimento ai dati del John Hopkins su tutti i paesi del mondo, i dati sulla mortalità sono stati affetti in diversi paesi dalla attribuzione ad altre patologie di decessi causati dal Covid, ovvero adottando un criterio di prevalenza tra Covid e altre patologie con diverse soglie in diversi paesi. Inoltre, il giorno in cui i dati sulla mortalità venivano rilevati è il giorno dell'accertamento della causa di morte, non il giorno del decesso, causando in tal modo una variabilità dei valori all'interno della settimana. Il basso dato della mortalità in molti paesi africani è probabilmente ascrivibile alla fragilità dei sistemi statistici nazionali.

2. Per quanto riguarda invece le persone risultate positive al Covid, e facendo riferimento alla situazione italiana, un indicatore che sembrerebbe permettere un confronto sul grado di diffusione in un intervallo di tempo, ad esempio un mese, è il rapporto

persone risultate positive / tamponi effettuati

che possiamo chiamare *tasso di contagiosità*. Purtroppo questo rapporto rispetto al fenomeno che vuole misurare, la diffusione della epidemia tra la popolazione, presenta problemi di inaccuratezza sia nel numeratore che nel denominatore.

Riguardo al numeratore, è più rappresentativo misurare non tutte le persone positive, ma quelle che lo sono diventate per la prima volta, escludendo le persone che si sono fatte il tampone, magari più volte durante la malattia per vedere se erano guarite. Analogo discorso

vale per i tamponi, che tuttavia presentano ulteriori problemi, come esemplificato dal trafiletto di giornale di Figura 74.

Figura 74 – C'è tampone e tampone

I diversi tipi di tampone avevano un diverso grado di affidabilità e operavano su universi diversi, per cui mescolarli nel calcolo del tasso di positività dà luogo a valori che perdono di significato (si dice anche informalmente, è come mescolare le mele con le pere….).

Ma soprattutto, il tasso di contagiosità misurato senza nessun criterio di scelta del campione statistico, non è veramente rappresentativo della popolazione, e di questo si è avuto evidenza nell'andamento nel tempo dell'indicatore, soggetto ad ampie variabilità. Molti ricercatori hanno più volte sottolineato la importanza di misurare questo ed altri indicatori con i metodi propri della statistica, ma sono sempre rimasti inascoltati, si veda in nota un testo molto chiaro che speiga in modo chiaro questa problematica.[9]

3. i dati sui pazienti in terapia intensiva, per molto tempo, furono forniti attraverso un numero che era la differenza tra i *pazienti usciti* dalla terapia intensiva e i *pazienti entranti*. Ma il vero dato rilevante era quest'ultimo, perché nella differenza erano in gioco diversi fenomeni che andavano osservati separatamente, quali i guariti, i deceduti, e i bisognosi di cure meno intense; ed infatti da un certo momento in poi il dato sui pazienti entranti fu fornito a parte.

4. infine l'indice Rt di diffusione del contagio, che esprime per ogni persona quante ne infetta mediamente in un certo intervallo di tempo, come ho già osservato nel Capitolo 1 è stato a lungo affetto da stime molto imprecise che hanno portato l'intervallo di confidenza, cioè l'intervallo dove si collocano il 95% dei valori risultato della stima, ad avere valori molto elevati, vedi ancora la Figura 9.

Ulteriori considerazioni sulle statistiche prodotte nel corso della epidemia Covid e sui metodi di stima possono trovarsi su molti siti, tra cui uno dei più autorevoli è *Our World in data*, https://ourworldindata.org/.

Qualità dei dati nel Web

Quando pensiamo ai dati nel Web, cambia tutto. La grande crescita dei dati accessibili dal Web ha modificato profondamente il concetto di qualità dei dati. Le differenze sono molteplici:

[9] https://www.ilsussidiario.net/news/indice-rt-e-chiusure-criterio-sbagliato-i-tamponi-non-sono-un-campione-statistico/2122097/ (verificato gennaio 2021)

1. nel Web mancano spesso sorgenti di conoscenza certe con cui confrontare il dato. Abbiamo visto in precedenza un esempio positivo, i titoli dei film, e un esempio negativo, la durata dei film. Essendo spesso la qualità del dato ricondotta alla qualità della fonte (si usa dire: di Wikipedia ci possiamo fidare!), acquista rilevanza stabilire la *provenienza* del dato e il *processo* con cui si è formato, ma questo spesso è difficile da realizzare.
2. il costo di produzione e di trasmissione dei dati è quasi nullo; è praticamente gratis pubblicare un messaggio su Twitter o mandare una mail a un indirizzario di centinaia o migliaia di indirizzi; scrivere un messaggio su Twitter richiede uno sforzo cognitivo molto basso, come sappiamo, ma può avere una diffusione di milioni di follower. E' dunque molto semplice diffondere dati errati, come accade nel caso delle cosiddette fake news.
3. Non c'è nessun filtro, come accade ad esempio sui giornali con una certa tradizione. Negli ultimi tempi i provider di reti sociali hanno iniziato a monitorare i messaggi, ma non basta. E Il "rovescio della medaglia" sta nel fatto che queste forme di controllo possono facilmente tramutarsi in censura. Quanto accaduto in occasione dell'assalto al Campidoglio negli ultimi giorni della Presidenza Trump ha portato i proprietari di Twitter ad escludere Trump dalla rete sociale. Personalmente penso che questa decisione fosse sacrosanta, ma ci sono state anche opinioni contrarie, soprattutto perché questa esclusione ha creato un precedente per la libertà di opinione nel Web.
4. Nel processo di formazione del dato, possono essere coinvolte diverse fonti, per cui può risultare molto difficile o impossibile ricostruire quale ruolo abbia avuto ciascuna di esse, e quindi la loro rilevanza rispetto alla qualità.

Per esprimerci con una metafora, i dati si espandono nel Web come una sfera opaca, vedi Figura 75, in cui accanto a dati attentamente verificati compaiono con sempre maggiore frequenza dati imprecisi, non aggiornati, incompleti, volutamente falsi, rendendo più arduo ricostruirne la validità. Nella grande sfera opaca il concetto di *qualità* del dato, da concetto intrinseco al dato, diventa un criterio sempre più soggettivo e sempre più influenzato dal messaggio, dal mezzo con cui viene diffuso e da aspetti emotivi e non razionali.

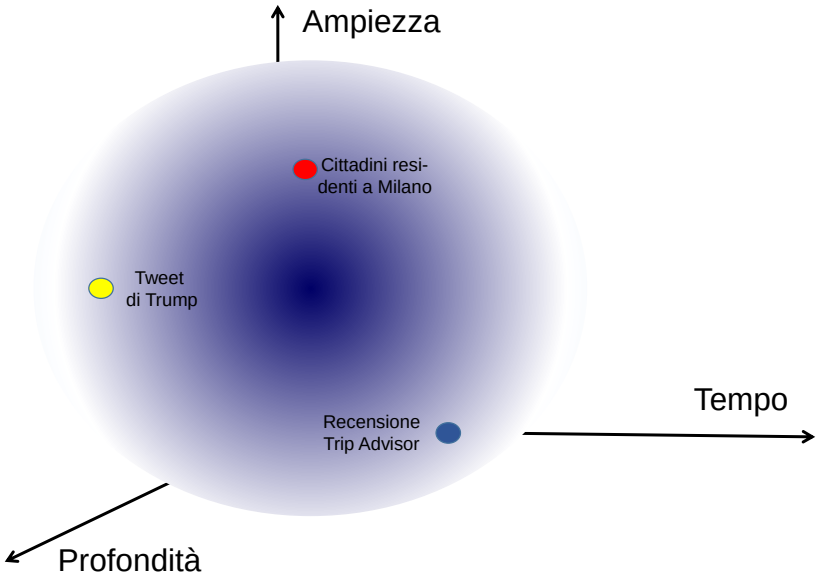


Figura 75 - La grande sfera opaca

La conclusione quale è, che non si può fare nulla? Se fosse così saresti un po' deludente…

No, per carità, si può fare tantissimo! Bisogna, però, munirsi di un po' di pazienza e *capacità investigativa*. Vediamo un esempio.

All'epoca della cerimonia di insediamento di Donald Trump come Presidente degli Stati Uniti, si diffusero foto che consideravano comparativamente le folle presenti all'insediamento di Obama e di Trump, vedi Figura 76. Apparentemente, la folla presente all'insediamento di Obama appariva essere di gran lunga più grande di quella di Trump, ma il portavoce di Trump 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 76 – I fatti alternativi

Ora, partiamo dall'assunto che la Conway avesse ragione, che la sua fosse una interpretazione alternativa, ma *plausibile*, delle foto rispetto al senso comune, e cerchiamo di confutarla. Per esempio, potrebbe essere accaduto che le due foto fossero stare fatte in momenti diversi rispetto all'inizio della cerimonia, per cui i sostenitori di Trump potevano essere arrivati in un tempo successivo alla foto.

Se ci procuriamo gli istanti temporali in cui le foto sono state scattate (vedi Figura 77), vediamo che così non è, le due foto sono state scattate in entrambi i casi mezz'ora prima. Ma potrebbe darsi che i sostenitori di Trump fossero arrivati tutti all'ultimo momento, in metropolitana.

Non è neanche vero questo, i dati sui biglietti utilizzati dicono che ci furono molti più biglietti utilizzati per Obama che per Trump.

Dobbiamo arrenderci, e con noi dovrebbe arrendersi anche Kellyanne Conway, che però, forse, aveva fatto quelle affermazioni consapevole che fossero false. Insomma, i dati sono *cocciuti*, o come diceva Umberto Eco, hanno una *tenuta*, reggono i colpi delle *confutazioni*.

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 77 – I dati sono cocciuti

Anche nelle elezioni americane del 3 novembre 2020 sono accaduti episodi simili. C'è stato un titanico sforzo da parte di Trump e dei suoi avvocati, in particolare Rudy Giuliani, di invalidare lo spoglio delle schede, sulla base di illazioni che furono tutte confutate.

qualche broglio ci sia stato davvero. Uno dei pezzi forti del repertorio trumpiano è che in alcune contee del Michigan ci siano stati più elettori che abitanti e questo dovrebbe provare che le elezioni sarebbero state truccate da un software speciale. Ma ieri è saltato fuori che nel ricorso legale gli avvocati di Trump hanno inserito per sbaglio anche molte contee del Minnesota (sigla: MN) perché credevano che fossero del Michigan (sigla: MI) e questo spiega perché i conti non tornano. I legali del presidente

Figura 78 – In Michigan ci sono più elettori che abitanti!!!

In particolare, Giuliani sostenne (vedi Figura 78) che in Michigan, ad esempio, c'erano più elettori che abitanti, e questo sulla base dei conteggi effettuati sulle persone presenti nelle anagrafi delle contee. Peccato che, come dice l'articolo del giornale "Il Foglio" nella edizione di quei giorni, riprodotto in Figura 78 a destra, fossero state confuse le sigle del Michigan (MI) e del Minnesota (MN), e di conseguenza gli elenchi prodotti contenessero anche gli abitanti del Minnesota! Ecco un altro esempio di dati cocciuti!

Questi episodi dimostrano che, anche senza essere esperti di dati, tuttavia possiamo sempre esercitare una sana curiosità, ragionare con il buon senso, leggere diverse fonti, diversi giornali, non dare nulla per scontato, e capire che, con l'avvento del Web, adattando una famosa frase di Umberto Eco, chiunque può dire qualunque cosa, e quindi è diventato molto più difficile del passato comprendere chi ha torto e chi ha ragione.

E nel diffondere nel Web qualunque falsità, e gli altri nel recepirlo, molti fanno leva soprattutto sulle emozioni, non sulla razionalità, come ci dice questo bellissimo articolo sul Guardian che potete leggere (verificato nel gennaio 2021) alla pagina

https://www.theguardian.com/us-news/2021/jan/01/disinformation-us-election-covid-pandemic-trump-biden

E' soprattutto questo equilibrio tra emozioni e razionalità che le società probabilmente devono ricercare perché il Web si trasformi nel tempo in una agorà tutta orientata alla ricerca della verità.

Riassumendo

I dati digitali rappresentano il mondo, ma lo rappresentano spesso, per tante ragioni, in modo ***impreciso***, si può dire anche con ***scarsa qualità***. La qualità dei dati è un concetto multiforme, ha tante sfaccettature o ***caratteristiche*** di qualità, che dipendono molto anche dalla ***forma*** che i dati digitali assumono. Le ***tabelle***, una delle forme di dati più usate, hanno come principali caratteristiche di qualità la ***accuratezza*** e la **completezza**. Accuratezza significa che i dati digitali rappresentano esattamente i valori reali nel mondo, completezza significa che li rappresentano tutti. Per le immagini le caratteristiche più importanti sono la ***fedeltà*** all'originale e la ***estensione del colore***, che è l'opposto della ***opacità***. Per le immagini abbiamo visto una caratteristica di qualità, la ***utilità***, che è rilevante per tutti i tipi di dati digitali; un dato è utile quando ci serve per la decisione o azione che intendiamo intraprendere. Abbiamo osservato per le immagini che fedeltà e utilità possono essere in contrasto o ***tradeoff***, quando aumenta l'una diminuisce l'altra, questo è tipico di molte caratteristiche di qualità.

Una volta scoperto von una ***valutazione di qualità*** che un dato ha una qualche caratteristica di qualità insoddisfacente, possiamo ***correggerlo, migliorandone la qualità***; ad esempio, per un nome di comune italiano visibilmente sbagliato, possiamo confrontarlo con i nomi dei comuni italiani, sostituendolo con il nome più vicino.

La ***qualità dei dati nel Web*** è molto più difficile da trattare, perché nel Web costa poco sforzo inviare dati, messaggi, opinioni, e ci vuole più tempo per verificare se un dato è falso che per accettarlo come vero. Tuttavia i dati hanno una ***tenuta***, come diceva Umberto Eco, insomma se facciamo delle ***verifiche*** anche ***empiriche***, e esercitiamo ***spirito critico***, prima o poi riusciamo a distinguere il ***dato falso*** dal ***dato vero***.

Capitolo 8

I dati che non abbiamo

Abbiamo visto nei capitoli precedenti che, quando utilizziamo dati digitali prodotti, ad esempio da una app sul telefono mobile, dobbiamo stare attenti al fatto che i dati siano accurati e aggiornati. Ma abbiamo anche visto che certe volte i dati *ci mancano*, ovvero sono *incompleti*.

Con la progressiva espansione dei big data, paradossalmente, più avremo a disposizione dati, più la dimensione della sfera di Figura 79 crescerà, e quindi crescerà la percezione della necessità di altri dati*, che non abbiamo*.

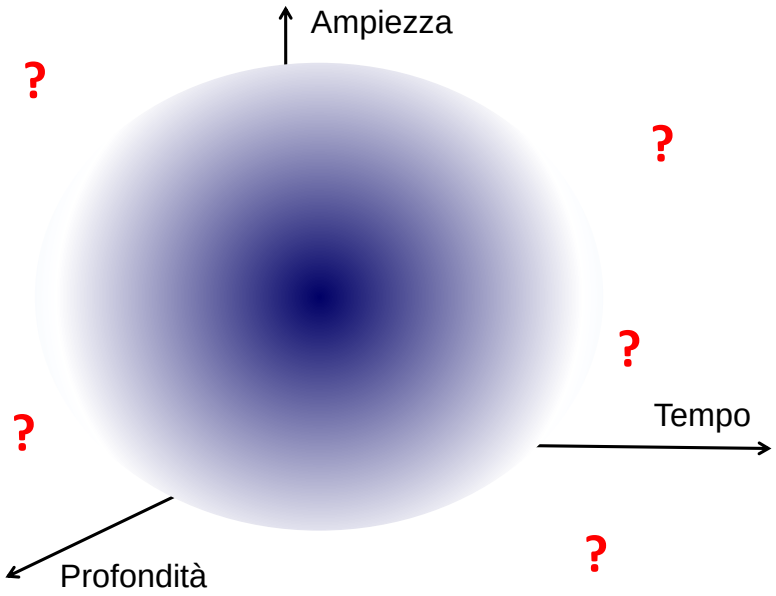


Figura 79 – I dati che ci mancano

E' quindi necessario investigare meglio il tema dei dati che ci mancano. In Figura 80 ho riportato una classificazione dei tipi di dati che non abbiamo. Questa classificazione è tratta e adattata da un libro molto bello, citato in nota [10].

[10] David Hand, *Il tradimento dei numeri, I dark data e l'arte di nascondere la verità*, Rizzoli, 2019

1. Dati che ci mancano, di cui conosciamo la mancanza
2. Dati di cui ignoriamo la mancanza
3. Dati volutamente o involontariamente non scelti
4. Dati che avremmo potuto avere
5. Dati che si sono modificati nel tempo
6. Definizioni di dati assunte da altri e che non conosciamo
7. Dati elementari nascosti nei dati aggregati
8. Errori di misurazione che introducono un margine di incertezza
9. Dati noti in modo asimmetrico
10. Dati intenzionalmente nascosti
11. Dati veri che si discostano da quelli che generiamo

Figura 80 - Tipi di dati che non abbiamo

Commentiamo ora alcune di queste tipologie, una trattazione completa sarà fatta nel volume della Enciclopedia dedicato alla qualità dei dati. L'esempio che segue è tratto dal libro di Hand.

Dati di cui ignoriamo la mancanza

Un esempio dei dati della tipologia 2: *Dati di cui ignoriamo la mancanza*, riguarda una iniziativa della città di Boston che aveva lo scopo di individuare con una app su telefono mobile le buche nelle strade della città, vedi Figura 81.

. Figura 81 - Dati di cui ignoriamo la mancanza – le buche a Boston

Per rilevare automaticamente le buche, la municipalità di Boston rilasciò una applicazione per telefoni mobili dotati di accelerometro, in grado di rilevare improvvisi sobbalzi nelle automobili i cui penumatici incontravano la buca; i sobbalzi generavano un improvviso cambiamento della velocità e accelerazione della massa della automobile.

La sperimentazione portò a risultati distorti dal fatto che i telefoni mobili dotati di accelerometro all'epoca avevano un costo che poteva essere sopportato solo da ceti agiati, e quindi erano diffusi in modo ineguale nella città.

Definizioni di dati assunte da altri e che non conosciamo

La tipologia 6, *Definizioni di dati assunte da altri e che non conosciamo* è interessante perché *non riguarda direttamente i dati ma il loro significato*, quella che viene chiamata la *definizione*. Capita molto spesso nell'uso del linguaggio naturale e anche nelle statistiche che si faccia riferimento a classi di dati cui vengono attribuiti significati diversi.

Un esempio rilevante è quello della epidemia di Covid, in cui nelle statistiche diffuse si usavano terminologie come "Il tasso di contagiosità è dato dal rapporto tra soggetti risultati positivi e tamponi effettuati", in cui, come abbiamo già visto nel Capitolo dedicato alla qualità dei dati, sia al numeratore che al denominatore possono essere dati diversi significati. Infatti, ricordo:

1. I soggetti risultati positivi possono essere tutte le persone che si sono sottoposte al tampone, ovvero le sole persone che si sono sottoposte al tampone la prima volta.
2. I tamponi possono essere di varie tipologie, il tampone nasofaringeo, il tampone (o test) sierologico, il tampone salivare, ecc. ognuno dei quali, è caratterizzato da errori diversi.

Nella prima fase della epidemia, provai a stilare un elenco dei diversi modi con cui le persone menzionate nelle statistiche, negli articoli di giornale, nei siti specializzati, nei provvedimenti del governo o delle regioni venivano citate, e venne fuori l'elenco di Figura 82.

1. Persona (fisica) vivente (qui e nel seguito: un certo giorno e in un certo luogo)
2. Persona cui è stato fatto un tampone ed è risultato positivo
3. Persona cui è stato fatto un tampone ed è risultato negativo
4. Persona ospedalizzata
5. Persona dichiarata guarita da coronavirus che in precedenza era stata diagnosticata affetta da coronavirus
6. Persona dimessa dichiarata guarita da coronavirus che in precedenza era stata diagnosticata affetta da coronavirus
7. Persona che si sente soggettivamente guarita da coronavirus
8. Persona in quarantena che vive da sola
9. Persona in quarantena che vive con almeno un'altra persona
10. Persona medico di base del SSN
11. Persona che lavora in ospedale
12. Persona con coronavirus
13. Persona con sintomi da coronavirus
14. Persona sottoposta a terapia intensiva
15. Persona che vive con familiare
16. Persona cui è stato fatto un tampone ed è risultato positivo che vive con altra persona
17. Persona cui è stato fatto un tampone ed è risultato positivo che vive da solo40.
18. Persona con sintomi da coronavirus che vive con altra persona
19. Persona con sintomi da coronavirus che vive da solo
20. Persona in quarantena che vive con altra persona
21. Persona in quarantena che vive da solo
22. Persona deceduta cui è stato fatto un tampone ed è risultata positiva a coronavirus
23. Persona deceduta che al momento della morte aveva sintomi da coronavirus
24. Persona deceduta che al momento della morte era affetta da coronavirus
25. Persona deceduta che al momento della morte aveva una patologia grave
26. Persona deceduta che al momento della morte aveva due patologie gravi
27. Persona deceduta che al momento della morte aveva più di due patologie gravi
28. Persona impiegata in una attività ritenuta strategica nel periodo della epidemia da coronavirus
29. Persona impiegata in una attività ritenuta non strategica nel periodo della epidemia da coronavirus
30. Persona con contratto temporaneo in una attività ritenuta non strategica nel periodo della epidemia da coronavirus
31. Persona con contratto temporaneo in una attività ritenuta non strategica nel periodo della epidemia da coronavirus
32. Persona considerata di categoria a particolare rischio
33. Persona che lavora in smart work
34. Persona che proviene da altro luogo l1 in cui era il giorno g-1
35. Persona georeferenziata (longitudine e latitudine)
36. Persona sottoposta a particolari restrizioni sociali
37 Asintomatico
38. Paucisintomatico
39. Lieve
40. Severo
41. Critico
42. Persona che vive in contesto a particolare rischio da coronavirus
43. Persona dotata di mascherina
44. Persona non dotata di mascherina
45. Persona che è stata a contatto in luogo l1 e giorno g1 con persona affetta da coronavirus
46. Persona che è stata a contatto in luogo l1 e giorno g1 con persona diagnosticata affetta da coronavirus

Figura 82 – I diversi tipi di persone nell'epoca del Covid

Questo elenco di tipi di persone può apparire eccessivamente pignolo, ma a me pare una fotografia della complessità dei problemi, e della enorme discontinuità e diversificazione nella società tra l'epoca pre-Covid e l'epoca post-Covid, oltre che dei differenti punti di vista e interpretazioni che possiamo dare alle diverse categorie di persone.

Questa diversificazione nel significato da attribuire a un dato, e le conseguenze che può portare il non conoscere e condividere il significato attribuito da un altro soggetto, ebbe

grande e decisiva importanza in una famosa lite tra Regione Lombardia e Istituto Superiore di Sanità nel gennaio 2021.

Il colore delle Regioni italiane – il caso della Lombardia del gennaio 2021

Nel periodo di fine anno 2020, le Regioni italiane nei giorni festivi e prefestivi furono tutte colorate di rosso, nel senso che a tutte indistintamente furono applicate le restrizioni più severe tipiche delle regioni rosse. Al termine del periodo, il 7 gennaio 2021, fu ripristinato nelle regioni il colore assegnato automaticamente dall'algoritmo di calcolo definito dall'Istituto Superiore di Sanità (ISS), in cui, ricordo, venivano considerati ai fini della attribuzione del colore, l'indice Rt di trasmissione del contagio e un fattore di rischio basato su ventuno indicatori.

Alla Regione Lombardia fu assegnato il colore rosso, suscitando le ire del Presidente della Regione Fontana e della da poco nominata vice presidente Moratti. Vi furono interlocuzioni tra la Regione e l'ISS, al termine delle quali la Regione inviò nuovo dati in merito alla evoluzione della pandemia, sostitutivi di quelli precedentemente inviati per lo stesso periodo, e l'ISS, a seguito di riapplicazione dell'algoritmo, assegnò alla Lombadia il colore arancione. La differenza tra rosso e arancione non era di poco conto, perché il colore arancione permetteva la apertura di attività commerciali che erano escluse nel caso di colore rosso.

Scoppiarono molte polemiche, con i vertici della Regione Lombardia che accusarono l'ISS di aver applicato in modo non corretto l'algoritmo di calcolo del colore e di aver usato dati *vecchi*, e in tal modo di aver procurato un danno economico ingente alla economia della Regione. L' Agenzia Adnkronos riportò una analisi dell'ufficio studi della Confcommercio che stimava un costo di almeno 200 milioni di euro (successivamente lievitato in altre stime a 600 milioni di euro) di mancato fatturato, considerando tutti gli esercizi commerciali, dai bar ai negozi per la settimana di zona rossa "superflua". L'ISS ribaltò le accuse alla Regione, per aver inviato per lungo tempo (quindi, anche ben prima della settimana finale) dati inaccurati e incompleti. Vediamo dunque di capire cosa è successo, partendo dalla Figura 83.

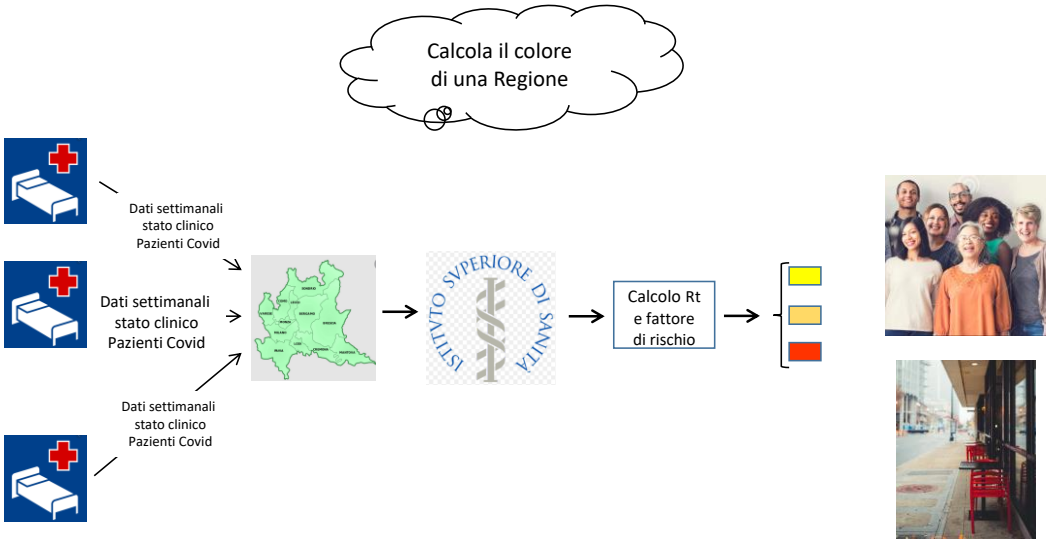


Figura 83 – Il flusso dei dati sottoposto al programma di calcolo del colore di una regione

In Figura 83 sono mostrati i flussi di dati e le organizzazioni coinvolte nel monitoraggio settimanale effettuato dall'ISS sulla base dei dati pervenuti dalle regioni. I dati venivano rilevati ogni settimana, inizialmente inseriti nel sistema per mezzo di appositi moduli dalle strutture territoriali (ospedali, case per anziani, ecc.), trasferiti digitalmente alle regioni, messi insieme in una unica tabella, inviati a Roma.

La figura 83 mette in evidenza un aspetto di straordinaria importanza; il processo di calcolo del colore ha un impatto fondamentale sulla vita delle persone e delle aziende. A seconda del colore, gli studenti delle varie scuole dovranno frequentare le scuole in presenza o con la didattica a distanza, gli esercizi commerciali saranno aperti oppure chiusi, i bar saranno chiusi ovvero potranno fornire servizi solo per asporto, oppure potranno far entrare i clienti in numero limitato. Come conseguenza, il processo di calcolo del colore richiede molta attenzione e cura, perché un dato sbagliato può avere conseguenze enormi sulla qualità della vita delle persone e delle aziende.

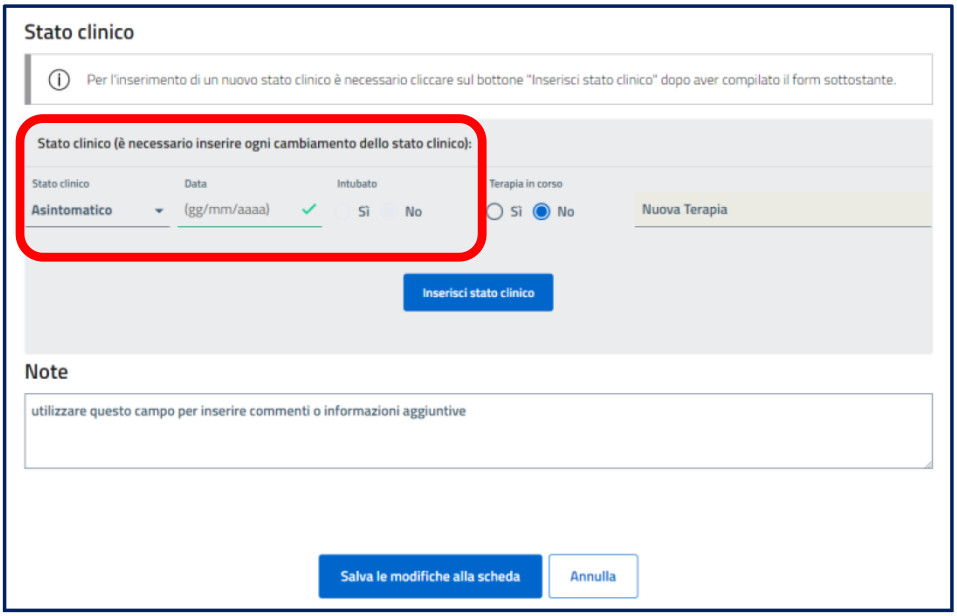


Figura 84 – Il modulo utilizzato per inserire i dati nelle tabelle presso gli enti territoriali.

I dati sono inseriti nel sistema a cura dagli enti territoriali, sulla base di un modulo di acquisizione dei dati mostrato in Figura 84. I due dati utilizzati per il calcolo dell'indice Rt (dato fondamentale per la determinazione del colore da attribuire) sono:

- lo stato clinico, che rappresenta un indicatore di gravità del quadro clinico del paziente, e
- la data di insorgenza dei sintomi.

Il significato delle diverse tipologie dello *stato clinico* è mostrato in Figura 85. Ai fini del calcolo dell'Rt *sono considerati solo i sintomatici*, per ragioni legate al modello matematico adottato nell'algoritmo di calcolo.

Lo stato clinico deve essere raccolto *per ogni paziente alla diagnosi* e *per ogni suo cambiamento, inserendo la data del cambiamento*. Quindi, ogni cambiamento dello stato

clinico del paziente deve essere inserito nel box “Stato clinico” (Figura 83, riquadro rosso) con valori possibili: asintomatico, pauci-sintomatico, lieve, severo, critico, guarito, deceduto; è inoltre necessario riportare la data.

DEFINIZIONE di STATO CLINICO

Asintomatico	Persona con assenza di segni o sintomi apparenti di malattia;
Pauci-sintomatico	Persona con sintomi lievi e generali (ad es. malessere, febbre, stanchezza, ecc.)
Lieve	manifestazioni cliniche a carico delle vie respiratorie/altri organi apparati che non necessiterebbe normalmente di ricovero;
Severo	manifestazioni cliniche a carico delle vie respiratorie/altri organi apparati che necessitano di ricovero (non in terapia intensiva);
Critico	manifestazioni cliniche a carico delle vie respiratorie/altri organi apparati che necessitano di ricovero in terapia intensiva.
Guarito	paziente con scomparsa dei sintomi di infezione da COVID-19 e a cui sono stati effettuati due tamponi consecutivi risultati negativi a distanza di 24 ore. Se si intende utilizzare la voce guarito anche per i pazienti guariti solo clinicamente, è necessario specificare nel campo note la voce “**guarito clinicamente**” in modo da poter eventualmente tenere distinte le due definizioni.

Figura 85 – Definizioni dei possibili stati clinici

Siamo arrivati al punto importante: gli operatori che inseriscono i dati nel modulo di Figura 83 *non sono obbligati a inserire dei valori*, possono lasciare il campo *vuoto*. Sembrerebbe che lasciare il campo vuoto corriponda al significato: *stato clinico non noto*, ma non è così.

Quel valore vuoto, non specificato, ha un significato. Infatti valgono le seguenti regole, *che fanno parte del significato del dato ma non sono esplicitamente scritte nel manuale*:

1. nel caso che, per una qualunque ragione: pigrizia dell’operatore, fretta nel compilare il modulo, o non disponibilità del dato, *alla data di inizio sintomi non sia associato uno stato clinico*, i pazienti *vengono inizialmente considerati sintomatici* perché in assenza di altre informazioni si riconosce il dato fornito dalla regione come indicativo della presenza di sintomi. Questo in base a un *principio di prudenza*: se tu non mi dici nulla, faccio prevalere il caso più pessimistico.

Possiamo sintetizzare il punto precedente in questo modo.

Se data inizio sintomi è *specificata* e stato clinico è *vuoto* ALLORA il paziente è *sintomatico*

La inclusione dei sintomatici dei pazienti per cui non è stato specificato lo stato clinico, li fa considerare automaticamente come una popolazione inclusa nel calcolo di Rt.

2. Nel caso in cui invece il campo stato clinico non venga mai compilato fino a quando sia documentata la guarigione o il decesso, il caso si considera *asintomatico* nonostante la presenza di una data di inizio sintomi. Non è infatti plausibile che ci sia una data di inizio sintomi di una persona senza alcun sintomo documentato fino alla guarigione.

Nella situazione emergenziale in cui possono verificarsi problemi nella qualità dei dati, questo approccio conservativo è volto ad evitare una *sovrastima* dei casi sintomatici su cui si calcola l'Rt.

Nel contesto descritto in precedenza, la Regione Lombardia ha continuato a fornire all'ISS per almeno due mesi dati in cui una significativa percentuale ricadeva nel caso 1. Afferma a tale proposito l'ISS:

"Da maggio 2020 a gennaio 2021 la Lombardia ha segnalato una grande quantità di casi, significativamente maggiore di quella osservata in altre regioni, *con una data di inizio sintomi a cui non ha associato uno stato clinico* e che pertanto si è continuato a considerare inizialmente sintomatici. Inoltre, nell'ultimo periodo ha classificato un gran numero di questi come guariti senza uno stato clinico sintomatico riportato.

La percentuale di casi incompleti per la sintomatologia (assenza del dato sullo stato clinico) è pari al 50,3% (per la Regione Lombardia) a fronte del 2,5% del resto d'Italia nel periodo 13 dicembre 2020-13 gennaio 2021."

Gli ultimi inserimenti da parte della Regione Lombardia risalgono alle ore 10.58 e alle ore 14.51 del 20 gennaio 2021 con una rettifica dei dati pregressi presenti alla data 13 gennaio 2021:

- *eliminando* la segnalazione di una data inizio sintomi in 4.875 casi segnalati;
- *diminuendo* di 17.654 casi quelli classificati in precedenza come sintomatici;
- *aumentando* di 12.779 casi quelli classificati come asintomatici.

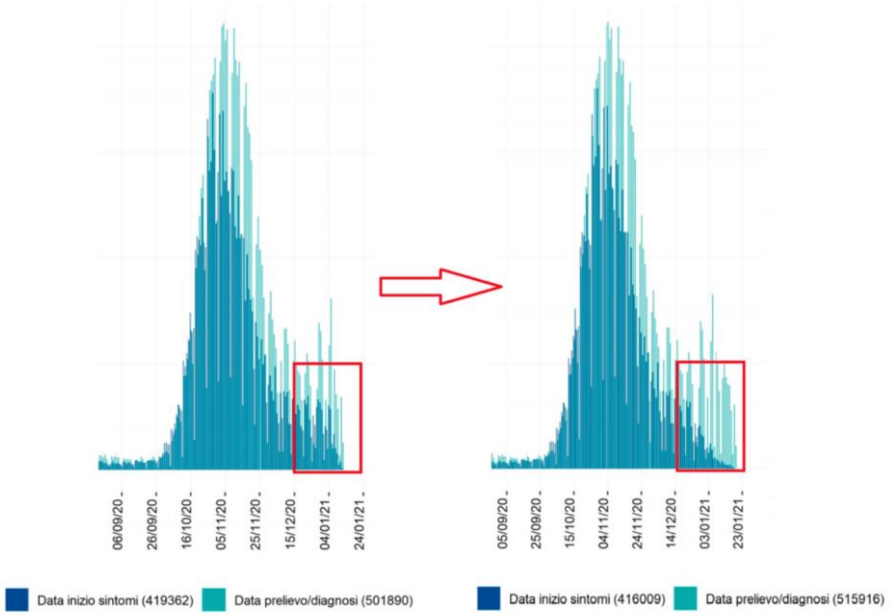


Figura 86 - Confronto tra i dati del 13 gennaio (sinistra) e quelli del 20 gennaio (destra) sui casi positivi riportanti una data inizio sintomi – Fonte: Iss

In Figura 86 vediamo un confronto tra i dati del 13 gennaio e quelli del 20 gennaio.

Siamo in grado di tratte conclusioni basate su una conoscenza adeguata di quanto successe: *da una parte* nel modulo non era specificato in maniera chiara che *il significato del campo vuoto* dovesse intendersi *sintomatico*, e quindi portasse a un aumento del valore stimato per Rt; ma *dall'altra* la Regione Lombardia è stata pressochè l'unica regione a non aver compilato quel dato in un elevato numero di casi (circa il 50%), e quindi è l'unica a non aver capito che in quel modo *firmava la propria sentenza di condanna al colore rosso*.

Abbiamo visto una vicenda in cui errori e incompletezze dei dati hanno comportato un danno economico; riprenderemo queste considerazioni quando parleremo del valore economico dei dati.

Dati elementari nascosti nei dati aggregati

La tipologia 7, *Dati elementari nascosti nei dati aggregati*, apre un intero mondo di considerazioni che si possono fare, e concetti che si possono utilizzare, nelle due discipline che operano sui dati, la nascente disciplina della Scienza dei Dati, giovanissima, i cui primi vagiti sono stati nei primi anni 2000, e la Statistica, le cui origini, nella concezione più moderna, vengono fatte risalire a quel a che un economista e matematico inglese, William Petty (1623 - 1687), chiamò "aritmetica politica", ovvero "l'arte di ragionare mediante le cifre sulle cose che riguardano il governo".

Una delle branche della Statistica è la statistica descrittiva, che rappresenta i fenomeni mediante vari tipi di aggregazione sui dati. Nell'ambito dei dati raccolti nella epidemia Covid, dire, ad esempio, che la età media dei pazienti Covid deceduti è di 81 anni non ci dice nulla sulla distribuzione per fasce d'età, ovvero sulla distribuzione tra uomini e donne; i dati aggregati nascondono per definizione dati elementari, per ogni problema che abbiamo, è sempre necessario capire sempre quale è il livello minimo di aggregazione dei dati necessario per conoscere il fenomeno.

Questa Enciclopedia non si occupa della Statistica, che pure è uno strumento di analisi indispensabile nel nostro mondo che possiamo conoscere solo in modo approssimato e incerto. Peraltro, la Statistica è una disciplina che nel corso del secolo ventesimo e di questo scorcio del ventunesimo ha visto comparire ottimi testi divulgativi, i cui riferimenti sono in nota[11].

Dati noti in modo asimmetrico

Menziono infine la tipologia 9, *Dati noti in modo asimmetrico*, che è alla base del concetto di *asimmetria informativa*, concetto cardine delle teorie economiche che si occupano di mercati. Quando andiamo in un negozio, ad esempio un venditore di tappeti, i dati che ha il venditore di un tappeto che ci interessa, della sua fattura, della sua provenienza, sono molti di più di quelli a nostra disposizione; se non siamo un esperto di tappeti, allora possiamo acquisire altri dati su quel tappeto, e farci una idea della sua qualità e della convenienza

[11] David Spiegelhalter, L'arte della statistica, cosa ci insegnano i dati, Einaudi 2020
Anna Ferrari, Io Statistica, le mie memorie, Capitolo 9 del testo "La Scienza dei dati", liberamente scaricabile dal link https://boa.unimib.it/handle/10281/295980

dell'acquisto, ma siccome esistono pochi esperti di tappeti in giro, spesso dobbiamo fidarci del venditore e della sua reputazione…

Insomma, i dati sono un bene troppo prezioso per trattarli superficialmente; i dati vanno rispettati, i dati vanno curati, ora, spero, possiamo dirlo con maggiore consapevolezza.

Riassumendo

Anche nella nostra epoca dei big data, sono di più le cose che non sappiamo di quelle che sappiamo, i **dati che non abbiamo** rispetto ai dati che abbiamo.

I dati che non abbiamo sono di tanti tipi. Possono essere **dati che ci mancano** di cui conosciamo la mancanza, ovvero ***dati che non sappiamo di non avere***, oppure ***dati che sono cambiati nel tempo***, oppure ***dati che non conosciamo e che il nostro interlocutore conosce bene***, oppure ***dati per i quali non sappiamo come il nostro interlocutore li definisce***. E altri ancora.

I ***dati che non abbiamo*** sono complessi da trattare perché la ricerca non li ha ancora studiati a sufficienza.

Capitolo 9

Gli occhiali non bastano, ci servono anche microscopi e cannocchiali

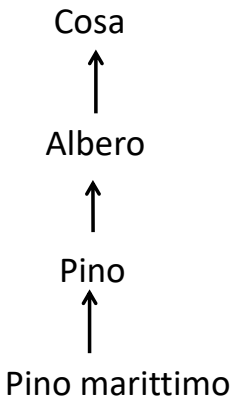
Le astrazioni dei dati

Immaginate di essere a Roma con un amico e supponete di visitare un'area archeologica (vedi Figura 87): se volete indicare al vostro amico il soggetto principale nella foto nella figura, quale espressione scegliereste tra:

- Guarda quella cosa;
- Guarda quell'albero;
- Guarda quel pino;
- Guarda quel pino romano?

Provate a rispondere alla domanda…..ne parleremo nella prossima pagina.

a. Una foto di un'area archelogica a Roma



b. Differenti nomi per lo stesso oggetto

Figura 87 – Come chiameresti quella cosa?

Proviamo a ragionare sulle possibili risposte.

Nel linguaggio verbale che usiamo quando comunichiamo con un'altra persona, per indicare un oggetto noi scegliamo un termine, una parola; nello scegliere tra i quattro termini di Figura 87 probabilmente non sceglieremo il termine *cosa*, a meno che il nostro vocabolario non sia molto limitato, mentre sceglieremo tra *albero*, *pino* e *pino marittimo* a seconda della nostra cultura in botanica.

Dei quattro termini: cosa, albero, pino, pino romano, *cosa* esprime un concetto intuitivamente *più generale* e astratto di *albero*, albero più astratto di *pino* e pino più astratto 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. Si usa dire che i quattro termini sono tra loro in una relazione di *astrazione*, più specificatamente di *generalizzazione*.

Possiamo affermare che, nel denotare un oggetto della realtà, noi, spesso inconsapevolmente, effettuiamo una operazione che risponde alla nostra idea intuitiva di *astrazione*, intesa come operazione mentale che nel denotare un oggetto del mondo, elimina gli aspetti di dettaglio ritenuti non rilevanti (ad esempio l'albero in Figura 87 ha otto rami visibili) per mettere in evidenza gli aspetti comuni (l'albero appartiene alla specie dei pini, ovvero, l'albero appartiene alla specie dei pini marittimi). La *generalizzazione* è una particolare forma di astrazione, con ogni probabilità la più usata nel linguaggio verbale e scritto. Noi diciamo che albero è generalizzazione di pino perché tutti i pini sono alberi, tra pini e alberi è definita una relazione di sottoinsieme.

Pizza e fichi

La caprese

Figura 88 - Due astrazioni: Pizza e fichi e la caprese

Per passare a una osservazione un po' scherzosa, non sono sempre chiari i meccanismi mentali usati dagli esseri umani nel dare un nome alle cose. Guardando la Figura 88, il cibo a destra, caprese, ha un nome astratto rispetto agli ingredienti che lo compongono, pomodori, mozzarella e basilico, mentre il nome a sinistra è la unione degli ingredienti, pizza e fichi. In entrambi i casi, è definita una astrazione di *aggregazione*, intendendo, ad esempio, che caprese ha come parti componenti, è quindi una aggregazione di, pomodori, mozzarella e basilico. Ad esempio, una bicicletta di cosa è astrazione di aggregazione? Las risposta nella prossima pagina.

Una bicicletta, vedi Figura 89, è l'aggregazione di un telaio, due ruote, due copertoni, una sella, un manubrio, due freni, due pedali, una catena.

Figura 89 – Una biciletta in riva al mare

Le astrazioni sono usate in tutte le scienze. Considera per esempio le seguenti uguaglianze:

$$(4 + 7)^2 = \quad 4^2 \; + 7^2 \; + 2 \times 4 \times 7$$

Quella uguaglianza è vera solo per 4 e 7 o per tutti i numeri interi?

Per tutte le coppie di numeri interi!

Certamente, e infatti possiamo affermare che per ogni coppia di numeri interi vale la uguaglianza:

$$\boxed{(a+b)^2 = a^2 + b^2 + 2ab}$$

Questa formula matematica afferma che hanno sempre valore uguale:

- il quadrato della somma di due numeri interi
- la somma dei quadrati dei due numeri più il prodotto dei due numeri moltiplicato per due.

La formula è quindi anch'essa una astrazione nel senso proposto poco fa, perchè elimina il dettaglio relativo ad una 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.). Nella matematica le astrazioni giocano un ruolo fondamentale.

Si, capisco le astrazioni in matematica, ma perché te ne occupi per i dati? A cosa servono le astrazioni per i dati?

Cominciano le domande difficili, come quando si risponde ad uno studente: se pazienta un pò, nelle prossime lezioni sarò in grado di rispondere in maniera più rigorosa… Ora preferisco risponderti con un esempio. Sei mai stato a Lione?

No….

Lione è una bella città nel sud della Francia, dove il nuovo e il moderno si fondono molto bene coni monumenti e gli edifici del passato. Se tu vuoi visitare Lione, puoi acquistare una mappa cartacea come quella di Figura 90. Riesci a raccapezzarti guardando questa mappa? Riesci a capire come andare da un posto a un altro, quali sono i punti di interesse più importanti della città?

E' tutto molto confuso…

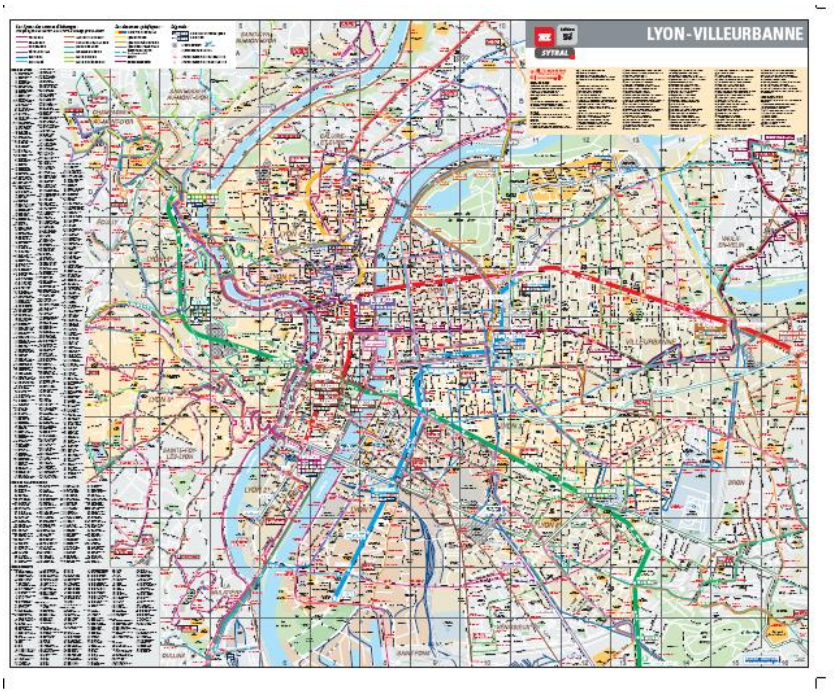


Figura 90 – Comprereste questa mappa per orientarvi a Lione?

Sono d'accordo: io penso che quella mappa di Lione sia *illeggibile*, perché fornisce troppi dati tutti insieme, è troppo fitta di dati, e alla fine non si riesce a capire niente, abbiamo come un senso di rifiuto!

Ora vorrei farti un esempio che mi permette di continuare a parlare di astrazioni, ma osservando il processo inverso alla astrazione, che chiamerò raffinamento. Usiamo in raffinamento quando partendo dalla descrizione di una bicicletta come oggetto che ci permette di spostarci velocemente su una strada, la esaminiamo come un insieme di parti interagenti, ognuna delle quali svolge una funzione. e

Guarda dunque la Figura 91, in cui le mappe sono tratte da Open Street Map. In questo caso immaginiamo di arrivare in automobile a Lanvollon, un'altra città francese. Man mano che ci avviciniamo, è utile focalizzarsi sulla zona che vogliamo visitare, entrando sempre più nel dettaglio del territorio.

Bada che ho definito la astrazione come eliminazione di dettagli irrilevanti, per giungere a una rappresentazione che si concentri sugli aspetti comuni. In questo caso, passando dalla mappa a sinistra con la cornice verde, alla mappa a destra con la cornice blu, noi effettuiamo la operazione concettuale opposta, inseriamo sempre più dettagli…chiameremo questa operazione *raffinamento*.

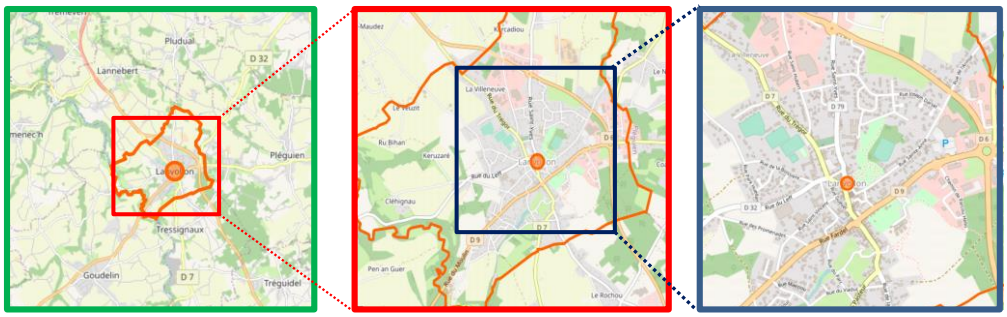


Figura 91 – Andare a Lanvollon

Spostandoci dalla rappresentazione nella cornice verde alla rappresentazione con cornice blu, noi rappresentiamo il territorio a *scala* sempre più piccola, introducendo sempre più dettagli.

Cosa significa il termine scala?

La scala è il rapporto tra una lunghezza, ad esempio un chilometro, nel territorio reale, e l'analoga lunghezza di quel territorio nella mappa. Per cui una scala 1:10.000 ci dice che un chilometro nella realtà è rappresentato con una lunghezza di 1/10.000 chilometri, cioè con una lunghezza di 10 centimetri (1000 metri/10.000).

Ebbene, vedi chiaramente come spostandoci dalla mappa rappresentata a sinistra a quella rappresentata a destra compaiono sempre più dettagli della città di Lanvollon, via via compaiono le strade principali, poi le strade secondarie, i loro nomi, le piazze, ecc.

I geografi hanno avuto da sempre il problema di rappresentare il territorio con diverse scale, a seconda dell'uso che si vuole fare delle mappe. Per esempio per andare in montagna, e individuare i sentieri, le diramazioni, le altitudini dei punti di interesse, è indispensabile una carta con scala *almeno* al 1:25.000.

Un altro esempio simile a quello della cartina in cui introduco i concetti di astrazione e di raffinamento è quello che voglio illustrarti qui di seguito. Supponi di essere ad Arabba, un paesino nelle Dolomiti, e di voler sapere che tempo farà la prossima settimana.

Per soddisfare la nostra esigenza, possiamo consultare tantissimi siti e tantissime app disponibili su telefono mobile; tra l'altro, Arabba ha un centro meteorologico di tutto rispetto,

gestito dalla Agenzia veneta ARPAV. Nella figura 92 vediamo quattro previsioni che possono essere acquisite da diversi siti; quelle con le icone del sole e delle nuvole sono acquisite da meteo.it, il testo proviene dalla agenzia ARPAV.

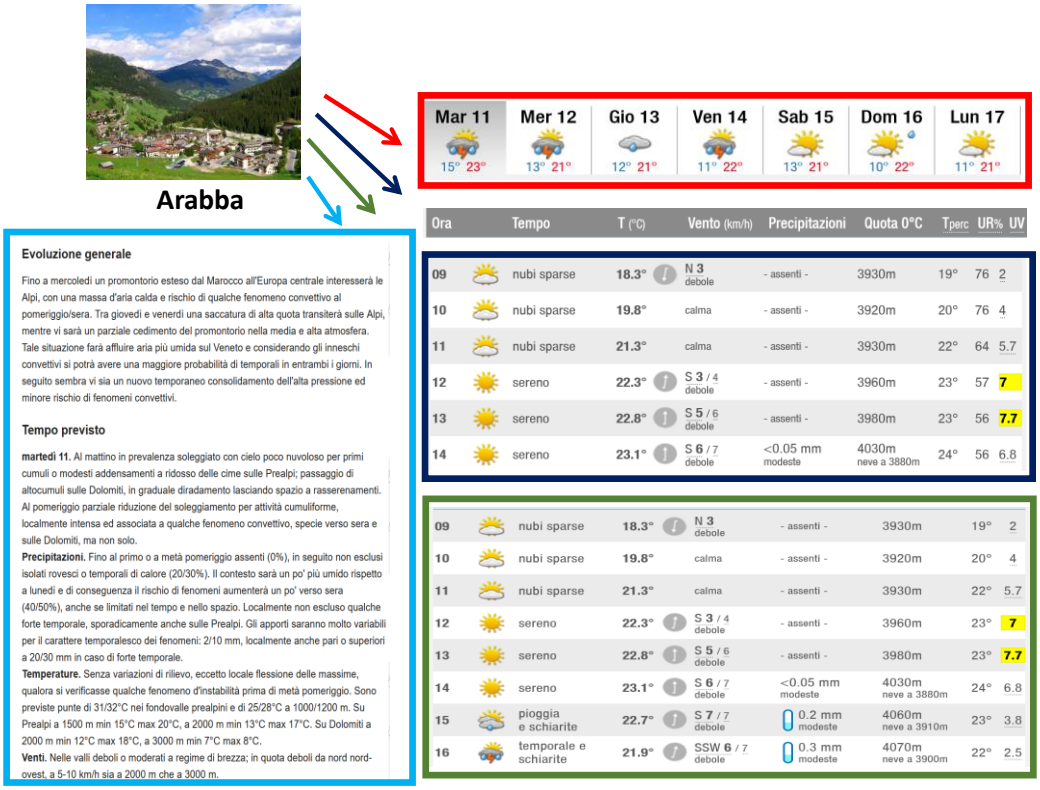


Figura 92 – Che tempo farà ad Arabba martedì prossimo?

La previsione con la cornice rossa è molto sintetica e ci fornisce le previsioni per i giorni che vanno da martedi 11 a lunedi 17 (di un certo mese di un certo anno), una previsione globale per ogni giorno. Questa previsioni vengono fornite per mezzo di un unico simbolo grafico, le cui possibili parti sono il sole, una nuvoletta bianca, una nuvoletta grigia, un simbolo per la pioggia e un fulmine.

Il significato di questi simboli si trova probabilmente sul sito di meteo.it, ma personalmente non sono mai andato a cercarlo, affidandomi a una interpretazione intuitiva: se c'è soltanto il sole significa che la giornata è bella e senza nuvole, se c'è sia il sole che la pioggia e il fulmine, interpreto che il tempo è variabile, con periodi soleggiati e periodi con temporale forte.

Le altre tre cornici forniscono per il solo giorno di martedì 11 il tempo, rispettivamente, dalle 9 alle 14 e dalle 9 alle 16, descritti con le icone già vista prima, più altre informazioni di contorno (racchiuse da cornice blu e verde), e infine quella in basso a sinistra (cornice azzurra) il tempo atmosferico descritto in maniera più analitica e documentata, mediante un testo in lingua italiana.

La situazione è molto simile a quella del Semaforo vecchio sul monte di Portofino, in cui *con il Semaforo* si trasmettono dati sintetici, codificabili con un alfabeto di segnali e *con il Telegrafo* dati più complessi e articolati, che si possono esprimere solo con un testo.

Dunque, tutte e tre le previsioni introducono dati di dettaglio rispetto alla previsione sintetica nella cornice rossa, abbiamo già chiamato questa operazione con il termine *raffinamento.* Un raffinamento è la operazione concettuale inversa della astrazione, una operazione, appunto, che introduce dettagli. I raffinamenti nelle cornici blu e azzurre ci possono essere utili quando abbiamo esigenze più specifiche, per esempio tutte le volte che, intendiamo fare una camminata con partenza alle 9, e vogliamo sapere che tempo farà per le prossime ore.

La relazione di *raffinamento* tra le quattro previsioni è meglio chiarita nella Figura 93, dove ho messo in evidenza le relazioni dirette di raffinamento. La previsione con cornice blu è un diretto raffinamento della previsione con cornice rossa; mentre la previsione con cornice verde è un raffinamento di quella con cornice blu.

Mentre la previsione blu può vedersi come un raffinamento *per disaggregazione*, in cui cioè io introduco dettagli sulle singole ore della giornata, dalle 9 alle 14, la verde può vedersi come un raffinamento *per estensione*, in cui estendo l'arco temporale, ma non introduco ulteriori dettagli. All'opposto, le due astrazioni definite in senso inverso si possono chiamare la prima, dalla cornice blu alla rossa, di *aggregazione*, astrazione che ho già introdotto, e la seconda di *contrazione*.

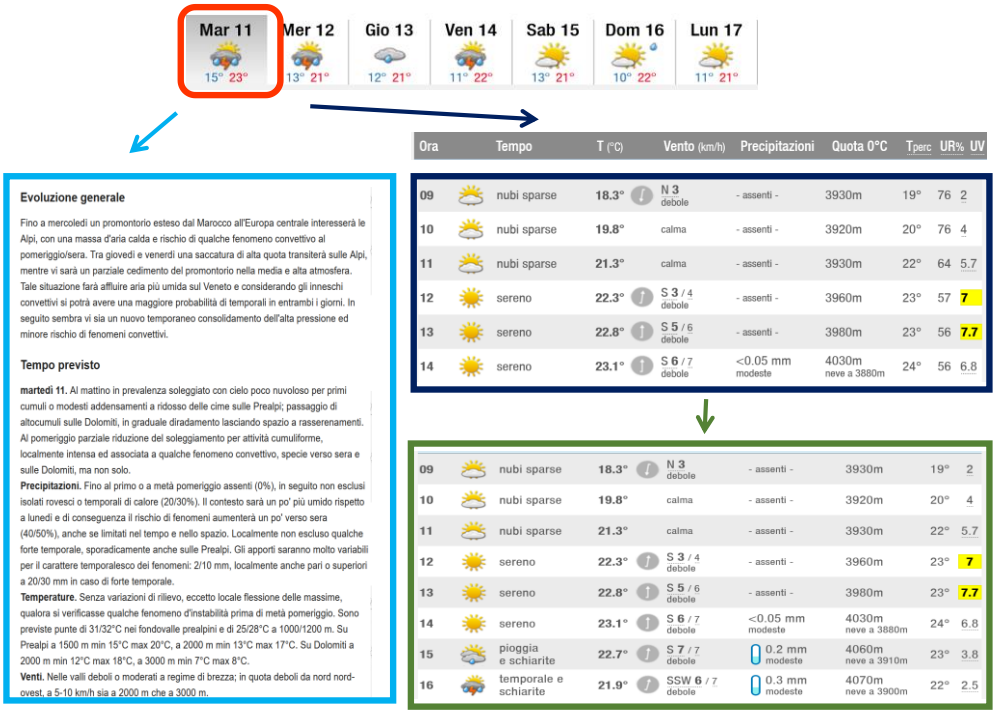


Figura 93 – Le previsioni messe in relazione di astrazione/raffinamento

Infine il testo nella cornice azzurra può vedersi come un raffinamento di natura diversa, non facilmente confrontabile con quelli blu e verde; dovremmo infatti estrarre tutti i dati inclusi nel testo e confrontare questi dati con quelli nelle cornici blu e verde. Ad esempio nel testo si

usa il termine *mattino* che può essere messo in relazione con l'intervallo temporale dalle 8 alle 13; mattino e 9-13 sono certamente confrontabili, ma non coincidenti.

Se ti interessa, puoi fare da solo questo esercizio di confronto tra i dati compresi nel testo e quelli nelle cornici blu e verse, non lo potrò controllare, ma è interessante anche farlo da soli, senza la soluzione.

Va bene, ci proverò.

Insomma, abbiamo imparato dal precedente esempio che i dati che troviamo nel Web o tramite le app dei nostri telefoni mobili descrivono un frammento di mondo a diversi livelli di sintesi (*astrazione*) o, visti all'inverso, a diversi livelli di dettaglio (*raffinamento*) e questi diversi livelli di astrazione/raffinamento possono essere *messi in relazione* tra loro (es., questo dato è una astrazione di questi altri dati).

Le astrazioni nelle mappe

Anche nelle mappe noi possiamo individuare vari tipi di relazioni di astrazione/raffinamento. Ad esempio, vedi Figura 94, se partendo da una scala al 25.000 il nostro scopo è quello di generare una mappa a scala maggiore, ad esempio al 40.000, noi possiamo astrarre il gruppo di caseggiati della parte sinistra della figura, caratterizzati da forme articolate con angoli e rientranze, nel gruppo di caseggiati nella parte destra, in cui le forme degli edifici sono più stilizzate.

Anche questa trasformazione rispetta la definizione di astrazione che ho dato in precedenza, astrazione come eliminazione di dettagli e messa in evidenza di caratteri comuni, che in questo caso sono le forme a rettangolo ovvero fatte come le lettere L o a T. Possiamo anche dire che gli edifici nella parte destra hanno forme più generali di quelle nella parte sinistra, e quindi la relazione tra le due mappe è la stessa che sussiste tra Albero e Pino, l'astrazione di *generalizzazione*.

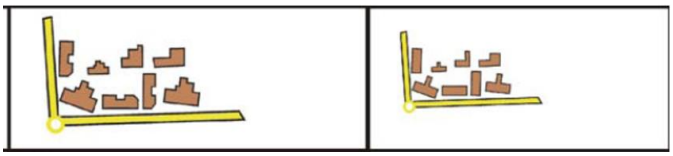


Figura 94 – Astrazione di generalizzazione

Un'altra trasformazione di astrazione/raffinamento compare in Figura 95; a sinistra abbiamo un disegno abbastanza realistico di stazione ferroviaria in cui sono evidenziati fasci di binari in ingresso e in uscita dalla stazione, a destra abbiamo due simboli di dimensioni più piccole. Il simbolo di locomotiva usa la metafora della sineddoche …

Cosa è una sineddoche? Se uno non ti mette a bada, certre volte fai dei voli pindarici…

La sineddoche è la figura retorica che usiamo quando rappresentiamo il tutto (la stazione ferroviaria) con una sua parte (la locomotiva) Questa astrazione è chiamata di *simbolizzazione*.

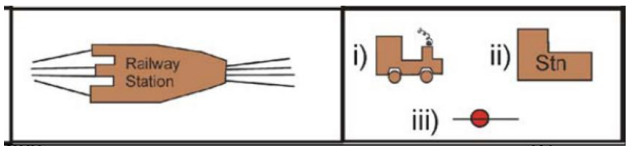


Figura 95 - Astrazione di simbolizzazione

L'astrazione di simbolizzazione viene utilizzata nella parte destra di Figura 96 per introdurre un altro importante concetto. In questo caso, partendo dal frammento di mappa che compare nella parte inferiore della figura, noi possiamo rappresentare il frammento in maniera più compatta con due filosofie completamente diverse.

Da una parte, caso a sinistra, possiamo attuare una *riduzione fotografica*, che cambia semplicemente la scala della mappa, *senza effettuare alcune trasformazione* sui simboli che rappresentano strade, piazze, e la stazione ferroviaria.

Ci sono due limiti percettivi legati a questa trasformazione di astrazione, uno è quello dell'occhio umano, che non riesce a distinguere oggetti o simboli più piccoli di una certa soglia, e l'altro è quello *della unità minima di colore* rappresentabile con un *pixel*, che è *l'unità minima di informazione* nelle foto o mappe stampate su carta e digitali.

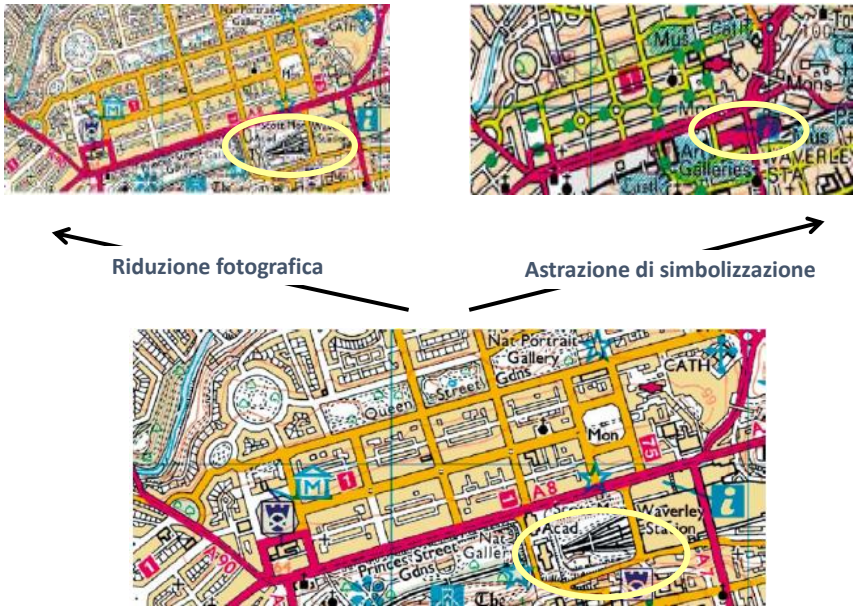


Figura 96 – Dire una piccola bugia per riuscire a dire parte della verità

Nel caso a destra è usata *l'astrazione di simbolizzazione*, nel senso che è stata applicata la trasformazione di Figura 95, ed è stata sostituita la rappresentazione realistica della stazione con quella simbolica geometrica che compare in figura. Inoltre, le strade hanno una

rappresentazione semplificata, e la dimensione delle strade principali è volutamente ingrandita rispetto all'originale.

Possiamo dire che la rappresentazione a sinistra, ottenuta con una riduzione fotografica, è *più fedele all'originale* di quella a destra, ma la riduzione vanifica questa fedeltà all'originale, perché alla fine la mappa diventerà simile a quella di Lione, troppo densa per capirci qualcosa.

La rappresentazione a destra è meno fedele all'originale, ma più chiara e informativa. I geografi tendono a dire: *certe volte nelle carte si deve dire una piccola bugia, per riuscire a dire parte della verità…..*

Il diluvio o sovraccarico di dati (data overload)

Vediamo ora un altro grande problema suscitato dai big data, rispetto al quale le astrazioni ci possono aiutare. Quando facciamo delle ricerche, per esempio stiamo scrivendo una tesina per la scuola, o vogliamo saperne qualcosa di più su un evento, un tema o un personaggio, lo spazio che via via si apre alla nostra ricerca può essere immenso.

Ad esempio quando scrissi un libro qualche tempo fa sulla qualità dei, le mie ricerche bibliografiche si allargavano via via che leggevo nuovi articoli, a un certo punto mi stancai di proseguire e misi un punto sulle ricerche; mi accorsi di aver raccolto 700 articoli, tutti citati nella bibliografia.

Insomma, spesso le nostre esplorazioni della realtà portano ad ampliare lo spazio della conoscenza secondo la dinamica di Figura 97.

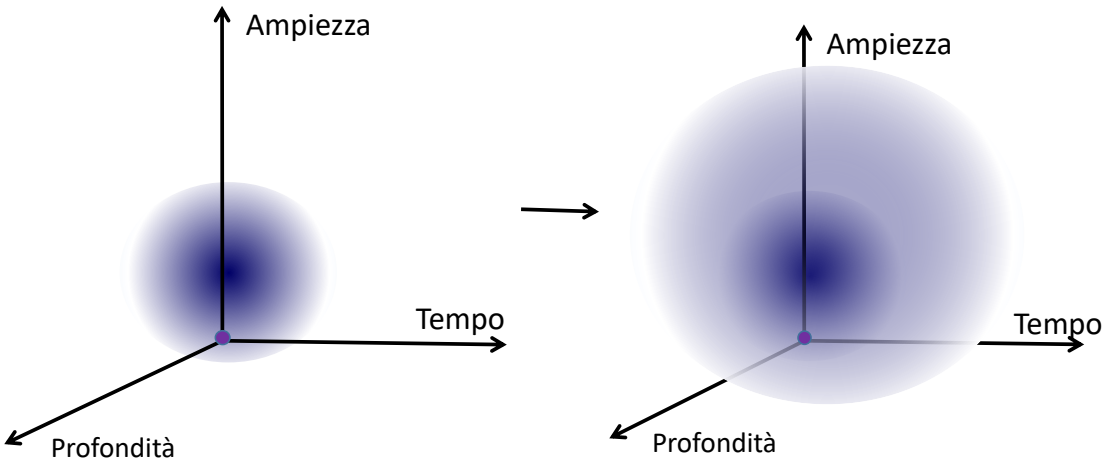


Figura 97 – Le nostre ricerche sul Web possono essere senza fine ….

All'inizio cominciamo con uno spunto, rappresentato simbolicamente a sinistra nella figura con un piccolo pallino all'origine dello spazio Ampiezza-Profondità-Tempo. Lo spunto può essere, ad esempio, il progetto di una tesina su Garibaldi. A questo punto, il nostro spazio di ricerca si amplia; cercando su Google vari possibili riferimenti su Garibaldi, il primo che consulteremo sarà con ogni probabilità Wikipedia.

Ma se vogliamo saperne di più, via via amplieremo lo spazio della ricerca, estendendo progressivamente la conoscenza, che nella figura è rappresentata simbolicamente con la sfera che si ingrandisce. Questo spazio si amplierà, e, man mano che si amplia, cresceranno le connessioni con altri frammenti di conoscenza; proprio come è accaduto a me con i 700 lavori di ricerca.

Insomma, molte ricerche connesse con la scoperta di nuovi dati utili presentano la dinamica di Figura 98. All'inizio, quando abbiamo pochi dati tendiamo a esplorare il Web, a consultare una biblioteca, e questa prima fase porta a nuove esplorazioni e nuove acquisizioni di dati utili.

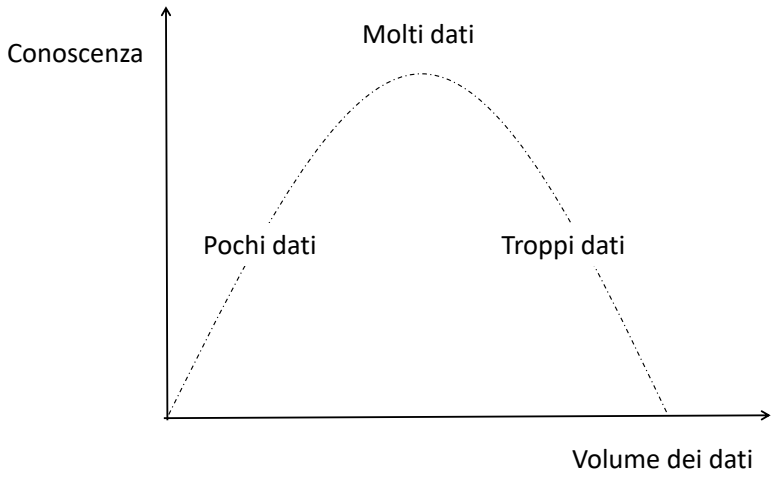


Figura 98 – Il sovraccarico di dati, o data overload

Ma c'è un momento, meglio, *ci può essere un momento,* in cui scopriamo che nuove esplorazioni cominciano a peggiorare la situazione, cominciano a disorientarci, e addirittura possono farci perdere il "bandolo della matassa" costituito dalla conoscenza che avevamo accumulato e organizzato. Come quando leggiamo un libro con tanti possibili personaggi, o un giallo con tanti potenziali assassini.

Il fenomeno, ben noto nelle scienze cognitive, è chiamato *sovraccarico di dati*, in inglese *data* o *information overload*. Attenzione, la curva di Figura 98 è un esempio limite di sovraccarico di dati, accade più spesso che, magari per stanchezza, ci fermiamo nella fase crescente…..

Insomma, il messaggio che mi piacerebbe condividere è che il Web è una grande fonte di dati, ma cerchiamo nelle nostre ricerche di difenderci dalla valanga di messaggi, di stimoli, di siti, di news, di blob, cerchiamo di filtrare ogni volta i dati più significativi e utili per noi.

Un modo con cui possiamo effettuare questo filtro sono le *astrazioni*. Cerchiamo ogni volta che accumuliamo conoscenza di eliminare i dettagli, di evidenziare le informazioni più importanti.

Io ad esempio, quando scrivo un libro o un articolo, scrivo prima un breve abstract degli articoli che raccolgo, poi divido gli abstract per tema, infine classifico i temi in un indice, insomma cerco di usare meccanismi basati sulle astrazioni.

La stessa cosa accade nella organizzazione delle cartelle nel nostro pc o tablet. Dapprima creiamo tante cartelle diverse tutte allo stesso livello di importanza, poi quando ci accorgiamo che il numero di documenti nelle diverse cartelle sta diventando troppo grande, suddividiamo le cartelle in sotto cartelle, oppure accorpiamo le cartelle in cartelle più grandi. Se vi va, e avete un computer vostro con tante cartelle, provate a comprenderne la struttura concettuale e provate a trasformare le cartelle attuali in una struttura più razionale che usi astrazioni del tipo di quelle che ho iontrodotto.

Potete anche provare a fare un esercizio, che consiste nel trovare un testo di cica cinque pagine su:

- i dati sulle previsioni metereologiche,
- i testi che forniscono le recensioni su un cantante o su film o su una serie televisiva,
- qualunque argomento vi piaccia.

Provate a sintetizzarlo in due pagine, e poi in una, e poi in dieci righe. Seguite il processo mentale di eliminazione dei dettagli, e interrogatevi sulle diverse forme di astrazione che state usando.

Le astrazioni nei grafi

Concludo questo capitolo con le astrazioni nei grafi. Abbiamo già visto esempi di grafi nel capitolo sui modelli, quando abbiamo parlato dei grafi semantici. Un *grafo* consiste in un insieme di vertici e un insieme di archi che li collegano, vedi in Figura 99 un esempio di *grafo non etichettato*, cioè senza nomi o etichette, con sei nodi e nove archi.

Vediamo nella Figura 99 tre grafi. Il primo, quello con sei nodi e nove archi, è disegnato in maniera esteticamente gradevole, i nodi sono distanti tra loro, gli archi sono segmenti diretti, senza piegamenti, non ci sono sovrapposizioni tra archi, si suole dire che il grafo è *planare*, cioè è rappresentato sul piano senza incroci tra gli archi.

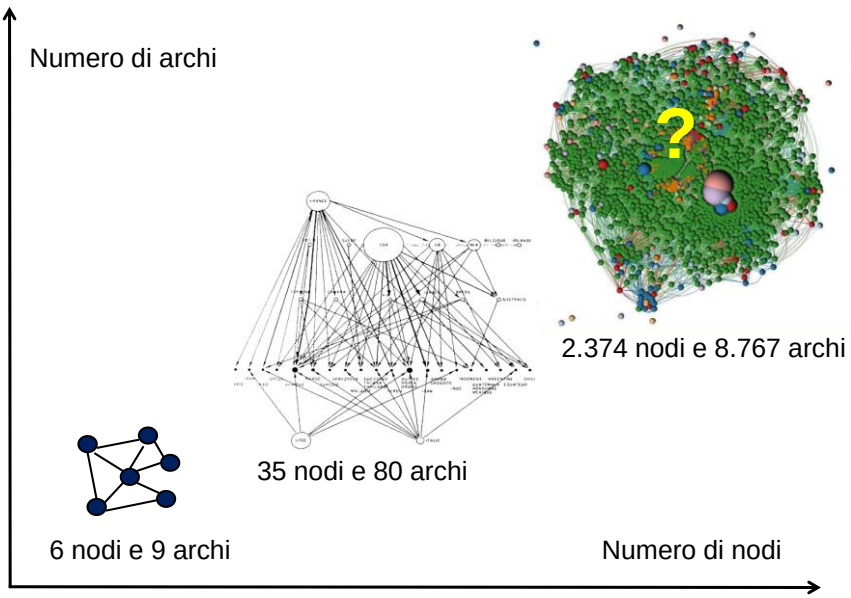


Figura 99 – Come controllare la complessità dei grafi sempre più grandi

Vi può sembrare una pignoleria discutere di grafi senza incroci tra gli archi, ma sono stati fatti molti studi sulla maggiore leggibilità dei grafi planari rispetto ai grafi con incroci Pensate solo come sarebbe poco leggibile il grafo semantico di Figura 57, che rappresenta le relazioni tra i personaggi dei Demoni, se fosse disegnato con tutti gli archi sovrapposti: non si capirebbe niente!

Se passiamo al secondo grafo, quello al centro della figura, con 35 nodi e 80 archi, vediamo che il disegnatore qui si è arreso, non ha più cercato di evitare gli incroci, era impossibile. Ha seguito però altre regole per controllare la complessità: ha rappresentato più grandi i nodi da cui esce un maggior numero di archi, ha rappresentato i nodi su fasce verticali, per dare loro una qualche forma di ordine. Ha cercato di contrastare la complessità, senza discostarsi dallo stile nodi/cerchi e archi/segmenti.

Se invece guardiamo il grafo in alto a destra, la rappresentazione è veramente caotica. D'altra parte, cosa ci potevamo aspettare nel caso di un grafo di 2374 nodi e 8767 archi?

Rappresentare il terzo grafo è impossibile. Avremmo bisogno di un foglio grandissimo, ci sarebbero tanti archi che devono fare un percorso caotico per collegare due nodi, e avremmo un numero innumerevole di incroci che non ci farebbero capire niente.

In questo caso le astrazioni sono una via obbligata. L'astrazione più usata per i grafi è il *clustering,* termine che in italiano si può tradurre *raggruppamento*. In Figura 100 vediamo un esempio di uso del clustering, in cui dapprima decidiamo un raggruppamento dei nodi secondo un determinato criterio, e poi fondiamo in un unico nodo l'insieme dei nodi di uno stesso gruppo.

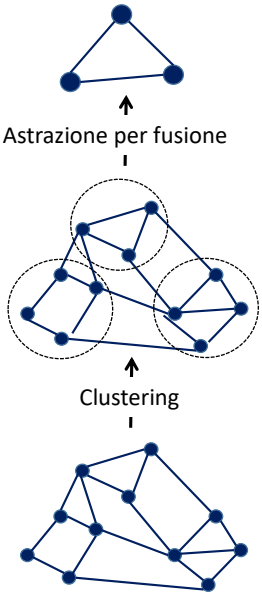


Figura 100 – L'astrazione di clustering

Proviamo ora ad applicare il clustering all'esempio dei Demoni, utilizzando la seconda parte della descrizione dei personaggi che abbiamo visto nel Capitolo 6, e che per comodità riproduco in Figura 101.

> Nella Russia del secondo '800 vive la nobile Varvara Petrovna Stavrogina, che è legata da profonda e platonica amicizia e mantiene economicamente lo scrittore e poeta incompreso Stepan Trofimovič Verchovenskij. Il figlio di Stepan, Petr Stepanovic Verchovenskiji, cresce lontano dal padre che è il tutore del figlio di Varvara, Nikolaj Vsevolodovic Stavrogin.
>
> *Entrambi i figli, cresciuti e dopo una lunga assenza all'estero, tornano a casa, ove tramano per compiere attentati. Nikolaj, detto anche Nicolas, è l'ideologo ispiratore di Petr, il quale coordina e comanda una "cinquina" di cospiratori composta da Virginsky, Sigalef, Liputin, Tolkacenko e Ljamsin.*

Figura 101 – Descrizione sintetica dei principali personaggi dei Demoni, e dei legami tra i personaggi

Nel secondo paragrafo compaiono cinque nuovi personaggi, detti i *cospiratori*. Ora, se dovessimo rappresentare la relazione di comando tra Petr e i cinque cospiratori dovremmo creare cinque relazioni, ma lo stesso testo ci suggerisce di aggregarli nella "cinquina". Il risultato è mostrato nella Figura 102, dove *Cinquina* è appunto un cluster dei cinque nuovi personaggi. Notate che in questo modo ho arricchito il modello utilizzato per il grafo semantico con un nuovo concetto, che corrisponde al *clustering di un insieme di concetti*, in questo caso cinque nodi, ognuno dei quali eredita tutti i legami con il resto del grafo semantico, in pratica la relazione *comanda*.

Per completezza, ho anche aggiunto la relazione tra Stepan e Petr e il soprannome di Nicolaj.

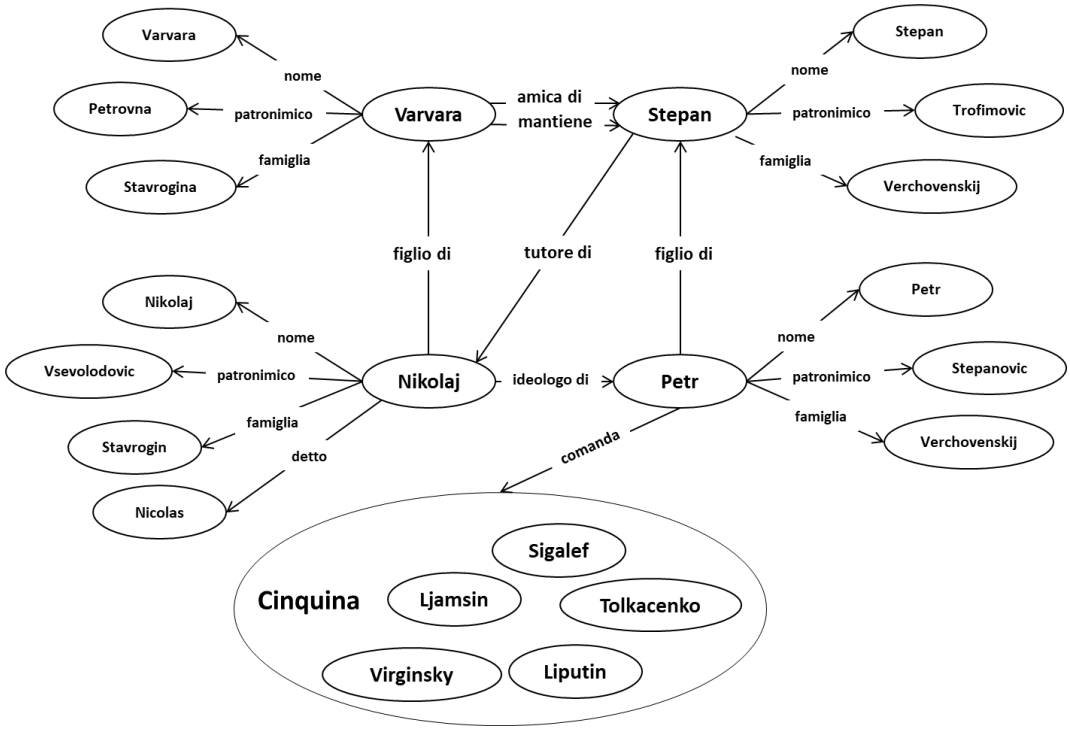


Figura 102 - La rete semantica dei Demoni con i nuovi personaggi aggregati nella *Cinquina*

Nella Russia del secondo '800 vive la nobile Varvara Petrovna Stavrogina, che è legata *da profonda e platonica amicizia e mantiene economicamente lo scrittore e poeta incompreso Stepan Trofimovič Verchovenskij. Il figlio di Stepan, Petr Stepanovic Verchovenskiji, cresce lontano dal padre che è il tutore del figlio di Varvara, Nikolaj Vsevolodovic Stavrogin.*
Entrambi i figli, cresciuti e dopo una lunga assenza all'estero, tornano a casa, ove tramano per compiere attentati. ***Nikolaj****,* ***detto anche Nicolas****, è l'ideologo* ***ispiratore di Petr****, il quale coordina e* ***comanda una "cinquina" di cospiratori composta da Virginsky, Sigalef, Liputin, Tolkacenko*** *e* ***Ljamsin****.*

Figura 103 - Descrizione annotata in ***neretto/corsivo***

Trovi in Figura 103 il testo corrispondente al grafo semantico annotato secondo l'usuale convenzione.

Come comprendere meglio un grafo semantico attraverso un processo di raffinamento

Se osserviamo con attenzione il grafo semantico di Figura 102, ci rendiamo conto che non è facile coglierne il significato complessivo con un solo sguardo; questo perché rappresenta 23 nodi, di cui 22 tutti dello stesso tipo e il ventitreesimo costituito dal cluster.

Per *comprendere* il grafo semantico, possiamo *generarlo per passi successivi*. L'idea che applichiamo è che i concetti complessi possono essere costruiti per passi, partendo da una rappresentazione astratta, e quindi semplice da capire, e generando le varie parti tramite trasformazioni di raffinamento.

Osserviamo la Figura 104. In questo caso il processo generativo è rappresentato per mezzo di due grafi semantici, quelli nelle cornici rossa e blu, il secondo dei quali può essere ottenuto dal primo sostituendo il nodo *Cinquina* con il cluster che già abbiamo descritto.

Nella parte sinistra della figura ho rappresentato il processo di raffinamento in forma astratta mediante un cono, in cui sono rappresentate due superfici, corrispondenti ai due grafi. L'area delle due superfici è orientativamente proporzionale alla dimensione del grafo, intesa come numero di concetti rappresentati. In questo caso le due aree hanno valore simile, perché il raffinamento riguarda il solo concetto *Cinquina*.

Nella Figura 105 compare un secondo raffinamento, che adotta convenzioni simili. In questo caso, il secondo grafo semantico è significativamente più complesso del primo, perché è ottenuto raffinando oltre il concetto Cinquina anche i concetti corrispondenti ai quatto personaggi. Corrispondentemente la superfice associata al primo grafo è significativamente più piccola rispetto al caso precedente.

Bello, elegante, e davvero semplice da utilizzare….

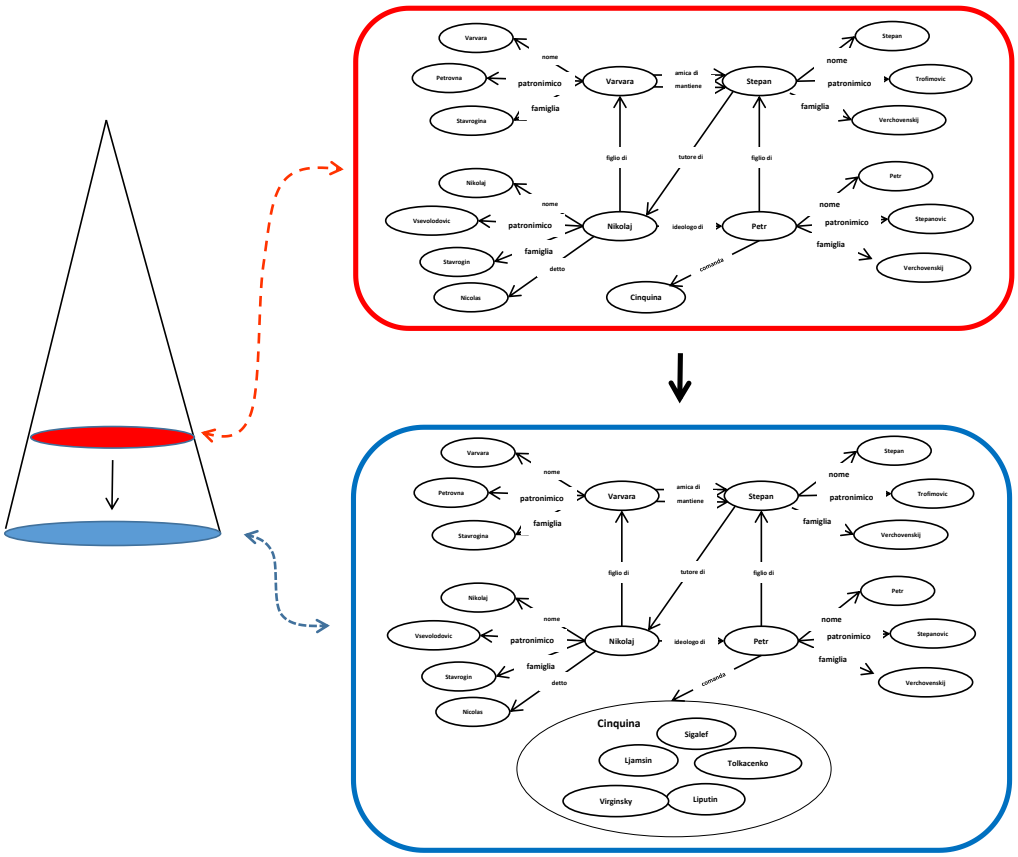


Figura 104 - Grafo semantico rappresentato con due piani di raffinamento – versione con piani vicini

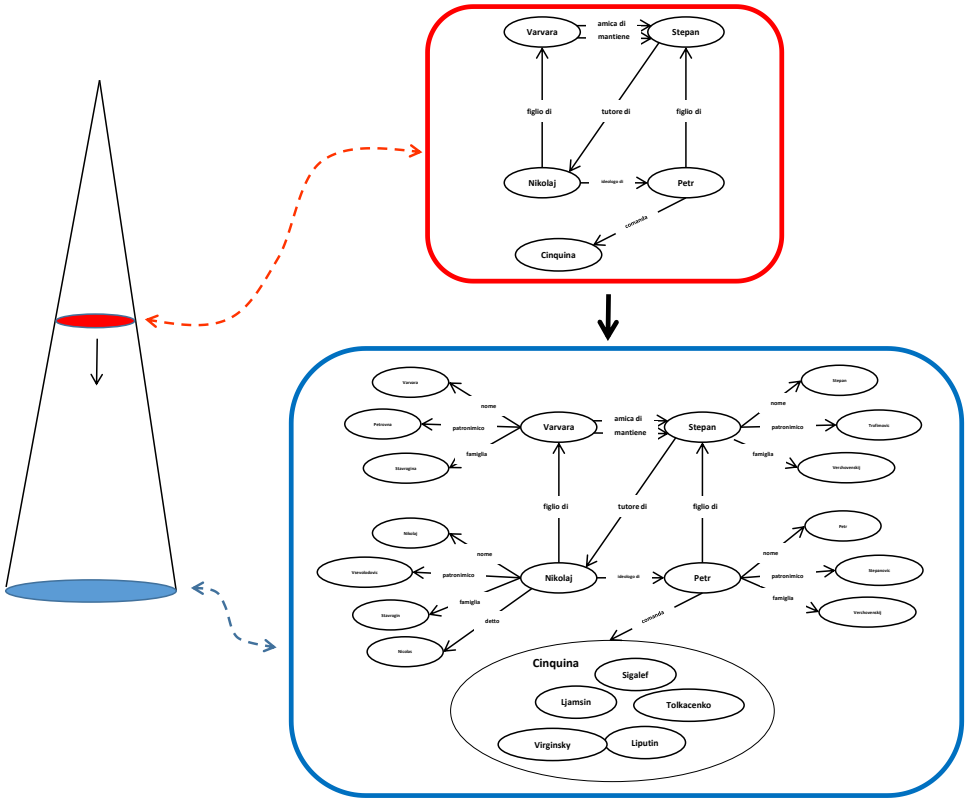


Figura 105 - Grafo semantico rappresentato con due piani di raffinamento – versione con piani distanti

Sì, i casi reali tuttavia non sono sempre così semplici. Provate a calcolare quante operazioni di clusterizzazione servono per un grafo di 1.024 nodi nel caso i diversi cluster siano sempre formati da due nodi. Provate poi a fare il calcolo nel caso generale in cui il grafo abbia m nodi e i cluster contengano n nodi. Le risposte nella prossima pagina.

Per la prima domanda, dobbiamo cercare il minimo k per cui 2^k è maggiore o uguale a 1.024; in questo caso k è uguale a 10.

Per la seconda domanda, il valore cercato è il minimo k per cui n^k è maggiore o uguale a m.

Per concludere, le astrazioni, nonostante il nome che ce le fa immaginare "tra le nuvole", poco concrete, sono uno strumento mentale che in realtà siamo abituati a usare ogni giorno, e che risulta molto utili per focalizzare l'attenzione sugli aspetti rilevanti di un problema che riguardi i dati, trascurando i dettagli non rilevanti; inoltre, le astrazioni ci forniscono una difesa efficace per il fenomeno del data overload.

Riassumendo

Le ***astrazioni*** sono uno strumento che usiamo spessissimo nella nostra vita, soprattutto nel linguaggio verbale o scritto, quando dobbiamo dare un nome a un oggetto o a un concetto. Quando usiamo una ***astrazione*** o un ***processo di astrazione***, noi trascuriamo gli ***aspetti non rilevanti*** (ad esempio per un pino le varie tipologie, pino romano, pino marittimo ecc.) per mettere in evidenza gli ***aspetti comuni*** (sono tutti pini). Le astrazioni più usate sono la ***generalizzazione*** (es tutti i pini romani sono pini, quindi pino è una generalizzazione di pino romano), la ***aggregazione*** (un pino è formato da una radice, un fusto, un insieme di rami, un insieme di foglie, quindi pino è aggregazione di radice, fusto, rami, foglie), e il ***clustering***, in cui, ad esempio per i ***grafi***, un insieme di nodi caratterizzati da qualche similitudine sono sostituiti da un unico nodo.

Il procedimento opposto alla astrazione è il ***raffinamento***, in cui, al contrario della astrazione, un concetto è sostituito da un altro concetto in cui sono introdotti dettagli. Il raffinamento associato alla generalizzazione è la ***estensione***, il raffinamento opposto alla aggregazione è la ***disaggregazione***.

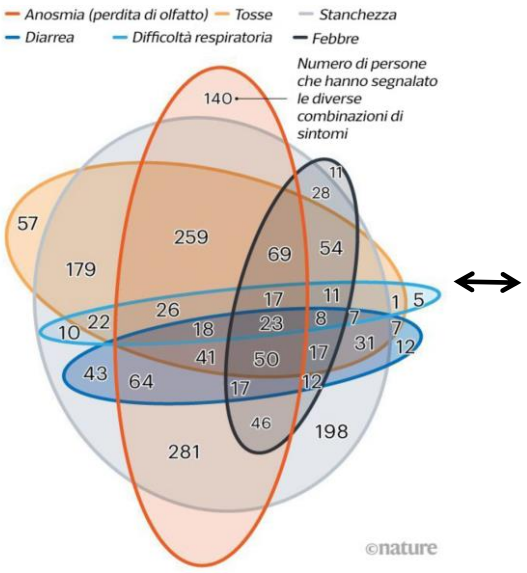
Usando astrazioni e raffinamenti, noi possiamo rappresentare un oggetto complesso (ad esempio il grafo semantico dei personaggi dei Demoni) attraverso un processo generativo, in cui (continuando nell'esempio) un grafo semplice viene via raffinato in grafi sempre più dettagliati.

L'astrazione è anche utile per contrastare il ***data overload***, quando cioè, ad esempio in una ricerca bibliografica, abbiamo troppi dati. I dati disponibili possono essere organizzati in livelli di astrazione, per esempio rappresentando ognicarticolo con un breve riassunto o abstract.

Capitolo 10

Una immagine è meglio di mille parole
La visualizzazione dei dati

Si suole dire che una immagine è meglio di mille parole; si vuole intendere che le strutture grafiche, i simbolismi, le metafore usate nelle immagini richiamano alla mente gli aspetti del mondo reale in modo molto più efficace di quanto accada con una descrizione testuale. La Figura 106 conferma in maniera inequivocabile il motto.



Anosmia	Tosse	Stanchezza	Diarrea	Difficoltà respiratoria	Febbre	Numero di persone
si	no	no	no	no	no	140
no	no	no	no	no	si	11
no	no	si	no	no	si	28
no	si	No	no	no	no	57
no	no	no	no	no	si	179
no	no	si	no	no	si	259
no	si	No	no	no	no	69
no	no	no	no	no	si	140
no	no	si	no	no	si	11
no	si	No	no	no	no	28
no	no	no	no	no	si	57
no	no	si	no	no	si	179
no	si	No	no	no	no	259
no	no	no	no	no	si	69
no	no	no	no	no	si	57
no	no	si	no	no	si	179
no	si	No	no	no	no	259
no	no	no	no	no	si	69
no	no	si	no	no	si	140
no	si	No	no	no	no	11
no	no	no	no	no	si	28
no	no	si	no	no	si	57
no	si	No	no	no	no	179
no	no	no	no	no	si	259
no	no	no	no	no	si	57
no	no	si	no	no	si	57
no	no	no	no	no	si	179
no	si	No	no	no	no	259
no	no	no	no	no	si	69
no	no	si	no	no	si	140
no	si	No	no	no	no	11
no	no	no	no	no	si	28
no	no	si	no	no	si	57
no	si	No	no	no	no	179
no	no	no	no	no	si	259

Figura 106 – Un diagramma e una tabella che rappresentano gli stessi dati

Nella parte sinistra della figura un diagramma, chiamato anche *diagramma di Venn*, mostra i risultati di una ricerca su un gruppo di malati della prima ondata di marzo – maggio 2020 della epidemia di Covid, che intendeva contare il numero di casi affetti da sintomi di varia natura, per singoli sintomi e per combinazioni di sintomi.

Le diverse ellissi di vario colore corrispondono ai diversi sintomi; ogni spazio che si trova alla intersezione tra diverse ellissi ha associato un numero di casi, e ogni singolo frammento di

ellissi o intersezione di ellissi corrisponde a una riga della tabella a destra. Ti prego di non munirti di lente di ingrandimento, perché i numeri e le combinazioni si/no che compaiono a destra sono un po' messi lì a caso…..

La tabella a destra, dunque, rappresenta esattamente gli stessi dati del diagramma a sinistra; ma se la guardo per cercare di capire qualcosa, non so voi, ma a me viene il mal di testa! Infatti la presenza o assenza di sintomi è ora segnalata in ogni riga da *si/no*, e quindi per capire quanti sono i casi associati, per esempio, a *stanchezza* e *febbre* insieme io devo cercare la riga in cui ci sono i sì sono solo in corrispondenza dei due simboli.

Tipi di visualizzazioni

Ci sono tantissimi tipi di visualizzazioni usate nei libri, nei giornali, nei siti Web, in cui la fantasia dei grafici si esprime in maniera straordinariamente ricca; ci vorrebbe una Enciclopedia da dedicare solo alle visualizzazioni dei dati digitali! Nella Figura 107 vediamo una classificazione dei principali tipi di visualizzazioni; per ogni tipologia si individua un esempio che la rappresenta.

Tabelle – Sono le più semplici tipologie di visualizzazioni, le abbiamo viste per la prima volta nella Figura 7, e le abbiamo discusse con qualche dettaglio nel Capitolo 6. Sono costituite da un insieme di righe e di colonne dove vengono inseriti i valori, la prima riga nelle varie colonne fornisce le proprietà dei dati, o attributi, a cui vanno associati i valori.
Grafici (chart in inglese) – Mostrano funzioni rappresentate mediante linee in un piano cartesiano, ovvero elementi geometrici, quali rettangoli, cerchi, sfere, altre superfici, in piano o in prospettiva, per descrivere dati statistici di varia natura. La Figura 7 mostra diversi esempi di grafici. I grafici sono anche utilizzati per rappresentazioni grafiche di modelli di dati, come ad esempio i grafi semantici.
Diagrammi – Adottano un alfabeto di simboli e collegate rappresentazioni grafiche, che devono rispettare determinate regole di composizione. Un esempio compare nella Figura 108 ; il diagramma rappresenta un grafo di flusso, che esprime in maniera visiva l'algoritmo per il calcolo della somma dei primi dieci numeri interi rappresentato mediante un linguaggio programmativo nella Figura 21.
Mappe – Descrivono rappresentazioni del territorio in cui vengono evidenziati un ampio insieme di caratteristiche del territorio, ad esempio gli edifici o le strade, mediante simboli specifici. Abbiamo visto esempi di mappe per la prima volta nella Figura 7 e nel Capitolo 7 sulla qualità dei dati.
Icona – Nella accezione di questo testo, è un simbolo che descrive attraverso un legame di somiglianza o una trasposizione metaforica un frammento di mondo (ad esempio un simbolo di ospedale come in Figura 83, dove le metafore adottate sono il letto e il simbolo della croce rossa).
Visualizzazione mista – Visualizzazione che utilizza due o più delle precedenti tipologie. Per esempio, nella Figura 7 sono rappresentate una mappa dell'Italia e un insieme di grafici.

Figura 107 – Tipi di visualizzazioni

In Figura 108 vediamo diversi tipi di visualizzazioni; ti propongo di individuare per ogni visualizzazione il tipo corrispondente (il diagramma è stato già individuato nella definizione precedente). La soluzione nella prossima pagina.

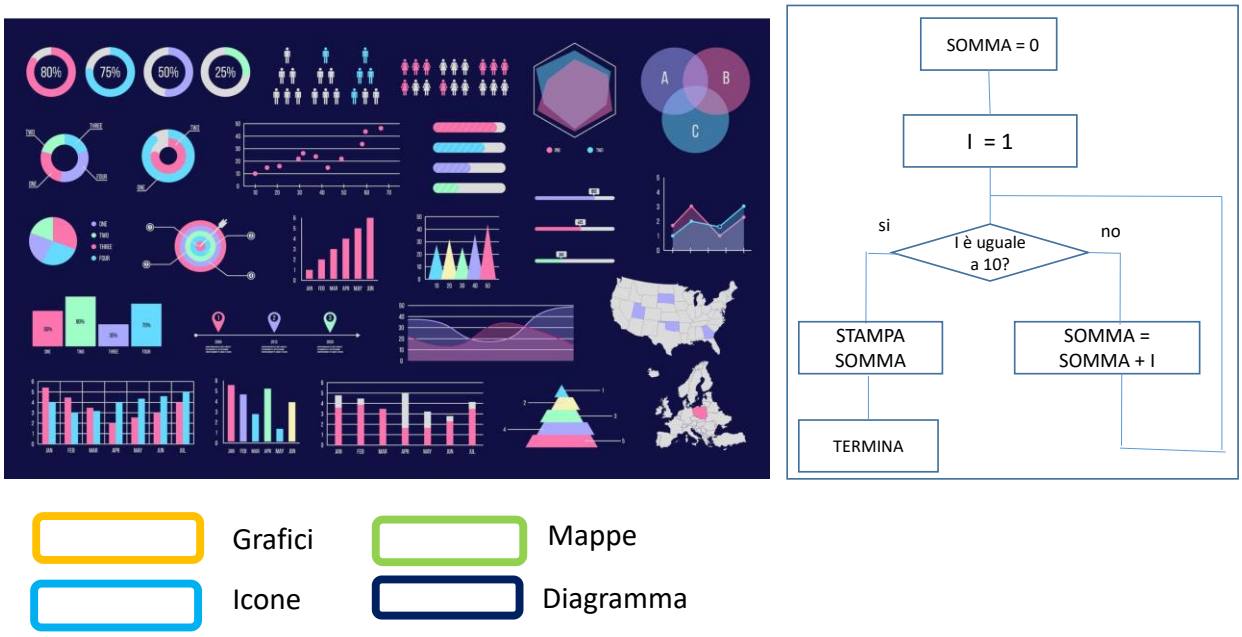


Figura 108 – Visualizzazioni oggetto della domanda
(ripresa in parte da https://boostlabs.com/blog/10-types-of-data-visualization-tools/)

Ecco la soluzione nella Figura 109.

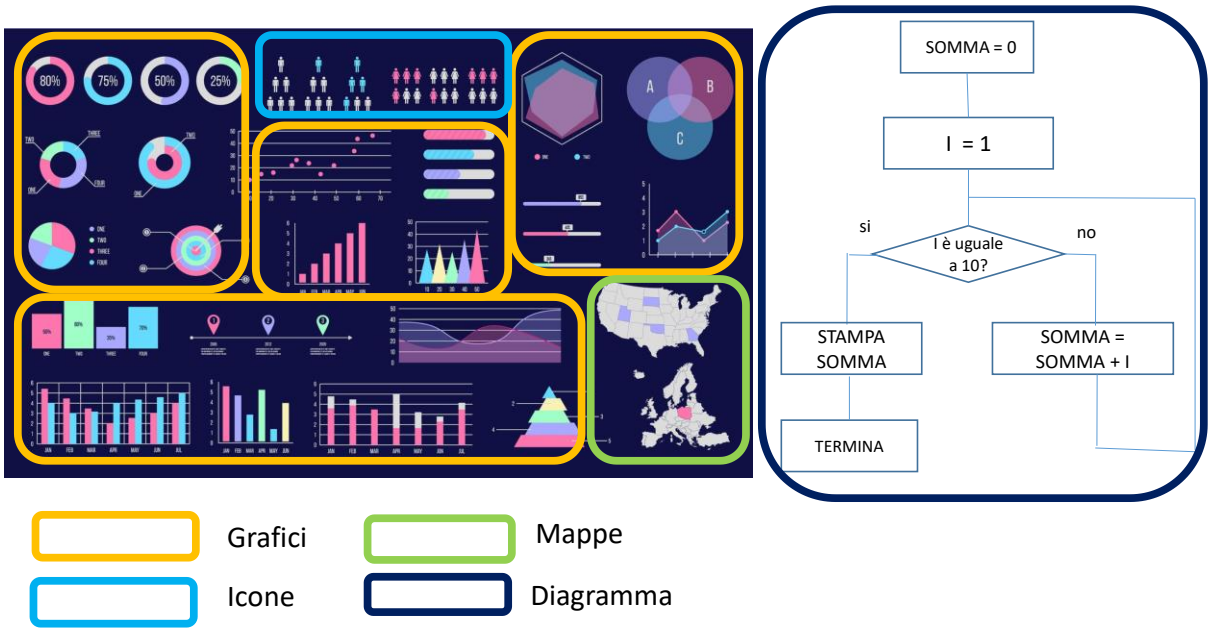


Figura 109 - Esempi di diversi tipi di visualizzazioni - soluzione

Una disciplina che usa molte visualizzazioni è la *Statistica descrittiva*, di cui abbiamo già parlato nel Capitolo 8, e che studia i criteri di rilevazione, classificazione, sintesi e rappresentazione dei dati acquisiti e analizzati nello studio di una *popolazione* o di una parte di essa, chiamata *campione*.

Molte visualizzazioni usate nella statistica descrittiva e nella matematica, sono rese disponibili da Excel, una applicazione diffusa in tutti i personal computer e tablet. In Figura 110 vediamo i tipi di grafici e di diagrammi (nella accezione di Figura 107) rappresentabili in Excel.

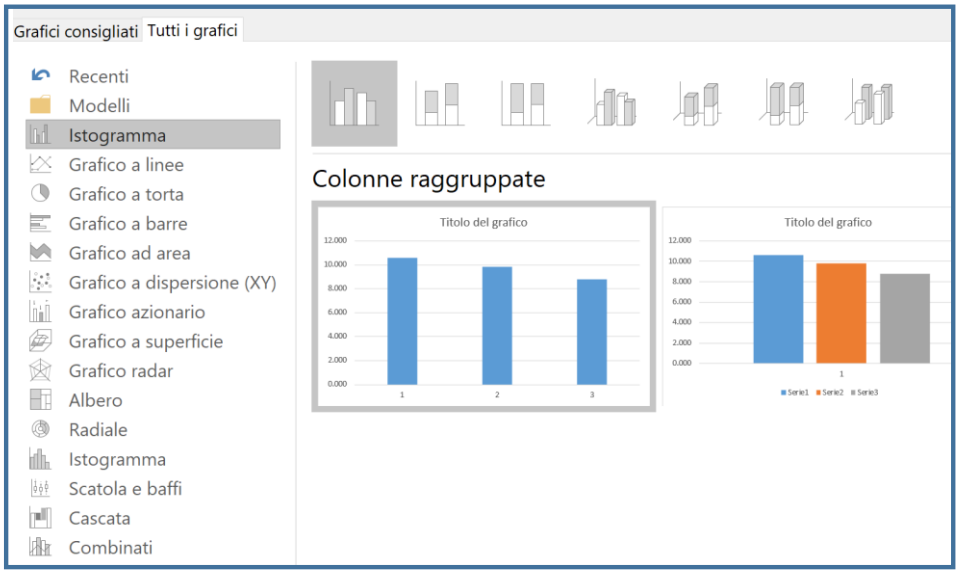


Figura 110 - Visualizzazioni per dati statistici in Excel

Le visualizzazioni, attraverso l'uso di figure geometriche, simboli grafici e icone hanno un grande potere espressivo, e ci forniscono una visione di insieme dei dati che rende più immediato effettuare vari tipi di analisi. Vediamone alcune "fatte in casa".

Le visualizzazioni durante la pandemia Covid

Durante la seconda ondata del Covid, a cavallo tra l'anno 2020 e il 2021, ogni pomeriggio io rilevavo e trasferivo su foglio excel vari tipi di indici, ed in particolare il già citato rapporto tra positivi e tamponi effettuati, chiamato anche *indice di contagiosità*. In Figura 111 vediamo, accanto alla serie storica dei valori misurati dal 18 dicembre 2020 al 24 gennaio 2021 (colonna sinistra) due grafici:

- nel grafico che compare nella parte superiore l'indice di contagiosità è rappresentato mediante una funzione discreta per punti, che esprime i valori giornalieri (linea continua) e una funzione continua con linea tratteggiata cosiddetta *polinomiale,* perché espressa da un polinomio che approssima la prima funzione.
- nel secondo (parte inferiore) è rappresentata la media a sette, cioè il valor medio degli ultimi sette giorni dell'indice.

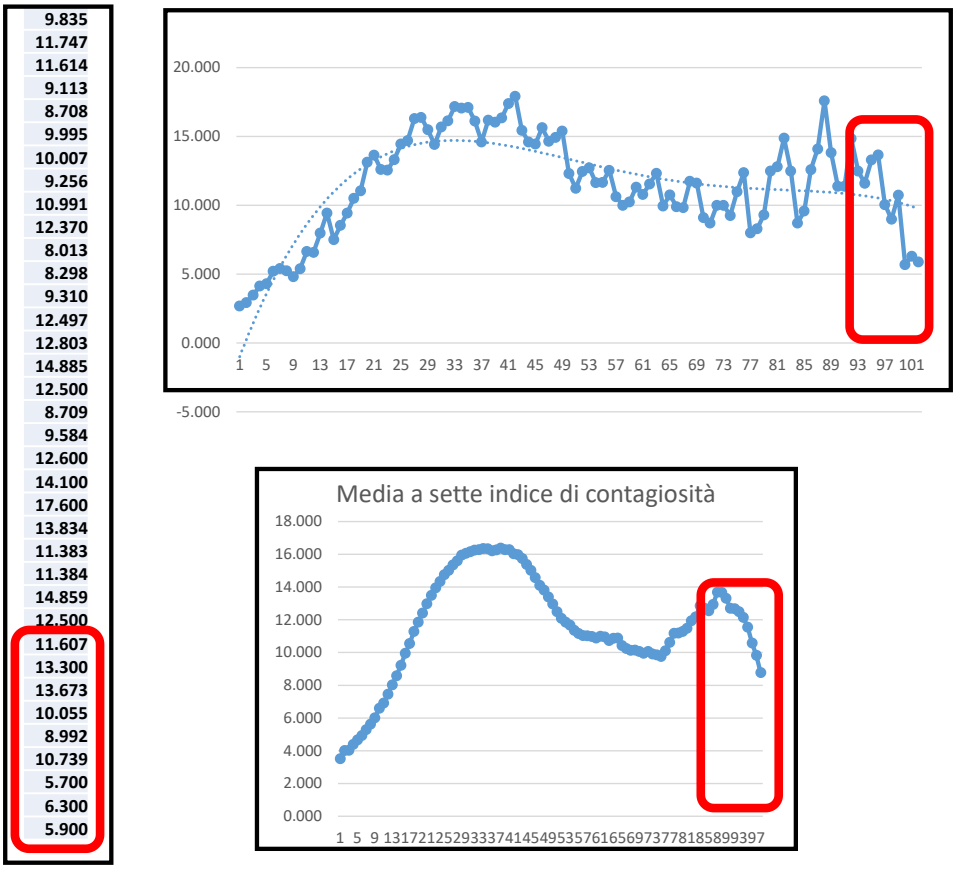


Figura 111 - Serie, medie, polinomiali, discontinuità

Osservo anzitutto che la media a sette giorni fornisce una informazione molto più efficace rispetto alla funzione discreta dei valori giornalieri, perché depura i fattori che provocano momentanee brusche variazioni, come ad esempio il fatto che i giorni festivi venivano fatti meno tamponi su soggetti in genere più a rischio che negli altri giorni, dando luogo a improvvisi picchi.

Inoltre, la brusca diminuzione intervenuta negli ultimi giorni a seguito della estensione dei tamponi considerati, prima solo naso-faringei successivamente anche antigenici, è certamente osservabile dalla serie storica, ma risulta più evidente sia nel grafico giornaliero, sia nella media a sette giorni, in questo caso la discontinuità è visivamente rivelata dai pallini blu separati e distanti l'uno dall'altro.

Insomma, produrre e usare le visualizzazioni è veramente alla portata di tutti coloro che abbiano un minimo di conoscenza di statistica e matematica elementari, e un po' di dimestichezza con strumenti come Excel.

Visualizzazioni di fenomeni nel tempo e visualizzazioni dinamiche

Le visualizzazioni sono anche molto efficaci per descrivere l'evoluzione di fenomeni nel tempo. In Figura 112 vediamo un esempio di app che serve per misurare l'inquinamento in un territorio. Se ad esempio in un viaggio arriviamo in una città e vogliamo fare un po' di corsa, utilizzando la applicazione Breezometer possiamo vedere nella zona attorno alla nostra residenza i diversi livelli di inquinamento mediante una heat map, o mappa di calore, una mappa dove i singoli dati sono rappresentati da colori di diversa intensità; in questo caso i diversi colori, dal chiaro allo scuro, rappresentano la intensità dell'inquinamento.

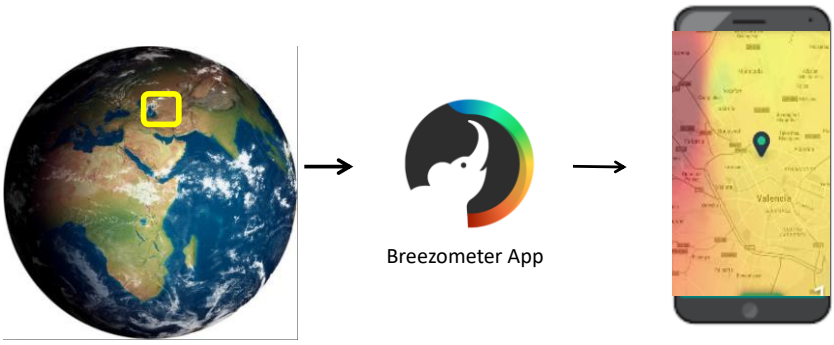


Figura 112 – La app Breezometer per misurare l'inquinamento

Chiaramente, la variazione dei livelli di inquinamento è mostrata nel tempo, vedi Figura 113, e quindi possiamo anche capire in quali ore della giornata si verificano i livelli più bassi.

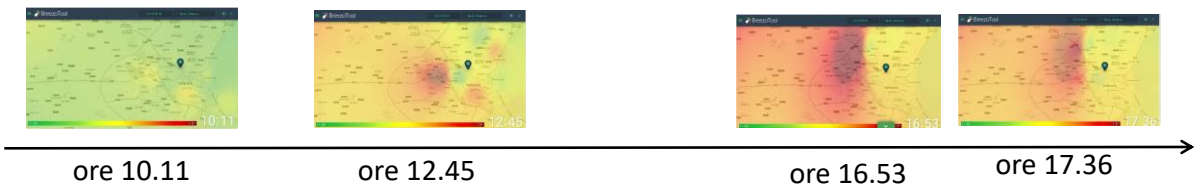


Figura 113 - Visualizzazione di Breezometer in diversi momenti della giornata

Le visualizzazioni dinamiche mostrano attraverso un insieme di immagini continue l'evoluzione di un fenomeno. Le più note e probabilmente le più belle possono essere viste sul sito https://www.gapminder.org/ (verificato gennaio 2021).

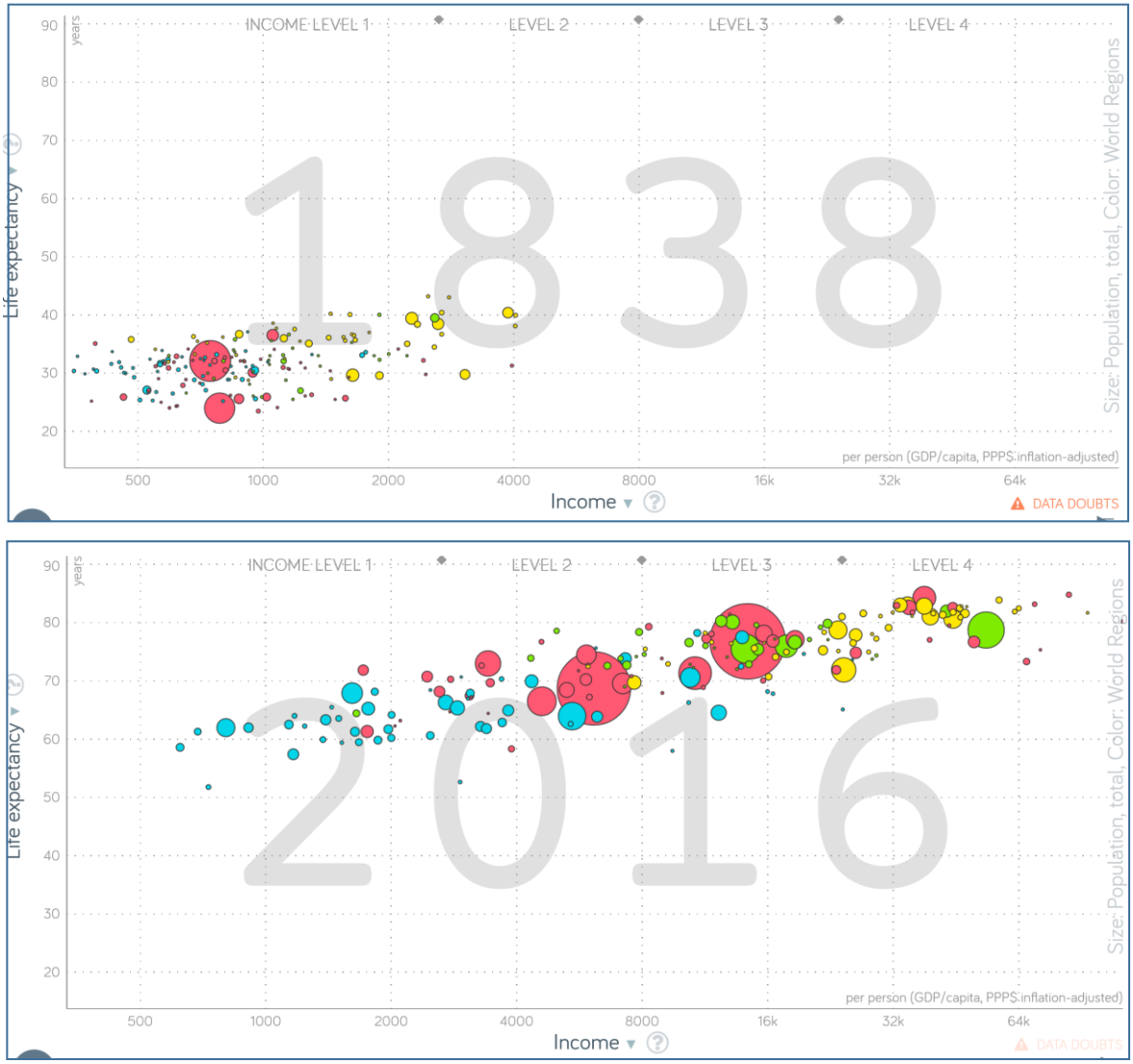


Figura 114 – Aspettativa di vita in funzione del reddito in diversi paesi del mondo (non viene riportata la legenda) fotografata nel 1838 e 2016, un paese per ogni cerchio; la dimensione del cerchio è proporzionale alla popolazione.

Visualizzazioni, icone e evoluzione socio-culturale

In tutti gli aeroporti del mondo, in molte stazioni ferroviarie e altri luoghi pubblici viene usata una icona per informare i visitatori su dove si trovino le toilette. Queste icone devono essere le più generali possibili, per essere comprese da persone di tutte le nazioni del mondo, far capire in modo chiaro e rapido dove si trovino i bagni associati alle due categorie classiche dei generi, le donne e gli uomini, e per non incorrere in distorsioni dovute a culture specifiche.

Figura 115 - Icone per il simbolo della toilette e evoluzione culturale

In Figura 115 sono mostrati diversi simboli adottati nel mondo per rappresentare le toilette. I tre simboli nella parte superiore rappresentano i due sessi tradizionali, donna e uomo, ma lo fanno in modo molto diverso, usando icone che nel primo caso evidenziano la differenza tra i sessi con la metafora della gonna e dei pantaloni, nel secondo tramite un riferimento ai duecromosomi femminile e maschile, e nel terzo caso attraverso un riferimento un po' azzardato al diverso modo di usare un fiocco come vestiario.

Nella parte inferiore della figura compiono icone che testimoniano la evoluzione socio culturale del concetto di genere, attraverso la introduzione di due simbolismi diversi per rappresentare un terzo tipo di genere, accanto a quelli tradizionali.

Qualità delle visualizzazioni

Abbiamo iniziato ad affrontare questo aspetto nel Capitolo 7 con particolare riferimento a foto e immagini dal vivo. Chiaramente esistono specifiche qualità per i diversi tipi di visualizzazioni definite nella Figura 109. Fissato un tipo di visualizzazione, ad esempio un grafico, esistono poi specifiche qualità, e regole per il loro raggiungimento, per specifici tipi di grafici. Ad esempio, in Figura 116 mostriamo quattro tra le tante regole definite sul sito [12].

[12] https://www.data-to-viz.com/caveats.html

Figura 116 – Esempi di regole per produrre grafici di qualità

Le visualizzazioni basate su icone vanno considerate con attenzione, perché essendo le icone basate su metafore, possono distorcere i dati rappresentati, ad esempio, mediante tabelle.

Osserviamo la Figura 117; la figura mostra a sinistra una tabella che riporta le soglie di consumo per la benzina stabilite per alcuni anni dal Transportation Department statunitense. A destra compare una visualizzazione che utilizza la metafora della strada, mostrata in prospettiva, con una larghezza crescente al crescere delle miglia per gallone.

Year	Miles per gallon
1978	18
1979	19
1980	20
1981	22
1982	24
1983	26
1984	27
1985	27,5

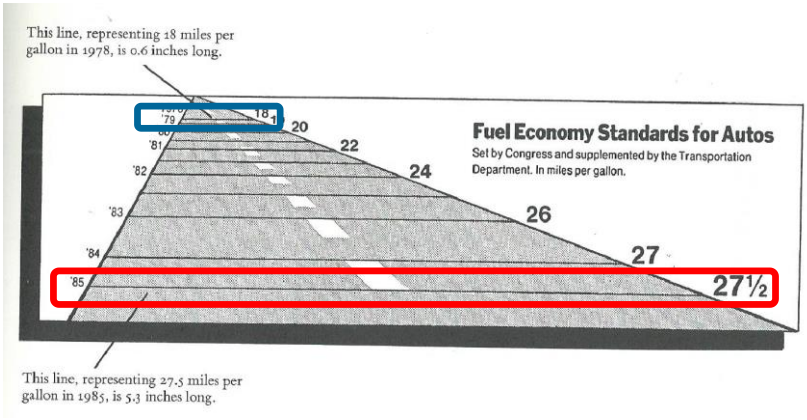


Figura 117 - Misurare le bugie nelle visualizzazioni
(tratta da New York Times, August 9, 1978)

La larghezza della strada corrisponde nella metafora al numero di miglia per gallone di benzina per gli anni dal 1978 al 1985; l'anno è riportato a sinistra della linea, il numero di miglia a destra. Ora, è pur vero che la strada è vista in prospettiva, ma è evidente la disproporzione tra numeri a sinistra e lunghezza della strada a destra.

Se avete un decimetro a disposizione, potete verificare facilmente che la proporzione numerica, ad esempio, tra il valore 18 e il valore 27,5, che è circa una volta e mezza, non è per nulla rispettata nella rappresentazione della strada. Se prendete esattamente le misure il

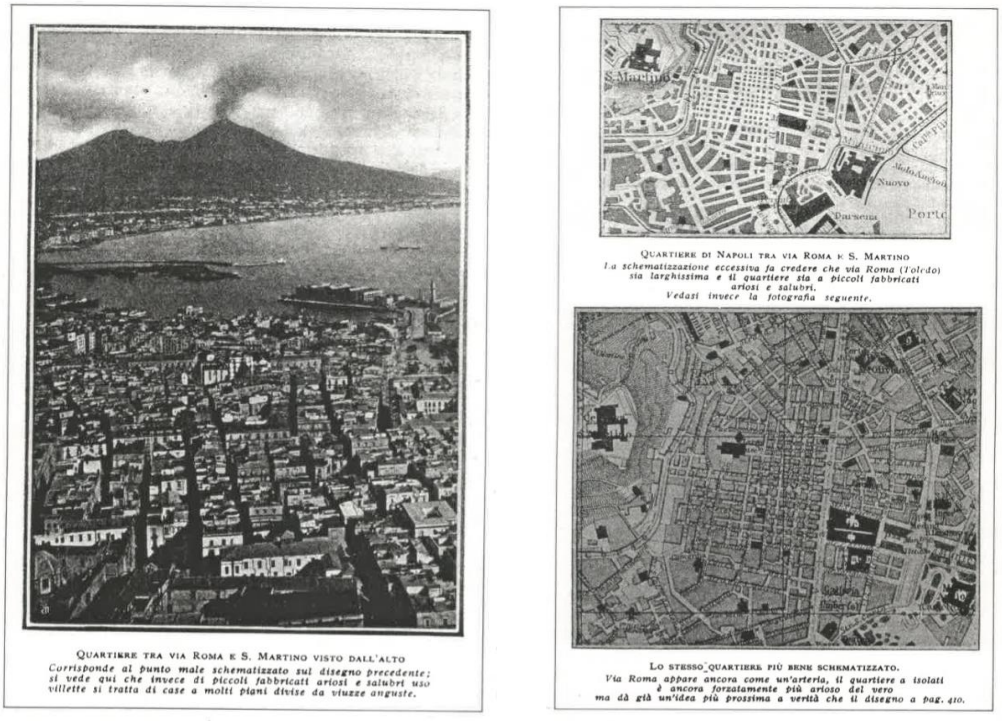
rapporto tra le due linee è di circa 15! Nel suo bellissimo libro in cui compare la figura[13], Edward Tufte introduce un livello di bugia definito come nel riquadro qui sotto.

Livello di bugia = rapporto tra valori numerici delle miglia per gallone nella tabella / rapporto tra lunghezze della strada nella visualizzazione

L'utente consapevole e critico è in grado di accorgersi di queste distorsioni e quindi riesce a interpretare in maniera oggettiva la realtà e accorgersi che la distorsione serve meramente ad "attirare l'attenzione".

Il rischio insito nelle visualizzazioni e nelle deformazioni che esse provocano rispetto alla realtà non è nuovo; è il problema delle rappresentazioni cartografiche, equivalenti, equidistanti, ecc. Nella Figura 118 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 sono comparse in successive edizioni della guida del Touring Club Italiano 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.

Possiamo toccare con mano, nell'esempio della rappresentazione deformata dei quartieri spagnoli, il tema della deformazione ottica che le visualizzazioni possono effettuare sulla realtà; in questo caso di ci troviamo di fronte a una deformazione di natura diversa dalla miopia e la presbiopia citate nella Figura 32, la cosiddetta deformazione *fisheye*, a occhio di pesce. La deformazione ci fa vedere innaturalmente ingrandito il fenomeno di interesse rispetto al contesto in cui esso è collocato.



[13] E. Tufte, The visual display of quantitative information, Graphic Press LLC, 2001

Figura 118 – I quartieri spagnoli di Napoli

In conclusione di questo capitolo, possiamo dire che le visualizzazioni sono di grande aiuto per *vedere* e analizzare i dati, ma a volte ci possono portare fuori strada; però, l'utente consapevole e critico è in grado di accorgersi di queste distorsioni e quindi ad interpretare in maniera oggettiva la realtà rappresentata; tornando alla Figura 117, può essere meglio tornare a usare le tabelle, più asettiche delle visualizzazioni e quindi più noiose, ma meno manipolabili.

Due libri eccellenti sulla visualizzazione in generale e la visualizzazione nelle mappe sono citati in nota[14].

Riassumendo

Una immagine (che mostra un dato) è meglio di 1.000 parole (che lo descrivono), perché l'effetto visivo delle **visualizzazioni** costitute da **tabelle**, **grafici**, **diagrammi**, **mappe**, **icone** riassume un **concetto**, un **messaggio**, un **insieme di dati risultato di un calcolo**, molto meglio di un **testo scritto**. Le tabelle sono le più semplici visualizzazioni, rispetto ad esse grafici, diagrammi, mappe e icone sono più espressive; d'altra parte, la **espressività** può essere accompagnata da distorsioni, che dobbiamo riconoscere per evitare che la percezione del dato sia sbagliata. Inoltre, le icone usate per rappresentare un dato (ad esempio dove si trovi la toilette in un aeroporto) hanno sempre inevitabilmente un **contenuto** ispirato da una **cultura** retrostante, che deve essere riconosciute per evitare una percezione, appunto, distorta. Per cui talvolta può essere importante partendo da un **grafico** e da una **icona**, osservare la **tabella** o il **testo** che rappresentano, per verificare la fonte, origine della visualizzazione.

[14] T. Munzner – Visualization Analysis and design, A.K. Peters Visualization Series, 2014.
The World of Maps accessibile sul sito https://mapyear.icaci.org/the-world-of-maps-book/

Capitolo 11

I dati per prevedere il futuro: il Machine Learning
I dati parlano da soli?

La *Hybris* è (da Wikipedia) un *topos* (tema ricorrente) della tragedia greca e della letteratura greca presente anche nella *Poetica* di Aristotele. Significa letteralmente "tracotanza", "eccesso", "superbia", "orgoglio" o "prevaricazione".

Quando i dati digitali sono diventati *tanti* dati, sono diventati i *big data*, come spesso accade di fronte a una ricchezza che ci giunge improvvisa, è insorto il sentimento della Hybris, la tracotanza, la superbia. E' ciò che è accaduto quando nel 2008 Chris Anderson, l'editor in chief della prestigiosa rivista Wired, ha fatto l'affermazione riportata in Figura 119.

This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear…. Forget taxonomy, ontology, and psychology. Who knows **why people do what they do**? The point is **they do it**, and we can track and measure it with unprecedented fidelity. With enough data, **the numbers speak for themselves.'**

Questo è un mondo in cui le grandi quantità di dati e la matematica applicate sostituiscono ogni altro strumento di analisi che possa essere proposto….Dimenticate le tassonomie, le ontologie, e le scienze congnitive. Chi sa **perchè le persone fanno ciò che fanno**? La sostanza è che lo fanno, a noi possiamo tracciare e misurare i comportamenti con una precisione senza precedenti nella storia. Avendo a disposizione una quantità adeguata di dati, **i numeri parlano da soli**.

Figura 119 – I dati parlano da soli

Per Anderson bastavano i dati, un po' di matematica, e si poteva spiegare tutto. Non si doveva più interrogarci sul perché delle cose, bastava osservare e tracciare asetticamente i comportamenti delle persone e i fenomeni della realtà, e tutto sarebbe stato compreso. Man mano che nuove applicazioni dei big data venivano sviluppate, si scoprì che le cose non erano così semplici.

Una delle applicazioni più controverse sviluppta più o meno nello stesso periodo delle affermazioni di Anderson è stata *Google Flu Trends*. Come sappiamo, molte delle interazioni con il Web avvengono attraverso i motori di ricerca; le ricerche vengono effettuate dagli utenti proponendo alcune parole chiave, che possono essere collegate tra loro mediante vari operatori.

Nel 2009 Google ritenne che si potesse prevedere l'andamento delle epidemie di influenza nei vari paesi in cui il motore di ricerca è utilizzato, sulla base della correlazione, cioè della

similitudine tra le parole utilizzate nel passato in occasione di epidemie di influenza (ad es. le tre parole: *influenza febbre aspirina*, oppure le quattro parole: *febbre alta raffreddore tosse*) e quelle utilizzate nel periodo sotto osservazione. Sulla base della similitudine, venivano prodotte previsioni sulla percentuale della popolazione colpita dalla epidemia nel tempo.

Veniva anche fornito da Google un confronto delle previsioni con le rilevazioni effettuate dagli osservatori epidemiologici, in inglese Centers for Disease Control and Prevention (CDC); questa è la modalità usuale di monitorare l'andamento della epidemia, modalità che abbiamo imparato a conoscere in occasione delle fasi della epidemia Covid. Il confronto era fatto anche rispetto ad una seconda fonte, il sito Flu Near You, un sito collaborativo che raccoglieva informazioni sulla epidemia fornite da una comunità di utenti.

Vedi in Figura 120 gli andamenti delle tre curve a partire dal gennaio 2011, nella cornice azzurra tratteggiata i simboli usati per le tre curve.

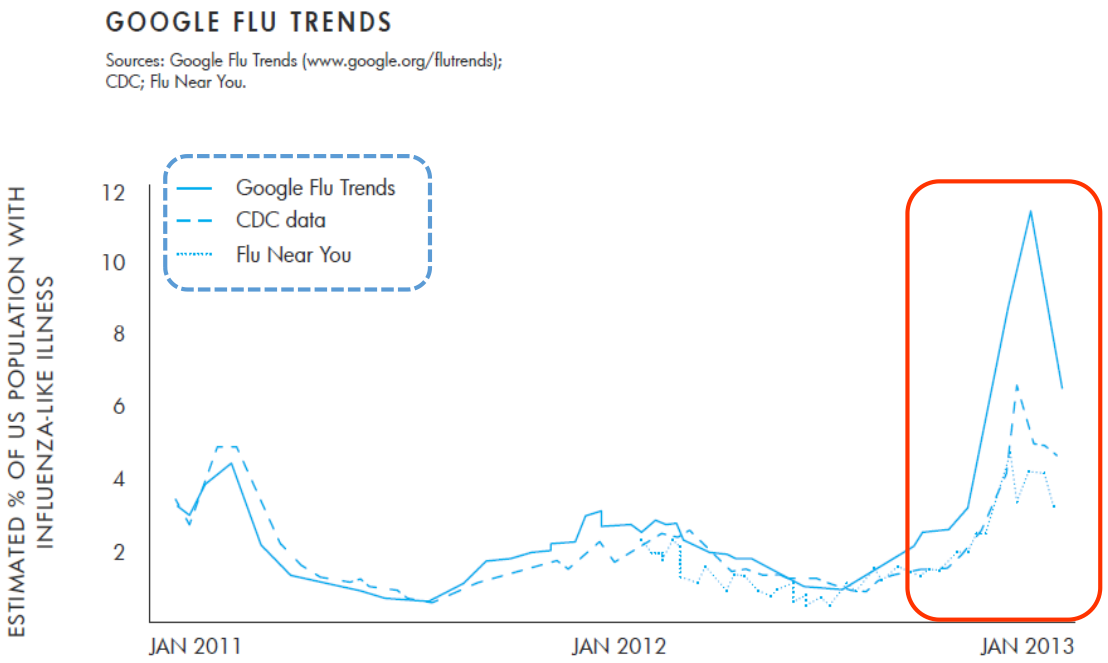


Figura 120 – Le previsioni di Google Flu Trends

Nel primo anno e mezzo la previsione fu molto vicina ai dati dei CDC, che possono considerarsi la fotografia esatta del fenomeno reale, in quanto misuravano l'epidemia partendo dai dati raccolti dai medici sul territorio. Ma a partire dall'autunno del 2012 (vedi l'area rossa in Figura 120), la curva di Google Flu Trends iniziò a discostarsi dalla curva dei CDC, fornendo previsioni largamente sovrastimate. A questo punto Google decise di chiudere il sito, non riuscendo più a correggere il modello di stima, e riconoscendo implicitamente il fallimento di Google Flu Trends.

Una storia di successo è invece quella di Oren Etzioni, un professore e imprenditore di origine israeliana che vive negli Stati Uniti, che a partire dai primi anni 2000 investigò il seguente problema.

Quando noi vogliamo *effettuare un viaggio aereo partendo un certo giorno*, *ad esempio il 20 gennaio 2021*, da un aeroporto di partenza a uno di arrivo, abbiamo a disposizione diversi siti che ci permettono di conoscere tutti gli itinerari possibili con le diverse compagnie aeree, gli eventuali scali intermedi, la durata del viaggio, e il prezzo del biglietto *se lo acquistiamo nel giorno della ricerca, ad esempio il 20 dicembre 2020*; inoltre, possiamo avere tutti i risultati ordinati dal biglietto più economico al biglietto più costoso, vedi Figura 121.

Quando parleremo di *modelli di dati*, sarà più chiaro che questo risultato si può ottenere con una *interrogazione* su una base di dati che rappresenta per mezzo di tabelle i viaggi delle compagnie, gli aeroporti di partenza e di arrivo, i prezzi per le varie classi.

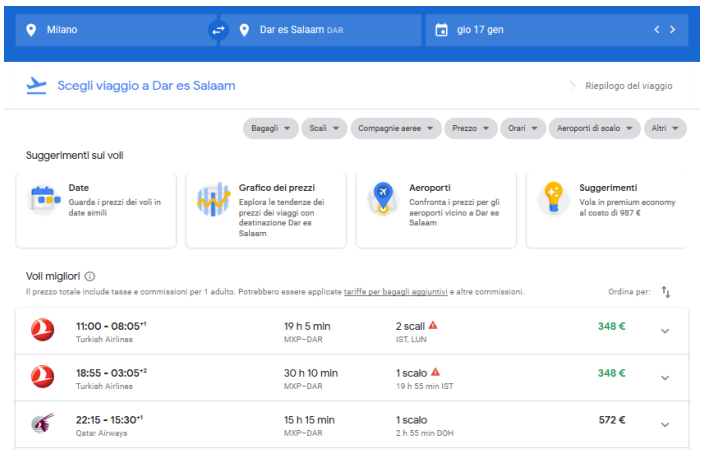


Figura 121 – Scegliere il biglietto più economico (dal sito eDreams)

Etzioni si pose l'obiettivo di risolvere *un altro problema*, quello in cui noi conosciamo come in precedenza il giorno di partenza, il 20 gennaio 2021, l'aeroporto di partenza e l'aeroporto di arrivo, e vogliamo conoscere *quale è il giorno compreso tra quello della ricerca (ad esempio il 20 dicembre 2021) e quello della partenza (20 gennaio 2021)* in cui è più conveniente acquistare il biglietto. Il problema precedente risolveva un'altra esigenza: noi sappiamo il prezzo più conveniente assumendo che *acquistiamo* il biglietto *il giorno della ricerca*, e, se vogliamo sapere il prezzo del biglietto nei giorni successivi, dobbiamo ripetere ogni volta la ricerca. Ti è chiara la differenza tra giorno della partenza, giorno della ricerca e giorno di acquisto del biglietto?

Vediamo, provo a ragionare a voce alta, e faccio un esempio diverso dal tuo. Io devo partire il 20 ottobre, e faccio la ricerca il 5 settembre; il giorno della ricerca è il 5 settembre, e i prezzi sono validi se il giorno di acquisto del biglietto è il 5 settembre; se ripeto la ricerca il 6 settembre, i prezzi sono riferiti al giorno di acquisto pari al 6 settembre. L'algoritmo di Etzioni mi dice: ti conviene acquistare il biglietto il 15 di settembre. Miracoloso! Come fa l'algoritmo di Etzioni a prevedere quale sarà il giorno migliore per acquistare il biglietto, senza calcolarlo il giorno stesso? Ha una sfera di cristallo?

Adesso ti spiego. Ti sei reso conto che il problema che si è posto Etzioni non si può risolvere come il caso precedente, perché noi *non sappiamo come evolverà il prezzo*. Alcune volte le compagnie fanno oscillare il prezzo verso l'alto e verso il basso, poi nella imminenza del

viaggio, se ci sono molti posti liberi fanno prezzi last minute da saldo, e se invece ne sono rimasti pochi, fanno salire i prezzi per coloro che si sono ridotti all'ultimo minuto e devono assolutamente viaggiare quel giorno.

Etzioni applicò un *algoritmo predittivo* su un insieme di dati, o *dataset*, che comprendeva 12.000 biglietti di *viaggi effettuati nel passato*, vedi Figura 122. Anche in questo caso, come nel caso di Google Flu Trends per la diffusione della influenza, i dati descrivevano il prezzo di viaggi passati, e l'algoritmo *imparava dal passato, per prevedere il futuro*. Approfondiremo tra poco le fasi di produzione di un algoritmo predittivo. Etzioni provò ad applicare l'algoritmo sui 12.000 viaggi del passato, e ottenne risultati affetti da errori molto significativi.

Come fece a capire che c'erano degli errori?

Lo vedremo in dettaglio tra poco, ma provo a giustificare subito la affermazione. Etzioni, nel produrre l'algoritmo di previsione, mise da parte alcuni viaggi del passato, e poi eseguì l'algoritmo su questi, e trovò che in molti casi l'algoritmo prevedeva un prezzo sbagliato…

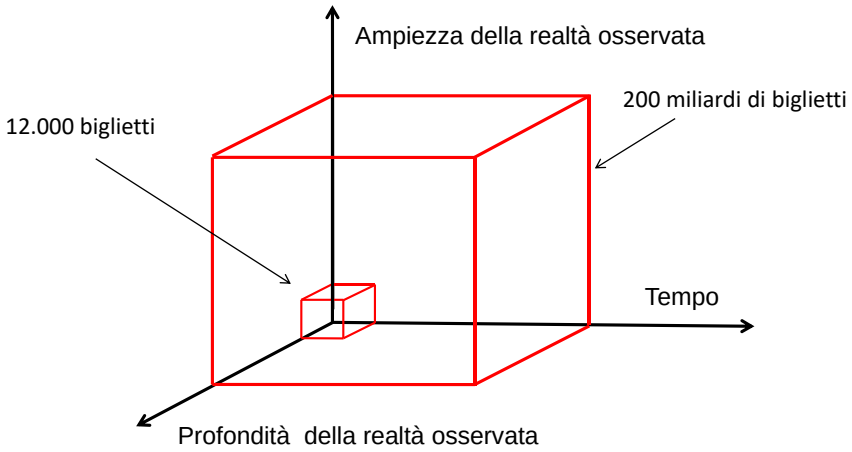


Figura 122 – Quantità di biglietti utilizzati per scegliere il giorno in cui acquistare il biglietto

Etzioni si procurò allora un dataset composto da 200 miliardi di biglietti (!), che riguardavano viaggi relativi a un intervallo di tempo molto più ampio del precedente, riguardavano più tratte, più compagnie, *espandendo* dunque lo spazio dei dati nelle tre coordinate che compaiono in Figura 112: dunque, un intervallo temporale più ampio, più biglietti, più caratteristiche dei biglietti.

Il nuovo algortimo predittivo si comportò benissimo, e la riprova fu che la app prodotta fece in media risparmiare i viaggiatori 50 dollari per ogni biglietto. La seconda riprova del successo, fu che Etzioni vendette nel 2008 alla Microsoft il brevetto al prezzo di circa 120 milioni di dollari!

Abbiamo visto quindi un fallimento (Google Flu Trends) e una storia di successo (Etzioni).

Il machine learning

Insomma, vediamo ora più in dettaglio come funziona il *machine learning*, o apprendimento automatico, la disciplina che produce algoritmi predittivi, imitando in questo modo gli esseri umani nella attività di apprendimento. La strategia di apprendimento che abbiamo visto nei precedenti esempi e che approfondiremo nelle pagine che seguono, fa riferimento all'apprendimento a partire da esempi del passato, o *apprendimento supervisionato*. Altre forme di apprendimento verranno discusse nel libro della Enciclopedia dedicato al machine learning.

Per comprendere meglio il funzionamento di un algoritmo di apprendimento supervisionato, consideriamo il caso di Figura 123, che fa riferimento all'applicazione Compas, commercializzata negli Stati Uniti dalla azienda Northpointe. Compas è stato utilizzato in diverse corti di Giustizia americane per permettere ai giudici di sorveglianza di decidere se concedere la libertà provvisioria a detenuti in attesa di giudizio.

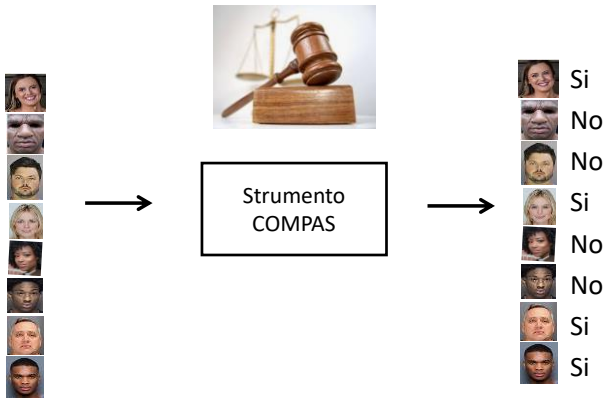


Figura 123 – Decisione se concedere la libertà provvisoria

La decisione se concedere o meno la libertà provvisoria può avere una grande importanza per l'accusato di un reato, perché anche pochi giorni passati in prigione possono influire molto sulla sua immagine sociale, e perché l'ambiente del carcere presenta problemi di tensione sociale che possono influire sulla psiche e sui comportamenti futuri.

Uno degli elementi fondamentali presi in considerazione dai giudici è il *rischio di recidiva*, cioè il rischio di ripetizione di reati dopo la concessione della libertà provvisoria, ovvero dopo il periodo passato in carcere. Compas assume che siano noti i dati su quanto accaduto nel passato, in particolare gli eventi di recidiva di diversi soggetti detenuti.

Per tutti i soggetti, sia quelli che hanno commesso recidiva sia quelli che non la hanno commessa nel passato, Compas conosce diverse caratteristiche, quali la *età*, il *numero di arresti precedenti*, il *genere*, e molte altre; Compas non raccoglie dati sulla etnia. In Figura 124 vengono mostrate le tre caratteristiche precedenti insieme al rischio corrispondente.

Figura 124 – Raccolta di dati da decisioni nel passato

L'insieme delle fasi del *machine learning supervisionato* è mostrato in Figura 125. I passi sono tre, e sono segnati con piccoli cerchi blu. Le caratteristiche e il rischio di recidiva sono rappresentate con un breve acronimo (es. E per Età).

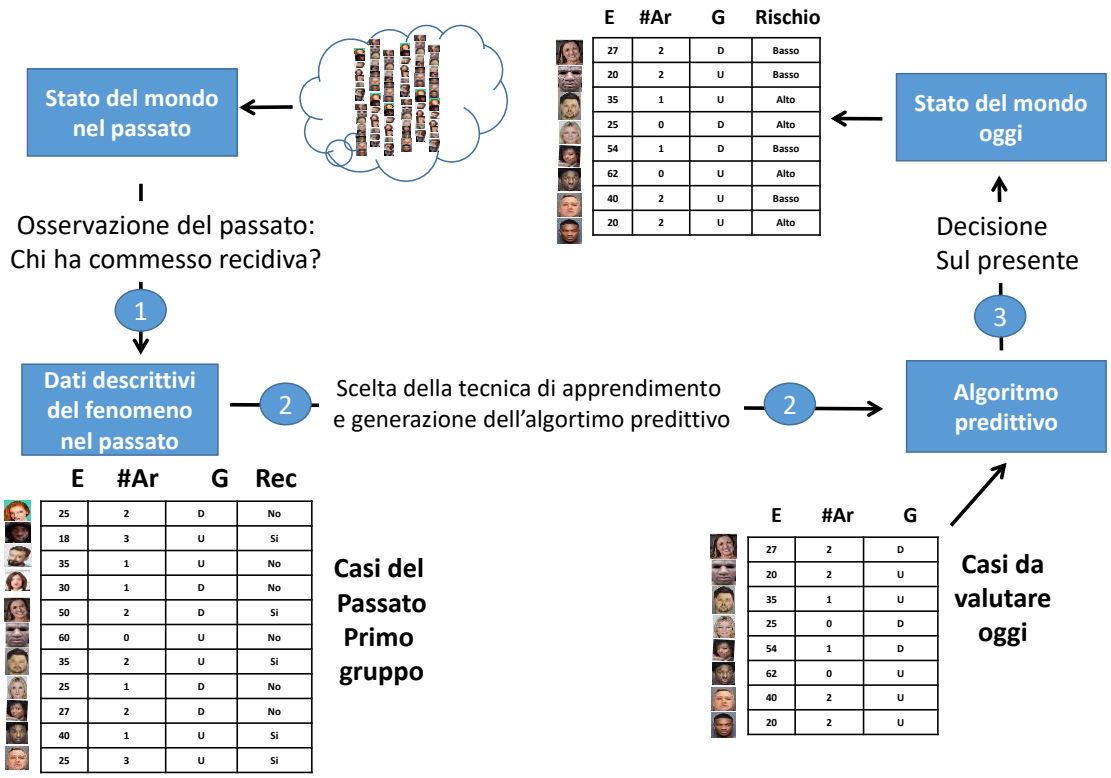


Figura 125 – Fasi del machine learning

Il *passo 1* di *osservazione del fenomeno della recidiva nel passato* costruisce tabelle come quella di Figura 124. Come stiamo cominciando a percepire, la filosofia del machine learning supervisionato è quella di *prevedere il futuro osservando e apprendendo dal passato*.

Osserviamo la scritta *Casi del passato primo gruppo*; questi sono i casi del passato che vengono usati nel seguito per produrre l'algoritmo predittivo; accanto ad essi, come discusso per l'algortimo di Etzioni, dobbiamo ricordarci di accantonare un *secondo gruppo*, che verrà usato per verificare la bontà dell'algoritmo predittivo.

La generazione dell'algoritmo predittivo è effettuata nel *passo 2* della procedura; in sostanza, vengono utilizzate varie tecniche che, operando sui casi del passato, *apprendono a prevedere il futuro*. Una di queste tecniche sono gli *alberi di decisione*, un esempio è mostrato in Figura 126.

La figura mostra una procedura che possiamo seguire per *decidere come andare vestiti se vogliamo uscire in una giornata di primavera con tempo incerto.* Rispondendo mentalmente alle domande sulla pioggia e sul sole, arriviamo a una conclusione sul nostro abbigliamento.

Un *albero di decisione* è una *tecnica* che ci permette di prendere una decisione; possiamo usarlo come *struttura portante* di un algoritmo predittivo, ottenendo così *un algoritmo predittivo basato su un albero di decisione*. Accanto agli alberi di decisione ci sono molte altre tecniche attorno a cui possiamo *cucire, costruire un algoritmo predittivo*, come vedremo meglio quando parleremo dell'etica dei dati digitali. Durante il passo 2, dunque, viene *prima* scelta la tecnica utilizzata per costruire l'algoritmo, e successivamente viene prodotto un algoritmo predittivo basato sulla tecnica scelta.

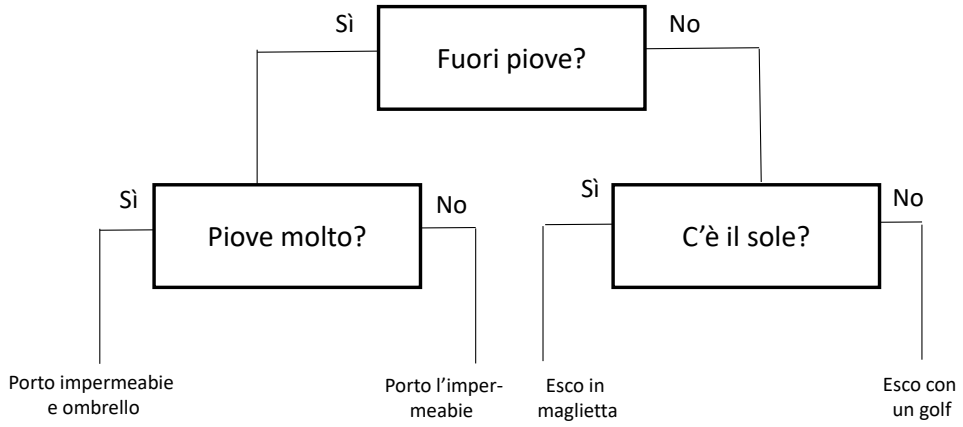


Figura 126 – Un albero di decisione

Nel caso di Compas, la decisione è *ben più impegnativa e delicata* rispetto al decidere come andare vestiti quando usciamo; Compas insegna a costruire una *procedura decisionale* o *algoritmo predittivo* che valuta il grado di rischio associato a ciascun caso di concessione della libertà provvisoria, grado di rischio che per semplicità possiamo esprimere in modo binario: *alto, che porta alla decisione di libertà non concessa, e basso, che porta alla decisione di libertà concessa*.

Costruito l'algoritmo, esso nel *passo 3* di Figura 125 viene applicato ai *casi attuali*, arrivando per ciascun detenuto a determinare la decisione.

Qualità degli algoritimi predittivi

Una frase attribuita al Premio Nobel in Fisica Niels Bohr afferma: "E' difficile fare previsioni, soprattutto sul futuro". Abbiamo qualche strumento che ci possa dare una misura della qualità dell'algoritmo di previsione, cioè della capacità dell'algoritmo di prevedere correttamente il futuro? La risposta è sì, perché possiamo sottoporre all'algoritmo di decisione *i casi del passato relativi al secondo gruppo di detenuti*, *quello che abbiamo accantonato*, dando luogo a quattro casi possibili:

- *Veri negativi*, quando si tratta di persone valutate dall'algoritmo a basso rischio di commettere recidiva, e che in effetti *non hanno commesso recidiva* nel loro comportamento futuro.
- *Veri positivi*, quando si tratta di persone valutate dall'algoritmo ad alto rischio di commettere recidiva, e che *hanno commesso recidiva* nel loro comportamento futuro.
- *Falsi negativi*, quando si tratta di persone valutate dall'algoritmo a basso rischio di commettere recidiva, e che *hanno commesso recidiva* nel loro comportamento futuro.
- *Falsi positivi*, quando si tratta di persone valutate dall'algoritmo ad alto rischio di commettere recidiva, e che *non hanno commesso recidiva* nel loro comportamento futuro.

Chiaramente in generale noi siamo interessati ad algoritmi che hanno il 100% di veri positivi e il 100% di veri negativi, ma il lettore che vorrà leggere il libro della Enciclopedia dedicato a questo tema vedrà che ciò non è praticamente mai garantito dagli algoritmi predittivi.

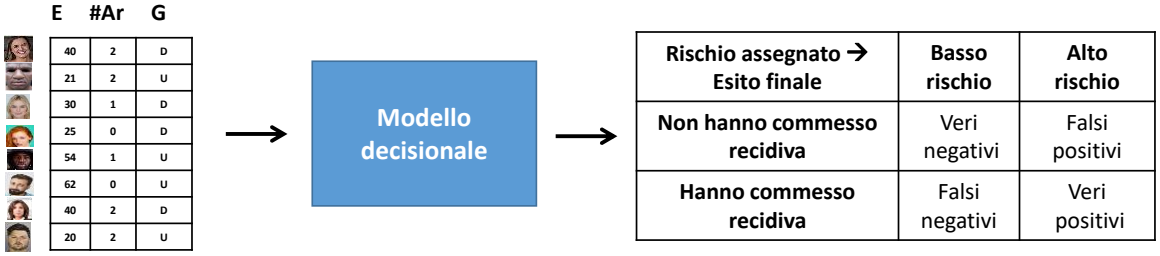


Figura 127 – Positivi e negativi, veri e falsi

Eseguendo l'algortimo sui casi del passato che avevamo messi da parte (*Casi del passato secondo gruppo*, vedi Figura 127) noi possiamo calcolare i veri negativi, i veri positivi, i falsi negativi, i falsi positivi prodotti da Compas.

Nel 2016 il sito di giornalismo investigativo Propublica (https://www.propublica.org/) pubblicò una indagine effettuata sui dati pubblici resi disponibili dal Ministero della Giustizia e dalla azienda Northpointe distributrice di Compas, in cui mostrò (vedi Figura 128) che, anche se Compas non usava dati sulla etnia, tuttavia Compas *discriminava i detenuti afroamericani*,

i quali avevano *un maggior numero di falsi positivi e falsi negativi* (vi prego di riguardare bene le definizioni) rispetto ai detenuti di etnia bianca.

Propublica

Esiti	Bianchi	Afro Americani
Valutato rischio **alto**, ma **non hanno** successivamente commesso recidiva	23,5%	44,9%
Valutato rischio **basso**, ma **hanno** successivamente commesso recidiva	47,7%	28,0%

Northpointe

Esiti	Bianchi	Afro Americani
Valutato rischio **alto**, e **hanno** successivamente commesso recidiva	63 %	59 %
Valutato rischio **basso**, e non **hanno** successivamente commesso recidiva	71 %	65 %

Il punto di vista di Propublica: e' discriminata la comunità di etnia Afro Americana

Rischio assegnato → / Esito finale ↓	Basso rischio	Alto rischio
Non hanno commesso recidiva	Veri negativi	Falsi positivi
Hanno commesso recidiva	Falsi negativi	Veri positivi

Il punto di vista di Northpointe: la capacità predittiva è circa la stessa

Figura 128 – Punti di vista sulla equità

Northpointe risposte dopo pochi mesi a Propublica, affermando che le conclusioni di Propublica erano a loro volta affette da un ragionamento non equo; se si andava a vedere la *capacità predittiva*, cioè quanto Compas era riuscito a prevedere correttamente sia i veri positivi sia i veri negativi (anche in questo caso occorre riguardare bene le definizioni) Compas faceva previsioni sostanzialmente simili tra afroamericani e bianchi.

Chi aveva ragione? Una discussione approfondita verrà fatta nel libro della Enciclopedia dedicato al machine learning, ma possiamo subito affermare che:

- entrambi i punti di vista hanno ragion d'essere
- Propublica è interessata a salvaguardare un principio di *equità sociale*, che non discrimini nelle decisioni gli afro americani;
- Northpointe è interessata al *meccanismo decisionale*, essendo il suo *committente* il giudice.

Ma, soprattutto, occorre dire che *decidendo sul presente, prevedendo il futuro sulla base del passato* (*questo è ciò che fa Compas*) si assume una società ingessata e incapace di concepire nuove forme di protezione e emancipazione sociale; siccome nel passato gli afro americani in percentuale sono stati più a rischio di recidiva, e hanno quantitativamente commesso più reati nel periodo di vita successivo alla prigione, allora ciò sarà anche vero nel futuro.

Scelta e integrazione delle fonti per prevedere l'inquinamento

Vediamo un ultimo caso, quello di Breezometer, menzionato nel Capitolo 10 sulle visualizzazioni. Come fa Breezometer a prevedere l'inquinamento in una certa zona di territorio?

Fino ad ora abbiamo visto casi (Google Flu Trends, Etzioni, Compas) in cui la fonte che ci forniva i dati sul fenomeno per la previsione era sostanzialmente *unica*. Breezometer invece deve prevedere un fenomeno complesso, in cui l'inquinamento è influenzato da una *molteplicità di fattori,* quali il riscaldamento delle abitazioni, i gas di scarico del traffico automobilistico, l'energia spesa per le fabbriche; tutto ciò deve poi essere associato al territorio cui fa riferimento (si usa dire che va *geolocalizzato*). Come conseguenza, il primo passo del processo di apprendimento (vedi Figura 129) è costituito da una fase di scelta delle fonti di dati, acquisizione delle caratteristiche e conservazione in una memoria digitale comune.

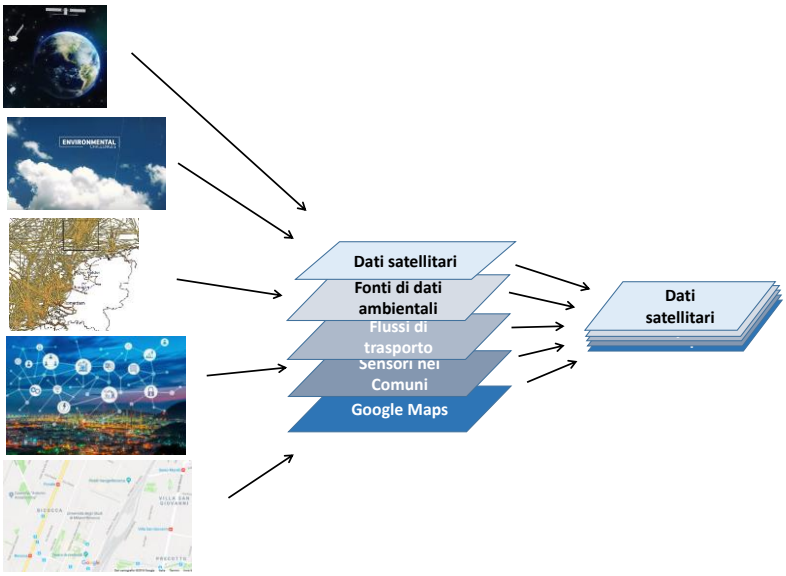
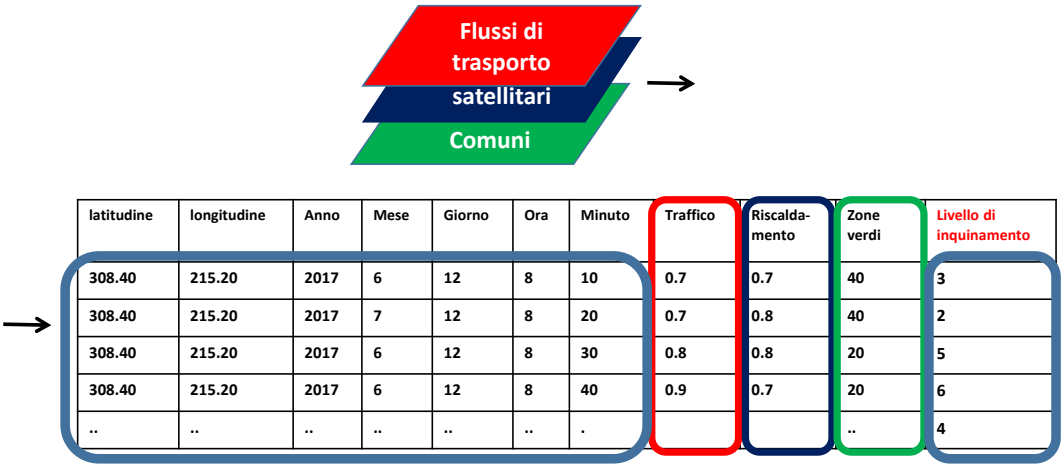


Figura 129 – Raccolta di dati nello strumento Breezometer

Successivamente (vedi Figura 130) tali caratteristiche (qui ne sono mostrate solo tre, *traffico, riscaldamento e zone verdi*) sono integrate in un'unica grande tabella, in cui per ogni latitudine e longitudine, momento temporale e valori delle tre caratteristiche vengono rappresentati i valori di inquinamento.



latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscalda-mento	Zone verdi	Livello di inquinamento
308.40	215.20	2017	6	12	8	10	0.7	0.7	40	3
308.40	215.20	2017	7	12	8	20	0.7	0.8	40	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	20	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	20	6
..	..	..	..	..	..	.			..	4

Figura 130 – Caratteristiche dell'inquinamento sulle quali si basa l'apprendimento

Il cosa e il perchè

Torniamo infine sul tema sollevato da Chris Anderson, del *cosa* e del *perché*; vediamo in Figura 131 una vignetta del bellissimo libro *The Book of Why*, scritto da Judea Pearl, in cu un robot, viene redarguito da una persona svegliata dai rumori; *perché mi hai svegliato!*

Il robot della figura era stato addestrato con un algoritmo di apprendimento in cui ci si era scordati di inserire dati relativi alle 6 la mattina, ora in cui non si può usare l'aspirapolvere. In quel *perché mi hai svegliato* è condensata tutta la grande questione del machine learning supervisionato, che non sa nulla del perché, e delle sue conseguenze, e agisce quindi senza consapevolezza.

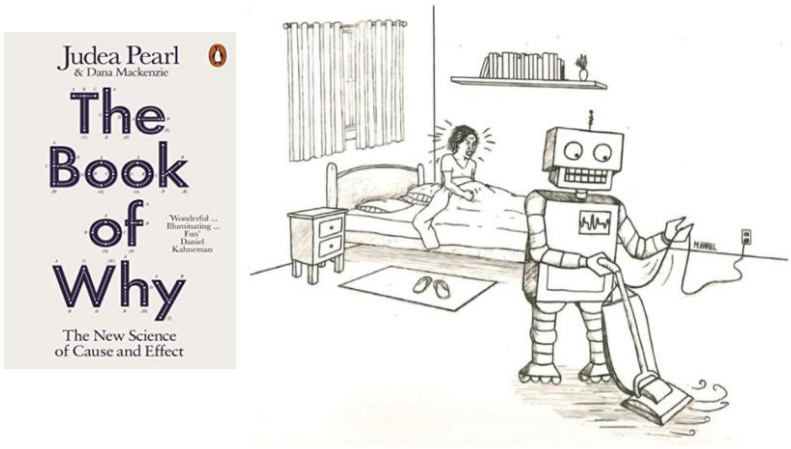


Figura 131 – Perché mi hai svegliato alle 6 di mattina?

Per dirla con il mio amico Walter Tocci (comunicazione privata, con permesso di renderla pubblica): "La domanda del perché ha sorretto la Scienza moderna dal Seicento in avanti, e ha legittimato il primato della legge scientifica, l'equazione che è in grado di prevedere i fenomeni in determinate condizioni al contorno. Il primato non è rimasto confinato in ambito scientifico ma ha improntato l'intero Codice moderno: la legge dello Stato come secolarizzazione della legge divina, le categorie dell'intelletto, le leggi della macroeconomia, e in generale ha influenzato tutte le scienze sociali. Tutto ciò per sottolineare che la nuova scienza dei dati proprio rifiutando il perché, paradossalmente, ci spiega perché ci troviamo ormai al di là del Codice moderno".

Riassumendo

Il ***machine learning (ML) o apprendimento automatico supervisionato*** ha lo scopo di costruire algoritmi predittivi che apprendono a predire e decidere sul futuro osservando dati del passato. Per esempio un giudice concede la libertà provvisoria a un detenuto valutando il rischio futuro di recidiva, sulla base di quanto accaduto nel passato con detenuti ***simili***. Accanto al ML supervisionato esistono altre tipologie, come il ***ML non supervisionato*** e il ***ML per rinforzo***. Gli algoritmi di ML supervisionato sono caratterizzati da una ***accuratezza*** che si può misurare valutando i ***veri positivi, veri negativi,falsi positivi, falsi negativi*** su ***dati di test***. Un'altra caratteristica del ML supervisionato è la equità dell'algoritmo nei confronti di tutti i soggetti cui viene applicato. Esistono diverse definizioni di equità. Un limite del ML è nel fatto che per **decidere sul nuovo si basa sul passato**.

Capitolo 12

I dati possono darci valore e dis-valore

Il termine *valore* è molto usato nei siti Web, nei giornali, nella pubblicità, nelle conversazioni: "quel prodotto è di valore", "un investimento di valore", "il valore di quel messaggio". Ma raramente, credo di poter dire, si comprende precisamente quale sia il suo vero significato.

Il valore è legato, nel seguito, al soddisfacimento di un scopo o bisogno. Nella nostra vita, nella vita delle comunità e delle società, accanto agli affetti, alle emozioni, i tradizionali oggetti che ci permettono di soddisfare i nostri bisogni sono i beni e i servizi; un esempio di *bene* sono due chili di arance, un esempio di *servizio* è un taglio di capelli dal barbiere o dal parrucchiere.

La tradizionale differenza che si fa tra beni e servizi riguarda la loro *materialità*; due chili di arance *pesano*, il servizio consistente in un taglio di capelli è di natura più *immateriale*. Mentre il concetto di bene ci è in genere molto chiaro, conviene spendere alcune parole in più sul concetto di servizio.

Un *servizio* consiste in una attività che dà luogo a uno scambio tra un produttore e un consumatore del servizio, dove l'oggetto dello scambio è per l'appunto il soddisfacimento dell'obiettivo del consumatore. Ad esempio, un viaggio in treno da Milano a Roma è un servizio, al termine del quale viene soddisfatta la necessità di recarsi a Roma.

Al contrario dei beni, ad esempio due chili di arance che diventano nostre quando le acquistiamo, quando fruiamo di un servizio alla fine non ci rimane in mano niente di materiale, se non il raggiungimento del nostro obiettivo, negli esempi precedenti avere i capelli in ordine o arrivare a Roma essendo partiti da Milano. [15] Molti settori economici sono essenzialmente basati sulla offerta o vendita di servizi: ad es. la formazione, la sanità, le banche.

In effetti, occorre dire che la differenza tra beni e servizi non è così netta. E' stata fatta una indagine su circa 50 prodotti della attività umana, ad esempio un libro, una seduta dal dentista, un parere di avvocato in cui la domanda era: questo prodotto ti appare più un bene o più un servizio?.

i prodotti che meglio rappresentano rispettivamente il concetto di bene e il concetto di servizio sono stati individuati in un paio di jeans e una seduta di supporto psicologico (vedi Figura 132); gli altri prodotti formano un continuo tra l'essere visti come totalmente un bene e l'essere totalmente un servizio.

[15] Se siete curiosi di saperne di più sui servizi e sulla loro differenza rispetto ai beni potete liberamente scaricare dal link http://hdl.handle.net/10281/98632 il testo di C. Batini e altri autori The Smart Methodology for the Life Cycle of Services, licenza Creative Commons.

Un paio di jeans

Una seduta psicolanalitica

Figura 132 – Come si distingue un bene da un servizio?

Per iniziare a parlare del concetto di valore, facciamo ora un esempio, *l'acquisto*, questa volta, *del biglietto del treno* che ci fa andare da Milano a Roma. Supponiamo di acquistare il biglietto tramite il sito Web di una compagnia che gestisce il trasporto ferroviario.

Acquistare il biglietto tramite il sito Web ci fa risparmiare tempo rispetto all'andare alla stazione ferroviaria o a una agenzia. Nel decidere l'acquisto del biglietto tramite sito Web, noi mettiamo sulla bilancia da una parte i *vantaggi* che abbiamo rispetto ad altre soluzioni, e dall'altra i *costi* in tempo (ad es. dobbiamo imparare a usare una app Web) ed eventualmente in denaro che dobbiamo sopportare (es. acquistare un biglietto tramite sito Web può costare una somma in più rispetto allo sportello fisico).

Quando si fa riferimento al *valore* di un bene, di un servizio, e tra poco di un insieme di dati, si assume in genere che la sua fruizione porti una qualche forma di *utilità* a una persona, a una comunità, a una intera società; e questa utilità è connessa all'uso che la persona, la comunità, la società fa del bene, del servizio, dei dati che ha acquisito.

Inoltre, questa utilità può essere messa in relazione con le risorse che dobbiamo mettere in campo per poter acquisire il bene, il servizio, i dati. Ecco perché, in una prospettiva generale, possiamo vedere il valore come *valore d'uso*.

Il valore d'uso consiste in generale nei *benefici* che traiamo dall'uso dei dati digitali, che possono, o meno, essere commisurati ai *sacrifici* che facciamo per averli disponibili; dunque, il valore d'uso è espresso da una qualche relazione tra benefici e sacrifici. Tornando all'acquisto del biglietto su un sito Web, i benefici che otteniamo sono, oltre che la prenotazione e il biglietto (benefici che sono chiamati *funzionali, cioè connessi alla funzione o scopo del servizio*), il tempo risparmiato nel non fare la fila alla stazione e il tempo che risparmiamo se dobbiamo modificare la prenotazione, modofica che possiamo fare anch'essa dal sito.

I sacrifici sono: il tempo che impieghiamo per capire la prima volta come fare la prenotazione sul sito Web, che però è una tantum, il tempo che impieghiamo a farla, di gran lunga minore che andare alla stazione e fare la fila, e l'eventuale costo supplementare. C'è un altro sacrificio che io vivo qualche volta, l'ansia che mi dà certe volte interagire con la applicazione sul Web, che non sempre mi fornisce messaggi o comportamenti completamente amichevoli e rassicuranti.

Iniziamo a parlare a questo punto del *valore dei dati*, mostrando la serie temporale in Figura 133, tratta dal testo citato in nota[16]. Le serie temporale rappresenta come sono cambiati i lavori da quando è iniziata la civiltà, due milioni di anni fa, fino ai giorni odierni e nel prossimo futuro. Agli inizi della civiltà, gli esseri umani erano tutti cacciatori/raccoglitori; circa 20.000 anni fa nacque l'agricoltura, e pian piano la coltivazione soppiantò la caccia fin quasi a farla scomparire. Il diciannovesimo secolo, con la invenzione della macchina a vapore, ha spostato i lavori verso la manifattura e la produzione industriale, centrata sulla produzione di beni fisici come il treno e l'automobile. Il 900' è invece stato il secolo della grande diffusione dei servizi.

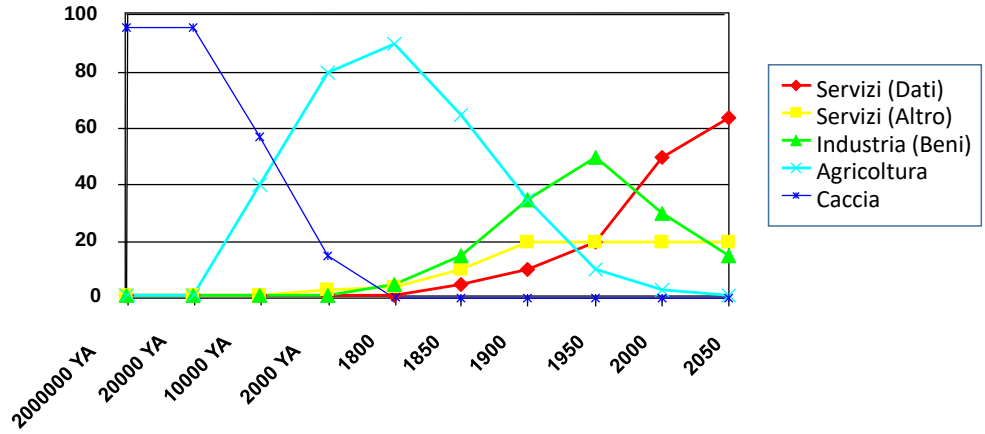


Figura 133 - I lavori dall'homo sapiens al giorno d'oggi, e oltre

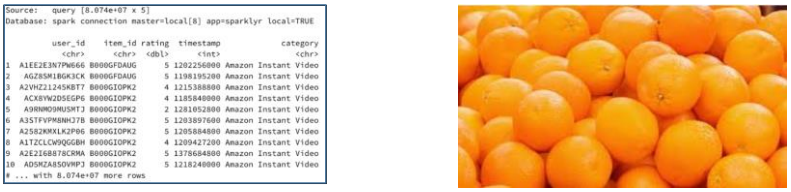
Il nostro secolo vede la esplosione dei lavori riguardanti la produzione e erogazione di servizi *basati sui dati digitali*, ad esempio tutti quelli forniti dalle app disponibili sul nostro telefono mobile; nella previsione della figura, questi servizi arriveranno a raggiungere entro il 2050 oltre il 60 per cento dei lavori.

Quindi, i dati digitali stanno ormai diventando il più importante fattore produttivo e la più importante fonte di servizi per la umanità. E lo stanno diventando perché, come abbiamo già notato, è sempre più grande la disponibilità di dati digitali, i cosiddetti big data, prodotti e alimentati dalle reti sociali, dall'internet delle cose, dai telefoni cellulari, e memorizzati e gestiti nel cloud. I dati sono il nuovo petrolio, disse il 12 dicembre 2011, la Commissaria europea Neelie Kroes lanciando un importante progetto europeo sui big data.

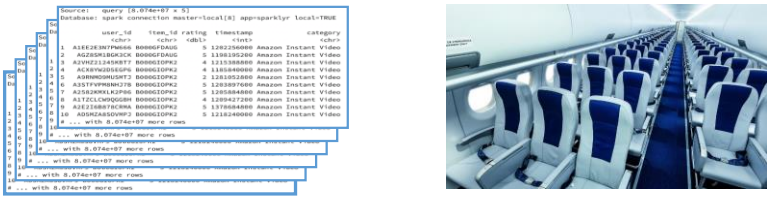
Ma dove si collocano i dati tra i beni e i servizi? Lo iniziamo a vedere con gli esempi della Figura 134; al contrario dei beni, i dati digitali non hanno fisicità, in questo sono simili ai servizi. Riguardo al confronto tra dati e servizi, i dati possono essere continuamente riusati, non si esauriscono mai; mentre, al contrario, i posti liberi su un aereo in un viaggio, che rappresentano un servizio, svaniscono quando il viaggio è terminato.

Insomma, i dati digitali sono un fenomeno nuovo e originale rispetto ai beni e ai servizi, e che ha impatto su diversi ambiti della nostra vita, come la economia e la società.

[16] M. Porat - *Info Economy: Definitions and Management*, 1977.



I dati non hanno la fisicità dei beni



I dati non svaniscono come i servizi

Figura 134 - Dati, beni, servizi

Nel seguito del capitolo approfondiamo il valore dei dati digitali, osservandolo secondo diverse prospettive e sfumature:

- anzitutto, parleremo di *valore d'uso dei dati digitali in un contesto generale*, discutendo alcune leggi che disciplinano il loro ciclo di vita, dalla raccolta dei dati alla loro elaborazione e analisi; una di queste leggi ci farà incontrare tra poco il *valore conoscitivo*.
- approfondiremo poi la tipologia di valore che abbiamo visto fino ad ora in questo capitolo, il *valore economico*.
- vedremo poi il *valore sociale*, che consiste, attraverso l'uso dei dati digitali, nel raggiungimento di un obiettivo che migliora la nostra vita.

Le leggi di Moody

Come per le leggi della fisica, ad esempio la legge di gravità che tutti conosciamo, anche i dati hanno le loro leggi di funzionamento, leggi che sono state formulate da Daniel Moody, vedi Figura 108.

Prima legge: I dati sono infinitamente condivisibili Seconda legge: Il valore dei dati cresce con il loro utilizzo Terza legge: I dati sono deperibili Quarta Legge: Il valore dei dati cresce con la loro accuratezza Quinta Legge: Il valore dei dati cresce quando sono combinati con altri dati Sesta Legge: "di più" non è necessariamente meglio, ovvero il fenomeno del data overload

Figura 135 - Le leggi di Moody sul valore dei dati

Vediamo adesso in maniera più approfondita la prima, la quinta e la sesta legge, una trattazione più approfondita verrà fatta nel volume della Enciclopedia dedicato al valore dei dati.

La *prima legge* dice che, al contrario dei beni, i dati sono infinitamente condivisibili (vedi Figura 136). Un esempio di questa legge è l'enciclopedia libera Wikipedia (www.wikipedia.org) che è consultabile da chiunque disponga di una connessione a Internet, senza limiti di accesso. Il valore d'uso in questo caso ha l'aspetto di *valore conoscitivo*.

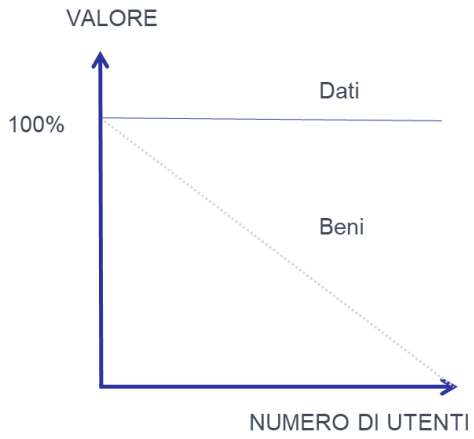


Figura 136 - I dati possono essere condivisi senza limite, non così i beni

Per quanto riguarda la *quinta legge*: Il valore dei dati cresce quando sono combinati con altri dati, un primo esempio riguarda una azienda di e-commerce che vende libri. Per una tale azienda, i dati di vendita e i dati sui clienti sono già di per sé, separatamente, dati di valore. Tuttavia, correlare questi due dataset, usarli insieme, rende i dati che se ne possono estrarre assai più preziosi, perché permette di concepire campagne pubblicitarie mirate.

La capacità di correlare le caratteristiche dei clienti con i profili di acquisto consente infatti di orientare le esperienze di acquisto in modo da promuovere I prodotti più adatti ai clienti più disponibili alla spesa nel momento giusto; è quello che fanno con enormi profitti i grandi players come Amazon, Google, Facebook, che, avendo a disposizione tantissimi dati su di noi, riescono a costruire un profilo per ciascuno di noi *mettendo insieme*, o come possiamo dire, *integrando* i dati disponibili.

Un secondo esempio riguarda il contrasto all'evasione fiscale o contributiva; nella Figura 137 supponiamo che *Persona, Società, Patrimonio, Reddito dichiarato* siano quattro tabelle, qui descritte per mezzo dei soli nomi. I legami tra tabelle sono rappresentati con *linee continue*; qui per legami intendiamo il fatto, ad esempio, che la tabella Persona e la tabella Reddito dichiarato *abbiano in comune il codice fiscale* della persona dichiarante, così che sia possibile formulare interrogazioni che utilizzino entrambe le tabelle insieme.

Nello schema a sinistra rappresentiamo con Q1, Q2, e Q3 le interrogazioni che si possono eseguire separatamente sulle coppie di tabelle. Se integriamo le tre coppie di tabelle in una unica base di dati, formata da quattro tabelle, potremo eseguire nuove interrogazioni come quelle indicate con Q23, che consentono di elaborare la *situazione patrimoniale complessiva* del contribuente, individuando eventuali anomalie, ad es. tra patrimonio e reddito dichiarato.

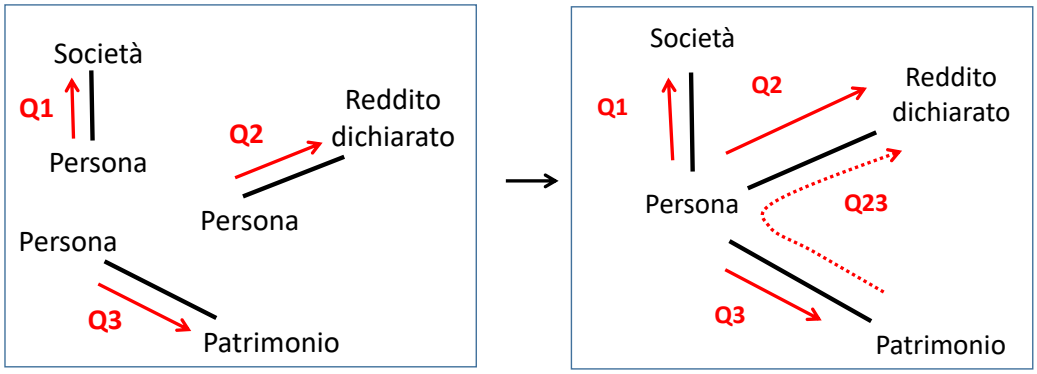


Figura 137 – Integrare porta valore

Infine, della *sesta legge* abbiamo parlato in occasione della discussione sul data overloal nel capitolo sulle astrazioni, a cui rimando.

L'economia digitale

I dati digitali sono un prodotto economico, o scambiati in quanto tali, ad esempio un giornale di carta o l'abbonamento per l'accesso alla copia digitale, ovvero scambiati tramite servizi basati sui dati (per es. sapere quanto devo aspettare il tram a una fermata).

La natura molto diversa dei dati digitali rispetto a beni e servizi, ha profondamente modificato il mercato dei dati digitali; non è questa la sede per esaminare sia pure superficialmente tutte le novità insite in questa rivoluzione[17]; anche qui ci soffermiamo su alcuni esempi. In Figura 138 vediamo uno dei mutamenti fondamentali intervenuti tra economia tradizionale e economia digitale: produrre due copie di un libro cartaceo costa circa il doppio che produrne una copia (al netto dei costi fissi per pagare i diritti d'autore e per l'avvio della produzione), mentre fare una copia di un file digitale non costa praticamente nulla.

Figura 138 – Un libro, due libri

[17] Due libri molto chiari sul tema sono
C. Shapiro e H. Varian – Information Rules: A strategic guide to information Economy, Harvard Business Reviw Press, 1999
S. Quintarelli – Capitalismo Immateriale: le tecnologie digitali e il nuov oconflitto sociale, Bollati Boringhieri, 2019
Si può anche leggere nel testo *La Scienza dei dati*, scaricabile dal sito https://boa.unimib.it/handle/10281/295980, il capitolo 13 *L'economia Digitale* di R. Masiero.

Un confronto più ampio tra economia tradizionale basata su beni e servizi prevalentemente fisici e economia digitale dei beni e servizi basati sui dati, e quindi immateriali, è mostrato in Figura 139.

Criterio di confronto	Beni e servizi prevalentemente fisici	Beni e servizi basati su dati digitali
Produrre	$$$$$$	$$
Riprodurre	$$$$	Quasi nulli
Conservare	$$	Quasi nulli
Trasferire	$$	Quasi nulli
Tempo per trasferire	Elevato	Immediato
Orario di lavoro	Orario lavorativo	24/365
Ritorni economici	Decrescenti	Crescenti
Interconnnessione	ridotta	Sempre connessi

Figura 139 – Confronto tra beni materiali e beni e servizi immateriali basati sui dati digitali (adattata dal testo di S. Quintarelli citato in nota)

Produrre beni fisici costa di più perché sono coinvolte materie prime. L'esempio del libro di Figura 138 fa riferimento alla seconda riga relativa al *riprodurre*. *Conservare* comporta costi comparativi analoghi ai precedenti; così pure il *trasferire* in una rete di trasporti materiali incide significativamente sui costi totali dei prodotti, mentre trasferire dati digitali in rete non costa praticamente nulla. Analogamente, con riferimento al *tempo per trasferire*, avete mai provato ad acquistare un e-book? Dopo pochi secondi lo avete sul vostro lettore e lo potete cominciare a leggere, al contrario del libro cartaceo che vi arriva dopo alcuni giorni, o che dovete andare ad acquistare in libreria.

Proprio l'esempio del libro ci dice che il commercio elettronico fa scomparire molte attività e luoghi fisici attraverso cui tradizionalmente avviene la distribuzione, come ad esempio le edicole, oramai spesso relegate a vendere figurine o riconvertire a uffici postali, e, insieme, fa scomparire molti posti di lavoro.

Ancora sulla Figura 139, *l'orario di lavoro* è rigido per i beni e servizi tradizionali, mentre è senza interruzioni nella economia digitale; riguardo a questo aspetto occorre riflettere sul fatto che, durante la epidemia Covid, essere forzati a rimanere a casa ha fatto crescere significativamente lo smart working, che ricade nella colonna a destra della tabella comparativa.

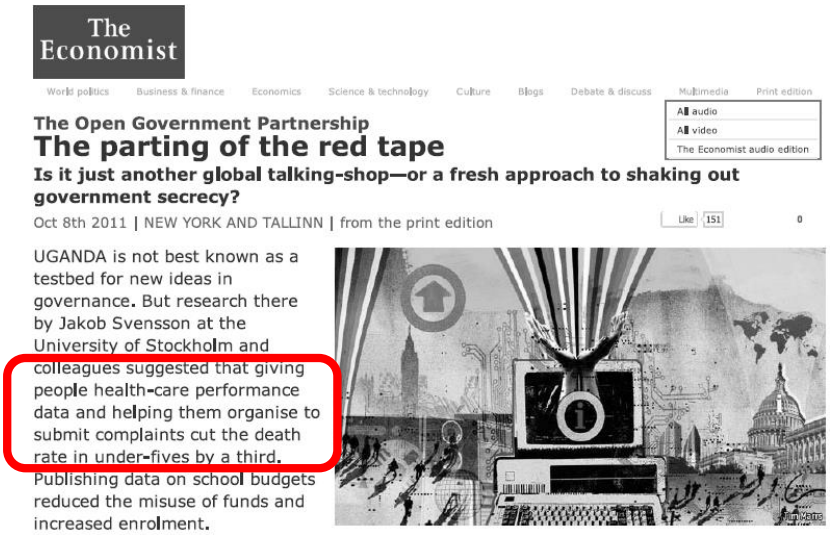
I *ritorni economici*, in virtù della profonda modifica della struttura dei costi vista nelle prime righe, sono decrescenti nel mondo dei prodotti fisici con l'aumento dei prodotti venduti, son o invece crescenti in presenza di costi di riproduzione quasi nulli. Infine la *interconnessione* tra le aziende nella filiera produttiva è costosa nelle economie tradizionali, mentre è continua (always on) nell'ambito della economia digitale, portando ad una enorme espansione delle aziende interconnesse in rete rispetto alle aziende tradizionali, fenomeno significativamente

più di più ampia portata del passaggio nella economia tradizionale tra negozi di prossimità e grandi ipermercati.

Attenzione, per concludere: la Figura 139 *non ci dice che le opportunità e i posti di lavoro cresceranno con il digitale* (su questo aspetto le previsioni sono ancora non univoche) ma ci dice che in percentuale aumenteranno i lavori in ambito digitale, rispetto a quelli più tradizionali.

Il valore sociale

Parliamo ora del valore sociale dei dati. Per introdurlo, osserviamo la Figura 140, 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

World politics Business & finance Economics Science & technology Culture Blogs Debate & discuss Multimedia Print edition

All audio
All video
The Economist audio edition

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

Like 151 0

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 140 – Il valore sociale dei dati: da un numero dell'Economist del 2011

Il *valore sociale dei dati* può essere definito come la capacità che hanno i dati di fornire una risposta alle esigenze delle persone e 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 in termini di miglioramento della qualità della vita. Gli ambiti riguardano: lo stato di salute (in cui ricade il caso dell'Uganda), la formazione, le competenze, il lavoro, la sicurezza personale ed altri ancora.

Un secondo esempio di valore sociale dei dati, legato alla epidemia Covid, è mostrato nella Figura 141.

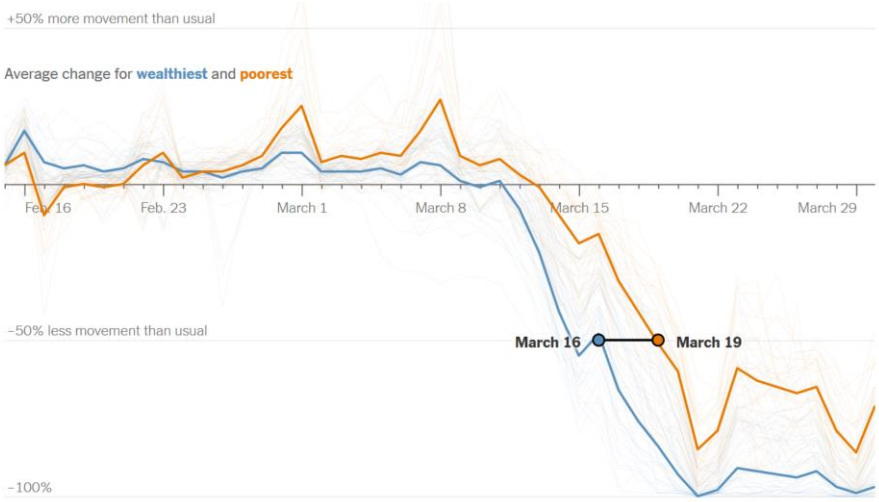


Figura 141 – La mobilità durante il Covid è ineguale tra le classi sociali

Nella figura vengono confrontate la mobilità dei più poveri e dei più ricchi durante il mese di marzo 2020 in alcune aree urbane degli Stati Uniti; la figura, a mio parere, dimostra che il lockdown per alcune classi sociali è stato un lusso che non potevano permettersi, in virtù della necessità di andare a lavorare per poter vivere. Il valore sociale della elaborazione mostrata in figura è per i governi degli stati e per il governo centrale, che possono tenerne conto nelle politiche sociali.

Vediamo ora cosa ci aspetta nel seguito del capitolo. Afferiscono al tema del valore sociale due importanti aspetti connessi all'uso dei dati digitali:

- l'effettiva possibilità di raccogliere ed elaborare dati sulla condizione sociale e sui servizi necessari alla vita di una popolazione o comunità, aspetto che collochiamo nell'ambito del *data divide*, vedi tra poco.
- l'effettiva possibilità di accedere e collegare tra loro i dati che vengono diffusi nel Web, che riguarda il tema dei *dati aperti*.

Il data divide

Con il diffondersi dei dati digitali, si intende con *data divide* la relazione asimmetrica tra coloro che nel settore pubblico e soprattutto nel settore privato raccolgono, archiviano e analizzano grandi quantità di dati e la vastissima comunità di persone e comunità che vorrebbe poter utilizzare in maniera estesa i dati nella propria vita personale, *ma non vi riesce*. Ciò avviene per mancanza anzitutto *dela possibilità di accesso* ai dati, e, secondariamente, per la non conoscenza delle tecniche e modelli con cui poter comprendere il significato dei dati, e delle tecniche e dei linguaggi per accedere ai dati e analizzarli.

La forma più critica di "big data divide" è nella asimmetria tra I dati di profilazione disponibili ai grandi operatori privati mondiali, come Facebook e Google, e *coloro che forniscono i dati*, cioè tutti noi, che in questo scambio non riceviamo nessun vantaggio. Questi dati permettono ai grandi operatori di fare enormi guadagni con la pubblicità, che può essere fornita agli utenti in modo mirato, basandosi sulle loro preferenze.

Un'altra asimmetria, questa volta tra paesi ricchi e paesi poveri, è nella conoscenza resa disponibile sul territorio. Osserviamo la Figura 142, 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 sulle strade, gli edifici, i servizi pubblici, le sedi degli esercizi commerciali, i ristoranti, ecc. è infinitamente maggiore rispetto alla piatta e poco informativa rappresentazione di Kibera.

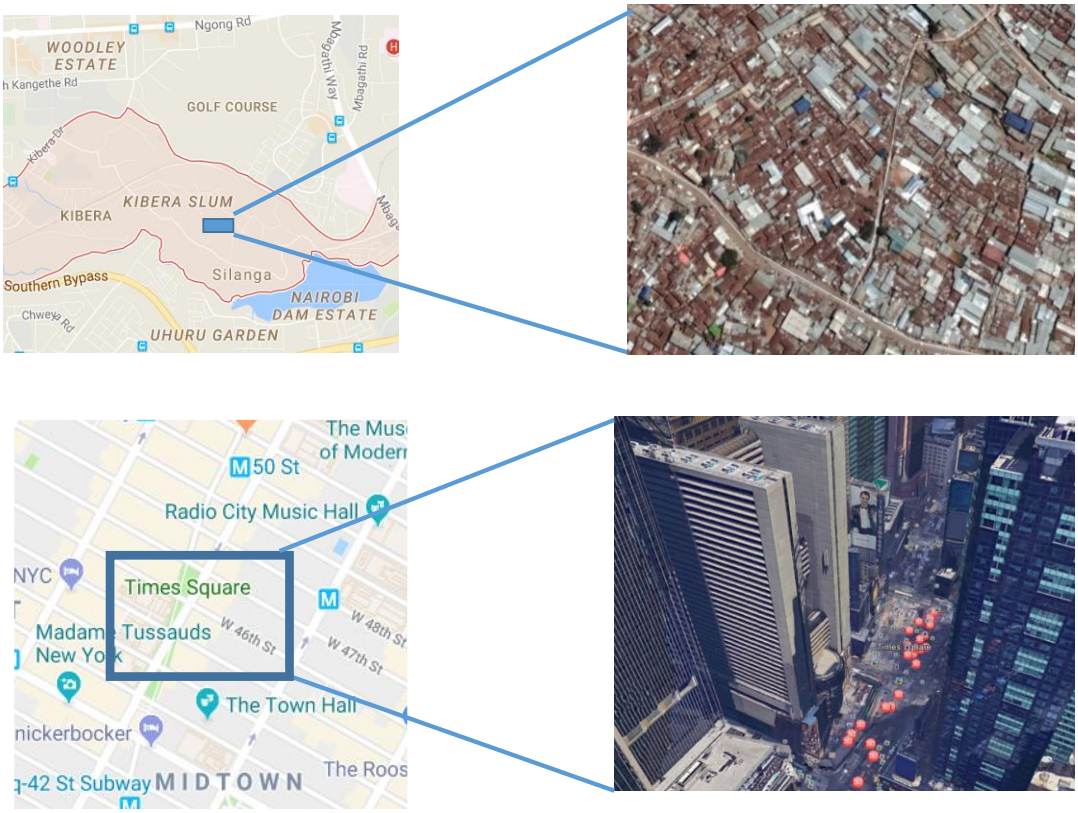


Figura 142 – Data divide in Google Earth

Anche I big data alimentati attraverso il "crowdsourcing", cioè il contributo volontario degli utenti, come ad esempio Open Street Map, rappresentano in modo ineguale il territorio, perche' anche in questo caso alcuni tematismi, ad esempio i luoghi turistici, vengono rappresentati in modo molti più ricco di altri, come, ancora, gli slum e le zone povere del territorio.

Infine, essendo molti dataset accessibili solo a pagamento, la loro disponibilità, ad esempio per la ricerca e per le analisi sociali, è garantita in modo ineguale. Questo porta a un grande disequilibrio nelle possibilità di sfruttamento espansione economica, turistica, commerciale tra i diversi territori.

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 fortemente sottostimato.

I dati aperti

Nel Web i dati possono essere illimitatamente condivisi, non ci sono gerarchie, tutti gli utenti sono potenzialmente parte di una comunità tra pari. D'altra parte, quando il Web viene utilizzato per condividere dati digitali, non possiamo immaginare che ciò possa essere fatto, banalmente, pubblicando delle tabelle di dati strutturati.

Se una persona, come in Figura 143, pubblica sul Web una tabella in linguaggio italiano sui premi Nobel italiani, e in Australia, a Sydney, qualcuno pubblica una tabella in inglese sui premi Nobel in Letteratura, come si fa a collegare sul Web le due tabelle per i *dati comuni*, visto che nel modello relazionale i collegamenti tra dati avvengono per mezzo di valori?

Premi Nobel italiani

Cognome	Nome	Ambito	Data Nascita
Deledda	Grazia	Letteratura	28091871
Natta	Giulio	Chimica	26021903
Pirandello	Luigi	Letteratura	28061867
Rubbia	Carlo	Fisica	31031934
…	……	…….	…..

Milano, Italia

← ?→

Nobel Prizes in Literature

Given Name	Last Name	Date of Birth	Place of Birth
……	…	…	…
Luigi	Pirandello	28/06/1867	Agrigento
Grazia	Deledda	28/09/1971	Nuoro
……..	….	….	….

Sydney, Australia

Figura 143 – Due tabelle pubblicate nel Web

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 aperti pubblicati sul Web, corrispondenti ai cinque livelli (chiamati anche *stelle*, indicate con il simbolo *), mostrati in Figura 144.

Una * - Rendere disponibili i dati sul Web, in qualunque formato siano rappresentati, con una licenza open che li rende utilizzabili da tutti.

Due ** - Rendere disponibili i dati sul Web come dati strutturati, ad esempio in Excel invece che in formato scannerizzato.

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).

Quattro **** - Usare identificatori universali di risorsa per denotare gli oggetti descritti dai dati.

Cinque ***** - Collegare i dati ad altri dati nel Web per condividerli ed integrarli nella comunità mondiale degli utenti.

Figura 144 – Le cinque * di Berners Lee

Utilizzando il quarto livello e il quinto livello, per collegare le due tabelle possiamo usare (vedi Figura 145) *identificatori universali di risorsa* (un termine complicato per indicare gli indirizzi nei browser che iniziano con http://....) e i collegamenti tipici dei grafi semantici (la freccia che indica l'identificatore di risorsa nella figura).

https://www.semoromani.it/nobel/deledda

Cognome	Nome	Ambito	Data Nascita	URL
Deledda	Grazia	Letteratura	28091971	

https://www.sydenylibraryi.it/nobelprize/deledda

Given Name	Last Name	Date of Birth	Place of Birth
Grazia	Deledda	28/09/1971	Nuoro

Figura 145 – Collegamento tra i due dati mediante identificatori universali di risorsa

Questa modalità, consistente nel pubblicare dati in formato aperto e nel collegarli usando il quarto e quinto livello di Figura 144, ha dato luogo alla creazione e progressiva espansione del *Linked Open Data Cloud*, una immensa fonte di dati condivisi, tra cui ad esempio l'enciclopedia Wikipedia, che possono essere utilizzati per tantissime applicazioni. Vedi la sua evoluzione dal 2007 al 2017 in Figura 146.

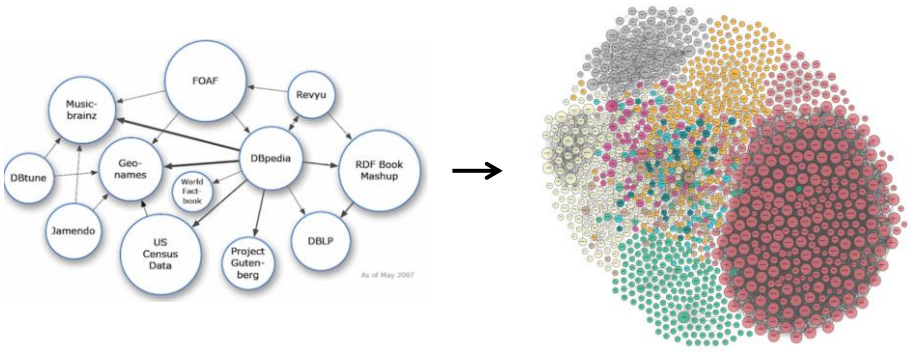


Figura 146 - Condivisione di dati nel Web, il Linked Open Data Cloud

Valore sociale e dis-valore economico

Finora abbiamo considerato il valore economico e valore sociale separatamente, come fossero due modi alternativi di misurare il valore dei dati. In realtà sappiamo che le scienze economiche e le scienze sociali sono profondamente legate, e l'emancipazione di gruppi sociali economicamente sfavoriti può passare solo da una più equa distribuzione della ricchezza.

Come semplice esempio di tensione tra valore sociale e valore economico, consideriamo l'iniziativa della polizia inglese che pubblica sul Web per ogni città e per ogni quartiere della città nel territorio inglese (https://www.police.uk/) una mappa dei reati compiuti in un determinato arco temporale, con numerosità e tipo di reato, vedi Figura 147. Se 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 di residenza; la polizia inglese infatti pubblica sul sito https://data.police.uk/ i dati relativi alla frequenza di varie tipologie di reati commessi nelle strade e quartieri delle città inglesi, vedi Figura 147, che rappresenta statistiche sui reati nella città di Leicester.

La sicurezza è tra gli ambiti che l'OCSE considera rilevanti per il miglioramento della qualità della vita; quindi possiamo dire che i dati sui reati resi disponibili dalla polizia inglese hanno

valore sociale. Allo stesso tempo, essi possono influenzare negativamente il prezzo degli affitti delle case, perché la domanda di case e il prezzo che i locatari sono disposti a pagare è inferiore quando cresce il rischio di subire furti o aggressioni.

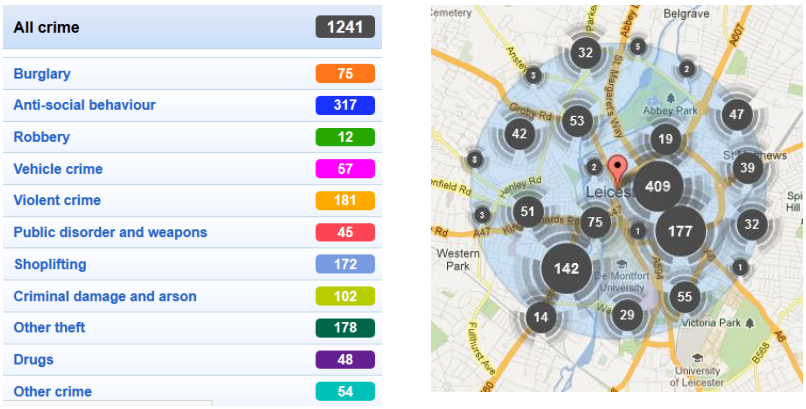


Figura 147 – Reati a Leicester

C'è dunque un contrasto evidente tra il valore che quei dati hanno per i locatari e il dis-valore che ne hanno gli affittuari. Questa peraltro non è una novità, è difficile che l'osservazione di un fenomeno sociale possa accontentare tutti, e il valore d'uso è valutato dalle persone, dalle comunità e dalle istituzioni secondo il proprio punto di vista soggettivo.

Gli stessi dati possono essere usati da tutti coloro (vigili del fuoco, ambulanze, pronto soccorso, etc.) che svolgono attività di servizio in caso di grandi eventi, manifestazioni, concerti, ovvvero che possono pianificare l'impiego delle risorse per interventi di ordine pubblico, o, negli stati totalitari, per interventi repressivi della libertà di manifestazione.

Insomma, ***i dati non sono neutrali***, possono essere utilizzati per tanti fini, e sta a noi essere comprendere questi diversi fini, per essere consapevoli e decidere quale valore abbiano per noi, e comportarci di conseguenza.

Riassumendo

I dati digitali sono usati a vario titolo da aziende, organizzazioni pubbliche, comunità e singole persone. A seguito di questa grande varietà di usi, i dati possono avere diversi tipi di ***valore***, come il ***valore d'uso***, ***valore economico***, ***valore sociale***, ***valore conoscitivo***.
I dati sono diversi rispetto ai due tradizionali ***oggetti di scambio***, i ***beni*** e i ***servizi*** e rispettano diverse ***leggi economiche***, negli aspetti che riguardano la produzione, la riproduzione, lo scambio, ed altre, dette ***leggi di Moore***, come ad esempio il fatto che al contrario dei beni e dei servizi i dati scambiati non si esauriscono. Il ***valore d'uso*** può vedersi come un bilancio tra ***benefici*** che abbiamo nell'usare i dati, e i **sacrifici** che dobbiamo compiere per ottenerli, tra cui il ***costo economico*** dei dati ***e lo sforzo*** che dobbiamo compiere per ottenerli***.*** Il ***valore sociale*** dei dati fa riferimento a quanto la conoscenza che ci portano migliora la nostra ***qualità della vita***. ***Valore economico*** e ***valore sociale*** possono creare valori d'uso in contrasto tra di loro, se aumenta il valòore sociale per una persona, può diminuire il valore economico per un'altra.
I dati possono rappresentare il mondo e possono essere acceduti in modo diseguale (fenomeno del ***data divide***), accrescendo in tal modo le diversità tra nazioni e tra comunità. I ***dati aperti*** sono dati che per le loro caratteristiche e modalità di pubblicazione possono essere acceduti da tutti e collegati tra di loro nel Web.

Capitolo 13

La sfera cognitiva ed emozionale

Il Web e la sfera cognitiva

Molti esempi portati nei capitoli precedenti mostrano che l'attività di interpretazione dei dati digitali che ci arrivano dalle reti sociali, dal nostro telefono mobile, dai siti Web cui accediamo per trovare una informazione o accedere a un servizio, richiede uno sforzo cognitivo. Tale sforzo cognitivo trova i suoi limiti nel concetto di razionalità limitata di Herbert Simon; Simon afferma che la razionalità di un individuo è limitata da vari fattori: la conoscenza che possiede, i limiti cognitivi della sua mente, la quantità finita di tempo di cui dispone per prendere una decisione.

Sebbene il Web abbia ridotto i costi connessi alla ricerca dei dati, accrescendo la loro accessibilità, rimangono nella interpretazione dei dati, e sono relativamente incomprimibili, costi significativi di interazione, in virtù della vastissima area di dati disponibili, e in virtù della loro eterogeneità e rapidità di variazione nel rappresentare i fenomeni della realtà.

Il 14 gennaio del 2021, una settimana circa dopo il drammatico assalto al Campidoglio USA dei facinorosi pro-Trump, ho sottoposto al motore di ricerca Google le parole in italiano "assalto al campidoglio usa" e, come si vede in Figura 148, ho ricevuto in risposta 3.290.000 pagine. Le parole "Capitol attack" hanno prodotto in risposta 405.000.000 di pagine. E' esperienza comune che quando si sottopone a Google o altro motore una ricerca per parole chiave, a malapena di tali pagine riusciamo a esaminarne le prime dieci o venti; il loro ordine, peraltro, è deciso da Google secondo un algoritmo che adotta criteri e tecniche molto complesse, e influenzate dal gettito pubblicitario.

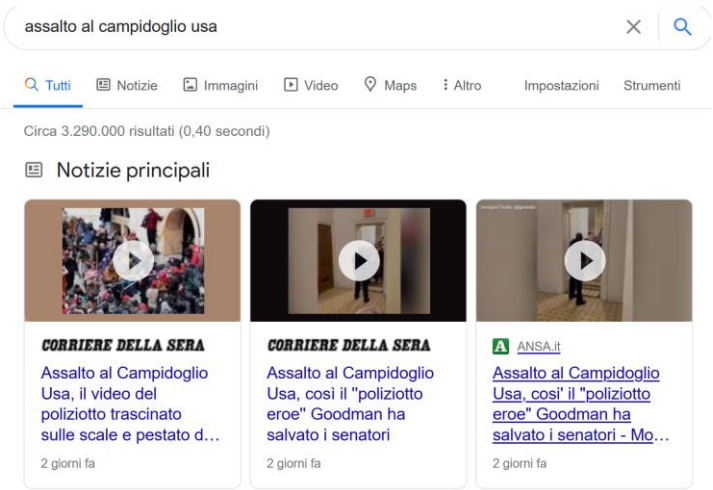


Figura 148 – I risultati di una ricerca su Google

Lo sforzo cognitivo che possiamo mettere in campo quando scorriamo i risultati di una ricerca sul Web, o quando interagiamo con una rete sociale è limitato, e quasi sempre, nella suddivisione tra pensiero lento e pensiero veloce di cui al libro di D. Kahneman[18], noi optiamo per il pensiero veloce. E' perciò ragionevole che dalla psicologia cognitiva non arrivino buone notizie sulla nostra capacità di filtrare notizie vere e notizie false. Mentre accettare per credibile un dato che ci arriva da una fonte non richiede particolare sforzo, riconoscere la cattiva informazione richiede processi cognitivi complessi, ricerche, confronti.

Abbiamo visto nel Capitolo 7 dedicato alla qualità dei dati che i dati sono "cocciuti", ma quante volte siamo disponibili a indagare sulla provenienza del dato, a cercare riscontri in altre fonti, a cercare conferme o confutazioni? 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 addirittura rafforzare le loro convinzioni. Di conseguenza, non è tanto rilevante *ciò che* la gente pensa, ma *come* pensa.

Parafrasando un po' scherzosamente la prima pagina di Anna Karenina, mentre tutte le notizie vere sono vere allo stesso modo, le notizie false (quelle che vengono chiamate con un termine molto abusato *fake news*) sono false in modo diverso; intendo dire che una notizia può essere deformata rispetto alla sua versione vera in tanti modi, e quindi esistono tante versioni diverse e false di una notizia che possono essere spacciate per una notizia vera.

D'altra parte, come scrive Cherubini nel testo [19]"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) sia nel loro diffondersi nella società. Non è certo una novità che le persone tendano a credere in massa alle fandonie più implausibili, totalmente prive di fondamenti empirici."

Se andiamo a vedere la voce "List of cognitive biases[20]" di Wikipedia, troviamo che le distorsioni o pregiudizi che possono influire sulla nostra capacità di obiettiva interpretazione della realtà possono riguardare:

- la *formazione delle credenze*, i processi di ragionamento, le decisioni che prendiamo.
- I *comportamenti sociali*, nella interazione con gli altri.
- I *processi attraverso cui ricordiamo o riportiamo a coscienza* fatti e avvenimenti.

Ebbene, Wikipeda nella sua voce "List of biases" elenca oltre 140 tipi di distorsioni delle tre tipologie! Si può comprendere che la ricerca su quali siano e come si differenzino rispetto alle precedenti le distorsioni nella nostra mente derivanti dalla interazione con reti sociali e siti Web e dalla acquisizione di dati digitali, siano solo molto recenti. Quali sono le specificità cognitive che caratterizzano i dati digitali rispetto alle tradizionali forme di acquisizione di conoscenza?

[18] D. Kahneman, Pensieri lenti e veloci, Mondadori 2020.
[19] P. Cherubini – Capitolo 17 Big Data e psicologia – Luci e ombre del testo di C. Batini e altri "La Scienza dei dati" liberamente accessibile e scaricabile dal sito https://boa.unimib.it/handle/10281/295980.
[20] https://en.wikipedia.org/wiki/List_of_cognitive_biases

Per Lorusso [21] "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".

La filosofia del linguaggio, vedi ancora Lorusso, ci dice anche che nel mondo contemporaneo sempre di più la costruzione del senso si dà per via *narrativa, attraverso narrazioni*; 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 *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."

Accanto a questo inquadramento nella filosofia del linguaggio, la ricerca recente si è soffermata sulle *euristiche,* cioè sui procedimenti approssimati, che in virtù della nostra razionalità limitata tendiamo ad utilizzare nella valutazione della credibilità nel Web, vedi Figura 149.

- SE LO DICE LUI/LEI, MI FIDO!- basato sulla *reputazione*
- NON MI FIDO PIU'! – basato sulla *delusione*
- C'E' QUALCOSA CHE NON MI TORNA ?!? basato sul *confronto tra fonti*
- MI STA IMBROGLIANDO, NON GLI CREDO! - tipico della *informazione commerciale*.
- QUESTO GIORNALE, SITO, MESSAGGIO TWITTER HA RAGIONE, DICE QUELLO CHE DICO IO! – basato sulla *auto-conferma*.

Figura 149 – Le tipiche euristiche che usiamo per valutare la credibilità dei dati

Le euristiche di Figura 149 sono basate su:

- la *reputazione*, che porta a privilegiare alternative riconoscibili rispetto a quelle meno familiari.
- la *violazione delle aspettative*, che assume una fonte non credibile se essa ha violato le aspettative in precedenti circostanze.
- la *consistenza*, che riguarda il confronto tra fonti per evidenziarne le differenze. Nel caso di inconsistenze, sono proposte varie tecniche per la scelta tra le alternative.
- *l'intento persuasivo*, che tende a non considerare credibile il dato che viene percepito come affetto da un pregiudizio (bias); questa euristica è tipica della informazione commerciale.
- *l'auto-conferma*, che misura la credibilità di un sito o di un messaggio, sulla base della conferma delle precedenti credenze. Rientrano in questa tematica, cui dedichiamo un po' più di attenzione, le analisi sulle *camere dell'eco*, investigate in Italia dal gruppo di ricerca

[21] A.M.Lorusso - Postverità, Editori Laterza, 2017.

di Walter Quattrociocchi. Una analisi effettuata dal gruppo di ricerca 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, porta alla seguente conclusione: gli utenti che 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 o *camera dell'eco.*

L'euristica di auto-conferma è una delle cause della polarizzazione delle convinzioni politiche negli Stati Uniti, la cui evoluzione dal 1994 al 2017 è rappresentata nella ricerca di cui alla visualizzazione di Figura 150, pubblicata dall'Economist.

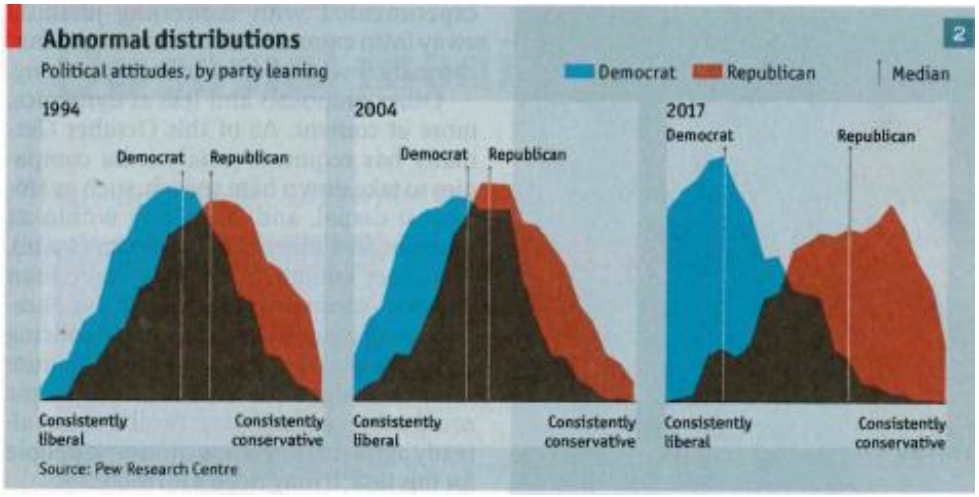


Figura 150 - Polarizzazione delle convinzioni politiche negli Stati Uniti

Informazione diffusa nelle reti sociali e nei giornali

Focalizziamo ora l'attenzione su uno strumento molto usato nella comunicazione sociale e politica, come dimostrato in maniera veramente stupefacente dall'uso che l'ex Presidente Trump ne fece nella relazione con i suoi sostenitori in tutto il suo mandato e nella drammatica giornata del 6 gennaio 2021 in cui ci fu l'assalto al Campidoglio: sto parlando di Twitter. Le grandi novità dei messaggi Twitter e in generale delle tecnologie di comunicazione che si sono sviluppate negli ultimi venti anni rispetto alla comunicazione tradizionale sta nei seguenti aspetti:

- il cambio di ordine di grandezza (1 a 1.000, 1 a 10.000, 1 a 100.000) nella quantità di dati che viene diffusa,
- la velocità con cui questa informazione si diffonde,
- il suo carattere digitale e quindi elaborabile e ri-diffondibile,
- la crescita nella dimensione dei soggetti che possono generarla,
- la crescita nella dimensione dei soggetti che possono acquisirla,
- la riduzione della sua qualità e delle risorse cognitive investite nel generarla,

- la grande frammentazione e la scarsa contestualizzazione con cui viene generata e fruita,
- l'assenza di costi economici nel produrla e nel diffonderla,
- la difficoltà di ricostruirne la storia, e la sequenza degli eventi e dei soggetti che hanno contribuito a generarla.

Ormai nella comunicazione politica i messaggi Twitter stanno sostituendo le interviste o i contributi ai giornali. Una descrizione comparativa di alcune caratteristiche tra un articolo di giornale e un messaggio Twitter compare nella Figura 151; per una analisi dei meccanismi economici nella produzione di giornali e nelle strategie di quelli che chiama *nuovi editori* (reti sociali, sistemi di condivisione, ecc.) si veda anche l'articolo di Stefano Quintarelli riportato in nota [22].

Caratteristica	**Articolo di giornale**	**Messaggio Twitter**
Tempo richiesto per la produzione	Alto	Basso
Tempo richiesto per la percezione	Alto	Basso
Numero di utenti potenziali	Milioni	Miliardi
Livello di qualità	Alto	Basso
Costo della qualità	Alto	Basso
Professionalità richiesta	Alta	Basso
Possibilità di ridiffusione	Bassa	Alta
Livello di emootività	Bassa	Alta

Figura 151 – Confronto tra caratteristiche di un articolo di giornale e di un messaggio Twitter

L'uso delle reti sociali come forma di comunicazione oramai prevalente tra persone sta provocando un progressivo declino delle forme di comunicazione tradizionali, quali i giornali quotidiani; Rusbridger[23] 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, e con alto costo umano: 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.

22 Stefano Quintarelli sul Foglio del 13 gennaio 2019 - Perché internet è tutto un articolo acchiappa click? Follow the money
https://www.ilfoglio.it/tecnologia/2019/01/13/news/perche-internet-e-tutto-un-articolo-acchiappa-click-follow-the-money-232291/

23 A. Rusbridger - Breaking News: The Remaking of Journalism and Why It Matters Now, Casnongate 2018 snongate 2018

Nelle grandi testate giornalistiche la cura per fornire notizie precise è sempre stata molto attenta. E tuttavia, per il Guardian[24] 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*. Le notizie grezze e non verificate sono gratuite e le notizie verificate sono molto costose.

Infine, mentre la responsabilità dei contenuti nei giornali tradizionali è identificata in una precisa figura, il direttore, non accade la stessa cosa nei social network, che a lungo hanno rivendicato il diritto di ospitare qualunque tipo di contenuti in virtù di una presunta libertà di opinione; solo in occasione dell'ultimo convulso periodo della Presidenza Trump alcune reti sociali hanno dapprima sottolineato i suoi messaggi come potenziali fake news e poi bloccato l'account della persona più potente del mondo.

Alcuni analisti ritengono che il governo dei contenuti sul Web non dovrebbe essere svolto dai players privati, che sono sempre guidati nel loro business da criteri economici, quanto piuttosto da autorità indipendenti; a mio parere, ciò è estremamente complesso da attuare, vista la dimensione mondiale del Web.

Il Web e le emozioni

La discussione precedente ha mostrato come l'estensione raggiunta dalle reti sociali abbia incrementato enormemente il tasso di comunicazione diretta tra esseri umani, rischiando allo stesso tempo di distorcere profondamente i rapporti sociali e la comunicazione interpersonale.

Uno studio, i cui risultati sono rappresentati nella visualizzazione di Figura 152, ha analizzato la diffusione in una rete sociale dei *sentimenti*. Dapprima è stato determinato il sentimento prevalente di ogni tweet nel database predisposto, analizzando gli *emoticon* che contenevano; altri studi hanno analizzato le parole o altre caratteristiche semantiche delle frasi.

I ricercatori hanno diviso i tweet in quattro categorie, esprimenti *gioia, tristezza, rabbia o disgusto*. Hanno a questo punto studiato il modo in cui i sentimenti si diffondono attraverso la rete. Ad esempio, se una persona ha inviato un tweet arrabbiato, hanno calcolato la frequenza dei messaggi dei destinatari caratterizzati dalla stessa emozione, e da emozioni diverse.

I risultati sono stati sorprendenti; Quando si fa riferimento a tristezza e disgusto, è stata trovata pochissima correlazione tra gli utenti; la tristezza e il disgusto non si diffondono facilmente attraverso la rete. E' stata trovata maggiore correlazione tra gli utenti che hanno twittato messaggi gioiosi, e la correlazione più alta è stata di gran lunga tra gli utenti arrabbiati; la rabbia ha una correlazione sorprendentemente più alta di tutte le altre emozioni.

[24] 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.

Nella Figura 152 vediamo la struttura della rete così costruita, in cui ogni nodo corrisponde a un utente, e i link corrispondono alle interazioni tra diversi utenti. Ogni nodo è colorato con la emozione prevalente, cioè l'emozione a cui è associato il massimo numero di tweet nel periodo temporale considerato. Le regioni dello stesso colore indicano che nodi vicini condividono la stessa emozione.

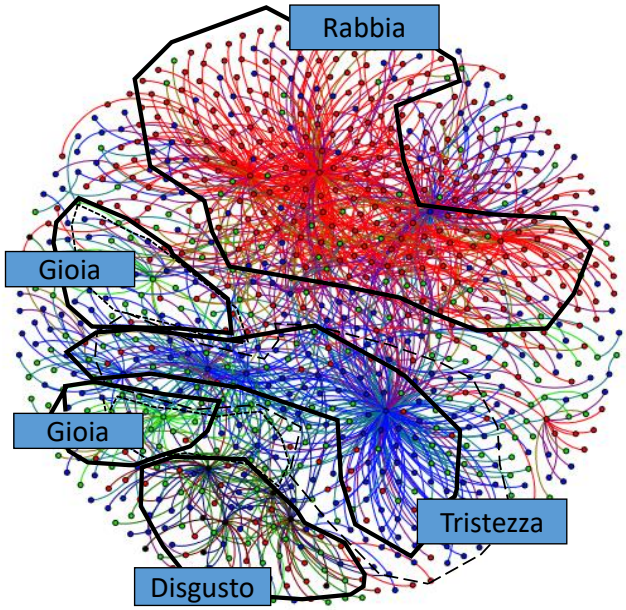


Figura 152 – Diffusione della rabbia, gioia, disgusto e tristezza
nei Tweet scambiati in un periodo temporale analizzato

Con riferimento specifico alla rabbia, come sostiene Slodertijck[25] la rabbia è un sentimento insopprimibile, che attraversa tutte le società, alimentato da coloro i quali, a torto o a ragione, ritengono di essere esclusi, discriminati o poco ascoltati.

Secondo Sloterdijk, storicamente in occidente è stata la Chiesa a dare uno sbocco a questa enorme accumulazione di rabbia e successivamente, a partire dalla fine dell'ottocento, i partiti della sinistra. Che hanno svolto, secondo Sloterdijk, la funzione di "banche della collera", accumulando energie che, anziché essere spese nel momento, potevano essere investite per costruire un progetto più grande. Negli anni recenti la rabbia è cresciuta in maggior misura per motivi religiosi ed è stata favorita da politiche populiste e nazionaliste.

Altri lavori considerano un insieme di emozioni più ampio delle quattro viste in precedenza, includendo la paura, la sorpresa, la fiducia, e generiche emozioni con polarità positiva e negativa. Inoltre, associano a ciascuna emozione un insieme di contesti e di parole a cui essa può essere associata, come ad esempio per la *sorpresa* i due contesti costituiti dalla *sorpresa estetica* nella visione di musei e opere d'arte, e la *sorpresa costituita dall'assistere a una magia*, allargando in questo modo l'insieme delle parole le cui occorrenze nel testo del

[25] Peter Sloterdijk – Ira e tempo, Marsilio, 2019.

messaggio vengono conteggiate; i risultati della analisi linguistica sono utilizzati per valutare la presenza, insieme alle emozioni, anche di sintomi di disturbi psichici come la depressione.

Concludo questa sezione osservando che i risultati descritti mostrano chiaramente una attività di ricerca ancora di tipo esplorativo, che non ha raggiunto una chiara maturità.

Valore emozionale

Oltre ai valori analizzati nel Capitolo 12, il valore conoscitivo, economico e sociale, c'è un altro valore che i dati ci possono trasmettere, e che discutiamo in questo capitolo, quello *emozionale*. Guardate in Figura 153 la copertina e una pagina del libro *Dear data* di Stefanie Posavec e Giorgia Lupi, ed. Particular Data.

In questo libro sono riportate tutte le cartoline che le due autrici si sono scambiate in una corrispondenza durata diversi anni, cartoline in cui rappresentano tanti tipi di dati con vari tipi di visualizzazioni, che ne accrescono il valore emozionale, cioè la intensità delle emozioni che proviamo nello sfogliare il libro. Nel testo i dati diventano *cari dati*, assumono una dimensione più vicina alla nostra sensibilità, e un po' si umanizzano rispetto ai dati digitali.

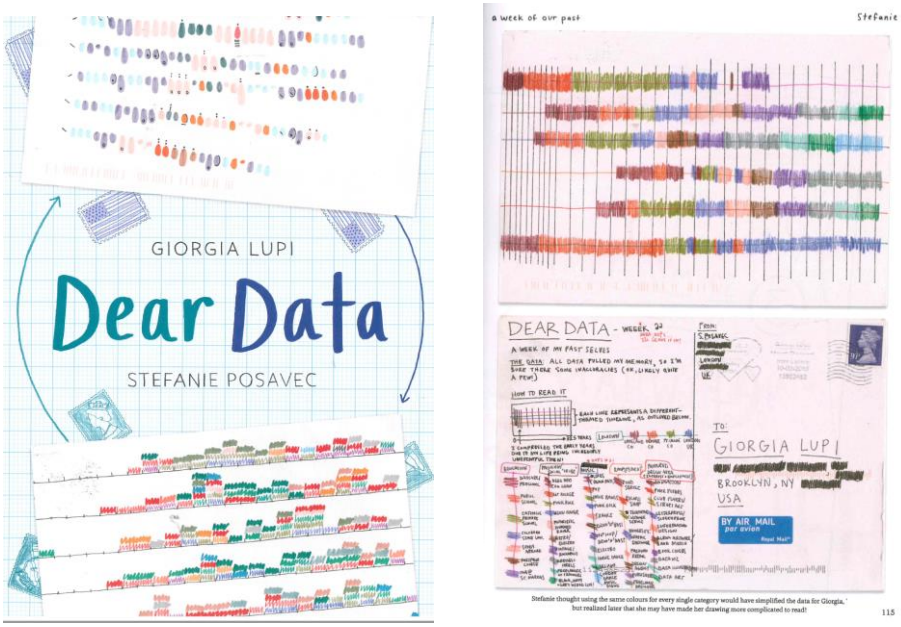


Figura 153 – Dear data

Attraverso il Web, i dati possono trasmettere tante emozioni positive: le poesie, le fotografie, tutto quello che possiamo trovare online, è un dato digitale che ci può trasmettere conoscenza o emozioni. Naturalmente, è più bello ed emozionante vedere il lago Antermoia dal vivo (vedi Figura 154), è più bello sfogliare i libri di carta, ma nel caso in cui non avessimo queste possibilità, i dati digitali ci consentono di estendere la nostra conoscenza, e vivere e condividere con gli altri su Instagram o con Whatsapp emozioni simili a quelle provate nella nostra vita analogica.

 che

Figura 154 – Il lago Antermoia nelle Dolomiti di Fassa

Riassumendo

La ***psicologia cognitiva*** ci dice che quando ***comunichiamo*** con altri per mezzo di dati digitali ovvero ***acquisiamo*** dati digitali dal Web noi possiamo dedicare solo uno sforzo limitato per comprenderli (***razionalità limitata***), per cui spesso usiamo ***euristiche***, cioè procedimenti approssimati, che ci possono portare a ***distorsioni*** (***bias***) nel processo cognitivo, basate sulla ***reputazione,*** la ***violazione delle aspettative,*** la ***consistenza,*** o assenza di contraddizioni***, l'intento persuasivo,*** e ***l'auto-conferma*****.**

La psicologia cognitiva ci dice anche che l'accesso al Web e alle reti sociali tende a polarizzare le opinioni politiche; inoltre la progressiva sostituzione dei giornali con il Web e le reti sociali come fonti di dati, e riduce la ***qualità delle analisi*** e degli ***approfondimenti*** sui ***fatti***; ciò anche perché produrre notizie verificate ha un costo, mentre inviare un messaggio su Twitter non costa nulla e raggiunge in alcuni casi molte più persone.

Il Web e le reti sociali sono luoghi virtuali dove noi scambiamo e viviamo ***emozioni***, è la ***rabbia*** quella che tende di più a diffondersi, rispetto a ***gioia***, ***disgusto*** e ***tristezza***, e altre ancora. Tutto questo non è scontato, per cui se li sappiamo usare, i dati hanno anche un grande ***valore emozionale*** positivo.

Capitolo 14

L’etica dei dati digitali

Abbiamo visto nei capitoli precedenti che i dati digitali sono un artefatto che è al tempo stesso tecnologia, servizio, 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 l’influenza che su di esse esercitano i dati, e gli algoritmi che ne fanno uso.

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 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).

Considerando la letteratura sull’etica dei dati, si vede in maniera chiara che le riflessioni tendono a concentrarsi non su temi generali (che relazione c’è tra etica e dati digitali? Quale è la specificità dei dati digitali riguardo a temi etici? ecc.) ma su tematiche e concetti specifici, che sono *elementi costituenti* l’etica dei dati, tematiche che chiameremo *determinanti* dell’etica dei dati digitali. Un esempio di questi determinanti lo abbiamo visto nel Capitolo 11 sul machine learning a proposito dello strumento Compas, quando abbiamo iniziato a discutere la *equità* delle tecniche di learning, tema che riprenderemo in questo capitolo.

Tuttavia, prima di ragionare sui determinanti, è necessario approfondire la questione delle diverse modalità con cui i dati digitali sono usati nelle società, nelle comunità, nella vita delle singole persone.

Nella più ampia generalità, abbiamo visto, quando abbiamo parlato di valore, che i dati possono essere considerati, analogamente ai beni e i servizi, come oggetto di scambi e relazioni tra persone, ovvero come oggetti di scambio tra persone e fornitori di beni, di servizi, di dati, fornitori che possiamo chiamare *provider* con un termine unico. Concentriamoci su questo aspetto rilevante, che tutti noi consapevolmente o meno viviamo quotidianamente; come avvengono questi scambi? Quali elementi sono coinvolti in questi scambi? La Figura 155 cerca di individuare tutti gli elementi che sono significativi in questa relazione.

Quando noi abbiamo necessità di comprare uno spazzolino da denti (bene acquistabile in un supermercato), ovvero abbiamo necessità di affittare una bicicletta vicino a casa nostra (servizio offerto dal comune o da un provider privato), ovvero vogliamo sapere chi ha vinto

ieri una partita di pallacanestro (dato acquisibile da un sito di giornale on line ovvero da un giornale in vendita in edicola), noi interagiamo con un *sistema organizzativo* di un provider.

Nella Figura 155 mostro come una qualunque *sistema organizzativo* di un provider è strutturato per fornire un bene, un servizio, un dato, a una persona che li richiede.

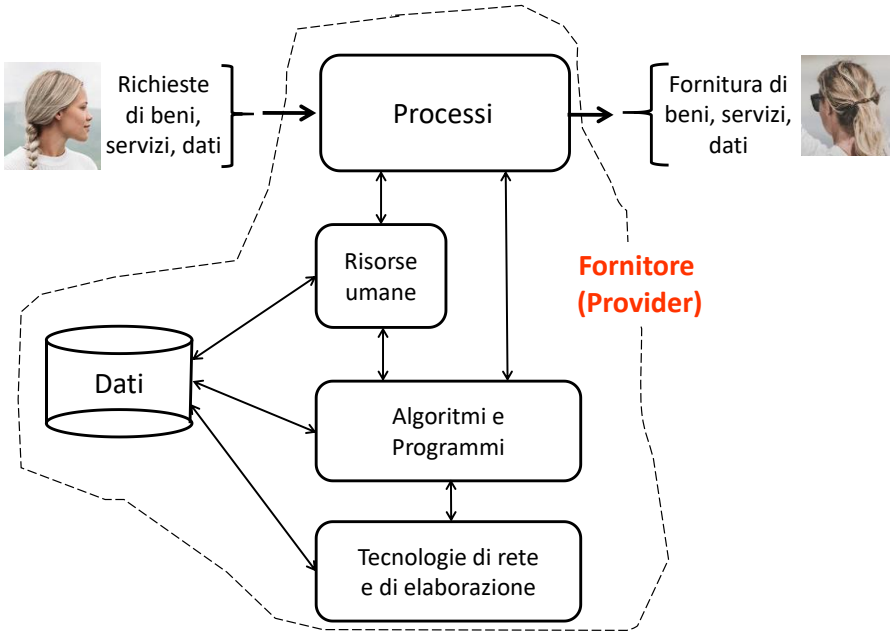


Figura 155 – I sistemi informativi che soddisfano le richieste di beni, servizi, dati

Anzitutto, nella organizzazione vengono eseguiti *processi*, cioè *insiemi di attività tra loro collegate*. L'edicolante per vendere i giornali cartacei deve scaricare i giornali dal furgone, metterli in ordine, porgere le copie agli acquirenti, dare il resto ecc.

I processi sono eseguiti da *risorse umane* o da *tecnologie*. Se compri un giornale all'edicola interagisci con l'edicolante, una persona, se leggi un giornale digitale, interagisci con un sito Web.Le tecnologie, in termini molto generali, possono essere di tre tipi: algoritmi e programmi, cioè metodi di calcolo astratti ovvero scritti in un linguaggio programmativo, reti per trasmettere dati, computer per elaborarli, e infine i *dati digitali*.

Ho solo un dubbio, ma se io chiedo una informazione a una persona che passa per la strada, questa persona è un sistema organizzativo?

Noto un po' di ironia…no quella persona non è un sistema organizzativo, è una persona, ma vedrai che le cose che diremo si applicano anche ai rapporti interpersonali. Per il resto è chiaro?

Si, chiaro. L'unica cosa che non capisco è cosa c'entri tutto questo con l'etica…

Ci arriviamo ora! Le questioni di cui ora parlerò, i determinanti dell'etica, gli aspetti della nostra vita che determinano l'avere un comportamento etico o non etico, nascono tutti dalle

relazioni che si instaurano tra una persona che ha necessità di un bene, di un servizio, di un insieme di dati, e un provider, un fornitore di tutte queste cose.

Nel seguito vediamo una lista di questi determinanti che, direttamente o indirettamente, fanno riferimento alle interazioni tra esseri umani e sistemi, che in risposta a richieste, forniscono loro beni, servizi, dati.

- *Trasparenza*, esprime la proprietà di un processo o di un provider che fa uso di dati digitali; gli utenti coinvolti nei risultati del processo possano comprenderne il funzionamento o le scelte.
    - Un esempio riguarda le domande per l'asilo nido per i figli; il processo decisionale è trasparente se sono pubblicati i criteri adottati e i punteggi parziali associati a ciascun criterio, l'elenco delle domande e la graduatoria finale.
    - Un secondo esempio riguarda i dati inviati dalle regioni all'istituto superiore di sanità per determinare i "colori" delle regioni nella epidemia Covid.
- *Responsabilità,* ovvero *capacità di rispondere dei propri atti*, *accountability* in inglese, esprime la esistenza e la messa a disposizione da parte di un processo o di un fornitore di strumenti conoscitivi per identificare chi e quando abbia preso una decisione o abbia effettuato una azione, e le ragioni di tale decisione o azione.
    - Un esempio sono i dati di tracciamento sui vaccini effettuati nel corso della epidemia Covid, così che sia possibile verificare e ricostruire eventualmente in forma anonima (cioè senza conoscere dati personali dei vaccinati) il rispetto delle priorità definite.
- *Equità o imparzialità* di un algoritmo di machine learning utilizato un processo di natura predittiva o decisionale: le decisioni prodotte dall'algoritmo devono essere indipendenti da aspetti sensibili, quali, ad esempio, i dati riferibili al genere, alla etnia, alle convinzioni religiose. E' anche la proprietà di un algoritmo di machine learning consistente nel comportarsi e prendere decisioni come parte terza, anche al di là delle propensioni, interpretazioni e sensibilità individuali.
    - Ne discuteremo tra poco.
- *Spiegabilità o Capacità di spiegare*, i dati dovrebbero permettere di chiarificare e rendere comprensibile il funzionamento interno di un'algoritmo.
    - Deve essere possibile per tutti capire come l'algoritmo produce i risultati. Ne parleremo nel seguito del capitolo.
- *Generalizzazione verso Personalizzazione*, equilibrio tra la messa a disposizione di dati per tutti, ovvero le esigenze che singole persone o comunità o aree territoriali hanno di personalizzazione e adattamento dei dati.
    - I dati sull'inquinamento raccolti dalle regioni possono essere restituiti in due forme: dati comuni a tutti i territori in modo tale da consentire un'analisi comparativa generale; dati differenziati a seconda delle diverse realtà territoriali; ad esempio a comuni fortemente urbanizzati possono essere distribuiti dati specifici sulle discariche di rifiuti urbani o industriali
- *Consapevolezza*, comprensione profonda di un fenomeno, che può riguardare un processo o una risorsa umana o una tecnologia, che abbia conseguenze etiche.
    - Tutto questo libro è pensato per dare strumenti che aumentino la consapevolezza. Un esempio riguarda l'atteggiamento critico con cui leggere una statistica pubblicata su un sito o sentita in televisione; vedi ad esempio la disussione sull'indicatore di Figura 8.
- *Inclusione*, nella raccolta e utilizzo dei dati digitali in una decisione o azione di un processo devono essere prese in considerazione le esigenze di tutti i soggetti interessati, soprattutto nel caso di decisioni che riguardano l'intera collettività in modo che a tutti gli

individui possano essere offerte le migliori condizioni possibili conseguenti alla decisione o azione.
    - Numerose norme italiane ed europee dispongono l'utilizzo delle consultazioni pubbliche prima dell'emanazione di norme e regolamenti che hanno impatto sulla vita dei cittadini. La normativa italiana tutela questa forma di partecipazione democratica[26], prevedendola nel Codice dell'amministrazione digitale, ma questa prassi dovrebbe essere adottata da tutti i soggetti, pubblici o privati.
    - Nel caso di decisioni che riguardano specifici settori economici dovrebbero essere interpellate tutte le parti coinvolte: imprese, sindacati dei lavoratori, consumatori dei prodotti/servizi di quel settore economico.
    - Nella emissione di un regolamento che ha impatto sulla qualità della vita di una popolazione, è necessario l'avvio di una consultazione pubblica di cui si tiene conto nella versione definitiva.
- *Parità di opportunità (Non discriminazione, egualitarismo*), I dati digitali resi disponibili in un sistema dovrebbero permettere a tutti di competere su base egualitaria.
    - In un concorso pubblico possono essere diffuse le domande poste nei concorsi precedenti, così da fornire a tutti i partecipanti gli stessi elementi informativi.
    - I siti pubblici debbono essere consultabili anche da parte di persone diversamente abili, così come prevede una specifica legge dello Stato italiano[27].
- *Qualità dei dati*, proprietà dei dati utilizzati da esseri umani, processi o algoritmi di learning di essere corretti, completi, aggiornati, essendo in tal modo aderenti alla realtà e non distorcendo i risultati.
    - A questo tema è dedicato il Capitolo 7.
- *Privacy*, la condizione per cui i dati personali devono essere protetti dall'accesso pubblico
    - Ad esempio, una organizzazione a cui comunichiamo il nostro codice fiscale o il nostro indirizzo di residenza lo deve tenere riservato.
    - Una applicazione che usa la nostra posizione a fini di georeferenziazione ci deve chiedere l'autorizzazione prima di usarla.
    - Ogni utente deve essere in grado di conoscere quali suoi dati personali sono mantenuti da ogni singola organizzazione, poterli consultare ed eventualmente chiederne la modifica o la cancellazione.
- *Condivisione*, la possibilità di mettere in comune i dati digitali e non considerarli come un bene privato.
    - Per una regione nel corso della epidemia Covid, diffondere a tutti i comuni nella regione i dati sullo stato clinico dei malati Covid, così da permettere la individuazione e correzione dei dati anomali.
- *Affidabilità*, gli algoritmi devono essere in grado di funzionare in modo corretto.
    - Algoritmi e programmi devono eseguire le funzioni e rispettare gli scopi per cui sono stati progettati. Al fine di massimizzarne l'affidabilità, gli algoritmi devono essere sottoposti a numerose prove di funzionamento, utilizzando anche dati di test presi da casi reali. Ad esempio, nel caso di un algoritmo che esamina fatture cartacee e dopo la loro scansione estrae, mediante tecniche OCR (Optical Character Recognition), i dati economici, sarà

[26] D.lgs. 82/2005, Art. 9 (Partecipazione democratica elettronica). [Le amministrazioni e le società pubbliche] favoriscono ogni forma di uso delle nuove tecnologie per promuovere una maggiore partecipazione dei cittadini, anche residenti all'estero, al processo democratico e per facilitare l'esercizio dei diritti politici e civili e migliorare la qualità dei propri atti, anche attraverso l'utilizzo, ove previsto e nell'ambito delle risorse disponibili a legislazione vigente, di forme di consultazione preventiva per via telematica sugli schemi di atto da adottare)).

[27] Legge 9 gennaio 2004, n. 4 "Disposizioni per favorire e semplificare l'accesso degli utenti e, in particolare, delle persone con disabilità agli strumenti informatici"

bene verificare la correttezza dei dati rilevati sottoponendo all'algoritmo numerosi insiemi di fatture reali di formato diverso, e confrontando i risultati con quanto rilevato visivamente o registrato in precedenza da operatori umani in un sistema contabile informatizzato.

E' impossibile in una breve sezione entrare nel dettaglio e fornire esempi di tutti i precedenti temi, che verranno esaminati nel capitolo della Enciclopedia dedicato all'etica. Nel seguito ci concentriamo sulla equità di un algoritmo di learning, la spiegabilità e la privacy.

Equità (Fairness)

La Figura 156 mostra un insieme di classificazioni prodotte da Google Photos, in cui due persone africane vengono classificati come gorilla. Sia nel caso di deliberata scelta nella classificazione, sia nel caso che derivi da errori progettuali o nei dati, certamente possiamo dire che il modello di classificazione non è *equo*, secondo la definizione vista in precedenza, nei confronti di persone provenienti dall'Africa.

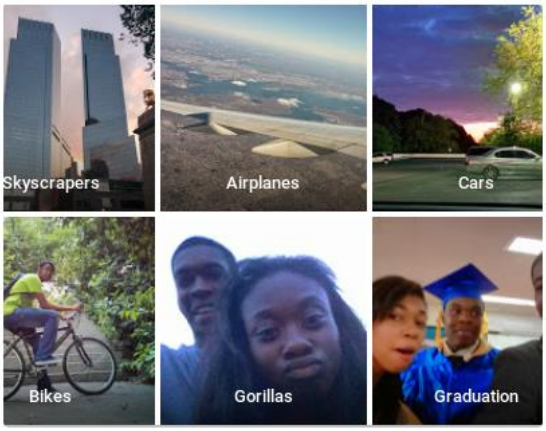


Figura 156 - Diverse classificazioni prodotte la Google Photos

Per fare un altro esempio, questa volta nell'ambito degli algoritmi di traduzione automatica da una lingua naturale a un'altra, provate a tradurre dall'inglese all'italiano usando Google Translator (https://translate.google.it/) le frasi:

1. she works, in an hospital, my friend is a doctor
2. she works in the hospital, my friend is a nurse

In cui *doctor* e *nurse* sono parole inglesi senza coniugazione di genere, e significano in italiano rispettivamente dottore/dottoressa e infermiere/infermiera.

In una prova fatta a gennaio 2021, gli esiti delle traduzioni in italiano sono quelli riportati in Figura 157. Come si vede, il fatto che le due frasi inizino con *she* in entrambi le frasi non ha influenza sulla traduzione di *friend*, tradotto *amico*, mentre la presenza della parola *nurse* nella seconda frase porta alla traduzione del soggetto della prima frase con *lei*, al femminile, e produce la versione femminile italiana del termine corrispondente a *nurse*, cioè *infermiera*; la traduzione risulta in una chiara discriminazione di genere.

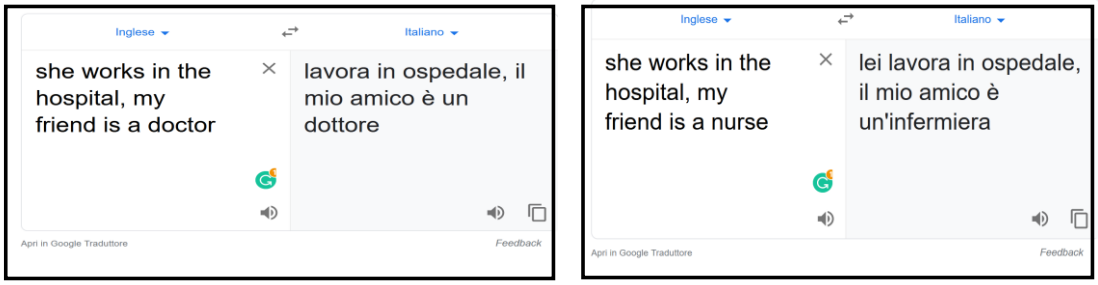


Figura 157 - Discriminazione nel natural language processing

L' equità è stata investigata nella filosofia dell'etica, nei sistemi giuridici e nelle scienze sociali, in virtù della sua rilevanza in tutti i *processi decisionali* che coinvolgono *esseri umani*. Una norma giuridica che si occupa di equità può riguardare un certo *ambito* della vita sociale, e può fare riferimento a una determinata *categoria da proteggere*. Vediamo ora in successione gli ambiti della vita sociale e le categorie protette in una delle legislazioni più avanzate esistenti nelle moderne democrazie, la legislazione statunitense.

Gli *ambiti* possono essere disciplinati da specifiche leggi, ovvero anche lasciati alla libera interpretazione di coloro che ne fanno uso. A titolo di esempio, la *forthy fifth rule* valida negli Stati Uniti afferma che se in un concorso viene violata la parità statitistica, cioè la attribuzione dei posti ai vincitori in proporzione ai gruppi sociali protetti, e viene violata per più del 20 per cento dei posti, occorre giustificare questa violazione della equità sulla base delle comprovate esigenze del soggetto che ha promosso il concorso. In Figura 158 compaiono gli ambiti disciplinati da norme nella legislazione americana e per alcuni di essi la norma d iriferimento.

- Prestiti (Equal Credit Opportunity Act, 1974)
- Formazione (Civil Rights Act of 1964; Education Amendments of 1972)
- Impiego (Civil Rights Act of 1964)
- Alloggi pubblici (Civil Rights Act of 1968; Fair Housing Act)
- Sentenze penali
- Concessione della libertà provvisoria agli imputati/detenuti
- Concorsi
- Assunzioni

Figura 158 - Ambiti di applicazione della equità nella legislazione americana

Riguardo alle *categorie da proteggere*, la legge sui diritti civili degli Stati Uniti del 1964 mise al bando le discriminazioni sulla base della etnia, il colore della pelle, la religione, il sesso o origine nazionale delle persone.

La legge conteneva due importanti disposizioni che esprimevano la comprensione della comunità dei cittadini su cosa significava essere non equo: il titolo VI, che impediva alle agenzie governative (comprese le università) di ricevere fondi federali che discriminavano in base alla etnia, 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 etnia, colore della pelle,

religione, sesso o origine nazionale. In Figura 159 vediamo le categorie protette e le relative leggi che le proteggono negli Stati Uniti.

- Etnia (Civil Rights Act of 1964);
- Colore della pelle (Civil Rights Act of 1964);
- Genere (Equal Pay Act of 1963; Civil Rights Act of 1964);
- Religione (Civil Rights Act of 1964);
- Origine nazionale (Civil Rights Act of 1964);
- Cittadinanza (Immigration Reform and Control Act);
- Età (Age Discrimination in Employment Act of 1967);
- Essere in Gravidanza (Pregnancy Discrimination Act);
- Stato familiare (Civil Rights Act of 1968);
- Disabilità (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990);
- Veterani (VieVNam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act);
- Dati genetici (Genetic Information Nondiscrimination Act)

Figura 159 – Categorie protette negli Stati Uniti

Tornando alle tecniche predittive, interroghiamoci ora su una questione veramente fondamentale: possiamo immaginare che venga sviluppato un algoritmo che, dato una algoritmo predittivo ci dica per ogni possibile dataset di ingresso se l'algoritmo sia o meno equo? Se un tale algoritmo esistesse, avremmo risolto la sfida morale insita nella equità: non ci dobbiamo più preoccupare di indagare sul comportamento di una commissione di concorso, sul comportamento di una banca che concede ad alcuni un prestito e lo nega ad altri, perché l'algoritmo stabilirebbe la equità o meno del comportamento del decisore.

Ebbene, un tale algoritmo *non esiste*. Torniamo infatti al problema della concessione della libertà provvisoria, e all'algortimo Compas che permette di misurare il rischio di recidiva si un detenuto, sulla base di quanto è accaduto nel passato. Nella Figura 128 abbiamo potuto constatare che esistono almeno due definizioni di equità che danno risultati contrastanti tra di loro riguardo alla accuratezza dell'algortimo predittivo.

La prima definizione, quella adottata dal sito ProPublica, sostanzialmente assume il punto di vista del detenuto, che, se Afro Americano, ha una maggiore probabilità che l'algoritmo si sbagli nel negare la libertà provvisoria, generando così falsi positivi e falsi negativi in maggiore percentuale tra gli AfroAmericani.

La seconda definizione di equità adottata da Northpointe si concentra sulla capacità predittiva, mostrando che l'algoritmo è stato equo tra Bianchi e AfroAmericani nel predire i veri positivi e veri negativi.

Dunque, due punti di vista alternativi, così come sono alternativi i due punti di vista che in un concorso con dieci vincitori, a cui partecipano 90 uomini e dieci donne, ritengono equo un algoritmo decisionale che rispettivamente fa vincere nove uomini e una donna, ovvero cinque uomini e cinque donne.

Nella letteratura del secondo decennio del ventunesimo secolo (2010 – 2020) sono state proposte tante definizioni di equità (tra cui le due prececenti), che adottano punti di vista molto differenti. Insomma, un concetto apparentemente semplice da definire, la equità, per cui sentiamo affermare "bisogna essere equi! Io sono sempre equo! Non guardo in faccia a nessuno!" è in realtà molto complicato da investigare, ed è un concetto con molteplici interpretazioni. E la soluzione non è neanche individuabile nel dialogo paradossale che ci propone Judea Pearl nella Figura 160!

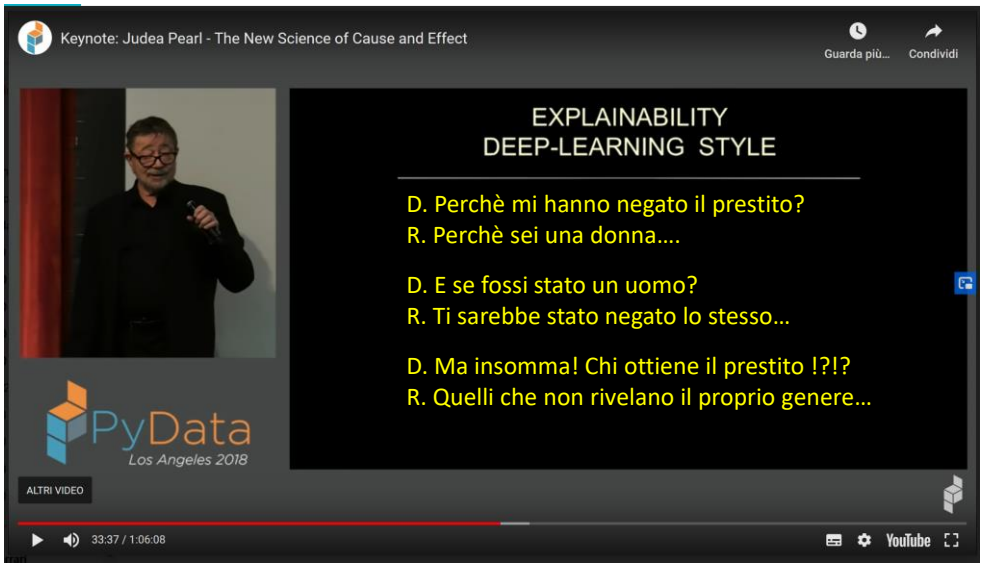


Figura 160 – Una visione scherzosa della equità da parte di Judea Pearl

Dalla precedente discussione emergono tre importanti considerazioni.

1. Gli algoritmi di machine learning di tipo supervisionato, che cioè apprendono da casi noti, per loro natura prevedono il futuro *partendo dal passato*; si basano su casi passati, e si applicano nella previsione dei casi futuri. Questa "filosofia" è a mio parere psicologicamente conservatrice, *non innova*, stabilisce una continuità tra passato e futuro, *se è stato così nel passato, sarà così anche in futuro*.
2. Quando si utilizza un algoritmo predittivo, occorre sempre determinare e rendere trasparente il suo livello di accuratezza ed equità. Per la equità, se essa è stabilità per legge, l'agortimo deve rispettare la specifica stabilita dalla legge.
3. *Peraltro, esistono molti punti di vista e definizioni di cosa significa per un algoritmo predittivo fare scelte eque*, e questo è sicuramente un *bene*: pensate quanto pericoloso sarebbe un mondo in cui per ogni fenomeno, per ogni scelta, decide per noi un algoritmo predittivo *unico*, depositario della verità assoluta. Al contrario, sta sempre a noi decidere alla fine, *siamo noi che ci dobbiamo assumere la responsabilità delle scelte*, non le macchine. E in un mondo in cui sempre più spesso gli umani si fanno aiutare e addirittura sostituire dalle macchine, gli umani devono sapere quali definizioni di equità siano garantite dagli algoritmi, *e fare una scelta su quale sia quella ritenuta migliore, e motivarla.*

Un'ultimo punto sulla equità, anche esso di capitale importanza. Torniamo alla discussione sugli algoritmi predittivi del Capitolo 11. Supponiamo che i dati sul numero di arresti (*#A* nella

Figura 125) abbiano molti errori di inaccuratezza, perchè nella ricostruzione degli arresti sono stati considerati dati a loro volta imprecisi o incompleti. Intuitivamente, l'algoritmo predittivo fornirà predizioni sbagliate perché influenzate da questi errori nei dati di ingresso. Quindi, è importante discutere tutte le possibili cause di questi errori.

Cause che influiscono sulla equità

Molteplici possono essere le cause della scarsa equità di un algoritmo predittivo/decisionale. Il ciclo di vita del machine learning riprodotto in Figura 125 è composto di un insieme di fasi; in queste fasi vi sono tre diversi *attori* che concorrono nella decisione finale, e nei possibili errori che si possono commettere:

1. l'algoritmo;
2. i dati utilizzati;
3. gli esseri umani che sono coinvolti in alcune attività specifiche, vedi tra poco.

Nella Figura 161 sono mostrati diversi aspetti che possono mettere a repentaglio la qualità del modello predittivo, associabili a uno o all'altro dei tre attori. Esaminiamo questi diversi aspetti critici, ognuno dei quali in figura è preceduto da una A, una D o una U a seconda che ne siano responsabili l'Algoritmo, i Dati o gli Umani.

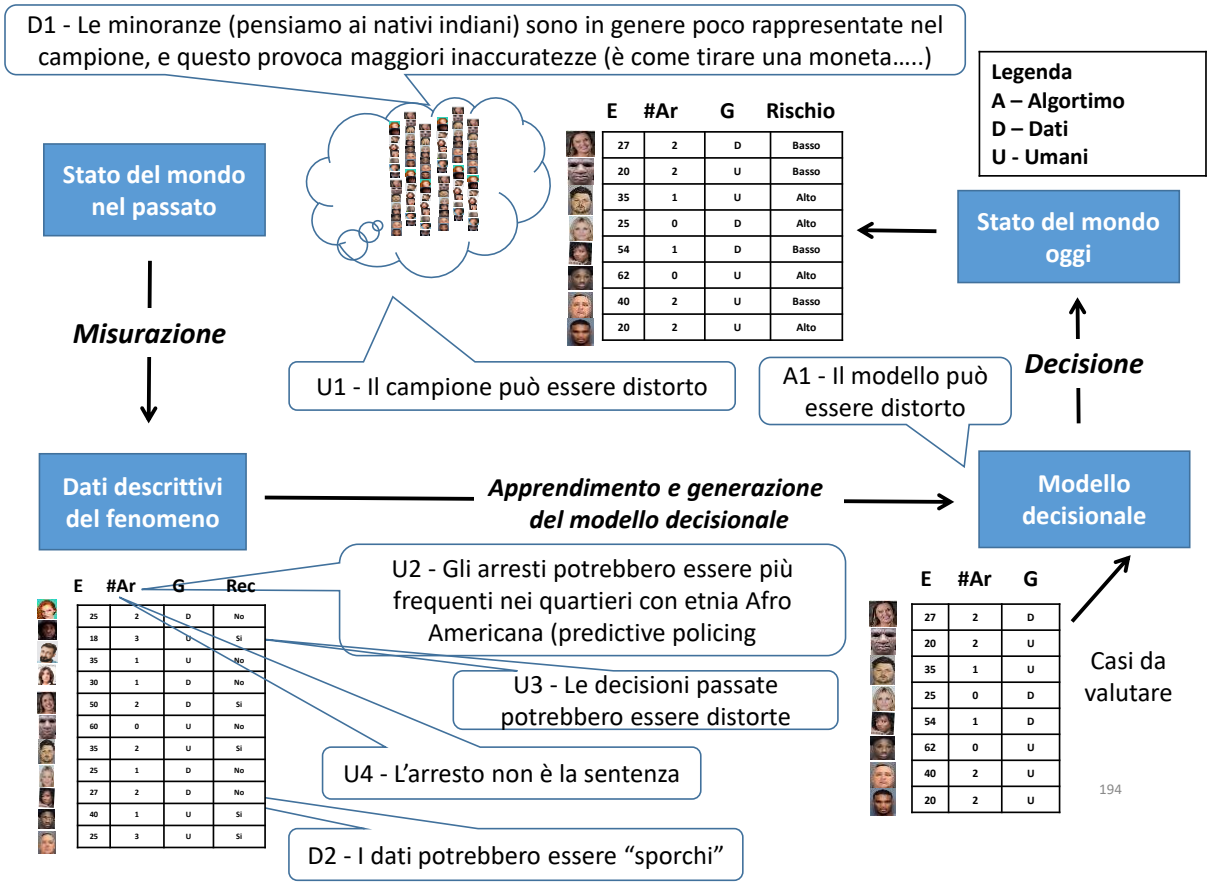


Figura 161 – Aspetti che influiscono sulla equità

A1 - Il modello può essere distorto, cioè può non rispettare il tipo di equità che è stata assunta come riferimento. Abbiamo discusso questo aspetto nelle pagine precedenti.
D1 – Nel campione su cui ci siamo basati per addestrare l'algoritmo, le minoranze (pensiamo negli Stati Uniti ai nativi indiani) sono in genere poco rappresentate, e quindi nell'addestramento sono pochi i casi che vengono presi in considerazione. Come abbiamo già visto nel caso dell'algoritmo predittivo di Etzioni, basarsi su pochi casi porta intuitivamente a ridurre la accuratezza dell'algoritmo. Diciamo che è un pò come tirare una moneta…
D2 - I dati potrebbero essere inaccurati, ad esempio, potremmo avere diversi casi in cui il numero di arresti è stato calcolato per difetto, o per eccesso.
U1 - Il campione può essere distorto, cioè, ad esempio, potrebbero essere stati presi in considerazione in proporzione più afro americani di bianchi.
U2 – Per una sorta di distorsione nel controllo del territorio, potrebbe accadere che le pattuglie abbiano mandato di andare solo o prevalentemente in alcuni quartieri abitati dai ceti meno abbienti, per cui gli arresti potrebbero essere più frequenti nei quartieri con etnia Afro Americana.
U3 - Le decisioni passate sulla concessione della libertà provvisoria potrebbero essere state influenzate da una distorsione cognitiva o un atteggiamento prevenuto verso gli Afro Americani; l'algoritmo eredita questa distorsione.
U4 - L'arresto non può essere confuso con la condanna passata in giudicato; il vero elemento per valutare il rischio di recidiva è la condanna o assoluzione, e nel caso di condanna la gravità del reato ascritto, non l'arresto; nel procedimento penale vale infatti la presunzione di innocenza. L'obiezione è che la condanna arriva dopo molto tempo, e non può essere assunta come dato acquisibile per il processo di apprendimento. Il vero fenomeno rilevante da stimare è la pericolosità sociale, che è associata solo ai procedimenti penali terminati con la sentenza definitiva.

Spiegabilità

"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 una certa multa o notifica, ad esempio quella volta che da una nota azienda fornitrice di elettricità mi arrivò una bolletta immotivata di circa 850 euro per due mesi di utenza.

In questa sezione indaghiamo il tema della *spiegabilità* nelle tecniche di machine learning, intesa come esistenza di una spiegazione dell'algoritmo espresso dalla tecnica decisionale e dei risultati della previsione. Ma, prima, cerchiamo di capire cosa si intende per spiegabilità di un programma software.

Spiegabilità di un programma software

Torniamo a considerare i due programmi per la somma dei primi 10 numeri interi, che riproduco nella Figura seguente.

```
SOMMA = 10 /2  * (10  +  1)
STAMPA SOMMA
TERMINA
```

Programma 1

```
SOMMA = 0
PER I CHE VA DA 1 A 10 ESEGUI
SOMMA = SOMMA + I
STAMPA SOMMA
TERMINA
```

Programma 2

Figura 162 – I due programmi per la somma dei primi 10 numeri interi mostrati nel Capitolo 2

Che sensazione provi nel guardare questi programmi?

Ho una doppia sensazione contrastante. Da una parte mi dico: quando mi hai fatto vedere il programma 1, dovrei averlo capito da solo che il programma calcolava la somma dei primi 10 numeri, non so come dire: dovrei avere la cultura per capire da solo quali calcoli esegue il programma, e quali regole matematiche ci siano "dietro". Dall'altra, capisco che non posso sapere tutto e quindi mi servirebbe un aiuto da parte dello stesso programma, ma non so come si faccia?

Certamente, hai ragione, non puoi sapere tutto. Per rispondere alla tua ultima domanda, sarebbe utile che all'inizio dei due programmi comparisse una frase, detta *commento*, così concepita:

COMMENTO – IL PROGRAMMA CALCOLA LA SOMMA DEI PRIMI DIECI NUMERI INTERI

Il precedente commento chiarisce "cosa" fa il programma, non *come* lo fa. Certe volte, può essere utile conoscere oltre al *cosa* anche il *come*, ad esempio per poter valutare la efficienza dell'algoritmo, misurata dal numero di istruzioni eseguite. Possiamo perciò aggiungere:

al Programma 1 il commento
LA SOMMA E' CALCOLATA IN BASE AD UNA SEMPLICE FORMULA MATEMATICA CHE LEGA IL NUMERO 10 ALLA SOMMA DEI PRIMI 10 NUMERI INTERI

al Programma 2 il commento
LA SOMMA E' CALCOLATA SOMMANDO AL VALORE 0 SUCCESSIVAMENTE I VALORI 1, 2, …, FINO A 10.

E' chiaro che il secondo programma esegue molte 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 100, 1.000, 10.000, ecc. numeri è sempre lo stesso, mentre invece nel secondo programma il numero di istruzioni cresce linearmente con il numero dei valori da sommare; per sommare 1.000 numeri devo eseguire 2.000 istruzioni, per sommarne 10.000 numeri ne devo eseguire 20.000, ecc. Possiamo dire che per sommare N numeri nel primo caso basta una sola istruzione, nel secondo ne servono 2 x N.

A proposito di N numeri. Puoi provare a modificare il Programma 2 nel senso di far leggere al programma il valore N con una istruzione LEGGI N, facendogli fare la somma non dei primi 10 numeri ma dei primi N numeri?

Ci provo, nella prossima pagina.

E' questo?

```
LEGGI N
SOMMA = 0
PER I CHE VA DA 1 A N ESEGUI
SOMMA = SOMMA + i
STAMPA SOMMA
TERMINA
```

Certo, bravissimo!

Io ti ho seguito passo passo, e credo di aver capito tutto. Ma mi dici perchè hai fatto tutta questa trattazione sui programmi software? Cosa c'entrano i programmi software con il machine learning?

Ottima domanda! C'entrano. Infatti, i programmi software sono scritti da esseri umani, vedi Figura 163, e quindi per capirne il significato possiamo chiedere a chi ha scritto il programma, sperando di sapere chi è, e sperando che se lo ricordi. Insomma, c'è una responsabilità umana diretta.

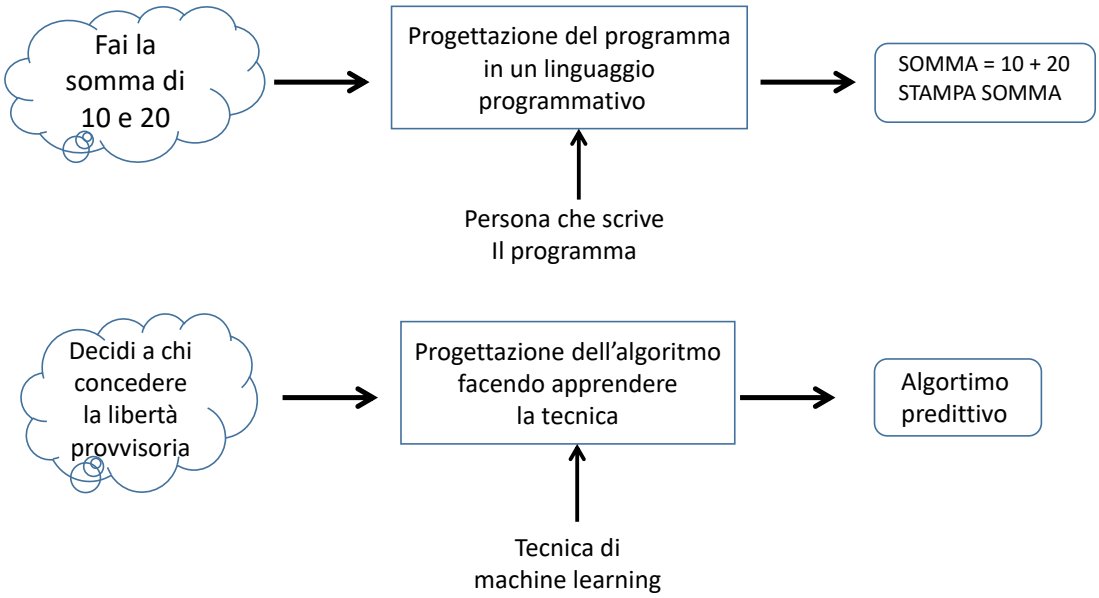


Figura 163 – La differente progettazione di un programma in un linguaggio programmativo e un algoritmo predittivo basato su una tecnica di machine learning

Le tecniche di machine learning generano *da sole* l'algoritmo predittivo, vedi ancora Figura 163, per cui noi umani siamo estromessi dal processo creativo; bada, come abbiamo visto nella Figura 161, gli esseri umani hanno sempre grandi responsabilità nella scelta delle caratteristiche utilizzate per l'apprendimento, nella scelta del campione, e nelle decisioni compiute nel passato, a cui si ispira il processo di apprendimento. Ma *l'algortimo*, l'algortimo è generato da una tecnica di learning.

Il problema di rendere gli algoritmi comprensibili agli umani, già presente nella informatica da quando esiste il software, diventa quindi ancor più critico nell'epoca del machine learning, perchè ora l'algoritmo che prima veniva concepito e prodotto da esseri umani, è autonomamente prodotto da una tecnica di apprendimento.

C'è quindi un rischio in più, che l'algoritmo predittivo sia considerato *obiettivo* per definizione, che il machine learning venga considerata una scienza esoterica in mano a scienziati inaccessibili. Il grande rischio che corriamo è quello che viene chiamata la *black box society* (vedi Figura 164), una società in cui le decisioni sono prese basandosi su analisi e algoritmi i cui meccanismi di funzionamento sono noti solo a chi li ha concepiti, soggetti che per ragioni di business, di concorrenza o sicurezza, non intendono o non vogliono condividerli.

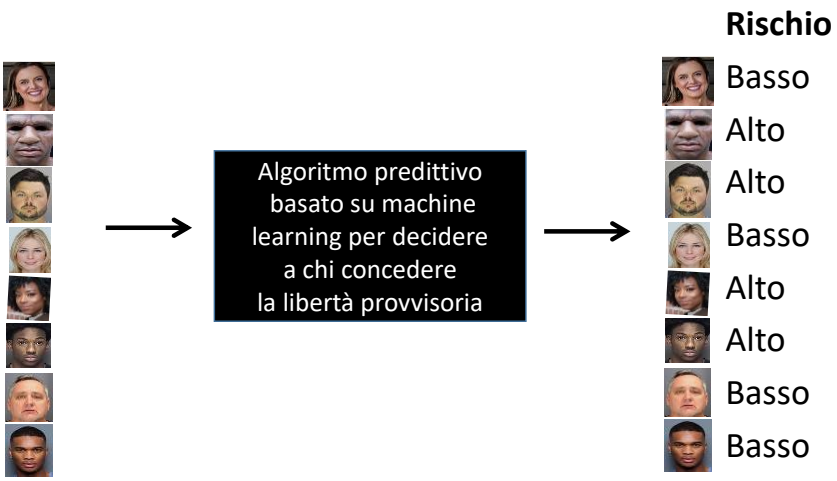


Figura 164 – La black box society

Abbiamo già visto poco fa nella discussione sulla equità come possiamo tentare di difenderci. Ma c'è un problema in più: oltre a voler verificare quale tipo di equità è rispettata, noi dobbiamo anche sapere *come* l'algoritmo decide. La equità garantisce le categorie protette, ma *non sappiamo nulla sulla equità con riferimento alle singole persone*, perché, ad esempio sia stata data la libertà provvisoria a Rossi e non Verdi.

La *spiegabilità* si occupa di rendere esplicito il processo decisionale/predittivo adottato dall'algoritmo di learning. Supponiamo, per esempio, che un modello decisionale produca una graduatoria per ottenere l'asilo nido per i figli. Se una persona inserisce i suoi dati e riceve come risultato il punteggio ottenuto, 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.

Nel Capitolo 10 abbiamo visto in Figura 126 un esempio di albero di decisione, una delle tante tecniche che sono usate nel machine learning supervisionato per produrre un modello predittivo; l'esempio riguardava la decisione su come uscire vestiti in una giornata dal tempo incerto.

Gli alberi di decisione sono considerati una tecnica di learning facilmente *comprensibile*: ad ogni nodo decisionale dobbiamo porci una domanda, basata sui dati a disposizione, a seconda della risposta percorriamo il ramo di destra o il ramo di sinistra, fino ad arrivare a una foglia dell'albero, a cui è associata una scelta: a quel punto scegliamo la soluzione cui siamo giunti nel percorrere l'albero. Le domande sono chiare, i rami delle decisioni altrettanto.

Ebbene, nel machine learning è possibile usare molte tecniche diverse, accenniamo a una tra le tante, i sistemi a regole. I *sistemi a regole* esprimono il procedimento decisionale per mezzo di formule logiche come la seguente:

se la febbre è superiore a 38 *e* la gola è rossa *e* il paziente starnutisce frequentemente *allora* il paziente ha una influenza

Quando applichiamo un sistema a regole, dobbiamo applicare l'antecedente della regola (*se la febbre è superiore a 38 e la gola è rossa e il paziente starnutisce*) ai dati in ingresso al sistema, e se essi rispettano la formula logica, allora associamo a tali dati il valore che compare a destra della regola (nel nostro caso, "ha una influenza"). Anche i sistemi a regole esprimono un processo decisionale che, essendo simile a uno di quelli che noi adottiamo usualmente, risulta comprensibile.

Esistono altre tecniche oltre agli alberi di decisione e i sistemi a regole, che possono essere addestrate per creare un modello predittivo. Gli alberi di decisione sono comprensibili, altre tecniche che qui non introduco, molto meno. Si può vedere che per determinati problemi esiste una tecnica ottimale, talvolta gli alberi di decisione, altre volte un'altra tecnica meno comprensibile.

Il problema della spiegabilità si tramuta allora nel seguente. Data una tecnica utilizzata per generare un algoritmo predittivo, se la tecnica non è comprensibile, *trovare una spiegazione* per la tecnica *consiste nel trovare una nuova tecnica* tra quelle considerate spiegabili (alberi di decisione o sistemi a regole), che imita il comportamento della tecnica non comprensibile e che porta a risultati simili. Vedi Figura 165.

Il problema ancora non risolto (2021) in modo maturo riguardante la spiegabilità, come introdotta poco fa, è che il nuovo algoritmo deve avere un livello di accuratezza (falsi positivi e falsi negativi o loro combinazioni) con qualità simile a quella dell'algoritmo non spiegabile. Insomma, non basta spiegare, la nuova tecnica *non deve divergere* rispetto alla tecnica originaria, costruendo un algoritmo con risultati diversi o meno accurati rispetto ai risultati forniti dalla prima.

La seconda questione non risolta nella ricerca (2021) riguarda il fatto che si riscontra una relazione inversa tra spiegabilità e accuratezza; quanto più una tecnica è spiegabile, tanto più tende a essere meno accurata. Quindi in certo senso, la spiegazione è in parte ingannevole.

Nel capitolo della Enciclopedia sull'etica entreremo nel dettaglio su tutti questi aspetti.

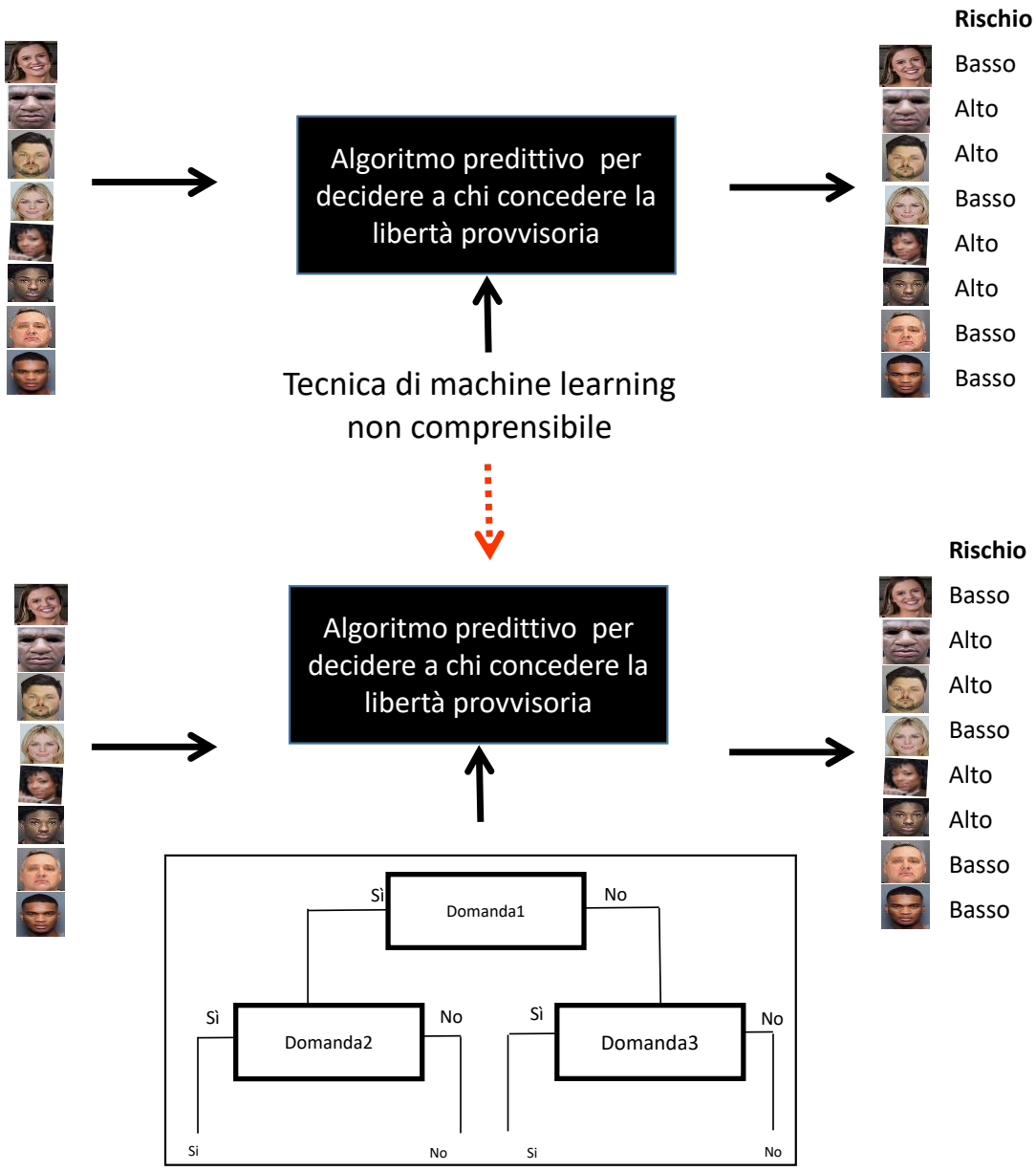


Figura 165 – Come affrontare la spiegabilità

La privacy e Il Regolamento generale sulla protezione dei dati (GDPR)

Riguardo alla privacy, accenniamo brevemente al Regolamento generale sulla protezione dei dati (General Data Protection Regulation o GDPR), la legge sulla privacy nella Unione Europea (UE) entrata in vigore il 25 maggio 2018. I principi su cui si basa il GDPR fanno riferimento 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 del GDPR riguardano:

- Il principio del *consenso informato*, vengono stabilite regole rigide per ottenere il consenso come base legale per l'elaborazione dei dati.
- La *portabilità* dei dati, che corrisponde al diritto di spostare i dati personali da un fornitore di servizi a un altro senza oneri.

- La *opacità* dei dati personali, cioè il diritto di cancellare le informazioni su *quali* dati personali 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.
- 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 algoritmi basati su tecniche predittive automatiche.
- Una classificazione più ampia, rispetto alla precedente legislazione, dei dati personali e sensibili, che include gli identificatori on line, 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 un alto rischio per i diritti degli individui.
- La *sicurezza* dei dati, che consiste nel conservare i dati in modo tale da garantire la integrità e non modificabilità, attraverso appropriate misure tecniche e organizzative.
- Il *trasferimento* dei dati personali fuori dalla Unione Europea può avvenire solo se sono garantite adeguate salvaguardie. Questa misura è particolarmente innovativa perché riconosce ai dati la dimensione globale che ad essi è data dai grandi players planetari come Google e Amazon, e individua norme di contrasto al loro monopolio sui dati personali.
- L'obbligatorietà di istituire un *data protection officer*, una figura organizzativa che si occupa di protezione dei dati, se l'organizzazione:
  1. è pubblica;
  2. ha relazioni con utenti su vasta scala;
  3. elabora dati sensibili, cioè dati relativi alla religione, 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.

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 in Europa
- *Organizzazioni esterne alla UE* che vendono beni e servizi a cittadini residenti in Europa.
- *Service provider* che elaborano dati personali come servizio ad altra organizzazione.

Le sanzioni prevedono multe pesanti e rimborsi per i danni subiti.

E infine il GDPR afferma il fondamentale problema del diritto alla spiegabilità degli algoritmi di machine learning discusso poco fa, stabilendo che coloro che sono direttamente oggetto delle tecniche *predittive* (viste fino ad ora) e *interpretative* (ad es. l'algoritmo ha individuato una persona come responsabile di un reato) hanno diritto a una spiegazione su come l'algoritmo funziona. Tra i giuristi vi sono opinioni contrastanti su questo punto; alcuni hanno parlato in modo contrario anche solo sulla possibilità che un tale diritto esista, altri sostengono che tale diritto scaturisce direttamente dal diritto naturale.

Riassumendo

L' ***etica dei dati*** viene trattata facendo riferimento a un insieme di ***principi o determinanti***, così come la ***qualità dei dati*** è tradotta in termini di ***caratteristiche***; essi riguardano: la ***trasparenza, responsabilità,*** ovvero ***capacità di rispondere dei propri atti (accountability) equità*** o ***imparzialità di un algoritmo di machine learning, spiegabilità*** o ***capacità di spiegare, generalizzazione verso personalizzazione, consapevolezza, inclusione, parità di opportunità (non discriminazione, egualitarismo), qualità dei dati, privacy, condivisione, affidabilità.***

L' ***equità*** di ***un algoritmo di machine learning*** è un determinante che ha varie possibili definizioni e ***significati***. Tende sempre a essere definita come ***non discriminazione di*** specifiche ***comunità*** di persone a cui si applica l'algoritmo predittivo decisionale. Queste comunità possono essere associate a un ***genere***, tipicamente le donne, a una ***etnia***, ad es. negli USA gli Afroamericani, ecc. L'equità può essere o meno ***disciplinata da una legge***. L'equità (o la non equità) può derivare dall' ***algoritmo predittivo***, dalle ***persone coinvolte***, ad esempio, ***nella scelta dei dati usati nell'apprendimento***, e dai ***dati stessi***.

La ***spiegabilità*** è la caratteristica per cui la tecnica usata nell'algoritmo di apprendimento, ad esempio gli ***alberi decisionali***, è ***comprensibile*** a ***chi applica l'algoritmo*** per le proprie decisioni (ad es. un giudice) e a ***chi subisce la decisione*** (ad es. un detenuto). Se un algoritmo adotta una tecnica non spiegabile, è possibile cercare per l'algoritmo una nuova tecnica che accresce la spiegabilità, verificando però che questa non porti a un peggioramento della accuratezza dell'algoritmo.

Riguardo alla ***privacy***, il regolamento per la protezione dei dati personali promulgato dalla Unione Europea sancisce diversi principio tra cui: il ***consenso informato***, la ***portabilità*** dei dati, la ***opacità*** dei dati personali, la ***qualità*** dei dati, il **diritto alla *cancellazione***, il ***diritto alla spiegazione***, una ***classificazione*** più ampia dei ***dati personali e sensibili***, la ***responsabilità*** delle organizzazioni che usano i dati, l' ***analisi di impatto*** della protezione dei dati, il ***trasferimento*** dei dati personali fuori dalla Unione Europea, l'istituzione nelle organizzazioni di un ***data protection officer***, ***l'ambito territoriale*** di applicazione.

Epilogo - Sintesi

E dunque, cosa sono i dati digitali? Dimmelo, in sintesi…

Il secondo novecento è stata l'epoca della ***alfabetizzazione linguistica***. Oggi, in virtù della grande importanza che i dati stanno assumendo nella nostra vita, nasce la esigenza di una **alfabetizzazione in tema di dati digitali**, fornendo a tutti gli strumenti per acquisire una consapevolezza critica sul loro uso. I dati vengono prodotti e mostrati a noi attraverso un itinerario spesso complesso, simile a un **iceberg** di cui noi vediamo la parte emergente.

Il grande uso che facciamo dei dati digitali deriva anzitutto dalla recente comparsa delle ***tecnologie*** del ***telefono mobile***, le **reti sociali**, ***l'internet delle cose***, il ***cloud***, che producono e fanno uso di una quantità crescente di ***grandi quantità di dati*** detti anche ***big data***.
Le altre tecnologie fondamentali legate ai dati digitali sono la ***rete Internet***, che permette di trasmettere dati in tutto il mondo a costi molto bassi, e il ***computer***, che permette di elaborare dati digitali producendo altri dati.

Il computer è costituito da una ***unità di elaborazione***, da ***organi di ingresso e di uscita***, che fanno interagire i computer con il mondo esterno, e da diverse memorie, la ***memoria centrale***, dove risiedono i dati che sono elaborati da programmi, e la **memoria secondaria** dove risiedono i dati che sono memorizzati permanentemente.
I dati digitali sono rappresentati con due cifre, ***0 e 1***. I ***programmi software*** sono definiti in un ***linguaggio programmativo***, che può essere il ***linguaggio macchina***, direttamente eseguibile dalla unità di elaborazione, ovvero ***in linguaggio ad alto livello***, comprensibile da esseri umani. Un ***algoritmo*** è un programma descritto in maniera più astratta.

Gli esseri umani hanno sempre convissuto con i ***dati analogici***, i dati che ci forniscono informazioni sul mondo. Un dato analogico è una codifica di fenomeno del mondo, una temperatura letta da un termometro o il semaforo rosso a un incrocio.
Con il diffondersi dei dati digitali, accanto al ***mondo*** dei dati analogici o ***analogico*** si sta sempre più espandendo il ***mondo digitale***. Sia nel mondo analogico che nel mondo digitale i dati hanno un ***significato*** che noi possiamo comprendere se abbiamo conoscenza del fenomeno rappresentato.

Quando un dato analogico viene trasformato in un dato digitale, il significato risiede insieme al dato nel mondo digitale. Il significato è influenzato e modificato dal percorso che il dato segue e dalle ***trasformazioni*** che il dato subisce. La comprensione dei **dati digitali** e del loro **significato** sono ***entrambi*** rilevanti per utilizzare in modo ***consapevole*** e ***corretto*** i dati digitali.

I dati digitali, come anche i dati analogici, per cercare di rappresentare il mondo possono assumere tante **forme** diverse, dalle ***tabelle*** ai ***grafici*** ai ***diagrammi***, le ***mappe***, le ***immagini***, i ***video***. Queste diverse forme servono per rappresentare mediante dati digitali ciò che percepiamo con i sensi della ***vista***, ***dell'udito***, ***l'olfatto***, il ***tatto***. Vi sono anche altri tipi di dati nati con l'avvento del Web, che permettono di collegare tra di loro dati su siti diversi.
Le tabelle e i testi sono i dati più utilizzati nelle organizzazioni, le immagini, i video e i testi sono utilizzati nei telefoni mobili e nelle reti sociali.

Le ***tabelle*** sono composte di ***righe*** e ***colonne*** dove vengono rappresentati i ***valori*** dei dati, e per questa caratteristica sono chiamati **dati strutturati**, i ***testi*** sono costituito da ***frasi in una lingua naturale***, e in genere hanno una struttura meno strutturata rispetto alle tabelle, chiamata anche ***debolmente strutturata***, costituita da capitoli, sezioni, articoli, commi, a seconda della natura del testo, ad esempio una legge o un articolo di ricerca o un documento amministrativo o un rapporto.

Per dare un ordine ai dati, e poterli usare, è opportuno rappresentarli per mezzo di un insieme di ***strutture***, che tutte insieme costituiscono un ***modello dei dati***, ad esempio la ***tabella*** e ***l'attributo*** per il ***modello relazionale*** dei dati. Un modello dei dati può essere visto come un **paio di occhiali** attraverso cui poter osservare **il mondo dei dati digitali**, che a sua volta permette di rappresentare il **mondo analogico**, sia pure solo in parte.
Anche i ***grafi semantici*** sono un modello dei dati, che, rispetto al modello relazionale, rappresenta il mondo per mezzo di ***entità*** e ***relazioni tra entità***; al contrario del modello relazionale, i grafi semantici hanno una ***rappresentazione grafica*** o ***mediante diagrammi,*** sia per le entità, i **nodi** del grafo semantico, sia per le **relazioni tra entità**, gli **archi** del grafo semantico. Esistono dunque ***modelli di dati*** e loro ***rappresentazioni grafiche***, i due concetti vanno tenuti distinti.

I dati digitali rappresentano il mondo, ma lo rappresentano spesso, per tante ragioni, in modo ***impreciso***, si può dire anche con ***scarsa qualità***. La qualità dei dati è un concetto multiforme, ha tante sfaccettature o ***caratteristiche*** di qualità, che dipendono molto anche dalla ***forma*** che i dati digitali assumono. Le ***tabelle***, una delle forme di dati più usate, hanno come principali caratteristiche di qualità la ***accuratezza*** e la **completezza**. Accuratezza significa che i dati digitali rappresentano esattamente i valori reali nel mondo, completezza significa che li rappresentano tutti. Per le immagini le caratteristiche più importanti sono la ***fedeltà*** all'originale e la ***estensione del colore***, che è l'opposto della ***opacità***. Per le immagini abbiamo visto una caratteristica di qualità, la ***utilità***, che è rilevante per tutti i tipi di dati digitali; un dato è utile quando ci serve per la decisione o azione che intendiamo intraprendere. Abbiamo osservato per le immagini che fedeltà e utilità possono essere in contrasto o ***tradeoff***, quando aumenta l'una diminuisce l'altra, questo è tipico di molte caratteristiche di qualità.

Una volta scoperto von una ***valutazione di qualità*** che un dato ha una qualche caratteristica di qualità insoddisfacente, possiamo ***correggerlo, migliorandone la qualità***; ad esempio, per un nome di comune italiano visibilmente sbagliato, possiamo confrontarlo con i nomi dei comuni italiani, sostituendolo con il nome più vicino.

La ***qualità dei dati nel Web*** è molto più difficile da trattare, perché nel Web costa poco sforzo inviare dati, messaggi, opinioni, e ci vuole più tempo per verificare se un dato è falso che per accettarlo come vero. Tuttavia i dati hanno una ***tenuta***, come diceva Umberto Eco, insomma se facciamo delle ***verifiche*** anche ***empiriche***, e esercitiamo ***spirito critico***, prima o poi riusciamo a distinguere il ***dato falso*** dal ***dato vero***.

Anche nella nostra epoca dei big data, sono di più le cose che non sappiamo di quelle che sappiamo, i **dati che non abbiamo** rispetto ai dati che abbiamo.

I dati che non abbiamo sono di tanti tipi. Possono essere **dati che ci mancano** di cui conosciamo la mancanza, ovvero ***dati che non sappiamo di non avere***, oppure ***dati che sono cambiati nel tempo***, oppure ***dati che non conosciamo e che il nostro interlocutore conosce bene***, oppure ***dati per i quali non sappiamo come il nostro interlocutore li definisce***. E altri ancora. I ***dati che non abbiamo*** sono complessi da trattare perché la ricerca non li ha ancora studiati a sufficienza.

Le ***astrazioni*** sono uno strumento che usiamo spessissimo nella nostra vita, soprattutto nel linguaggio verbale o scritto, quando dobbiamo dare un nome a un oggetto o a un concetto. Quando usiamo una ***astrazione*** o un ***processo di astrazione***, noi trascuriamo gli ***aspetti non rilevanti*** (ad esempio per un pino le varie tipologie, pino romano, pino marittimo ecc.) per mettere in evidenza gli ***aspetti comuni*** (sono tutti pini). Le astrazioni più usate sono la ***generalizzazione*** (es tutti i pini romani sono pini, quindi pino è una generalizzazione di pino romano), la ***aggregazione*** (un pino è formato da una radice, un fusto, un insieme di rami, un insieme di foglie, quindi pino è aggregazione di radice, fusto, rami, foglie), e il ***clustering***, in cui, ad esempio per i ***grafi***, un insieme di nodi caratterizzati da qualche similitudine sono sostituiti da un unico nodo.

Il procedimento opposto alla astrazione è il ***raffinamento***, in cui, al contrario della astrazione, un concetto è sostituito da un altro concetto in cui sono introdotti dettagli. Il raffinamento associato alla generalizzazione è la ***estensione***, il raffinamento opposto alla aggregazione è la ***disaggregazione***. Usando astrazioni e raffinamenti, noi possiamo rappresentare un oggetto complesso (ad esempio il grafo semantico dei personaggi dei Demoni) attraverso un processo generativo, in cui (continuando nell'esempio) un grafo semplice viene via raffinato in grafi sempre più dettagliati.

L'astrazione è anche utile per contrastare il ***data overload***, quando cioè, ad esempio in una ricerca bibliografica, abbiamo troppi dati. I dati disponibili possono essere organizzati in livelli di astrazione, per esempio rappresentando ognicarticolo con un breve riassunto o abstract.

Una immagine (che mostra un dato) è meglio di 1.000 parole (che lo descrivono), perché l'effetto visivo delle ***visualizzazioni*** costitute da ***tabelle, grafici, diagrammi, mappe, icone*** riassume un ***concetto***, un ***messaggio***, un insieme di dati risultato di un calcolo, molto meglio di un **testo scritto**. Le tabelle sono le più semplici visualizzazioni, rispetto ad esse grafici, diagrammi, mappe e icone sono più espressive; d'altra parte, la ***espressività*** può essere accompagnata da distorsioni, che dobbiamo riconoscere per evitare che la percezione del dato sia sbagliata. Inoltre, le icone usate per rappresentare un dato (ad esempio dove si trovi la toilette in un aeroporto) hanno sempre inevitabilmente un **contenuto** ispirato da una **cultura** retrostante, che deve essere riconosciute per evitare una percezione, appunto, distorta. Per cui talvolta può essere importante partendo da un ***grafico*** e da una **icona**, osservare la **tabella** o il **testo** che rappresentano, per verificare la fonte, origine della visualizzazione.

Il ***machine learning (ML) o apprendimento automatico supervisionato*** ha lo scopo di costruire algoritmi predittivi che apprendono a predire e decidere sul futuro osservando dati del passato. Per esempio un giudice concede la libertà provvisoria a un detenuto valutando il rischio futuro di recidiva, sulla base di quanto accaduto nel passato con detenuti ***simili***.

Accanto al ML supervisionato esistono altre tipologie, come il ***ML non supervisionato*** e il ***ML per rinforzo***. Gli algoritmi di ML supervisionato sono caratterizzati da una ***accuratezza*** che si può misurare valutando i ***veri positivi, veri negativi, falsi positivi, falsi negativi*** su ***dati di test***. Un'altra caratteristica del ML supervisionato è la equità dell'algoritmo nei confronti di tutti i soggetti cui viene applicato. Esistono diverse definizioni di equità. Un limite del ML è nel fatto che per **decidere sul nuovo si basa sul passato**.

I dati digitali sono usati a vario titolo da aziende, organizzazioni pubbliche, comunità e singole persone. A seguito di questa grande varietà di usi, i dati possono avere diversi tipi di ***valore***, come il ***valore d'uso***, ***valore economico***, ***valore sociale***, ***valore conoscitivo***.
I dati sono diversi rispetto ai due tradizionali ***oggetti di scambio***, i ***beni*** e i ***servizi*** e rispettano diverse ***leggi economiche***, negli aspetti che riguardano la produzione, la riproduzione, lo scambio, ed altre, dette ***leggi di Moore***, come ad esempio il fatto che al contrario dei beni e dei servizi i dati scambiati non si esauriscono. Il ***valore d'uso*** può vedersi come un bilancio tra ***benefici*** che abbiamo nell'usare i dati, e i **sacrifici** che dobbiamo compiere per ottenerli, tra cui il ***costo economico*** dei dati ***e lo sforzo*** che dobbiamo compiere per ottenerli***.*** Il ***valore sociale*** dei dati fa riferimento a quanto la conoscenza che ci portano migliora la nostra ***qualità della vita***. ***Valore economico*** e ***valore sociale*** possono creare valori d'uso in contrasto tra di loro, se aumenta il valore sociale per una persona, può diminuire il valore economico per un'altra.

I dati possono rappresentare il mondo e possono essere acceduti in modo diseguale (fenomeno del ***data divide***), accrescendo in tal modo le diversità tra nazioni e tra comunità. I ***dati aperti*** sono dati che per le loro caratteristiche e modalità di pubblicazione possono essere acceduti da tutti e collegati tra di loro nel Web.

La ***psicologia cognitiva*** ci dice che quando ***comunichiamo*** con altri per mezzo di dati digitali ovvero ***acquisiamo*** dati digitali dal Web noi possiamo dedicare solo uno sforzo limitato per comprenderli (***razionalità limitata***), per cui spesso usiamo ***euristiche***, cioè procedimenti approssimati, che ci possono portare a ***distorsioni*** (***bias***) nel processo cognitivo, basate sulla ***reputazione*****,** la ***violazione delle aspettative*****,** la ***consistenza*****,** o assenza di contraddizioni**,** ***l'intento persuasivo*****,** e ***l'auto-conferma*****.**

La psicologia cognitiva ci dice anche che l'accesso al Web e alle reti sociali tende a polarizzare le opinioni politiche; inoltre la progressiva sostituzione dei giornali con il Web e le reti sociali come fonti di dati, e riduce la ***qualità delle analisi*** e degli ***approfondimenti*** sui ***fatti***; ciò anche perché produrre notizie verificate ha un costo, mentre inviare un messaggio su Twitter non costa nulla e raggiunge in alcuni casi molte più persone.

Il Web e le reti sociali sono luoghi virtuali dove noi scambiamo e viviamo ***emozioni***, è la ***rabbia*** quella che tende di più a diffondersi, rispetto a ***gioia***, ***disgusto*** e ***tristezza***, e altre ancora. Tutto questo non è scontato, per cui se li sappiamo usare, i dati hanno anche un grande ***valore emozionale*** positivo.

L' ***etica dei dati*** viene trattata facendo riferimento a un insieme di ***principi o determinanti***, così come la ***qualità dei dati*** è tradotta in termini di ***caratteristiche***; essi riguardano: la ***trasparenza*****,** ***responsabilità,*** ovvero ***capacità di rispondere dei propri atti*** **(*accountability)***

equità o ***imparzialità di un algoritmo di machine learning*****,** ***spiegabilità*** o ***capacità di spiegare*****,** ***generalizzazione verso personalizzazione*****,** ***consapevolezza*****,** ***inclusione*****,** ***parità di opportunità (non discriminazione, egualitarismo)*****,** ***qualità dei dati*****,** ***privacy*****,** ***condivisione*****,** ***affidabilità*****.**

L' ***equità*** di ***un algoritmo di machine learning*** è un determinante che ha varie possibili definizioni e ***significati***. Tende sempre a essere definita come ***non discriminazione di*** specifiche ***comunità*** di persone a cui si applica l'algoritmo predittivo decisionale. Queste comunità possono essere associate a un ***genere***, tipicamente le donne, a una ***etnia***, ad es. negli USA gli Afroamericani, ecc. L'equità può essere o meno ***disciplinata da una legge***. L'equità (o la non equità) può derivare dall' ***algoritmo predittivo***, dalle ***persone coinvolte***, ad esempio, ***nella scelta dei dati usati nell'apprendimento***, e dai ***dati stessi***.
La ***spiegabilità*** è la caratteristica per cui la tecnica usata nell'algoritmo di apprendimento, ad esempio gli ***alberi decisionali***, è ***comprensibile*** a ***chi applica l'algoritmo*** per le proprie decisioni (ad es. un giudice) e a ***chi subisce la decisione*** (ad es. un detenuto). Se un algoritmo adotta una tecnica non spiegabile, è possibile cercare per l'algoritmo una nuova tecnica che accresce la spiegabilità, verificando però che questa non porti a un peggioramento della accuratezza dell'algoritmo.

Riguardo alla ***privacy***, il regolamento per la protezione dei dati personali promulgato dalla Unione Europea sancisce diversi principio tra cui: il ***consenso informato***, la ***portabilità*** dei dati, la ***opacità*** dei dati personali, la ***qualità*** dei dati, il **diritto alla *cancellazione***, il ***diritto alla spiegazione***, una ***classificazione*** più ampia dei ***dati personali e sensibili***, la ***responsabilità*** delle organizzazioni che usano i dati, l' ***analisi di impatto*** della protezione dei dati, il ***trasferimento*** dei dati personali fuori dalla Unione Europea, l'istituzione nelle organizzazioni di un ***data protection officer***, ***l'ambito territoriale*** di applicazione.

Conclusioni

Perchè una Enciclopedia sui dati digitali

Siamo alla fine di questo primo viaggio esplorativo alla scoperta dei dati digitali. Abbiamo visto quante grandi possibilità per la nostra vita crei il poter disporre del grande oceano di dati che le tecnologie digitali ci mettono a disposizione, ma abbiamo visto anche quali nuovi problemi e nuove sfide essi ci pongano.

E abbiamo visto quanti siano i fili della tela e le tessere del mosaico che è necessario costruire per raggiungere una consapevolezza e una maturità nel loro uso. Accanto alle due tradizionali competenze definite per gli esseri umani, con termini inglesi la *numeracy*, in poche parole il saper usare le astrazioni della matematica per risolvere problemi, e la *literacy*, il saper comprendere un testo parlato o scritto e il sapersi esprimere oralmente e tramite un testo scritto, nasce nella nostra cultura il tema della *datacy*, il saper usare i dati digitali per risolvere problemi, per capire il mondo, per comunicare con gli altri.

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 le tabelle, i testi, le immagini, essendo in grado di:

- ricostruirne e modellarne il significato,
- valutare il loro livello di qualità,
- applicare tecniche e algoritmi basati su apprendimento per costruire modelli decisionali, interpretativi, predittivi e comrenderne il funzionamento,
- visualizzare i dati per comprenderne meglio la natura 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 digitali.

Riguardo alle similitudini tra la datacy e la literacy, ho iniziato questo libro ricordando la storia della alfabetizzazione linguistica. Che vi siano grandi somiglianze tra il modo con cui usiamo l'Italiano scritto e verbale e il modo con cui dobbiamo usare i dati digitali ce lo ricorda George Orwell nel pensiero di Figura 166 tratto dal testo in nota [28]. Così come c'è un effetto reciproco tra ragionamenti e pensieri e lingua che usiamo, c'è una relazione anche tra ragionamenti e pensieri e come usiamo i dati.

[28] George Orwell, *Politics and English language* – Penguin Classics, 2013, tradotto in italiano come George Orwell, La neolingua della politica, Garzanti Classici, 2021.

La datacy è poi profonda debitrice dalla numeracy dell'immenso catalogo di modelli che la matematica ha costruito nei millenni; e tuttavia ha qualcosa di diverso e di originale e di nuovo che la contraddistingue, perché i dati digitali rappresentano il mondo, e quindi aspirano forse un po' presuntuosamente a descrivere e elaborare il significato delle cose del mondo, in qualunque forma siano rappresentate digitalmente.

A man may take to drink because he feels himself to be a failure, and then fail all the more completely because he drinks. It is rather the same thing that is happening to the English language. It becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts

…. L' inglese diventa brutto e impreciso perchè i nostri pensieri sono stupidi, ma a sua volta la sciatteria della lingua ci rende più facili ai pensieri stupidi

Figura 166 – Un pensiero di Goerge Orwell

In questo libro abbiamo iniziato ad assaporare i problemi che devono essere affrontati nell'uso dei dati digitali e in alcuni casi abbiamo cominciato a vedere soluzioni a questi problemi. Ma, come diceva Einstein, "Everything should be made as simple as possible, but not simpler", ogni cosa dovrebbe essere resa la più semplice possibile, ma non più semplice, insomma non dovrebbe essere (mia traduzione) *banalizzata*.

So bene quale sia il grande rischio di questa Enciclopedia, che io banalizzi problemi complessi e ancora lontani dall'essere risolti. Non ci sono risposte semplici a domande complicate, come ci dice la vignetta di Figura 167, che un po' ci fa sorridere, un po' ci inquieta.

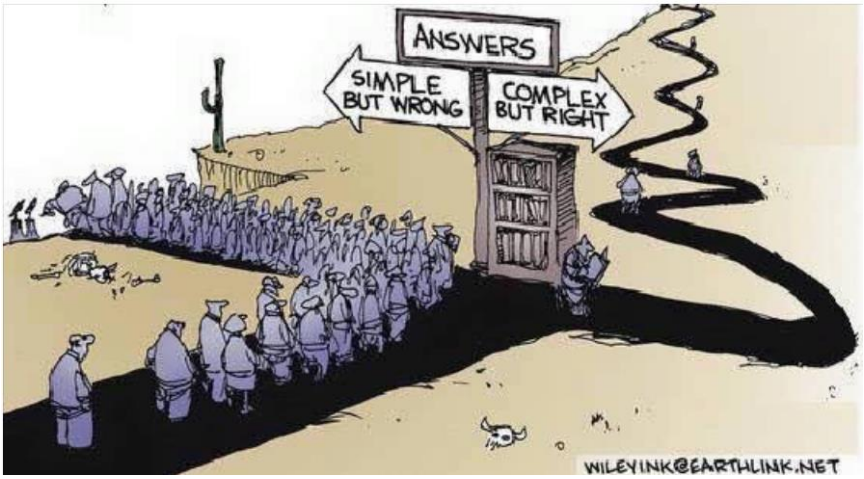


Figura 167 – Non ci sono risposte semplici a domande complicate

Il fenomeno della digitalizzazione e dei big data è caratterizzato da ineguale sviluppo, e possiamo constatare che la crescita del *volume* dei dati digitali ha un andamento esponenziale, e che il *costo* delle elaborazioni sui dati diminuisce rapidamente (pensiamo ad es. al sequenziamento del genoma, i cui costi sono diminuiti di diversi ordini di grandezza negli

ultimi anni), così che i due fenomeni visti congiuntamente provocano profonde modifiche al modo stesso con cui comprendiamo il mondo, comunichiamo, viviamo le nostre emozioni, e alle leggi della economia digitale e allo sviluppo delle società.

Il ritmo con cui la ricerca scientifica e le scienze informatiche, statistiche, sociali, cognitive, economiche evolvono e studiano il fenomeno dei dati digitali segue necessariamente leggi lineari, per cui la distanza tra le domande poste dalla digitalizzazione e le risposte fornite dalla ricerca scientifica tende talvolta ad approfondirsi più che a colmarsi, vedi Figura 168.

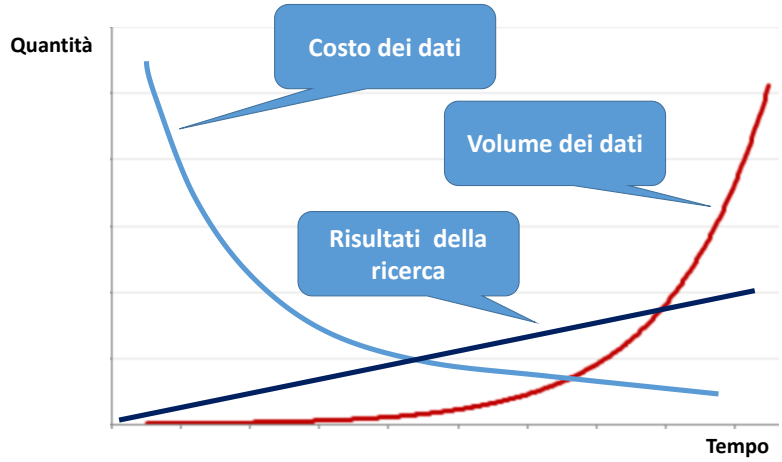


Figura 168 – I dati digitali evolvono velocemente in dimensione e costi,
la ricerca è più lenta e fa fatica a seguirne le implicazioni per la nostra vita

Ebbene, arrivato a questo punto della mia vita, dedicata per molta parte allo studio e alla riflessione sui dati digitali, penso che valga la pena accettare la sfida, percorrendo lo strettissimo sentiero tra rigore, chiarezza, e banalizzazione. Questo libro è servito per introdurre ai diversi aspetti del grande affresco dei dati digitali. Nella pagina successiva compare il piano dell'opera; caro lettore, incrocio le dita, vorrei percorrere questo sentiero fino alla sua meta finale, la cima della montagna.

Carlo Batini, febbraio 2021

Enciclopedia dei dati digitali

Piano dell'opera

1. **Introduzione ai dati digitali** - I dati digitali sono una finestra sul mondo
2. **Modelli** - I modelli dei dati ci aiutano a rappresentare e comprendere il mondo
3. **Forme** e **Significato** dei dati digitali
4. **Qualità** - I dati vanno curati
5. **Integrazione** - La Babele dei dati: i dati vanno riconciliati
6. **Astrazioni** - La Babele dei dati: la potenza delle astrazioni
7. **Valore** - Il valore e il dis-valore dei dati
8. **Apprendimento** - Imparare con i dati
9. **Visualizzazione** - Meglio una immagine di 1.000 dati: la visualizzazione dei dati
10. **I dati che ci mancano** – Come operare sui dati quando non li abbiamo
11. **Dati, psicologia cognitiva e emozioni**
12. **L'etica** dei dati

Ringraziamenti

Desidero ringraziare con tutto il cuore le tante persone che mi hanno aiutato nel leggere versioni iniziali di questo libro e nel fare osservazioni che hanno contribuito moltissimo a migliorare il testo.

Mi riferisco a

Chiara Batini
Giuseppe Batini
Laura Batini
Chiara Damiani
Anna Ferrari
Valentina Fortichiari
Gaetano Santucci
Claudio Salone
Fabio Stella

Gaetano Santucci in particolare ha fornito contributi per il Capitolo 6 sui modelli, per il Capitolo 8 sulle astrazioni e per il Capitolo 14 sull'etica dei dati.

Naturalmente la responsabilità delle cose scritte è solo mia.