Alessandro Viviani

«BIG DATA»

Firenze 23 gennaio 2020

Molte delle nostre teorie sono state concepite e sviluppate quando:

- i dati erano rari e costosi;

- le scienze procedevano con cautela e con lentezza.

Dati strutturati e non strutturati: la loro misurazione.

Dalle biblioteche (Alessandria 500000 rotoli, quella del Congresso 158.000.000 volumi) al **bit** (*binary digit).*

Dal bit al **Byte** (8 bit rappresentano 256 sequenze), nuova unità di misura. Da qui alle migliaia di miliardi di byte.

Ogni giorno, più o meno consapevolmente, produciamo, e incontriamo, un’immensa quantità di dati. Dati che seguono i nostri comportamenti.

Molte operazioni lasciano traccia!

Mai prima d’ora i dati ci giungono con tale ritmo, varietà e volume.

Testi, immagini, suoni, numeri, video, sensori, pdf, email, social network, influenzano il nostro modo di conoscere e di prendere decisioni (?).

Quanti dati sono necessari?

Cosa sono i dati

I dati sembrano di per sé oggettivi: **data** è plurale di ***datum,*** cioè fornito da un Agente (il produttore) in un determinato tempo e spazio e con un mezzo (***medium***).

Ogni dato è una rappresentazione della realtà, anche quando è dato numerico, verbale, contestuale all'evento, contabilizzazione del presente e immediatamente disponibile, come nel caso dei Big Data. Il valore del dato come strumento di valida conoscenza alla sua modalità di rappresentazione della realtà nei suoi aspetti di raccolta (ideazione)dei dati e loro analisi.

Dati sperimentali e dati osservazionali.

Ciò che non è sperimentale (in area economica, sociale, comportamentale)

sta diventando sempre più guidato dai dati.

Quale è la quantità di dati necessaria per uno specifico problema?

Scarsità ed eccesso di dati: improduttivi (!)

La rappresentazione della realtà, oggi legata alla tecnologia usata nella produzione del dato, deve esser scientificamente condotta e le modalità di raccolta e analisi svelate.

Evoluzione tecnologica non automaticamente accompagnata da adeguata evoluzione culturale, che metta in grado i non specialisti di capire e trarre vantaggio dall'innovazione.

Il controllo sociale sull'innovazione non è automatico né garantito.

La grammatica è la Statistica (Data Science)

Cosa sono (!?)

Dati che provengono, in generale, via web

Big Data sono caratterizzati da almeno 3 V:

VOLUME

VELOCITA’

VARIETA’

Non è un oggetto ben definito e la capacità di usarlo per risolvere problemi dipende dal problema stesso.

BIG: il Volume

Dalla trasmissione (dati e informazioni) orale a quella scritta e quella digitale.

Papiri, biblioteche e ora computer sempre più potenti.

«Accumulare dati non implica aumentare l'informazione per la conoscenza…possiamo avere dati senza informazione ma non informazioni senza dati»

Un flusso inarrestabile di dati:

La VELOCITA‘

I dati che produciamo e che si si accumulano

sono trasmessi sostanzialmente in tempo reale.

Banche, like, post, registri medici, registri amministrativi e commerciali: tanti esempi.

Come il Volume, la Velocità ha imposto con l’evoluzione digitale una nuova unità di misura:

Il processore.

Il mondo è bello perché è vario: la Varietà’

Dalle tabelle al foglio excel (righe e colonne): per ottenere analisi descrittive, sintetiche o modelli più sofisticati (dati strutturati).

Molte fonti, eterogenee e non convenzionali, coesistono contemporaneamente: immagini, mail, what’up, tweet, video, dati testuali, post, like…(dati non strutturati).

Come raccoglierli, immagazzinarli, trattarli..

I data base tradizionali:inadeguati

NoSQL (Not OnlyStructured Query Language):Hadoop

Problemi vecchi e nuovi: rottura col passato

La soluzione «muscolare»:dalla caverna al grattacielo (+sforzi e + persone).

Per Big Data: è un problema informatico e quindi sistemi di calcolo e memorie più potenti e veloci. Il sistema evolve e cresce fino a occupare spazi più ridotti e modi di lettura più versatili (es. analisi semantica)

La soluzione «cerebrale»: cambiare approccio

Uso congiunto di strumentazione informatica e di metodo statistico, un approccio basato su dati empirici, che tende alla quantificazione e semplificazione della realtà per giungere a decisioni empiricamente fondate:

un metodo che da secoli informa la conoscenza umana in tutti i campi.

«non ci si può innamorare di pratica senza scienza»

Un’ultima V: Veridicità

La grande varietà delle fonti dei dati non consente di controllarne la qualità:

errori, imprecisioni (es. dati di natura osservativa), fake news.

Quando i dati non sono raccolti per una finalità conoscitiva occorre separare i dati «sporchi» dall’accumulo: ad esempio a volte si ha un campionamento di convenienza (!) come dati raccolti su base volontaria .

Anche qui vale la consueta distinzione tra segnale e rumore.

- i social network;
-i dati crowdsourced (trasmessi dagli utenti);
-imprecisione di strumenti;
-dati mancanti;
-dati raccolti senza campionamento.

Validità - Volatilità -Privacy

- i dati devono essere modificati e corretti;

- per quanto tempo sono validi i dati e per quanto devono essere conservati;

-alcuni dati sono modificati per proteggere la riservatezza degli individui.

I Dati sono il nuovo petrolio ?

Disponibilità dei dati : vantaggio competitivo.

Implementazione di Big Data per sfruttarli a pieno: registrare operazioni, relazioni con clienti.. algoritmi che imparano dal passato..

(Amazon, Google, Facebook..).

Ancora: target per pubblicità, credit scoring..

Le economie più forti sono motivate all'analisi di enormi quantità di dati; 4,6 miliardi di *smartphone* attivi e più di 2 miliardi accedono alla rete.

Consapevolezza sui metodi di generazione dati e per quanto riguarda la loro analisi?

La natura persuasiva dei dati a disposizione rischia di appannare le nostre capacità di scelta consegnandole per comodità ad elaborazioni fatte da altri. Una profilazione e sintesi di dati «eterodiretta» sulla base di modelli e *machine learning* proprietarie.

Tutto questo è riassunto in:

«non puoi gestire ciò che non misuri»

I dati non hanno valore di per sé:

-combinazione con altri dati;

-modello interpretativo.

Accresciute esigenze di analisi dei dati:

gestione dei dati e loro valorizzazione.

Una nuova figura professionale: **data scientist**

Tecniche di analisi e visualizzazione,

Programmare disegni di esperimenti,

Legami causa-effetto nei fenomeni empirici.

Sexy Job : lo statistico, come l’ingegnere informatico negli anni 90 (Varian)

Il metodo quantitativo :

misurare per conoscere.

Senza Big Data siete come ciechi e sordi nel bel mazzo di un’autostrada (G.Moore).

Qualcuno che guida, un metodo di ricerca in quell’autostrada; pur vedendoci e sentendoci benissimo rischiamo di perire investiti dalla gran massa di dati.

Di fronte alla complessità, con l’incapacità di comprenderla nella sua interezza:

circoscrivere i fenomeni, definirli e misurarli.

I concetti di media, correlazione, variabilità campionaria, errore sistematico, errore casuale

sono termini trascurati dalla contemporaneità.

Con tanti, troppi dati questo non è più possibile.

Aristotele e Newton

«del particolare non si da scienza»

il ragionamento deduttivo

«dai particolari all’universale»

il ragionamento induttivo e il metodo sperimentale

Milan Kundera

Qualsiasi studente nell'ora di fisica può provare con esperimenti l'esattezza di un'ipotesi scientifica. L'uomo, vivendo una sola volta non ha possibilità di verificare un'ipotesi mediante esperimento e perciò (Thomas) non saprà mai se avrebbe dovuto o no dare ascolto al proprio sentimento.

Noi non possiamo osservare tutta la realtà, ma solo una parte e cerchiamo di trarre da essa considerazioni di tipo generale.

POPOLAZIONE

CAMPIONE

(probabilità)

Nel diluvio di dati come selezionarne alcuni e scartarne altri piuttosto che bulimicamente accumularli fino a perdere traccia del loro contenuto informativo?

«E’ meglio non guardare che guardare malamente» (Poincarè)

POPOLAZIONE

-analisi descrittiva (sintesi,grafici…)

CAMPIONE

- stima (calcolo delle probabilità,induzione)

BIG DATA (?)

Che fare?

Come trattarli?

Quali informazioni «estrarre»?

Non tanto affermare una verità assoluta, ma approssimare meglio le spiegazioni ai fenomeni empiricamente osservati.

La perduta arte della semplicità

«Mi scuso per la lunghezza di questa lettera, ma non ho avuto il tempo di renderla più breve»

Mark Twain

BIG DATA: il Sesto Potere

Chi detiene Big Data ed è in grado di trarne informazioni detiene potere.

Cambridge Analytica

Ogni nuova tecnologia nel tempo ha spazzato via il vecchio mondo (vapore, elettricità, energia atomica...Big Data).

Prendere decisioni su dati empirici di fonti diversissime. I dati sono ampiamente disponibili; e la capacità di estrarne conoscenza?

Dopotutto una rivoluzione non chiede il permesso: da una mentalità logico-deduttiva ad una empirico-induttiva.

Attenzione alla vulnerabilità!

Grazie per l’attenzione