Lezione 8

Le evoluzioni in atto:
Big Data e Data Science

Metodi statistici per l'analisi socio-economica

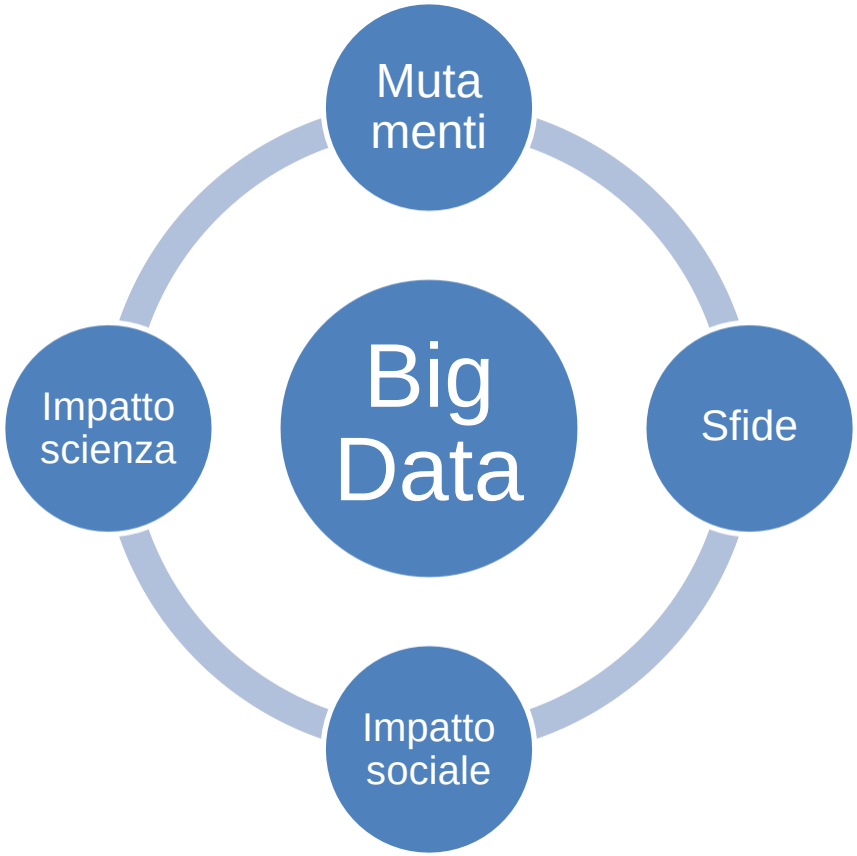
Docente: Giovanni Giuseppe Ortolani

Corso di Laurea Magistrale in Economia dei settori produttivi e dei mercati internazionali

a.a. 2021/2022

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Di cosa parleremo?



Data is the new OIL …

or maybe the new SUN

The Economist

Menu | Weekly edition | Search

Leaders | Regulating the internet giants

The world's most valuable resource is no longer oil, but data

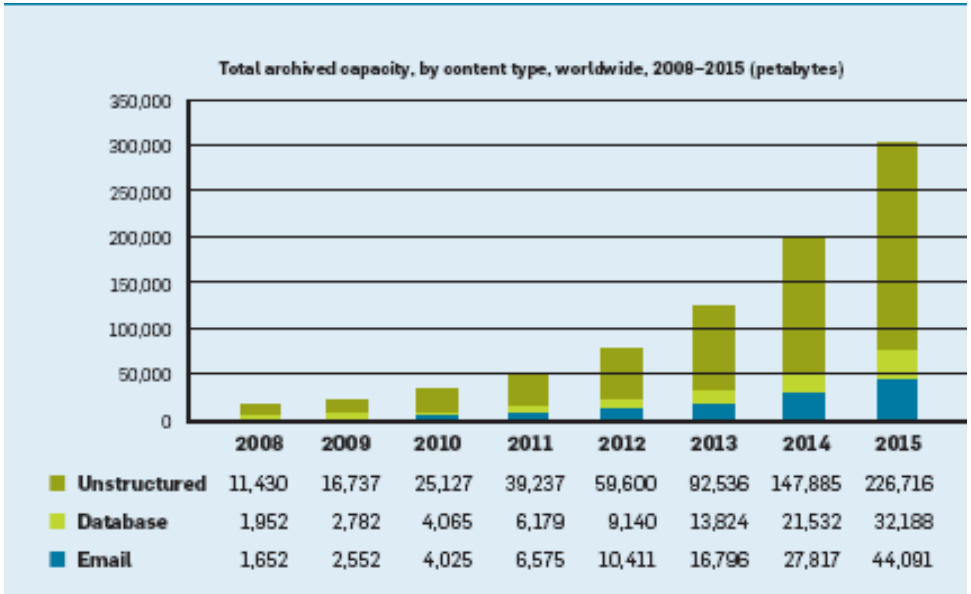
The data economy demands a new approach to antitrust rules

May 6th 2017 *(Updated May 11th 2017)*

Share

Crescita esponenziale

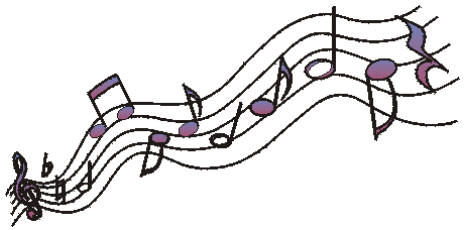
- I dati crescono in media del 30-40% annuo
- Ogni 2,5 anni si raddoppia il volume
  - 2,7 ZB (10^{21} bytes) nel 2012!
  - 35 ZB nel 2020

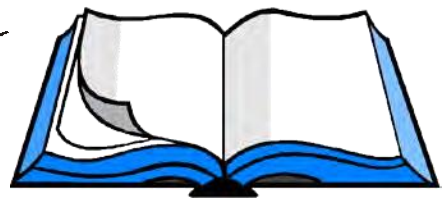


Nome	Simbolo	Multiplo
kilobyte	kB	10^3
megabyte	MB	10^6
gigabyte	GB	10^9
terabyte	TB	10^{12}
petabyte	PB	10^{15}
exabyte	EB	10^{18}
zettabyte	ZB	10^{21}
yottabyte	YB	10^{24}

Datizzazione

- Neologismo che indica la conversione in formato digitale (dati) di:
- Film, musica, libri, etc. (contenuti che fino a qualche anno fa viaggiavano su pellicole, carta, vinili e altri supporti)
- Conversazioni telefoniche, mail, trasmissioni televisive e radiofoniche





Datizzazione

- *Facebook* ha “datizzato” le relazioni,

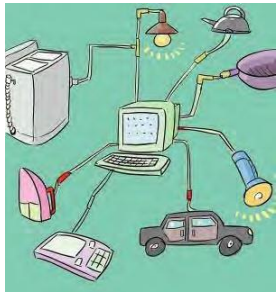
- *Twitter* ha reso possibile la “datizzazione” dei sentimenti,

- *LinkedIn* ha “datizzato” le nostre esperienze professionali

I dispositivi generano dati …

- **Internet of Things** (IoT): Gli oggetti si rendono riconoscibili e acquisiscono intelligenza grazie al fatto di poter comunicare dati su se stessi e accedere a informazioni aggregate da parte di altri.

I dispositivi generano dati …



Esempi:

- Le sveglie suonano prima in caso di traffico,
- Le piante comunicano all'innaffiatoio quando è il momento di essere innaffiate,
- i vasetti delle medicine avvisano i familiari se si dimentica di prendere il farmaco.

Tutti gli oggetti possono acquisire un ruolo attivo grazie al collegamento a Internet.

Noi generiamo dati …

- Grazie alla nostra forte simbiosi con le tecnologie digitali, siamo diventati dei "sensori" viventi.
- 7 miliardi di persone e 6,8 miliardi di cellulari

La scienza genera dati …

- Le tecnologie digitali hanno permesso di fare passi da gigante, in questi anni, nel campo della **genomica**, dove le moli di dati da analizzare sono enormi.

- mappatura del DNA di un individuo

  da 3 miliardi di dollari e 13 anni di ricerca (1990-2003) ➔ poche migliaia di dollari per un processo che dura un paio di settimane.

La scienza genera dati …

- ***Human Brain Project***
- Un osservatorio del cervello che monitora 1 milione di neuroni (o 100.000 neuroni in 10 soggetti) per 1.000 volte al secondo genererebbe:
  - 1 gigabyte di dati al secondo,
  - 4 terabytes all'ora,
  - 100 terabytes al giorno
  - 4 petabyte all'anno (ipotizzando un fattore di compressione di 1/10).

Le aziende generano dati …

- Oggi ogni grande business è un digital business:
  - **Alibaba** è il più grande negozio al mondo, ma non ha nemmeno un magazzino.
  - **Uber** è la più grande compagnia di noleggio veicoli, ma non possiede nemmeno un'auto.
  - **Airbnb** è il più esteso network dedicato alla ricettività, ma è del tutto privo di strutture.

Le aziende generano dati …

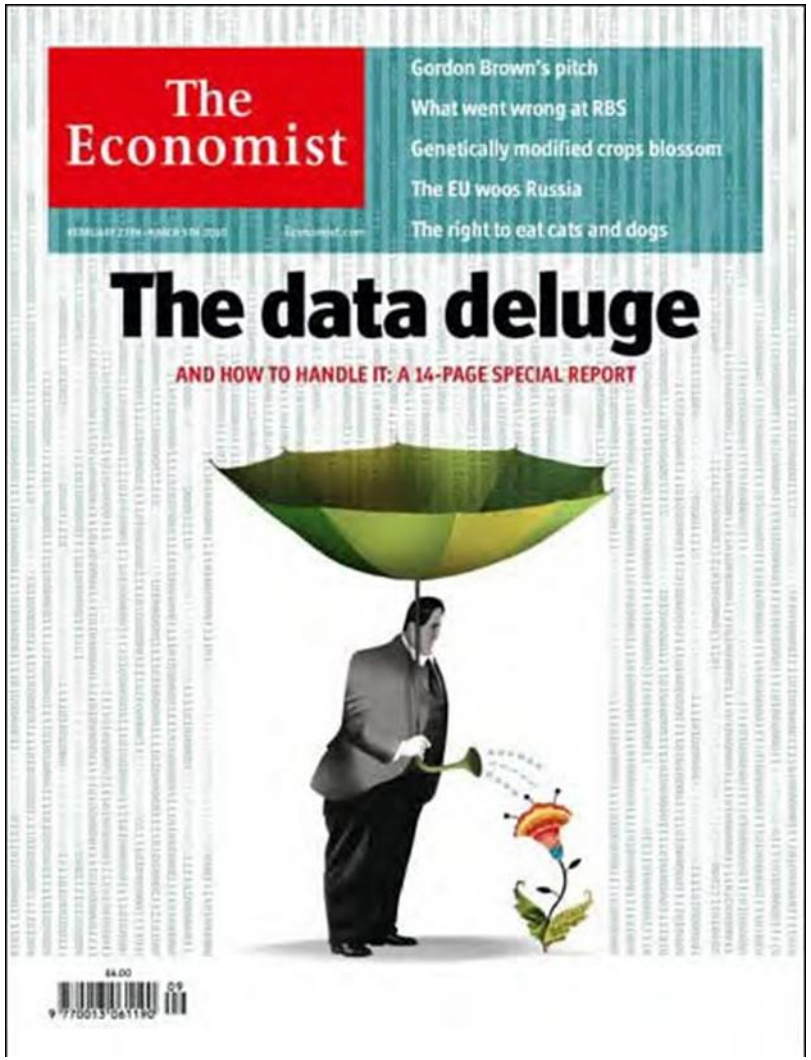
- Ordini, acquisti, vendite, spedizioni, difetti di produzione, …
- I dati sono raccolti nei sistemi informatici delle aziende. Sono considerati un *asset (intangibile).*
- *Facebook:* dichiara *asset* (tangibili) per 6,3 miliardi ma venne valutata in Borsa 104 miliardi il giorno del suo debutto.

Le aziende generano dati …

- Nonostante i dati siano un asset, oggi viene elaborato solo il **5‰** dei dati aziendali
- Perché?
  - mancanza di competenze sull'analisi computazionale dei dati;
  - sovversione dei poteri generati da un'informazione così tempestiva

Il Diluvio dei Dati

Il termine "**diluvio dei dati**" si riferisce alla situazione in cui le incredibili dimensioni dei dati generati sta sopraffacendo la capacità delle istituzioni nel gestirli e dei ricercatori nel farne uso nei loro studi.



Big Data

- Raccolta di dati così estesa in termini di volume, velocità e varietà da richiedere **strumenti non convenzionali** per estrapolare, gestire e processare informazioni entro un tempo ragionevole.

Big Data: una rivoluzione?

- La vera rivoluzione non sta nelle tecnologie per elaborare i dati, ma nei dati in sé e nel modo in cui li usiamo.

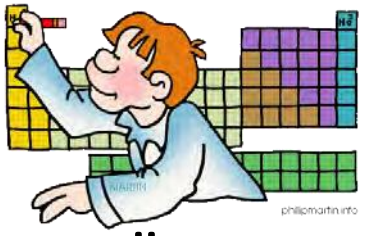
- Aumentando la scala dei dati con cui si lavora, si possono fare cose nuove che non sono possibili con minori quantità dei dati.

Big Data vs. Data Science

- Data Science
  - *La scienza dei dati studia i metodi per estrarre la conoscenza dai dati.*
    - *Dati di qualunque natura*
  - Un approccio olistico alla creazione di prodotti e servizi basati sull'estrazione di conoscenza dai dati
    - La conoscenza estratta è immediatamente utilizzabile (*actionable*) nei processi decisionali.

https://thispersondoesnotexist.com/

Big Data Analyst vs. Data Scientist



- *Data Scientist*
  - Figura professionale dotata di abilità “integrate” che spaziano dalla matematica, all’apprendimento automatico, alla statistica, al data mining, ai database e all’ottimizzazione
  - capace di ingegnerizzare delle soluzioni efficaci alla creazione di nuovi prodotti e servizi.
  - **Carenza di professionalità**: 150.000 data scientist richiesti solo negli USA

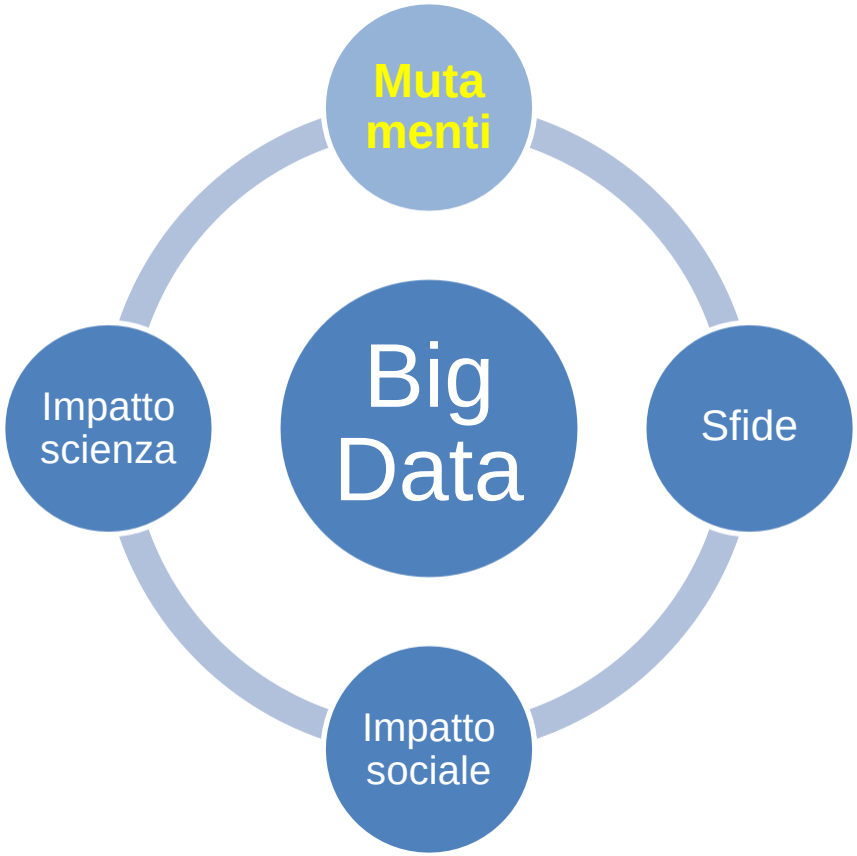
Big Data vs. Open Data

- Open Data
  - dati liberamente accessibili a tutti, privi di brevetti o altre forme di controllo che ne limitino la riproduzione
  - gli eventuali copyright eventualmente si limitano all'obbligo di citazione della fonte o al rilascio delle modifiche con stesso copyright.

Big Data vs. Open Data

- Open Data: iniziative governative
  - direttiva sull'Open government, amministrazione di Obama 2009
    - sito pubblico Data.gov portale che raccoglie i dati resi disponibili dagli enti statunitensi in formato aperto
  - Open Data Institute, UK
    - Incoraggia gli enti pubblici a rilasciare i loro dati in formato aperto e aiuta le start-up a sviluppare prodotti commerciali sulla base degli open-data
  - Portale italiano dell'Open data dati.gov.it

Di cosa parleremo?



Big Data: *Un esempio*

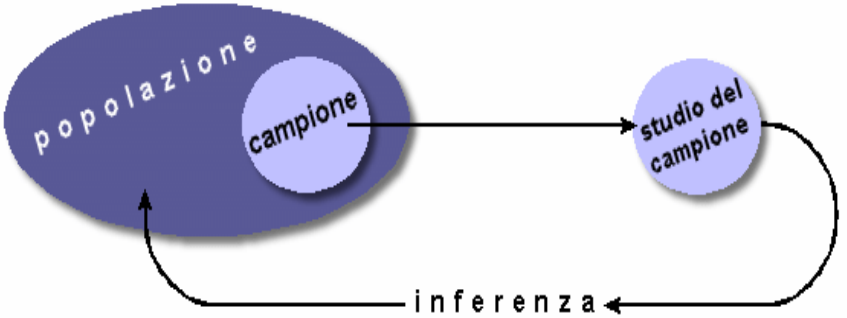
Google Flu Trends: previsione in base all'oggetto delle ricerche condotte con *Google search* ➔ ugualmente accurate ma in tempo reale

Detecting influenza epidemics using search engine query data. ***Nature 457, 1012-1014 (19 February 2009)***

Big Data: Un cambio di prospettiva

- L'ascesa dei Big Data evidenzia tre mutamenti nel modo in cui analizziamo le informazioni:
  1. Analizzare tutti i dati disponibili
  2. Accantonare il desiderio di esattezza
  3. Abbandonare la tendenza a ricercare la causalità

Big Data: Di più



- Analizzare tutti i dati disponibili
  - Assuefazione al campionamento statistico ➔ autolimitazione nell'uso delle informazioni
  - Il campionamento casuale è solo un ripiego
  - È poco utile quando si vuole scavare in profondità
  - Il campionamento trascura i dettagli

Big Data: Confusione

- Rinunciare all'esattezza
  - L'incremento dei volumi ➔ inesattezza
  - È importante avere *small data* (campioni) accurati
  - L'esattezza può essere sacrificata in favore dell'ampiezza o della frequenza
  - Accettare l'inesattezza dei modelli estratti dai dati o nella struttura dei dati
- Nell'epoca dei Big Data, ***la quantità è più importante della qualità***.
- L'abbondanza permette di tollerare un certo livello di imprecisione, di confusione

Big Data: Confusione

- L'esattezza vs. ampiezza / frequenza

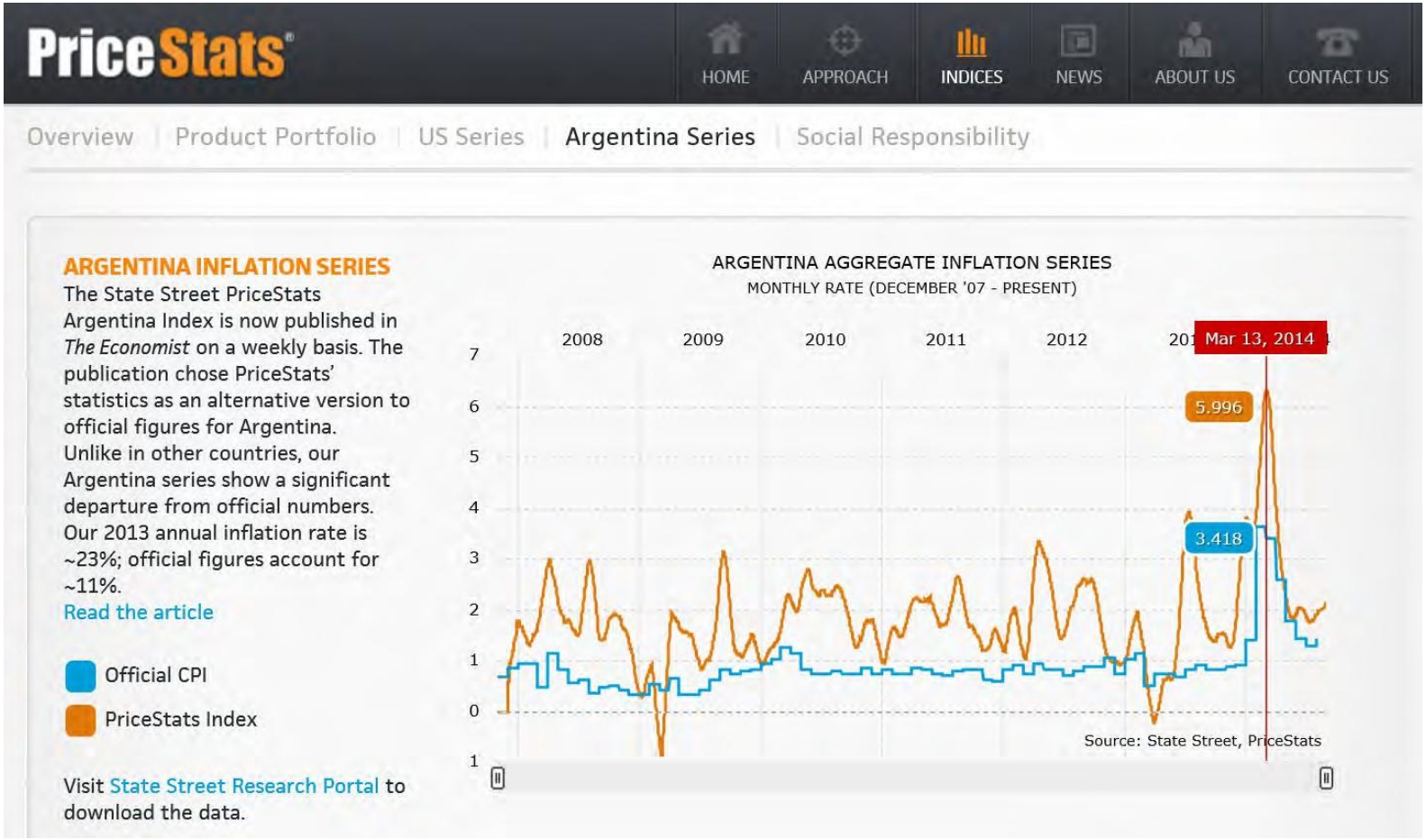
Vigna wireless della Intel. La rete di sensori wireless rileva la presenza di parassiti e permette di selezionare l'insetticida.

[*J. Burrell, T. Brooke, and R. Beckwith, "Vineyard computing: Sensor networks in agricultural production," IEEE Pervasive Computing, vol. 3, no. 1, pp. 38–45, 2004.*]

Big Data: Confusione

- Indice dei prezzi al consumo: PriceStats

Usa solo prezzi online



Big Data: *Correlazione*

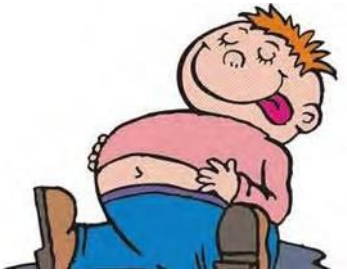
- Rinunciare alla causalità in favore della correlazione
  - Non conta sapere perché (*why*) vendo un libro online, ma cosa (*what*) fa aumentare le vendite
    - In previsione di un uragano aumentano le vendite di torce elettriche, ma anche di merendine e dolci

Big Data: *Correlazione*

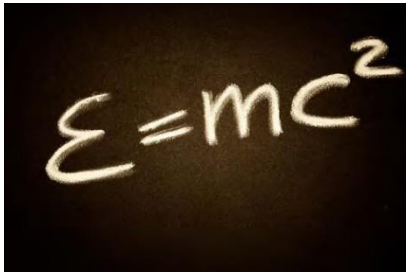
- Rinunciare alla causalità in favore della correlazione
  - La dimostrazione di una causalità è molto più costosa della individuazione di una correlazione.

Big Data: *Correlazione*

Esempio: il peso dei bambini di scuola elementare è correlato positivamente al quoziente intellettivo



peso ↔ QI



Facile da scoprire.

Ma direste che mangiare fa aumentare il QI?

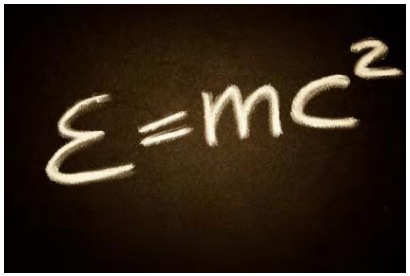
O che il QI influisce sul peso?

Big Data: *Correlazione*

Esempio: Se tenessimo sotto controllo il fattore età giungeremmo a conclusioni diverse.

età

peso QI

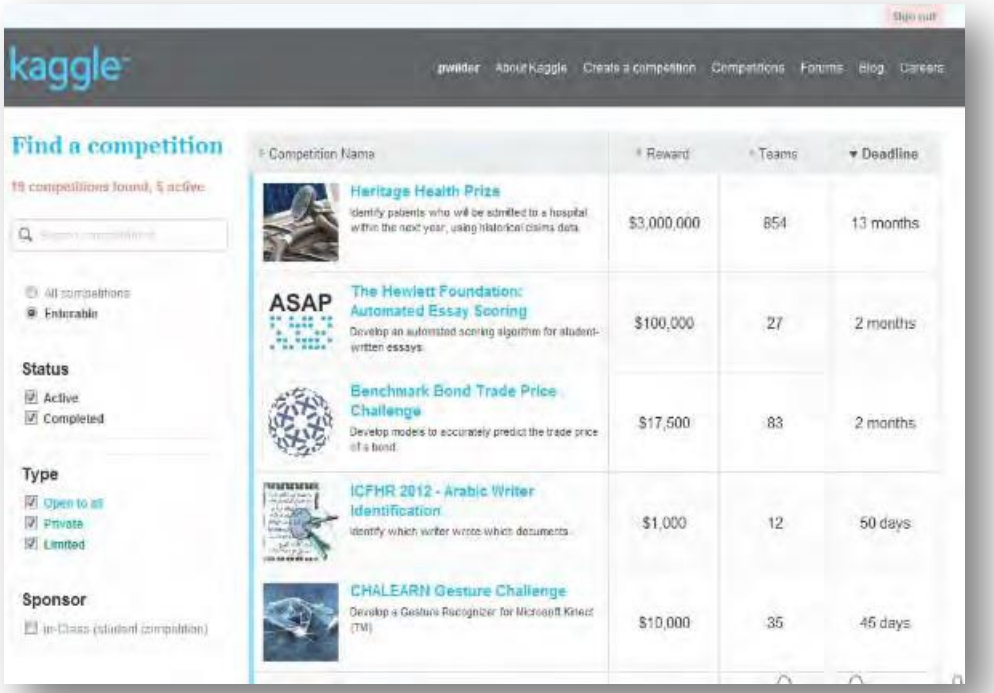


Tenere sotto controllo un singolo fattore costa

… e non sempre è possibile

Big Data: *Correlazione*

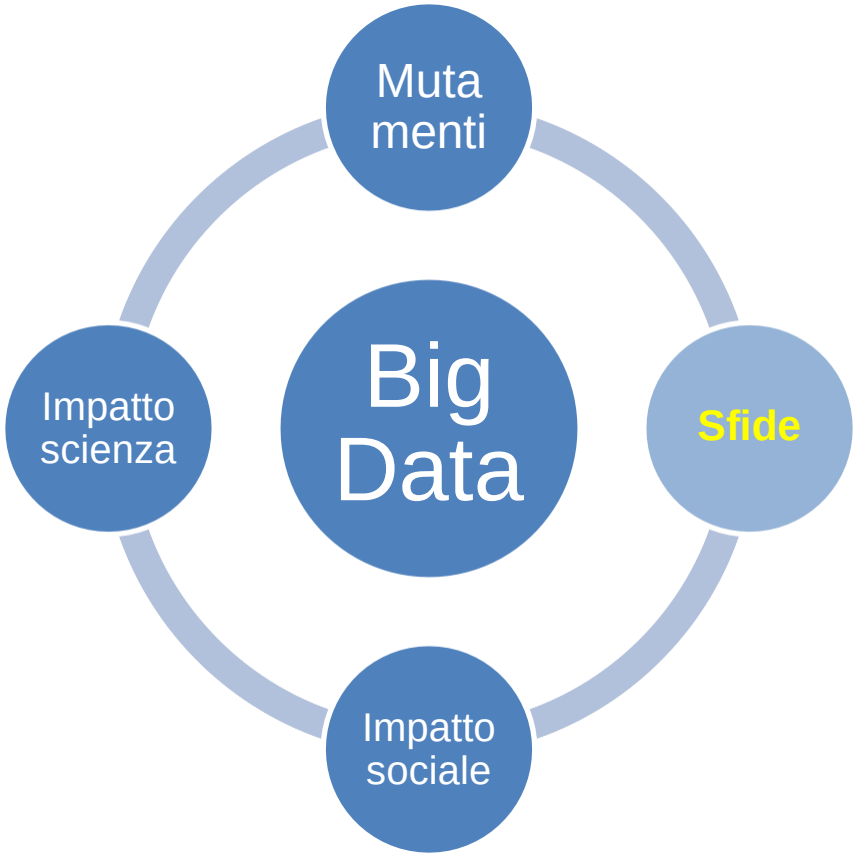
- Se scopriste che le auto usate arancioni sono meno soggette ad avere difetti, che auto usata comprereste?
- Vi porreste il problema di spiegare perché avete fatto quell'acquisto?



Big Data*: Correlazione*

- Avvertimento: Avere una gran quantità di dati a disposizione non significa saper comprendere la realtà.
- Le correlazioni ci dicono cosa, ma non perché.
- I Big Data ci aiuteranno a individuare il colore che andrà di moda il prossimo anno, ma non sono in grado di spiegarci perché.

Di cosa parleremo?



Caratteristiche: *Volume*

- Principali caratteristiche dei Big Data
  - Volume: dimensione dei data set (oltre le capacità degli odierni DBMS – *Data Base Management Systems*)

Caratteristiche: *Velocità*

- Principali caratteristiche
  - Velocità: rapidità con cui i dati arrivano e devono essere elaborati
    - Real-time o , almeno, near-time
    - Spesso in stream

    Non c'è tempo per importare i dati in un DBMS per forzarne una rappresentazione uniforme.

Caratteristiche: *Varietà*

- Altre caratteristiche
  - Varietà: tipologia di dati e sorgenti
    - Semi-strutturati (XML, tweets, …)
    - Destrutturati (documenti, pagine web)

    Scarsa adattabilità alle restrizioni dei DBMS moderni

Caratteristiche: *Veridicità*

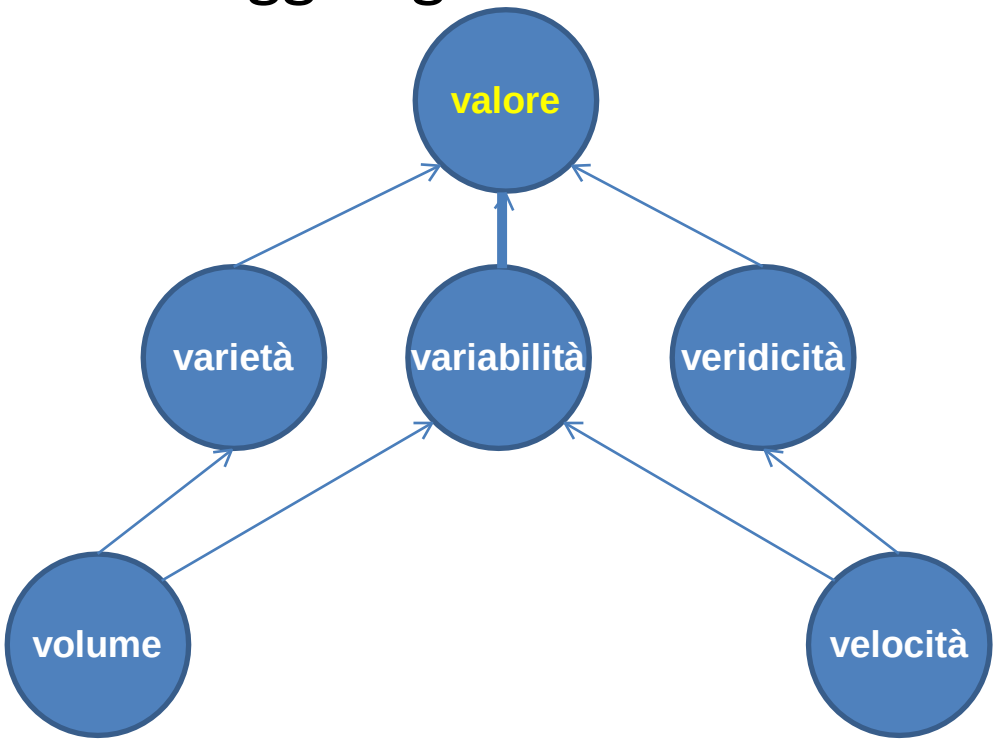
- Altre caratteristiche
  - Veridicità:
    - Le sorgenti dei dati sono non controllate e/o controllabili.
    - C'è incertezza sulla singola informazione
      - Incompleta, vaga, …

Caratteristiche: *Variabilità*

- Altre caratteristiche
  - Variabilità: ci sono variazioni sia nella struttura dei dati che nella semantica sottostante;
  - Insito nella datizzazione (es. le nostre parole sono dati)

Caratteristiche: *Valore*

- Altre caratteristiche
  - Valore: potenzialità dei dati in termini di vantaggi competitivi raggiungibili con la loro analisi



Caratteristiche: *Valore*

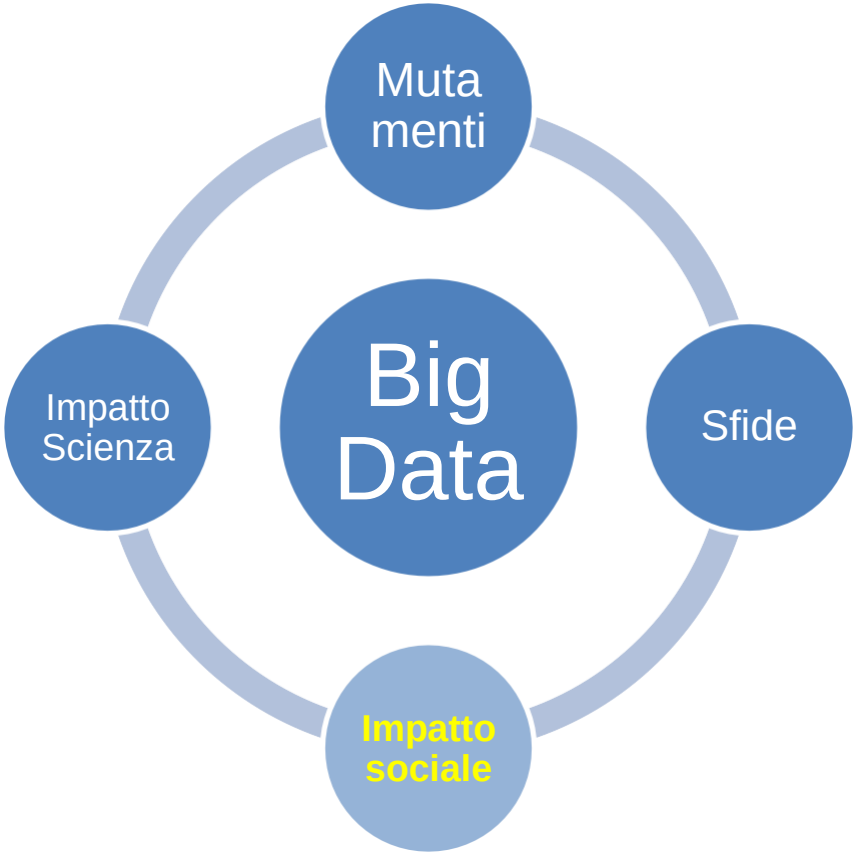
- I Big Data dovrebbero creare “valore”
  - Scoprendo esigenze, aiutandoci a migliorare le performance di una organizzazione
  - Segmentando meglio la clientela
  - Rimpiazzando/supportando i decisori umani con algoritmi
  - Innovando i nuovi modelli / servizi aziendali

Caratteristiche: *Valore*

- Chi dovrebbe beneficiare del “valore” creato con i Big Data?
  - Le imprese
  - La comunità
  - Il singolo cittadino

Dovrebbe valere il principio che “chi genera dati” ne deve beneficiare in primis.

Di cosa parleremo?



Quali limitazioni a uno sviluppo economico data-driven?

- Non sono solo tecnologiche
- Innanzitutto c'è una **carenza di talenti** necessari affiché le organizzazioni possano avvantaggiarsi dei Big Data.
- Occorrono competenze specifiche
  - statistica, apprendimento automatico, data mining
  - ma anche manager e analisti in grado di cogliere il valore nei dati

(Sorgente: McKinsey Global Institute)

Il lato oscuro dei Big Data

- Con i Big Data si affaccia l'incubo della distopia orwelliana del "Grande fratello" (George Orwell, *1984*, 1949).
- I Big Data rendono antiquato il concetto di privacy

Il lato oscuro dei Big Data

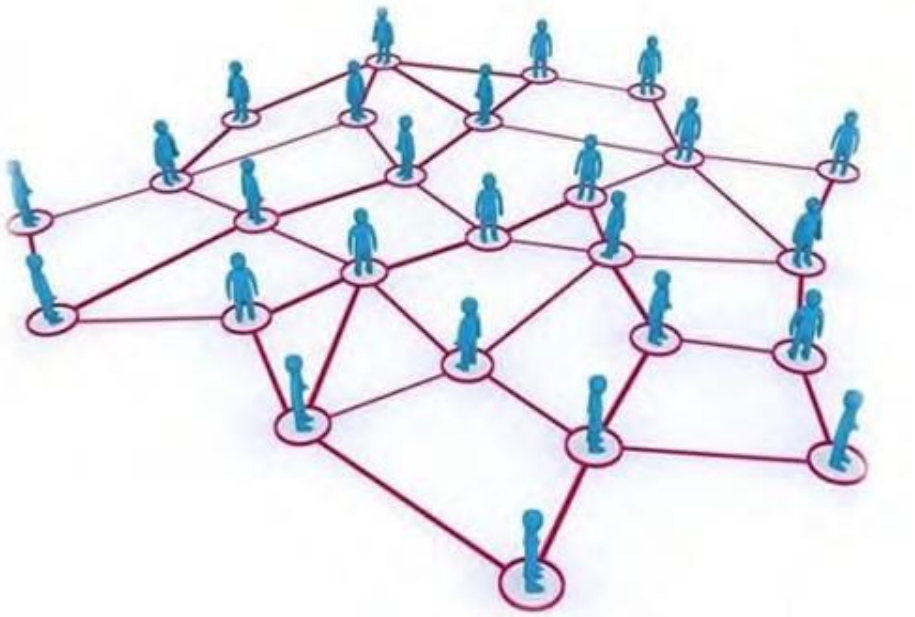
- Esempio: la deCODE Genetics ha raccolto le caratteristiche genomiche di 2636 islandesi
- Il sequenziamento più vasto mai effettuato su una popolazione
- Ha ottenuto informazioni su geni che predispongono a malattie neurodegenerative, cardiocircolatorie, etc.
- Tali informazioni possono essere utilizzate, ad esempio, per personalizzare i premi delle polizze assicurative.

Il lato oscuro dei Big Data

- Esempio: l'adozione di strumenti di scatole nere o black box (*event data recorder*) da installare sugli automezzi per avere in cambio sconti sulle polizze assicurative.

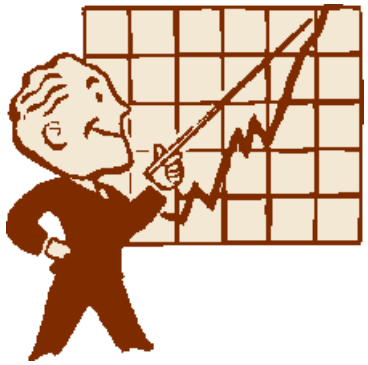
I Big Data per il bene comune

- I big data ci danno anche la possibilità di osservare la rete delle relazioni sociali e dei movimenti, mettendo a nudo il tessuto sociale in cui siamo immersi.

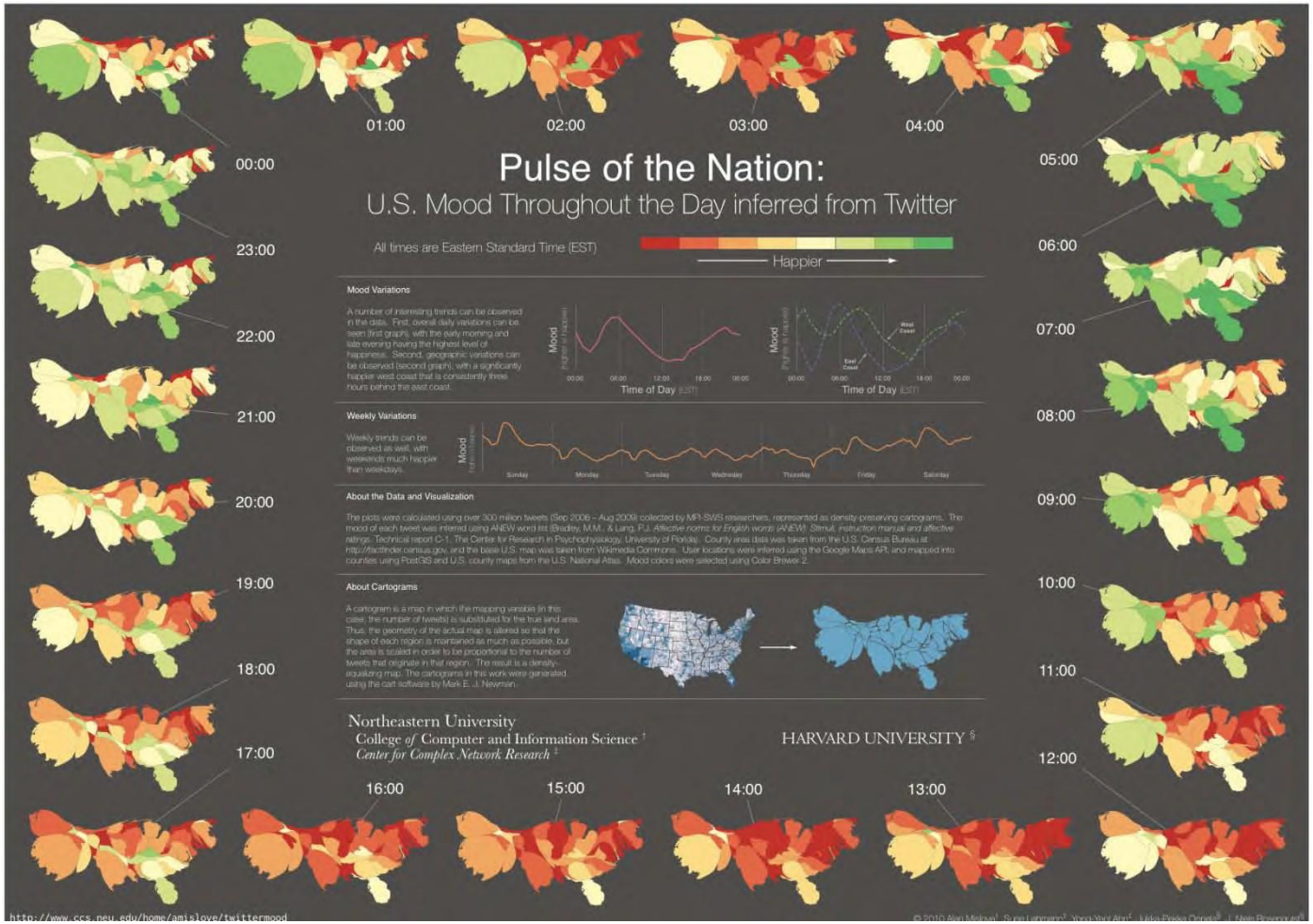


I Big Data per il bene comune

- I Big Data possono essere utilizzati per migliorare la democrazia
  - Offrendo informazioni basate sull'evidenza
  - Consentendo di verificare il successo di una politica pubblica.

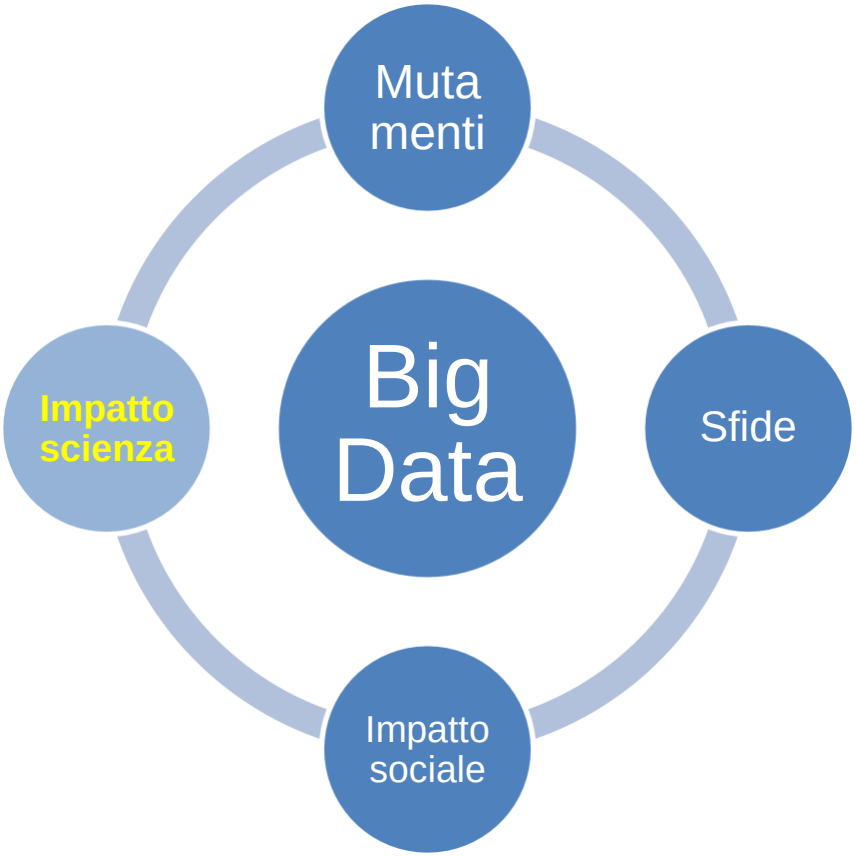


I Big Data per il bene comune



https://mislove.org/twittermood/

Di cosa parleremo?



Un nuovo paradigma scientifico?

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database / files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$

Conclusioni

- Big Data ha un enorme potenziale
  - per la società

    democrazia, cultura, sport, …
  - per l'economia

    il 91% delle Fortune 100 companies hanno almeno una iniziativa big data in corso
  - per la scienza
    - nuovo paradigma, genomica,…

Conclusioni

In un mondo guidato dai dati,
che spazio resta per le persone?

Conclusioni

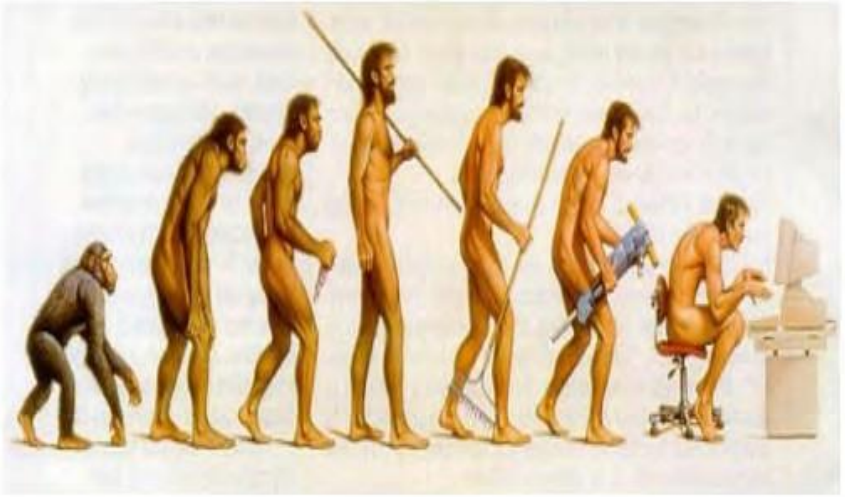
- I Big Data sono solo uno strumento: servono a informare più che a spiegare.

- Il più grande apporto degli esseri umani sta proprio in quello che gli algoritmi non rivelano e che non possono rivelare, perché non è incapsulato nei dati.

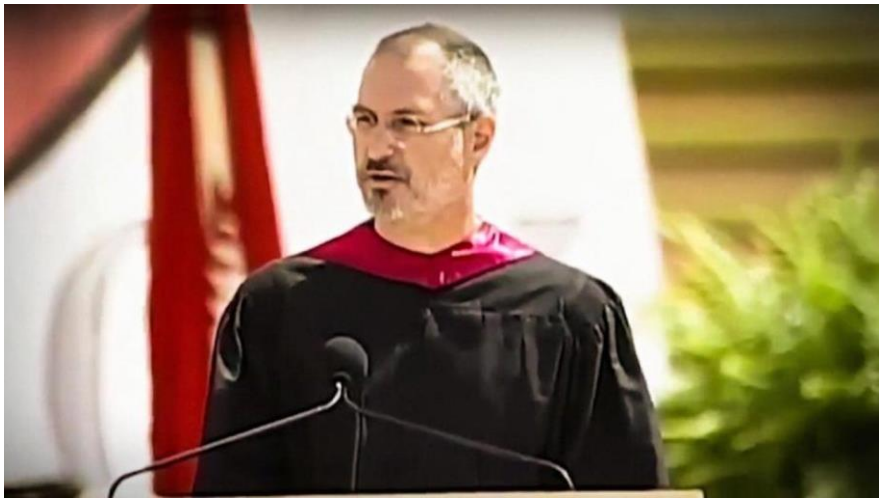
- Lo spazio per l'uomo sta nel «ciò che non c'è», nel «non detto», nel «non ancora pensato».

Conclusioni

La scintilla dell'invenzione sta in quello che non dicono i dati.



In un mondo dominato dai Big Data, dovremo far leva sulle nostre caratteristiche più specifiche: creatività, intuito, ambizione intellettuale.



Non potete sperare di unire i puntini guardando avanti, potete farlo solo guardandovi alle spalle: dovete quindi avere fiducia che, nel futuro, i puntini che ora vi paiono senza senso possano in qualche modo unirsi.
……

E l'unico modo di fare un gran bel lavoro è amare quello che fate. Se non avete ancora trovato ciò che fa per voi, continuate a cercare, non fermatevi, come capita per le faccende di cuore, saprete di averlo trovato non appena ce l'avrete davanti. E, come le grandi storie d'amore, diventerà sempre meglio col passare degli anni. Quindi continuate a cercare finché non lo trovate. Non accontentatevi

(Steve Jobs, Discorso all'Università di Stanford, 12 giugno 2005)