Il contesto: i dati

I dati del Web:

un reticolo caotico e globale di informazioni in continua espansione che aspira a diventare l'infrastruttura mondiale della conoscenza

Il contesto: i dati

I dati delle biblioteche:

un ecosistema ricco di informazioni preziose, ma in contrazione e separato dalla Rete globale

IL DILUVIO DEI DATI

Cosa sono i Big Data

FABIO DI GIAMMARCO

Diluvio di informazioni

- **Dopo l'invenzione della stampa le informazioni in Europa impiegavano 50 anni a raddoppiare**
  - **dal 1453 al 1503 furono stampati in Europa circa 8 milioni di libri, più di tutti quelli che avevano prodotto gli amanuensi europei nei 1200 anni precedenti.**
  - **Oggi raddoppiano in 3 anni**
- **Informazioni immagazzinate in un anno nel mondo**
  - **Nel 2007 è stato calcolato che sono stati archiviati circa 300 exabyte di dati al mondo**
    - **Un film corrisponde circa a un gigabyte, un exabyte è un miliardo di gigabyte**
  - **Nel 2013 la quantità di informazione immagazzinata al mondo è stata di circa 1200 exabyte, dei quali meno del 2% in formato non digitale.**

La «spinta» dell'innovazione tecnologica

- **Negli ultimi cinquant'anni il costo dell'archiviazione digitale si è dimezzato**
- **mentre la densità dei dati in memoria è aumentata di circa 50 volte**
  - **Nel 1969 la memoria del computer di bordo dell'Apollo 11 era di 4 Kb**
    - **4 Kb corrispondono a circa 2 pagine di testo**

La «spinta» dell'innovazione tecnologica

- **L'innovazione tecnologica cambia il mondo:**
  - **Il telescopio ha consentito di esplorare l'universo**
  - **Il microscopio ha permesso di scoprire i batteri**
  - **L'Information Technology ha rivoluzionato la società con Internet, il Web, i Social Media, i Big Data**

Dove i dati hanno cominciato a diventare «Big»

- **Astronomia**
  - **il telescopio del New Mexico nell'ambito della Sloan Digital Sky Survey ha raccolto nel corso dell'anno 2000 più dati di quanti ne erano stati accumulati nell'intera storia dell'astronomia**
- **Genoma**
  - **Vengono sequenziate 3 miliardi di coppie di basi che formano il genoma umano**
- **Finanza**
  - **A Wall Street ogni giorno sono contrattate circa 7 miliardi di azioni di cui i 2/3 attraverso algoritmi**

Comparsa e diffusione del termine «Big Data»

Nel 2011 comincia a circolare sistematicamente per indicare un settore del mercato dell'informatica che mira alla gestione di enormi archivi digitali

Nel 2013 si diffonde sulla stampa italiana a seguito del «datagate», vale a dire le rivelazioni dell'ex CIA Edward Snowden sulla raccolta massiva e indiscriminata di informazioni digitali effettuata dal NSA (National Security Agency)

Big Data, definizioni...

- **Non esiste una definizione rigorosa e universalmente accettata…**
  - ***Termine usato per descrivere una raccolta di dati così estesa in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore [Wikipedia]…***
  - ***Si parla di Big Data quando si ha un dataset talmente grande da richiedere strumenti non convenzionali per estrapolare, gestire e processare informazioni entro un tempo ragionevole***

Big Data: la regola 3V + 1V

- **La definizione di big data è anche legata alle tre caratteristiche che il dataset deve avere**
  - **Volume: nel dataset devono essere presenti grandi quantità di dati, nell'ordine degli zettabyte**
  - **Varietà: i dati provengono da fonti eterogenee, quindi hanno natura diversa e non strutturata. Oltre che valori numerici possono essere comprese immagini, parole, video, etc.;**
  - **Velocità: si riferisce alla velocità con cui i dati vengono generati, svolgendo l'analisi in tempo reale.**
- **Con l'aggiunga di una quarta V, il Valore/Veridicità**
  - **Dati devono essere affidabili e capaci di fornire analisi utili e interessanti**

Big Data vs Small Data: definizione «Small»

- **Small Data corrisponde a una quantità di dati abbastanza piccola per essere gestita da un essere umano**
  - **I dati sull' uso dell'energia da parte di un nucleo familiare**
  - **quelli sugli orari degli autobus o sulla spesa pubblica sono tutti ''small data''**
  - **Tutto quello che può essere elaborato tramite Excel può essere definito ''small data''.**

L’azienda paradigmatica nello sfruttamento dei Big Data: Google

- **Google Translate**
- **Google Flu trends (GFT)**
- **La forza dei Google Big Data**

Google Translate

- **Nel 2006 Google entra nel business della traduzione automatica**
- **Invece di puntare come fatto in precedenza su algoritmi performanti, punta sui dati**
- **Crea un sistema basato sull'acquisizione di tutti i testi possibili tra cui la scansioni dei libri del progetto Google Books**
- **Raccolti più dati possibili, tratta il linguaggio come una massa caotica di dati a cui applicare il calcolo delle probabilità**
- **Google Translate supporta oltre 100 lingue e serve oltre 200 milioni di persone al giorno (dati 2013)**

Google Flu trends (GFT)

- **Servizio creato da Google per monitorare, attraverso l'analisi delle ricerche effettuate dagli utenti, il diffondersi in tempo reale delle sindromi influenzali, cercando di anticipare le forme di diffusione dell'epidemia**

- **L'idea alla base è che attraverso il monitoraggio di milioni di query di ricerca fatte su Google riguardanti sintomi o rimedi per l'influenza sia possibile rivelare se è in atto una diffusione epidemica di infezioni influenzali in una popolazione**

La forza dei Google Big Data

- **Google è stato in grado di processare ben 450 milioni di modelli matematici per testare le queries sull'influenza e poi confrontare le proprie previsioni con la casistica dei CDC (Centers for Disease Control and Prevention) degli anni precedenti**

- **E poi a mettere a punto un software con una combinazione di 45 parole-chiave che quando venivano impiegate insieme a un modello matematico generavano una forte correlazione tra la loro previsione e i dati ufficiali relativi al territorio nazionale**

Le “Big Player» del mondo digitale: da aggregatori, fornitori di servizi, rivenditori on line,

a “Big Data Company»

- **Google processa oltre 24 petabyte di dati al giorno: un volume pari a mille Library of Congress ogni giorno**

- **Amazon Big Data sul comportamento d’acquisto di circa 152 milioni di clienti on line**
  - **Ancora Amazon Big Data per monitorare, tracciare e proteggere i suoi 1,5 miliardi di articoli gestiti attraverso i 200 centri logistici sparsi in tutto il mondo**

Le “Big Player» del mondo digitale: da aggregatori, fornitori di servizi, rivenditori on line,

a “Big Data Company»

- **Su Facebook si caricano ogni ora oltre 10 milioni di fotografie**
  - **Datizzazione delle relazioni**
    - **1miliardo di utenti che interagiscono attraverso oltre 100 miliardi di amicizie = grafico sociale che copre il 10% della popolazione mondiale (dati 2012)**

- **Su Twitter si pubblicano oltre 400 milioni di tweet al giorno**
  - **Datizzazione dei sentimenti**
    - **Analisi 509 milioni di tweet inviati da 2,4 milioni di persone nell’arco di 2 anni**
      - **Registrazione stati d’animo quotidiani e settimanali**

La ricerca Big Data dei «Big Player»

- **I Big Data sono importanti per mantenere alto il livello di competitività**
- **I Big Data sono importanti perché consentono di creare nuovi prodotti e/servizi**
- **Inoltre, i Big Data utilizzati dalle imprese commerciali spesso possono essere usati per rispondere a domande che interessano non solo il marketing, ma anche l'economia, la sociologia, la politica:**
  - **I temi non commerciali, possono toccare: identità, comportamenti sociali, atteggiamenti culturali ecc.**
- **Ricerche avanzate si svolgono:**
  - **Microsoft Research Labs**
  - **Yahoo! Labs**
  - **Google Research**
  - **Facebook data science**

Big Data: l'idea di leggere il mondo in maniera diversa

Non più una successione di eventi, ma piuttosto un universo (caotico) composto da informazioni

Big Data: l'era dell'incertezza

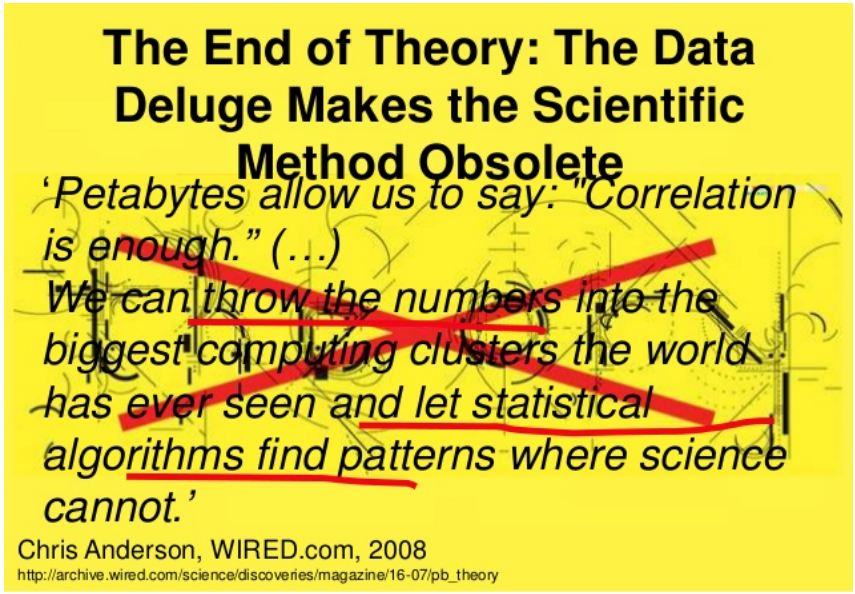
- **L'ossessione per l'esattezza è un costrutto dell'era analogica caratterizzata da una costante carenza di informazioni (small data)**

- **Rinuncia dell'esattezza a livello micro per comprensione generale (trend) dei fenomeni a livello macro: i Big Data trasformano le cifre in qualcosa di probabilistico piuttosto che di preciso**

- **Per massimizzare i benefici dei Big Data dobbiamo accettare un certo grado di confusione come ineludibile e non come un elemento di disturbo da eliminare**

L'era dell'incertezza... ma predittiva

Nel 2008, Chris Anderson – giornalista e saggista statunitense – pubblica sulla rivista Wired un articolo dove dichiara la «fine della teoria» in quanto il diluvio di dati reso possibile dai Big Data renderebbe obsoleto il metodo scientifico tradizionalmente inteso



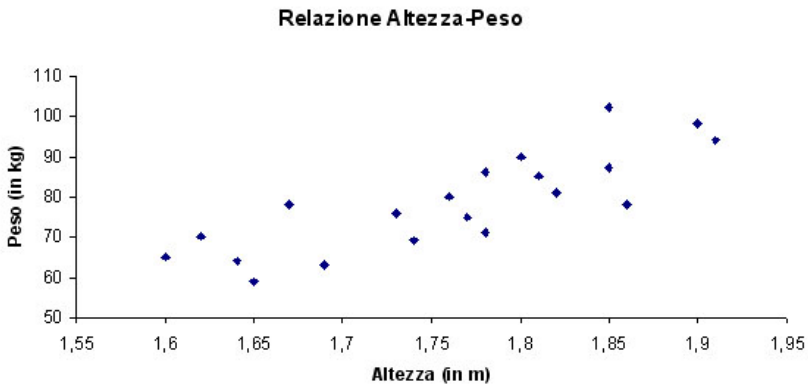
L'era dell'incertezza... ma predittiva

"lasciare cadere le proprie ossessioni sui fenomeni di causalità in cambio di correlazioni semplici in quanto in futuro interesserà sempre di più sapere il cosa succede e non il perché succede……"

Correlazione

- **Per correlazione si intende una relazione tra 2 variabili statistiche tale che a ciascun valore della prima variabile corrisponda «con una certa regolarità» un valore della seconda**

- **Non è un rapporto di causa-effetto, ma semplicemente la tendenza di una variabile a variare in funzione di un'altra**



L’era dell’incertezza… ma predittiva

L’idea alla base della «fine della teoria» e del libro di Viktor Mayer-Schonberger e Kenneth Cukier è che la possibilità di scoprire, attraverso i Big Data, un gran numero di correlazioni tra i dati, renderebbe possibile la scoperta di processi nascosti con la probabilità di estrapolare informazioni rispetto al futuro con una certa affidabilità.

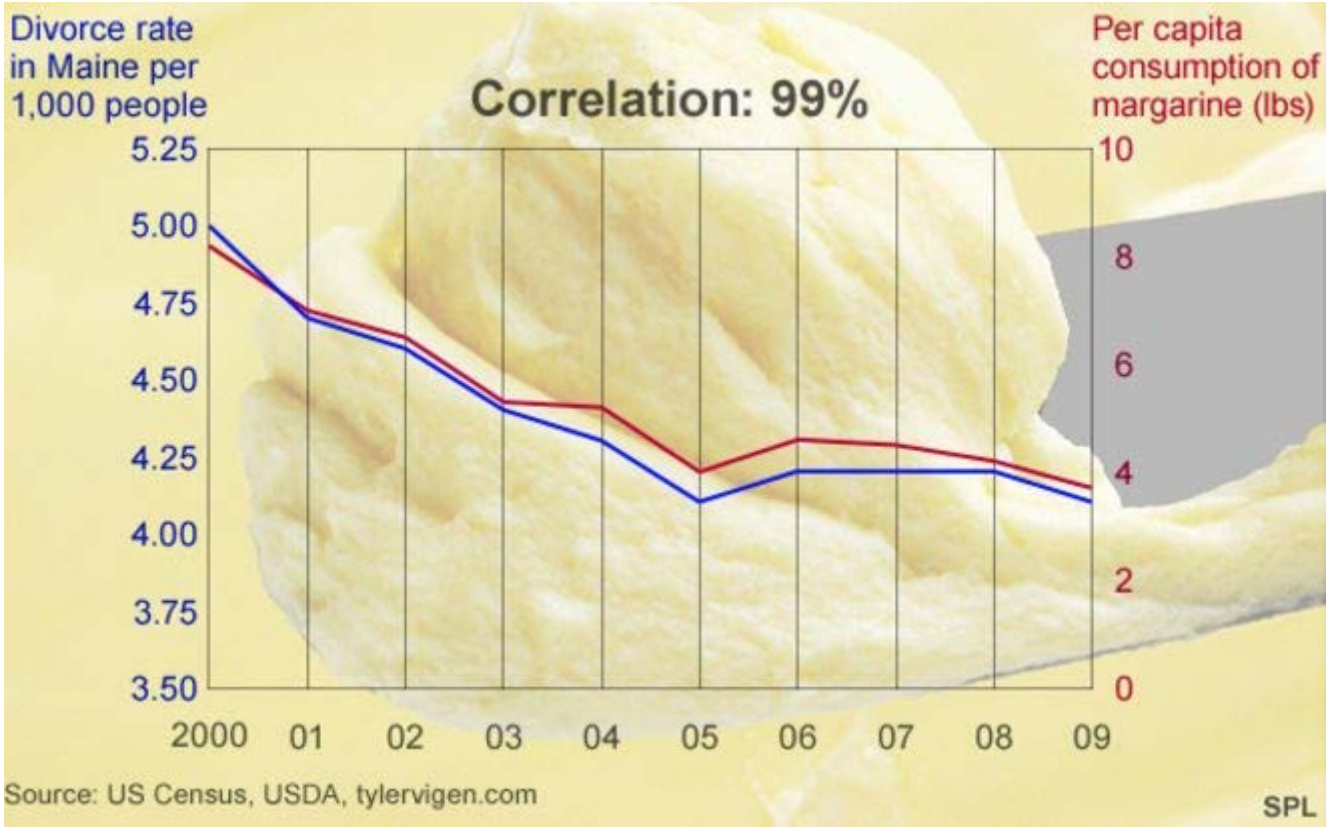
“Big Data hubris” (arroganza dei Big Data)

- **Il fatto è che ormai è relativamente facile usare enormi quantità di dati per far emergere accattivanti correlazioni**

- **Al contrario è invece molto più difficile trasformare queste correlazioni in qualcosa di scientificamente valido**

- **Il problema è che la maggior parte dei dati che vengono processati è danno certe correlazioni come output potrebbero non avere all’origine una valida impostazione di ricerca scientifica.**

Il consumo di margarina può causare il divorzio: esempio di falsa correlazione



“Big Data hubris” l’esempio Google Flu Trends

- **Le correlazioni si sono basate sulle interrogazioni di Google fatte a seguito del clamore delle notizie sull’influenza.**
- **Di conseguenza il numero crescente delle persone interessate non corrispondeva al numero reale delle persone infette.**
- **Risultato: Google Flu Trends ha avuto risultati discordanti rispetto al conteggio reale effettuato dai Centri di controllo della malattia sul territorio.**

I Big Data come gli astronomi Maya?

Rischiamo di utilizzare i Big Data come i Maya utilizzarono i loro dati astronomici? Cioè, cercando regolarità nelle sterminate sequenze di dati senza un metodo scientifico?

Big Data, ma con metodologia

- **Nel 2012, una ricerca di Danah Boyd e Kate Crawford "Critical questions for big data", si incaricava di ridimensionare i "Big Data fondamentalismi" confutando che i Big Data *"possano offrire di per sé una più alta forma di intelligenza e conoscenza dando vita a intuizioni e/o rivelazioni prima impossibili"***

- **L'accumulazione dei dati può rappresentare una preziosa fonte di informazione, ma senza una comprensione teorica, senza una metodologia è impossibile aggiungere tridimensionalità alle interpretazioni. In pratica, è impossibile studiare e capire i dati.**

Un altro Big Data – Rischio: la privacy

- **I Big Data possono rendere inefficaci i principali meccanismi tecnici e legali con cui oggi tentiamo di tutelare la privacy**
  - **I Big Data scavalcano il «Consenso informato»**
    - **Finora le leggi che tutelano la privacy hanno lasciato il controllo del rischio agli individui consentendo loro di decidere se come e da parte di chi potevano essere processate le proprie informazioni personale**
- **Ora con i Big Data il grosso del valore dei dati si è spostato sull'utilizzo secondario, cioè su utilizzi che neanche si potevano immaginare al momento della raccolta**
  - **Per questi motivi deve essere pensato un sistema di tutela della privacy meno focalizzato sul consenso individuale e più sugli utilizzatori dei dati, cioè una "privacy tramite responsabilizzazione" di coloro che acquisiscono il diritto di sfruttare i dati.**

Un altro Big Data – Rischio: le propensioni individuali

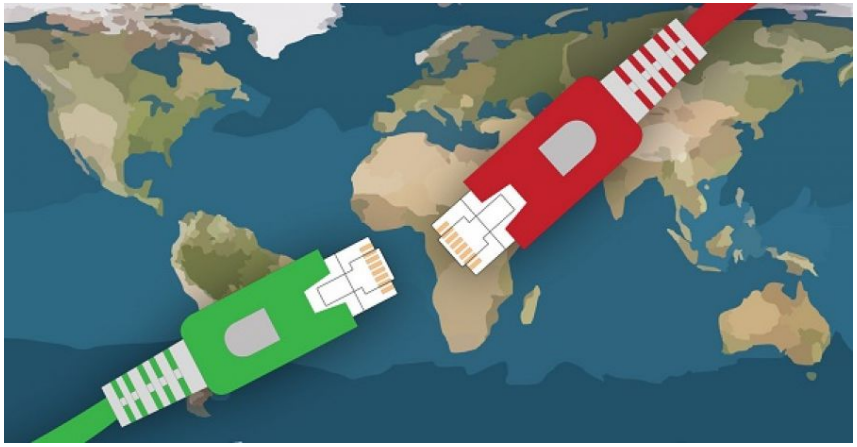
- **Nel 2054 la città di Washington ha cancellato gli omicidi grazie a un sistema chiamato *Pre-crime* la polizia riesce a impedire gli omicidi prima che essi avvengano e ad arrestare i potenziali "colpevoli". In questo modo non viene punito il fatto (che non avviene), bensì l'intenzione di compierlo**

- **L'elaborazione dei Big Data può fornire anche giudizi probabilistici sui comportamenti umani**

- **Dovranno prevedersi forme di garanzia rispetto al fatto che i giudizi sulle persone non possano essere formulati sulla base di elaborazioni probabilistiche di grossi volumi di dati**

Definire nuovi diritti per l'era digitale

La soluzione generale è definire nuovi diritti per l'era digitale, come ad esempio:

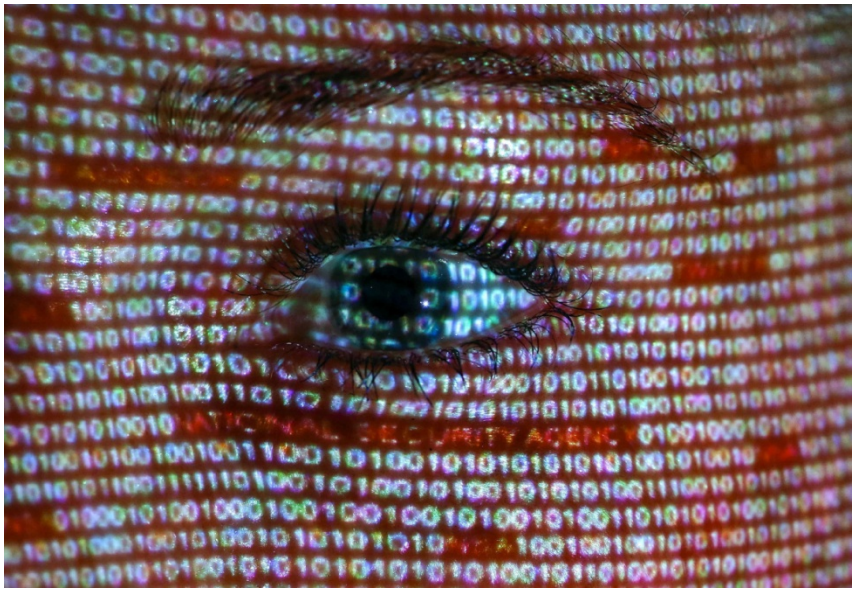


- **Il diritto all'oblio**
  - **Essere cancellati dai database dei motori di ricerca**
  - **Essendo un diritto emergente ha bisogno di essere spinto da una scelta politica**
  - **Come nel caso della Corte Europea che ha imposto a Google la rimozione dei contenuti indicizzati**

Datification

Trasformare i fenomeni della vita in dati informatici

- **convertire qualsiasi fenomeno in una forma quantitativa in modo da poterlo:**
  - **Tabulare**
  - **Analizzare**

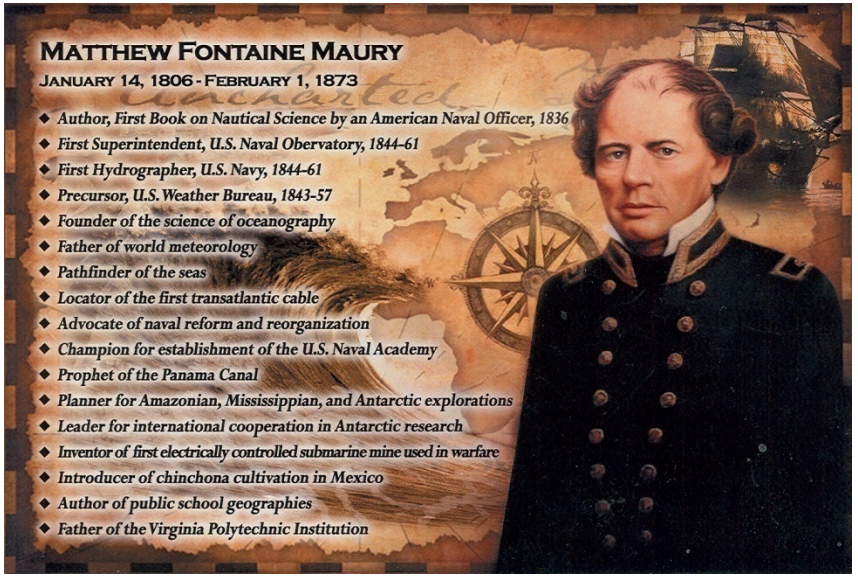


Le origini della Datification: i dati della navigazione

- **Matthew Fontaine Maury (1806-1873) creatore del sistema di rotte che è rimasto in uso per tutto il XX secolo**

- **Sfruttando tutti i dati che le navi e i loro equipaggi nel tempo avevano raccolto e trascritto sui loro giornali di bordo riuscì a costruire, con pochissimi mezzi di calcolo, per la prima volta nella storia, una mappa dei venti, delle temperature e delle correnti di tutti i mari e gli oceani, regalando all'umanità un nuovo strumento per la navigazione più sicura, su cui costruire una nuova oceanografia**

- **Ovvero: prendere un contesto (un fenomeno) considerato fino a quel momento privo di valore informativo e quantificarlo estraendo in questo modo un insieme di dati da utilizzare per scopi completamente nuovi**

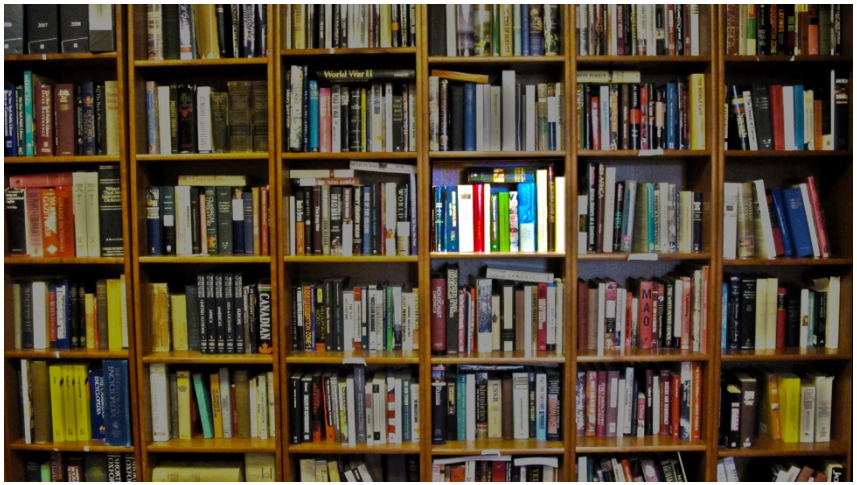


La trasformazione delle parole (dei libri) in dati: la datafication di Google Books

- **La digitalizzazione massiva mette le ali alla datizzazione permettendo di estrarre, da grandi quantità di dati, valori fino a quel momento rimasti occulti: trasformando ad esempio in dati le parole racchiuse in milioni di libri**

- **Google Books è anche un esempio paradigmatico di "economia dell'abbondanza" tipica del mondo digitale, dove l'abbondanza (di informazioni e dati) eccede di gran lunga la domanda**

  - **All'opposto rispetto al modello economico d'abbondanza di Google ci sono le biblioteche che invece fanno parte di quell'economia della scarsità" tipica del mondo analogico, e quindi non «Big» ma "Small Data"**



La Datafication che si fa scienza: la Culturomica

- **Cos'è la culturomica? Studiare la storia, la cultura, il comportamento umano mediante analisi quantitative dei testi digitalizzati: l'applicazione dei Big Data alla cultura umana**

- **Il termine è la traduzione italiana di un neologismo inglese apparso per la prima volta nel 2010 in un articolo della rivista *Science* dal titolo *Quantitative Analysis of Culture Using Millions of Digitized Books*, firmato da diversi ricercatori dell'Università di Harvard**

- **I ricercatori di culturomica estraggono dati da archivi digitali per investigare come i fenomeni culturali vengono riflessi nel linguaggio e nell'utilizzo delle parole**

“Quantitative Analysis of Culture Using Millions of Digitized Books”

“Abbiamo costituito un corpus di testi digitalizzati che comprende circa il 4% di tutti i libri mai stampati. L’analisi di questo corpus ci permetterà di indagare dal punto di vista quantitativo le tendenze culturali.

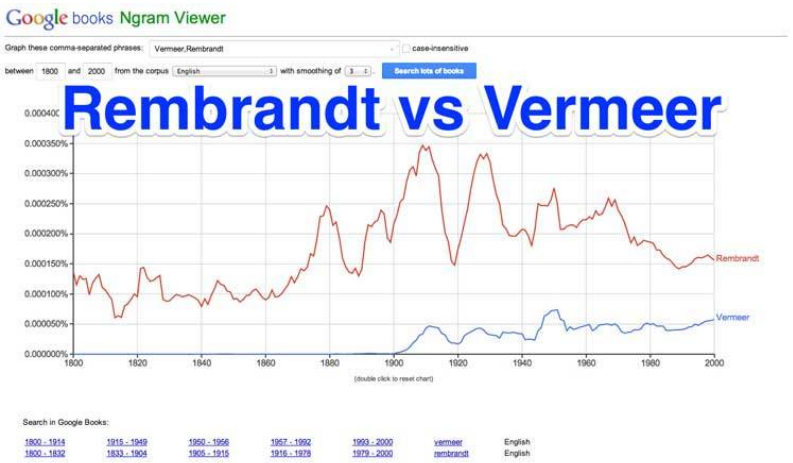
Abbiamo esaminato il vasto campo della “Culturomica”, e ci siamo concentrati su aspetti linguistici e culturali: un fenomeno che ha riflessi nella lingua inglese tra il 1800 e il 2000.

Vogliamo mostrare come questo approccio possa fornire interessanti approfondimento in diversi campi: lessicografia, grammatica, memoria collettiva, sviluppo della tecnologia, censura, epidemiologia storica.

La Culturomica può estendere i confini di questi campi incrociando fenomeni concernenti le scienze sociali e gli studi umanistici.”

Un esempio di applicazione della culturomica: Ngram Viewer di Google Books

- **Sito che genera un grafico dell'andamento nel tempo degli n-grammi**
  - **Stringhe di caratteri separate da uno spazio:**
    - **Mela, 1972: unigrammi**
    - **Umberto Eco: digramma**
    - **Martin Luter King: trigramma**
- **Utilizzando come fonte, i dati scansionati in (OCR) degli oltre 25 milioni di libri (dati ottobre 2015) del programma Google Books**



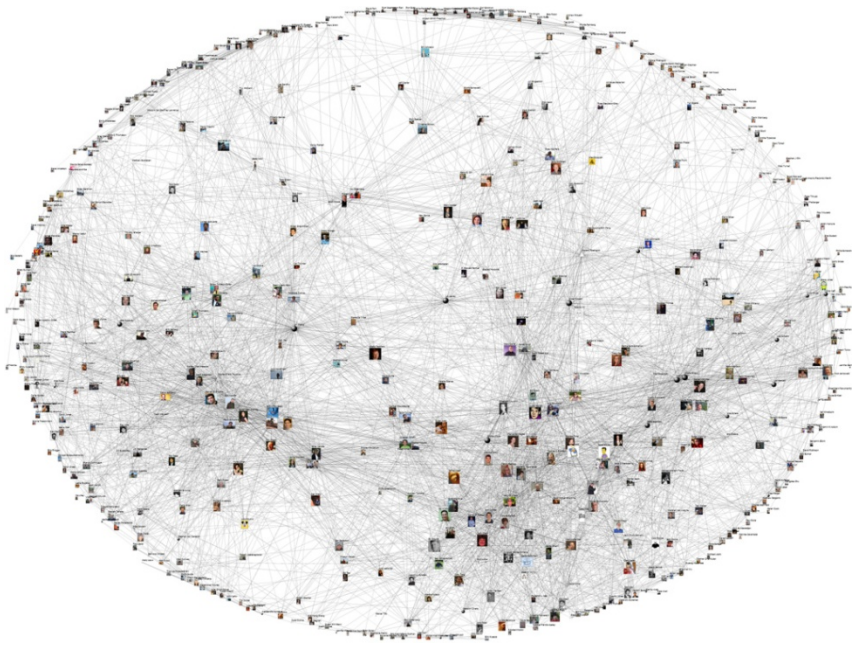
Datafication e raccolta dati: esempio dei social media

Esistono diverse strategie di raccolta dati sui social media:

- **Tramite API, cioè una serie di istruzioni informatiche per accedere a informazioni messe a disposizione sulle piattaforme dei social media**

- **Tramite servizi sviluppati da terze parti, ovvero strumenti che consentono di raccogliere dati ed elaborare semplici metriche per effettuare delle analisi**

- **Tramite Data Reseller, cioè l'acquisto dei dataset direttamente dalle piattaforme dei social media**



Datafication e raccolta dati: il modello Twitter

È leader nella fornitura di dataset da Big Data

- **Possiede un ecosistema di offerta dati completo**
  - **API**
  - **Terze Parti**
  - **Reseller**

- **Il suo flusso dati verso il Web ha un nome preciso: Firehose**

- **Per questi motivi è il social media più studiato tramite approcci computazionali**

- **Per queste sue caratteristiche è considerato un modello riguardo i Big Data originati dai social media**

Datafication

Valore dei dati

Riuso dei dati

Datafication

Il valore nascosto dei dati

- **Le informazioni sono – come dicono gli economisti - un bene "non competitivo" nel senso che se una persona li usa non impedisce ad un'altra di utilizzarli**

- **Nell'era dei Big Data, tutti i dati sono considerati preziosi, di per se, in quanto dati: il loro valore non diminuisce anche se vengono utilizzati**

I dati il nuovo "petrolio"

- **Studio della Commissione Europea sul valore del mercato dei dati nell'EU (aggiornato a dicembre 2016)**

- **Nel 2016, il valore è stato di 59,53 miliardi, in crescita rispetto al 2015 (54,3 miliardi)**

- **negli Stati Uniti è valutato dallo stesso report ben 129 miliardi di euro**

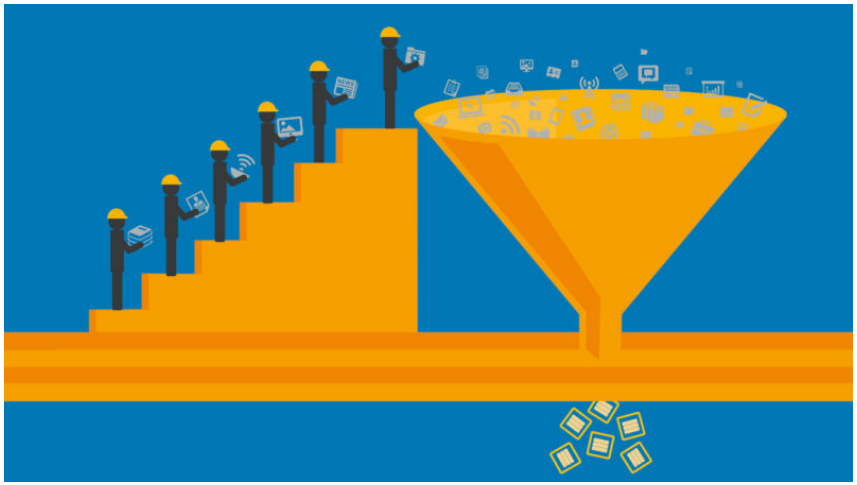
- **il suo valore nel 2020 potrebbe raggiungere, in uno scenario di massima crescita, i 106 miliardi di euro**

“Data exhaust”: il valore del riuso

“Data exhaust”:

- **Dati che si creano come sottoprodotto delle azioni svolte dagli utenti in rete**
  - **Molte aziende progettano i propri sistemi in modo che possano recuperare i dati residui e riciclarli**
    - **Google è il leader incontrastato anche in questo campo**

- **Il recupero dei dati residui è un meccanismo che sta dietro molti servizi come**
  - **Il riconoscimento vocale**
  - **I filtri anti-spam**
  - **La traduzione automatica**



I “Data exhaust” della lettura

Gli eReader rilevano enormi quantità di dati

- **Preferenze letterarie**
- **Abitudini di lettura**
- **Registrano sottolineature e annotazioni a margine**

- **Una volta aggregati tutti questi dati residui sono preziosi**
  - **Hanno un valore commerciale**
    - **Possono essere vendute alle case editrici tradizionali**
      - **Che possono utilizzarle per migliorare struttura e contenuto dei libri cartacei**

- **L’esempio dei dati raccolti dall’eReader Nook di Barnes & Noble**
  - **Ha rivelato che i lettori tendevano a lasciare a metà i libri di saggistica**
    - **L’azienda come risposta ha lanciato una nuova collana “Nook Snaps”: delle opere sintetiche su temi prioritari**

FlyOnTime.us:
un esempio “virtuoso” nel riuso dei dati

Informazioni interattive su incidenza delle condizioni meteorologiche sui ritardi dei voli in un determinato aeroporto (Stati Uniti)

- **Il sito combina informazioni sui voli e sul tempo raccolte da fonti ufficiali che sono liberamente disponibili e accessibili in Rete**