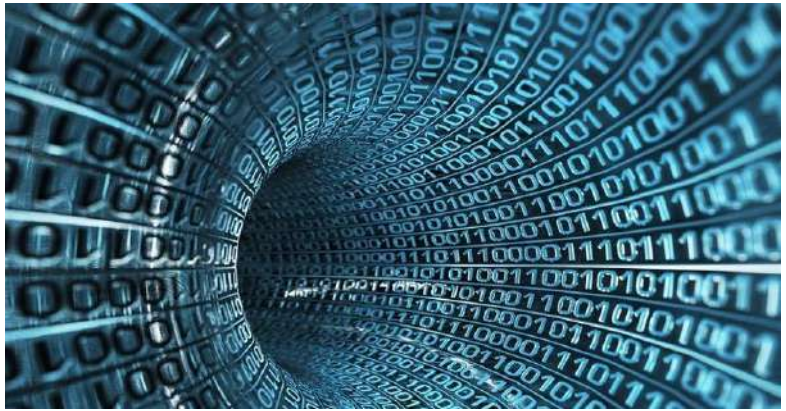
Sistemi informativi Evoluti e Big Data



Big Data e Data Science: concetti introduttivi

Prof. Devis Bianchini

Università degli Studi di Brescia

Dipartimento di Ingegneria dell'Informazione

UNIVERSITY OF BRESCIA

Big Data

Tante definizioni diverse

"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." -Teradata Magazine article, 2011

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." - The McKinsey Global Institute, 2012

"Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools." - Wikipedia, 2014

Quando i dati diventano “Big”

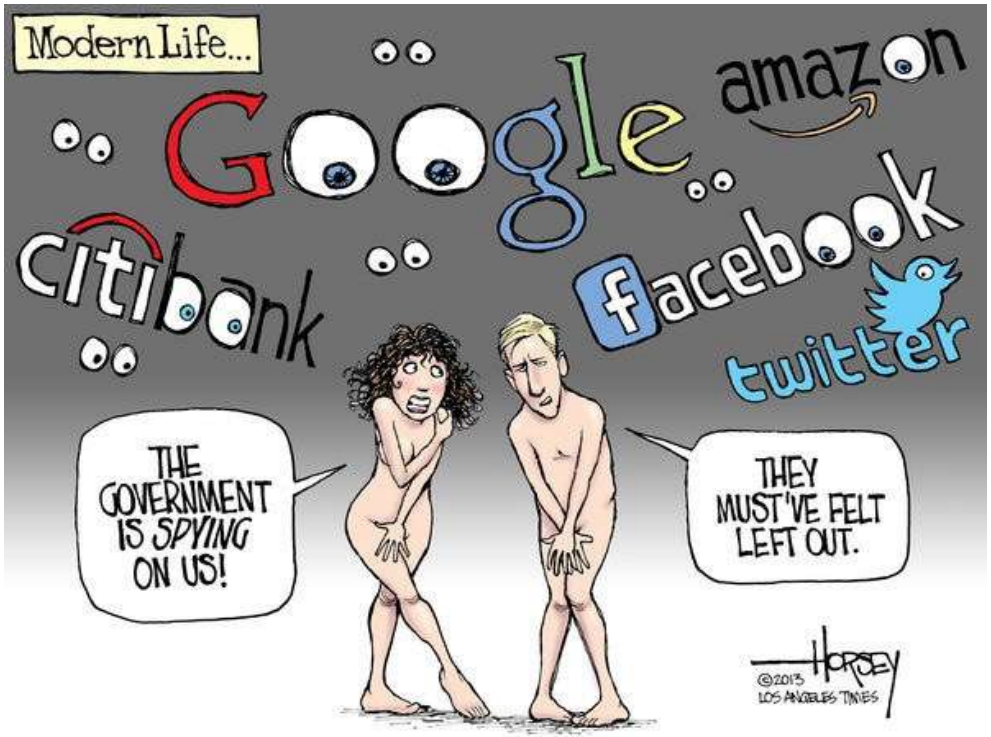
Come si è arrivati ai Big Data

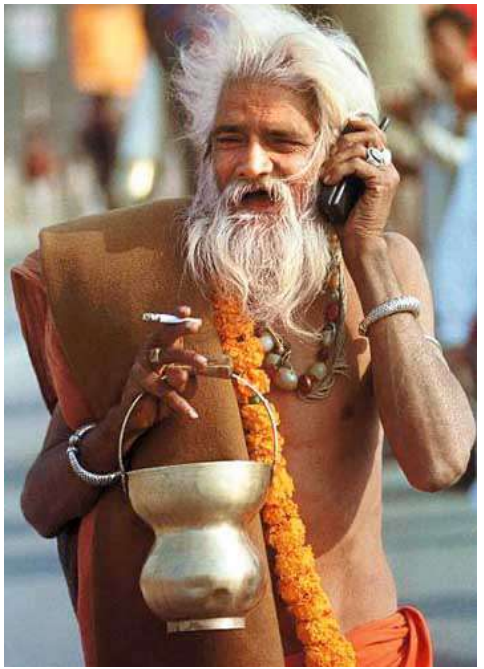
Tra le principali fonti di dati, che nel tempo hanno contribuito allo sviluppo del fenomeno dei Big Data, troviamo:

- Fonti operazionali (e.g., gestione della produzione, gestione degli acquisti, contabilità, gestione del personale, gestione dei clienti)
  - *in alcuni casi, i dati operazionali arrivano a creare dei volumi rilevanti*

- Sensori, DCS (Distributed Control Systems) e strumenti scientifici
- Dati non-strutturati e semi-strutturati provenienti da viarie fonti (per esempio, le applicazioni Web 2.0)

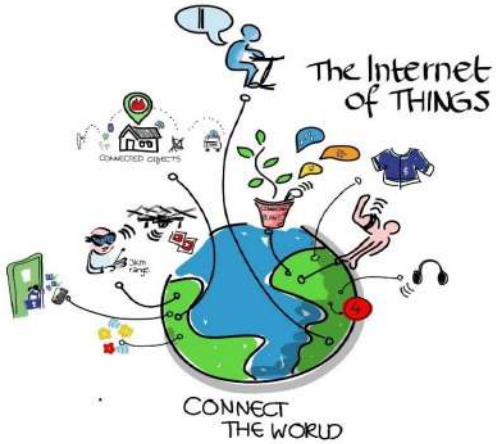
UGC – User Generated Content





7 miliardi di persone e 6,8 miliardi di cellulari

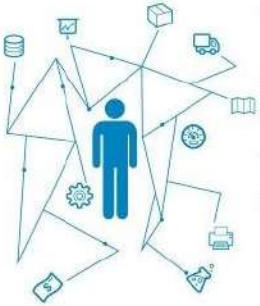
Internet of Things (IoT)



'There will be as many as

40 TO 80 BILLION

connected objects by 2020.



There will be **10** connected objects for every man, woman, and child on the **PLANET.**

Vehicle,asset,person & pet monitoring & controlling

Agriculture automation

Energy consumption

Security & surveillance

Building managment

Embedded Mobile

Internet of things

Everyday things get connected for smarter tomorrow

M2M & wireless sensor network

Everyday things

Smart homes & cities

Telemedicine & helthcare

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) -> a poche migliaia di dollari per un processo che dura un paio di settimane

Human Brain Project

- Un osservatorio del cervello che monitora 1 milione di neuroni (o 100.000 neuroni in 10 soggetti) per 1000 volte al giorno genererebbe 1 GB di dati al secondo, 4 TB di dati all'ora, 100 TB di dati al giorno, 4 PB di dati all'anno (immaginando un fattore di compressione di 1/10)

Il **Large Hadron Collider (LHC)**, generatore di particelle presso il CERN di Ginevra, è utilizzato per ricerche sperimentali nel campo della fisica delle particelle e può produrre 30 PB di dati all'anno

L'**Agenzia Spaziale Europea** genera più di un PB di dati all'anno

Le aziende generano dati (I)

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 veicolo
- **Airbnb** è il più esteso network dedicato alla ricettività, ma è del tutto privo di strutture

Le aziende generano dati (II)

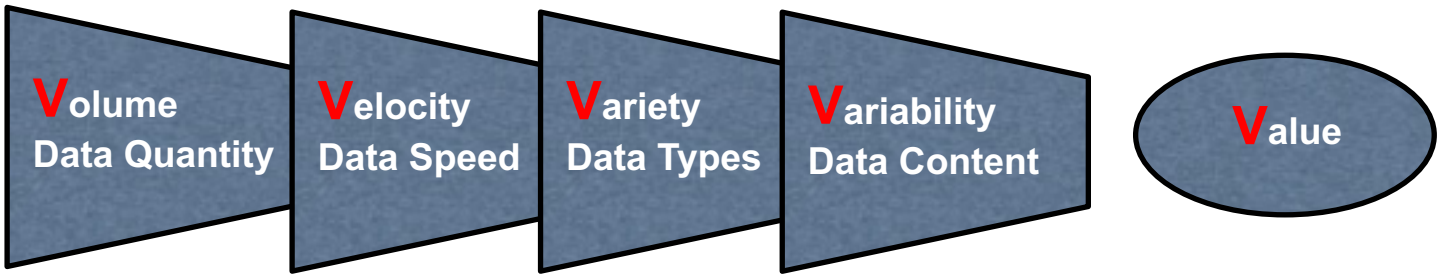
- Ordini, acquisti, vendite, spedizioni, difetti di produzione…
- I dati sono raccolti nei sistemi informatici delle aziende e sono considerati un asset (***intangibile***)
- Facebook dichiara asset (***tangibili***) per 6,3 miliardi, ma viene valutata in Borsa in 104 miliardi il giorno del suo debutto

- Nonostante i dati siano un asset, oggi viene elaborato solo lo 0,5% dei dati aziendali
- Perché
  - Mancanza di competenze sull'analisi computazionale dei dati
  - Sovversione dei poteri generati da un'informazione così tempestiva

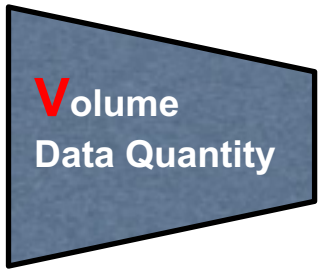
Caratteristiche distintive dei Big Data

Sono **dati** solitamente disponibili in **grandi volumi**, che si presentano in **differenti formati** (spesso privi di struttura) e con **caratteristiche eterogenee**, prodotti e diffusi generalmente con una **elevata frequenza**, e che **cambiano spesso** nel tempo

Le 5+ V dei Big Data:



Big Data: Volume

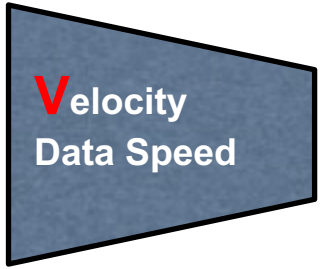


- Alcune tipologie di Big Data sono **transitorie**:
  - Dati generati dai sensori
  - Log dei web server
  - Documenti e pagine web

- Il **primo passo** quando si opera con i Big Data è quindi l'**immagazzinamento**; l'**analisi** (e la **pulizia**) avvengono in una fase successiva (per evitare di perdere potenziali informazioni)

- Ciò richiede **importanti investimenti in termini di storage e di capacità di calcolo** adatta all'analisi di grandi moli di dati

Big Data: Velocità (I)



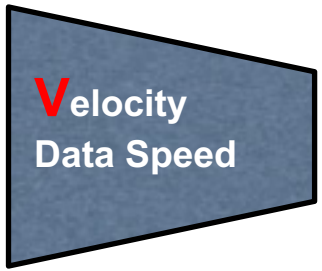
- È una delle caratteristiche che ha più significato
  - Si riferisce in primis alla elevata frequenza con cui i dati vengono generati – si ripercuote sulla quantità (Volume)
  - Si riferisce in secondo luogo anche alla velocità con cui le nuove tecnologie permettono di accedere e di analizzare questi dati

Maggiore è la velocità di accesso ai dati

Maggiore sarà la velocità in un processo decisionale

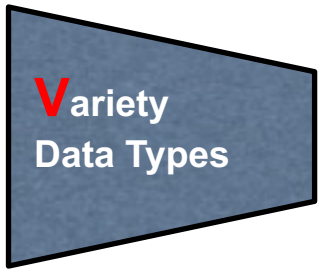
Maggiore/migliore sarà la competitività sui diversi panorami del mercato

Big Data: Velocità (II)



- Particolarmente adatte per gestire la velocità dei Big Data sono le **architetture distribuite**
  - Gestione di **strutture dati anche complesse**
  - **Accesso ai dati real-time** o, almeno, near-real-time
  - Spesso elaborazione di dati in **streaming**
  - **Velocità di elaborazione** grazie a tecniche di calcolo distribuito
- Potrebbe non esserci tempo per importare i dati in un DBMS relazionale per forzarne una rappresentazione uniforme (**tecnologie NoSQL/NewSQL**)

Big Data: Varietà



- Varietà nelle tipologie di dati e di sorgenti
  - **Strutturati** (DBMS tradizionali)
  - **Semi-strutturati** (XML, JSON, …)
  - **Destrutturati** (tweet, documenti, pagine web, …)
- Variabilità sia nella **struttura** dei dati che nella **semantica** sottostante
- Scarsa adattabilità alle restrizioni dei DBMS relazionali
  - Nel contesto Big Data i dati da trattare non sono sempre adatti ad essere lavorati con le tecniche tradizionali dei database relazionali
  - Dati come **email, immagini, video, audio, stringhe di testo** a cui dare un significato non si possono memorizzare in una tabella
  - Adozione delle tecnologie NoSQL/NewSQL, che non impongono uno schema rigido (***schemaless databases***)

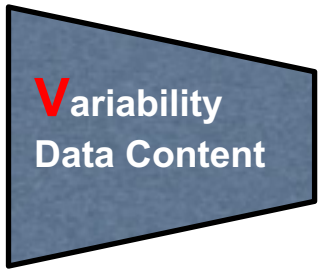
Struttura dei dati: dati semi-strutturati

Dati strutturati	Dati semi-strutturati
Netta divisione tra schema e istanze. Schema rigido. Schema-on-write.	Schema flessibile. Schema-on-read.
Basato sul concetto di insieme.	Basato sul concetto di lista.
Ordinamento irrilevante.	Ordinamento significativo sintatticamente e semanticamente.
Normalizzazione.	Annidamento.

Struttura dei dati: dati non strutturati

Dati strutturati	Dati non strutturati
Netta divisione tra schema e istanze. Schema rigido. Schema-on-write.	Nessuno schema, né in fase di lettura del dato, né in fase di scrittura.
Linguaggi di interrogazione/query dei dati (per esempio, SQL).	Tecniche ricerca delle informazioni da sorgenti non strutturate, tipicamente basate su parole chiave (*keyword*).
Modello booleano (completezza + correttezza).	Modello probabilistico (grado di adeguatezza dei risultati della ricerca rispetto a quanto cercato, precisione + recall).
Aggiornamento possibile.	Strategie cancella-tutto, riscrivi-tutto.

Big Data: Variabilità (I)

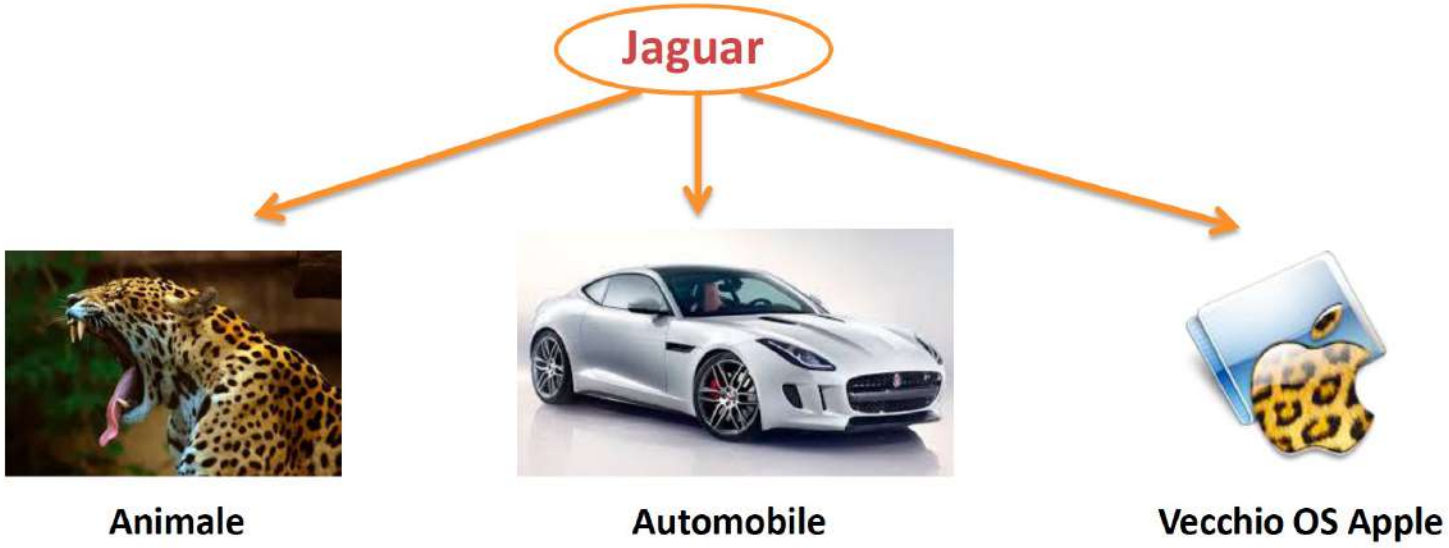


- Le sorgenti dei dati **non sono controllate e/o controllabili**
- C'è **incertezza** sulla singola informazione
  - Incompleta, vaga, …
  - Il significato o l'interpretazione della stessa informazione **può variare in base al contesto** in cui esso viene raccolto e analizzato
    - Per esempio, la frase "leggete il libro" avrà un *significato positivo* in un blog che parla di letteratura, mentre potrà avere una *connotazione negativa* in un blog per appassionati di cinema
    - Il significato di un dato può essere *differente anche in base al momento* in cui viene fatta l'analisi, spesso è fondamentale l'analisi in tempo reale (velocità)

Big Data: Variabilità (II)

Variability
Data Content

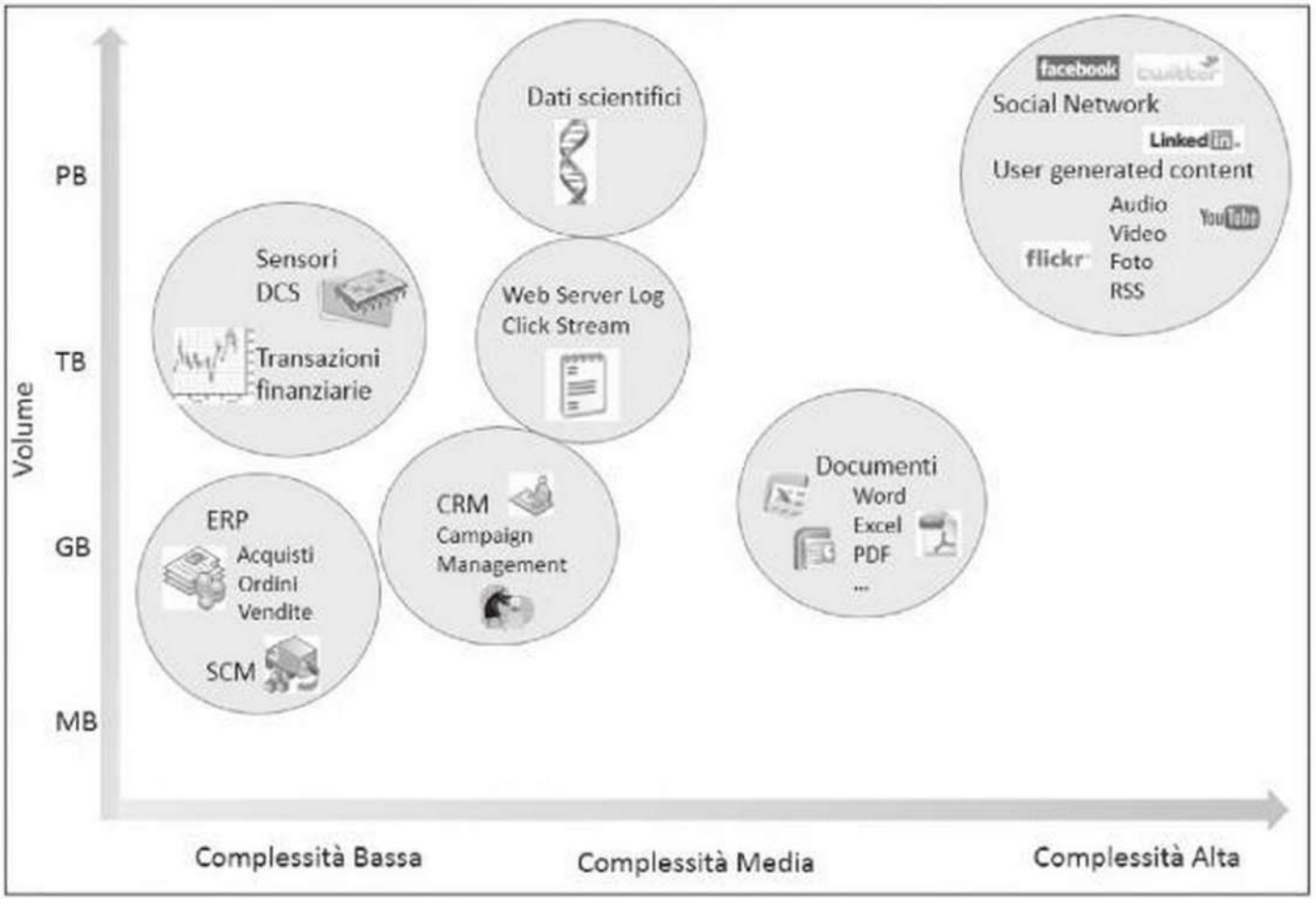
È importante trovare meccanismi che riescano a dare una **semantica ai dati** in base al contesto in cui sono espressi



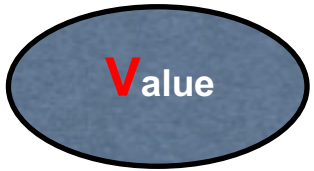
Altre V…

- **Veracità**: caratteristica che riguarda l'*affidabilità delle informazioni* con cui si ha a che fare (trustworthiness)
- **Viralità**: caratteristica che ha a che fare con *quanto e come i dati si diffondono* (propagazione dei dati)
  - Esempio: una notizia o un evento diffusi tra diversi canali, diffusione amplificata con i collegamenti nei vari social network
  - Virale è anche la crescita del volume dei dati generati dalle attività digitali dell'uomo (*user-generated content*)
  - *Influencer*: persone, organizzazioni o aziende ritenuti "esperti" in uno specifico settore e in grado di raggiungere con i contenuti che pubblicano un maggior umero di utenti "targettizzati" e interessati a determinate informazioni

Classificazione dei dati per volume e complessità



Big Data: Valore



- Potenzialità dei dati in termini di vantaggi competitivi raggiungibili con la loro analisi
- 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 e i servizi aziendali
  - Integrando continuamente nuove informazioni per costruire una base di conoscenza sempre più ampia
- Chi potrebbe 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

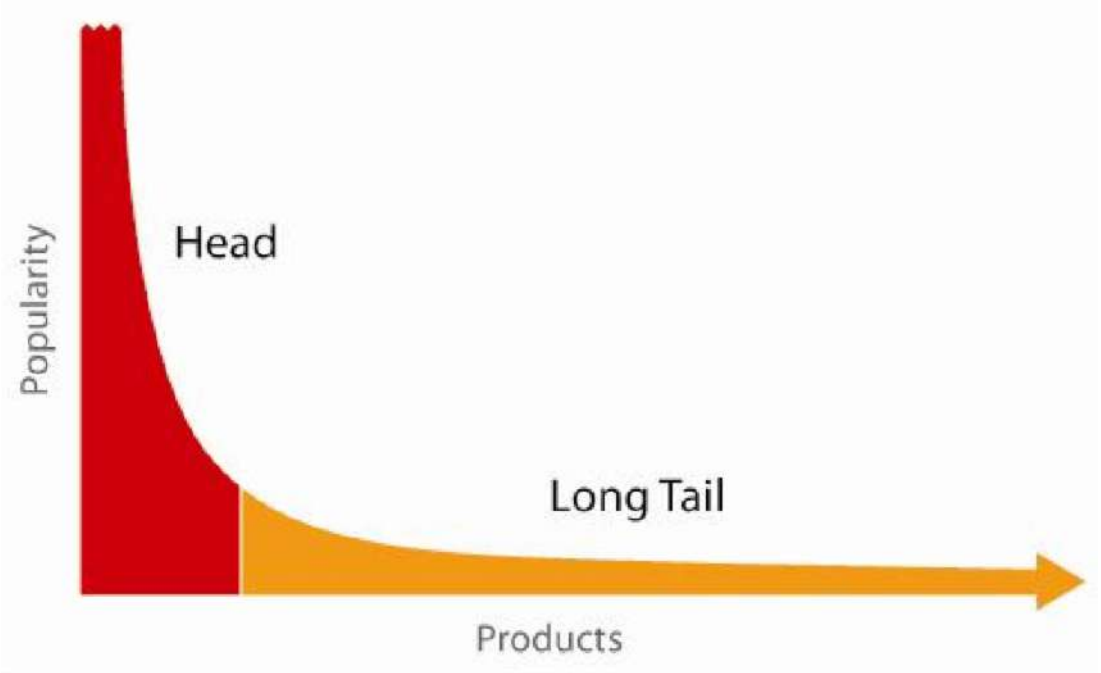
Big Data vs Data Science

- Data Science
  - La scienza dei dati studia i metodi per estrarre la conoscenza dei dati
    - Dati di qualunque natura e dimensione
- 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
- Data Science non necessita sempre di Big Data, tuttavia la costante crescita dei dati fa sì che i Big Data siano un aspetto importante della Data Science

Big Data: una rivoluzione

Big Data: raccolta dei 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 (**diluvio dei dati**)

- 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



Un cambio di prospettiva

L'ascesa dei Big Data evidenzia tre mutamenti nel modo in cui analizziamo le informazioni:

- Analizzare tutti i dati disponibili
- Rinunciare all'esattezza
- Abbandonare la tendenza a ricercare la causalità

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!!

L'identità "N=tutti" non comporta necessariamente l'analisi di una gran massa di dati

Rinunciare all'esattezza

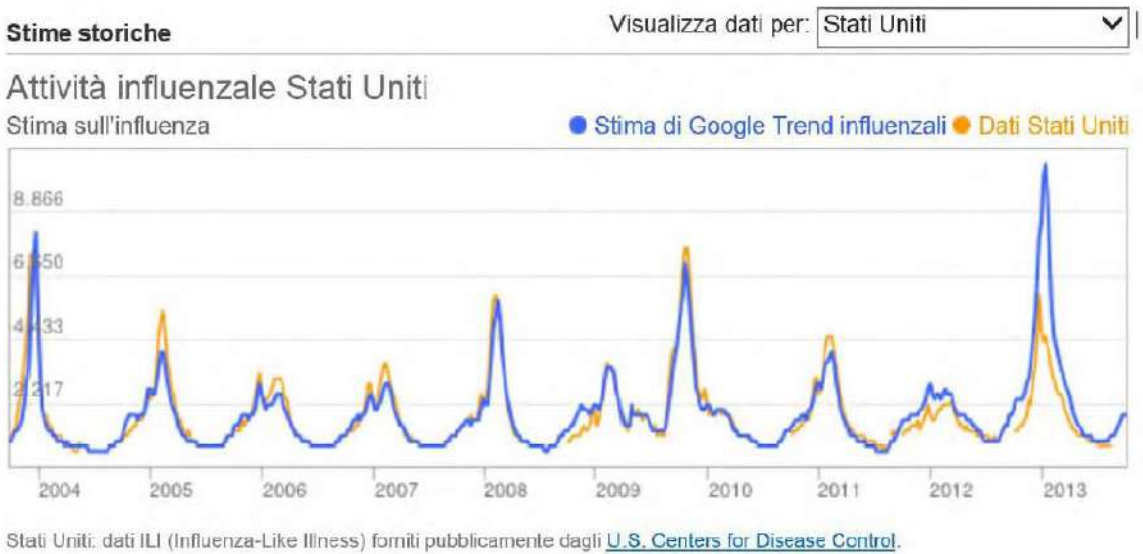
Nell'epoca dei Big Data, **la quantità è più importante della qualità**

L'abbondanza permette di tollerare un certo livello di imprecisione, di confusione

- Il traduttore di Google prende le informazioni di cui ha bisogno per le sue traduzioni da pagine Web non filtrate, piene di errori ortografici e sintattici e a volte incomplete, ma la sterminata quantità di dati a disposizione gli permette di essere più affidabile di tutti i suoi predecessori, che si basavano su dizionari corretti e redatti da esperti, ma con il limite di contenere un numero limitato di informazioni

Rinunciare all'esattezza

Google Flu Trends – previsione in base all'oggetto delle ricerche condotte on *Google Search* -> ugualmente accurate, ma in tempo reale



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

Meno causalità più correlazione

Non conta sapere perché (*why*) vendo un libro online, ma cosa (*what*) fa aumentare le perdite

- In previsione di un uragano aumentano le vendite di torce elettriche, ma anche di merendine e dolci
- La dimostrazione di una causalità è molto più costosa della individuazione di una correlazione

Esempio – Il peso dei bambini che frequentano la scuola elementare è correlato positivamente al quoziente intellettivo

- Facile da scoprire
- Direste che mangiare fa aumentare il quoziente intellettivo
- Oppure che il quoziente intellettivo influisce sul peso

Se tenessimo sotto osservazione il fattore età giungeremmo a conclusioni diverse – ma **tenere sotto osservazione un singolo fattore costa e non è sempre possibile**

Alcuni casi d'uso interessanti

Dominio e sfide	Nuovi dati	Nuove opportunità
Healthcare Costi per visite pazienti	Remote patient monitoring	Assistenza sanitaria preventiva, riduzione delle ospedalizzazioni
Manufacturing Supporto per gli operatori	Dati da sensori	Diagnosi automatica, manutenzione predittiva
Location-Based Services Basati sulla posizione	Real time location data	Ricerca di info geolocalizzate, traffico, geo-advertising
Settore pubblico Servizi per il cittadino	Dati raccolti dai cittadini	Servizi personalizzati, riduzione dei costi
Retail e politica Prodotti mirati al singolo	Social media	Sentiment analysis per la segmentazione della clientela, key influencer identification
GIS Servizi geo-localizzati	Dati raccolti con coordinate geografiche	Disaster management, advertising geo-localizzato, fraud detection

Problematiche legate alla caratteristiche dei Big Data (I)

- Costituiscono nuove sorgenti di dati, da integrare con quelle tradizionali, dove la gestione dei dati costituisce di per sè una sfida (per dimensione, velocità di raccolta)

- Non sono pensati per essere user-friendly (e.g., data streaming), anche perchè sono spesso generati automaticamente (per esempio, dati provenienti dai sensori sulle macchine)

- Permettono di analizzare la realtà al massimo livello di dettaglio, ma
  - non tutti i dati sono importanti
  - non è sempre possibile sapere quali vanno scartati e quali opportunamente processati

Problematiche legate alla caratteristiche dei Big Data (II)

- Elevato numero di campi applicativi diversi tra loro
- Differenti canali attraverso i quali i dati vengono raccolti
- Impossibile identificare un'unica architettura adattabile a tutte le aree
- Come è possibile scoprire il "valore" dei Big Data

- Utilizzo di complesse analisi e processi di modellazione e organizzazione dei dati
- Formulazione di ipotesi -> implementazione di modelli semantici, visuali, statistici -> validazione

Criticità e rischi dei Big Data – Qualità dei dati

La **qualità dei dati** è determinata da un insieme di caratteristiche:

- **Completezza:** la presenza di tutte le informazioni necessarie a descrivere un oggetto, entità o evento (es. anagrafica)
- **Consistenza:** i dati non devono essere in contraddizione (ad esempio, il saldo totale e movimenti, disponibilità di un prodotto richiesto da soggetti differenti, etc.)
- **Accuratezza:** i dati devono essere corretti, cioè conformi a dei valori reali (ad esempio, un indirizzo mail non deve essere solo ben formattato *nome@dominio.it*, ma deve essere anche valido e funzionante)
- **Assenza di duplicazione:** Tabelle, record, campi dovrebbero essere memorizzati una sola volta, evitando la presenza di copie; le informazioni duplicate comportano una doppia manutenzione e possono portare problemi di sincronia (consistenza)
- **Integrità:** è un concetto legato ai database relazionali, in cui sono presenti degli strumenti che permettono di implementare dei **vicoli di integrità**; per esempio, un controllo sui tipi di dato (presente in una colonna), o sulle chiavi identificative (impedire la presenza di due righe uguali)

Criticità e rischi dei Big Data – Qualità dei dati

Nei contesti applicativi che coinvolgono l'uso di database tradizionali, la **qualità complessiva dei dati** può essere minata da:

- **Errori nelle operazioni di data entry** (campi e informazioni mancanti, errati o malformati)
- **Errori nei software di gestione dei dati** (query e procedure errate)
- **Errori nella progettazione delle basi di dati** (errori logici e concettuali)

Criticità e rischi dei Big Data – Qualità dei dati

Nel mondo Big Data invece:

- **Dati operazionali**: i problemi relativi alla qualità sono noti ed esistono diversi strumenti per realizzare in modo automatico la pulizia dei dati
- **Dati generati automaticamente**: i dati scientifici o provenienti dai sensori sono privi di errori di immissione, ma sono spesso “deboli” a livello di contenuto informativo, è necessario integrarli con dati provenienti da altri sistemi per poi analizzarli
- **Dati del web**: social network, forum, blog generano dati (semi-)strutturati; la parte più affidabile è costituita dai metadati, mentre il testo è soggetto ad errori, abbreviazioni, etc.

Criticità e rischi dei Big Data – Qualità dei dati

Nel mondo Big Data invece:

- **Disambiguare le informazioni**: uno stesso dato può avere significati diversi (es. calcio), la sfida è quella di trovare il significato più attinente al contesto in esame
- **Veridicità**: notizie, affermazioni, documenti non sempre veri o corrispondenti alla realtà

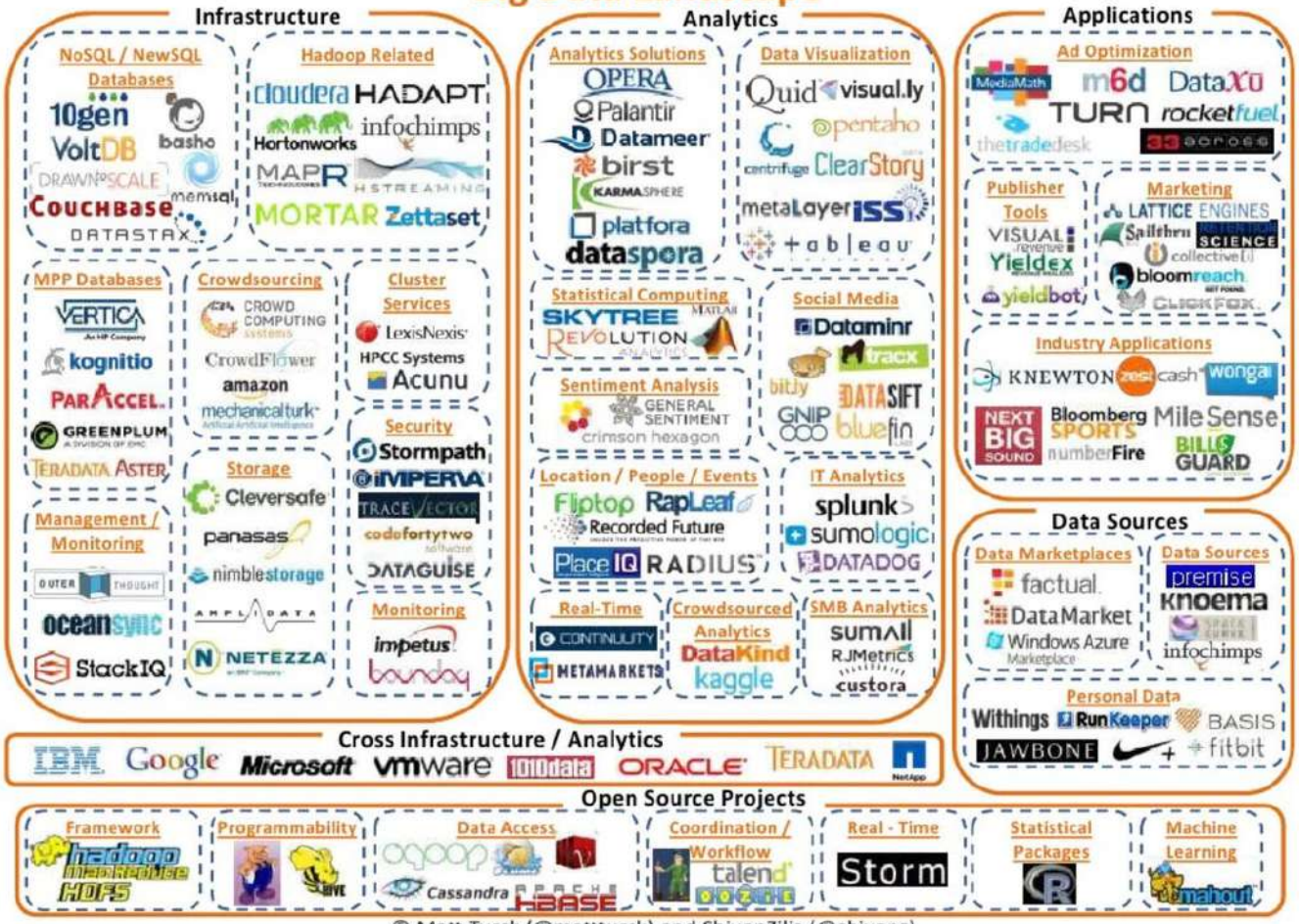
Osservazione: *la qualità dei dati è spesso legata al contesto in cui essi sono analizzati; le operazioni di filtraggio e pulizia devono essere effettuate per gradi, onde evitare di eliminare dati potenzialmente utili*

Criticità e rischi dei Big Data – Privacy

Il tema Big Data si apre a problemi di privacy, proprietà ed utilizzo dei dati da parte di terzi:

- **Dati del web**: gli *user-generated content* sono condivisi e accessibili a tutti, è etico il loro utilizzo?
- **Dati sensibili**: i dati relativi alla storia degli utenti sono opportunamente trattati e protetti; per esempio, l'uso di smartphone, GPS, sistemi di pagamento elettronico, ma anche social network lasciano delle tracce da cui è possibile ricavare gli spostamenti degli utenti

Big Data Landscape



© Matt Turck (@mattturck) and ShivonZilis (@shivonz)