CORSO I.F.T.S

“TECNICHE PER LA PROGETTAZIONE E LA GESTIONE DI DATABASE ”

Ing. Mariano Di Claudio

Lezione del 10/09/2014

Chi sono

- Ing. ***Mariano Di Claudio***
- ***Assegnista di Ricerca*** presso il **DISIT Lab**. *dell'Università degli Studi di Firenze*
- 2° anno di ***Dottorato di Ricerca in "Informatica, Sistemi e Telecomunicazioni"***
- Email: **mariano.diclaudio@unifi.it**

Programma del Corso

- **Panoramica sul tema Big Data**
  - *Definizione, evoluzione e concetti chiave*
  - *Principali contesti applicativi*
- **Problematiche, tecnologie e metodologie per la gestione e l'analisi dei Big Data**
  - *Aspetti architetturali e di Data Management*
  - *NoSQL Database tipologie ed esempi*
- **Ecosistema Hadoop**
  - *Descrizione ed esempi implementativi*

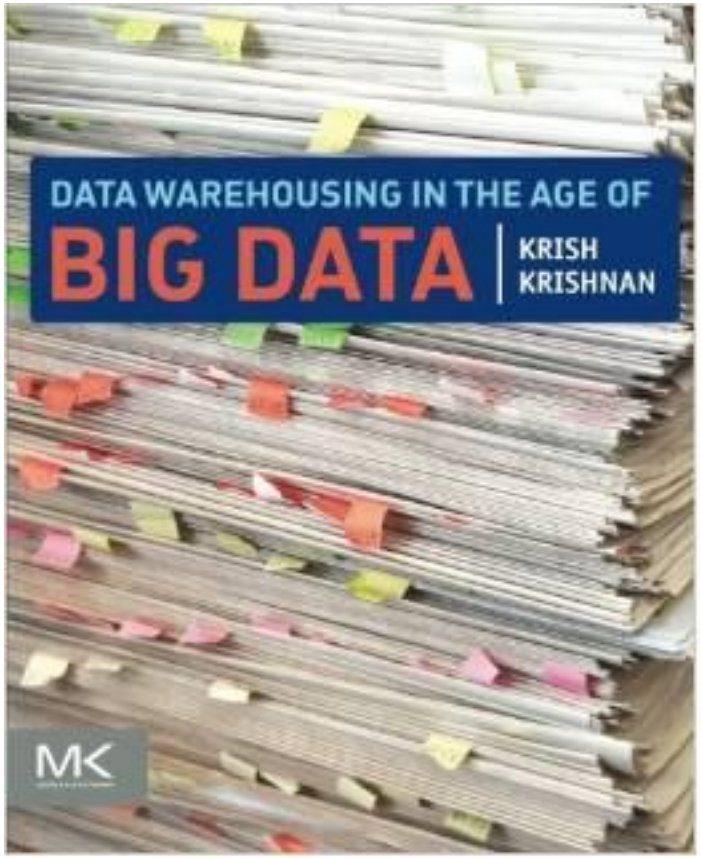
Testi di riferimento

- **Big Data**, *Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati*

- ***Alessandro Rezzani***

- ***Apogeo Editore***

Testi di riferimento



- *Data Warehousing in the age of* **Big Data**

- ***Krish Krishnan***

- ***Morgan Kaufmann*** *Editore*

Indice

Evoluzione dei dati e delle tecniche di analisi

- Partendo dai **dati** i processi di analisi vogliono **trasformarli** in **informazioni utilizzabili per supportare i processi decisionali** (in contesti aziendali e non).

- Negli **anni ’60** i dati erano immagazzinati su **dischi e supporti magnetici**. Si svolgevano **analisi statiche e limitate** (es. il numero di vendite dell’ultimo semestre…)

- Negli **anni ‘80** i **database Relazionali e SQL** (*Structured Query Language*) permette di realizzare **analisi più dinamiche**.

- Analisi svolte su DB **operazionali,** su cui è registrata ad esempio l’attività giornaliera di un’azienda.

Evoluzione dei dati e delle tecniche di analisi

Problemi delle basi di dati **operazionali**

- L'analisi è svolta da applicativi differenti:
  - Gestione degli ordini.
  - Gestione delle anagrafiche.
  - Contabilità e fatturazione.

- Applicazioni differenti non garantiscono **l'uniformità** e la **coerenza dei dati**:
  - **Dati replicati** e manipolati in sw differenti.
  - **Possibili differenze di formato**.
  - **Aggiornamenti dei dati non garantiti**.

Evoluzione dei dati e delle tecniche di analisi

Problemi delle basi di dati **operazionali**

- Sono di tipo **OLTP** (*On Line Transaction Processing*), e presentano un modello dati fortemente ***normalizzato.***

- **(+)** La normalizzazione favorisce **inserimenti**, **cancellazioni e modifiche dei dati** (attività transazionali).

- **(-)** Non è però adatta alle letture.
- **(-)** Incremento notevole del numero di tabelle.
- **(-)** Molte operazioni di JOIN per *denormalizzare* (ricostruire la forma tabellare) e quindi estrazione dei dati complessa.
- **(-)** Mancanza di una profondità storica dei dati.

Evoluzione dei dati e delle tecniche di analisi

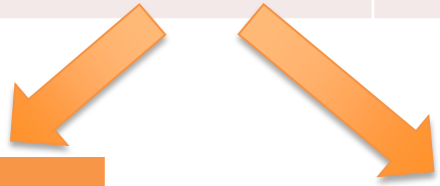
Si consideri il **DB di una palestra** in cui sono raccolti il ***Codice Fiscale*** dell'iscritto (chiave primaria), il ***Codice Corso*** e il nome dell'***Insegnante.***

Codice Fiscale	Codice Corso	Insegnante
GBYNJU76B15H345F	BB01	Gianni
BABGAF89U12J564F	BB01	Gianni
NCLACG25H02G563N	BB01	Gianni
ACVSGB12F11K764G	AA03	Federica

Evoluzione dei dati e delle tecniche di analisi

Il DB **non è in forma normale** perché il campo ***Insegnante*** non dipende dalla *chiave CF*, ma dal campo ***Codice Corso***.

Codice Fiscale	Codice Corso	Insegnante
GBYNJU76B15H345F	BB01	Gianni
BABGAF89U12J564F	BB01	Gianni
NCLACG25H02G563N	BB01	Gianni
ACVSGB12F11K764G	AA03	Federica



Codice Fiscale	Codice Corso
GBYNJU76B15H345F	BB01
BABGAF89U12J564F	BB01
NCLACG25H02G563N	BB01
ACVSGB12F11K764G	AA03

Codice Corso	Insegnante
BB01	Gianni
AA03	Federica

Evoluzione dei dati e delle tecniche di analisi

- A causa di questi limiti a partire dagli anni ’90 si inizia a parlare di ***data warehouse,*** cioè db che integrano dati provenienti da diversi sistemi operazionali.

- I dati sono **integrati**, **certificati** e **consistenti** ossia il punto di partenza perfetto per le attività di analisi dei sistemi di **BI**.

- **BI (***Business Intelligence***) =** è un *insieme di metodi, modelli, processi, persone e strumenti che permettono una raccolta dati regolare e organizzata*.

- **Dati** che possono poi essere e**laborati, aggregati, analizzati e trasformati (valorizzati) in informazioni**, che vengono conservate e rese accessibili in modo semplice e flessibile.

Evoluzione dei dati e delle tecniche di analisi

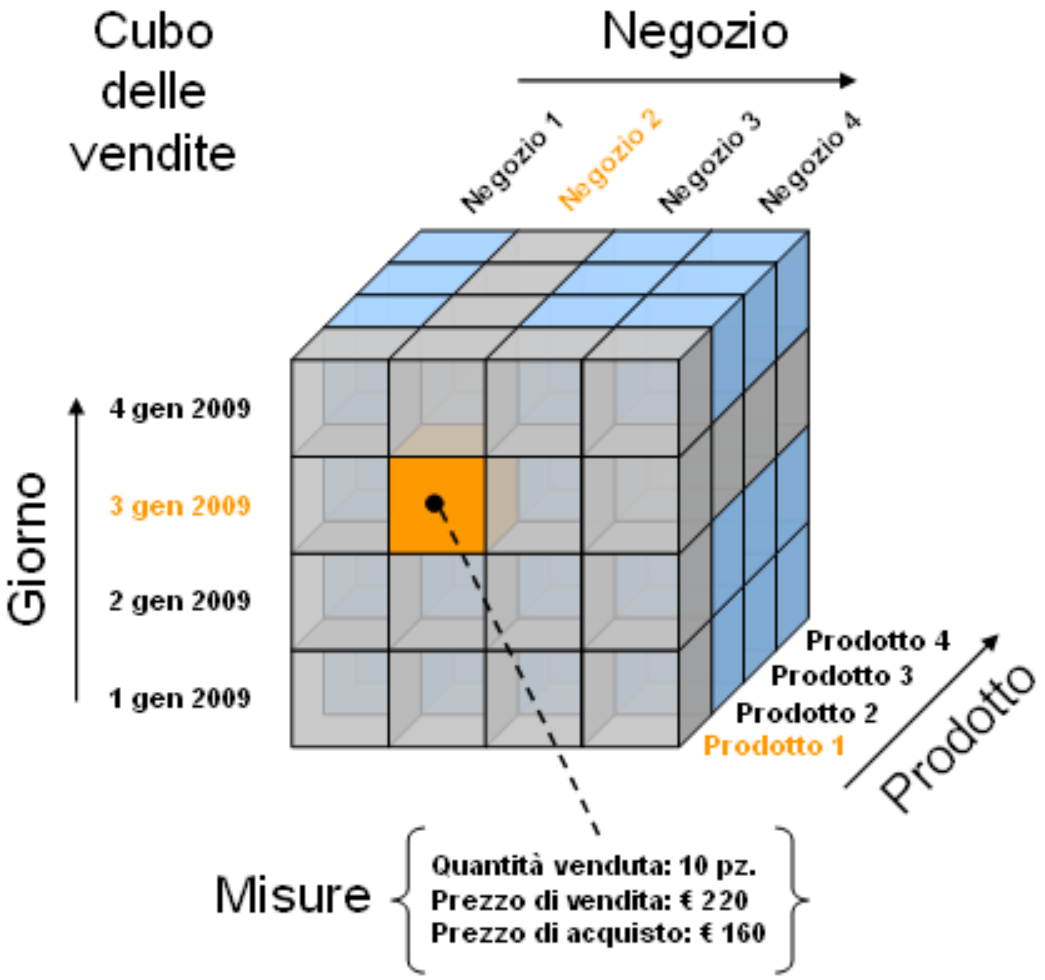
Evoluzione dei sistemi di **BI**

- Possibilità di analisi su data warehouse, con query SQL su basi di **dati multidimensionali** (dati e metadati insieme).

- Questi db sono sistemi di tipo **OLAP (***On Line Analytical Processing***).** Hanno una struttura multidimensionale, chiamata ***Ipercubo*** (spesso semplificata in tre dimensioni).

(-) Questi sistemi offrono comunque una visione storica:
  - Valutazioni di ciò che è accaduto o che sta accadendo.
  - Valutazione statica.

Evoluzione dei dati e delle tecniche di analisi



Navigazione dei dati più semplice grazie ad operazioni di:

- Drill down
- Drill-up
- Slicing
- Dicing

Evoluzione dei dati e delle tecniche di analisi

- Dai primi anni Duemila viene fuori la necessità di **un'analisi dei dati in grado fare previsioni e dare suggerimenti per anticipare gli eventi.**

- Si inizia a parlare di ***data mining,*** termine che identifica un insieme di tecniche in grado di ***"scavare "*** nei dati per **estrarre nuove informazioni e significati**, non evidenti immediatamente.

- Queste tecniche portano spesso alla definizione ***pattern*** (cioè un modello di rappresentazioni di alcune informazioni) e **relazioni tra i dati**.

- *Numerose applicazioni*: la segmentazione della clientela, market basket analysis, campagne pubblicitarie mirate, previsioni etc.

Evoluzione dei dati e delle tecniche di analisi

Dal **2010** le principali tendenze evolutive nell'analisi dei dati e BI sono:

- **Sviluppo di strumenti di business analytics.**

Tecnologie e applicazioni che fanno uso di modelli matematici e statistici per operazioni di data analysis e data mining.
Solitamente offrono **funzionalità per migliorare la visualizzazione dei dati e favorirne la navigazione**, e strumenti di **ottimizzazione nella gestione dei processi** (suddivisione carico di lavoro).

- **Collaboration e information sharing**

La collaborazione e la condivisione delle informazioni (report, documenti, modelli, valutazioni e analisi già svolte) è un requisito sempre più importante, soprattutto in un contesto aziendale.
Es. *Microsoft Share Point*, portale web per pubblicare informazioni.

Evoluzione dei dati e delle tecniche di analisi

Dal **2010** le principali tendenze evolutive nell'analisi dei dati e BI sono:

- **Cloud Computing**

- Risorse HW e SW disponibili come servizi su internet.
- Accesso alle risorse da diversi luoghi e con diversi dispositivi.
- Basti costi iniziali di investimento (determinabili a priori).
- Architettura scalabile.
- Gestione e manutenzione piattaforma (aggiornamenti sw, backup, fault-tollerance..) sono a carico del provider.

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**
- **Sensori, DCS (Distributed Control System) e strumenti scientifici**
- **Dati non-strutturati e semi-strutturati**

Basi di dati Operazionali

Sono quei **dati** che hanno a che fare con **l'attività giornaliera di un'azienda** (industrie, banche o GDO). Alcuni esempi sono:

- *Applicativi di gestione della produzione* (materie prime, consumi…)
- *Applicativi di gestione degli acquisti* (prodotti, ordini, magazzino…)
- *Applicativi di contabilità* (fatture, saldo, movimenti…)
- *Applicativi di gestione del personale* (anagrafica, premi, malattie…)
- *Applicativi di gestione del cliente* (abitudini, marketing mirato…)

In alcuni casi i **dati operazionali arrivano a creare dei volumi rilevanti**. Esempio consideriamo una banca di grandi dimensioni:

Basi di dati Operazionali

- Le basi di dati operazionali in genere fanno riferimento ai **database relazionali** o **RDBMS** (*Relational Data Base Management System*), tra i più famosi ci **sono MySQL**, **IBM DB2**, **Oracle**, **Microsoft SQL Server**.

- **Aumento dei dati** = Gestione e storicizzazione complessa e onerosa in termini di risorse.

- Gli RDBMS mettono a disposizione alcune tecniche di ottimizzazione:
  - **Indicizzazione**
  - **Compressione**
  - **Partizionamento**

Basi di dati Operazionali

- **Indicizzazione**: Utilizzo di Indici (strutture ordinate).
  - **(+)** Recupero rapido di informazioni.
  - **(-)** Scritture lente e aumento dello spazio occupato dal DB.

- **Compressione:** Applicazione di algoritmi di compressione.
  - **(+)** Meno spazio per il salvataggio dei dati.
  - **(-)** Tempo di esecuzione degli algoritmi e decompressione dei dati dopo averli recuperati.

- **Partizionamento:** Suddivisione di una tabella in più parti sulla base di uno specifico criterio.
  - **(+)** Query limitate ad una parte limitata del DB (es. tutti i record da una certa data in poi).
  - **(-)** I vantaggi si perdono se le query impattano più partizioni.

Sensori, DCS e strumenti scientifici

- **Dati** prodotti da sistemi computerizzati utilizzati per il monitoraggio e controllo di impianti industriali.
- Gli impianti generalmente sono costituiti da **numerosi componenti (e sensori) distribuiti**, che inviano i dati ad una postazione centralizzata.
- Le rilevazioni vengono realizzate in intervalli temporali molto piccoli, anche **meno di 1 secondo**.

Dati non-strutturati e semi-strutturati

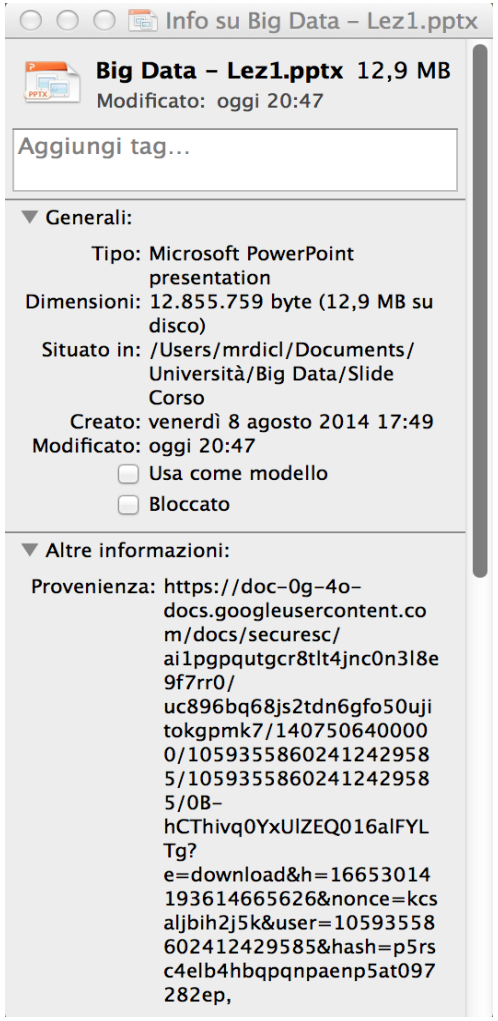
Sono dei **dati che non presentano una struttura predefinita** e che quindi **non** si prestano ad essere gestiti con uno **schema tabellare**. Alcuni esempi presenti in un contesto aziendale sono:

- **Documenti di varia tipologia** (PDF, Word, Excel, PowerPoint etc.)
- **E-mail**
- **Immagini in vari formati** (JPEG, TIFF, GIF, RAW etc.)
- **Strumenti Web 2.0** (Forum e Wiki)

Alcuni di questi documenti in realtà non sono del tutto privi di struttura, si potrebbero definire **semi-strutturati,** per la presenza di informazioni aggiuntive rappresentabili in tabella, **i metadati**.

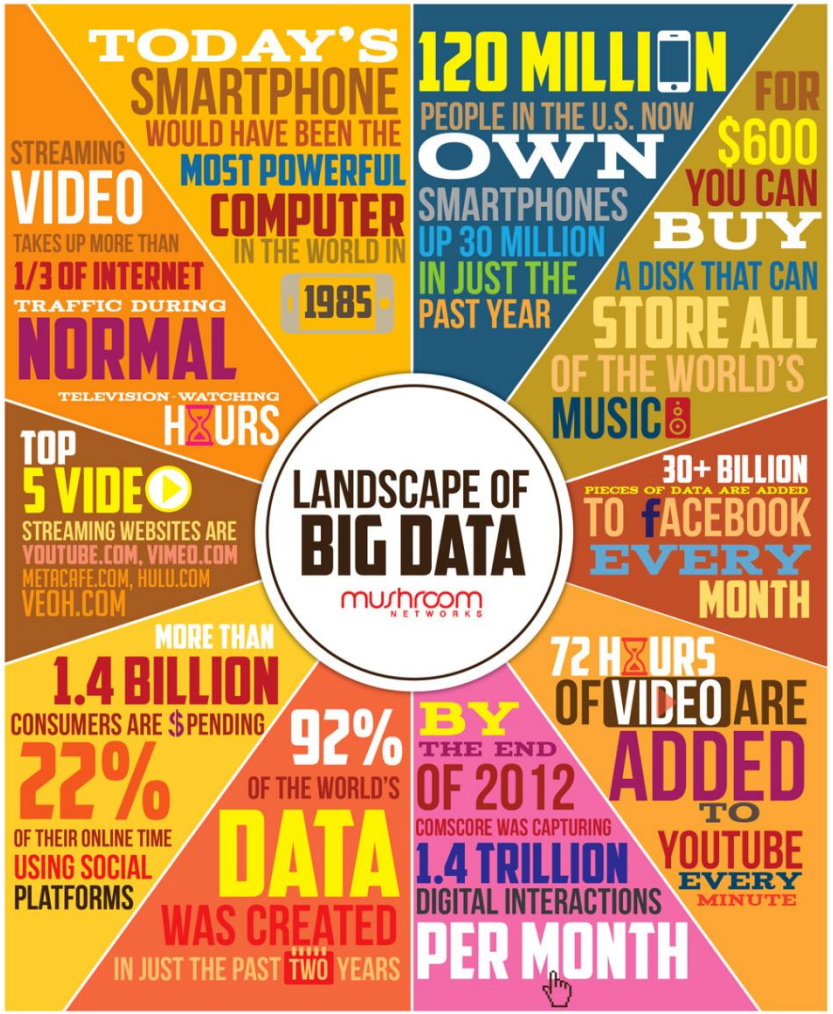
Dati non-strutturati e semi-strutturati

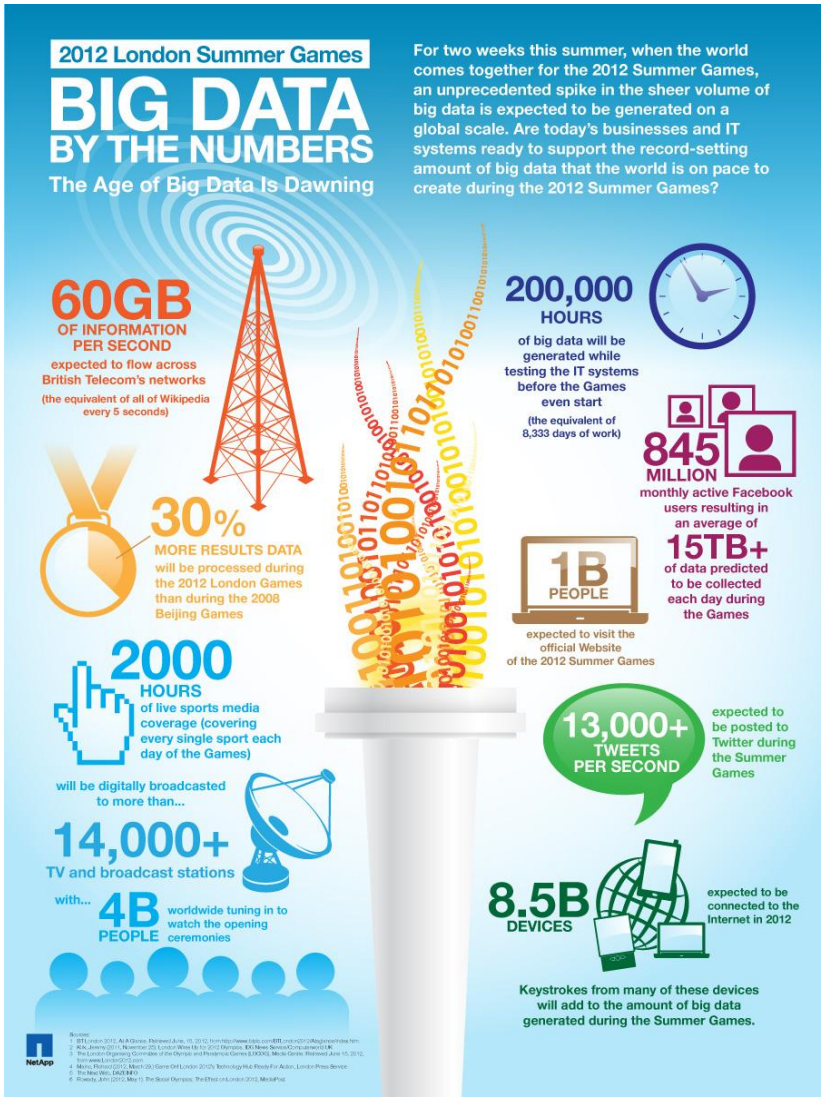
- **Metadati:** sono dei dati utilizzati per descrivere altri dati.
- Sono **facilmente estraibili** dai documenti che descrivono e messi in tabelle.
- Possono essere utilizzati per operare delle ricerche di e sui documenti.



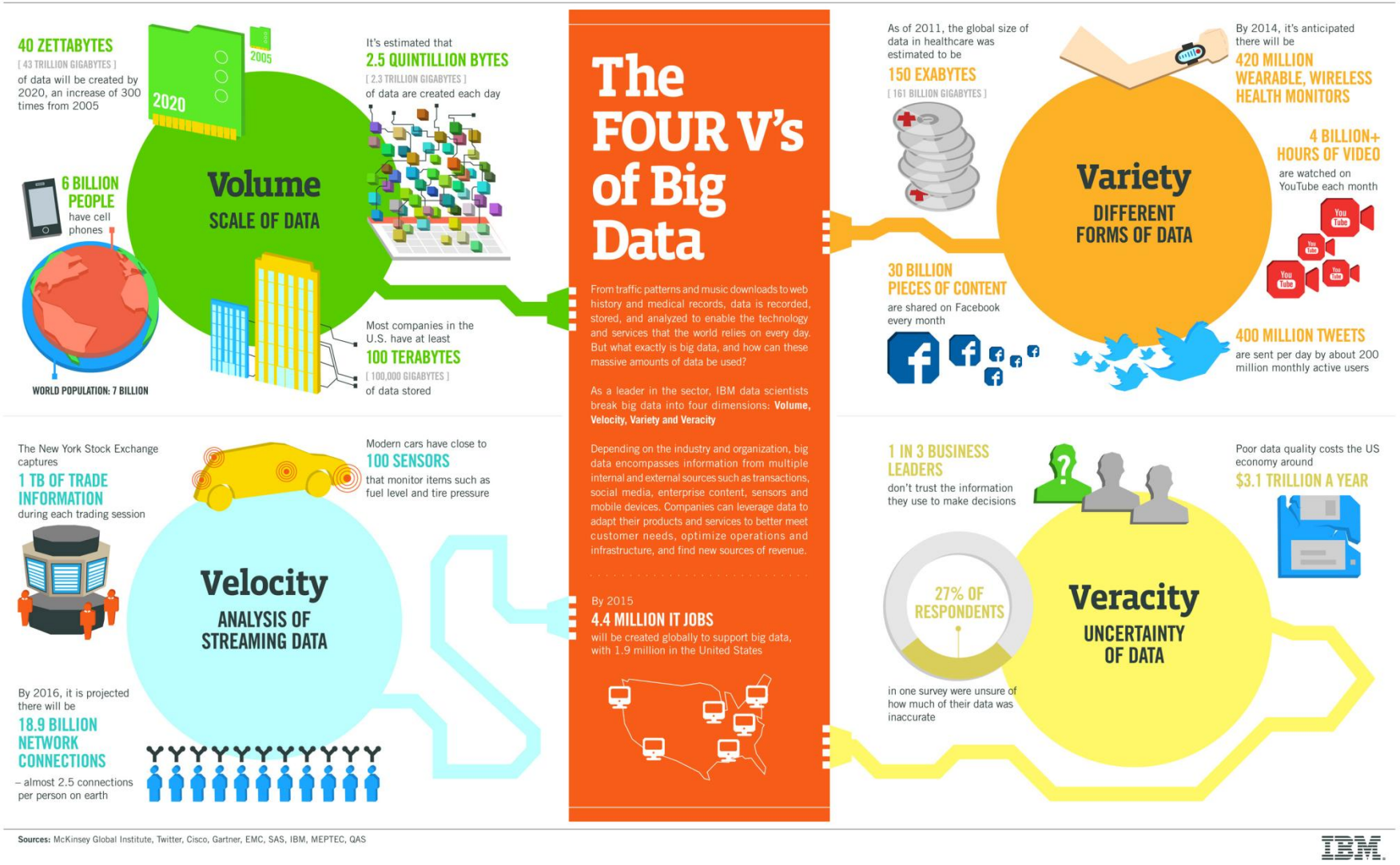
Evoluzione dei dati e delle tecniche di analisi

Big Data

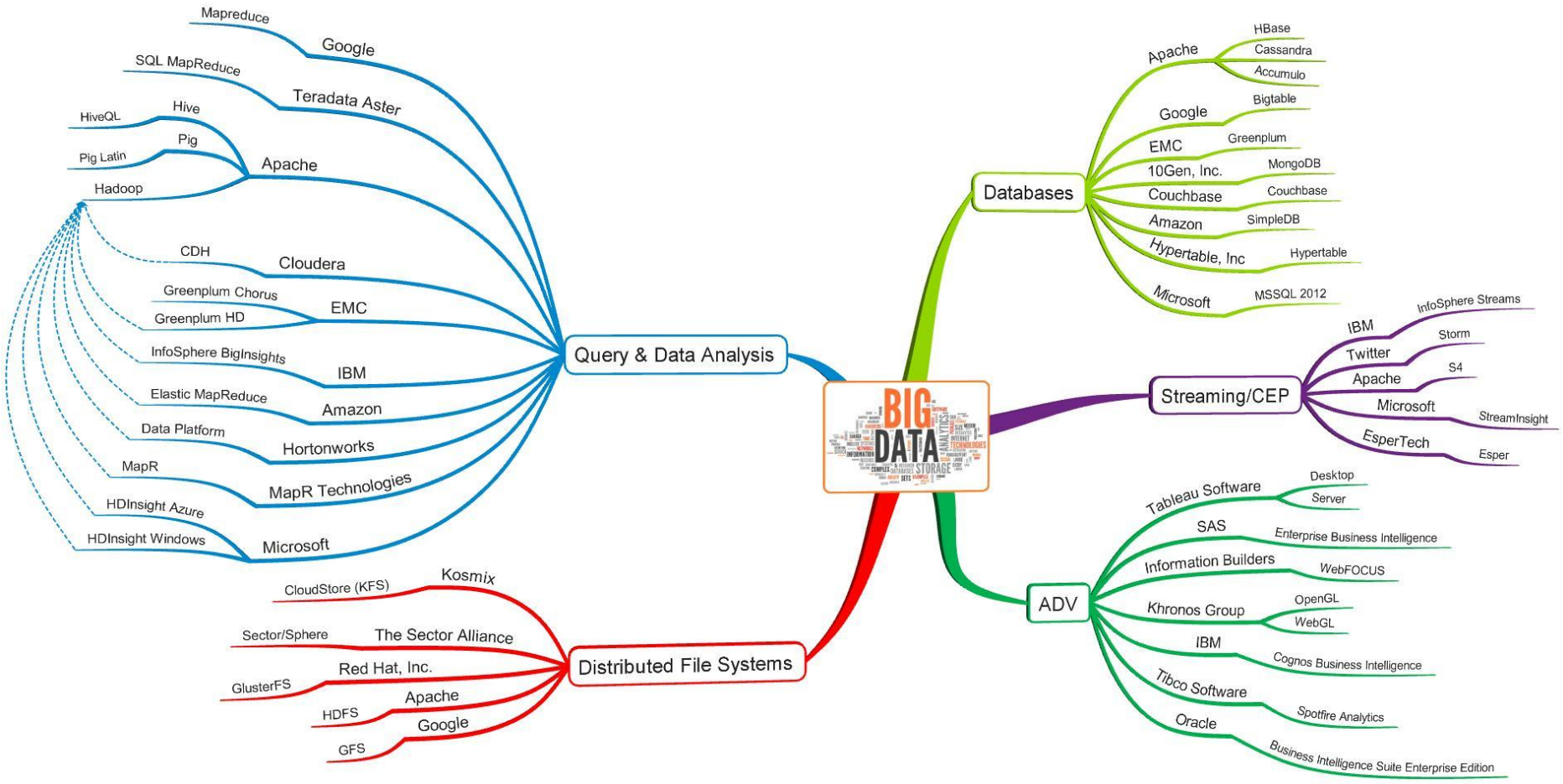




Evoluzione dei dati e delle tecniche di analisi



Evoluzione dei dati e delle tecniche di analisi



Indice

1. **Big Data**
   - *Evoluzione dei dati e delle tecniche di analisi*
   - *Le 5V dei Big Data* 
   - *Teorema CAP*
   - *Pipeline dell'analisi dei Big Data*

2. **Principali Contesti Applicativi**

3. **Criticità e Rischi dei Big Data**

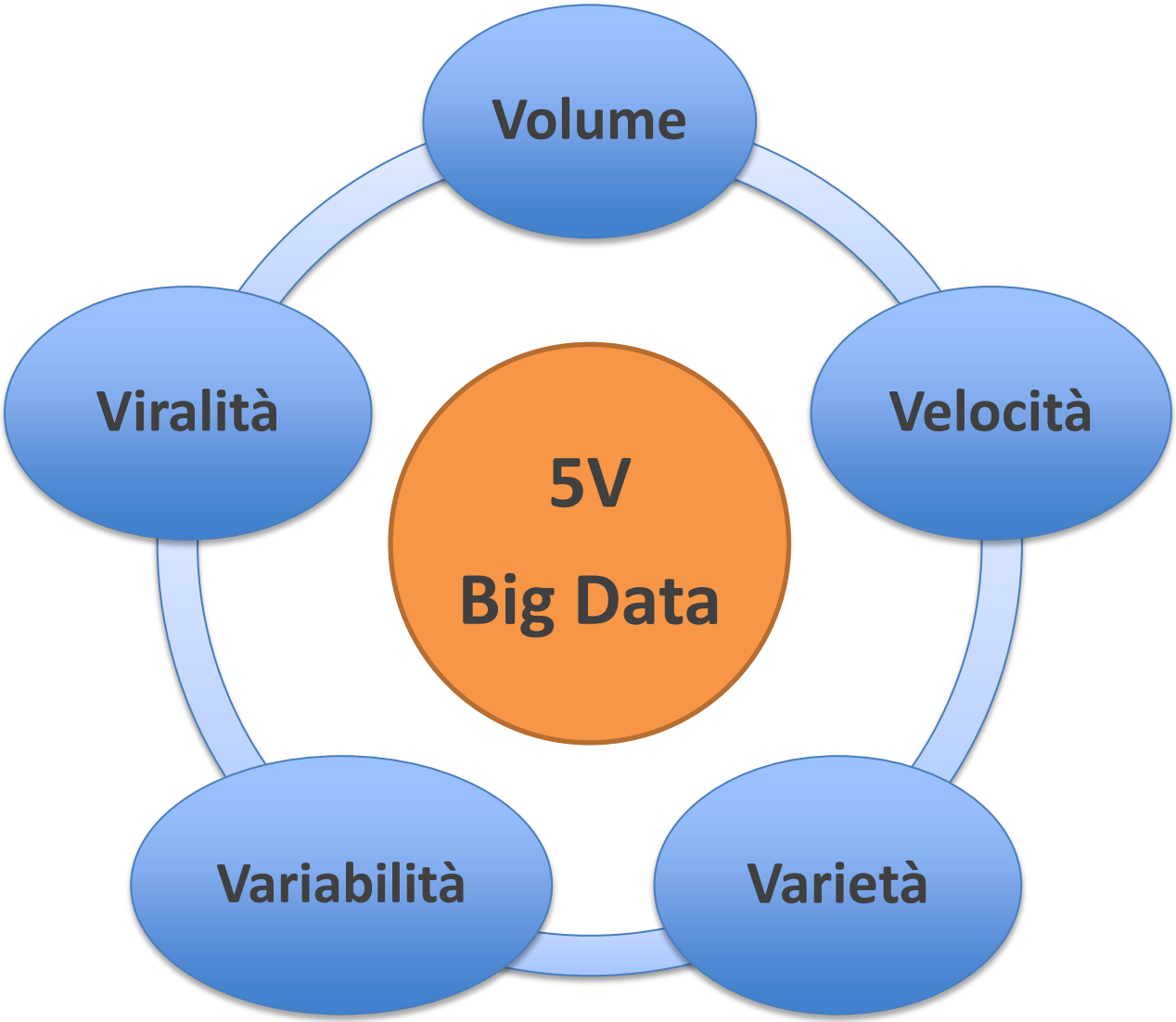
Big Data

I Big Data sono dati che superano i limiti degli strumenti tradizionali.

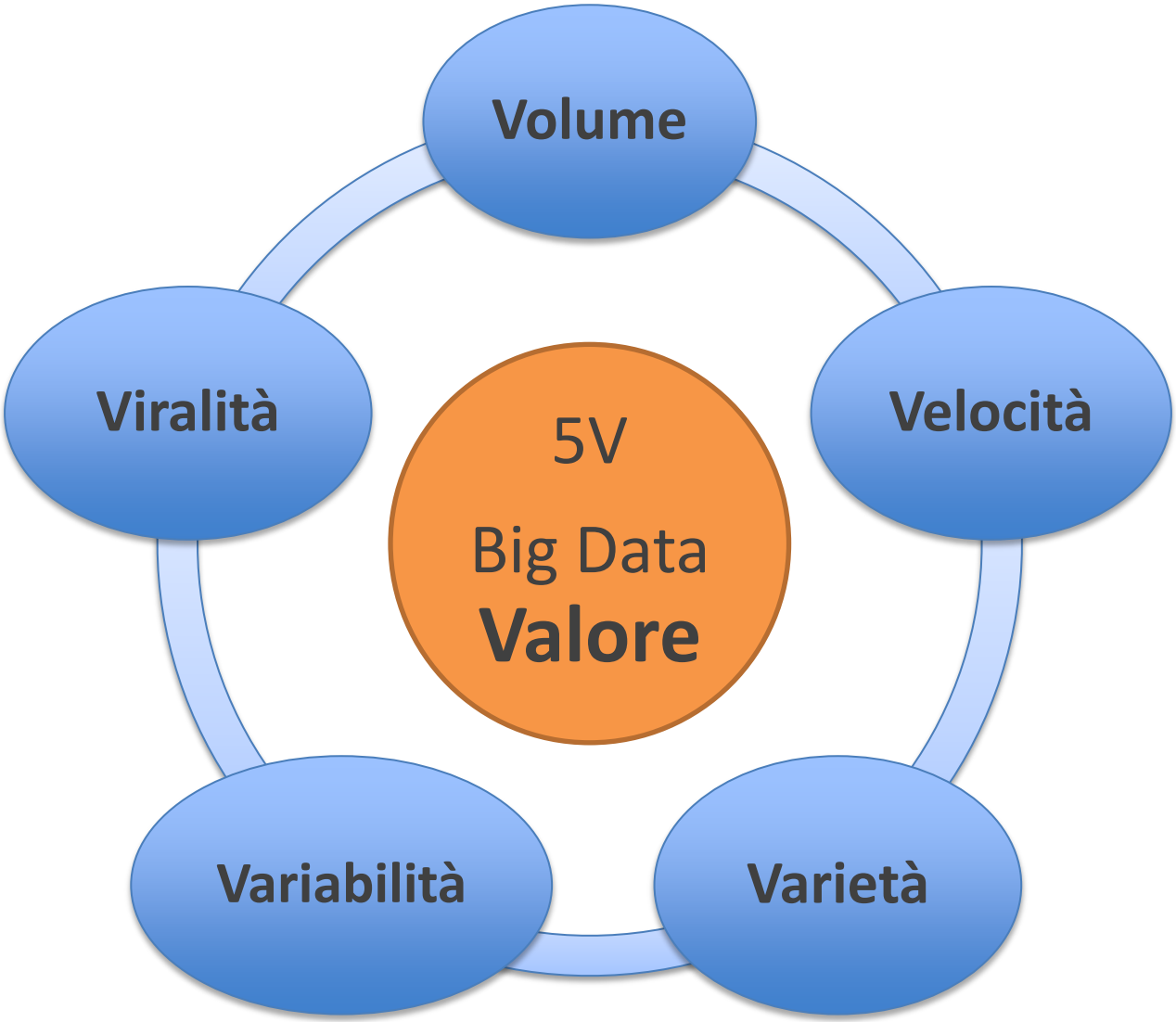
- *Sono **dati** solitamente disponibili in **grandi volumi**, che si presentano in **differenti formati** (spesso privi di struttura) e con **caratteristiche eterogenee**, prodotti e diffusi generalmente con una **elevata frequenza**, e che **cambiano spesso** nel tempo.*

- Per questo motivo sono identificati con le **5V (+1).**

I Big Data



I Big Data


Volume
Velocità
Varietà
Variabilità
Viralità
5V
Big Data
Valore

Big Data: le 5 V - Volume

Volume: forse la caratteristica più immediata, dal momento che si tratta di dati presenti in **grandi quantità**. In **1 minuto** infatti:

- **100 mila tweet** trasmessi nel mondo.
- **35 mila "Like" FB** a siti ufficiali di organizzazioni.
- **160 milioni (circa) di email** inviate.
- **2 mila check-in su 4square** effettuati.

- Ciò va aggiunto alle restanti **"attività digitali"**, generando una enorme mole di dati e informazioni a loro volta incrociabili.

- Aziende, martketeers, analisti (ma anche la politica) sono le figure più ingolosite dalle potenzialità di tutto ciò.

Big Data: le 5 V - Volume

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

- Il **primo passo** quando si opera con i Big Data é allora 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.

- Tecnologia open source più diffusa e utilizzata: **Apache Hadoop**.

Big Data: le 5 V - Velocità

Big Data: le 5 V - Velocità

Velocità: è una caratteristica che ha più di un significato.

- Si riferisce in primis alla **elevata frequenza con cui i dati vengono generati** – si ripercuote sulla quantità (Volume).
- Il secondo aspetto riguarda la **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 competitività sui diversi panorami del mercato

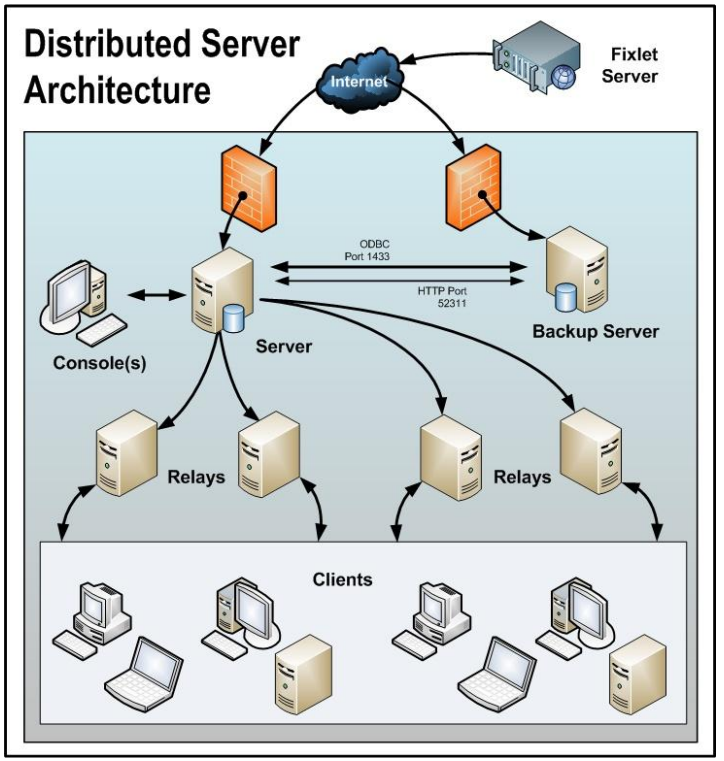
Quali tecnologie?!

Big Data: le 5 V - Velocità

Velocità: è una caratteristica che ha più di un significato.

- Particolarmente adatte sono le **architetture distribuite.**

- **Gestione** di **strutture dati anche complesse**.
- **Accesso ai dati in tempo reale**.
- **Velocità di elaborazione** grazie a tecniche di calcolo distribuito.
- Database non relazionali come i ***column DB*** e ***key/value DB*** (NoSQL).



Big Data: le 5 V - Varietà

Varietà: caratteristica che ha a che fare con la forma in cui i dati si presentano.

- Nel contesto **Big Data** le **informazioni** da trattare sono dati non-strutturati (o semi-strutturati). Non 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.

- Per la gestione e il salvataggio di questi dati si ricorre spesso ai ***database NoSQL***. Non impongono uno schema rigido per organizzare i dati (*schemaless database*).

Big Data: le 5 V - Variabilità

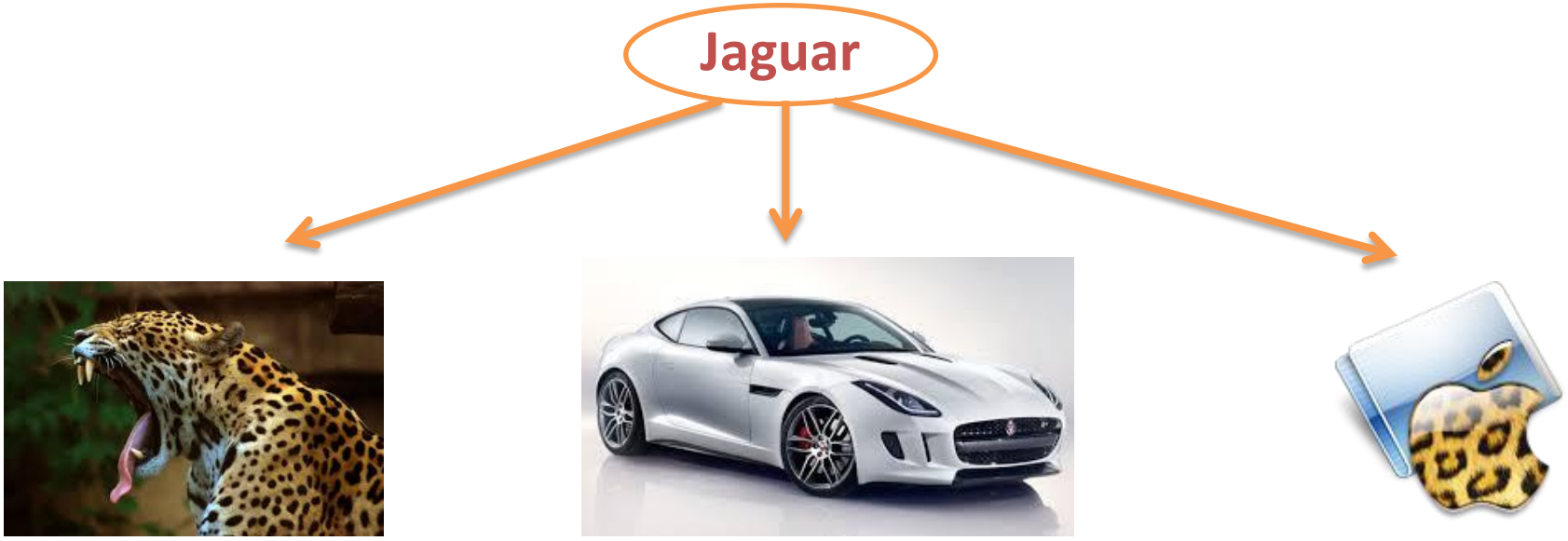
Variabilità: caratteristica relativa alla **contestualizzazione** di un dato.

- Il significato o l'**interpretazione di uno stesso dato** può **variare** in base al contesto in cui esso viene raccolto e analizzato.
- Esempio la frase "*leggete il libro*", essa avrà un **significato positivo** in un blog che parla di letteratura, mentre avrà 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: le 5 V - Variabilità

Variabilità: caratteristica relativa alla **contestualizzazione** di un dato.

E' importante trovare dei meccanismi che riescano a dare una **semantica ai dati** in base al contesto in cui sono espressi.



Animale **Automobile** **Vecchio OS Apple**

Big Data: le 5 V - Viralità

Viralità: caratteristica che ha a che fare su **quanto e come i dati si diffondono (**Propagazione dei dati**)**.

- La **grande quantità** di dati (spesso correlati tra loro) e **l'alta velocità con cui sono prodotti** implica una **diffusione virale** delle informazioni.

- **Esempio**: **una notizia o un evento** diffusi tra diversi canali. **Diffusione amplificata** con i collegamenti nei vari **social network**.

Big Data: le 5 V - Viralità

Istituzioni e alcune organizzazioni sfruttano questa caratteristica/potenzialità per migliore attività di pronto intervento.

Big Data: le 5 V - Viralità

Virale è anche la **crescita del Volume** dei dati generati dalle attività digitali dell'uomo (user-generated content):

- Nel **2010** è stata stimata una produzione di **1,2 zettabyte** di dati (**1ZB corrisponde a mille miliardi di GB**).
- Nel **2011** è cresciuta a **1,8ZB**.
- Nel **2013** si è arrivati a **2,7ZB**.
- La proiezione per il **2015** parla di **8ZB**.

Big Data: le 5 V - Viralità

MULTIPLI DEL BYTE

Nome	Simbolo	Multiplo	byte
Kilobyte	kB	10^3	1.000
Megabyte	MB	10^6	1.000.000
Gigabyte	GB	10^9	1.000.000.000
Terabyte	TB	10^{12}	1.000.000.000.000
Petabyte	PB	10^{15}	1.000.000.000.000.000
Exabyte	EB	10^{18}	1.000.000.000.000.000.000
Zettabyte	ZB	10^{21}	1.000.000.000.000.000.000.000
Yottabyte	YB	10^{24}	1.000.000.000.000.000.000.000.000

Tabella dei nomi e simboli dei multipli del byte

Big Data: le 5 V – Viralità (Curiosità)

Buzzsumo, una società di analisi dati, ha analizzato recentemente *milioni di contenuti in rete* per capire quali sono le **caratteristiche che rendono un contenuto virale**.

Dimensioni

- **Maggiore è la lunghezza** dei contenuti, **maggiori saranno le condivisioni**. Contenuti lunghi e ricchi di informazioni (***> 2000 parole***) ottengono più share rispetto ai contenuti brevi (***< 2000 parole***).

Emozioni

- Un contenuto deve *generare emozioni*, **le persone amano condividere elementi che facciano ridere e stupiscano i lettori** (42% dei contenuti studiati). Di contro, le emozioni ***meno gradite*** sono la ***tristezza e la paura***, che arrivano al 7%.

Big Data: le 5 V – Viralità (Curiosità)

Buzzsumo, una società di analisi dati, ha analizzato recentemente *milioni di contenuti in rete* per capire quali sono le **caratteristiche che rendono un contenuto virale**.

Immagini

- **I contenuti visivi attirano l'attenzione** degli utenti, favoriscono una **comprensione immediata** e quindi tendono ad avere **maggiori interazioni**. Per la loro natura, ***le immagini aumentano le condivisioni sui social.***

- **Il 65%** delle persone usa **Facebook** per condividere **post che contengano almeno un'immagine**. Questi post sono quelli con cui poi si interagisce maggiormente.

- **Più del 20%** degli **utenti su Twitter** preferisce **pubblicare contenuti in cui sia presente un'immagine.**

Big Data: le 5 V – Viralità (Curiosità)

Buzzsumo, una società di analisi dati, ha analizzato recentemente *milioni di contenuti in rete* per capire quali sono le **caratteristiche che rendono un contenuto virale**.

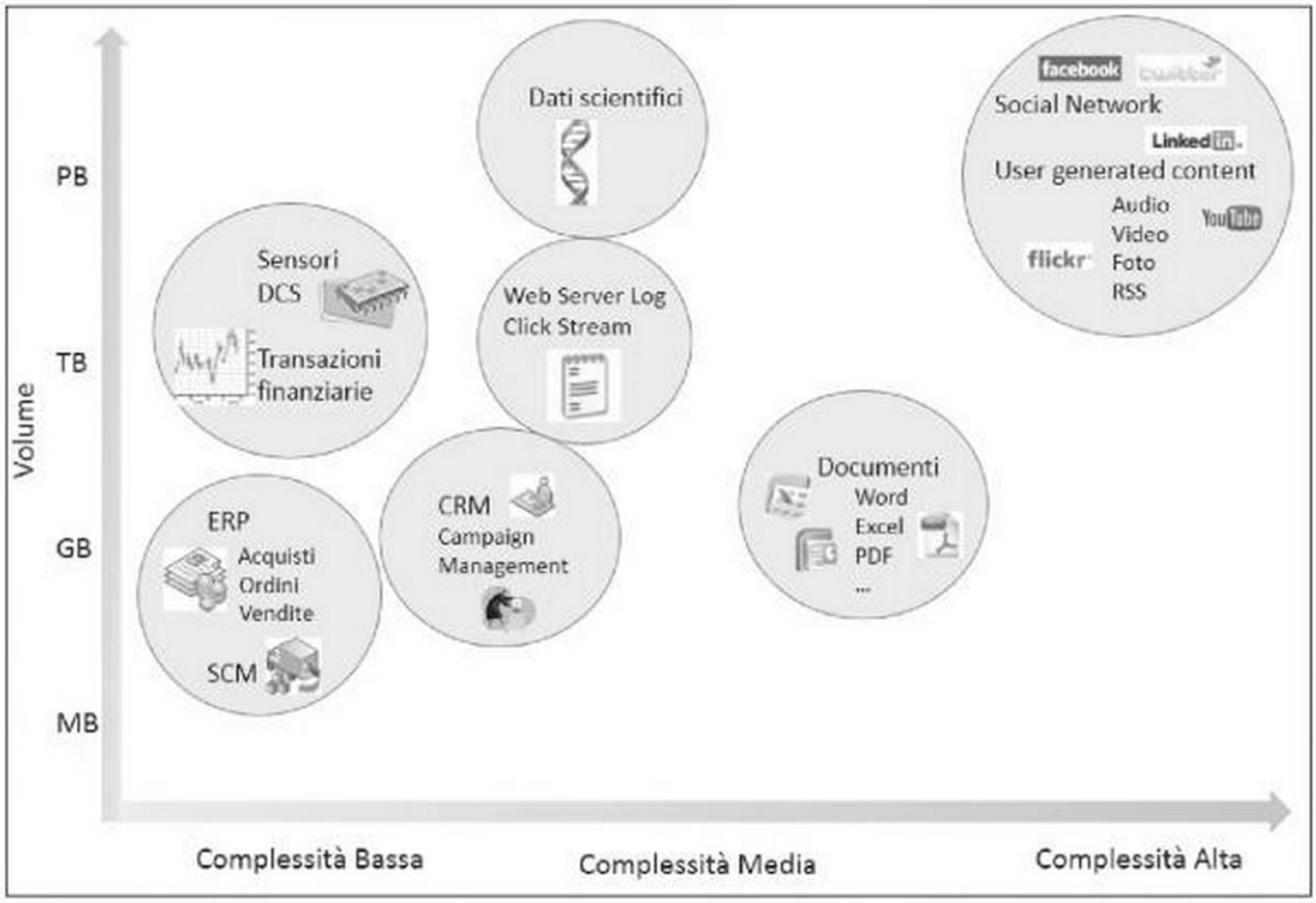
Elenchi puntati

- **Lettori e utenti web amano gli elenchi puntati, le infografiche e gli how to**. Ciò dipende dal fatto che questi contenuti permettono di **sintetizzare in forma visiva gli aspetti salienti di un post**, facilitando la comprensione.

Influencer

- Contenuti condivisi da **persone, organizzazioni e aziende ritenuti "esperti" in uno specifico settore** raggiunge un maggior numero di **utenti "targettizzati" e interessati a determinate informazioni**.

Big Data: le 5 V



Classificazione dei dati per volume e complessità

Big Data: le 5 V - Valore

Valore: è necessario comprendere e gestire in modo adeguato i dati e tutti questi aspetti ad essi legati in modo da riuscire ad **estrarre il potenziale informativo**.

- I Big Data **nascondono un grande valore**. Al primo utilizzo di solito se ne estrae soltanto una parte, il *valore rimanente rimane "dormiente"* fino ad un successivo utilizzo.

- E' quindi importante adottare metodologie e tecnologie che permettano la **continua integrazione di nuove informazioni**, in seguito ad un utilizzo reiterato, con l'obiettivo di **costruire una base di conoscenza sempre più ampia**.

Problematiche

- **Elevato numero di campi applicativi** diversi tra loro.
- I **differenti canali** attraverso i quali i dati vengono raccolti.
- Identificare una possibile **architettura adattabile** a tutte le aree.
- Come è possibile scoprire il "**Valore**" dei Big Data?

- **Utilizzo di complesse analisi e processi di modellazione.**
- **Formulazione di ipotesi -> implementazione di modelli semantici, visuali e statistici -> validazione.**

Indice

1. **Big Data**
   - *Evoluzione dei dati e delle tecniche di analisi*
   - *Le 5V dei Big Data*
   - *Teorema CAP*
   - *Pipeline dell'analisi dei Big Data*

2. **Principali Contesti Applicativi**

3. **Criticità e Rischi dei Big Data**

Teorema di Brewer (o Teorema **CAP**)

- Il **Teorema CAP** (Consistency – Availability – Partition tolerance) è fondamentale per capire il comportamento di **sistemi SW distribuiti**, e progettarne l'architettura in modo da rispettare requisiti non funzionali stringenti, tra cui:
  - ***Elevate prestazioni.***
  - ***Continua disponibilità.***
  - ***Sistemi geograficamente distribuiti.***
- Il **Web 2.0**, è popolato da applicazioni che lavorano su bilioni e trilioni di dati ogni giorno . La **scalabilità** è un concetto chiave.
- A tal proposito si stanno sviluppando **database che sono distribuiti** sulla rete per realizzare una ***scalabilità orizzontale***.

Teorema di Brewer (o Teorema **CAP**)

***Il Teorema CAP** afferma*

"sebbene sia altamente desiderabile per un sistema software distribuito fornire simultaneamente **totale coerenza** (*Consistency*), **continua disponibilità** (*Availability*) e **tolleranza alle partizioni** (*Partition tolerance*), ciò **non è possibile.**

E' necessario stabilire, di volta in volta in funzione dei requisiti di una specifica applicazione, quali di queste tre garanzie sacrificare.

- E' importante tenere a mente questo teorema, perchè specie in applicazioni Web2.0, fornire agli utenti una pessima esperienza può avere una diffusione virale a causa dei vari social network (fonti Amazon e Google).

Teorema di Brewer (o Teorema **CAP**)

Consistency *(Totale coerenza)*
Un sistema distribuito è **completamente coerente** se preso un dato che viene scritto su un **nodoA** e viene letto da un altro **nodoB,** il sistema ritornerà l'ultimo valore scritto (quello *consistente*).

- Se si considera la cache di un singolo nodo la **totale consistenza è garantita**, così come la tolleranza alle partizioni.
- **Non** si hanno però **sufficiente disponibilità** (fault-tolerance) e **buone performance**.

- Se la **cache è distribuita** su due o più nodi, aumenta la disponibilità, ma vanno **previsti dei meccanismi complessi** che permettano ad ogni nodo di accedere ad un repository virtuale distribuito (e leggere lo stesso valore di dato).

Teorema di Brewer (o Teorema **CAP**)

Availability *(disponibilità)*
Un sistema (distribuito) è **continuamente disponibile** se ogni nodo **è sempre in grado di rispondere ad una query o erogare i propri servizi** a meno che non sia *indisponibile.*

- Banalmente **un singolo nodo non garantisce la continua disponibilità**.

- Una **cache distribuita** mantiene nei vari nodi delle aree di backup in cui sono memorizzati i dati presenti su altri nodi.

- Per realizzare la **continua disponibilità** si ricorre alla **ridondanza dei dati (su più nodi)**. Ciò però richiede meccanismi per garantire la consistenza e problematiche riguardo la tolleranza alle partizioni.

Teorema di Brewer (o Teorema **CAP**)

Partition-Tolerance *(Tolleranza alle partizioni)*
È la capacità di un sistema di essere tollerante ad una aggiunta o una rimozione di un nodo nel sistema distribuito (*partizionamento*) o alla perdita di messaggi sulla rete.

1. Si consideri una **configurazione** in cui **un solo cluster** è composto da **nodi su due diversi data center**.
2. Supponiamo che i **data center perdano la connettività di rete**. I **nodi** del cluster **non riescono più a sincronizzare** lo stato del sistema.
3. I **nodi si riorganizzano in sotto-cluster**, tagliando fuori quelli dell'altro data center.

Il sistema continuerà a funzionare, in modo non coordinato e con possibile perdita di dati (es. assegnazione della stessa prenotazione a clienti diversi).

Teorema di Brewer (o Teorema **CAP**)

Poiché non è possibile garantire simultaneamente **completa consistenza**, **continua disponibilità** e **tolleranza alle partizioni,** quando si progetta un sistema distribuito è necessario valutare attentamente quale soluzione di compromesso accettare tra le seguenti coppie possibili **CA**, **CP** e **AP**

Consistency/Availability (CA)

- E' il compromesso offerto solitamente dai **RDBMS.**
- **I dati sono coerenti** su tutti i nodi (attivi e disponibili).
- **Scritture/letture sempre possibili**, e dati aggiornati propagati tra i nodi del cluster (**dati sempre aggiornati**).

(-) Possibili problemi legati alle performance e alla scalabilità.
(-) Possibile disallineamento tra i dati nel caso di partizioni di nodi.

Teorema di Brewer (o Teorema **CAP**)

Poiché non è possibile garantire simultaneamente **completa consistenza**, **continua disponibilità** e **tolleranza alle partizioni,** quando si progetta un sistema distribuito è necessario valutare attentamente quale soluzione di compromesso accettare tra le seguenti coppie possibili **CA**, **CP** e **AP**

Consistency/Partition-Tolerance (CP)

- Compromesso preferito da soluzioni come **Hbase, MongoDB, BigTable.**
- **I dati sono coerenti** su tutti i nodi e sono garantite le partizioni, assicurando la sincronizzazione dei dati.

(-) Possibili problemi di disponibilità, dati non più disponibili se un nodo va giù.

Teorema di Brewer (o Teorema **CAP**)

Poiché non è possibile garantire simultaneamente **completa consistenza**, **continua disponibilità** e **tolleranza alle partizioni,** quando si progetta un sistema distribuito è necessario valutare attentamente quale soluzione di compromesso accettare tra le seguenti coppie possibili **CA**, **CP** e **AP**

Availability/Partition-Tolerance (AP)

- Compromesso usato da soluzioni come **CouchDB, Riak, Apache Cassandra.**
- **I nodi restano online** anche se impossibilitati a parlarsi.
- E' necessario un processo di risincronizzazione dei dati per eliminare eventuali conflitti quando la partizione è risolta.

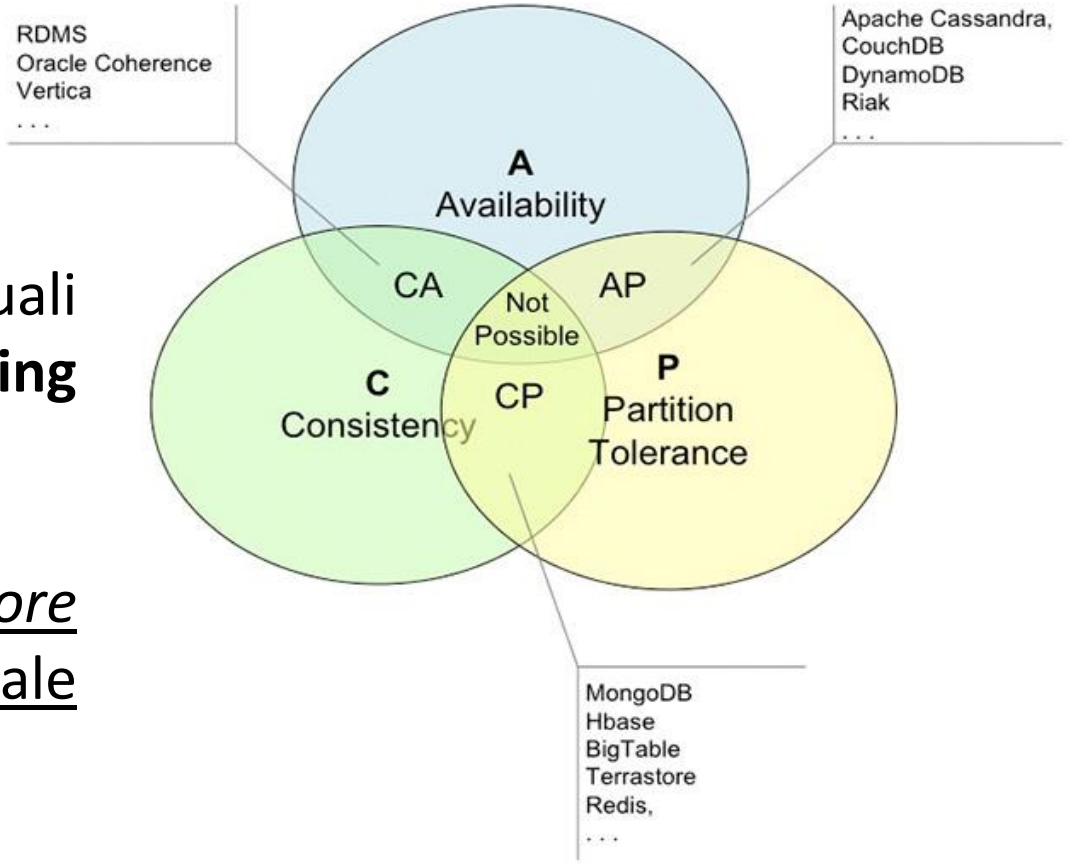
(+) Buone prestazioni in termini di latenza e scalabilità.

Teorema di Brewer (o Teorema **CAP**)

Osservazione

La maggior parte delle attuali soluzioni prevedono il **tuning** della modalità operativa.

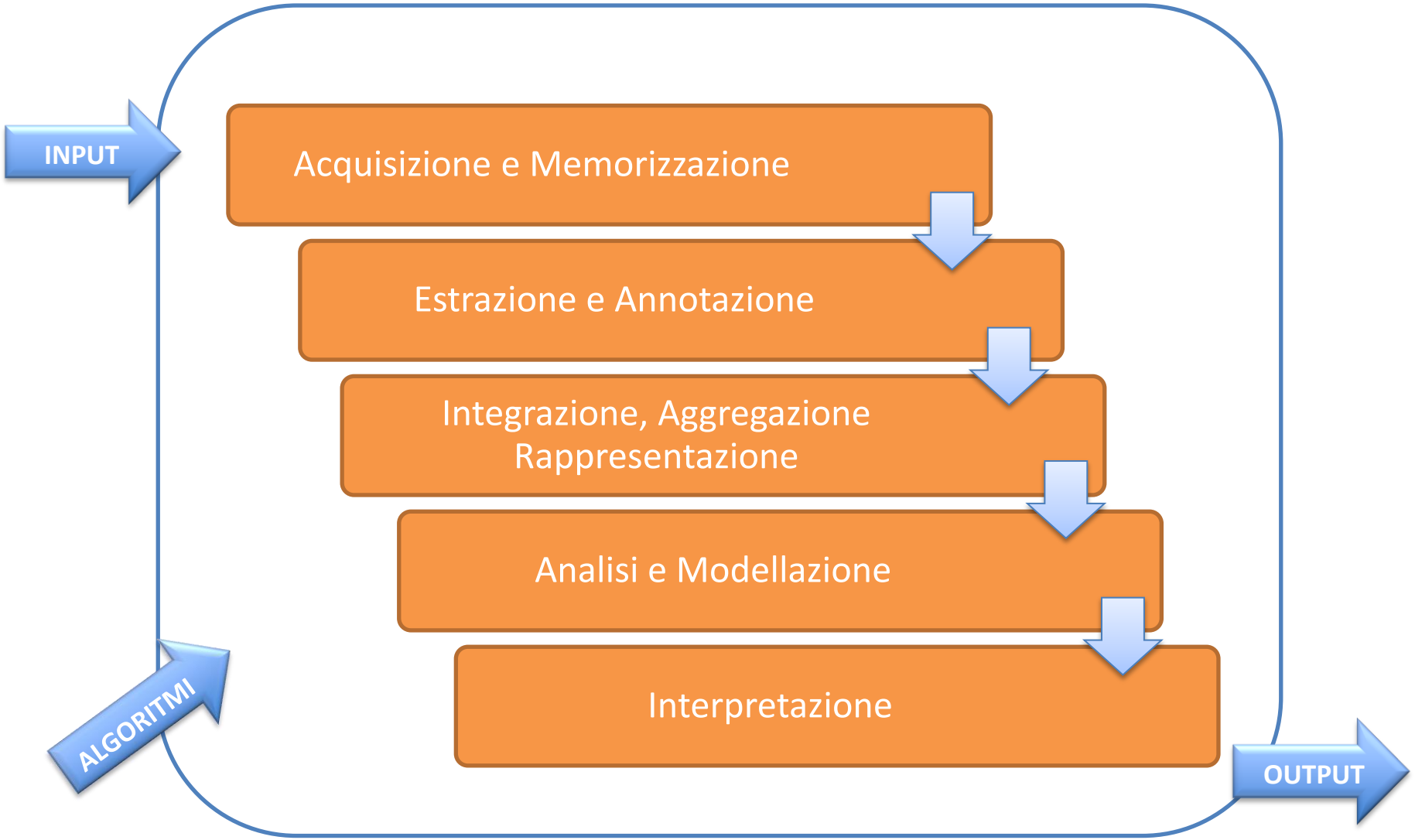
<u>Cioè lasciano allo *sviluppatore* la possibilità di scegliere quale *garanzia sacrificare*.</u>



Indice

1. **Big Data**
   - *Evoluzione dei dati e delle tecniche di analisi*
   - *Le 5V dei Big Data*
   - *Teorema CAP*
   - *Pipeline dell'analisi dei Big Data* 
2. **Principali Contesti Applicativi**
3. **Criticità e Rischi dei Big Data**

Pipeline di Analisi dei Big Data



Pipeline: Acquisizione dei Dati e Memorizzazione

- Grandi quantità di dati possono essere **filtrati e compressi** a diversi ordini di grandezza.
  - Sfida: Definire dei filtri opportuni in modo che non vadano perse informazioni di interesse.

- **Dettagli** inerenti a **condizioni sperimentali e procedure** possono essere richiesti per interpretare i risultati correttamente.
  - Sfida: Generazione automatica dei metadata corretti.

- Possibilità di ricerca sia all'interno dei metadata che nei dati di sistema.
  - Sfida: Creare e utilizzare delle strutture dati ottimizzate che consentano le ricerche in tempi accettabili.

Pipeline: Estrazione delle Informazioni e Pulizia

- Le **informazioni** raccolte spesso **non** sono in un **formato pronto per l'analisi** (es. immagini di sorveglianza VS immagini scattate da fotografi)
  Sfida: Realizzare un processo di estrazione delle informazioni che le fornisca in un formato adatto alla fase di analisi.

- I Big Data sono **incompleti** a causa di **errori** commessi durante la fase di acquisizione.
  Sfida: Definire dei vincoli e modelli per la gestione e correzione automatica di errori in diversi domini Big Data.

Pipeline: Integrazione, Aggregazione e Rappresentazione dei Dati

- I **Dati sono eterogenei** e può non essere abbastanza raccoglierli all'interno di repository.
  Sfida: Creare delle strutture dati di memorizzazione che siano in grado di adattarsi alle differenze nei dettagli sperimentali.

- I **modi di memorizzare dati** sono **diversi**, alcuni modelli hanno dei vantaggi rispetto ad altri per determinati scopi.
  Sfida: Creare dei tool di supporto al processo di progettazione dei database e alle tecniche di sviluppo tenendo conto del contesto applicativo e d'uso dei dati.

Pipeline: Query, Modellazione Dati e Analisi

- I **metodi per investigare e interrogare** i Big Data sono **differenti** dalle tradizionali analisi statistiche.

  Sfida: Creare delle tecniche per l'elaborazione di query complesse e scalabili (sull'ordine dei TeraByte), considerando delle risposte interattive nel tempo.

- **Big Data interconnessi** formano delle **reti di dati eterogenee**, in cui la **ridondanza dei dati** può essere sfruttata per compensare l'assenza di alcune informazioni, per verificare situazioni di conflitto e evitare che ci siano relazioni nascoste.

  Sfida: Rendere coordinati i sistemi DB e le interrogazioni SQL, con i tool di analisi che realizzano diverse forme di elaborazione non-SQL (data mining, analisi statistica).

Nuovi Concetti

Datification

- Prendere **informazioni su qualsiasi cosa** e **trasformarle in un qualsiasi formato dati** in modo da renderle **quantificabili.**

- **Utilizzare queste informazioni** in un nuovo modo con l'obiettivo di **tirar fuori il loro valore implicito e nascosto**.

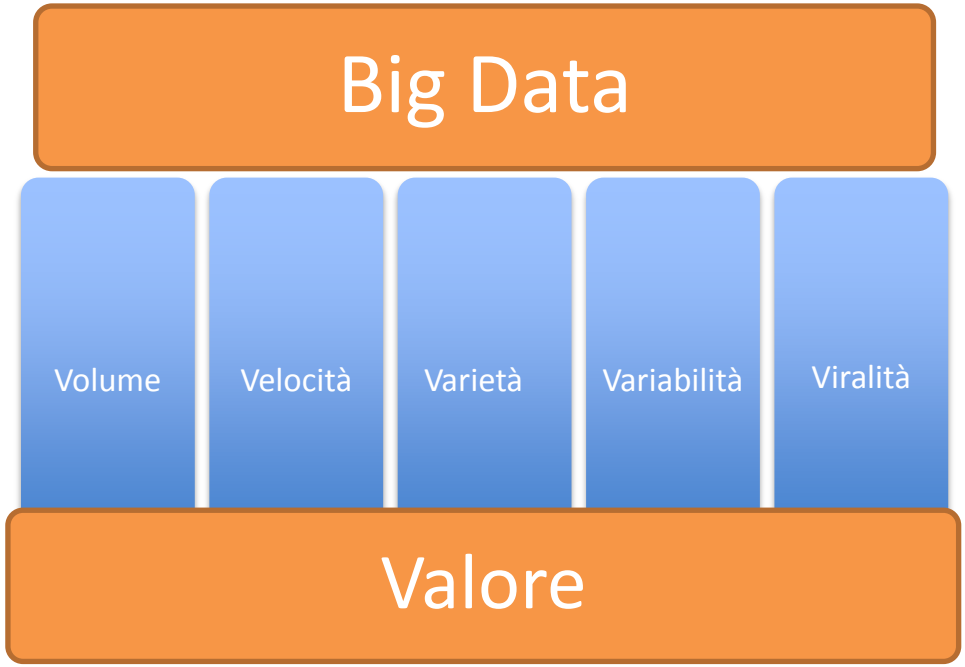
- *Quando i dati sono pochi è desiderabile che siano accurati* (campionamento random). I Big Data hanno cambiato il concetto di aspettativa della precisione: *Trattare queste grandi quantità di dati spesso imprecise e imperfette permette di fare delle previsioni superiori* (Analisi Predittiva).

Indice

1. **Big Data**
   - *Evoluzione dei dati e delle tecniche di analisi*
   - *Le 5V dei Big Data*
   - *Teorema CAP*
   - *Pipeline dell'analisi dei Big Data*
2. **Principali Contesti Applicativi** 
3. **Criticità e Rischi dei Big Data**

Campi di Applicazione

- Il Problema dei Big Data si riferisce alla combinazione di un **grande volume di dati** che deve essere **trattato in tempi abbastanza rapidi**.



- Sono molte aree applicative in cui i Big Data sono attualmente utilizzati con risultati interessanti ed **eccellenti prospettive future** per affrontare le **principali sfide** come **Analisi dei Dati**, **Modellazione**, **Organizzazione** e **Ricerca** (Data Retrieval).

Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- **Sanità e Medicina**
- Ricerca Scientifica (Analisi dei Dati)
- Istruzione
- Settore Energetico e dei Trasporti
- Social Network – Servizi Internet – Web Data
- Finanza/Business – Marketing
- Sicurezza

Sanità e Medicina

- Nel campo Medico/Sanitario molte delle informazioni raccolte provengono da:
  - Fascicolo Sanitario Elettronico (FSE)
  - Sintomatologie
  - Diagnosi
  - Terapie e Risposte a trattamenti

- In **12 giorni** circa **5000 pazienti** arrivano in un pronto soccorso.

- Nella ricerca medica due importanti applicazioni sono:
  - **Analisi di sequenze genomiche** (In un singolo esperimento sono coinvolte circa 100 milioni di piccole sequenze).
  - **Analisi delle neuroimmagini** (Memorizzazione di dati intermedi ~1.8 PetaBytes).

Sanità e Medicina

- I **processi ospedalieri** sono caratterizzati dal fatto che spesso **diversi reparti e unità sono coinvolti** nel trattamento di un paziente, a volte ognuno di essi ha delle **proprie applicazioni IT**.

- Allo stesso tempo, specie nella Sanità pubblica, c'è una forte pressione da parte degli enti governativi di **rivedere i processi clinici** con l'obiettivo di **migliorarne l'efficienza e ridurne i costi**.

Una possibile soluzione è quella di utilizzare **dati "real-time"** in modo da **supportare analisi e decisioni dei processi esistenti**, applicativi e strutture dati comuni.

Sanità e Medicina

- Implementare **Tecniche di Data Mining** per **estrarre conoscenza** da questi dati, ad esempio per identificare nuovi interessanti **modelli nello sviluppo delle infezioni** o definire delle pratiche di intervento.

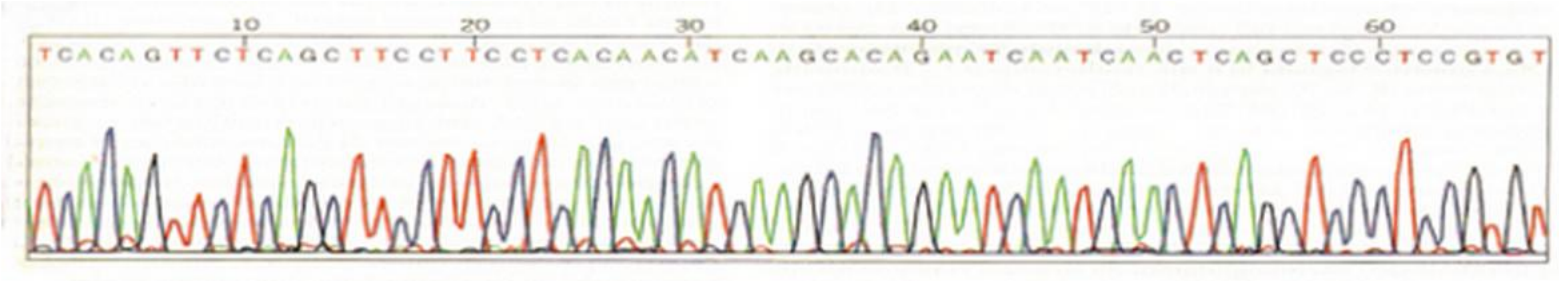
- Il **Processo di Mining** si realizza attraverso **tecniche di analisi e valutazione di eventi/processi memorizzati in file di log**. Negli ospedali con il FSE sono state studiate tecniche per l'accesso rapido e l'estrazione di informazioni dai log di eventi, in modo da produrre modelli facilmente interpretabili, attraverso tecniche di **partizionamento**, **clustering** e **pre-elaborazione**.

- La costruzione di un **modello predittivo**, potrebbe essere utile per fornire **supporto decisionale** nel triage e in diagnosi specifiche o per la produzione di piani efficaci per la gestione delle malattie croniche, migliorare la qualità dell'assistenza sanitaria e abbassarne i costi.

Sanità e Medicina

- Alla base della genomica ci sono tecniche di clonazione di geni e **sequenziamento del DNA**, con l'obiettivo di conoscere l'intero **genoma** degli organismi.

- La conoscenza dell'intero genoma permette di identificare più facilmente i **geni coinvolti** e osservare **come questi interagiscono**, in particolare nel caso di malattie complesse come **tumori.**



Sanità e Medicina

Grandi quantità di dati, conoscenza genetica, pratiche cliniche consolidate, appropriati DB

Permettono di definire **un'ampia base di conoscenza** con la quale è possibile **effettuare studi predittivi** sull'incidenza di alcune malattie

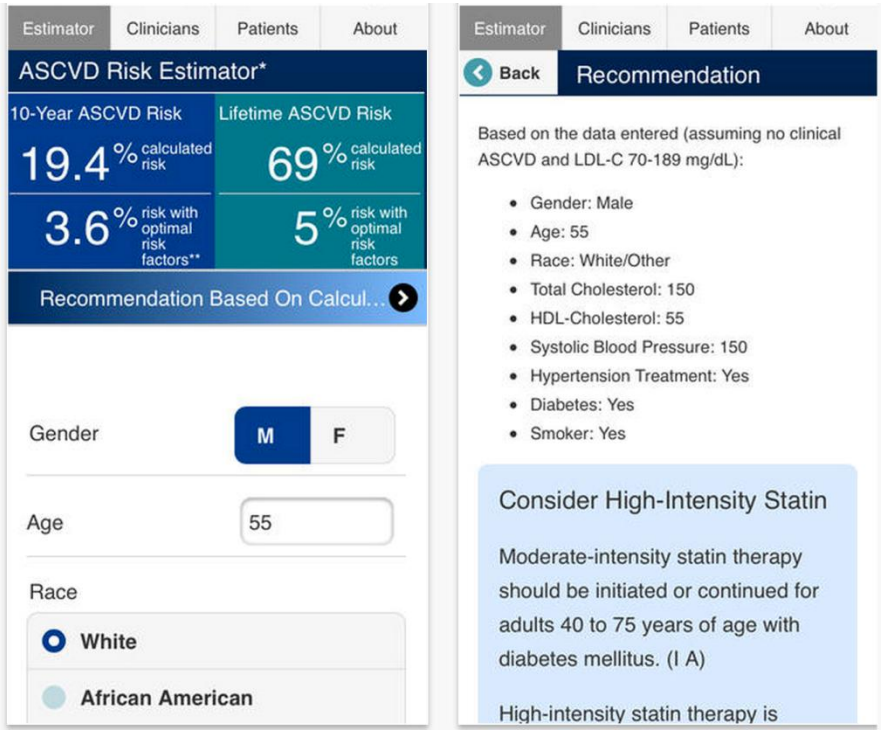
- Una interessante opportunità sono i **Gk-arrays** (combinazione di tre array), soluzione intelligente per indicizzare grandi collezioni di piccole sequenza genomiche.

http://www.atgc-montpellier.fr/gkarrays

Sanità e Medicina

- **Ascvd Risk Estimator** è un'app lanciata dall'***America College of Cardiology e The American Heart Association***.

- Consente di **monitorare in modo costante i rischi di infarto** e problemi cardiovascolari dei pazienti in dieci anni. L'app **raccoglie informazioni relative al paziente** fra cui ***età, sesso, razza, colesterolo, ipertensione, pressione sanguigna***. I medici, analizzano questi dati, **stimano le possibilità di rischio e poi comunicano ai loro pazienti le cure e le terapie da seguire**, secondo un processo evolutivo ed in continuo aggiornamento.



Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- Sanità e Medicina
- **Ricerca Scientifica (Analisi dei Dati)** 
- Istruzione
- Settore Energetico e dei Trasporti
- Social Network – Servizi Internet – Web Data
- Finanza/Business – Marketing
- Sicurezza

Ricerca Scientifica e Dati sperimentali

Ci sono diverse aree della ricerca scientifica in cui si può parlare di Big Data, l'obiettivo dell'analisi dei dati è quello di **estrarre significato dai dati** e determinare le **azioni da intraprendere**.

Astronomia (Osservazione Automatica del Cielo)

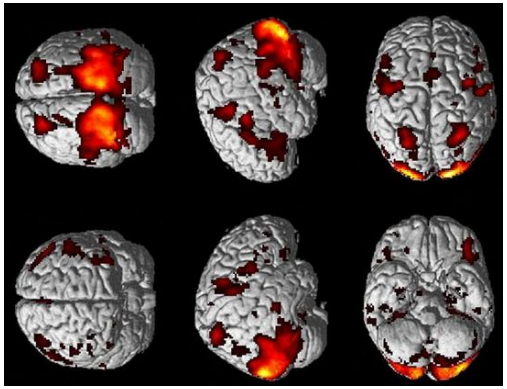
- Circa **200 GB** di nuovi dati ottici ad alta risoluzione vengono catturati ogni sera da dispositivi ad accoppiamento di carica (CCD) collegati a telescopi.

Sociologia (Analisi dei Web log e di dati comportamentali)

- Circa **2 Mld e mezzo** di utenti internet nel mondo.
- Circa **5h di navigazione online tradizionale** (pc e notebook) e **2h di navigazione mobile** giornaliera solo in Italia.

Ricerca Scientifica e Dati sperimentali

Ci sono diverse aree della ricerca scientifica, in cui l'obiettivo dell'analisi dei Big Data è quello di **estrarre significato dai dati** e determinare le **azioni da intraprendere**.



Biologia
(Sequenziamento DNA – Codifica dei geni)

Neuroscienze
(Aspetti molecolari, cellulari, Neuroimaging funzionale)

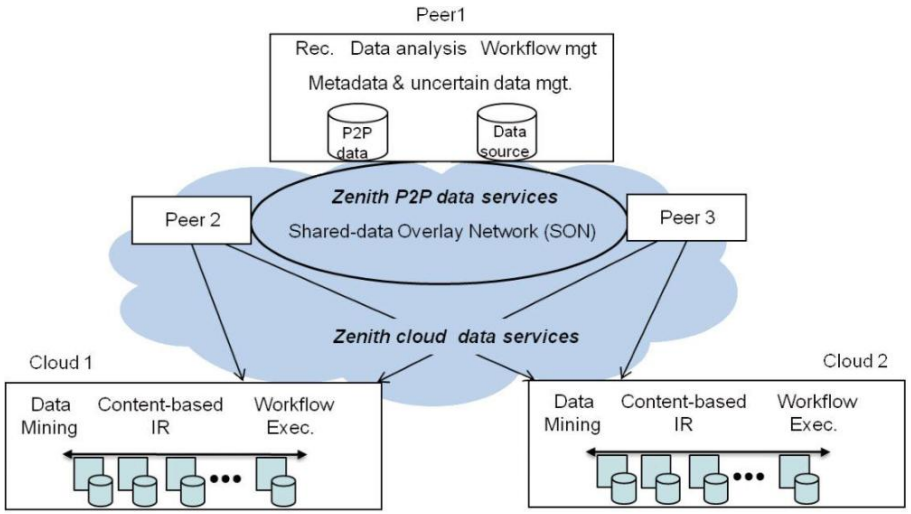
- La ricerca scientifica è caratterizzata dall'elevata **collaboratività**, team formati da scienziati di diverse nazioni e specializzati in diverse discipline.

Ricerca Scientifica e Dati sperimentali

- Per far fronte alla grande quantità di dati sperimentali prodotti da discipline moderne, l'Università Montpellier avviato il progetto **ZENITH**.

- Zenith adotta un'**architettura ibrida p2p/cloud.**

- Natura collaborativa dell'attività di ricerca. L'idea è di adottare un approccio p2p per la condivisione dei dati, mantenendo comunque un controllo decentralizzato, e di sfruttare le pontenzialità del cloud in termini di computazione memorizzazione nell'elaborazione di quantità di dati considerevoli.



http://www.sop.inria.fr/teams/zenith

Ricerca Scientifica e Dati sperimentali

- **Europeana** è una piattaforma (service platform) per la gestione e la condivisione di contenuti multimediali (Testi, Video, Immagini).

- **Milioni di contenuti sono indicizzati** e quindi posso essere ricercati in tempo reale.

- Dati e risorse inizialmente modellati attraverso un **metadata model** denominato **ESE** (Europeana Semanitc Elements).

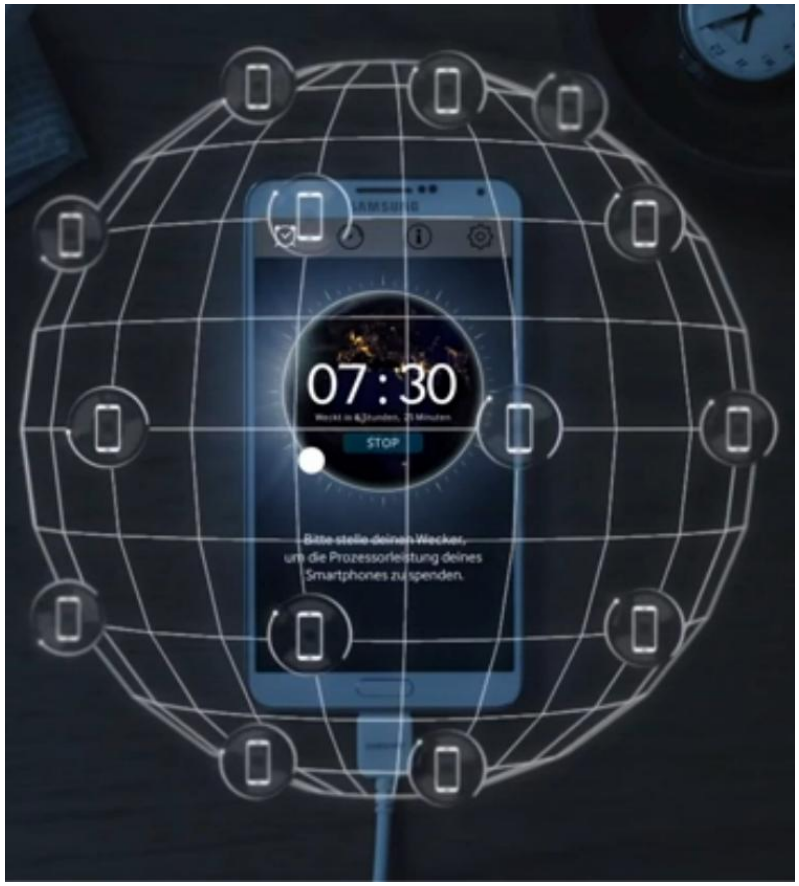
- E' in via di adozione un nuovo modello più complesso che include un **set di relazioni semantiche**, denominato **EDM** (Europeana Data Model)**.**

europeana
think culture

www.europeana.eu

Ricerca Scientifica e Dati sperimentali

- **Samsung Power Sleep,** applicazione *Android* sviluppata da *Samsung Austria e dall'Università di Vienna*.
- L'utente **imposta l'ora della sveglia** sull'app e mette il **telefono sotto carica** con **Wi-Fi attivo**.
- **Power Sleep elabora i dati** e li **invia a uno specifico database**, *Similarity Matrix of Protein* (***Simap***): qui vengono decifrate sequenze di proteine utili per varie ricerche in campo medico-scientifico, fra cui genetica, biochimica, contro il cancro e l'Alzheimer. Tutto ciò é possibile perché ***l'app è connessa al Berkeley Open Infrastructure Network Computing*** (**Boinc**), che oltre a connettere i computer di tutto il mondo con lo scopo, appunto, di elaborare dati scientifici, ora collega anche i mobile devices.



Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- Sanità e Medicina
- Ricerca Scientifica (Analisi dei Dati)
- **Istruzione**
- Settore Energetico e dei Trasporti
- Social Network – Servizi Internet – Web Data
- Finanza/Business – Marketing
- Sicurezza

Istruzione

- I Big Data hanno il potere di **rivoluzionare** non solo la ricerca scientifica, ma anche il sistema di **Istruzione.**

- Alcuni dei principali dati nel campo dell'istruzione sono:
  - Performance degli studenti (Project KDD 2010*)
  - Meccanismi di apprendimento
  - Risposte a diverse strategie pedagogiche

Un **nuovo approccio di insegnamento** può essere definito sfruttando la gestione dei Big Data

Big Data utilizzati per definire dei **modelli che permettano di capire** le **conoscenze** attuali degli studenti (come aumentarle), e i loro **progressi**.

*** https://pslcdatashop.web.cmu.edu/KDDCup/**

Istruzione

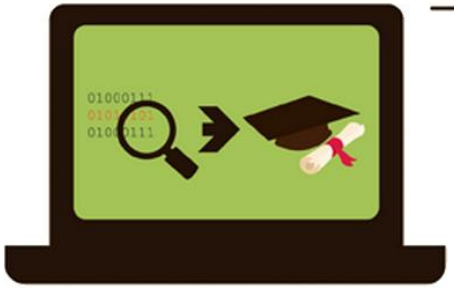
- I **nuovi modelli di insegnamento** sfruttano le potenzialità dell'**informatica** e della **tecnologia** combinate con **tecniche di analisi dei dati.**
Obiettivo: risolvere le problematiche considerando aspetti pedagogici, meccanismi psicologici e di apprendimento, e definire **un'istruzione personalizzata**, soddisfacendo le esigenze dei singoli o di gruppi di studenti.

- Altro campo di interesse in questo contesto è l'**e-learnig**, in cui sono definiti due principali tipologie di utenti:

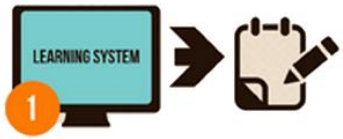
- Tutti i dettagli personali dei **learners** e le informazioni fornite dai **learning providers** sono memorizzate in appositi DB.

- Applicando opportune tecniche di data mining è possibile attuare programmi di insegnamento **customizzati** sugli effettivi interessi e bisogni degli studenti.

Istruzione

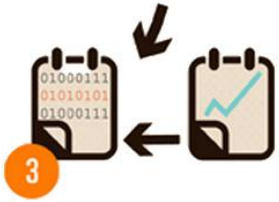


come LAVORA QUESTO APPROCCIO?



Un sistema di apprendimento interagisce con uno studente, fornendo contenuti e raccogliendo risposte e dati personali.

Dati dettagliati relativi all'esperienza dello studente sono raccolti e memorizzati in un DB.



Questi dati sono usati per fare delle previsioni sulle future performance dello studente.

Predizioni e feedback sono visualizzati sulla console di monitoraggio e analisi.

Gli studenti ricevono il materiale didattico adeguato al loro livello di apprendimento e ai propri interessi

Insegnanti, tutor e sviluppatori possono intervenire per aiutare in base ai vari bisogni.

Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- Sanità e Medicina
- Ricerca Scientifica (Analisi dei Dati)
- Istruzione
- **Settore Energetico e dei Trasporti**
- Social Network – Servizi Internet – Web Data
- Finanza/Business – Marketing
- Sicurezza

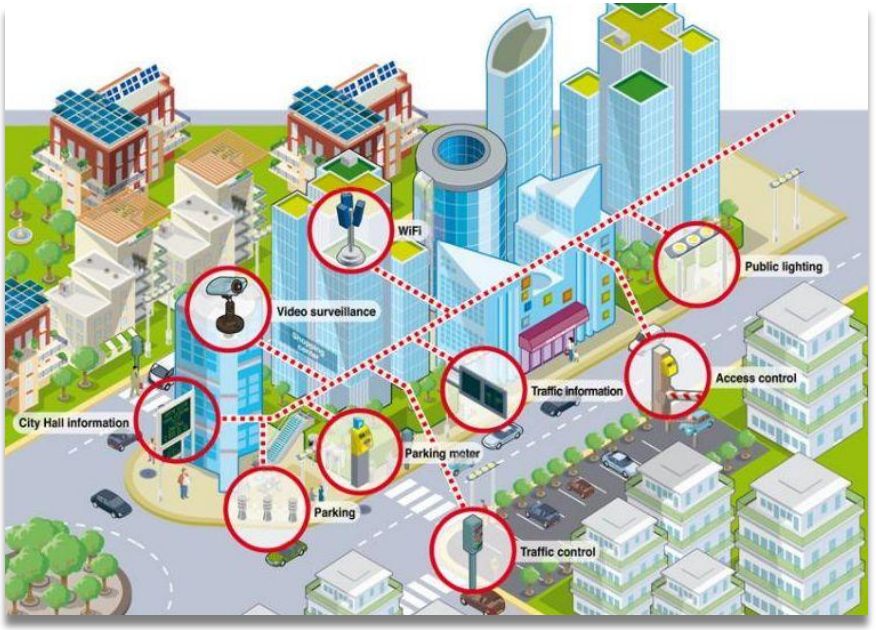
Energia e Trasporti

Gli **spostamenti delle persone** (trasporto pubblico e privato) tra e all'interno delle **aree metropolitane** è uno dei fattori chiave relativamente alla qualità della vita e **coinvolge una grande varietà e quantità di dati**.

- **Organizzazione del servizio di treni, tram e autobus**
  - Circa **33.000 turisti** visitano Roma ogni giorno
  - Circa **1.600.000 lavoratori** arrivano a Roma ogni giorno
- **Informazioni GPS** (Autobus, taxi, infopoint, PI, etc.)
- **Interruzioni del traffico** (informazioni temporanee)
- **Dati metereologici**
- **Sensori parcheggi e servizi sharing mobility** (RFID)
- **Orari apertura/chiusura di attività e servizi**

Energia e Trasporti

- Un **approccio data-centrico** può aiutare ad incrementare l'efficienza e l'affidabilità del sistema di trasporto Pubblico (e privato).



- L'ottimizzazione di **infrastrutture di trasporto multimodale** e il loro utilizzo intelligente può **migliorare l'esperienza di viaggio** e l'efficienza operazionale, con un impatto positivo anche sui costi e problematiche ambientali.

Energia e Trasporti

- Attraverso **l'analisi e la visualizzazione** di dati dettagliati della rete stradale e mediante **l'uso di un modello predittivo** è possibile realizzare un ambiente di trasporto intelligente.

- L'integrazione di **dati geografici** memorizzati di elevata precisione con i **dati real-time provenienti da reti di sensori** sparsi, può favorire la realizzazione di un efficiente sistema di pianificazione urbana che mescola **trasporto pubblico e privato**, offrendo alle persone soluzioni di mobilità più flessibili (**Smart Mobility**).

  - Amsterdam – Copenaghen – Berlino – Boston – Singapore – Venezia – Firenze – Bologna…
  - **http://www.smartcityexhibition.it**

Energia e Trasporti

- Relativamente all'ottimizzazione delle risorse energetiche e al monitoraggio ambientale, molto importanti sono i **dati relativi ai consumi energetici**: elettricità, gas, acqua, emissioni CO_2…

- Analisi di un insieme di **profili di carico** e di **informazioni georeferenziate**.

- Appropriate tecniche di data mining e la costruzione di modelli predittivi a partire da questi dati.

Definizione di **intelligenti strategie di distribuzione dell'energia** con l'obiettivo di ridurre i costi e aumentare la qualità della vita.

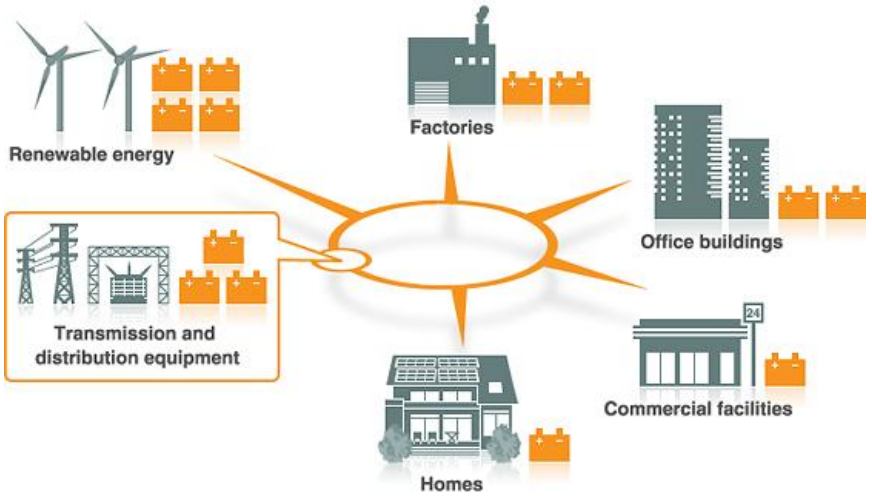
Energia e Trasporti

- Ricercatori hanno mostrato notevoli vantaggi ottenibili installando nelle abitazioni, uffici, fabbriche già solo tre tipologie di sensori:

Elettricità **Gas Naturale** **Aqua**

In questo modo è possibile determinare la quantità di risorse realmente utilizzate in una singola abitazione.

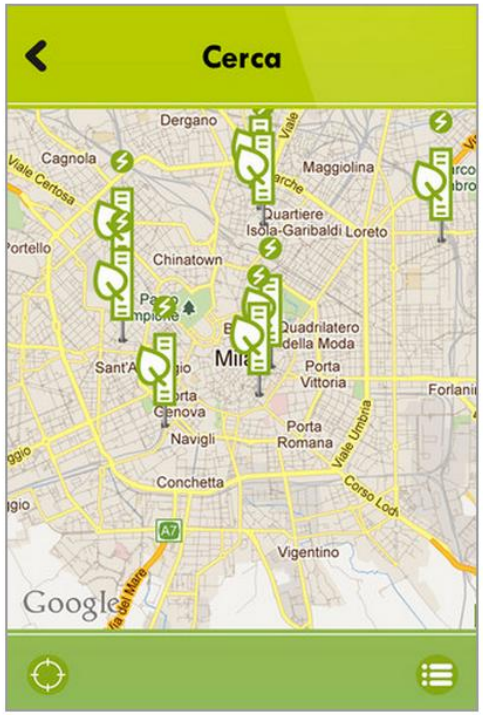
Nasce l'opportunità di trasformare **case e complessi residenziali in "ricche" reti di sensori,** integrare le informazioni relative al consumo effettivo con quelle relative alla disponibilità energetica, e definire così delle strategie ottimizzate di gestione dell'energia (**Smart Grid**).



Energia e Trasporti

http://app.comune.fi.it/app/a0011.html

- **iPhev** è un *progetto ambientale indipendente* che nasce con ***l'obiettivo di geolocalizzare tutti gli impianti di rifornimento per veicoli elettrici***, ideato per essere *implementato ed aggiornato attraverso le segnalazioni degli utenti* grazie all'utilizzo di un'applicazione gratuita per Smartphone.



Energia e Trasporti

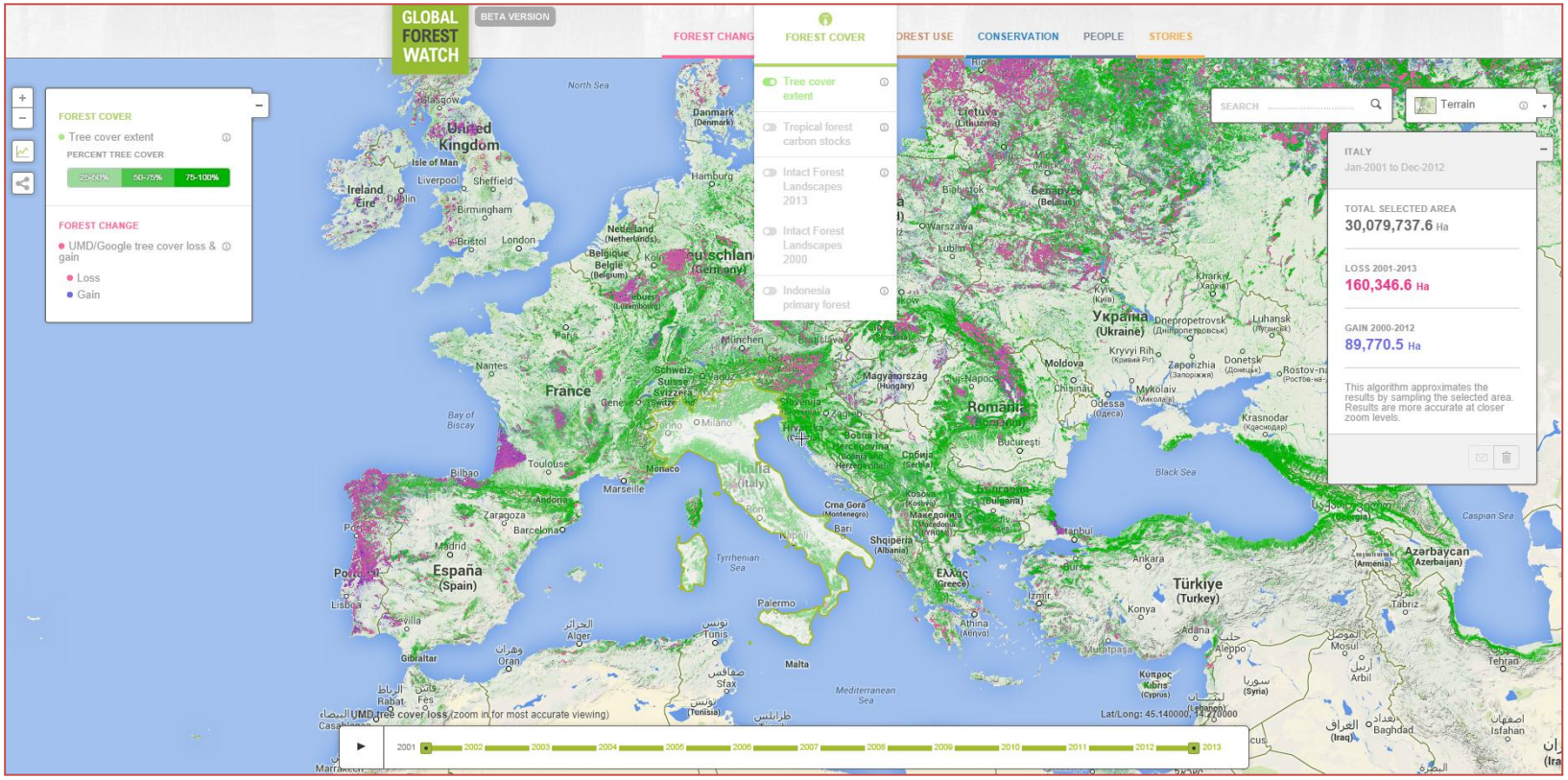
- Big Data Climate Challenge è un'iniziativa proposta dalla Segreteria Generale del Climate Summit che si terrà a settembre 2014.

- Progetti da tutto il mondo che **usano i Big Data e l'analisi per affrontare problematiche legate ai cambiamenti climatici** e al loro impatto sul mondo reale.

Energia e Trasporti

http://www.globalforestwatch.org/

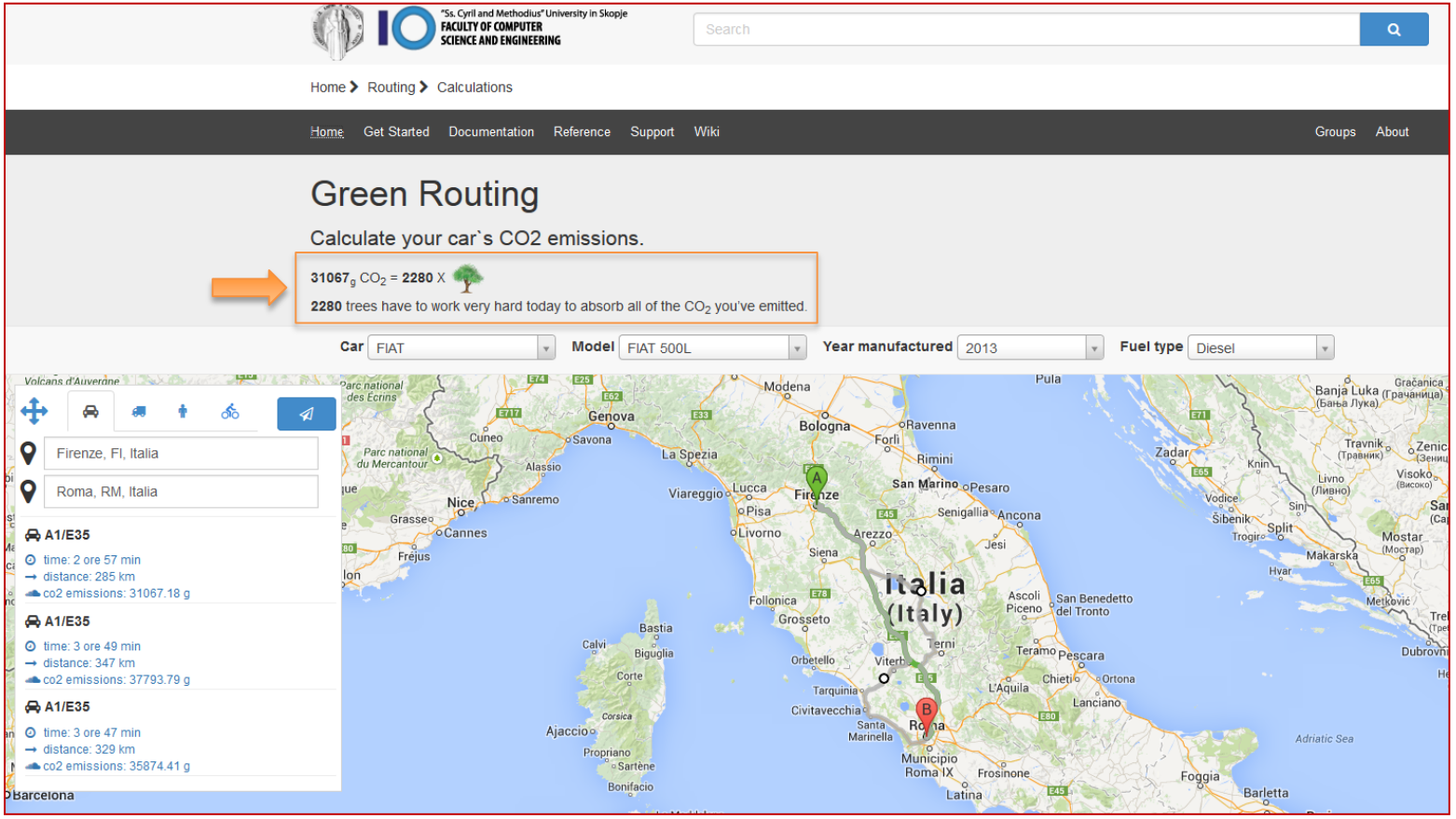
- **Global Forest Watch** (GFW – progetto vincitore) è una piattaforma che permette di **gestire e analizzare la "*situazione forestale*"** mediante l'uso combinato di **immagini satellitari, open data e aggiornamenti/informazioni tempestive proveniente da Governi o organizzazioni non governative**.



Energia e Trasporti

http://contest.msdev.finki.ukim.mk/

- **Green Routing** è un progetto *dell'Università di Skopie (facoltà di Ingegneria Informatica)* in cui sono stati utilizzati **Big Data e le mappe di Google** per ***determinare le emissioni di CO2*** di un veicolo durante un tragitto.



Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- Sanità e Medicina
- Ricerca Scientifica (Analisi dei Dati)
- Istruzione
- Settore Energetico e dei Trasporti
- **Social Network – Servizi Internet – Web Data**
- Finanza/Business – Marketing
- Sicurezza

Social Network Big Data

2012

- **Facebook**: più di 10 milioni di foto caricate ogni ora, 3 miliardi di "like" e commenti ogni giorno.

2009

- **Facebook**: 3 milioni di foto caricate ogni mese, implementazione del pulsante "like"

- **Google Youtube**: 800 milioni di utenti caricano ~1h di video ogni secondo.

- **Google Youtube**: tutti gli utenti caricavano 24h di video ogni minuto.

- **Twitter**: più di 400 milioni di tweet ogni giorno.

- **Twitter**: 50 milioni di tweet ogni giorno.

- **Instagram**: 7.3 milioni di utenti univoci ogni giorno.

- **Instagram**: è stato creato nel 2010.

Social Network – Servizi Internet – Web Data

- Il **volume di dati** generati dai **servizi internet**, **siti web**, **applicazioni mobili** e **social network** è grande, la velocità di produzione è invece variabile, a causa del **fattore umano**.

- Da queste grandi quantità di dati raccolti in particolare attraverso i social network, aziende e ricercatori cercano di **prevedere il comportamento collettivo** e analizzare i **trend topic**.

- Ad esempio attraverso il **monitoraggio degli hashtag (#)** di Twitter è possibile identificare dei **modelli di influenza**.

- In senso più ampio da tutte queste informazioni è possibile **estrarre conoscenza** e evidenziare le **relazioni tra i dati**, in modo da migliorare l'**attività di query-answering**.

Social Network – Servizi Internet – Web Data

Alcuni ricercatori hanno proposto un **utilizzo alternativo** di tali dati per creare una **nuova forma di vivibilità urbana**, in un'iniziativa/progetto chiamato ConnectiCity.
Gli aspetti chiave sono:

- Creare un **set di tool** per catturare **in real-time** differenti forme di **contenuti** rilevanti **generati dai cittadini/utenti**, provenienti da diverse tipologie di sorgenti:
  - *Social network*
  - *Siti web*
  - *Applicazioni mobile*

Social Network – Servizi Internet – Web Data

Alcuni ricercatori hanno proposto un **utilizzo alternativo** di tali dati per creare una **nuova forma di vivibilità urbana**, in un'iniziativa/progetto chiamato ConnectiCity.
Gli aspetti chiave sono:

- **Mettere in relazione questi contenuti al territorio** utilizzando tecniche di *Geo-Referencing*, *Geo-Parsing* e di *Geo-Coding*. **Analizzarli e classificarli** utilizzando tecniche di *Natural Language Processing* per identificare:
  - *Topic di interesse*
  - *Espressioni emozionali e sentimenti*
  - *Analisi della rete per capire la propagazione delle informazioni e i modelli di comunicazione*

Social Network – Servizi Internet – Web Data

Alcuni ricercatori hanno proposto un **utilizzo alternativo** di tali dati per creare una **nuova forma di vivibilità urbana**, in un'iniziativa/progetto chiamato ConnectiCity.
Gli aspetti chiave sono:

- Rendere queste **informazioni disponibili** e **accessibili** sia a livello centrale e periferico, per consentire la creazione di **nuove forme di processi decisionali**, nonché di sperimentare modelli innovativi di partecipazione, peer to peer, iniziative generate dai cittadini/utenti.

Project link - http://www.connecticity.net/
http://www.opendata.comunefi.it

Social Network – Curiosità

- L'università di Cambridge ha pubblicato uno studio nel 2013 dove emerge come **è possibile descrivere i tratti della personalità dalla semplice analisi dei Like degli utenti di Facebook.**

http://www.pnas.org/content/early/2013/03/06/1218772110.full.pdf

- Tra gli **attributi analizzabili** troviamo:
  - *Orientamento Politico*
  - *Orientamento Sessuale*
  - *Orientamento Religioso*
  - *Aspetti caratteriali*
  - *Livello di soddisfazione della propria vita*

Social Network – Curiosità

- Il modello proposto può essere applicato a qualsiasi **insieme di dati** in grado di **esprimere una preferenza dell'utente**.

- I dati di Facebook sono pubblici e con **Facebook Connect**, sono facilmente ottenibili, sempre previa autorizzazione da parte dell'utente.

- L'algoritmo è stato sviluppato implementato da una start up Italiana, **Cube You** e può essere testato all'indirizzo:

http://youarewhatyoulike.com/

Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- Sanità e Medicina
- Ricerca Scientifica (Analisi dei Dati)
- Istruzione
- Settore Energetico e dei Trasporti
- Social Network – Servizi Internet – Web Data
- **Finanza/Business – Marketing** 
- Sicurezza

Finance/Business e Marketing

- Il compito di **trovare modelli nei dati aziendali** non è nuovo. Tradizionalmente gli analisti di business usano tecniche statistiche.

- Oggi l'uso diffuso di **PC e tecnologie di rete** ha creato grandi **repository elettronici** che memorizzano numerose transazioni commerciali.
  - La grandezza di questi dati varia tra **50-200 PBs** al giorno
  - Gli **accessi ad Internet** in Europa sono circa **381 milioni di visitatori unici.**
  - **40% dei cittadini Europei** fa shopping online.

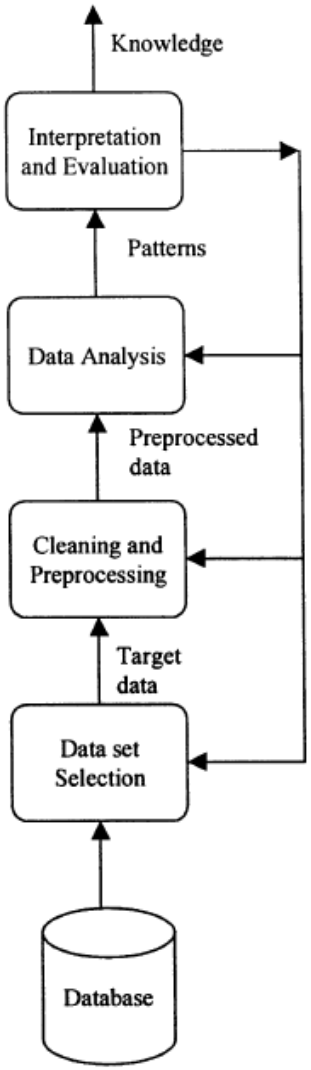
- Questi dati possono essere analizzati per definire:
  - **Previsioni sul comportamento degli utenti.**
  - **Identificare modelli di acquisto di clienti individuali o gruppi.**
  - **Fornire nuovi servizi personalizzati.**

Con l'uso di tecnologie di data warehousing e tecniche di apprendimento automatico mature.

Finance/Business e Marketing

In campo finanziario, invece, si possono creare **piani di investimento e di business grazie a modelli predittivi** ottenuti con tecniche di ragionamento o per scoprire modelli interessanti e significativi dai dati aziendali.

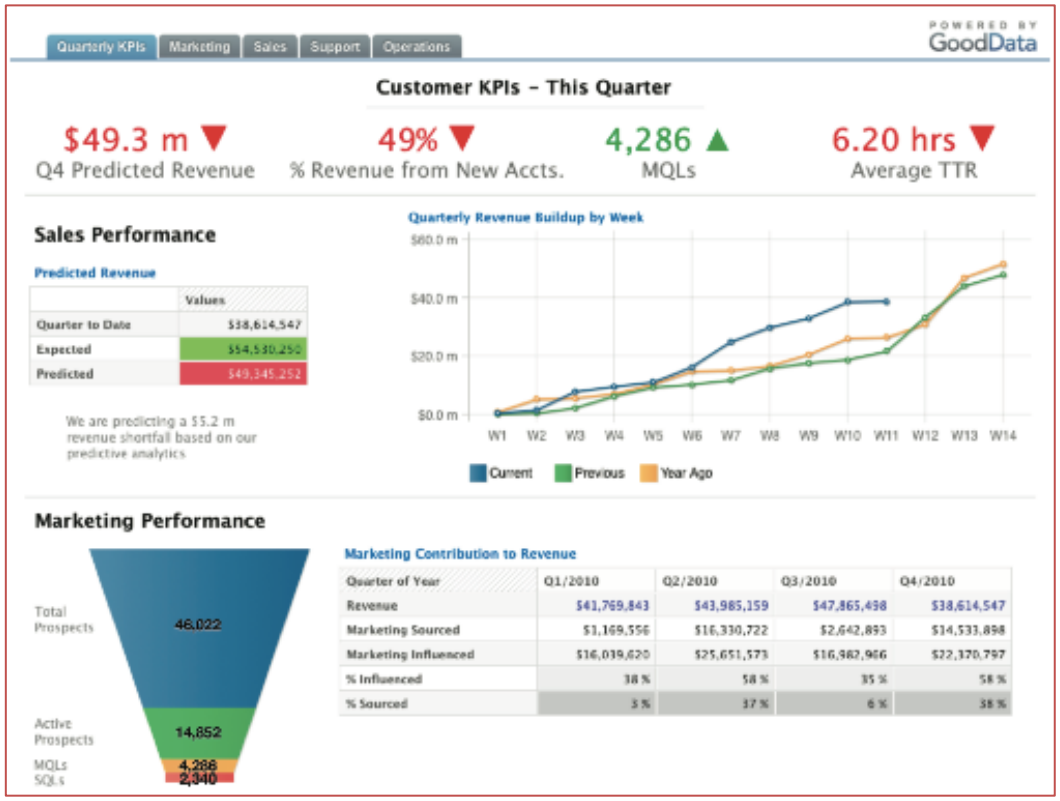
1. **Selezione dei dati per l'analisi** (da una rete di DB).
2. **Operazioni di raffinamento** per rimuovere discrepanze e inconsistenze.
3. I **dati sono analizzati** per identificare dei **Pattern** (modelli che mostrano la relazione tra i dati).
4. Dovrebbe essere possibile la t**raduzione del modello in business plan praticabile**, che aiuti l'azienda a raggiungere il suo obiettivo.
5. Modelli/Pattern che soddisfano queste condizioni diventano **business knoweldge**.



Finance/Business e Marketing

http://www.gooddata.com/

- **GoodData**, società di San Francisco, che fornisce una ***piattaforma con un insieme di tool di BI***. L'obiettivo è di supportare le aziende ad analizzare la loro **enorme mole di dati** (indagini di mercato, resoconto vendite, costi etc.) per ***favorire il processo decisionale***.



Campi di Applicazione

Investimenti crescenti nei Big Data possono portare fondazioni, enti e organizzazioni di nuova generazione a interessanti **scoperte** *in campo scientifico, nella medicina*, **vantaggi e guadagni** nel *settore ICT e in contesti Business,* **nuovi servizi e opportunità** per *cittadini digitali e utenti web*.

- Sanità e Medicina
- Ricerca Scientifica (Analisi dei Dati)
- Istruzione
- Settore Energetico e dei Trasporti
- Social Network – Servizi Internet – Web Data
- Finanza/Business – Marketing
- **Sicurezza** 

Sicurezza

- **Intelligence, Sorveglianza, e Recognition (ISR)** definiscono argomenti che sono adatti per analisi computazionali di tipo data-centrico.
  - Dimensione vicina ad uno **zettabyte (10^{21}bytes o un miliardo di TeraByte)** di dati digitali sono generati ogni anno.

- Importanti fonti di dati per i sistemi di intelligence sono
  - **Immagini satellitari e aeree (da veicoli UAV**).
  - **Comunicazioni intercettate**: civili e militari, tra cui voce, e-mail, documenti, i registri delle transazioni (log) e altri dati elettronici – **5 miliardi di telefoni cellulari in uso in tutto il mondo.**
  - **Dati di tracciamento radar.**

Sicurezza

- Importanti fonti di dati per i sistemi di intelligence sono
  - **Sorgenti di dominio pubblico** (siti web, blog, tweet e altri dati Internet, televisione, carta stampata e radio).
  - **Dati di Sensori** (dai meterologici, oceanografici, riprese di telecamere di sicurezza).
  - **Dati Biometrici** (Immagini facciali, DNA, impronte digitali, scansioni dell'occhio, registrazione del portamento).
  - **Informazioni strutturate e semi-strutturate** fornite da aziende e organizzazioni: *log delle compagnie aeree, carte di credito e transazioni bancarie, registrazioni telefoniche, elenco del personale dipendente, cartelle cliniche elettroniche, rapporti investigativi e dati nei registri di polizia.*

Sicurezza

- La **sfida** per i servizi segreti è quello di **trovare, combinare e definire modelli e tendenze nelle tracce di informazioni ritenute importanti.**

- **Occorre trovare modelli di evoluzione significativi in modo tempestivo tra diverse informazioni potenzialmente offuscate provenienti da fonti multiple.** Necessità di metodi sofisticati per individuare modelli accurati, **senza generare un gran numero di falsi positivi** in modo che non emergano cospirazioni o allarmi dove non esistono.

- *Un esempio di come sfruttare efficacemente fonti che producono dati su larga scala, sono le principali aziende in ambito web, come Google, Yahoo e Facebook.*

Sicurezza

- All'interno del mondo dei servizi di Intelligence le **tecnologie informatiche** e le consolidate **tecniche di apprendimento automatico** devono essere considerate come un elemento per aumentare le capacità degli analisti piuttosto che come un modo per sostituirli.

- L'idea chiave dell'apprendimento automatico è:

- **Applicare** ad un certo dataset **prima un'analisi statistica strutturata** al fine di generare un ***modello predittivo***.

- **Poi applicare questo modello** a *diversi flussi di dati* per **supportare diverse forme di analisi** e ottenere nuovi risultati.

Sicurezza – Esempio

http://www.phlcrimemapper.com/

- Nel dicembre 2012 il comune di **Philadelphia** ha rilasciato un **dataset con l'elenco dei crimini** dal 1° gennaio 2006.
- Ogni **crimine** (*furto, rapina, omicidio*…) è **taggato nella posizione esatta** in cui è stato commesso.
- Con questi **dati** è possibile la **creazione di tool e statistiche** utili sia al cittadino che alla pubblica amministrazione.

Indice

1. **Big Data**
   - *Evoluzione dei dati e delle tecniche di analisi*
   - *Le 5V dei Big Data*
   - *Teorema CAP*
   - *Pipeline dell'analisi dei Big Data*
2. **Principali Contesti Applicativi**
3. **Criticità e Rischi dei Big Data** 

Criticità e rischi dei Big Data

Come ogni "nuova tecnologia" **i Big Data offrono grandi prospettive e potenzialità**, ma non presentano esclusivamente caratteristiche positive. Vi sono alcuni **aspetti critici** che è bene prendere in considerazione:

- Problematiche legate alla **qualità e all'affidabilità dei dati.**
- Problematiche relative alla **privacy e alla proprietà dei dati.**

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.

Criticità e rischi dei Big Data – Qualità dei dati

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

- **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**. 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 conosciuti e esistono diversi strumenti per realizzare in modo automatico la pulizia dei dati.

- **Dati generati automaticamente:** i dati scientifici o provenienti da sensori sono privi di errori di immissione. Spesso però sono "deboli" a livello di contenuto informativo, c'è la necessità di integrarli con dati provenienti da altri sistemi per poi analizzarli.

- **Dati del Web:** Social network, forum, blog generano dati semistrutturati. La parte più affidabile sono i metadati (se presenti), il testo invece è soggetto a 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 differenti (es. calcio). La sfida è cerca di trovare quello più attinente al contesto in esame. Un aiuto sono i ***tag***, etichettando i dati si cerca di evidenziare l'ambito di pertinenza.

- **Veridicità:** Notizie, affermazioni, documenti non sempre veri o corrispondenti alla realtà.

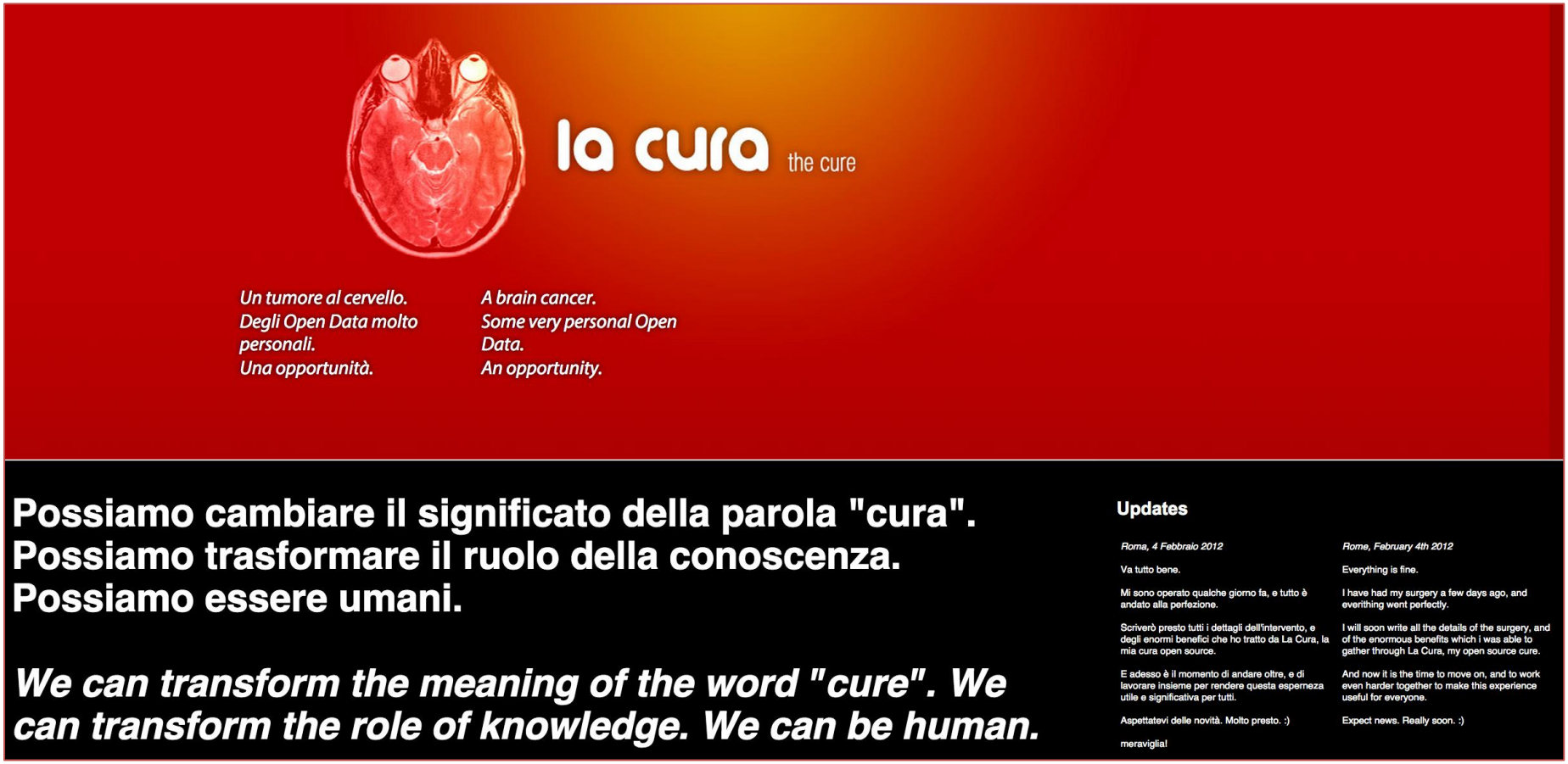
OSS. *La qualità dei dati è però legata anche al contesto in cui essi sono analizzati. Operazioni di filtraggio e pulizia devono essere fatte procedendo per gradi per evitare di eliminare dati potenzialmente utili.*

Criticità e rischi dei Big Data – Privacy

Il tema Big Data si apre a problemi di Privacy, proprietà e utilizzo dei dati da parte di terzi.

- **Dati del Web:** gli *user-generated-content* sono condivisi accessibili a tutti. E' etico il loro utilizzo?
- **Dati sensibili:** i dati presenti nei DB degli ospedali relativi alla storia clinica dei pazienti sono opportunamente protetti?
- **Dati di posizione:** 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.

Criticità e rischi dei Big Data – Privacy



http://opensourcecureforcancer.com/