Management delle informazioni e gestione della conoscenza AA 2021-22

Big Data

What is big data?

In altre parole, Big Data sono dati il cui volume è troppo grande per i tradizionali sistemi di analisi, ma non solo…

Big Data

Dati prodotti in Internet:

- 2003: 5 exabyte dati (=5G di gigabyte)
- 2010: 800 exabyte dati (=800G di gigabyte)
- 2011: 1,8 zettabyte (trilioni di gigabyte)
- Dati prodotti da acquisti e vendite online, dai cellulari, spostamenti (telepass, aerei), dai sensori RFID (IoT), ecc.
- Ogni minuto: 200M email, 2M ricerche Google, 48h video, 100000 tweet, +2M azioni in FB …

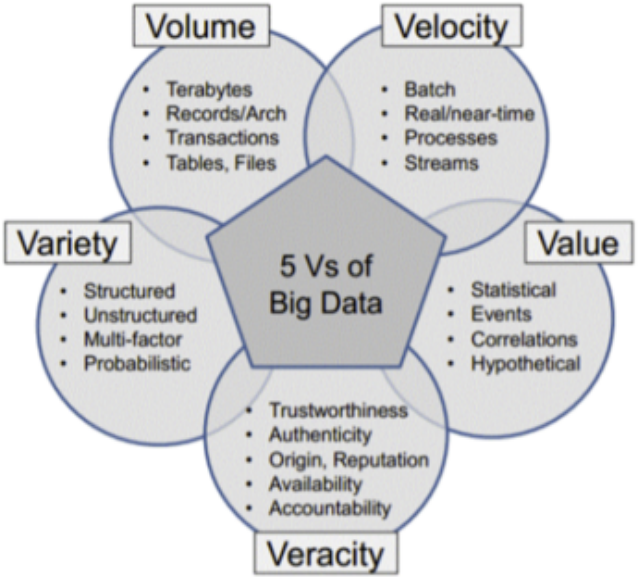
Non ci sono più unità di misura, si parla di **Big Data**

Google Trends: “Big Data”

Inizia l’era degli smartphone

Big Data: le 5V

- I database per Big Data sono caratterizzati da:
  - **Volume**: non gestibili da db tradizionali
  - **Velocità**: processati in tempo reale
  - **Varietà**: dati di diversa natura e soprattutto non strutturati
  - **Veridicità**: incertezza, inconsistenza, complessità …
  - **Valore**: economico, informativo, statistico…



Visualizzazione Big Data sul Web

- Esempi di Big Data (in real time):
  - https://www.flightradar24.com/ : traffico aereo
  - http://earth.nullschool.net/ : condizioni meteo sulla Terra
  - https://www.worldometers.info/watch/world-population/region.php : popolazione mondiale
  - https://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/ : una giornata degli americani
  - http://www.internetlivestats.com/ : visualizzazione dei numeri di Internet
  - https://medium.com/echelon-indicators/the-year-in-news-2017-2b832594b4a6 : un anno di storie su Twitter USA, 2017

Le sfide dei Dati (1/2)

- **Data Availability:** quale è il livello di disponibilità dei dati? Sono disponibili a tutti?
- **Data Quality:** quanto sono "buoni" i dati (rilevanti e consistenti)? Quale è il livello di copertura? Quanto sono aggiornati?
- **Data Discovery:** come troviamo alta qualità di dati dalla vasta collezione di dati che si trovano nel web? è una grande sfida
- **Combining multiple data sets:** come connettere e/o integrare differenti e grandi basi dati?

Le sfide dei Dati (2/2)

- **Completeness of the data**: ci sono aree senza copertura? Quale è l'implicazione?
- **Privacy:** molte informazioni riguardano informazioni personali, possiamo estrarre sufficienti informazioni per effettuare analisi a supporto delle persone, senza compromettere la privacy?
- ….

Process Challenges

- Le sfide del processo di gestione dei Big Data includono:
  - L'acquisizione dei dati
  - L'allineamento dei dati derivanti da diverse sorgenti (es. quando due oggetti sono gli stessi in differenti database)
  - Trasformazione dei dati in una forma adatta all'analisi
  - Definizione e scelta di Modelli di analisi
  - Capire l'output, condivisione e visualizzazione di risultati

Management Challenges

- Data Privacy, Security e Governance
  - Assicurarsi che i dati siano usati correttamente (rispetto degli utilizzi previsti e delle normative)
  - Gestione del traking dei dati utilizzati, trasformati e raccolti ecc.
  - E gestione del loro ciclo di vita

BI vs. Big Data Analytics

- **BI Analytics**: Tecniche, tecnologie, sistemi, pratiche, metodologie e applicazioni per analizzare dati di business al fine di comprendere meglio il mercato e prendere decisioni velocemente
  - Applicabile a vari ambiti: e-commerce, e-government, sanità, sicurezza ecc.
- **Big Data Analytics**: applicazione tecniche analitiche a enormi moli di dati (da terabyte a exabyte) generati da sistemi complessi (dai sensori ai social media), per comprendere dei trend o per fare previsioni

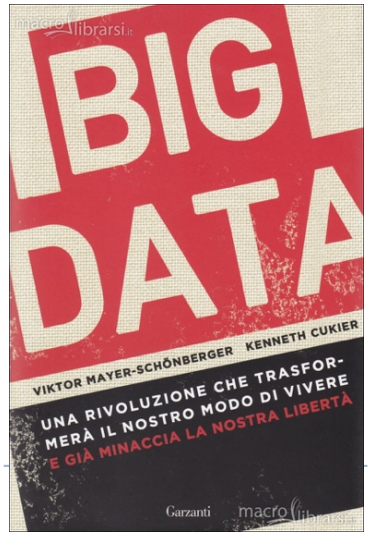
Sfruttamento dei Big Data

- Insegnare ad un computer a pensare come gli esseri umani? No!
- Applicazione della matematica ad enormi quantità di dati per ***desumere delle probabilità:***
  - ***La probabilità che una email sia spam;***
  - ***Che le lettere digitate teh siano l'inversione di the;***
- *Creare sistemi che funzionano bene perché alimentati da enormi quantità di dati su cui basare le proprie previsioni; costruiti per:*
  - *Automigliorarsi (Google è in grado di selezionare il sito più pertinente; Linkedin di indovinare chi conosciamo)*

Sfruttamento dei Big Data

"Così come internet ha cambiato radicalmente il mondo aggiungendo la capacità di comunicazione ai computer, i BIG DATA modificheranno aspetti fondamentali della vita dandole una dimensione quantitativa che non ha mai avuto prima"

(Shonberger – Cukier, Big data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà, Garzanti, 2012)



Big data e analisi delle informazioni

- **Esattezza-precisione *vs* imprecisione-tendenza; causalità *vs* correlazione**
  - Da strumenti fondati sull'esattezza: misurare nel modo più preciso possibile ciò che vogliamo quantificare

    (es. motori di ricerca finalizzati a recuperare con precisione i record corrispondenti esattamente alle ns query);
  - A strumenti fondati su "cogliere una tendenza": rinuncia ad un po' di esattezza; ciò che perdiamo a livello micro lo recuperiamo in comprensione a livello macro;
  - Abbandono della tendenza a ricercare la **causalità** per scoprire nei dati **correlazioni** che offrono indicazioni originali e preziose (non sempre serve conoscere la causa di un fenomeno si può lasciare che i dati parlino da sè)

Dati e decision-making

- In passato i dati erano scarsi e costosi, e ci si affidava all'esperienza degli esperti
- Oggi con i Big Data i dati sono accessibili e possono guidare le decisioni
- I Big Data richiedono nuove competenze analitiche, di business e di decision-making
- In USA servono da 140 a 190 mila persone per sfruttare i Big Data nei processi decisionali, e +1,5M di analisti e data manager
- Però 80% dei responsabili marketing prendono decisioni basandosi ancora su ricerche di mercato e benchmarking tradizionali, non sui Big Data

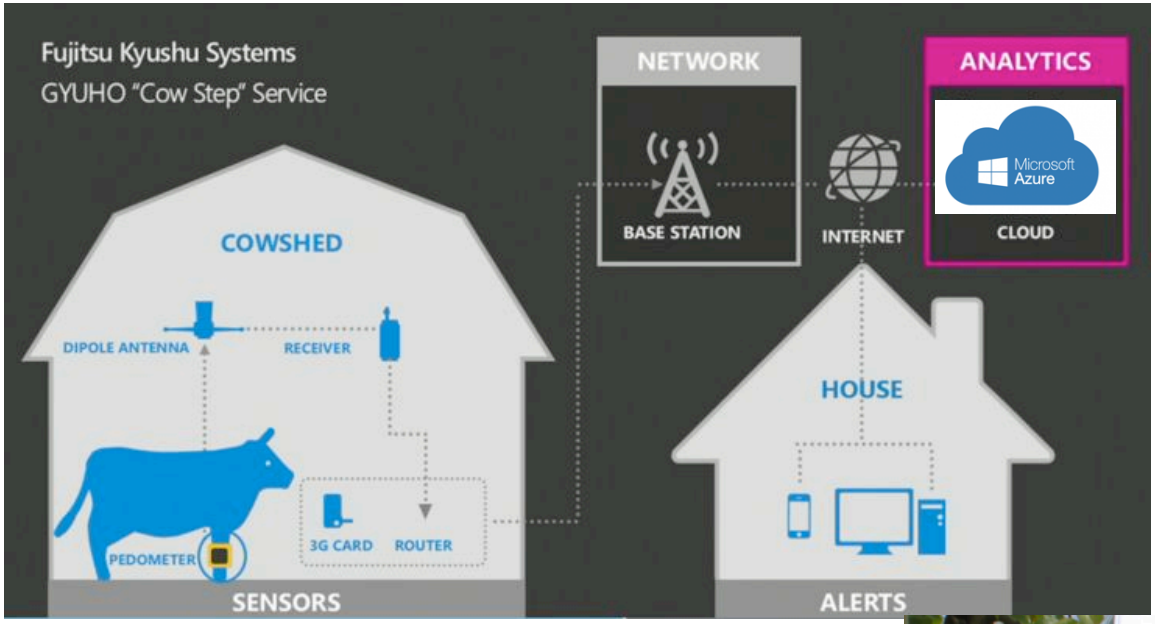
Utilizzo Big (Web) Data nel privato

- **Wal-Mart** ha creato @WalmartLabs per monitorare conversazioni su FB, Tw, Foursquare ecc. per rilevare le esigenze dei consumatori (patatina Takis in California)
- **Tesco** (supermarket UK) incrocia dati ordini con dati meteo e limitano sprechi nella distribuzione dei prodotti freschi
- **UPS** analizza i tempi di consegna della merce e il traffico incontrato dai suoi mezzi per individuare le rotte più veloci

ma anche…

- **Target** (supermercati USA) utilizza algoritmo predittivo sulle preferenze dei clienti (storia del padre di Minneapolis)
- **American Express** comportamento discriminatorio basato su deduzione scorretta (Storia di Kevin Johnson)

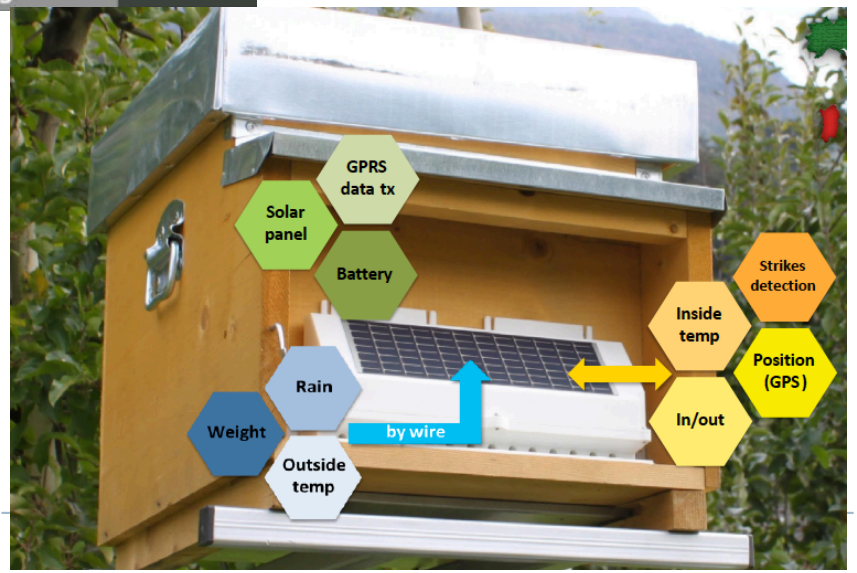
Big Data in agricoltura e allevamento (IoT)



Connected cow

Mucche e Fujitsu

Api e Melixa



Settore pubblico

Potenziali effetti positivi dall'analisi dei BD:

- Razionalizzazione delle spese (profilo sanitario elettronico)
- Controllo delle performance e sprechi
- Miglioramento dei servizi
- Individuazione degli evasori
- Ottimizzazione delle risorse
- Trasparenza (open data)
- …

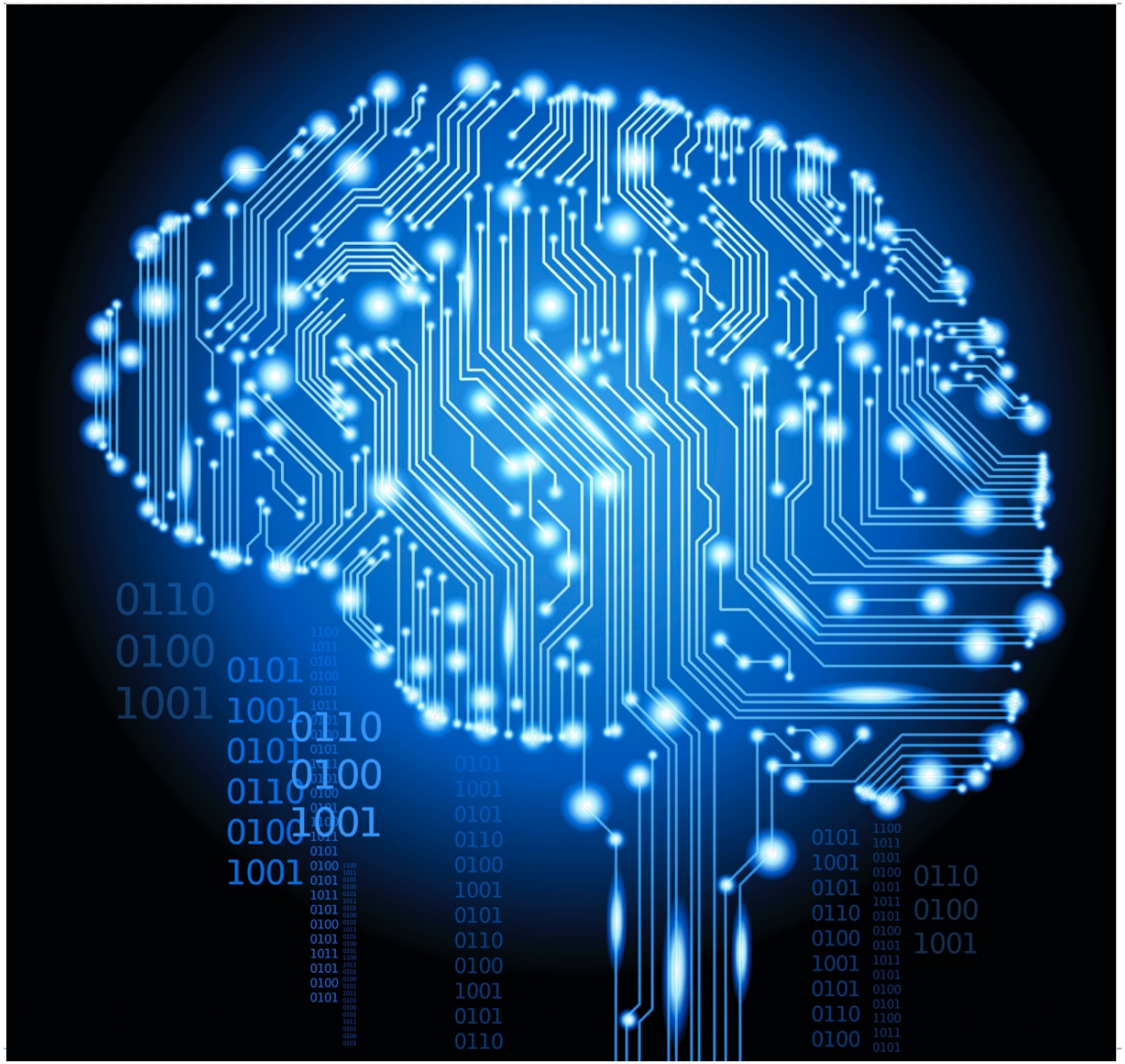
Esempi nel pubblico

- **Agenzia tasse svedese**: incrocia informazioni sui contribuenti da db diversi e invia modulo per le imposte già compilato che i cittadini devono confermare o modificare via web o SMS
- **Agenzia lavoro tedesca**: analizzando i dati storici su impiego e investimenti effettuati segmenta i disoccupati e offre interventi mirati ed efficienti, risparmiando 10G euro per il pubblico
- **Polizia di Los Angeles**: usa PredPol, software per analisi predittiva del crimine, che indica quali aree sorvegliare sulla base delle serie storiche dei reati, calo del 13% reati (simile a NYC stop & frisk)

Personal Health Data

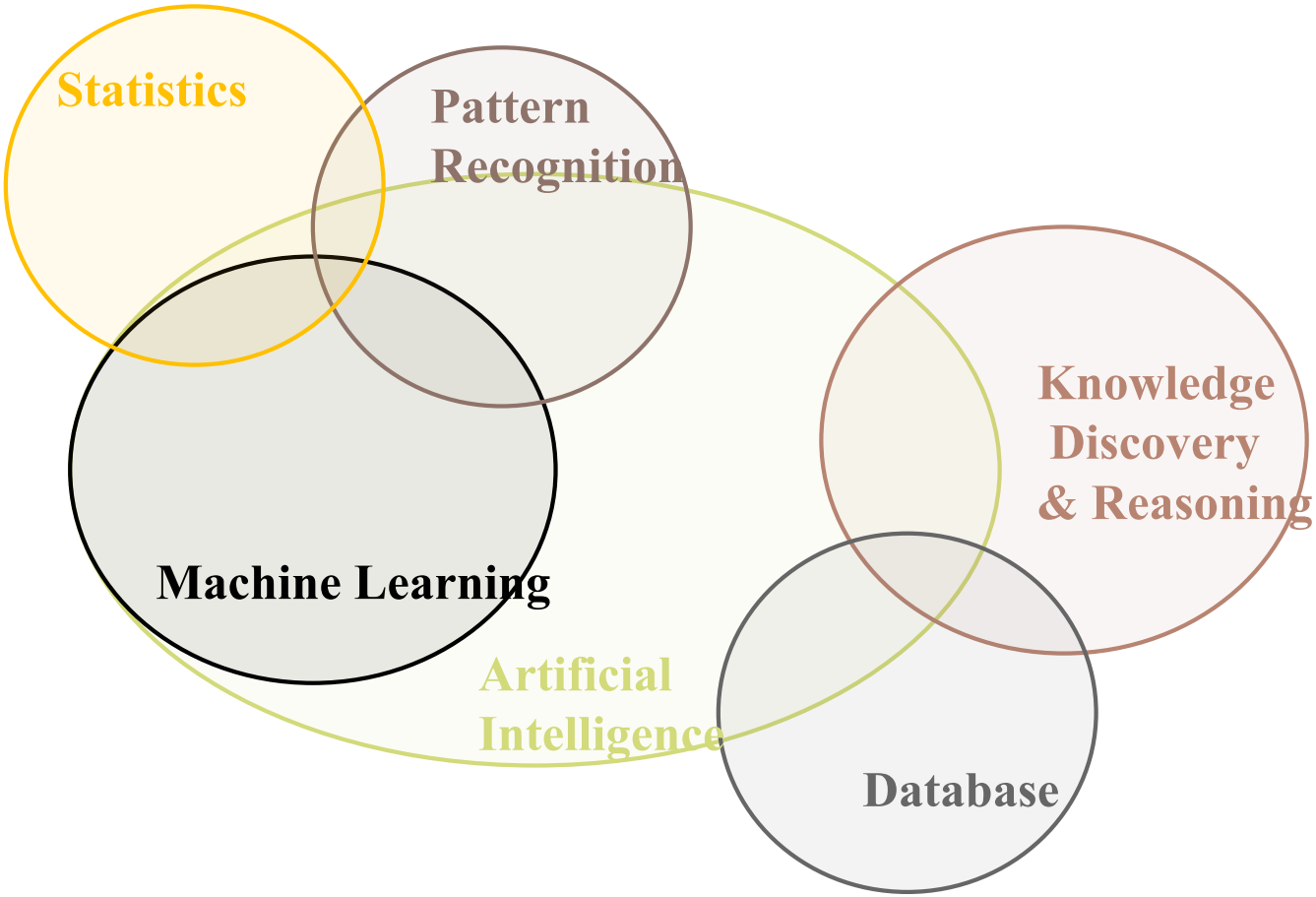
- Self-tracking: sensori e device da vestire collegati a siti/log dove raccogliere e analizzare i dati, per la salute, sport, umore, tracciamento delle abitudini, consumi energetici…
- Apps per la salute:
  - Ginger.io: per assistere i diabetici nel caso entrino in depressione e interrompano i medicinali, manda alert a medici e parenti
  - **Cardiio**: misura battiti cuore con la fotocamera frontale dello smartphone (rapporto battito-colore viso)
  - "Calzini intelligenti" sensori che rilevano pressione e/o se persona anziana è caduta

AI & Big Data

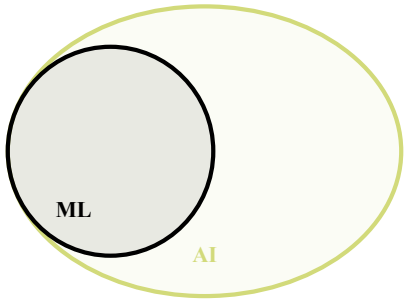

0110
0100
1001
0101
1001
0110
0100
1001
0101
1001
0110
0100
1001
0101
0110
0100
1001
0101
0110
0100
1001
0101
1001
0110
0100
1001

Big Data: the fuel of AI

AI: verso un approccio multidisciplinare

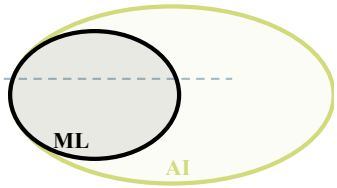


Machine Learning



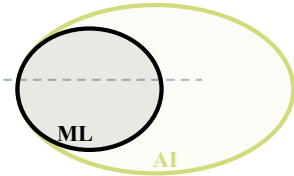
Un software che impara come svolgere un insieme di compiti attingendo dall'esperienza, e che incrementa la propria esperienza migliorando la sua capacità di svolgere il compito per il quale è stato progettato

Machine Learning: due categorie



Unsupervised: Il sistema tenta di classificare (raggruppare) elementi aventi caratteristiche (feature) comuni sulla base di un criterio di similarità. Al variare del criterio e della calibrazione dell'algoritmo varia il risultato del processo

Machine Learning: due categorie



Supervised Il sistema classifica **(raggruppa) elementi aventi caratteristiche comuni** (feature) sulla base delle caratteristiche che ha appreso durante il training. Il test serve per sapere «quanto bene» il sistema ha appreso. Non c'è modo di sapere quanto il sistema sia «bravo» in produzione

Supervised Learning (Fase di Addestramento)

Training Set

(the bigger, the better)

Balena

Squalo Bianco

Machine Learning Algorithm

Supervised Learning (Fase di Valutazione)

Test Set

Squalo Bianco (V)

Balena (X)

Machine Learning Algorithm

Score: 92% di accuratezza

Unsupervised Learning (in produzione)

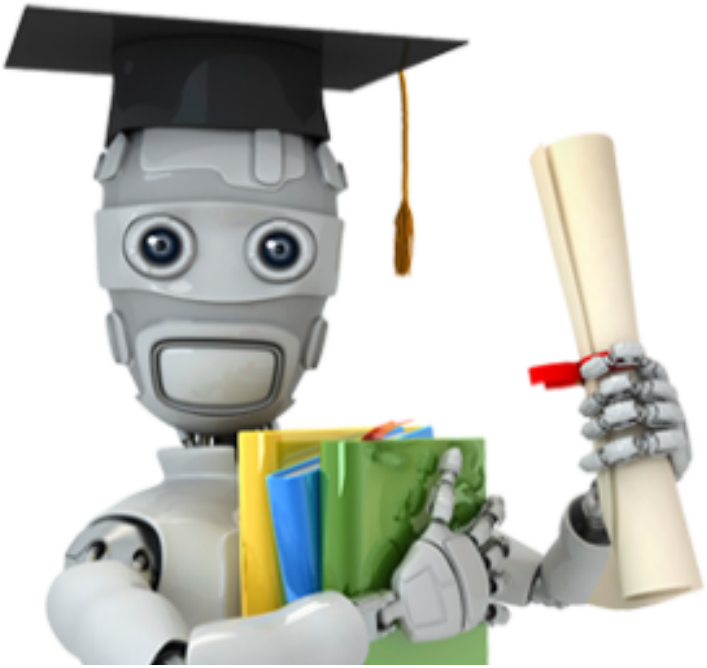
Squalo o Balena?

Machine Learning Algorithm

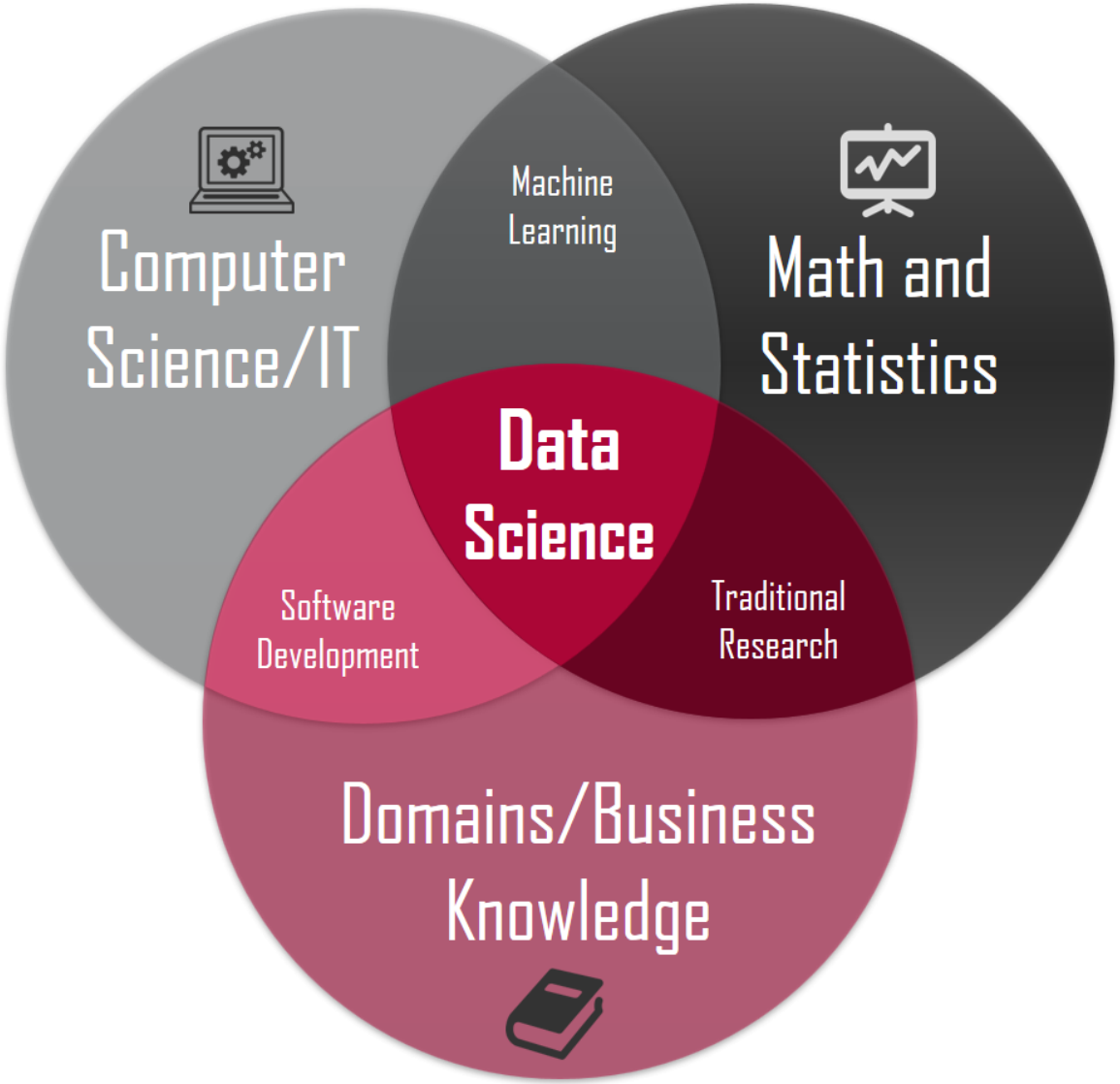
Butta i dati nell'algoritmo, e spera che ci prenda!

Due aspetti ortogonali

- Analytics / Machine learning
  - Apprendere dai dati per decidere
- Big data
  - Gestire enormi masse di dati
- Possono essere combinati, o usati separatamente

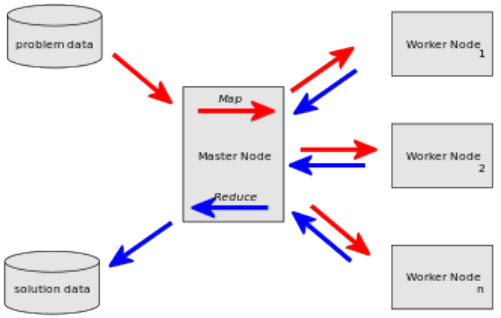


Data Science

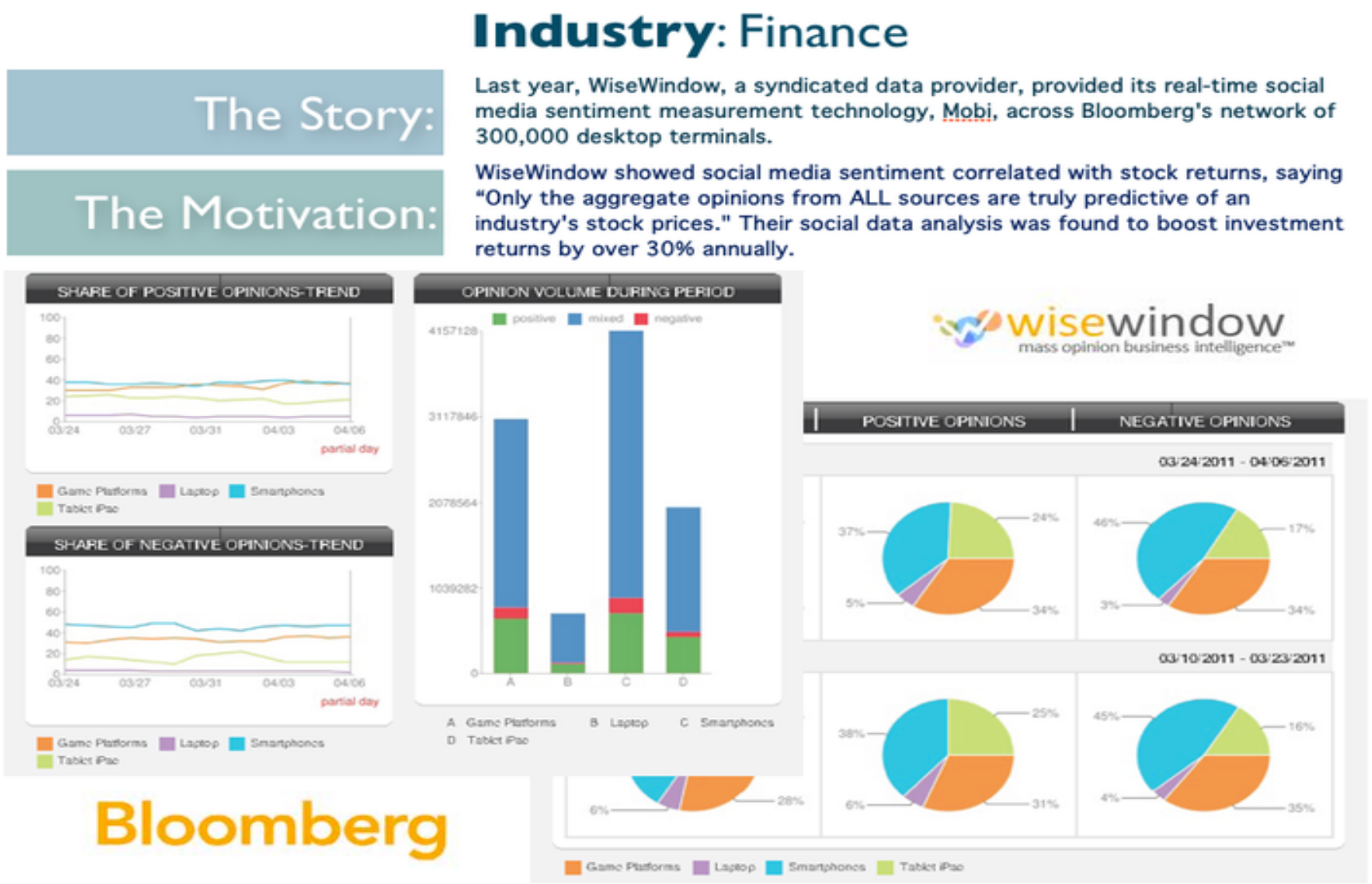


Quali software usare?

- Se i database relazionali non sono sufficienti, cosa serve?
- MapReduce: un framework (software) brevettato e introdotto da Google per supportare la computazione distribuita su grandi quantità di dati in cluster di computer
- NoSQL database: graph db, document db, a oggetti…



Esempi di utilizzo di Big Data e Social Media



From http://zdnet.com/blog/hinchcliffe on ZDNet.

Esempi di utilizzo di Big Data e Social Media

Industry: Disaster Management

The Story:

In 2011, when the 5.8 magnitude earthquake hit Virginia, Twitter turned out to be the first and richest source of data, even over the official U.S. Geologic Survey.

The Motivation:

Many visualizations have been created to show how the information about natural disaster (see the Wall Street Journal example below, on the Virginia earthquake) The U.S. Geologic Survey is now exploring how to use social media to augment its own reports, which take 2-20 minutes to issue.

Twitter Responses to the Virginia Earthquake

The visualization below (created by SocialFlow) replays a 90-second spread of earthquake-related Tweets across North America, sparked by the tremor that hit Mineral, Va., at 1:51 p.m. on Aug. 23, 2011. The colors represent distance from the epicenter and the circle size and opacity represent the number of tweets.

Sources: Twitter, SocialFlow.

From http://zdnet.com/blog/hinchcliffe on ZDNet.

Esempi di utilizzo di Big Data e Social Media

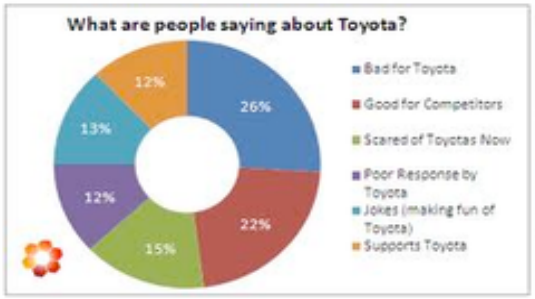
Industry: Automobile Quality Assurance

The Story:

Pamplin College of Business wondered if social media postings by consumers could be a source of useful information about vehicle safety and performance defects for automobile manufacturers.

The Results:

The first large-scale case study ultimately confirmed the value of social media for vehicle quality management. Pamplin's research shows that the existence of safety and performance defects is strongly predicted by the incidence of automotive problem reports in social media.



From http://zdnet.com/blog/hinchcliffe on ZDNet.

Esempi di utilizzo di Big Data e Social Media

Industry: Energy

The Story:

GE wanted to speed up the repair process and get customers back on a working electricity grid after an outage. It needed more up-to-date source of information about outages and problems.

The Results:

The result is GE's Grid IQ Insight tool, scheduled for release next year. It can mine social media for mentions of electrical outages. Since many social media posts are geotagged, utilities will use Grid IQ Insight to get an early notification of an outage in its area.

GE's Grid IQ Insight screenshot

From http://zdnet.com/blog/hinchcliffe on ZDNet.

Esempi di utilizzo di Big Data e Social Media

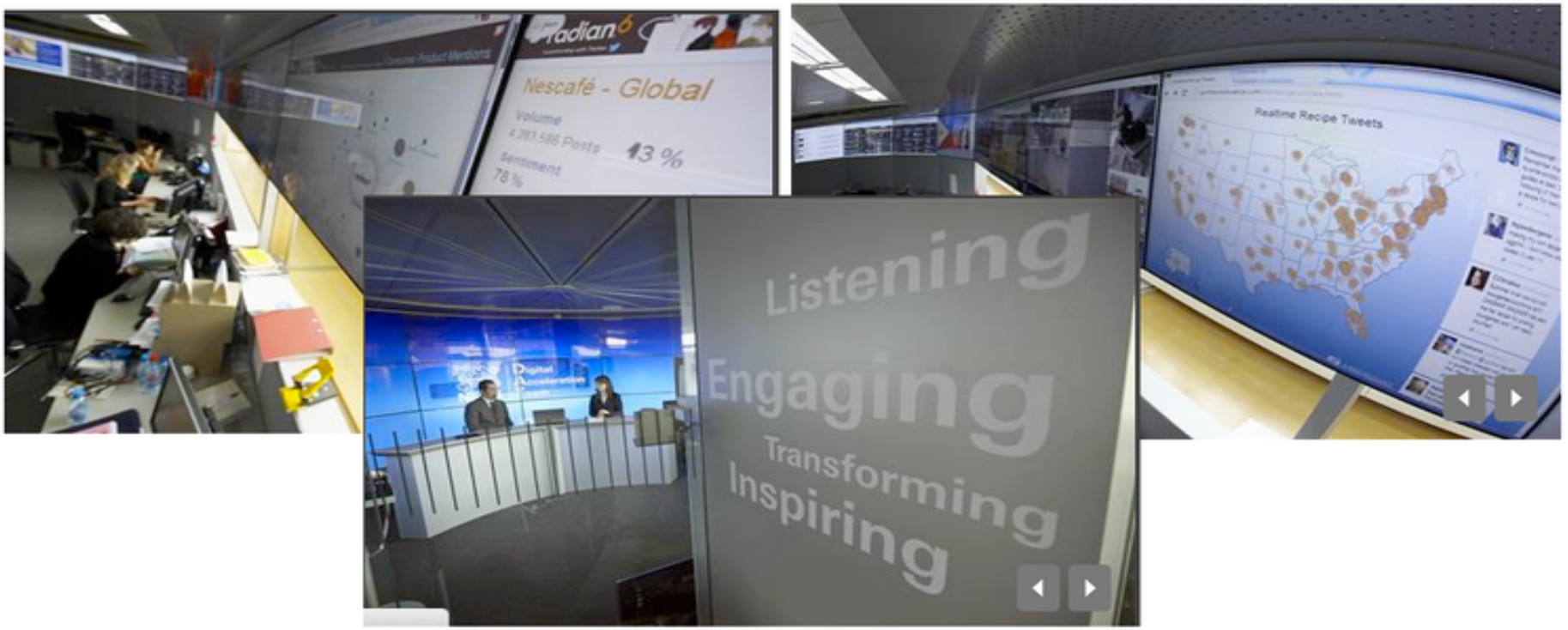
Industry: Marketing

The Story:

Nestle needed to have a better handle on customer sentiment, instead of relying on surveys and other periodic customer samplings which quickly went out of date.

The Results:

Their Digital Acceleration Team established a 24/7 monitoring center to listen to conversation about its products on social media. Staff members reach out and proactively engage customers to improve their experience with the company. Since then, Nestle has gone from #16 to #12 in the Reputation Institute's index of the world's most reputable companies.



Source: Reuters

From http://zdnet.com/blog/hinchcliffe on ZDNet.