# Sono un matematico...e ora?

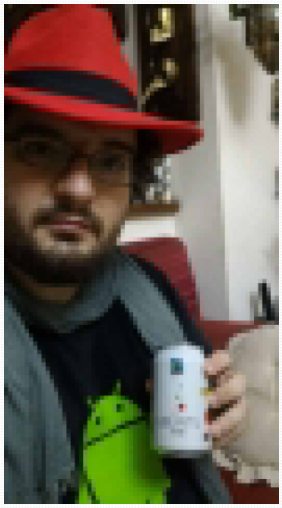
Guida su cosa succede nel mondo e come riciclare le nostre conoscenze

Gabriele Nocco, Senior Data Scientist

Catenate s.r.l.



**IO - STUDENTE**

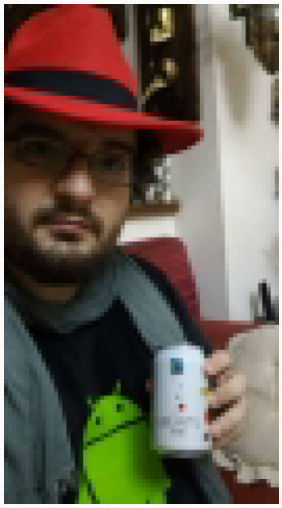


Gabriele Nocco, laureato in Matematica con specializzazione in Geometria Algebrica e Differenziale presso l'Università degli studi di Roma3.

- Luglio 2009, Tesi sulla teoria dei divisori con relatore Angelo Felice Lopez
- Iscritto presso l'Università degli studi di Roma, La Sapienza, nel master "Intelligenza Artificiale e robotica".
- Assiduo studente di corsi online di Machine learning, BigData, Statistica, Gamification e Social Network Analysis presso siti come Coursera e EdX.
- Co-fondatore del Meetup di Machine Learning e Data Science.

MACHINE LEARNING
DATA SCIENCE
Meetup

## IO - LAVORATORE



- Settembre 2009, impiegato presso Codin, prima come Sistemista applicativo, poi come Programmatore Java.
- Settembre 2014, impiegato presso Be Consulting come Ricercatore Machine Learning ed esperto nuove tecnologie come BigData e NOSQL.
- Settembre 2016, impiegato presso Catenate come Capo del reparto di Data Science e Advanced Analytics.



## AGENDA

- Importanza dei dati
- Approccio agnostico
- Cosa farci con i dati

## AGENDA

- Importanza dei dati
- Approccio agnostico
- Cosa farci con i dati





## IMPORTANZA DEI DATI - DA DOVE COMINCIO?


Google è senza dubbio una delle principali fonti di risposte a qualsiasi domanda noi possiamo porci!

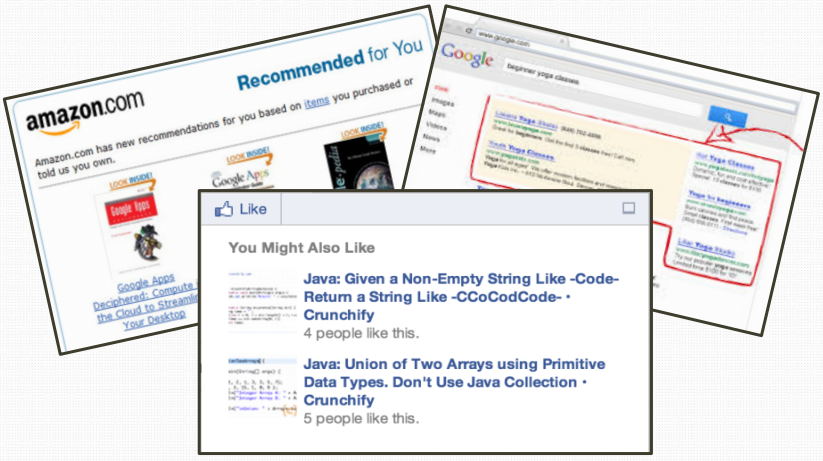
**Come fa a decidere quale risposta sia la migliore?**

## IMPORTANZA DEI DATI - DATI

- La centralità del dato è ormai una priorità per tutte le aziende. Le informazioni sono reperibili ovunque ed in ogni momento.
- Le informazioni sono un bene non contabilizzabile che però conferisce potere a chiunque le possegga. Tramite le informazioni si riesce ad essere più incisivi nel proprio mercato.



## IMPORTANZA DEI DATI - OPPORTUNITÀ




Conoscere la propria customer base amplia le possibilità di soddisfarla e di proporle nuovi prodotti.
Un sistema di recommendation raffinato è alla base di e-commerce o di networking

## IMPORTANZA DEI DATI - BIG-COSA?

**Che cos'è BigData? Cosa si intende?**


Tecnicamente si comincia a parlare di *BigData* per qualsiasi mole di dati non entri su una singola macchina. Comunemente si comincia a parlare di dati veramente big quando si sta sull'ordine di Petabyte (1025 TB).



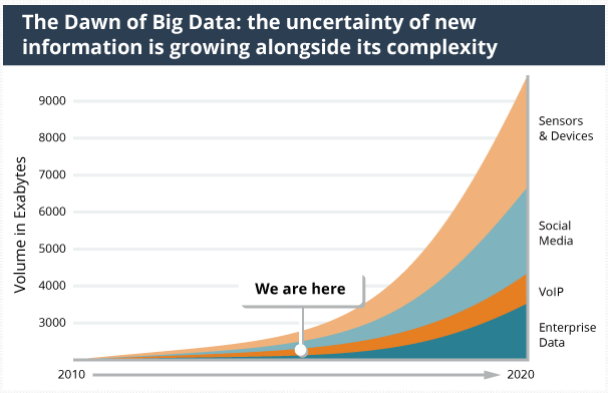
## IMPORTANZA DEI DATI - IOT

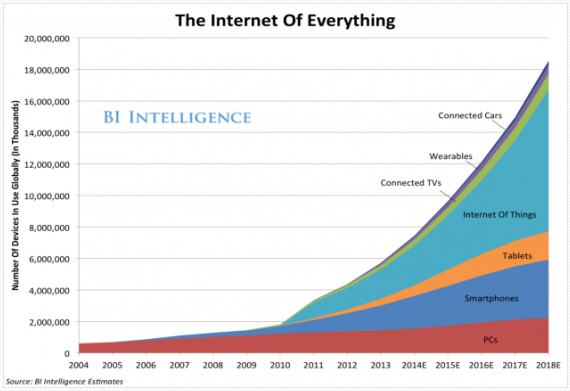

**IOT** (**I**nternet **O**f **T**hings) è uno dei filoni più innovativi del panorama informatico degl'ultimi anni.

Ogni oggetto intorno a noi si connette al mondo e raccoglie informazioni per aumentare la propria funzionalità e migliorare l'esperienza di utilizzo.

Ogni dispositivo deve avere coscienza del suo stato e dell'ambiente in cui si trova e tutte queste rilevazioni generano a loro volta dati.

## IMPORTANZA DEI DATI - PROSPETTIVE

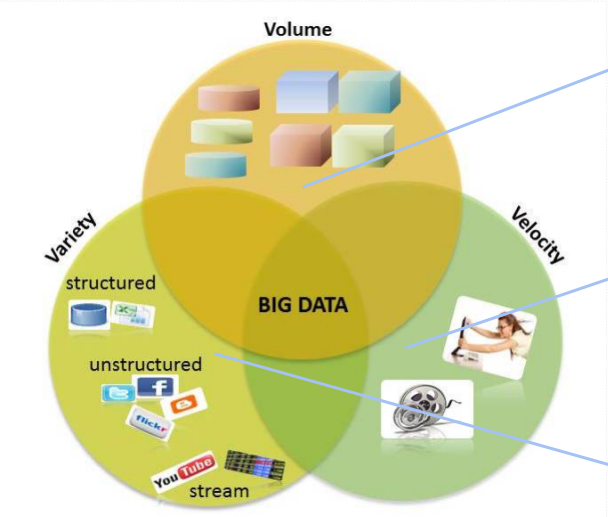



Nei prossimi anni si prevede una crescita esponenziale dei dati prodotti da fonti non Enterprise, con Social Network e sensoristica. È previsto che nei prossimi due anni il numero dei dispositivi IOT connessi in rete sarà maggiore del numero di Pc, Smartphone e Tablet.



## IMPORTANZA DEI DATI - LE TRE "V"




Se consideriamo la dimensione di tutti i dati generati nel mondo dagli albori dell'informatica fino al 2008, questo ammontare è stato eguagliato da quando è stata visualizzata questa slide!!!

Le nuove tecnologie ci permettono di analizzare i dati durante la loro generazione, senza che questi vengano neanche salvati su una base dati.

Più dell'80% dei dati generati fino ad ora sono non strutturati: messaggi, video, immagini o suoni. Nuove strutture dati ci permettono di mettere in relazione tutte queste fonti.

## IMPORTANZA DEI DATI - NUMERI

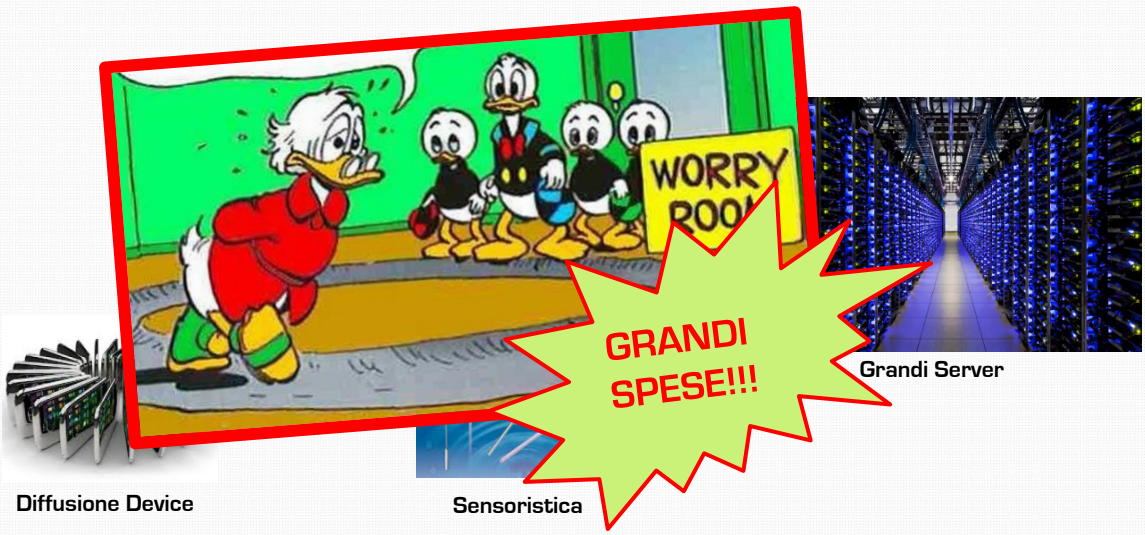
- Produce **300 TB al giorno**
- Il Database testuale è di circa **600 PetaBytes**
- Ogni mese vengono prodotti **7 PetaBytes** di sole fotografie

- Ogni tweet è composto da soli 140 caratteri
- Ogni giorno vengono prodotti 12 TB di tweet, ovvero **4 PetaBytes** all'anno

- Ogni singolo giorno Google gestisce **20 PetaBytes** di dati generati da utenti
- Più di **una trillione** di pagine web sono indicizzate correntemente



## IMPORTANZA DEI DATI - PROBLEMA




Diffusione Device

Sensoristica

Grandi Server

## AGENDA

- Importanza dei dati
- Approccio agnostico
- Cosa farci con i dati





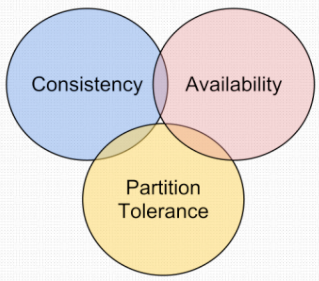
## APPROCCIO AGNOSTICO - TEOREMA CAP

**Consistency:** Ogni nodo del Database deve vedere ad ogni istante gli stessi dati.

**Availability:** Ogni richiesta deve avere una risposta, sia in caso di successo che in caso di fallimento

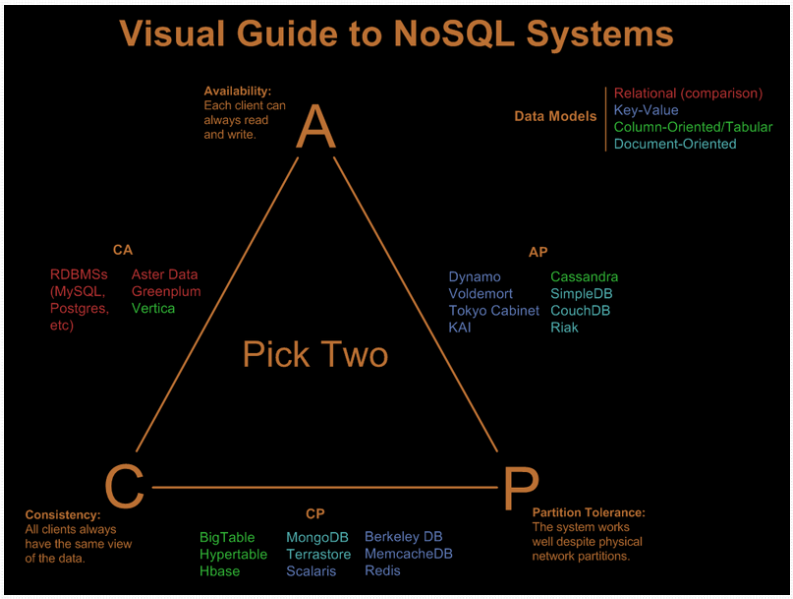
**Partition tolerance:** Il sistema deve continuare a funzionare anche in presenza di perdite di messaggi o di fault di parti dell'infrastruttura

**Teorema CAP:** Non ci sono database distribuiti che hanno contemporaneamente tutte e tre le proprietà, **al più due!**



## APPROCCIO AGNOSTICO - TEOREMA CAP

Storicamente i database tradizionali (RDBMS) quando hanno bisogno di scalare, decidono di concentrarsi sulla Consistency e sulla Availability, tralasciando la Partition Tolerance. Questo porta a non supportare una mole di dati "Big".






## APPROCCIO AGNOSTICO - DATI DESTRUTTURATI

Il tipo di dato che potre… …ro

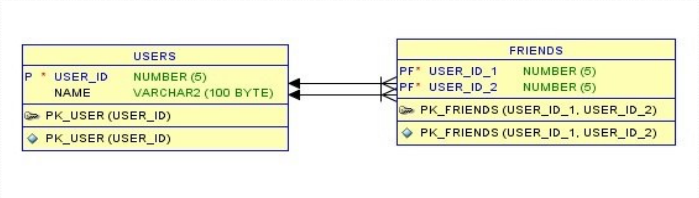
eterogeneo, non possia… …npi

VARCHAR… …o D…

E venir…

## APPROCCIO AGNOSTICO - RELAZIONALE FINO AD UN CERTO PUNTO

Poniamo il caso che Facebook decida di salvare le informazioni degli utenti e delle loro relazioni di amicizia in un DB relazionale:




Quindi per avere tutte le informazioni dei nostri amici dovremmo fare due join tra tutti gli utenti e tutte le relazioni.

Quindi per avere tutte le informazioni degli amici dei nostri amici dovremmo fare 3 join tra tutti gli utenti e tutte le relazioni.

Quindi per avere tutte le informazioni degli amici degli amici dei nostri amici dovremmo fare 4 join tra tutti gli utenti e tutte le relazioni.

.....



## APPROCCIO AGNOSTICO - RELAZIONALE FINO AD UN CERTO PUNTO

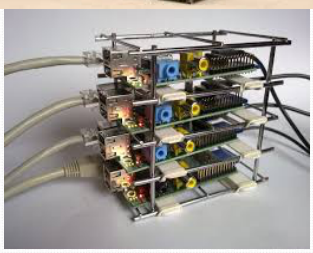
JSON
JavaScript Object Notation

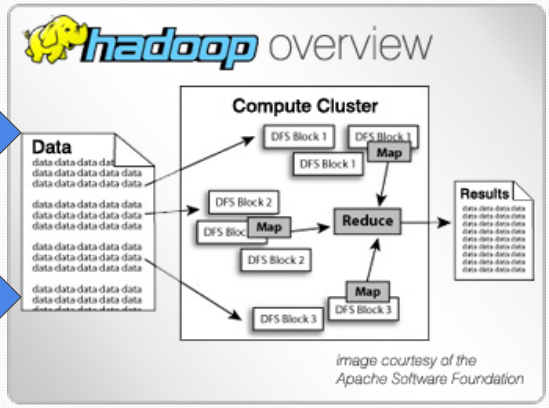
```
{
    hey: "guy",
    anumber: 243,
  - anobject: {
        whoa: "nuts",
      - anarray: [
            1,
            2,
            "thr<h1>ee"
        ],
        more: "stuff"
    },
    awesome: true,
    bogus: false,
    meaning: null,
    japanese: "明日がある。",
    link: http://jsonview.com,
    notLink: "http://jsonview.com is great"
}
```

I dati che vengono generati possono essere rinchiusi in formati innovativi.

Anche le forme di immagazzinamento di queste informazioni possono essere diverse dalla classica tabella.

## IMPORTANZA DEI DATI - HADOOP










## IMPORTANZA DEI DATI - HADOOP

The Apache Software Foundation
Community-led development since 1999.

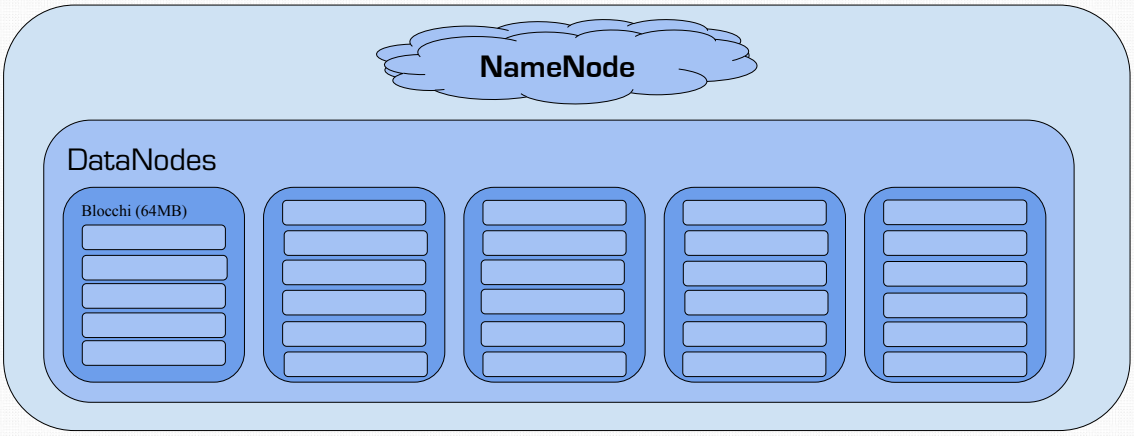
hadoop

Hadoop e il suo HDFS sono progetti opensource disponibili grazie all'Apache Software Foundation, all'indirizzo:

https://hadoop.apache.org/

- Può essere installato su qualsiasi sistema operativo Unix
- Unico prerequisito server: SSH
- Essendo scritto in Java, richiede una JDK 1.6 o superiore
- Non ci sono limiti di macchine su cui si può distribuire l'installazione
- Sono richieste discrete nozioni sistemistiche per completare l'installazione del pacchetto base

## IMPORTANZA DEI DATI - HADOOP




Struttura di un Cluster Hadoop



## AGENDA

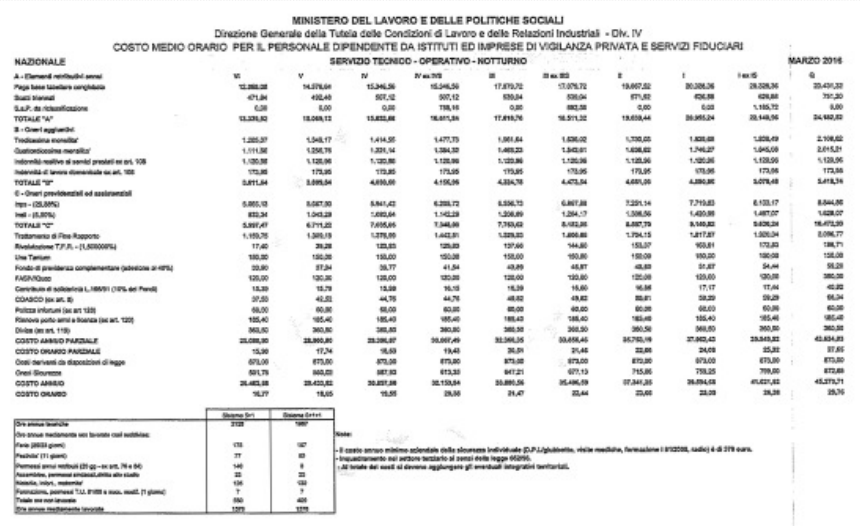
- Importanza dei dati
- Approccio agnostico
- Cosa farci con i dati

## COSA FARCI COI DATI

| Employee | City | State | Tenure |
|---|---|---|---|
| Brewer, Alan | Redmond | WA | 15 |
| Bento, Nuno | Redmond | WA | 10 |
| Ito, Shu | Redmond | WA | 15 |
| Han, Mu | Portland | OR | 3 |
| Hamilton, David | Eugene | OR | 1 |
| Bowen, Eli | Portland | OR | 15 |
| Brady,Mouton | Los Angeles | CA | 3 |
| Colleen,Marquez | San Diego | CA | 10 |
| Wendi,Conner | Los Angeles | CA | 2 |
| Ward,Cage | Phoenix | AZ | 3 |



## COSA FARCI COI DATI



MINISTERO DEL LAVORO E DELLE POLITICHE SOCIALI

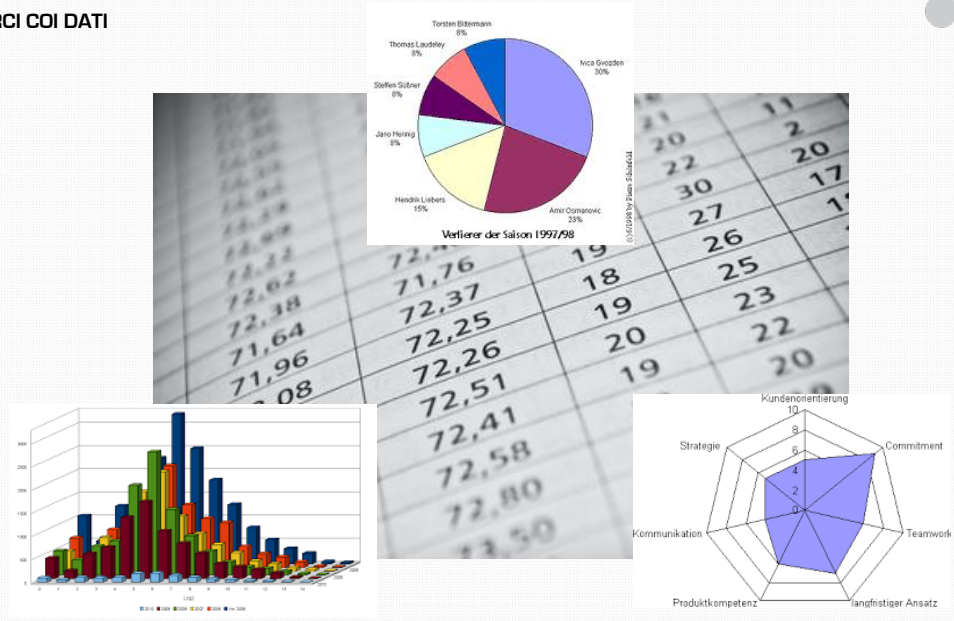
Direzione Generale della Tutela delle Condizioni di Lavoro e delle Relazioni Industriali - Div. IV

COSTO MEDIO ORARIO PER IL PERSONALE DIPENDENTE DA ISTITUTI ED IMPRESE DI VIGILANZA PRIVATA E SERVIZI FIDUCIARI

SERVIZIO TECNICO - OPERATIVO - NOTTURNO

NAZIONALE — MARZO 2016

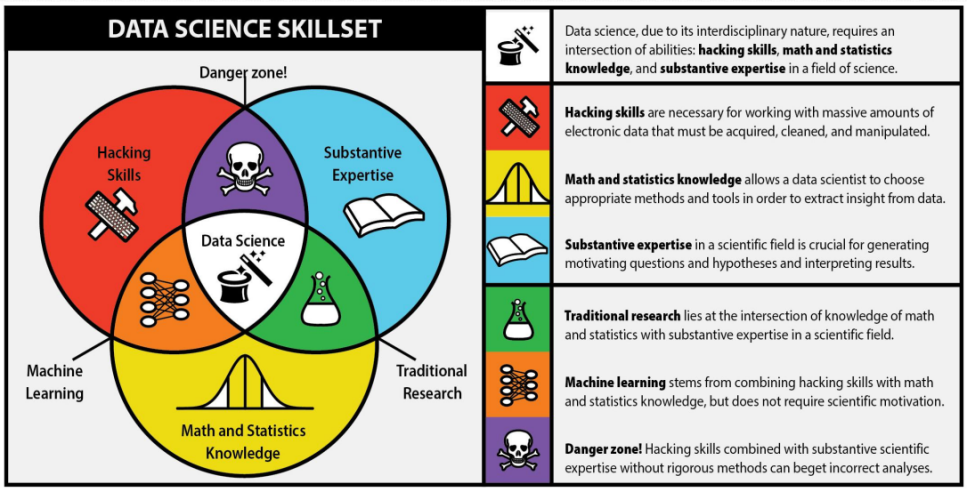
## COSA FARCI COI DATI






## COSA FARCI COI DATI

## COSA FARCI COI DATI - DATA SCIENTIST



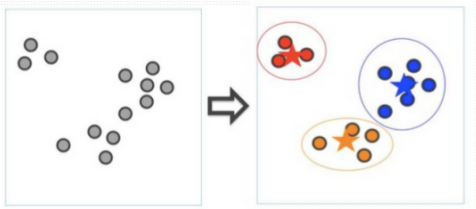
## COSA FARCI COI DATI - CHI PUNTA SUL MACHINE LEARNING



## COSA FARCI COI DATI - KMEANS

### KMeans

Algoritmo che permette di suddividere un dataset di *n* osservazioni in *k* cluster, individuando dei *k* centroidi e minimizzando iterativamente la distanza delle osservazioni dai centroidi.

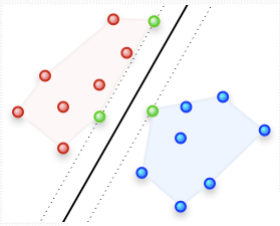


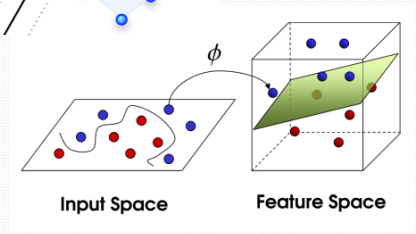
Algoritmo:

- Inizializzare i centroidi
- ripetere fino a convergenza:
  - per ogni elemento si ricalcola il centroide da associargli
  - per ogni centroide si ricolloca in base alla media degli elementi ad esso associati

## SVM

Uno degli algoritmi supervisionati più diffusi per la classificazione logistica, le **S**upport **V**ector **M**achines sono volte a trovare un iper-piano che sconnetta le osservazioni tra "good" e "bad" e massimizzare la distanza dei punti dall'iper-piano stesso.






Algoritmo:

- risolvere il problema di ottimizzazione convessa

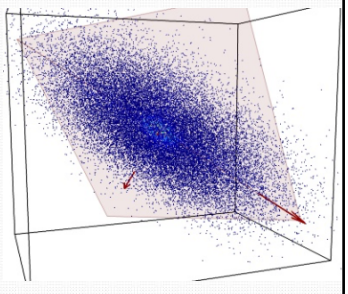
$$\min_{w,b} \sum_{i=1}^{n} \left(1 - Y_i(w^T X_i + b)\right)_+ + \lambda \|w\|^2$$





## PCA

**PCA** (**P**rincipal **C**omponent **A**nalysis) è uno degli algoritmi più diffusi in ambito scientifico per la riduzione della dimensionalità e per il preprocessing dei dati. Fa si che ci si possa restringere allo studio di variabili, iniziali o calcolate, che massimizzano la varianza e che quindi meglio caratterizzano il dataset.



Algoritmo:

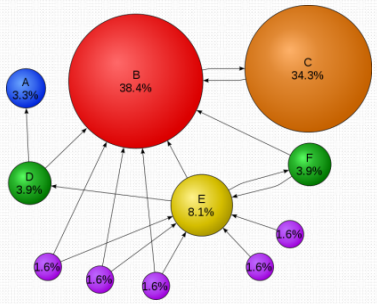
- sostanzialmente l'algoritmo consiste nel cercare, dato *k*, di ortogonalizzare una matrice associata al dataset, per poi prendere in considerazione come nuove direzioni solo i *k* autovettori con autovalori più alti.

## COSA FARCI COI DATI - PAGERANK

Algoritmo per l'assegnazione di un rank ad ogni nodo.

$$PR(A) = (1-d) + d\left(\frac{PR(P1)}{C(P1)} + \frac{PR(P2)}{C(P2)} + \ldots + \frac{PR(Pn)}{C(Pn)}\right)$$

- **PR(A)** sta per il ranking del nodo **A** che vogliamo calcolare
- **d** è il cosiddetto **Damping factor**: è un fattore che indica la probabilità che un visitatore decida di passare ad un altro nodo; alzando il valore **d** si abbasserà il valore del ranking totale del nodo
- **PR(P1), PR(P2),...** rappresentano i valori di ranking di **P1, P2,...** etc
- **C(P1), C(P2)** il numero complessivo di **link uscenti** dal nodo






## CONCLUSIONI


Google è senza dubbio una delle principali fonti di risposte a qualsiasi domanda noi possiamo porci!

**Come fa a decidere quale risposta sia la migliore?**

## CONCLUSIONI - Q&A




KEEP
CALM
AND
Grazie per
L'attenzione