

***dati***

## *parliamo di ...*

- perché ci interessano i dati?
- chi 'semina' dati?
- quanti dati?
- big data
- i nostri dati

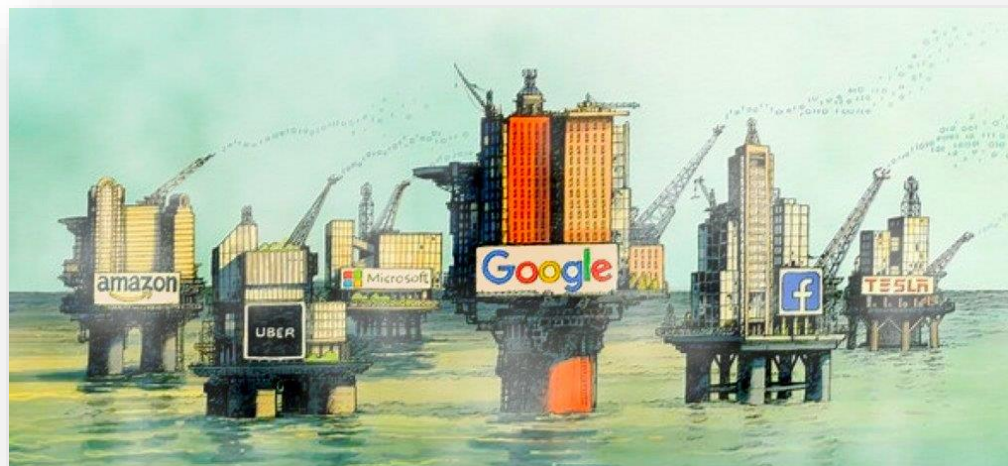


## *l'importanza dei dati*

*«i dati sono il nuovo petrolio»*

*Clive Humby, data scientist e matematico inglese (2006)*

- il **petrolio** ha permesso lo sviluppo socio-economico mondiale nel **XIX** e **XX** secolo
- le **connessioni**, le **tecnologie** ed i **dati** svolgono questo ruolo nel **XXI** secolo



*Alberto Ferrari – Analisi dei Dati*

## ***dati e petrolio***

- l'industria dei big data è un'industria **estrattiva**
  - il petrolio si ricava dalle profondità del suolo
  - il carbone si estrae dalle miniere
  - i dati personali vengono
    - **estratti in forma grezza** (es da internet)
    - poi vengono **raffinati** (aggregati per produrre informazione)



## *big data – una fra le tante definizioni*

- raccolta di dati così estesa in termini di *volume*, *velocità* e *varietà* da richiedere *strumenti non convenzionali* per estrapolare, gestire e processare informazioni entro un tempo ragionevole
- aumentando la scala dei dati di cui si dispone, *si possono fare cose nuove* che non sono possibili con minori quantità dei dati

## ***dato e informazione***

- ogni ***dato*** preso singolarmente è spesso ***privo di significato***
- l'organizzazione e la gestione di ***enormi quantità di dati*** suddivisi secondo un determinato criterio può fornire ***importanti informazioni***
  - queste informazioni possono poi essere utilizzate in modo da dare ***benefici***
  - ***o ...***
- scopo dei ***big data***:
  - analizzare enormi quantità di dati
  - estrapolare informazioni
  - in ***tempi*** ragionevoli
  - con ***risorse*** limitate







**POLLICINI DIGITALI**

*«siamo tutti pollicini digitali»  
Dino Pedreschi*

*Alberto Ferrari – Analisi dei Dati*

# Differenza con Pollicino

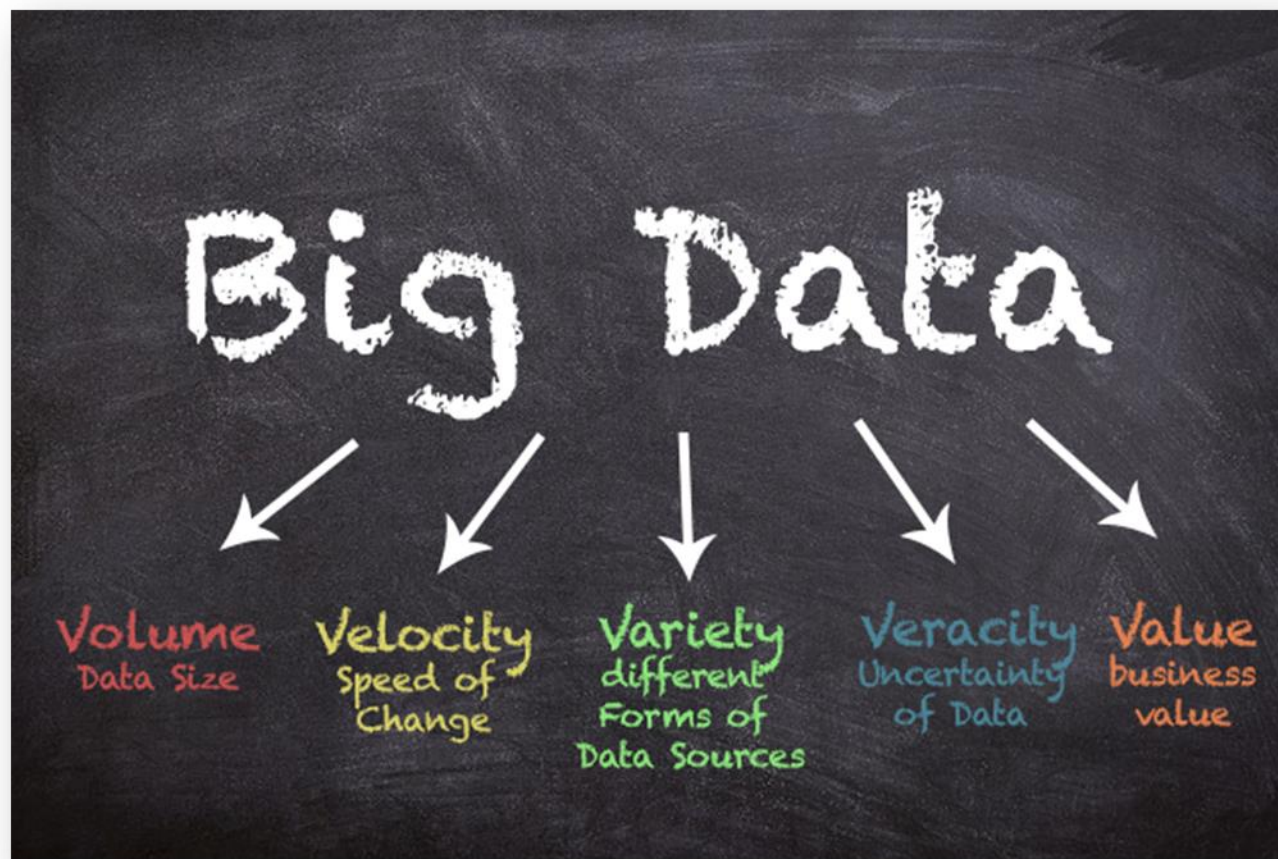
- ... Il giorno dopo, quando i genitori conducono i figli nella foresta con una scusa, Pollicino **lascia cadere i sassolini** dietro di sé; seguendo questa traccia riesce a riportare i fratelli a casa.
- e noi?
- siamo consapevoli dei dati che lasciamo lungo la strada?
- quali dati avete «lasciato lungo la strada» ieri?





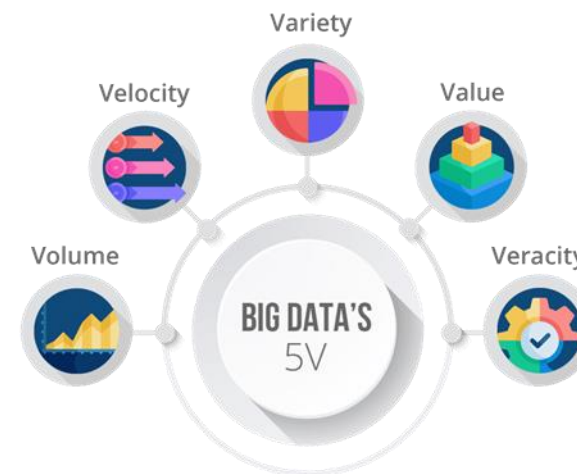
## *5 v – le caratteristiche dei big data*

- *volume*
- *velocità*
- *varietà*
- *veridicità*
- *variabilità*



## *volume*

- ***ogni giorno***, in moltissime attività della nostra vita quotidiana, ***generiamo dati***
- le tecnologie tradizionali non sono in grado di gestire l'ingente massa di informazioni che vengono generate
- il volume di dati è in continua ***crescita***
- è difficile identificare un valore limite al di sopra del quale si può parlare di Big Data



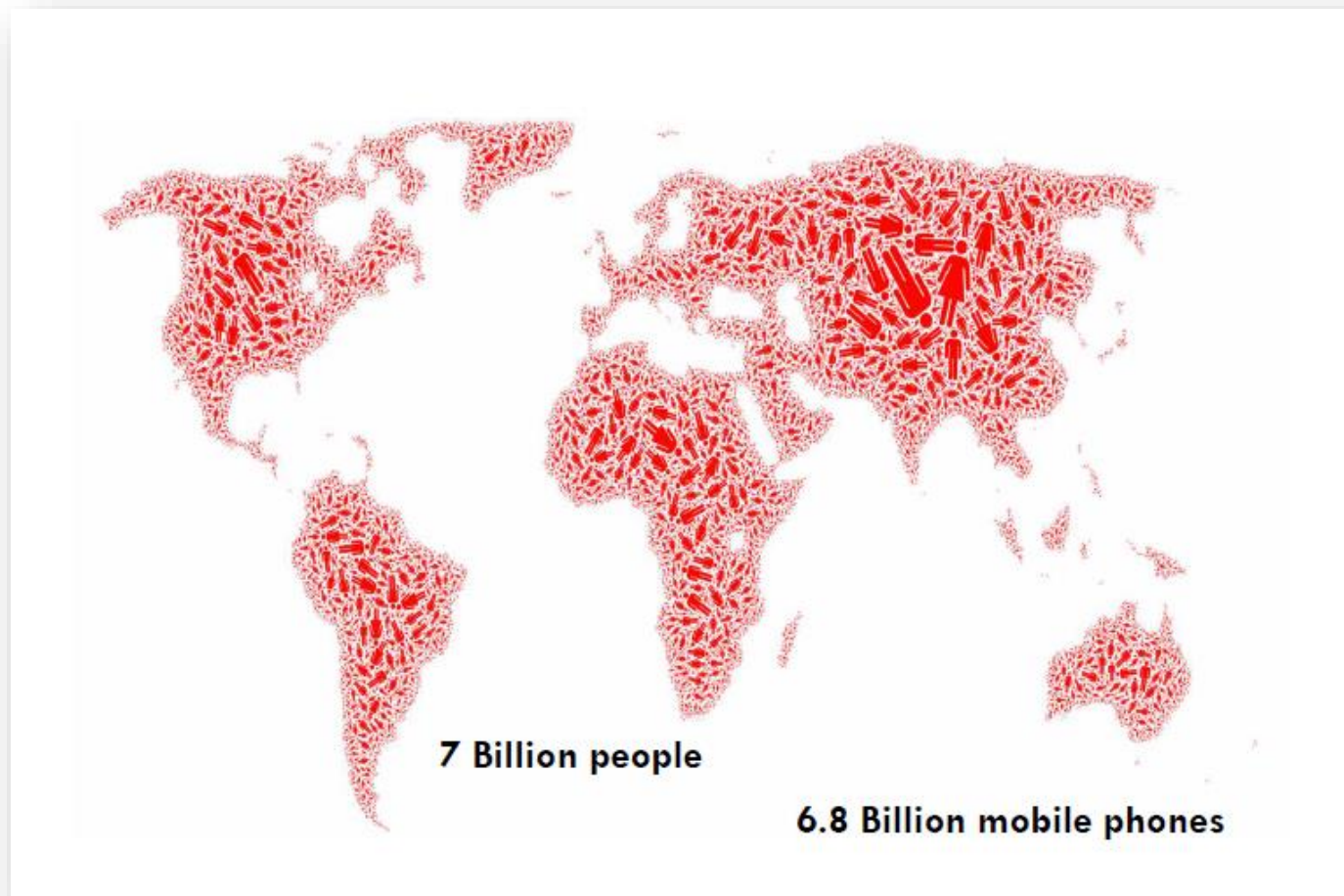
## *dove ‘seminiamo’ i nostri dati*

- ***Facebook***
  - testi, immagini, collegamenti (‘amici’) ...
- ***Google***
  - ricerche, cronologia, maps, ...
- informazioni sulla nostra attività fisica raccolte dagli ***smartwatch***
- gli spostamenti memorizzati dagli ***smartphone***
- la musica che ascoltiamo su ***Spotify***
- i film che vediamo su ***Netflix***
- ***tessere***
  - supermercati, librerie, ...
- ***acquisti***
  - carte di credito
- ...

## *conclave 2005 e 2013*



## *traffico telefonico*





<https://oggiscienza.it/2021/03/15/intelligenza-artificiale-calcio/>



*Alberto Ferrari – Analisi dei Dati*



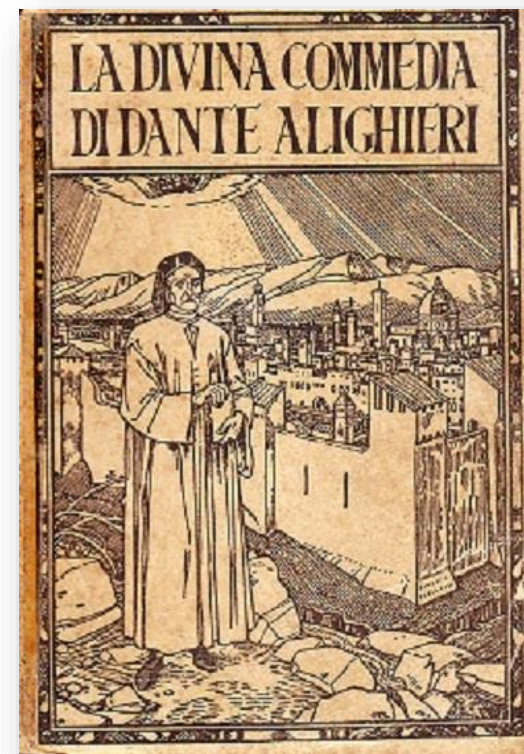
## *dati: unità di misura*

### 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 |

## *divina commedia*

- *La Divina Commedia* di Dante Alighieri è composta da 671.447 caratteri
- 1 carattere = 1 byte
- **670 Kb** = 1 Divina Commedia  $\simeq$  1 megabyte
- *universo digitale*
  - stima
    - attualmente **2.7 zettabyte**  
1 zettabyte equivale a un triliardo di byte
  - previsione
    - entro il 2025 **180 zettabyte**



## *origine dei dati – dati commerciali*

- Google ha Peta Byte di dati
- Facebook ha miliardi di utenti attivi
- Amazon gestisce milioni di visite/giorno
- Transazioni bancarie/carte di credito



... *numeri* ...

<https://www.infodata.ilsole24ore.com/2022/01/02/quanti-dati-generati-minuto/>

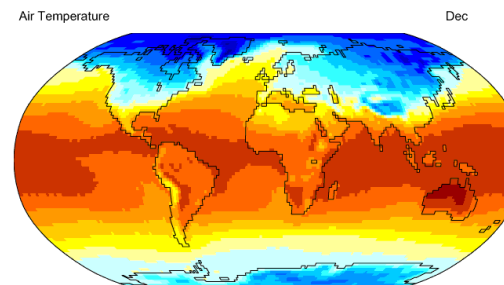
<https://www.wired.it/internet/web/2019/02/14/dati-internet-minuto/>

# 2021 *This Is What Happens In An Internet Minute*



## *origine dei dati – dati scientifici*

- dati raccolti e archiviati a *velocità enormi*
  - sensori remoti su satelliti
    - NASA EOSDIS genera più di un petabyte di dati ogni anno
  - telescopi che scrutano i cieli
  - simulazioni scientifiche
    - terabyte di dati generati in poche ore



## *dati - una stima di crescita*

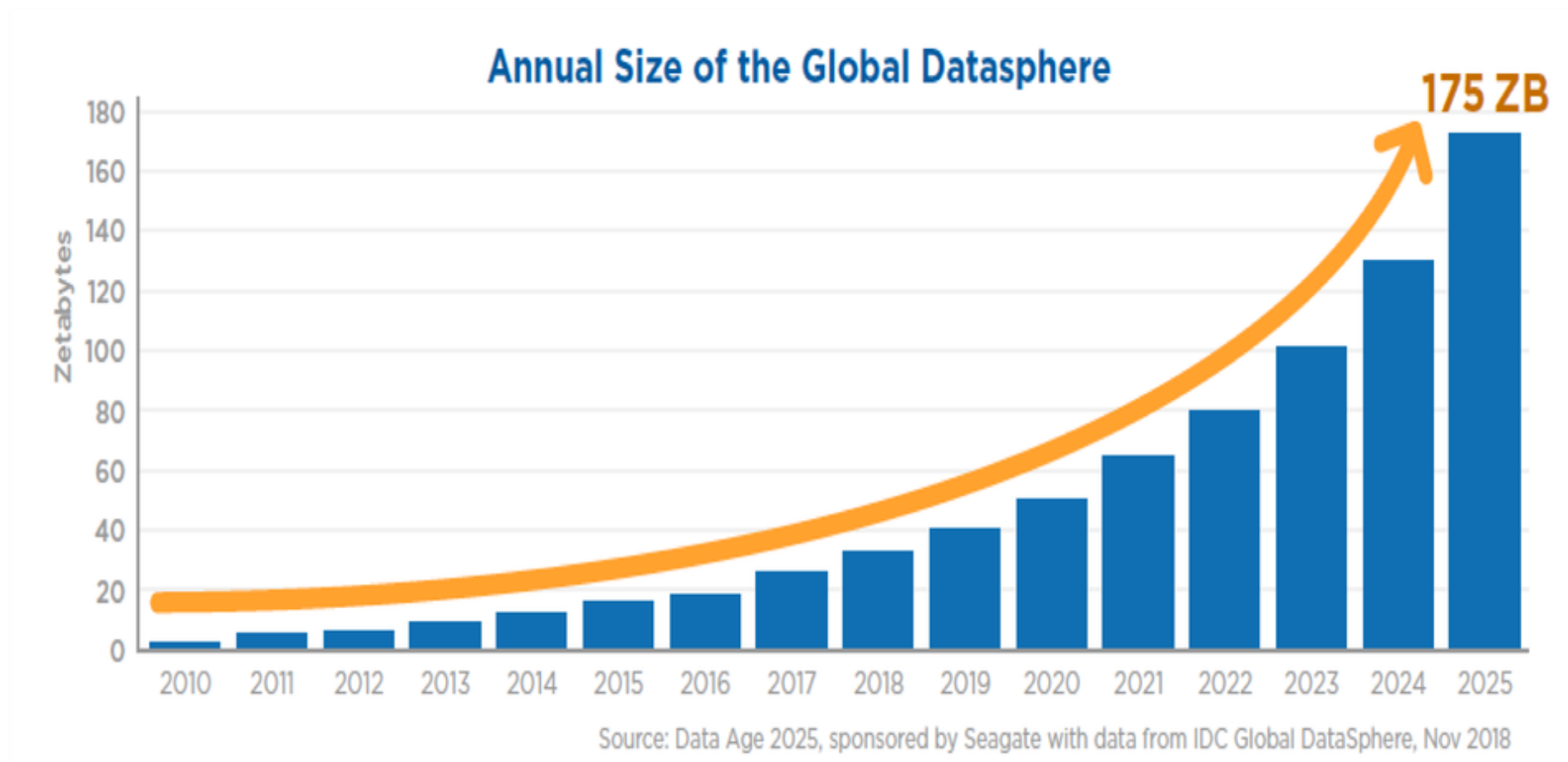
- i dati crescono in media del **30-40% annuo**
- ogni 2,5 anni si **raddoppia** il volume
  - oggi X
  - fra 2,5 anni  $X \cdot 2$
  - fra 5 anni  $X \cdot 4$
  - fra 7,5 anni  $X \cdot 8$
  - fra 10 anni  $X \cdot 16$
  - ...
  - fra 20 anni  $X \cdot 256$





## *Global DataSphere*

*quantità di dati creati, acquisiti e replicati in un dato anno in tutto il mondo*



## *enormi quantità di dati*

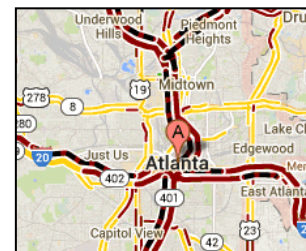
- nuovo *mantra*
  - *raccogli tutti i dati che puoi quando e dove possibile*
- *aspettative*
  - i dati raccolti avranno *valore* sia per lo scopo per cui sono stati raccolti sia per uno *scopo non previsto*



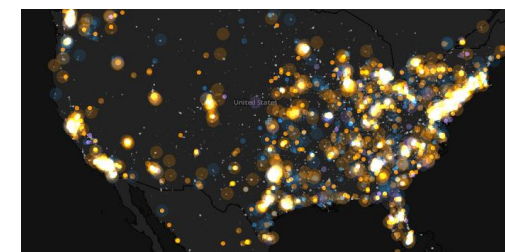
*Cyber Security*



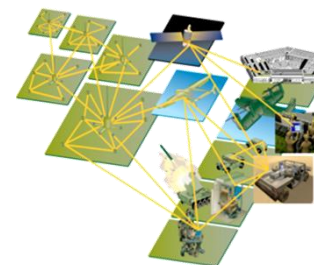
*E-Commerce*



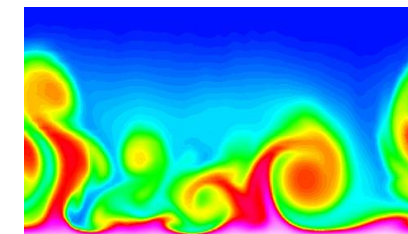
*Traffic Patterns*



*Social Networking: Twitter*



*Sensor Networks*



*Computational Simulations*

# *open data*

- dati ***liberamente accessibili*** a tutti, privi di brevetti o altre forme di controllo che ne limitino la riproduzione o l'utilizzo
- eventuali copyright si limitano all'obbligo di citazione della fonte o al rilascio delle modifiche con stessa tipologia di copyright



## *open data*

- <https://www.data.gov/>



- *The home of the U.S. Government's open data*
- *Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more*

- <https://www.dati.gov.it/>

- *Agenzia per l'Italia digitale*
- *i dati aperti della pubblica amministrazione*



**dati.gov.it**

*I dati aperti della Pubblica Amministrazione*

## *qualità - quantità*

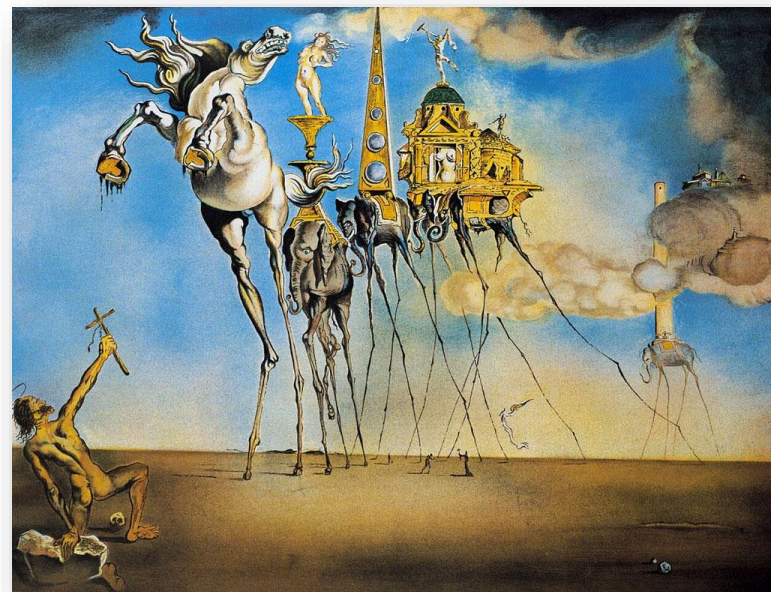
- nei big data, la **quantità** è più importante della qualità
- l'abbondanza permette di tollerare un certo livello di imprecisione



- es. google translate
  - prende le informazioni di cui ha bisogno per le sue traduzioni da pagine web non filtrate, piene di errori ortografici
  - ma l'enorme quantità di dati a disposizione gli permette di essere più affidabile di tutti i suoi predecessori, che si basavano su dizionari corretti e redatti da esperti, ma con il limite di contenere un numero limitato di informazioni

## *New York 1964*

- fiera dell'elettronica dimostrazione di un software di traduzione automatica dall'inglese al russo
- «lo spirito è forte ma la carne è debole»
- in russo il senso diventava «la vodka è forte ma la carne è marcia»



*«La tentazione di Sant'Antonio» Salvador Dalí*

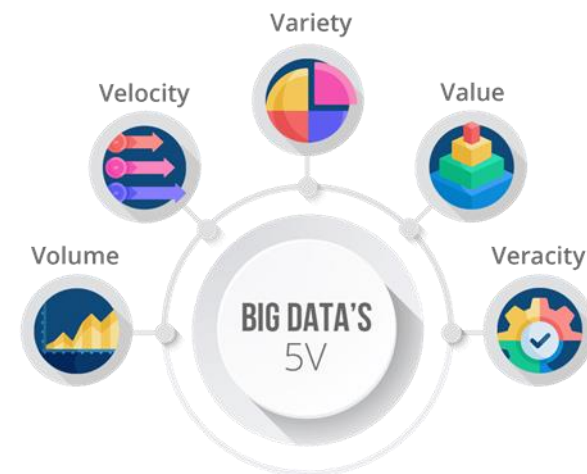


## *fattori determinanti per lo sviluppo dei big data*

- cloud computing
  - enormi quantità di dati memorizzabili in rete
  - servizi di elaborazione remota
- database più efficienti (NoSQL)
- machine learning verso deep learning
- disponibilità di tecnologie open source
  - Hadoop
  - Spark

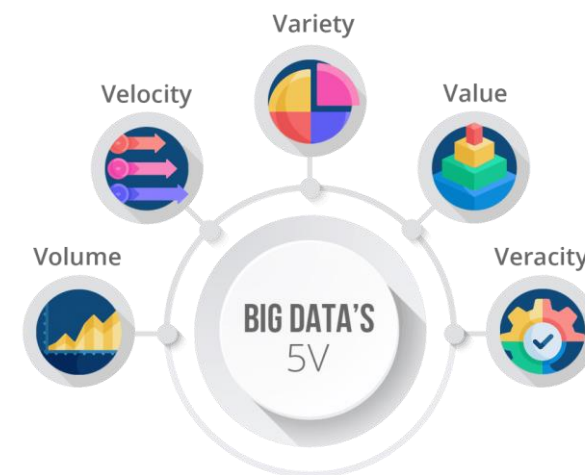
## *velocità*

- i dati vengono prodotti e acquisiti sempre più ***rapidamente***
  - dispositivi dotati di sensori capaci di raccogliere dati in ***tempo reale***
  - la ***sfida*** è avere la capacità di ***analizzarli in tempo reale*** per poter prendere decisioni con la maggiore tempestività possibile



## *varietà*

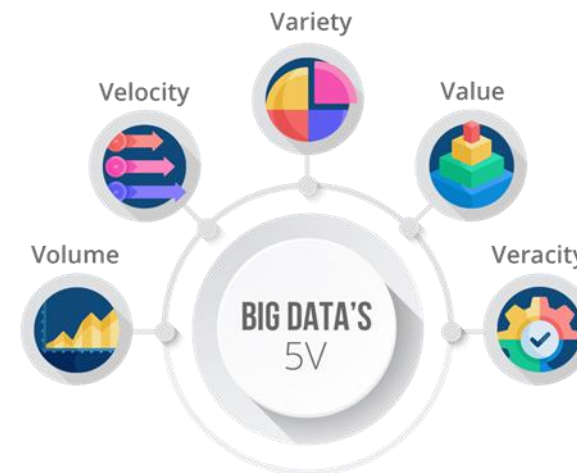
- i dati provengono da *fonti eterogenee*
- varie tipologie di dati
  - sensori
  - social network
  - open data
- dati *strutturati* o *non strutturati*
- *interni* o *esterni* all'organizzazione



*“More isn’t just more. More is different”  
Chris Anderson (Wired 2008)*

## *veridicità*

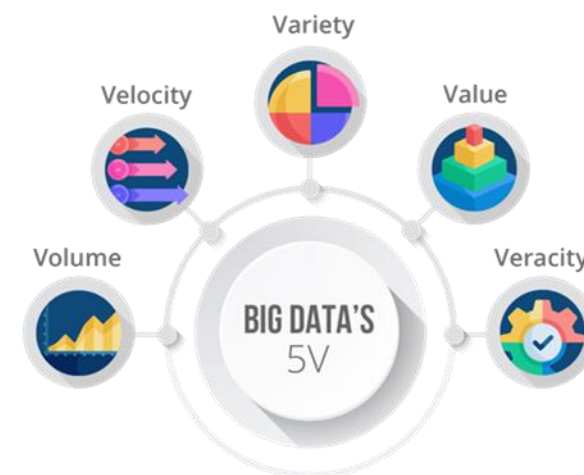
- i dati devono essere *affidabili*
- devono dire il vero
- la qualità e l'integrità delle informazioni rimane un pilastro imprescindibile per portare ad analisi utili e affidabili



*“Bad data is worse than no data”*

## *variabilità*

- molti dati
  - in *diversi formati*
  - provenienti da *diversi contesti*
- la *mutevolezza* del loro significato è un aspetto da tenere in considerazione nel momento in cui i dati vengono interpretati



## *data science e big data*

- *scienza dei dati*
  - studia metodi per estrarre **conoscenza** dai dati
  - opera con dati di qualunque natura
- data science non necessita sempre di big data
  - la costante crescita dei dati fa si che i big data siano un aspetto importante della data science



## *analisi dei big data - finalità*

- ***medicina***

- prevedere la diffusione delle malattie
- contrastare possibili epidemie

- ***business***

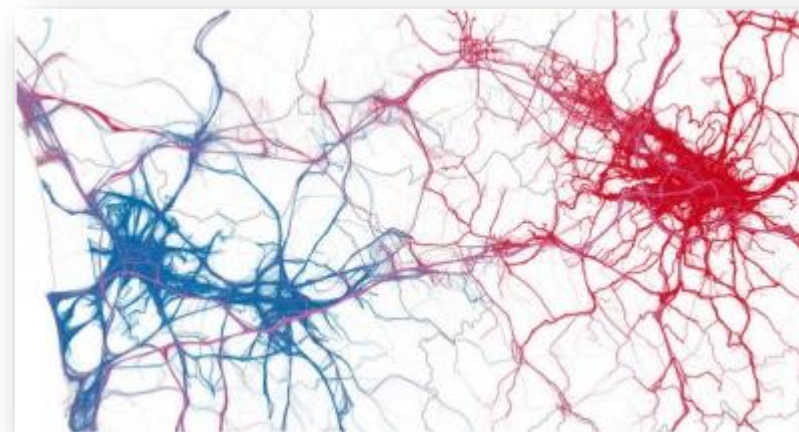
- analizzare comportamenti di acquisto dei consumatori
- monitorare feedback delle promozioni e offerte
- studiare le campagne di marketing

- ***ambiente***

- studiare eventi metereologici
- Giappone, big data del meteo e super computer per previsioni migliori
  - <https://tg24.sky.it/ambiente/2018/01/18/giappone-big-data-meteo>

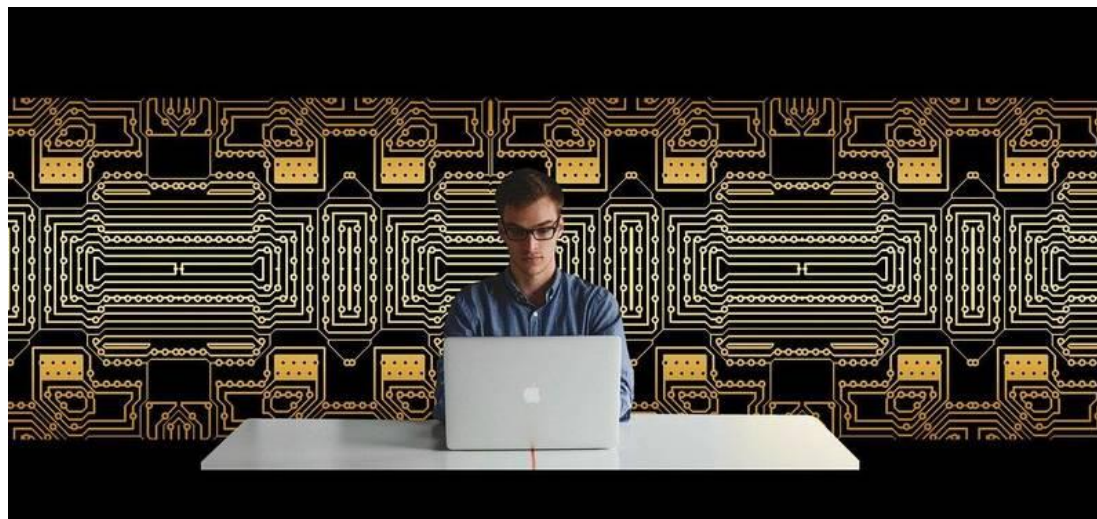
## *analisi dei big data - finalità*

- ***sport***
  - definire strategie di gioco
  - studiare strategie degli avversari
  - valutazione performance
  - <https://innovaformazione.net/big-data-e-calcio/>
- ***trasporti***
  - migliorare la gestione del traffico in tempo reale
- ***sicurezza***
  - prevenire attentati terroristici



# *big data*

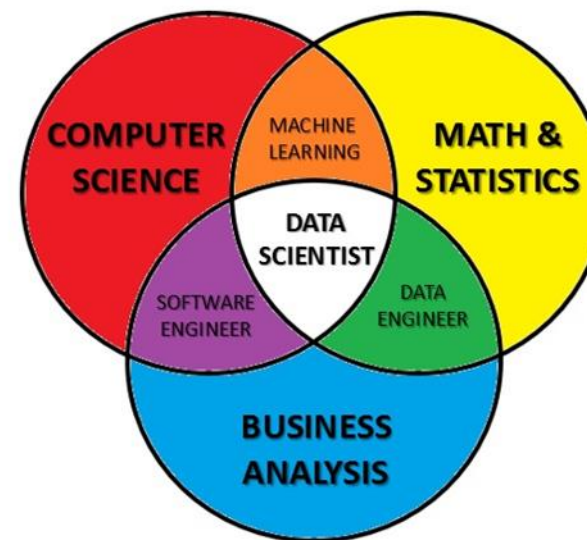
*le varie professioni*



*Alberto Ferrari – Analisi dei Dati*

## *data scientist*

- gestisce i big data (*dati grezzi*)
- *trae informazioni* rilevanti per
  - strategie di business
  - strategie di marketing e di vendita
  - definizione di nuovi prodotti e servizi, ecc.
- profilo:
  - conoscenza di *modelli matematico-statistici* e algoritmi di *machine learning*
  - conoscenza dei *linguaggi di programmazione* (R, Python)
  - competenze di business intelligence, di semantica, di ontologie per la gestione delle informazioni, di metodi e tecnologie per la gestione di progetti data-driven innovativi, di machine learning.
  - tecniche di data mining
    - clustering
    - analisi della regressione....
- laurea avanzata (Master, PhD) in informatica



## *data engineer*

- garantire la **disponibilità**, la qualità e la **fruibilità** dei dati a chi li utilizza
- gestire processi, individuare opportunità e rischi
- competenze informatiche e ingegneristiche per aggregare, analizzare e manipolare insiemi di big data
- creazione di algoritmi informatici, sviluppo di processi tecnici per migliorare l'accessibilità dei dati e la progettazione di report e strumenti per gli utenti finali
- competenza nella progettazione di **database**, padronanza di linguaggi di programmazione
- capacità di **comunicazione scritta e verbale**, capacità di lavorare sia in modo indipendente che in team

## *data analyst*

- analizza e interpreta i dati per *trasformarli* in informazioni utili al processo decisionale
- il data scientist è il data analyst avanzato
- lavora con i team di ingegneri per ottenere i dati corretti
- eseguire il *data munging*
  - trasforma i dati grezzi in dati nel formato utile per l'analisi/interpretazione e per ricavare informazioni dai dati
- lavora su database strutturati
- buona conoscenza di programmi informatici (Excel, Access...)
- buone capacità di comunicazione e di presentazione

## *security engineer*

- svolgono un ruolo di grande responsabilità: **difesa** rispetto a problemi informatici e possibili **attacchi**
- hacker buono: evita o risolve problemi di **sicurezza** sui dati
- definisce protocolli di **protezione** per le reti informatiche
- laurea in ingegneria, informatica e certificazioni di sicurezza industriale
- conoscenza tecnica dei linguaggi informatici e dei sistemi operativi, capacità di problem solving
- la capacità di lavorare in modo indipendente e rimanere costantemente aggiornati



## *database manager*

- responsabilità del *funzionamento* e del miglioramento dei *database*
- diagnostica e riparazione di database danneggiati
- aggiornare i sistemi di gestione di basi di dati in base agli *sviluppi tecnologici*
- laurea in tecnologia dell'informazione
- buona conoscenza dei software per la *gestione dei database* (MySQL, Oracle)

## *data architect*

- ***progettano i sistemi informativi***, i flussi e i repository dei dati in base alle necessità dell'azienda
- conoscenza dei linguaggi orientati ai dati per organizzare e mantenere i dati in database
- ***competenze tecniche*** avanzate (SQL, XML
- acume analitico e capacità di problem-solving
- laurea di primo livello (spesso laurea avanzata) in un campo legato all'informatica

## *big data problemi e opportunità*

- Così big data e intelligenza artificiale stanno battendo il coronavirus in Cina
  - <https://www.ilsole24ore.com/art/la-macchina-tech-xi-jinping-cosi-big-data-e-intelligenza-artificiale-stanno-battendo-coronavirus-cina-ADsL0XB>
- Big data per i servizi pubblici: tutti i vantaggi e i rischi
  - <https://www.agendadigitale.eu/cittadinanza-digitale/big-data-per-i-servizi-pubblici-tutti-i-vantaggi-e-i-rischi/>
- Il compromesso tra dati e libertà
  - <https://nova.ilsole24ore.com/frontiere/il-compresso-tra-dati-e-liberta/>

## *materiale di consultazione*

- Superquark – Big Data (8 min)
  - [https://www.youtube.com/watch?v=A2pUx5B\\_C4A](https://www.youtube.com/watch?v=A2pUx5B_C4A)
- Superquark - Big data e pandemia (5 min)
  - <https://www.youtube.com/watch?v=3WvaZoZhwjU>
- come funzionano i filtri di snapchat (5 min)  
[generare automaticamente i sottotitoli in italiano]
  - <https://www.youtube.com/watch?v=Pc2aJxnmzh0>
- Progetti di analisi di big data
  - Guida ai prodotti più acquistati durante l'emergenza Covid
    - <https://masterbigdata.it/project-files/2019-2020/panicBuying/>
  - EVERYBODY'S CHANGING (How has music changed in the past few years?)
    - <https://masterbigdata.it/sites/default/files/projects/2020-2021/musica/#progetto>

# Shoshana Zuboff – Il capitalismo della sorveglianza



cosa sono o cookies?



I cookie sono piccoli file di testo che i siti web salvano sul tuo computer o dispositivo quando li visiti. Servono per vari scopi, tra cui:

1. **Memorizzazione delle preferenze:** Ricordano le tue impostazioni, come la lingua o il layout della pagina.
2. **Tracciamento della sessione:** Aiutano a mantenerti connesso durante la navigazione, ad esempio quando accedi a un account.
3. **Analisi e pubblicità:** Raccolgono informazioni sulle tue abitudini di navigazione per offrire annunci personalizzati o migliorare l'esperienza utente.

I cookie possono essere "di sessione" (che si cancellano quando chiudi il browser) o "persistenti" (che rimangono finché non vengono eliminati). Puoi gestire le impostazioni dei cookie direttamente nel tuo browser.

## Chi può consultare le informazioni presenti in un cookie?



Le informazioni contenute in un cookie possono essere consultate principalmente da tre parti:

1. **Il sito web che ha creato il cookie:** Solo il sito che ha impostato il cookie può leggerne i dati. Questo consente al sito di ricordare le tue preferenze o le informazioni di accesso.
2. **Siti web di terze parti:** Se un sito incorpora contenuti di terze parti (come annunci pubblicitari o plugin social), questi possono impostare i propri cookie e accedere a quelli già presenti. Questo è comune per il tracciamento pubblicitario.
3. **L'utente stesso:** Gli utenti possono visualizzare e gestire i cookie attraverso le impostazioni del proprio browser, che permettono di vedere quali cookie sono stati memorizzati e di eliminarli se necessario.

In generale, l'accesso ai dati contenuti in un cookie è limitato al dominio che lo ha creato, per motivi di sicurezza e privacy.



## *cookies e privacy*

- <https://www.garanteprivacy.it/faq/cookie>
- <https://www.wired.it/article/fingerprinting-cosa-e-tracciamento/>



## *quali dati condividiamo?*

- Facebook
  - impostazione privacy
- Instagram
- Twitter
- ...

