Università degli Studi di Palermo

dipartimento di ingegneria unipa

Introduzione al Corso

CORSO DI BIG DATA

a.a. 2021/2022

Prof. Roberto Pirrone

LABORATORIO DI INTERAZIONE UOMO-MACCHINA
CHILAB

Sommario

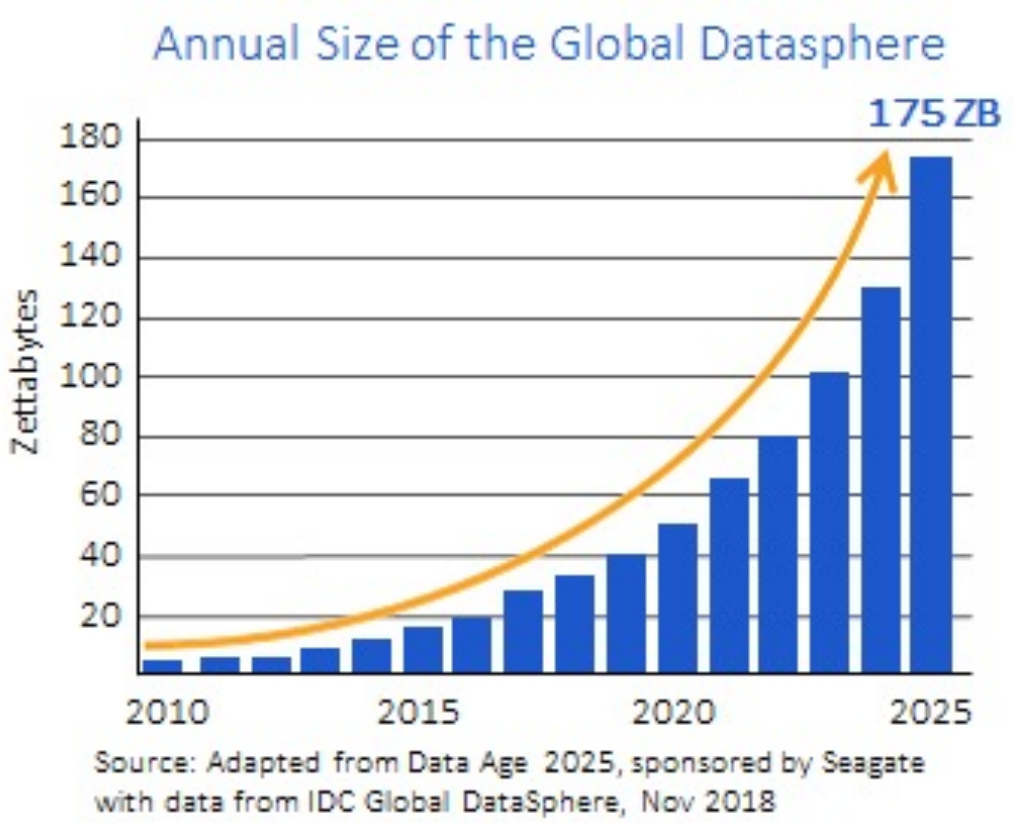
- Il docente
- Perché «Big Data»
- Cosa non è «Big Data»
- Cosa è «Big Data»
- Il Syllabus
- Il materiale didattico
- Gli esami
- Le tesi di laurea

Il Docente

- Roberto Pirrone
  - Studio: Edificio 6, terzo piano, stanza 8
  - Email: roberto.pirrone@unipa.it, roberto.pirrone@you.unipa.it (Microsoft) roberto.pirrone@community.unipa.it (Google)
  - Telefono studio: 091238.62625, laboratorio: .62643
  - Ricevimento: ogni giovedì dalle 11 alle 13 sul team con codice: 4rylimr

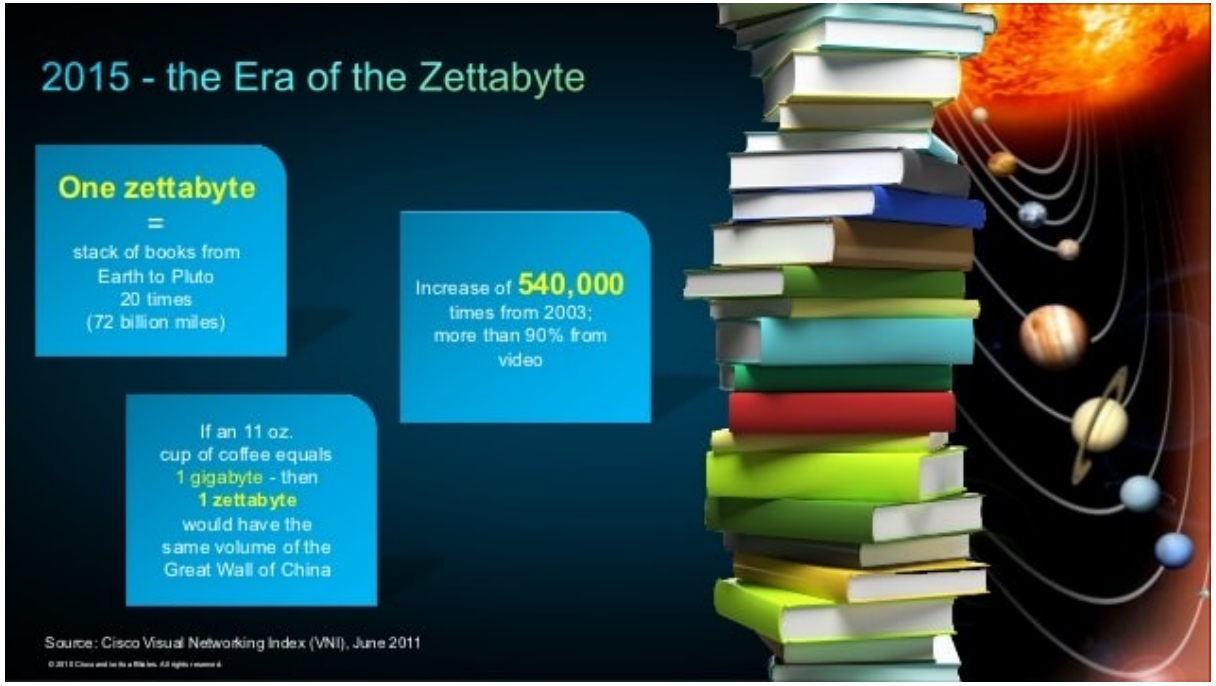
Perché «Big Data»

- Perché i dati sono diventati «Big»
  - Ad oggi si stima una produzione annua di dati di oltre 60 ZB nel 2021
    - 1 ZB = 10^{21} B
  - ***175 ZB nel 2025***



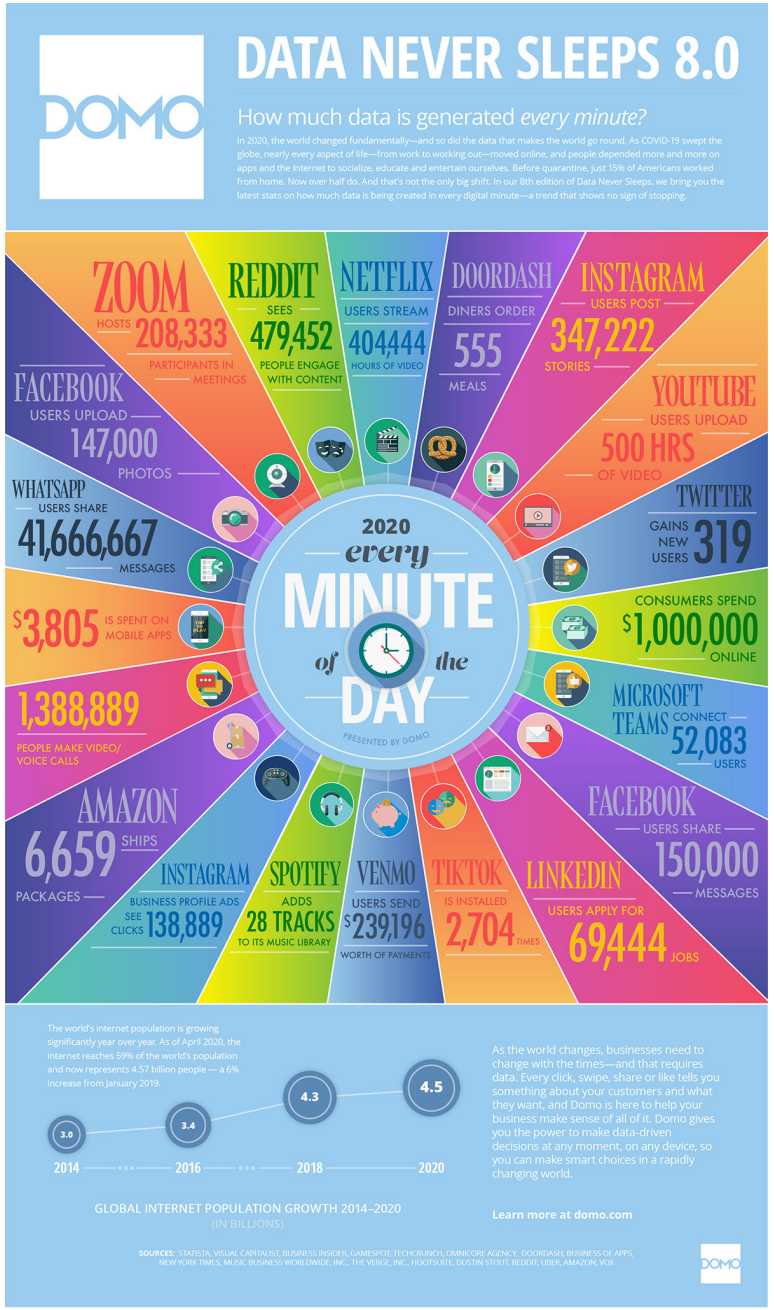
Perché «Big Data»

- Perché i dati sono diventati «*Big*»
  - Quanta informazione c'è in uno ZB?
    - ***Una catasta di libri 20 volte la distanza Terra-Plutone***
    - ***Il volume della Grande Muraglia Cinese***, posto che 1 GB == 1 tazza di caffè americano



Perché «Big Data»

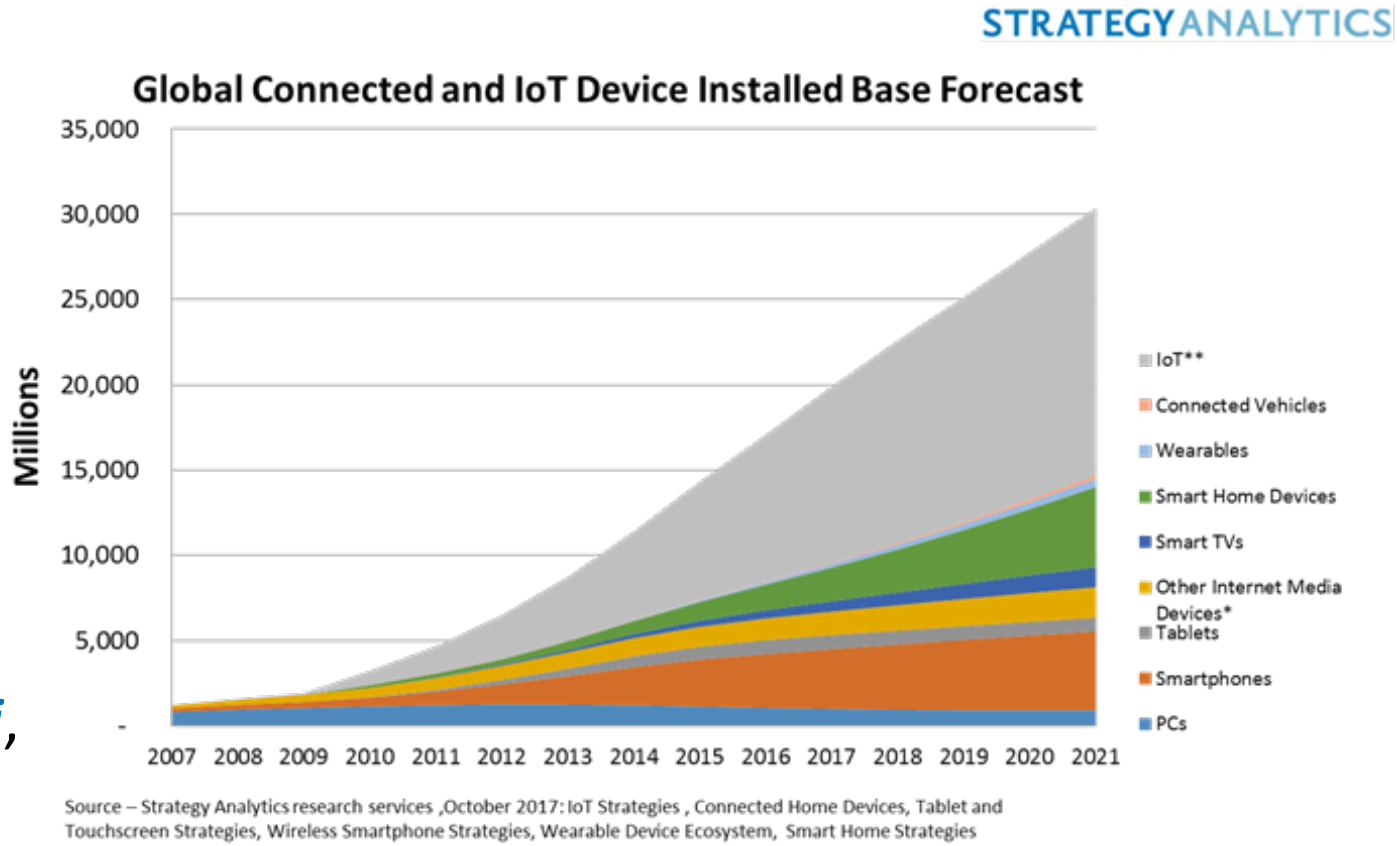
- Perché i dati sono diventati *vari ed eterogenei*
  - Internet e social media



Fonte https://www.visualcapitalist.com/every-minute-internet-2020/

Perché «Big Data»

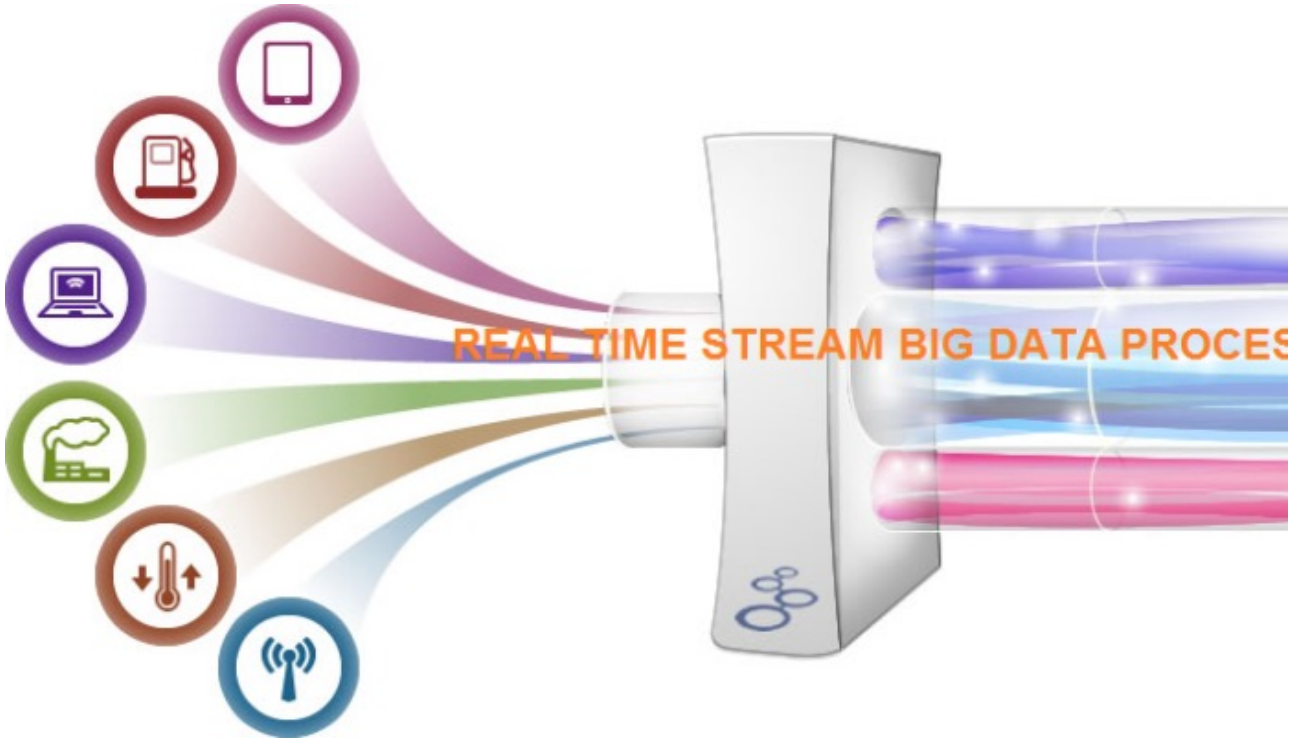
- Perché i dati sono diventati *vari ed eterogenei*
  - I device e i sensori connessi a Internet (IoT – *Internet of Things*)
  - Dati *strutturati*, *semi-strutturati*, *non strutturati*



Fonte https://www.digitaltveurope.com/2017/10/27/strategy-analytics-iot-to-reach-50-billion-devices/

Perché «Big Data»

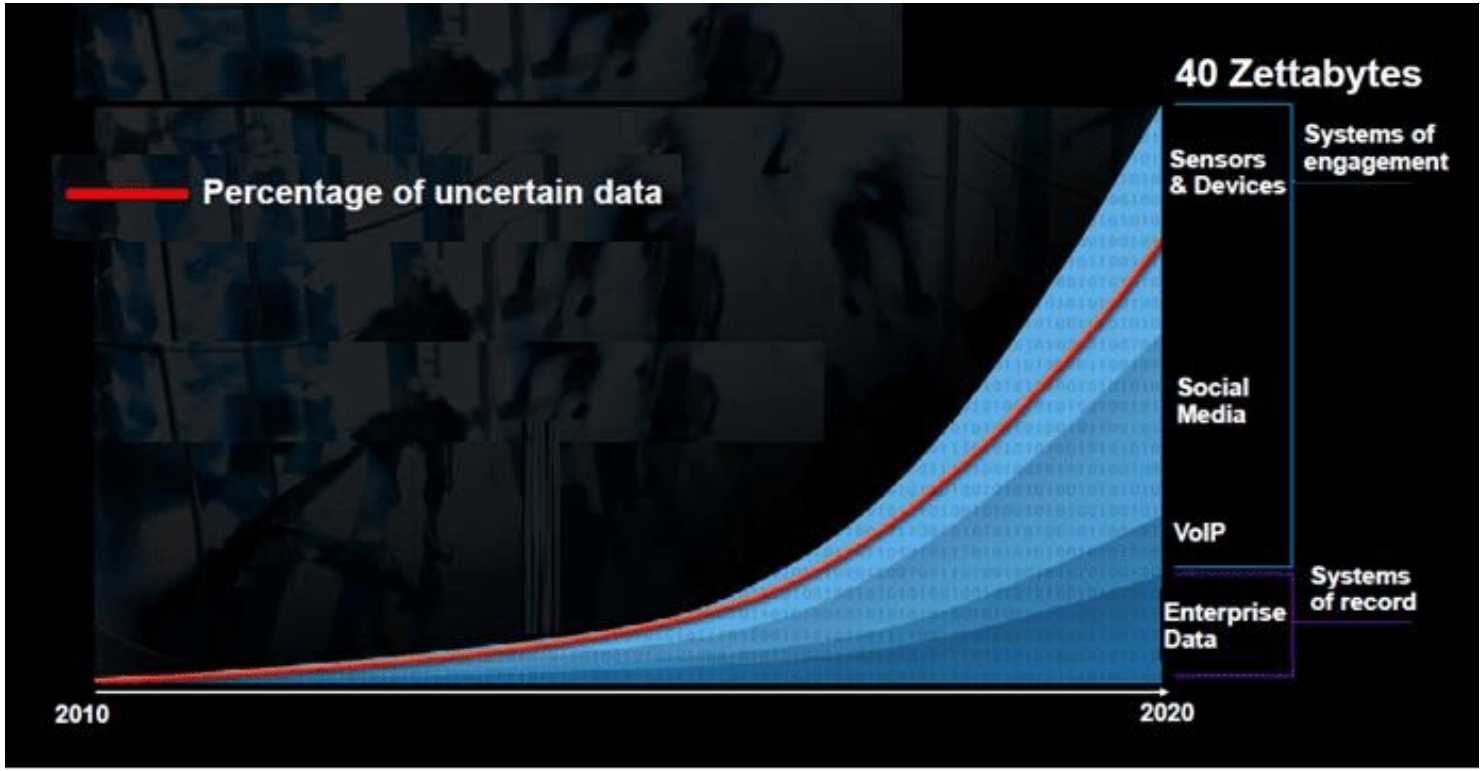
- Perché i flussi di dati sono quasi sempre in *real time*
  - IoT
  - User Generated Contents
  - Monitoraggio ambientale
  - Automotive
  - Monitoraggio della rete
  - Dati di cloud
  - …



Fonte https://www.thedigitaltransformationpeople.com/channels/enabling-technologies/real-time-stream-processing-in-big-data-platform/

Perché «Big Data»

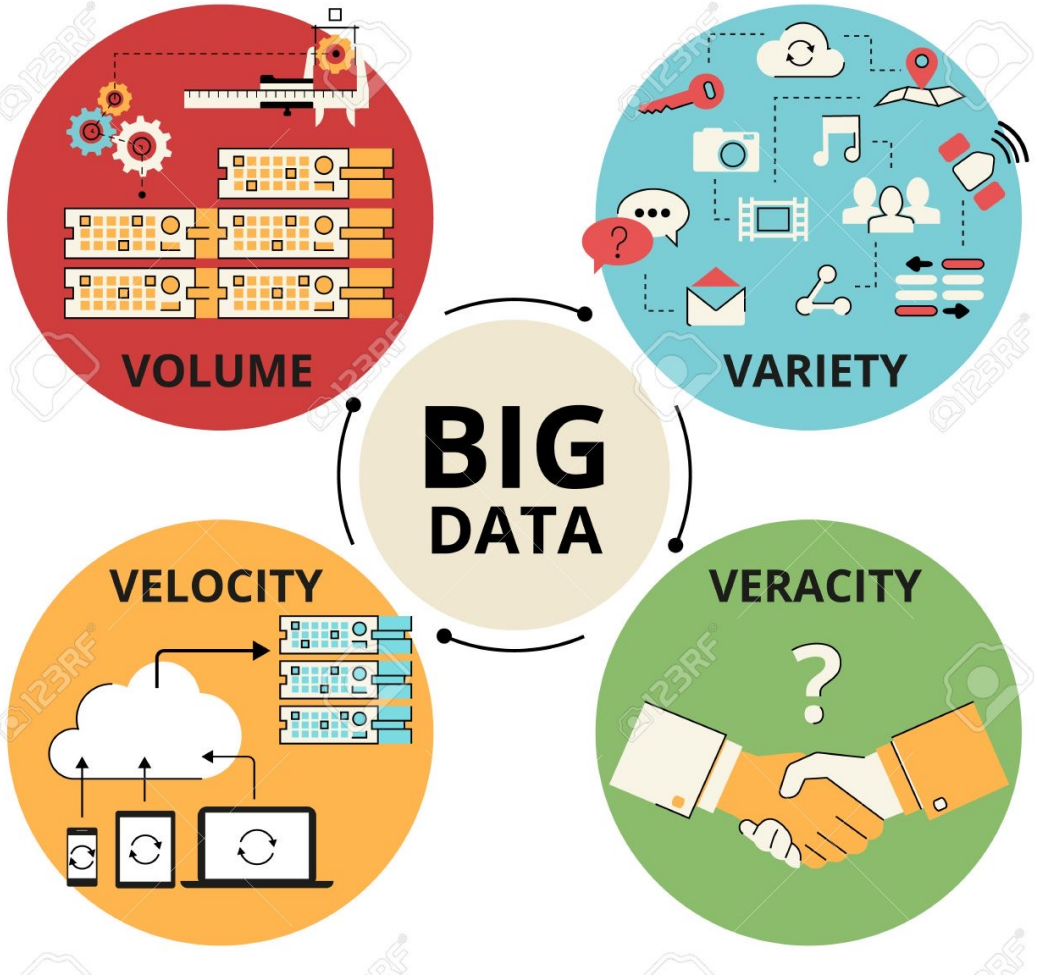
- Perché i flussi di dati sono quasi sempre in *di origine incerta*



Fonte https://www.researchgate.net/figure/Projected-Growth-of-Big-Data-based-on-1_fig2_272391443

Perché «Big Data»

- Le dimensioni rispetto alle quali si analizzano i Big Data vengono denominate *le quattro V*
  - *Volume*
  - *Velocità*
  - *Varietà*
  - *Veridicità*

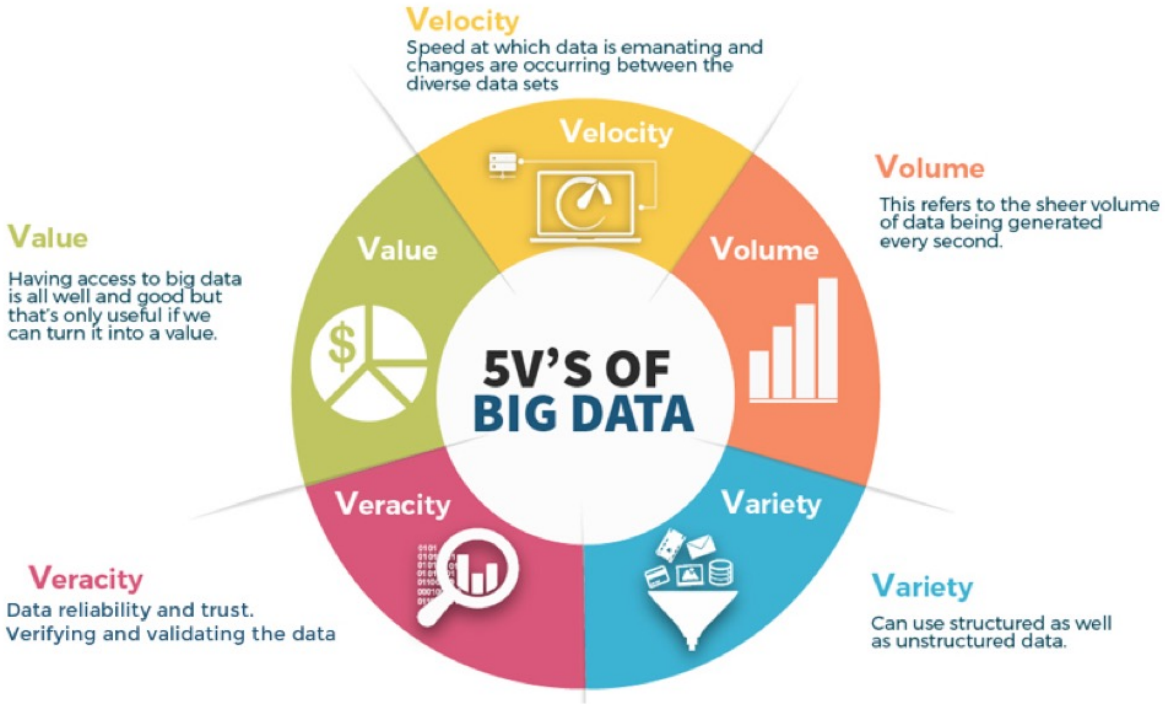


Fonte https://www.123rf.com/photo_44494927_stock-vector-infographic-flat-contour-concept-illustration-of-big-data-4v-visualisation-.html

Perché «Big Data»

- Alle dimensioni precedenti si aggiunge una quinta e quindi si parla de *le cinque V*

  - *Volume*
  - *Velocità*
  - *Varietà*
  - *Veridicità*
  - *Valore*



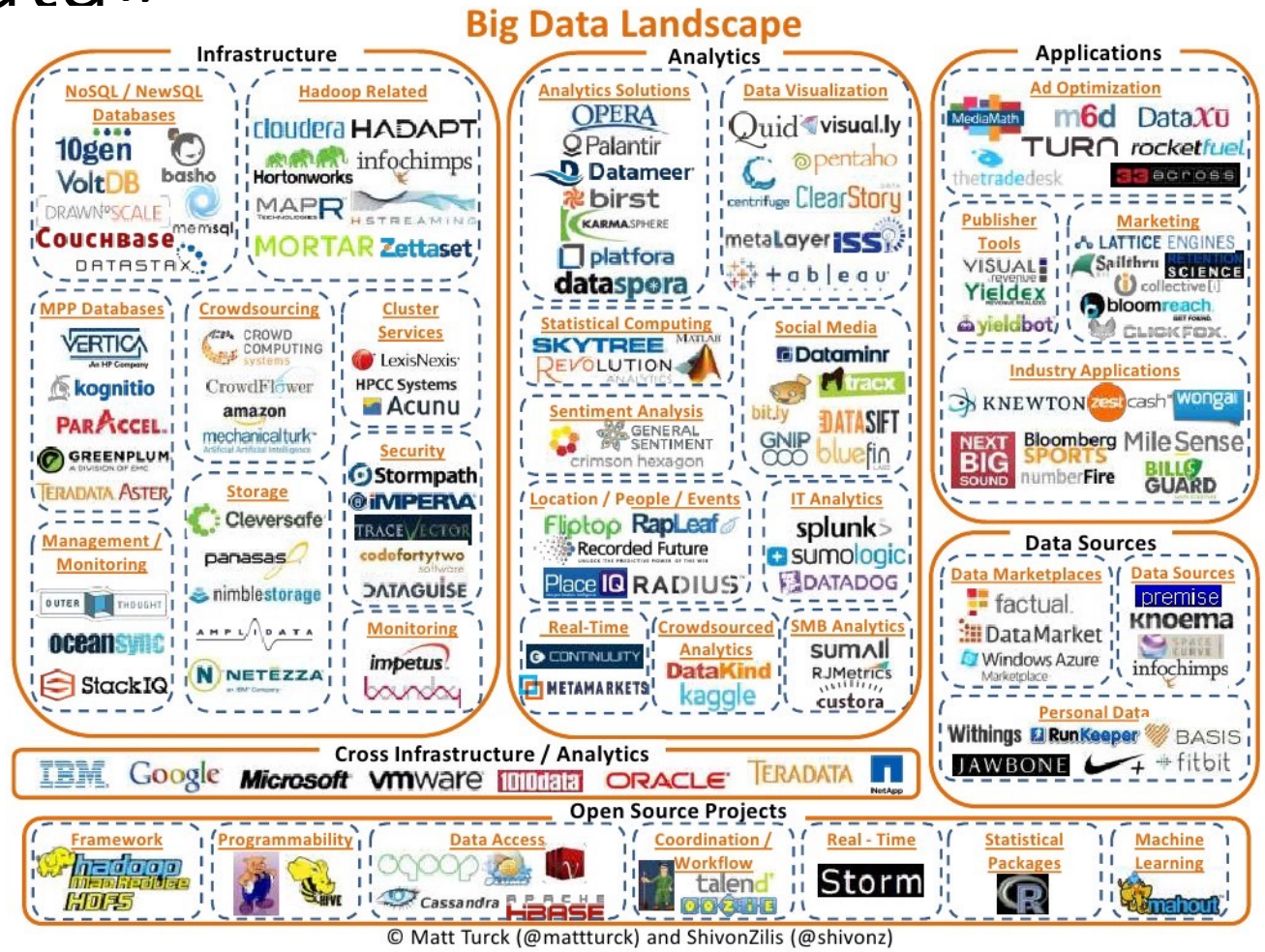
Fonte https://www.techentice.com/the-data-veracity-big-data/

Cosa non è «Big Data»

- Il corso di «Big Data» *non è*:
  - Un corso di Python (anche se lo studieremo abbastanza)
  - Una serie di tutorial su framework più o meno esoterici (anche se ne studieremo diversi)
  - Un corso di Machine Learning (anche se ne studieremo un bel po')

Cosa non è «Big Data»

- Non è possibile studiare nel dettaglio tutte le soluzioni software che gravitano nel mondo dei Big Data!!!



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

Fonte https://medium.com/be-data-driven/what-is-data-engineering-fe158db36c1e

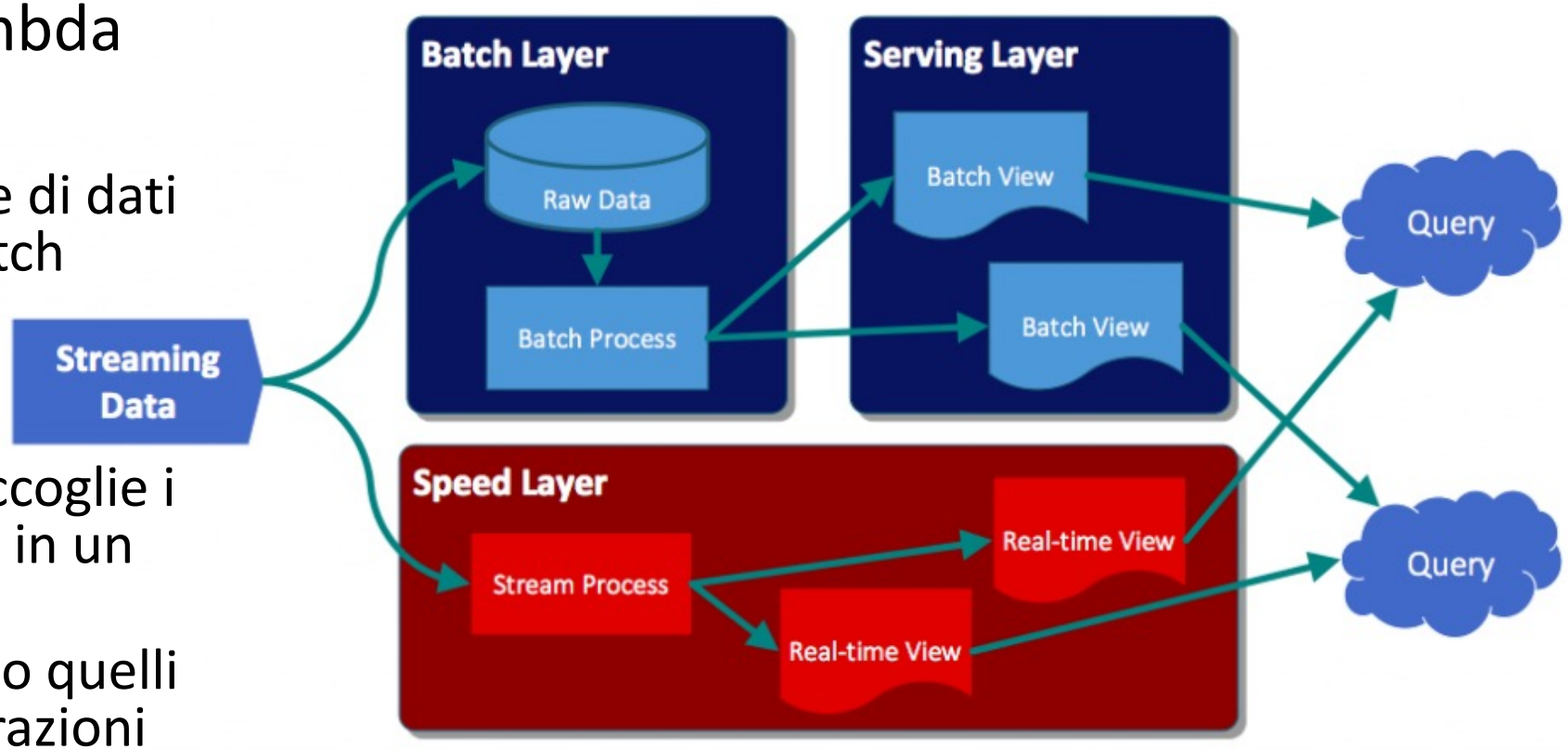
Cosa è «Big Data»

- Il corso di «Big Data» è un insieme degli argomenti visti prima, ma integrati opportunamente per consentirvi di progettare delle *pipeline di analisi dei dati*

- Un Ingegnere Informatico deve conoscere le architetture software per i Big Data e deve saperne scegliere i componenti giusti per il problema in esame

Cosa è «Big Data»

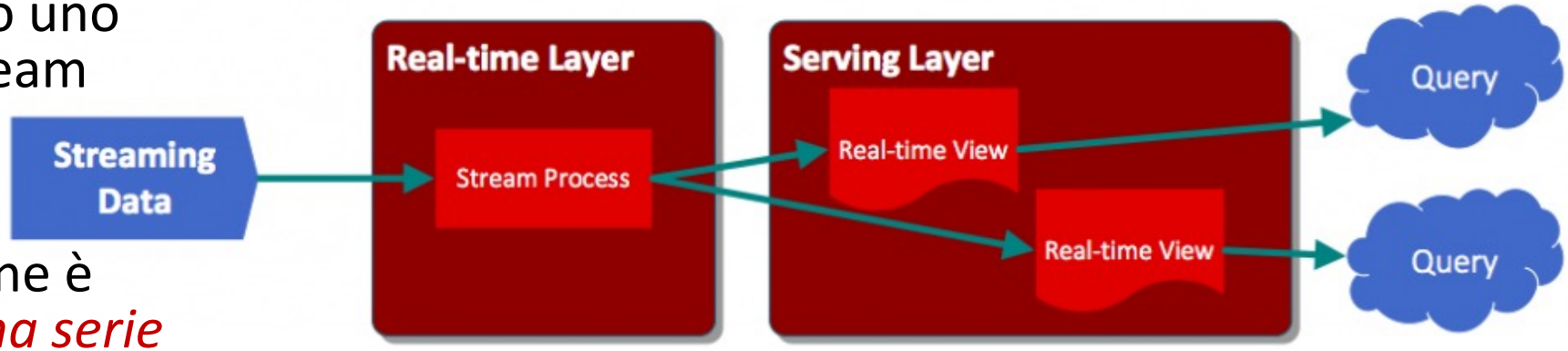
- Architettura Lambda
  - Analisi separate di dati streaming e batch
  - Il Batch layer accoglie i dati eterogenei in un *Data Lake*
  - I dati batch sono quelli legati ad elaborazioni più onerose



Fonte https://medium.com/@Talend/from-lambda-to-kappa-a-guide-on-real-time-big-data-architectures-fe63f3079d3e

Cosa è «Big Data»

- Architettura Kappa
  - Tutti i dati sono uno considerati stream
  - La computazione è intesa come *una serie di trasformazioni sullo stream* fino ad ottenere la view in output



Fonte https://medium.com/@Talend/from-lambda-to-kappa-a-guide-on-real-time-big-data-architectures-fe63f3079d3e

Cosa è «Big Data»

- Una corretta architettura per un problema Big Data richiede che
  - Si conoscano le caratteristiche numeriche e statistiche dei vari tipi di dati
    - Dati vettoriali
    - Grafi
    - Serie temporali
    - Dati categorici
    - …

Cosa è «Big Data»

- Una corretta architettura per un problema Big Data richiede che
  - Si determinino le corrette fasi di acquisizione e pre-processing in ingresso all'architettura
    - Cosa fare se mancano dei dati?
    - Cosa fare se ci sono errori nei dati?
    - Mi servono davvero tutti i dati che ho?
    - …

Cosa è «Big Data»

- Una corretta architettura per un problema Big Data richiede che
  - Si individuino i componenti software più adatti e quindi anche il modello lambda o kappa
    - MongoDB
    - Cassandra
    - Hadoop
    - Spark
    - Kudu
    - …

Cosa è «Big Data»

- Una corretta architettura per un problema Big Data richiede che
  - Si sappiano determinare *i giusti processi di analisi e predizione* sui dati stessi
    - Scelta delle tecniche di ML/DL
    - Tesorflow
    - Pytorch
    - …

Cosa è «Big Data»

- Tutto questo richiederà un po' di *appoggio esterno*
  - Le caratteristiche *statistiche* dei dati
  - Un linguaggio di programmazione che ci supporti in tutto il processo: *Python*
    - E' orientato all'analisi dei dati
    - Ha tutte le librerie necessarie
    - Supporta i principali framework per i Big Data e per il Machine Learning e Deep Learning

Il Syllabus

- Le informazioni complete sugli obiettivi didattici del corso, il programma delle lezioni e i libri di testo si trovano nella *Scheda di Trasparenza*

  - Big Data
  - Intelligent Data Analysis

Il Syllabus

ORE	Lezioni Frontali
1	Introduzione al Corso. Il processo di analisi dei dati: raccolta dei dati, pre-processing, applicazione delle tecniche di analisi ed estrazione della conoscenza.
2	Cenni di statistica, stimatori e campionamento.
3	Preparazione dei dati: tipi di dati, data cleaning, gestione dei dati mancanti, campionamento.
3	Riduzione della dimensionalita: Principal Component Analysis, Singular Value Decomposition, Trasformazioni Wavelet, Multi Dimensional Scaling, Embedding di grafi.
2	Distanze e similarita' per i diversi tipi di dati: dati quantitativi, dati categoriali, dati testuali, sequenze temporali, grafi.
2	Data cubes, Cenni sulla tecnologia OLAP e creazione di un datawarehouse.
2	Mining di pattern ricorrenti: algoritmo Apriori, misure statistiche di correlazione.
3	Introduzione al Machine Learning: capacità di un modello, generalizzazione, Errore di training e di test, tecniche di addestramento

Il Syllabus

ORE	Lezioni Frontali
5	Clustering: k-means e simili, clustering gerarchico, clustering density based e a griglia, clustering basato su grafi, clustering di dati ad elevata dimensionalita, validazione del clustering, analisi degli outlier.
5	Classificatori: feature selection, decision tree e classificatori a regole, Naive Bayes, regressione logistica, Support Vector Machines, Nearest Neighbor, valutazione dei classificatori.
2	Classificatori, concetti avanzati: Multi-class e rare class learning, regressione su dati numerici, semi-supervised learning, metodi di ensemble.
6	Introduzione al Deep Learning: CNN, Autoencoder, LSTM, GAN, Graph Neural Networks.
6	Architetture software per i Big Data: database noSQL, Apache Cassandra, MongoDB, Neo4j.
6	Architetture software per i Big Data: l'algoritmo MapReduce, Apache Hadoop e il suo ecosistema (Pig, Hive, HDFS).
6	Architetture software per i Big Data: Spark e le sue librerie.

Il Syllabus

ORE	Esercitazioni
13	Introduzione alla programmazione Python
3	Librerie di analisi dei dati in Python
3	Algoritmo Apriori e data cubes
3	Riduzione di dimensionalità
3	Stima e campionamento.
3	Uso di MongoDB
3	Uso di Cassandra
3	Apache Hadoop, Hive e Pig.
5	Sviluppo di una pipeline di analisi dei dati su Spark
5	Introduzione ai framework per Deep Learning Pytorch e Tensorflow
ORE	**Altro**
10	Sviluppo di un'intera pipeline di analisi di dati con tecnologie per i Big Data su un caso di studio proposto dal docente.

Il materiale didattico

- Le slide da sole *non sono* materiale didattico: esse sono a compendio dei libri di testo, della spiegazione orale del docente e degli *appunti* presi dallo studente

- *Suggerimento*: stampate le slide prima della lezione e annotatele con i vostri appunti

Il materiale didattico

- Libri di testo (consigliati)
  - Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, prezzo orientativo € 70,00
  - Deep Learning, (2016), di Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, ISBN 978-0262035613, prezzo orientativo €65,00
  - Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, (2017) Sebastian Raschka, Vahid Mirjalili, Packt Publishing, ISBN 978-1787125933, prezzo orientativo € 35,00
  - Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, prezzo orientativo € 45,00.

Il materiale didattico

- Repository GitHub del corso
  - https://github.com/fredffsixty/Big-Data
  - Contiene:
  - I file pdf di tutte le slide (incluse queste)
  - I Notebook Jupyter con i codici delle esercitazioni
  - I dati utilizzati nelle esercitazioni

Gli esami

- Gli esami constano di due parti
  - Presentazione di un progetto di analisi di dati su un caso di studio reale proposto dal docente
  - Colloquio sul programma di teoria svolto

Gli esami

- Il progetto
  - È un caso di studio reale: normalmente si utilizzano competizioni Kaggle
  - Vi verrà fornito un documento che specifica il progetto, dove si trovano i dati e cosa è richiesto
  - Si inizierà insieme nelle ultime ore del corso e poi si continuerà a casa durante lo studio della materia
  - Possono costituirsi dei gruppi di *max 3 persone*

Gli esami

- Il colloquio
  - Tutti gli argomenti che sono stati presentati nel Syllabus e che saranno effettivamente affrontati nel corso sono oggetto del colloquio
  - Si verificheranno le conoscenze teoriche, il rigore nell'esposizione, la capacità di collegare i diversi argomenti del corso
  - Il voto finale tiene conto in egual misura del progetto, che è una base di partenza nella valutazione, e del colloquio: *non ci sono valutazioni ponderate sulla base dei CFU*

Le tesi di laurea

- Vi verranno proposti dei possibili argomenti di tesi di laurea da condurre presso il nostro Laboratorio (CHILab – Laboratorio di Interazione Uomo-Macchina) su temi inerenti il Deep Learning e l'IA

- Altra alternativa sono le tesi aziendali che vertono su temi più legati alle architetture per i Big Data

- Lo studente che voglia sostenere una tesi all'interno del nostro laboratorio assolverà l'esame alla fine del suo percorso di tesi presentando il lavoro svolto