Università degli Studi di Palermo

dipartimento di ingegneria unipa

Introduzione al Modulo

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

a.a. 2022/2023

Prof. Roberto Pirrone

LABORATORIO DI INTERAZIONE UOMO-MACCHINA
CHILAB

Sommario

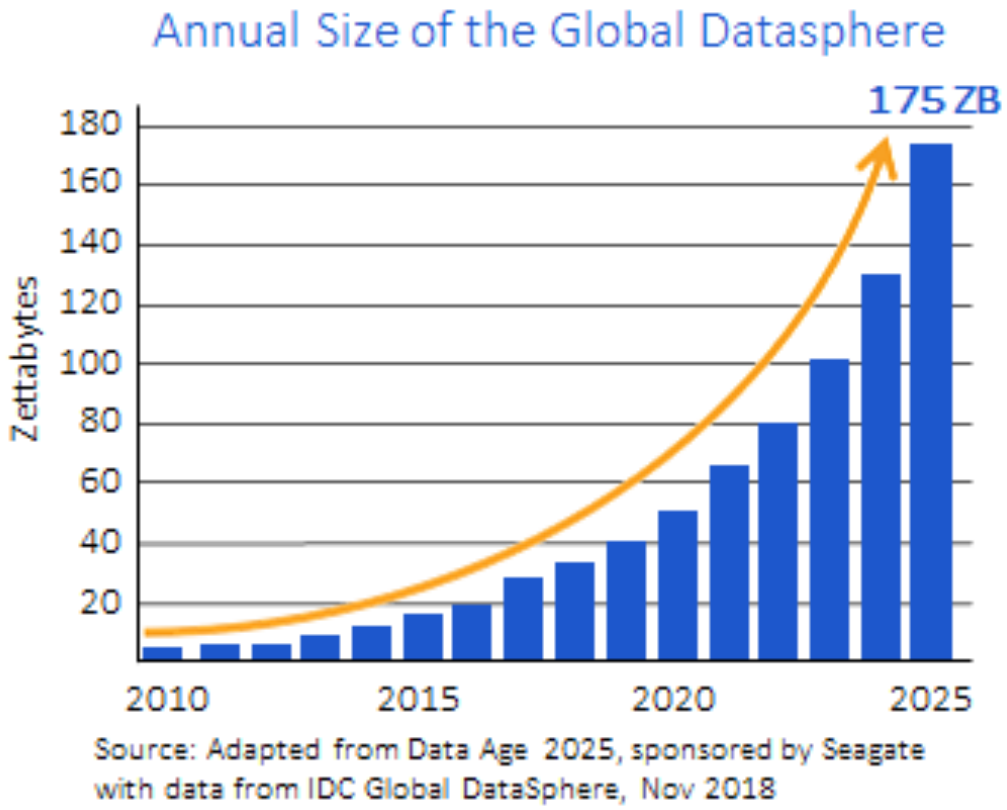
- Il docente
- Perché «(Analisi per i) Big Data»
- Cosa non è «Analisi per i Big Data»
- Cosa è «Analisi per i Big Data»
- Il Syllabus
- Il materiale didattico
- Gli esami
- Le tesi di laurea

Il Docente

- Roberto Pirrone
  - Studio: Edificio 6, terzo piano, stanza 3025
  - Email: roberto.pirrone@unipa.it, roberto.pirrone@community.unipa.it (Google)
  - Telefono studio: 091238.62625, laboratorio: .62643
  - Ricevimento: ogni mercoledì dalle 11:30 alle 13 presso il proprio studio

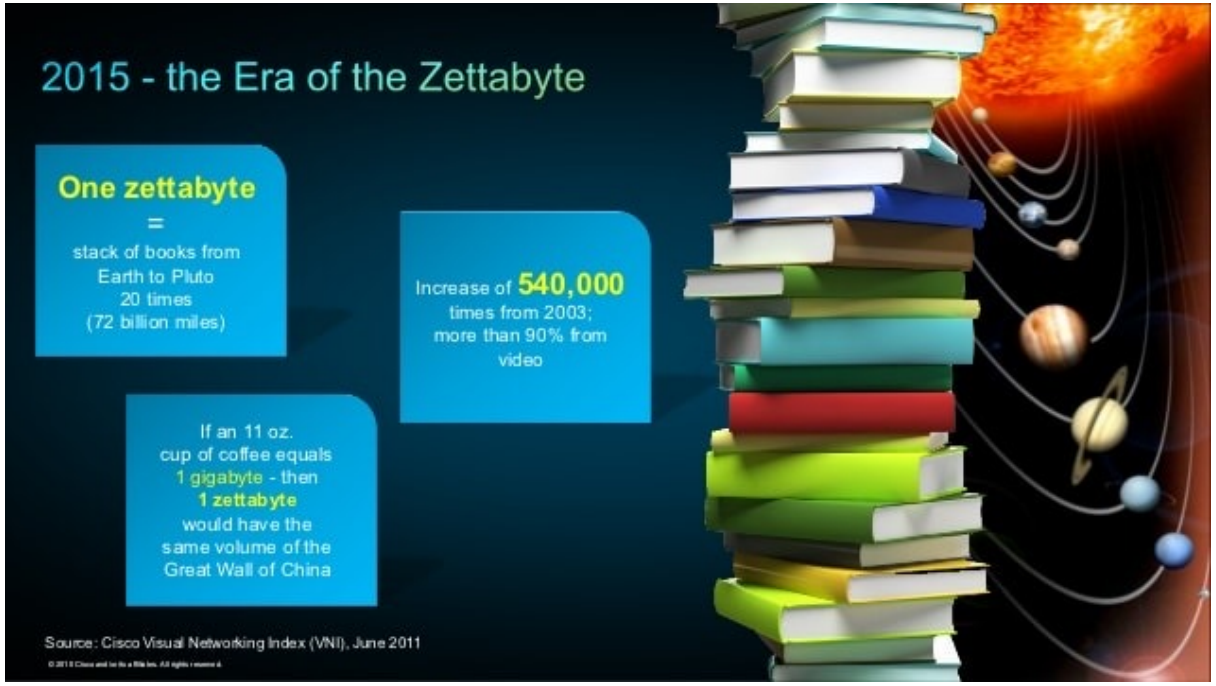
Perché «(Analisi per i) Big Data»

- Perché i dati sono diventati «Big»
  - Ad oggi si stima una produzione annua di dati di circa 80 ZB nel 2022
    - 1 ZB = 10^{21} B
  - ***175 ZB nel 2025***



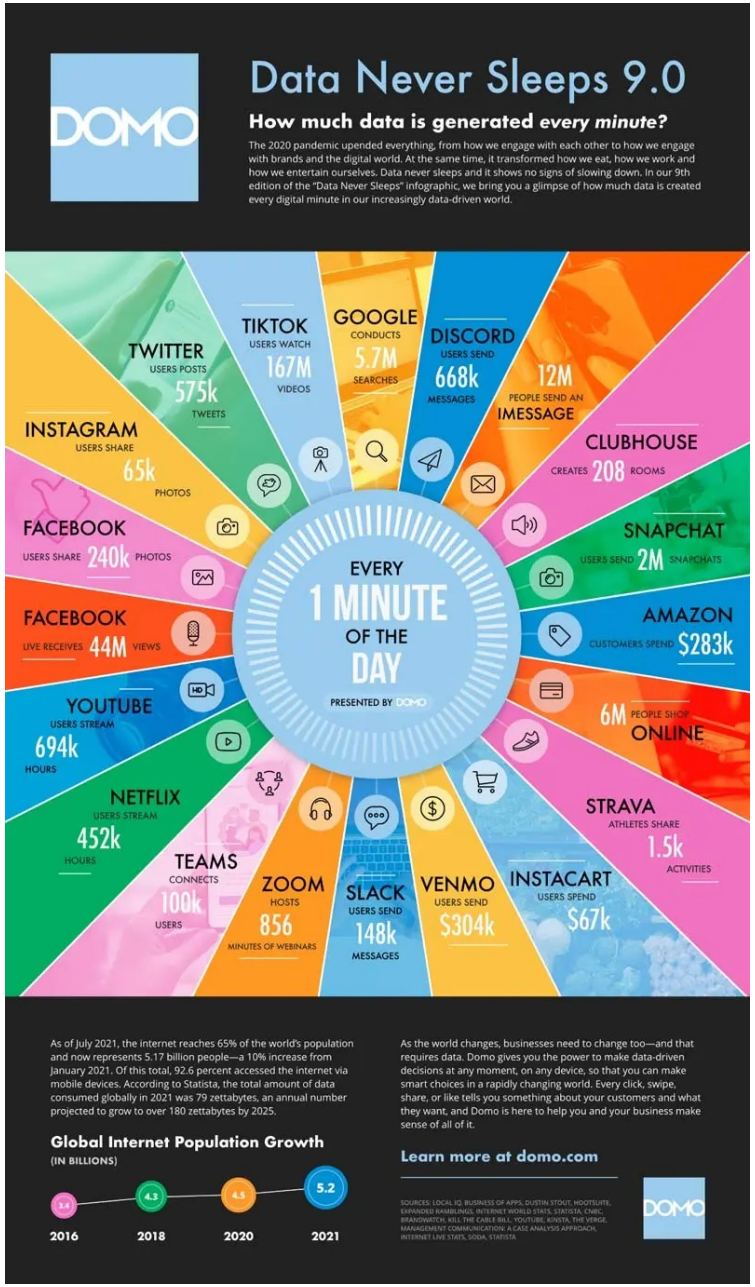
Perché «(Analisi per i) Big Data»

- 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 «(Analisi per i) Big Data»

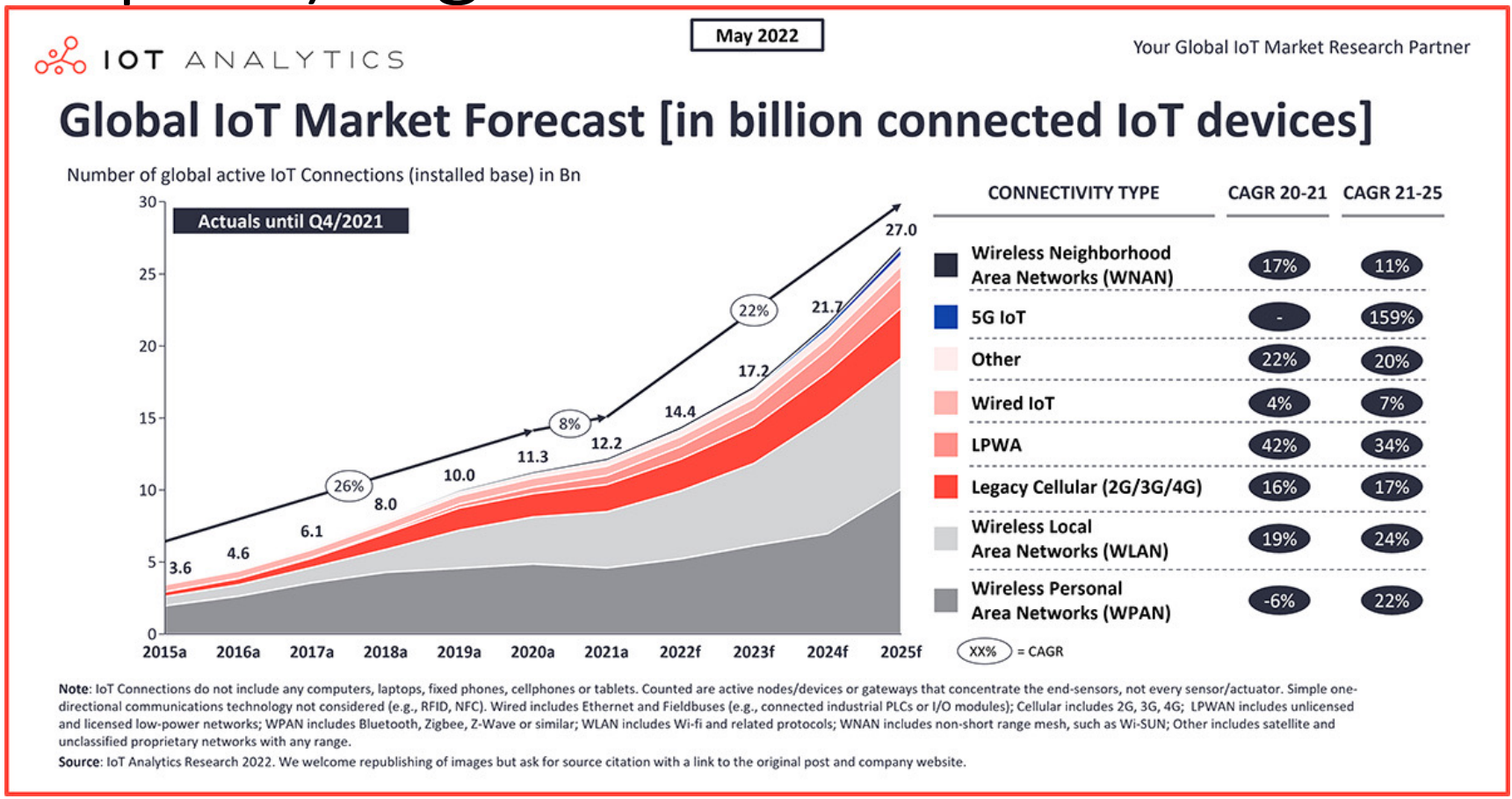
- Perché i dati sono diventati *vari ed eterogenei*
  - Internet e social media



Fonte https://www.domo.com/learn/infographic/data-never-sleeps-9

Perché «(Analisi per i) Big Data»

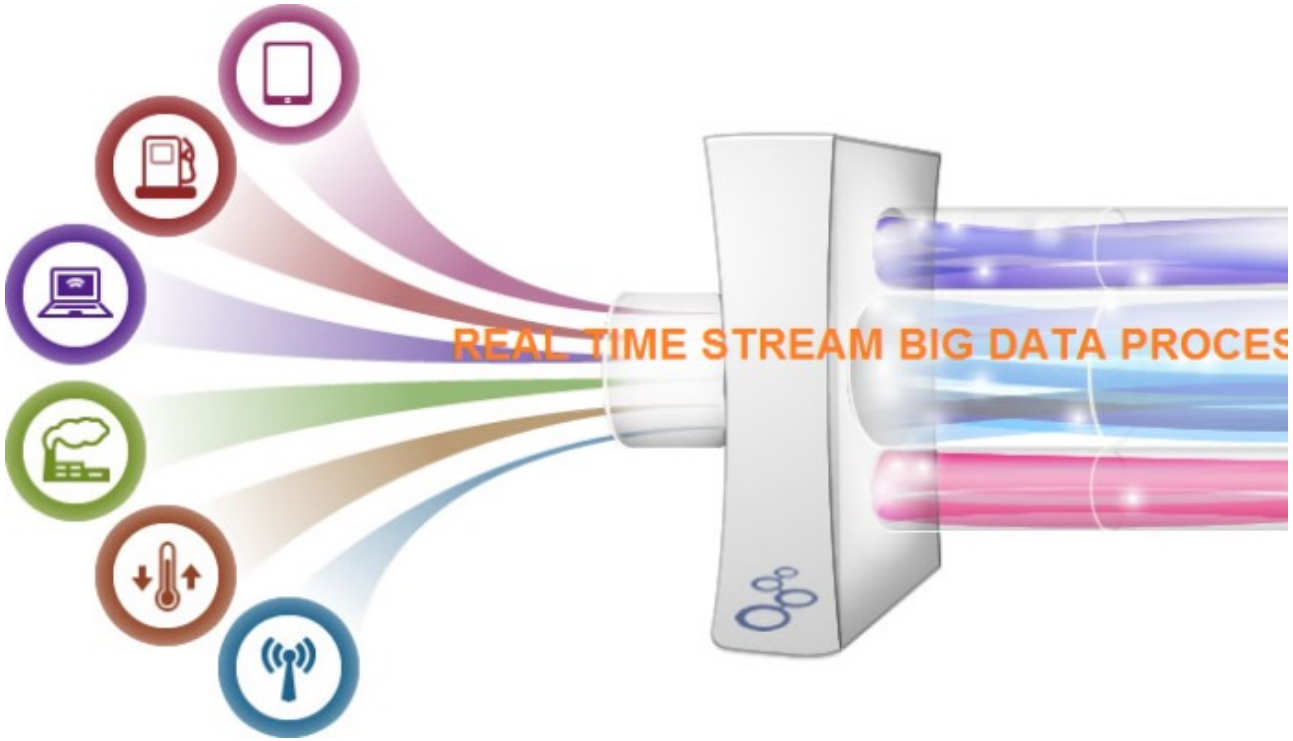
- Perché i dati sono diventati *vari ed eterogenei*
  - I device e i sensori connessi a Internet (IoT – *Internet of Things*)
  - Dati *strutturati*, *semi-strutturati*, *non strutturati*



Fonte https://bit.ly/3RjehFx

Perché «(Analisi per i) Big Data»

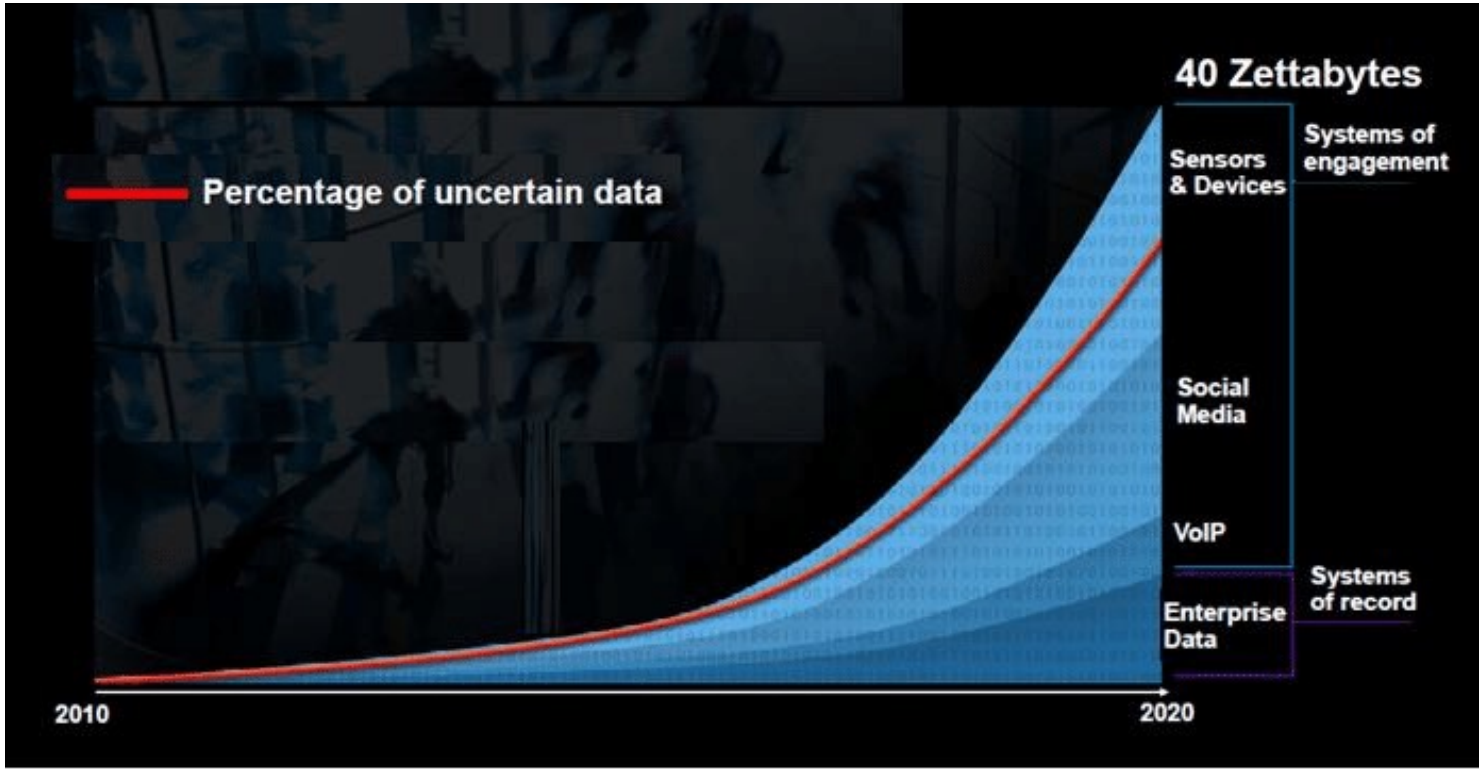
- Perché i flussi di dati sono quasi sempre in *real time*
  - IoT
  - User Generated Contents
  - Monitoraggio ambientale
  - Automotive
  - Monitoraggio della rete
  - Dati di cloud
  - …



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

Perché «(Analisi per i) Big Data»

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



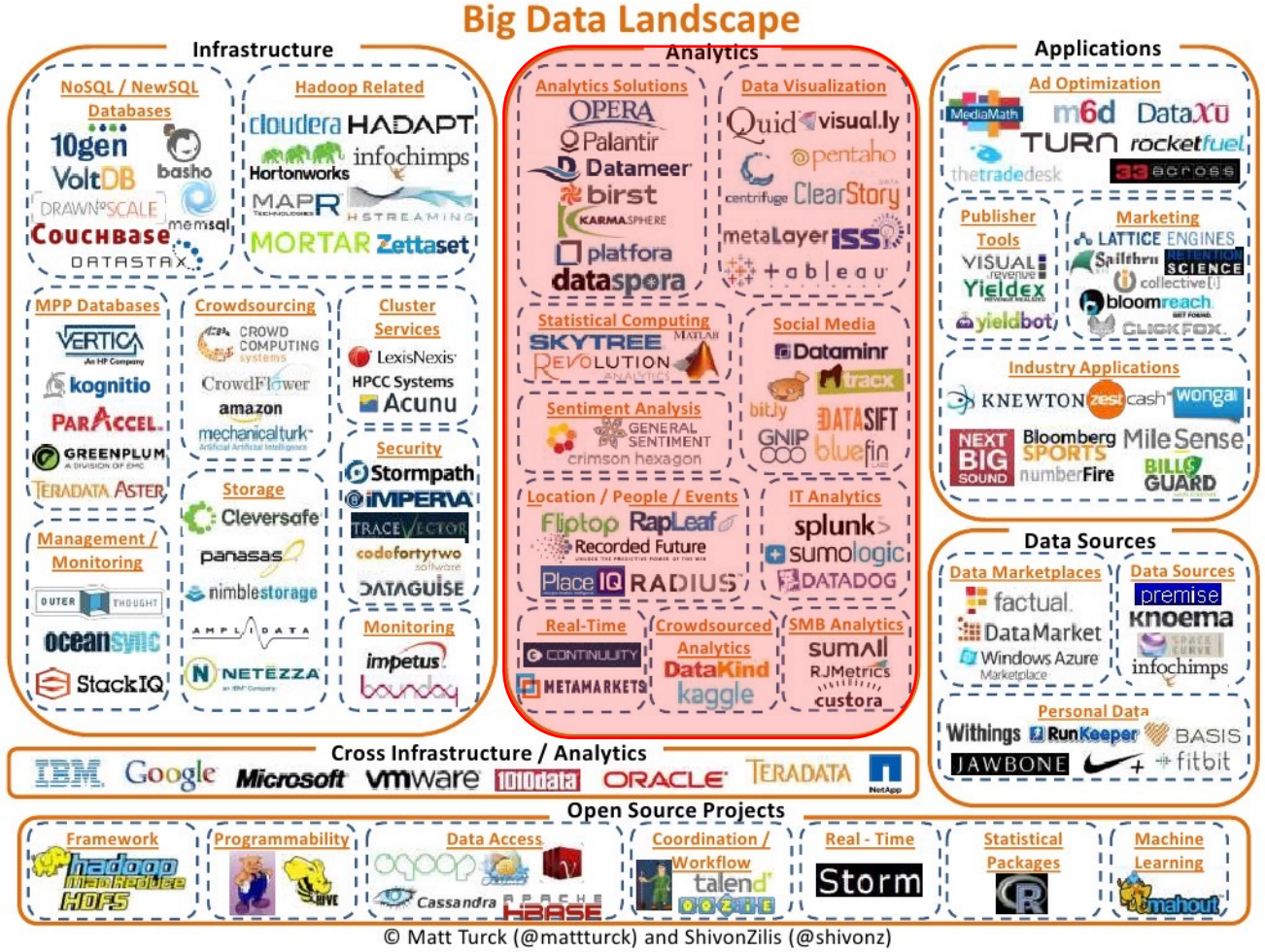
Fonte https://www.researchgate.net/figure/Projected-Growth-of-Big-Data-based-on-1_fig2_272391443

Cosa non è «Analisi per i Big Data»

- Il modulo di «Analisi per i Big Data» *non è*:
  - Un corso di Python (anche se lo useremo tantissimo)
  - Una serie di tutorial su framework più o meno esoterici (anche se ne abbiamo studiati e ne studieremo ancora diversi)
  - Un corso di Machine Learning (anche se ne studieremo un bel po')

Cosa non è «Analisi per i Big Data»

- Non è possibile studiare nel dettaglio tutte le soluzioni software che gravitano anche nel solo mondo dell'analisi dei Big Data!!!



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

Fonte http://bit.ly/40juPDk

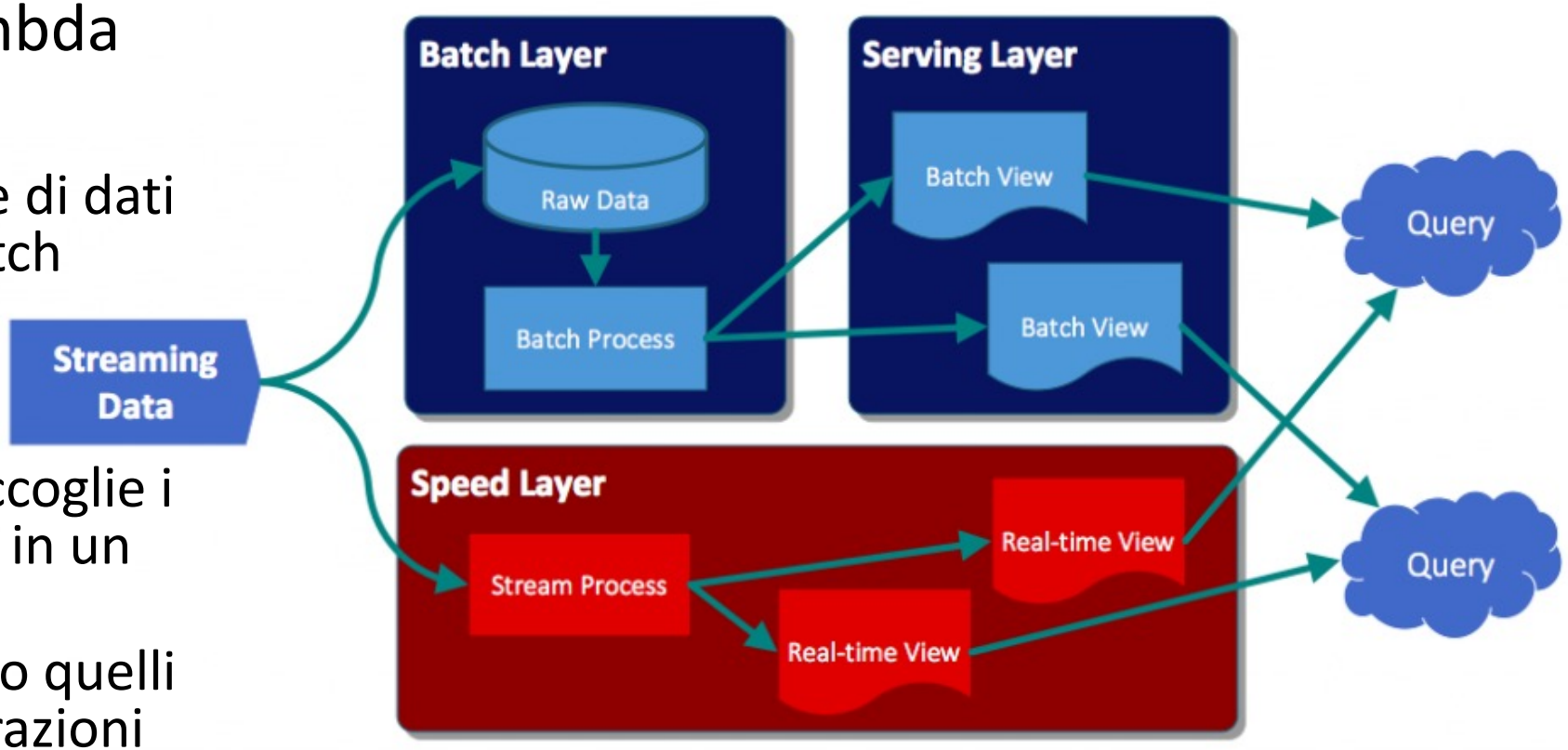
Cosa è «Analisi per i Big Data»

- Il modulo di «Analisi per i 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 (modulo precedente) e deve saperne scegliere i componenti giusti per il problema in esame (e qui ci aiuta questo modulo)

Cosa è «Analisi per i Big Data»

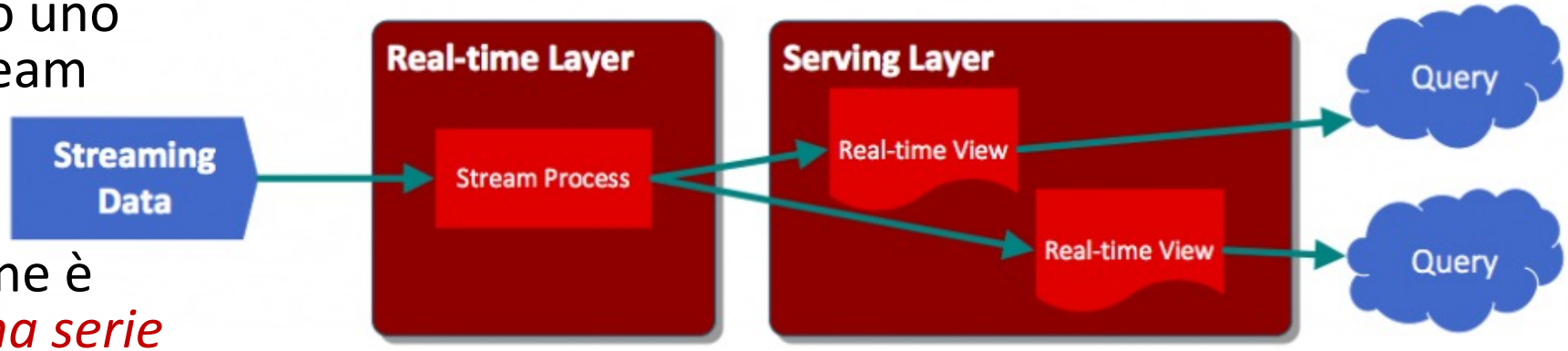
- Architettura Lambda
  - Analisi separate di dati streaming e batch
  - Il Batch layer accoglie i dati eterogenei in un *Data Lake*
  - I dati batch sono quelli legati ad elaborazioni più onerose



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

Cosa è «Analisi per i 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 «Analisi per i Big Data»

- Una corretta architettura per un problema Big Data richiede che
  - Si conoscano le caratteristiche numeriche e statistiche dei vari tipi di dati ✅
  - Si determinino le corrette fasi di acquisizione e preprocessing in ingresso all'architettura ✅
  - Si individuino i componenti software più adatti e quindi anche il modello lambda o kappa ✅

Cosa è «Analisi per i 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
    - Spark ML Pipeline
    - Tesorflow
    - Pytorch
    - …

Cosa è «Analisi per i 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*

  - Analisi per i Big Data

Il Syllabus

- Testi 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)
  - 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 Syllabus

ORE	Lezioni Frontali	Testo rif.
1	Introduzione al Corso.	Slide docente
2	Cenni di Teoria della Probabilità e Teoria dell'Informazione; stimatori statistici e tecniche di campionamento.	Estratti dal Bengio capp. 3 e 17.1-2
2	Introduzione al Machine Learning: apprendimento supervisionato, non supervisionato, apprendimento con rinforzo, capacità del modello, parametri e iperparametri, tipologie di errore, tecniche di addestramento.	Estratti dal Bengio cap. 5
5	Clustering: k-means e sue varianti, clustering gerarchico, clustering density based e a griglia, clustering basato su grafi, clustering di dati ad elevata dimensionalità, validazione del clustering, analisi degli outlier.	Aggarwal cap. 6
5	Classificatori: feature selection, decision tree e classificatori a regole, Naive Bayes, regressione logistica,Support Vector Machines, Nearest Neighbor, valutazione dei classificatori.	Aggarwal cap. 10

Il Syllabus

ORE	Lezioni Frontali	Testo rif.
2	Classificatori, concetti avanzati: Multi-class e rare class learning, regressione su dati numerici, semi-supervised learning, metodi di ensemble.	Aggarwal cap. 11
8	Deep Learning: struttura di una rete neurale, tipologia di unità nascoste e di uscita, funzioni di loss, concetto di grafo di computazione, stochastic gradient descent, ottimizzazione e regolarizzazione, CNN, Autoencoder, LSTM, GAN, Graph Neural Networks, fine tuning e transfer learning.	Estratti dal Bengio capp. 6, 7 e 8 Slide docente
3	Elaborazione di immagini mediche: classificazione di volumi TAC/RM con CNN 3D.	Slide docente
3	Elaborazione del linguaggio naturale: classificazione di testi con Word2Vec.	Slide docente
5	Analisi di dati web: algoritmo PageRank, recommender systems, web usage analysis, social network analysis.	Estratti dal Aggarwal capp. 18 e 19

Il Syllabus

ORE	Esercitazioni
3	Uso dei database NoSQL: il caso di MongoDB.
3	Stima statistica di una distribuzione Gaussiana al variare del campionamento.
3	Uso di sci-kit learn per classificazione e clustering.
3	Creazione di una pipeline con Spark ML per il clustering.
3	Creazione di una pipeline con Spark ML per la classificazione.
3	Uso di Tensorflow ed esempi di implementazioni di DNN.

- Il riferimento per le esercitazioni saranno dei Notebook ed eventuali slide predisposte dal docente

- Il libro di riferimento per Spark è lo Zaharia

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/BigDataAnalytics
  - Contiene:
  - I file pdf di tutte le slide (incluse queste)
  - I codici delle esercitazioni
  - I dati utilizzati nelle esercitazioni

Gli esami

- *Gli esami sono analoghi a quelli del primo modulo*

- Compito scritto che consta di:
  - Domande aperte sugli argomenti teorici affrontati durante il modulo
  - Quesiti con semplici esercizi da risolvere su carta legati alle tematiche di analisi dei dati
  - Prova di programmazione su carta per lo sviluppo di una semplice pipeline Spark di analisi dei dati

Gli esami

- *Gli esami sono analoghi a quelli del primo modulo*
- Il compito dura due ore
- Il voto del compito costituisce proposta di voto finale

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 possono essere le tesi aziendali che abbiano attinenza con la Big Data Analytics e *siano di interesse per il nostro laboratorio*

Le tesi di laurea

- Vincoli sull'assegnazione della tesi
  - Che ci sia uno slot libero (max 4 tesisti in contemporanea, altrimenti deve prima laurearsi qualcuno per poter avere la tesi)
  - Che siano garantiti almeno sei mesi effettivi di lavoro
    - Consecutivamente al netto delle altre materie e del tirocinio
    - Non ha senso chiedere la tesi un anno prima, quando ancora si devono sostenere altri esami e si «sparisce» per mesi
    - Manifestate comunque il vostro interesse!!
    - L'anno prossimo c'è anche «Natural Language Processing»

Le tesi di laurea

- Vincoli sull'assegnazione della tesi
  - Che si concordi su una delle tematiche proposte e su un livello minimo di obiettivi concordato al momento dell'assegnazione
  - Che l'argomento di tesi industriale risulti di interesse per il nostro gruppo di ricerca
- Si verrà affidati ad uno o più dottorandi del laboratorio con i quali si dovranno avere incontri (anche on line) *al più bi-settimanali* sullo stato di avanzamento lavori