Università degli Studi di Palermo

dipartimento di ingegneria unipa

Introduzione al Modulo

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

a.a. 2023/2024

Prof. Roberto Pirrone

CHILAB COMPUTER HUMAN INTERACTION LABORATORY Deep Learning and NLP in Healthcare

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

Il Laboratorio

- Laboratorio di Interazione Uomo-Macchina
  - Edificio 6, terzo piano, a sx dalle scale
  - Email: chilab@unipa.it
  - Telefono: 091238.62643

E adesso ….

Cosa vi aspettate da questo corso?

Cosa pensate che sia «Analisi per i Big Data»?

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»

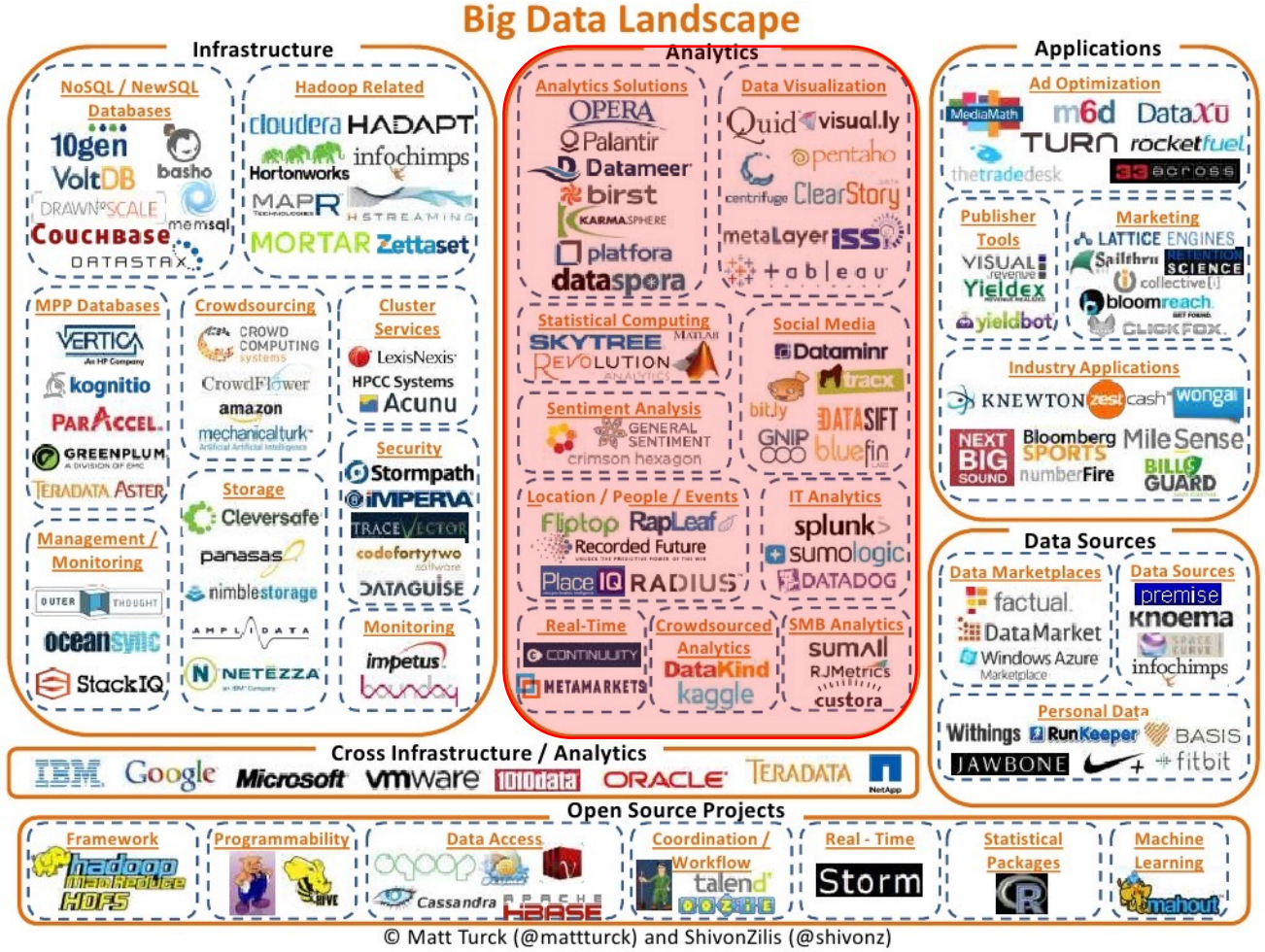
- *Non è possibile* studiare analiticamente tutte le librerie Python che useremo durante il corso!!!

```python
# Define Sequential model with 3 layers
model = keras.Sequential(
    [
        layers.Dense(2, activation="relu", name="layer1"),
        layers.Dense(3, activation="relu", name="layer2"),
        layers.Dense(4, name="layer3"),
    ]
)
# Call model on a test input
x = ops.ones((3, 3))
y = model(x)
```

```python
keras.layers.Dense(
    units,
    activation=None,
    use_bias=True,
    kernel_initializer="glorot_uniform",
    bias_initializer="zeros",
    kernel_regularizer=None,
    bias_regularizer=None,
    activity_regularizer=None,
    kernel_constraint=None,
    bias_constraint=None,
    lora_rank=None,
    **kwargs
)
```

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*

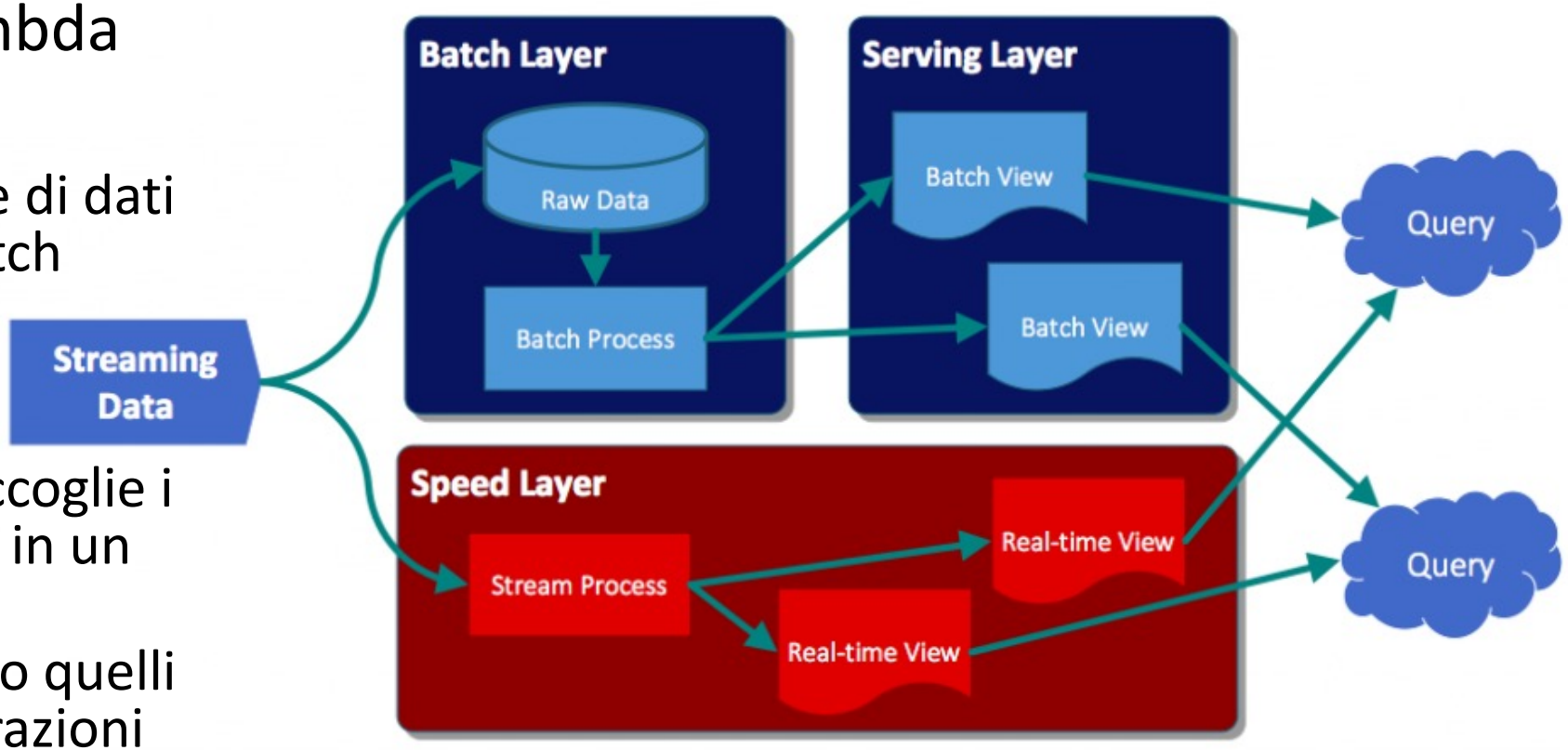
- Un Ingegnere Informatico deve conoscere:

  - Le architetture software per i Big Data (modulo precedente)

  - I componenti giusti per il problema in esame (e qui ci aiuta questo modulo)

Cosa è «Analisi per i Big Data»

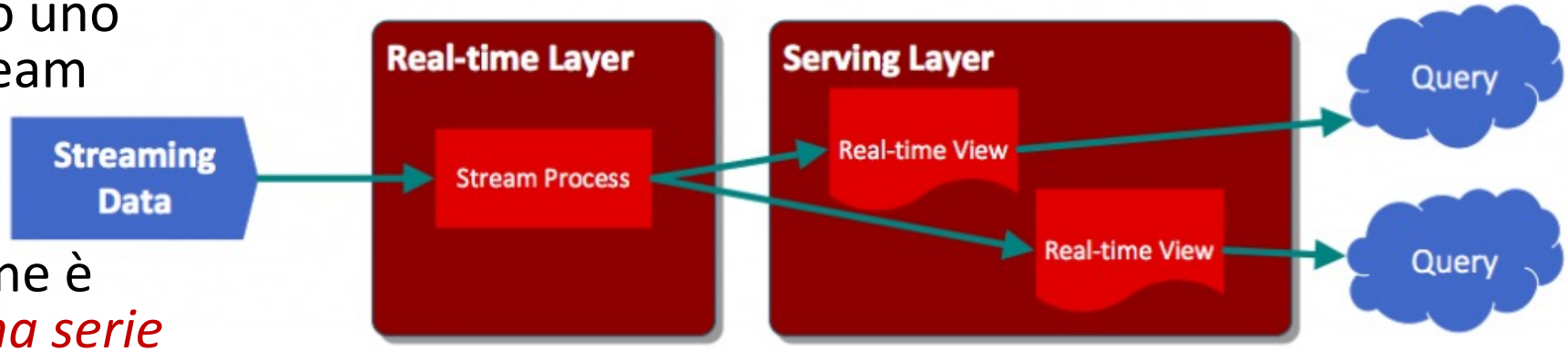
- Architettura Lambda
  - Analisi separate di dati streaming e batch
  - Il Batch layer accoglie i dati eterogenei in un *Data Lake*
  - I dati batch sono quelli legati ad elaborazioni più onerose



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*

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
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	Richiami sull'uso di Spark
3	Esercitazione su stima e campionamento
3	Uso di Sci-kit Learn per il clustering.
3	Creazione di una pipeline con Spark ML per il clustering.
3	Uso di Sci-kit Learn per la classificazione
3	Creazione di una pipeline con Spark ML per la classificazione.
3	Uso di Tensorflow ed esempi di implementazioni di DNN.
3	Svolgimento di una prova di esame

- Le esercitazioni saranno svolte in aula, usando Google Colab per comodità
- I Notebook risultanti verranno condivisi dal docente nel repository del corso
- 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

- Repository GitHub del corso
  - 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 al calcolatore con allegato un data set su cui vi verrà richiesto di effettuare pre-processing e analisi di tipo ML/DL
  - 1 quesito di difficoltà bassa (4/30)
  - 2 quesiti di difficoltà media (8/30 ciascuno)
  - 1 quesito di difficoltà alta (10/30)
- *Per ogni quesito il voto viene attribuito «gradualmente» e non in maniera binaria*

Gli esami

- *Gli esami sono analoghi a quelli del primo modulo*
- Il compito dura due ore
- Il voto del compito, mediato con quello del primo modulo, costituisce proposta di voto finale
- Nei casi dubbi, a insindacabile giudizio della commissione, si svolge la prova orale

Le tesi di laurea

- Lepossibili tesi di laurea presso il nostro Laboratorio riguardano le applicazioni dell'IA alla salute:
  - Analisi di strutture molecolari per il supporto intelligente al Drug Discovery/Repurposing
  - Analisi di strutture polimeriche per il supporto allo sviluppo di nuovi (bio-)materiali
  - Integrazione degli algoritmi di AI nei sistemi radiologici ospedalieri (PACS)
  - Sistemi di interazione Medico-sistema diagnostico basati sulle Intelligenze Artificiali Generative (Chat-GPT e simili)
  - Segmentazione di scansioni TAC, RM, PET usando le reti neurali

Le tesi di laurea

- Altra alternativa possono essere le tesi aziendali che abbiano attinenza con la Big Data Analytics e *siano di interesse per il nostro laboratorio*
  - Vengono pubblicate sul sito del Corso di Laurea
  - Cercate quelle in cui io sono relatore

Le tesi di laurea

- Vincoli sull'assegnazione della tesi
  - Che ci sia uno slot libero (*max 5 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
  - Si verrà affidati ad uno o più ricercatori e/o dottorandi del laboratorio
  - Ci si incontrerà con i propri referenti (anche on line) *al più settimanalmente* sullo stato di avanzamento lavori