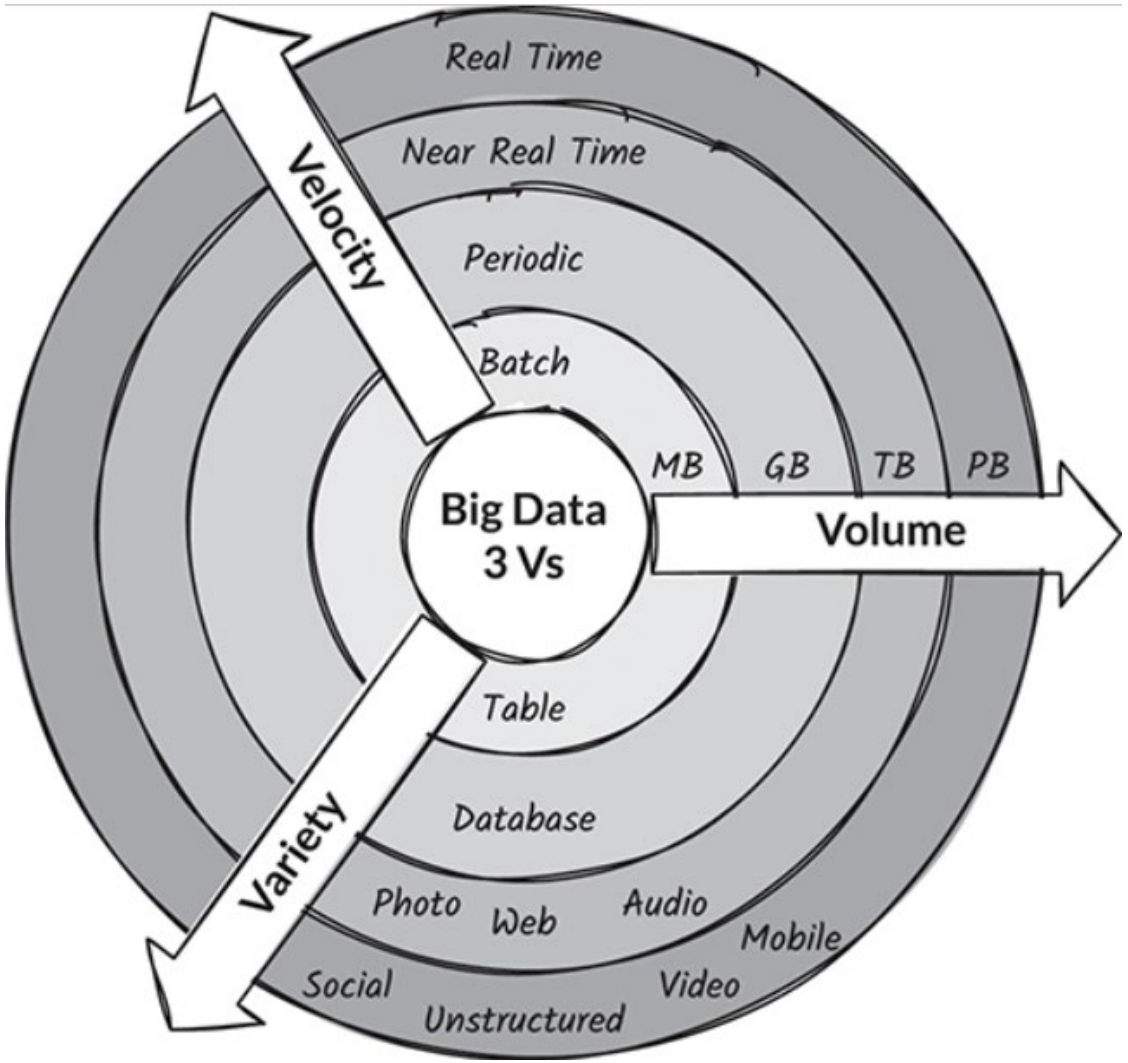
Big Data
a.a. 24/25

Prof. Alfredo Pulvirenti

Obiettivi del corso

- Comprendere i concetti fondamentali di Big Data
- Esplorare le tecnologie e gli strumenti principali
- Acquisire competenze pratiche per l’analisi e la gestione di grandi volumi di dati



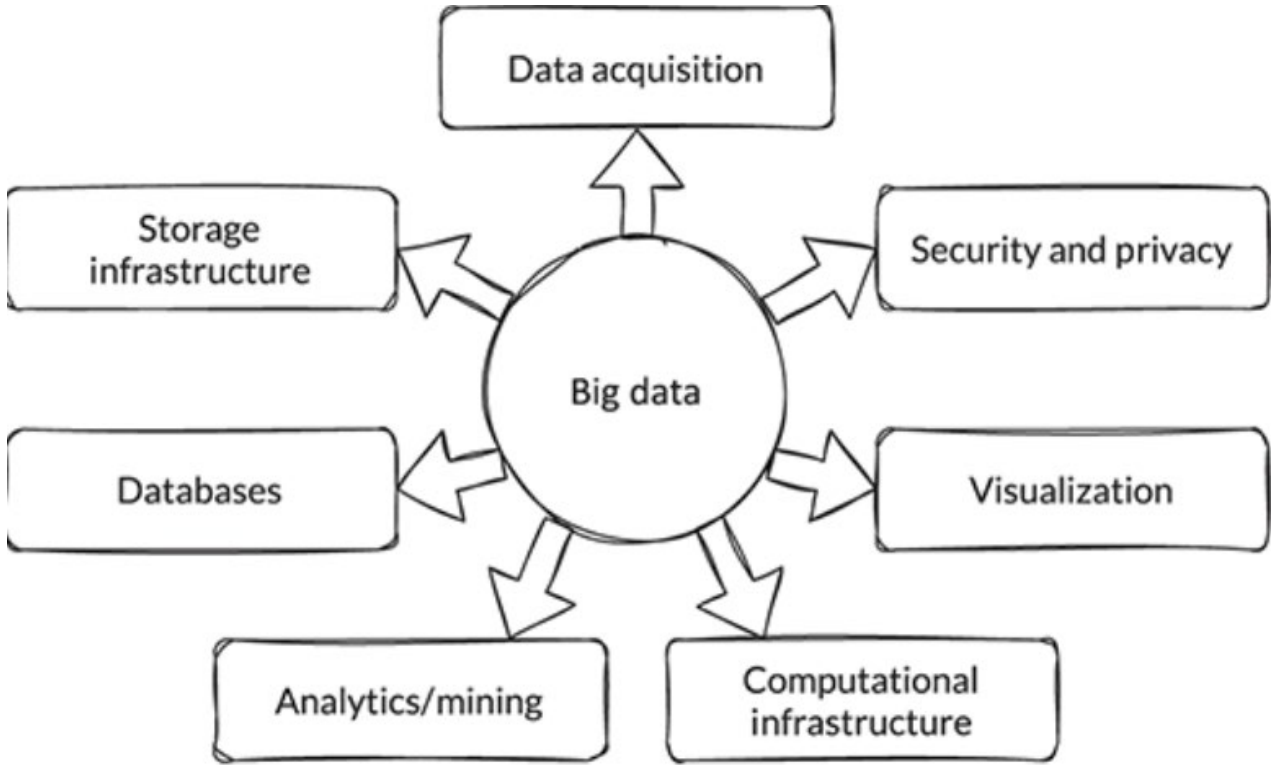
L’era dei dati

- Prodotti velocemente
- Eterogenei
- Potenzialmente ricchi di informazioni
- Volume, Velocità, Varietà, Veridicità, Valore

Big data: Qualsiasi processo di dati in cui la dimensione dei dati stessi è un problema: conservare, trasmettere, elaborare su scala.

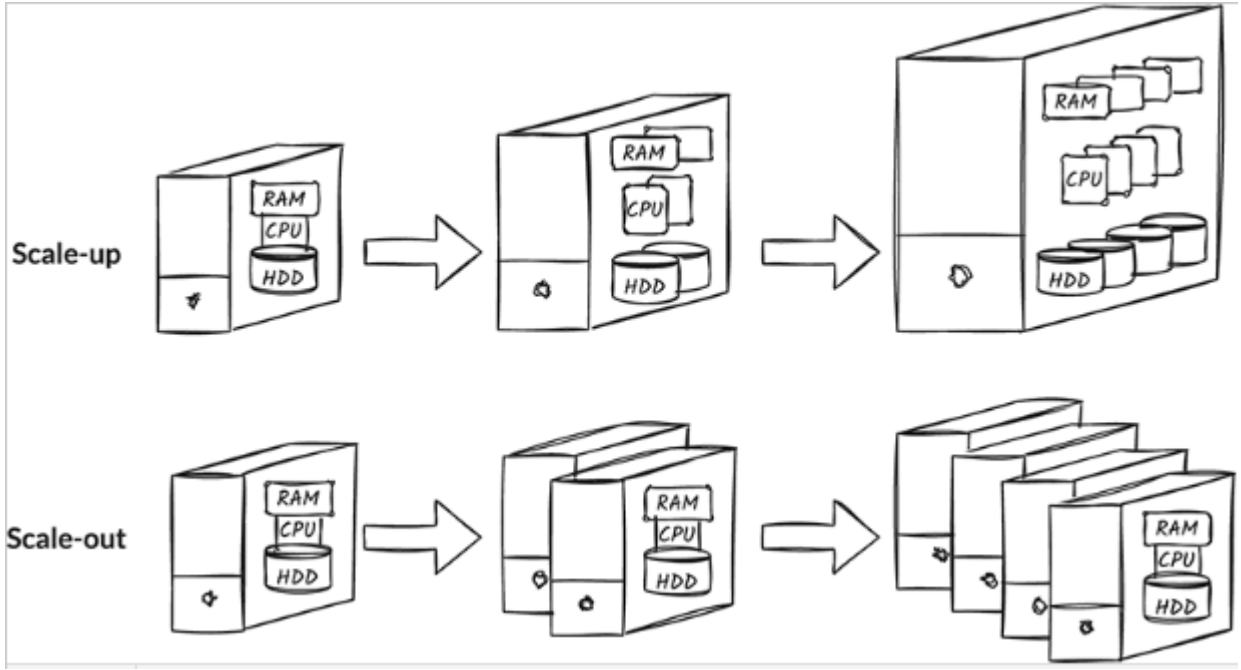
Diversi aspetti dei Big Data

- Concetti spesso confusi perché coinvolgono diversi aspetti



Come interagire con grandi quantità di dati

- Supponiamo di dover analizzare 100 terabyte di dati.
  - Non siamo in grado di memorizzare tutto in una sola macchina
  - Non abbiamo la possibilità di processare questi dati in un tempo ragionevole

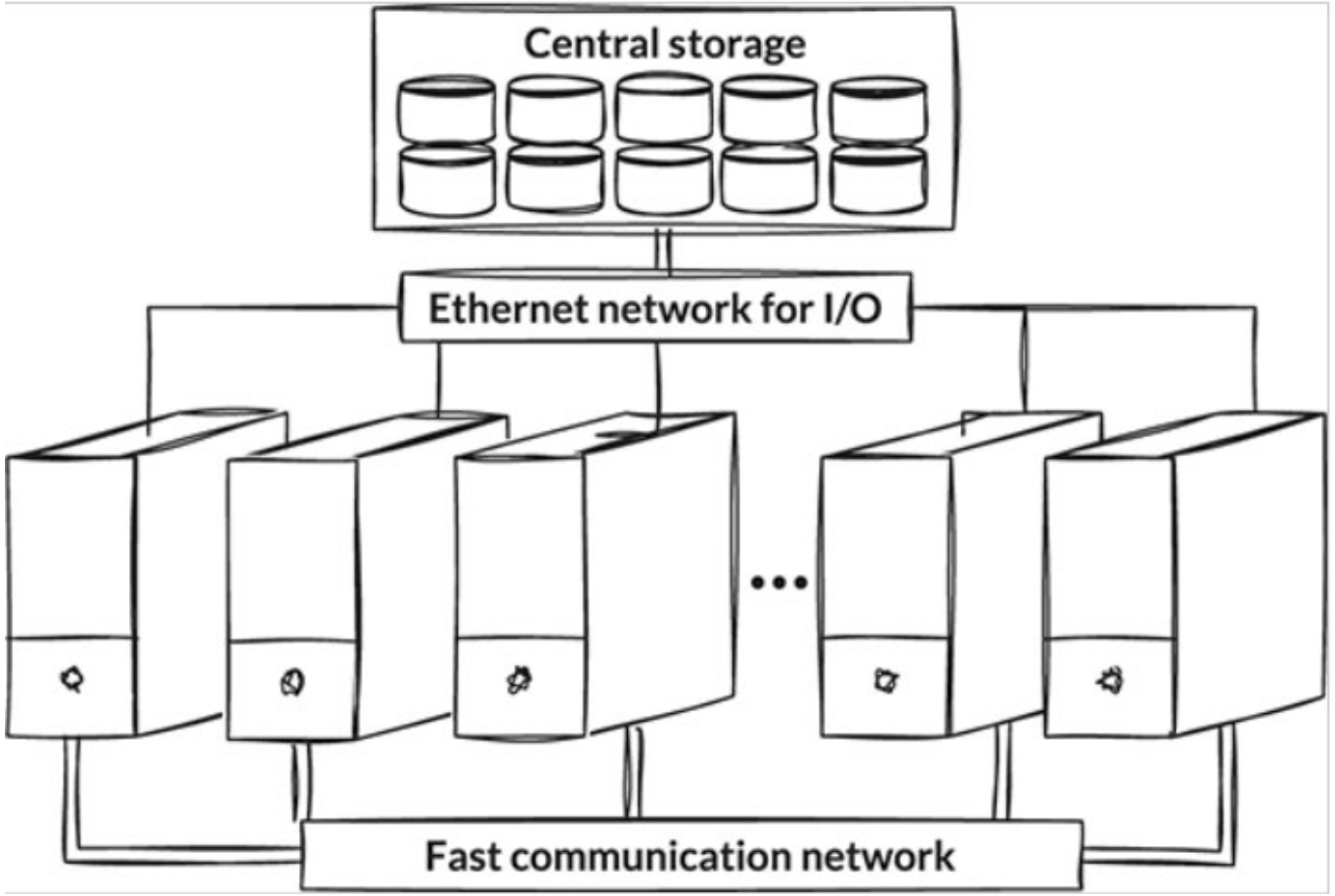


Scale-up	Scale-out
Key idea	*Key idea*
Add more memory, processors	Add more (cheaper) nodes
Advantages	*Advantages*
✓ Less energy consumption	✓ Cheaper
✓ Less expense on cooling systems	✓ Fault tolerance possible
✓ Easier to implement solutions	✓ Easy to grow
Disadvantages	*Disadvantages*
× Price	× More physical space
× No fault tolerance	× Energy costs (electricity and cooling)
× Limited hardware upgrades	× Network equipment required

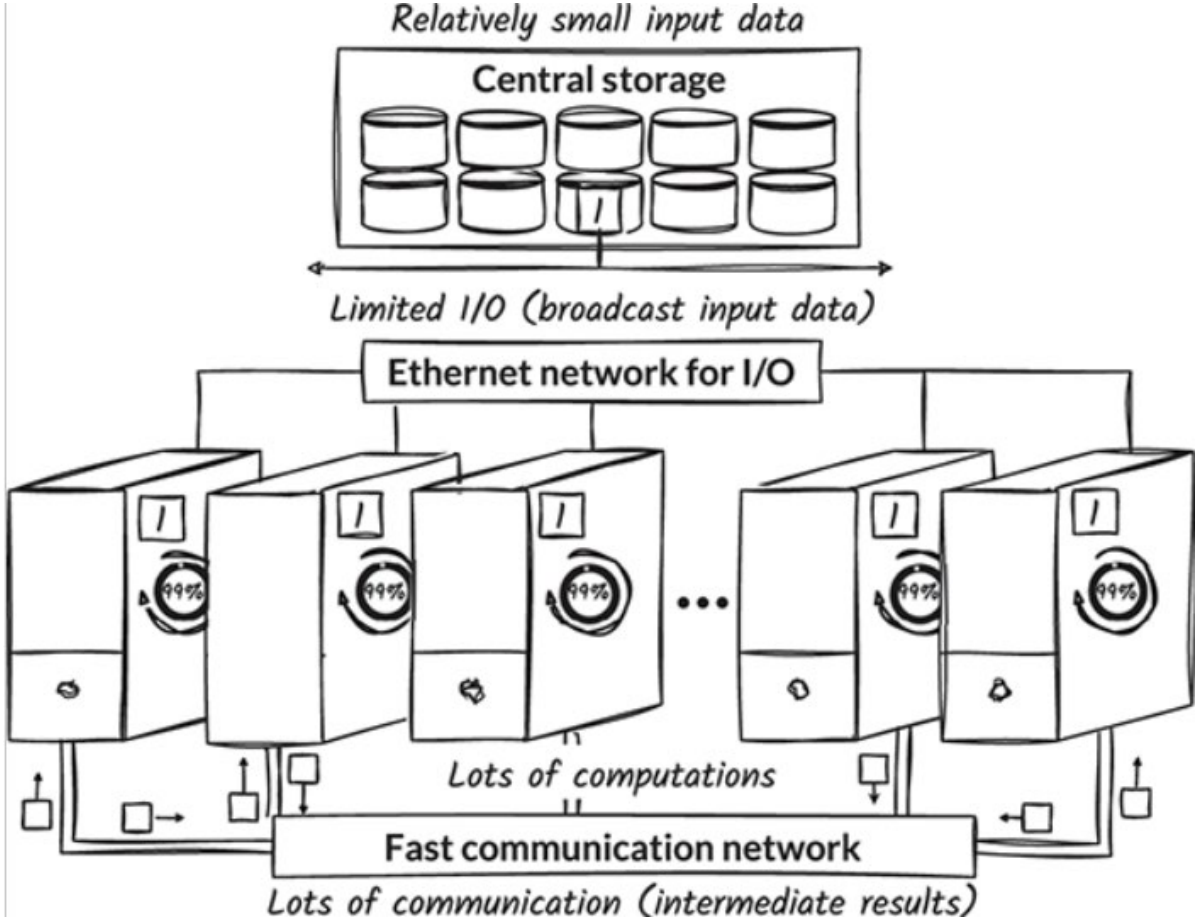
High Performance Computing Vs Big Data Computing

- HPC possiamo immaginare ad un cluster di macchine indipendenti (nodi) connessi tramite una rete di comunicazione ad alta velocità che hanno accesso ad uno storage centralizzato.
- Ogni nodo ha una CPU multicore e la sua RAM e HD.
- Ogni applicazione che viene eseguita su un nodo accede alla sua RAM (non quella degli altri nodi), parliamo di architettura a memoria distribuita.

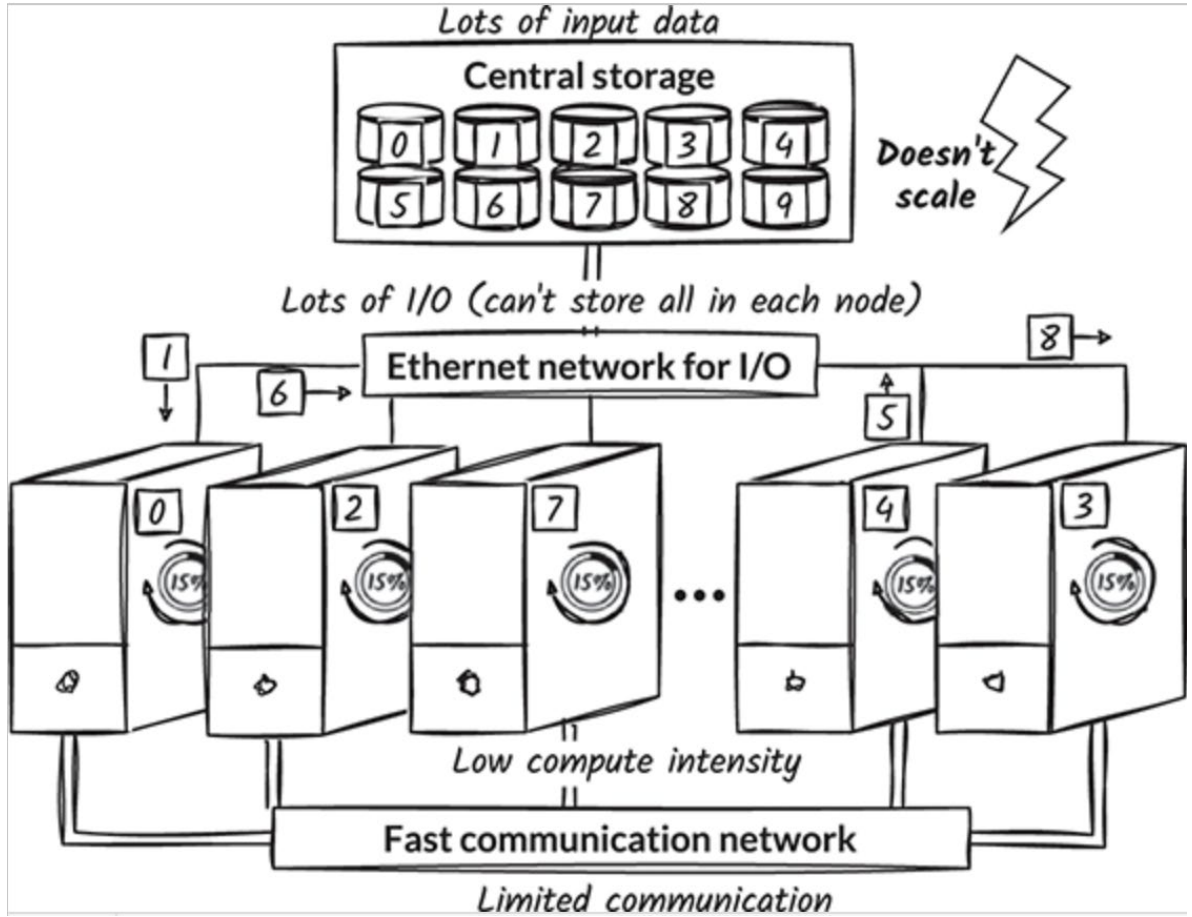
Architettura HPC semplificata



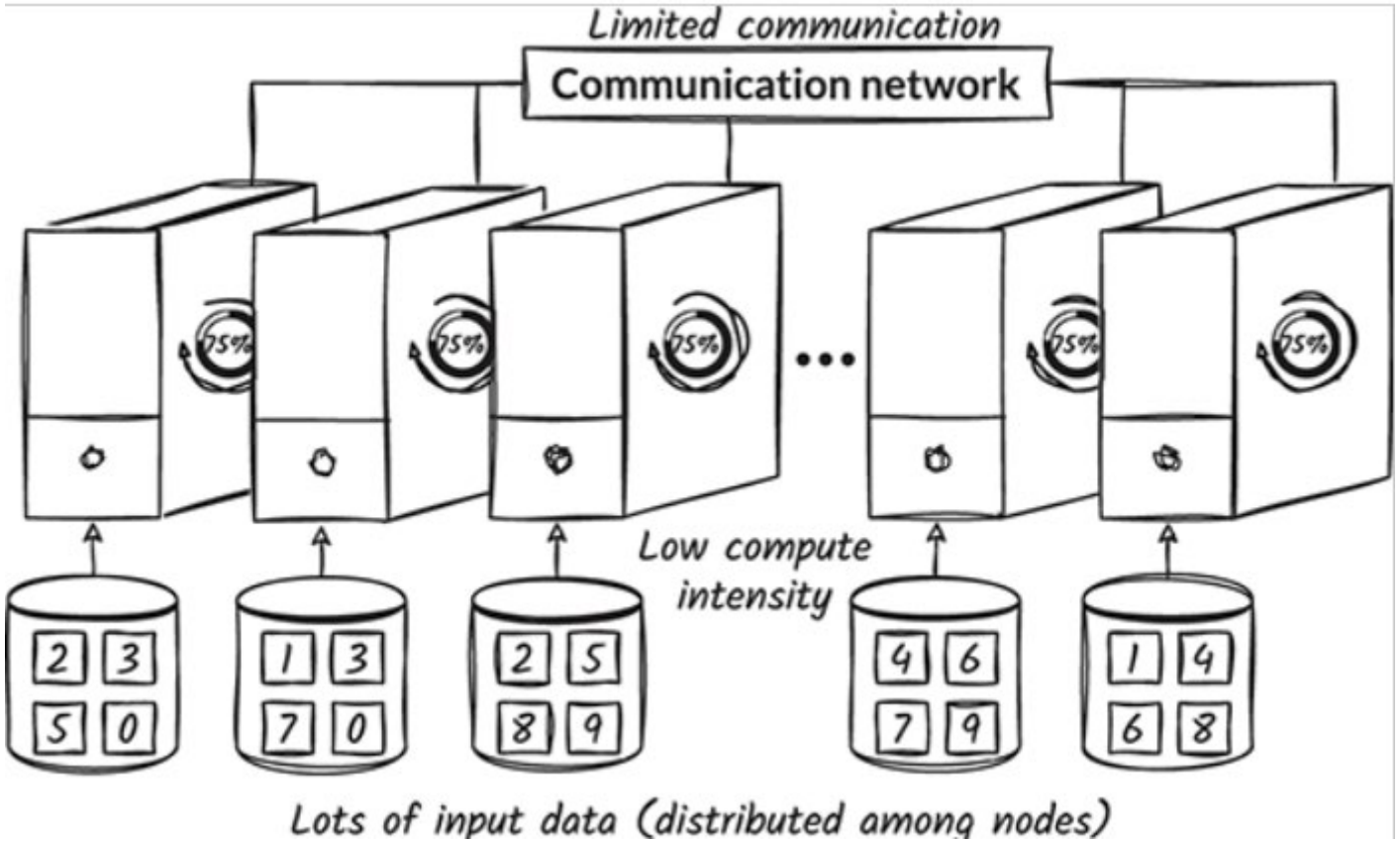
HPC


Relatively small input data
Central storage
Limited I/O (broadcast input data)
Ethernet network for I/O
Lots of computations
Fast communication network
Lots of communication (intermediate results)

- Se abbiamo diversi computer connessi dobbiamo implementare programmi distribuiti e paralleli.
- I nodi:
  - Nodo master si occupa dell’orchestrazione e comunicazione
  - Nodi worker si occupano del calcolo
- Un modo classico per interagire con distributed computing è MPI (Message Passing Interface).


Lots of input data
Central storage
0
1
2
3
4
5
6
7
8
9
Doesn't scale
Lots of I/O (can't store all in each node)
Ethernet network for I/O
1
6
8
5
0
2
7
4
3
15%
Low compute intensity
Fast communication network
Limited communication

Architettura dei Sistemi Big Data



Big data	HPC
Focus on data-intensive jobs	Focus on computation-intensive jobs
Hardware failure common	Surprised by hardware failure
Code: data science, graphs	Code: simulation, optimization
Usually mix CPU/GPU and data	Mix CPU/GPU
Job moved to where the data is located	Data moved to where it will be processed
SIMD model:[a] data parallelism	SIMD/MIMD[b] model (more general parallelism)
Commodity hardware acceptable	Needs specialized hardware

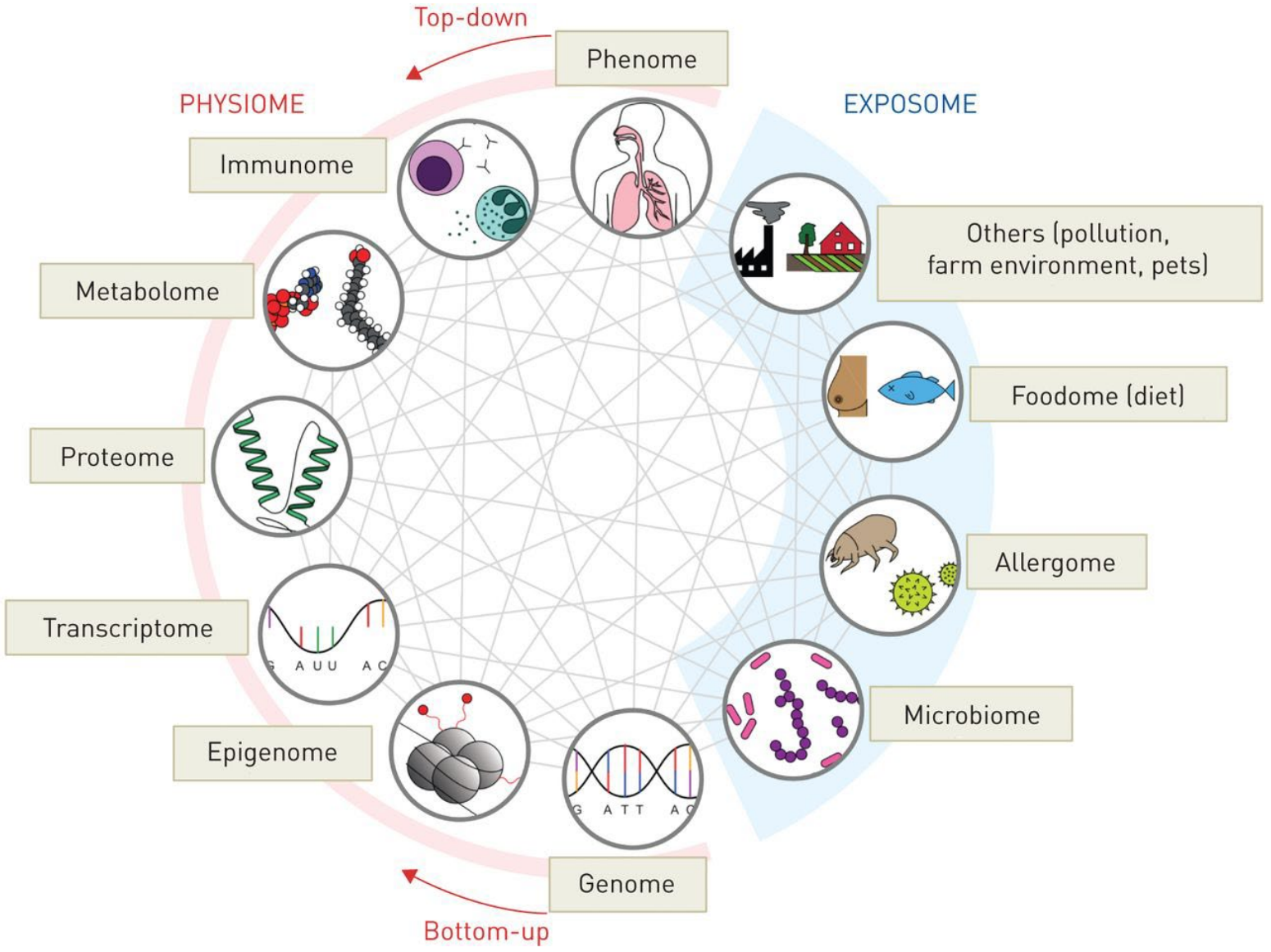
Applicazioni dei Big Data

- Settori che utilizzano Big Data:
  - Sanità,
  - Finanza,
  - Marketing,
  - Industria, etc.

Alcuni esempi

- **Sanità**: Analisi predittiva per diagnosi precoci, gestione delle pandemie, personalizzazione delle cure
- **Finanza**: Rilevamento delle frodi, algoritmi di trading ad alta frequenza, analisi del rischio creditizio
- **Marketing**: Pubblicità mirata basata su dati comportamentali, personalizzazione dell'esperienza utente, analisi delle tendenze di mercato
- **Industria**: Manutenzione predittiva, ottimizzazione delle catene di approvvigionamento, automazione basata su AI

Sanità


Top-down
Phenome
PHYSIOME
EXPOSOME
Immunome
Others (pollution, farm environment, pets)
Metabolome
Foodome (diet)
Proteome
Allergome
Transcriptome
Microbiome
Epigenome
Genome
Bottom-up

Teconlogie

- Hadoop e l’ecosistema (HDFS, MapReduce, YARN)
- Apache Spark e il calcolo distribuito
- Strumenti NoSQL (MongoDB, Cassandra, HBase)

Cosa impareremo?

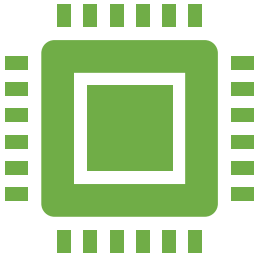
Impareremo a analizzare diversi tipi di dati:

high dimensional

graph

labeled

text



Impareremo a usare diversi modelli di computazione e tecnologie:

MapReduce

Single machine in-memory

Spark

Che problem risolveremo?

Probelmi della vita reale:

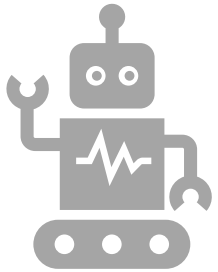
Recommender systems

Market Basket Analysis

Spam detection

Duplicate document detection

..



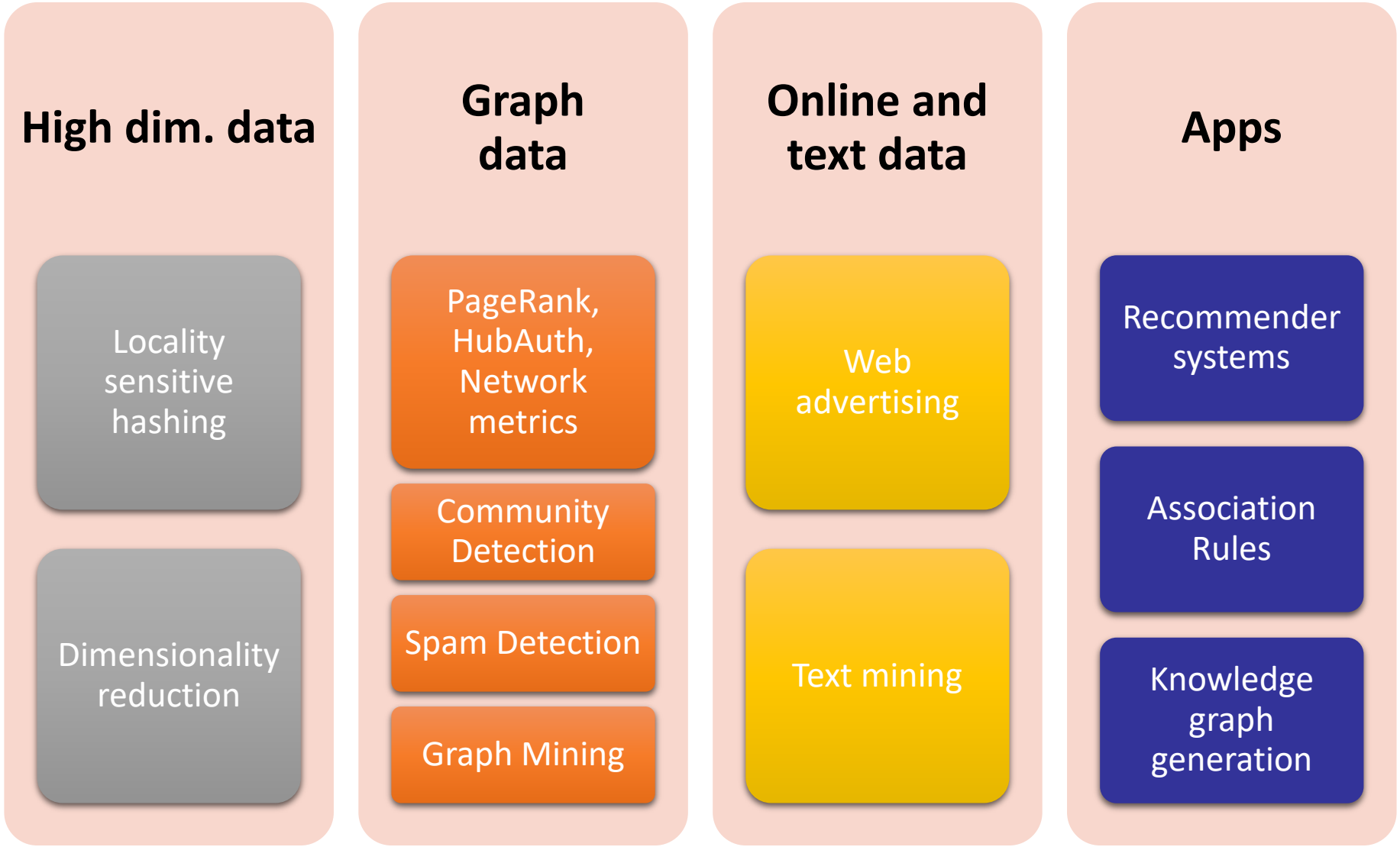
Impareremo diverse “metodologie”:

Linear algebra (SVD, Rec. Sys., Communities)

Optimization (stochastic gradient descent)

Hashing (LSH)

Tutto assieme



Programma /1

Sistemi di raccomandazione:

- Collaborative filtering,
- Modelli a semantica latente
- Network based inference
- Modelli ibridi

Map-Reduce:

- Concetti, motivazioni e algoritmi:
- conteggio parole documenti, prodotto vettore/matrice; Prodotto matrice/matrice; Join; Group By
- Beyond map-reduce: Spark

Programma /2

Ricerca di similarità su alte dimensioni:

- Shingling
- Min-Hashing
- Locality Sensitive Hashing (LSH)
- Min-LSH
- Applicazioni

Dimensionality reduction:

- PCA
- SVD
- CUR
- Proiezioni random e Teorema di Johnson-Lindenstrauss

Link Analysis:

- PageRank e sue estensioni
- Link spam
- Hub-Authorities
- Metriche per l'analisi di reti
- Applicazioni su Map-Reduce

Programma /3

Web Advertising:

- Algoritmi online
- Adword e sue implementazioni

Graph mining e network analysis:

- Network models
- community detection: overlapping communities
- Applicazioni
- Graph Neural Networks
- Knowledge graph generation

Text mining

- TF.IDF, Bag-Of-Word,
- Entity annotation based on AI
- Applicazioni

https://colab.research.google.com/

Informazioni

Contatti

- Prof. Alfredo Pulvirenti
  - Stanza 35 terzo blocco, Dipartimento di Matematica e Informatica
  - Tel. 095-7383087
  - e-mail: apulvirenti@dmi.unict.it
  - Homepage: http://www.dmi.unict.it/~apulvirenti/
  - Materiale: http://studium.unict.it

Ricevimento

- **Mercoledi’ 11-13 (in presenza)**

Logistica

- **Course website:**

  http://studium.unict.it

  - Slide
  - Esercizi, soluzioni
  - Letture

- Libri
  - **Mining of Massive Datasets**, Leskovec, Rajaraman, Ullman , **Free online:** http://www.mmds.org
  - **Large-Scale Data Analytics with Python and Spark** Isaac Triguero, Mikel Galar
  - **Introduction to Data Mining**, Tan, Steinbach, Kumar, Pearson Ed.
  - **Data Mining: Concepts and Techniques,** Han, Kamber, Morgan Kaufmann Ed.

Communicazioni

- **Studium**
- **email:**
  - apulvirenti@dmi.unict.it
- **Messaggi sul corso saranno pubblicati su studium**

Prerequisiti

Algoritmi

- Programmazione dinamica, strutture dati

Basic probability

- Momenti, distribuzioni, MLE, …

Programmazione

- C++/Java/R ecc. saranno utili

Esame

Scritto: 25%

Progetto: 75% (da consegnare entro 60 giorni dal superamento dello scritto)

Attività laboratoriale svolta in aula
3 punti extra