BIG DATA

INTRODUCERE

Modalitatea de evaluare

- Proiect individual sau de echipă, în funcție de dimensiunea problemei tratate
- **Structură Proiect**:

Alegeți un set de date și o aplicație ML, pe marginea cărora efectuați următoarele task-uri:

- Utilizați script-uri Spark pentru procesarea datelor, curățarea, transformarea acestora etc. Utilizați Dataframes și Spark SQL (ambele). (3p)
- Utilizati Spark ML/Tensorflow pentru clasificări, regresii (în funcție de problemă). Folosiți cel puțin două metode: o metodă clasică ML și una Deep Learning. (4p)
- Creați un proces de streaming (din orice sursă) folosind Spark Streaming. Aplicați modelele ML antrenate anterior pentru inferența în timp real sau realizați învățarea online a unui model. De exemplu: procesarea în timp real a imaginilor care provin de la diferite camere video într-un sistem distribuit. (2p)
  - Pentru aplicarea Spark Streaming în afara contextului ML punctajul maxim este 1p.

Predarea proiectului

- Proiectul va fi trimis până ~~pe 30 mai 2021~~ cel târz**iu cu 3 zile înainte de data programată a examenului**. Link-ul pentru încărcare va fi anunțat prin Moodle și Teams.
- Proiectul încărcat va conține:
  - un document pdf care va prezenta problema tratată și va explica detaliat soluția
  - codul sursă
- Doar proiectele complete (cu document și cod sursă) vor intra în examen
- Proiectul va fi susținut în ziua examenului

Bibliografie

- Bibliografia indicată în fișa cursului
- Link-urile indicate în timpul cursului și laboratorului

Plan General

- Hadoop
- Spark
- Tensorflow

Introducere în Big Data

Ce este Big Data?

Introducere în Big Data

- O definiție „naivă”:
  - Big Data se referă doar la dimensiunea datelor
  - 1 GB / 1 TB / 1 PB ... ?
- O interpretare de acest tip pierde din vedere aspecte importante precum:
  - Timpul – Analiza a 1GB de date într-o zi este diferită de analiza a 1GB de date în câteva secunde
  - Diversitatea – Analiza unor foi de calcul cu valori numerice este diferită de analiza de pagini web care conțin un amestec de text și imagini
  - Distribuirea – Analiza datelor ce provin dintr-o singură sursă este diferită de analiza datelor ce provin din surse multiple

Definiția Big Data

- Big data reprezintă sisteme de informații cu **volum**, **viteză** și **varietate** ridicate ce solicită forme accesibile și inovative de procesare a informației pentru a permite analiza, luarea deciziilor și automatizarea proceselor.
- Glosarul IT Gartner:

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

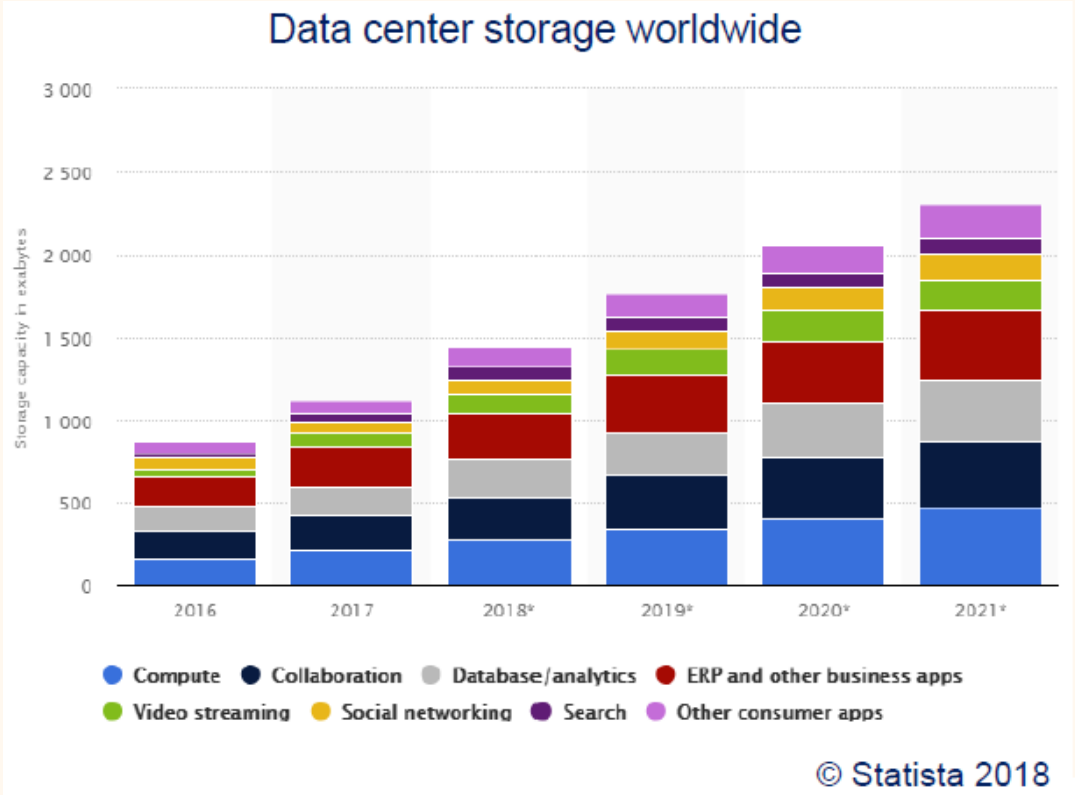
- 3 V:
  - Volum
  - Viteză
  - Varietate

Definiția Big data

- În unele lucrări se folosesc 10 V pentru a defini Big data:
  - Variabilitate
  - Veracitate
  - Validitate
  - Vulnerabilitate
  - Volatilitate
  - Vizualizare
  - Valoare

Volum

- Scala datelor trebuie să fie considerabilă („big”)
  - Nu este o definiție clară
  - „that demand[…] innovative forms of information processing“ (Gartner)

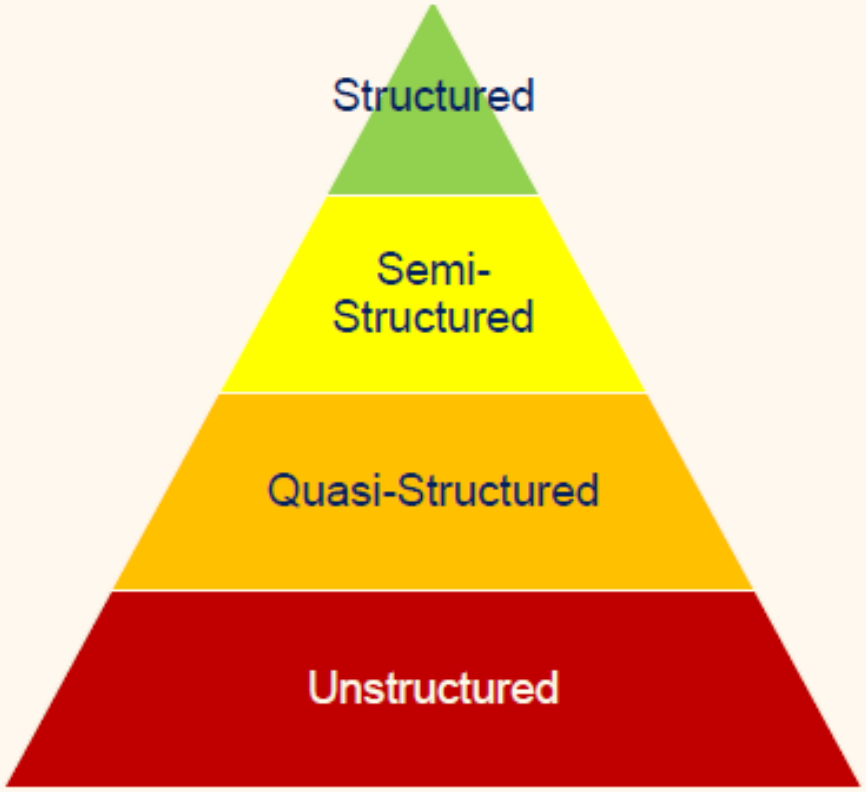


Velocitate

- Viteza cu care datele sunt create
- Viteza cu care datele trebuie procesate și analizate
  - De foarte multe ori – aproape de real-time

Varietate

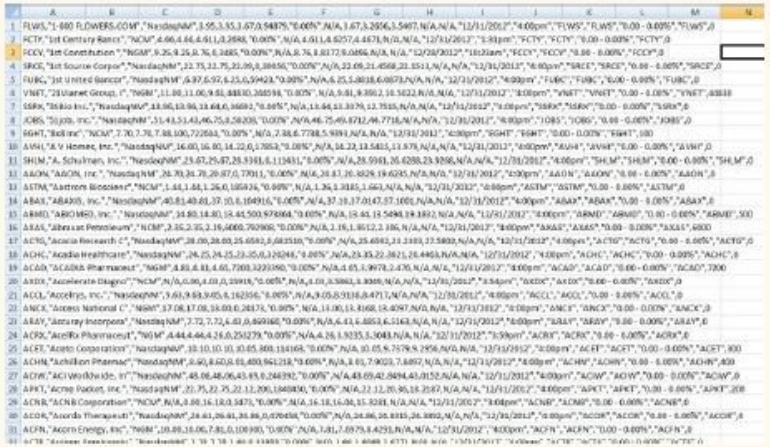
- Diversitate în tipurile de date și sursele de date



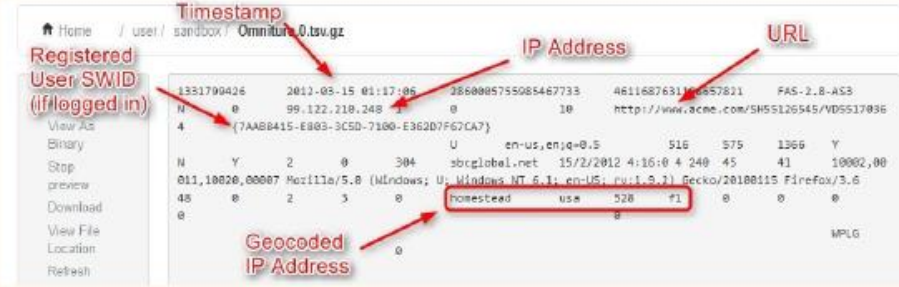
- Date cu tipuri și structuri definite (exemplu: csv)
- Date de tip text parsabile (exemplu: fișiere XML cu schemă XSD)
- Date de tip text care pot fi formatate (exemplu: datele *clickstream*)
- Date fără structură inerentă, adesea cu mai multe formate (exemplu: site-uri web, conținut video)

Tipuri de date

Structured



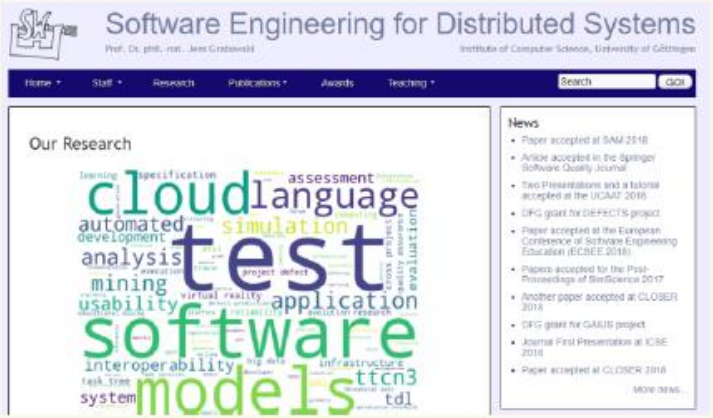
Quasi-Structured



Semi-Structured

```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

Unstructured



Data Science

- Nu există o definiție clară
- Scopul: extragerea cunoștințelor din date
- Combinarea tehnicilor din cadrul unor discipline diferie
- Analiza datelor este ghidată de principii științifice

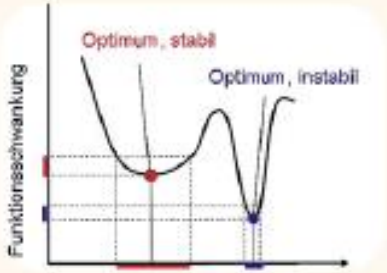
Data Science

- Nu există o definiție clară
- Scopul: extragerea cunoștințelor din date
- Combinarea tehnicilor din cadrul unor discipline diferie
- Analiza datelor este ghidată de principii științifice

- Metode, Big Data, Machine Learning... ?

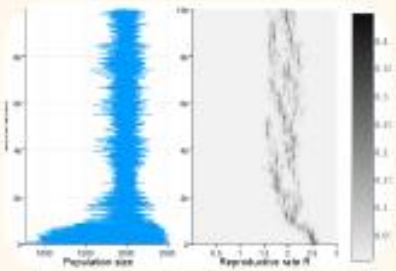
Aspecte matematice

Computational Geometry



Optimization

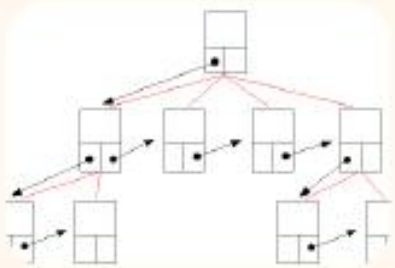
Stochastics



Scientific Computing

Machine Learning

Aspecte informatice



Data Structures and Algorithms

Data bases

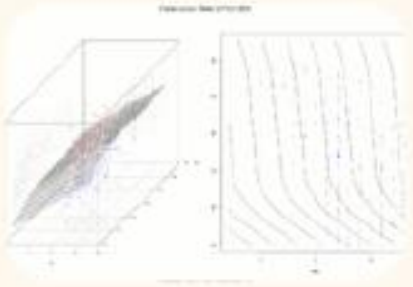
Distributed Computing

Software Engineering

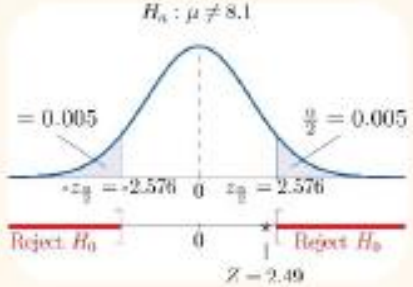
Artificial Intelligence

Machine Learning

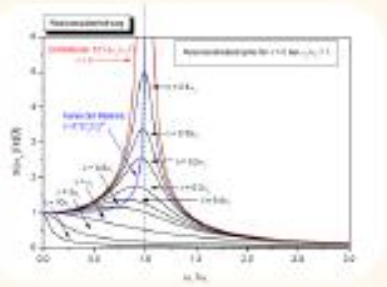
Aspecte statistice



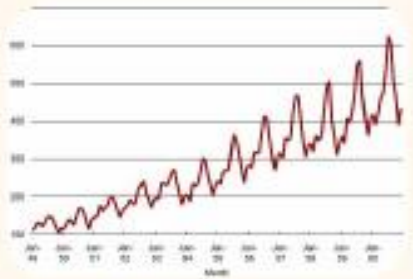
Linear Models



Statistical Tests



Inference



Time Series Analysis

Machine Learning

Aplicații

Intelligent Systems

Robotics

Marketing

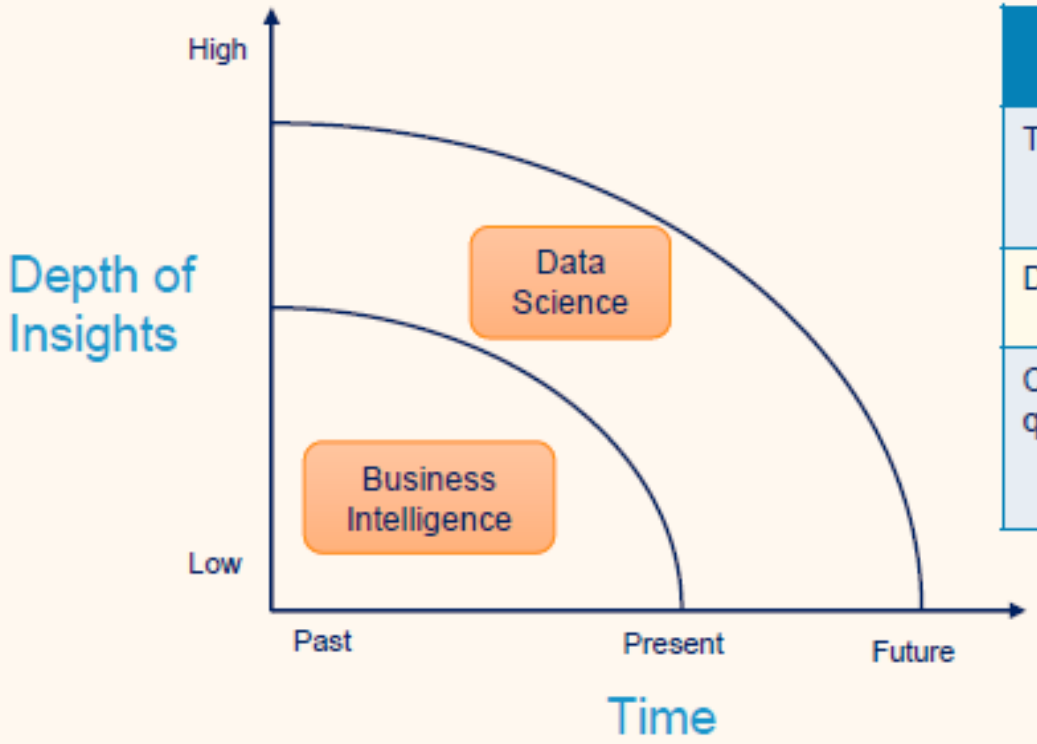
Medicine

Autonomous Driving

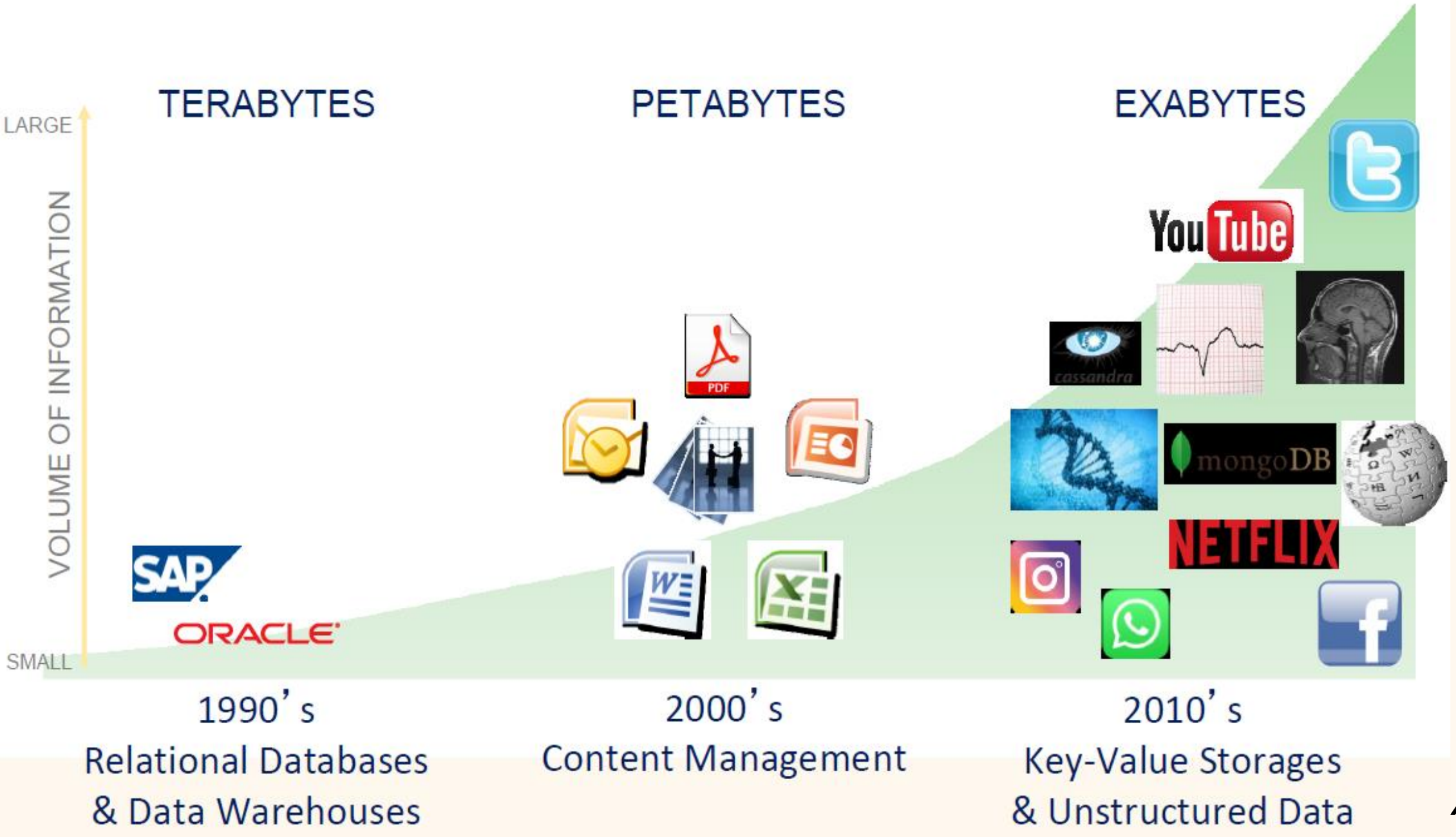
Social Networks

Data Science vs Business Intelligence

- BI = […] *best practices that enable access to and analysis of information to improve and optimize decisions and performance* (Gartner)



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optmization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened…? How much did…? When did…?	What if…? What will…? How can we…?


TERABYTES
PETABYTES
EXABYTES
LARGE
VOLUME OF INFORMATION
SMALL
SAP
ORACLE
PDF
YouTube
cassandra
mongoDB
NETFLIX
1990's
Relational Databases
& Data Warehouses
2000's
Content Management
2010's
Key-Value Storages
& Unstructured Data

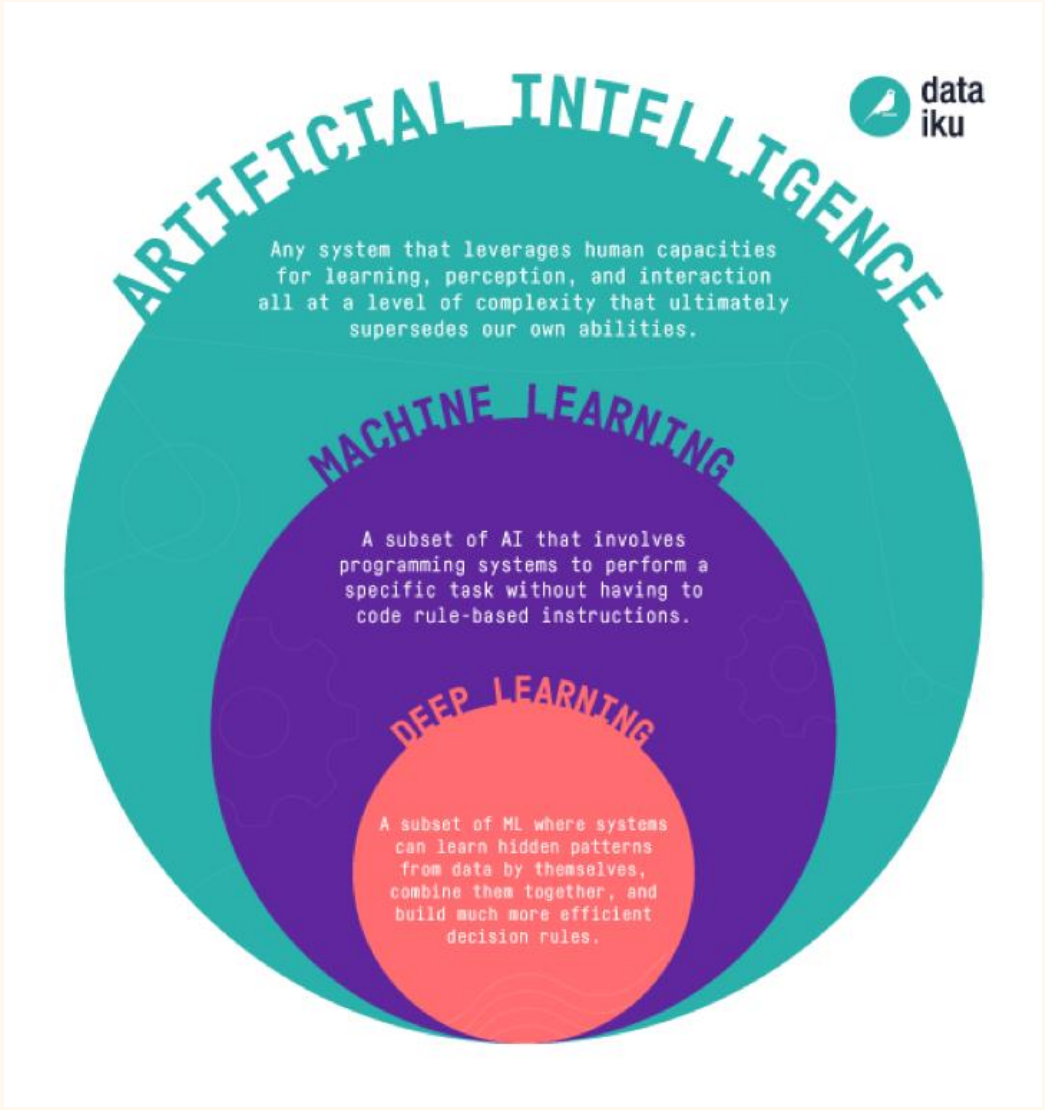
Inteligența Artificială

- **Machine Learning:** Programe informatice care permit unei mașini să învețe din date, cu scopuri de a efectua task-uri complexe care nu au fost programate explicit.

- **Deep Learning:** o clasă de algoritmi de ML specializată în învățarea pornind de la date nestructurate.

- **Inteligența Artificială:** o disciplină științifică ce include tehnici și algoritmi ce permit simularea inteligenței.

- **Data Science:** un domeniu pluridisciplinar ce vizează extragerea de informații utile pornind de la date, combinând tehnologia și metodele științifice.



Ce face un Data Scientist?

În cadrul activităților noastre cotidiene, generăm și consumăm cantități uriașe de date

De exemplu, căutarea pe site-urile web,

rețelele sociale, informațiile meteorologice etc.

Rolul său este de a analiza datele și a le utiliza cu scopuri definite, sprijinind luarea celor mai bune decizii