




# **B I G D A T A**

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 căroră efectuați următoarele task-uri:

- Utilizați script-uri Spark pentru procesarea datelor, curățarea, transformarea acestora etc. Utilizați Dataframes și Spark SQL (ambele). (3p)
- Utilizați 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 târziu **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





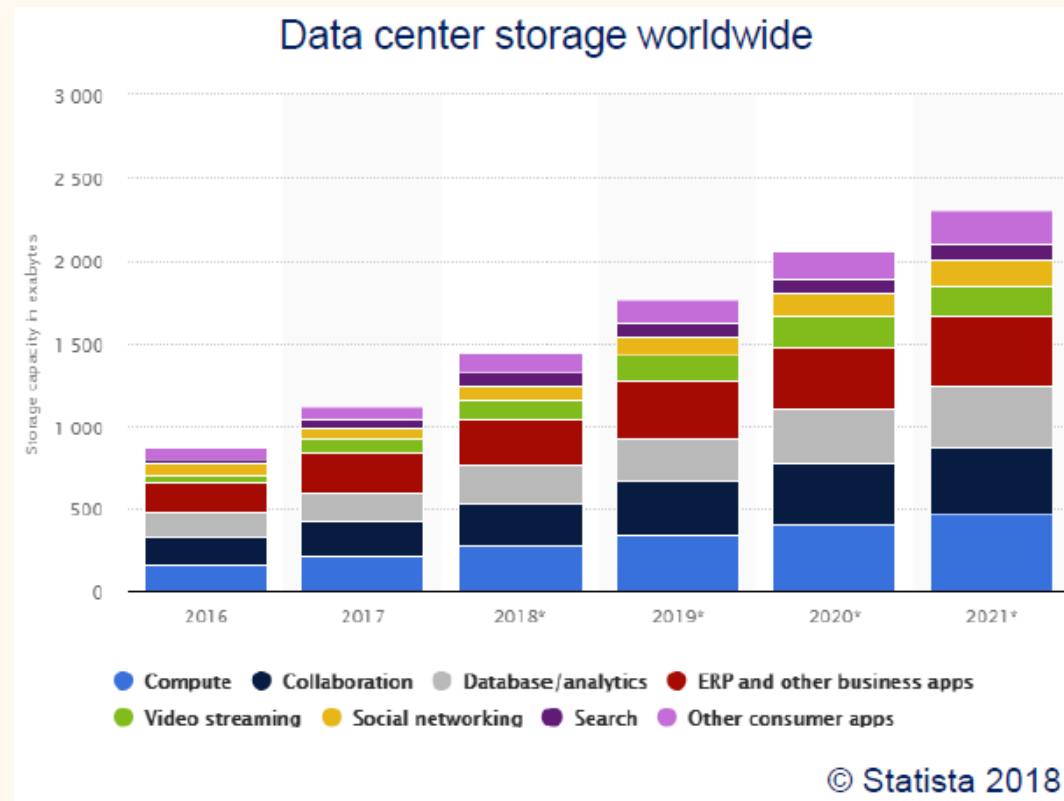
# ○ Definiția Big data

- 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)



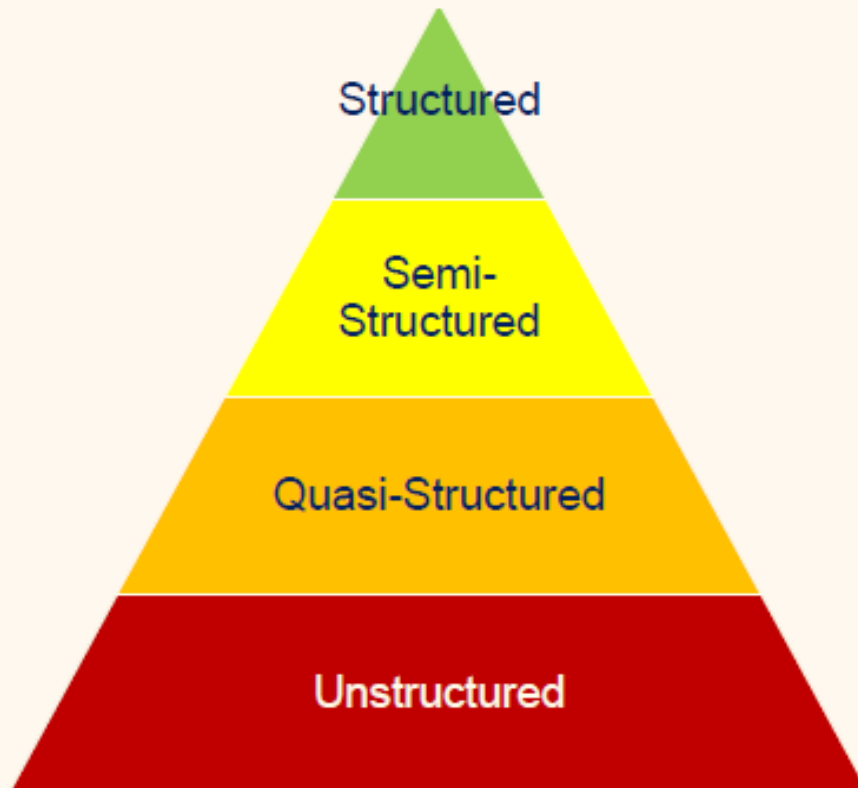
# ○ Velocitate

- 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

| A  | B   | C | D | E | F | G | H | I | J | K | L | M | N |
|----|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1  | FLWS,"1-800-FLWERS.COM","NasdaqNM","3.55,3.55,3.67,0.54875,"0.00%","N/A,1.67,3.2056,3.5497,"N/A,N/A,"12/11/2012","4.00pm","FLWS","FLWS","0.00 - 0.00%","FLWS",0           |   |   |   |   |   |   |   |   |   |   |   |   |
| 2  | ACTV,"1st Century Bank","NasdaqNM","4.66,4.66,4.61,0.2598,"0.00%","N/A,4.61,4.62,0.74447,0.00%","N/A,N/A,"12/11/2012","3.39pm","ACTV","ACTV","0.00 - 0.00%","ACTV",0      |   |   |   |   |   |   |   |   |   |   |   |   |
| 3  | FCV,"VIM Corporation","NasdaqNM","9.26,9.26,9.26,0.0000,"0.00%","N/A,9.26,9.27,0.0000,"N/A,N/A,"12/11/2012","10:00am","FCV","FCV","0.00 - 0.00%","FCV",0                  |   |   |   |   |   |   |   |   |   |   |   |   |
| 4  | SPKE,"Sun Source Corp","NasdaqNM","21.75,21.75,21.80,0.0000,"0.00%","N/A,21.80,21.80,0.0000,"N/A,N/A,"12/11/2012","9:00pm","SPKE","SPKE","0.00 - 0.00%","SPKE",0          |   |   |   |   |   |   |   |   |   |   |   |   |
| 5  | FURC,"3rd United Bancorp","NasdaqNM","6.97,6.97,6.97,0.0000,"0.00%","N/A,6.97,6.97,0.0000,"N/A,N/A,"12/11/2012","4:00pm","FURC","FURC","0.00 - 0.00%","FURC",0            |   |   |   |   |   |   |   |   |   |   |   |   |
| 6  | VNET,"Vietnam Group","NasdaqNM","11.88,11.88,11.88,0.0000,"0.00%","N/A,11.88,11.88,0.0000,"N/A,N/A,"12/11/2012","3:00pm","VNET","VNET","0.00 - 0.00%","VNET",0            |   |   |   |   |   |   |   |   |   |   |   |   |
| 7  | SEK,"Seiko Inc","NasdaqNM","13.86,13.86,13.86,0.0000,"0.00%","N/A,13.86,13.86,0.0000,"N/A,N/A,"12/11/2012","3:00pm","SEK","SEK","0.00 - 0.00%","SEK",0                    |   |   |   |   |   |   |   |   |   |   |   |   |
| 8  | JCRS,"JCRS Inc","NasdaqNM","51.43,51.43,51.43,0.0000,"0.00%","N/A,51.43,51.43,0.0000,"N/A,N/A,"12/11/2012","4:00pm","JCRS","JCRS","0.00 - 0.00%","JCRS",0                 |   |   |   |   |   |   |   |   |   |   |   |   |
| 9  | EGHT,"Eight Inc","NasdaqNM","1.76,1.76,1.76,0.0000,"0.00%","N/A,1.76,1.76,0.0000,"N/A,N/A,"12/11/2012","3:00pm","EGHT","EGHT","0.00 - 0.00%","EGHT",0                     |   |   |   |   |   |   |   |   |   |   |   |   |
| 10 | AVHL,"A V Homes Inc","NasdaqNM","14.00,14.00,14.00,0.0000,"0.00%","N/A,14.00,14.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","AVHL","AVHL","0.00 - 0.00%","AVHL",0            |   |   |   |   |   |   |   |   |   |   |   |   |
| 11 | SHLM,"A. Schulman Inc","NasdaqNM","23.67,23.67,23.67,0.0000,"0.00%","N/A,23.67,23.67,0.0000,"N/A,N/A,"12/11/2012","4:00pm","SHLM","SHLM","0.00 - 0.00%","SHLM",0          |   |   |   |   |   |   |   |   |   |   |   |   |
| 12 | AAON,"AAON Inc","NasdaqNM","28.70,28.70,28.70,0.0000,"0.00%","N/A,28.70,28.70,0.0000,"N/A,N/A,"12/11/2012","4:00pm","AAON","AAON","0.00 - 0.00%","AAON",0                 |   |   |   |   |   |   |   |   |   |   |   |   |
| 13 | ASTM,"ASTM International","NasdaqNM","1.44,1.44,1.44,0.0000,"0.00%","N/A,1.44,1.44,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ASTM","ASTM","0.00 - 0.00%","ASTM",0            |   |   |   |   |   |   |   |   |   |   |   |   |
| 14 | ABAX,"ABAX Inc","NasdaqNM","40.81,40.81,40.81,0.0000,"0.00%","N/A,40.81,40.81,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ABAX","ABAX","0.00 - 0.00%","ABAX",0                 |   |   |   |   |   |   |   |   |   |   |   |   |
| 15 | AKMD,"AKMD Inc","NasdaqNM","14.80,14.80,14.80,0.0000,"0.00%","N/A,14.80,14.80,0.0000,"N/A,N/A,"12/11/2012","4:00pm","AKMD","AKMD","0.00 - 0.00%","AKMD",0                 |   |   |   |   |   |   |   |   |   |   |   |   |
| 16 | ABAX,"ABAX Inc","NasdaqNM","40.81,40.81,40.81,0.0000,"0.00%","N/A,40.81,40.81,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ABAX","ABAX","0.00 - 0.00%","ABAX",0                 |   |   |   |   |   |   |   |   |   |   |   |   |
| 17 | ACLS,"Accellis Pharmaceuticals","NasdaqNM","28.00,28.00,28.00,0.0000,"0.00%","N/A,28.00,28.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACLS","ACLS","0.00 - 0.00%","ACLS",0 |   |   |   |   |   |   |   |   |   |   |   |   |
| 18 | ACHC,"Acadia Healthcare","NasdaqNM","24.25,24.25,24.25,0.0000,"0.00%","N/A,24.25,24.25,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACHC","ACHC","0.00 - 0.00%","ACHC",0        |   |   |   |   |   |   |   |   |   |   |   |   |
| 19 | ACAD,"Acadia Pharmaceuticals","NasdaqNM","4.85,4.85,4.85,0.0000,"0.00%","N/A,4.85,4.85,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACAD","ACAD","0.00 - 0.00%","ACAD",0        |   |   |   |   |   |   |   |   |   |   |   |   |
| 20 | AKDX,"Accelerate Diagnostics","NasdaqNM","10.00,10.00,10.00,0.0000,"0.00%","N/A,10.00,10.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","AKDX","AKDX","0.00 - 0.00%","AKDX",0   |   |   |   |   |   |   |   |   |   |   |   |   |
| 21 | ACCL,"Accelrys Inc","NasdaqNM","1.60,1.60,1.60,0.0000,"0.00%","N/A,1.60,1.60,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACCL","ACCL","0.00 - 0.00%","ACCL",0                  |   |   |   |   |   |   |   |   |   |   |   |   |
| 22 | AKIS,"Accuris National Corp","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","AKIS","AKIS","0.00 - 0.00%","AKIS",0         |   |   |   |   |   |   |   |   |   |   |   |   |
| 23 | ABAY,"Accuris Inc","NasdaqNM","1.72,1.72,1.72,0.0000,"0.00%","N/A,1.72,1.72,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ABAY","ABAY","0.00 - 0.00%","ABAY",0                   |   |   |   |   |   |   |   |   |   |   |   |   |
| 24 | ACRF,"Accuris Pharmaceuticals","NasdaqNM","1.44,1.44,1.44,0.0000,"0.00%","N/A,1.44,1.44,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACRF","ACRF","0.00 - 0.00%","ACRF",0       |   |   |   |   |   |   |   |   |   |   |   |   |
| 25 | ACST,"Acetate Corporation","NasdaqNM","10.00,10.00,10.00,0.0000,"0.00%","N/A,10.00,10.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACST","ACST","0.00 - 0.00%","ACST",0      |   |   |   |   |   |   |   |   |   |   |   |   |
| 26 | ACHN,"Achillion Pharmaceuticals","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACHN","ACHN","0.00 - 0.00%","ACHN",0     |   |   |   |   |   |   |   |   |   |   |   |   |
| 27 | ACW,"ACW Inc","NasdaqNM","4.00,4.00,4.00,0.0000,"0.00%","N/A,4.00,4.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACW","ACW","0.00 - 0.00%","ACW",0                           |   |   |   |   |   |   |   |   |   |   |   |   |
| 28 | APKT,"Acropac Inc","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","APKT","APKT","0.00 - 0.00%","APKT",0                   |   |   |   |   |   |   |   |   |   |   |   |   |
| 29 | ACNB,"ACNB Corporation","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACNB","ACNB","0.00 - 0.00%","ACNB",0              |   |   |   |   |   |   |   |   |   |   |   |   |
| 30 | ACOR,"Acorda Therapeutics","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACOR","ACOR","0.00 - 0.00%","ACOR",0           |   |   |   |   |   |   |   |   |   |   |   |   |
| 31 | ACFN,"Acorn Energy Inc","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACFN","ACFN","0.00 - 0.00%","ACFN",0              |   |   |   |   |   |   |   |   |   |   |   |   |
| 32 | ACR,"ACR Inc","NasdaqNM","1.00,1.00,1.00,0.0000,"0.00%","N/A,1.00,1.00,0.0000,"N/A,N/A,"12/11/2012","4:00pm","ACR","ACR","0.00 - 0.00%","ACR",0                           |   |   |   |   |   |   |   |   |   |   |   |   |

## 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>
```

## Quasi-Structured

| Timestamp                           |            | IP Address                             |                    | URL        |        |
|-------------------------------------|------------|--|--------------------|------------|--------|
| Registered User SWID (if logged in) | 1331790426 | 2011-05-15 01:17:05                    | 286008755085467733 | 4611687631 | 557821 |
|                                     | N          | 0                                      | 99.122.218.248     | 0          | 10     |
|                                     | 4          | (7AAB8415-E803-3C5D-7100-E362D7F67CA7) |                    |            |        |
| View As                             |            |  |                    |            |        |
| Binary                              |            |  |                    |            |        |
| Stop                                |            |  |                    |            |        |
| preview                             |            |  |                    |            |        |
| Download                            |            |  |                    |            |        |
| View File                           |            |  |                    |            |        |
| Location                            |            |  |                    |            |        |
| Refresh                             |            |  |                    |            |        |

## Unstructured

### Software Engineering for Distributed Systems

Prof. Dr. phil. rer. Jens Grieschke  
Institute of Computer Science, University of Göttingen

Home • Staff • Research • Publications • Awards • Teaching •

Search  GO

#### Our Research

Learning, Specification, assessment, cloud, language, automated, development, analysis, test, usability, mining, virtual reality, application, software, interoperability, system, models, ttcn3, td1

#### News

- Paper accepted at SAM 2018
- Article accepted in the Springer Software Quality Journal
- Two Presentations and a Tutorial accepted at the UIC 2018
- DFG grant for the project
- Paper accepted at the European Conference of Software Engineering Education (ECSEE 2018)
- Papers accepted for the Post-Proceedings of SimScience 2017
- Another paper accepted at CLOSER 2018
- DFG grant for the project
- Journal First Presentation at ICSE 2018
- Paper accepted at CLOSER 2018

MORE NEWS



# ○ Data Science

- Nu există o definiție clară
- Scopul: extragerea cunoștințelor din date
- Combinarea tehnicilor din cadrul unor discipline diferite
- 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 diferite
- Analiza datelor este ghidată de principii științifice
- Metode, Big Data, Machine Learning... ?

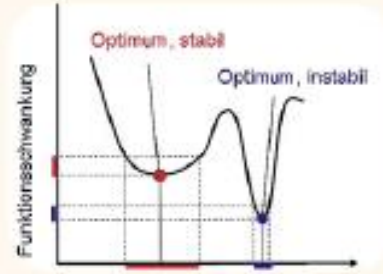




# ○ Aspecte matematice



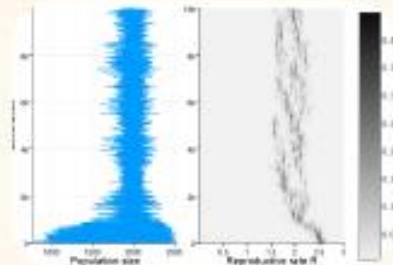
Computational  
Geometry



Optimization



Stochastics



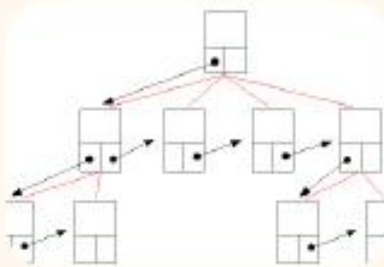
Scientific  
Computing



Machine  
Learning



# ○ Aspecte informatice



Data Structures and Algorithms



Data bases



Distributed Computing



Software Engineering



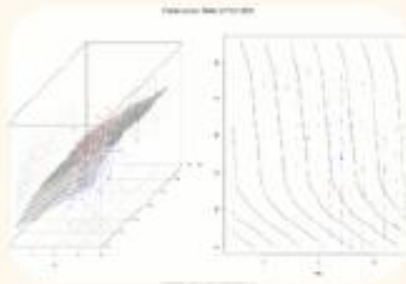
Artificial Intelligence



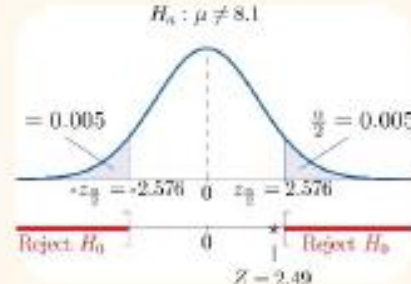
Machine Learning



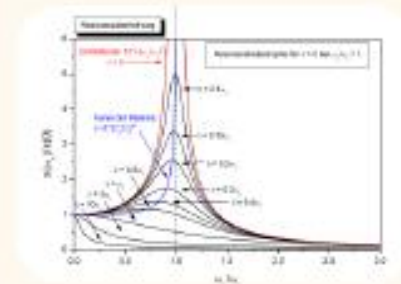
# ○ Aspecte statistice



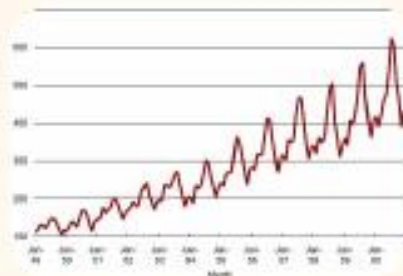
Linear Models



Statistical Tests



Inference



Time Series Analysis



Machine Learning



# ○ Aplicații



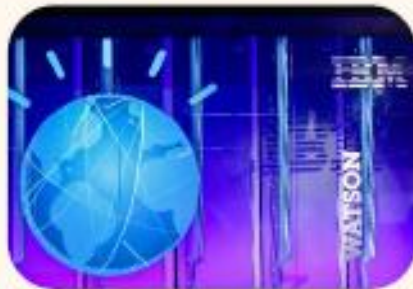
Intelligent Systems



Robotics



Marketing



Medicine



Autonomous Driving

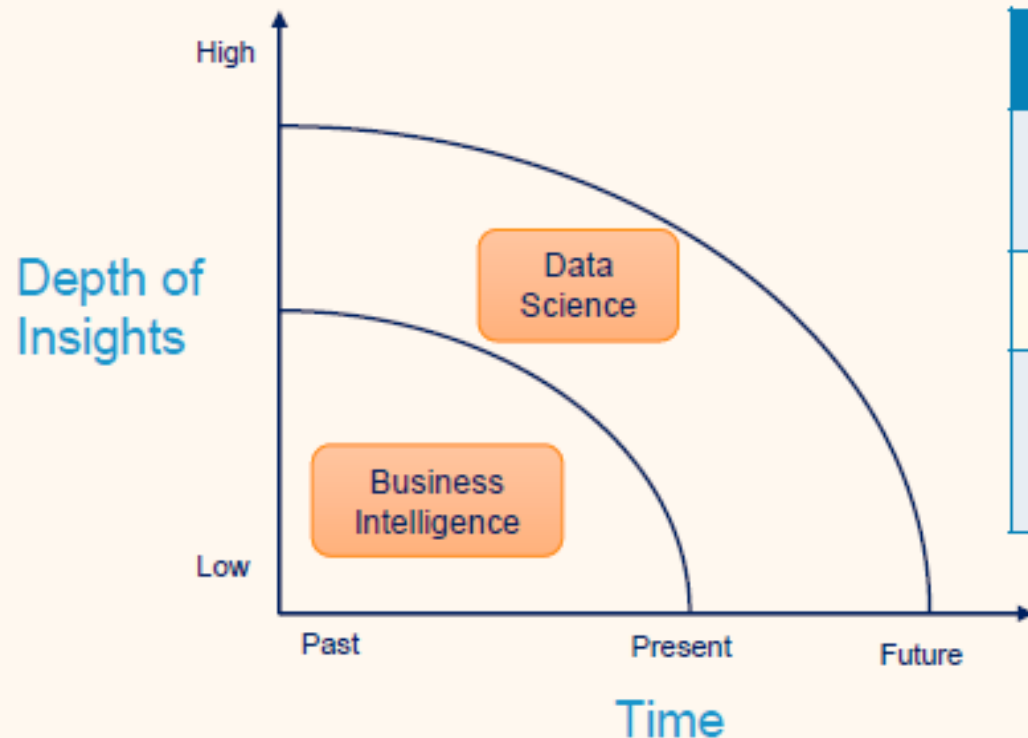


Social Networks



# ○ Data Science vs Business Intelligence

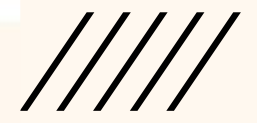
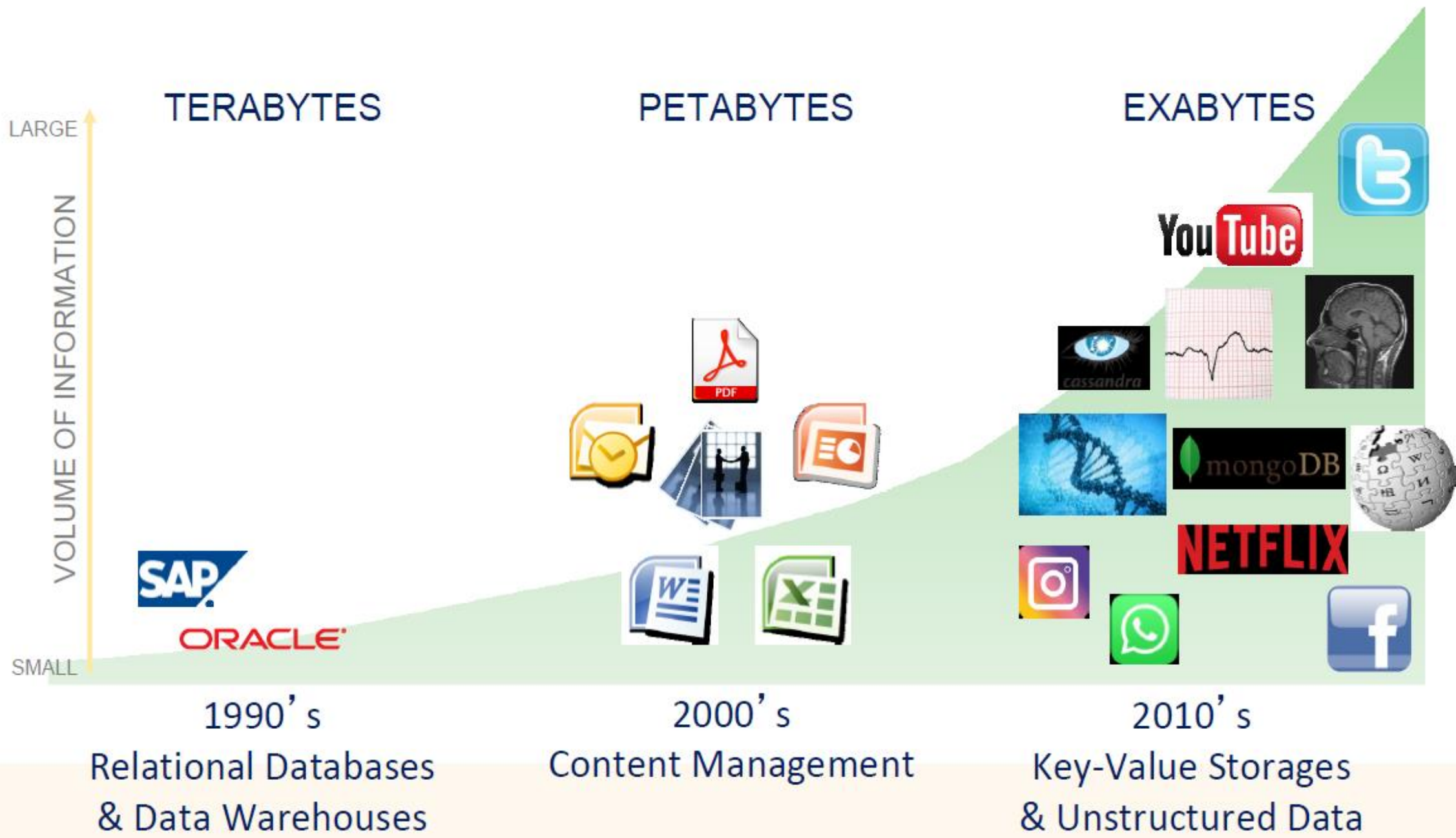
- BI = [...] *best practices that enable access to and analysis of information to improve and optimize decisions and performance* (Gartner)



|                  | 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...?<br>How much did...?<br>When did...? | What if...?<br>What will...?<br>How can we...? |

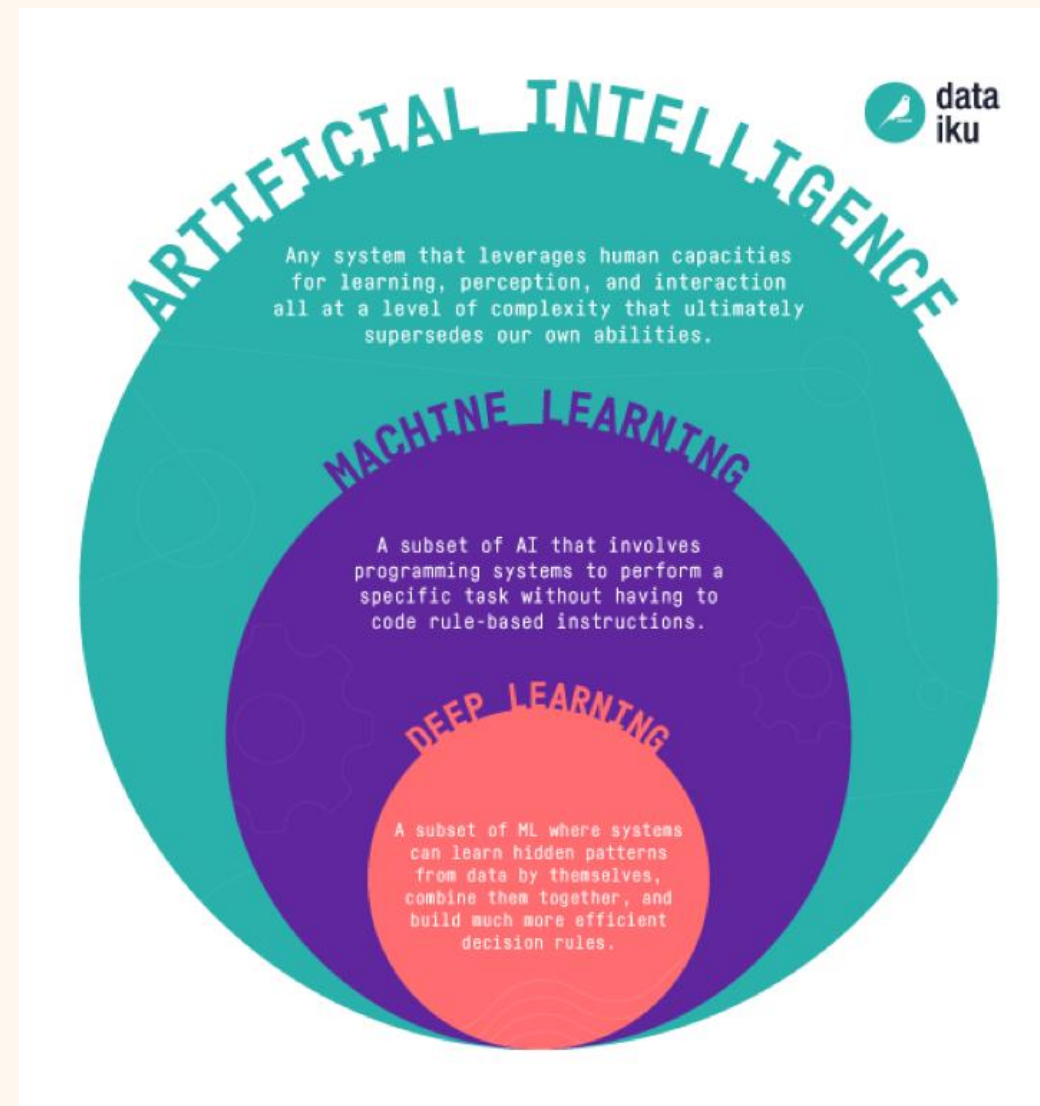






# Inteligența Artificială

- **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**