Data is the Glue

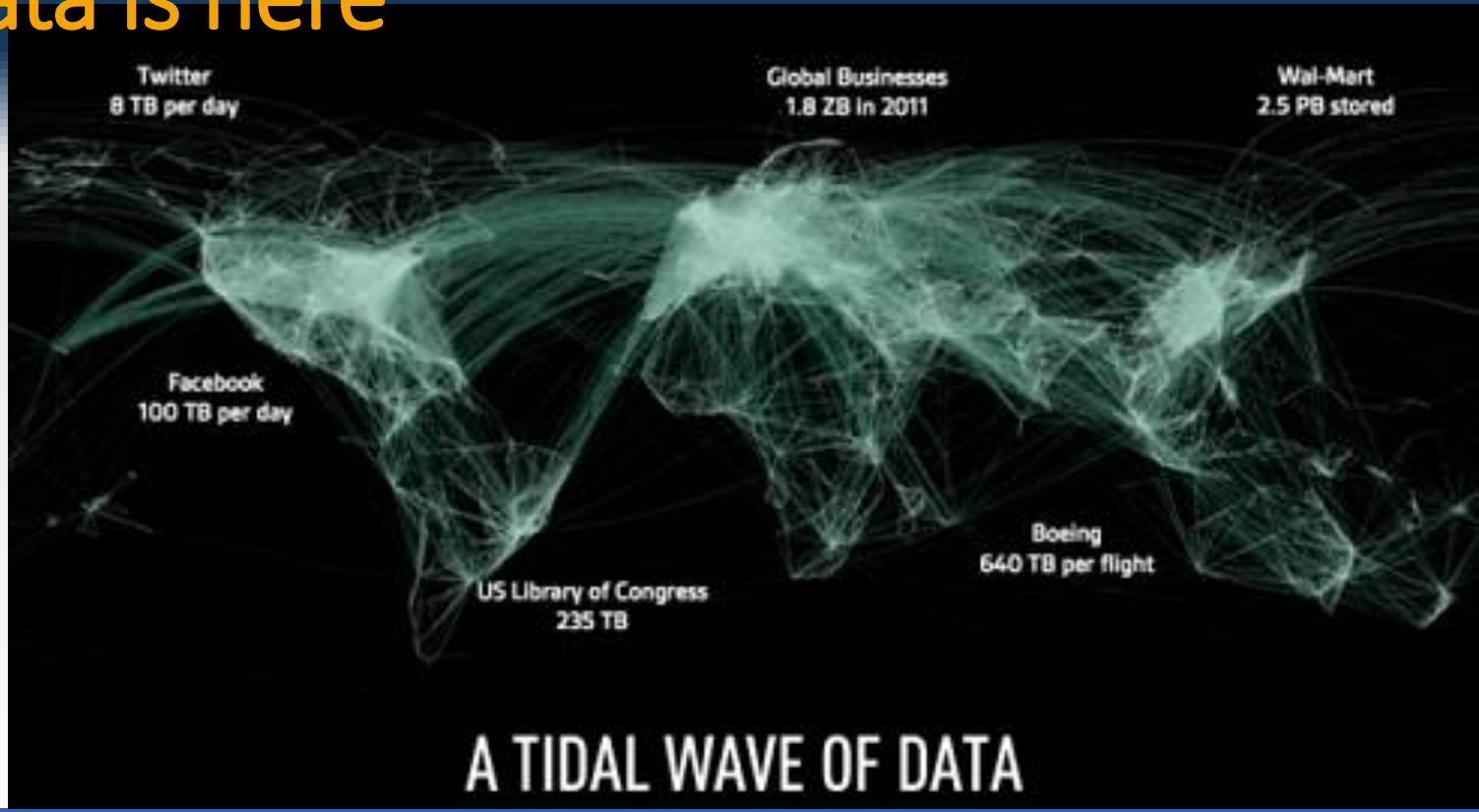
ELENA BARALIS

POLITECNICO DI TORINO

Data Base and Data Mining Group of Politecnico di Torino

Big data is here

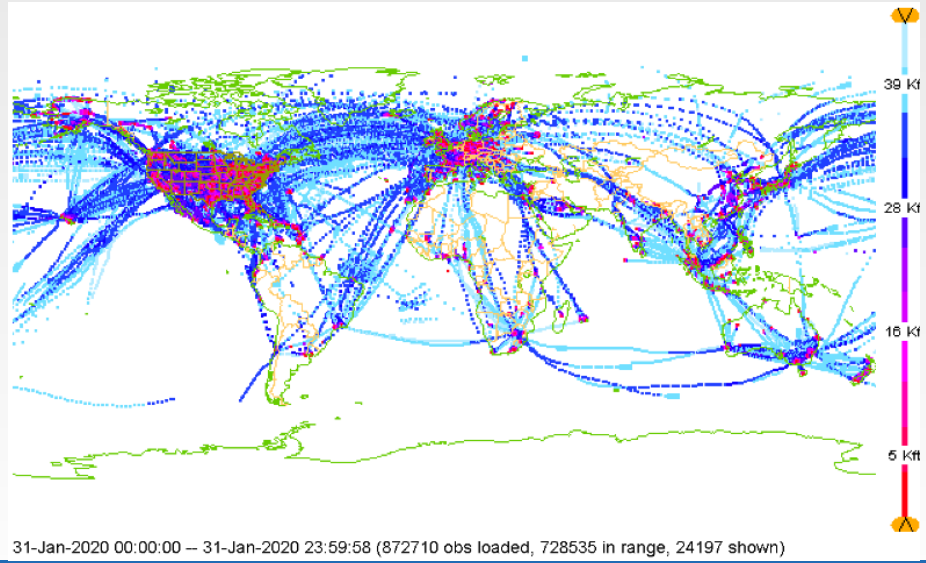
- Volume
- Velocity
- Variety
- Veracity
- … Value



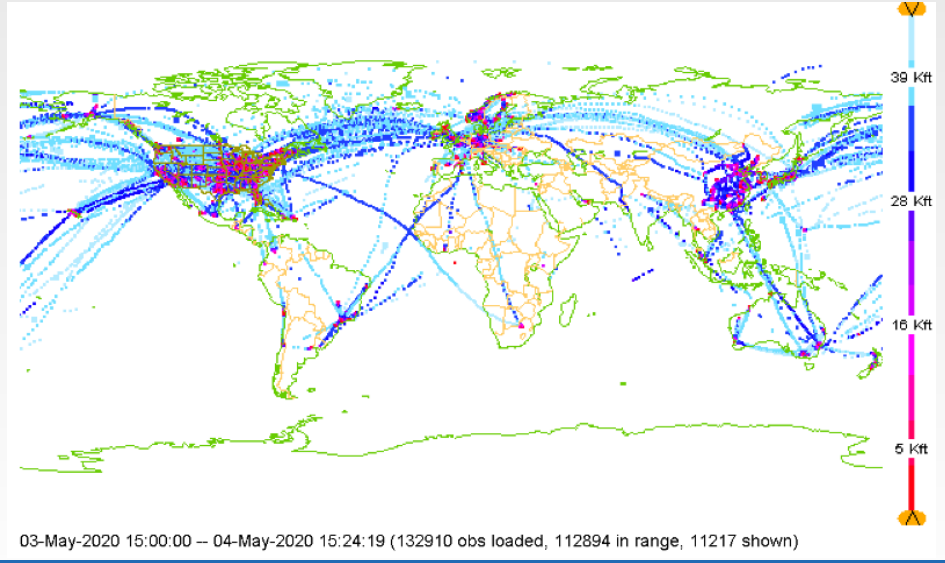
… but not always available!

WORLD METEOROLOGICAL ORGANIZATION

January 2020

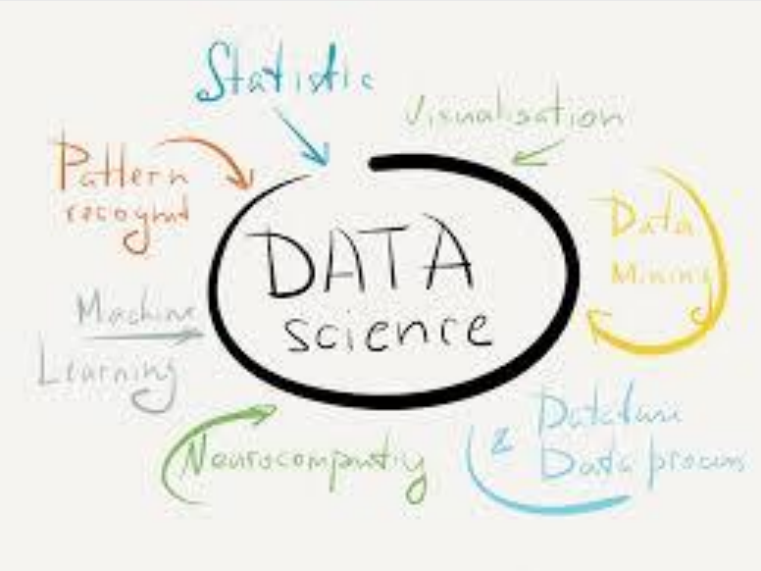


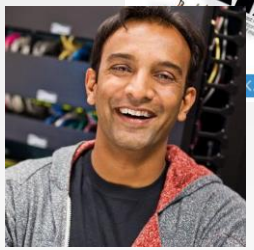
May 2020



Data science

"Extracting meaning from very large quantities of data"

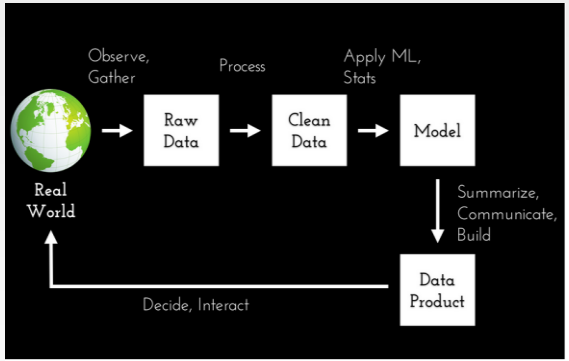




D.J. Patil coined the word *data scientist*

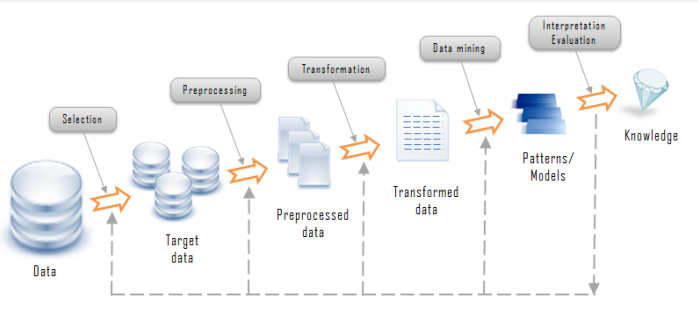
The data science process

Generation → Acquisition → Storage → Analysis

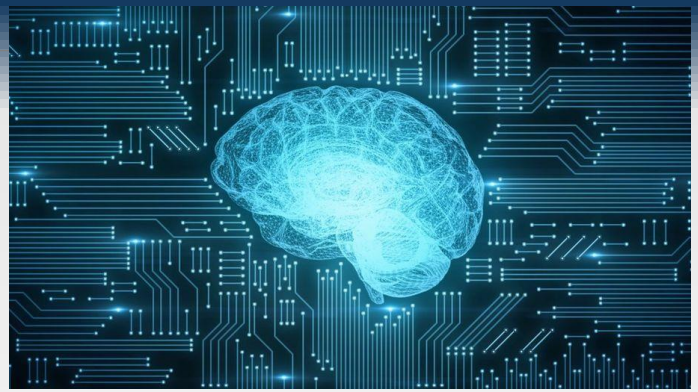


AKA *KDD* process

Knowledge Discovery in Databases



Analysis



- Objectives
  - Descriptive analytics
  - Predictive analytics
  - Prescriptive analytics
- Methods
  - Statistical analysis, data mining, text mining, network and graph data mining
  - Association analysis, classification and regression, clustering
- Diverse domains call for customized techniques

A word from practitioners

- At least 80-90% of their work involves not machine learning, but
  - Working with experts to understand the domain, assumptions, questions
  - Trying to catalog and make sense of the data sources
  - Wrangling, extracting, and integrating the data
  - Cleaning the wrangled data

Photo by Oliver Hale on Unsplash

Association rules

- Objective
  - extraction of frequent correlations or pattern from a transactional database

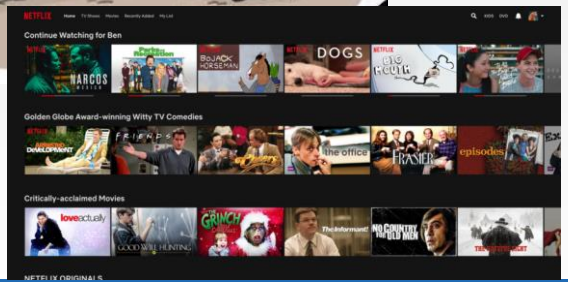
Tickets at a supermarket counter

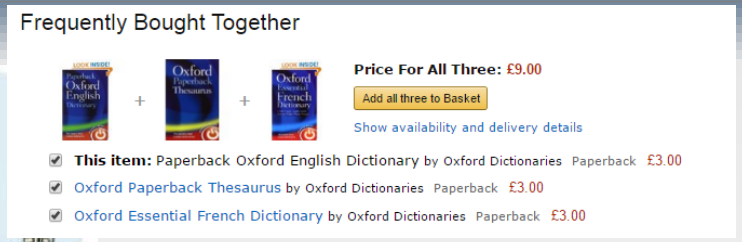
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk
…	…

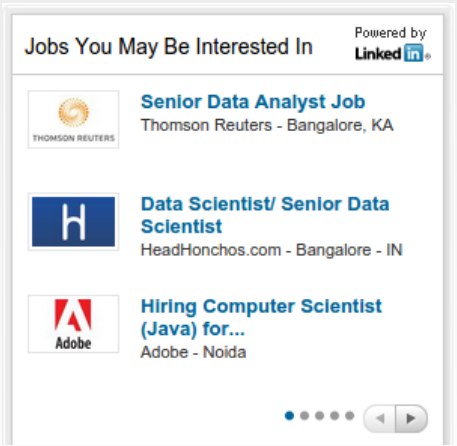
- Association rule

  diapers $\Rightarrow$ beer
  - 2% of transactions contains both items
  - 30% of transactions containing diapers also contain beer

Association rules





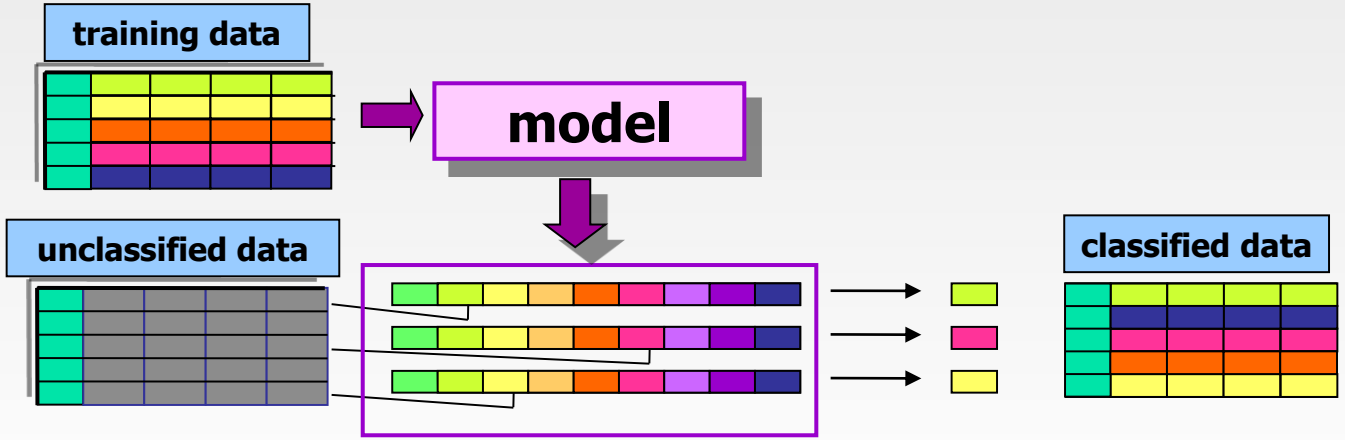


Spam or Ham?

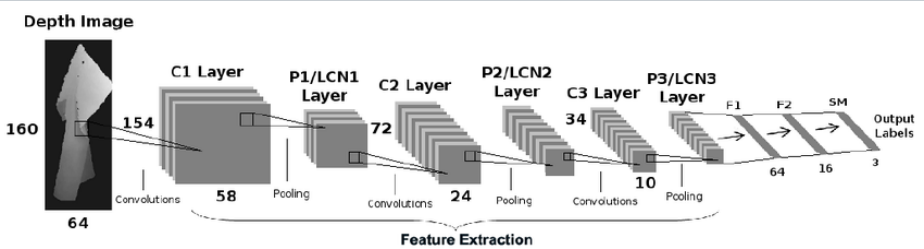
- What is the difference between *ham* and *spam*?
- How to *classify* incoming messages into ham and spam?

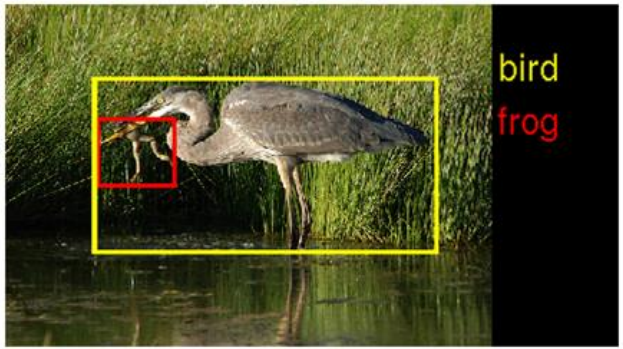
Classification

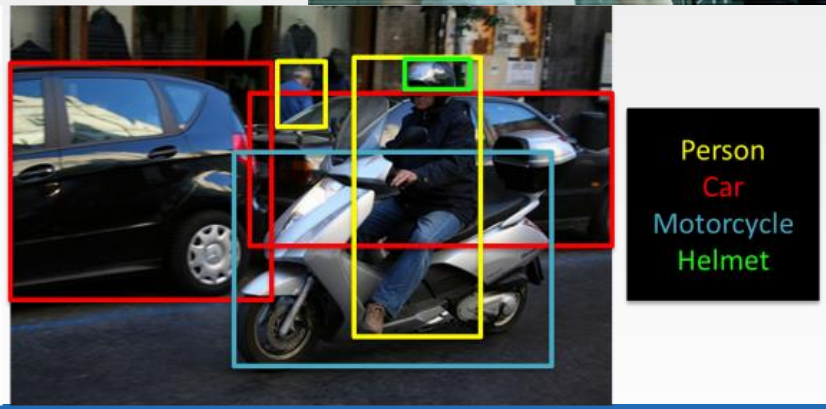
- Objectives
  - prediction of a class label
  - definition of an interpretable model of a given phenomenon



Classification

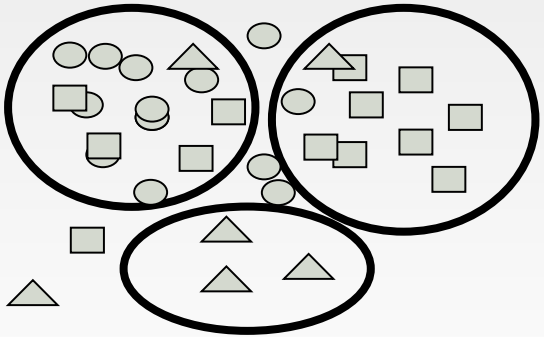






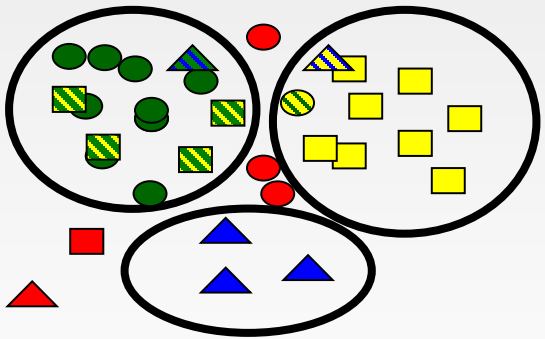
Clustering

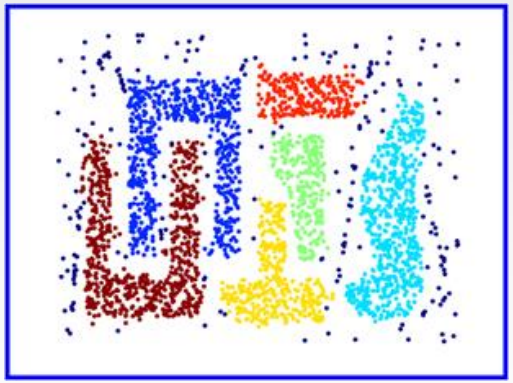
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers



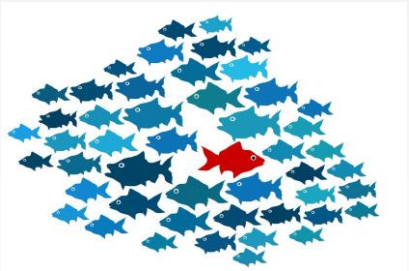
Clustering

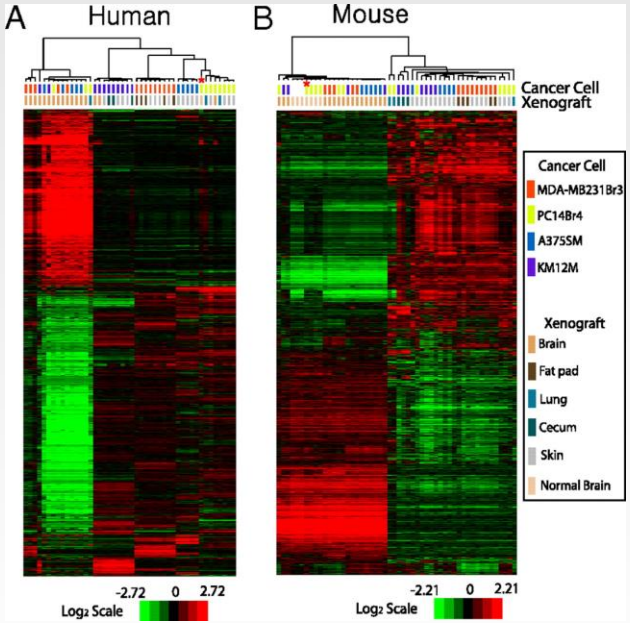
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers



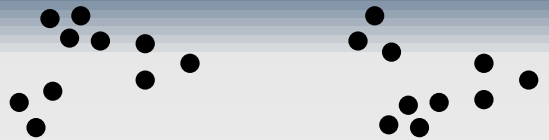


Clustering

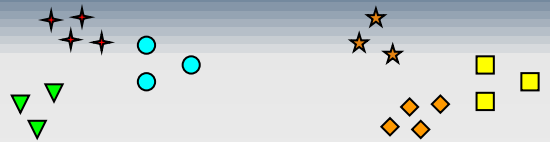



A
Human
B
Mouse
Cancer Cell Xenograft
Cancer Cell
MDA-MB231Br3
PC14Br4
A375SM
KM12M
Xenograft
Brain
Fat pad
Lung
Cecum
Skin
Normal Brain
-2.72
0
2.72
Log2 Scale
-2.21
0
2.21
Log2 Scale

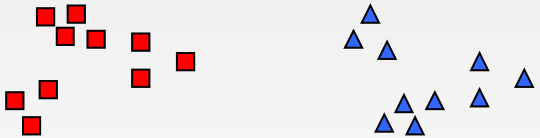
Notion of a Cluster can be Ambiguous



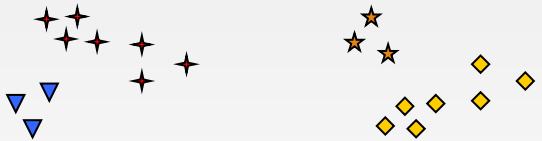
How many clusters?



Six Clusters



Two Clusters

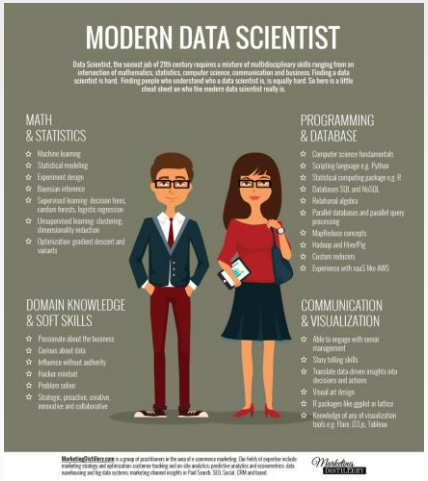


Four Clusters

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

The data science recipe

- Different ingredients needed
  - Data expert
    - Data processing, data structures
  - Data analyst
    - Data mining, statistics, machine learning
  - Visualization expert
    - Visual art design, storytelling skills
  - Domain expert
    - Provide understanding of the application domain
  - Business expert
    - Data driven decisions, new business models



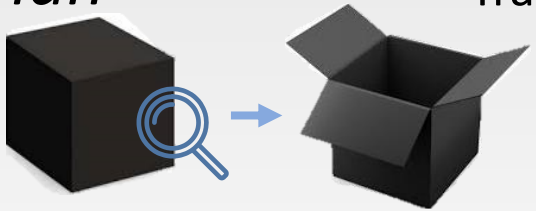
Some open issues

- Social impact of analysis is very important
  - Interpretability and transparency of the analysis process
  - Bias in algorithms and data

Interpretability in machine learning

"The ability to explain or to present in understandable terms to a human"

Trade-off Accuracy-Interpretability



Open the black box

- Model explanation: global understanding of how a model works
- Prediction explanation: local understanding of why a prediction is made

Interpretability

- Learned decision rule in pneumonia patients dataset from USA hospital

  history of asthma → lower chance of dying from pneumonia

- MD consider asthma as a serious risk factor for people who get pneumonia
- Analysis
  - asthmatics probably notice earlier the symptoms of pneumonia
  - a healthcare professional is going to provide earlier pneumonia diagnosis
  - as high-risk patients, they're going to get high-quality treatment sooner than other people

  ➡ asthmatics actually have almost half the chance of dying of non-asthmatics
- Using a black box model, this model issue would *never* have been uncovered

Algorithmic and data bias

- Task: predict likelihood of an individual committing a future crime
  - Risk scores used by US criminal justice system
- Scores computed from
  - Questions answered by the defendants
  - Information pulled by criminal records
- Race was not among the questions
  - … however other items may be correlated (e.g., poverty, joblessness)
- Software product flagged black defendants as future criminals more frequently than white defendants

➡ Training data was biased by a larger black defendant population

The database and data mining group


Daniele Apiletti


Elena Baralis


Luca Cagliero


Tania Cerquitelli


Silvia Chiusano


Paolo Garza


Danilo Giordano


Giuseppe Attanasio


Elena Daraio


Jacopo Fior


Flavio Giobergia


Moreno La Quatra


Andrea Pasini


Eliana Pastor


Francesco Ventura

…. and many more!

Thank you!