Data Mining

Overview

Raquel Sebastião

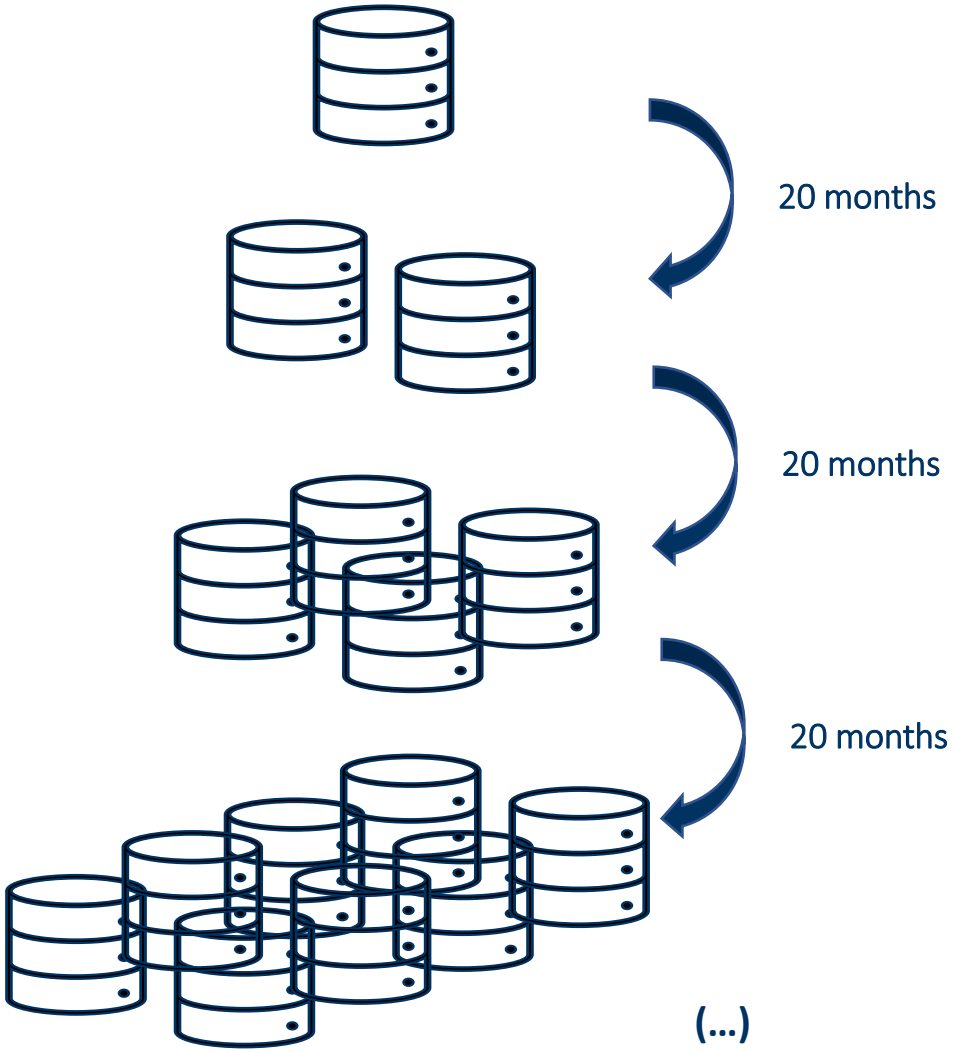
Departamento de Eletrónica, Telecomunicações e Informática

Universidade de Aveiro

raquel.sebastiao@ua.pt

2022/2023

Why data mining?

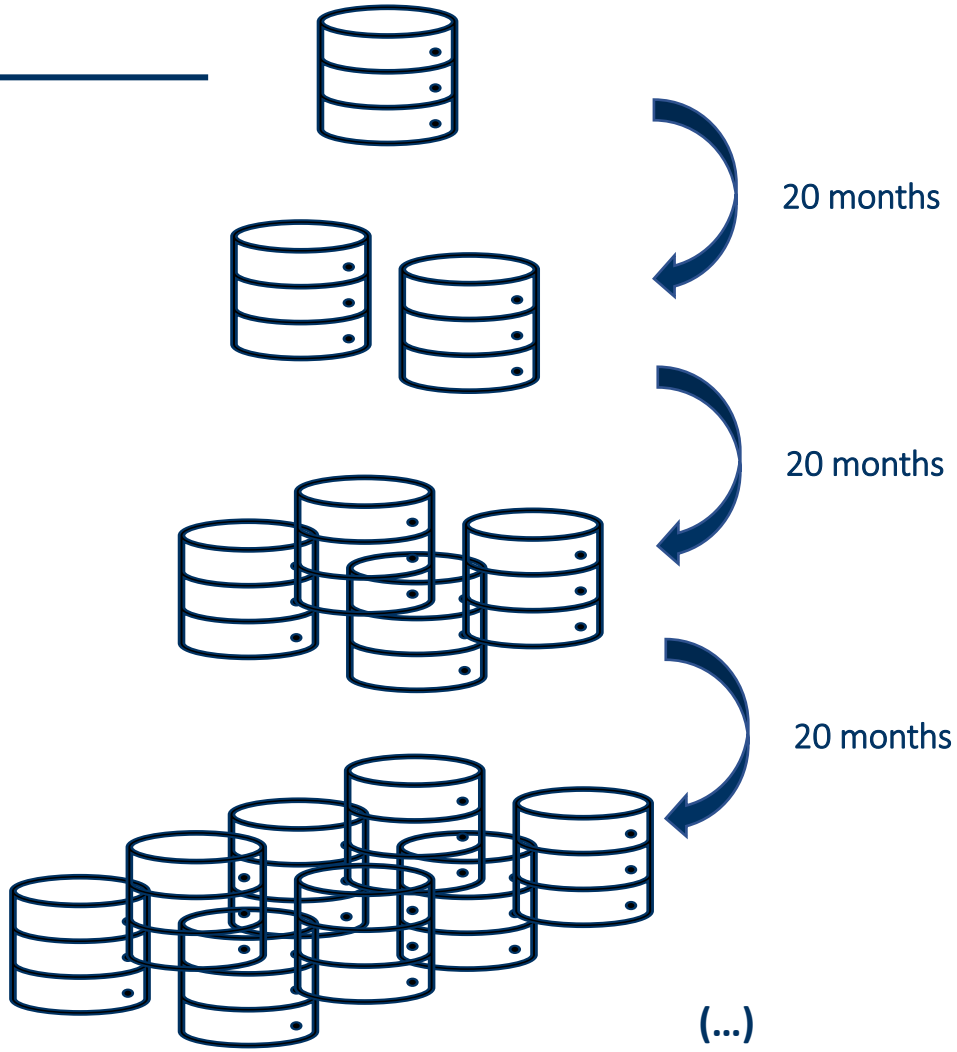


Overwhelmed with data

Why data mining?

"We are drowning in data,

but starving for knowledge."



Motivation

"Necessity is the mother of invention"
proverb (Plato)

The amount and type of data are ever-increasing

- Several **data collection** methods have been advanced
- Increase in the **storage capacity** and **computational power**

Data contain potentially useful (and interesting) information

Overwhelmed with data

- **Manual** inspection is almost impossible
- **Automatic** data analysis methods are required

From Data to Knowledge

Data

- Facts, numbers, or text that can be processed by a computer

Metadata

- Data about the data itself such as logical database design or data dictionary definitions

Information

- The patterns, associations, or relationships among all this data can provide information

Knowledge

- Information can be converted into knowledge about historical patterns and future trends

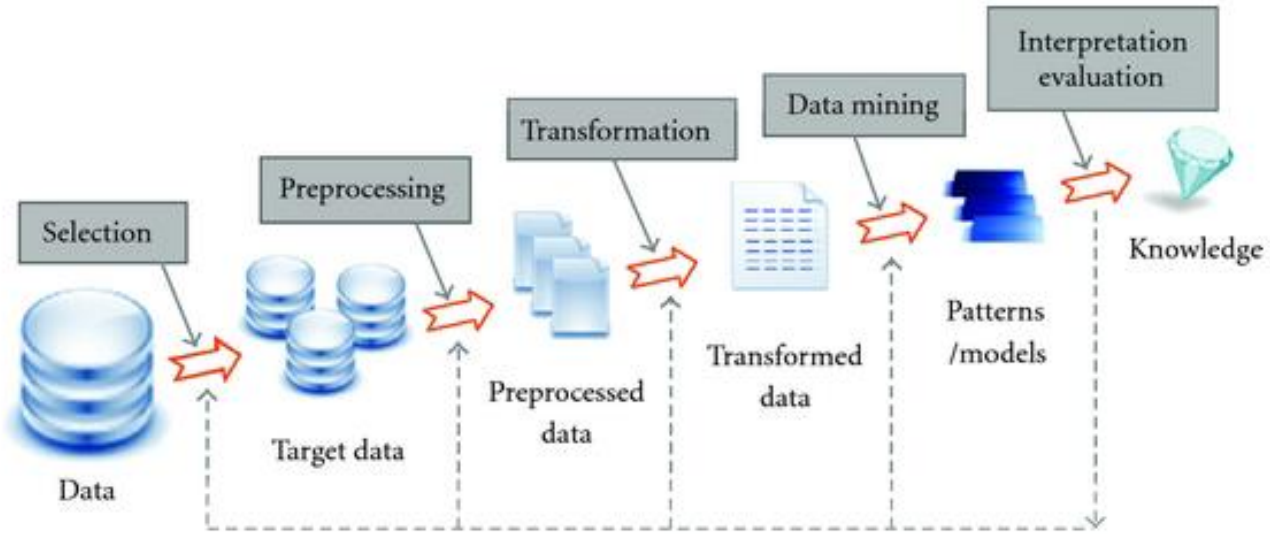
From Data to Knowledge

Criteria to assess Knowledge:

- correctness (probability, success in tests);
- generality (domain and conditions of validity);
- usefulness (relevance, predictive power);
- comprehensibility (simplicity, clarity, parsimony);
- novelty (previously unknown, unexpected)

Data Mining is the process of knowledge discovery from data!

Knowledge Discovery from Data (KDD)



Data Mining is the process of knowledge discovery from data!

Data mining: possible definitions

"is the process of automatically discovering useful information in large data repositories"

Introduction to Data Mining, Tan et al.

"is the process of discovering interesting patterns from massive amounts of data"

Data Mining: Concepts and Techniques, Han et al.

"is defined as the process of discovering patterns in data"

Data Mining: Practical Machine Learning Tools and Techniques, Witten et al.

"is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data"

Data Mining: The Textbook, Aggarwal

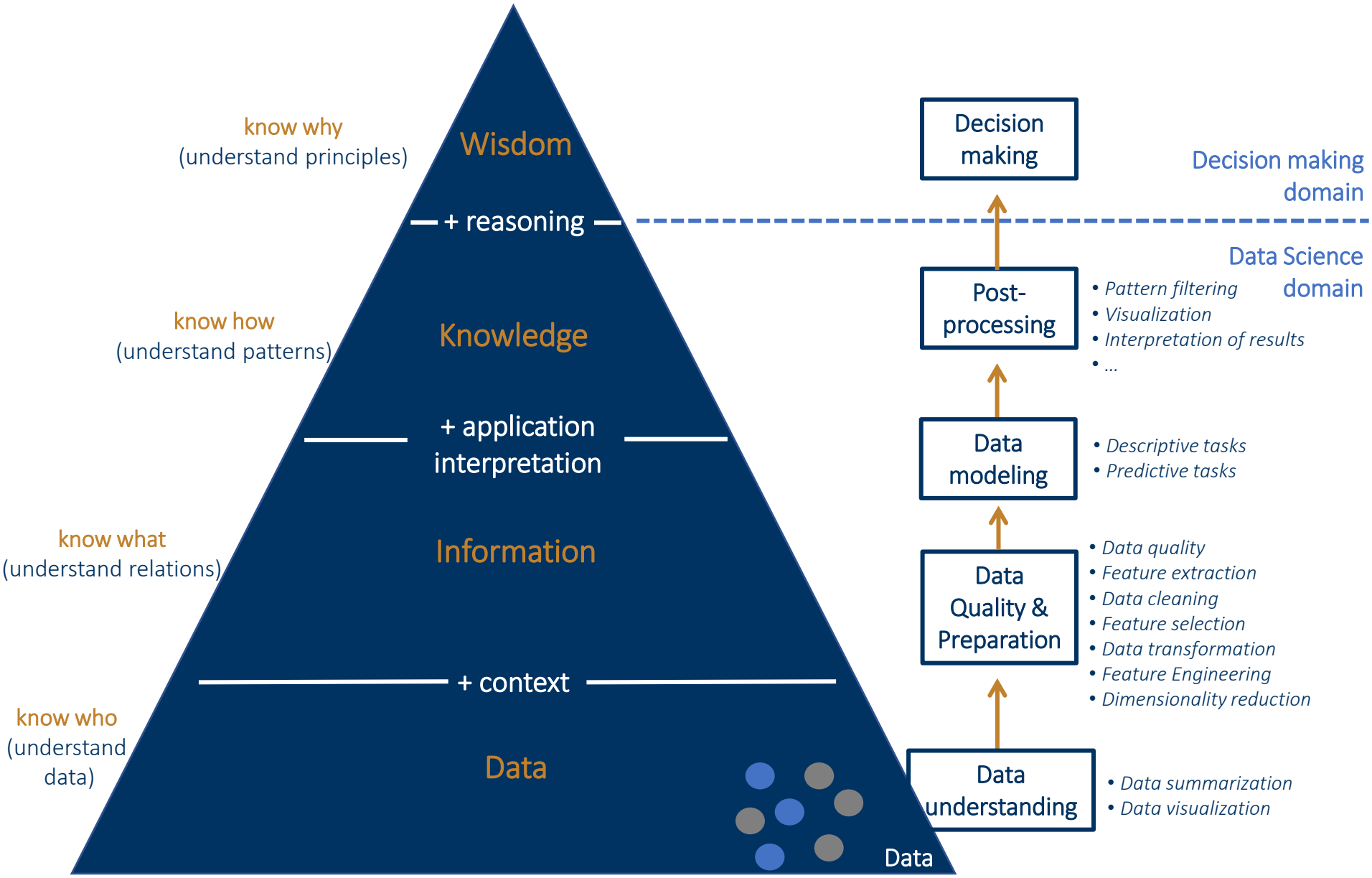
Humorous definition:

Data Mining, *noun* 1. Torturing the data until it confesses … and if you torture it enough, you can get it to confess to anything. ☺

ACM SIGKDD

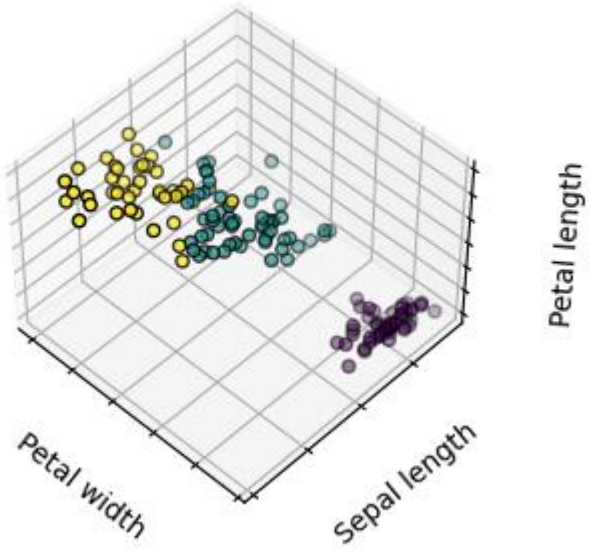
Data mining

Involves the manipulation of large amounts of data, usually stored in databases, in order to discover implicit, previously unknown, and potentially useful information that can be conveyed into knowledge


know why
(understand principles)
Wisdom
+ reasoning
know how
(understand patterns)
Knowledge
+ application
interpretation
know what
(understand relations)
Information
+ context
know who
(understand
data)
Data
Data
Decision
making
Decision making
domain
Data Science
domain
Post-
processing
• Pattern filtering
• Visualization
• Interpretation of results
• …
Data
modeling
• Descriptive tasks
• Predictive tasks
Data
Quality &
Preparation
• Data quality
• Feature extraction
• Data cleaning
• Feature selection
• Data transformation
• Feature Engineering
• Dimensionality reduction
Data
understanding
• Data summarization
• Data visualization

Clustering :example

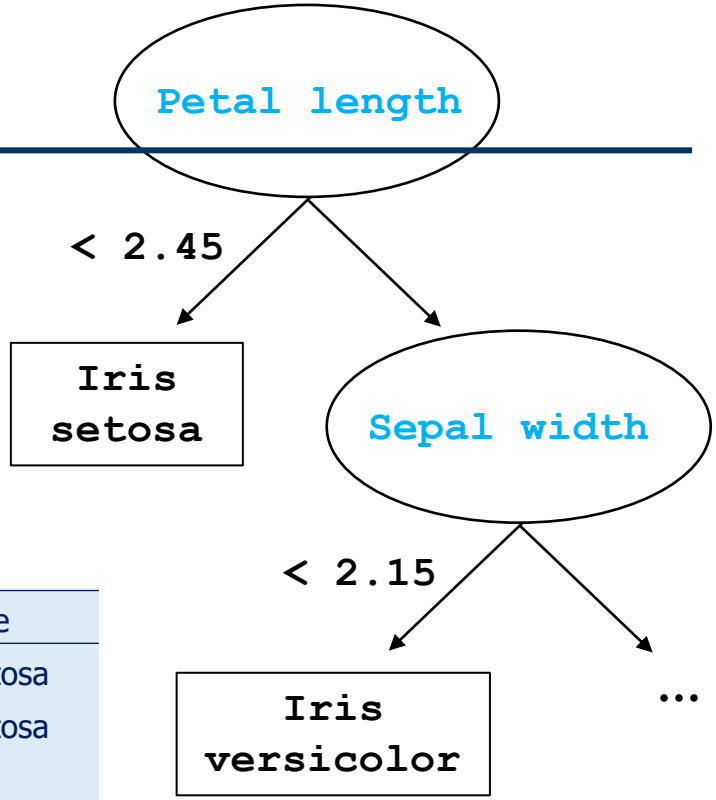
- Finding groups of items that are similar
- Clustering is *unsupervised*
  - The class of an example is not known
- Success often measured subjectively



	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

Classification: example

- Predicting the target/class
- Classification is *supervised*
  - The class of an example is known
- Success measured objectively



	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

Practical assignment: key issues

- **Presentation:**
  - Introduction to the problem and application domain, description of the dataset, and summarization of results, conclusions, and further research of the reference literature.

- **Project (and analysis of the report)**
  - Data Structure
    - what to measure? pre-processing steps? ...
  - Model Structure
    - what type of model(s) should we build? ...
  - Score Function
    - how to evaluate the obtained models? ...
  - Optimization and Search Method
    - how to search and optimize the models in the context of the selected structure? ...
  - Data Management Strategy
    - how to handle the data efficiently during model construction/evaluation? ...

Data mining: origins

Originally proposed to solve perceptual tasks like:

- Optical character recognition
- Face recognition
- Voice recognition
- …

which the **humans can perform** well, however there is **no mathematical model** to address the problems

Data mining: the role of machine learning

Learning computational models from examples

- WHY?
  - the representation of the problem is hard (or even impossible) to define with equations
- WHAT?
  - The free parameters of the model
- WHAT FOR?
  - New examples can be processed by the models to help in new decisions

Example: in medical image analysis is the classification of objects such as lesions into certain categories (e.g., abnormal or normal, tumor or non-tumor)

Data mining: main disciplines

Machine Learning:

- learning from examples to construct models
  - Computer Science Community

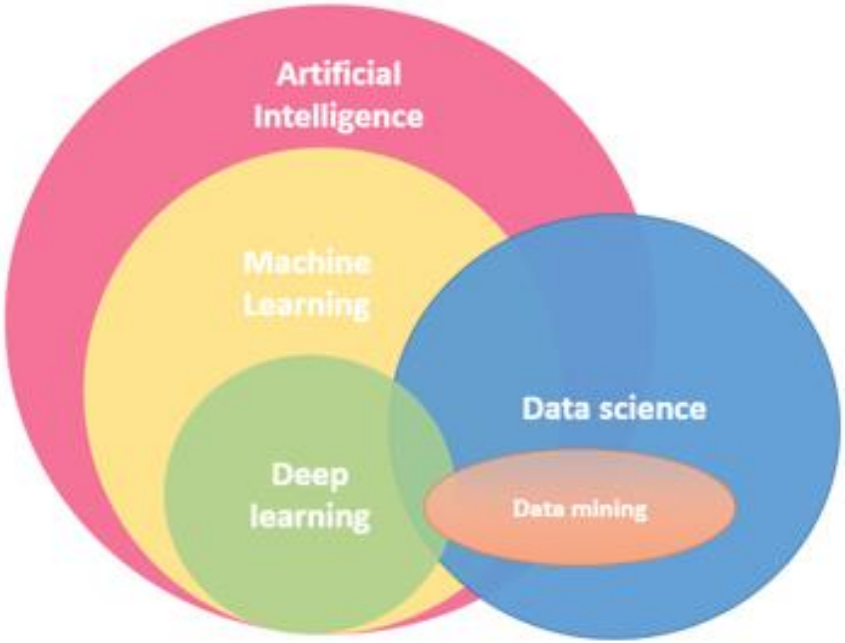
Pattern Recognition

- identifying patterns not necessarily with a learning phase
  - Electrical Engineering Community

Data Mining

- involves the manipulation of large amounts of data, usually stored in databases, in order to discover patterns

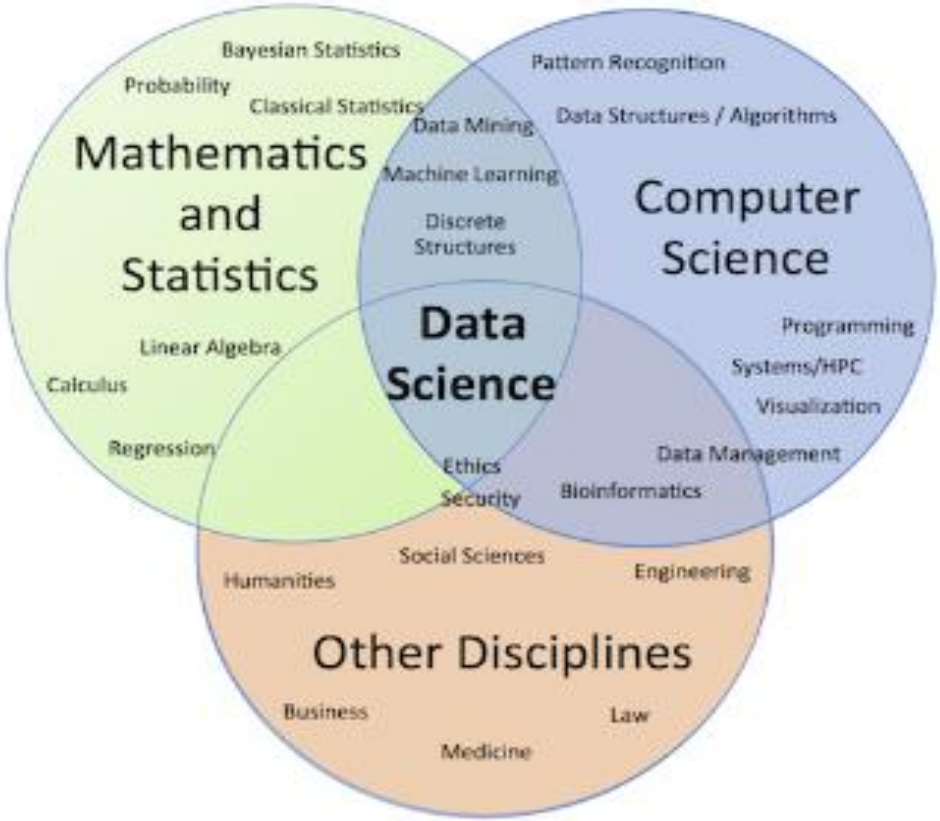
Generally, all use similar algorithms or methods



Data Science

Kulin, M.; Kazaz, T.; De Poorter, E.; Moerman, I.
A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer.
Electronics 2021, *10*, 318.
https://doi.org/10.3390/electronics10030318

Data Science



ACM Data Science Task Force
https://dstf.acm.org/

Data mining: tasks

Exploratory Data Analysis

- Summarization
- Visualization tools

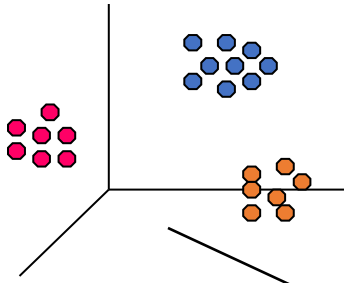
Descriptive tasks

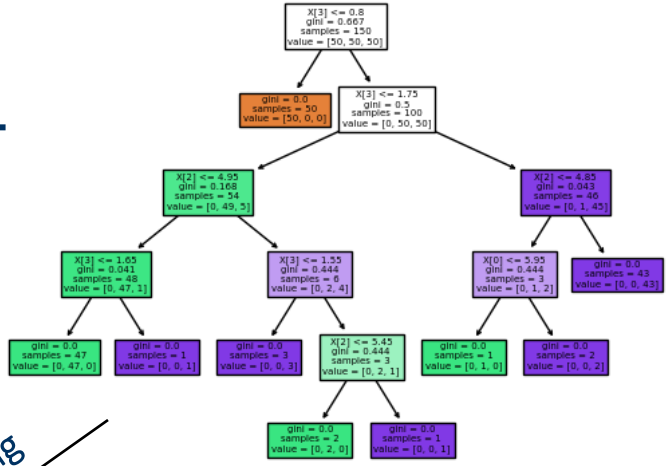
- Find human-interpretable patterns that describe the data.
  - Clustering
  - Association analysis
  - Anomaly detection

Predictive tasks

- Use some variables (features) to predict unknown or future values of other variables.
  - Classification
  - Regression

Data mining: tasks





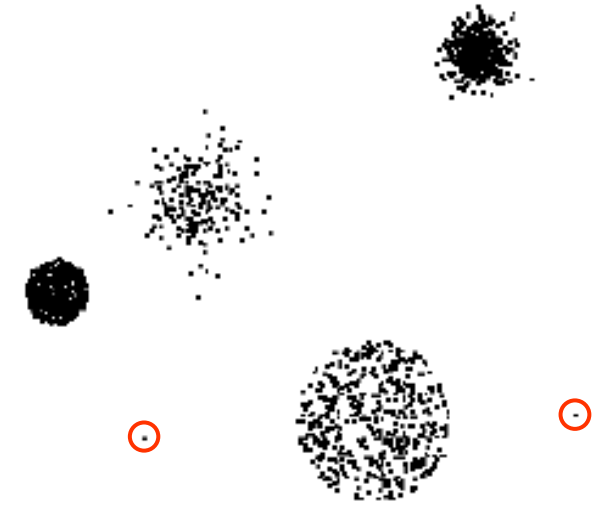
Clustering

Data

Predictive Modeling

Association Rules

Anomaly Detection



Classification vs. Association rules: examples

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

Classification rule:

- predicts value of a given attribute (the classification of an example)

```
If outlook = sunny and humidity = high
   then play = no
```

Association rule:

- predicts value of arbitrary attribute (or combination)

```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```

The role of domain knowledge: example

```
If leaf condition is normal
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

```
If leaf malformation is absent
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

- But in this domain, “leaf condition is normal” implies “leaf malformation is absent”!

Key issues in a Data Mining Project

Data Structure

- what to measure? pre-processing steps?

Model Structure

- what type of model(s) should we build?

Score Function

- how to evaluate the obtained models?

Optimization and Search Method

- how to search and optimize the models in the context of the selected structure?

Data Management Strategy

- how to handle the data efficiently during model construction/evaluation?

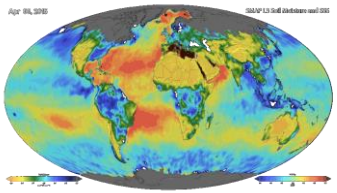
BIG data

E-Commerce

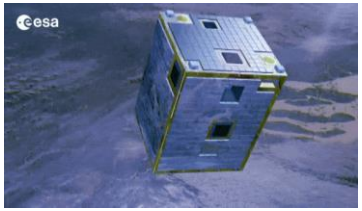
The amount and type of data are ever-increasing

- Several **data collection** methods have been advanced
- Increase in the **storage capacity** and **computational power**

Bio-informatics



Surface Temperature of Earth

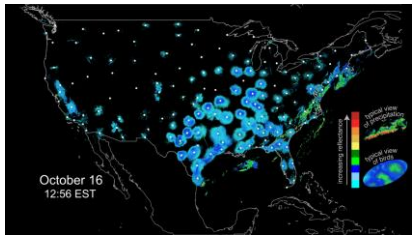


Sky Survey Data

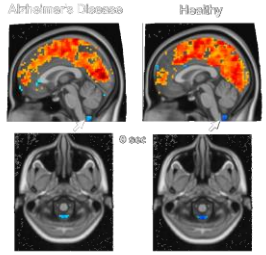
Data is useless!

convert it to useful information and into knowledge

Cyber Security



Bird migration



fMRI Data from Brain

Traffic Patterns

Data mining and Big data

The amount and type of data are ever-increasing

Big Data has **three dimensions** described by the Three V's Gandomi and Haider, 2015:

- Volume: massive, high dimensional, distributed data sets
- Velocity: generated at high-speed
- Variety: heterogeneous, complex

Data mining and Big data

Traditional techniques may be **unsuitable** as new as applications provide data that is:

- Large-scale
- High dimensional
- Complex
- Heterogeneous

A key **challenge** for data mining is to develop techniques that can cope with Big Data.

Data mining tasks: challenges

- Scalability: capacity of dealing with massive data sets (large number of objects);
- Dimensionality: capacity of dealing with lots of attributes/features for each object;
- Complex and Heterogeneous Data: Different type of attributes;
- Data Quality: Usually data does not result of a designed data collection as in traditional statistical experiment;
- Data Ownership and Distribution: data stored and owned by various organizations;
- Privacy Preservation: development of data mining has the capacity to compromise privacy in ways not previously possible;
- Streaming Data: extracting information of an ordered sequence of data records.

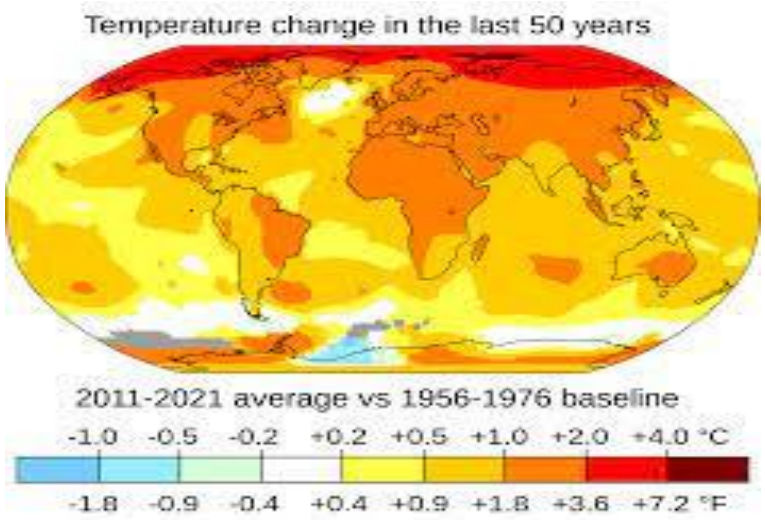
Data mining: applications

- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Banking, fraud analysis, stock market analysis
- Telecommunications
- Diagnosis of machine faults

Data mining: solving society's major problems?

Improving health care and reducing costs

Data mining: solving society's major problems?



Predicting the impact of climate change

Data mining: solving society's major problems?

Finding alternative/ green energy sources

Data mining: solving society's major problems?

Reducing hunger and poverty by increasing agriculture production

Summing up…

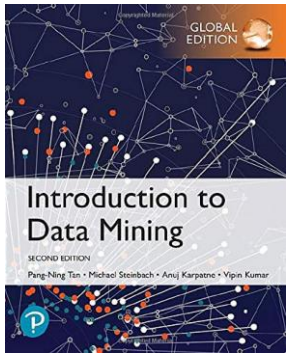
- **Data to be mined**
  - Database data, data warehouse, heterogeneous data, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, social data, information networks, …
- **Data mining tasks**
  - Exploratory Data Analysis
  - Descriptive vs. predictive
    - Clustering
    - Association analysis
    - Anomaly detection
    - Prediction (classification/regression)
- **Techniques**
  - Data-intensive, data warehouse (OLAP), pattern recognition, statistics, machine learning, visualization, …
- **Applications**
  - Retail, telecommunications, banking, fraud analysis, bio-data mining, stock market analysis, text mining, web mining, fault diagnosis, …

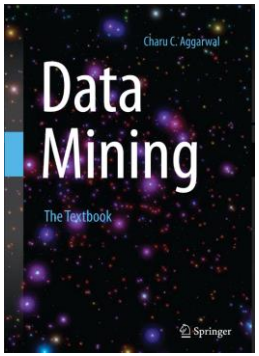
Bibliography

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, *Pearson*, 2019

Data Mining, the Textbook, Charu C. Aggarwal, *Springer*, 2015

Sebastian Raschka, **Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning**. https://arxiv.org/abs/1811.12808





Some online resources

- SIGKDD
- Data Science, Machine Learning, AI & Analytics – Kdnuggets
- Category: Machine Learning - VideoLectures.NET
- UCI Machine Learning Repository
- Kaggle: Your Machine Learning and Data Science Community
- Dataset Search (google.com)

Lectures and Laboratories

Weekly session (3h)

- Exposition (lecture): Slides will be available
- Paper and pencil: Solving a couple of exercises from the booklet exercises
- Programming: Jupyter Notebooks to practice the use of different packages

Mandatory tools

- Jupyter Environment
- Python Programming (NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Mlxtend)
- ANACONDA could be used to manage all the facilities

Assessment

- Practical assignment (groups with 3 students)
  - Presentation & project
    - Presentation of the application and the available data set: **10%**
    - Project Notebook. Using the data set developing an exploratory data mining project: **25%**
- Participation (lab exercises, lectures and discussion): **10%**
- Written exam: **55%**

Practical assignment: groups

Please, send an email (**up to 5th October**) with the **names of the group members**.