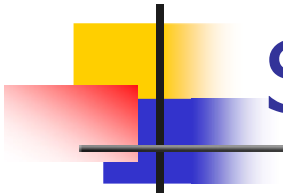
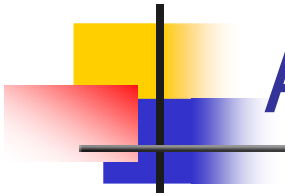
Introductory Background

(1IK172 Introduction to Data Analytics)



Summary

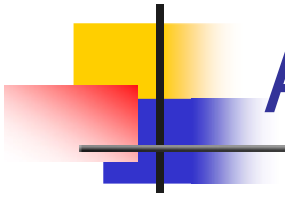
- About Artificial Intelligence
- About the course
- Analytics
- Big data, small data & data science
- Big data architectures
- What is data?
- Taxonomy on data analytics
- Methodologies for projects of data analytics
- A project on data analytics
- Credits
- Further Reading



About Artificial Intelligence

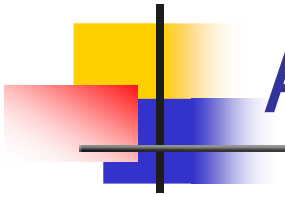
"Any sufficiently advanced technology is indistinguishable from magic"

- Arthur C. Clarke



About Artificial Intelligence

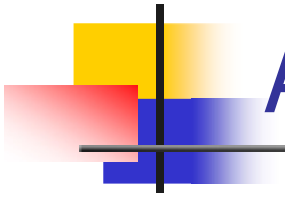
- Intelligence
- Artificial Narrow Intelligence
- Artificial Generic Intelligence



About Artificial Intelligence

What is behind the magic?

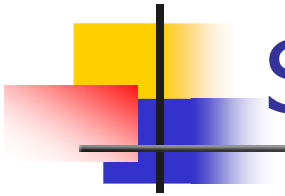
- Lots of data
- Lot of computing power
- Decades of research
- Parameter optimalization



About the course

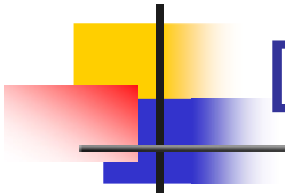
Goals:

- Have a good general understanding of the concepts of data analytics
- Familiarity with most common statistical/ML models
- Have practical knowledge about the most common ML algorithms
- Know where to move on for improving



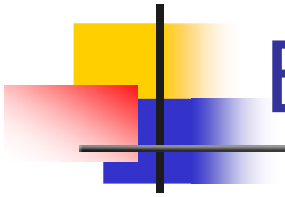
Success stories

- Content, item recommender systems
- Smart services
- Speech recognition
- Autonomous trading
- Self-driving cars
- Medical anomaly detection
- Augmented Reality applications



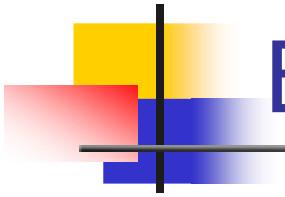
Data Analytics

- Definition
  - The science that analyze crude data to extract useful knowledge (patterns) from them
- Includes
  - Statistics -> Inductive learning
  - Reproducing the human behaviour: artificial intelligence
  - Learning from databases
  - Machine Learning
  - Data Mining



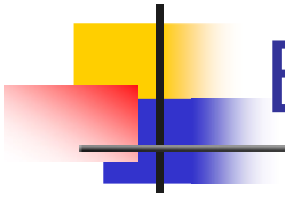
Big data, small data and data science

- Big Data
  - **V**olume: How to store large amounts of data whose structure is not known in advance?
  - **V**elocity: How to guarantee that we can process the incoming data before the new data arrives?
  - **V**ariety: how to use together information arriving in different moments, different granularities, different sources?
  - Others: Veracity, Validity, Volatility, Variability, ...



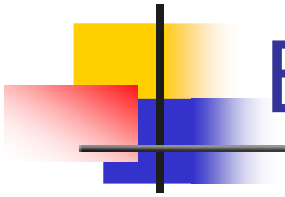
Big data, small data and data science

- Small Data
  - Data set whose volume and format allows its processing and analysis by a person or a small organization



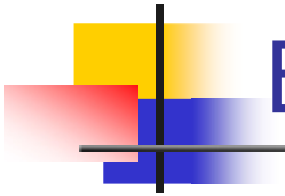
Big data, small data and data science

- Data Science
  - Data science extracts meaningful and useful knowledge from data, with the support of suitable technologies
  - It has a strong relation to analytics and data mining
  - Data science goes beyond data mining by providing a knowledge extraction framework that also includes statistics and visualization



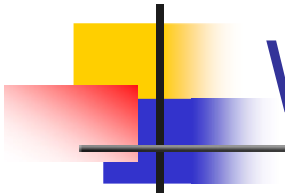
Big data architectures

- Distributed systems
  - the most popular big data processing technique using clusters of computers is MapReduce
    - Hadoop: is its most famous implementation
    - Is a programming model
    - Has two steps: map & reduce
    - Divide the data into chunks and split them by the computers in the cluster



Big data architectures

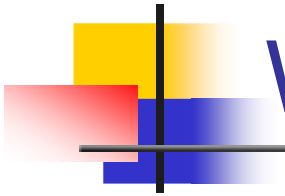
- Expected characteristics of distributed systems
  - resource sharing
  - openness
  - concurrency
  - scalability
  - fault tolerance
  - transparency



What is data?

- Tabular data
  - Rows: represent instances also named objects; an instance per row
  - Columns: represent attributes also named features; an attribute per column
- Instances
  - Are examples of the concept we want to characterize
- Attributes
  - Are characteristics present in the instances

Name	Age	Educational level	Company
Andrew	55	1	Good
Bernhard	43	2	Good
Carolina	37	5	Bad
Dennis	82	3	Good
Eve	23	3.2	Bad
Fred	46	5	Good
Gwyneth	38	4.2	Bad
Hayden	50	4	Bad
Irene	29	4.5	Bad
James	42	4.1	Good
Kevin	35	4.5	Bad
Lea	38	2.5	Good
Marcus	31	4.8	Bad
Nigel	71	2.3	Good

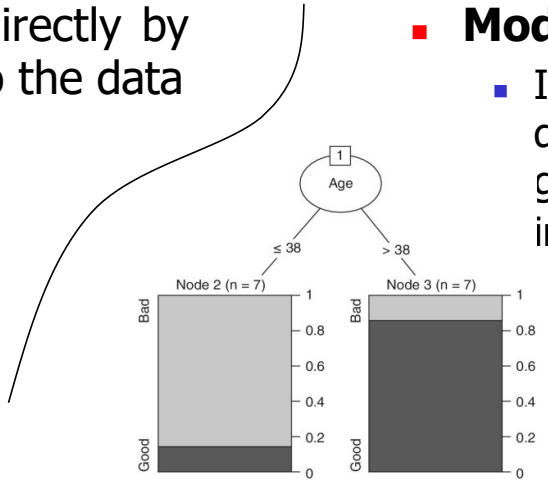


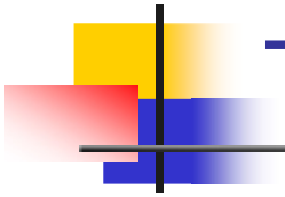
What is data?

- Relational data
  - There are data that are not possible to represent in a single table
  - Data sets represented by several tables, making clear the relations between these tables, are named relational data sets
  - Data sets represented by a single table but where there are relations between their instances, are also named relational data sets

Taxonomy on data analytics

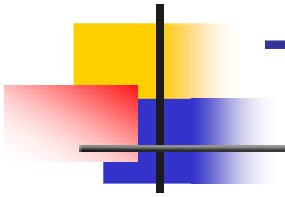
- Descriptive analytics
  - Summarize or condensate data to extract patterns
  - The result of a given method or technique is obtained directly by applying an algorithm to the data
- Predictive analytics
  - The result of applying an algorithm on a predictive method to given data, is typically a model
  - **Model**
    - Is a generalization obtained from data that can be used afterwards to generate predictions for new given instances





Taxonomy on data analytics

- Algorithm
  - A self-contained step-by-step set of instructions easily understandable by humans, allowing the implementation of a given method to an arbitrary programming language
- Method or technique
  - Is a systematic procedure that allows to achieve an intended goal



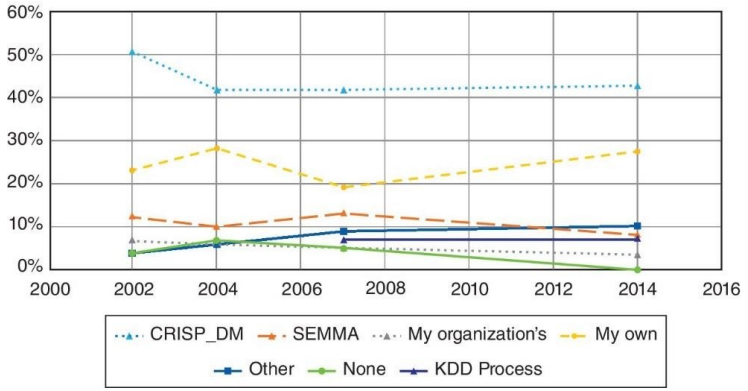
Taxonomy on data analytics

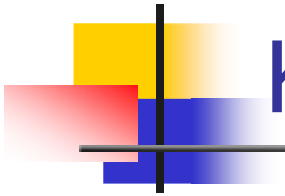
- Hyper-parameters
  - The values of the hyper-parameters are set by the user, or some external optimization method
    - E.g. the number of clusters in k-means
- Parameters
  - The parameter values are model parameters whose values are set by a modeling or learning algorithm in its internal procedure
    - E.g. the slope parameter of multivariate linear regression

Methodologies for projects of data analytics

- The KDD process: a nine-step methodology
- The CRoss-Industry Standard Process for Data Mining (CRISP-DM): a six-step methodology
- SEMMA - Sample, Explore, Modify, Model and Assess

- Surveys conducted in 2002, 2004, 2007 and 2014 by kdnuggets on the use of planning and developing methodologies for projects on data analytics

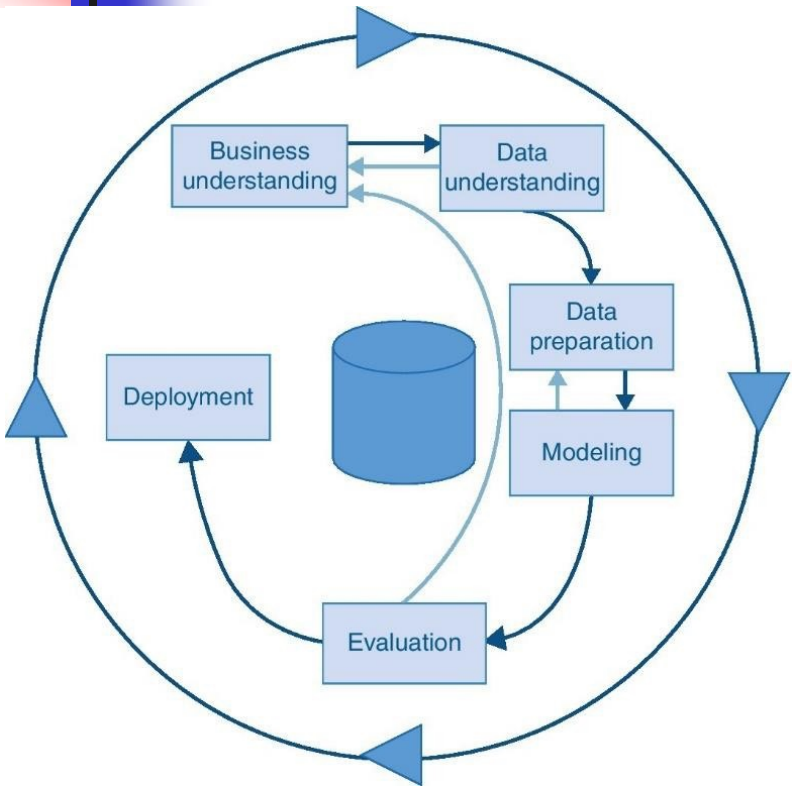


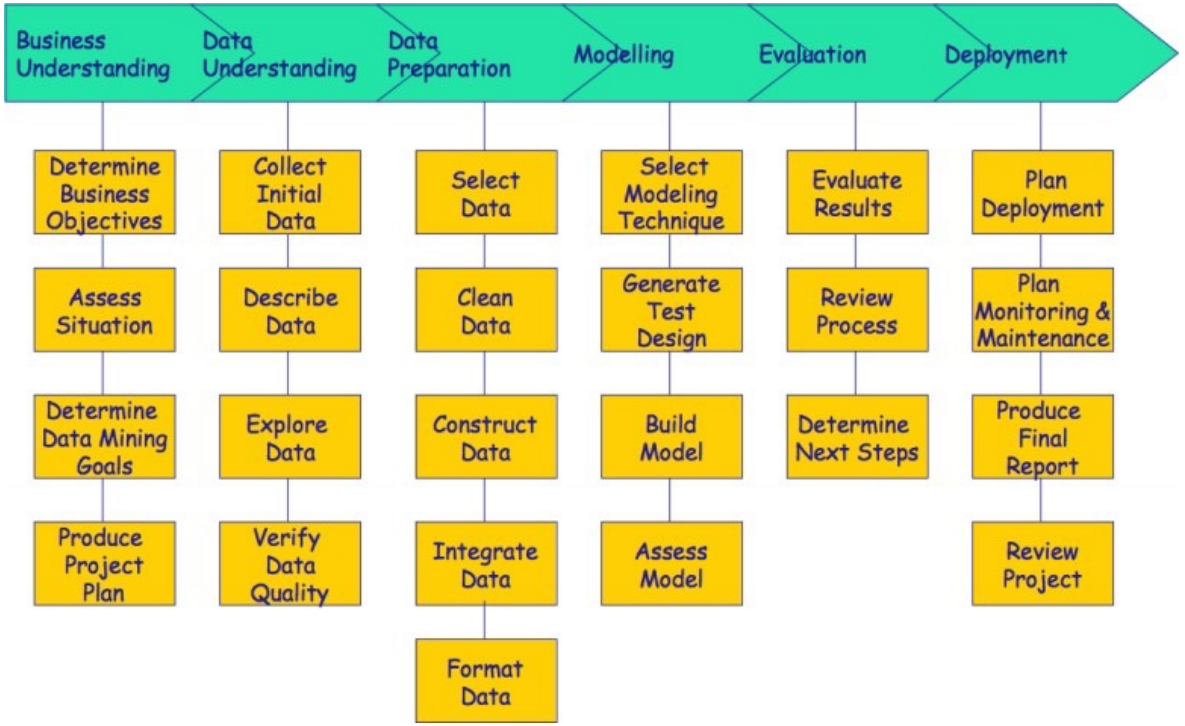


KDD Method

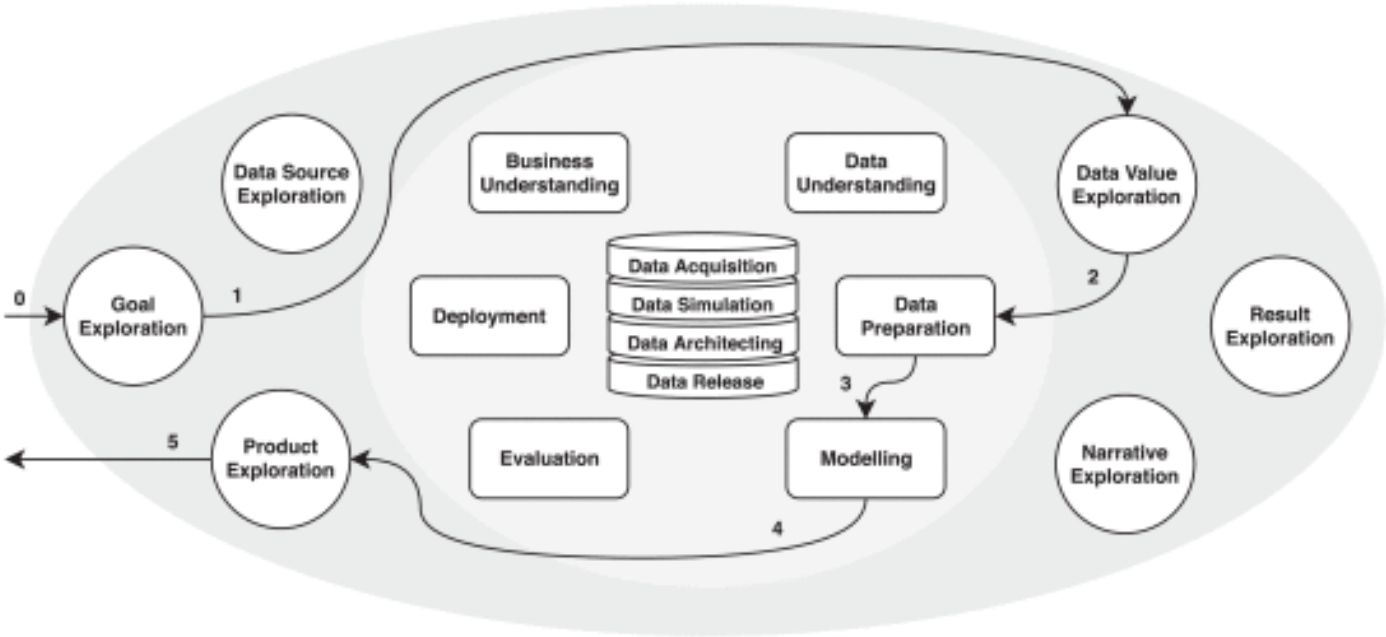
1. Developing an understanding of the domain, goals
2. Creating a target data set
3. Data cleaning and preprocessing
   - Removal of noise or outliers
4. Data reduction and projection
5. Choosing the data mining task
   - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithm
7. Data mining
8. Interpreting mined patterns
9. Consolidating discovered knowledge

The CRISP-DM methodology

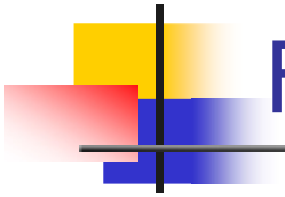




Example trajectory through a data science project



https://ieeexplore.ieee.org/document/8943998



Further reading

- **A General Introduction to Data Analytics**

  by João Mendes Moreira, André C. P. L. F. de Carvalho and Tomáš Horváth

- **Introduction to Data Mining**

  by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar

  https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item3

Thank You!