Department of Mathematics
University of Calabria

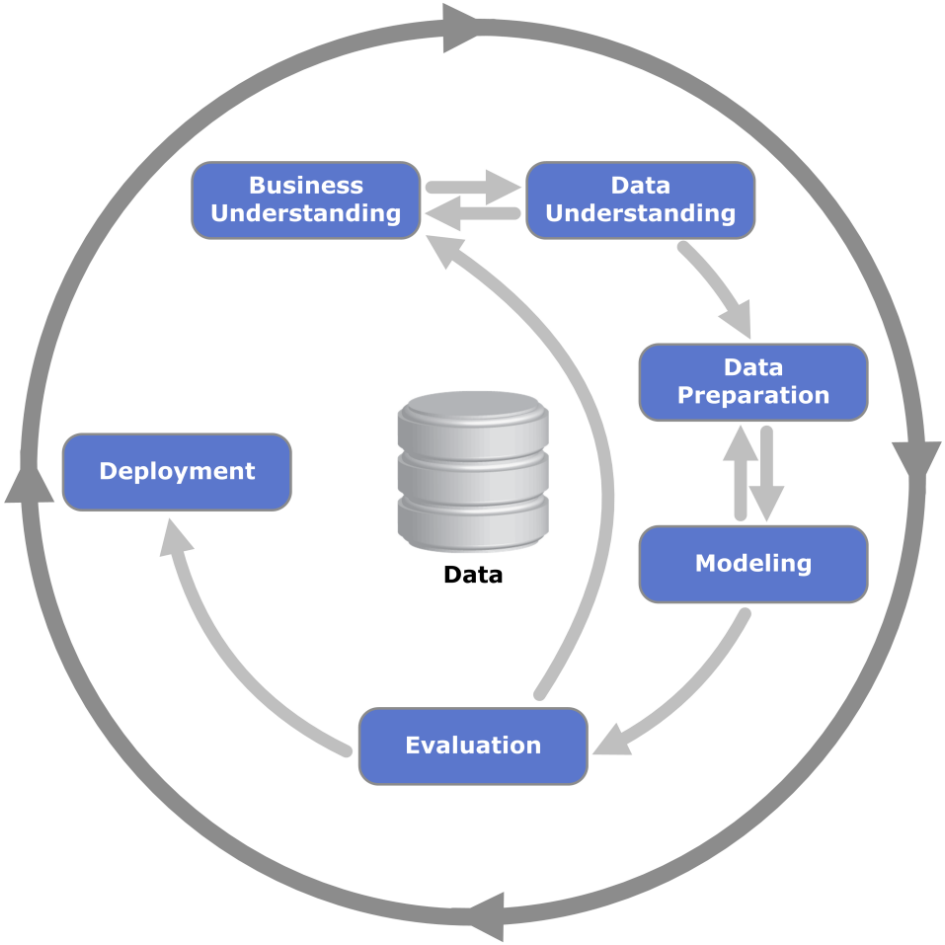
Business Intelligence and Analytics

Data Mining

Case study: Drug

Ph.D. Ettore Ritacco

The Knowledge Discovery Process (CRISP-DM)



The software

- Programming Language:
  - Python3.7 (http://www.python.org)

- How to install on:
  - Windows (http://www.youtube.com/watch?v=ndrCfBJkkvE)
  - Linux (http://www.youtube.com/watch?v=ndrCfBJkkvE)
  - Mac OS (http://www.youtube.com/watch?v=8BiYGIDCvvA)

The software

- Modules (packages) to install/update:
  - SciPy library (http://www.scipy.org/scipylib/index.html)
    - *The SciPy* library is one of the core packages that make up the *SciPy stack*. It provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization
  - NumPy (http://www.numpy.org)
    - *NumPy* is the fundamental package for scientific computing with Python
  - Matplotlib (http://matplotlib.org)
    - *Matplotlib* is a Python plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms
  - Pandas (http://pandas.pydata.org)
    - *Pandas* is a set of easy-to-use data structures and data analysis tools for the Python
  - SciKit-Learn (http://scikit-learn.org)
    - *SciKit-Learn* is a repository rich of simple and efficient tools for data mining and data analysis

The software

- How to install modules:
  - Windows (http://www.youtube.com/watch?v=FKwicZF7xNE)
  - Linux (http://www.youtube.com/watch?v=UKXx4e9PotI)
  - Mac OS (http://www.youtube.com/watch?v=q_3doIUZTFg)

The software

- Integrated development environment (IDE):
  - PyCharm (http://www.jetbrains.com/pycharm/)
  - How to install:
    - Windows (http://www.youtube.com/watch?v=SZUNUB6nz3g)
    - Linux (http://www.youtube.com/watch?v=cVROiVgR_qg)
    - Mac OS (http://www.youtube.com/watch?v=mDqxeCqVsOg)
- Sharing server: Jupyter Notebook (http://jupyter.org)
  - *The Jupyter Notebook* is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
  - How to install:
    - Windows (http://www.youtube.com/watch?v=5NU6w5VhmMc)
    - Linux (http://www.youtube.com/watch?v=dpQ9yKn0Y1s)
    - Mac OS (http://www.youtube.com/watch?v=HW29067qVWk)

Business Understanding

- **Scenario**:
  - A medical division collected some data from its patients
  - All the target patients contracted the same disease
  - The therapy consists in 5 different and exclusive cures
    - Each cure depends on the patients' conditions
- **Goal**:
  - Define an automatic procedure for the cure assignment

Data Understanding

Attribute	Description
Instance_number	Incremental tuple ID (INTEGER)
ID	Patient's ID (INTEGER)
Age	Patient's age (INTEGER)
Sex	Patient's gender: F or M
BP	Blood Pressure: HIGH, NORMAL or LOW
Cholesterol	Concentration of cholesterol in the blood: NORMAL or HIGH
Na	Concentration of sodium in the blood (REAL)
K	Concentration of potassium in the blood (REAL)
Drug	The chosen cure: drugY, drugC, drugX, drugA, drugB