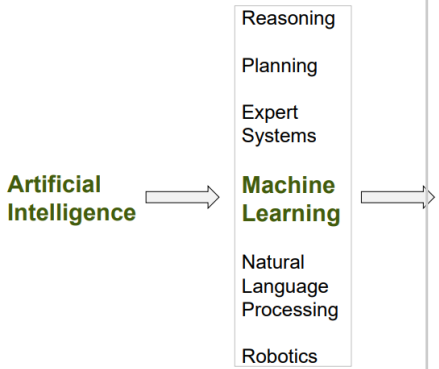
Machine Learning and Data Mining

Jil Zerndt
FS 2025

01 - Intro



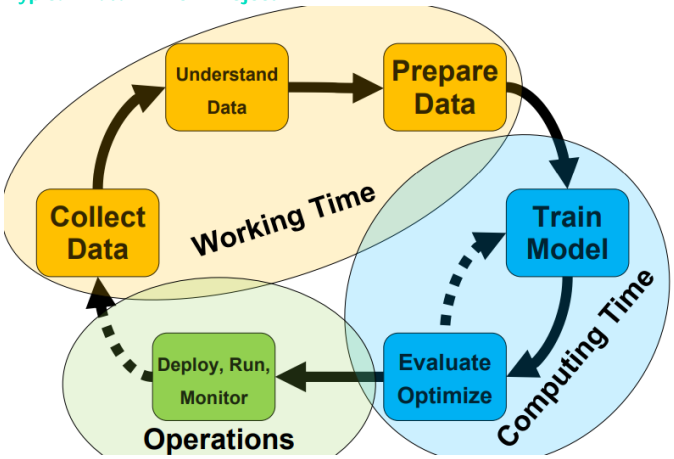
What's behind the Magic?

Data Processing

Learning Objectives:
- Understand fundamental importance of data preprocessing
- Know basic algorithms for data cleaning, (near) duplicate detection and filling missing values

Data

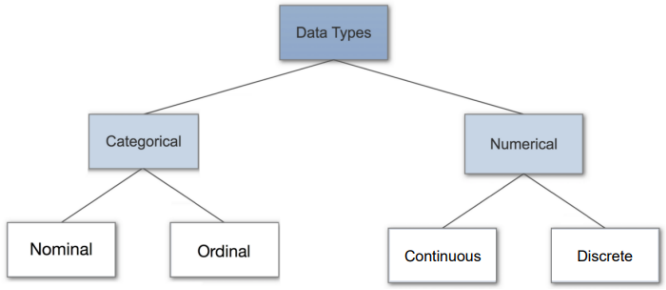
Typical Data Driven Project



Data has many sources, e.g.:
- sensor data
- survey data
- simulation data
- social media data
- textual data
- financial data
- multimedia data
- ERP systems data

Independent of the data source, each data point has a data type

Data Types



Nominal Data
- Nominal scales are used for **labelling** variables, without any quantitative value
- No numerical significance
- Nominal data has no order
- Scales could simply be called labels
- Examples: gender, hair colour, race, marital status

Ordinal Data
- Represents **discrete and ordered** units
- Nearly the same as nominal data, but **order matters**
- No distance between the different categories
- Examples: military rank, star rating, education level

Discrete Numeric Data
- Represents items that can be **counted**
- Values may go from 0, 1, 2, on to infinity (making it countably infinite)
- Examples: number of persons in a room, number of "headsïn 60 coin flips, time elapsed in minutes

Continuous Numeric Data
- Also known as **interval data**
- Often measurements
- Possible values **cannot be counted** and can only be described using intervals on the real number line
- Examples: temperature, weight, height, time, ...