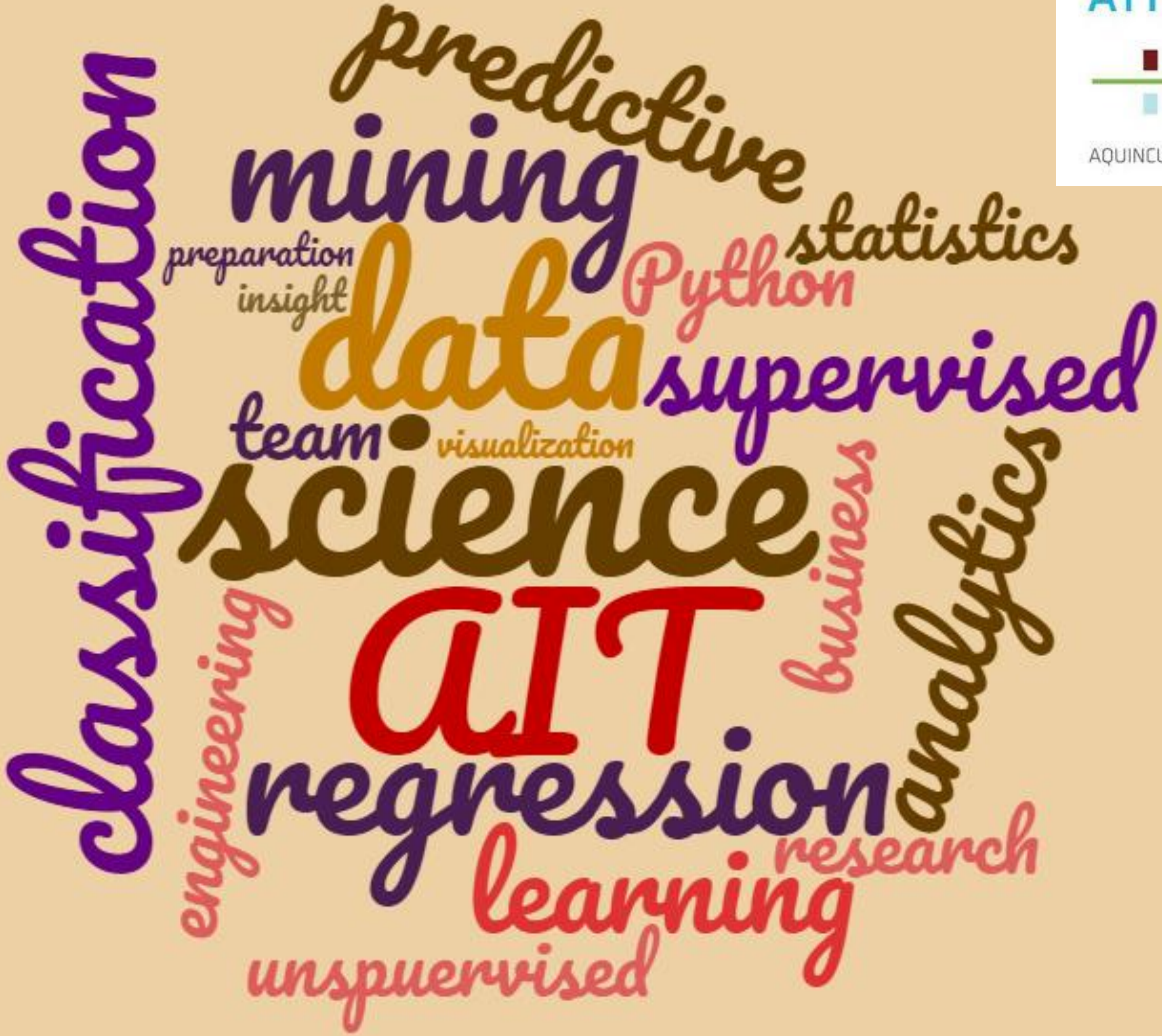
# Data Science

AIT-BUDAPEST
AQUINCUM INSTITUTE OF TECHNOLOGY




Dr. Roland Molontay

February 10, 2023
Process of data science

# Schedule of the semester

| | *Monday midnight* | *Tuesday class* | *Friday class* |
|---|---|---|---|
| *W1 (02/06)* | | | |
| *W2 (02/13)* | | HW1 out | |
| *W3 (02/20)* | | | |
| *W4 (02/27)* | HW1 deadline | HW2 out | |
| *W5 (03/06)* | PROJECT PLAN | | |
| *W6 (03/13)* | HW2 deadline | HW3 out | |
| *W7 (03/20)* | | | MIDTERM |
| SPRING BREAK | | SPRING BREAK | SPRING BREAK |
| *W8 (04/03)* | HW3 deadline | HW4 out | GOOD FRIDAY |
| *W9 (04/10)* | MILESTONE 1 | | |
| *W10 (04/17)* | HW4 deadline | | |
| *W11 (04/24)* | | | |
| *W12 (05/01)* | MILESTONE 2 | | |
| *W13 (05/08)* | | | |
| *W14 (05/15)* | | FINAL | PROJECT presentations |
| *W15 (05/22)* | | PROJECT presentations | |


# Reminder

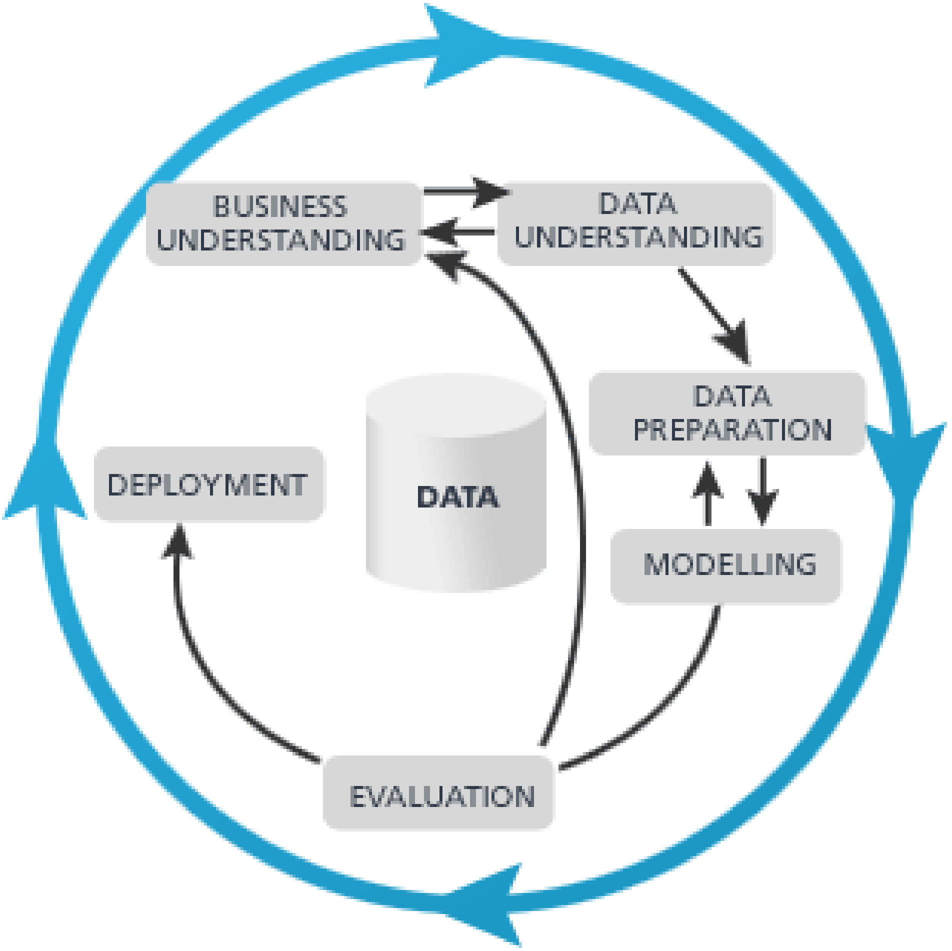
- Please refresh Python and download Jupyter Ipython notebook with Anaconda distribution
  - See the last slides of Lecture 01

# Process for data mining / data science

- CRISP-DM: **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
  - A technical standard with is own limitations but worth following
  - Back and forth effect
  - Cyclic

CRISP-DM

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING



# BU - Business understanding

- What is the aim of the project?
- What is its business relevance?
- What is the research question?
- How can it be translated to a data science question?

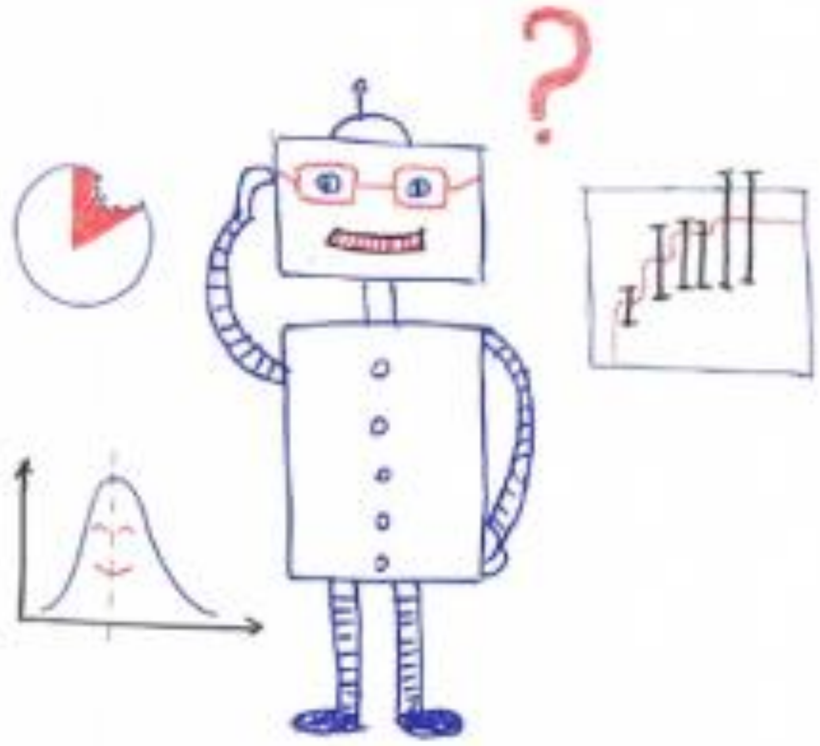
# DU - Data understanding

What data do we have?

Can we collect more data?

What is the quality of the data?

What do the features mean?



# Dataset

- Everything that carries information, and we would like to extract insights from
- In the simplest case the data is structured, i.e., it is like a table / data frame
  - Rows: records, observations, data points, instances
  - Columns: attributes, features
  - A record is described by the values of the attributes in its row
- The data can be inherently unstructured but in many cases we pursue to make it structured

# Representation of data

- Rows: record, object, data point, observation, entity, representative, item
- Columns: attribute, feature, dimension
  - Regarding regression, also called: explanatory variable, independent variable
- Target variable/output (for supervised learning):
  - For classification problems: label, class
  - For regression problems: response variable, dependent variable

attributes

records

| *Tid* | **Refund** | **Marital Status** | **Taxable Income** | **Cheat** |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Attribute types

- Continuous: real-valued (in most cases it is also considered to be „continuous” if it can take countably infinitely many values)
  - E.g.: temperature, height, weight
- Discrete: can take finitely many vales (sometimes variables with countably infinitely many possible values also)
  - Usually represented with integer values or category names
  - E.g.: ZIP code, marital status, (quantity)
- Binary: a special discrete attribute – possible values 0 and 1
  - Sometimes has asymmetric meaning: 0 may mean that something is not true, something is missing
  - Sometimes they can be found in sparse data matrices where the vast majority of the elements are 0
    - E.g.document-term matrices
    - Sparse data structures need special methods

# Attribute types – another partition

- Categorical / nominal variables
  - E.g. gender, marital status, place of birth, got treatment?, is overweight?
  - Reasonable operations: frequencies, mode
- Ordinal variables
  - May seem to be categorial, but can be ordered in a quantitative manner
  - E.g. stages (inchoative, advanced), military ranks (admiral, captain, commander), letter grades
  - Reasonable operations: median (but average is not), percentile, rank-correlation
- Quantitative (numerical) variables
  - Interval variables
    - The numerical values show both the ordinal relationship and the extent of deviation
    - E.g.: temperature (°C, °F), IQ score
    - Reasonable operations: average, difference, variance, correlation
  - Ratio variables
    - They have all the properties of an interval variable, and also have a clear definition of 0.0 (none of that variable)
    - E.g.: temperature (°K), height, weight, pieces
    - Reasonable operations: any operations that are defined for real numbers

# What to compute?

| OK to compute.... | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| frequency distribution. | Yes | Yes | Yes | Yes |
| median and percentiles. | No | Yes | Yes | Yes |
| add or subtract. | No | No | Yes | Yes |
| mean, standard deviation | No | No | Yes | Yes |
| ratio, | No | No | No | Yes |

Determine the type of the following attributes in two ways:
1: Continuous, discrete, binary? 2: Nominal, ordinal, quantitative (interval, ratio)?

1. Altitude
2. Total number of rooms in a hotel
3. Military ranks
4. Distance from the center of Heroes Square
5. International Standard Book Number (ISBN)
6. Degree: measurement of plain angle (between 0 and 360)
7. Degree of transparency: transparent, translucent, opaque
8. Cloakroom ticket numbers
9. Grades (from F to A+)
10. Medals (bronze, silver, gold)
11. Sex (male, female)
12. Age (in years)
13. pH (acidity or basicity of an aqueous solution)

# Data exploration

- What are the important features? Are there any interesting relations or redundancy?
- Are there any apparent problems with the data that need action?
  - Scaling, missing data, outliers
- Are there patterns that can be recognized using data visualization?

- Methods:
  - Summary statistics / descriptive statistics
  - Simple data viszalization, plots

# Summary statistics

- Purpose: to summarize the variables with numerical values
  - Easy to compute and informative
  - What are the typical values, how scattered are they, what are the frequencies?
  - They can be queried by a simple command in any programs
- Categorical variables: frequencies
- Numerical variables:
  - Percentiles: indicating the value below which a given percentage of observations in a group of observations (e.g. values in a column) falls, e.g. *p*-percentile denotes the $x_p$ value below which *p*% of the observations may be found
    - Usually considered values: min, 25, 50, 75, max
  - Mean: arithmetical average of the values: $mean(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$
    - Sensitive to outliers median is more robust
  - Median: the value separating the higher half from the lower half of a data sample
    - Similar to the 50-percentile but not the same

$$median(x) = \begin{cases} x_{r+1} & \text{if } m = 2r+1 \\ \frac{1}{2}(x_r + x_{r+1}) & \text{if } m = 2r \end{cases}$$

# Values describing deviation

- Range: what is the range of the possible values: *max - min*
- Sample variance:
  - Sensitive to outliers

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2$$

  - Standard deviation: root of the variance
- Average absolute deviation:

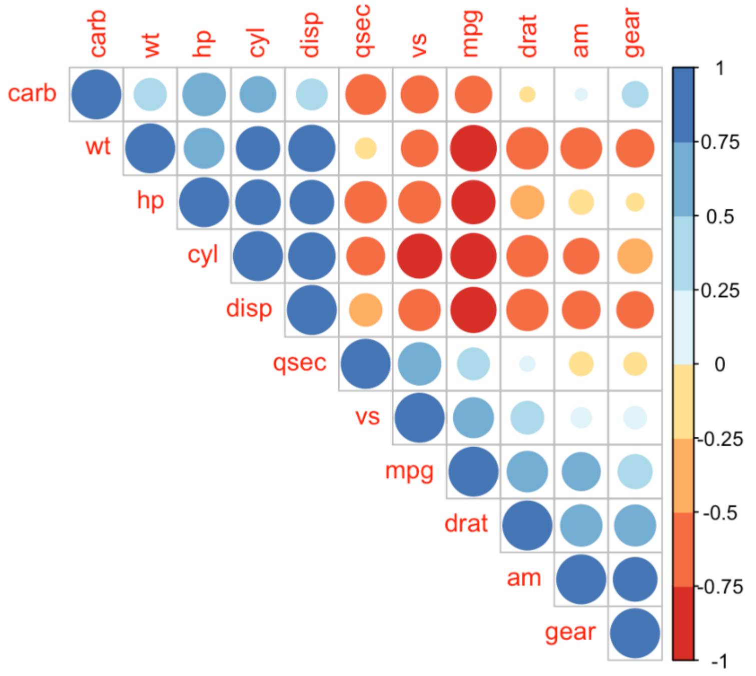
$$\frac{1}{m}\sum_{i=1}^{m}\left|X_i - \overline{X}\right|$$

# Covariance and correlation

- Sample covariance
  between values of $j$th and $k$th columns (attributes)

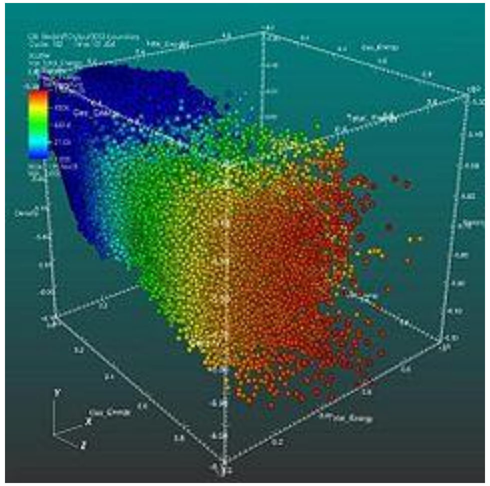
$$q_{jk} = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_{ij} - \overline{X_j}\right)\left(X_{ik} - \overline{X_k}\right)$$

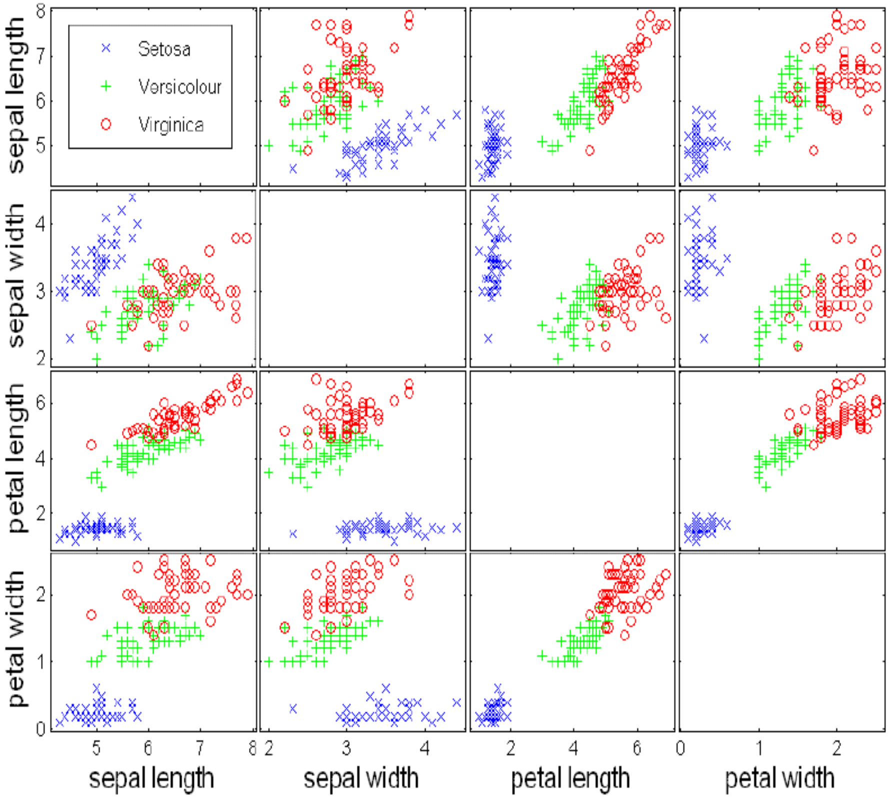
  - We can form a matrix: sample covariance matrix (symmetric)
- Sample correlation: $$r_{jk} = \frac{q_{jk}}{s_j s_k}$$
  - A sample correlation matrix can be formed
- Sensitive (not robust) against outliers
- Measure the strength of the linear relationship between variables



# Scatterplot

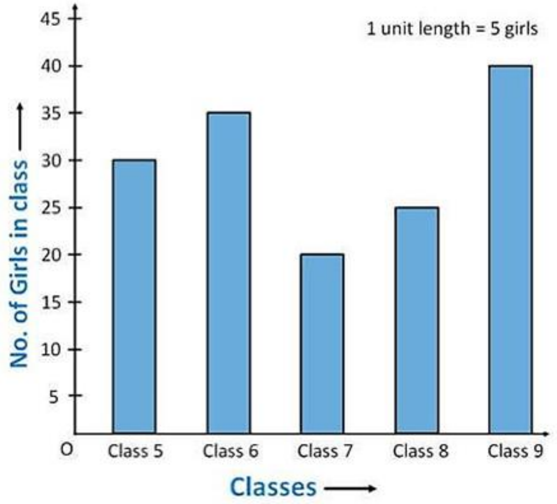
- The objects correspond to points on the plane / in the space
- The coordinates of the points correspond to the values of two/three attributes of the object
- Beyond the (max) three dimensions the point can also have color/shape/size
  - So we can visualize 5-6 dimension all together
    - It is hard to interpret above 4 dimensions

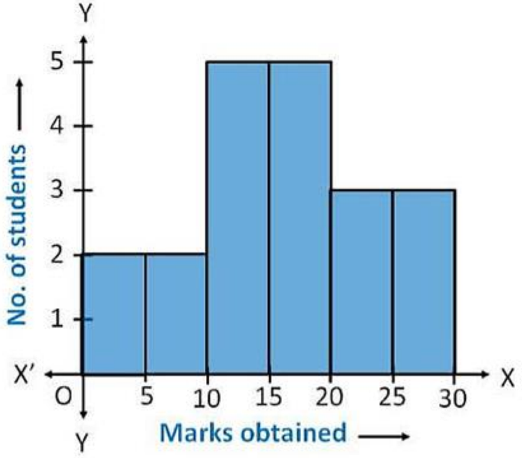


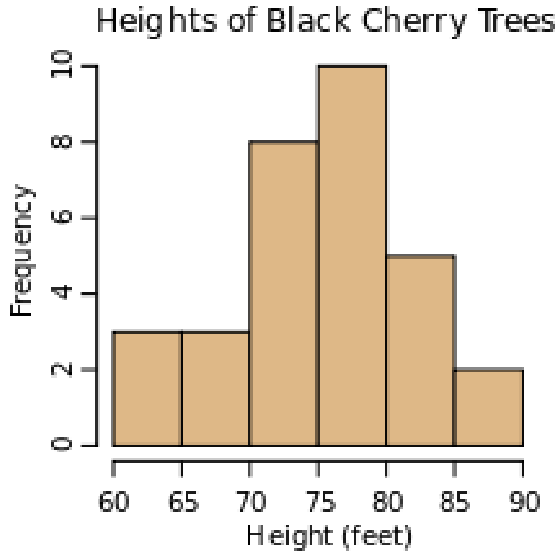


# Histogram

- Representation of the distribution of numerical data
- It is an estimate of the probability distribution of a continuous variable
  - Empirical distribution
- Binning the range of values: dividing the entire range of values into a series of intervals
- Counting how many values fall into each interval
  - A rectangle is erected over the bin with height proportional to the frequency

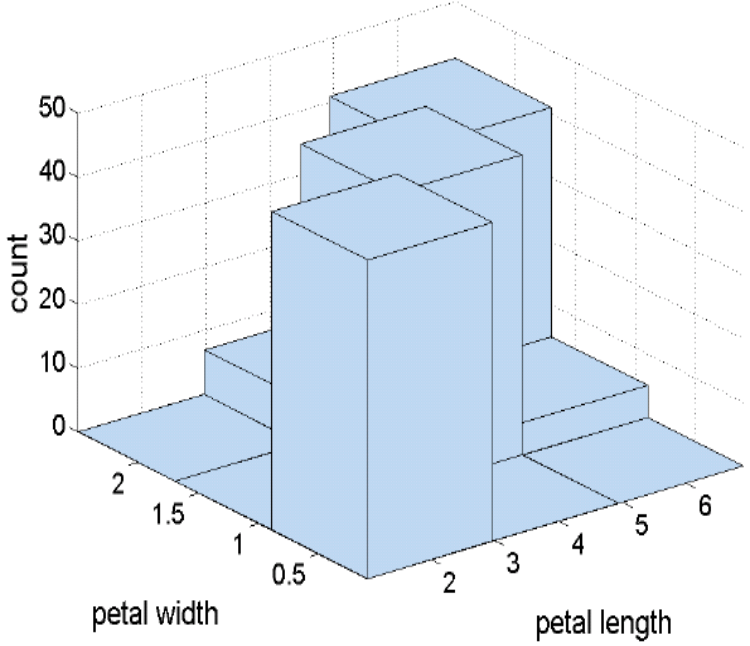


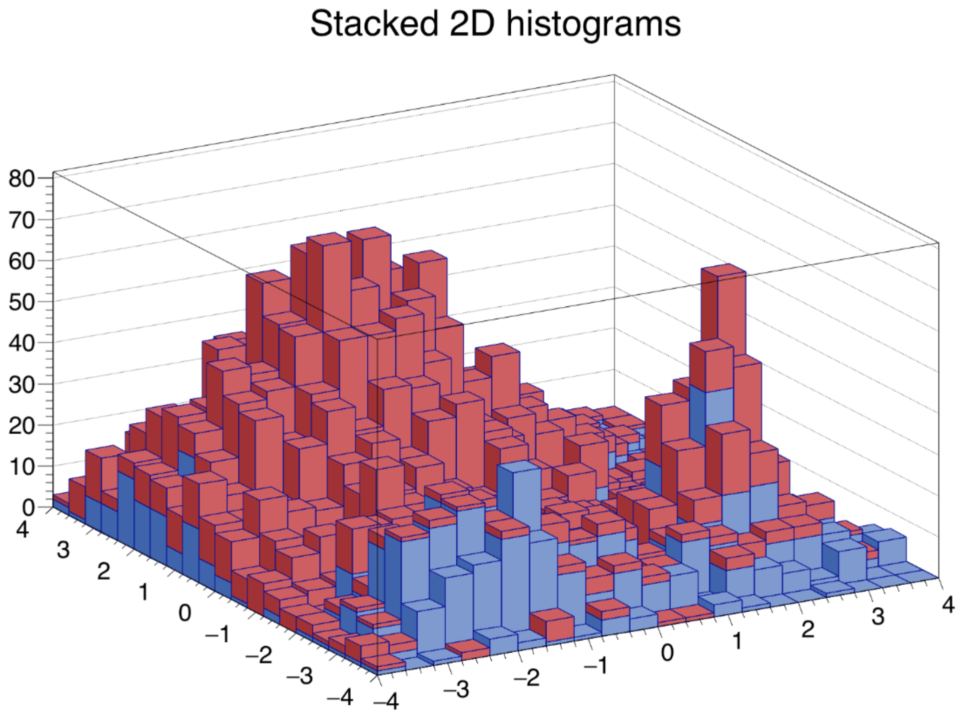



# Two-dimensional histogram

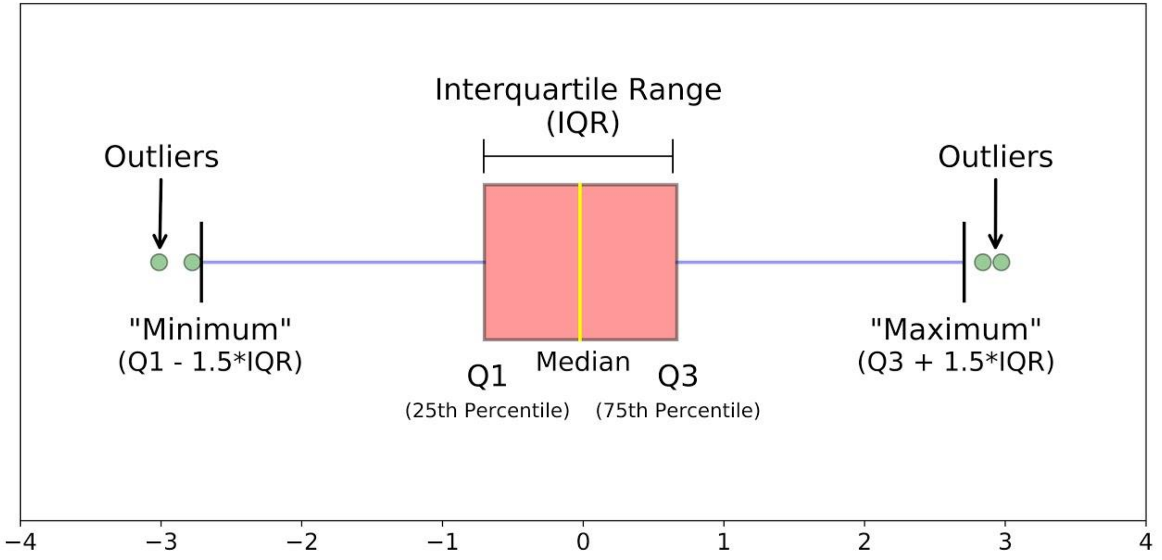
- It estimates the joint distribution of two variables

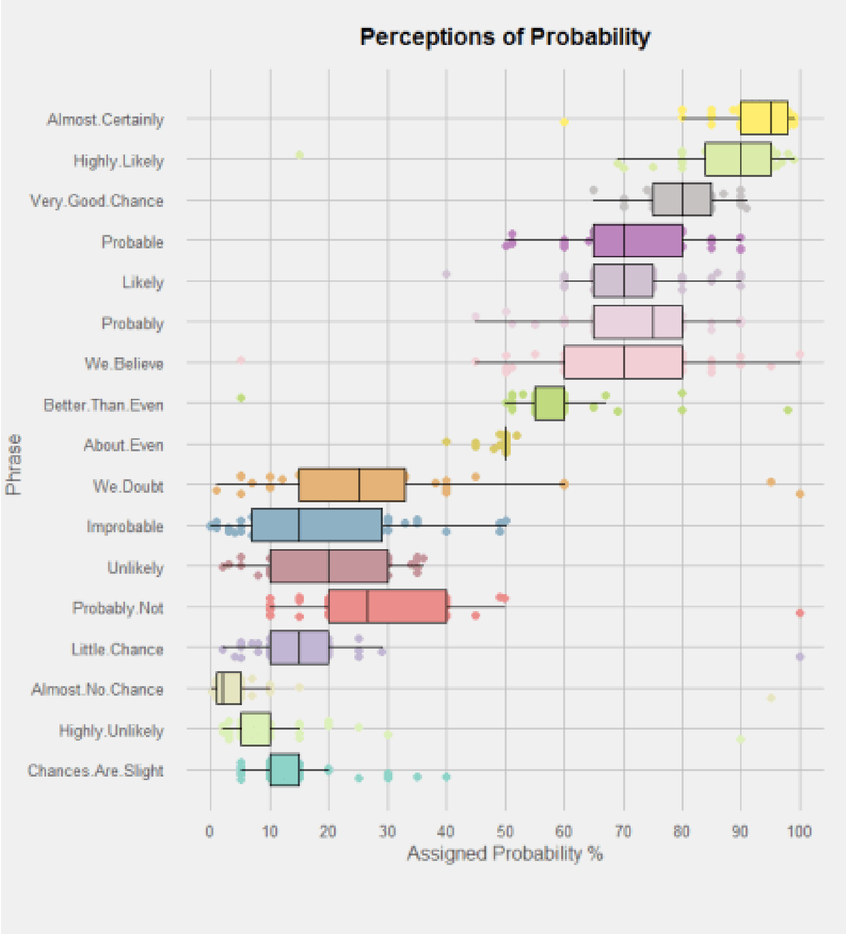


# Boxplot

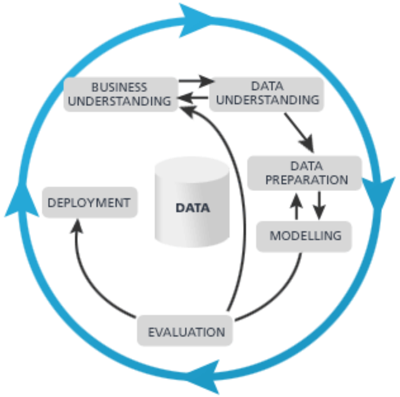
- Another method to visualize distribution
  - Attributed to J. Tukey



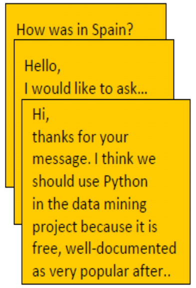



# DP – Data preparation

The process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for analytics.



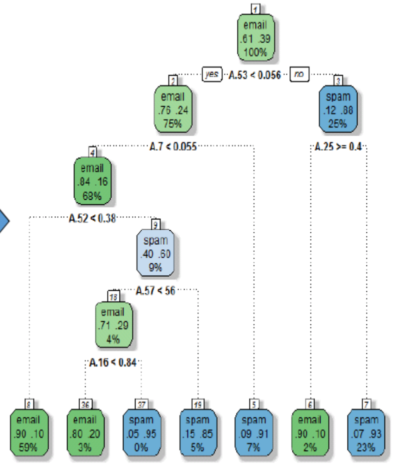

**We can estimate that 70% of resources (time, technology, personnel) used in the whole data science project are committed to DU + DP phases.**




Raw data

| „Black list" | # Capital letters | Total length | ... | Spam |
|---|---|---|---|---|
| 1 | 67 | 2078 | | Yes |
| 0 | 124 | 1279 | | No |
| 0 | 42 | 1538 | | No |
| ... | ... | ... | | ... |

Processed data



(Mathematical) model

Preprocessing

Machine learning algorithm

# Data preparation

## Values

- Identifying and correcting errors
- Imputing missing values

## Rows

- Remove duplicates
- Detect outliers

## Columns

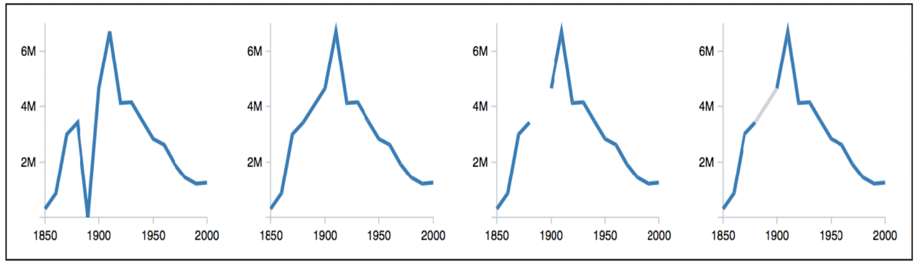
- Feature selection / dimension reduction
- Introducing new features
- Transforming features
- Scaling attributes
- Discretization
- One-hot-encoding

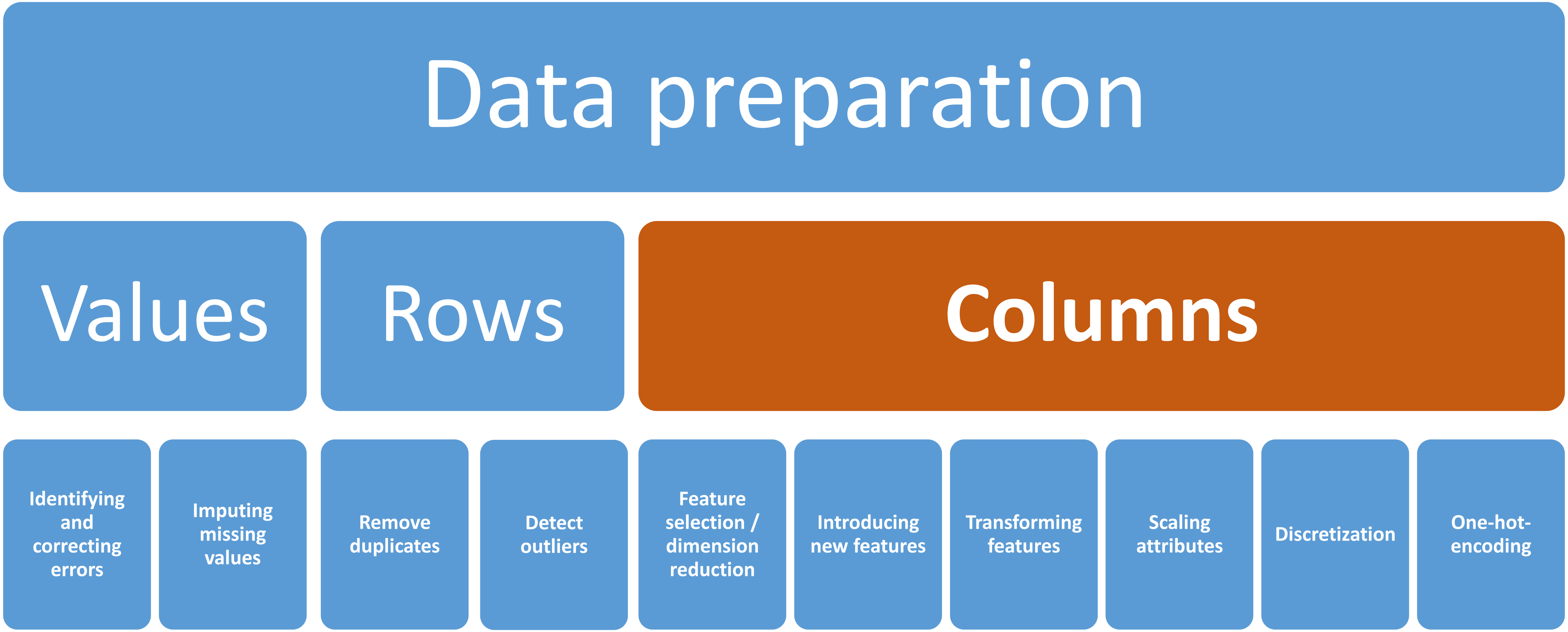
# Common problems with values and rows

- Measurement errors
- Inconsistency (mile, m, km; Budapest, Bpest)
- Not plausible data
  - Everybody has a six-figure salary
  - Everybody gets an A+ from Data Science at AIT ☺
- No header
- Missing apostrophe from text fields
- Missing data
- Duplicates (recurring rows)
  - Sometimes not completely identical, e.g. same person with more very similar addresses
- Outliers: point that is distant from other observations
  - Not a problem by itself, but may be

# How can the problems be fixed?

- Measurement error: can’t be fixed but can be excluded from data if it is detected
- Missing values (data imputation):
  - Not necessarily a problem (perhaps that attribute is not interpreted/defined for every row)
  - We can remove the entire row of the missing value (not a good solution if we have many missing values)
  - We introduce a new global constant, e.g. an „unknown” label
  - We substitute the missing value with the column average (global column average or average with a given label)
  - We impute the missing value with a smart guess based on a machine learning model
- Duplicates: to detect the (nearly) identical observations
- Outlier:
  - Detecting the outlier can be the aim of the project (e.g. freud detection)
  - Sometimes outliers should be excluded
    Sometimes outliers are organic part of the data and they should remain in the data




Data preparation
Values
Rows
Columns
Identifying and correcting errors
Imputing missing values
Remove duplicates
Detect outliers
Feature selection / dimension reduction
Introducing new features
Transforming features
Scaling attributes
Discretization
One-hot-encoding

# Reducing the number of attributes

## Aim: to have fewer columns

**Why?**

- Achieve faster running time
- Need less storage capacity
- Easier to visualize
- The results are easier to interpret
- In high dimension most of the models perform poorly (due to curse of dimensionality)

**How?**

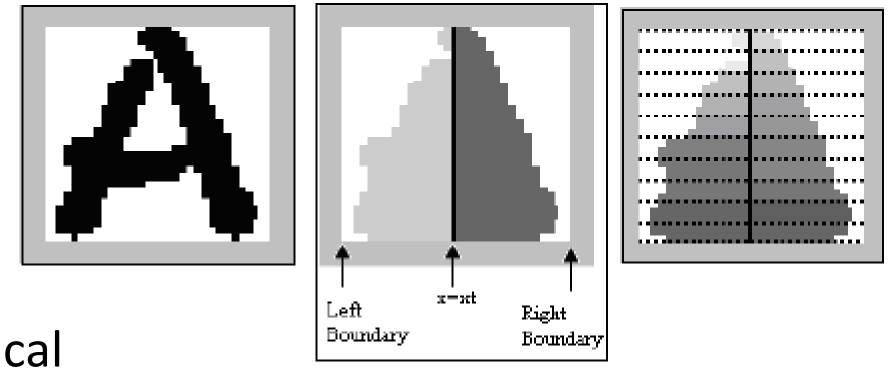
- Omit redundant columns
- Merging columns
- Introducing new (better) attributes and omitting the old ones
- Advanced dimension reduction methods (such as PCA)

# Methods for reducing the dimensionality

- Finding redundant columns
  - E.g. Column A: price of the product, Column B: paid VAT
- Finding irrelevant columns
  - E.g. the phone number of a person regarding their creditworthiness (are you sure?)
- Automatic filtering
  - If the correlation of two columns is too high, we omit one of them
- Embedded methods
  - The used machine learning method chooses the relevant variables itself (later)
- Advanced dimension reduction methods (such as PCA)

# Introducing new attributes

- Sometimes we don’t necessarily want to reduce the number of attributes but we want better, more expressive attributes
- Sometimes domain knowledge is needed
- Examples:
    - To extract features from pixel series of images
        - Number of „on” pixels, average of the horizontal coordinates of the „on” pixels, variance of the vertical coordinates, correlation between the horizontal and vertical positions of „on” pixels




    - Combining attributes based on domain knowledge
        - Introducing density instead of volume and mass

# Scaling attributes

- Feature scaling: sometimes it is necessary to standardize/normalize the range of independent variables (e.g. for visualization, for some machine learning algorithms)
  - Rescaling the range in [0, 1] (min-max normalization) : $x' = \dfrac{x - \min(x)}{\max(x) - \min(x)}$
    - Affected by outliers
    - Useful when we don’t know about the distribution
  - Standardization (zero-mean, unit-variance): $x' = \dfrac{x - \overline{x}}{s_x}$
    - Much less affected by outliers
    - Useful when the feature distribution is Normal

# Other transformations – logarithmic transformation

- In some application we can take the logarithm of the attribute
  - Especially salary/income/wealth-related
    - It makes more sense to measure „percent” changes in wage rather than absolute changes
    - It is usually more normally distributed)
    - „Diminishing marginal utility"
- Other (bijective) mappings of the attributes might also be useful

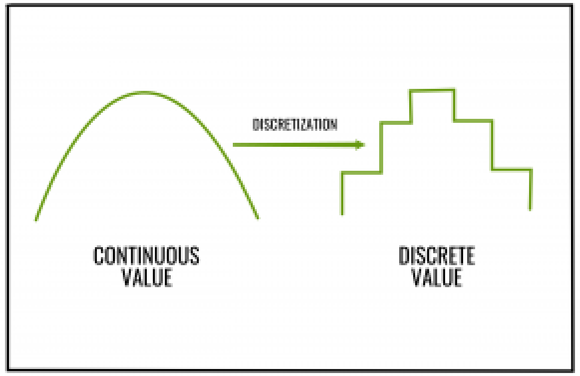

$$\log_b(MN) = \log_b(M) + \log_b(N)$$

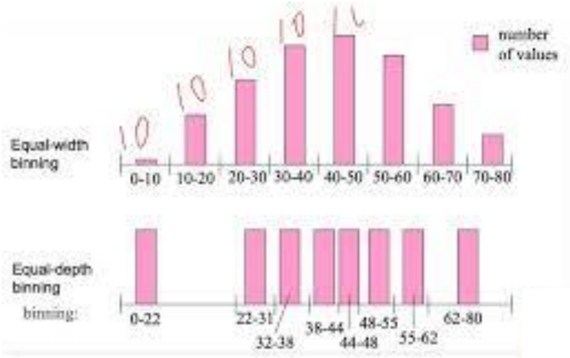
$$\log_b\left(\frac{M}{N}\right) = \log_b(M) - \log_b(N)$$

$$\log_b(M^p) = p\log_b(M)$$

# Discretization

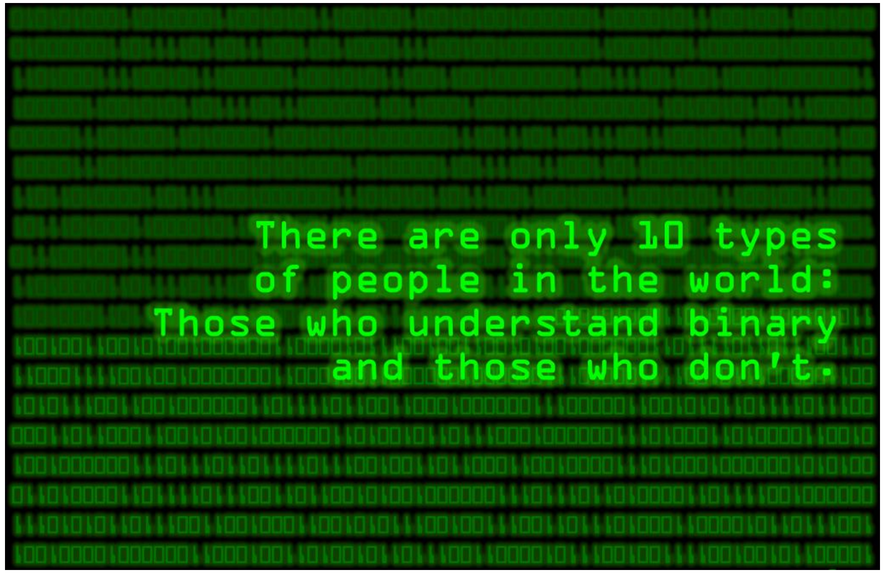
- Goal: transferring continuous variables into discrete counterparts
- Why?
  - Some algorithms need discrete variables
  - We need to store less values, for some algorithms the running time is much faster
  - Sometimes a rougher scale is sufficient, e.g. high, medium, low values, the data could be more  clear-out
- How to partition the data? What are the discretization cut points?
  - Divide the range of the continuous variable into intervals of the same length (equidistant division)
  - Dividing the range into intervals with the same number of observations (along quantile values) – equal frequency
  - Dividing along quantile values creating groups with differing size, e.g. along quantiles 10, 30, 70, 90
  - If there are some natural cut points where the data is rare, it is reasonable to choose these cut points



# From categorial to binary: one-hot-encoding

- If a nominal attribute has *k* possible values, it is replaced by *k* synthetic binary attributes (one attribute-per-value approach or one-hot encoding)
  - The *i*th being 1 if and only if the original value corresponds to the *i*th group



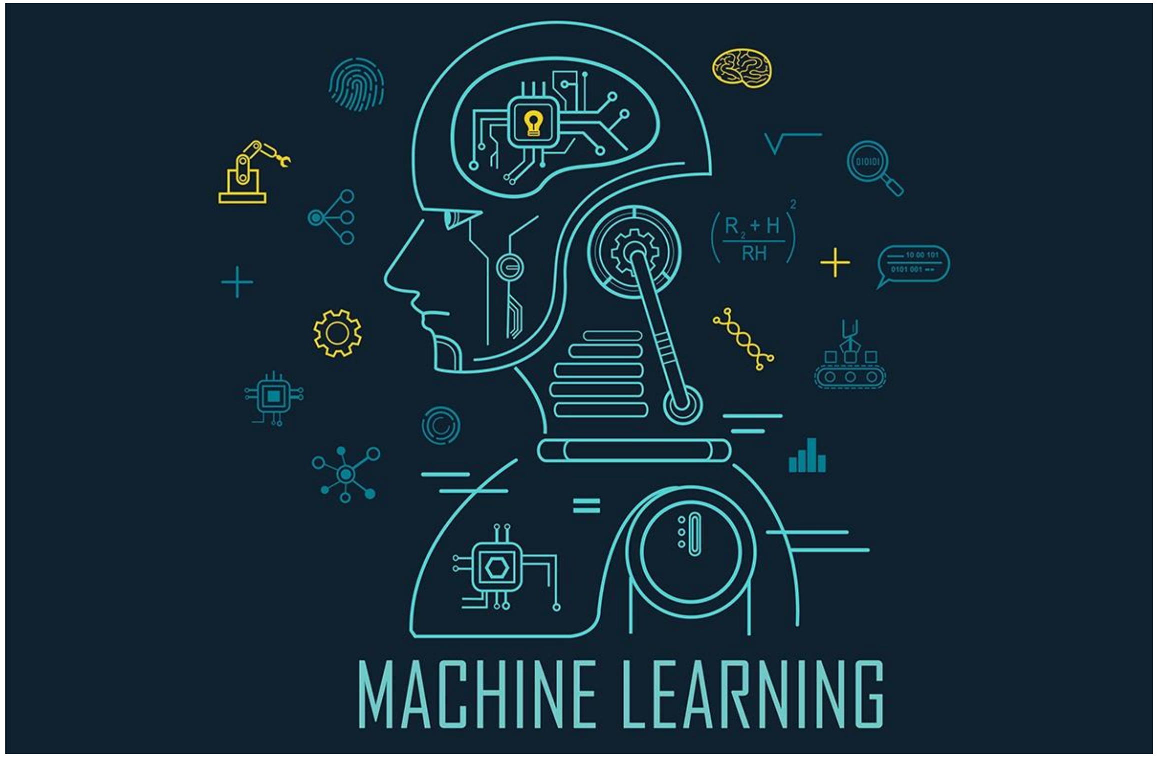

| id | color |
|---|---|
| 1 | red |
| 2 | blue |
| 3 | green |
| 4 | blue |

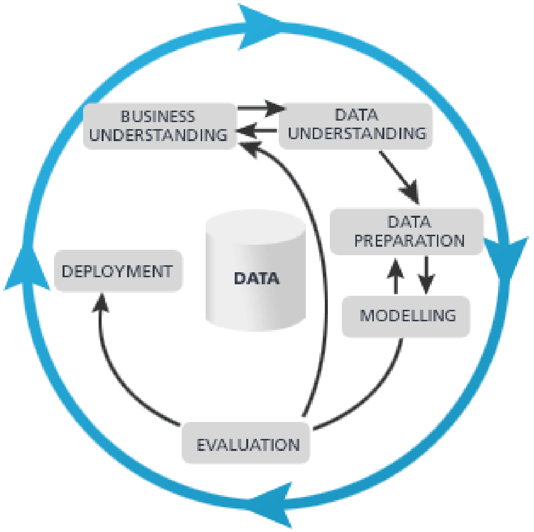
One Hot Encoding

| id | color_red | color_blue | color_green |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 1 | 0 |

# M-Modeling

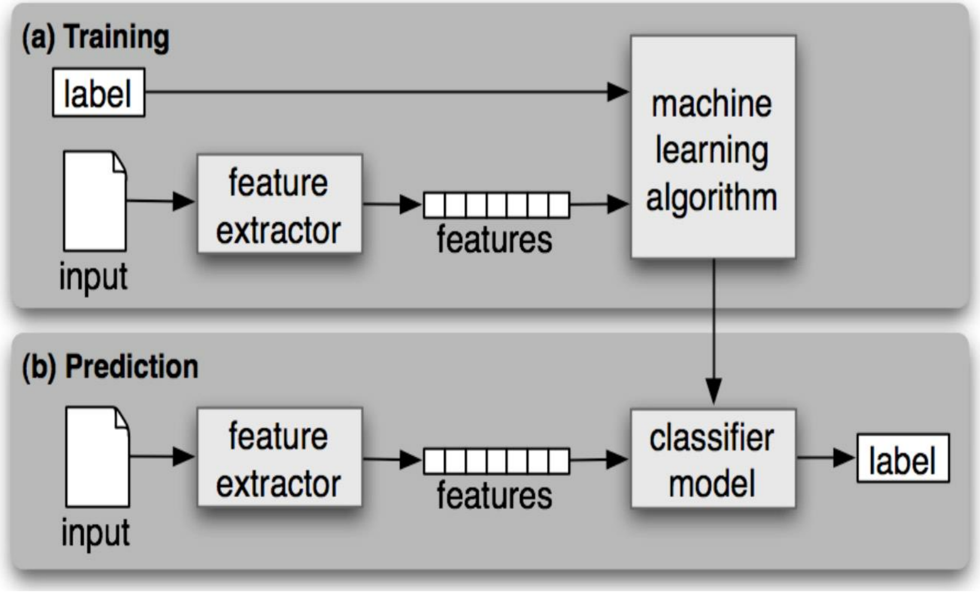
Finding the best performing model, fitting the parameters



# Supervised / unsupervised learning

- Supervised learning
  - We have a training set where the value of the target variable is known
  - Aim: based on the attributes predict the target when it is not known
  - Example: classification, regression
- Unsupervised learning
  - The target (label) is not known for any records (latent labels)
  - Aim: to associate useful labels to the records based on the attributes
  - In many cases our aim is to gain better understanding of the data or visualize the data
  - Example: clustering



# Regression

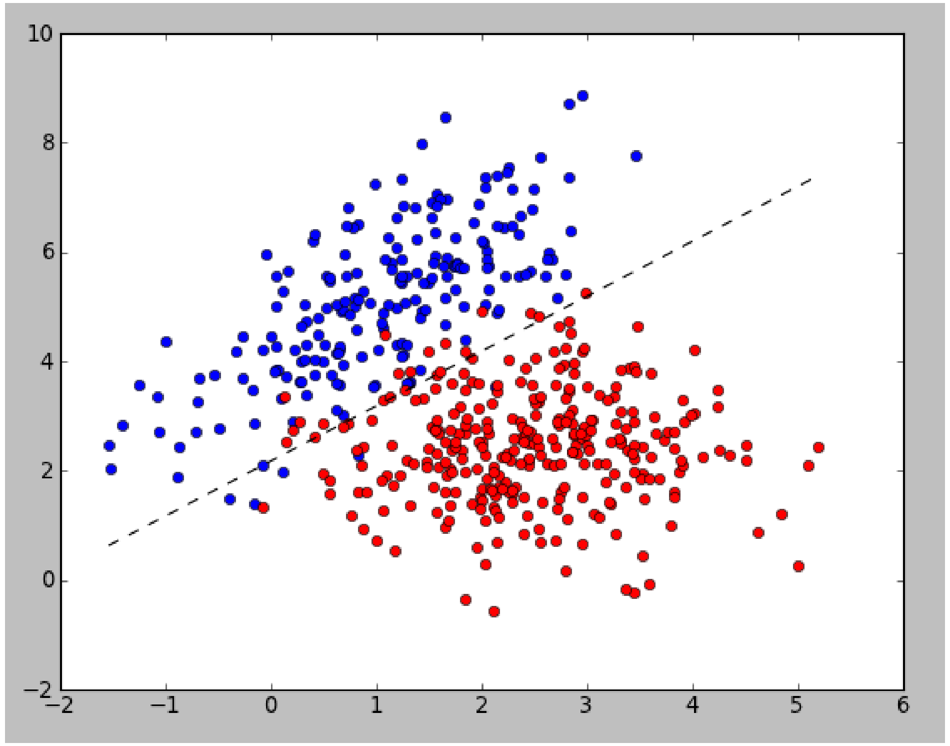
- We try to predict continuous valued output based on the values of other variables (via supervised learning)
- Examples:

| Explanatory (input) variables | Target (output) |
|---|---|
| Number of rooms, size, location (ZIP code), … | Market value of a house |
| Movie budget, film genre, popularity of the actors (based on their IMDB pages) | Box office result of a movie |
| Major, admission point score, gender, age, GMAT scores, …. | GPA |

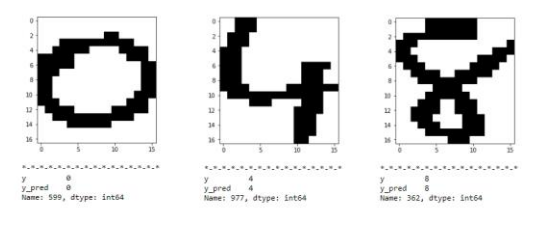
- Challenges: finding the right explanatory variables, the most suitable functional form/modelling approach

# Classification

- We try to predict discrete (sometimes binary) valued output based on the values of other variables (via supervised learning)
- It is also possible to do „classification via regression”
- Challenges: finding the right explanatory variables, the most suitable modeling approach, fitting the parameters of the model



# Classification - examples

| Input variables (features) | Target variable |
|---|---|
| Purchase history, age, gender | Should we send a targeted advertisement message to a customer? (0/1) |
| Number of „on" pixels, average of the horizontal coordinates of the „on" pixels, variance of the horizontal coordinates, correlation between the horizontal and vertical positions of „on" pixels, … | Handwritten digit recognition (0/1/2/3/4/5/6/7/8/9)  |
| Salary, marital status, address, profession, qualification, … | Is the customer creditworthy? (0/1) |
| Words/n-grams appearing in the e-mail, subject of the mail, sender, number of receivers, … | Is the email spam? (0/1) |
| Age, gender, profession, qualification, contents liked on Facebook, … | Psychological profiles/ temperaments (e.g.: sanguine, phlegmatic, choleric, and melancholic) |

# Fundamental task of regression

Let $X = \left(X_1, X_2, \ldots, X_p\right)$ be the feature vector and $Y$ is the target variable.

Regression: we suppose that there is a relationship between $X$ and $Y$, in general: $Y = f(X) + \epsilon$, where $\epsilon$ (the random error) is independent from $X$ and has zero mean

Aim: giving prediction: $\hat{Y} = \hat{f}(X)$

> In reality $\hat{f}$ is sometimes considered to be a black-box, we are not interested in the functional form, but in giving accurate enough prediction for $Y$

Learning: On the labeled data of the training set we estimate the function $f$, minimizing the „prediction error” on the training set

Prediction: using $\hat{f}$ for data that we have not seen before $\hat{Y} = \hat{f}(X)$

# Fundamental task for classification

Let $X = (X_1, X_2, \ldots, X_p)$ be the feature vector and $Y$ is the target variable.

For classification problems: $Y \in \{c_1, c_2, \ldots, c_k\}$

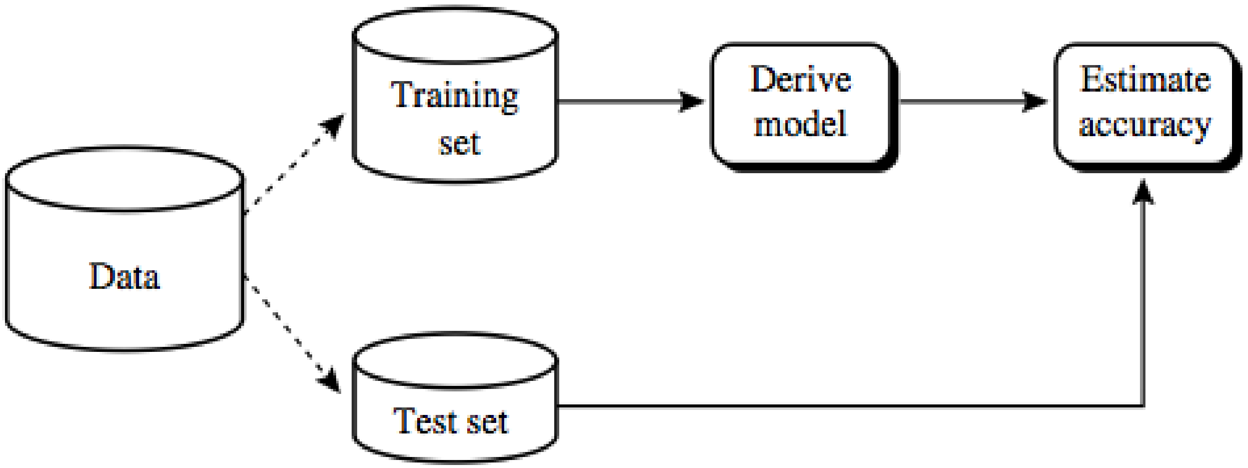
The real $p(X, Y)$ joint distribution (background distribution) is not known

Aim: finding $f$ such that $P\big(Y = f(X)\big)$ is maximal.

Learning: On the labeled data of the training set (independent identically distributed sample from the $p(X, Y)$) we estimate the function $f$, minimizing the „classification error” on the training set

# Generalization ability

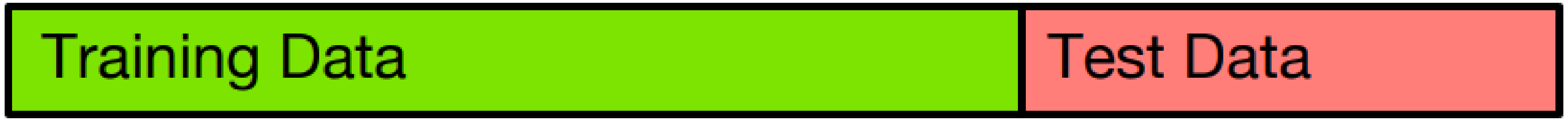
- Purpose: to build a model that predicts the target variable well in general not just on the available data set ➔ good generalization ability
- Dataset is divided into two (or later three) parts
- Cross validation: later
- To evaluate models, a numerical „goodness" notion is needed



# Training and test set

Spliting the data set into two parts

- Training set: fitting the model (i.e. optimizing its parameters) on the training set in such a way that it has a good performance on the training set and has a good generalization ability
- Test set: we test the model performance on data that were not seen by the model before
  - We choose the model that has the best performance on the test set

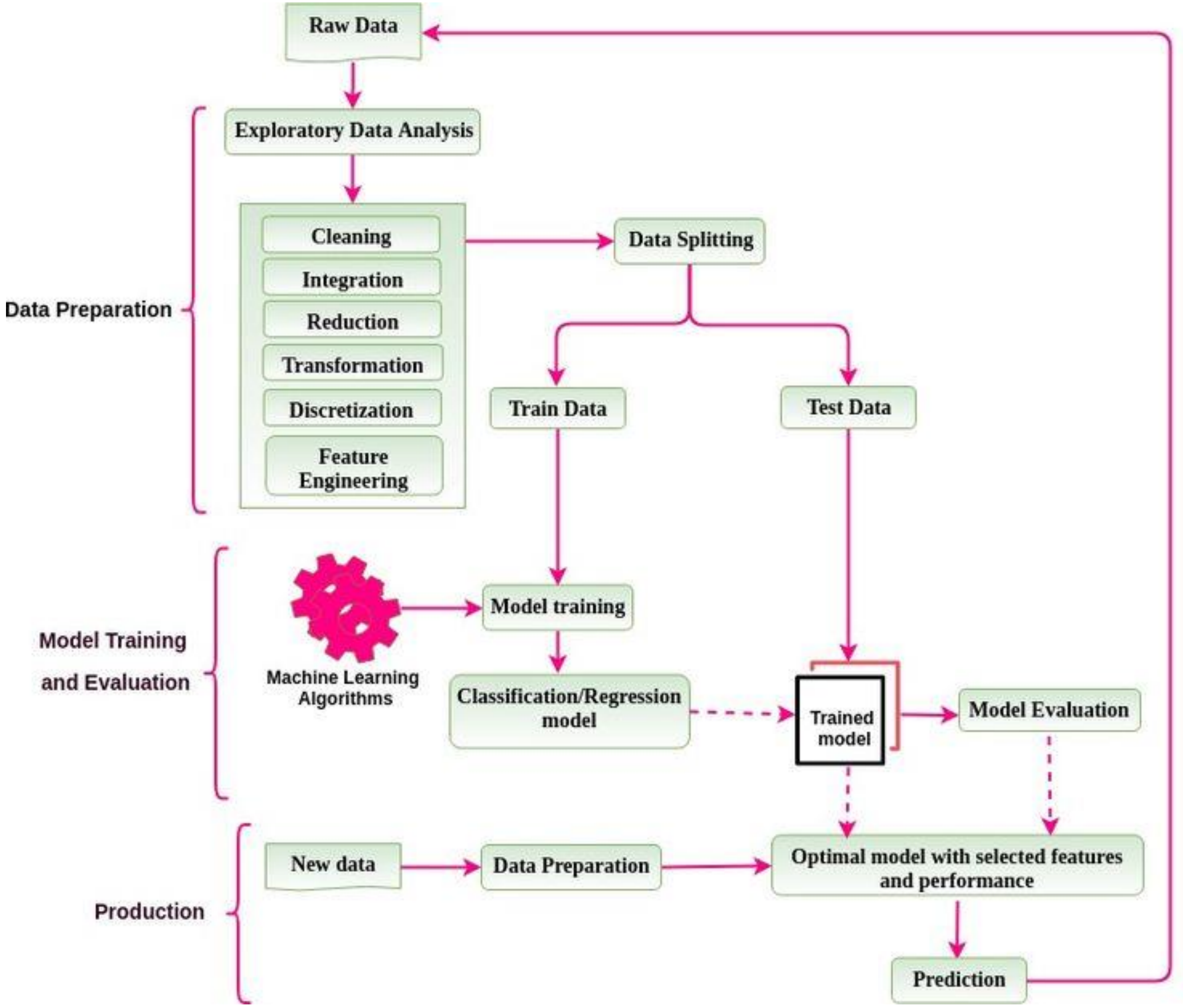


# E -Evaluation

- Evaluating the model. How does it perform? Is it good enough to achieve our goal?

# D -Deployment

- Implementing the model, embedding it to the system. Communicating the results. Writing the report/research paper.



# Requirements for successful data science projects

- Having domain knowledge or consulting with domain experts
- Big data (many observations)
  - Less likely to retrieve connections that is just in the data due to chance
  - (It can be computationally expensive!)
- Many features
  - Simple analytics bears with few features
- Clean data
  - Bad data encumber data analysis or leads to false results
  - GIGO: garbage in, garbage out

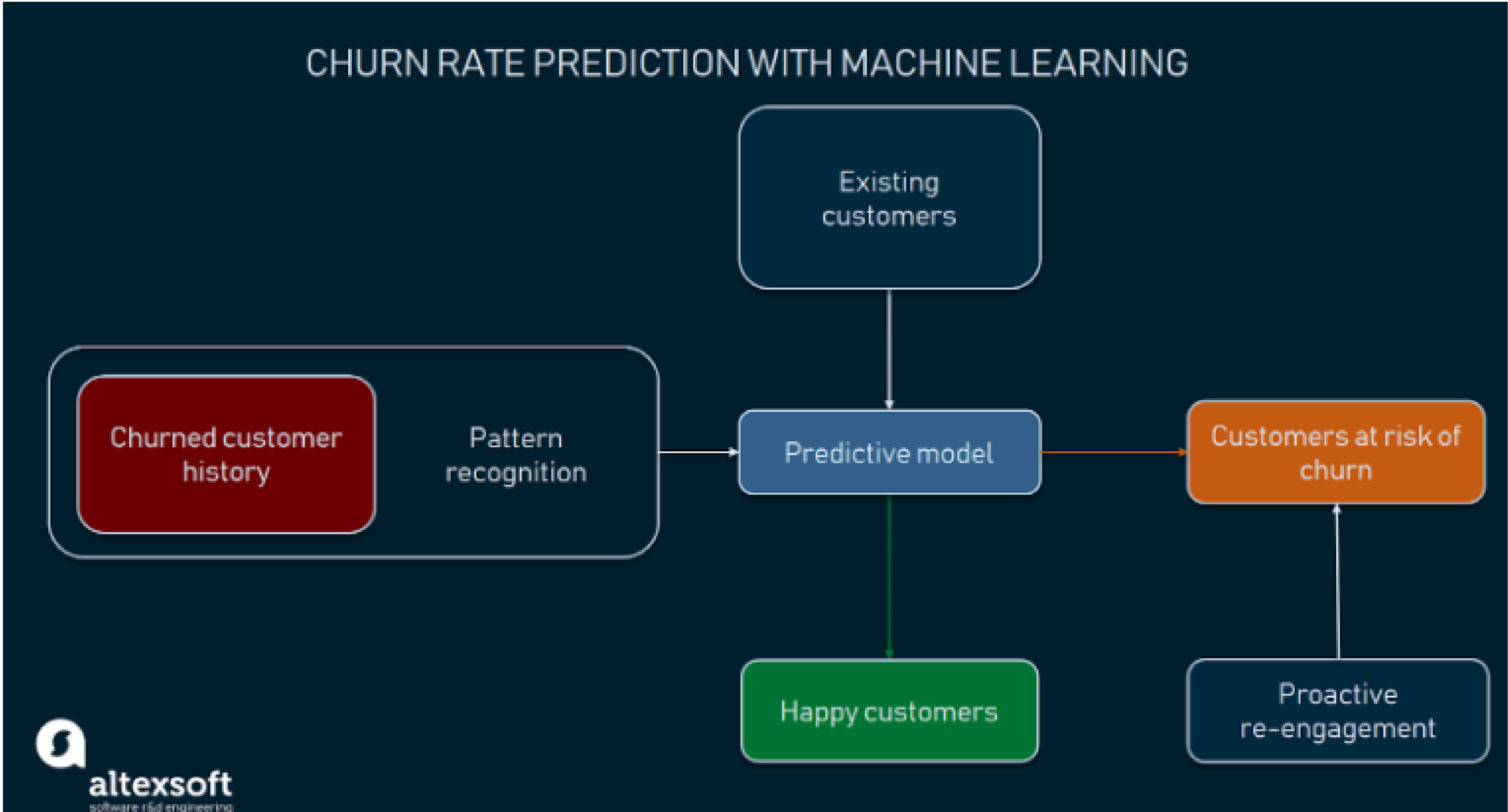
# Requirements for successful data science projects II.

- Unbiased data
  - The data (the sample) should be representative to the population itself
  - BIBO: bias in, bias out
- The capacity of act
  - Sometimes the knowledge is discovered, but it will not go into action (high costs, too rigid system)
- Measurability of Return of Investment (ROI)
  - It defines the success of a project

# Case study – customer churn detection in the telecommunication sector

- Churn: occurs when customers  unsubscribed or cancel their service contract
- A telecommunication company approached our (imaginary) data science consulting company to predict which customers are at risk of leaving our business
  - Customer retention campaign targeted on at-risk customers
  - Offering coupons or discounts to those most likely to churn

1. How would you formulate the task as a data science problem?
2. Plan the analysis based on the CRISP-DM methodolgy!
3. Do you think that the requirements of a successful data science project are met?


CHURN RATE PREDICTION WITH MACHINE LEARNING
Existing customers
Churned customer history
Pattern recognition
Predictive model
Customers at risk of churn
Happy customers
Proactive re-engagement
altexsoft
software r&d engineering

# Customer churn prediction

- BU
  - business objective is reducing customer churn by identifying potential churn candidates beforehand, and take proactive actions to make them stay
- DU
  - Personal data about the customers (age, address, …)
  - Information about their subscription plan
  - Call/text/data logs (who?, when? how much? etc.)
- DP
  - Feature engineering, transforming features etc.
- M
  - Binary classification problem (supervised learning)
- E
  - Test the performance of the model. Is it good enough to deploy?
- D
  - Design a retention campaign (probably with A/B testing)

What about the success requirements?

# Acknowledgement

- András Benczúr, Róbert Pálovics, SZTAKI-AIT, DM1-2
- Krisztián Buza, MTA-BME, VISZJV68
- Bálint Daróczy, SZTAKI-BME, VISZAMA01
- Judit Csima, BME, VISZM185
- Gábor Horváth, Péter Antal, BME, VIMMD294, VIMIA313
- Lukács András, ELTE, MM1C1AB6E
- Tim Kraska, Brown University, CS195
- Dan Potter, Carsten Binnig, Eli Upfal, Brown University, CS1951A
- Erik Sudderth, Brown University, CS142
- Joe Blitzstein, Hanspeter Pfister, Verena Kaynig-Fittkau, Harvard University, CS109
- Rajan Patel, Stanford University, STAT202
- Andrew Ng, John Duchi, Stanford University, CS229