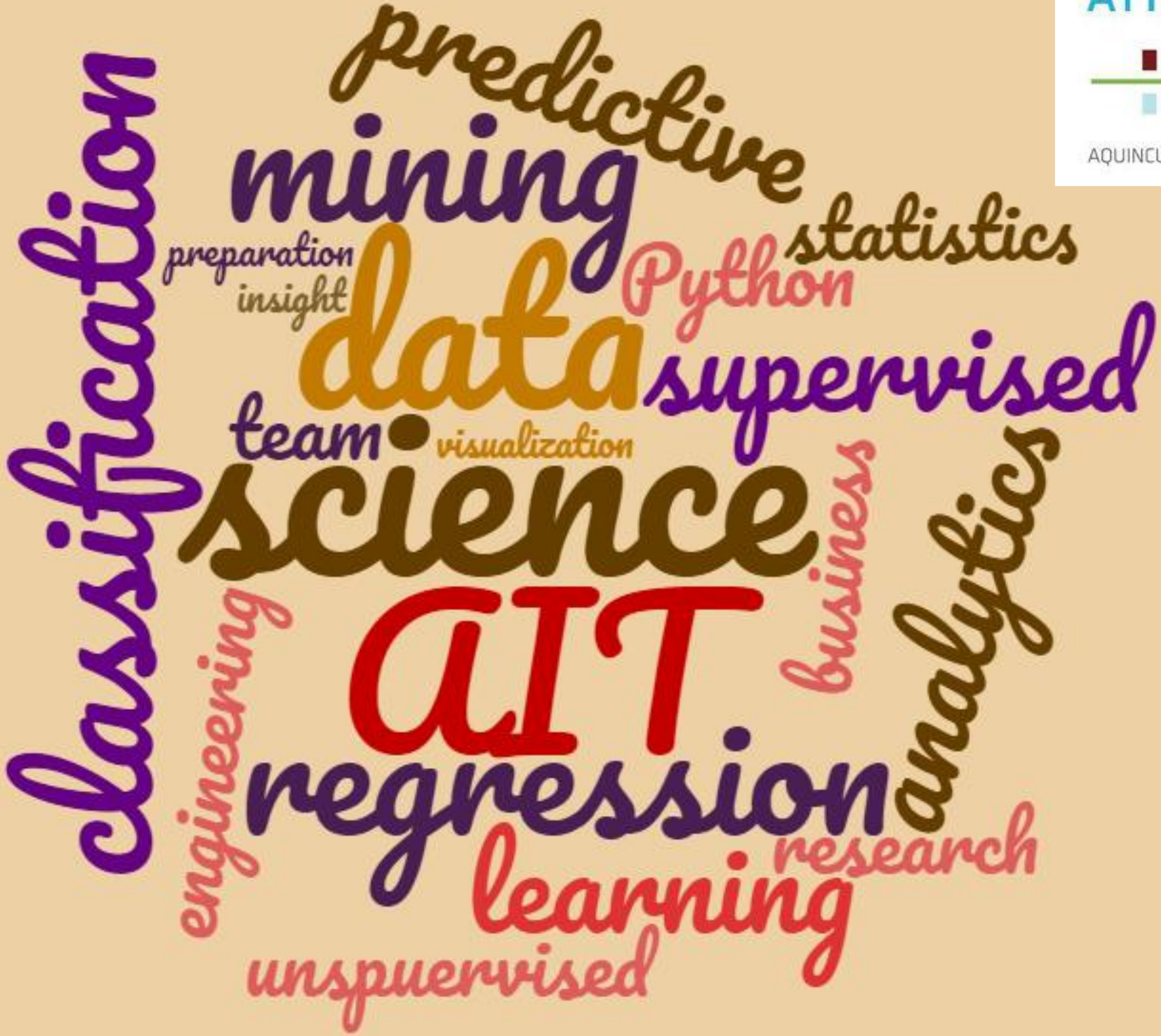
AIT-BUDAPEST
AQUINCUM INSTITUTE OF TECHNOLOGY

Data Science

February 7, 2020
Process of data science, fundamental tasks



Roland Molontay

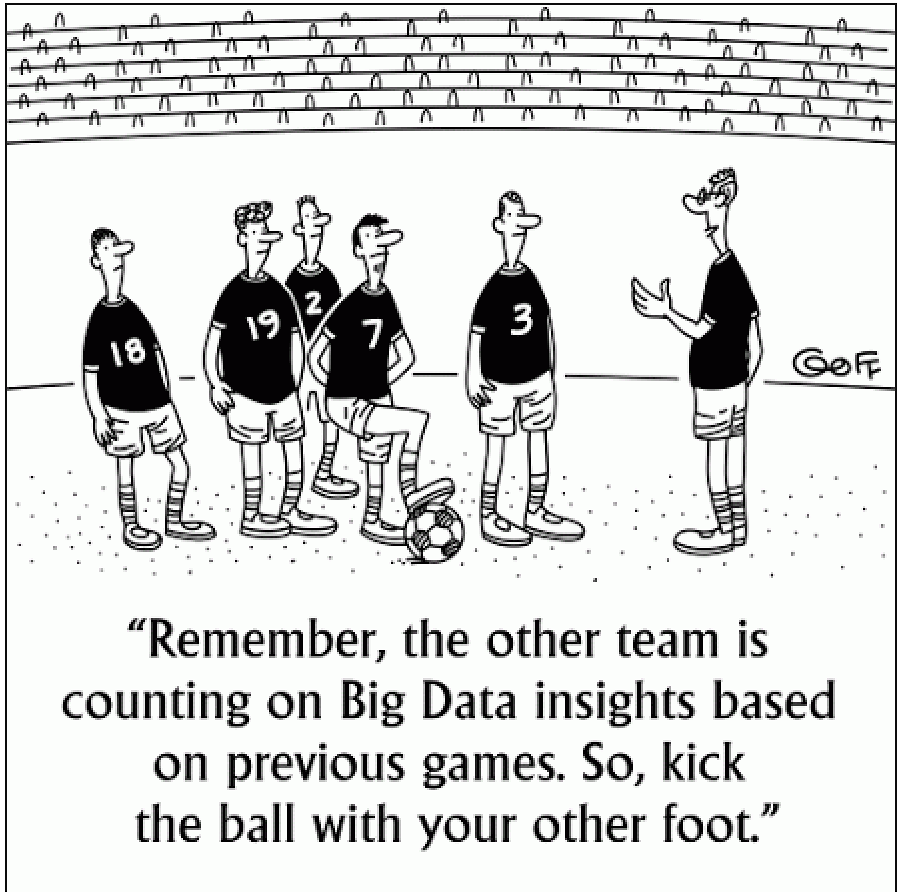
Fundamental tasks of data science

- Predictive analytics:
  - Based on the values of some variables (attributes/features) predict the value of a selected variable (target variable)
  - If the target variable is continuous: regression
  - If the target variable is discrete: classification
    - Binary target variable: binary classification
  - Supervised learning
  - Recommender systems

THIS YEAR, WE PREMADE YOUR GIFT BASED ON DATA-DRIVEN INSIGHTS FROM YOUR WISH HISTORY.
TOM FISH BURNE
© marketoonist.com

Fundamental tasks of data science II.

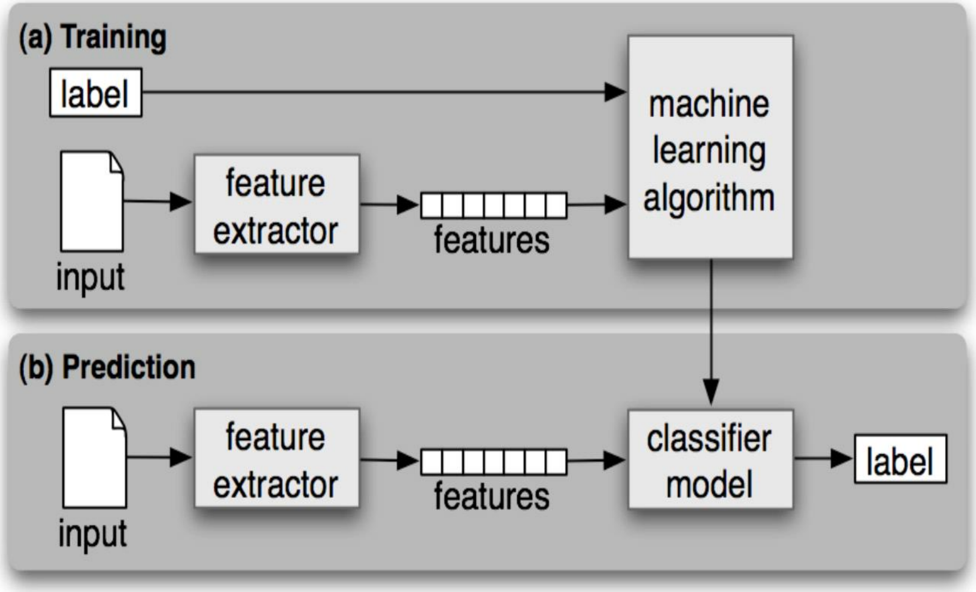
- Finding patterns, explorative, descriptive analysis
  - Finding connections between variables
    - Unsupervised
  - Anomaly / outlier detection
    - May be supervised/unsupervised
  - Feature selection, dimension reduction
    - Unsupervised
  - Clustering
    - Unsupervised
  - Association rule mining



"Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot."

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, dimension reduction

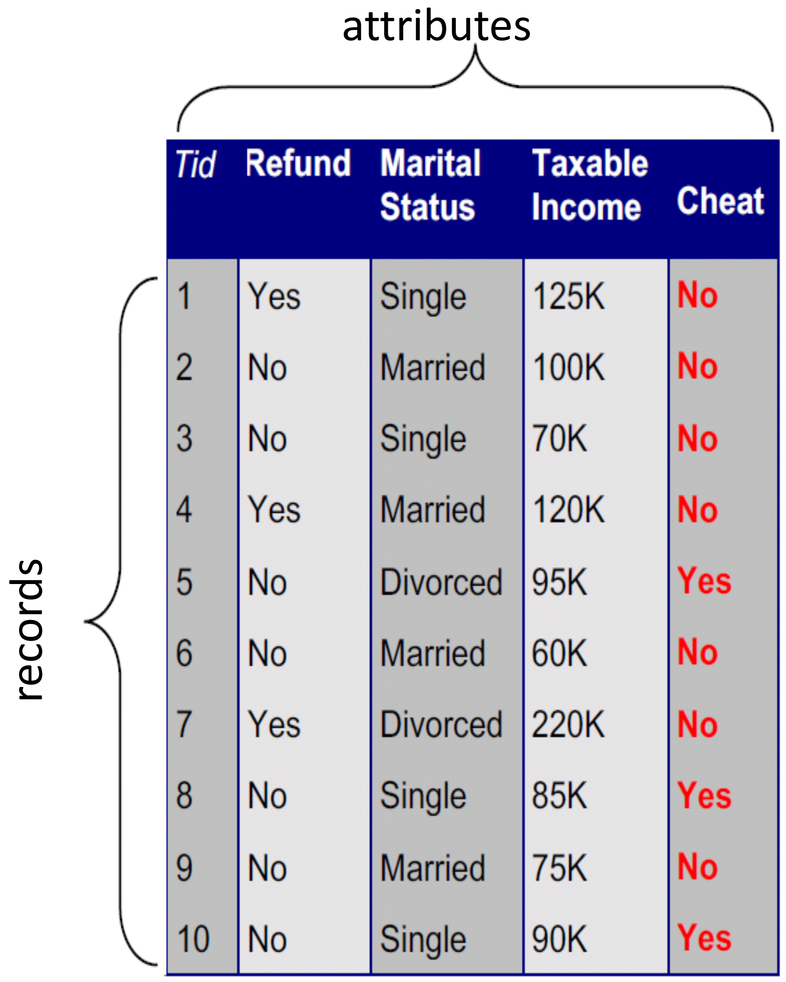


Dataset

- 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

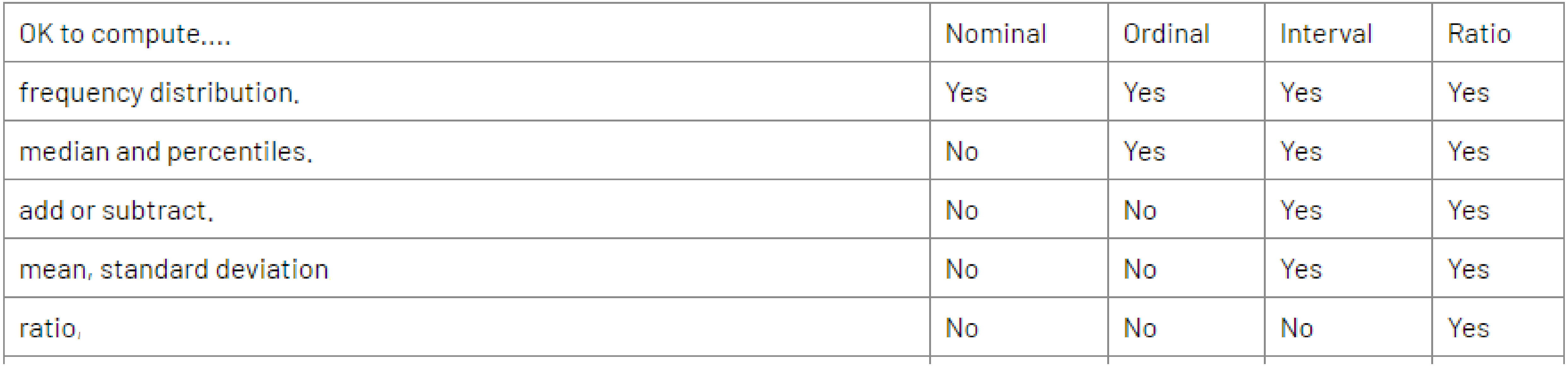
Attribute types

- 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 formats

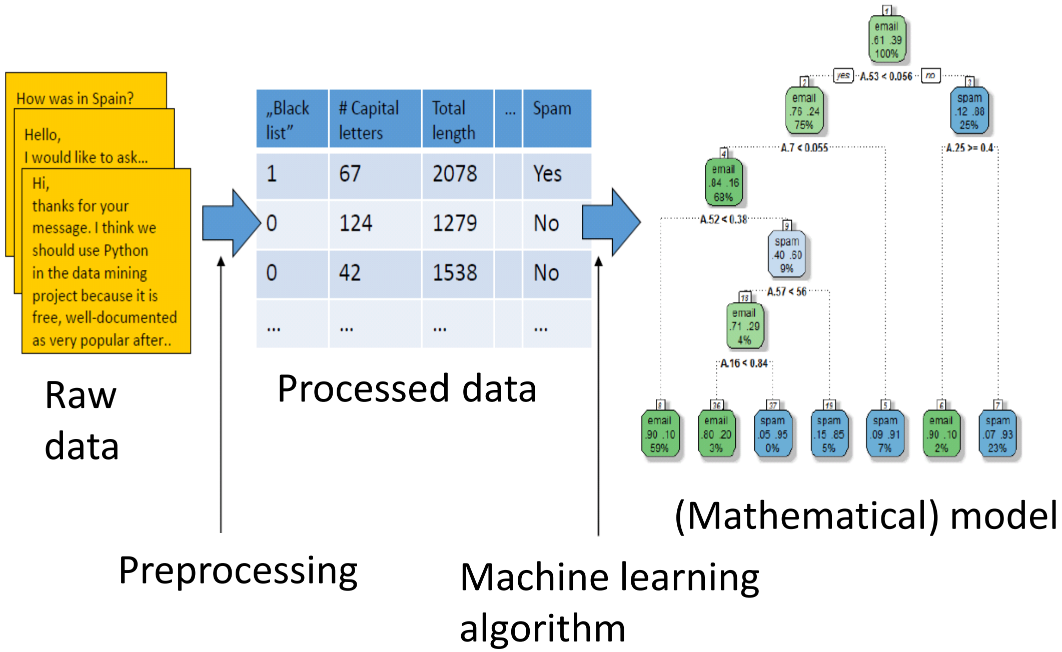
- Structured data (data frame, data table, matrix)
  - If it is a table of numbers with *m* rows and *n* columns the rows can be considered as points in an *n* dimensional space
    - Usually we have many columns ➔ high dimension
  - Special case: document-term matrix
    - Rows are documents, columns are keywords
    - Binary attribute shows if that certain keyword is present in the document (or a discrete attribute stands for the number of appearances)
  - Special case: transaction matrices
    - Rows are transactions, columns are products/items
- Unstructured data
  - Usually we pursue to make them structured
  - Graphical data: connections between molecules – which molecules are connected (bond length, bond angles)
  - Images: they can be reshaped to pixel series ➔numerical features can be extracted
  - Spatial and/or temporal connections between data points
    - E.g. meteorological measurements

Raw data

- The original form that we collect the data
- In most cases we can’t analyze the data in its original form ➔ preprocessing is necessary
- Data (pre)processing, data wrangling, data munging, data cleaning: 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
  - There are no ultimate solutions
  - „Best practice” exists but usually difficult and time-consuming

Processed data

- Data in a format that is ready for analytics
- Data (pre)processing consists of many steps
  - We will cover it later
- It is important that data (pre)processing should be documented as well
  - What is the source of the data? What preprocessing steps were conducted?



How was in Spain?

Hello,
I would like to ask...

Hi,
thanks for your message. I think we should use Python in the data mining project because it is free, well-documented as very popular after..

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

Expectations

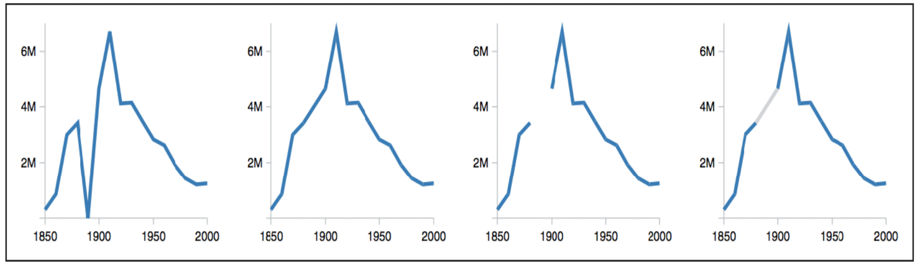
- A data frame should consist of rows of the same type
  - E.g.: Do not mix data about students with data about courses
- A row should correspond to one entity
- A column should correspond to a variable consistently
  - Do not mix naming (New York, New York City, City of New York, NY, NYC)
  - Do not mix formats (2020. 02. 07., 07/02/2019, February 07 2019)
  - Do not mix unity of measurements (mile, km, m)

Common problems with data

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

How can the problems be fixed?

- Measurement error: can’t be fixed but can be excluded from data if it is detected
- Missing values:
  - 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



Fundamental task of regression

Let $X = (X_1, X_2, \ldots, X_p)$ 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 ϵ (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 to give accurate enough prediction to 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 $X = \left(X_1, X_2, \dots, X_p\right)$ be the feature vector and Y is the target variable.

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

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

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

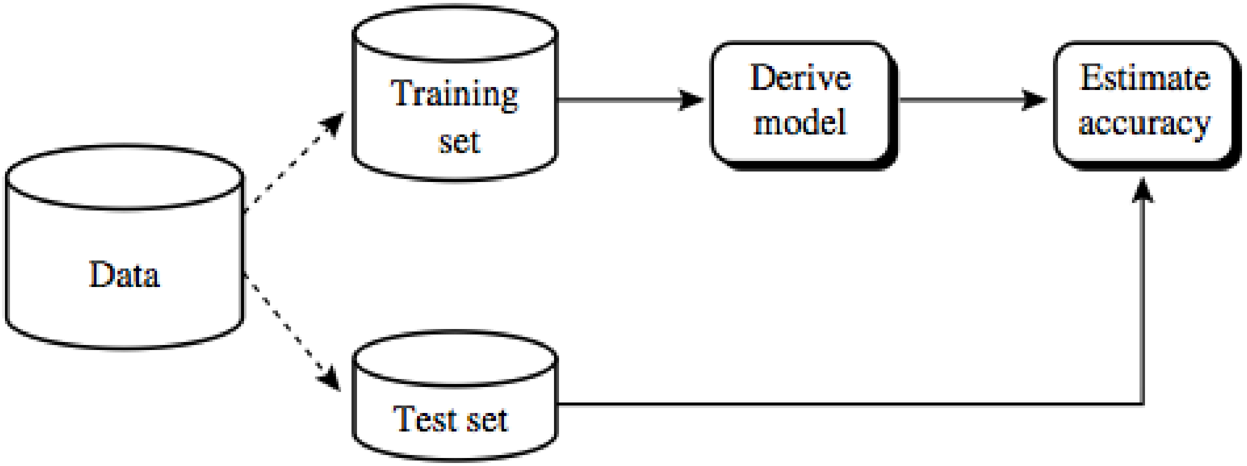
Learning: On the labeled data of the training set (independent identically distributed sample from the $p(X, Y)$) we estimate the function f, minimizing the „classification error" on the training set

Parametric and non-parametric models

- Parametric model
  - We find $f(X)$ in a predefined function family (functional form)
    - E.g. linear, polynomial form
  - The models can be described using a finite number of parameters
    - We optimize for these parameters
- Non-parametric model
  - No functional form is supposed
  - No assumption is needed for the functional form, there is no restriction, the model can better fit the data, the model is more flexible
  - „Infinite dimensional parameter space”
  - To achieve an accurate estimation, more data point are needed than in the parametric case
  - Usually more complex and slower that the parametric case
  - E.g: kNN, decision tree, SVM

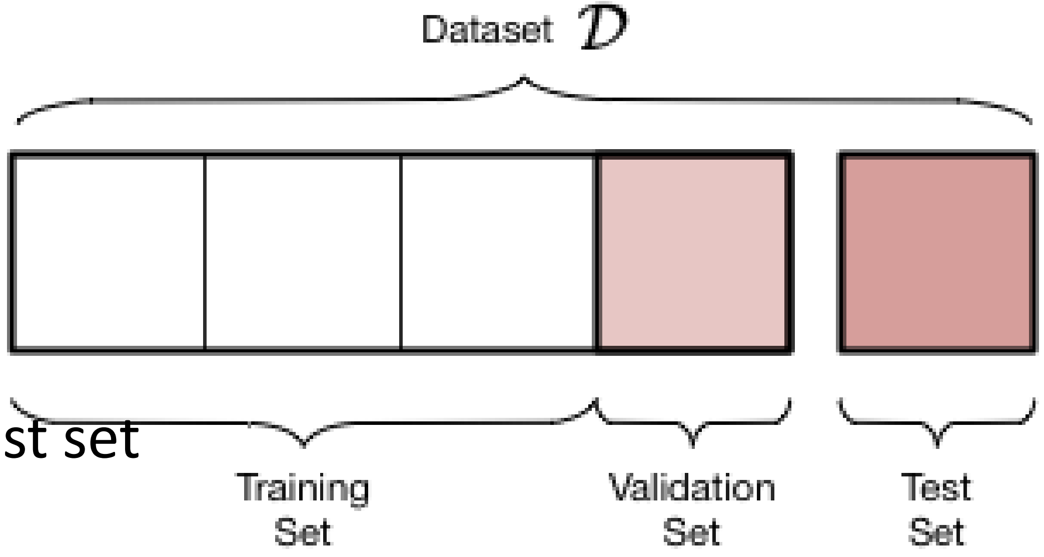
Generalization ability

- 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 to two (or three) parts
- Cross validation: later
- To evaluate models, a numerical „goodness" notion is needed



Training and test sets

- Dataset is divided to two (or three) parts
  - Training set: the model is trained on the training dataset (fitting the parameters) to ensure that it fits the training set well but also to have a good generalization ability
  - Test set: we test the fitted model on data that was not used for fitting
  - Validation set:
    if we have more models, we fit them on the training set, the fitted models are used to predict the targets for the observations in the validation set, finally we evaluate the best-performing final model on the test set



Regression

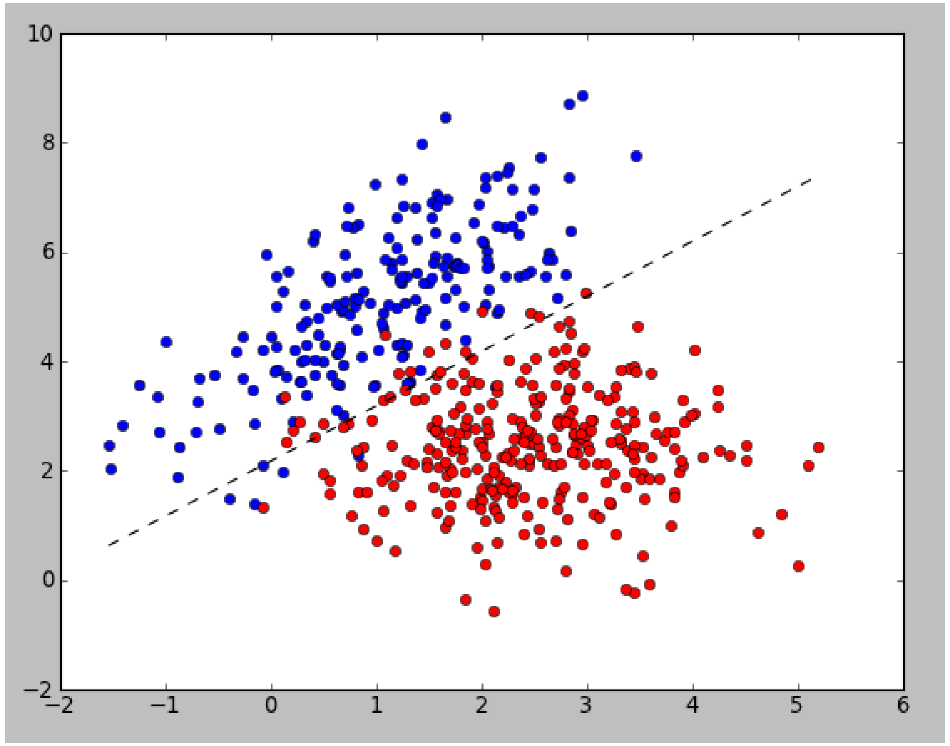
- We try to predict continuous valued output based on the values of other variables (via supervised learning)
- Examples:

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 vertical 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)

Anomaly detection

- Also know as: outlier detection
- If it is supervised, it is a special classification problem
  - Binary label: „good”, normal observation, anomalous observation
  - BUT: anomalous observations are rare ➔ different methods are needed
- Examples

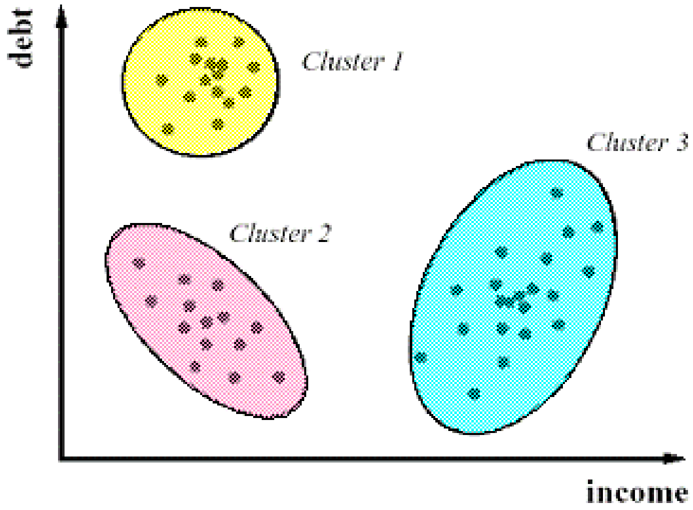
Input variables	Target variable
Amount and location of the purchase, purchase history,	Is it a stolen credit card transaction?
Memory usage, CPU usage, processor time, disk file access, …	Computer error occurred?

Clustering

- Unsupervised learning
- Grouping similar objects together
  - Aim: objects within a group should be more similar to each other compared to objects from different groups
- How to measure similarity? ➔ similarity measures (later)
- Challenges: What features are the clustering based on? How to measure similarity? How many clusters do we want? How to evaluate a clustering? How to visualize it?

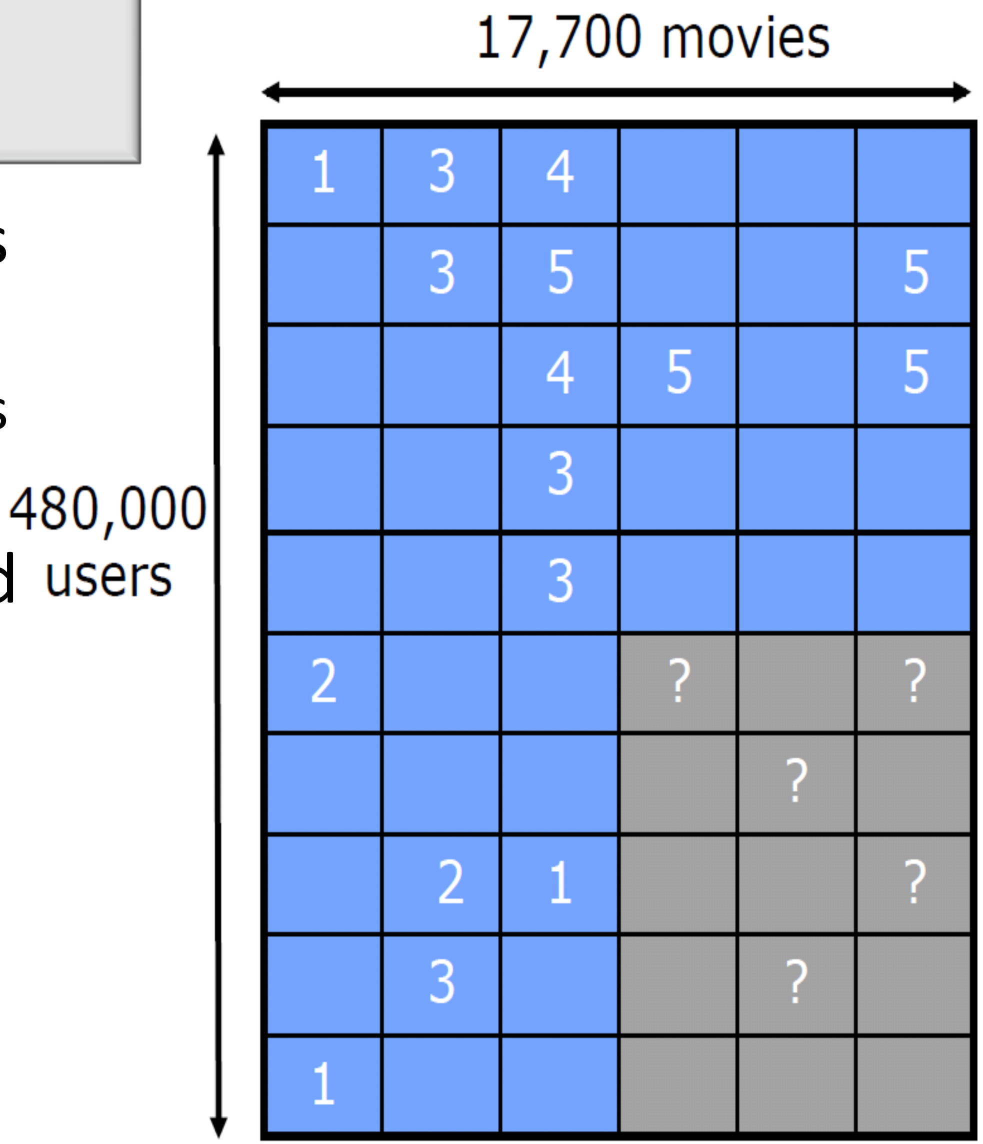
Clustering - examples

- Customer segmentation
  - Features: Customer data (sex, age, address, profession, …), purchase history
- Grouping documents based on their content
  - Features: words, n-grams appearing in the text
- Grouping pictures
  - Features: extracted from the pixels



Recommender systems

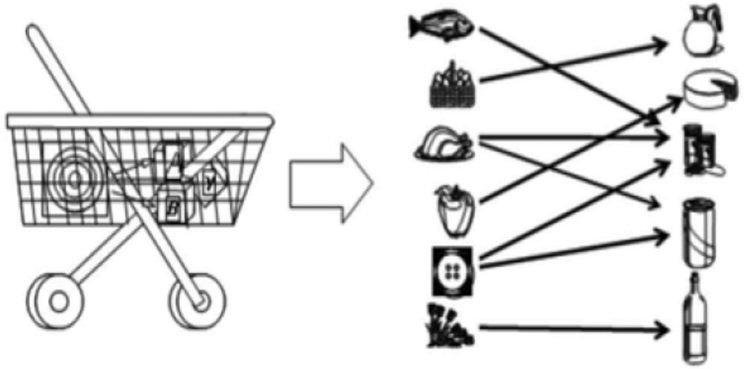
- Recommender system: it recommends items to users
  - Usually ranks the items and recommends items with high rank
- Known evaluations can be represented as a sparse matrix
  - Aim: predict the values in empty cells
    - Matrix factorization can be used



1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			?		?
				?	
	2	1			?
	3			?	
1					

Association rule mining (finding frequent item sets)

- A record contains a purchase transaction
  - E.g.: {bread, milk, diaper, beer}
- Aim: finding rules that describes if somebody buys {A_1, A_2, A_3, …, A_k} item set then it is likely that (s)he also buys item B
- The non-trivial rules are the interesting
- Benefit:
  - Pricing strategy
  - Improve retail store layout design
- Data mining started with this topic, then it went out of mind, nowadays it comes in fashion again

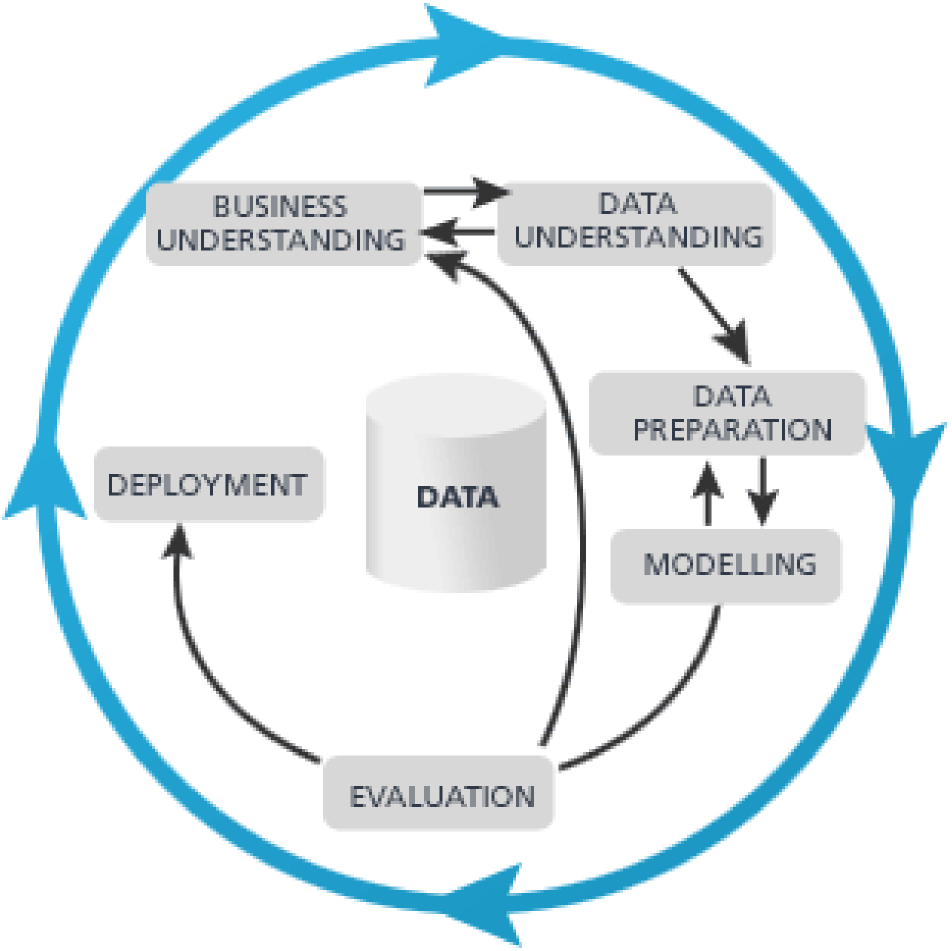


Process for data mining

- 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



CRISP-DM stages

- 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?
- 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. Data cleaning, finding relevant attributes.

CRISP-DM stages II.

- M -Modeling: Finding the best performing model, fitting the parameters.
- 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.

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

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.

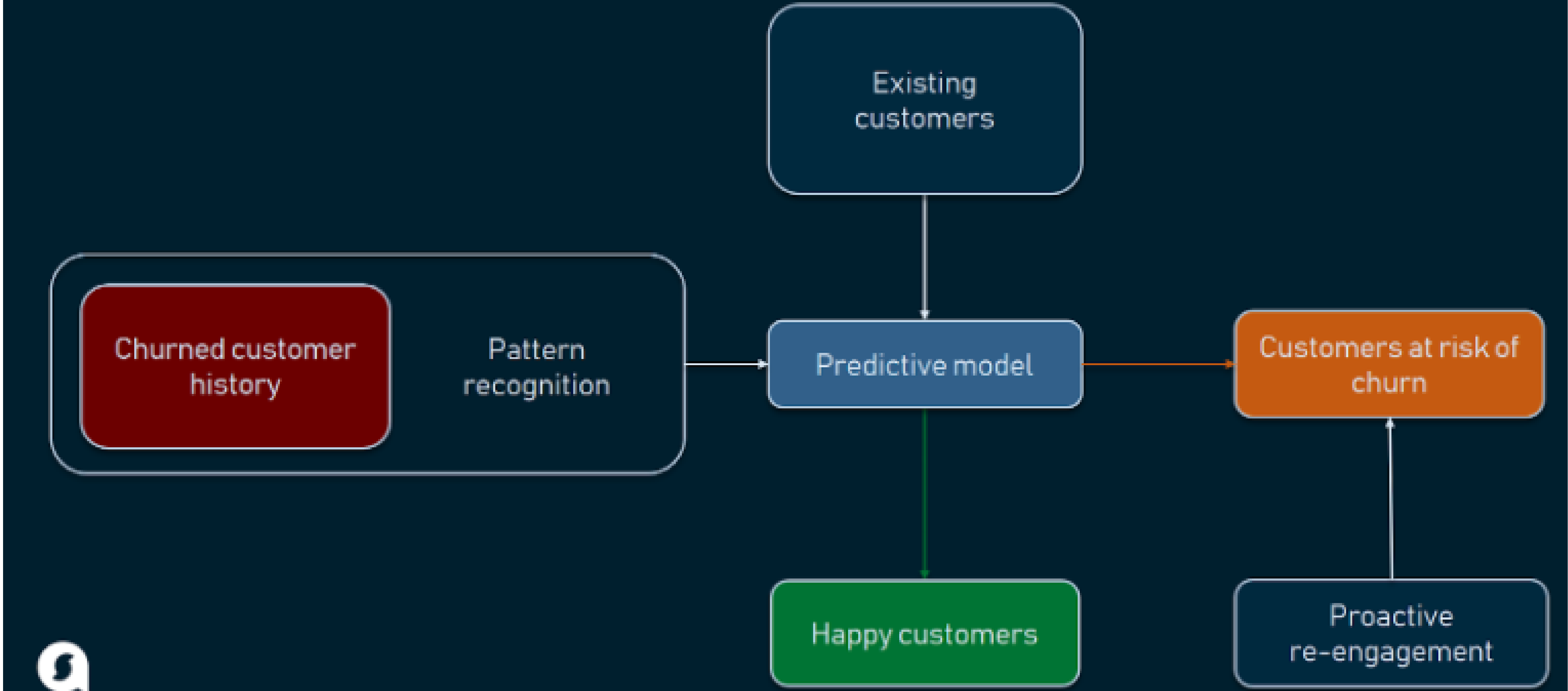
- Unbiased data
  - The data (the sample) should be representative to the population itself
  - BIBO: bias in, bias out
- The capacity of act
  - Sometimes the knowledge is discovered, but it will not go into action (high costs, too rigid system)
- Measurability of Return of Investment (ROI)
  - It defines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