Data Analysis and Knowledge Discovery

Data Analysis Process

Antti Airola

University of Turku
Department of Computing

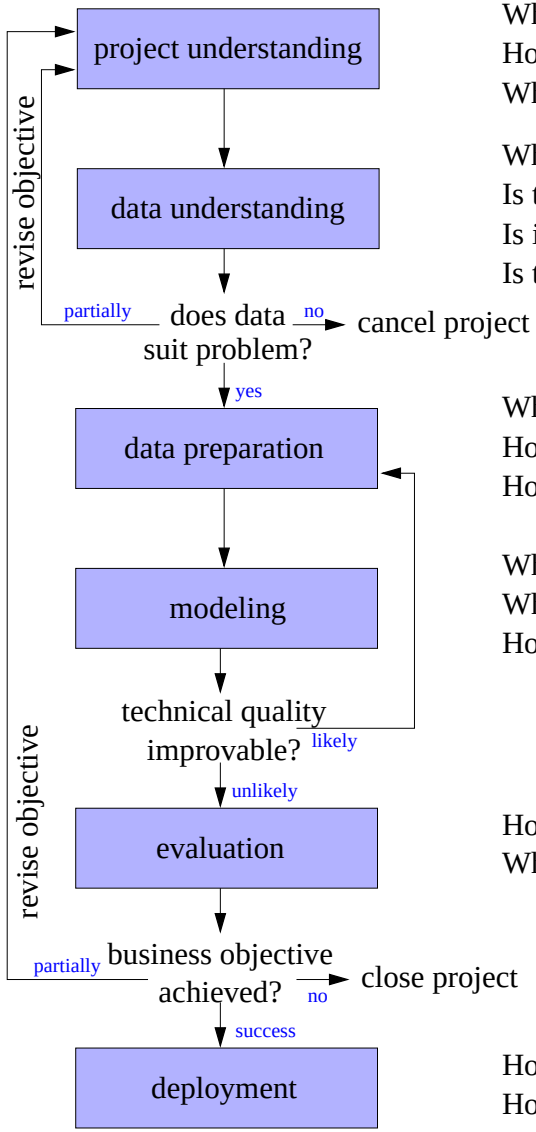
Antti.Airola@utu.fi

Outline

CRISP-DM Model

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- A commonly used methodology in planning data mining projects
- In the following, the term business understanding originally used in the CRISP-DM model is replaced by the term project understanding

CRISP-DM Overview



What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

1. Project understanding phase

- Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole
- Translate these goals and restrictions into the formulation of a data mining problem definition
- Prepare a preliminary strategy for achieving these objectives

2. Data understanding phase

- Collect the data
- Use exploratory data analysis to familiarize yourself with the data and discover initial insights
- Evaluate the quality of the data
- If desired, select interesting subsets that may contain actionable patterns

3. Data preparation phase

- Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive
- Select the cases and variables you want to analyze and that are appropriate for your analysis
- Perform transformations on certain variables, if needed
- Clean the raw data so that it is ready for the modeling tools

4. Modeling phase

- Select and apply appropriate modeling techniques
- Calibrate model settings to optimize results
- Remember that often, several different techniques may be used for the same data mining problem
- If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique

5. Evaluation phase

- Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field
- Determine whether the model in fact achieves the objectives set for it in the first phase
- Establish whether some important facet of the business or research problem has not been accounted for sufficiently
- Come to a decision regarding use of the data mining results

6. Deployment phase

- Make use of the models created: Model creation does not signify the completion of a project.
- Example of a simple deployment: Generate a report.
- Example of a more complex deployment: Implement a parallel data mining process in another department.
- For businesses, the customer often carries out the deployment based on your model.

CRISP-DM standard process

- Lessons learned from past projects should always be brought to bear as input into new projects
- In practice issues encountered during the evaluation phase can send the analyst back to any of the previous phases for amelioration

Case study: automobile warranty claims

- Quality assurance continues to be a priority for automobile manufacturers, including Daimler Chrysler.
- Jochen Hipp of the University of Tubingen, Germany, and Guido Lindner of DaimlerChrysler AG, Germany, investigated patterns in the warranty claims for DaimlerChrysler automobiles.

Project understanding phase

- Main objectives:
  - reduce costs associated with warranty claims
  - improve customer satisfaction
- Through conversations with plant engineers, who are the technical experts in vehicle manufacturing, the researchers are able to formulate specific business problems, such as
  - Are there interdependencies among warranty claims?
  - Are past warranty claims associated with similar claims in the future?
  - Is there an association between a certain type of claim and a particular garage?
- The plan is to apply appropriate data mining techniques

Data understanding phase

- Quality Information System (QUIS) contains information on over 7 million vehicles
- QUIS contains production details about how and where a particular vehicle was constructed, including an average of 30 or more sales codes for each vehicle.
- QUIS also includes warranty claim information, which the garage supplies, in the form of one of more than 5000 possible potential causes.

Data preparation phase

- It was found that the QUIS database had limited SQL access.
- They needed to select the cases and variables of interest manually, and then manually derive new variables that could be used for the modeling phase.
- For example, the variable number of days from selling date until first claim had to be derived from the appropriate date attributes.
- They then turned to proprietary data mining software, which had been used at DaimlerChrysler on earlier projects.

Data preparation phase

- Here they ran into a common roadblock-that the data format requirements varied from algorithm to algorithm.
- The result was further exhaustive preprocessing of the data, to transform the attributes into a form usable for model algorithms.
- The researchers mention that the data preparation phase took much longer than they had planned.

Modeling phase

- Since the overall business problem from phase 1 was to investigate dependence among the warranty claims, the researchers chose to apply the following techniques:
  - Bayesian networks
  - Association rules
- Bayesian networks model uncertainty by explicitly representing the conditional dependencies among various components
  - a graphical visualization of the dependency relationships among the components.
- The mining of associations is a natural way to investigate dependence among warranty claims

Modeling phase

- It was found that a particular combination of construction specifications doubles the probability of encountering an automobile electrical cable problem:
  - DaimlerChrysler engineers begun to investigate how this combination of factors can result in an increase in cable problems.
- The researchers investigated whether certain garages had more warranty claims of a certain type than did other garages:
  - Their association rule results showed that, indeed, the confidence levels for the rule "If garage X, then cable problem," varied considerably from garage to garage.

Evaluation phase

- The researchers were disappointed that the support for sequential-type association rules was relatively small, thus precluding generalization of the results
- Overall the researchers state:
  - "In fact, we did not find any rule that our domain experts would judge as interesting, at least at first sight."
- According to this criterion, then, the models were found to be lacking in effectiveness and to fall short of the objectives set for them in the project understanding phase.
- The researchers also pointed out to that the structures of the databases, were not designed for data mining:
  - They suggest adapting and redesigning the database to make it more amenable to knowledge discovery.

Deployment phase

- The project was counted as a pilot project, and as such, do not intend to deploy any large-scale models from this first iteration
- After the pilot project, however, they have applied the lessons learned from this project, with the goal of integrating their methods with the existing information technology environment at DaimlerChrysler.
- To further support the original goal of lowering claims costs, they intend to develop an intranet offering mining capability of QUIS for all corporate employees

Lessons learned from this case study?

- Intense human participation and supervision is required at every stage of the data mining process.
- Domain knowledge is essential.
- Regardless of what some software vendor advertisements may claim, you can't just purchase some data mining software, install it, sit back, and watch it to solve all your problems.
- There is no guarantee of positive results when mining data for actionable knowledge, any more than when one is mining for gold.
- When used properly, by people who understand the models involved, the data requirements, and the overall project objectives, data analysis can indeed provide actionable and highly profitable results.

Outline

Project understanding

Initial phase of the data analysis project

- Problem formulation
  - Objectives
  - Potential benefits
  - Constraints and assumptions (a priori knowledge)
  - Risks
- Mapping the problem formulation to a data analysis task
- Understanding the situation (available data, suitability of the data...)
- Average time spent for project and data understanding within CRISP-DM: 20 %
- Importance for success: 80 %

Project understanding

- Communication problems between domain and data analysis experts more of a rule than exception
- Project owner may not understand the technical terms of the analyst
- Analyst often does not understand well the domain
- Project owner may not understand the produced models or how to make use of them
- Formulating clear goals for the project difficult, requirements may have to be adapted later
- Is the customer committed to the project?
- Very similar to challenges encountered in standard software engineering projects...

Problems between domain and data analysis expert

problem source	project owner perspective	analyst perspective
communication	project owner does not understand the technical terms of the analyst	analyst does not understand the terms of the domain of the project owner
lack of understanding	project owner was not sure what the analyst could do or achieve models of analyst were different from what the project owner envisioned	analyst found it hard to understand how to help the project owner
organization	requirements had to be adopted in later stages as problems with the data became evident	project owner was an unpredictable group (not so concerned with the project)

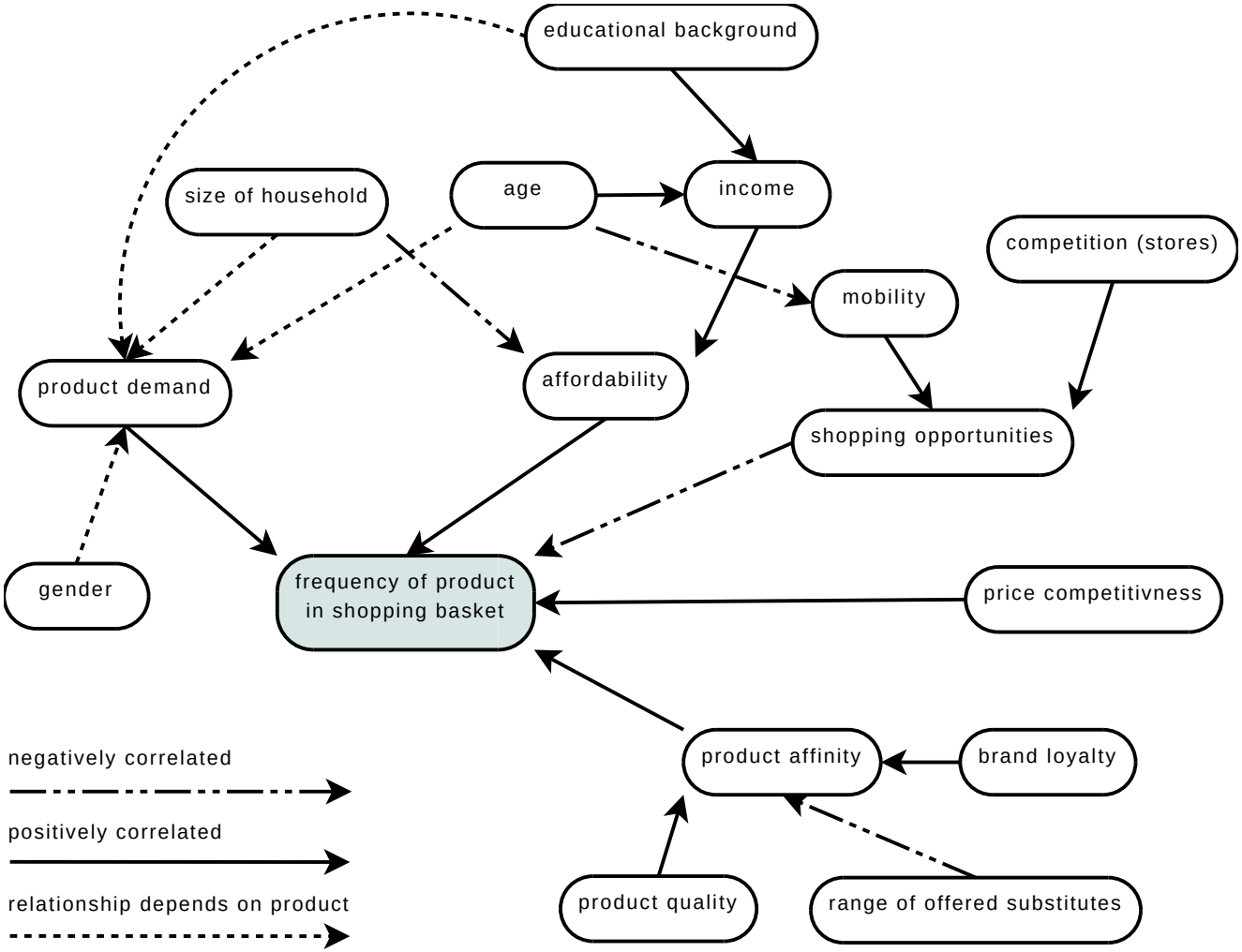
Determine the project objective

- Define
  - The aim of the project
  - Criteria to measure success
- Example
  - **Objective:** increase revenues (per campaign and/or per customer) in direct mailing campaigns by personalized offer and individual customer selection
  - **Deliverable:** software that automatically selects a specified number of customers from the database to whom the mailing to be sent, runtime max. half-day for database of current size
  - **Success criteria:** improve order rate by 5% or total revenues by 5%, measured within 4 weeks after mailing was sent, compared to rate of last three mailings

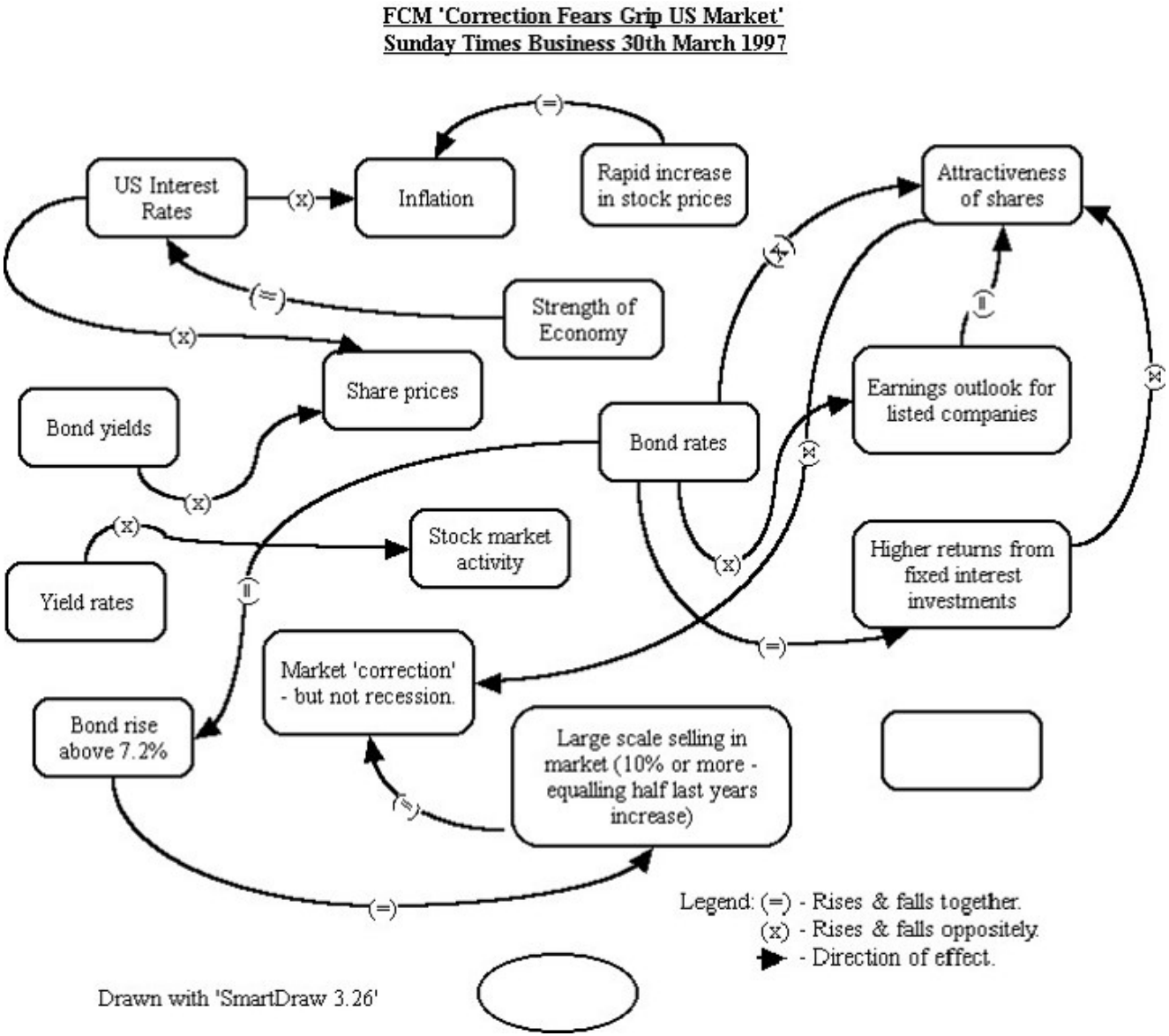
Cognitive maps

- Cognitive mapping, mind mapping, concept mapping,...
- A visual representation of problem that supports the domain understanding
- A way (tool) of capturing and structuring ideas visually
- Represented as nodes and arrows
  - Nodes: domain variables
  - Arrows: direction of influence of the variables and their type of influence (e.g. positive/negative correlation)

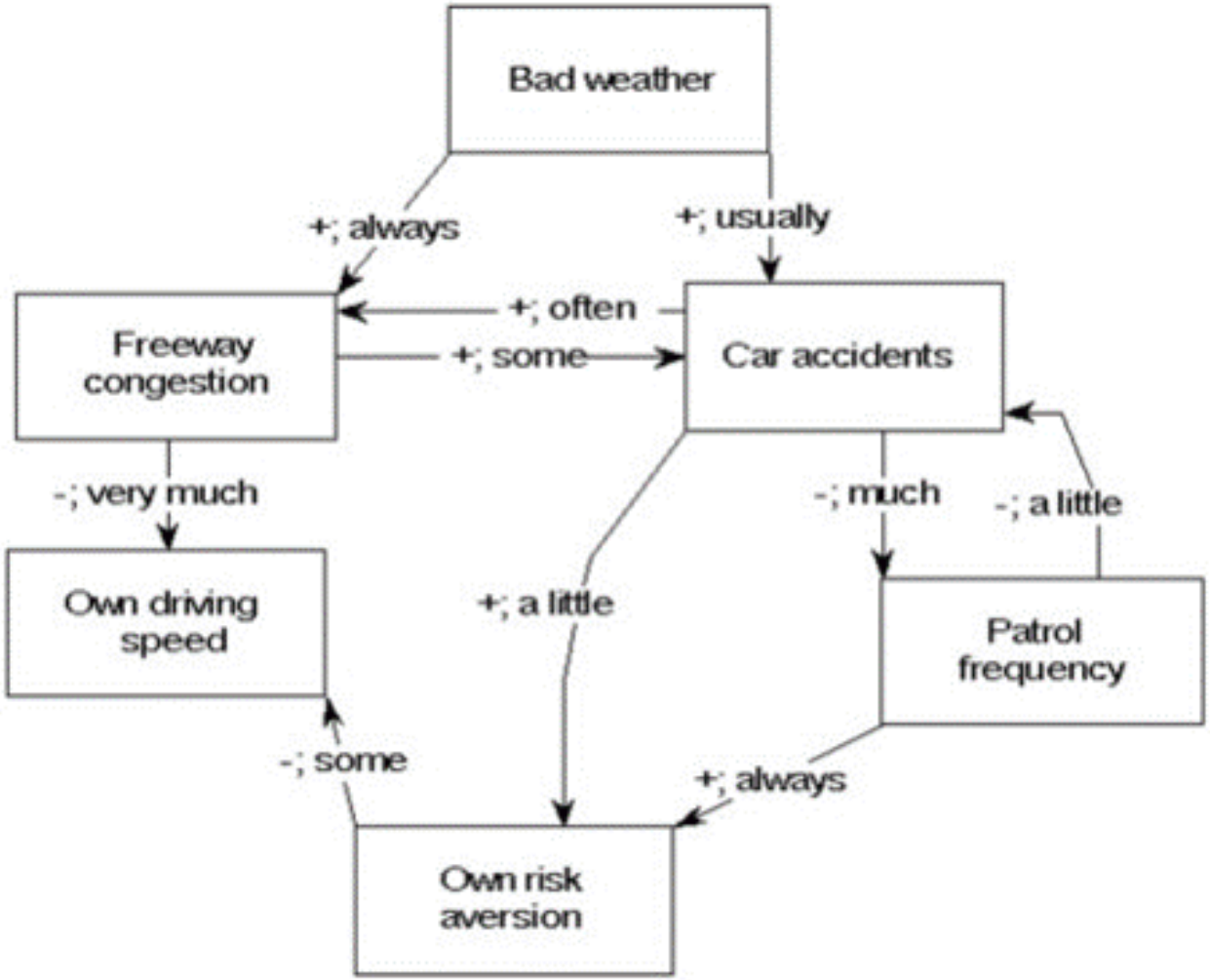
Mind maps


educational background
size of household
age
income
competition (stores)
mobility
product demand
affordability
shopping opportunities
gender
frequency of product in shopping basket
price competitivness
product affinity
brand loyalty
product quality
range of offered substitutes
negatively correlated
positively correlated
relationship depends on product

Mind maps


FCM 'Correction Fears Grip US Market'
Sunday Times Business 30th March 1997
US Interest Rates
Inflation
Rapid increase in stock prices
Attractiveness of shares
Strength of Economy
Share prices
Bond yields
Bond rates
Earnings outlook for listed companies
Stock market activity
Yield rates
Higher returns from fixed interest investments
Market 'correction' - but not recession.
Bond rise above 7.2%
Large scale selling in market (10% or more - equalling half last years increase)
Legend: (=) - Rises & falls together.
(x) - Rises & falls oppositely.
- Direction of effect.
Drawn with 'SmartDraw 3.26'

Mind maps


Bad weather
+; always
+; usually
+; often
+; some
Freeway congestion
Car accidents
-; very much
-; much
-; a little
Own driving speed
+; a little
Patrol frequency
-; some
+; always
Own risk aversion

Assess the situation

Requirements and constraints

- Model requirements
  - e.g. model has to be explanatory (because decisions must be justified clearly)
- ethical, political, legal issues
  - e.g. variables such as gender, age, race should be used with care
- technical constraints
  - e.g. applying the technical solution must not take more than seconds

Assess the situation

Assumptions

- Representativeness
  - If conclusions about a specific target group are to be derived:
    - The database should include a sufficiently large number of cases from this group
  - The database samples must be representative for the whole population
- Informativeness
  - To cover all aspects by the model, most of the influencing factors should be represented by the database attributes.

Assess the situation

Assumptions

- Good data quality:
  - The relevant data must be of good quality (correct, complete, up-to-date) and unambiguous thanks to the available documentation
- Presence of external factors:
  - We may assume (stationarity) that the external world does not change constantly , for instance
    - in a marketing project we may assume that the competitors do not change their current strategy or product portfolio at all.

Data analytics in practice

- In practical projects, it is quite usual that the data were not collected for the specific purpose of analysis
  - in contrast to experimental data where a suitable experiment is designed to collect the data.
- Example: Money transactions within a bank must be documented for safety reasons. However, valuable information can be deduced from such data, for instance, how much money might be needed daily for a new ATM
- Example: Electronic health records contain vast amount of data about patient diagnoses, treatments and outcomes. Could this data be used to make predictions about risk of complications after surgery?

Determine analysis goal

The primary objective must be transformed into a more technical data mining goal

- Determine data mining tasks
  - classification, regression, cluster analysis, finding associations, deviation analysis,...
- Specify the requirements for the models that will be constructed by the data mining tasks

Determine analysis goal

- Predictive accuracy
  - Is the goal of the analysis to produce a predictor that makes as accurate predictions as possible?
  - How should this be measure, are some mistakes more costly than others?
- Model flexibility/adequacy
  - A flexible model can adapt to more (complicated) situations than an inflexible model, which typically makes more assumptions about the real world and requires less parameters.
  - If the problem domain is complex, the model learned from data must also be complex to be successful.
  - With flexible models the risk of overfitting increases; the models learns noise, and it has not the best possible generalization ability to unseen data.

Determine analysis goal

- Interpretability
  - If the goal of the analysis is a report that sketches possible explanations for a certain situation, the ultimate goal is to understand the delivered model
  - For some models (e.g. neural networks) it may be hard to comprehend how the final decision is made and hence the models lack interpretability
- Runtime
  - If restrictive runtime requirements are given (either for building or applying the model), this may exclude some computationally expensive approaches

Determine analysis goal

- Interestingness and use of expert knowledge:
  - The more an expert already knows, the more challenging it is to surprise him/her with new findings.
  - Some techniques looking for associations are known for their large number of findings, many of them redundant and thus uninteresting.
  - Possibility of including any kind of previous knowledge
    - may ease the search for the best model considerably
    - may prevent us from re-discovering too many well-known artefacts.

Determine analysis goal

- The data mining process should produce a result that can be represented in a form that is understandable for the user of the data, not for a statistician.
- In the ideal case, the data mining methods are directly applied by the user of the data.
- In many cases, this does not work.
- Unfortunately there is no general purpose methodology

Data analysis and ethics

- Ethical issues arise in practical applications
- Anonymizing data is difficult
  - 85% of Americans can be identified from just zip code, birth date and sex
- Data mining can be used to discriminate
  - E.g. loan applications: using some information (e.g. sex, religion, race) is unethical
- Ethical situation depends on application
  - E.g. same information ok in medical application
- Attributes may contain problematic information
  - E.g. zip/area code may correlate with race
- Data analytics may be used for mass surveilance, feeding misinformation, military applications

Data analysis and ethics

- Important questions related to the law
  - Who has permitted access to the data?
  - For what purpose was the data collected?
  - What kind of conclusions can be legitimately drawn from it?
- When dealing with ethical issues purely statistical arguments are never sufficient
- Are resources put to good use?

EU Regulation

- EU Comission published on 21.4.2021 proposal for Artificial Intelligence Act
- Prohibited systems
  - Manipulative AI systems causing significant physical of psychological harm
  - Social scoring used disproportionally
  - Certain biometric recognition systems used for law enforcement
- Management based regulation
  - quality control, risk management, data governance systems
  - minimize risks to health, safety and fundamental rights
- Technology regulation
  - transparency, human control, event logs, accuracy, robustness, cybersecurity
- Additional regulation for high-risk AI systems