# Summary: Data Analytics in Organisations and Business

Fabian MARBACH, Autumn Semester 2015/16

## 1 Data Analytics

**Definition of data analytics** Data analytics is the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and add value.

**Facts about data analytics**

- The amount of data cannot anymore be grasped with he human brain nor proceeded for extracting the relevant information.
- Today's business world is changing from making decisions based on knowledge to **fact-based decisions**.
- The reasons for this development are: the world becomes more **global** and **complex** and there is an ongoing **decentralisation of data storage**.

### 1.1 Classes of analytics

**Descriptive analytics**

- Gathering and organising data
- Plotting the data and giving characteristics
- Classifying groups of data (e.g. customer groups)
- Reporting performance of organisations
- No statements about causes and consequences
- To determine influencing factors
- e.g. histograms, box plots, pie diagrams, bar plots, scatterplots, time series plots, tables with percentages, mean, median, quantiles

**Predictive analytics**

- Statistical modelling and forecasting
- Models and data from the past are used to forecast the future
- Associations among variables are identified and then the dependent variable is forecasted
- Causual effects are **not** necessarily assumed, since these are not always required to make accurate predictions
- e.g. Monte Carlo simulation, survival analysis
  e.g. to forecast market trends, to predict breakdowns of machines

**Prescriptive analytics**

- Optimisations, scenario testing and randomized tests
- Experimental design: makes causual inference by conducting experiments IOT find out why something happened
- Gives actions to perform
- e.g. to determine the optimal price of a product that maximizes profit, market share, margin etc.
  e.g. to test different marketing campaigns and marketing media against each other
  e.g. test and control design, Monte Carlo simulations, optimisation methods, operations research

**Quantitative analytics**

- Systematic empirical investigation of a phenomenon by statistical, mathematical or computational methods
- Data is collected in a structured manner out of a large representative sample and analysed
- *Methods:*
  Statistics, forecasting (estimation based on past data), data mining (automatic and semiautomatic extraction of patterns in large data sets via computational techniques), text mining (data mining focused on text), optimisation (mathematical techniques for optimal solutions), experimental design (test and control groups)

**Qualitative analytics**

- Gain an understanding of the underlying (qualitative) reasons or motivations for a behaviour
- Gain insight into causual effects from a behavioural perspective
- Collection of unstructured data of small and non-representative samples that are analysed non-statistically
- Part of exploratory research in early stages of an analytics process

## 2 Framing the Business Problem

How to frame the business problem:

1. Obtain or work out the **description of the business problem** and what should be the usability
2. Identifying all (direct and indirect) **stakeholders**
3. Analyse whether the business problem is **amenable** to an analytics solution
4. **Refinement** of the problem statement and if necessary depict known or possible **constraints**
5. Determine the **business benefits**
6. Obtain **stakeholder** agreement on the business problem statement

### 2.1 Description of the business problem

**Business problem statement** A description of a business problem that contains a description about the business opportunity or threat or an issue.

**Five W's**

- **Who** are the stakeholders who are . . .
  . . . sponsoring the project, using the results, making decisions based on the outcome, affected by the results?
  - **direct stakeholders:** affected by the actions and involved in the project but not necessarily part of the decision making process
  - **indirect stakeholders:** affected by the actions but neither involved in the project nor the decision making process
- **What** problem has to be solved?
  Perfect solution? What happens if the problem remains unsolved?
  What are the constraints? (e.g. financial, regulatory, organizational, political)
  What type of problem is it and to which area does it belong to? (e.g. economic, financial, marketing, strategic)
- **Where** does the problem occur?
  Where does the function requires to perform?
  (e.g. a price-cost problem)
- **When** does the issue occur?
  When does the project need to be completed?
  (Since actions normally require time to be implemented and to show effect, most problems should actually be solved by now.)
- **Why** does the problem occur?
  Why has this problem to be resolved?
  (Since the reason for a problem is usually unknown, data analytics has to answer also the why, which might be paramount for a solution.)

**Problem definition checklist**

1. **What problem are we addressing?**
   - *Importance of this question:*
     Having a clear understanding of the problem is key for success.
   - *Aim of this question/follow-up questions:*
     - definition of success and failure of the project
     - list of project deliverables, desired results and further requirement
     - state and quantify elements of performance
   - *Special considerations:*
     - Users might not clearly understand what data analytics is.
     - There are unspoken and maybe unreasonable expectations.
     - Consider as well questions from people who are evaluating the project and expectations of the sponsor or buyer.
2. **What would be the perfect solution?**
   - *Importance of this question:*
     This question directly implies unspoken and hidden expectations and can thus help omit these from the very beginning.
   - *Aim of this question/follow-up questions:*
     - What is the business case behind this problem?
     - What are the expected returns out of this project?
   - *Special considerations:*
     - Often there is some other unsolved problem.
     - There are some must-have features that drive the decision.
3. **How would you characterise the desired solution?**
   - *Importance of this question:*
     Everyone needs to understand the different solution possibilities IOT work together through this decision and to avoid solving the wrong problem.
   - *Aim of this question/follow-up questions:*
     - This answer drives the analytics process and methods used
     - What are special features needed as outcome?
   - *Special considerations:*
     - Most users have little knowledge about this aspect.
     - Look for indications that additional explanations are needed.
4. **What makes that problem difficult?**
   - *Importance of this question:*
     Often there are have been unsuccessful trials for solving the issue.
     Thus, find out what have been the methods that failed and why these methods failed.
   - *Aim of this question/follow-up questions:*
     - Knowing what has failed gives a lot of information about the problem.
     - Ask people who have worked before on that problem.
   - *Special considerations:*
     - Often, past failure is not fully understood.
     - Questions about past unsuccessful work can be problematic.
     - Be sensitive about the organisation's politics.
5. **What is the current level of performance?**
   - *Importance of this question:*
     The current level of performance is the lower bound of the result.
     Thus, results only as good as the existing are regarded as waste of effort.
   - *Aim of this question/follow-up questions:*
     - determination of the level of performance that is regarded as success
     - to ask for quantitative performance measures
   - *Special considerations:*
     - Often, people have unrealistic goals.
6. **What is good performance and what is bad performance?**
   - *Importance of this question:*
     Since just improving the performance does not mean success, the cost-benefit ratio also has to be considered.
   - *Aim of this question/follow-up questions:*
     - to find the benefit of an organisation
     - to ask for quantitative performance measures
   - *Special considerations:*

  - Distinguish between business answers vs. technical answers.

**Remarks:**

- *Likely responses to all questions above:*
  Often people tend to give vague, confusing and divergent answers to these questions and often they do not know what they actually want.
- Since vague responses and room for interpretation lead to failure, one has to press for more detailed answers in such cases.

**Conclusions**

- The full understanding of the problem is the most important aspect and is guiding the whole analytics process.
- Keep in mind that not all business problems are amenable to approaches employed by data analytics and thus might need very different approaches.
- Review previous analyses of the problem.
- Think about how the problem has been structured so far and how it should be newly structred.

## 2.2 Identification of all stakeholders

**Stakeholder analysis worksheet/checklist**

- **Identifying all stakeholders**
  Is it clear which executives have a stake in the success of the project?
- **Documenting stakeholder needs**
  Have they been briefed on the problem and the outline of the solution?
  What are the needs of each and every stakeholder?
  Do they understand the problem and the possible solution? If not, what should be the follow-up questions?
  Are there any special considerations? Which ones?
- **Assessing and analysing stakeholder interest/influence**
  Do they have the ability *to provide the necessary resources* and *to bring about the business changes* needed to make the project successful? (e.g. budget, personnel, IT)
  What are the stakeholders' interest and/or influence?
  Who would be a supporter?
  Who could have a potential negative impact on the project?
- **Managing stakeholders' expectations**
  Do they *generally support* the use of analytics and data for decision making?
  Are they clear about the possibilities and limitations of data analytics?
  Are all expectations realistic?
- **Take actions**
  Does the proposed analytical story and method of communication coincide with their *typical way of thinking and deciding*?
- **Reviewing status and repeating**
  Do you have a plan for *providing regular feedback and interim results to them*?

*Remarks:*

- The staff often perceives such data analytics projects only as job cutting exercises.
- Most executives are too focused on their own area and take their area too important.
  (e.g. CFO's often request cost cutting while marketing officials request more budget)
- There is often an expectation mismatch if executives are prejudiced.
- The most important question IOT help stakeholders to frame their needs and expectations is:
  What is the **decision** they want to make as a result of the analyis?

## 2.3 Amenability of a business problem to an analytics solution

**Cost-benefit analysis**

- Is the answer of the analytics process and the implementation **within the organisation's control**?
  Or are there external factors (e.g. of the economy, regulations) that prevent implementation?
- Would be **data available** to perform the analysis?
  (e.g. some (detailed) past data that needs to be available)
- How likely is it that the problem **can be modelled and solved**?
  (e.g. information about economic interactions/possible influences)
- Will the organisation **accept the solution** and deploy it?
  (e.g. if the implementation would require the stakeholders to admit past failure in decision making)

## 2.4 Refinement of the problem statement and constraints

**Refinement/redefinition of the problem statement** It may be necessary to refine or redefine the problem statement (i.e. the problem, expectations, needs, decision making process, tools, data) IOT:

- make it more accurate and appropriate to the stakeholders.
- make it more amenable to available analytical tools, methods and data.

**Constraints** Define the analytical, financial or political constraints the project will operate under.

- *Analytical:* e.g. optimisation problem with either no solution (with all constraints satisfied) or inappropriate results (with weaked constraints)
- *Financial:* e.g. bad cost-benefit analysis
- *Political:* e.g. may the results reveil past failure in decision making?

**Worksheet/checklist for framing the business problem**

1. Have you defined a **clear problem or opportunity** that is important to your organisation?
2. Have you considered **multiple alternative ways** to solve the problem?
3. Have you identified the **stakeholders** and their will to use the results to make a decision?
4. Will the way to solve the problem **resonate with the stakeholders**?
5. Is it clear **what is the decision to be made** and **who will make it**?
6. Have you started with a broad definition of the problem but then narrowed it down to a very **specific problem** with a clear phrasing of the question, the **data to be applied** and the **possible outcomes**?
7. Are you able to describe the type of **analytical story** you want to tell in solving this question?
8. Do you have **someone who can help you** in solving this particular problem?
9. Have you looked systematically to see whether there are **previous findings or experiences** related to this problem, within or outside your organisation?
10. Have you **revised your problem definition** based on what you have learned from previous findings?

## 2.5 Determining business benefits

- A project will only be conducted if it brings more benefits that it costs (e.g. if there is a positive ROI).
- Different **business benefits:**
  - *Qualitative benefits:* e.g. transparency w.r.t. certain figures/spendings, to work out actionable improvements, to support communication, to help some executives keeping their position
  - *Quantitative benefits:* key argument: additional revenues
- Usual **measures of financial analyses:**
  Return on Investment (ROI), Net Present Value (NPV), Internal rate of return (IRR), Cost of Capital (CoC), Payback period
- **Special considerations:**
  - additional cost/revenues
  - How can assumptions for scenarios be verified?
  - Are there any options which have not been considered?
- It is often very difficult to quantify the return or cash flows out of data analytics results.
  - e.g. cost of the IT infrastructre used and the support of the IT experts?
  - e.g. cost of the people involved in the project and for providing support?
  - e.g. how to quantify better compliance with regulations?

## 2.6 Obtaining stakeholder agreement on the problem statement

**Written stakeholder agreement**

- the project's **objectives** (and clarification)
  - **affirmative:** e.g. often analysis of a decrease in revenues, sensitivity analysis, time of impact between a certain action and its impact (e.g. an increase in revenue), suggestion of actionable results
  - **distinction:** only analysis vs. implementation of the actions
  - **negative:** results provide options with a high likelihood of impact (in contrast to actions with a sure impact), if only parts of the project can be performed then the likelihood of success decreases
- the **definition of the problem**
  - e.g. a deterioration of revenue or another performance measure, inconsistencies in figures and related strategies
  - distinguish between indisputable facts and speculations of the stakeholder (e.g. the management)
  - mention the given time frame (e.g. in the next few weeks actionable results have to be produced)
- the **resources**
  - distinguish between own resources and provided resources of the stakeholder
  - 1x data analytics professional costs app. 10,000–15,000 CHF per week
- the **time frame**
  - first results, final results (e.g. 3/5 weeks)
  - implementation of the actions (e.g. in weeks 3–8)
  - expected visible outcomes of the actions
    (e.g. in marketing app. 4 months after implemention)
- the **performance measures**
  - project measures: e.g. adherence to the project plan and project deliverables, can be measured during the project
  - measures of actionable results: only after a reasonable time period measurable, define measure
- the **budget** to get there
  - budget for data analytics team and budget for resources provided by the stakeholder
  - conduct some plausibility checks
- further **considerations** and **constraints**
  - A good project planning is required if many resources within a short period of time need to elaborate many deliverables.
  - Resources of the stakeholder need to be informed about their tasks, their being available and the time frame.
  - Depending on the time frame, one maybe has to focus on trends rather than detailed analyses.
  - Often also external data (e.g. from a census office, the internet) may be considered.
  - Often a rather holistic approach is chosen in contrast to a narrow cost-benefit analysis.

# 3 Framing the Analytics Problem

How to frame the analytics problem:

1. Translate the business problem statement into an **analytics problem**
2. Propose a set of **drivers and relationships to inputs**
3. State the set of **assumptions** related to the problem
4. Define **key metrics of good performance**
5. Obtain **stakeholder agreement** on the approach

## 3.1 Translating a business problem statement into an analytics problem

One has to translate the "what" of the business problem into the "how" of the data analytics problem

**(Un-)supervised analytics problems**

- **Unsupervised** analytics problem: No specific purpose nor target
- **Supervised** analytics problem: Specific target, segmentation for a specific goal for actionable results

### 3.1.1 Quality function deployment

**Quality**

- Quality must be designed into the product.
- Quality is meeting customer needs and providing superior value.

**Quality function deployment** Quality function deployment is a systematic approach to design products ...

- based on **customer needs and desires**.
- with the **integration of the different functions** within a company (e.g. marketing, business, accounting, controlling, manufacturing).
- is used to translate often subjective quality criteria into **objective characteristics**.
- which can then be **measured and quantified**.

**Phases of quality function deployment**

1. **Product planning** / House of quality
   Documentation of the customer requirements, competitive advantage, product measurements, technical ability of the organization, does the product meet customer needs?
2. **Product design**
   Design phase (creativity and innovation are of importance); development of product concepts and parts of the specifications are defined and documented
3. **Process planning**
   Flowcharts of the intended processes and the process parameters (or target values) are documented
4. **Process control**
   Performance indicators are set up to monitor the production process and maintenance; decisions about which process poses the most risk and controls are put in place IOT prevent failures

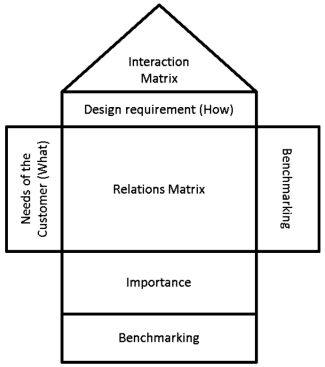



Figure 1: House of quality as in the quality function deployment method.

**House of quality**

1. Define the **customer needs** (what).
   - e.g. What are actionable results for improving ...?
   - e.g. What is the influence per region to ...?
   - What is the cost benefit ratio per region?
   - What is the influence of the marketing budget on ...?
   - e.g. Project has to be completed in $X$ weeks.
2. Define the **analytics project requirements** (how).
   - e.g. fast methods which reveal trends vs. detailed analyses
   - e.g. appropriate data (internal or external)
   - e.g. sufficient resources for performing the project
   - e.g. actionable results within time constraint
   - e.g. proper project planning (resources & time constraint)
3. Link them/determine their relationship, i.e. define the **relations matrix**.
   - 0: no relationship, 1: weak, 2: medium, 3: strong
   - e.g. 3 between *questions asking for causes / influences / relations* and *methods and data*
   - e.g. 2–3 between *actionable results* and *appropriate data*, as well as between *actionable results* and *project planning*
   - e.g. 2–3 between all customer requirements and project planning
   - e.g. 2–3 between project completion and project planning
4. **Benchmark** the customer needs (what) against the *current status*, *full project* and *parts of the project*.
   - e.g. *current status* generally low but high for project completion
   - e.g. *full project* generally high but low for project completion
   - e.g. *parts of the project* only high for the part(s) it covers and project completion
5. Define the **interaction matrix** (interactions of the design requirements (how)).
   - (– strong negative, - negative, 0 neutral, + positive, ++ strong positive relationship)
   - generally high interactions/correlations, i.e. + to ++
6. Determine the **importance** of the design requirements (how).
   - generally high importance, i.e. 60–100%
   - e.g. 90–100% of actionable results and project planning
7. **Benchmark** the design requirements (how) against the project options.
   - e.g. current status generally low but high for sufficient resources
   - e.g. full project generally high but low for sufficient resources
   - e.g. parts of the project quite similar to current status but a bit more balanced

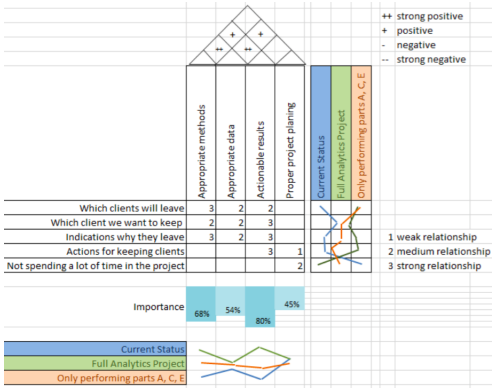



Figure 2: Example of a house of quality.

**Requirements**

- **Expected or basic requirements:** are considered as given and are often unspoken
- **Normal or performance requirements:** directly mentioned by the customer and which thus can be verbalised
- **Exciting or emotional requirements:** reflect a need that the client has not appreciated before

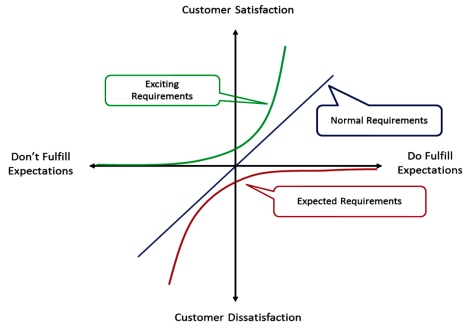



Figure 3: Customer requirements as in Kano's model and the quality function deployment model.

**Conclusion**

- Very rigorous process that can be applied to any data analytics process
- Maps the translation or requirements from the business level to the following analytics levels
- Requires time
- Recommended for larger projects where a detailed design and process have to be set up in advance with high reliability

### 3.1.2 Kano's model

In the Kano model, a **four step approach** is carried out.

**Requirements**

- **Expected requirements** or **must-be requirements**
  Although not explicitly stated they are taken for granted and constitute basic attributes of each product; if not fulfilled they lead to customer dissatisfaction and if fulfilled they may only lead to non-dissatisfaction
- **Normal requirements**
  Connect the customer satisfaction and the level of fulfilment proportionally; they are usually explicitly demanded by the customer
- **Exciting requirements** or **attractive requirements**
  These are the differentiating factor; have the greatest influence on customer satisfaction; are usually not explicitly stated nor expected; give additional experience with a product; if not fulfilled there is no dissatisfaction

**1) Identification of product requirements**

- What are the associations of the customer when using the product?
  (e.g. high expectation of immediate actionable results, or expecting just a simple cost analysis)
- Which problems are associated by the customer with the use of the product?
  (e.g. tight time schedule, openness to new insights vs. predefined opinions, supporters vs. non-supporters)
- Which criteria are taken into account when buying the product?
  (e.g. focus on strategic & organizational impact (economic buyers), focus on cost, focus on analytical results (technical buyers))
- Which new features or services would better meet the expectatioins of the customer? Or: What would the customer change in the product?
  (e.g. quick analyses with actionable outcomes ("quick wins"), just cost analyses, confirmation of existing opinions)

**2) Construction of the Kano questionnaire**

- A pair of questions is formulated for each product/service feature:
  A **functional form** which concerns a reaction if a product has a feature
  A **dysfunctional form** which conerns a reaction if a product does not have a feature
- *Possible answers:*
  I like it that way / It must be that way / I am neutral / I can live with it that way / I dislike it that way
- *Examples:*
  - *Actionable results:*
    If we have within two weeks only actionable results which additionally require an increase in the marketing budget in the short run, how do you feel?
    If we have within two weeks no actionable results which require an increase in the marketing budget in the short run, how do you feel?
  - *Willingness to action on the results:*
    If you have actionable results from the analysis which require a change of your corporate strategy, how do you feel?
    If you have no actionable results from the analysis which

require a change of your corporate strategy, how do you feel?

– *Full project vs. parts of the analysis:*
If the full data analytics project is performed (with all costs, resources and time), how do you feel?
If not the full data analytics project is performed (with all costs, resources and time), how do you feel?

**3) Getting the data from the customers**

- Standardised questionnaire vs. interview
- Mail vs. online questionnaire
- Customer panel vs. randomised potential customer

| Customer requirements | | Dysfunctional question | | | | |
|---|---|---|---|---|---|---|
| | | Like | Must-be | Neutral | Live with | Dislike |
| Functional questions | Like | Q | A | A | A | N |
| | Must-be | R | I | I | I | M |
| | Neutral | R | I | I | I | M |
| | Live with | R | I | I | I | M |
| | Dislike | R | R | R | R | Q |

Figure 4: Categorization IOT calculate the customer satisfaction index.

**4) Analyse the results**

- **Categories:**
**A:** attractive, **N:** neutral, **M:** must-be, **R:** reverse, **I:** indifferent, **Q:** questionable
- Evaluation order: If there is no clearly dominant category:
M > N > A > I
- **Customer satisfaction index:** how strongly a product feature may influence satisfaction or dissatisfaction:

$$\text{Satisfaction} = \frac{A+N}{A+N+M+I}$$

$$\text{Dissatisfaction} = -\frac{N+M}{A+N+M+I}$$

- Plot results as frequencies.
- The level of **satisfaction** quantifies the impact of **fulfilling** a certain requirement on overall satisfaction.
- The level of **dissatisfaction** quantifies the impact of **not fulfilling** a certain requirement on overall satisfaction.

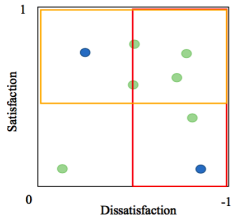



Figure 5: Example plot of the correlations between satisfaction and dissatisfaction in the Kano model.

## 3.2 Set of drivers and relationships to inputs

- Main goal: define **inputs** and **outputs**
- This can be conducted **informally** or **formally**
"Formally" is prefered since it supports communication with the stakeholder.
- At this stage one is not looking for causual relationships, one is just collecting ideas IOT build up hypotheses (against which the model can later be tested)
- Determine increasing or decreasing factors (and evaluate their impact on scale).
- Draw a **Black-box sketch**
Make the inputs visible s.t. it can be discussed and challenged, e.g. by technical experts.

## 3.3 Set of key assumptions related to the problem

- In each input/output factor there are **assumptions** implied.
- These are often **common practice assumptions** which are nowhere explicitly stated.
- In a project all assumptions have to be listed and assessed.

## 3.4 Defining key metrics of good performance

- The key metric always have to be **measurable**.
- *Examples:* revenue has to be improved by X%, no net loss of customers, the production process has to be Y minutes faster per quantity, . . .
  - *Analysis of a supply chain w.r.t. supply chain risk:*
avoidance or reduction of possible business interruption (e.g. to a certain number of days), certain level of manufacturing material available, price stability of manufacturing material
  - *Analysis of insurance claims for fraud cases:*
amount of detected fraud cases where the insurance company does not have to pay for claims
  - Reduction of the amount of chemicals in a production process without loss of quality and efficiency:
amount of chemical used per quantity compared to a certian threshold, processing time per quantity, number of quantity that does not pass the quality check
  - *Improvement of transport routes of a cargo firm:*
reduction in tonne-kilometre while delivering the same or more services/transports, delivering the same tonne-kilometre with less resources
  - *Analysis of non-compliant product sales of a bank:*
number of non-compliant product sales found per $X$ customers/per client advisor $Y$
  - *Improvement of the production process of a manufacturing company:*
production cost per quantity, production time per quantity, production capacity

## 3.5 Obtaining stakeholder agreement on the approach

- This can either be separated from the business problem statement or be integrated.

# 4 Data

**What has already been done**

- *Business problem statement:* first assessment of available data, the "what"
- *Data analytics problem statement:* the "how", set of drivers (i.e. input variables and outcomes), set of relationships between the variables

## 4.1 Identification and priorization of data

**Identification and priorization**

1. Decide about the **characteristic** of the input variable (e.g. mean, distribution, text, ect.).
2. Determine **which data can cover this characteristic** (e.g. personal data for determination of the age/distributions of age).
3. Determine which **data types** are most preferable and prioritize them.
4. List all **available data** one already has or one knows they are available.
5. In case of **missing data**, either create an inventory of **additional data** together with the customer or collect it directly.
6. If data are neither available nor collectable in due time, one has to **refine or redefine** the analytical or the busines problem eventually.

*Remarks:*

- One must become clear about the type of data one needs and one can finally receive/collect.
- A thorough understanding of the data is paramount.

**Data definitions**

- **Population:** includes all the members/items of interest in a study (i.e. the population w.r.t. a data analytics project rarely coincides with the population in the common sense)
- **Sample:** a subset of the population; is determined randomly and is preferably representative of the whole population
- **Data set:** (of structured data) is usually an array of data with variables in columns and observations in rows
- **Variable:** (or field or attribute) is a characteristic of the items of the populaltion
- **Observation:** (or case or record) is a list of all variable values for a single member/item of a population

### 4.1.1 Data types

**Hard data vs. soft data**

- **Hard data:** is collected by scientific observation and measurement (e.g. experiments)
- **Soft data:** is explored from inverviews and reflects subjective opinions and preferences
- *Translation from soft data into hard data:* develop a set of rules to achieve the same behaviour
(e.g. operating room optimization: from inverviews with surgeons and management to a computer simulation)

**Numerical vs. categorical data**

- **Numerical data:** meaningful arithmetical operations can be performed
(e.g. it is possible to compute the mean and standard deviation)
**Discrete vs. continuous data:**
  - **Discrete data:** results from a count (e.g. number of customers buying a certain product)
  - **Continuous data:** results from an essentially continuous measurement (e.g. the waiting time of a patient)
  - **Binned data:** was categorized into discrete categories, i.e. bins
- **Categorical data:** otherwise
**Ordinal vs. nominal data:**
  - **Ordinal data:** there is a natural ordering of its possible categories
  - **Nominal data:** otherwise (i.e. there is no such ordering)

**Cross-sectional data vs. time series**

- **Cross-sectional data:** on a cross-section of a population at a distinct point in time, i.e. without any time-dependeny
- **Time series:** collected over time (e.g. stock quotes)

**Structured vs. unstructured data**

- **Structured data:** can be put into rows and columns
- **Unstructured data:** otherwise (e.g. a text or a movie)
- However, data analytics can only be applied to structured data (since its algorithms cannot directly access unstructured data). Thus, unstructured data first has to be structured IOT make it an input for a data analytics process.
(e.g. text mining, big data (hadoop, map-reduce), matching fingerprints)

**Primary vs. secondary data**

- **Primary data:** is not yet available and has to be measured and collected first
(e.g. new reporting data (due to new production processes), new regulatory data (e.g. on how to quantify expected credit losses on financial instruments))
- **Secondary data:** has already been collected by someone else (e.g. internet, accounting, statistical data (e.g. census), log data in IT systems)
  - Advantage of secondary data: safes time as data is already available; some data is already structured and cleaned
  - Disadvantage of secondary data: may be outdated; may already be processed (or manipulated)

**Meta data**

- Data about the data, i.e. describes the data
- (e.g. data when data was collected, purpose of the collection, how the data was collected, size/volume of the data, image resolution, dates of changes in the data, dates of access to the data, tags in social media)

**Dummy variable**

- A 0-1 coded variable for a specific category.
- 1 labels the observations in this category and 0 labels all other observations not in that category.

**Examples**

- *book prices on amazon:* hard, numerical, discrete, time series, structured, primary, meta
- *multiple choice questionnaire:* soft, categorical, ordinal/nominal, cross-sectional, structured, primary, meta
- *Twitter tweets:* soft, categorical, nominal, cross-sectional/time series, unstructured, primary, meta
- the following examples may comprise almost all types of data: claims & customer data of an insurance company, clinical tests, web page access

## 4.2 How to collect and get data

**Decision to collect data**

- One has to identify which kind of data collection will have the most favourable impact on the quality of the actions and recommendations supported by data analytics.
  This is typically done by a **decision tree**.
- Without data gathering, a change will be made if:

$$pU + (1-p) = p(U-L) + L > u$$

where:

$p$ probability of getting a good outcome, if a change is made

$U$ value (utility) of making a change with good outcome

$L$ value (utility) of making a change with bad outcome

$u$ value (utility) of continuing with the present course

- But: one can gather data and make decisions based thereon.
- While the value of information is nonnegative, gathering data also incurs cost.
- Data will be gathered if:

$$d(qu^* + (1-q)u) > u$$

where:

$p^*$ probability of getting a good outcome, if getting favourable information

$p^{**}$ probability of getting a good outcome, if getting unfavourable information

$p$ probability of getting a good outcome, if a change is made
$p = qp^* + (1-q)p^{**}$

$q$ probability of getting favourable information

$u^*$ value (utility) of implementing a change, given getting favourable information:
$u^* = p^*(U-L) + L$

$u$ value (utility) of continuing the present course, given getting unfavourable information

$d$ factor by which cost decrease utility

- $u^*$ depends on how much $p^*$ differs from $p$. Thus, the degree of change of $p$ with the new information dpends on:
  - the confidence we have in the original value $p$
  - the impact of the data
- Finally, the decision to collect data and which data is very **subjective**, is biased on our **beliefs** and depends on how confident we are in the original probability $p$.

### 4.2.1 Data collection

**Overview: Data collection**

1. **Sample design:** determine how to identify the subjects/items
2. **Sample plan:** determine how many subjects/items to identify
3. Determine the **questions** to be asked
4. **Granularity of the experiment:** determine the possible answers to the questions
5. Determine the **control group**

**1. Sample design**

- **(Simple) random sampling**
  - To each subject/item the same chance is allocated to be selected for the sample.
  - Advantage: unbiasedness
  - Disadvantage: if events/subjects/items are very unlikely, it is also very unlikely to have them in a (simple) random sample
  - Thus, it may be advantageous to bias the sampling towards those subject that are of interest.
    Nevertheless, the analysis has to take into account this bias and the formulas have to be corrected for this bias.
- **Stratified random sampling**
  - divide population into homogenous subpopulations (strata)
  - strata have to be mutually exclusive (i.e. as partitions)
  - to each stratum sampling is applied
  - correction for bias has to be implemented
  - e.g. testing the correctness of Swiss DRG cost code allocations (health care)
- **Full factorial design:** gives the possibility to identify the impact of each factor as well as of possible interactions between the factors in an efficient manner
  (e.g. regression analysis, ANOVA)

**2. Sampling plan**

- **Depends on:**
  - Amount of uncertainty
  - How much does this uncertainty need to be reduced?
  - Degree of error contained in the responses of a subject/item
- **Candidates:** width of the confidence interval, confidence level, power of a test
- Usual **assumptions:**
  - *Independence:* often given in practice
    e.g. buying decisions of customers, car accidents
  - *Identically distributed:* often not given in practice
    Thus, homogenous subclasses are needed for the analysis, but then the sample might be too small for performing meaningful data analytics
  - *No systematic error:* very difficult to detect in practice
    e.g. drifts or if the subsample is very small

**3. Determination of the questions to be asked / 4. Granulatiry of the experiment**

- **Primary data:** interview questions and questionnaires
  - *Categorical:*
    Format: {black, silver, red, white, ...}, {yes, no}, {fully agree, partially agree, ...}
    e.g. preferred characteristics of a new product (such as preferred colour)
    e.g. to find out how many foreign languages people are speaking
    e.g. to find out what people think about the political agenda of a party
  - *Semantic differential:*
    Format: user friendly vs. cumbersome, slow vs. fast, confusing vs. clear
    e.g. to determine personality traits
  - *Rank-order:*
    e.g. to rank the most important benefits of a product
    e.g. to find out the five most important benefits of a new product
    e.g. to find out the most and least useful feature of a new product
  - *Multiple choice:*
    e.g. to test analytical thinking of job candidates
- **Secondary data**
  - Since secondary data is already available, one has to find and assess it for appropriateness.
  - Questions to be asked:
    kind of data, data inventories, source of data, primary purpose for collecting the data, who collected the data, when was it collected, raw data vs. cleaned data, meta data, coverage of the data, aggregate basis vs. single basis, type of data, volume, who is responsible for the data, who has access to the data, can the data be downloaded/is it accessible in due time, is there a description of the recorded data fields, from how many systems/sources is the data, cost for accessing the data, who are the users of the data, are there reports using the data, are there identifiers that facilitate merging data, who has the rights to change data afterwards, is there a delay in recording and/or transferring data?

**5. Determination of the control group**

- To compare the test results with the current or standard status, a control group is chosen out of the same population.
- e.g. as in clinical studies or in marketing

## 4.3 Harmonizing, rescaling and cleansing of data

**Reasons for data cleaning**

- **Secondary data (pre-existing databases):** data is collected for other purposes, i.e. quality of data is driven by what was originally important, and thus there is so far no need to satisfy the quality requirements for the new analysis
- **Primary data (data collected via surveys):** individuals may get fatigued when filling out an extensive survey and thus may simply put default values (usually neutral or extremely positive or negative); false answers if individuals got offended by questions; most people simply refuse to fill out a survey

**Process of data cleaning**

1. **Raw Data**
2. **Technically correct data**
   - each value can be recognized belonging to a certain variable
   - each value is stored in a data type that represents the value domain of the real-world
3. **Consistent data**
   - missing values, special values, (obvious) errors and outliers are either removed, corrected or imputed
   - data is consistent with constraints based on real-world knowledge about the subject
   - i.e. data is fit for the analysis

**Common issues in data cleaning**

- Assess **technical correctness**
  - range of valid responses
  - invalid data responses (e.g. letters in number fields)
- Assess **consistency**
  - inconsistent data encodings (e.g. different abbreviations for countries)
  - suspicious/questionable data responses (do the outliers make sense?)
  - suspicious distribution of values (use descriptive statistics to identify, e.g. histogramms, box plot, etc.)
  - suspicious relationships between fields (identify correlations, e.g. factor analysis, etc.)
- Handling **null or missing values**
  - Deletion: always drop
  - Deletion when necessary: only drop observation unless required for analysis
  - Imputing a value: use of regression to predict an answer (but this might understate the uncertainty)
  - Randomly imputing a value: rerun analysis for all possible outcomes and weight by regression-based probability

**Issues when combining data from different sources**

- The **level of granularity** may be different.
  *Solution:* either skip inconsistent variables unless needed or apply methods for missing values
- The **data architecture** may be different (e.g. NA, na, space, etc.)
  *Solution:* harmonize before aggregating the data
- Data may be **stored in different ways** (e.g. in several databases with different structure).
  *Solution:* define a record which has enough fields to contain the information of each of the databases
- There may be **different weights** of the observations (e.g. based on 10,000 vs. 100 responses).
  *Solution:* introduce a weighting field

**Helpful tips and tricks for data cleaning**

- create a **date stamp**
- create a field identifying the **source**
- create new fields with **computed values** (i.e. if information is required that is not yet in the database)
- delete fields that have the same value accross all datasets
- store input data for each stage (i.e. raw, technically correct, consistent, aggregated, formatted)

**Final checks on data quality**

- **the 10 C's**
  - **Completeness**
  - **Correctness**
  - **Currency:** is the data obsolete?
  - **Collaborative:** is the data based on one opinion or on a consensus of experts?
  - **Confidential:** is the data secure from unauthorized use?
  - **Clarity:** is the data legible and comprehensible?
  - **Common format:** is the data in a format easily to be used?
  - **Convenient:** convenient and quick access to data?
  - **Cost-effective**
- **ACCURATE**
  - **Accurate:** fair and free from bias? arithmetical or grammatical errors?
  - **Complete**
  - **Cost-beneficial**
  - **User-targeted:** style, format, detail and complexity address the the nedds of the users?
  - **Relevant:** communicated to the right person?
  - **Authoritative:** reliable source?
  - **Timely:** does the receiver of the information have enough time to decide appropriate actions?
  - **Easy to use:** understandable to users?

## 4.4 Discovery of relationship in data

**Overview**

- **Describe:** How do I develop an understanding of the content of my data?
  - **Processing:** How do I clean and separate my data?
    - Filtering
    - Imputation
    - Dimensionality reduction
    - Normalization and transformation
    - Feature Extraction
  - Aggregation
  - Enrichment
- **Discover:** What are the key relationships in the data?
  - Clustering
  - Regression

**Discovery of relationship in data**

- **Filtering:** identify data based on its absolute or relative values
  - use *relational algebra projection and selection* IOT add or remove data based on its value
  - use *outlier removal, exponential smooting* and either *Gaussian or median filters*
- **Imputation:** fill in missing values in data
  - generate values for missing data using *random sampling* or *Monte Carlo Markov Chain methods*
  - use *mean, regression models* or *statistical distributions* based on existing observations
- **Dimensionality reduction:**
  - determine correlation accross different dimensions using *principle component analysis* or *factor analysis*
  - in unstructured text data: use *term frequency/inverse document frequency* IOT identify the importance of a word
    (use the inverse frequency since the important terms usually have a low frequency)
  - create a fixed number of features which form the indices of an array using *feature hashing*
  - use *sensitivity analysis* and *wrapper methods* when important features are not known
  - use *self-organizing maps* and *Bayes nets* IOT understand the probability distribution of data
- **Normalization and transformation:** reconcile duplication representations
  - correct duplicate data elements with *de-duplication*
  - *normalization* ensures your data stays within a common range
  - *format conversion* is typically used when data is in binary format
  - for frequency data: use *fast Fourier transforms (FFT)* and *discrete wavelet transforms*
  - for geometric data defined over Euclidean: use *coordinate transformations*
- **Feature Extraction:** determine the set of features with the highest information value to the model
  - features can be filtered out and tested with statistics, which is independent of the finally applied model
  - with *wrapper methods* (e.g. regression, k-nearest neighbour), one evaluates feature sets by constructing models and measuring performance
- **Aggregation:** collect and summarize the data
  - use *basic statistics* (raw counts, means, medians, standard deviations, ranges) IOT summarize data
  - use *box plots* and *scatter plots* IOT provide compact representation
  - use *"baseball card" aggregation* IOT summarize all the information available on an entity
- **Enrichment:** add new information to the data
  - use *annotation* IOT track source information and other user-defined parameters (e.g. which ZIP code belongs to which canton)
  - use *relational algebra rename and feature addition* (e.g. geography, weather) IOT process certain data fields together or use one field to compute the value of another
- **Clustering:** segment the data to find natural groupings
  - *connectivity-based* methods: hierarchical clustering
  - *centroid-based* methods: use *k-means* when the number of clusters is known, use *x-means* or *Canopy clustering* when the number is unknown
  - *distribution-based* methods: *Gaussian mixture models*
  - *density-based* methods: *fractal and DB scan* are useful for non-elliptical clusters
  - *graph-based* methods: useful when you only have knowledge of how one item is connected to another
  - *topic modeling:* for segmentation of text data
- **Regression:** determine which variables are important
  - *tree-based methods:* when structure of data is unknown
  - *generalized linear models:* when statistical measure of importance is needed
  - *regression with shrinkage* (e.g. LASSO, elastic net) and *stepwise regression* when statistical measure of importance is not needed

## 4.5 Documentation and reporting of findings

- Raw data and relationships will not attract attention of your stakeholders.
- Tie your findings to the analytics problem and from there to the business problem.
- State the impact of your findings to the business and complement this with recommendations.
- Document exceptional or more pronounced relationships.
- Document as well the data warehouses IOT make them usable for external parties (there might be requests to revisit the data months or years after the analysis is done).

## 4.6 Re-definition of the business and analytics problem statement by use of the data analytics result

**Solid data and relationships** will allow you to find that:

- the *true constraints of the system* are not as originally assumed and thus the analytics problem needs to be **reframed**. Also, maybe only a part of the original question can be answered.
- the business problem itself missed a key facet (e.g. unexpected relationship, time-series effect, etc.) that needs to be **included**.

# 5 Identification of problem solving approaches and appropriate tools

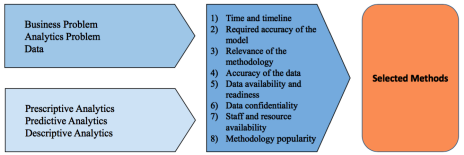



Figure 6: Process of selecting methods.

## 5.1 Identify available problem solving approaches/methods

1. **Time and timeline**
   - typically tight deadlines
   - thus often no time for experiments or testing new methods
   - consequently, fast and familiar methods are preferred
2. **Required accuracy of the model**
   - accuracy of the results depends on the aggregation/granularity of the model and the quality and readiness of the data
   - thus, if quality and accuracy of data is poor, sophisticated models do not provide any additional value
3. **Relevance of the methodology**
   - the business problem statement and the analytics problem framing often point the direction to a methodology
   - thus it is important to understand the goal of the analytics project (e.g. is it descriptive, predictive or prescriptive)
   - also, one has to adapt the methodology (or maybe even the goal) to the data analytics awareness of the stakeholder
4. **Accuracy of the data**
   - accuracy of the data restricts the pool of possible methods
   - e.g. granular time series models vs. simple regression analysis
5. **Data availability and readiness**
   - data might be available but might not be readily accessible, might only be available in very poor quality, or the quality and relevance of the data might not be known at all
   - e.g. data about standard products vs. special products
   - e.g. data stored on old IT infrastructure
6. **Data confidentiality**
   - classification of data might be internally regulated or by data protection laws
   - thus, data might be available only highly anonymized and/or aggregated
   - e.g. patient data from hospitals
7. **Staff and resource availability**
   - there are differend experts for different areas, e.g. experts on statistics, operations research, pricing, ect.
   - lack of experts can often delay a project for several weeks
   - there might also be restrictions due to limited available licenses of analytics software (especially if one has to work on the premises of the customer)
8. **Methodology popularity**
   - sometimes the customer asks for a certain method to be applied
   - however, it is in the responsibility of the data analytics professional to apply the most accurate/adequate methods but not only the most popular

## 5.2 Descriptive methods

**Examples**

- mean, median, mode, variance, standard deviation
- scatterplot, histogram, boxplot, steam-leaf plot

**When to use descriptive methods?**

- for reducing, summarizing and grouping data
- there is no dependent variable, i.e. the aim is to only describe what the data shows
- to simplify large amounts of data
- if one is unfamiliar with the data set and thus a feeling for the data first needs to be build up

## 5.3 Predictive methods

- **Regression:** techniques for estimating relationships among variables and building up an understanding of which variables are important in predicting future values
  - consists of a dependent variable (to be estimated) and of independent variables (predictors)
  - **Linear regression**
  - **Step-wise regression:** inclusion or deletion of the independent variable step by step based on some statistical measure (e.g. t-test or F-test); *forward selection* (starting with no independent variable and adding then the variable which improves the model most by its inclusion) vs. *backward selection* (start with all variables included and deleting then the variable which improves the model most by its deletion)
  - **Shrinkage regression** (e.g. Ridge regression or Lasso): if there are more variables than observations, least-square estimators do not exist; a *penalty* (additional constrations on the coefficients) can overcome the issue of $(X^T X)^{-1}$ being singular
  - **Logistic regression:** the regression case where the dependent variable is categorical, especially *binary*; important in credit scoring models and generally in the financial risk modelling area
    e.g. also to determine which customers will respond to a new marketing campaign
  - Test of performance: $R^2$, $R^2_{\text{adj}}$ ($R^2_{\text{CS}}$ or $R^2_{\text{McF}}$ or likelihood ratio test for logistic regression), AIC, BIC, accuracy/error rate, ROC
- **Clustering:** techniques for the segmentation of data into naturally similar groups
  - **Hierarchical clustering:** for an ordered set of clusters with observation precision
  - **k-means clustering:** to partition $n$ observations into $k$ clusters, where $k$ is known (e.g. centroid base method, where one aims to minimize the distance (as sum of the squares) of each point in the cluster to the cluster centre)
  - **x-means clustering:** if the number of clusters $x$ is unknown
  - Test of performance: ROC, sum of squared error (SSE), measures the compactness of a cluster or comparison with a given cluster structure (rand index)
- **Classification:** techniques for the prediction of the group membership of the observations
  - **Decision trees:** the algorithms start at the top and at each node a variable is chosen/determined that splits best the sample of observations; typically used when a transparent model is needed; e.g. CART, MARS, random forest, bagging
  - **k-nearest neighbours:** non-parametric alogrithm that classifies a point under consideration of the $k$-nearest neighbours to this point; typically used when the data dimension is not too high
  - **Neural networks:** the algorithms learn features in the data by changing the weights between the nodes based on learning rules
  - if one does not know where to start: use **Support Vector Machine (SVM)** or **Naïve Bayes**
  - Test of performance: accuracy/error rate, ROC
- **Statistical inference:** drawing conclusions based on data
  - e.g. confidence intervals, hypotheses testing, analysis of variance (ANOVA)
  - **Design of experiments:** planning, conducting, analyzing and interpreting control test; it aims to quantify the effects of values of the output parameters by controlled variation of the input factors
- **Simulation:** the design of a model of a real-world system or process, executing the model and analyzing the output of the model; the intention is learning by doing
  - Test of performance: analysis of the difference between the model and the status quo.

## 5.4 Prescriptive methods

- **Mathematical optimisation**
  - generall mathematical optimization problem: minimize $f(z)$ (objective function), subject to $f_i(x) \leq b_i$ (constraint functions)
  - **Linear programming:** optimum in a linear mathematical model subject to linear constraints
  - **Integer programming:** special case of linear programming where variables take integer values only
  - **Nonlinear programming:** either the objective function or the constraints are non-linear
  - **Metaheuristics:** uses intensification and diversification
  - also: calculus of variations, e.g. Euler-Lagrange equations, the generalization of the optimal control theory, dynamic programming (derives overall solution from solutions of the sub-problems)
- **Stochastic optimisation**
  - Generally: minimize the loss function $L = L(\theta)$ where $\theta$ is the $n$-dimensional vector of parameters that are being adjusted
  - but: either there is a random noise in the measurement, i.e. $y(\theta) = L(\theta) + \epsilon(\theta)$ and/or there is a random choice made in the search direction of the iteration algorithm
  - measurement with noise: **stochastic approximation** that is a recursive update rule (e.g. Robbins-Monro algorithm)
  - random search: **simulated annealing**
- **Simulation**
  - **Discrete event simulation** (e.g. patients running through the operating room process)
  - **Markov models** or **queuing models:** analyze queues where the queue lenghts and the waiting time can be determined (e.g. production processes)
  - **Agent-based modelling (ABM):** simulates the actions and interactions of autonomous agents which are assigned with rules (e.g. product launch in the market where the behaviour of the customers and the competitors are simulated)
  - **Monte Carlo simulation:** based on generated random samples which follow (parametric or non-parametric) statistical distributions and interdependencies (e.g. financial risk models or the development of required capital or equity)
  - **System dynamics:** understand the interactions over time in a complex dynamic system; often implying feedback and reinforcing loops (e.g. simulation of a location strategy of a country)
  - General examples: optimisation of a production line in a plant of a manufacturing company

## 5.5 Selection of software tools

**Dimensions of software** statistical capabilities, data mining, simulation, optimization, visualization/reporting, user-friendliness, costs, maintenance, transparency, ...

**Available software**

- **Microsoft Excel:** regression; all companies have it, VBA, add-ons like @Risk
- **R:** statistical computing and graphics, regression, clustering, classification; scientific standard, open source, many packages
- **Python:** high-level programming language, e.g. clustering; open source
  **Pandas:** high-performance, easy-to-use data structures and data analysis tools for Python, especially to manipulate numerical tables and time series, clustering; open source
- **MATLAB:** simulations; proprietary
- **KNIME** (Konstanz Information Miner): data mining, machine learning, reporting/visualization, clustering, classification; R and WEKA integration, Java based, open source
- **Rapidminer:** machine learning, data/text/web mining, predictive and business analytics; partly free
- **WEKA:** data preprocessing, regression, clustering, classification, feature/attribute selection, and visualization; open source
- **SAS:** statistical programming language for data analytics, data management, business intelligence, risk management, supply chain management, regression, classification
- **IBM SPSS Statistics:** addresses the entire analytical process from planning and data collection to analysis, reporting and development, regression, classificatioin; professional vendor suite, licensing system
- **SQL:** structured query language; language for managing structure data (i.e. data in a relational database management system); used in all database types like oracle, SAP, IBM DB, Microsoft etc.
- **Apache Hadoop:** open-source software framework in Java for distributed storage and distributed processing of very large data sets
- **.NET framework:** developed by Microsoft, e.g. C#
- **Java**
- **Julia:** high-level dynamic programming language used for scientific computing, machine learning, data mining, large-scale linear algebra, distributed and parallel computing; very efficient and effective language
- **System Dynamics/ABM/discrete events software:** PowerSim, Vensim, AnyLogic

# 6 How to set up and validate models

**Choosing a good model**

- Good models depend on all previous steps, i.e. framing the business and analytics problems, acquiring, exploring and cleansing the data, identification of problem solving approaches.
- The tools should actually be chosen based on the requirements of the approaches and models.
  In practice, though, tools can also be given which then define which models have to be applied.
- Typically, there is a class of models that seems most appropriate to model the data.
  In practice, one chooses several types of models from such a class, fits them to the data, and selects the best.

**Collaboration of involved people**

- **Business expert:** should have a clear view of the required characteristics to be modelled
  (e.g. a practioneer, such as a doctor responsible for the operating room management)
- **Data owner:** should known how to bring the data with the needed characteristics together and create the required data structure for the data analyst
  *Remark:* often a lot of data cleansing occurs during the model development since each modelling type has its own data obstacles.

**Qualities of a model** After several models are fit, one has to assess their performance and consider how the models will be used later.

- **Precision:**
  - Regression: $R^2$ has to be as close as possible to 1.
  - Classification: The error rate, i.e. the proportion of incorrectly classified items must be as close as possible to 0.
- **Robustness:**
  - Low sensitivity to random fluctuations, missing data or data that changes over time
  - Little dependencies on the training data
  - e.g. regressions or when developing simulations
- **Concision/parsimony:**
  - Rules (conditions, constraints) of a model should be as simple as possible and there should be as few such rules as possible
- **Explicit results:**
  - Rules of a model should be understandable, accessible and explicit (i.e. easy to implement)
  - Very often of importance (maybe except for simulations)
- **Diversity of types of data processed:**
  - A model should be applicable to the available type of data (e.g. discrete, categorical, time dependent, missing)
  - e.g. when analysing messages on social media platforms
- **Speed of the model development:**
  - Fast application of the model (i.e. (near) real-time)
  - Model's training, testing and adaptions should be processed within a reasonable time.
  - e.g. if model should be used as trading strategy

- **Possibility of parameter setting:**
  - Possibility to influence the parameter setting
  - e.g. in classification, clustering or when developing simulations

**Parameter learning/parametric modelling**

- **Goodness-of-fit:** shows how well a model is fitting the data
  - $R^2$ or $R^2_{\text{adj}}$ for regression
  - Akaike information criterion (AIC) or Bayes information criterion (BIC) for time series models
  - likelihood ratio test (expansions of a model compared to the simpler base model)
  - $R^2_{\text{CS}}$ or $R^2_{\text{McF}}$ or likelihood ratio test for logistic regression
  - linear discriminant analysis (Wilk's lambda)
  - accuracy/error rate, ROC
  - p-value in inference
- **Accuracy** and **error rate:**
  Accuracy = percentage of correct classifications
  Error rate = percentatage of incorrect classifications
- **Generalization vs. overfitting:**
  - **Generalization:** property of a model where the model has a generalized application to data that was not used to build the model
    I.e. when the model is not adequately complex, it becomes very inaccurate.
  - **Overfitting:** the tendency of data mining procedures to tailor models to the training data, at the expense of generalization to data that are not used to build the model
    i.e. when the model is too complex and incorporates idiosyncracies of the training set
    This causes performance to decline.

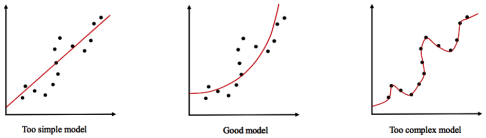



Figure 7: Generalization vs. overfitting.

- **Holdout set generalization performance:** the data set (the population) is split into (at least) two subsets.
  - **Training set:** set of data on which the model is fitted and the model parameters are determined
  - **Holdout** or **test set:** used to test the performance of the previously developed model
  - Rule of thumb: app. 70% of the data are allocated to the training set, while app. 30% of the data are allocated to the holdout set.
  - Size of the training set:
    A small training set can easily result in a low error rate in the training phase but will probably results in a high error rate in the testing phase.
    Contrary, if the training set is too large a model can be less efficient in the training phase but will perform better in the testing phase.
  - Disadvantage: it gives only a single estimate (i.e. a point estimate) of the generalization performance.
- **Cross-validation:**
  - Split the data set into $k$ partitions/folds (usually $k \in \{5, 10\}$)
  - Cross-validation iterates the training set and holdout set $k$ times, i.e. in each iteration of the cross-validation a different fold is chosen as the holdout set.
  - In each iteration $\frac{k-1}{k}$ of the data is used for training and $\frac{1}{k}$ is used for testing.
  - Based on the performance estimates of all the $k$ folds one can compute the average and standard deviation of the generalization performance.
    Rule of thumb: the training set should be at least 1,000 items, but for a sufficiently robust model one needs at least 350–500 items.

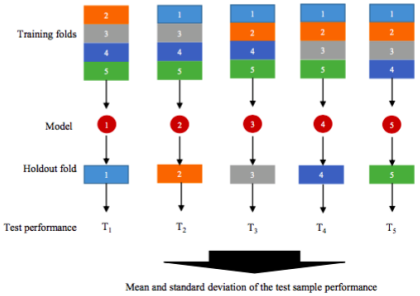



Figure 8: The method of k-fold cross-validation.

**Receiver Operating Characteristic (ROC)/Gini Index**

- Measures of the **Receiver Operating Characteristic (ROC):**
  - Accuracy $= (\alpha + \delta)/(\alpha + \beta + \gamma + \delta)$
  - True positive rate $= \delta/(\gamma + \delta)$
    (sensitivity)
  - True negative rate $= \alpha/(\alpha + \beta)$
    (specificity)
  - False positive rate $= \beta/(\alpha + \beta)$
    (1-specificity)
  - False negative rate $= \gamma/(\gamma + \delta)$
    (1-sensitivity)

| | | predicted | |
|---|---|---|---|
| | | negative | positive |
| actual | negative | true negative α | false positive (type I error) β |
| | positive | false negative (type II error) γ | true positive δ |

Figure 9: Error types of the ROC.

- **Gini index:**

$$\text{Gini-index} = \frac{A}{A+B}$$

The Gini index is in the range of $[0, 1]$. The closer the Gini index to 1, the closer is the stastical model to the perfect model.

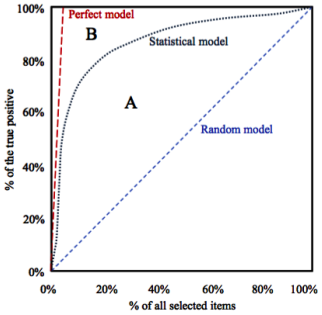



Figure 10: ROC curves of differen statistical models IOT compute the Gini index.

# 7 The deployment of a model

**Model deployment** Deployment is the implementation of the data analytics model into an IT or computer system for its use on a regular basis.

**Implementation in IT systems** IOT elaborate the implementation plan:

- What is the concept of operations (CONOP) of the current system/process?
  (feedback and acceptance of end users? a solution should interact with the current system that is maximally helpful but minimally disruptive)
- What is the legacy/computing/processing/operational environment?
  (familiarize yourself with the system's configuration management system and any leftover software, hardware and methods)
- What are the available interface mechanisms/processes?
  (existing application program interface as part of the system specification; ask for documentation; coordinate with system developer IOT avoid conflict)
- At what point in the processing stream can data be injected?
  (when is it possible to read/write/update data in the system architecture; synchronize with actions)
- What are the political/organizational considerations for interaction with the systems?
  (clearly understand organizational policies regarding the system)

*Further remarks:*

- The four most common ways of deploying models in data mining are:
  - Data mining tool or cloud (app. 45%)
  - Programming language, e.g. Java, C, VB (app. 15%)
  - Database and SQL script (app. 25%)
  - PMML: Predictive model and mark up language (app. 15%)
- Before requesting data and developing a model, make sure that the data will continue to be captured in the future.
- Failing in the implementation phase means failing in the project.

**Availability for users** Make the model available to the users and train the users.

The model will only be available for users after:

- model was tested in production for several months with volunteers, in parallel to the previous tool.
- champion-challenger strategies are implemented and evidently yield better results.
- a proper take-over phase and training.

*Further remarks:*

- The specification of the model should not be revealed to the end users (otherwise the model might not anymore serve its purpose).
- Users need to be well trained and understand the background of the models and tools.
- Training involves: objectives of the model, principles of the tool, limitations, methods for using the tool, contributions of the tool, operational and organizational consequences

**Monitoring and maintenance**

- **Monitoring and maintenance plan:**
  - which results may require updating and why
  - how will updating be triggered (regular updates, trigger event, performance monitoring)
  - how will updating be performed
  - summary of the results of the updating process
- **Types of monitoring:**
  - *One-off monitoring:* whenever a new data mining application is put in use
  - *Ongoing monitoring:*
    *Population stability:* understand how the target population changes over time
    *Scorecard performance:* understand the benefits of having the right information available when making decisions
    *Decision management:* understand how the model degradation may affect the quality of the portfolio

**Model reports** Producing reports out of the model:

- A the end of the project and depending on the deployment plan, this might be a written summary of the project or a final presentation of the data mining results.
- The actual content depends on the target audience.
- *General contents:*
  obtained results, process, incurred cost, deviations from the original plan, implementation plans, recommendations for future work
- *Content of data analytics reports:*
  data received, data treatment, data mining result, performance assessment, implementation, deployment plan and consideration, monitoring and maintenance strategy, reference received, clarifications from the client, reference to intermediary reports and presentations
- Identify which reports are needed, outline the structure and select findings to be included.

- Often, also a final presentation for the project/management sponsor has to be held.

**Project reports** Producing a final project report that summarizes the entire project and its results:

- business understanding: background, objectives, success criteria
- data mining process, results and evaluation
- deployment and maintenance plan
- cost/benefit analysis
- conclusions for the business and for future data mining projects

*Further remarks:*

- Perform also a project review and create an experience documentation IOT determine potential improvements.

# 8 Model lifecycle

**Importance of model lifecycle** Such a lifecycle is important because:

- data and the environment are changing over time.
- more data or better quality of data becomes available.
- there is technological and methodological advancement.
- the company and the users of models progress in data analytics.
- the business benefit might change over time (e.g. due to advancing competitors, new products).

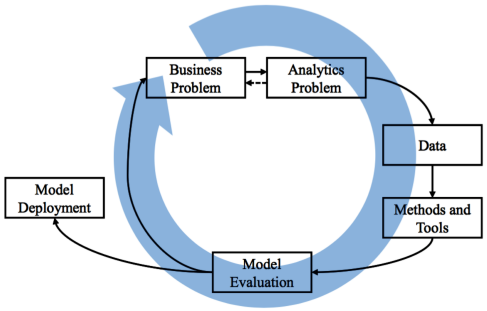



Figure 11: Model lifecycle.

**Lifecycle governance** A good lifecycle process:

- helps minimizing the cost and efforts for developing and maintaining the model.
- *allocates the users in a company clear roles.*
- defines the roles of different departments involved.
- defines the *escalation processes* and *decision making processes.*

**Best practices**

- Do the documentation during the project (since things get forgotten quickly, there is no time for documenting after the project, people may leave the project early).
- Define the measure for model quality in advance in the context of the model.
- Have the old data available in a form s.t. one can readily recompute the different measures.
- Define the frequency with which the model should be evaluated.
- Do not only train users but also re-assess how they are using the model and the results and if necessary train the users again.
- Evaluate the business benefit on a regular basis as it can diminish.

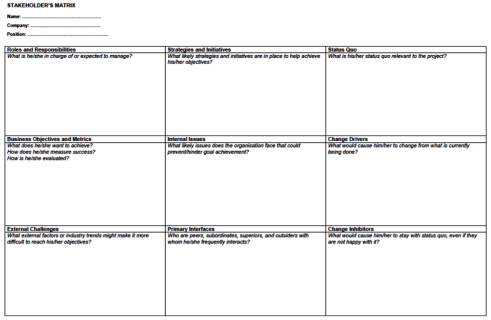
# 9 Soft skills

A data analytics professional needs the ability to convince, or explain the problem, problem solution and implications.
Thus a data analytics professional needs the following:

- **Communication skills:** the ability to communicate with a client/employer/stakeholder regarding the framing of a business and an analytics problem.
  - find the deep underlying motives of any client engagement
  - question until it is clear what the problem is and how a solution can be attempted
  - do not come up with proposals or solutions too early
- **Background and position:** the understanding of the background of the client/employer/stakeholder regarding his/her organization and specific industry focus.
  - an organizational chart can help but might be insufficient
  - observe the inter- and intra-office communication since it often reveils the "pecking order" or internal group dynamics
  - take note of the people in the project management meetings and about their presence and behaviours since these are indicators of their status within the organization
    $\Rightarrow$ stakeholder matrix
- **Clarifying the analytics process:** the ability to explain the findings of a the analytics process in sufficient details to ensure clear understanding by the client/employer/stakeholder.
  - the data analytics professional is at the heart of the whole process and thus needs to have a complete understanding
  - ensure that your questions and comments are seen as necessary to the process, not as intrusive and time wasting
  - thus, be transparent and explain in simple words why you need certain information
  - translate between technical jargon and acronyms

**Stakeholder matrix**

- *Roles and responsibilities:*
  What is he/she in charge of or expected to manage?
- *Business objectives and metrics:*
  What does he/she want to achieve? How does he/she measure success? How is he/she evaluated?
- *External challenges:*
  What external factors or industry needs might make it more difficult to reach his/her objectives?
- *Strategies and initiatives:*
  What likely strategies and initiatives are in place to help achieve his/her objectives?
- *Internal issues:*
  What likely issues does the organisation face that could prevent/hinder goal achievement?
- *Primary interfaces:*
  Who are peers, subordinates, superiors, and outsiders with whom he/she frequently interacts?
- *Status quo:*
  What is his/her status quo relevant to the project?
- *Change drivers:*
  What would cause him/her to change from what is currently being done?
- *Change inhibitors:*
  What would cause him/her to stay with the status quo, even if they are not happy with it?


STAKEHOLDER'S MATRIX

Name: .....
Company: .....
Position: .....

| Roles and Responsibilities | Strategies and Initiatives | Status Quo |
|---|---|---|
| What is he/she in charge of or expected to manage? | What likely strategies and initiatives are in place to help achieve his/her objectives? | What is his/her status quo relevant to the project? |
| **Business Objectives and Metrics** | **Internal Issues** | **Change Drivers** |
| What does he/she want to achieve?<br>How does he/she measure success?<br>How is he/she evaluated? | What likely issues does the organisation face that could prevent/hinder goal achievement? | What would cause him/her to change from what is currently being done? |
| **External Challenges** | **Primary Interfaces** | **Change Inhibitors** |
| What external factors or industry trends might make it more difficult to reach his/her objectives? | Who are peers, subordinates, superiors, and outsiders with whom he/she frequently interacts? | What would cause him/her to stay with status quo, even if they are not happy with it? |

Figure 12: Structure of a stakeholder matrix.

# 10 Excurses

## 10.1 Sentimental analysis

**Introduction**

- **user-generated content:** reviews of products, blogs, forums, groups (all in the internet)
- Nowadays, instead of conducting surveys, one simply has to mine the corresponding weg pages and extract the relevant information IOT get to know about the public opinion of a product or company.
- *Difficulties to find this information:* huge volume of data, unstructured data, data might be hidden in a certain blog, text and opinions are not straight-forward (e.g. irony, reversing expressions)

**Definitions**

- **Sentiment or opinion analysis:** extracting subjective information out of data by the use of natural language processing or text mining.
- **Natural language processing:** analyzing, understanding, generating and interacting with the language that humans use for interactions with computers in both written and spoken contexts using natural human languages instead of computer languages; area of computer science and artificial intelligence.
- **Text mining:** analyzing text and gathering information out of it by using pattern analysis techniques.
- **Object ($o$):** an entity which can be a product, person, event, organization or topic.
  The object $o$ is associated with a pair: $o : (T, A)$, where $T$ is a hierarchy of components (or parts), sub-components, etc. and $A$ is a set of attributes (properties) of $o$.
  Each component has its own set of sub-components and attributes.
- **Opinionated document ($d$)**: a product review, a forum post or a blog that evaluates a set of objects.
  In the most general sense, $d$ consists of a sequence of sentences $d = \{s_1, s_2, \ldots, s_m\}$.
- An **opinion passage** on a feature ($f$) of an object $o$ evaluated in $d$ is a group of consecutive sentences in $d$ that expresses a positive or negative opinion on $f$.
- **Explicit feature:** $f$ is called an explicit feature in $s$ if $f$ or any of its synonyms appears in a sentence $s$.
- **Implicit feature:** if neither $f$ nor any of its synonyms appear in $s$ but $f$ is implied.
- **Opinion holder ($h$):** the person or organization that expresses the opinion.
- An **opinion** on a feature $f$ is a positive or negative view, attitude, emotion or appraisal on $f$ from an opinion holder.
- The **opinion orientation ($oo$)** on a feature $f$ indicates whether the opinion is positive, negative or neutral.
- **Model of an object:** an object $o$ is represented with a finite set of features, $F = \{f_1, f_2, \ldots, f_n\}$, which includes the object itself as a special feature.
  Each feature $f_i \in F$ can be expressed with any one of a finite set of words or phrases $W_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$, which are synonyms of the feature, or indicated by any one of a finite set of feature indicators $I_i = \{i_{i1}, i_{i2}, \ldots, i_{iq}\}$ of the feaute.

**Model of an opinionated document**

- A general opinionated document $d$ contains opinions on a set of objects $\{o_1, o_2, \ldots, o_q\}$ from a set of opinion holders $h_1, h_2, \ldots, h_p$.
  The opinions on each object $o_j$ are expressed on a subset $f_j$ of features of $o_j$.
  An opinion can either be direct or comparative.
- **Direct opinion:** a quintuple $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$
  For feature $f_{jk}$ that opinion holder $h_j$ comments on, he/she chooses a word or phrase from either the corresponding synonym set $W_{jk}$, or a word or phrase from the feature indicator set $I_{jk}$ to describe the feature.
- **Comparative opinion:** expresses a relation of similarities or differences between two or more objects, and/or object preferences of the opinion holder based on some of the shared features of the objects.
  Usually the **comparative or superlative form** of an adjective or adverb is used.

**Objective of mining direct opinions:** Given an opinionated document $d$,

1. Discover all opinion quintuples $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$ in $d$.
2. Identify all synonyms $(W_{jk})$ and feature indicators $(I_{jk})$ of each feature $f_{jk}$ in $d$.

**Erroneous sentimental analysis** E.g. denials and irony might be interpreted erroneously.

## 10.2 European Statistic Code of Practice

- **Vision:** becoming a world leader in statistical information services; programme of harmonized European statistics as basis for democratic processes and progress in society
- **Mission:** provide the EU, the world and the public with independent high quality information on the economy and society on all levels and make information available to everyone for decision-making purposes, reasearch and debate

**Code of Practice**

- **Institutional environment**
  - Professional independence
  - Mandate for data collection
  - Adequacy of resources
  - Commitment to quality
  - Statistical confidentiality
  - Impartiality and objectivity
- **Statistical processes**
  - Sound methodology
  - Appropriate statistical procedures
  - Non-excessive burden on respondents
  - Cost effectiveness
- **Statistical output**
  - Relevance
  - Accuracy and reliability
  - Timeliness and punctuality
  - Coherence and comparability
  - Accessibility and clarity

## 10.3 System dynamics

**Definition**

- System Dynamics (SD) is a methodology for framing, modelling and understanding the dynamcis of a complex system.
- It used for the understanding of non-linear behaviour of systems and to support system thinking.
- SD is suitable for systems that:
  - are dynamic and evolving over time.
  - are non-linear.
  - have accumulations and delays.
  - have feedback loops.

**Patterns of behaviour**

- Typical variables are:
  cost, sales, revenue, profit, market share, risk, etc.
- Typical patterns of behaviour over time are:
  exponential, goal-seeking, S-shape, oscillation
- Often one of the first three patterns is combined with oscillation whose amplitude gradually declines over time.

**Loops**

- **Feedback loop** or **causual loop:** a closed sequence of causes and effects, i.e. a closed path path of action and information (thus each element part of this subsystem influences itself)
  - **Self-reinforcing (R):** positive feedback
  - **Self-correcting/balancing (B):** negative feedback
    Remark: a balancing loop has always a goal/natural level
- **Open loop:** a linear chain of causes and effects which does not close back on itself

**Loop diagrams**

- **Elements:** variables, describe as nouns
- **Actions:** represented by arrows
- Ensure the definition makes it clear which direction is up or down, i.e. prefer the *positive sense in naming* a variable/action.
- Causal links should imply a direction of causation, but not simply a time sequence.
- Think about possible *unexpected side effects*.
- Negative feedback loops always have **a goal**.
- Include elements for both the **actual value** of an element and the **perceived value**.
- Distinguish between **short-term** and **long-term consequences** with different loops.
- If a link requires some explanation, then add intermediate elements inbetween.
- Keep the diagramm as simple as possible.

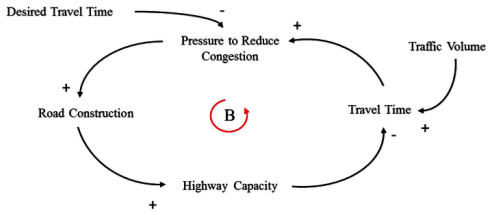



Figure 13: Loop diagram for the relationship between highway capacity and road congestion.

**Stock and flow diagrams**

- Consider **stocks**, **in**- and **outflows**, and **auxiliary variables**.
- Every stock and flow diagramm has *at least one source*, *one sink* (i.e. stocks outside the model) and *one stock*.
- Usually every stock has at least one in- and one outflow, each regulated by a valve.
- *Translation of elements from the loop diagram to the stock and flow diagram:*
  - **Stock:** an inventory which contains items where the amount of items can decrease or increase
  - **Valve:** a variable regulating the in- or outflow of a stock, might be itself regulated by a stock or an auxiliary variable
  - **Auxiliary variable:** elements that depict a certain dynamics but are neither stocks nor valves
- **Feedback loops:**
  - Valves that regulate stocks but which are in return also regulated by stocks may replicate feedback loops from the loop diagram.
  - Such feedback loops may also incorporate auxiliary variables (i.e. elements of causual loop diagrams that are neither stocks nor valves).
  - Usually, every stock direcly or indirectly (via an auxiliary variable) reinforces its outflow valve.
    *The logic behind:* since the valves depict rates but not ratios and assuming that the outflow ratios remain constant, it follows that if a stock increases also its outflow rate (i.e. decrease of elements per time unit) increases. Thus, each stock/outflow valve subsystem usually generates a balancing feedback loop (but which might be incorporated into a larger reinforcing loop).

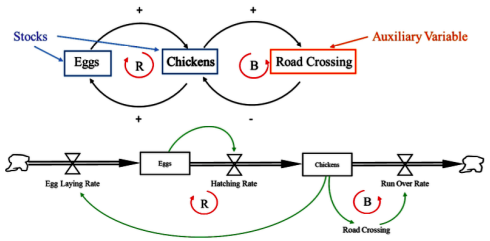



Figure 14: Loop diagram translated into a stock and flow diagram of a chicken and egg dummy system.

*intentionally left blank*

## Abbreviations

| | |
|---|---|
| **IOT** | in order to |
| **RV** | random variable |
| **s.t.** | such that |
| **w.r.t.** | with respect to |

## Disclaimer

- This summary is work in progress, i.e. neither completeness nor correctness of the content are guaranteed by the author.
- This summary may be extended or modified at the discretion of the readers.
- Source: Lecture Data Analytics in Organisations and Business, autumn semester 2015/16, ETHZ (lecture slides, script and exercises). Copyright of the content is with the lecturer.
- The layout of this summary is mainly based on the summaries of several courses of the BSc ETH ME from Jonas LIECHTI.