Prof. Dr. Bastian Amberg
School of Business & Economics
Department of Information Systems

Business Intelligence

05b CRISP-DM – Project Understanding

Prof. Dr. Bastian Amberg

(summer term 2024)

15.5.2024

Schedule

		Wed., 10:00-12:00		Fr., 14:00-16:00 (Start at 14:30)		Self-study		
	W1	17.4.	(Meta-)Introduction	19.4.		Python-Basics	Chap. 1	
	W2	24.4.	Data Warehouse – Overview & OLAP	26.4.	*[Blockveranstaltung SE Prof. Gersch]*		Chap. 2	
Basics	W3	1.5.		3.5.	Data Warehouse Modeling I		Chap. 3	
	W4	8.5.	Data Warehouse Modeling I & II	10.5.	Data Mining Introduction			
	W5	15.5.	CRISP-DM, Project understanding	17.5.	Python-Basics-Online Exercise	Python-Analytics	Chap. 1	
	W6	22.5.	Data Understanding, Data Visualization	24.5.	*No lectures, but bonus tasks 1.) Co-Create your exam 2.) Earn bonus points for the exam*		Chap. 2	
	W7	29.5.	Data Preparation	31.5.				
Main Part	W8	5.6.	Predictive Modeling I	7.6.	Predictive Modeling II (10:00 -12:00)	BI-Project	Start	
	W9	12.6.	Fitting a Model I	14.6.	Python-Analytics-Online Exercise		|	
	W10	19.6.	*Guest Lecture*	21.6.	Fitting a Model II		|	Case Study
	W11	26.6.	How to avoid overfitting	28.6.	What is a good Model?		|	
Deep-ening	W12	3.7.	Project status update Evidence and Probabilities	5.7.	Similarity (and Clusters) From Machine to Deep Learning I		|	
	W13	10.7.		12.7.	From Machine to Deep Learning II		|	
	W14	17.7.	Project presentation	19.7.	Project presentation		End	
Ref.					*Klausur 1.Termin ~ 22.7. bis 3.8. Klausur 2.Termin ~ 23.9. bis 5.10.*	Projektbericht		

CRISP-DM

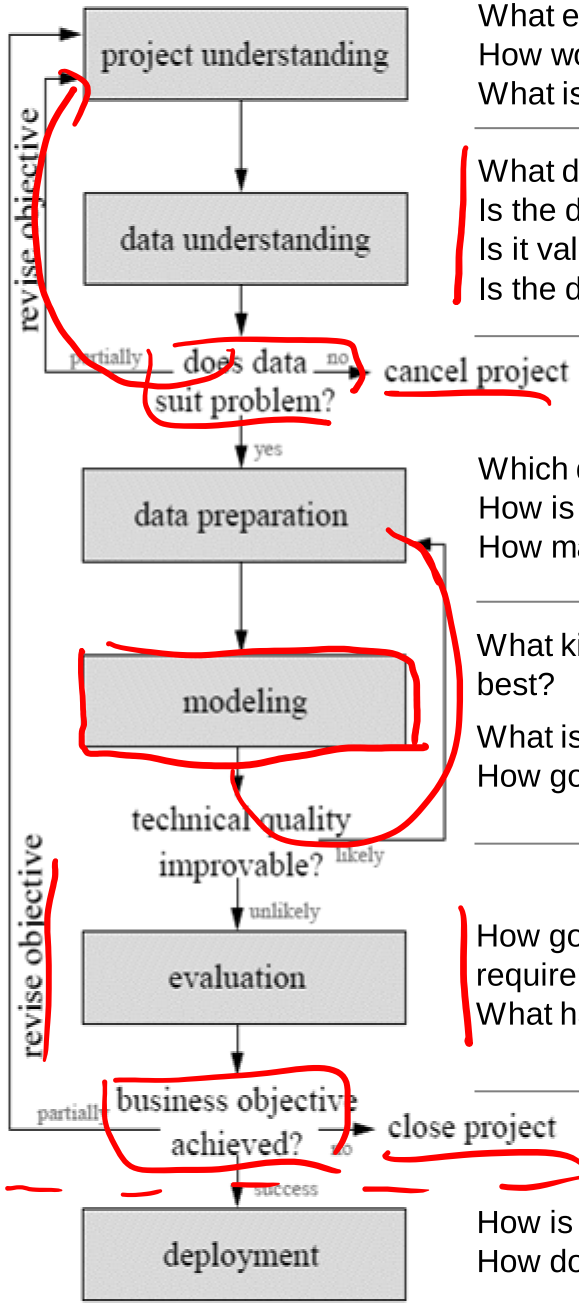
Daimler, SPSS

Cross
Industry
Standard
Process for
Data
Mining

Iteration as a rule

Process of data exploration

Implementation of the KDD Process



Ref. Wirth / Hipp (2000), Azevedo (2008)

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

Understand the problem to be solved, its context, and the subsequent requirements for a solution.

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?

Data are the available **raw materials** from which the solution will be built.
Match business problem to one or several data mining tasks

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

Data often need to be manipulated and converted into forms that yield better results

Match **data** and **requirements** of DM techniques

Select the relevant **variables**

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?

This is the primary place where **DM techniques** are applied to the data

Select a Model, generate a test design, build the model and assess it.

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

Assess the DM results **rigorously** (Gain confidence that results are valid and reliable)

Ensure that the model satisfies the original **business goals** (support decision making)

Ensure **comprehensibility** of the model to stakeholders
Evaluation framework needed (tested environments)

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

Models are put into **real use** in order to realize some return on investment.

Agenda

(1) Project Understanding

Assess the situation

Determine analysis goals

Ref.

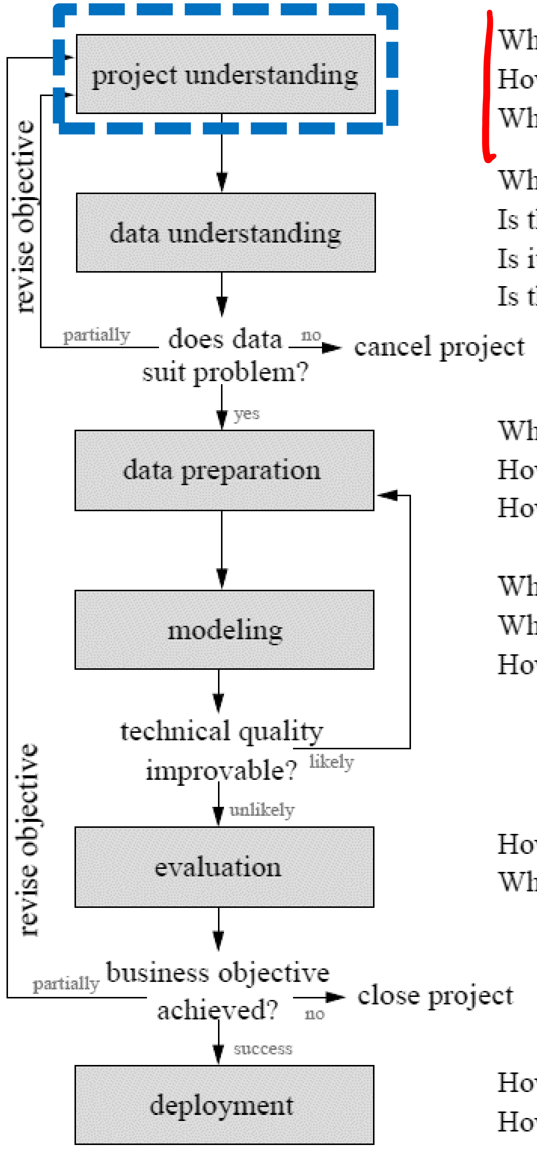
CRISP-DM

Cross
Industry
Standard
Process for
Data
Mining

Iteration as a rule

Process of data exploration

Implementation of the KDD Process



project understanding

data understanding

does data suit problem?

partially

no

cancel project

yes

data preparation

modeling

technical quality improvable?

likely

unlikely

evaluation

business objective achieved?

partially

no

close project

success

deployment

revise objective

revise objective

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?

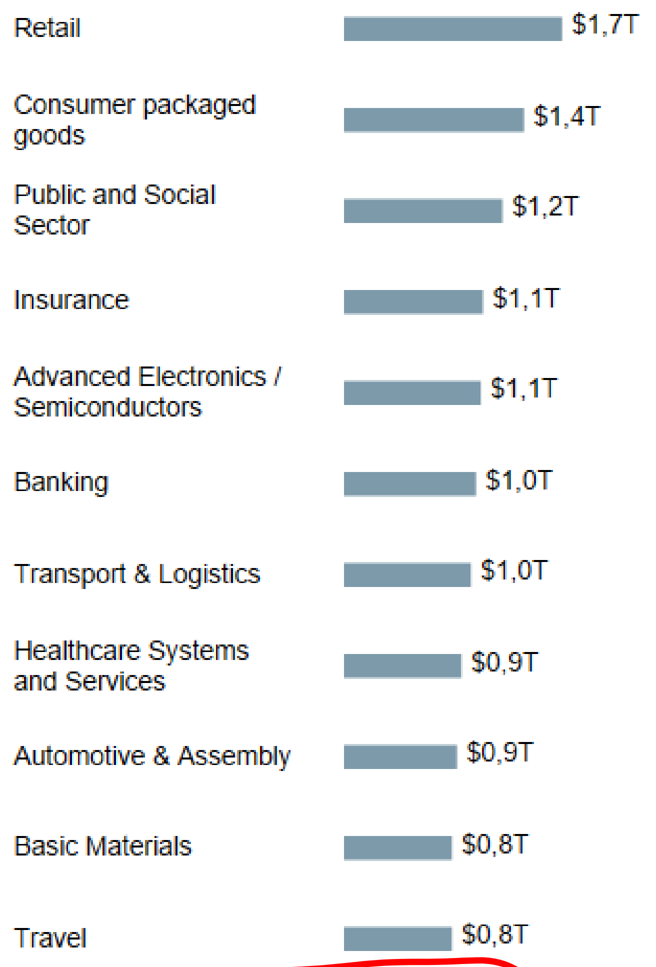
How good is the model in terms of project requirements?
What have we learned from the project?

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

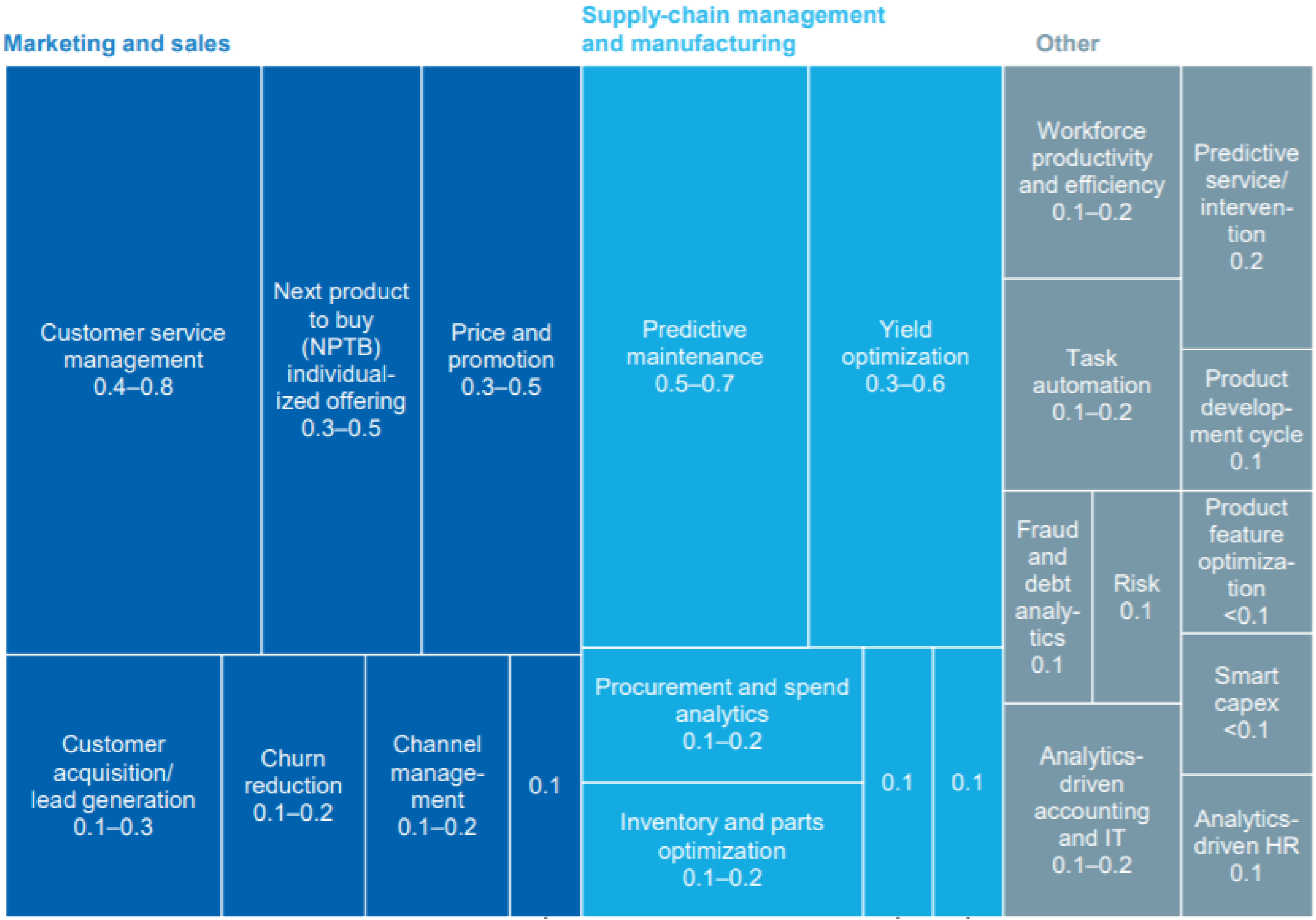
Ref. Wirth / Hipp (2000), Azevedo (2008)

Excursus: Business Problem Domains

AI and other analytics impact by industry, function and business problem.



Impact of AI per industry in trillion $



Impact of AI per business problem in trillion $

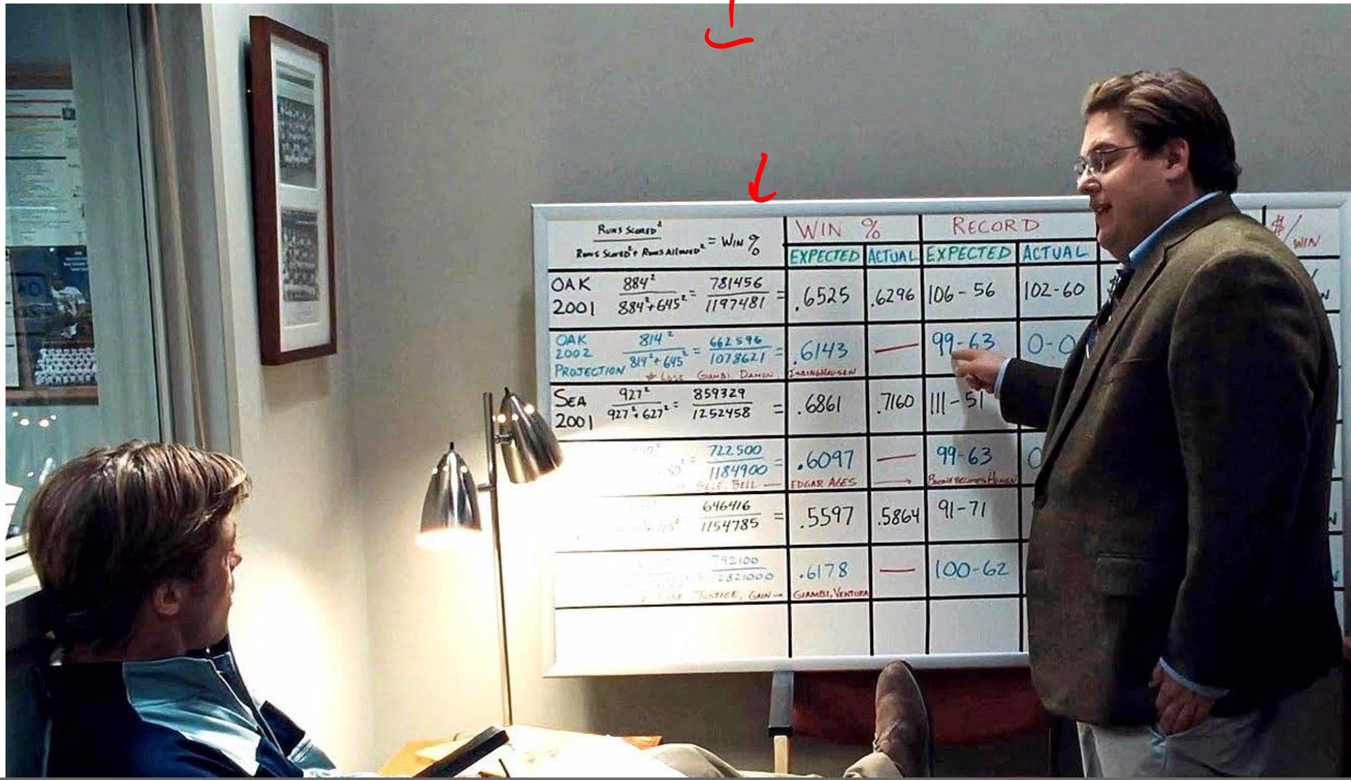
Ref. McKinsey (2018a), McKinsey (2018b)

Project understanding

Assess **the main objective**, the potential benefit, as well as the constraints, assumptions and risks.

- ○ Problem formulation
- ○ Map the problem formulation to a data analysis task → Classification → Clusting
- ○ Understand the situation
  (available data, suitability of the data, …)

- ➢ **Assess the situation**
- ➢ **Determine analysis goals**



Average time spent for project and data understanding within the CRISP-DM model: 20%
Importance for success: 80%

Ref.

Image: "Moneyball"/Columbia Pictures, Video

Agenda

(1) Project Understanding

Assess the situation

Determine analysis goals

Ref.

Assess the situation

Project‘s success

Estimate chances of a successful data analysis project

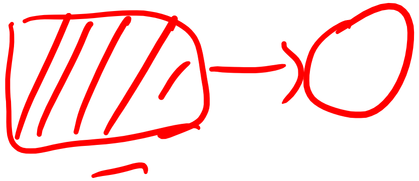
- Resources (data!), requirements and risks

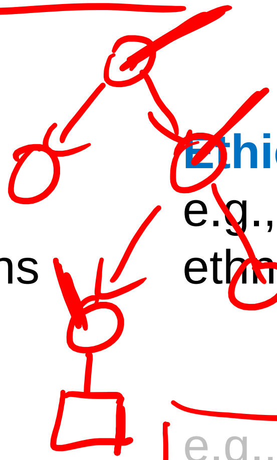
Does the given data satisfy the project‘s needs?

Typical requirements and constraints:

Model requirements
e.g., model has to be explanatory, because decisions must be justified clearly

vs. blackbox behavior





Ethical, political, legal issues
e.g., variables such as gender, ethnicity must not be used

e.g., no racial profiling
(Example from Antidiskriminierungsstelle des Bundes)

Technical constraints
e.g., applying the technical solution must not take more than *n* seconds

e.g., spam detection

Ref.

Assess the situation

Determine the project objective

The aim of the project should be clearly defined

Criteria to measure the success of the project should be agreed upon

1.

Aim/ Objective: increase revenues (per campaign and or/per customer) in direct mailing campaigns by personalized offer and individual customer selection

2.

Success criteria: improve order rate by 5% or total revenues by 5%, measured within 4 weeks after mailing was sent

Deliverable: software that automatically selects a specified number of customers from the database to whom the mailing shall be sent, runtime max. half a day

Ref.

Assess the situation

Assumptions

Representativeness:
Sample in the database must be representative for the whole population for which we intend to generalize.*

Good data quality:
The relevant data must be correct, complete, up-to-date and unambiguous thanks to the available documentation.

Informativeness:
To cover all aspects by the model, most of the influencing factors (e.g. identified in the *cognitive map*) should be represented by attributes in the database

Presence of external factors:
We may assume that the external world does not change constantly

560 Mio

4,4 Mio → ~3,6 Mio Deutschland

2018

*Excursus: Not representative "84% want to abolish the time changeover"
Link to European Commission, *Link to newspaper article*

Ref.

Assess the situation

Cognitive map for domain knowledge

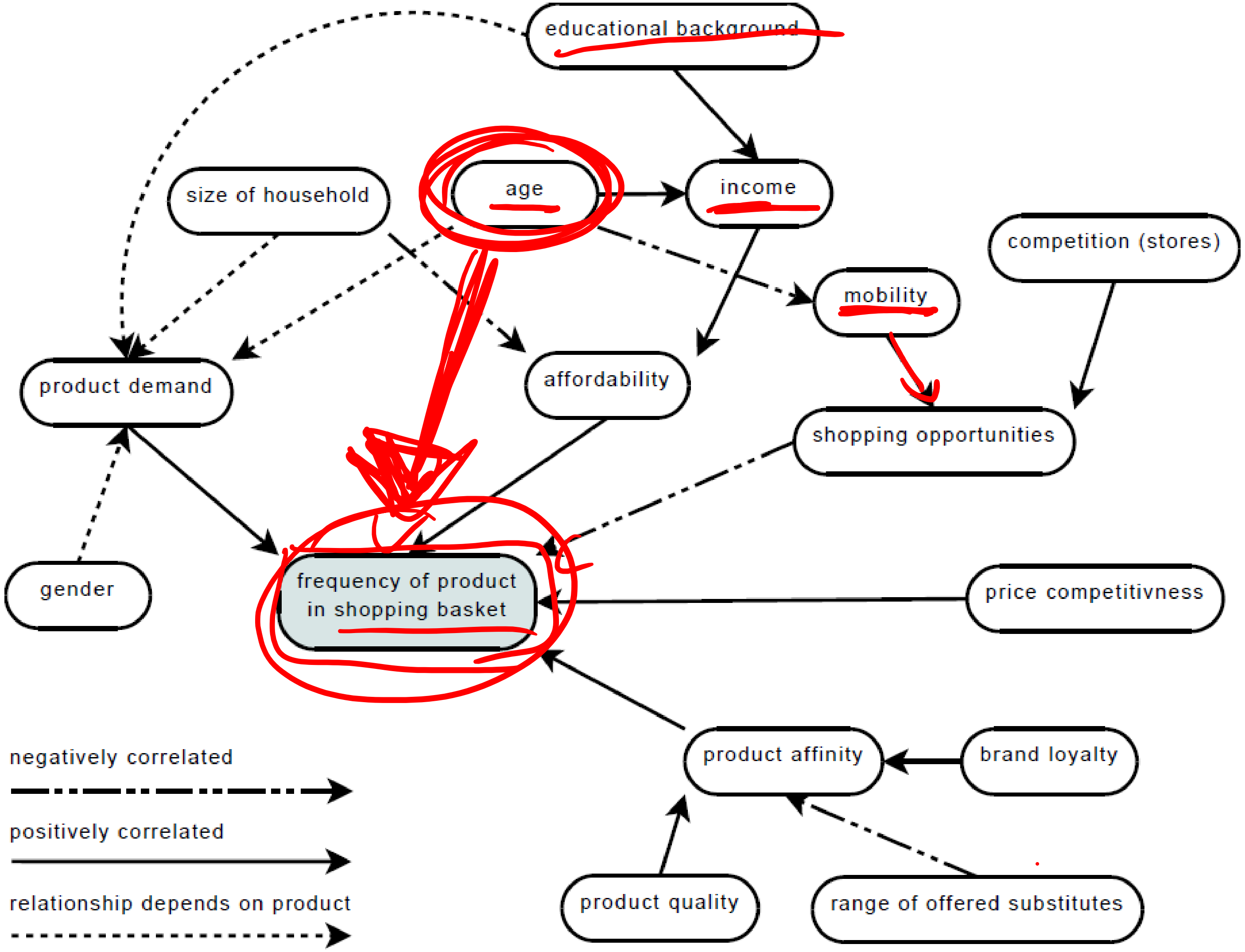
je höher ... desto ...

Perception of a reality

Directed graph of variables (causal concept) and relations (causal connections) in the decision problem domain and their strength (causal value).

The development of a cognitive map supports domain understanding and adjustment of expectations

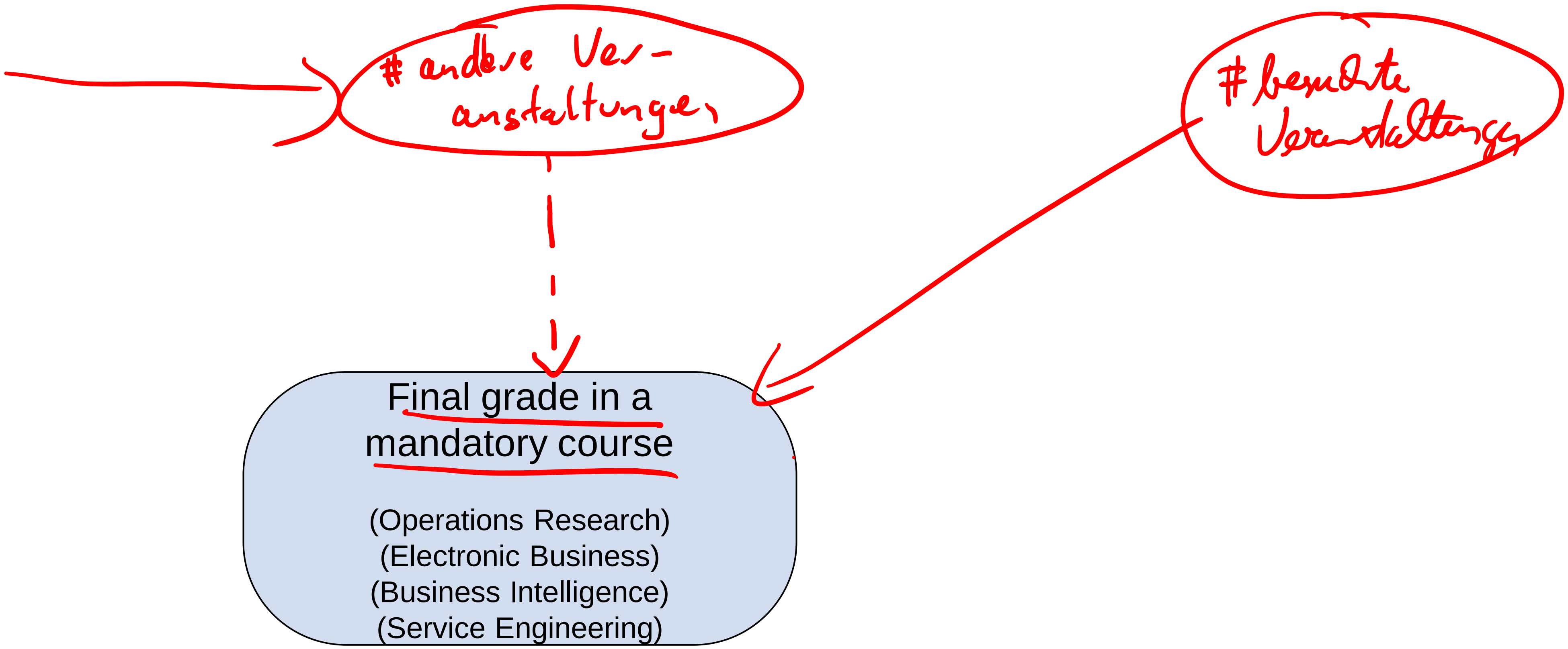
- ○ Include only direct dependencies to keep the map clear
- ○ Choose labels of nodes carefully so they are easily interpretable
- ○ Stick to the labels in project communication



Ref. Nadkarni/Prakash (2000)

Exercise: Cognitive map

Domain knowledge?



andere Veranstaltungen

besuchte Veranstaltungen

Final grade in a mandatory course

(Operations Research)
(Electronic Business)
(Business Intelligence)
(Service Engineering)

5 Min.

Ref.

Assess the situation

Risks: Domain experts and data analysis experts

Problem source	Project owner perspective	Analyst perspective
Communication	Does not understand the *technical terms* of an analyst	Does not understand the *terms of the domain*
Lack of understanding	Is not sure *what* the analyst *could do* or achieve	Finds it hard to understand *how to help* the project owner
Organization	*Requirements* have to be adopted *in later stages* as problems with data become evident	Project owner is an *unpredictable group* (not so concerned with the project)

Glossar

Interviews

-> Possible solutions?

Ref.

Images: V. Hanacek | Picjumbo

Agenda

(1) Project Understanding

Assess the situation

Determine analysis goals

Ref.

Determine analysis goals

Problem decomposition

Determine DM tasks and decompose problem

- Classification, regression, cluster analysis, …

Specify the requirements for the models that will be constructed by the DM tasks

There is no unique best method for a task

Decision Trees

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 **black box models** it is hard to comprehend how the final decision is made, and their model lacks interpretability. (i.e., deep learning)

Ref.

Images: Financial Times (cc-by 2.0)

Determine analysis goals

Stability and Flexibility

Reproducibility / stability

If the analysis is carried out more than once, we may achieve similar performance – but not necessarily similar models.

This does no harm if the model is used as a black box, but hinders a direct **comparison** of subsequent models to investigate their differences.

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.

Ref.

Image: Creative Tools (2015) | Flickr (cc-by 2.0)

Determine analysis goals

Runtime

If restrictive runtime requirements are given (either for building or applying the model), this may exclude some computationally expensive approaches.

Interestingness and use of expert knowledge

The more an expert already knows, the more challenging it is to **surprise** her with new findings. Some techniques are known for their large number of findings, many of them redundant and thus uninteresting.

So if there is a possibility of including any kind of previous knowledge, this may ease the search for the best model considerably and may **prevent us from re-discovering** too many well-known artefacts.

Ref.

Images: Thomas Leuthard (cc-by 2.0)

Fragen?

- ✓ The data mining process – CRISP-DM
- ✓ Business / Project understanding

  Starting in week W8, you will continue to deepen this content by working on your project

Recommended reading (for this week)

Berthold et al.	Guide to Intelligent Data Analysis Chapter 3, 4
Provost, F., Fawcett, T.	Data Science for Business Chapter 2
Pyle, D.	Business Modeling and Data Mining. Morgan Kaufmann, San Mateo (2003)

Ref.

Bibliography

- Nadkarni, Sucheta, and Prakash P. Shenoy. "A causal mapping approach to constructing Bayesian networks." *Decision support systems* 38.2 (2004): 259-281.
- Tukey JW. “Exploratory data analysis”. Reading, MA’ Addison-Wesley Publishing Company (1977): p. 1 – 688.

Ref.