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

Business Intelligence

05a Data Mining Introduction (continued)

Prof. Dr. Bastian Amberg
(summer term 2024)
15.5.2024

Schedule

14.30 - 16^{00}

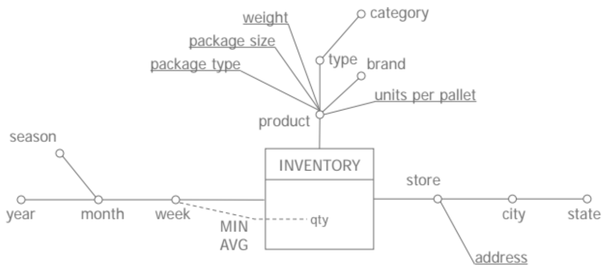
		Wed., 10:00-12:00		Fr., 14:00-16:00 (Start at 14:30)		Self-study		
Basics	W1	17.4.	(Meta-)Introduction	19.4.		Python-Basics	Chap. 1	
	W2	24.4.	Data Warehouse – Overview & OLAP	26.4.	*[Blockveranstaltung SE Prof. Gersch]*		Chap. 2	
	W3	1.5.		3.5.	Data Warehouse Modeling I		Chap. 3	
	W4	8.5.	Data Warehouse Modeling I & II	10.5.	Data Mining Introduction			
Main Part	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.				
	W8	5.6.	Predictive Modeling I	7.6.	Predictive Modeling II (10:00 -12:00)	BI-Project	Start	Case Study
	W9	12.6.	Fitting a Model I	14.6.	Python-Analytics-Online Exercise		|	
	W10	19.6.	*Guest Lecture*	21.6.	Fitting a Model II		|	
	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	
					Klausur 1.Termin ~ 22.7. bis 3.8. Klausur 2.Termin ~ 23.9. bis 5.10.	Projektbericht		

Ref.

Last lesson

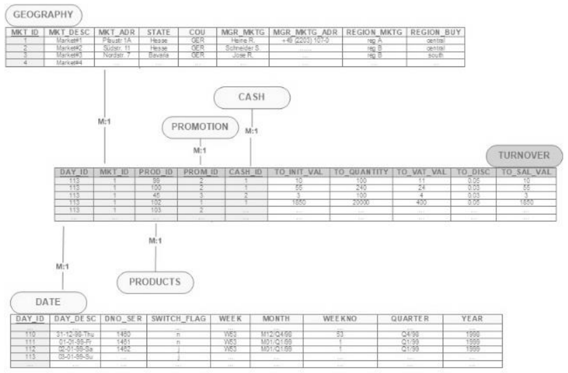
✓ How can **multidimensional data models** be **developed** and **stored**?

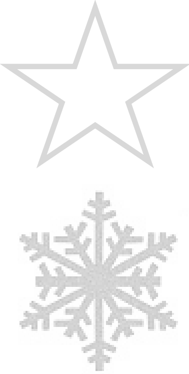
Conceptual modeling

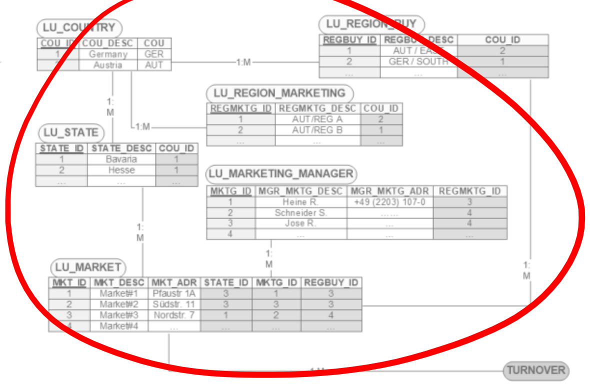


Logical modeling

Physical modeling







○ Introduction **Data Mining**

Where are the limits of the handling of data considered so far?

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."
(Hand, Mannila, Smyth (2001), Principles of Data Mining

Ref.

Agenda

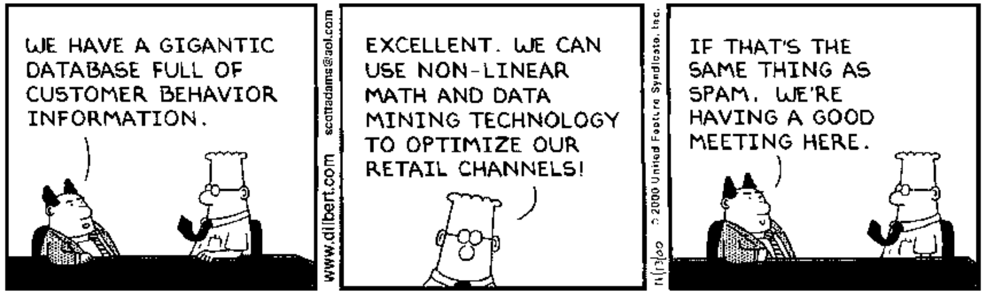
Freie Universität Berlin

Wdh.

(1) The need for data mining

(2) From business problems to data mining tasks

(3) Supervised vs. unsupervised methods

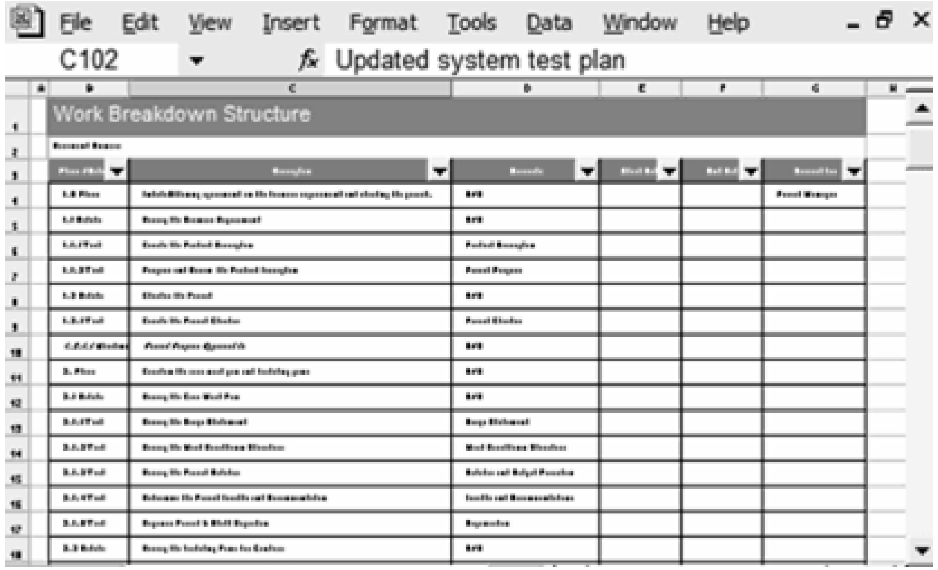


The need for data mining

Wdh.

“80% of the knowledge of interest in a business context can be extracted from data using **conventional tools**.” (Lusti)

- reporting
- query-languages (SQL, QBE, …)
- OLAP and spreadsheets



Disadvantages of conventional tools:

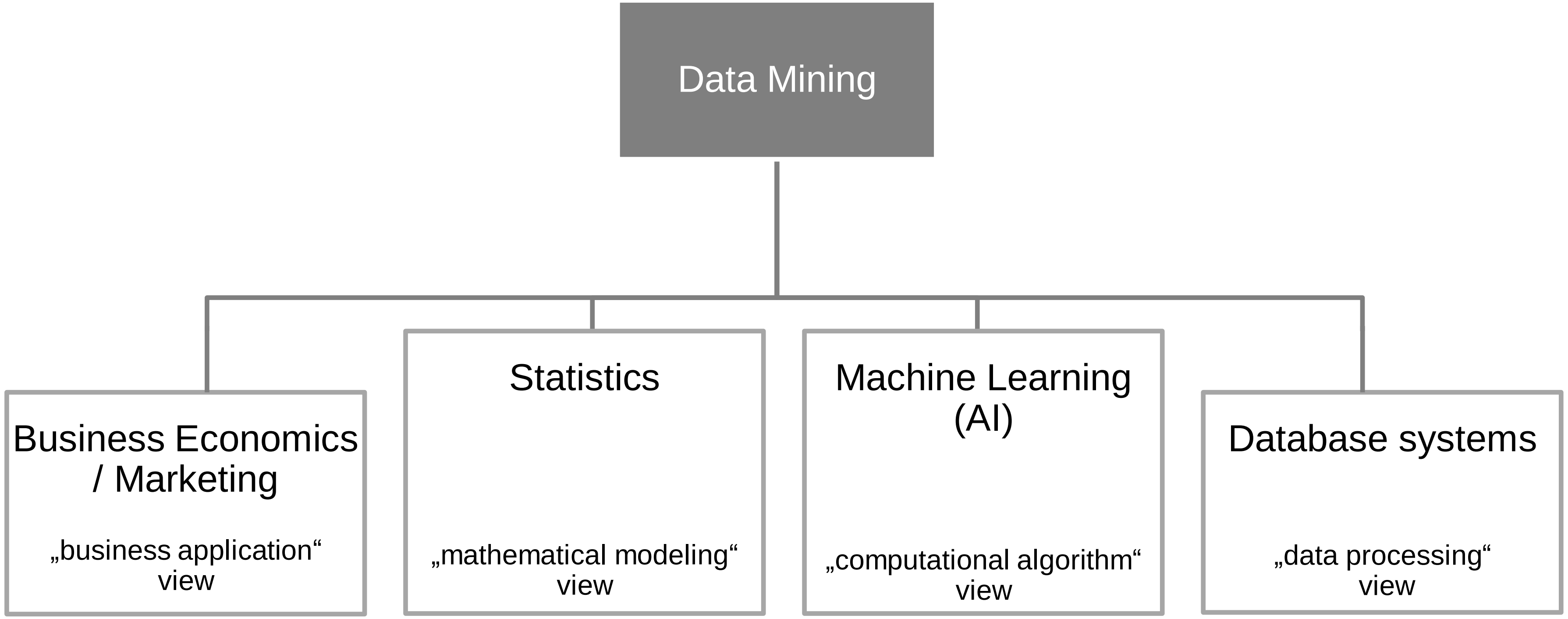
- Often, merely simple questions can be answered
- OLAP: query-focused and low complexity of analysis
  e.g., performance changes are visible, but what about the overall context/ reasons?
- Automation of knowledge discovery is difficult
  hypothesis needed
- Only small amounts of data may be handled (esp. spreadsheets)
  but exploding amount of raw data available

Ref.

Images: WBS excel (2009) | Flickr (cc by 2.0)

Major roots of data mining

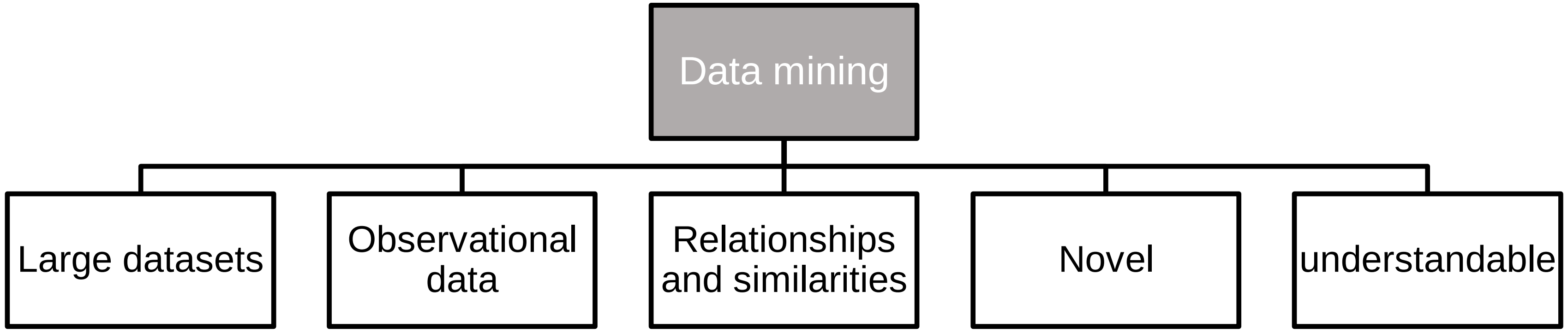
Wdh.



Ref.

Data mining: definition (1/3)

Wdh.



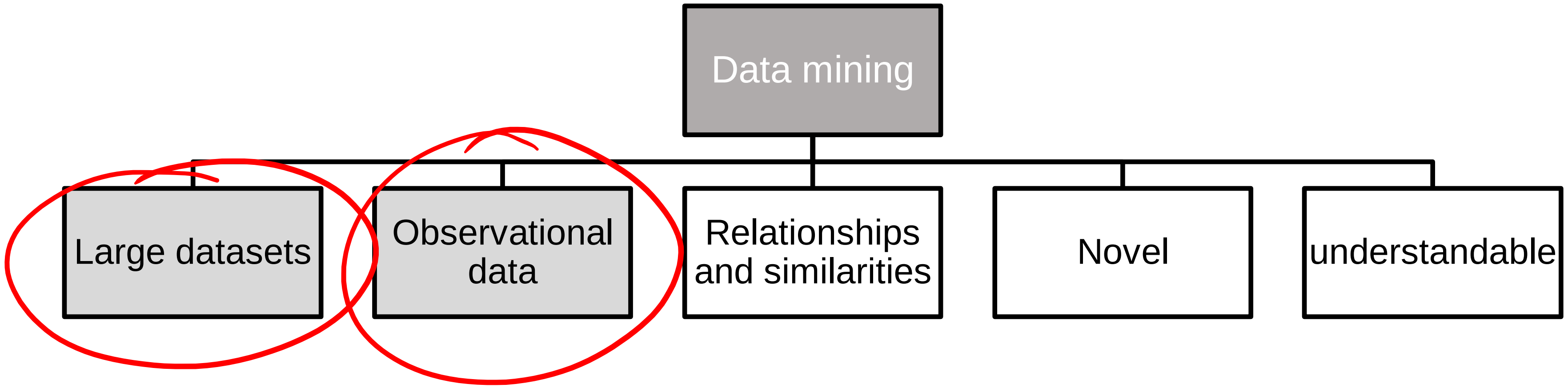
“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

(Hand, Mannila, Smyth (2001), Principles of Data Mining

Ref.

Data mining: definition (2/3)

Wdh.



Often large datasets:

Small datasets $\Rightarrow$ exploratory data analysis in statistics

Large datasets (as they exist in DWHs) provoke new problems

- Storage and access of data
- Runtime issues
- Determination of representativeness of data
- Difficulty to decide whether an apparent relationship is merely a chance occurrence or not

Observational data:

Data often collected for some other purpose than data mining

Objectives of the data mining exercise play no role in data collection strategy

e.g., DWH data relying on an airline reservation system or a bank account administration system

opposite: experimental data (as it is used quite often in statistics)

Ref.

Data mining: definition (3/3)

Wdh.

$y = f(x) + u$

Modell

Pattern

$p(y > 1{,}2 \wedge x < 50 \mid 0{,}5)$

Data mining

- Large datasets
- Observational data
- Relationships and similarities
- Novel
- understandable

Relationships and summaries:

often referred to as **models** or **patterns**

e.g., linear equations, tree structures, clusters, patterns in time series, …

Understandable:

Novelty is not sufficient to qualify relationships worth finding

Simple relationships may be preferred to complicated ones

Novel:

Novelty should be measured relative to users prior knowledge

Problem understanding

Ref.

Exercise: Data Mining vs. OLAP

Typical questions

Historie -> Zukunft

Historie

Fragestellung	Data Mining	OLAP
Kundenwert	Welche Kunden bieten uns das größte Deckungsbeitragspotenzial?	Wer waren letztes Jahr unsere 10 besten Kunden?

Kahoot-Fragen
www.kahoot.it
(über Smartphone oder Laptop)
PIN folgt

(Diese Folie ist nach der Vorlesung mit Lösungen verfügbar

Ref.

A typical Data Mining Process

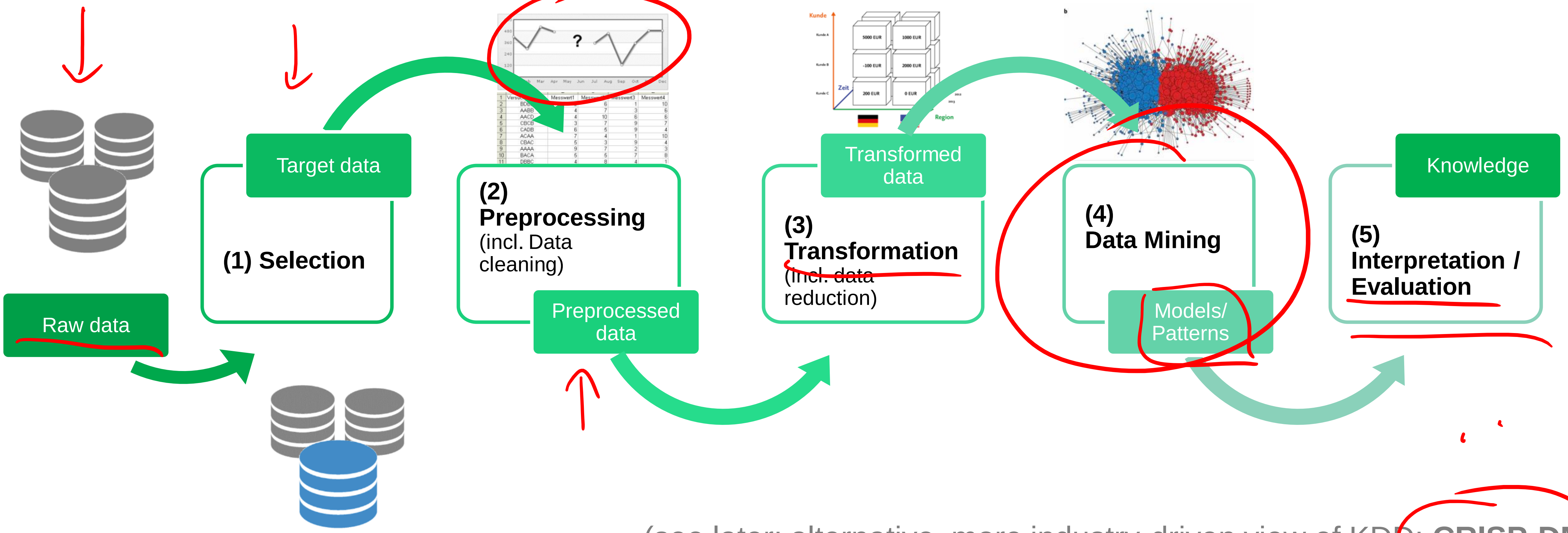
Knowledge discovery in databases (KDD)

100 → Nofe → 80% ⇒ Modell
Testen ← 20%

Data mining is often set in the broader context of **knowledge discovery in databases** (KDD)

The precise boundaries of the data mining part within the KDD process are not easy to state

Binär 0/1



(see later: alternative, more industry-driven view of KDD: **CRISP-DM**)

Ref. Fayyad, Piatetsky-Shapiro und Smyth 1996, S. 39 ff.

Agenda

Freie Universität Berlin

(1) The need for data mining

(2) From business problems to data mining tasks

(3) Supervised vs. unsupervised methods

Ref.

From business problems to data mining

Data mining is a **process** with well-understood stages based on

- application of information technology
- analyst‘s creativity
- business knowledge
- common sense

Counterexample: Youtube ads, see e.g. The Guardian, 2017

“Major brands ... pulled their ads after they were found to be appearing next to videos promoting extremist views or hate speech”

We look at typical ***tasks*** and examples, then at the ***process***

Decompose a data analytics problem into pieces such that you can solve a known task with a tool

There is a large number of data mining algorithms available, but only a limited number of data mining tasks

We will illustrate the **fundamental concepts** based on

- Classification
- Regression

Ref.

Classification

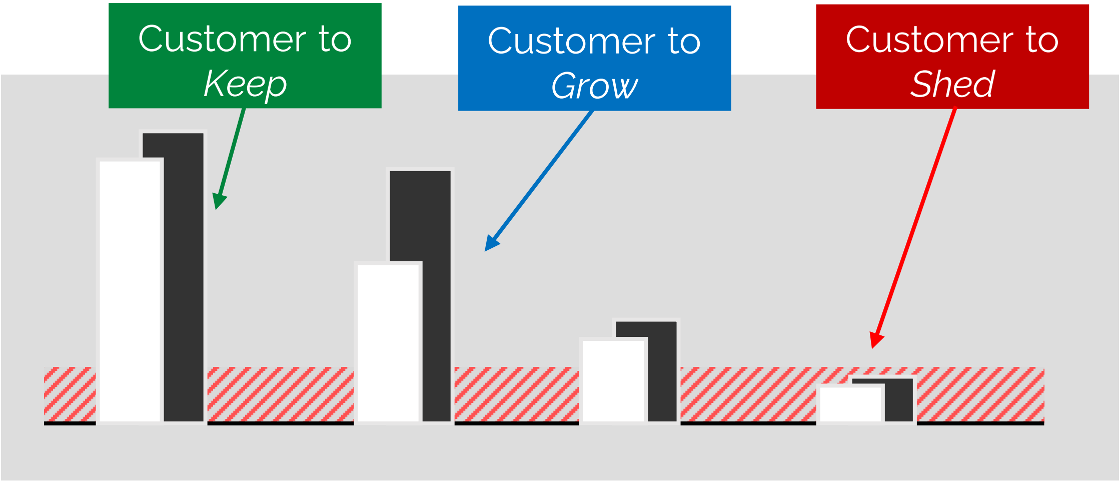
Freie Universität Berlin

Classification attempts to **predict**, for each individual in a population, which of a (small) *set of classes* that individual belongs to

"Among all the customers of a cellphone company, which are likely to correspond to a given offer?"

~ *predict whether something will happen*

Classification algorithms provide models that determine which class a **new** individual belongs to (and its probability).



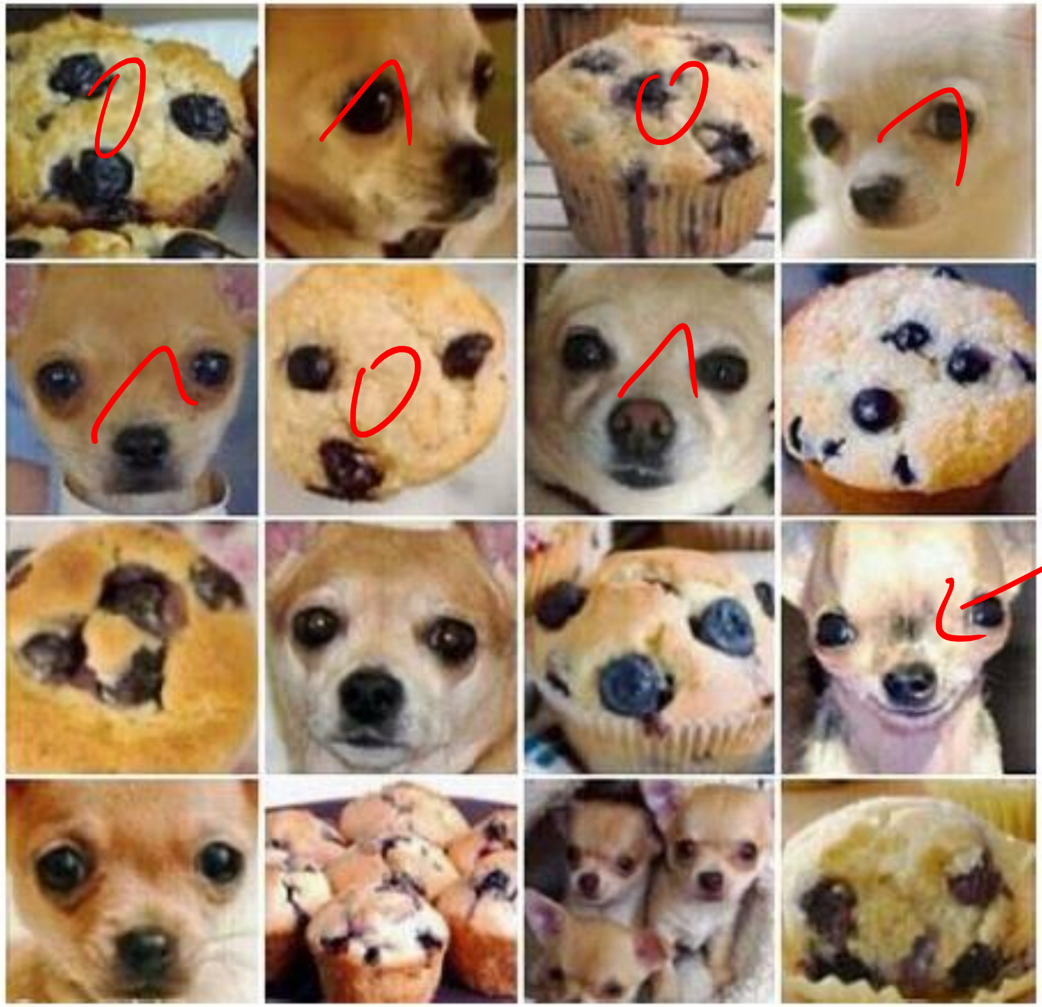
Classification is related to scoring (for instance in Customer Relationship Management) as the underlying value is categorical.

Ref.

Dog or Muffin?

Binär

1 0



Deep Learning required

(we will talk about this in July)

Ref.

Regression

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

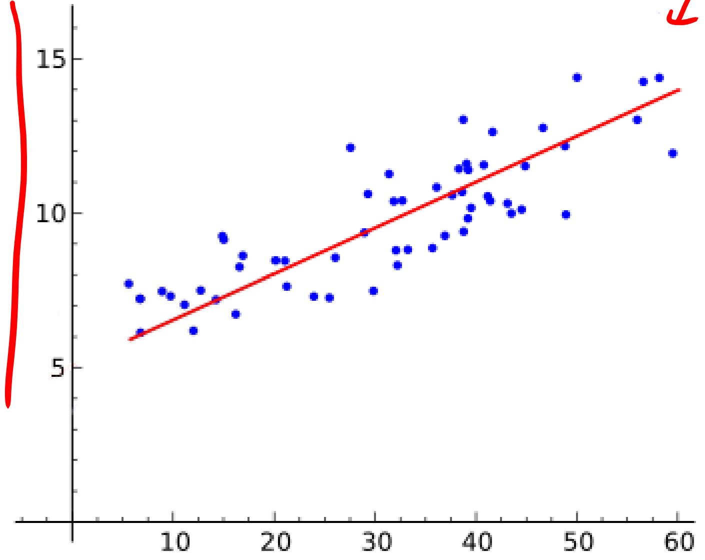
Regression (value estimation) attempts to estimate or **predict**, for each individual, the (continuous) *numerical value* for that individual

"How much will a given customer use the service?"

Predicted variable: service usage

~ *predict how much something will happen*

Generate regression model by looking at other, similar individuals in the population



Be careful:
linear regression vs. logistic regression

Ref.

Exercise: Classification or Regression problem?

Examples

a) Will this customer purchase service *S1* if given incentive *I*? C Binär

b) Which service package (*S1, S2*, or none) will a customer purchase if given incentive *I*? C

c) How much time will this customer spend on our web service? R

d) How long after buying a product can a repeat purchase by customer X be expected? R

e) Which potentially profitable customers are most likely to move to a competitor? →R C

~~Profitabilität~~

Webseite	Low	Med	High
Low	–	–	–
Med	–	–	S2
High	–	S2	S1

3 Min.

Ref.

Another fundamental data mining task: ….

Quelle: www.welt.de

Ref.

…. Similarity matching

Freie Universität Berlin

Distanzmaße

	Alter				Einkom
P1	20				100000
P2	21				80000

21 − 20 = 1

20.000

Similarity matching attempts to **identify similar individuals** based on the data known about the individuals

Find similar entitites

Basis for making **product recommendations**

- Find people who are similar to you in terms of the products they have liked or purchased

Ref.

Image: www.welt.de

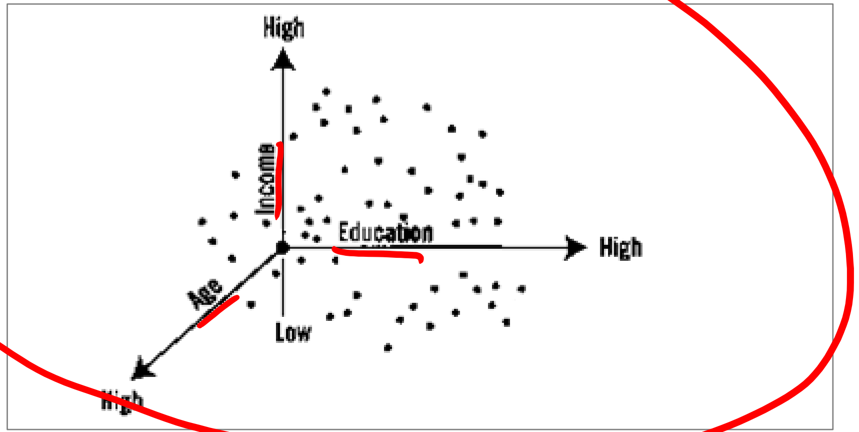
Clustering

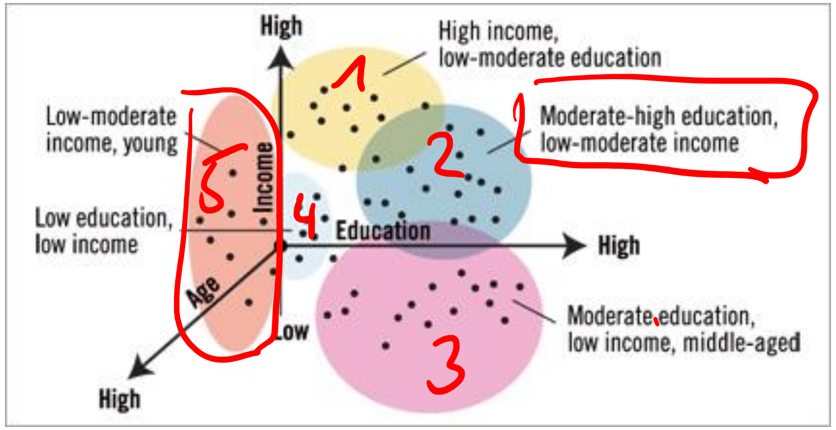
Clustering attempts to **group** individuals in a population together by their similarity, but *without regard to any specific purpose*

Do customers form natural groups or segments?

Result: groupings of the individuals of a population

Useful in preliminary domain exploration





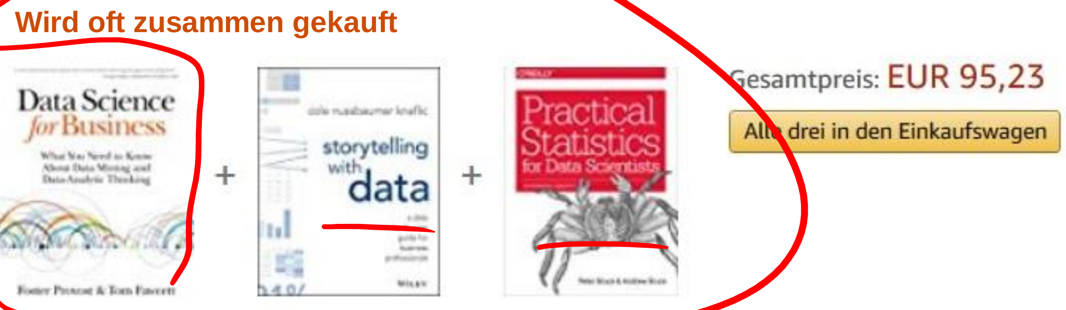
Ref.

Co-occurence grouping

A B

Attempts to find associations between entities based on transactions involving them aka **association rules** or **market-basket analysis**

> *"What items are commonly purchased together?"*



Considers similarity of objects based on their appearing together in transactions

Included in recommendation systems (people who bought X also bought Y)

Result: a description of items that occur together

Ref.

Image: Amazon.com,

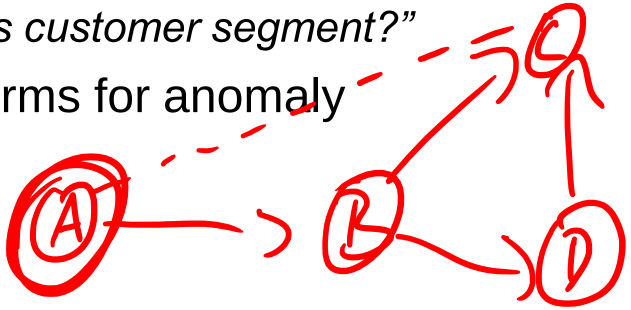
Some more data mining tasks

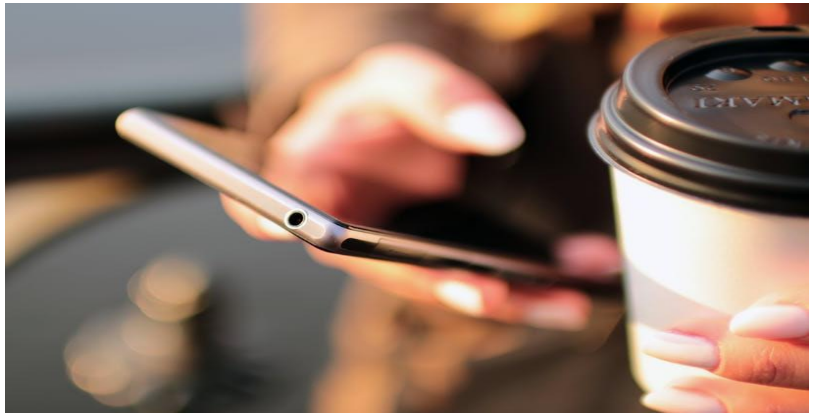
Freie Universität Berlin

- **Profiling** attempts to characterize the typical behavior of a group or population, aka **behavior description**

  “What is the typical cellphone usage of this customer segment?”

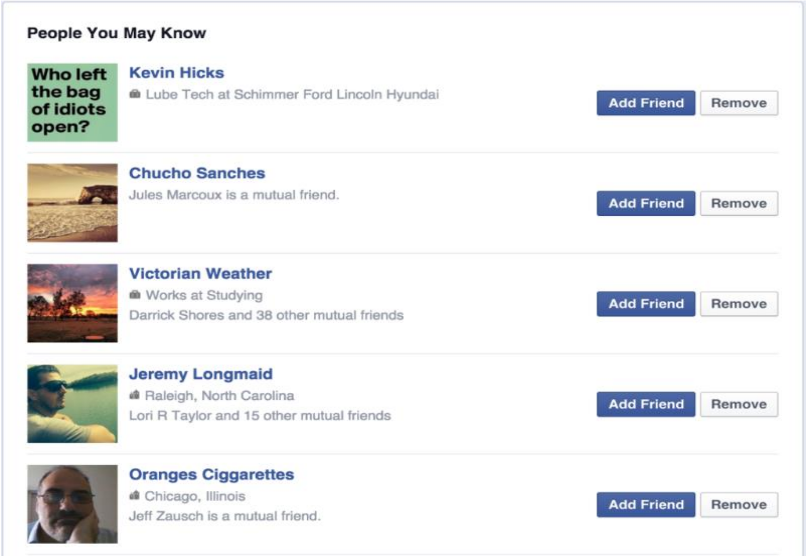
  Often used to establish behavioral norms for anomaly detection (fraud detection)





- **Link prediction** attempts to predict connections between data items (→ social network systems)

  “Since you and Karen share ten friends, maybe you‘d like to be Karen‘s friend?“



- **Data reduction** attempts to take a large data set of data and replace it with a smaller set of data that contains the relevant information
  (Easier processing, but often loss of information)

Ref.

Image: Briley (2015)

Agenda

(1) The need for data mining

(2) From business problems to data mining tasks

(3) Supervised vs. unsupervised methods

Ref.

Supervised vs. unsupervised

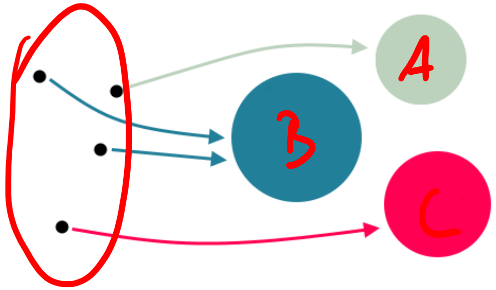
Überwachtes Lernen

Supervised Learning

“Can we find groups of customers who have particularly high likelihoods of cancelling their service soon after their contracts expire?”

→ specific target

Classification



input → output

Person → gut / schlechter Abschluss

Supervised and unsupervised tasks require different techniques

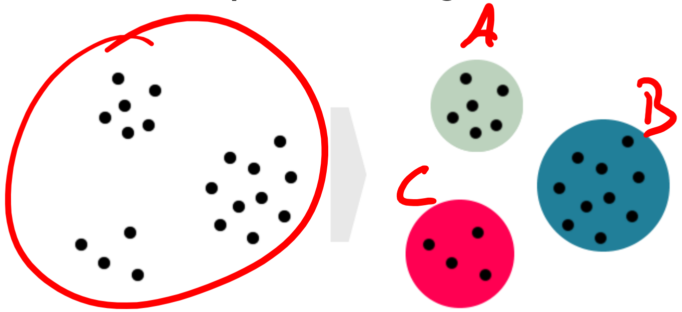
Unüberwachtes Lernen

Unsupervised Learning

“Do our customers naturally fall into different groups?”

→ no specific target

Clustering



Person

There is no guarantee that unsupervised tasks provide meaningful results

Ref.

Supervised and unsupervised techniques

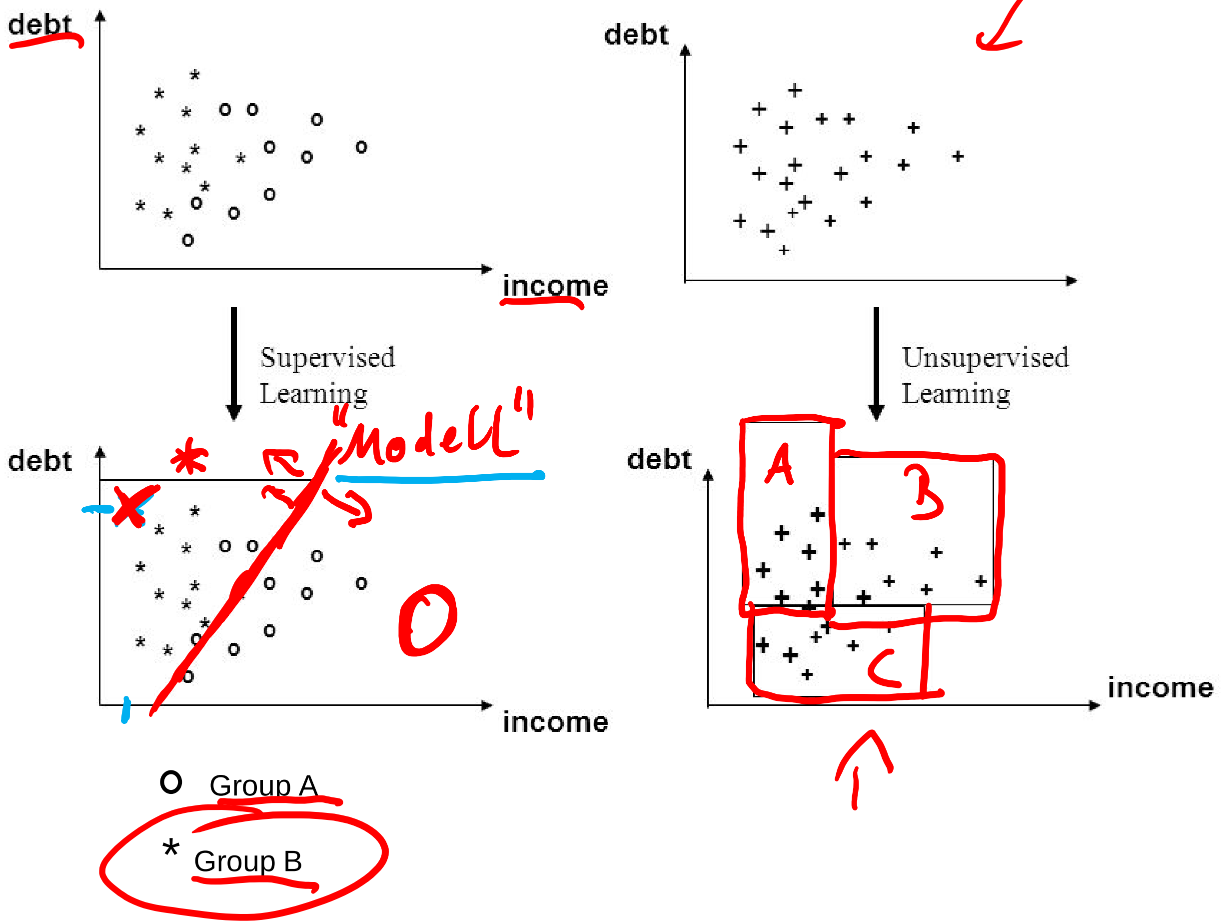
Classification and regression are generally solved with **supervised** techniques

Clustering, co-occurence grouping, and profiling are generally **unsupervised**

Similarity matching and link prediction could be either

Supervised

Unsupervised



Ref.

Image: [Datacafeblog (2015)]

Supervised vs. Unsupervised vs. Reinforcement Learning

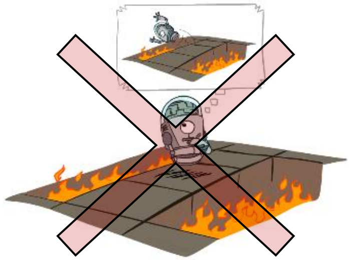
Search by yourself

e.g. https://towardsdatascience.com/machine-learning-101-supervised-unsupervised-reinforcement-beyond-f18e722069bc

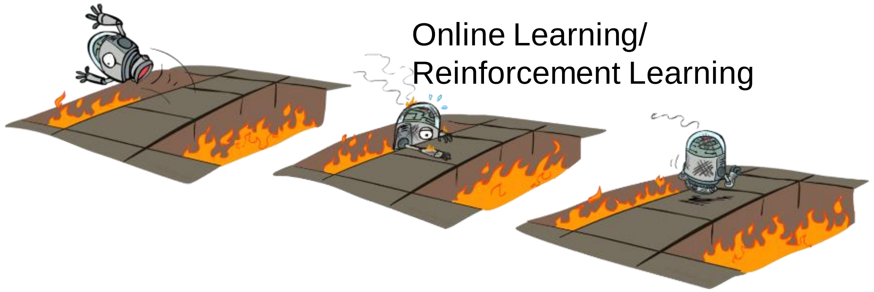
Learning Approach	Learn from…. (input? - output?)	Learn for…. (target?)	Example?
Supervised			
Unsupervised			
Reinforcement			

e.g., Spiegel Online, 2021
“KI zockt besser als der Mensch“
Corresponding article
Ecoffet, A., Huizinga, J., Lehman, J. et al. First return, then explore. Nature 590, 580–586 (2021).

- A.I. Learns to Drive From Scratch in Trackmania (very well explained youtube video)



Offline Solution (model based)



Online Learning/ Reinforcement Learning

Quelle: Course188 Intro to AI at UC Berkeley

10 Min.

- Demo Reinforcement Learning

Ref.

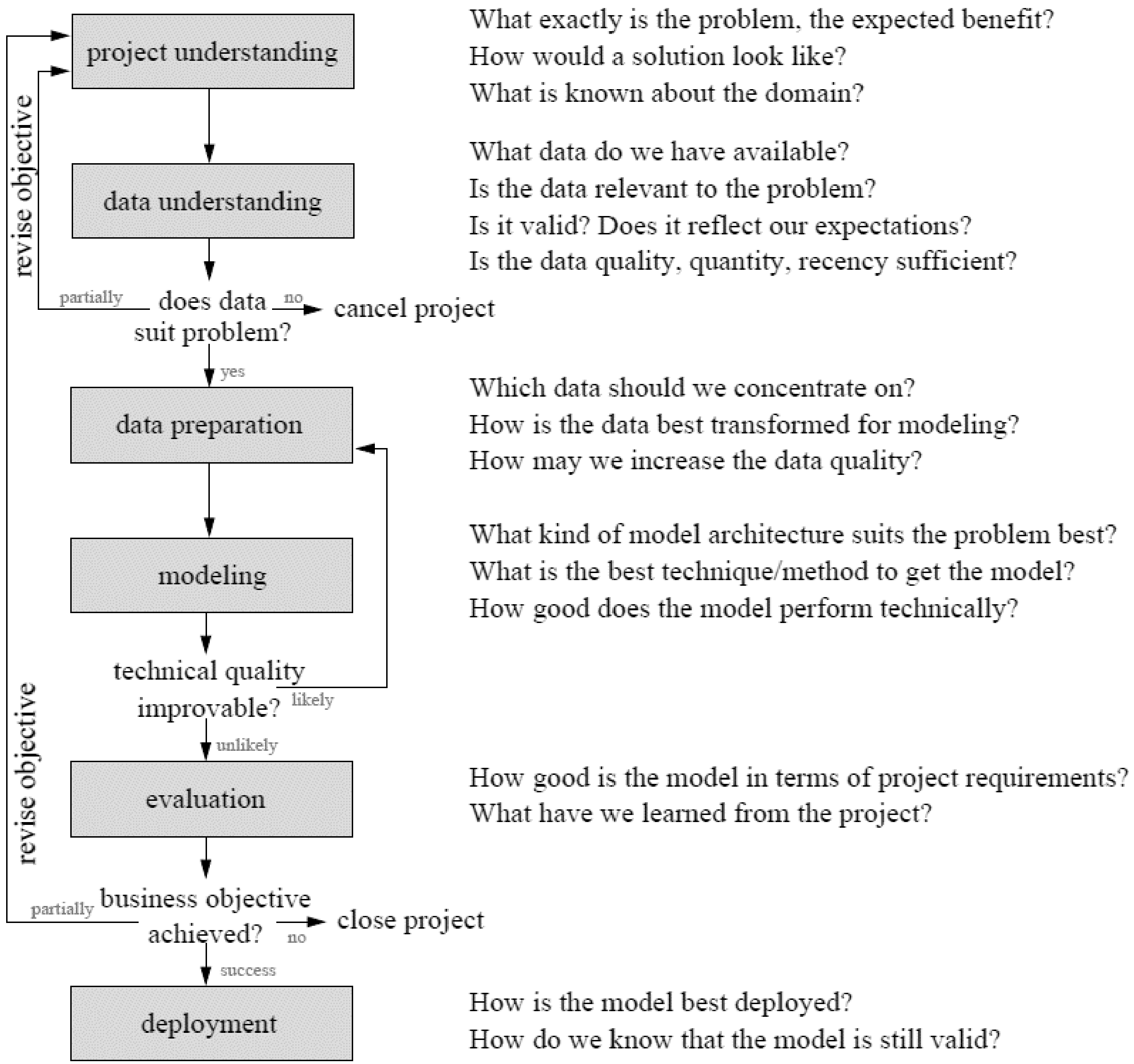
Outlook: The data mining process - CRISP-DM

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)

Fragen?

- ✓ The need for data mining
- ✓ From business problems to data mining tasks
- ✓ Supervised vs. unsupervised methods (vs. reinforcement learning)
- o The data mining process – CRISP-DM

Data Mining

Recommended reading

Provost, F. Fawcett, T.	Chapter 2
Berthold et al.	Chapters 1, B, C
Lusti, M.	Data Warehousing und Data Mining (Chapter 6)
Hand, D. et al.:	Principles of Data Mining (esp. Chapters 1, 5, 6 and 11)

Ref.

Bibliography

- Azevedo, A. I. R. L. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.
- Ecoffet, A., Huizinga, J., Lehman, J. *et al.* (2021) First return, then explore. *Nature* 590, 580–586. https://doi.org/10.1038/s41586-020-03157-9
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.
- Hand, David J., Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. MIT press, 2001.
- Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).

Ref.