

Business Intelligence

05a Data Mining Introduction (continued)

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-stud	dy
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 unders	tanding	17.5.	Python-Basics-Online Exercise	Python-Analytics	Chap. 1
	W6	22.5.	Data Understanding, Data Vis	sualization	24.5.	No lectures, but bonus tasks 1.) Co-Create your exam		Chap. 2
	W7	29.5.	Data Preparation		31.5.	2.) Earn bonus points for the exam		
	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		1
	W10	19.6.	Guest Lecture		21.6.	Fitting a Model II		1
Deep- ening	W11	26.6.	How to avoid overfitting	ng	28.6.	What is a good Model?		- 1
	W12	3.7.	Project status update Evidence and Probabili		5.7.	Similarity (and Clusters) From Machine to Deep Learning I	•	
	W13	10.7.			12.7.	From Machine to Deep Learning II		1
	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.	Projektberi	cht

Last lesson



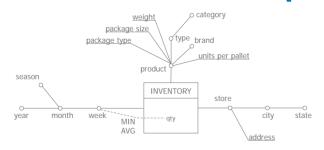
✓ How can multidimensional data models be developed and stored?

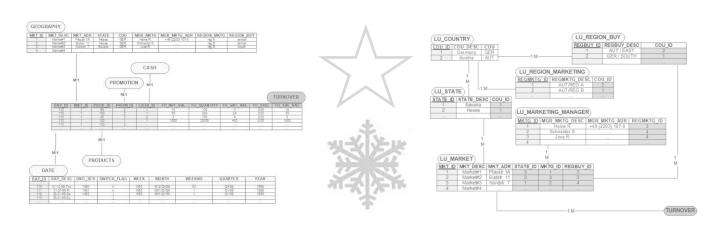
Conceptual modeling

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





(1) The need for data mining

(2) From business problems to data mining tasks

(3) Supervised vs. unsupervised methods

WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.



EXCELLENT. WE CAN
USE NON-LINEAR
MATH AND DATA
MINING TECHNOLOGY
TO OPTIMIZE OUR
RETAIL CHANNELS!





The need for data mining

Freie Universität

Berlin

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

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

_ & × Format Tools Window Updated system test plan Vork Breakdown Structure habitalities agreement on the foreconseparated and charles the proofs. Broom the Brown Brown and LAITHE Easts to Period Benegle Propertied From the Posted Secretor Personal Property Charles He Pound LAST OF Fresh Mr. Roant Shado Examine the core and pre-rad holiday gran B. C. B. Co. Co. Street, St. San World Str. BASTM Berry the Berry Blackware Berge Bildement Beng th West Bootless House Autobas and Autool Passacher 3.0.4Tm2 Belower the French treathers of Brown and des fundly self to mesself free Brong the budding Franchis Gradies

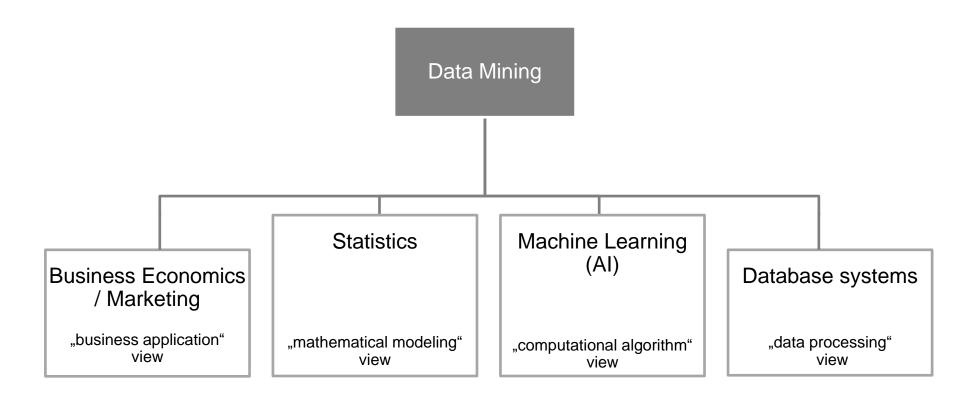
Disadvantages of conventional tools:

- 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)

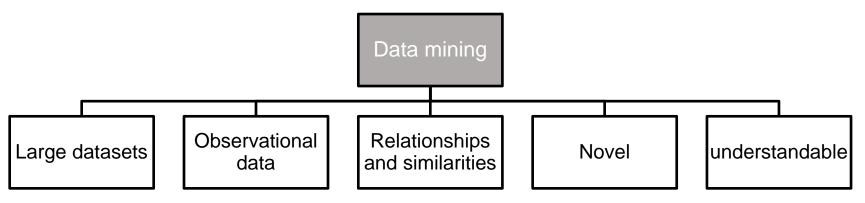
Major roots of data mining





Data mining: definition (1/3)



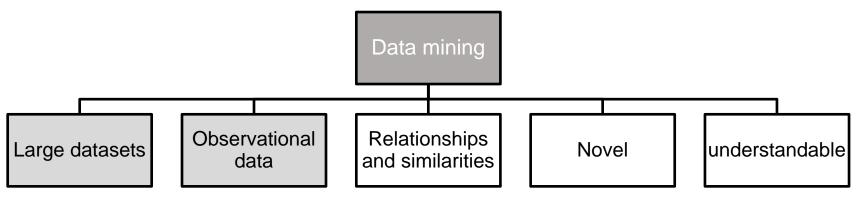


"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

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

Data mining: definition (2/3)





Often large datasets:

Small datasets ⇒ 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

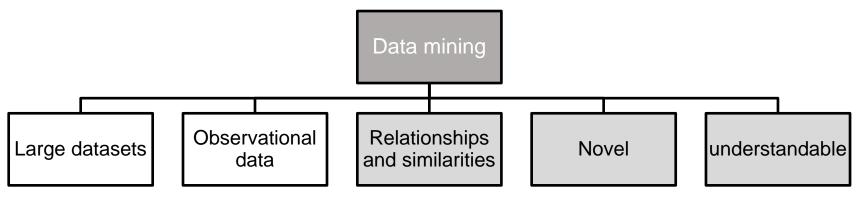
Objectives of the data mining exercise play no role in data collection strategy

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)

Data mining: definition (3/3)





Relationships and summaries:

often referred to as **models** or **patterns**e.g., linear equations, tree structures, clusters, patterns in time series, ...

Novel:

Novelty should be measured relative to users prior knowledge

Understandable:

Novelty is not sufficient to qualify relationships worth finding

Simple relationships may be preferred to complicated ones

Exercise: Data Mining vs. OLAP



Typical questions

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

Kahoot-Fragen

<u>www.kahoot.it</u>

(über Smartphone oder Laptop)

PIN folgt

(Diese Folie ist nach der Vorlesung mit Lösungen verfügbar

A typical Data Mining Process

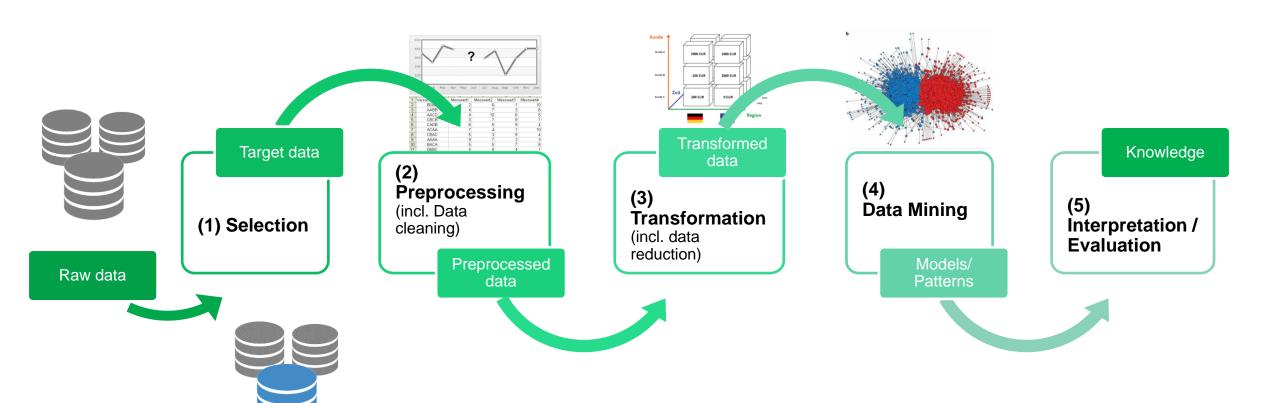
Freie Universität

Berlin

Knowledge discovery in databases (KDD)

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

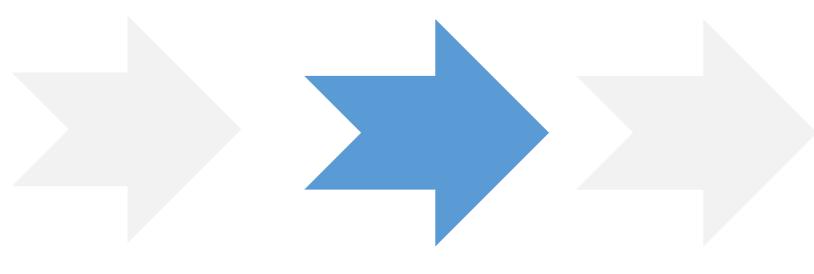


(see later: alternative, more industry-driven view of KDD: CRISP-DM)

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

Agenda





- (1) The need for data mining
- (2) From business problems to data mining tasks
- (3) Supervised vs. unsupervised methods

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. <u>The Guardian, 2017</u> "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

Classification

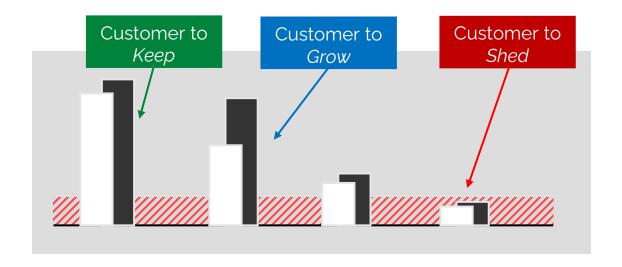


Classification attempts to **predict**, for each individual in a population, which of a (small) set of classes that individual belongs to

"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.

Regression



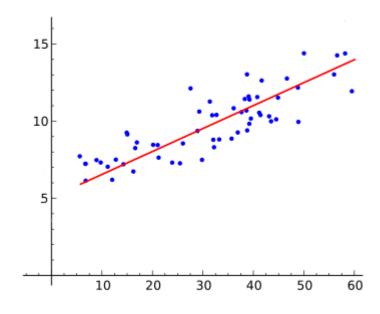
Regression (value estimation) attempts to estimate or **predict**, for each individual, the (continuous) *numerical* value for that individual

"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



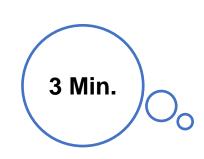
Be careful: linear regression vs. logistic regression

Exercise: Classification or Regression problem?



Examples

- a) Will this customer purchase service S1 if given incentive I?
- b) Which service package (S1, S2, or none) will a customer purchase if given incentive I?
- c) How much time will this customer spend on our web service?
- d) How long after buying a product can a repeat purchase by customer X be expected?
- e) Which potentially profitable customers are most likely to move to a competitor?



Another fundamental data mining task:





Quelle: www.welt.de

Clustering

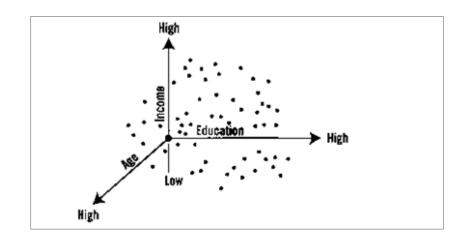


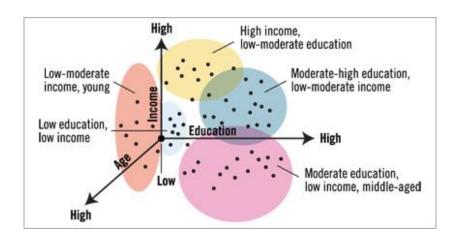
Clustering attempts to **group** individuals in a population together by their similarity, but *without* regard to any specific purpose

Do customers form natural groups or segments?

Result: groupings of the individuals of a population

Useful in preliminary domain exploration





Co-occurence grouping



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

Wird oft zusammen gekauft



Gesamtpreis: EUR 95,23

Alle drei in den Einkaufswagen

Customers who bought this item also bought







Cuisinart DLC-2A Mini-Prep Plus Food Processor, 24 Ounce, Pink DISCONTINUED BY...



KitchenAid KHM512PK 5-Speed Ultra Power Hand Mixer, Pink 1,921 Click for details Prime



Smeg 1.7-Liter Kettle-Pastel Green ★★★☆☆ 56 \$129.95 *Prime*

Ref. Image: Amazon.com,

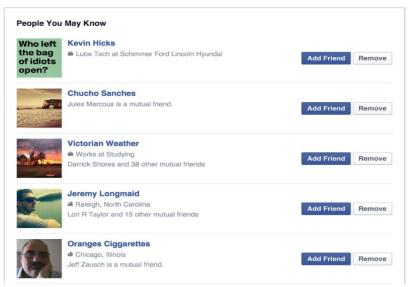
Some more data mining tasks



22

- 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

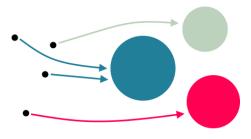
Supervised vs. unsupervised



Supervised Learning

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

→ specific target

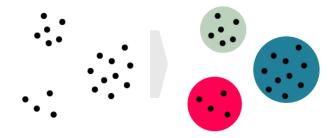


Supervised and unsupervised tasks require different techniques

Unsupervised Learning

"Do our customers naturally fall into different groups?"

→ no specific target



There is no guarantee that unsupervised tasks provide meaningful results

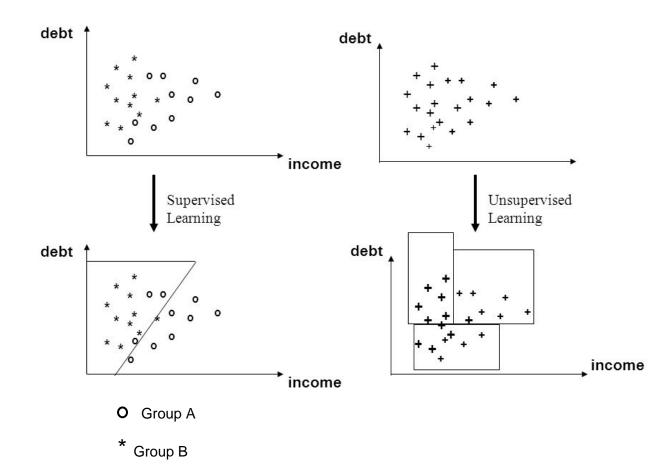
Supervised and unsupervised techniques



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



Ref. Image: <u>Datacafeblog (2015)</u>

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., <u>Spiegel Online</u>, <u>2021</u>
"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)



Quelle: Course188 Intro to AI at UC Berkeley



Demo Reinforcement Learning

Ref.

Outlook: The data mining process - CRISP-DM

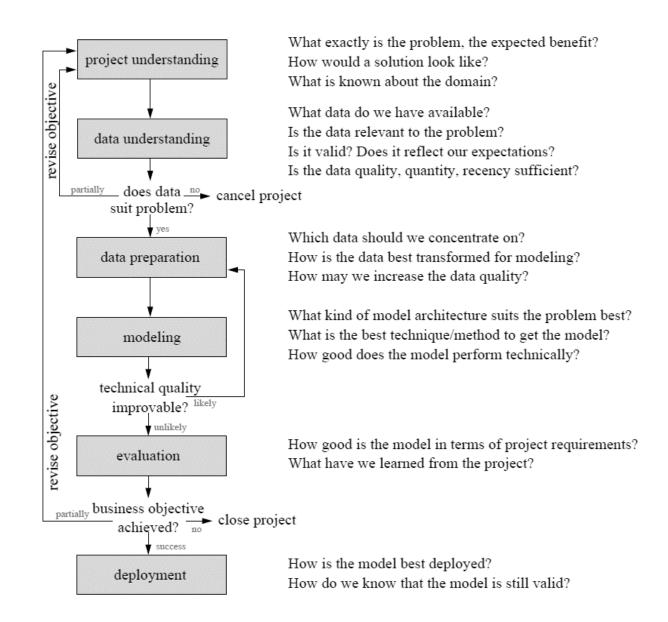


Cross
Industry
Standard
Process for
Data
Mining

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)
- The data mining process CRISP-DM

Data Mining



Recommended reading

Provost, F. Chapter 2

Fawcett, T.

Berthold et al. Chapters 1, B, C

Lusti, M. Data Warehousing und Data Mining (Chapter 6)

Principles of Data Mining (esp. Chapters 1, 5, 6 and 11) Hand, D. et al.:

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. *Al 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).