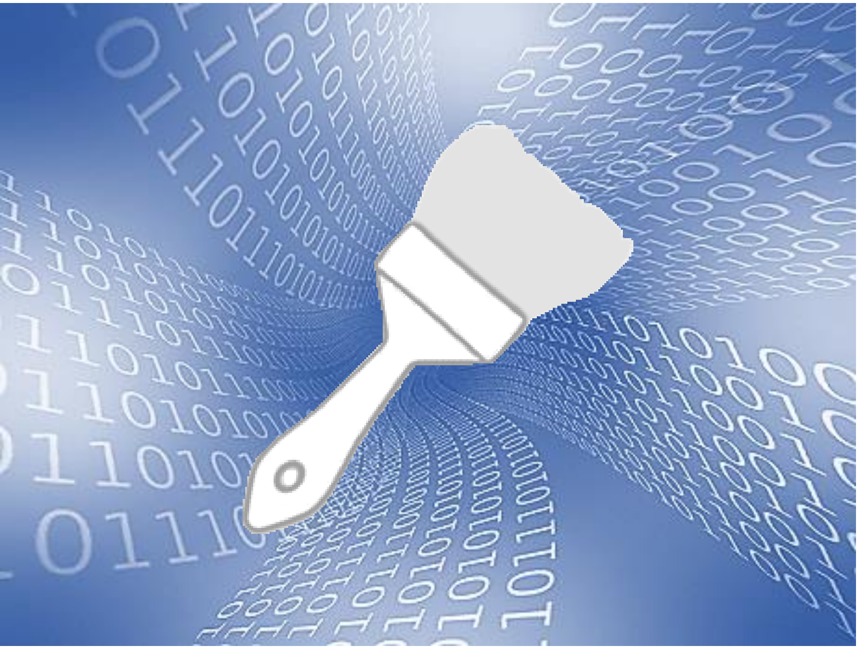
Prof. Dr. Bastian Amberg
School of Business & Economics
Department of Information Systems



Business Intelligence

08 Data Preparation

Prof. Dr. Bastian Amberg

(summer term 2024)

5.6.2024

Schedule

		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.	🚆		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 I	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 Visualization II	31.5.				
	W8	5.6.	Data Preparation	7.6.	Predictive Modeling I (10:00 -12:00)	BI-Project	Start	Case Study
	W9	12.6.	Predictive Modeling II, Fitting a Model I	14.6.	Python-Analytics-Online Exercise		|	
	W10	19.6.	*Guest Lecture Dr. Ionescu*	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	
Ref.					*Klausur 1.Termin, 31.7.'24* *Klausur 2.Termin, 2.10.'24*	Projektbericht		

Last Lesson

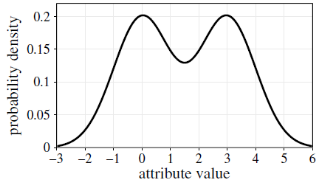
data visualization

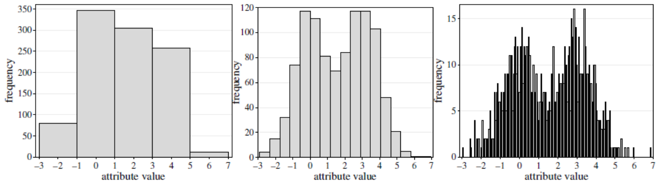
→A1 | A2 | A3

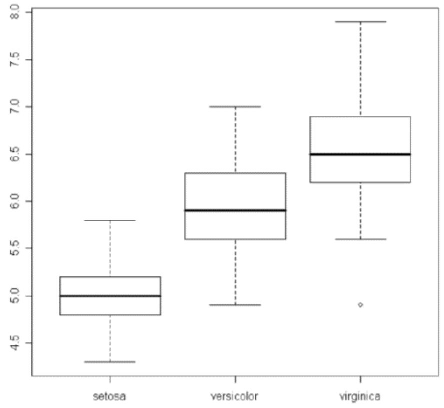
PCA1 $\overset{1,5}{\lambda_1} \cdot A_1 + \overset{2,75}{\lambda_2} \cdot A_2 + \overset{4}{\lambda_3} \cdot A_3$

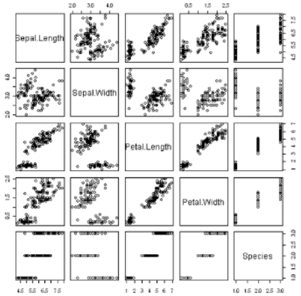
PCA2 $8{,}37 A_1 + 0{,}2 A_2 + 0{,}1 A_3$

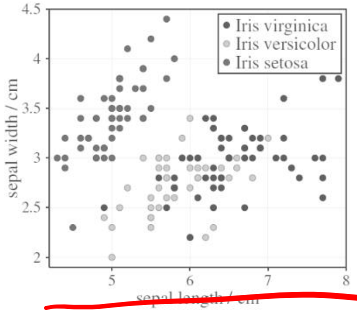
- ✓ Low-dimensional relationships
  - ✓ Univariate Analysis
  - ✓ Bivariate Analysis









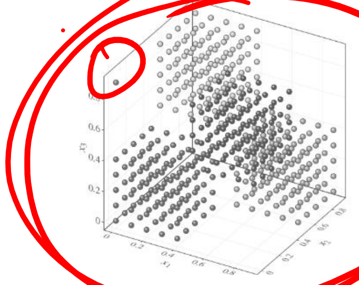


- ❖ Pearson's correlation coefficient
  >> video for explanation
- ❖ Rank correlation coefficients
  >> video for explanation
  Spearman's rho
  Kendall's tau

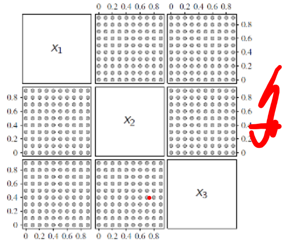
- ✓ Higher-dimensional relationships
  - ✓ Principal Component Analysis

Clusteranalyse

How to preserve original "structure" in lower dimensional representations?

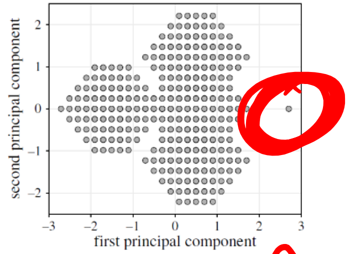


3D



(-)

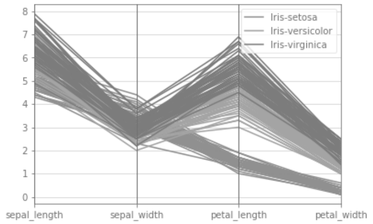
2D

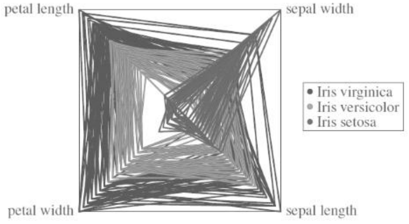


(+)

2D

- ✓ Parallel Coordinates





Examples python

Ref.

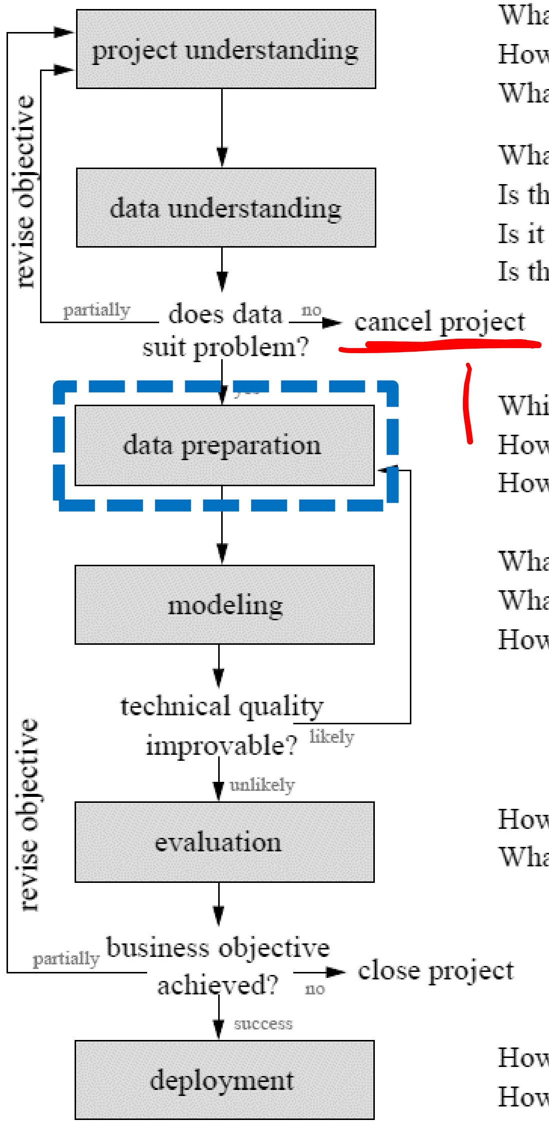
CRISP-DM

Cross
Industry
Standard
Process for
Data
Mining

Iteration as
a rule

Process of data
exploration

Implementation of the
KDD Process



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?

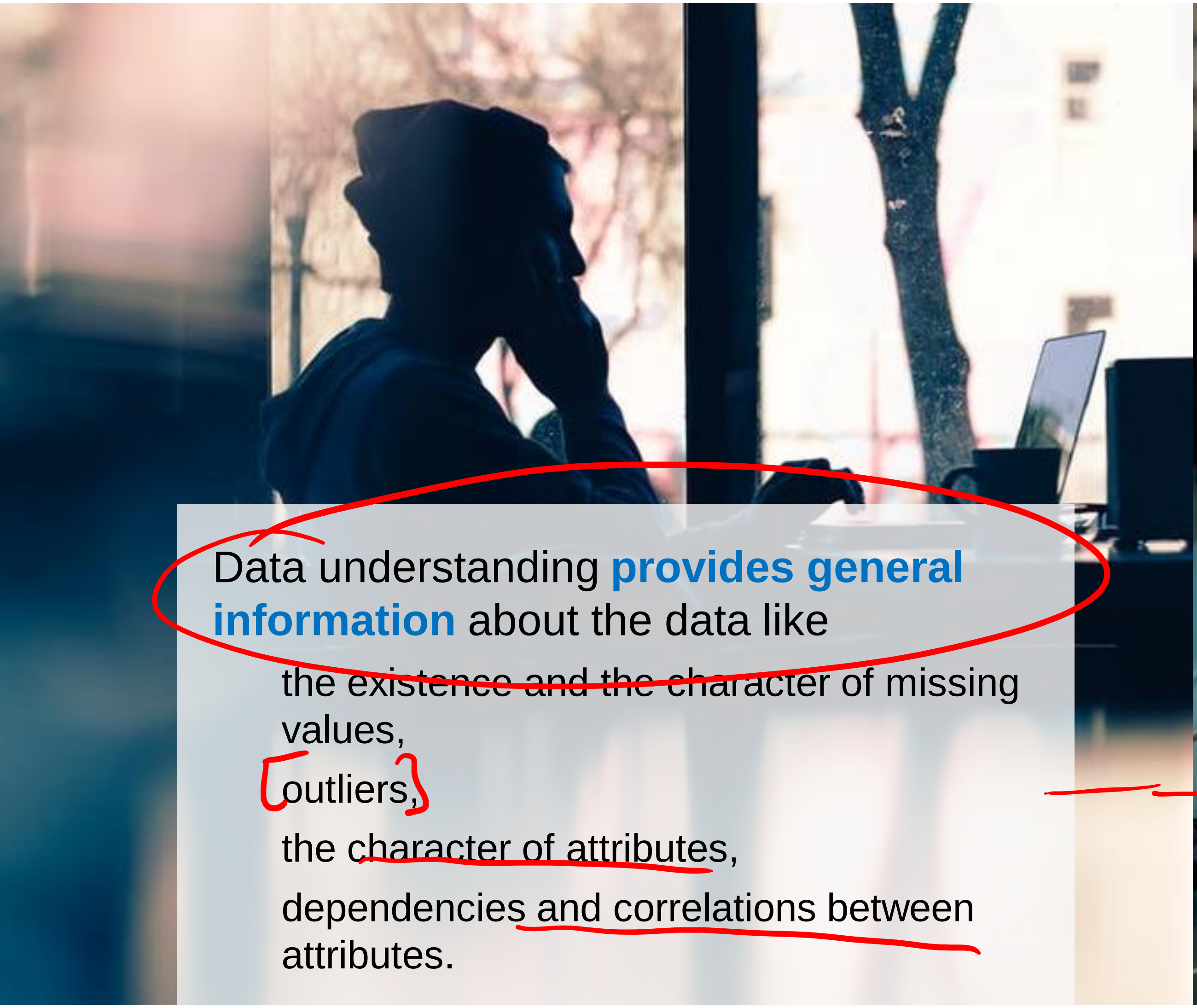
What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

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?

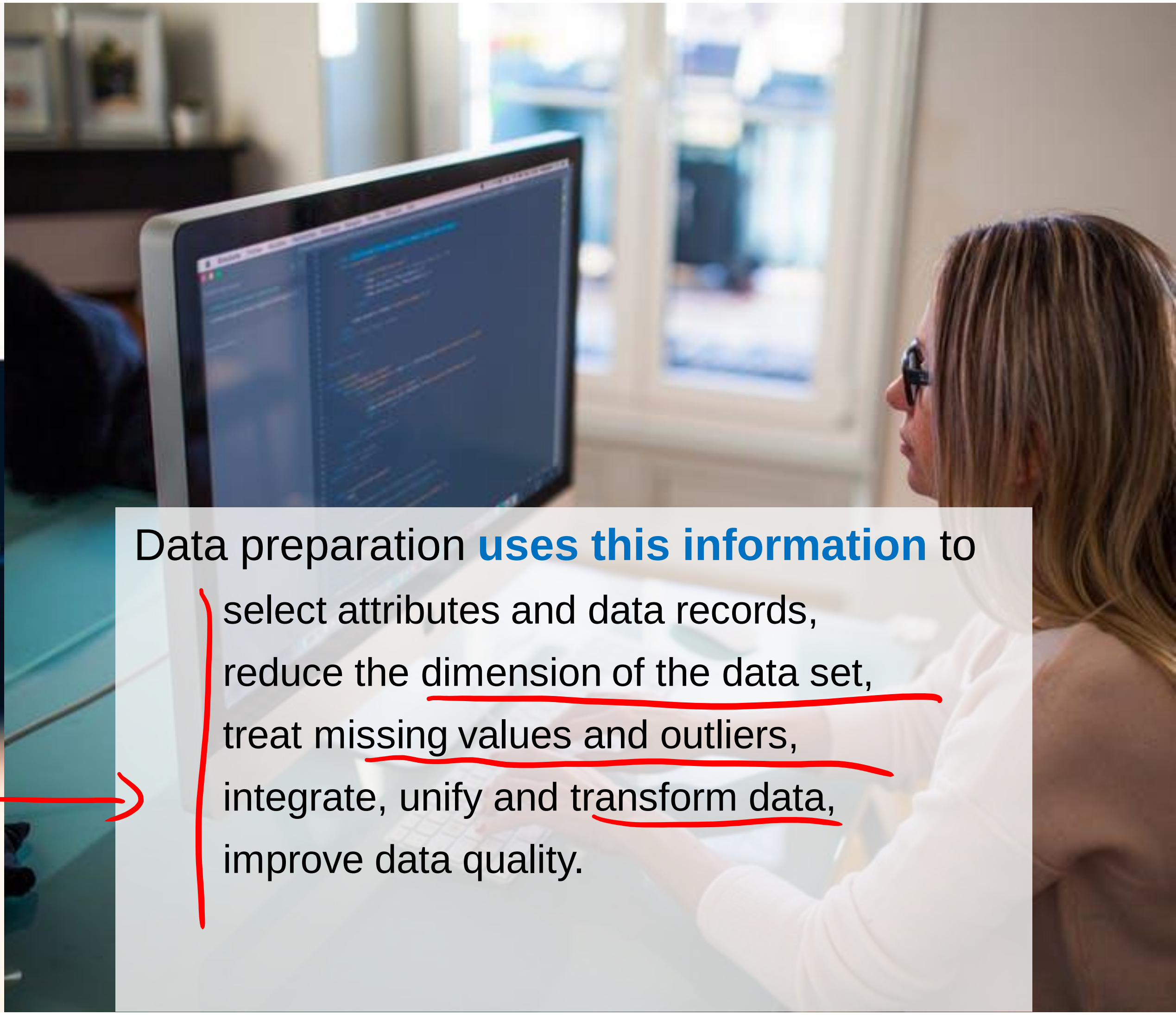
Ref. Wirth / Hipp (2000), Azevedo (2008)

Data understanding vs. Data preparation



Data understanding **provides general information** about the data like

- the existence and the character of missing values,
- outliers,
- the character of attributes,
- dependencies and correlations between attributes.

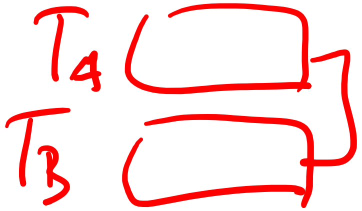


Data preparation **uses this information** to

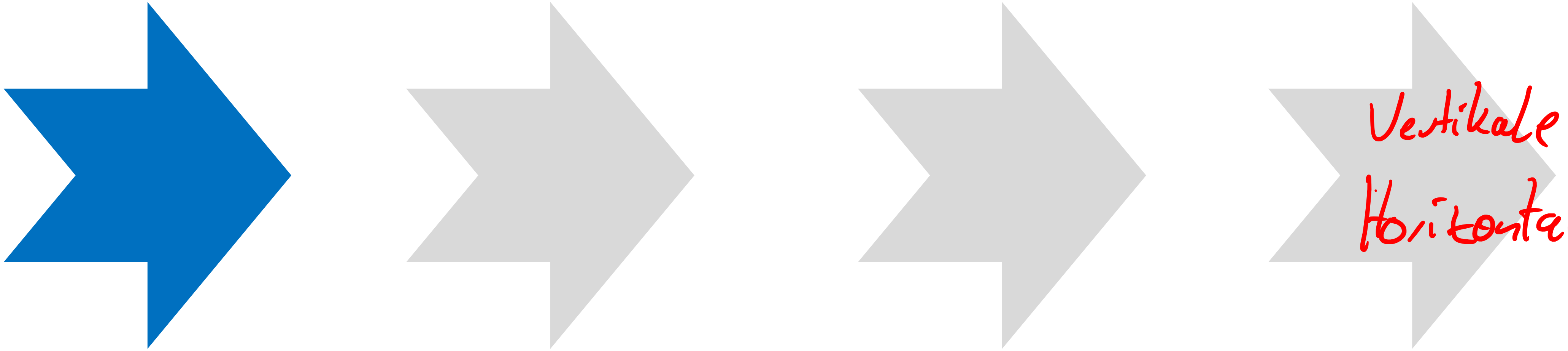
- select attributes and data records,
- reduce the dimension of the data set,
- treat missing values and outliers,
- integrate, unify and transform data,
- improve data quality.

Ref.

Agenda



Data Preparation



Vertikale

Horizontale

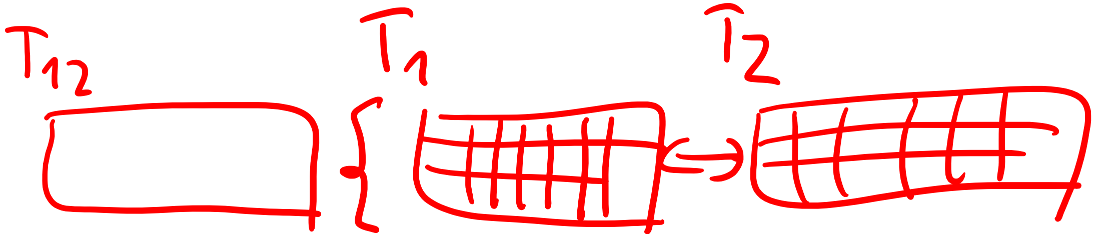
(1) Data selection

(2) Data cleaning

(3) Data transformation

(4) Data integration

DE



Ref.

Feature extraction

Constructing (new) features from the given attributes

$A_1 \mid \underbrace{A_2 \mid A_3}_{E_{23}}$

$A_1 \mid E_{23}$

Example:

We are interested in finding the best workers in a company.

There are attributes available like

- the *tasks a worker* has finished *within each month*,
- the *number of hours* he/she has *worked each month*,
- the *number of hours* that are *normally needed* to finish each task.

tasks

hours

In principle, these attributes contain information about the efficiency of the worker.

It might be more useful to **define a new attribute "efficiency"**, which is …. For example?

$\frac{\# tasks}{\# hours}$

The proportion of hours spent to finish a task to hours normally needed to finish a task.

Ref.

Image: Roosevelt Library (NLFDR(

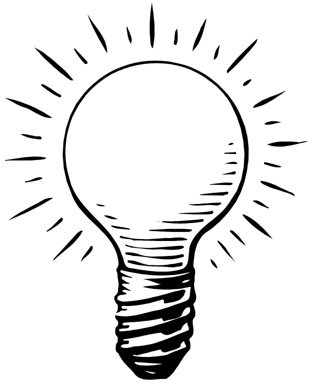
Dimensionality reduction for feature extraction

$\lambda_1 A_1 + \lambda_2 A_2 + \lambda_3 A_3$

Dimensionality reduction techniques like **PCA** can also be considered as feature extraction methods.

But such automatic feature extraction methods usually lead to features that **can no longer be interpreted** in a meaningful way.

How to understand a feature that is a linear combination of 10 attributes?



Therefore, in most cases, either knowledge-based, problem-dependent feature extraction methods or feature selection techniques are preferred.

Ref.

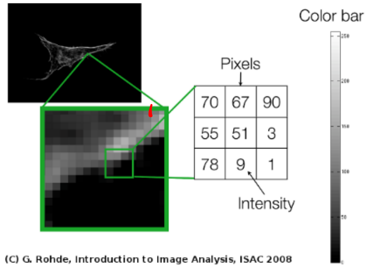
Feature extraction and selection

NEU

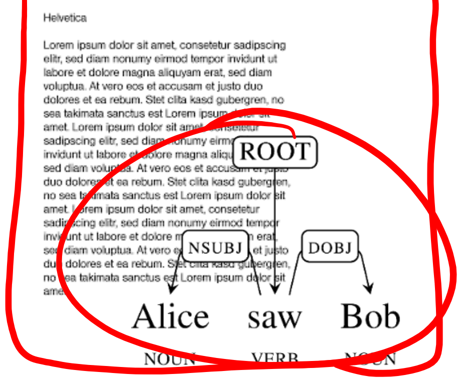
Selecting features and creating subsets

Feature extraction is especially relevant for complex data types:

- Text data analysis – frequency of keywords, …
- Time series and image data analysis – fourier or wavelet coefficients, …
- Graph data analysis – number of vertices and edges



(C) G. Rohde, Introduction to Image Analysis, ISAC 2008



Feature selection refers to techniques that **choose a subset of the features** (attributes) that is as small as possible and sufficient for data analysis.

A_1, A_2, A_3, A_4

- Remove (more or less) **irrelevant features**
  - For *removing irrelevant features*, a performance measure is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task.
- Remove **redundant features**
  - For *removing redundant features*, either a performance measure for subsets of features or a correlation measure is needed.

PS → ~~Verbrauch~~

Difference between irrelevant and redundant?

Ref.

Feature selection techniques (1/2)

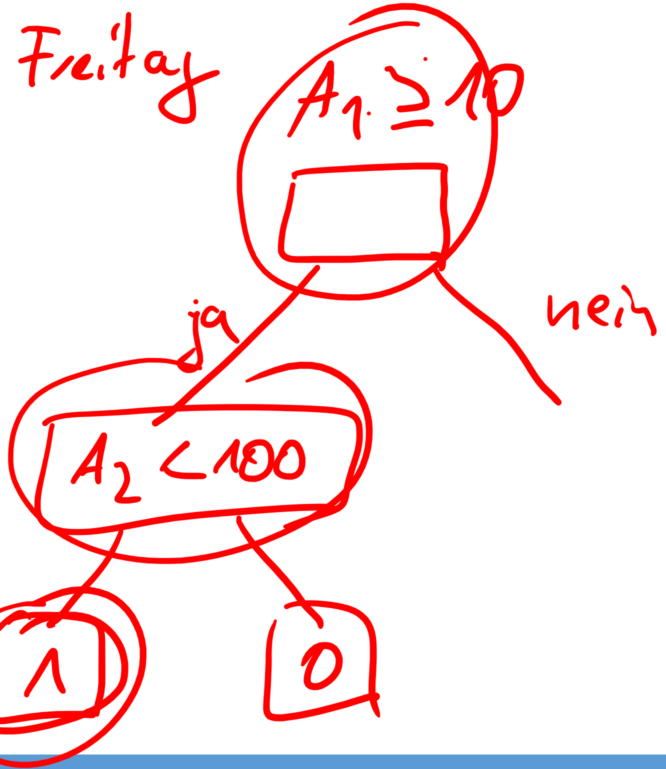
Selecting features with the highest performance

A_1	A_2	A_3	K
			1
			0

For **classification tasks** with a target attribute, typical performance measures are

- χ^2 **test** for independence. It measures the deviation of the sample marginal distributions from the marginal distribution one would obtain assuming the considered attribute and the target variable are independent.
- **Information gain.** Based on entropy reduction (we'll dive deeper into this later with Decision Trees)

→ am Freitag



Wrapper methods can be applied when the model class (e.g. decision trees) is already specified.

- Train the model with different subsets of features and choose the features that lead to the model with the best performance.

Ref.

Feature selection techniques (2/2)

Four ways to select features

- ○ Selecting the top-ranked features (single features)

  Choose the features with **the best evaluation**.

- ○ Selecting the top-ranked subset

  Choose **the subset of features** with the best performance. This requires exhaustive search and is impossible for larger numbers of features.

A_1, A_2, A_3

- ○ Forward selection

  Start with an empty set of features. **Add features one by one.** Consider which one yields the best improvement.

- ○ Backward elimination

  Start with the full set of features and **remove features one by one**. In each step, remove the feature that yields to the least decrease in performance.

A_1, A_2, A_3

A_1, A_2

A_2

Ref.

Feature selection

Example

40 Datensätze
A B C D

A + B

Freie Universität Berlin

Exercise

Consider the following classification task that consists of 9 repetitions of the four data records in the first table and the four records in the second table.

	A	B	C	D	target
9 ×	+	+	+	−	no
	+	−	+	−	yes
	−	+	+	−	yes
	−	−	−	+	no

	A	B	C	D	target
1 ×	+	+	+	−	no
	+	−	+	−	yes
	−	+	+	−	yes
	−	−	+	+	no

Performance of the single attributes

	target	
A	no	yes
+	10	10
−	10	10

	target	
B	no	yes
+	10	10
−	10	10

	target	
C	no	yes
+	11	20
−	9	0

	target	
D	no	yes
+	10	0
−	10	20

Which set of attributes should be selected for classification?

(Diese Folie ist nach der Vorlesung mit Erläuterungen verfügbar

A greedy strategy selecting those attributes with **the best performance** would choose attributes *C* and *D* first.

Attributes *C* and *D* together cannot perfectly predict the target value, though.

Attributes *A* and *B* alone provide no information about the target value.

However, attributes *A* and *B* together are sufficient to perfectly predict the target value (if A=B: no).

Evaluation of the performance of isolated attributes does usually not provide proper information about their performance in combination.

Ref.

Record (Instance) selection

Timeliness

If data have been collected over a long period, some of the **older data might not be useful** or even misleading for the data analysis task. Only the recent data should be selected.

Representativeness

The sample in the database might not be representative for the whole population. When we have information about the distribution of the population, we can **draw a representative subsample** from our database.

Rare events

When we are interested in predicting rare events (e.g. stock market crashes, failures of a production line), it can be helpful to incorporate this in the cost function or to **artificially increase the proportion** of these rare events in the data set.

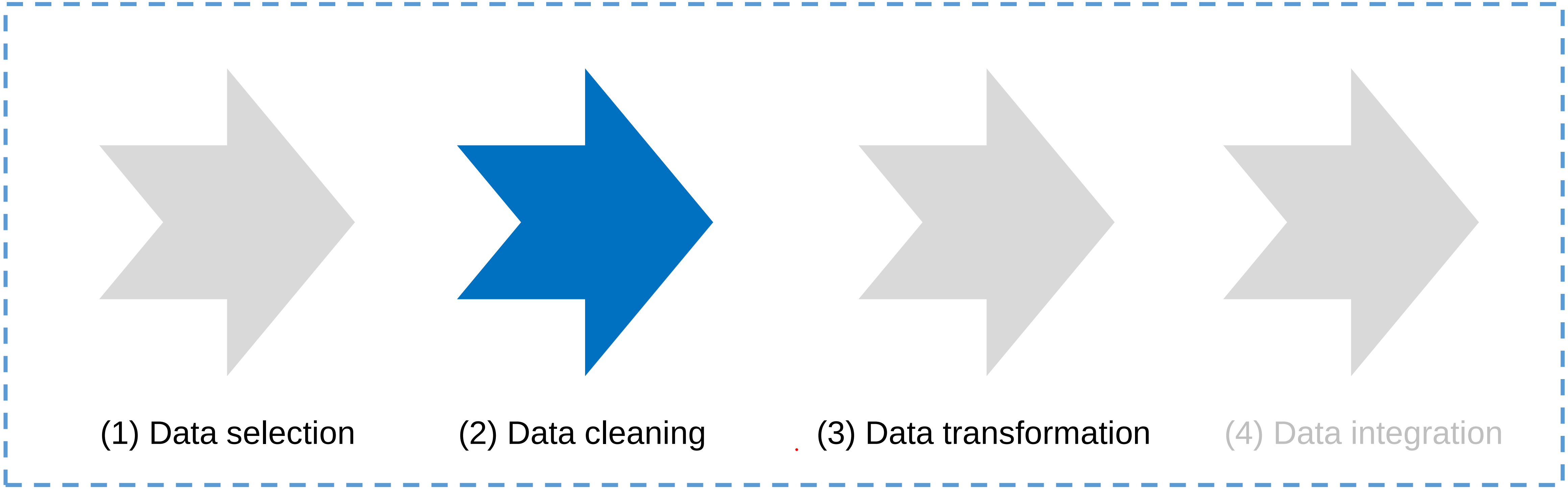
We will address these issues in depth when building predictive models and evaluating their performance.

Ref.

Image: Elliott Brown (2016) | Flickr

Agenda

Data Preparation



(1) Data selection

(2) Data cleaning

(3) Data transformation

(4) Data integration

Ref.

Please read on your own

Data clean(s)ing

Data clean(s)ing or data scrubbing refers to **detecting and correcting or removing inaccurate, incorrect or incomplete data records** from a data set.

Improve data quality

- Turn all characters into capital letters to **level case sensitivity**
- **Remove spaces** and nonprinting characters (\n, \t etc.)
- **Fix the format** of numbers, data and time (decimal point! Datetime objects or standard date format, i.e. YYYY-MM-DD)
- **Split fields** that carry mixed information into separate ones (“Chocolate, 100g” → “Chocolate” and “100.0”)

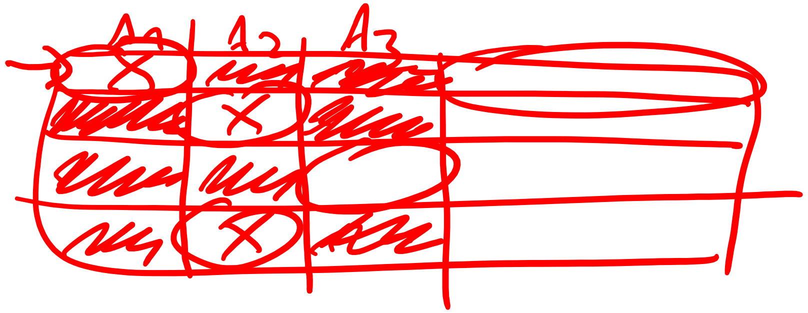
- Use **spell-checker** or stemming to normalize spelling
- **Replace abbreviations** by their long form (dictionary)
- Normalize the **writing of addresses** and names, possibly ignoring the order of title, surname, forename, etc. to ease their re-identification

Eu/ DM

- **Convert numerical values** into standard units, especially if data from different sources and different countries are used
- **Use dictionaries** containing all possible values of an attribute to assure that all values comply with the domain knowledge

Ref.

Missing values

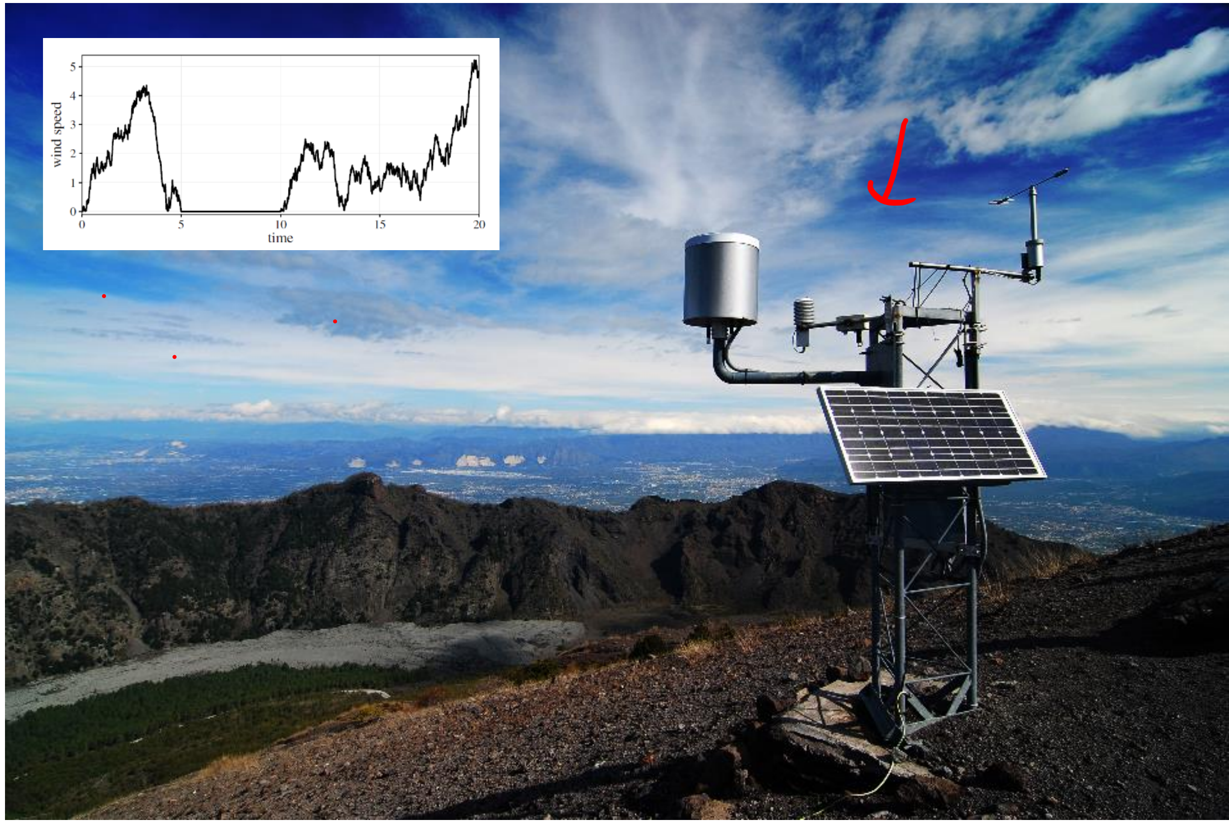


For some instances, values of single attributes might be missing.

Causes for missing values:

- Broken sensors
- Refusal to answer a question (pregnant (yes/no) in a job interview)
- Irrelevant attribute for the corresponding object (pregnant (yes/no) for men)

Missing value might not necessarily be indicated as missing, instead: it may have a default values (i.e., zero, 99 etc.).



Ref. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.

Types of missing values

Consider the attribute X_{obs}. A missing value is denoted by “?”.
X is the true value of the considered attribute, i.e., we have $X_{obs} = X$, if $X_{obs} \neq ?$.

Let Y be the (multivariate)(random) variable denoting the other attributes apart from X.

(1) Missing *completely* at random (MCAR)

The probability that a value for X is missing does neither depend on the true value of X nor on other variables → $P(X_{obs} = ?) = P(X_{obs} = ? \mid X, Y)$

Example:
The maintenance staff ***sometimes*** *forgets to change the batteries of a sensor so that the sensor at that times does not provide any measurements*

MCAR is also called Observed At Random (OAR).

Ref. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.

(2) Missing *at* random (MAR)

The probability that a value for X is missing does not depend on the true value of X
→ $P(X_{obs} = ? \mid Y) = P(X_{obs} = ? \mid X, Y)$

Example:
The maintenance staff does not change the batteries of a sensor ***when it rains****. Thus, the sensor does not always provide measurements when it rains.*

(3) *Nonignorable*

The probability that a value for X is missing depends on the true value of X.

Example:
A sensor for the temperature will not work ***when there is frost****.*

Exercise: Type of missing values?

Further examples

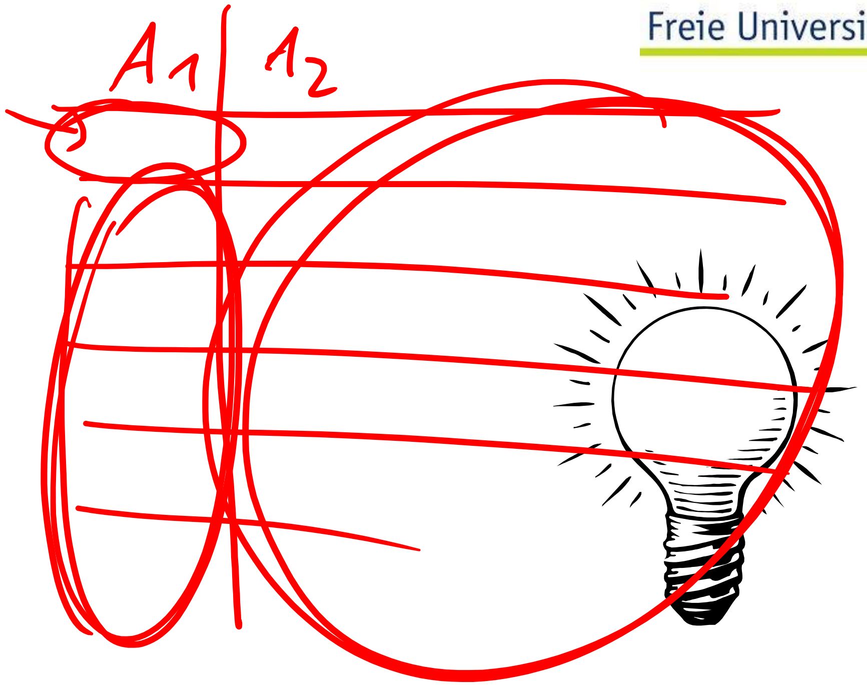
Kahoot-Fragen
www.kahoot.it
(über Smartphone oder Laptop)
PIN folgt

(Diese Folie ist nach der Vorlesung mit Lösungen verfügbar

Ref.

Types of missing values

Estimation of Missing Values



For MCAR and MAR, the missing values can be **estimated**

- At least in principle, when the data set is large enough – based on the values of the other attributes
- The cause for the missing values is ignorable

(1) For MCAR, it can be assumed that the missing values follow **the same distribution** as the observed values of X

(2) For MAR, the missing values might not follow the distribution of X. But by taking the other attributes into account, it is possible to derive **reasonable imputations** for the missing values.

(3) For *nonignorable* missing values it is **impossible to provide sensible estimations** for the missing values

Ref. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.

How to determine the type of missing values

1 ✓
0 ✓

$\otimes \to 1, 0$ | y_1, y_2, y_3 | y_3

0% -> 100% $\frac{4}{8}$

≥ 10%

If *domain knowledge* does not help which kind of missing values can be expected, the following strategy can be applied

- Turn the considered attribute X into a **binary attribute**, replacing all measured values by “yes” and all missing values by “no”
- **Build a classifier** with the now binary attribute X as the target attribute, and use all other attributes for the prediction of the class values “yes” and “no”
- **Determine the misclassification rate**, which is the proportion of data objects that are not assigned to the correct class by the classifier.

MAR

- For MCAR, the other attributes should not provide any information, whether X has a missing value or not. Therefore, the misclassification rate should not differ significantly from **pure guessing**
- If there are 10% missing values for the attribute X, the misclassification rate of the classifier should not be much smaller than 10%.
- **If the misclassification rate is significantly better** than pure guessing, this is an indicator that there is a correlation between missing values for X and the values of the other attributes. The missing values are not MCAR.

- MAR and nonignorable cannot be distinguished in this way.

Ref.

How to handle missing values

Ignorance/Deletion

If only a few records have missing values, and it can be assumed that the values are MCAR, these records can be deleted for the following data analysis step.

Imputation

The missing values may be replaced by some estimate.

Single Imputation
Mean, median or mode of the attribute (MCAR required)

Multiple Imputation
By an estimation based on the other attribute,
e.g., max-likelihood estimation, bayesian procedure …
(MAR required!)

Further reading

- How to Handle Missing Data, https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
- Flexible Imputation of Missing Data https://stefvanbuuren.name/fimd/

Explicit value

Missing values are characterized by a specific value („MISSING“). The selected model must be able to hand these specified missing values (most models assume MCAR)!

Ref. e.g., Efromovich, S. (2014).

Outlier detection - Single attributes

Please read on your own

An **outlier** is a value or data object that is far away or very different from all or most of the other data

Categorical attributes: an outlier is a value that occurs with an extremely lower frequency than the frequency of all other values

In some cases, the outliers can even be the target objects of the analysis

- Example: automatic quality control system
- Goal: train a classifier, classifying the parts as correct or with failures based on measurements of the produced parts
- The frequency of the correct parts will be so high that the parts with failure might be considered as outliers

Numerical attributes: outliers can be identified in boxplots or by statistical tests.

Problems: asymmetric distribution, large data sets

Statistical test that a sample following a normal distribution does not contain outliers (Grubb's test):

Define the statistic $G = \frac{max\{|x_i - \bar{x}|| 1 \leq i \leq n\}}{s}$ where $x_i, \dots, x_n$ is the sample, $\bar{x}$ its mean value and s its empirical standard deviation.

For a given significance level α, the null hypothesis that the sample coming from a normal distribution does not contain outliers is rejected if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{1-\frac{\alpha}{2n}, n-2}}{n-2+t^2_{1-\frac{\alpha}{2n}, n-2}}}$$

where $t^2_{1-\frac{\alpha}{2n}, n-2}$ denotes the $(1 - \frac{\alpha}{2n})$-quantile of

the t-distribution with $(n-2)$ degrees of freedom.

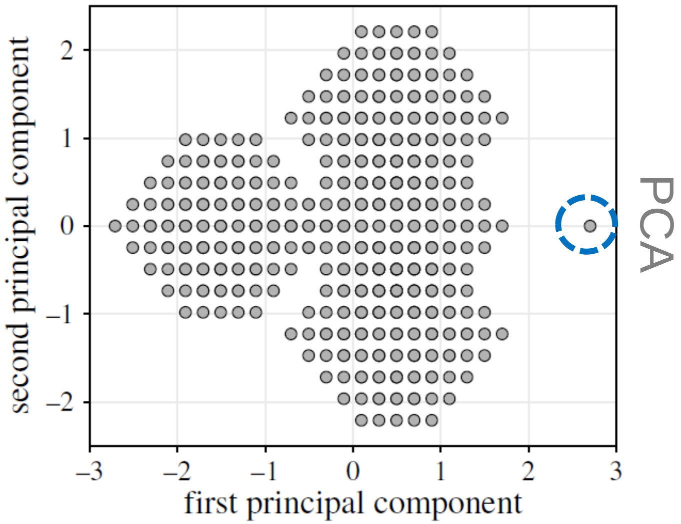
Ref.

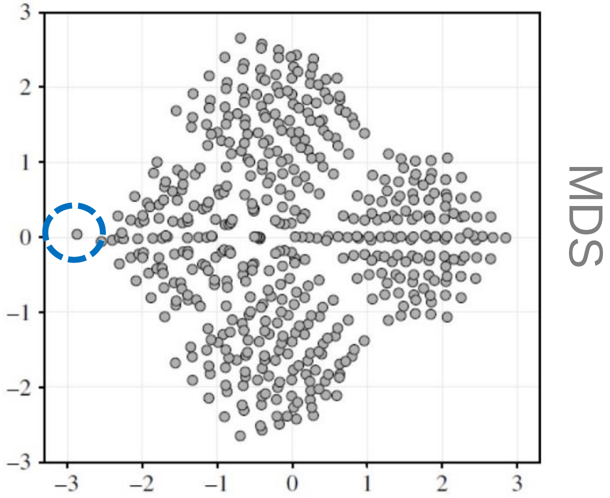
Outlier detection - Multidimensional data

Scatter plots for (visually detecting) outliers w.r.t. two attributes.

PCA (or multi-dimensional scaling) for (visually) detecting outliers.

Cluster analysis techniques:
outliers are those points which cannot be assigned to any cluster/which are far away from other clusters.

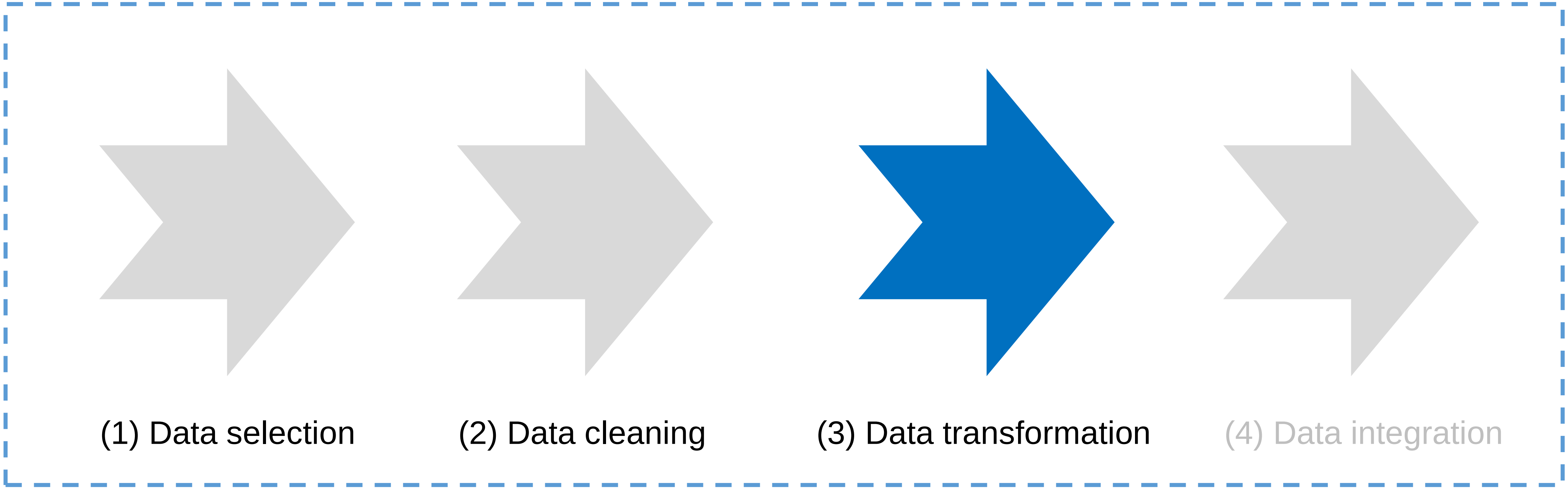




Ref.

Agenda

Data Preparation


(1) Data selection
(2) Data cleaning
(3) Data transformation
(4) Data integration

Ref.

Data transformation

Categorical -> Numerical attributes

Some models can only handle numerical attributes, other models only categorical attributes.

In such cases, categorical attributes must be transformed into numerical ones or vice versa.

Categorical attribute -> Numerical attribute:

A binary attribute can be turned into a numerical attribute with the values 0 and 1 (aka dummy variable)

A categorical attribute with more than two values, say $a_1, \ldots, a_k$, **should not be turned into a single numerical attribute** with the values $1, \ldots, k$, unless the attribute is an ordinal attribute. It should be turned into k attributes $A_1, \ldots, A_k$ with values 0 and 1 (dummies). a_1 is represented by $A_i = 1$ and $A_j = 0$ for $i \neq j$.

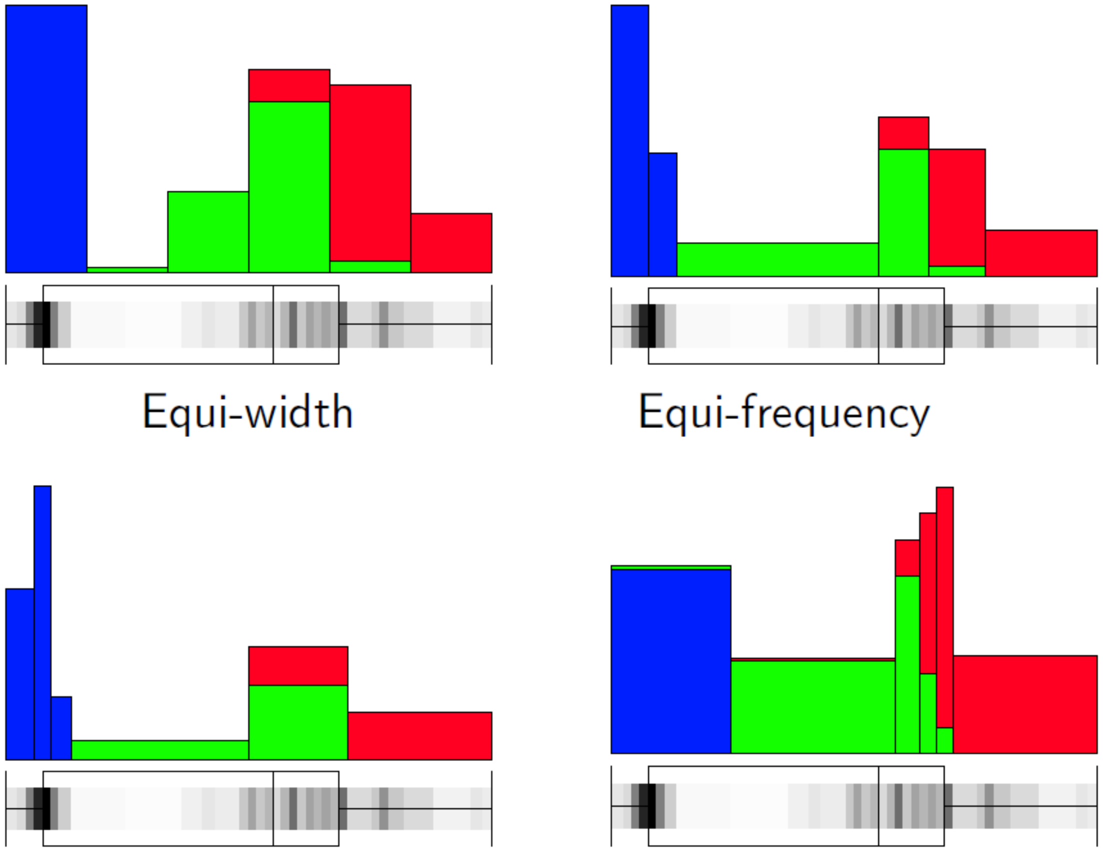
Ref.

Data transformation

Discretization: Numerical -> Categorical attributes

Discretization techniques refer to splitting a numerical range into a number of finite bins.

- **Equi-width discretization.** Splits the range into intervals (bins) of the *same width*.
- **Equi-frequency discretization.** Splits the range into intervals such that each interval (bin) contains (roughly) the *same number of records*.
- **V-optimal discretization.** Minimizes $\sum_i n_i V_i$ where n_i is the *number of data objects* in the ith interval and V_i is the sample *variance* of the data in this interval.
- **Minimal entropy discretization.** Minimizes the *entropy*. (Only applicable in the case of classification problems, we'll dive deeper into this with decision trees)



Equi-width

Equi-frequency

V-optimal

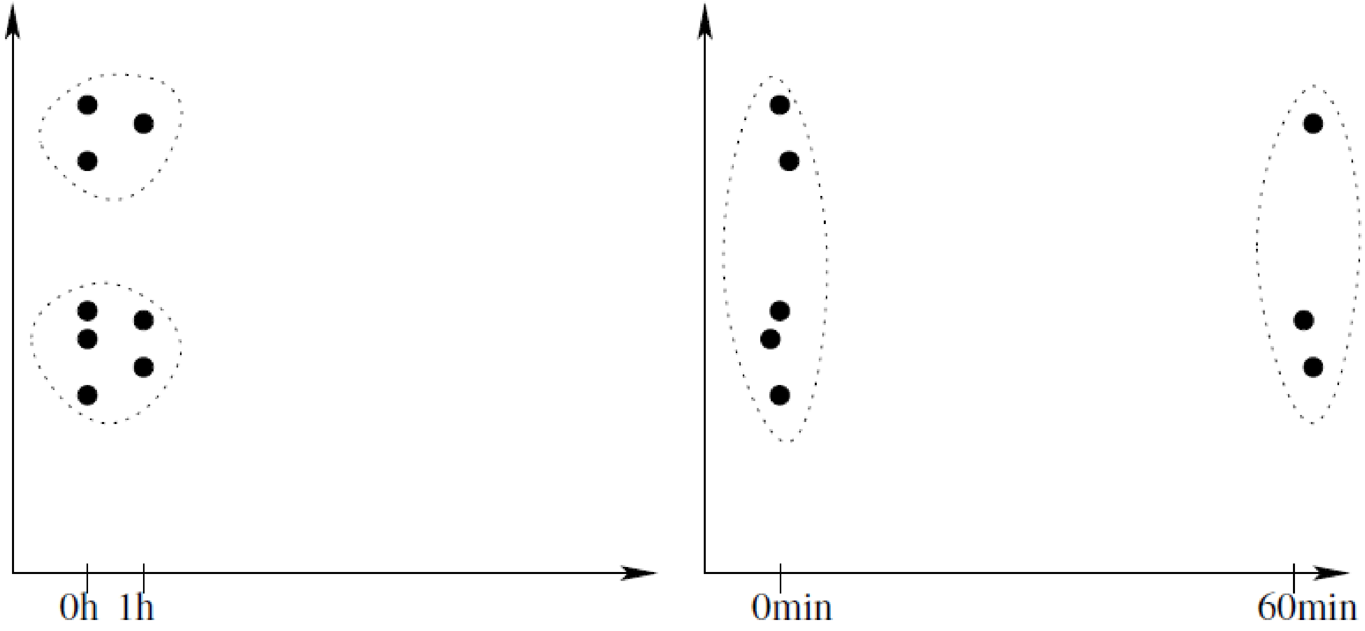
Minimal entropy

Ref.

Normalization | Standardization (1/2)

For some data analysis techniques (PCA, MDS, cluster analysis) the influence of an attribute depends on the **scale** or measurement unit.

To guarantee impartiality, some kind of standardization or normalization should be applied.



Ref.

Normalization | Standardization (2/2)

Min-max normalization:

For a numerical attribute X with min_x and max_x being the minimum and maximum value in the sample, the min-max normalization is defined as

$$n: domX \rightarrow [0,1], \qquad x \rightarrow \frac{x - min_X}{max_X - min_X}$$

Z-score standardization:

For a numerical attribute X with sample mean $\hat{\mu}_X$ and empirical standard deviation $\hat{\sigma}_X$, the z-score standardization is defined as

$$s: domX \rightarrow \mathbb{R}, \qquad x \rightarrow \frac{x - \overline{\hat{\mu}_X}}{\hat{\sigma}_X}$$

Robust z-score standardization:

The sample mean and empirical standard deviation are easily affected by outliers. A more robust alternative is (see also boxplots):

$$s: domX \rightarrow \mathbb{R}, \qquad x \rightarrow \frac{x - \bar{x}}{IQR_X}$$

Decimal scaling:

For a numerical attribute X and the smallest integer value s that is larger than $log_{10}(max_X)$, the decimal scaling is defined as

$$d: domX \rightarrow [0,1], \qquad x \rightarrow \frac{x}{10^s}$$

Ref.

Fragen?

- ✓ Data preparation

- ✓ Data selection
- ✓ Data cleaning
- ✓ Data transformation
- ○ Data integration

Recommended reading

Data Preparation

Berthold et al. Chapter 4, 6

Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011

Ref.

Bibliography

- J. Bertin (1983) *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin Press. Originally in French: *Semiologie Graphique*, 1967
- Cairo, A. (2012). *The Functional Art: An introduction to information graphics and visualization*. New Riders.
- Mertens, P., & Meier, M. (2009). *Integrierte Informationsverarbeitung*. Wiesbaden: Gabler.
- Woolman, M. (2002). *Digital information graphics*. Watson-Guptill Publications, Inc..

Ref.