Technology Arts Sciences
TH Köln

About Data, Preparation, PCA

Unknown Image Source

Technology Arts Sciences TH Köln

Data Mining (praktisch) SoSe 2025

Datum	Thema
01.04.25	Einführung ins Modul und Thema
08.04.25 (flex)	Python Einführung via Data Camp (Zeitlich frei einteilbarer Onlinekurs)
15.04.25	V&Ü Business Understanding und wirtschaftliche Grundlagen
22.04.25	Online Sprechstunde
29.04.25	Online Einzelabnahme Business Understanding Meilenstein
06.05.25	V&Ü Data Understanding and Visualization
13./14.05.25	Probelehrveranstaltungen für Statistische Methoden in der KI – DataCamp
20.05.25	Online Einzelabnahme Data Understanding Meilenstein
27.05.25	V&Ü Data Distributions and Transformations
03.06.25	Online Gruppenabnahme Data Preparation Meilenstein
10.06.25	V&Ü Clustering algorithms and evaluation
17.06.25	Online Sprechstunde
24.06.25	Online Gruppenabnahme Modeling and Evalation Meilenstein
01.07.25	Online Sprechstunde
08.07.25	Finale Projekt-Präsentationen (alle Gruppen durchgängig anwesend)

Goals today

- Learn some data preprocessing / preparation

- Get to know dimension reduction methods

RECAP OF LAST WEEK

…to remember the most important concepts / ideas

Other Categorization: Discrete & Continuous Attributes

- Discrete Attribute
  - Has only a finite or countable infinite set of values
    - Examples: zip codes, counts or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
    - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

1. Excercise about different types of attribute values

Which operations are allowed ?

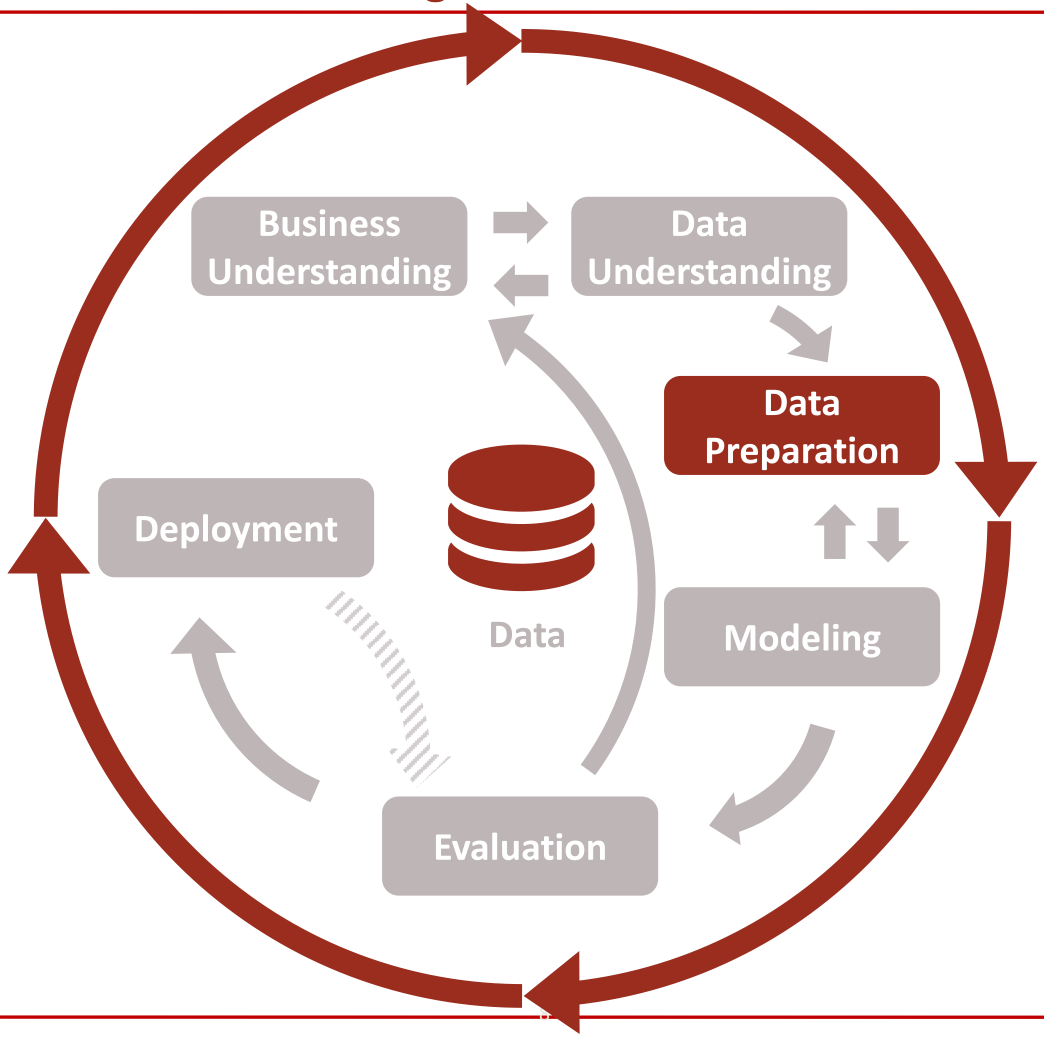
Attribut Type	Mathem. Operation (two values)	Aggregation (many values)
nominal	= ≠	Mode, count
ordinal	all from nominal & <>	Median, count
interval	all from ordinal & +, -	Mean, sum
ratio	all from intervall & *, /	Mean, sum, division

DATA PREPARATION

…how to get more value to the data

CRISP-DM

Cross-Industry Standard Process for Data Mining



Why Data Preprocessing is Crucial

- The problem: all data analysis approaches are dumb. They rely on statistics, but
  - they do not understand data
  - they can not be creative
- Data pre-processing is the step where the data scientist tells the algorithm what it needs to know about the data

Coding of the Application Domain – Mapping of Real World Objects

- Real world objects / real world context
- **Instance**, example, sample, object
- Coding is composed of **attributes** (features, characteristics)
- Different possible types of attributes
  - Numerical (age: 10, 50, 100)
  - Ordinal - ordered categories (weight: underweight, normal weight, overweight, very overweight)
  - Nominal - categories (profession: computer scientist, analyst, teacher, …)

An Explanatory Example

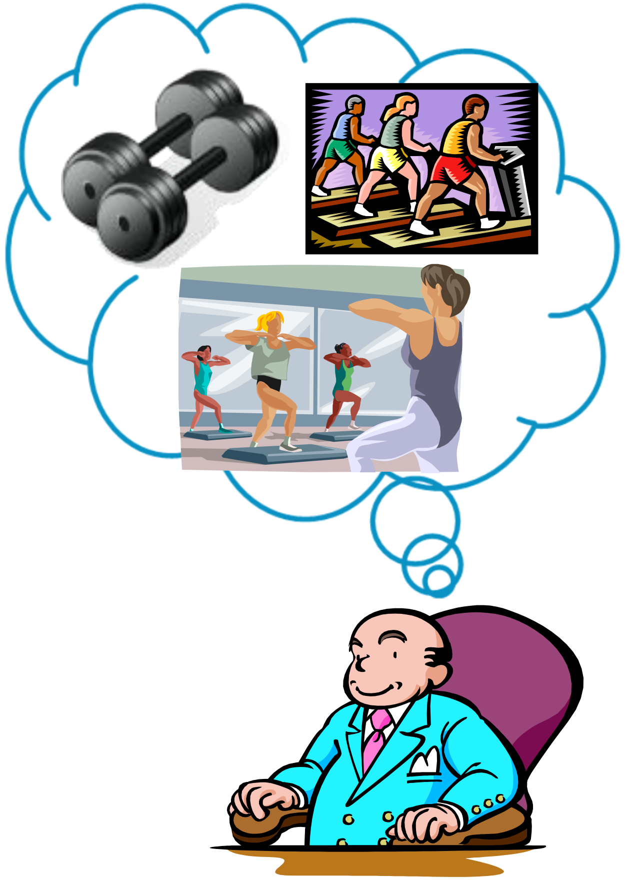
Manager of gym chain

Goals:

- New members
- Less contract cancellations

The road to success:

- New individual pricing structure
- More targeted advertising and offers



How to Code the Members?

ID - key

attribute

effect variable/ label

Effect variable (label) for advertising in example: stated goal of fitness

ID	gender	age	weight	running contract?	contract period (until today)	stated goal of fitness
1	F	51	normal	yes	5	general fitness
7	F	16	over	no	1	loss of weight
D1	M	26	normal	yes	0.5	muscle growth
19	M	60	under	yes	4	muscle growth
93	M	40	very over	yes	8	loss of weight
46	F	37	normal	no	6	general fitness
56	M	50	over	yes	12	?
4	F	19	under	yes	3	?
58	F	29	normal	yes	0.2	?

labelled data

unlabelled data

A few more words about attributes

Naming attributes

Descriptive attributes

Referential attributes

ID	gender	age	weight	running contract?	contract period (until today)	stated goal of fitness	Studio ID
1	F	51	normal	yes	5	general fitness	1
7	F	16	over	no	1	loss of weight	2
11	M	26	normal	yes	0.5	muscle growth	2
19	M	60	under	yes	4	muscle growth	1
93	M	40	very over	yes	8	loss of weight	3
46	F	37	normal	no	6	general fitness	3

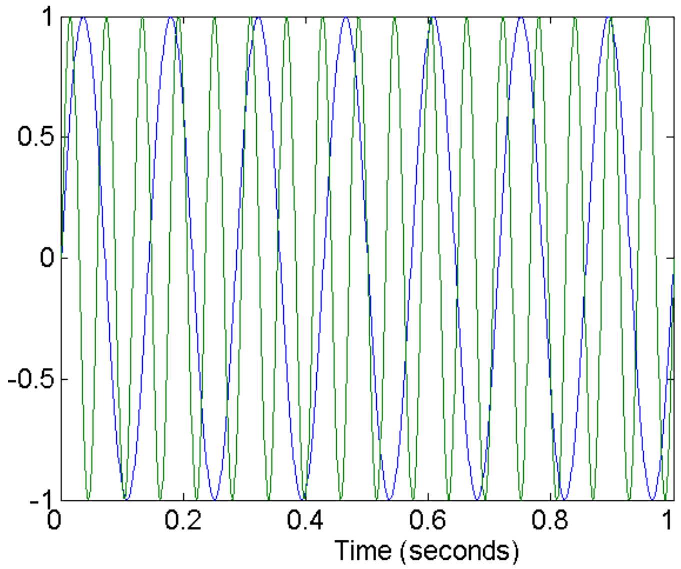
Derived attributes

Data Quality

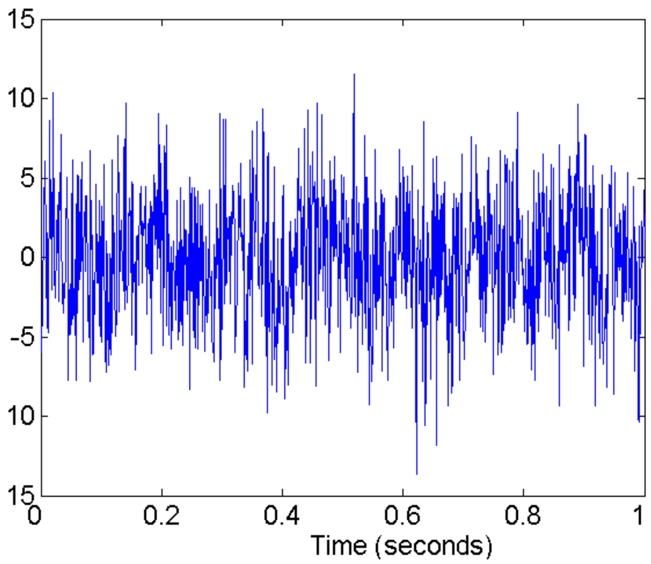
- What kinds of data quality problems exist?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



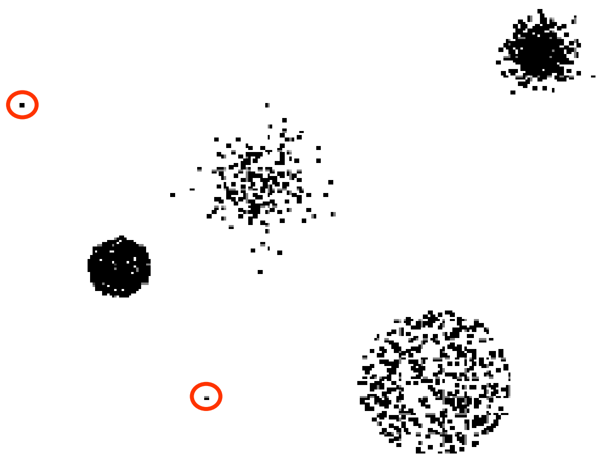
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Technology
Arts Sciences
TH Köln

Missing Values

- Reasons for missing values
  - Information is not collected
    - e.g., people decline to give their age and weight
  - Attributes may not be applicable to all cases
    - e.g., annual income is not applicable to children
- Handling missing values
  - Eliminate Data Objects
  - Estimate missing values
  - Ignore the missing value during analysis
  - Replace with all possible values
    - weighted by their probabilities

How do I Cope with Missing Data?

- Delete instance
- Delete attribute
- Insert standard value
- „Mean“value
- Striking value
- Use suitable model-building

– Frequent problems

- Sometimes a missing value has been replaced with a hard-coded value („9999“).
- But sometimes a missing value also has a meaning

ID	gender	age	weight	running contract?	contract period (until today)	stated goal of fitness
1	F	51	normal	yes	5	general fitness
7	F	16	?	no	1	loss of weight
11	F	26	normal	yes	0.5	muscle growth

Technology Arts Sciences TH Köln

Duplicate Data

- Data set may include data objects that are duplicates or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
  - Sea Thomas Krause, Entwurf und Implementierung einer effizienten Dublettenerkennung für große Adressbestände in http://epb.bibl.fh-koeln.de/frontdoor/index/index/docId/270

Derived Features

- The algorithm does not know about attribute dependencies
  - Construct quotients, differences, etc. of attributes
- The algorithm does not know about meaning of data
  - Date → weekday, holiday, …
- Also: joining with external data, e.g. ZIP code → population, location & day → weather, …
- Construct meaningful attributes by hand
- Also remove unnessary attributes, e.g. identifiers

Note: different philosophy between database as archive and analytical database!

Special case: Time-dependent Data

- Many data mining algorithms work only with spreadsheet data
- Connections between rows – as with time-dependent data – are usually not made automatically
- Construct meaningful attributes
  - Time since last visit
  - Number of visits in last month
  - Average number of visits per week
  - Weight difference since last visit
  - …

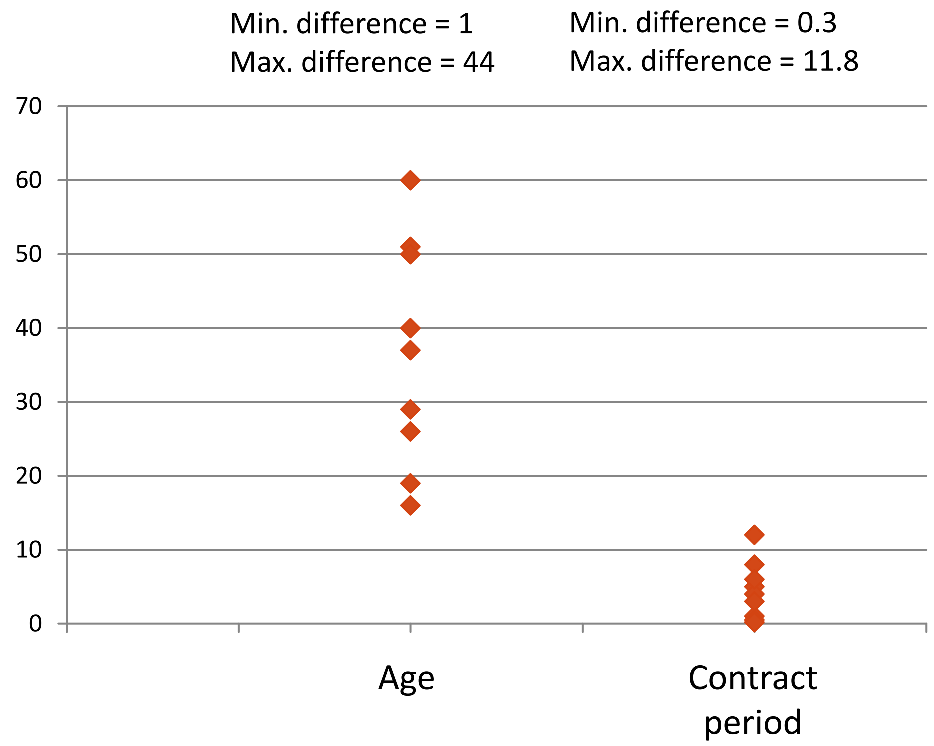
Technology Arts Sciences TH Köln

Standardization / Normalization

Numerical attributes may be in different intervals

→ varying influence on model

age	contract period (until today)
51	5
16	1
26	0.5
60	4
40	8
37	6
50	12
19	3
29	0.2



Standardization / Normalization

- Numerical attributes maybe in different intervals → varying influence on model
- Harmonization by standardization
- Z-Transformation (mean value to 0, standard deviation to 1)

$$\tilde{a}_i = \frac{a_i - \mu_a}{\sigma_a}$$

- Linear normalization (values between 0 (-1) and 1)

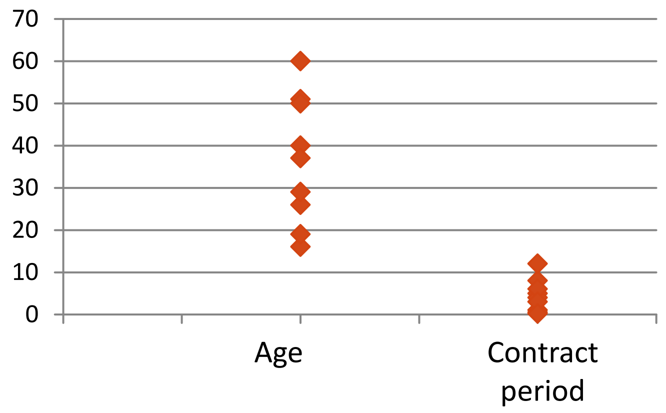
$$\tilde{a}_i = \frac{a_i - \min(a)}{\max(a) - \min(a)}$$

Gym Example – Z-Transformation

age	contract period (until today)
51	5
16	1
26	0.5
60	4
40	8
37	6
50	12
19	3
29	0.2

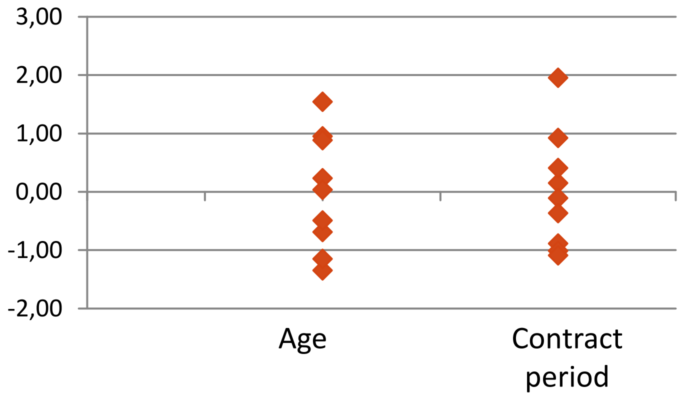
$$\tilde{a}_i = \frac{a_i - \mu_a}{\sigma_a}$$

age	contract period (until today)
0.96	0.15
-1.34	-0.88
-0.69	-1.01
1.55	-0.11
0.23	0.93
0.04	0.41
0.89	1.96
-1.15	-0.36
-0.49	-1.09



Min. difference = 0.07
Max. difference = 2.89

Min. difference = 0.08
Max. difference = 3.04

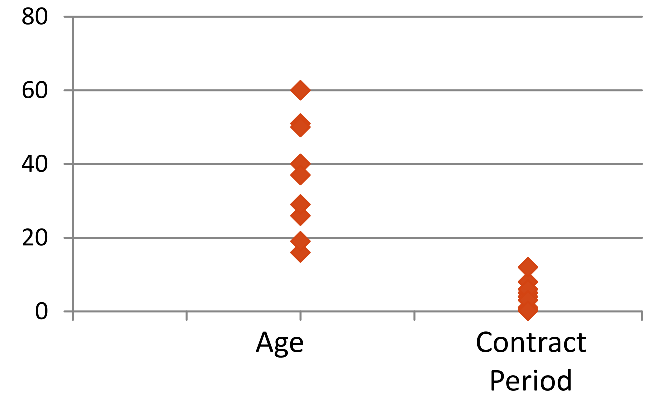


Gym Example – Linear Standardization

age	contract period (until today)
51	5
16	1
26	0.5
60	4
40	8
37	6
50	12
19	3
29	0.2

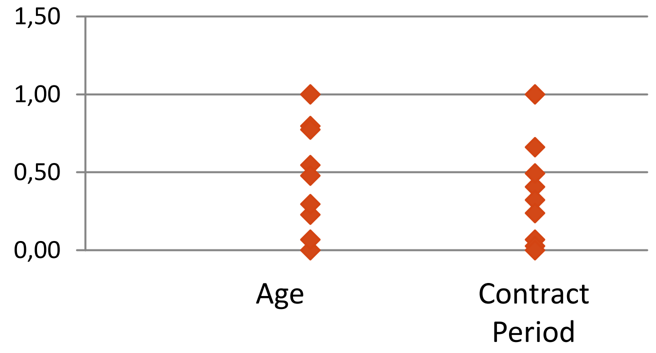
$$\tilde{a}_i = \frac{a_i - \min(a)}{\max(a) - \min(a)}$$

age	contract period (until today)
0.80	0.41
0.00	0.07
0.23	0.03
1.00	0.32
0.55	0.66
0.48	0.49
0.77	1.00
0.07	0.24
0.30	0.00



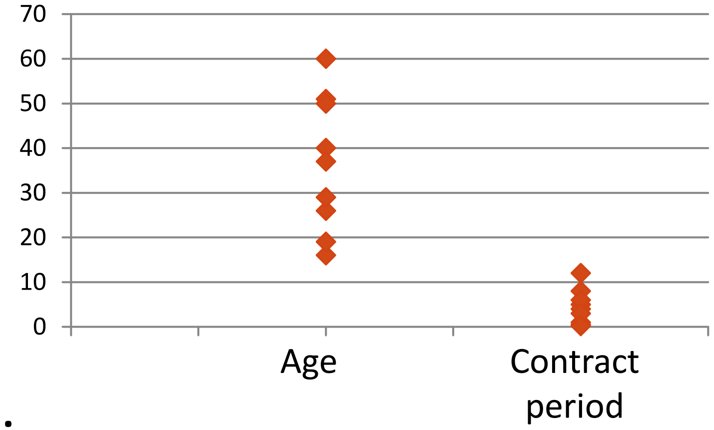
Min. difference = 0.02
Max. difference = 1.00

Min. difference = 0.03
Max. difference = 1.00

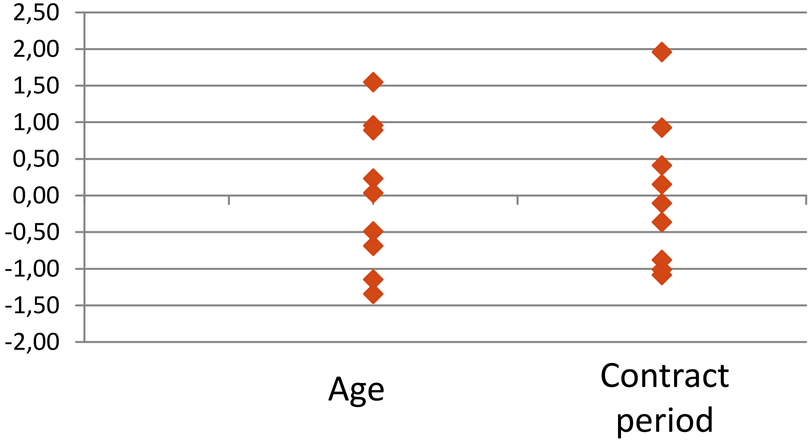


Technology Arts Sciences TH Köln

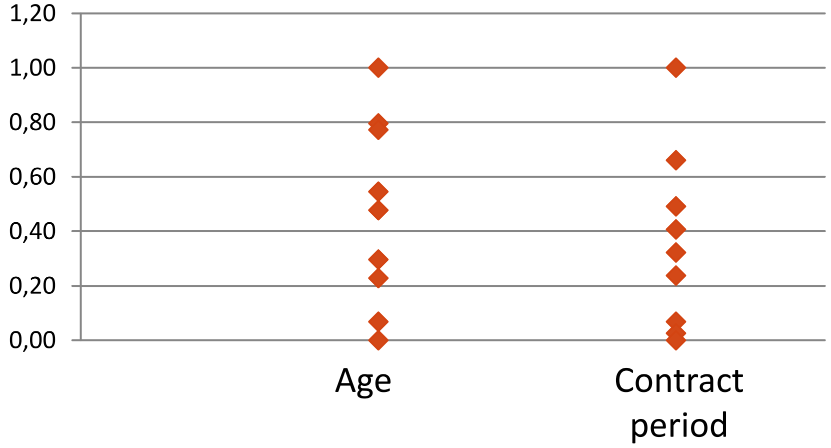
Gym Example – Standardization Comparisons



Z-Transformation



Linear Standardization



Technology
Arts Sciences
TH Köln

Most algorithms assume data to be normal-distributed

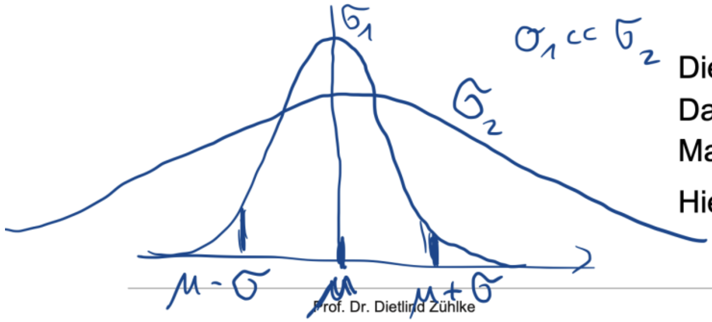
- Normal distribution – recap from mathematics 2

Normalverteilung = Gaußverteilung

Eine stetige Zufallsvariable $X\colon \Omega\colon \mathbb{R}$ heißt normalverteilt mit Mittelwert μ und der Standardabweichung σ oder kurz $N(\mu,\sigma)$-verteilt, wenn ihre Dichtefunktion lautet

$$\omega(t) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

Die Normalverteilung hat die typische Form der Gauß'schen Glockenkurve.

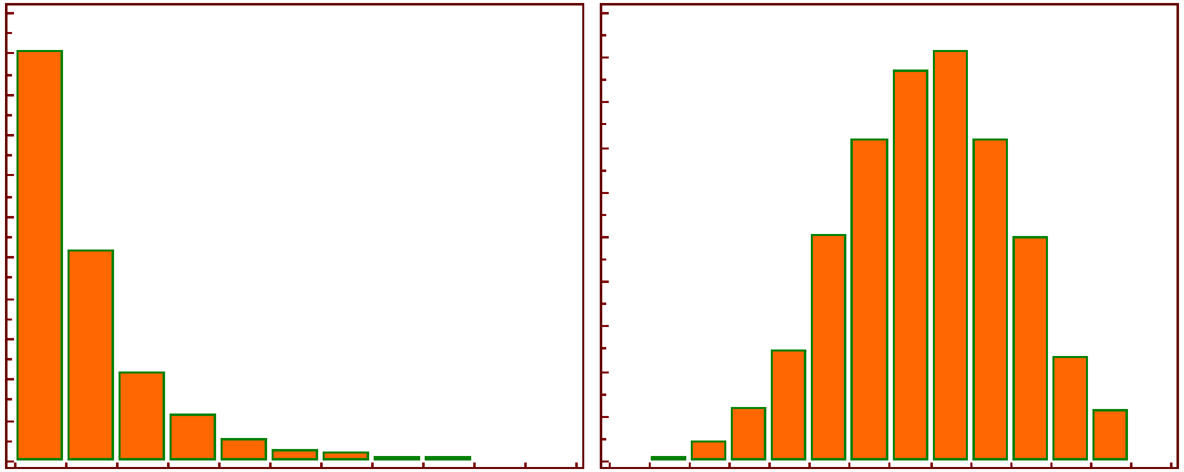


Die Parameter μ und σ lassen sich unmittelbar aus der grafischen Darstellung der Dichtefunktion ablesen: Die Gaußglocke hat ihr Maximum bei $t = \mu$ und ihr Wendepunkt bei $\mu - \sigma$ und $\mu + \sigma$.

Hier liegen 68,2% der Daten im Intervall $[\mu - \sigma, \mu + \sigma]$.

Log-transformation to help with skewed data

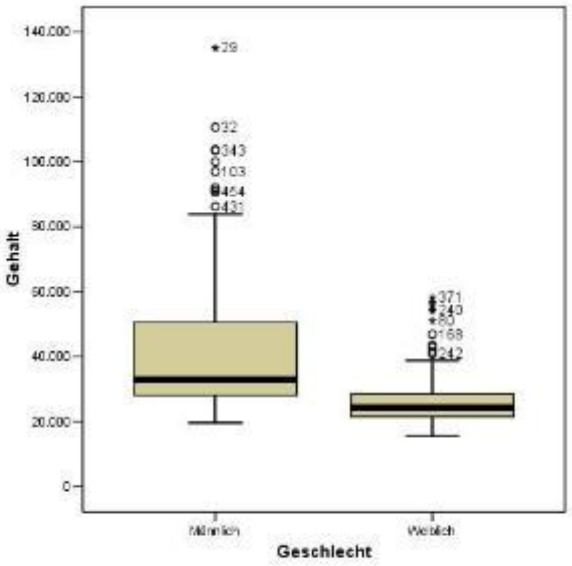
- Skewed data means a lot of small values and a long „tail“ of larger ones
- Log-transformation makes data more similar to normal-distributed data
  - Original number = x
  - Transformed number $x'=\log_{10}(x)$
  - Back-transformed number = $10^{x'}$
- For zeros or negative numbers log is not defined
  - Add a constant to each number to make them positive and non-zero.



https://www.medcalc.org/manual/log-transformation.php

Technology Arts Sciences TH Köln

Exploratory Data Analysis

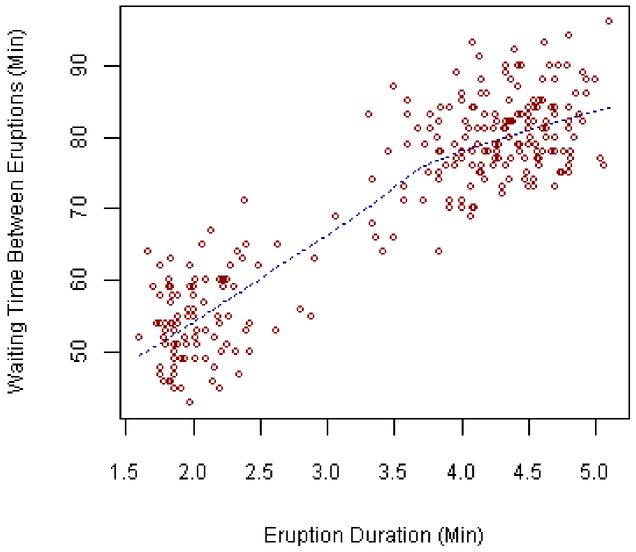


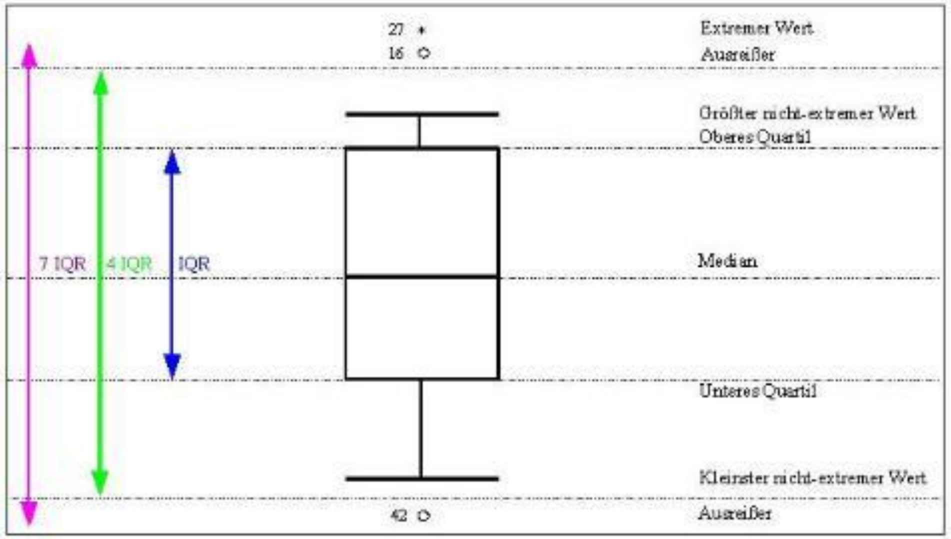
Boxplots

- Categorical target attributes

Scatterplots

- Numerical target attributes
- Suspected correlation





Technology Arts Sciences TH Köln

Re-Use of Pre-Processing
Making your life easier in the long run

- Data preprocessing is a manual, time-consuming process, but it can often be reused for new projects
- Report quality problems!
- If possible, turn off the root cause
- Add pre-processing to data warehouse, e.g. derived features
- Document insights

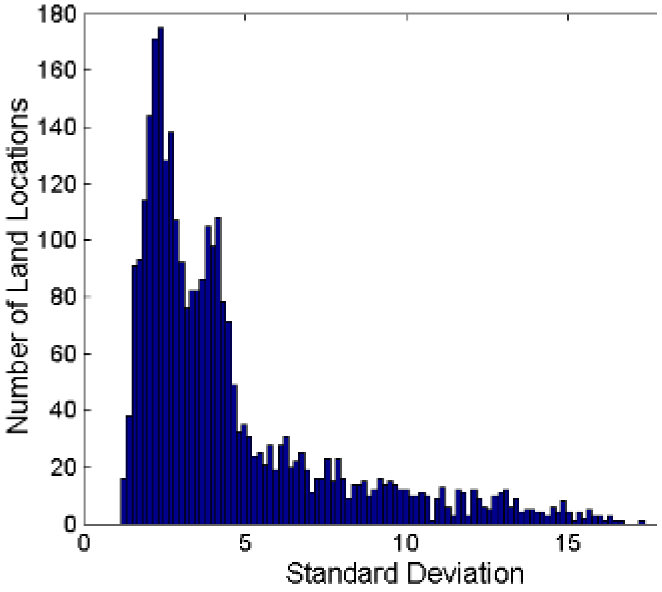
Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

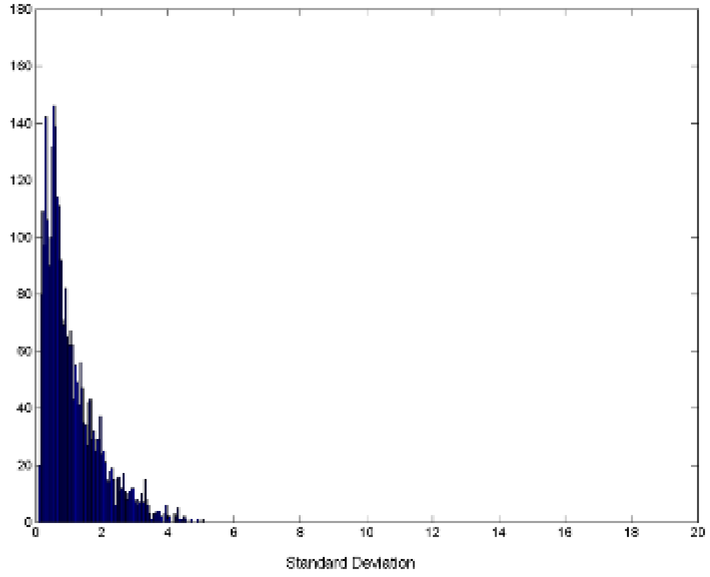
SQL:
Group By

Aggregation

Variation of precipitation (of rain) in Australia



Standard Deviation of Average Monthly Precipitation



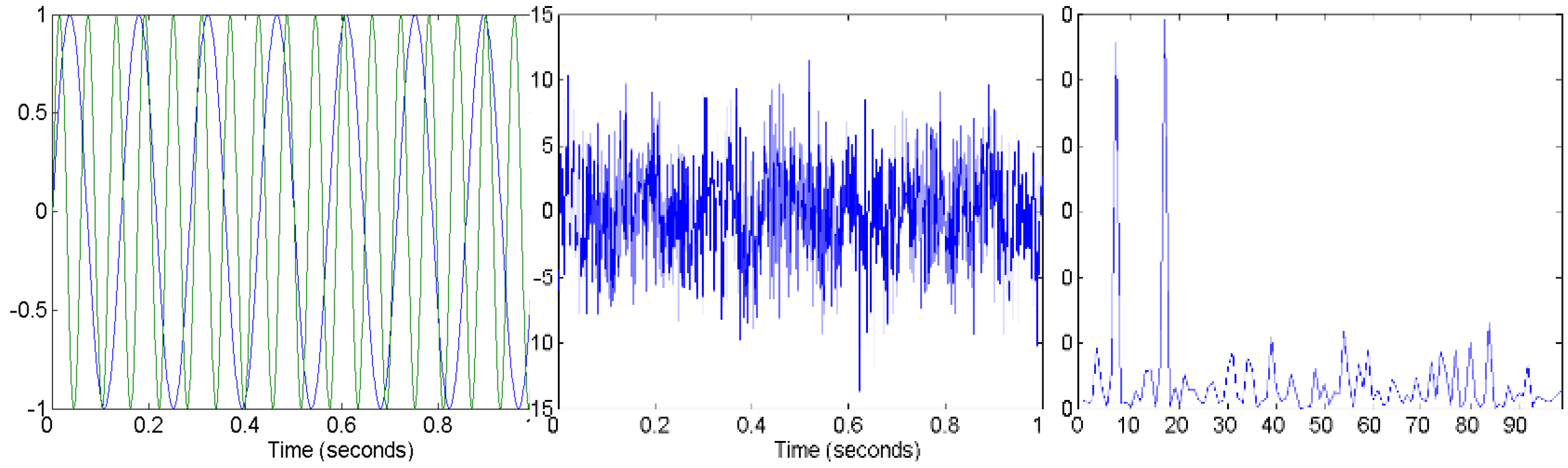
Standard Deviation of Average Yearly Precipitation

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Mapping Data to a New Space

- Fourier transform



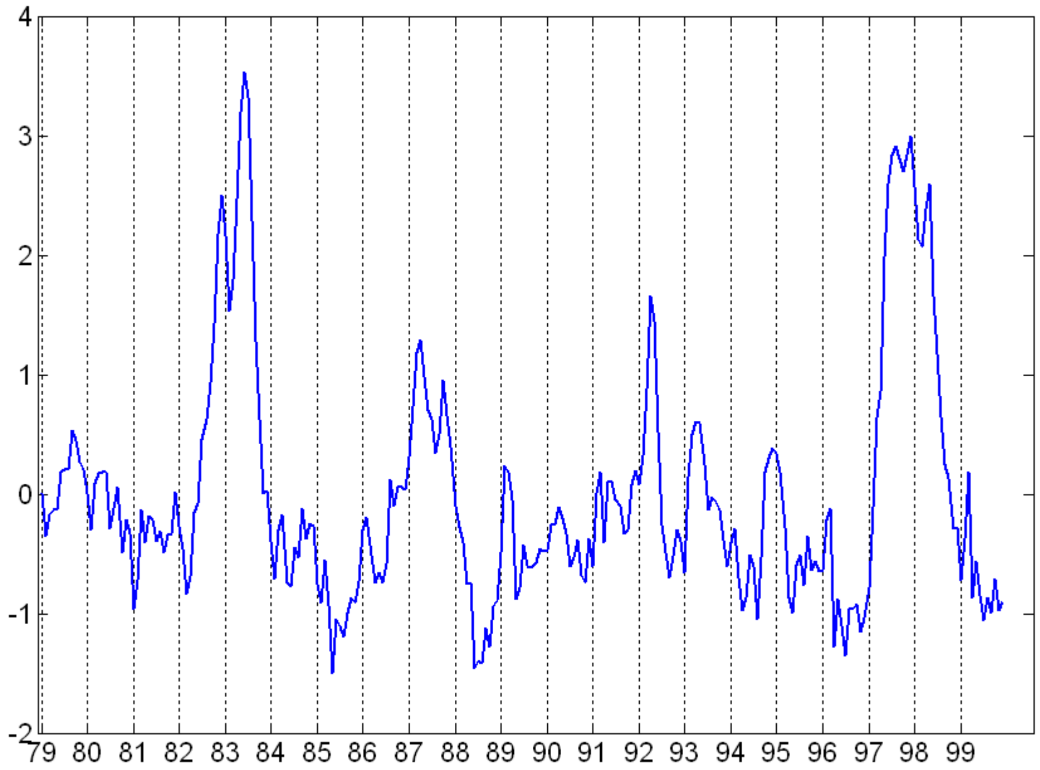
Two Sine Waves

Two Sine Waves + Noise

Frequency

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$
  - Standardization and Normalization



Why dimensionality reduction?

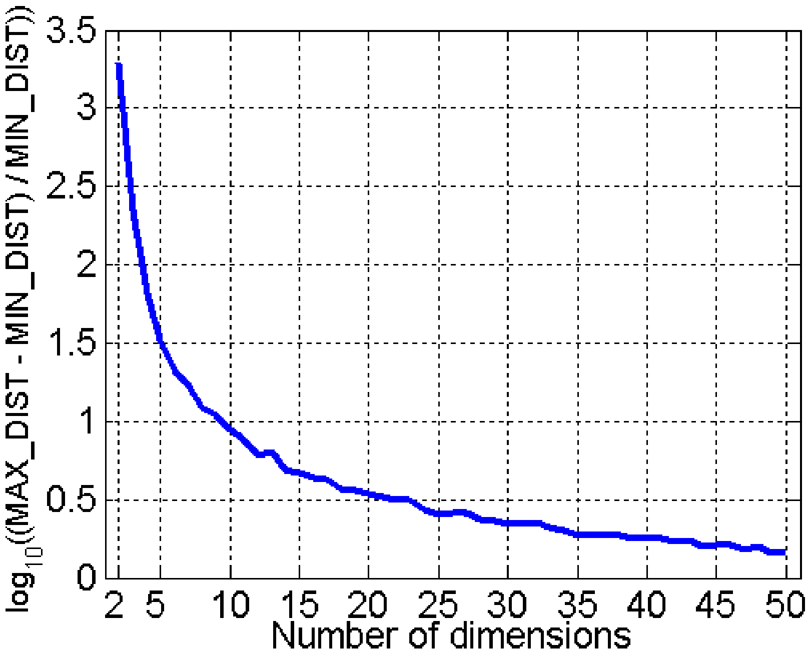
- Some features may be irrelevant
- We want to visualize high dimensional data
- “Intrinsic” dimensionality may be smaller than the number of features
- Applications
  - Digital image and speech processing
  - Gene expression
  - Visualization of large networks

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

- Term "Curse of Dimensionality" was introduced by Richard E.Bellmann,
  - https://en.wikipedia.org/wiki/Curse_of_dimensionality



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principle Component Analysis
  - t-SNE
  - UMAP
  - Others: e.g. supervised

Feature Subset Selection

- Another way to reduce dimensionality of data
  - Redundant features
    - duplicate much or all of the information contained in one or more other attributes
    - Example: purchase price of a product and the amount of sales tax paid
  - Irrelevant features
    - contain no information that is useful for the data mining task at hand
    - Example: students-ID is often irrelevant to the task of predicting students-GPA (Grade Point Average)

Feature Subset Selection

- Techniques:
  - Brute-force approach:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm runs
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

Unsupervised feature selection – dimensionality reduction

- Idea:
  - Given data points in n-dimensional space,
  - Project into lower dimensional space while preserving as much information as possible
  - In particular, choose projection that minimizes the squared error in reconstructing original data

PCA

…how to boil it down to lower dimensionality

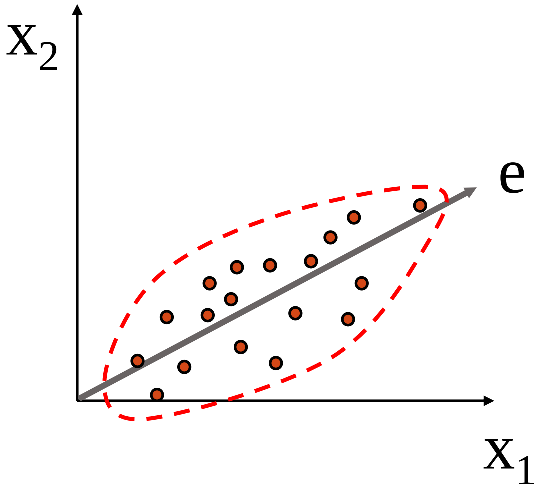
Principle Component Analysis

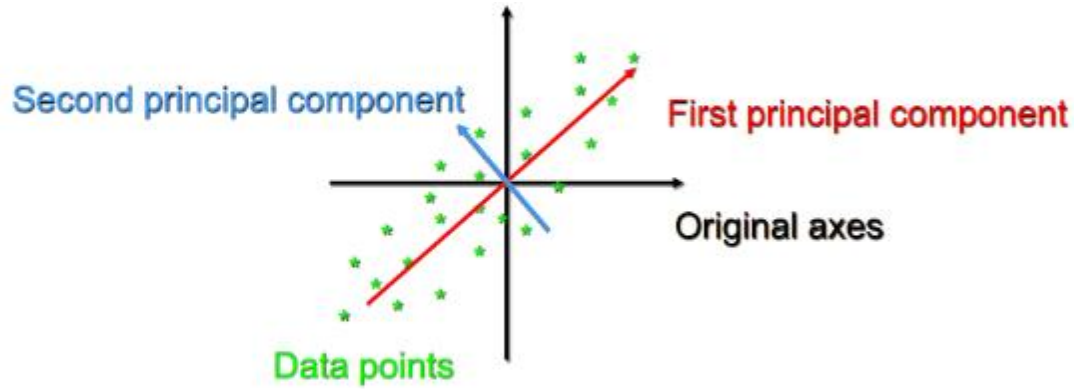
- Type: **Linear method**
- Aim: Projection of the data onto orthogonal axes (principal components) that explain maximum variance
- Preservation: Preserves global structures (variance)
- Computational effort: Very efficient, can handle large data sets and high dimensionality
- Deterministic: Always the same result (if no random component)
- Interpretability: Each component is a linear combination of the original characteristics → easy to interpret

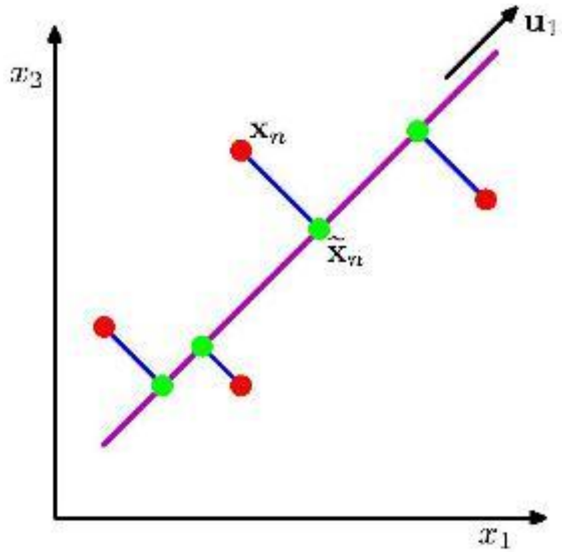
- Use: First choice when the relationship between variables is predominantly linear and fast, reproducible results are required.

Principle Component Analysis

- Goal is to find a projection that captures the largest amount of variation in data
- Approximating a high-dimensional data set with a lower-dimensional linear subspace
- Orthogonal projection of data into lower-dimension linear space that
  - minimizes mean squared distance between
    - data point and
    - projections (sum of blue lines)



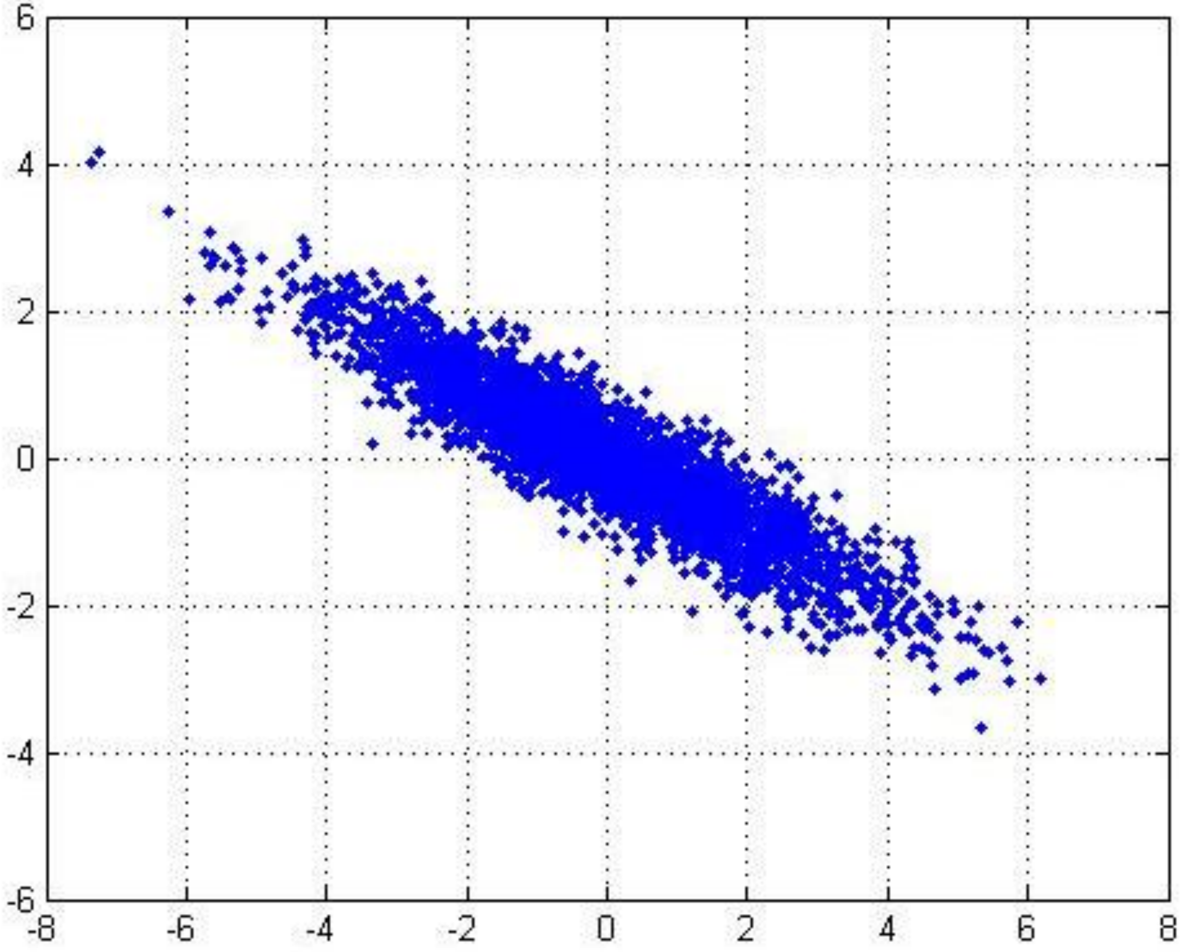




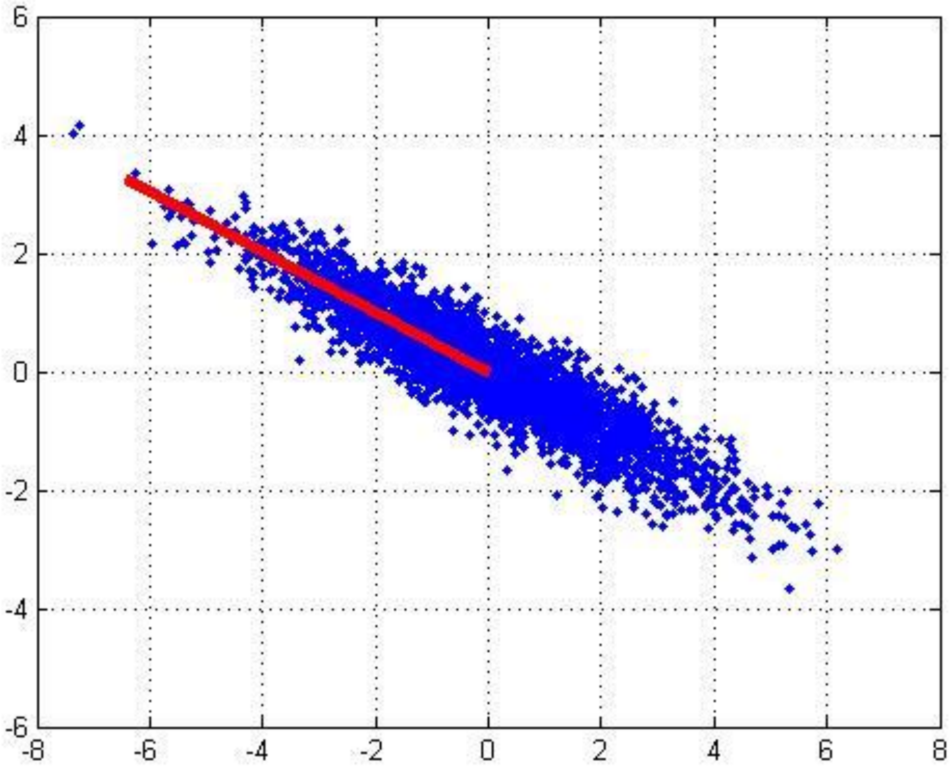
Technology Arts Sciences TH Köln

PCA: Two dimensional Values

2D Gaussian dataset

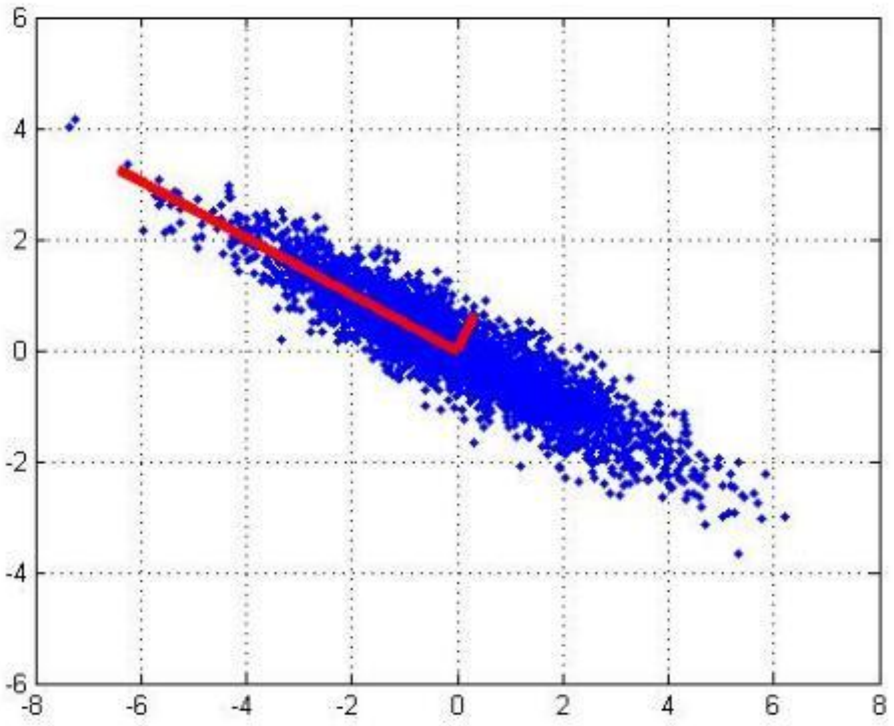


1st PCA axis



Technology Arts Sciences TH Köln

2^{nd} PCA axis



PCA algorithm: Covariance Matrix

- Given vectors {$\mathbf{x}_1$, …, $\mathbf{x}_n$}, compute covariance matrix Σ

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

for j, k = 1..n and $\overline{\mathbf{x}}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{k,i}$

- **PCA** basis vectors = the eigenvectors of Σ
- Larger eigenvalue $\Rightarrow$ more important eigenvectors

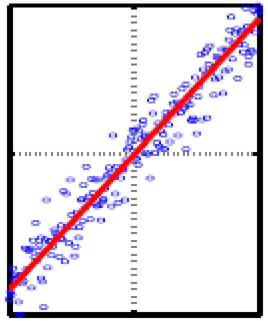
Technology Arts Sciences TH Köln

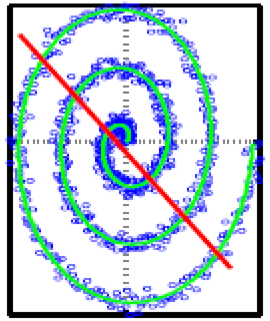
PCA Algorithm

1. Create n x m data matrix X, with one row vector x_k per each data point
2. Subtract mean $\bar{\mathbf{x}}$ from each row vector x_k in X
3. compute Σ the covariance matrix of X
   - Σ is square and symmetrical, grade m
4. Find all eigenvectors and eigenvalues of Σ
5. PCA‘s are the eigenvectors with the largest eigenvalues

Properties of PCA

- Strengths
  - Eigenvector method
  - No tuning parameters
  - Non-iterative
  - No local optima
- Weaknesses
  - Limited to linear projections





PCA is for linear, fast analyses.

Dimensionality Reduction: PCA

Technology Arts Sciences TH Köln

Dimensions = 206

T-SNE

…when the local similarity of data is important

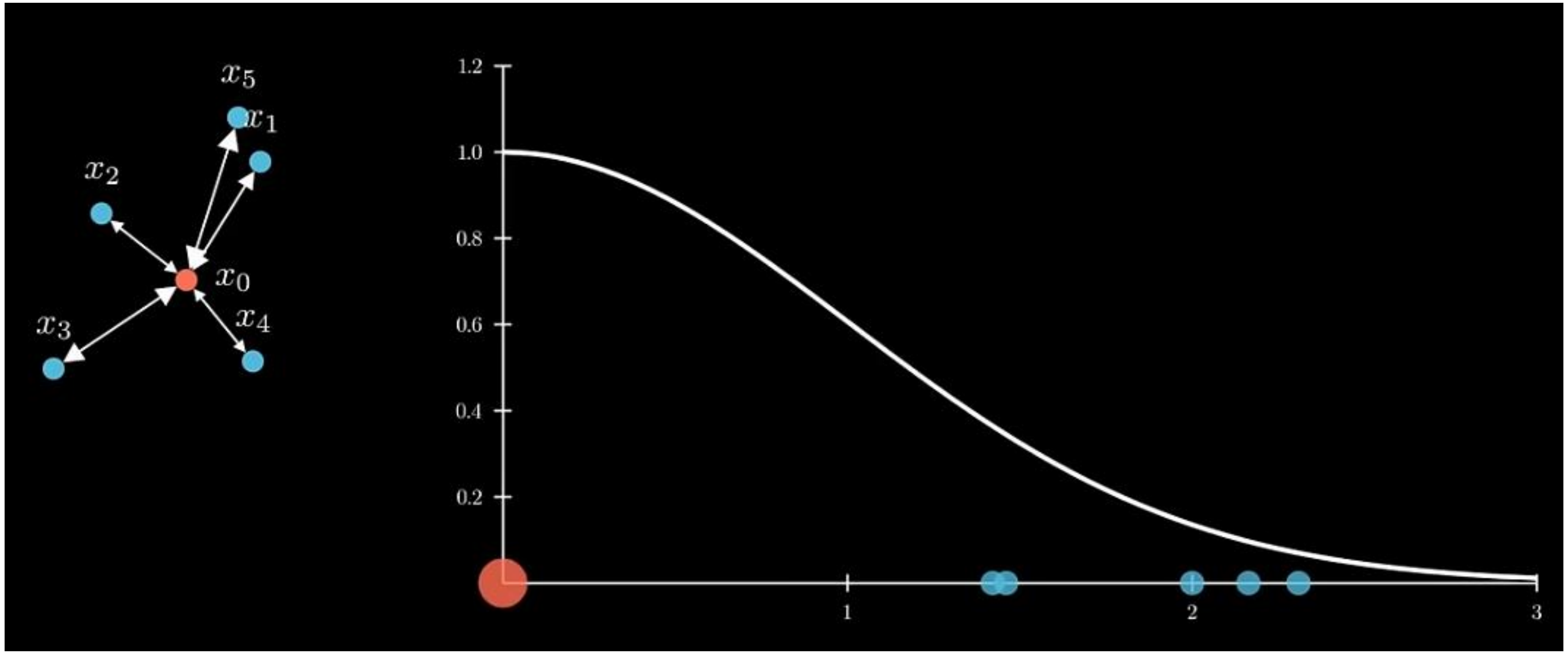
t-distributed Stochastic Neighbor Embedding

- Type: **Non-linear, probabilistic method**
- Goal: Preservation of local neighborhoods; similar points in the original space remain close in the 2D/3D projection space
- Preservation: **Strong local focus** (clusters become clearly visible), global structure usually distorted
- Computational effort: Relatively high ($O(N^2)$), scales poorly to very large data sets
- Stochastic: Results vary slightly with different initializations
- Parameters: "Perplexity", learning rate, etc., must be fine-tuned

- Use: Ideal for explorative data analysis when the focus is on cluster structures. Not suitable for directly interpreting distances between distant clusters.

t-SNE is excellent for detecting fine clusters but is computationally intensive and not globally accurate.

t-distributed Stochastic Neighbor Embedding



UMAP

…when the global structure of data is important

Uniform Manifold Approximation and Projection

- Type: **Non-linear, graph-based method**
- Goal: Reconstruction of a weighted graph of the data and optimization of a low-dimensional embedding
- Conservation: Balances local and global structures better than t-SNE
- Computational effort: Significantly faster than t-SNE, scales up to millions of points
- Deterministic (with fixed random seed): Reproducible with the same seed
- Parameters: n_neighbors (local vs. global focus), min_dist (cluster density)

- Use: If you want to see both cluster structures and rough global relationships and need to process large data sets efficiently.

UMAP combines the strengths of both approaches: fast calculation, good cluster representation and at the same time a certain global coherence

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

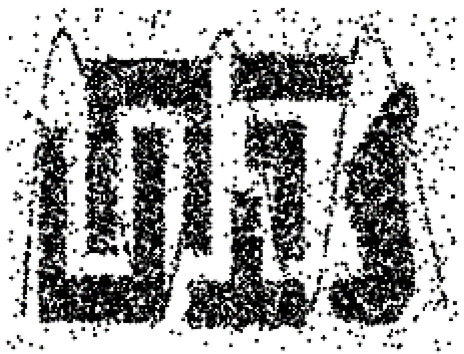
Sampling …

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data
- Example:
  - Extrapolation in political votings

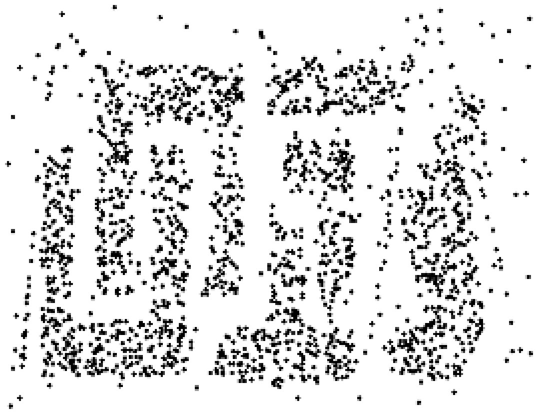
Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions;
  - then draw random samples from each partition

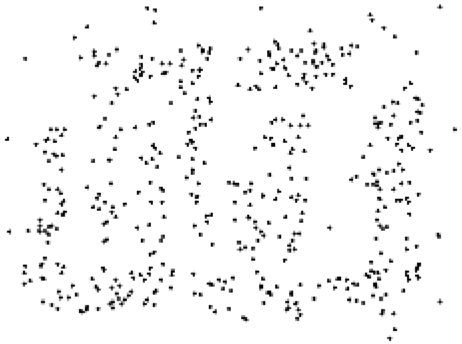
Sample Size: Patterns



8000 points



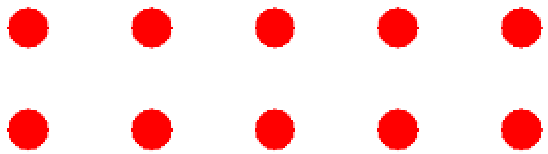
2000 Points

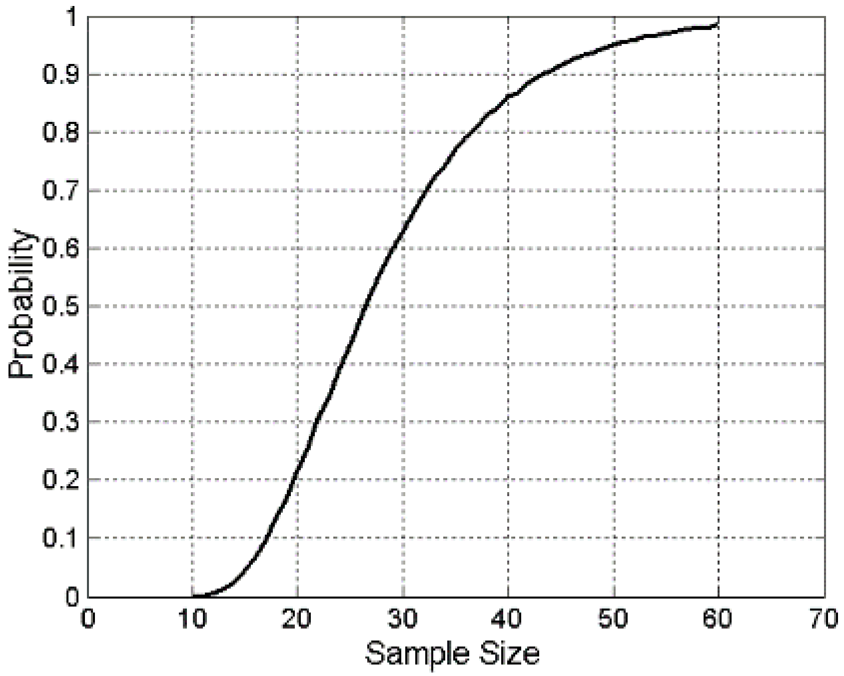


500 Points

Sample Size

- What sample size is necessary to get at least one object from each of 10 groups?





Technology Arts Sciences TH Köln

Literature

1. Bing, Liu, Web Data Mining, Springer, 2008.
2. Han, J. und Kamber, M. "Data Mining. Concepts and Techniques", Morgan Kaufmann, 2011.
3. Smith, L.I,: A tutorial on Principal Components Analysis, 2002 in http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
4. Tan, Steinbach, Kumar: „Introduction to Data Mining", Pearson Education Limited, 2013.
5. Witten I.H., Eibe, F., Data Mining, Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.
6. https://scikit-learn.org/stable/modules/decomposition.html