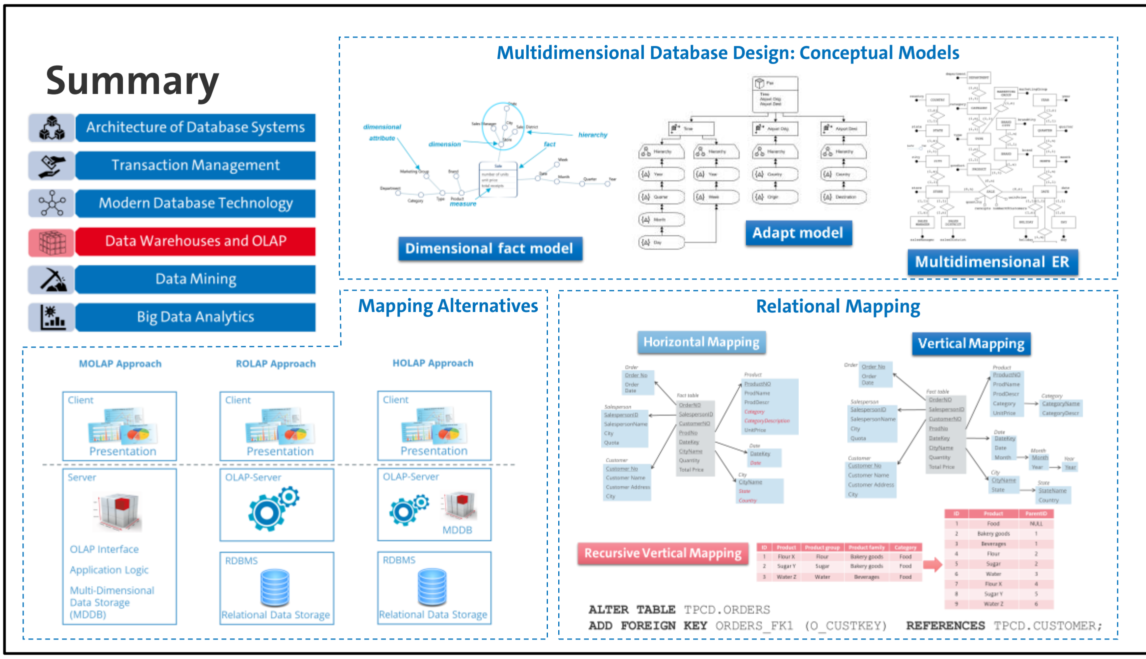
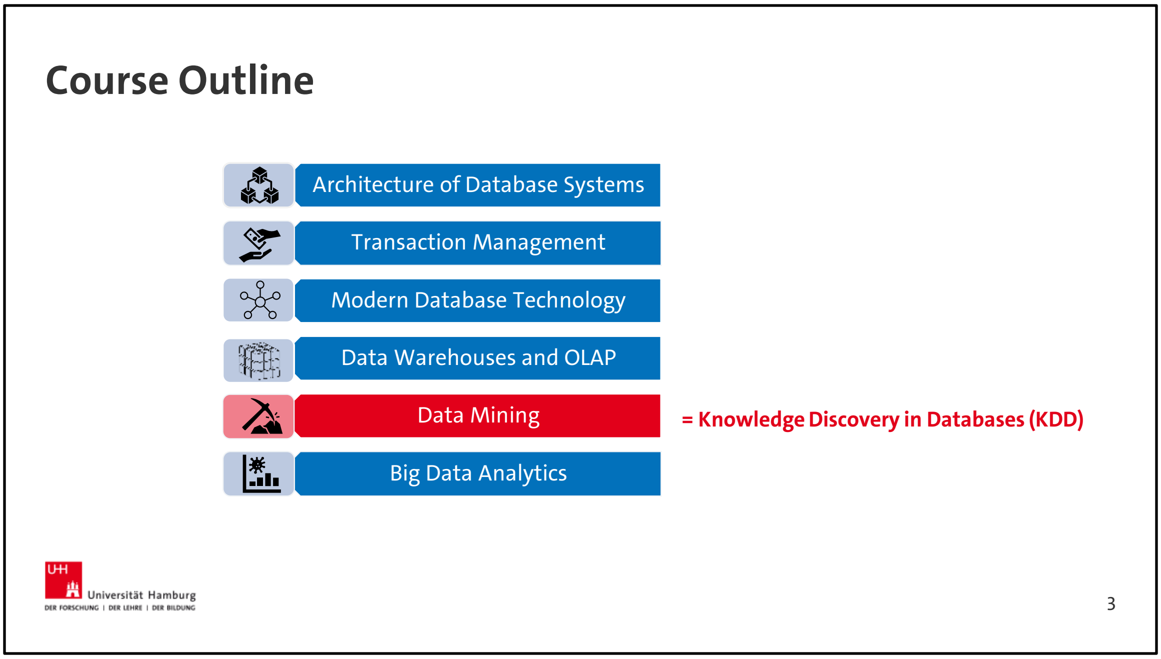
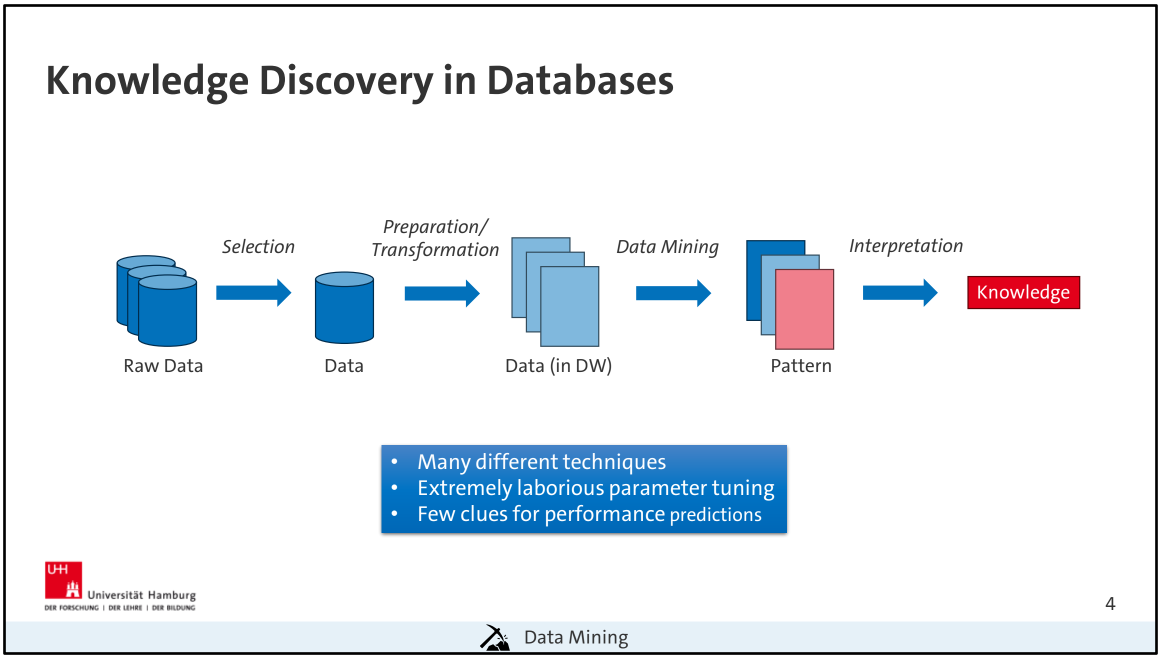

Summary
Architecture of Database Systems
Transaction Management
Modern Database Technology
Data Warehouses and OLAP
Data Mining
Big Data Analytics
Multidimensional Database Design: Conceptual Models
Dimensional fact model
Adapt model
Multidimensional ER
Mapping Alternatives
MOLAP Approach
ROLAP Approach
HOLAP Approach
Client
Presentation
Server
OLAP Interface
Application Logic
Multi-Dimensional Data Storage (MDDB)
OLAP-Server
RDBMS
Relational Data Storage
MDDB
Relational Mapping
Horizontal Mapping
Vertical Mapping
Recursive Vertical Mapping
ALTER TABLE TPCD.ORDERS
ADD FOREIGN KEY ORDERS_FK1 (O_CUSTKEY) REFERENCES TPCD.CUSTOMER;



Mining as an explorative process

- Finding cues
- Making hypotheses
- Evaluating Hypotheses
- Getting interesting/useful information



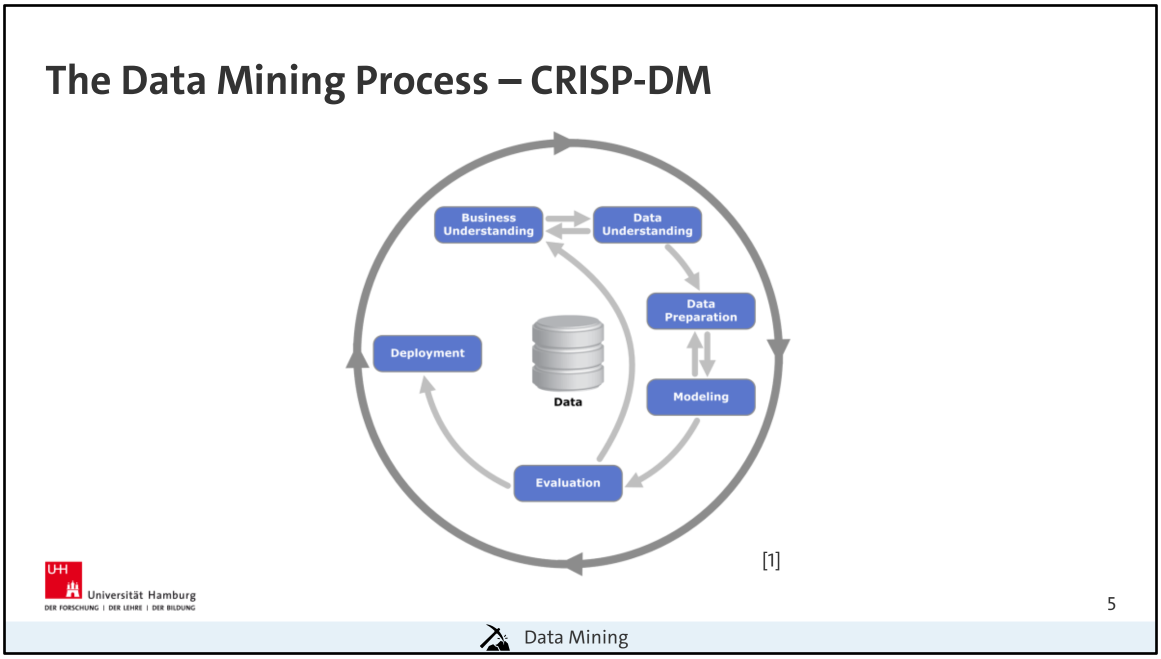
- (Semi-)automatic extraction of knowledge from databases which is:
    - Valid (in the statistical sense)
    - Unknown so far
    - Potentially useful
- Combination of approaches from databases, statistics, and machine learning
- Differences to querying a database:
    - No precise semantics
    - Not 100% perfect results
    - No perfect results
    - Solutions hard to port to similar application domains

Example tasks

- Customer relationship management
  - Grouping of customer populations (tailored marketing)
  - Prediction of customer behaviour (individualized marketing)
  - Risk assessment (risky credits, fraudulent credit card use)
- Fault analysis
  - Interdependencies between faults
  - Interdependencies between production processes or maintenance

procedures and faults

- Time-series analysis
  - Trend detection
  - Stock market development
  - Event prediction (stock market crashes, bankruptcies, natural disasters)
  - Intrusion detection
- Web Usage and Text Mining



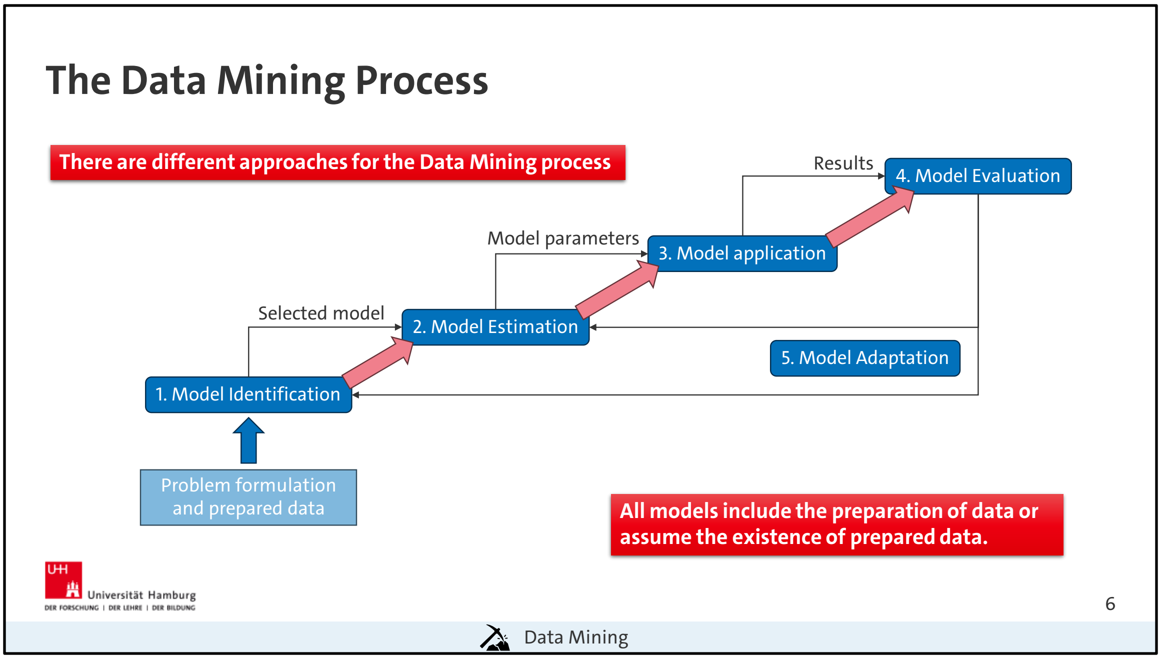
[1] source: Kenneth Jensen@Wikipedia based on ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf

- Cross industry standard process for data mining
- Life-cycle model

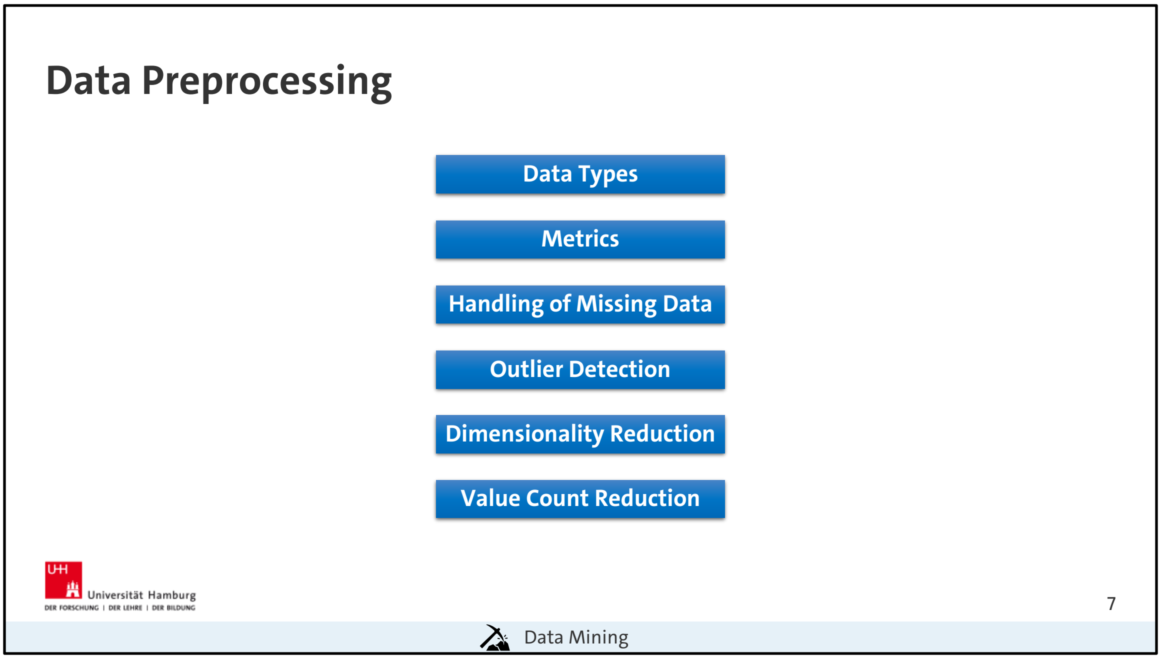
1. **Business understanding phase**
   - Analysis of objectives and requirements
   - Problem definition
   - Initial strategy development
2. **Data understanding phase**
   - Data collection
   - Exploratory data analysis
   - Assessment of data quality
3. **Data preparation phase**
   - Cleansing, transformation etc.
4. **Modelling phase**
   - Selection of modelling techniques and tools

- Parameter tuning / optimization
- Data analysis

5. **Evaluation Phase**
    - Evaluation of the model
    - Comparison of the outcome to the initial objectives
    - Deployment decision
6. **Deployment phase**
    - Reporting
    - Transfer to other application cases
    - If applicable: introduction into day-to-day business



1. **Model Identification** – Choose the optimal model type
2. **Model Estimation** – Instantiation of the model by training its model parameters
3. **Model Application**– Usage of the model to calculate the next results
4. **Model Evaluation** – Compare the model‘s results with real values using an error measure
5. **Model Adaption** – Adaption of the model parameters or the model type

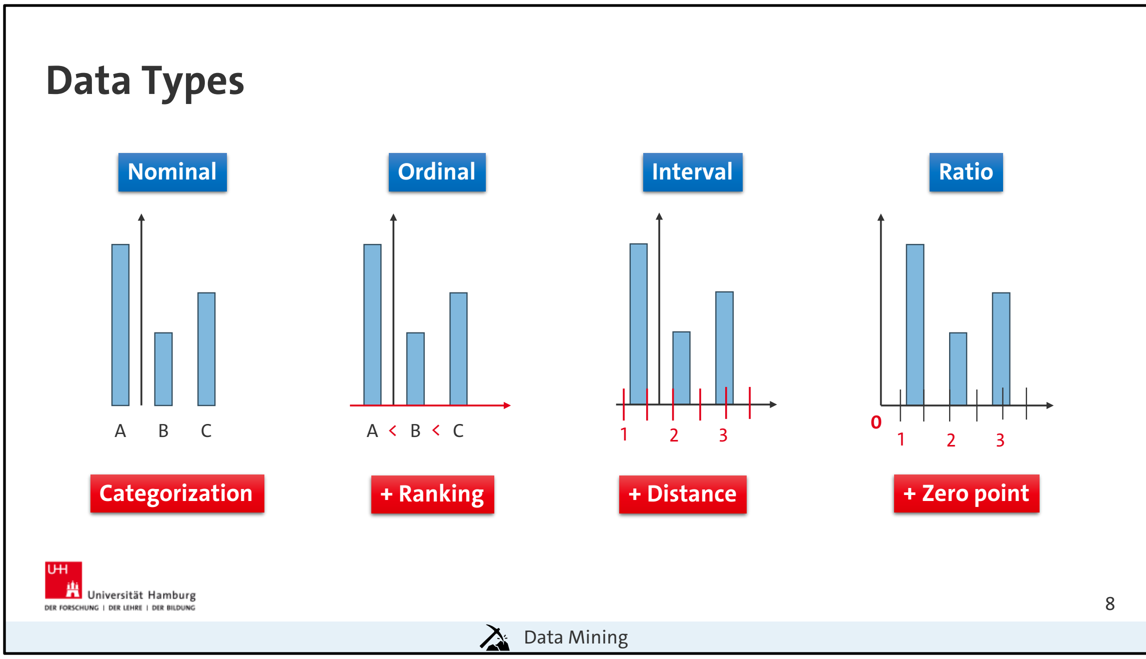


Goals

- Decrease runtime of data mining process
- Decrease resource requirements of data mining process
- Increase quality of mining result

Important aspects

- Handling of missing data
- Detecting outliers
- Reducing the number of dimensions
- Reducing the number of values
- Transforming data values (e.g. binning, rescaling)
- Depends on data type
- Uses similarity/distance measures



Nominal scale

- No problem-specific order and distance relation
- Mathematical Operators: =, !=
- Central Tendency: mode (most often occurring value)
- Examples: color, zip-code

Ordinal scale

- Problem-specific order relation
- No problem-specific distance relation
- Mathematical Operators: =, !=, >, <
- Central Tendency: mode, median
- Examples: income classes, medal ranks, age

Interval scale

- Problem-specific order and distance relation
- No problem-specific zero point
- Differences meaningful, ratios meaningless
- Mathematical Operators: =, !=, >, <, +, −
- Central Tendency: mode, median, arithm. mean, deviation

- Examples: temperature (Celsius), date (B.C./A.D.)

Ratio scale
- Problem-specific zero point (absence of the feature)
- Differences and ratios meaningful
- Mathematical Operators: =, !=, >, <, +, −, ·, /
- Central Tendency: mode, median, arithm./geom. mean, deviation
- Examples: temperature (Kelvin), speed, length, age, quantity

Metrics

Many data mining techniques are based on some notion of distance, similarity or dissimilarity.

- Data Types
- **Metrics**
- Handling of Missing Data
- Outlier Detection
- Dimensionality Reduction
- Value Count Reduction

Examples

Numeric values/points: Minkowski distance

$$d(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^m\right)^{1/m}$$

m=1: Manhattan distance
m=2 Euclidian distance
m=∞ Tchebychev distance

Sets/Bags/Vectors:
- **Cosine similarity** = cos(θ)
- **Jaccard Coefficient**

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Signatures, histograms, Probability distributions:
- **Earth Mover's distance/Wasserstein metric**

Strings:
- **Soundex (phonetic)**
- **Monge-Elkan similarity (sequence- and set-based)**
- **Levenshtein distance (sequence based)**

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Levenshtein Distance

String edit distance = number of insertions/deletions/exchanges necessary to transform one string into another

Examples
- “Bear” and “Bar” → 1 (insert “e”)
- “Bear” and “Goal” → 3 (exchange “B”, “e”, and “r” for “G”, “o”, and “l”)

Algorithm for distance between word x and word y

1. Create initial matrix of size (|x|+1, |y|+1)
2. Fill 1st row and column with ascending numbers starting at 0
3. Compute the rest pf the matrix according to the following rule:

$$D_{i,j} = min \begin{cases} D_{i-1,j-1} \; if \; x_i = y_i \\ D_{i-1,j-1} + 1 \; (exchange) \\ D_{i,j-1} + 1 \; (insert) \\ D_{i-1,j} + 1 \; (delete) \end{cases}$$

4. Get the distance from the lower right cell in the matrix

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Finding the minimum distance is an optimization problem
- Space and time requirements O($m \cdot n$)

Levenshtein Distance: Example

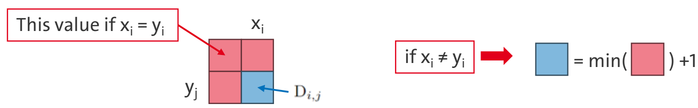
1. Create initial matrix of size (|x|+1, |y|+1)
2. Fill 1st row and column with ascending numbers starting at 0
3. Compute the rest pf the matrix according to the following rule:

$$D_{i,j} = min \begin{cases} D_{i-1,j-1} \; if \; x_i = y_i \\ D_{i-1,j-1} + 1 \; (exchange) \\ D_{i,j-1} + 1 \; (insert) \\ D_{i-1,j} + 1 \; (delete) \end{cases}$$

4. Get the distance from the lower right cell in the matrix

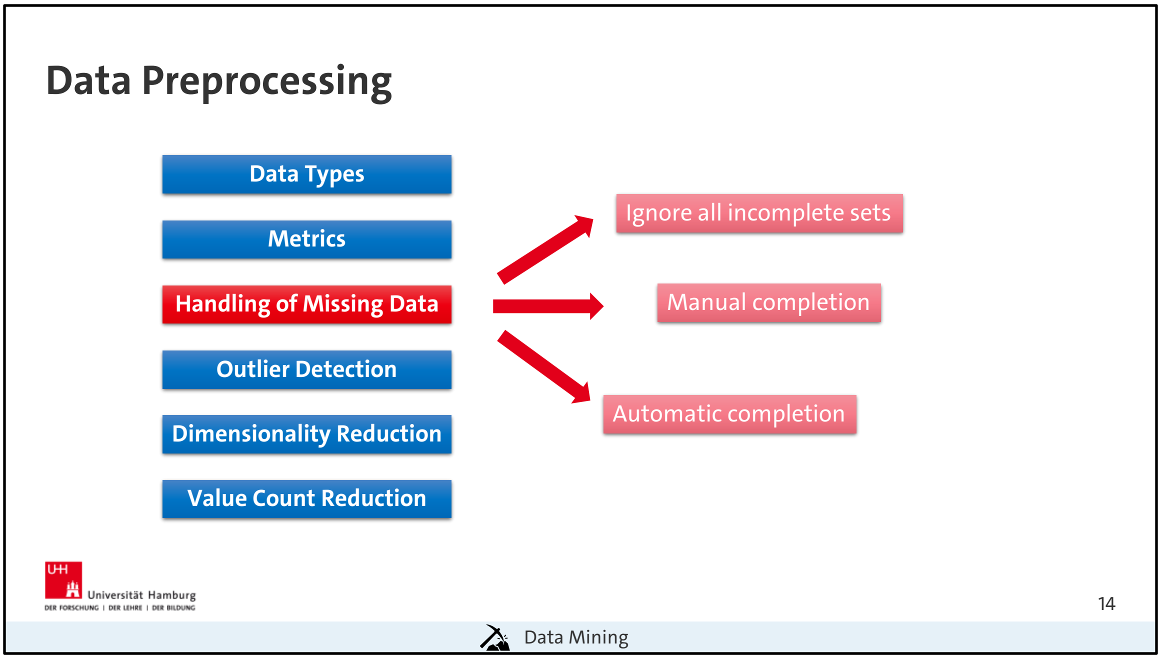
		b	e	a	r
	0	1	2	3	4
g	1	1	2	3	4
o	2	2	2	3	4
a	3	3	3	2	3
l	4	4	4	3	3

Translation for step 3



Exercise: Levenshtein Distance

Determine the Levenshtein distance (and the according matrix) between “english” and “danish”



Some data mining tools are insensitive to missing data

Ignore all incomplete objects

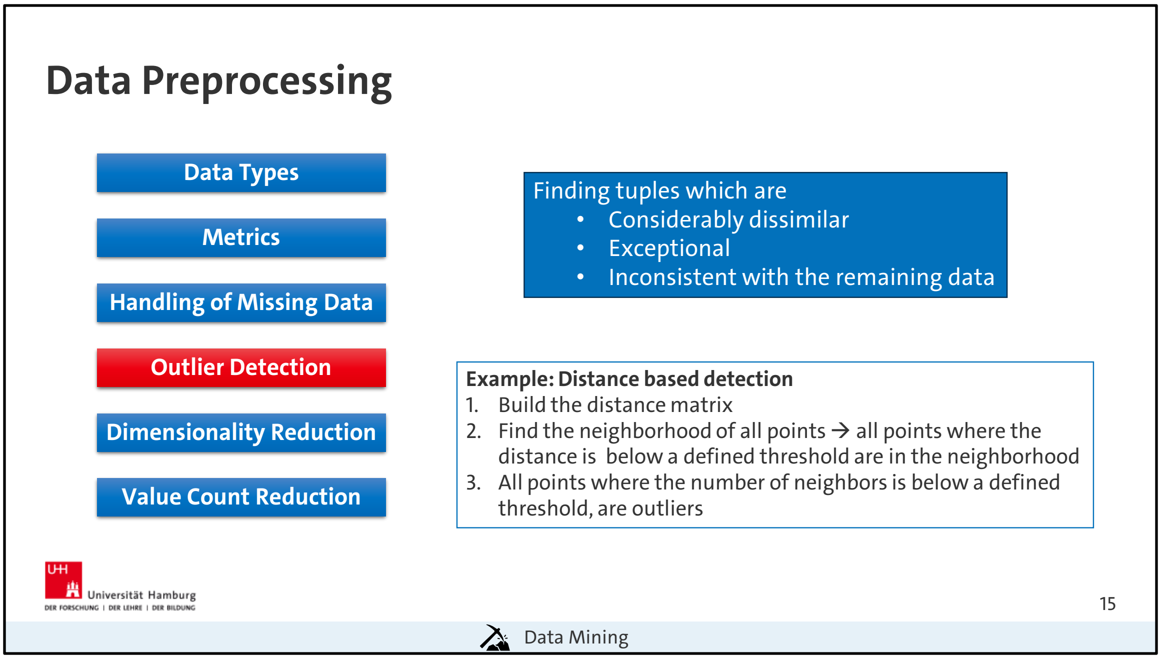
- Might result in the loss of a substantial amount of data
- Problem: maybe a systematic correlation between target of mining process and missing values (e.g. customers who do not answer a specific question of a survey)

Manual completion

- Expensive in time and money

Automatic completion

- Using a global constant
- Using the global mean value
- Using a class-dependent mean value
- Use a predictive model, e.g. based on feature correlations



- Not applicable if the application is aimed at outlier detection, e.g. fraudulent credit card transactions

Manual detection supported by visualization tools

- Only for low dimensional data

Statistical methods

- Threshold for the variance, e.g. two times variance
- Only applicable if the distribution is known

Using domain knowledge

- Value restrictions, e.g. $0 < age < 150$
- Less applicable for multi-dimensional data

Distance-based detection

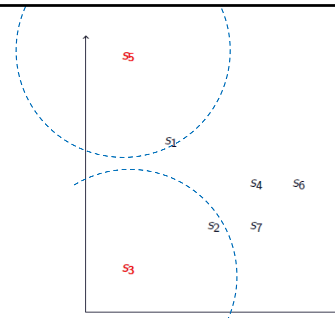
- A sample is an outlier if it has not enough neighbours

Deviation-based methods

- Measure the dissimilarity of a data set (e.g. variance)
- Determine the smallest subset of data that if removed results in the largest reduction of dissimilarity
- Combinatorics of subset selection → extremely expensive

Outlier Detection: Example

Data Set $S = \{s_1, \ldots, s_7\} = \{(2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2)\}$



Distance matrix **Neighborhood: d ≤ θ = 3**

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1		2.236	3.162	2.236	2.236	3.162	2.828
s_2	2.236		2.236	1.414	4.472	2.236	1.000
s_3	3.162	2.236		3.605	5.000	4.472	3.162
s_4	2.236	1.414	3.605		4.242	1.000	1.000
s_5	2.236	4.472	5.000	4.242		5.000	5.000
s_6	3.162	2.236	4.472	1.000	5.000		1.414
s_7	2.828	1.000	3.162	1.000	5.000	1.414	

Number of neighbors per point

Sample	s_1	s_2	s_3	s_4	s_5	s_6	s_7
	4	5	1	4	1	3	4

If threshold for outliers is set to < 2 neighbors, S_3 and S_5 are outliers

Exercise Outlier Detection

S = {S1,...,S4} = {(2,3),(1,1),(3,3),(3,2)}

Data Preprocessing

- Data Types
- Metrics
- Handling of Missing Data
- Outlier Detection
- **Dimensionality Reduction**
- Value Count Reduction

High-dimensional spaces are sparse
→ Keeping the same object density in a space with more dimensions requires exponentially more objects
→ To enclose a prespecified portion of objects, an increasingly large part of the hypercube needs to be “encircled”

Example

Portion p	Dimensions n	Edge length e	$e^n = p$
0.1	1	0.1	$0.1^1=0.1$
0.1	2	0.316	$0.316^2 = 0.1$
0.1	3	0.464	$0.464^3 = 0.1$
	…		
0.1	10	0.794	$0.794^{10} = 0.1$
	…		
0.1	100	0.977	$0.977^{100} = 0.1$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

18

Data Mining

- Almost every object is closer to an edge of the cube than to another sample object
- Almost every object is an outlier

Which data can be discarded without sacrificing the quality of the data mining results?

Too many dimensions:
- Mining results degrade (insufficient amount of data)
- Resulting model is incomprehensible
- Problem becomes intractable

Too few dimensions:
- Data dependencies are lost
- Mining results degrade (limited expressiveness)

Feature selection approaches:
Idea: discard features (i.e. attributes, dimensions) which
- have many inaccurate/inconsistent values
- have many missing values

- do not provide much (relevant) information
- contribute least to the overall class distinction capability (task specific criterion)

Approaches: Feature ranking, minimum subset selection

Feature composition approaches:

Idea: features are composed into a new feature set with reduced dimensionality
→ Given feature space is transformed into a more compact feature space without losing relevant information

Approach: Principal component analysis (PCA)

Data Preprocessing

- Data Types
- Metrics
- Handling of Missing Data
- Outlier Detection
- Dimensionality Reduction
- **Value Count Reduction**

Increase Performance (less values to process)
Simplify mining process (e.g. finding of rules)

One dimensional
Feature discretization (binning)
→ Mapping values to intervals

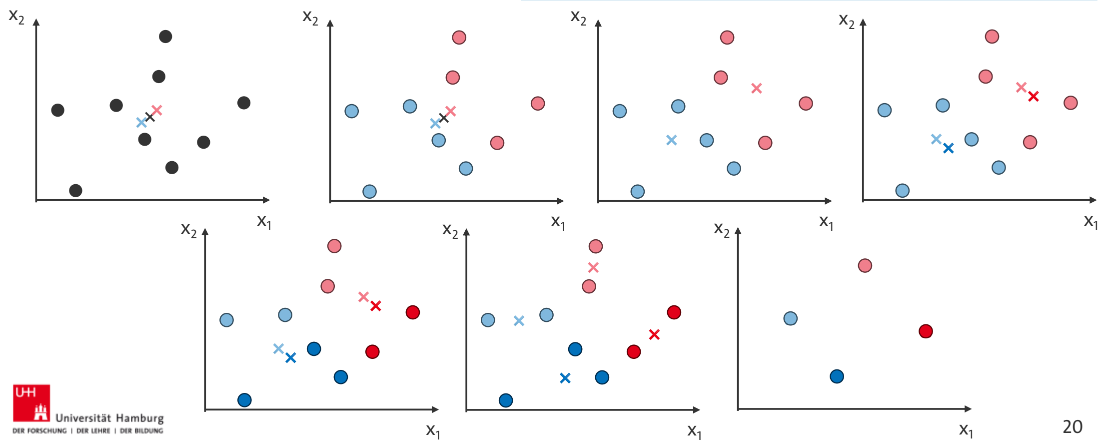
Multi dimensional
Clustering of feature vectors
→ Splitting techniques

Splitting Techniques: Splitting of bins (vector quantization)
1. Start: All values in a single bin/cluster
2. Compute the mean of all data values (centroid)
3. Split each centroid into 2 or more centroids (bins)
4. Assign data points to the nearest centroid (bin)
5. Continue until enough bins have been generated

Splitting of bins

Input: 9 values
Output: 4 values

1. Start: All values in a single bin/cluster
2. Compute the mean of all data values (centroid)
3. Split each centroid into 2 or more centroids (bins)
4. Assign data points to the nearest centroid (bin)
5. Continue until enough bins have been generated



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG