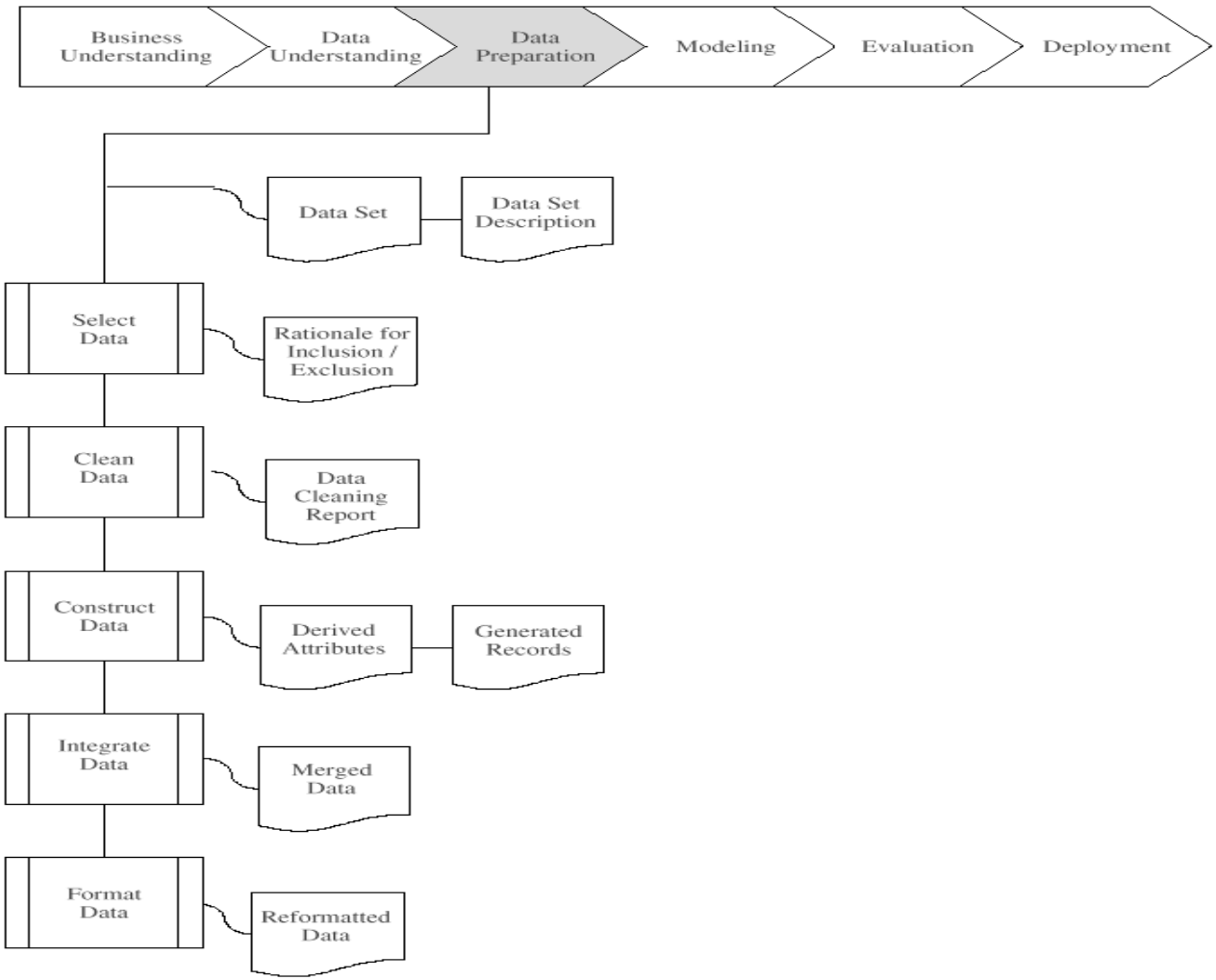
Kapitel IV

Preprocessing

IV Preprocessing



IV. Preprocessing

IV.1 Introduction in preprocessing

Purpose of preprocessing

- transform datasets so that their information context is best exposed to the mining tool

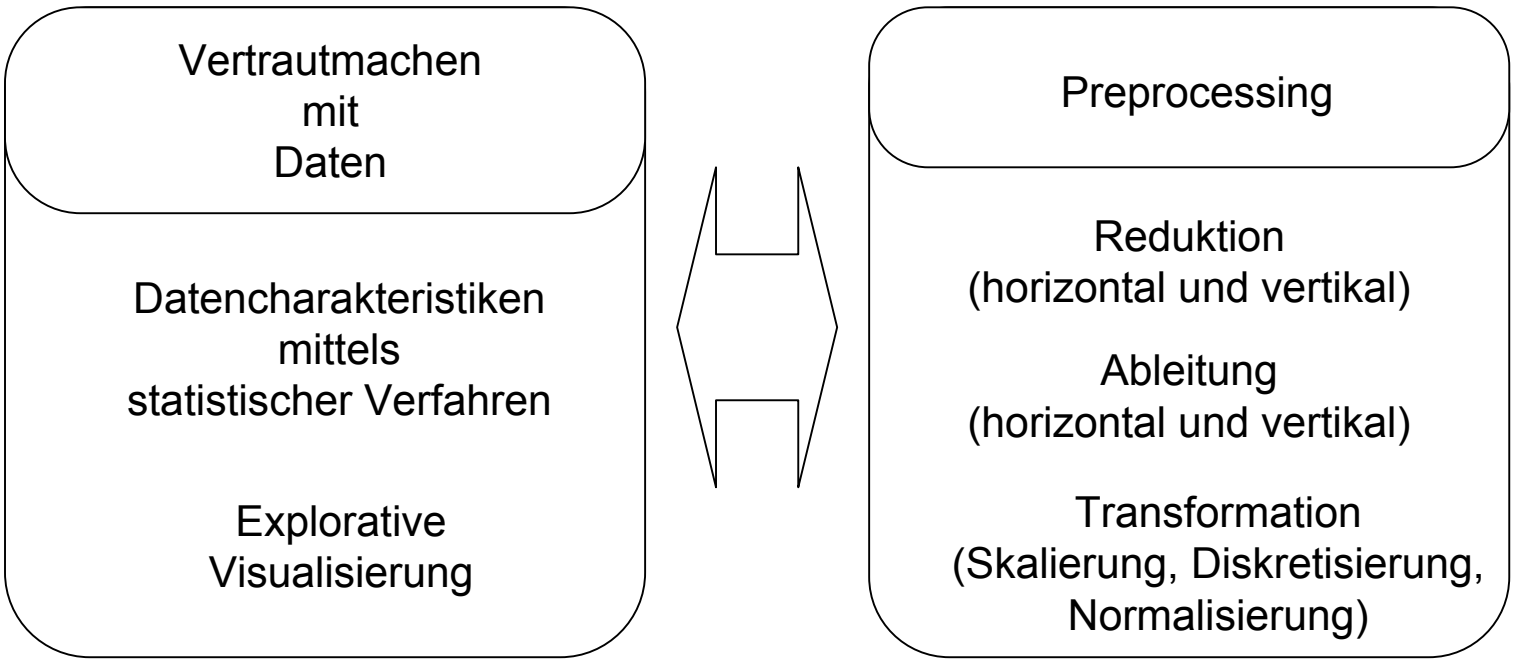
Problem: - learn the „true“**relationship**
- do not learn **noise**

2 Types of Preprocessing

prepare the data

prepare the miner

Verknüpfung von Preprocessing und Datenverständnis



Weitere Aspekte des Preprocessing:

Gute Vorverarbeitung benötigt das Wissen eines Domänenexperten

Daten-Kontext

- Falsche Verteilung
- nominal vs. ordinal
- Korrelation

Domänen-Kontext

- richtige Daten im Kontext?
- Repräsentieren die Daten gesuchte Zusammenhänge?
- weitere Daten notwendig?

Beispiel:

Repräsentieren meine Daten die gesuchten Zusammenhänge aus Sicht der Domäne?

OLAP

- Visualisiert schnell die Zusammenhänge in den Daten
- Daten können schnell manipuliert werden

- Domänenexperte bekommt schnell einen Überblick über den aktuellen Stand der Daten und kann die gestellte Frage beantworten.

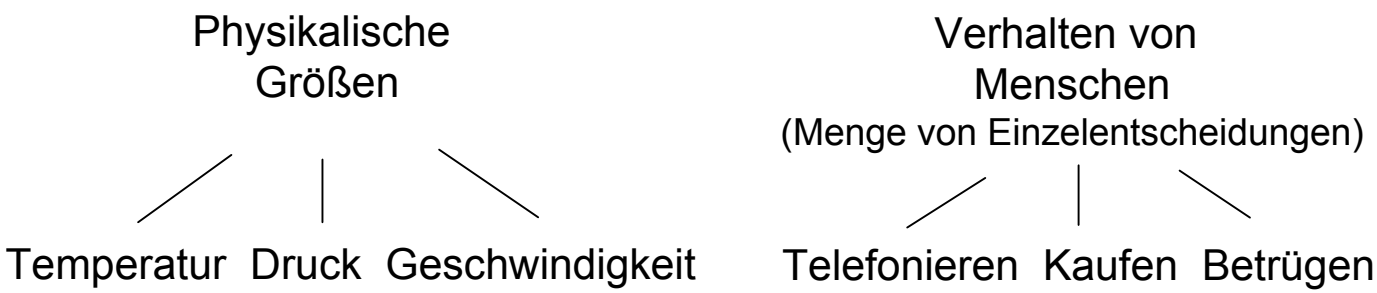
Telekom:

Bei der Deutschen Telekom AG sammelt man „nur“ über die eigenen Kunden schon länger Informationen über das Gesprächsverhalten. Für eine wieder einmal durchzuführende Analyse stellt sich die Frage:

Kann man mit den Daten aus dem Jahr 1997 eine allgemeingültige Aussage z.B. über das Telefonverhalten der Deutschen machen? Geht dies auch noch 1999?

- 1997 ja
- aber 1999 nicht mehr, da nicht mehr alle Telefonbesitzer auch Kunden der Deutschen Telekom AG sind.
- Außerdem ist zu beachten, daß auch 1997 mit den gesammelten Daten nur Aussagen über die Telefonbesitzer gemacht werden können, nicht aber über alle Deutschen.

Prinzipelle Unterschiede beim Preprocessing



Es ergeben sich unterschiedliche Probleme bei den Daten:

- komplexe Zusammenhänge
- meist nicht-linear
- konsistent

- häufig fehlende Werte (missing values)
- sehr große Datenmenge
- häufig inkonsistent

Beispiel:

Prozessoptimierung in einem Chemieunternehmen

Analyse des Verhaltens der Kunden einer Telefongesellschaft

IV.2 Preprocessing Steps

Data Cleansing (Datenbereinigung)

- consistency (Konsistenz)
- detail / aggregation level (Aggregationsniveau)
- pollution (Verunreinigung)
- relationship (Beziehungen)
- range (Definitionsbereich)
- defaults
- duplicate or redundant variables
- missing and empty values (fehlende Werte)

Data Manipulation (Datenmanipulation)

- reverse pivoting
- reducing dimensionality
- increasing dimensionality
- sparsity (schwach besetzte Werte)
- monotonicity (Monotonie der Daten)
- outliers (Ausreisser)
- numerating categorical values
- anachronisms
- relation between variable via pattern in the variable
- combinatorial explosion

a) Data Cleansing

Consistency

- different things are represented by the same name in different systems
- same things are represented by different names in different systems

Detail / Aggregation level

- transaction record (detailed) vs. summarized transaction record (aggregated)
- general rule for data mining: detailed data is preferred to aggregated data
- level of detail in the input stream is one level of aggregation more detailed than the required level of detail in the output

Pollution

- garbage in the data, e.g. comma delimited data with comma in the data
- Human resistance, e.g. data fields are blank, incomplete, inaccurate

Relationship

- merging multiple input streams (use for example keys)
- find the right keys, eliminate double keys

Range

- variable has a particular domain, a range of permissible values
- detect outliers

Defaults

- the miner must know the default values of data capturing programs
- conditional default values can create seemingly significant patterns

Duplicate or redundant variables

- identical information in multiple variables, e.g. „date of birth“ and „age“
- problem for neural network with colinearity of variables

Missing and empty values

- empty values may not have a corresponding real-world value or have a real-world value but it was not captured
- miner should differentiate between both types of values
- data mining tools have different strategies to handle these values

b) Data Manipulation

Reverse pivoting

- modelling important things under the right point of view
- Example: Database with detailed call records
  Task is to analyse customers
  Problem: the focus of the database is not the customer

Reducing dimensionality

- eliminate features, which are not important for your task

Increasing dimensionality

- expand one dimension to represent the information in a better way
- example: Zip code can be transformed in “Lat” and “Lon”

Sparsity

- individual variables are only sparsely populated with instance values
- miner must decide e.g. to remove or to collapse the variable

Sample

- take only a part of the collected data (population)
- do not lose any information

Monotonicity

- a monotonic variable is one that increases without bound
- example variable: date, time, social number
- must be transformed in a non-monotonic form, since prediction can only be performed within the range of the learning data set

Outliers

- single or very low frequency occurrence of the value of a variable
- far away from the bulk of the values of the variable
- Is the outlier a mistake or very important information?

• **Anachronisms**

• something out of place in time
• information not actually available in the data when a prediction is needed

Relation between variable - pattern in the variable

• enough instance values to represent the variable's features are needed to detect patterns
• interactions between variables are interesting too

Combinatorial Explosion

• if you are interested in the interactions between variables you must check every combination of your variables

Number of variables	Number of combination
5	26
9	502
25	33.554.406

IV.3 Preprocessing example for categorical data

How categorical values are best represented depends very much on the needs of the modelling tool.

Enumeration of categorical values

Time period	Rate of pay ($)
Half-day	100
Day	200
Half-week	500
Week	1000
Half-month	2000
Month	4000

Time period	...	Rate of pay ($)
Day	1	200
Half-day	2	100
Half-month	3	2000
Half-week	4	500
Month	5	4000
Week	6	1000

- both tables show the rate of pay for a period in dollars
- you don‘t see a structure if you order the time periods alphabetically (right table)

IV.3 Preprocessing example for categorical data

Problem:

At worst (if the scale niveau is nominal) enumeration of categorical values introduces and creates patterns in the data that are not natural and that reflect throughout the data set, wreaking havoc with the modelling process.

- **Domain Knowledge can help**
- **do as little damage as possible to the natural structure**

Measure of distance for categorical data

Example: call detail record

customerID	distance	type of day	date/time	comm. minutes
1	Ort	Mo-Fr	19.11.98/9:55	20 min
1	Ort	Mo-Fr	20.11.98/10:10	18 min
2	Regional	Mo-Fr	19.11.98/21:00	120 min
2	Regional	Mo-Fr	20.11.98/17:00	2 min

Problem: Which records are similar to each other?

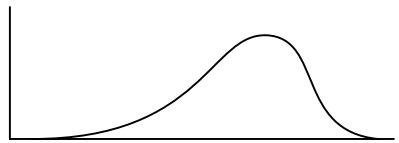
- the simple answer is: all are different
- a person would say: the first two records are similar, the last two not

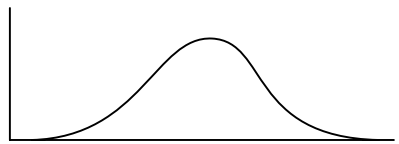
Most tools need a similarity measure to process the data automatically.
But always check that your measure is meaningful!

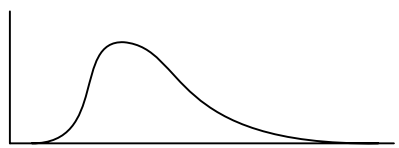
IV.4 Preprocessing example for numerical data

Normalising a variable‘s Range with ladder of power (Tukey 1977)

p	Transformation $T(x_i)$	Name
...	...	...
10	x_i^{10}	dezimal
...	...	...
3	x_i^3	kubisch
2	x_i^2	quadratisch
1	x_i	**Rohdaten**
$\frac{1}{2}$	$\sqrt{x_i}$	Wurzel
0	$\log(x_i)$	logarithmisch
$-\frac{1}{2}$	$-\frac{1}{\sqrt{x_i}}$	reziproke Wurzel
1	$-\frac{1}{x_i}$	reziprok
...	...	...





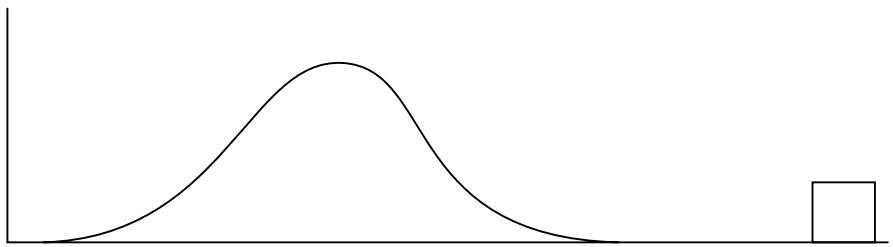


IV.4 Preprocessing example for numerical data

Discretize numerical variable

- take a range of values and map it to a new value
- understand the underlying distribution

Select the right range



Why?

- all variables have a particular resolution limit in practice
  - accuracy of measurement
  - precision of representation
- if value is out of range - two different input values become the same

IV.5 Preprocessing example for missing and empty values

Difference between missing and empty values

- empty values have no corresponding real-world value
- missing values have a real-world value but it was not captured

Tools can have difficulties in handling such values

- ignore missing and empty values
- use some metric to determine „suitable“ replacement
- automated replacement techniques are critical
  - Does the miner know the problems of the technique?
  - Does the miner know the replacement method being used?
  - What are its limitations?

Task for miner

- replacement must be as **neutral** as possible
- use a method understood and controlled by the miner

Replacement: Problems and Aspects

- some modelling techniques cannot deal with missing values
- default replacement methods may introduce distortion
- know and control the characteristics of any replacement method
- important information is sometimes in the missing-value patterns

Example

- the data had carefully been prepared for warehousing, including the replacement of the missing values
- data warehouse data resulted in a remarkable poor quality of the learned model
- quality was improved when the original source data was used
  - most predictive variable was the missing-value pattern

Übersicht über die weitere Vorlesung

VI. **Überwachte Data und Text Mining Verfahren**
- Entscheidungsbaumverfahren C4.5
- Induktives Logisches Programmieren (ILP)
- Künstliche Neuronale Netzwerke

VII. **Unüberwachte Data und Text Mining Verfahren**
- Clustering: Self Organizing Maps
- Formale Begriffsanalyse
- Assoziationsregeln
- Generalisierte Assoziationsregeln mit Taxonomien

V. **Text Mining**

VIII. **Modellierung (Zusammenfassung)**

IX. **Evaluierung**

X. **Anwendung**