Data Mining

1 Introduction

2 Data Mining methods

1 Introduction

1.1 Motivation

1.2 Goals and problems

1.3 Definitions

1.4 Roots

1.5 Data Mining process

1.6 Epistemological constraints

1.1 Motivation

- Which goods are bought at the same time by which customers?
- What goods should be offered a certain customer?
- Will a customer pay his invoice?
- What is the probability that a customer will cancel his contract?
- What will the trend be next season?

1.2 Goals

- optimize business processes
- find competitive advantages
- analyze customer behavior
- derive theories about future developments

1.2 Problems

- huge amount of data
- important correlations
  cannot be found by humans

1.3 Definitions

"Data Mining means different *methods*, which allow to use computer-aided algorithms
which analyze huge *amounts of data* for internal *relations* and discover new, unknown correlations in them"

(Kral, Data Mining, 1998, 11).

1.3 Definitions

"Knowledge discovery in databases is the non-trivial process to identify valid, new, potentially useful and finally understandable patterns in data"

(Fayyad, U. et al. 1996).

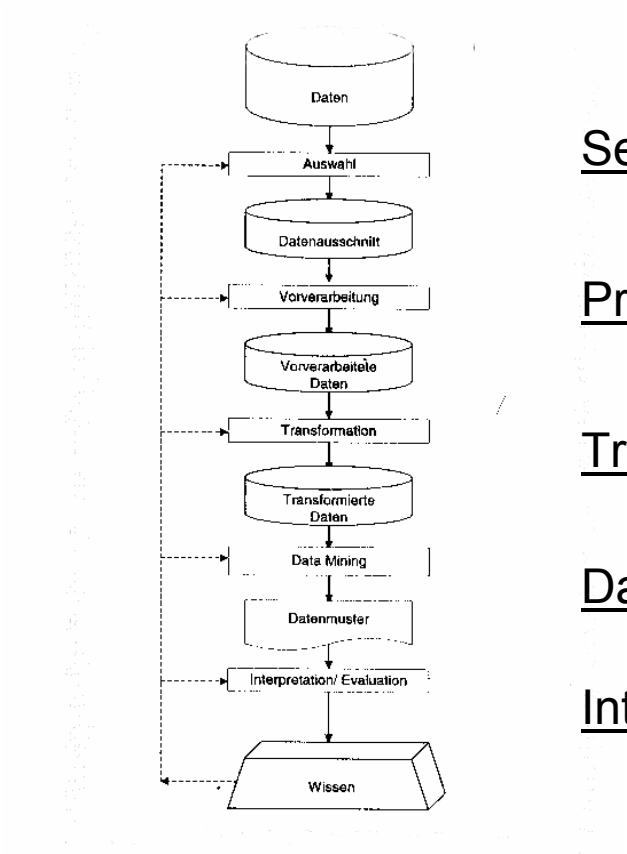
1.4 Roots

- Statistics
  (analysis of data relationships)

- Data base research
  (handling of huge amounts of data)

- Artificial Intelligence
  (data → hypothesis)

1.5 The Data Mining process: overview



Data

Selection

Extracted data

Preliminary filtering

Filtered data

Transformation

Transformed data

Data mining

Data patterns

Interpretation / Evaluation

Knowledge

1.5 The Data Mining process: overview

- Step 1: pre-processing
  - data selection
  - preliminary filtering
  - transformation
- Step 2: processing
  - execution of the Data Mining algorithm
- Step 3: post-processing
  - interpretation, evaluation

1.5.1 The Data Mining process

- Step 1: pre-processing
  - data selection
    - understanding the application area
    - identify goals
    - define which data is relevant

1.5.1 The Data Mining process

- Step 1: pre-processing
  - preliminary filtering
    - complete data
    - make data consistent
    - integrate data

1.5.1 The Data Mining process

- Step 1: pre-processing
  - transformation
    - select attributes
    - replace attributes by discrete attributes

1.5.1 The Data Mining process: important constraints of step 1

- Pre-processing is 85% of the total work.
- Data integration can be supported by using Data Warehouses.
- Def. Data Warehouse:

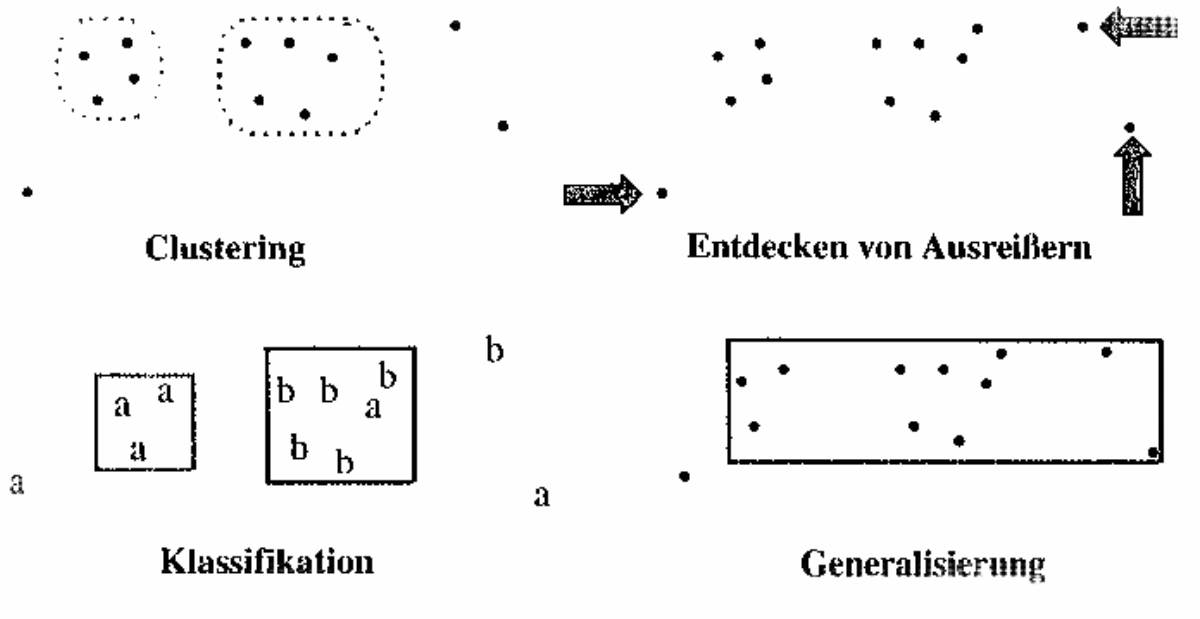
  ”A company-wide enterprise concept with the goal to build a logically central, uniform and consistent database for the different applications supporting the analytical tasks of managers.“

1.5.2 The Data Mining process

- Step 2: processing
  - execution of the Data Mining algorithm
    - find clusters
    - find anomalies
    - classify
    - generalize

1.5.2 The Data Mining process

- Step 2: processing - typical tasks



1.5.3 The Data Mining process

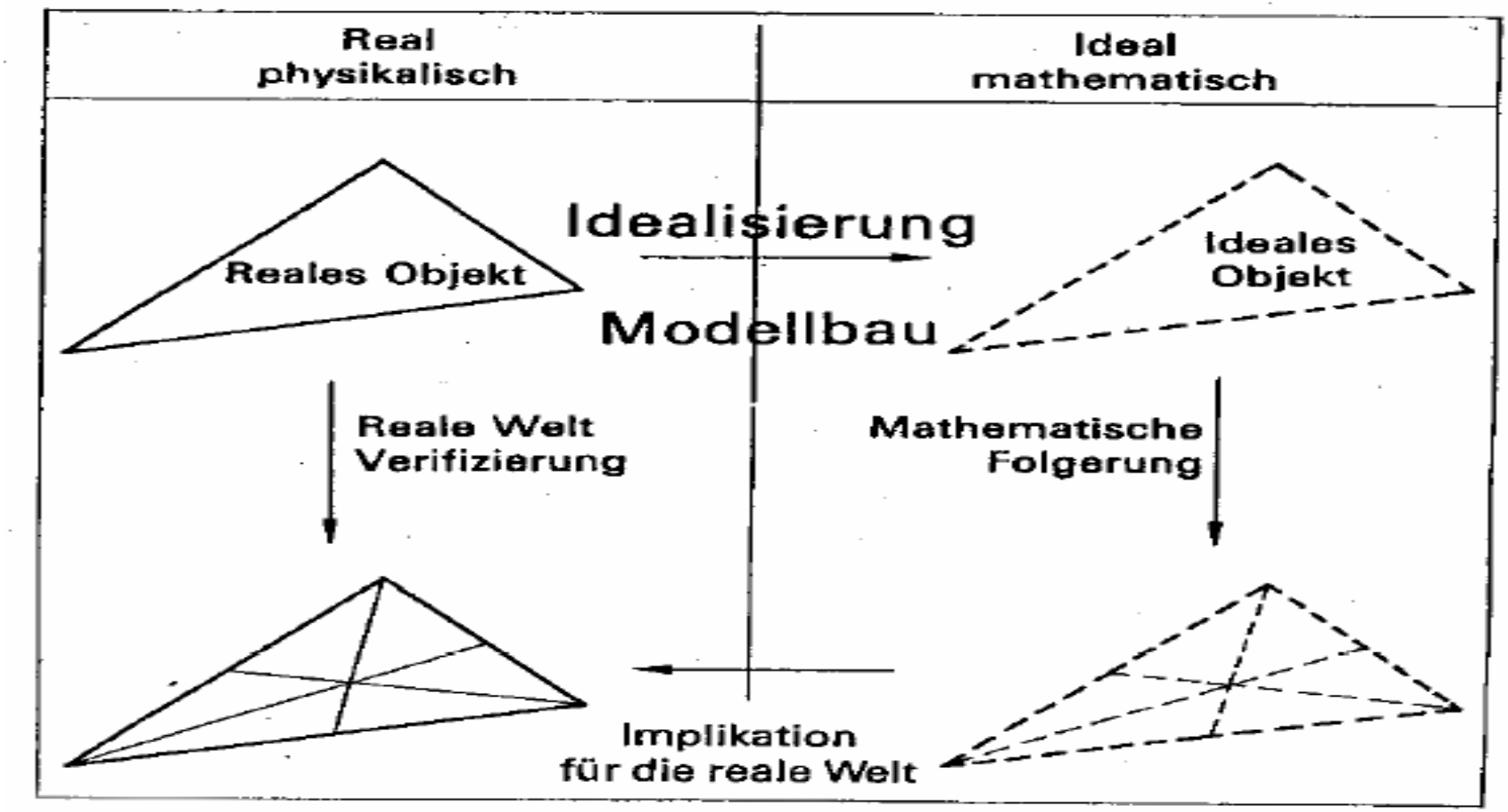
- Step 3: post-processing
  - interpretation, evaluation

    - show patterns found
    - evaluate patterns in comparison to goals
    - predict future developments and behavior

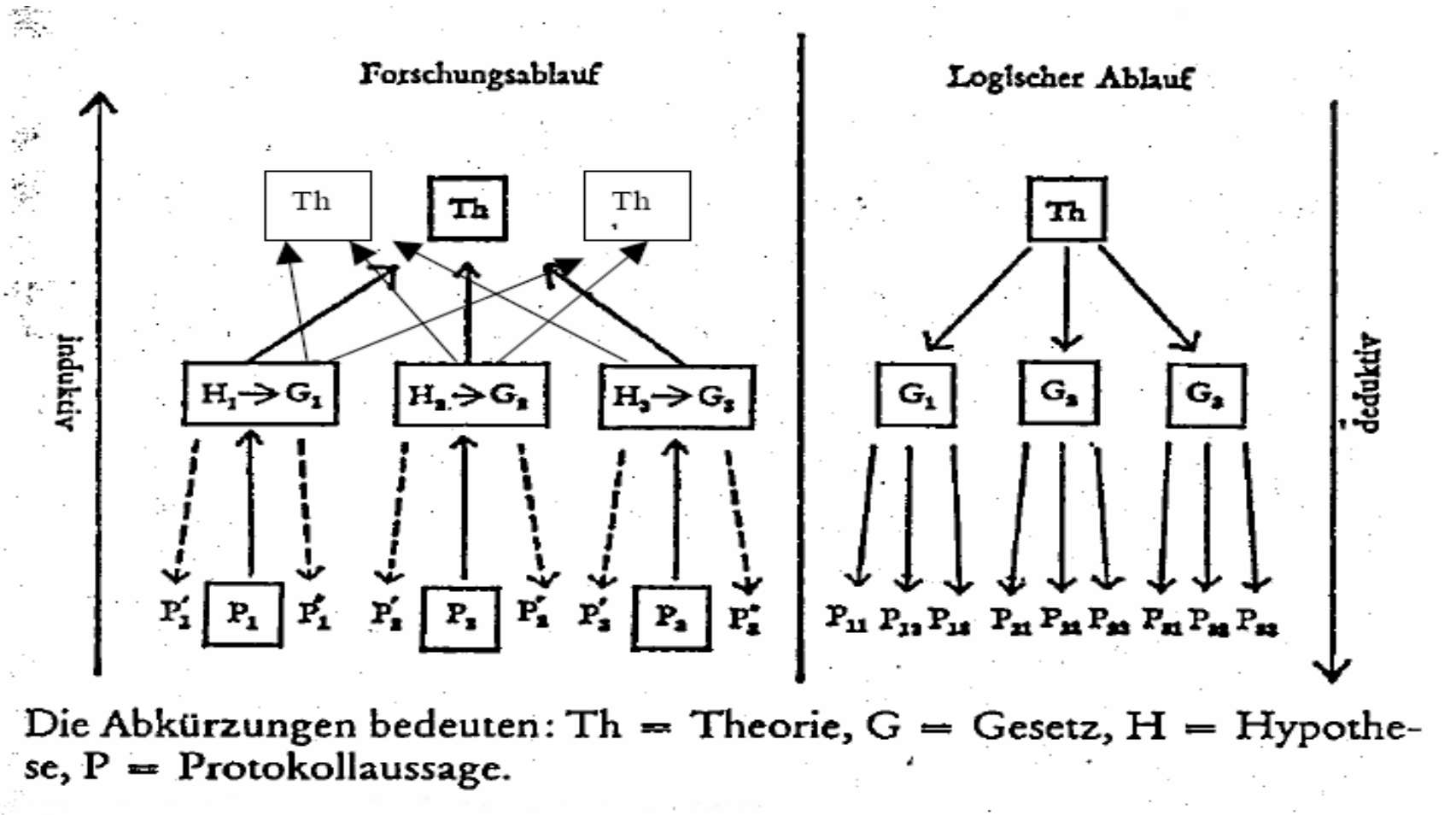
1.6 Epistemological constraints

- Data Mining is always based upon reduction and abstraction of reality.
- Only connections between gathered data can be found in the final result.
- Not all of the theories derived are necessarily correct.
- The selection of data influences the result.
- The goals influence the selection of the data.

1.6 Epistemological constraints



1.6 Epistemological constraints



Die Abkürzungen bedeuten: Th = Theorie, G = Gesetz, H = Hypothese, P = Protokollaussage.

2 Data Mining Methods

2.1 Non-supervised learning

2.2 Supervised learning

2.1 Non-supervised learning

- no training examples

  → segmentation and association

  – segmentation:

    search for global partition of segments of data

  – association:

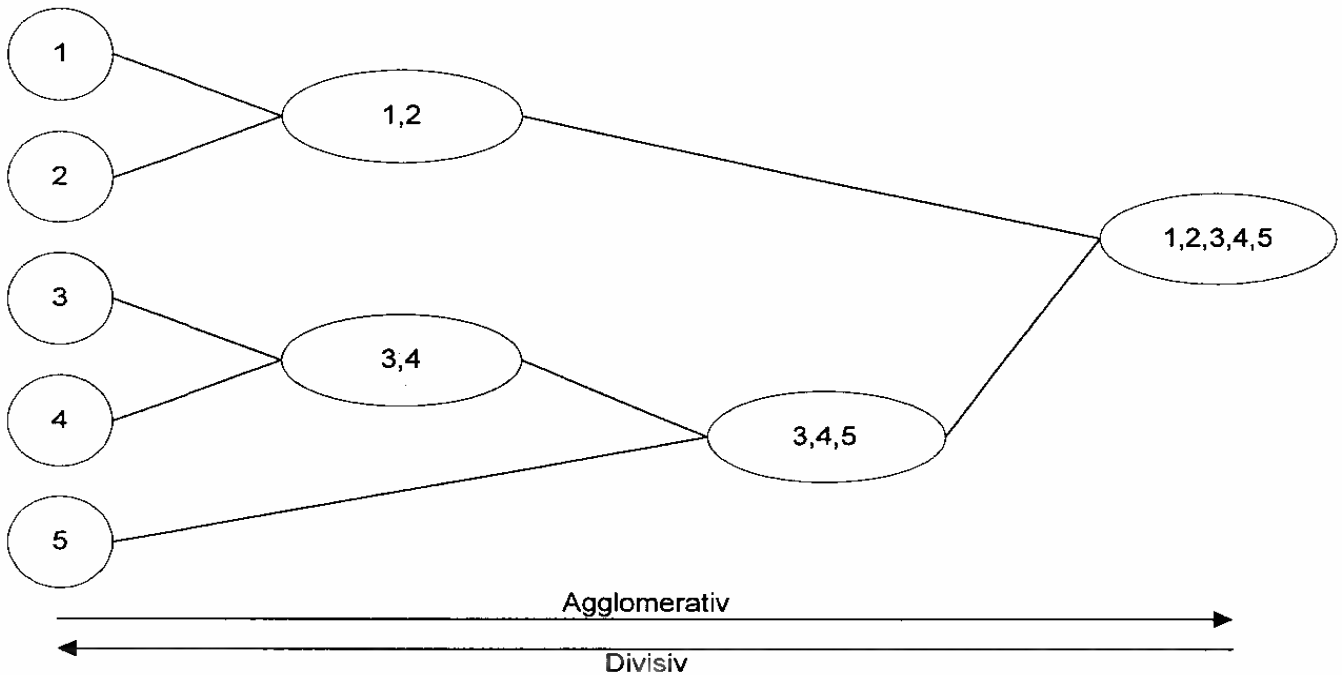
    search for relations between data

- methods

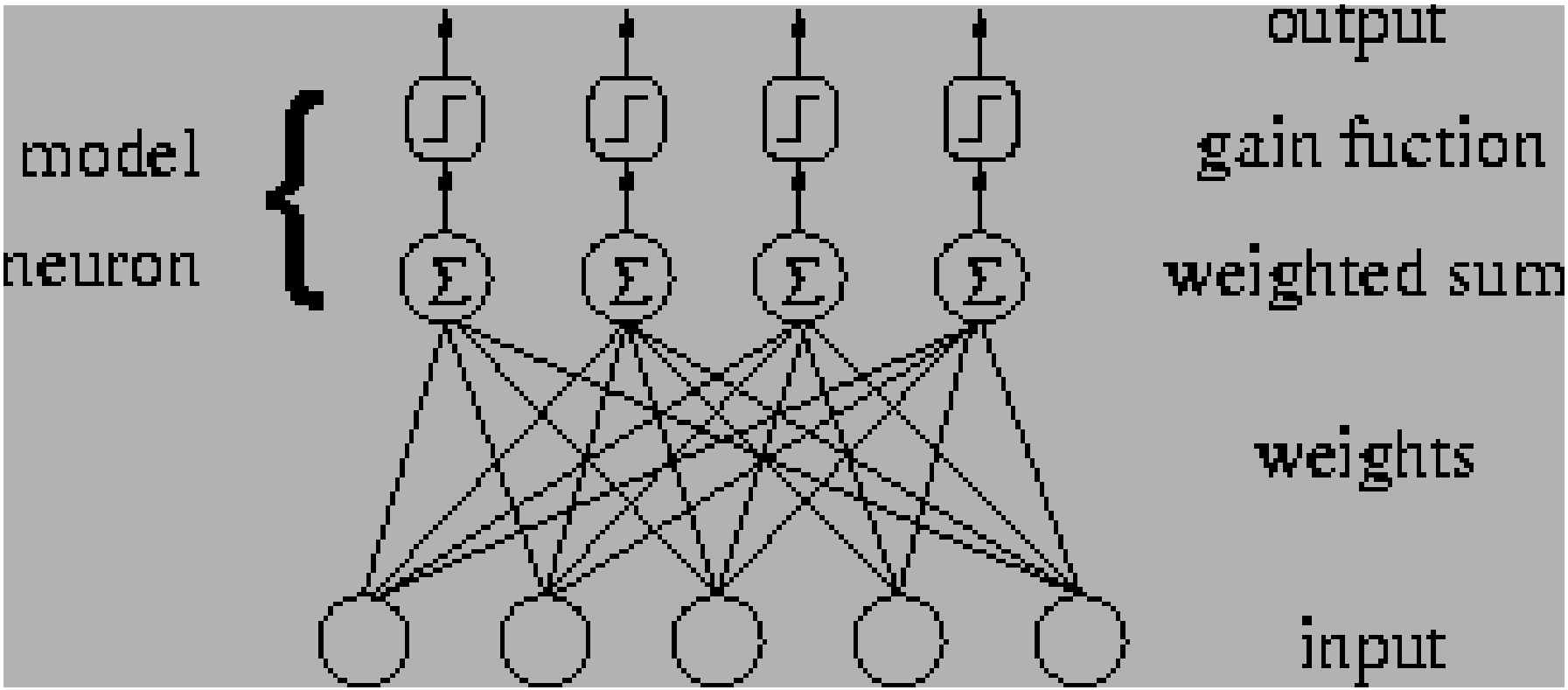
  – demographic, *k*-means, hierarchic clustering

  – neural networks

2.1.1 Cluster analysis



2.1.2 Neural networks



2.2 Supervised learning

- starts from training examples with known classifications
- learns the classification via training examples
- uses the classification learnt

- methods:
  - decision trees (ID3)
  - neural networks
  - rule induction
  - *k*-nearest neighbors

2.2.1 Decision trees

- Decision tree:
  Visualization of a classification rule
- Each node tests attributes.
- Decision is made when a leaf is reached.
- Leaf:
  node without any children

2.2.1 Decision trees

- The basis of the construction of a decision tree is training data.
- Training data records consists of examples containing several attributes each.
- One attribute is the target attribute.

2.2.2 ID3: Induction of Decision Trees

Example: insurance company

Training data

CustomerID Contract-term Occupation Cancellation

Kundennummer	Vertragsdauer	Berufsstatus	Kündigung
1	Mittel	Nicht-Erwerbstätige	Nein
2	Niedrig	Beamte	Ja
3	Niedrig	Angestellte	Ja
4	Mittel	Selbständige	Nein
5	Mittel	Angestellte	Nein
6	Hoch	nicht-Erwerbstätige	Nein
7	Mittel	Angestellte	Nein
8	Hoch	Beamte	Nein
9	Hoch	Selbständige	Ja
10	Mittel	nicht-Erwerbstätige	Nein

2.2.2 ID3

Entropy (S, Z) = $\sum_{z \in Z}$ - p(z) $\log_2$ p(z)

- S sample
- Z target attribute
- z target attribute value
- p(z) = $\frac{\#S_z}{\#S}$ probability that Z has value z

In the example:

Entropy (customers1-10, cancellation) =

-0,3 * $\log_2$ 0,3 – 0,7 * $\log_2$ 0,7 = 0,881

2.2.2 ID3

InfoGain (*S*, *Z*, *A*) =

Entropy (*S*, *Z*) - $\sum_{v \in A} \frac{\#S_v}{\#S}$ * Entropy (S_v, *Z*)

- *A* attribute whose InfoGain is calculated
- *v* index for all possible values of *A*
- S_v subset of *S* whose elements have the value *v* in *A*

InfoGain (customers1-10, cancellation, contract-term) =

Entropy (customers1-10, cancellation) –

$$\sum_{v \in Vertragsdauer} \frac{\#S_v}{\#S} \quad * \quad \text{Entropy } (S_v, \text{ cancellation})$$

2.2.2 ID3

For contract-term = short:
2/10 * Entropy (*Scontracttermshort*, cancellation) = 0

For contract-term = medium:
5/10 * Entropy (*Scontracttermmedium*, cancellation) = 0

For contract-term = long:
$3/10 * (- 1/3 * \log_2 1/3 – 2/3 * \log_2 2/3) = 0{,}275$

In the example, the InfoGain is:

InfoGain (customers1-10, contract-term) = 0,881 – 0,275 = 0,606

2.2.2 ID3

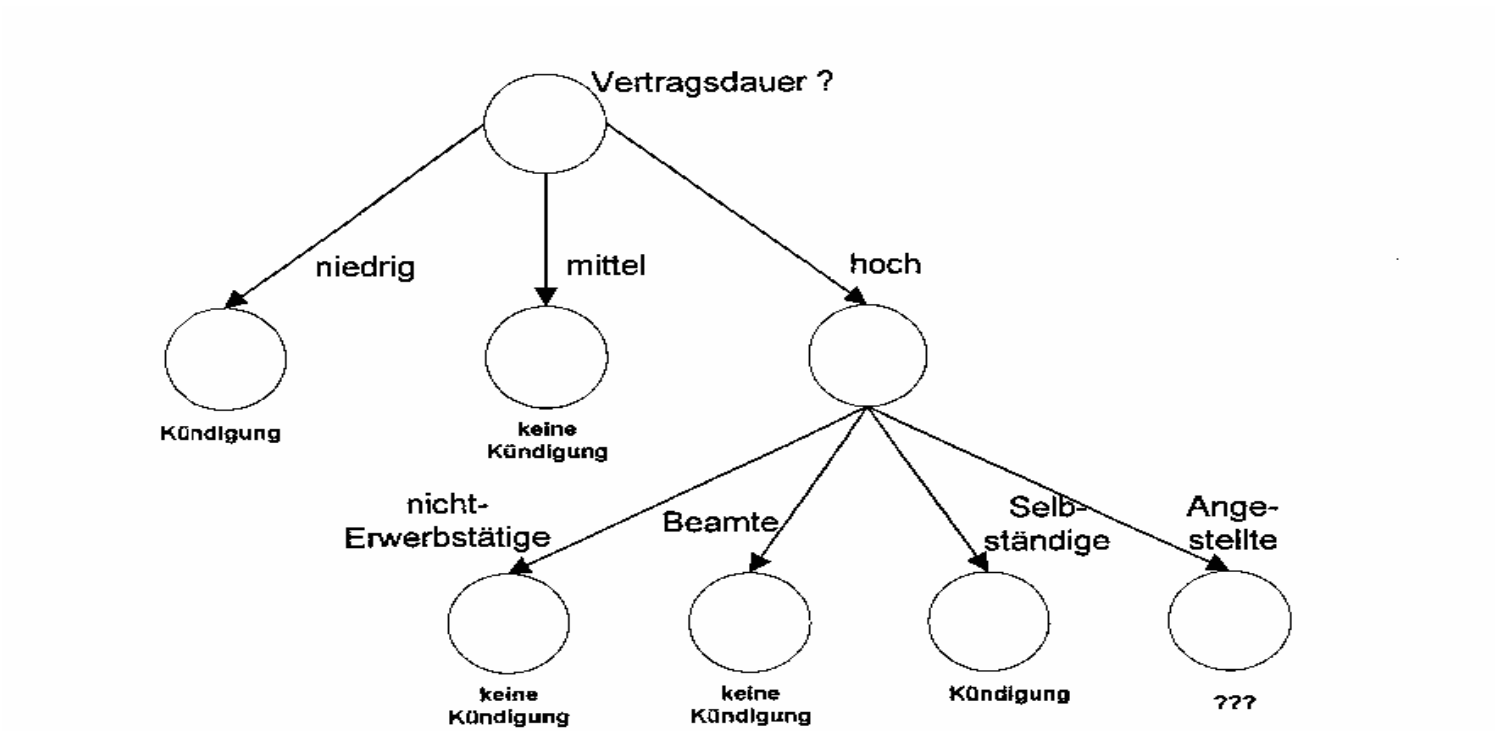
For occupation:

InfoGain (customers1-10, cancellation, occupation) =
0,881- (#*Sunemployed* / #*S* * Entropy (*Sunemployed*) +
#Sofficial / #S * Entropy (Sofficial) +
#Semployee / #S * Entropy (Semployee) +
#Sfreelance / #S * Entropy (Sfreelance)) =
3/10 * 0 +
2/10 * (-½ * $\log_2$ ½ - ½ * $\log_2$ ½) +
3/10 * (-⅓ * $\log_2$ ⅓ – 2/3 * $\log_2$ 2/3) +
2/10 * (-½ * $\log_2$ ½ - ½ * $\log_2$ ½) =
2/10 * 1 + 0,275 + 2/10 * 1 = 0,675

InfoGain (customers1-10, cancellation, occupation) = 0,881 – 0,675 = 0,206

2.2.2 ID3

InfoGain (customers1-10, cancellation, contract-term) >
InfoGain (customers1-10, cancellation, occupation)



2.2.2 ID3

New customers (no training data)

CustomerID Contract-term Occupation

Kundennummer	Vertragsdauer in Monaten	Berufsstatus
11	mittel	Beamte
12	niedrig	Selbständige

2.2.2 ID3

Decision rules:

- With a short-term contract, the customer is insecure.
- With a medium-term contract,
  the customer is not insecure.
- With a long-term contract and occupation unemployed,
  the customer is not insecure.
- With a long-term contract and occupation official,
  the customer is not insecure.
- With a long-term contract and occupation freelance,
  the customer is insecure.