Introduction

Contents of this Chapter

Basic concepts and relationship to other disciplines

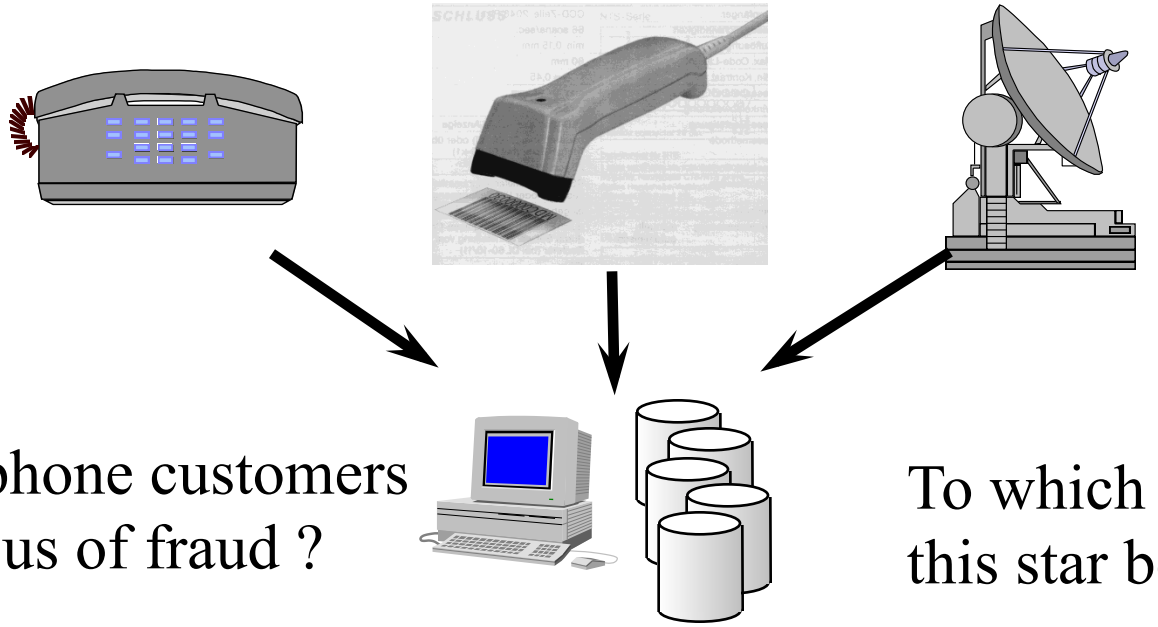
Knowledge discovery process

Overview of the course

References

Motivation

huge amounts of data are automatically collected



Which telephone customers are suspicious of fraud ?

To which class does this star belong?

Which associations exist between different products in a supermarket?

⇨ such an analysis can no longer be conducted manually

Definition KDD

Knowledge discovery in databases (*KDD*) is the process of (semi-)automatic extraction of knowledge from databases which is

- *valid,*
- *previously unknown,* and
- *potentially useful*. [Fayyad, Piatetsky-Shapiro & Smyth 96]

Remarks

- *(semi)-automatic*: different from manual analysis.
  Often, some user interaction is necessary.
- *valid*: in the statistical sense.
- *previsouly unknown*: not explicit, no „common sense knowledge“.
- *potentially useful*: for a given application.

Relationship to Other Disciplines

Contributions from Database Systems

- scalability for large datasets
- integration of data from different sources (data warehouses)
- novel datatypes (e.g. text and web data)

Contributions from Statistics

- probabilistic knowledge
- model-based inferences
- evaluation of knowledge

Contributions from Machine Learning

- different paradigms of learning
- supervised learning
- hypothesis spaces and search strategies

Relationship to Other Disciplines

Database Systems

+ discovery of implicit (not explicit) patterns
+ learning capabilities

Statistics

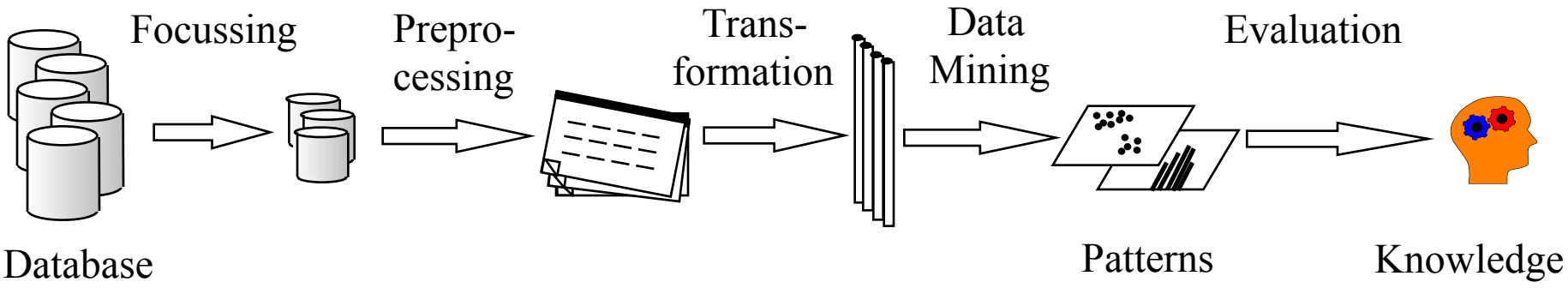
+ analysis of existing databases (not designed for task)
+ automatic generation of plausible hypotheses
+ efficient algorithms

Machine Learning

+ dealing with imperfect data
+ very large datasets
+ understandabality of knowledge

KDD Process

KDD Process Model [Fayyad, Piatetsky-Shapiro & Smyth 1996]



iterative and
interactive process

Focussing

Understanding the application

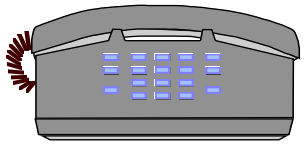
Ex.: make new telecommunication rates

Definition of the KDD goal

Ex.: customer segmentation

Data acquisition

Ex.: from operational billing DB



example application

Data management

file system or DBS?

Selection of relevant data

Ex.: 100‘000 important customers with all calls in 2002

Focussing

"File mining"

- Data typically in database systems (DBS)
- Data mining often on specially prepared files

Integration of data mining with DBS

- avoids redundancies and inconsistencies
- exploits DBS-capabilities (e.g. index structures)

Data mining primitives

- basic operations for a class of KDD algorithms or for some datatype
- efficient DBS - support
  - Faster development of new KDD methods
  - Better portability of algorithms

Preprocessing

Integration of data from different sources

- Simple conversion of attribute names (e.g. CNo → CustomerNumber)
- Use of domain knowledge for duplicate detection (e.g. spatial match based on ZIP codes)

Consistency check

- Test of application specific consistency constraints
- Resolution of inconsistencies

Completion

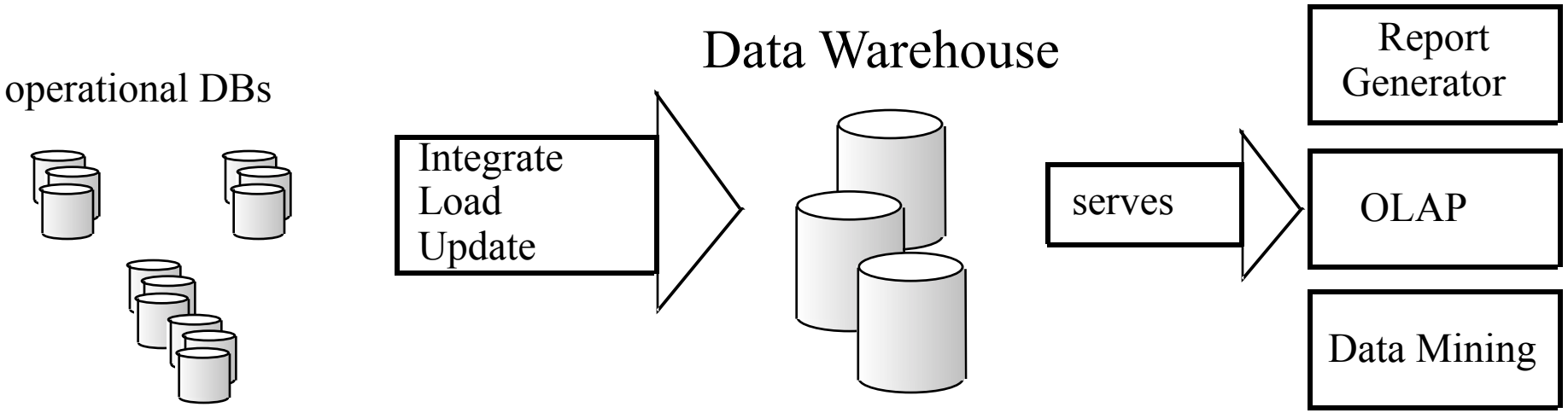
- Substitution of unknown attribute values by defaults
- Distribution of attribute values shall not be changed

preprocessing is often the most expensive KDD step

Preprocessing

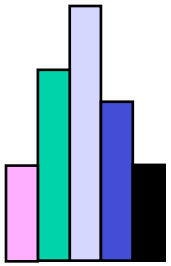
Data Warehouse [Chaudhuri & Dayal 1997]

- persistent
- integrated collection of data
- from different sources
- for the purpose of analysis or decision support



Transformation

Discretization of numeric attributes



- Independent from the data mining task

  Ex.: partitioning of the attribute domain in equal-length intervals

- Specific for the data mining task

  Ex.: partitioning in intervals such that the information gain w.r.t. class membership is maximized

Generation of derived attributes

- Aggregation over sets of data records

  Ex.: from single call records to

  „Total minutes daytime / evening, weekday / weekend“

- Combination of several attributes

  Ex.: revenue change = revenue 2000 - revenue 1999

Transformation

Selection of attributes

- *manual*
  if domain knowledge available on the attribute semantics and on the data mining task
- *automatic*
  bottom-up (starting from the empty set, add one attribute at a time)
  or top-down
  (starting from the set of all attributes, remove one attribute at a time)
  e.g. optimizing the discrimination between the different classes

⇒ too many attributes can lead to inefficient and ineffective data mining

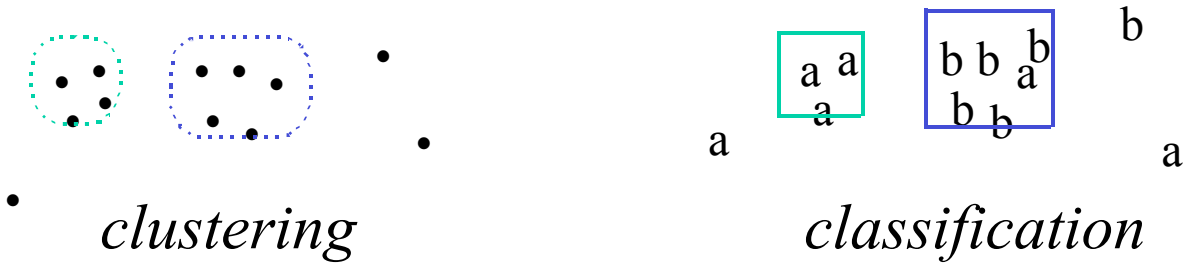
⇒ some transformations can be realized by OLAP-systems

Data Mining

Definition [Fayyad, Piatetsky-Shapiro, Smyth 96]

Data Mining is the application of efficient algorithms that determine the patterns contained in a database.

Data mining tasks



clustering *classification*

A and B → C

association rules

other tasks: regression, outlier detection . . .

Data Mining

Applications

- Clustering

  customer segmentation, structuring sets of web documents,

  determining protein families and superfamilies

- Classification

  automatic credit check, automatic interpretation of astronomical images,

  prediction of protein function

- Association rules

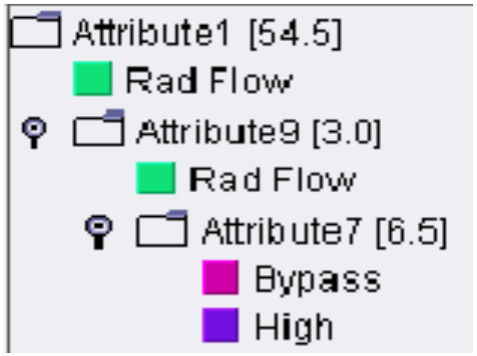
  redesign of supermarket layout, improving cross-selling,

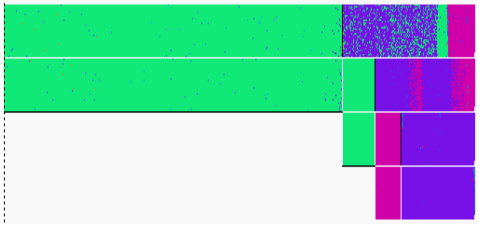
  improving the structure of a website

Evaluation

Procedure

- Presentation of discovered patterns supported by appropriate visualizations
- Evaluation of the patterns by the user
- If evaluation not satisfactory: repeat data mining with
  - Different parameters
  - Different methods
  - Different data





- If evaluation o.k.:

  Integration of discovered knowledge in the enterprise knowledge base

  Use of the new knowledge for future KDD-processes

Evaluation

Evaluation of discovered patterns

Interestingness

- Pattern already known?
- Pattern surprising?
- Pattern relevant for the application?

Predictive power

- How accurate is the pattern? (*accuracy*)
- For how many cases does the pattern apply? (*support*)
- How well does the pattern generalize to unseen cases?

Overview of the Course

- Prerequisites

  Basic Algorithms (algorithms, data structures, complexity, efficiency)

  Basic Statistics (means, standard deviation, probability, probability distributions, . . .)

- Goals of this course
  - Understanding of the main KDD concepts
  - Knowledge of the most important data mining tasks and methods
  - Selection and implementation of methods for a given application
  - Research on new data mining methods

- Focus on
  - Cluster analysis,
  - Classification,
  - Applications to genomics.

Overview of the Course

Outline

1. Introduction
2. Cluster Analysis

   types of data and distance functions, representative-based clustering, probabilistic model-based clustering, hierarchical clustering, density-based clustering, non-negative matrix factorization, consensus clustering, high-dimensional clustering, semi-supervised clustering
3. Classification

   classifier evaluation, decision trees, Naïve Bayes classifier, logistic regression, Bayesian networks, support vector machines, nearest neighbor classifier, ensemble methods, regression analysis
4. Applications to Genomics

   motif discovery, gene expression clustering / patient stratification, protein function prediction, reconstruction of biological networks

References

Textbook

- Aggarwal, C.: „*Data Mining: The Textbook*“, Springer, 2015.

Further recommended books

- Han J., Kamber M., Pei J.: „*Data Mining: Concepts and Techniques*“, Morgan Kaufmann Publishers, 3rd ed., 2011.
- Leskovec J., Rajaraman A., Ullman J.D.: „*Mining of Massive Datasets*“, Cambridge University Press, 2nd ed., 2014.

Research articles

will be provided in class

References

Other resources

- KDNuggets: a very comprehensive resource of KDD software, companies, publications and more.

  (http://www.kdnuggets.com/)

- ACM SIGKDD: ACM's special interest group on Knowledge Discovery in Databases.

  (http://www.acm.org/sigkdd/)

Open source data mining tools resources

- WEKA (Java): http://sourceforge.net/projects/weka/
- Orange (Python): http://orange.biolab.si/
- RapidMiner (data mining as a service):

  https://rapidminer.com/news-posts/rapidminerembracesitscommunity/