Knowledge Discovery in Databases

Erich Schubert, Michael Gertz

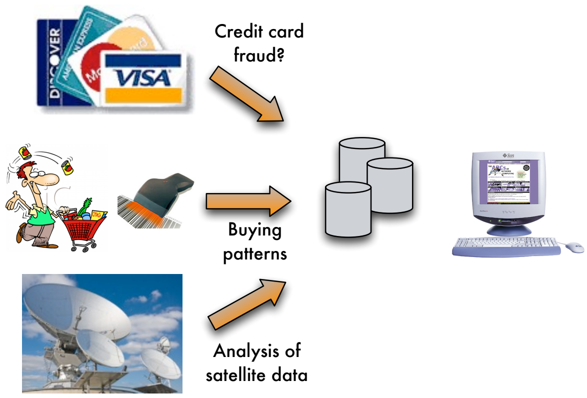
Heidelberg University
Institute of Computer Science
Database Systems Research Group

Winter Semester 2017/18

Part I

Introduction

Overview



Credit card fraud?

Buying patterns

Analysis of satellite data

Large amounts of data need to be analyzed ➡ Analysis and exploration of large amounts of data have to be supported by appropriate (automated) methods and systems

KDD: Commercial Sector

Collecting and storing large data sets

- product data, inventory, goods movement (RFIDs), vendor data
- sales transactions, credit card transactions
- customer surveys

Objectives of evaluating data

- optimizing processes
- improvement of services
- cost reduction
- increase in profit

KDD: Scientific Sector

Automated observations and capturing

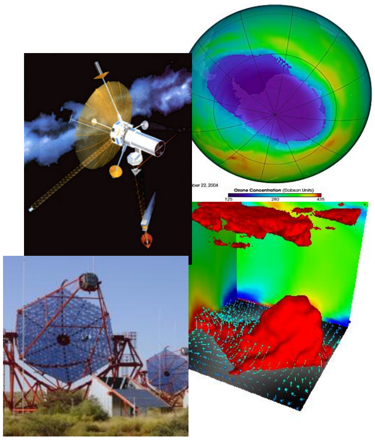
- telescopes
- simulations models (climate, earthquakes, . . .)
- microarrays in genetic research
- sensor networks (environmental data)

Large data sets are produced (GB/TB per hour)

- manual processing and evaluation almost impossible

Objectives of analysis

- classification / segmentation of the data
- formulating/verification of hypotheses

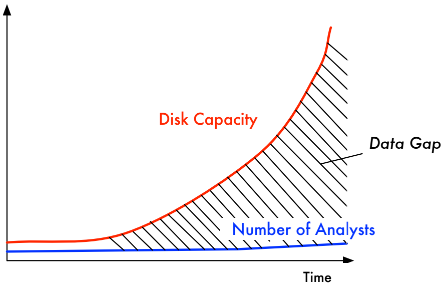


Analysis of Large Data Sets

Analysis of large data volumes that are managed in some data store (GB . . . PB of data)

Hidden or latent information: patterns, correlations, . . .

- cannot be identified manually
- because of the data volume often no analysis is possible at all



KDD: A First Definition

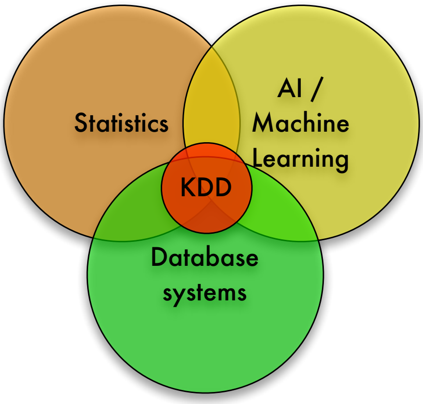
Fayyad, Piatetsky-Shapiro, and Smyth [FPS96]:

> Knowledge discovery in databases (KDD) is the process of (semi-) automatic extraction of knowledge from databases which is **valid**, **previously unknown**, and **potentially useful**.

Notes

- (semi-) automatic: in comparison to a manual analysis; but includes user interaction
- valid: in a statistical sense
- previously unknown: not explicitly known so far, no "general knowledge"
- potentially useful: for a given application or domain

KDD: Origins & Areas of Influence

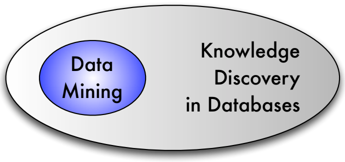


➨ New challenges: very large data volumes, high-dimensional data, unstructured data, heterogeneity, parallelization, distributed processing

KDD & Data Mining

- Basic idea of knowledge discovery: derive knowledge from data
- KDD is an iterative process in which hypotheses of the data mining step are verified and/or interpreted

- Data Mining
  - misnomer: we do not search data, but for knowledge
  - no verification of statistical assumptions
  - "autonomous" generation of hypotheses, e.g., in the form of a rule



- Flood of terms:
  - in the commercial field: Data Mining = KDD
  - also: Knowledge Mining, Knowledge Extraction, Data Dredging, Data Science, …

Data Mining: Further Definitions

> “ *Data mining is the process of discovering meaningful new* ***correlations, patterns and trends*** *by sifting through large amounts of data stored in repositories, using* ***pattern recognition*** *technologies as well as* ***statistical and mathematical techniques****.* ” — Gartner Group

> “ *Data mining is the exploration and analysis, by automatic and semiautomatic means, of large quantities of data in order to discover* ***meaningful patterns and rules****.* ” — M.J.A. Berry, G. Linoff

> “ *Data mining is the analysis of (often large) observational data sets to find* ***unsuspected relationships*** *and to summarize the data in* ***novel ways*** *that are both* ***understandable and useful*** *to the data owner.* ” — D. Hand, H. Mannila, P. Smyth

> “ *The automated extraction of* ***predictive*** *information from large databases.* ” — K. Thearling

KDD: Delineation

KDD tasks

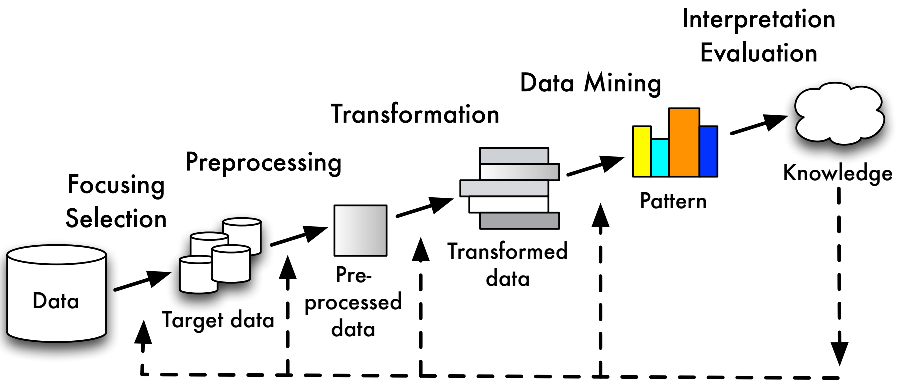
- determine products that are frequently bought together
- determine criteria for creditworthiness of customers
- determine stars or galaxies that have similar features
- determine unusual behavior of users in a social network
- …

Non-KDD tasks

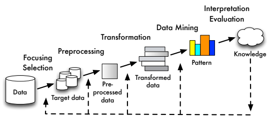
- data collection, e.g., web scraping
- web search (also: search for documents on the PC, in the intranet, …)
- sales figures of a particular product last month
- average age of customers buying product X
- …

KDD Process: Overview

Fayyad, Piatetsky-Shapiro, and Smyth [FPS96] proposed:



KDD Process: Overview /2

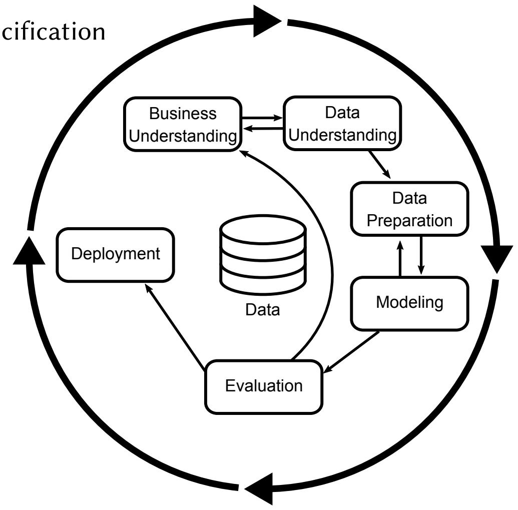


Iterative process

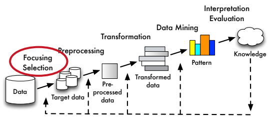
- **Focusing** and **Selection**: choosing data from data source(s)
- **Preprocessing**: search for and remove glitches such as data errors or incomplete data
- **Transformation**: reduce quantity
  - remove attributes that occur rarely in the data
  - transform data into format appropriate for analysis
- **Data Mining**: analyze the data
  The exact method used depends on the task we are trying to solve.
- **Interpretation** and **Evaluation**
  - Validity of results
  - How to use this knowledge
- Iterate as necessary.

KDD Process: Overview /3

CRISP-DM: Industry Specification



KDD Process: Focusing



“File Mining”

- data typically reside in database systems (DBS)
- Data Mining approaches are typically applied to pre-arranged files

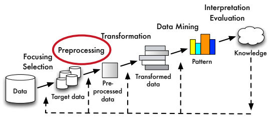
Integration of data mining with DBS [IM96; AS96; Fay+95]

- avoid redundancies and inconsistencies
- employ functionality of DBS, e.g., index structures

Base operations in support of data mining

- standard operations for a class of KDD algorithms
- efficient support by DBS
- rapid development of new KDD algorithms
- improved portability of algorithms

KDD Process: Preprocessing



Integration of data from different sources

- simple mapping of attribute names, e.g., CustomerKey ⟶ CustID
- Employ domain knowledge in order to merge similar data, e.g., regional attribution of zipcodes

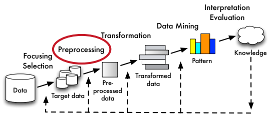
Consistency checks

- verify domain specific consistency constraints
- removal of inconsistencies

Completion

- replacing unknown attribute values by default values
- distribution of attribute values should be preserved

KDD Process: Preprocessing /2



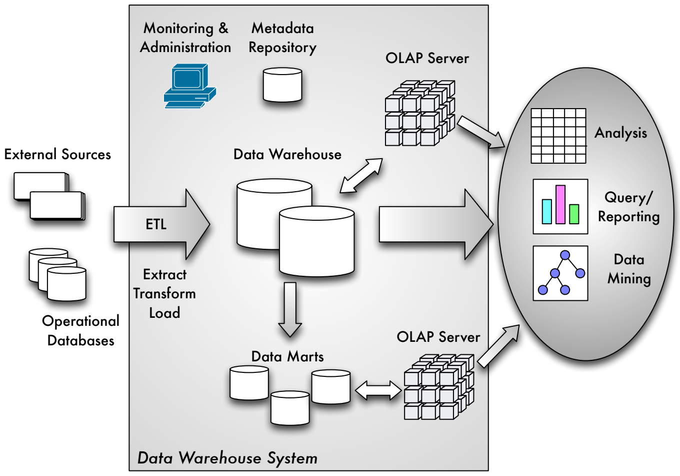
Preprocessing is typically the most complex and time-consuming step

Preprocessing is often executed as part of a Data Warehouse / OLAP system

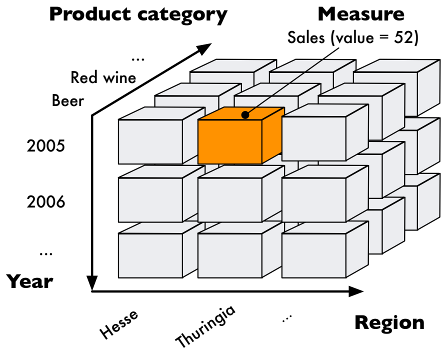
Data Warehouse :=

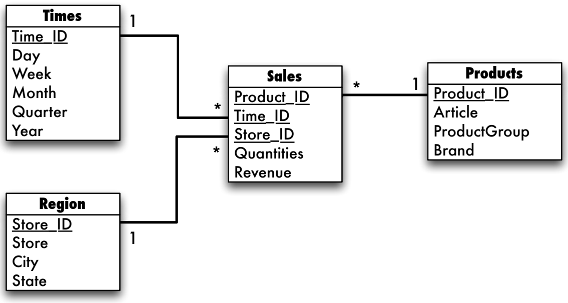
- **non-volatile**,
- **integrated** collection of data
- from **different** sources
- in support of **data analysis** and decision making

Data Warehousing: Process



Data Warehousing: Data Model





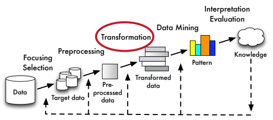
Times
Time_ID
Day
Week
Month
Quarter
Year

Region
Store_ID
Store
City
State

Sales
Product_ID
Time_ID
Store_ID
Quantities
Revenue

Products
Product_ID
Article
ProductGroup
Brand

KDD Process: Transformation



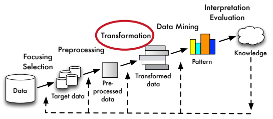
Discretize numeric attributes

- independent of the data mining task, e.g., partitioning of a value range into equi-length intervals
- dependent on the data mining task, e.g., partitioning into intervals such that the information gain is maximized with respect to class membership

Generating derived attributes

- by aggregating sets of data records (data values), e.g., from single sales transactions to daily sales, weekly sales, monthly revenue etc.
- by combining multiple attributes, e.g., change in revenue (revenue_change = revenue_2014 − revenue_2013)

KDD Process: Transformation /2



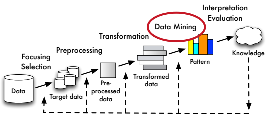
Selection of attributes

- manually:
  - if there is domain knowledge about the role of the attribute and the given data mining task
- automatically:
  - Bottom-up (starting from the empty set, expand set iteratively by "most relevant" attribute)
  - Top-down (starting from all attributes, iteratively remove attribute such that, e.g., discrimination of classes is optimized)

Problem

- too many attributes may lead to inefficient and possibly ineffective (∼ poor quality)
- only some transformations can be realized by OLAP systems (➡ use file mining?)

KDD Process: Data Mining



Data Mining is considered the application of **efficient algorithms** that determine patterns from the data in a database

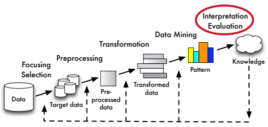
Predictive approaches:

- using features (variables) of objects for predicting unknown or future values of features of other objects
- example: classification, regression

Descriptive approaches:

- extraction of human-interpretable patterns that describe the data
- example: association rules, clustering, outlier detection

KDD Process: Evaluation



Presentation of patterns:
often done with the help of appropriate visualizations

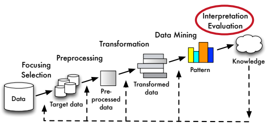
In case of a poor rating/evaluation (done by the user): repeat data mining step with

- other parameters, other approach, other data

In case of a good rating/evaluation (done by the user)

- integration of the found knowledge into the knowledge base
- use this new knowledge for future KDD processes (learning)

KDD Process: Evaluation /2



Evaluating the patterns that have been determined:

Patterns' utility for prediction

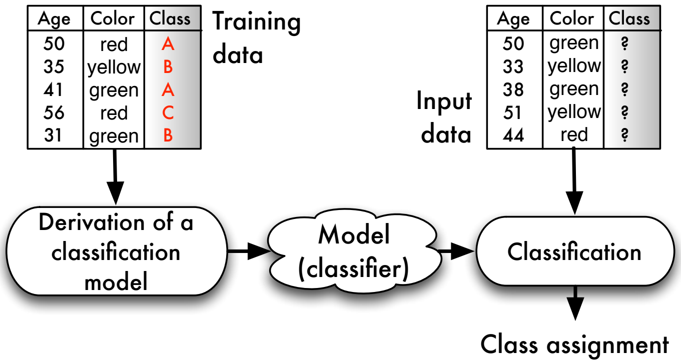
- data used for mining are a sample from the basic population of all data
- how well can the patterns found in the "training data" be generalized to future/other data?
- quality of patterns for prediction grows with the size and representativeness of the data

Interestingness of a pattern

- Is the pattern already known?
- Is the pattern surprising?
- Can the pattern be applied for many cases?

Classification: Example

- Assigning objects to classes, that is, prediction of features (class assignment, class label) based on some other features/characteristics of objects
- Derivation of a **classification model (classifier)** from the training data



Training data

Age	Color	Class
50	red	A
35	yellow	B
41	green	A
56	red	C
31	green	B

Input data

Age	Color	Class
50	green	?
33	yellow	?
38	green	?
51	yellow	?
44	red	?

Derivation of a classification model

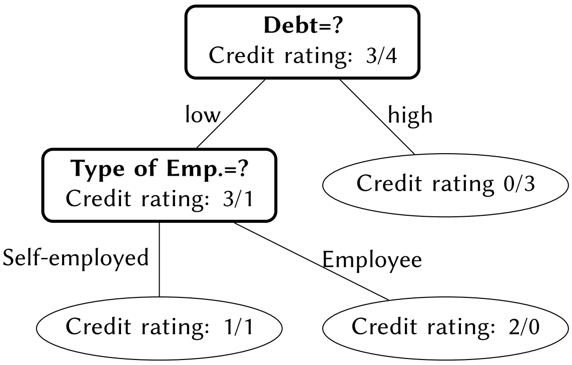
Model (classifier)

Classification

Class assignment

Classification: Example /2

CustomerID	Debt	Income	Type of employment	Credit rating
1	High	High	Self-employed	poor
2	High	High	Employee	poor
3	High	Low	Employee	poor
4	Low	Low	Employee	good
5	Low	Low	Self-employed	poor
6	Low	High	Self-employed	good
7	Low	High	Employee	good



Debt=?
Credit rating: 3/4

low

high

Type of Emp.=?
Credit rating: 3/1

Credit rating 0/3

Self-employed

Employee

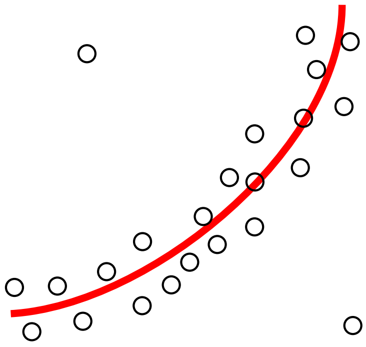
Credit rating: 1/1

Credit rating: 2/0

Regression: Example

Regression aims at

- detecting a trend in a data set (linear, piecewise linear, non-linear, . . .)
- predicting one variable from the others



Association Rules: Example

- Discovery of (significant) statistical correlations between variables
- Model: association rules or “frequent itemsets”
- Example: market basket analysis

Transaction-ID	Products
1	milk, butter
2	milk, honey, butter
3	milk, bread, butter
4	milk, bread, honey
5	diapers

Association Rules: Example /2

Frequent Itemsets:

Products	Support
{ milk }	4
{ butter }, { milk, butter }	3

{ honey }, { bread }, { honey, bread }
{ honey, milk }, { honey, butter }
{ bread, milk }, { bread, butter }
	2

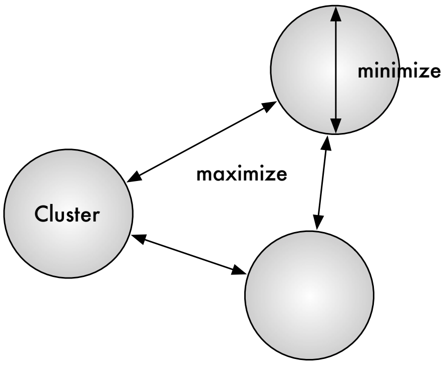
Rule generation: $\{\text{milk}\} \rightarrow \{\text{butter}\}$ = **Customers who buy milk also buy butter.**

- Support (number of transactions containing that set): 3
- Confidence (percentage of “milk” transactions that also contain “butter”; here 3 out of 4): 75%

Clustering: Example

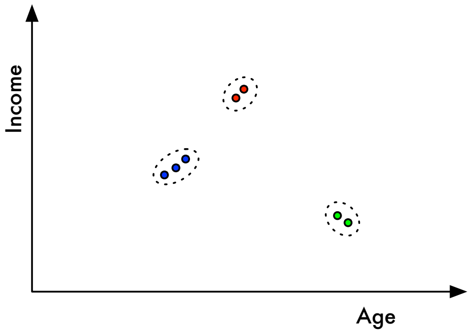
Grouping similar objects into new groups (clusters) such that

1. similarity between objects **within** a group is **high**, and
2. similarity between objects from **different** groups is **low**



Clustering: Example /2

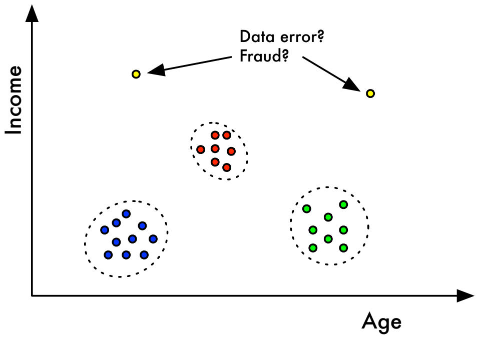
Age	Income
25	50.000
27	55.000
26	58.000
40	85.500
42	90.000
57	38.000
59	40.000



Outlier Detection: Example

Outlier detection (aka anomaly detection) aims at

- determining data that are untypical, or
- dissimilar to all other objects



Sequence Analysis

Also known as **Sequential Pattern Discovery**
Find frequent episodes or event sequences in data set that has (temporal) order

- set of event types E, sequence of pairs (e, t) with $e \in E$ and t is a timestamp
- episode α: partial order of event types
- frequency of an episode α: number of partitions of the sequence of a given length containing the event types in α and their order

Applications

- analysis of alert messages (events) in telecommunication systems
- Web usage mining / clickstream analysis
- (temporal) buying behavior, e.g.,
  "Lord of the Rings (DVD)" ⟶ "Lord of the Rings (book)" ⟶ "Silmarillion (book/collection)"
  If a customer bought "Silmarillion", he probably already bought the other somewhere else!

Aspects of Data Security and Data Privacy

- Danger of misuse of data mining techniques
- For example, in cases where personal data is collected and analyzed without the consent and/or knowledge of the persons
- Aspects of data privacy in the context of KDD
- Examples
  - Monitoring telecommunication (Echelon, PRISM, Tempora, GCHQ, NSA, BND)
  - “Election campaign software” of US parties with about 160 million data records (Demzilla (Democrats), Voter Vault (Republicans))

Aspects of Data Security and Data Privacy /2

Example: Amazon.com

- “Purchase Circles are highly specialized bestseller lists. They let you know what people are buying around the world and in your hometown, at your workplace and at your alma mater.”
- collect personal data, allows for verification and correction
- anonymization by at least 200 customers

Example: IMS Health

- collects information about all prescriptions converted in pharmacies since 1969 (see most recent cases in Germany!)
- identification of drugs and physician
- classification in terms of geographic aspects, field of specialization, pharmaceutical companies

Summary

Basics of KDD process

- motivation
- several phases and iterative approach

Use cases

- analysis of empirically collected data
- management analysis, analysis of scientific data, network data (Intrusion Detection, Web Mining, Web Log Mining), …
- but also text analysis/mining (text classification), image analysis (image clustering), …

Data security and privacy

- important aspect when analyzing personal data (just check the news …)

Bibliography I

[AS96] Rakesh Agrawal and Kyuseok Shim. “Developing Tightly-Coupled Data Mining Applications on a Relational Database System”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. 1996, pp. 287–290. URL: `http://www.aaai.org/Library/KDD/1996/kdd96-049.php`.

[Fay+95] Usama M. Fayyad, Padhraic Smyth, Nicholas Weir, and S. George Djorgovski. “Automated Analysis and Exploration of Image Databases: Results, Progress, and Challenges”. In: *J. Intell. Inf. Syst.* 4.1 (1995), pp. 7–25. URL: `https://doi.org/10.1007/BF00962819`.

[FPS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “Knowledge Discovery and Data Mining: Towards a Unifying Framework”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. 1996, pp. 82–88. URL: `http://www.aaai.org/Library/KDD/1996/kdd96-014.php`.

[IM96] Tomasz Imielinski and Heikki Mannila. “A Database Perspective on Knowledge Discovery”. In: *Commun. ACM* 39.11 (1996), pp. 58–64. URL: `http://doi.acm.org/10.1145/240455.240472`.