# CMPT 740 – Database Systems
# Foundations of Data Mining

*Martin Ester*

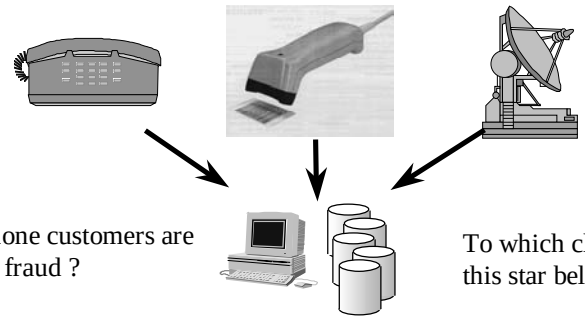
Simon Fraser University
School of Computing Science
03-3

## 1. Introduction

*Contents of this Chapter*

# 1. 1 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



# 1.1 Definition KDD

[Fayyad, Piatetsky-Shapiro & Smyth 96]

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

- *valid*
- *previously unknown*
- and *potentially useful*.

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.

## 1.1 Relationship to other disciplines

Database Systems
+ discovery of implicit patterns
+ learning capabilities

Statistics
+ automatic generation of plausible hypotheses
+ efficient algorithms

Machine Learning
+ dealing with imperfect data
+ very large datasets
+ understandability of knowledge



## 1.1 Relationship to other disciplines

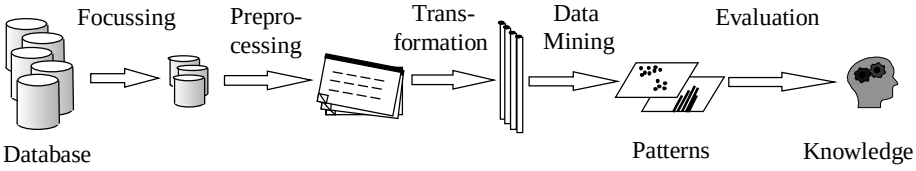
Database Systems [Han & Kamber 2000]
• scalability for large datasets
• integration of data from different sources (data warehouses)
• novel datatypes (e.g. text and web data)

Statistics [Hand, Mannila & Smyth 2001]
• probabilistic knowledge
• model-based inferences
• evaluation of knowledge

Machine Learning [Mitchell 1997]
• different paradigms of learning
• supervised learning
• hypothesis spaces and search strategies

## 1.2 KDD process

KDD process model [Fayyad, Piatetsky-Shapiro & Smyth 1996]




*iterative* and
*interactive* process



## 1.2 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

## 1.2 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



## 1.2 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

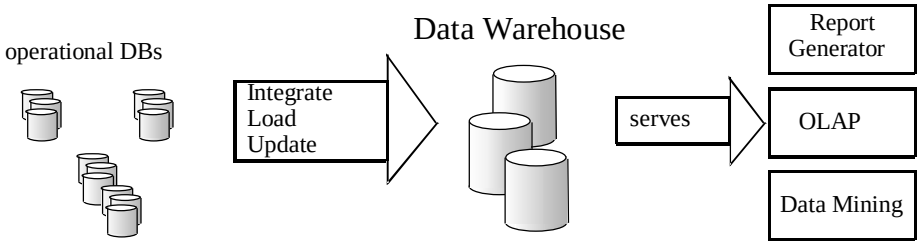
Completion
- Substitution of unknown attribute values by defaults
- Distribution of attribute values shall, in general, be satisfied.

preprocessing is often the most expensive KDD step

## 1.2 Preprocessing

Data Warehouse [Chaudhuri & Dayal 1997]

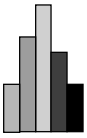
- persistent
- integrated collection of data
- from different sources
- for the purpose of analysis or decision support






## 1.2 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, weekday, local calls“
- Combination of several attributes

  Ex.: revenue change = revenue 2000 - revenue 1999

## 1.2 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)
  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 lead to inefficient and ineffective data mining

⇨ some transformations can be realized by OLAP-systems



## 1.2 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 . . .

# 1.2 Data Mining

## *Applications*

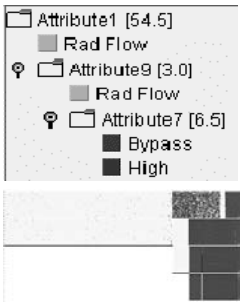
- clustering
  customer segmentation, structuring sets of web documents,
  determining protein families and superfamilies
- classification
  running a credit check, automatic interpretation of astronomical images,
  prediction of protein function
- association rules
  redesign of supermarket layout, improveing cross-selling,
  improving the structure of a website



# 1.2 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

## 1.2 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?



## 1.3 Typical KDD Applications

*Astronomy*

*SKICAT* System [Fayyad, Haussler & Stolorz 1996]

- Architecture


- Classification method: decision tree classifier
- Evaluation
  - Much faster than manual classification
  - Classifies also very faint celestial objects

# 1.3 Typical KDD Applications

## *Marketing*

Customer segmentation [Piatetsky-Shapiro, Gallant & Pyle 2000]

- Goal
  partitioning of the customers into segments with similar purchasing behavior
- Purpose
  - Ideas for product packages (Product Bundling)
  - Design of a new pricing policy (Pricing)
- Project
  - Development of an automatic model (Bayesian Clustering)
    too complex, no consideration of domain knowledge
  - Manual development of a decision list
    based on the insights gained
  - Application of the insights in the company‘s marketing
  - Integration of decision list in the existing software system



# 1.3 Typical KDD Applications

## *Electronic Commerce*

Electronic Awarding of Credit Cards [Himelstein, Hof & Kunii 1999]

- Traditionally
  - Manual credit checks
  - Requires access to several databases
  - Takes up to several weeks
- with data mining
  - Analys of a new customer in a few seconds
  - Much prompter service
  - Allows a credit company to attract numerous new customers
- data mining method: decision tree classifier

# 1.3 Typical KDD Applications

Purchase Circles [Beck 1999]

- Idea


- Method
  - Grouping purchases w.r.t. ZIP code and domain
  - Aggregation of these data
  - Construction of bestseller lists which are significantly more popular in a given customer group than in the set of all customers



# 1.4 Case Study: prediction of outer membrane proteins

*Proteins*

1D Structure

- chains of amino-acids: AVFAMLCNFQDMAQSWKKKAVFAAGDE . . .
- 20 different amino-acids (one / three letter codes)
- typical length of proteins: 3 to 400 amino-acids

Physico-chemical properties

- hydrophic / hydrophile
- charged / uncharged
- polar / non-polar

same properties imply similarity of proteins

| Amino-acid | Three-Letter Code | One-Letter Code | Physico-chemical Properties |
|---|---|---|---|
| Alanine | Ala | A | Hydrophobic |
| Lusine | Lys | K | Charged |
| Glutamine | Gln | Q | Polar |
| . . . | . . . | . . . | . . . |

## 1.4 Case Study: prediction of outer membrane proteins

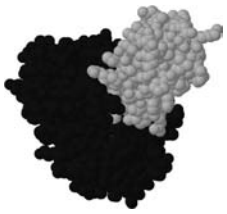
### *Proteins*

2D Structure

- subsequences of the 1D structure form 2D structures such as sheets, strands, . . .

3D Structure

- coordinates of the atoms in 3D space
- known only for small subset of all sequenced proteins
- protein surface important for many biological processes



Protein-Protein Docking



## 1.4 Case Study: prediction of outer membrane proteins

### *Protein Data*

- sequences from a 20-letter alphabet

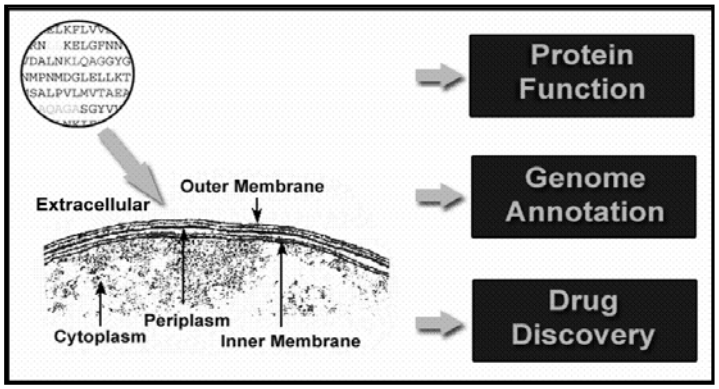
  AVFAMLCNFQDMAQSWKKKAVFAAGDE

- 3D structure
- textual annotations
- databases such as Swiss-Prot (120 607 entries)

sequences are (more or less) automatically determined
analysis requires data mining methods

## 1.4 Case Study: prediction of outer membrane proteins

*The Problem* [She, Chen, Wang, Ester, Gilardy & Brinkman 2003]




*Gram-negative bacterial cell*



## 1.4 Case Study: prediction of outer membrane proteins

*The Problem*

Importance of outer membrane proteins

- many gram-negative bacteria are pathogens
- outer membrane proteins (OMPs) are drug / vaccine targets

KDD goals

classification of protein localization based on sequence data with
(1) highly confident identification of OMPs
(precision > 90%)
(2) good recall of OMPs (> 50%)

fast prioritization of proteins for further (expensive!) analysis
(literature search / wet lab experiments)

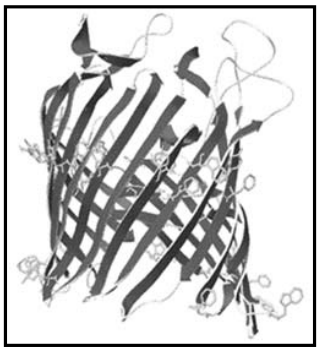
## 1.4 Case Study: prediction of outer membrane proteins

*Domain Knowledge*

Hypotheses

(1) Information determining the protein localization is primarily coded in the amino-acid sequence.

(2) β-barrels (anti-parallel β-strands) are indicators for OMPs .





## 1.4 Case Study: prediction of outer membrane proteins

*The Approach*

Frequent-pattern-based data mining

- determine frequent patterns in the dataset
- use these patterns as features for the classification task

Biological motivation

- subsequences crucial for the function are very likely conserved
- large part of sequences are noise

→ search for *frequent* patterns

→ frequent patterns provide biological insights

## 1.4 Case Study: prediction of outer membrane proteins

### *Data Mining Methods*

State-of-the art method in molecular biology
- Hidden Markov Model (HMM)
- requires large training effort
- precision / recall do not meet our requirements

New frequent-pattern-based methods
- association rules
  - use frequent *confident* subsequences as features
  - select subset of features for classifier (greedy method)
- SVM (Support Vector Machine)
  - use *all* frequent subsequences as features
  - SVM chooses / combines *relevant* features



## 1.4 Case Study: prediction of outer membrane proteins

### *Experimental Evaluation*

| Data | Number of Sequences | Percentage | Min. Length | Max. Length | Avg. Length |
|---|---|---|---|---|---|
| OMP | 427 | 27.4% | 91 | 3705 | 571.1 |
| Non-OMP | 1132 | 72.6% | 50 | 1034 | 256.8 |
| Total | 1559 | | | | 342.9 |

→ experimentally verified localisations

| | Actual OMP | Actual Non-OMP |
|---|---|---|
| classified as OMP | TP (true positive) | FP (false positive) |
| classified as Non-OMP | FN (false negative) | TN (true negative) |

**Precision** w.r.t. OMP = TP / (TP+FP) **Recall** w.r.t. OMP = TP / (TP+FN)

## 1.4 Case Study: prediction of outer membrane proteins

*Experimental Evaluation*

Classification Quality of SVM

| MinSup (%) | 0.8 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Precision (%) | 97 | **98** | 95 | 94 | 92 | 92 | 91 | 92 |
| Recall (%) | 79 | **81** | 82 | 82 | 83 | 82 | 83 | 82 |
| Number of Features | 115028 | 53879 | 14058 | 6611 | 5042 | 4252 | 3561 | 3111 |

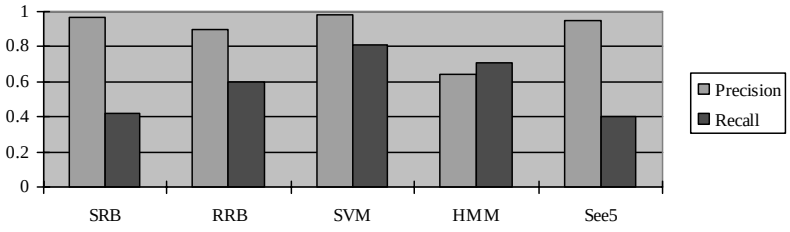
all *frequent* patterns are good compromise for feature selection



## 1.4 Case Study: prediction of outer membrane proteins

*Experimental Evaluation*

Comparison of the different data mining methods




SVM clearly outperforms all other methods w.r.t. precision / recall
RRB (association rules) is second best method and better understandable

## 1.5 Overview of the Course

*Goals*

- Prerequisites

  Basic Database Systems (relational data model, SQL, efficient implementation)

  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
  - Development of new data mining methods
- Focus on the original KDD challenges
  - scalability for large datasets
  - understandability of knowledge
  - novel tasks and datatypes



## 1.5 Overview of the Course

*Outline* (1)

1. Introduction
2. Principles of Data Mining

   Learning from training data, probabilistic knowledge,

   hypothesis search space and search methods
3. Data Preprocessing

   data cleaning, data integration and transformation,

   basic data reduction, discretization
4. Association Rules and Frequent Pattern Analysis

   basic concepts, frequent pattern mining methods,

   mining various kinds of frequent patterns

## 1.5 Overview of the Course

*Outline* (2)

5. Classification and Regression
   basic concepts, classifier accuracy, Bayesian classification, decision tree induction, neural networks, support vector machines, k-nearest neighbor classifier, regression
6. Cluster and Outlier Analysis
   basic concepts, types of data and distance functions, partition-based clustering, hierarchical clustering, density-based clustering, the EM algorithm, outlier analysis
7. Mining Biological Data
8. Mining Text and Web Data
9. Case Study
10. Summary and Outlook



## 1.6 Literature

Textbook

- Han J., Kamber M., „*Data Mining: Concepts and Techniques*“, Morgan Kaufmann Publishers, August 2000.

Further books

- Hand D.J., Mannila H., Smyth P., „Principles of Data Mining“, MIT Press, 2001.
- Mitchell T. M., „*Machine Learning*“, McGraw-Hill, 1997.
- Witten I. H., Frank E., „*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*“, Morgan Kaufmann Publishers, 2000.

Research articles

will be referenced in class

# 1.6 Literature

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/)

- DBLP Bibliography: very comprehensive resource of CS

  (in particular DB) articles

  (http://www.informatik.uni-trier.de/~ley/db/index.html)

- SFU‘s database and data mining lab: people, projects, library, events

  http://www.cs.sfu.ca/~ddm/