DATABASE SYSTEMS GROUP

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

LMU

Knowledge Discovery in Databases

SS 2016

Chapter 1: Introduction

Lecture: Prof. Dr. Thomas Seidl

Tutorials: Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

What is new?

- Professor
  - Thomas Seidl
  - Short bio

- Master Program „Data Science“ (as of WS 2016/17)
  - Funded by program „Elitestudiengang Bayern“
  - Apply until end of May

Schedule and People

- Weekly Schedule
  - Lecture (begins: Apr. 12th):
    - Tuesday, 09:30 – 12:00 h, Raum B 138 (Theresienstr. 39)
  - Tutorials (begins: Apr. 20th):
    - Wednesday, 14:15 – 15:45 h, Raum S 007 (Schellingstr. 3)
    - Thursday, 14:15 – 15:45 h, Raum B 106 (Hauptgebäude)
    - Friday, 14:15 – 15:45 h, Raum A 015 (Hauptgebäude)
  - Exam: tba

- People
  - Prof. Dr. Thomas Seidl
  - Julian Busch, Evgeniy Faerman, Florian Richter, Klaus Schmid

Credits, Material, Tutorial

- Material (Slides, Exercises, etc.) available on:
  - Course Webpage:
    http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)_16
  - Tutorial:
    - First exercise sheet available for download around April 13th
    - Prepare at home
    - Presentation and discussion one week after
  - Exam:
    - Written exam at the end of semester
    - All material discussed in the lecture and tutorials
    - Registration via UniWorX

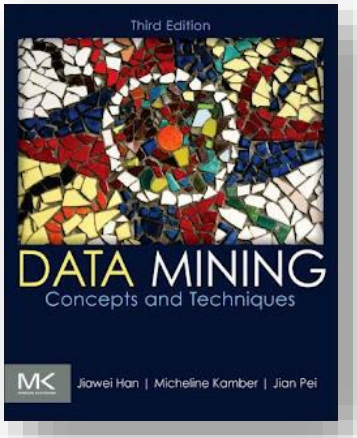
Content of this Course

1. Introduction
2. Data Representation
3. Frequent Pattern Mining
4. Clustering
5. Outlier Detection
6. Classification
7. Regression
8. Privacy Issues for Data Mining
9. Further Topics

Textbook / Acknowledgements

The slides used in this course are modified versions of the copyrighted original slides provided by the authors of the adopted textbooks:

© Jiawei Han, Micheline Kamber, Jian Pei:
Data Mining – Concepts and Techniques,
3rd ed., Morgan Kaufmann Publishers, 2011.
http://www.cs.uiuc.edu/~hanj/bk3



© Martin Ester and Jörg Sander:
Knowledge Discovery in Databases – Techniken und Anwendungen
Springer Verlag, 2000 (in German).

Motivation

- Data Mining = extraction of patterns from data
- Patterns
  - Regularities – examples: frequent itemsets, clusters
  - Irregularities – examples: outliers
- Salient patterns
  - Many patterns are trivial or known
    - „all mothers in our database are female“
  - Many patterns are redundant
    - „{bread, butter} is frequent“ given „{bread, butter, salt} is frequent“
- Aggregation of data may help: basic statistics

What is Data Mining

- Knowledge Discovery in Databases (Data Mining):
  - Extraction of interesting (*non-trivial*, *implicit*, *previously unknown* and *potentially useful*) information or patterns from data in *large databases*

- Alternative names and their "inside stories":
  - Data mining: a misnomer?
  - knowledge extraction, data/pattern analysis, data archeology, data dredging ("Ausbaggern"), information harvesting, business intelligence, etc.

- Roots of data mining
  - Statistics
  - Machine learning
  - Database systems
  - Information visualization

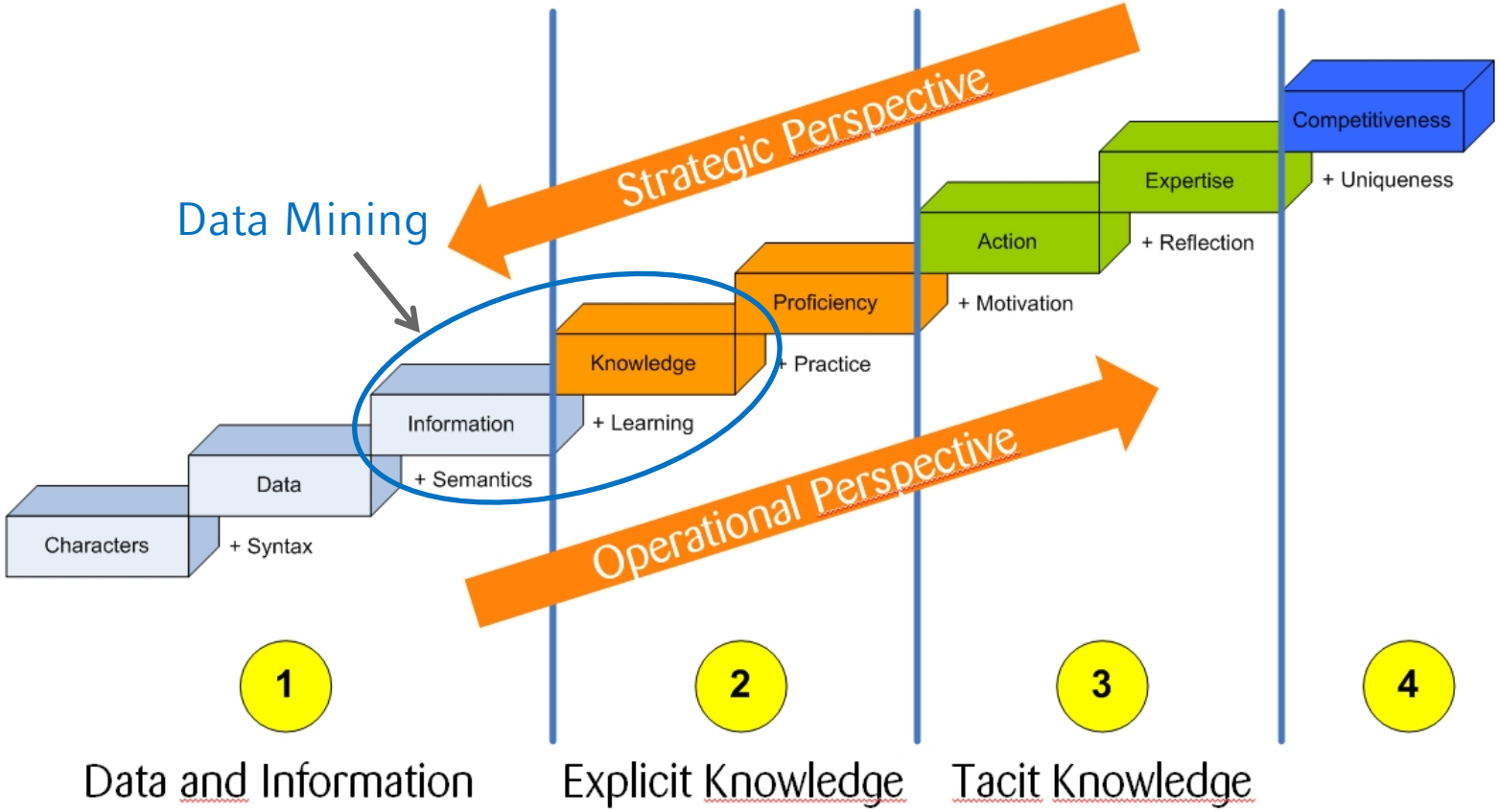
Data Mining: Motivation

"Necessity is the mother of invention"

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- We are drowning in data, but starving for knowledge!
- Solution: data warehousing and data mining
  - Data Warehousing and on-line analytical processing (OLAP)
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data Mining: Motivation

- Stairs of Knowledge (K. North):



Stairs of Knowledge: North, K.: Wissensorientierte Unternehmensführung - Wertschöpfung durch Wissen. Gabler, Wiesbaden 1998.
Picture from: http://wissensarbeiter.wordpress.com/2012/10/29/information-wissen-und-expertise-dazwischen-liegen-welten/

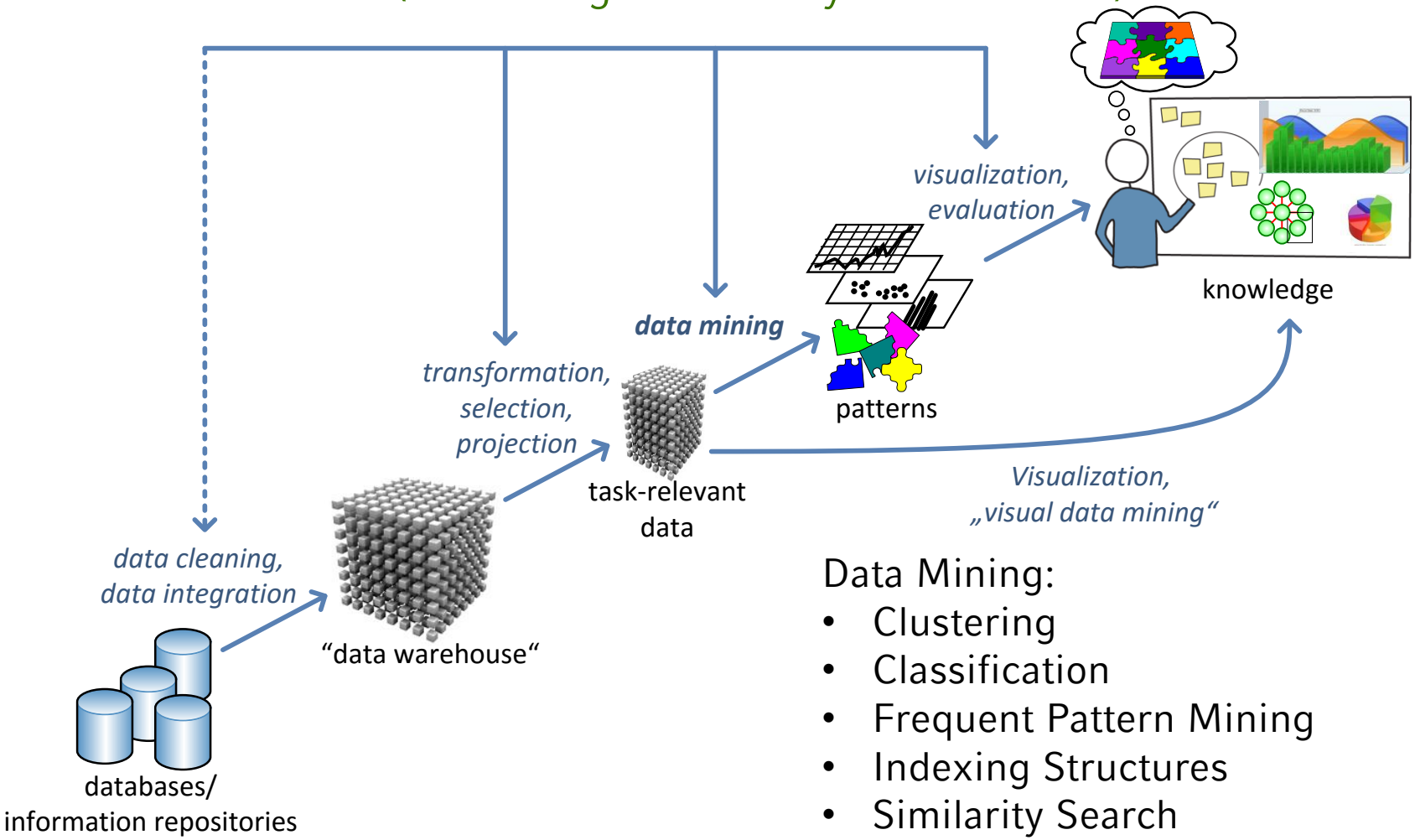
Data Mining: Potential Applications

- Database analysis and decision support
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention (“Kundenbindung”), improved underwriting, quality control, competitive analysis
  - Fraud detection and management

- Other Applications
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering

The Knowledge Discovery Process

- The KDD-Process (**K**nowledge **D**iscovery in **D**atabases)

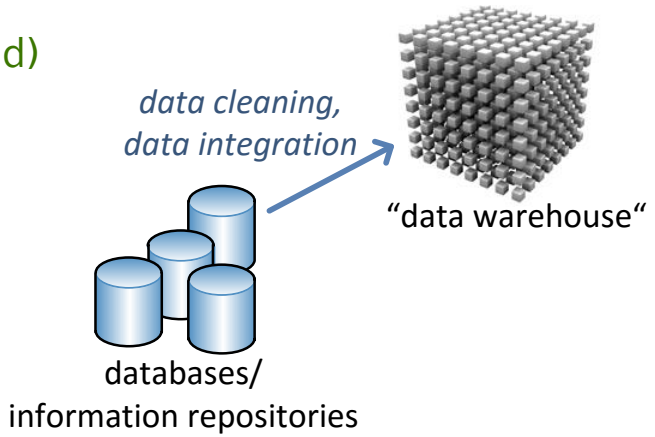


Data Mining:

- Clustering
- Classification
- Frequent Pattern Mining
- Indexing Structures
- Similarity Search

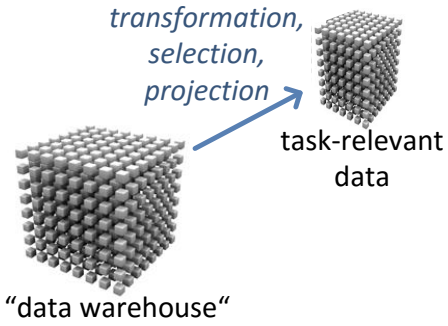
Steps of a KDD Process: Data Cleaning and Integration

- …may take 60% of effort
- Integration of data from different sources
  - Mapping of attribute names (e.g. C_Nr → O_Id)
  - Joining different tables
    (e.g. Table1 = [C_Nr, Info1]
    and Table2 = [O_Id, Info2] ⇒
    JoinedTable = [O_Id, Info1, Info2])
- Elimination of inconsistencies
- Elimination of noise
- Computation of Missing Values (if necessary and possible)
  - Fill in missing values by some strategy (e.g. default value, average value, or application specific computations)



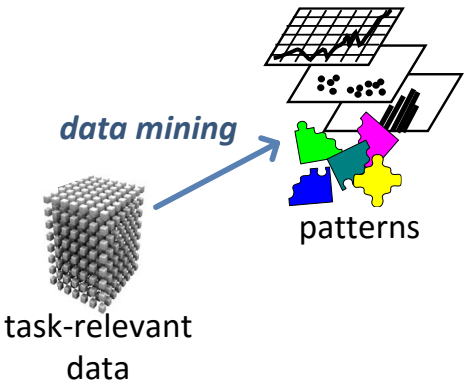
Steps of a KDD Process: Focusing on Task-Relevant Data

- Find useful features, dimensionality/variable reduction, invariant representation
- Creating a target data set
- Selections
  - Select the relevant tuples/rows from the database tables (e.g., sales data for the year 2001)
- Projections
  - Select the relevant attributes/columns from the database tables (e.g., "id", "date" "amount" from (Id, name, date, location, amount))
- Transformations, e.g.:
  - Normalization (e.g., age:[18, 87] → n_age:[0, 100]
  - Discretization of numerical attributes (e.g., amount:[0, 100] → d_amount:{low, medium, high}
  - Computation of derived tuples/rows and derived attributes
    - aggregation of sets of tuples (e.g., total amount per months)
    - new attributes (e.g., diff = sales current month – sales previous month)



Steps of a KDD Process: Basic Data Mining Tasks

- Searching for patterns of interest
- Choosing functions of data mining:
  - Clustering
  - Classification
  - Frequent Patterns
  - Concept Characterization and Discrimination
  - Other methods
    - Outlier detection
    - Sequential patterns
    - Trends and analysis of changes
    - Methods for special data types, e.g., spatial data mining, web mining
    - …
- Choosing the mining algorithm(s)



Basic Data Mining Tasks: Frequent Itemset Mining

- Find frequent patterns in transaction databases
  - Frequently co-occurring items in the set of transactions (*frequent itemsets*): indicate correlations or causalities

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

- Applications:
  - Market-basket analysis
  - Cross-marketing
  - Catalog design
  - Also used as a basis for clustering, classification
  - Association rule mining: Determine correlations between different itemsets

Examples:

buys(x, “diapers”) → buys(x, “beers”) [support: 0.5%, confidence: 60%]

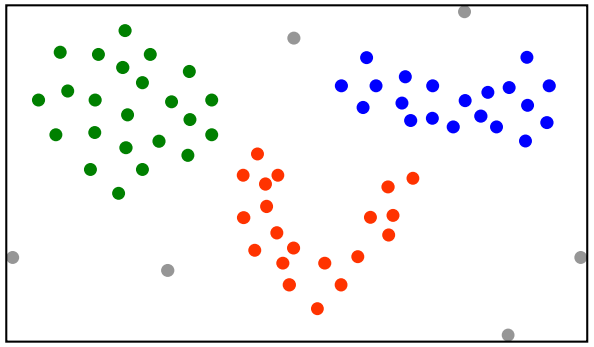
major(x, “CS”) ^ takes(x, “DB”) → grade(x, “A”) [support: 1%, confidence: 75%]

Basic Data Mining Tasks: Clustering

- Class labels are unknown:

  Group objects into sub-groups (clusters)
  - Similarity function (or dissimilarity function = distance) to measure similarity between objects
  - Objective: "maximize" intra-class similarity and "minimize" interclass similarity

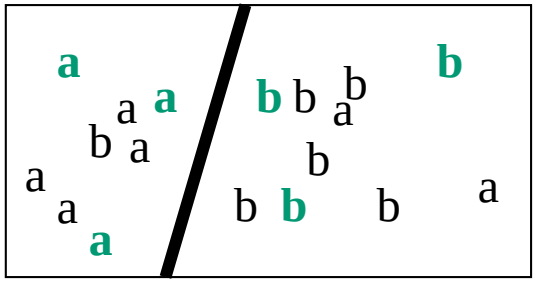
- Applications
  - Customer profiling/segmentation
  - Document or image collections
  - Web access patterns
  - . . .



Basic Data Mining Tasks: Classification

- Class labels are known for a small set of "training data":

  Find models/functions/rules (based on attribute values of the training examples) that
  - describe and distinguish classes
  - predict class membership for "new" objects

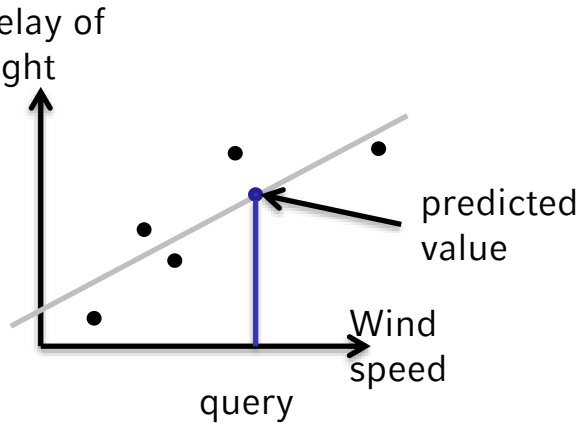


- Applications
  - Classify gene expression values for tissue samples to predict disease type and suggest best possible treatment
  - Automatic assignment of categories to large sets of newly observed celestial objects
  - Predict unknown or missing values (→ KDD pre-processing step)
  - . . .

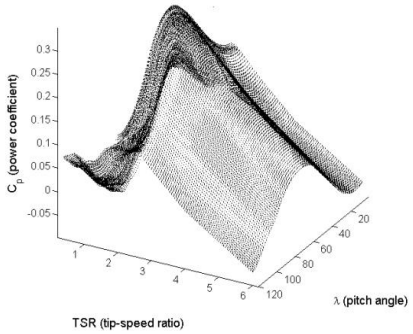
Basic Data Mining Tasks: Prediction

- Numerical output values are known for a small set of "training data":

  Find models/functions (based on attribute values of the training examples) that
  - describe the numerical output values of the training data (Major method for prediction is regression)
  - predict the numerical value for "new" objects



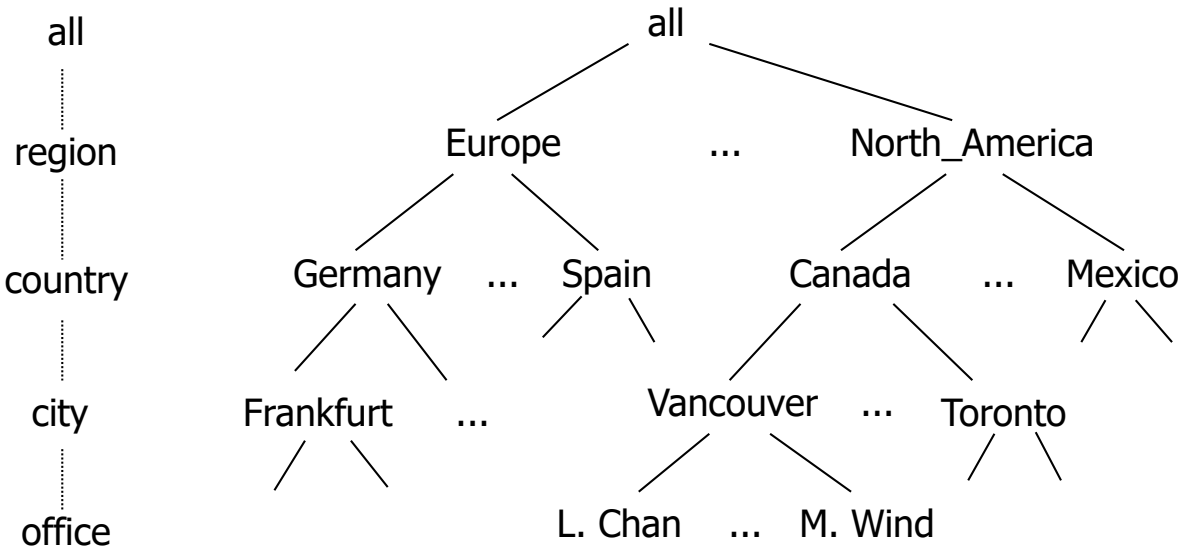
- Applications
  - Build a model of the housing values, which can be used to predict the price for a house in a certain area
  - Build a model of an engineering process as a basis to control a technical system
  - . . .



Wind turbine

Basic Data Mining Tasks: Generalization Levels

- Generalize, summarize, and contrast data characteristics
  - Based on attribute aggregation along concept hierarchies
    - Data cube approach (OLAP)
    - Attribute-oriented induction approach

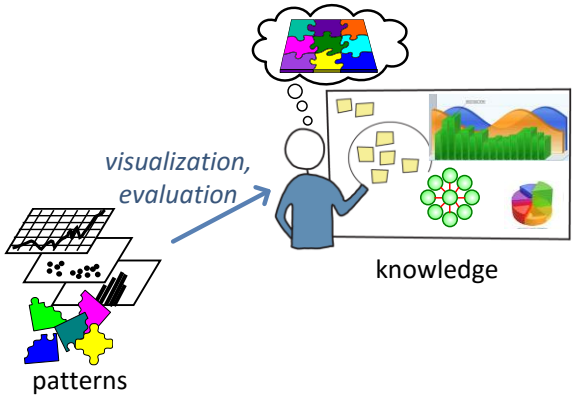


Basic Data Mining Tasks: Other Methods

- Outlier detection
  - Find objects that do not comply with the general behavior of the data (fraud detection, rare events analysis)
- Trends and Evolution Analysis
  - Sequential patterns (find re-occurring sequences of events)
- Methods for special data types, and applications e.g.,
  - Spatial data mining
  - Web mining
  - Bio-KDD
  - Graphs
  - . . .

Steps of a KDD Process: Evaluation and Visualization

- Pattern evaluation and knowledge presentation:
  - Visualization, transformation, removing redundant patterns, etc.
- Integration of visualization and data mining
  - data visualization
  - data mining result visualization
  - data mining process visualization
  - interactive visual data mining



- Different types of 2D/3D plots, charts and diagrams are used, e.g.:
  - box-plots, trees, X-Y-Plots, parallel coordinates
- Use of discovered knowledge

Summary

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, machine learning, statistics, visualization, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

Outline of the following chapters

1. Introduction
2. Data Representation
3. Frequent Pattern Mining
4. Clustering
5. Outlier Detection
6. Classification
7. Regression
8. Privacy Issues for Data Mining
9. Further Topics

References

- Data mining and KDD:
  - Conference proceedings: KDD, PKDD, PAKDD, SDM, ICDM etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field:
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, CIKM
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, VLDBJ, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization:
  - Conference proceedings: CHI (Comp. Human Interaction), etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

References

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. 2nd ed., Morgan Kaufmann, 2006.
- T. Imielinski and H. Mannila. *A database perspective on knowledge discovery*. Communications of the ACM, 39:58-64, 1996.
- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. *From data mining to knowledge discovery: An overview*. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
- G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
- M. Ester and J. Sander. *Knowledge Discovery in Databases: Techniken und Anwendungen*. Springer Verlag, 2000 (in German).
- M. H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2003.
- D. Hand, H. Mannila, P. Smyth. *Principles of Data Mining*. MIT Press, 2001.