TIES443

Introduction to Data Mining

Instructors: Mykola Pechenizkiy & Sami Äyrämö
Contact: mpechen@cs.jyu.fi samiayr@mit.jyu.fi
Mailing list: ties443@korppi.jyu.fi
Course webpage: http://www.cs.jyu.fi/~mpechen/TIES443

November 1 – December 12, 2006

Department of Mathematical Information Technology
University of Jyväskylä

Objectives of the Course

- Provide basic introduction into key areas such as
  - OLAP (that stands for On Line Analytical Processing) Design,
  - Data Warehousing (DW), and
  - Data Mining (DM)
- Provide an overview of most common tasks and application areas of DM
  - Prediction and knowledge discovery
- Provide an overview of most common techniques used in DM
  - Building and evaluating predictive and descriptive models
- Ensure that students of the course will gain the necessary background and skills
  - to turn available data into valuable and useful information

Course overview

- Lectures 15*2 = 30 hours
  - Wed 8:15 – 10:00, Thu 12:15 – 14:00, Fri 10:15 – 12:00
  - all in Ag Beeta
  - Nov 17, 12.00 – 14.00 Sami's public examination of PhD (after the lecture)
- Tutorial followed by an assignment 5*2 = 10 hours
  - Tue 14.15 – 16.00, Ag B212.2 (Mountains),
  - but week 50: Wed 8:15 – 10:00
- Seminar 2 hours
  - Ag B212.2 (Mountains)
  - 5-10 min presentation by each student about the final assingment
- Final assignment (no final exam)
  - to be sent to mpechen@cs.jyu.fi and samiayr@mit.jyu.fi by the end of Jan'07 (always use TIES443 keyword in the subject field)

Credits, Passing the Course/Grading

- Credits
  - 5 ECTS (3 ov),
- Grading
  - Five assignments + final assignment = 5*5 + 3*5 = 40 points as max
    - 1-2 pages report for each of five assignments should be submitted by e-mail to mpechen@cs.jyu.fi and samiayr@mit.jyu.fi within a week of the day of the assignment
    - Report on final assignment should be submitted by the end of Jan'07
    - I will tell you more during the first lab
- Communication outside the classes
  - ties443@korppi.jyu.fi
  - Appointment by sending a request to mpechen@cs.jyu.fi or samiayr@mit.jyu.fi

Course Contents: BI & DW Part

- Introduction to introduction :
  - **Basic definitions**
    - DM and KDD
  - **History of DM**
    - Motivation for DM, reference disciplines, DM community
  - **Major DM tasks and application**
    - Prediction, knowledge discovery
  - **Major issues in DM**
- Introduction to Business Intelligence
  - **DM in BI context**
    - DM myths, OLAP vs DM
- Introduction to Data Warehousing
  - **DW architecture, design, implementation**
    - Data cubes, OLAP operations

Course Contents: DM Part

- **DM: Input and Output**
  - **Input: Concepts, instances, attributes**
    - What is a concept, an example an attribute?
  - **Output: Knowledge Representation**
    - Decision tables, trees, and rules; relations, CBR
- **DM: Techniques**
  - **Data preparation**
    - Cleaning, missing values; transformation, Curse of dimensionality
  - **Clustering, Classification, Associations, Visualization**
    - The largest part of the course
- **DM: Evaluation and Credibility**
  - **Predicting Performance**
    - Train, test, and validation sets, cross-validation; unbalanced data
  - **Comparing Data Mining Schemes**
    - ROC; cost-sensitive learning; Occam's razor; parameters tuning
- **DM: KDD process**
  - **Iterative, interactive**
- **DM: Miscellaneous issues**
  - **Privacy, ethics, distributed DM**

Course Contents: Tutorials

- Prototyping DM techniques and solutions
  - WEKA and YALE open-source software
  - MATLAB environment
- Mining time-series data
  - Review of basic techniques
- Mining image data
  - Review of basic techniques
- Mining text data
  - Review of basic techniques
  - ExtMiner (Miika Nurminen)
- An assignment will follow each tutorial
  - Application of DM to benchmark/real world data

Resourses

- Witten I., Frank E. 2000. Data Mining: Practical machine learning tools with Java implementations", Morgan Kaufmann, San Francisco. (book & software page)
- Crawford D. 1996 Special Issue on Data Mining. Communications of the ACM, Volume 39, Number 11, November, 1996
- Reinartz, T. 1999, Focusing Solutions for Data Mining. LNAI 1623, Berlin Heidelberg.
- Han J. and Kamber M. 2000, Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 550 pages. ISBN 1-55860-489-8 (ppt slides to the book)
- Data Mining: A Practitioner's Approach, ELCA Informatique SA, 2001
- CRISP-DM 1.0: Step-by-step data mining guide, SPSS Inc.
- Check TIES443 homepage for more. It will be updated regularly

More on DM and KDD

KDnuggets.com

- News, Publications
- Software, Solutions
- Courses, Meetings, Education
- Publications, Websites, Datasets
- Companies, Jobs
- …

Special Acknowledgments

for many adopted/adapted ppt slides used in the lectures:

- Piatetski-Shapiro (KDnuggets)
- Witten & Frank's book http://www.cs.waikato.ac.nz/~ml/weka/book.html
- Eamon Keogh http://www.cs.ucr.edu/~eamonn/
- Han's DM book http://www.cs.sfu.ca/~han/dmbook
- and many many DM-related courses available in www

Topics for this week

- Introduction to the DM field:
  - definitions, motivation, brief history
  - DM tasks and application examples
- Business Intelligence
  - DM in BI context
    - DM myths, OLAP vs DM
- Data Warehousing (DW)
  - DW architecture, design, implementation
    - Data cubes, OLAP operations

Topics for today

- What is Data Mining?
  - **Basic definitions**
    - DM and KDD
  - **History of DM**
    - Motivation for DM, reference disciplines, DM community
  - **Major DM tasks and application**
    - Prediction, knowledge discovery
  - **Major issues in DM**

What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (*non-trivial, implicit, previously unknown* and *potentially useful*) information or patterns from data in *large databases (Fayyad)*
  - ***the process of selecting, exploring, and modeling*** large amounts of data to uncover previously unknown patterns for a ***business advantage*** *(SAS Institute)*
  - Data mining is ***an area in the intersection of machine learning, statistics, and databases.*** *(Holsheimer et al.)*
- Alternative names and their "inside stories":
  - Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- What is not data mining?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

Data Mining in the BI Context

- Data Mining is a business-driven process, supported by adequate tools, aimed at the discovery and consistent use of meaningful, profitable knowledge from corporate data
- A kind of operationalization of Machine Learning with emphasis on process and actions

- Hand (2000), "Data Mining is the process of seeking interesting or valuable information in large data bases
- Large commercial data bases could be used to increase profitability by pinpointing target classes of clients and pointing clients toward desirable options.
- The result – large data mining programs sold by SAS, SPSS, etc.

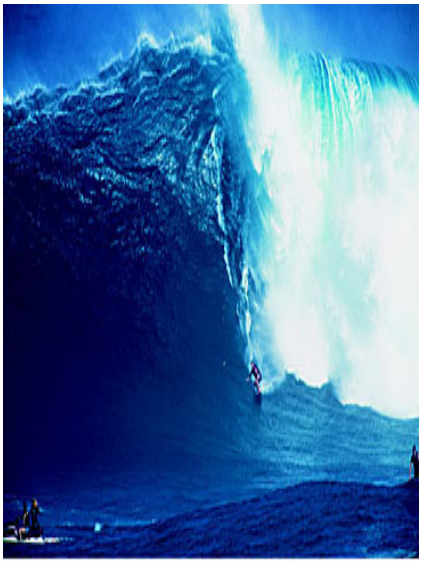
Motivation for DM

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
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Trends leading to Data Flood

- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce



Big Data Examples

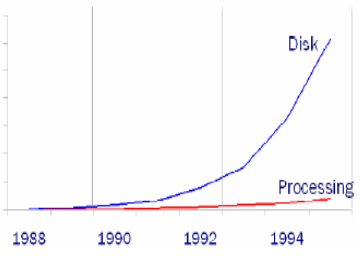
- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
  - storage and analysis a big problem
- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done "on the fly", on streaming data

Largest databases in 2003

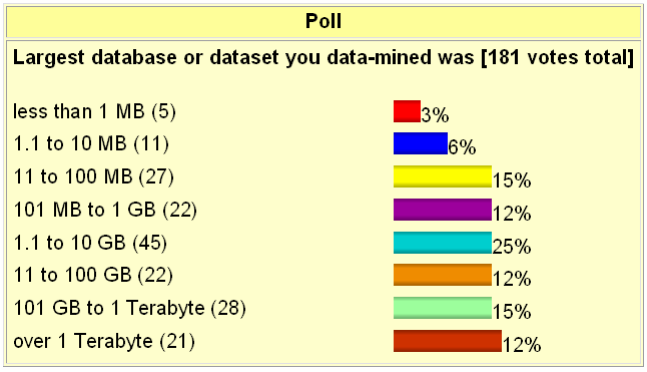
- Commercial databases:
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB
- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - IBM WebFountain, 160 TB (2003)
  - Internet Archive (www.archive.org),~ 300 TB

Growth Trends

- Moore's law
  - computer speed doubles every 18 months
- Storage law
  - total storage doubles every 9 months
  - exabytes (million terabytes) of new data are created every year
  - huge DBs (telecom, AT&T, astronomy,…)
- Consequence
  - very little data will ever be looked at by a human
  - data flood / information overload
- DM/KDD is **needed** to make sense and use of data.

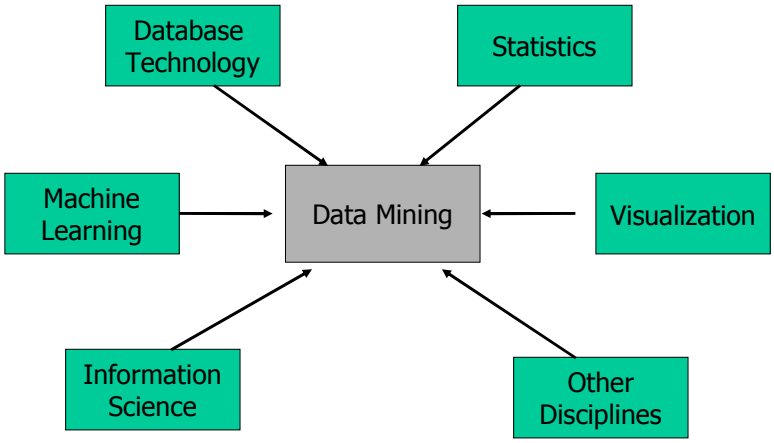


Largest Database Data-Mined (June 2006)



"Data miners are tackling much larger databases in 2006. The median value for the largest database size is between 1.1 and 10 Gigabytes, and 12% report mining terabyte size databases."
http://www.kdnuggets.com/polls/2006/largest_database_mined.htm

Data Mining: Reference Disciplines

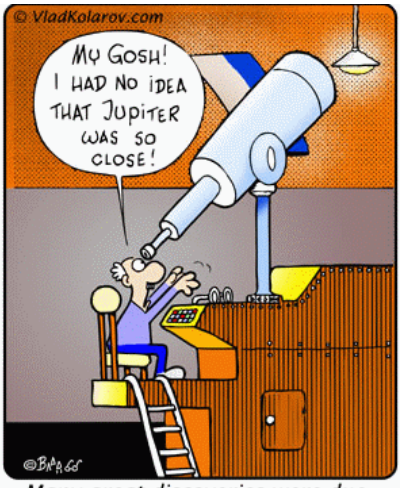


Brief History of Time …

Brief History

- 1800 Statistics starts
- Benjamin Disraeli, *later quoted by Mark Twain*, said, "**There are three kinds of lies: lies, damned lies, and statistics.**"
  - And now comes …Data Mining
- 1985 machine learning starts
- 1990 data mining starts

False Positives in Astronomy



Many great discoveries were due to Professor McDowell's shortsightedness.

cartoon used with permission

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

Many Names of Data Mining

- Data Fishing, Data Dredging: 1960-
  - used by Statistician (as bad name)
- Data Mining :1990 --
  - used DB, business
  - in 2003 – bad image because of TIA
- Knowledge Discovery in Databases (1989-)
  - used by AI, Machine Learning Community
- also Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...

Currently: Data Mining and Knowledge Discovery are used interchangeably

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on KDD (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on KDD
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

Statistics vs. Machine Learning

- *Answering different questions*
  - Inside nature associates the predictor variables with the response variables
- Statistics
  - Infer the mechanism of the inside of the box.
- Machine Learning
  - Find f(**x**) that is a good predictor of the outputs **y.**
  - Don't worry about the inside of the box.
  - Look at what's outside--the inputs and outputs.
  - The function f(**x**) is an algorithm - lines of code that direct the computer how to operate on **x** to produce f(**x**)**.**

A common goal –construct accurate prediction algorithms.

Statistics, Machine Learning and Data Mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Witten&Eibe

Potential Data Mining Applications

Data Mining Application Areas

- Science
  - astronomy, bioinformatics, drug discovery, …
- Business
  - advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care, …
- Web:
  - search engines, bots, …
- Government
  - law enforcement, profiling tax cheaters, anti-terror

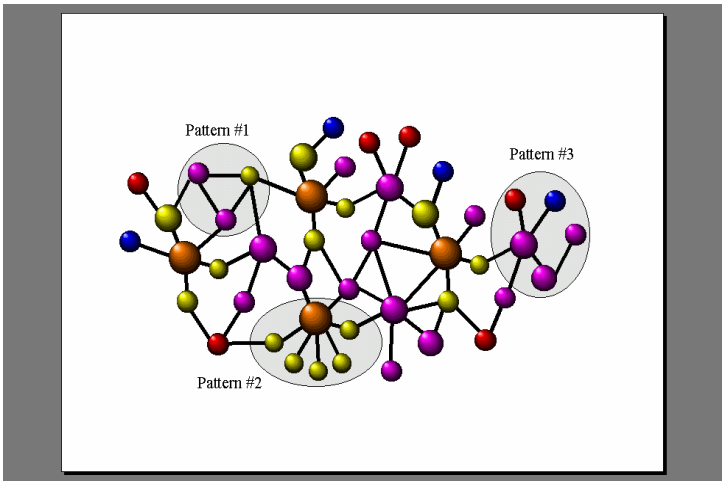
Assessing Credit Risk

- Situation:
  - Person applies for a loan
- Task:
  - Should a bank approve the loan?
    - Note: People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle
- DM solution:
  - Banks develop credit models using variety of machine learning methods.
  - Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan
- Widely deployed in many countries

e-Commerce

- A person buys a book (product) at Amazon.com.
- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful

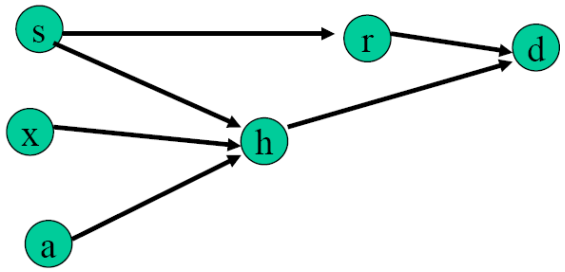
Link Analysis



Can find unusual patterns in the network structure

Link Analysis

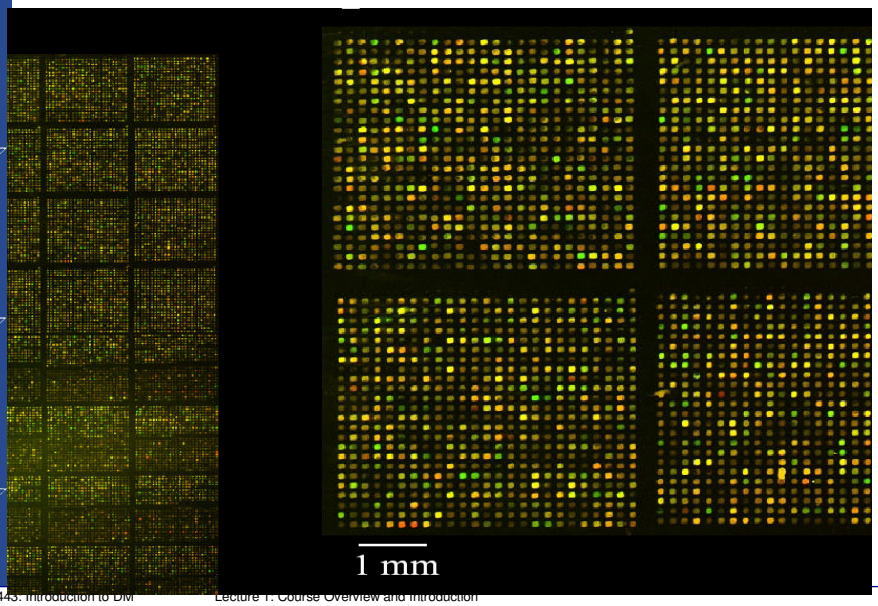
Discovering interdependencies



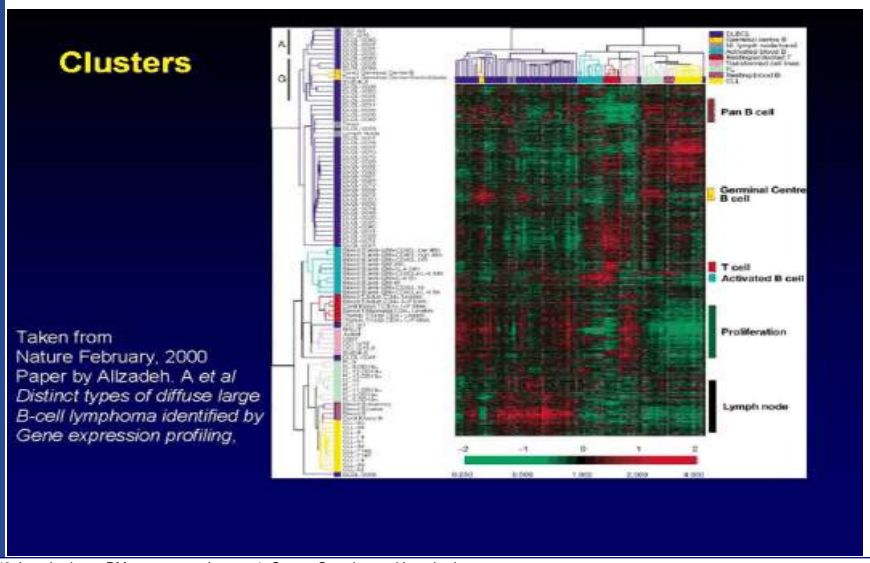
s=smoker
x=sex
a=age
h=health
r=resistance
d=live/death

What influences what and to which extent?
Bayesian networks produce graphical models of knowledge

Microarray (an example)

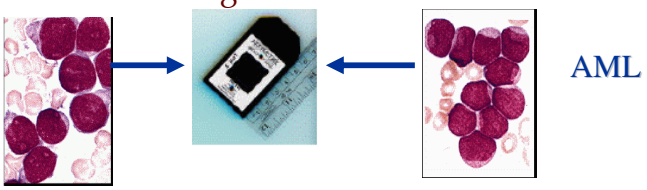


Graphical Representation of Microarray Data



Biology: Molecular Diagnostics

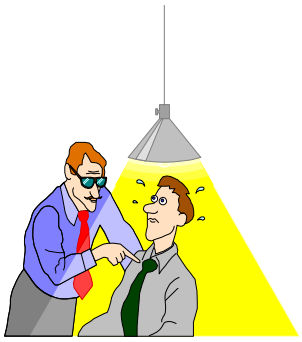
- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model

ALL  AML

Results on test data:
33/34 correct, 1 error may be mislabeled

Security and Fraud Detection

- Credit Card Fraud Detection
- Money laundering
  - FAIS (US Treasury)
- Securities Fraud
  - NASDAQ Sonar system
- Phone fraud
  - AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terrorism detection at Salt Lake Olympics 2002



Fraud Detection and Management

- Detecting inappropriate medical treatment
  - Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian $1m/yr).
- Detecting telephone fraud
  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
- Retail
  - Analysts estimate that 38% of retail shrink is due to dishonest employees.

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Problems Suitable for Data-Mining

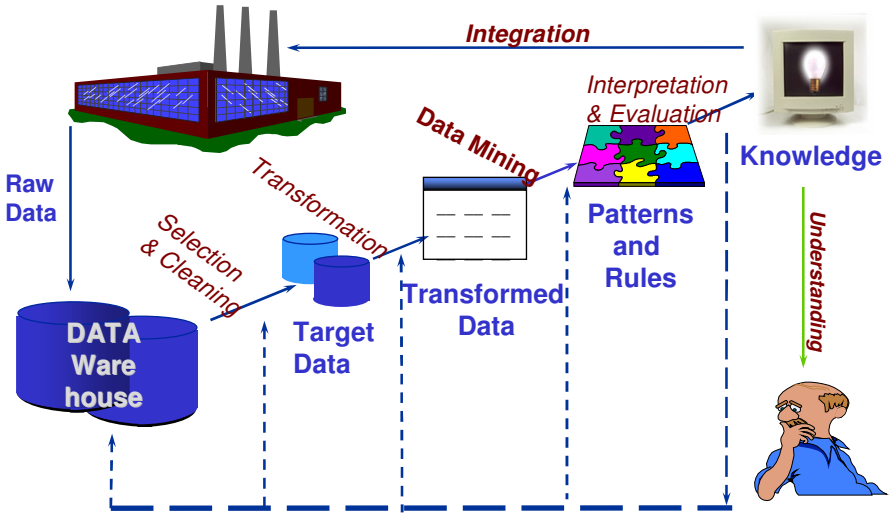
- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!

Privacy considerations important if personal data is involved

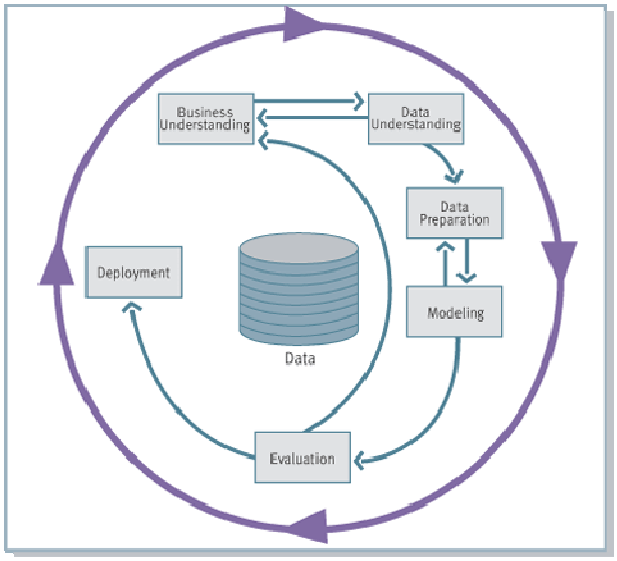
Examples of Problems Suitable for DM: Summary

- Medical diagnosis: soft or hard contact lenses
- Credit application scoring: grant a loan or not?
- Fraud detection: is the transaction suspicious or not?
- Direct mailing: who should be offered a given product?
- CPU-performance: how to configure computers?
- Remote sensing: determine water pollution from spectral images
- Load forecasting: predict future demand for electric power
- Intelligent ATM's : how much cash will be there tomorrow?
- identify groups of similar credit card users
- automatically organize incoming e-mails
- characterize interests of an Internet user

Knowledge Discovery Process



CRISP-DM



see www.crisp-dm.org for more information

Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Major DM Tasks

Major Data Mining Tasks

- Classification: predicting an item class
- Clustering: finding clusters in data
- Associations: e.g. A & B & C occur frequently
- Visualization: to facilitate human discovery
- Estimation: predicting a continuous value
- Outlier analysis
- Deviation Detection: finding changes
- Link Analysis: finding relationships
- Sequential pattern mining, periodicity analysis
- …

Interestingness of "Discovered" Patterns

- A DM system/query may generate thousands of patterns, not all of them are interesting.
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**:
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures:**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Finding All and Only Interesting Patterns

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering
- Search for only interesting patterns: Optimization
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

DM Facets

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Major Issues in Data Mining (1)

- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- Performance and scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods

Major Issues in Data Mining (2)

- Issues relating to the diversity of data types
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- Issues related to applications and social impacts
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy

Summary

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, statistics and AI, - in great demand, with wide applications
- DM is a process
- Mining can be performed in a variety of information repositories
- DM functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Classification of data mining systems
- Major issues in data mining

What should you get from this lecture?

- Course contents
  - BI & DW part
  - DM part
  - Tutorials and assignments
  - Seminar and final assignment
- Idea of what DM is
  - Motivation
  - Where DM can be used
    - Major tasks and applications
  - Reference disciplines
    - Distinctive characteristics of DM
  - DM facets and major issues
- What did you remember?