
Data Mining

Analiza i eksploracja danych



Lecturer: JERZY STEFANOWSKI
Institute of Computing Science
Poznań University of Technology
Poznań, Poland

Software Engineering – Master Course
Computer Science, PUT, 2008

Course mission

- Data mining?
 - Extraction of useful information patterns from data.
 - Huge or complex data challenge (DB community).
 - Trend to data warehouses but also flat table files.
 - More than typical data analysis, machine learning or classical decision support!
 - Still „young” field, although quite „fashionable”.
 - Teaching materials and course book ...



Course mission - 2

- Thus, our aims:
 - To give you insight in this field, typical methods, examples of applications.
 - Rather focus on algorithms and methodology aspects, not so much on dealing with massive data.
 - Choice of methods.
 - Case study and applications.
 - Comments to available software (WEKA, etc.)
 - Data Sets – UCI Repository and others

Course information

- The planned schedule of lectures (15 weeks):
 - Introduction
 - Data Preprocessing
 - Prediction models (multivariate regression and ANN)
 - Classification (evaluation)
 - Symbolic methods (Decision Trees and Rules)
 - Other on-symbolic methods
 - Association rules and sequence patterns
 - Clustering
 - Mining complex data
 - KDD Process and summary

Background literature

- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001 (1 ed.), there is 2d
- Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001.
- Kononenko I., Kukar M., Machine Learning and Data Mining: Introduction to Principles and Algorithms. Horwood Pub, 2007.
- Maimon O., Rokach L., The data mining and knowledge discovery Handbook, Springer 2005.
- Witten I., Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.
- Bramer M., Principles of Data Mining, Springer 2007.
- Weiss S., Indurkha N., Predictive data mining, Morgan Kaufmann, 1998.



Background literature [Polish translations]

- Larose D., Odkrywanie wiedzy z danych. Wprowadzanie do eksploracji danych, PWN, 2006.
- Larose D., Metody i modele eksploracji danych, PWN 2008.
- Hand D., Mannila H., Smyth P. Eksploracja danych, WNT, 2005.

Polskie książki

- Dobry podręcznik jeszcze nie istnieje ...
- Koronacki J., Ćwik J., Statystyczne systemy uczące się, WNT 2005 (kolejne wydanie w drodze).
- Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. PP, 2003.
- Cichosz P., Systemy uczące się. WNT, 2000.
- Lasek M., Data mining: Zastosowanie w ocenach i analizach klientów bankowych. Biblioteka Menadżera, 2003.



Acknowledgements:

- Many of the slides are based on my earlier courses:
 - Data mining and advanced data analysis; Knowledge discovery (PUT CS; M.Sc. Course) more at <http://www.cs.put.poznan.pl/jstefanowski>
- Some slides are based on ideas „borrowed” from:
 - WEKA teaching materials (Witten & Frank Waikato University; Morgan Kaufmann)
 - Gregory Piatetsky – Shapiro: Data mining course.
 - Jiawei Han: Knowledge discovery in databases.
 - T.Mitchell and P.Flach courses on ML (see their WWW).
- Other course books – see the previous slides

Lecture 1 a.

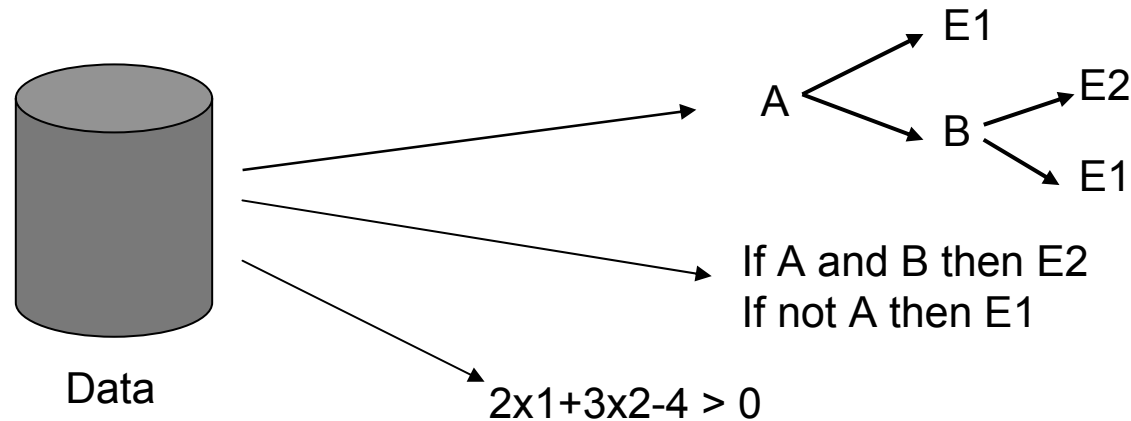
Data Mining: Introduction

L1. Introduction to Data Mining: Outline

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data and what kind of knowledge representation?
- Basic tasks and methods.
- Examples of applications.
- WEKA and Statistica – software frameworks for this course.



Data mining: what is it?



- Data mining is
 - Extraction of useful patterns from data sources, e.g., databases, texts, web, images.
- Patterns (knowledge representation) must be:
 - Valid, novel, potentially useful, understandable to the users.

What is data mining? More ...

- Data mining is the analysis of data for relationships that have not previously been discovered or known.
- A term coined for a new discipline lying at the interface of database technology, machine learning, pattern recognition, statistics and visualization.
- The key element in much more elaborate process called „**Knowledge Discovery in Databases**”.
- The efficient extraction of previously unknown patterns in very large data bases.
- 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 (**Hand, Mannila, Smyth**).

Motivations - 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.
- More data is generated:
 - Bank, telecom, other business transactions ...
 - Scientific data: astronomy, biology, etc
 - Web, text, and e-commerce
- Very little data will ever be looked at by a human!
- We are drowning in data, but starving for knowledge!
- Knowledge Discovery is **NEEDED** to make sense and use of data.

Big Data Examples

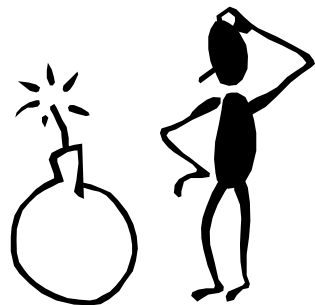
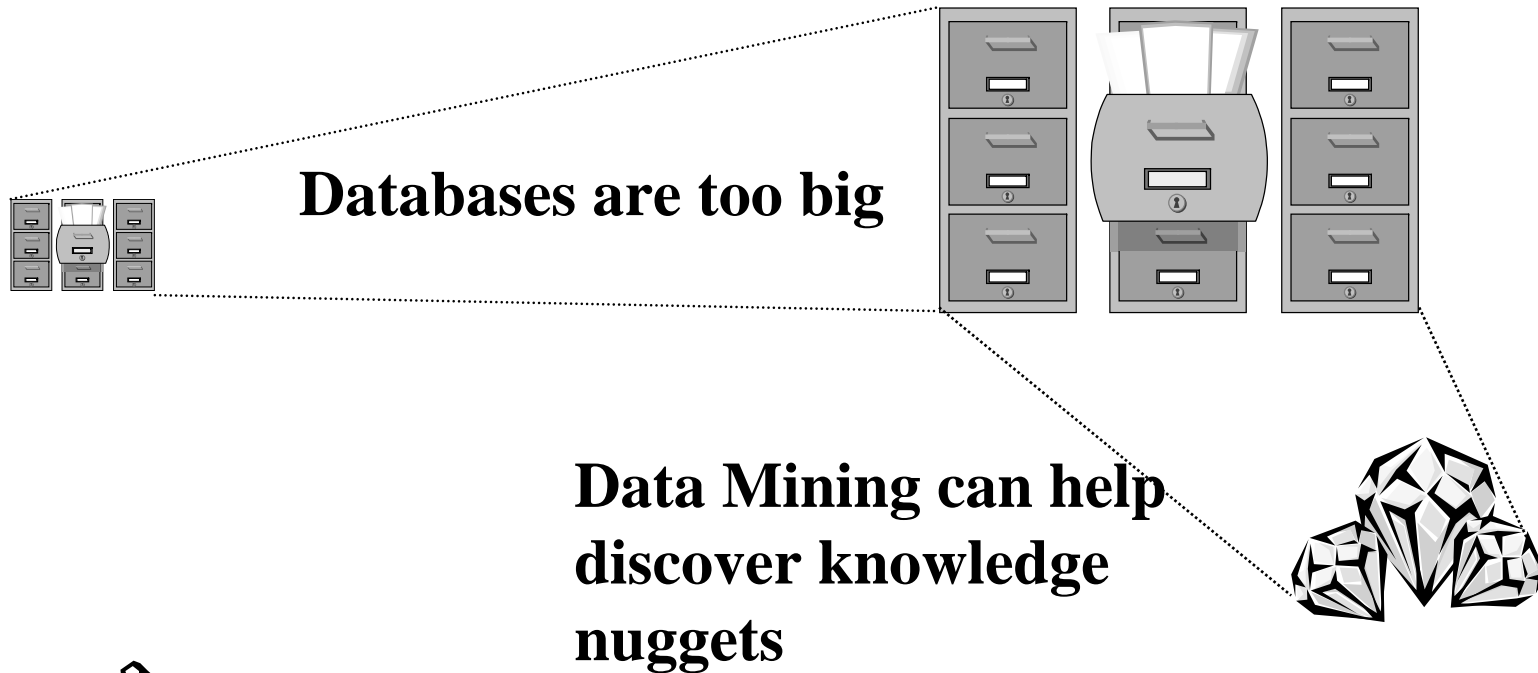
- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
 - storage and analysis a big problem
- AT&T handles billions of calls per day
 - so much data, it cannot be all stored -- analysis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

Largest databases in 2003



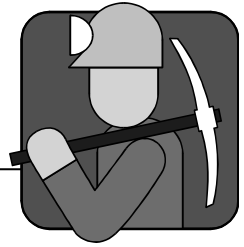
- Commercial databases:
 - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB
- Web
 - Alexa internet archive: 7 years of data, 500 TB
 - Google searches 4+ Billion pages, many hundreds TB
 - IBM WebFountain, 160 TB (2003)
 - Internet Archive (www.archive.org), ~ 300 TB
- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002.
www.sims.berkeley.edu/research/projects/how-much-info-2003/
- US produces ~40% of new stored data worldwide

We are Data Rich but Information Poor



„Terrorbytes“

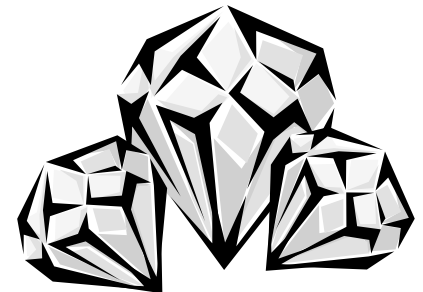
Knowledge Discovery Definition



- Knowledge Discovery in Data is the *non-trivial* process of identifying
 - *valid*
 - *novel*
 - potentially *useful*
 - and ultimately *understandable patterns* in data.

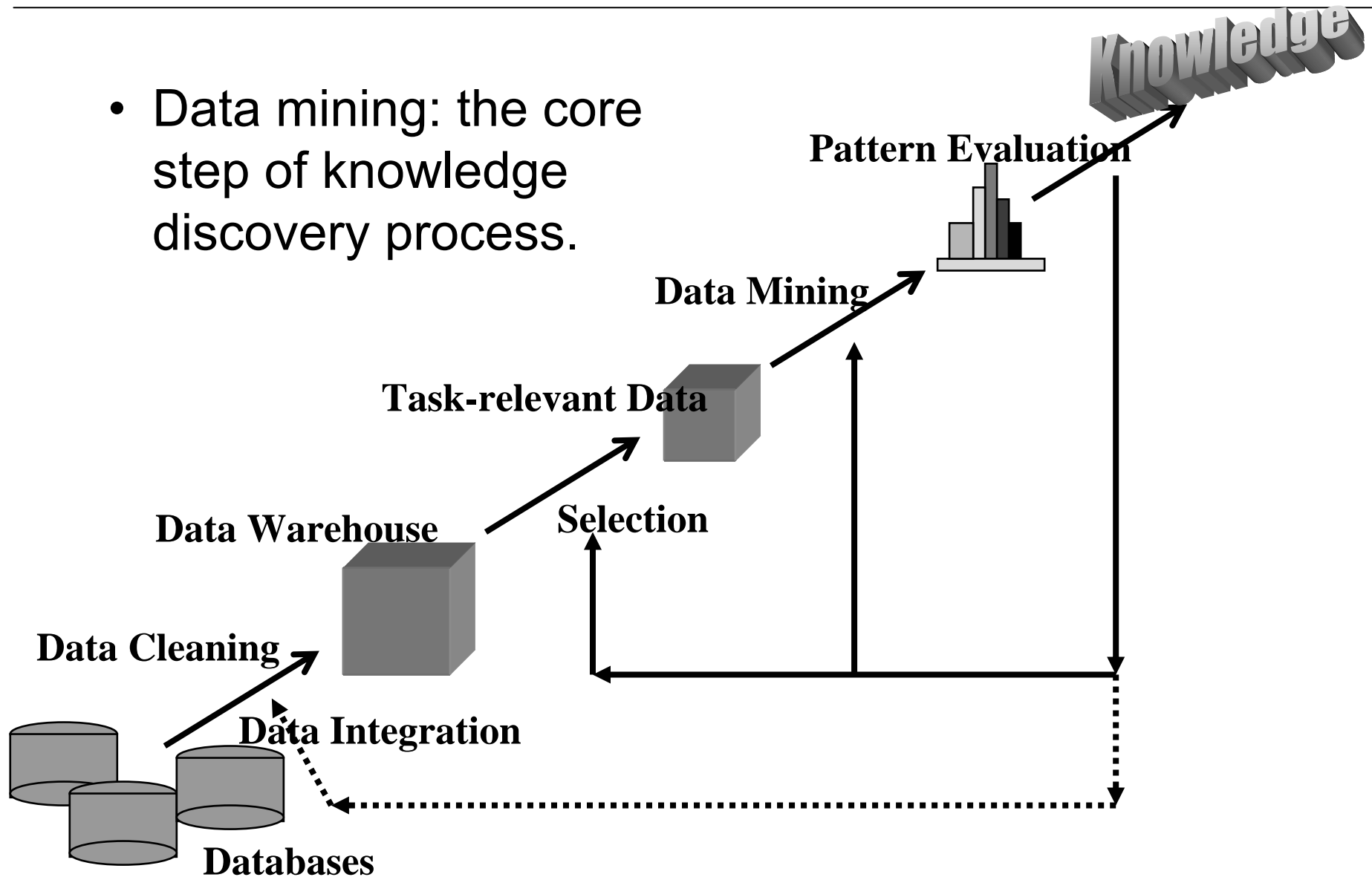
from *Advances in Knowledge Discovery and Data Mining*,
Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter
1), AAAI/MIT Press 1996.

The name first used by AI, Machine Learning Community
in 1989 Workshop at AAAI Conference.



Data Mining as a step in A KDD Process

- Data mining: the core step of knowledge discovery process.



Steps of a KDD Process

- Learning the application domain:
 - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and projection:
 - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing a task of a data mining step
 - summarization, classification, regression, association, clustering.
- Choosing proper mining algorithm(s)
- Data mining: searching for patterns of interest
- Interpretation: analysis of results.
 - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

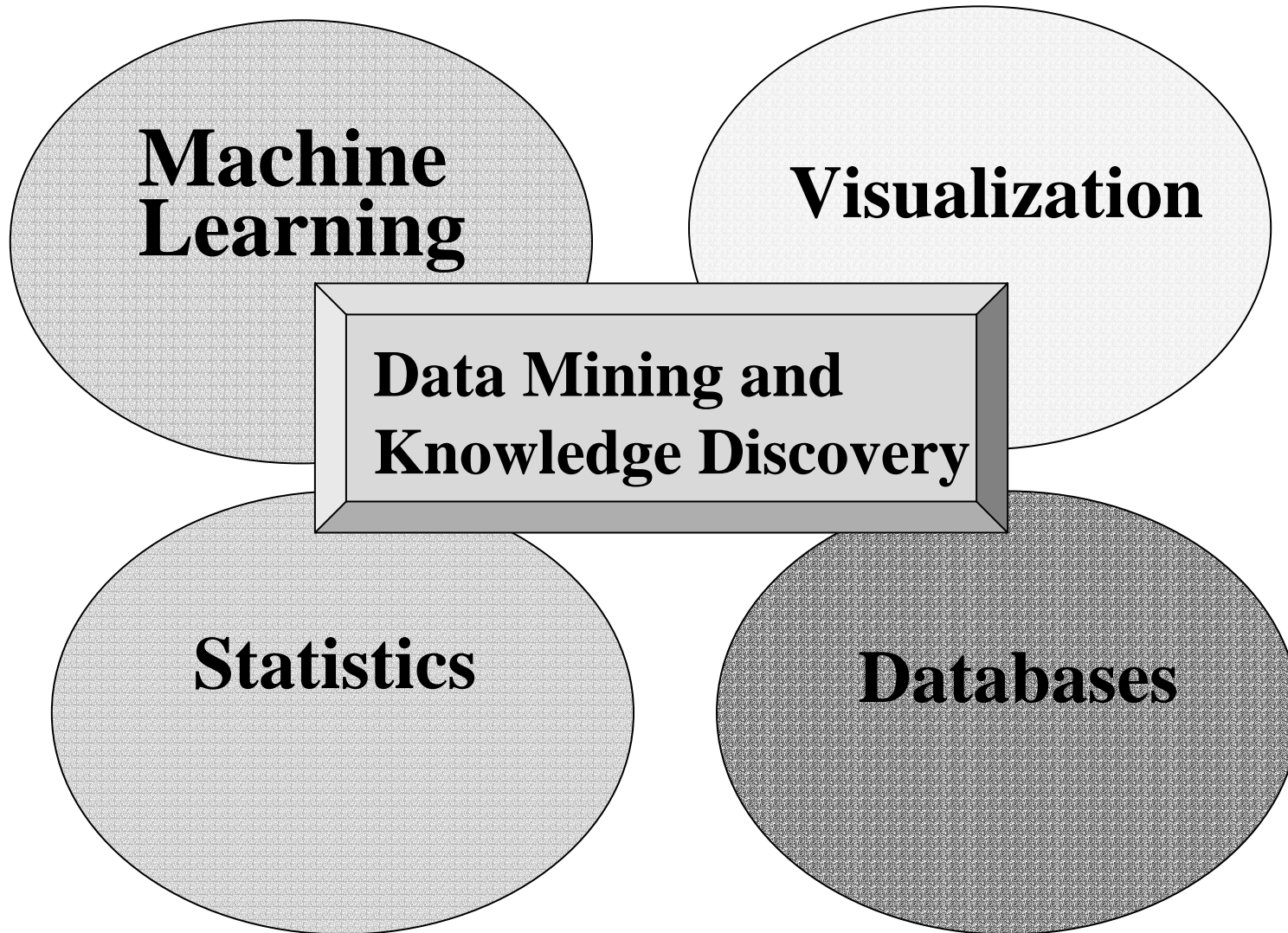
SKICAT – an example of KDD process

- **Sky Image Cataloging and Analysis Tool (SKICAT)**
- Developed by NASA's Jet Propulsion Laboratory and the California Institute of Technology in the 90's (Fayyad, Djorgowski, Weir et al.) .
- **Aim:** a software system to catalog and analyze the estimated half billion sky objects in the second Palomar Observatory sky survey
 - Task 1 – general classes (galaxis, stars, quasars, etc.)
 - Task 2 – find some interesting clusters of objects (quasars with redshift..)
- The survey of the northern sky includes more than 3,000 digitized photographic plates produced by Palomar, located in San Diego.
 - Over 3 terrabytes of images 13000×13000 pixels
- The SKICAT system will produce a comprehensive survey catalog database containing about one-half billion entries by automatically processing about three terabytes (24 trillion bits, 8-bits to a byte) of image data.
- SKICAT is based on state-of-the-art machine learning, high performance database and image processing techniques.
- SKICAT has a correct sky object classification rate of about 94 percent, which exceeds the performance requirement of 90 percent needed for accurate scientific analysis of the data.

SKICAT – KDD steps

- *Dostępne dane i wiedza początkowa*: kilkadziesiąt tysięcy fotografii o różnych rozdzielczości ręcznie skatalogowanych przez ekspertów.
- *Wybór docelowych danych* – identyfikacja właściwych atrybutów charakteryzujących poszczególne klasy
 - konieczność uwzględnienia dodatkowej wiedzy astronomicznej and image analysis FOCAS.
- *Krok czyszczenia* – identyfikacja różnych obserwacji odstających i błędów w danych.
- *Krok redukcji* – wybór tylko części z dostępnych danych.
- *Wybór zadania i algorytmu*: klasyfikowanie – drzewa decyzyjne (modyfikacja C4.5, ale także alternatywne klasyfikatory).
- *Ocena rezultatów* (trafność klasyfikacji – tutaj ponad 94%).
- *Zastosowanie* – wspomaganie tworzenia elektronicznego katalogu gwiazd i galaktyk wraz z ich opisami.

Related Fields



Related fields

- Statistics – model driven approach based on a strong mathematical background; model-hypothesis testing or estimation methods.
 - Modern statistical learning.
- Machine learning – improving performance of intelligent systems; knowing concepts to be learned; strongly related to AI paradigms.
- Databases systems – efficient data management and extractions; data warehouses.
- Knowledge based system – representation and AI methods.

Data mining and related disciplines

What is not data mining?

- Another statistical approach shifted into a new context!
- It is not only machine learning!
- Moreover, it is not
 - (Database) query processing,
 - Expert systems software.



Data Mining: On What Kind of Data?

- **Attribute-value tables (standard form / data table)**
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, ...)
- Sequence data bases
- Other more complex data repositories
 - Object-oriented and object-relational databases
 - Spatial databases
 - Time-series data and temporal data
 - Text databases and multimedia databases
 - WWW resources
 - ...



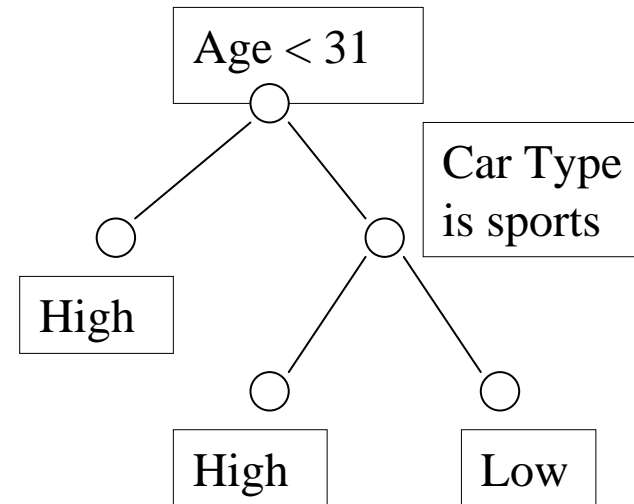
What can be discovered?

- Symbolic knowledge representations
 - Decision trees
 - Rules
 - Relations / Logic forms (ILP)
 - Attribute generalizations
 - Probability distributions
 - Conceptual clusters and taxonomies
- Sub-symbolic
 - Artificial neural networks
 - Instance based learners
 - Functions / equations
- Others, ...

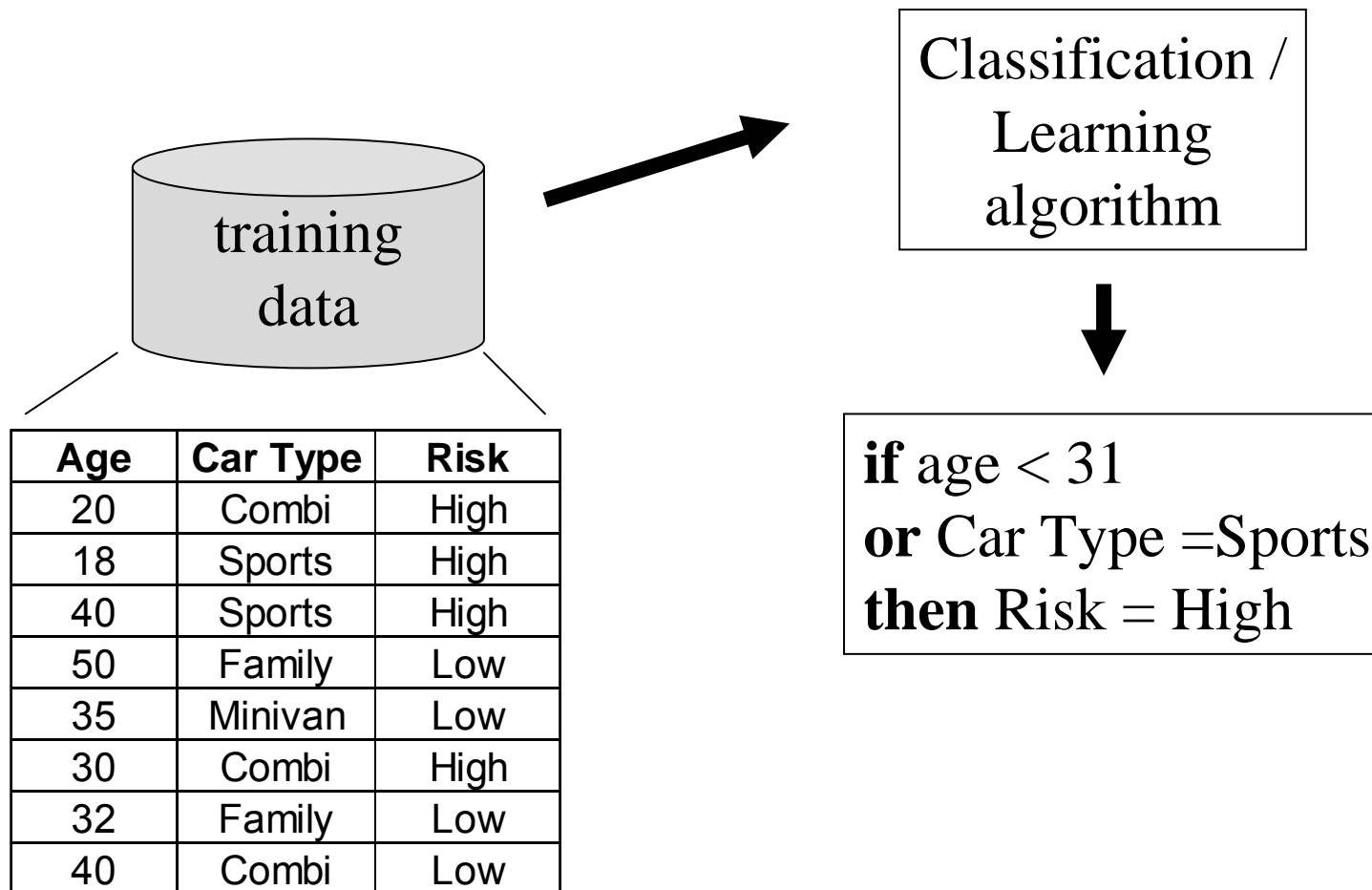
Decision trees

- Typical approach to the classification task.

Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Decision rules



Rules with variables - ILP

- Using variables and multiple relations:

```
If height_and_width_of(x,h,w) and h > w  
then standing(x)
```

- The top of a tower of blocks is standing:

```
If height_and_width_of(x,h,w) and h > w  
and is_top_of(x,y)  
then standing(x)
```

- The whole tower is standing:

```
If height_and_width_of(z,h,w) and h > w  
and is_top_of(x,z) and standing(y)  
and is_rest_of(x,y)  
then standing(x)  
  
If empty(x) then standing(x)
```

Numeric prediction – regression function

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

$$\text{PRP} = -55.9 + 0.0489 \text{ MYCT} + 0.0153 \text{ MMIN} + 0.0056 \text{ MMAX} \\ + 0.6410 \text{ CACH} - 0.2700 \text{ CHMIN} + 1.480 \text{ CHMAX}$$

Association rules

- Transaction data
- Market basket analysis



TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

- {Cheese, Milk} → Bread [sup=5%, conf=80%]
- Association rule:
„80% of customers who buy *cheese* and *milk* also buy *bread* and 5% of customers buy all these products together”

Sequential Patterns

- Sequential pattern mining is the extracting of frequently occurring patterns related to time or other sequences.
- A sequential rule: $A \rightarrow B \rightarrow C$, says that event A will be followed by event B and this by event C with a certain confidence
- An example:
“A customer who bought a TV three months ago is likely to order a new DVD player within one month”

Cluster

- A group of examples similar in the sense of some distance measure.
- Unsupervised data

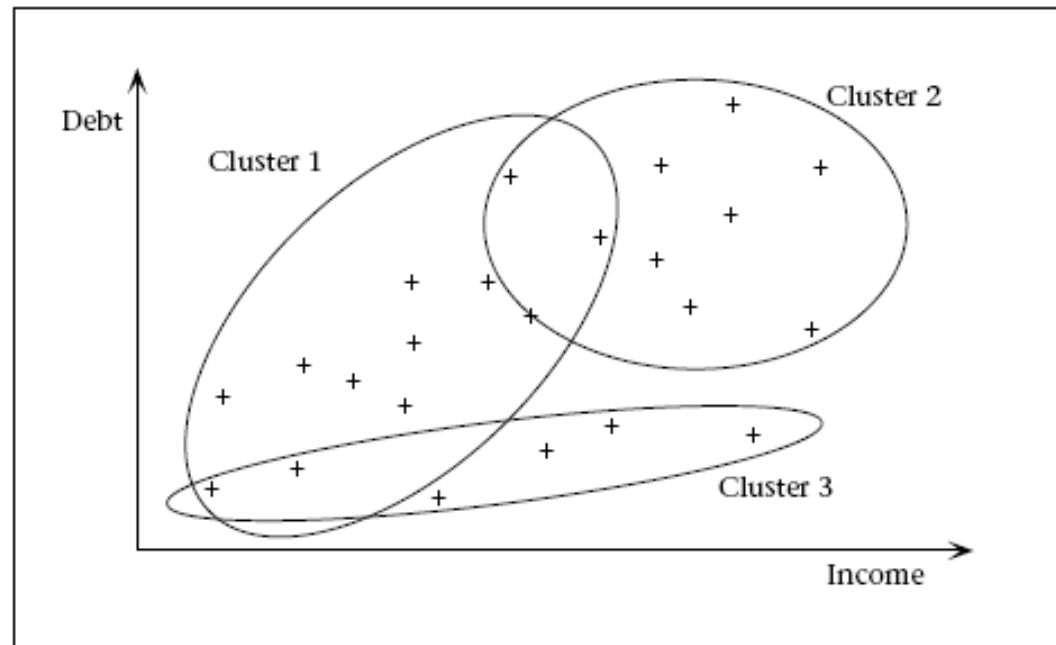
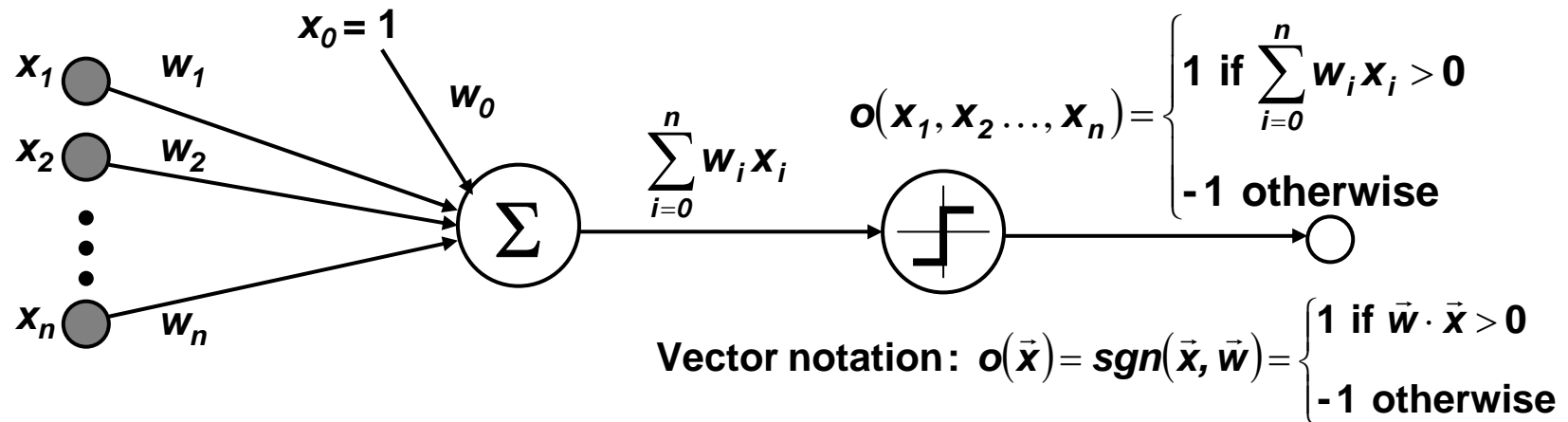


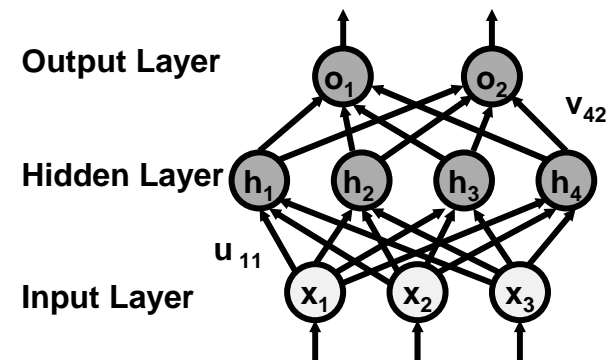
Figure 5. A Simple Clustering of the Loan Data Set into Three Clusters.

Note that original labels are replaced by a +.

Non –symbolic? Neural networks



- Single Neuron Model
 - Linear Threshold Unit (LTU)
inputs to unit: defined as linear combination
 - Output of unit: threshold (activation) function on net input (threshold $\theta = w_0$)
- Neural Networks
 - Neuron is modeled using a unit connected by weighted links w_i to other units
 - Multi-Layer Perceptron (MLP): future lecture

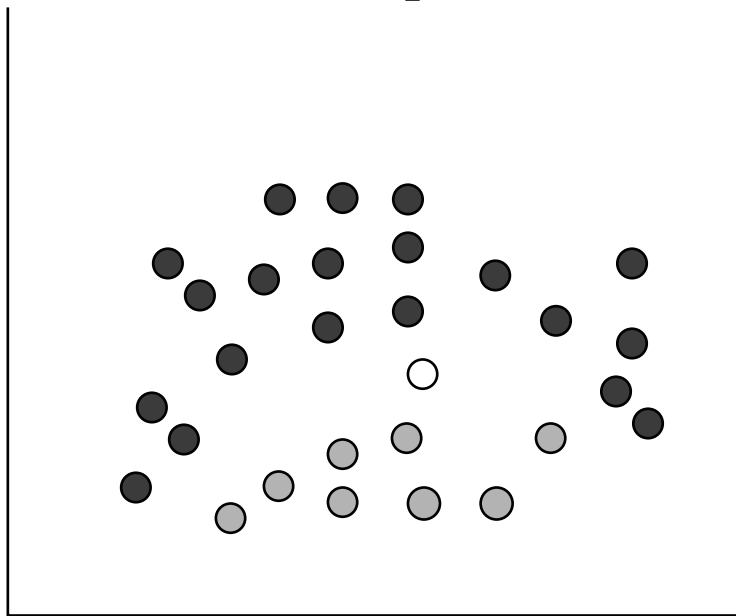


Major Data Mining Tasks

- **Classification:** predicting an instance class on the basis of its description.
- **Prediction:** predicting a continuous value.
- **Clustering:** finding similarity groups in data.
- **Associations:** e.g. A & B & C occur frequently.
 - Sequence analysis
- **Summarization:** describing a group.
- **Visualization:** graphical methods to show patterns in data.
- **Deviation or Anomaly Detection:** finding important changes or anomalies
- ...

Supervised classification

- Discover a method / function that maps a data element / instance into one of predefined classes
- Learning how to predict the instance class from pre-labeled (classified) examples - classifiers



Many approaches:

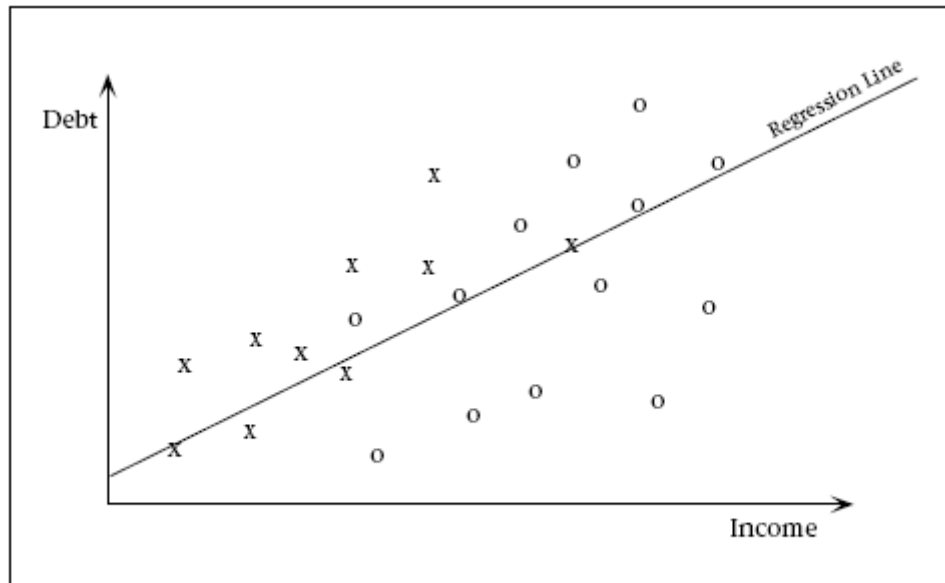
- Discriminant Analysis,
- Decision Trees or Rules,
- Bayesian Approaches,
- SVM

...

Given a set of points from classes ● ●
what is the class of new point ○?

Numeric prediction

- Supervised task, but “output” is numeric,
 - Scheme is being provided with target value
- Evaluation - good statistical background.

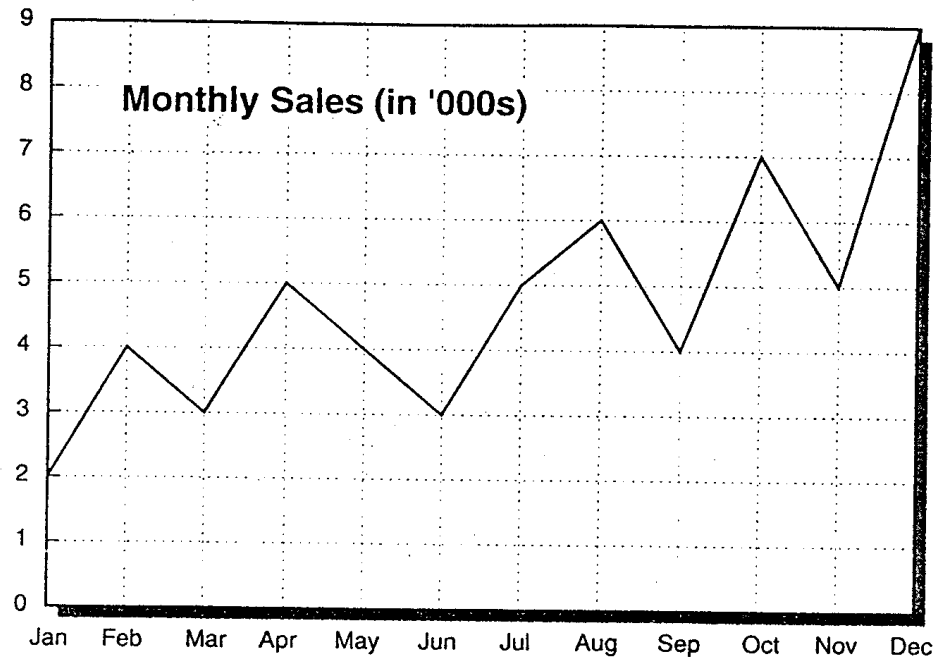


Outlook	Temp.	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
	Hot	High	True	0
	Hot	High	False	55
	Mild	Normal	False	40
...

Figure 4. A Simple Linear Regression for the Loan Data Set.

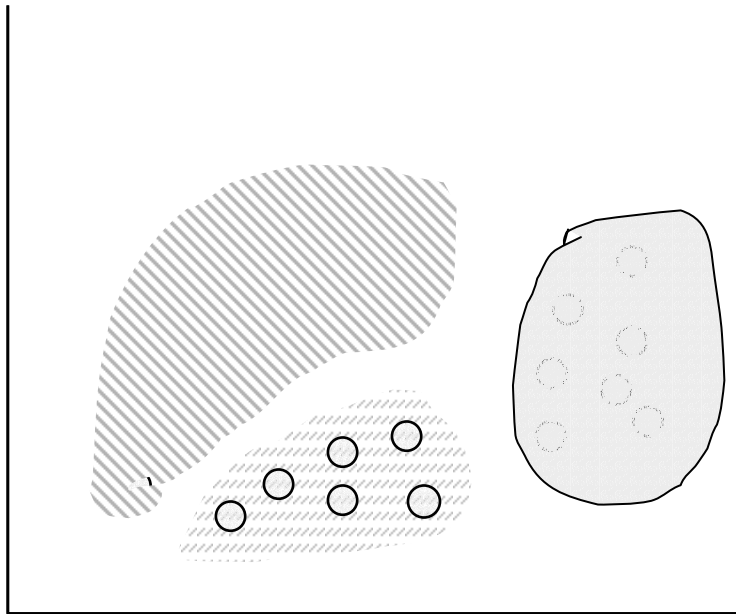
Time series – prediction of time stamped variable

- Predictive data mining – time stamped variable; typical example – stocks, financial data, production planning.



Clustering

Find “natural” grouping of un-labeled examples
– similar in some well defined sense.



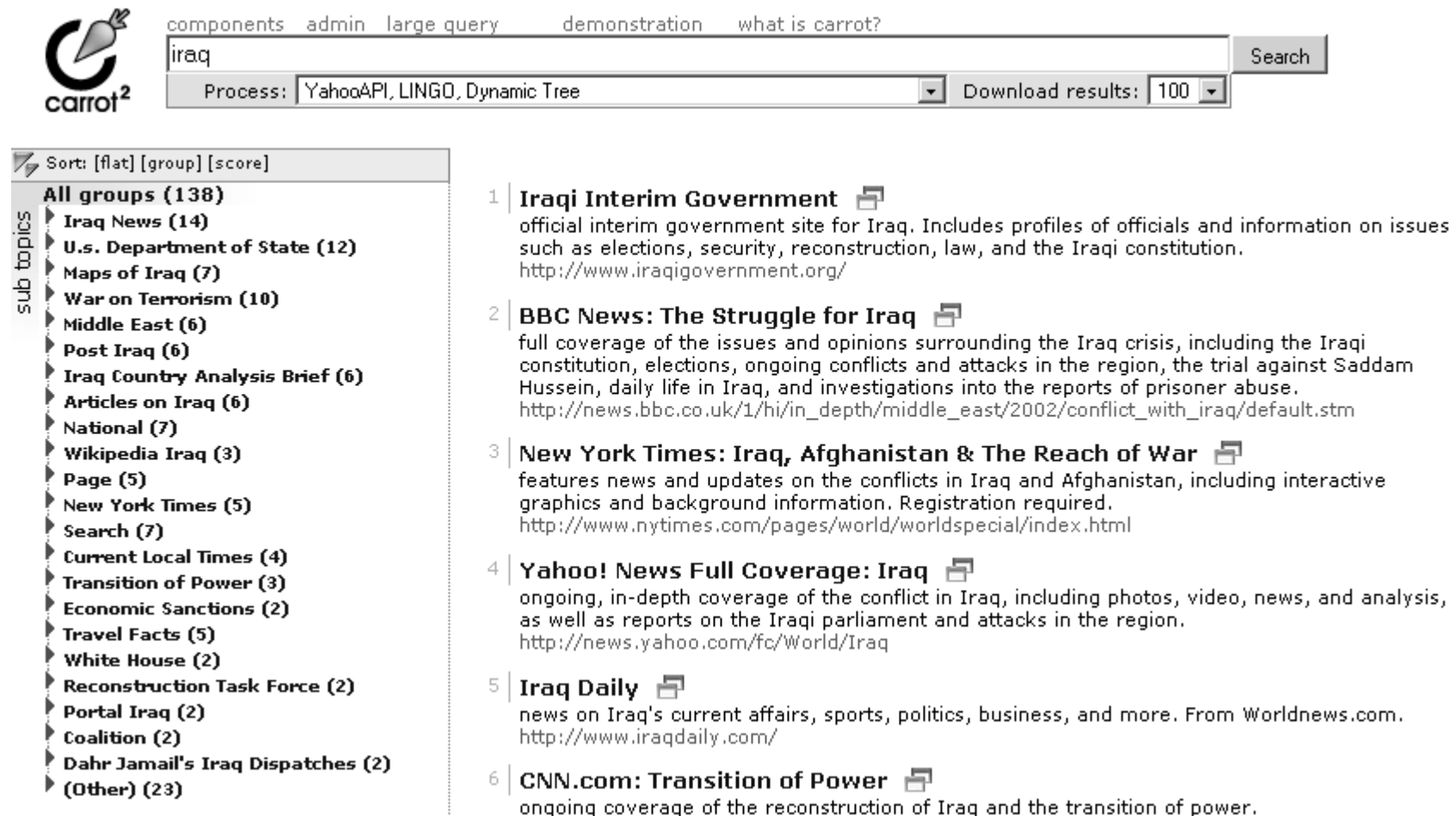
Clustering

- Examples: customer grouping
- Finding groups of items that are similar
- Clustering is *unsupervised*
 - The class of an example is not known
- Success often measured subjectively

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

Descriptive clustering – text mining

- Discovering diverse groups of semantically related documents described with meaningful, comprehensible and compact text labels → LINGO available in Carrot2 project.



The screenshot displays the Carrot2 web interface. At the top left is the Carrot2 logo. The search bar contains the text 'iraq'. Below the search bar, the 'Process' dropdown is set to 'YahooAPI, LINGO, Dynamic Tree' and the 'Download results' dropdown is set to '100'. A 'Search' button is located to the right of the search bar. Below the search bar, there is a sidebar on the left with a 'sub topics' label and a list of clusters. The main content area on the right shows a list of search results, numbered 1 through 6, each with a title, a brief description, and a URL.

components admin large query demonstration what is carrot?

iraq

Process: YahooAPI, LINGO, Dynamic Tree







Download results: 100

Search

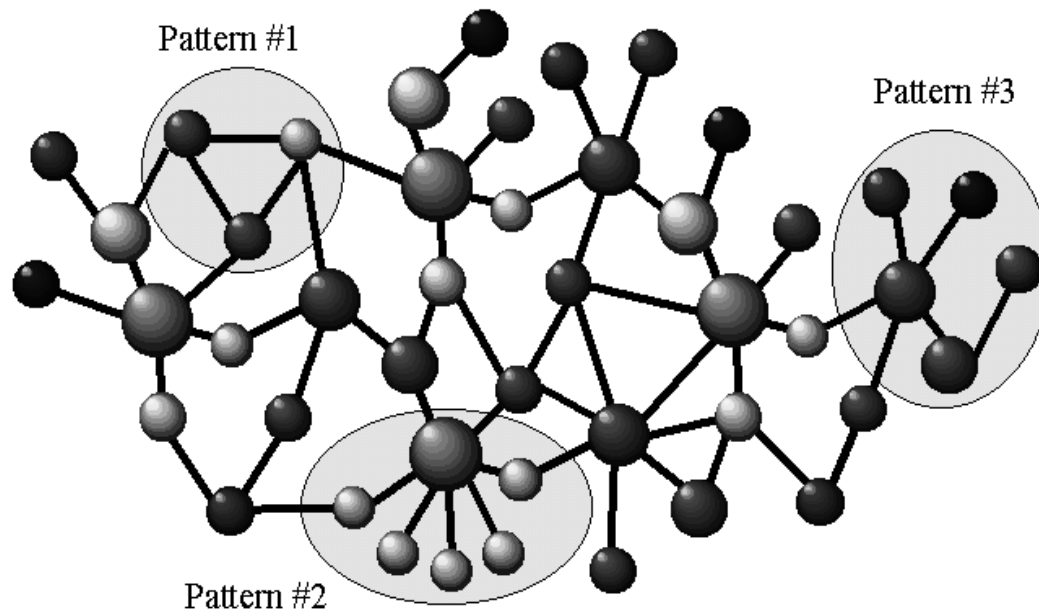
Sort: [flat] [group] [score]

sub topics

- All groups (138)
- Iraq News (14)
- U.s. Department of State (12)
- Maps of Iraq (7)
- War on Terrorism (10)
- Middle East (6)
- Post Iraq (6)
- Iraq Country Analysis Brief (6)
- Articles on Iraq (6)
- National (7)
- Wikipedia Iraq (3)
- Page (5)
- New York Times (5)
- Search (7)
- Current Local Times (4)
- Transition of Power (3)
- Economic Sanctions (2)
- Travel Facts (5)
- White House (2)
- Reconstruction Task Force (2)
- Portal Iraq (2)
- Coalition (2)
- Dahr Jamail's Iraq Dispatches (2)
- (Other) (23)

- Iraqi Interim Government** 
official interim government site for Iraq. Includes profiles of officials and information on issues such as elections, security, reconstruction, law, and the Iraqi constitution.
<http://www.iraqigovernment.org/>
- BBC News: The Struggle for Iraq** 
full coverage of the issues and opinions surrounding the Iraq crisis, including the Iraqi constitution, elections, ongoing conflicts and attacks in the region, the trial against Saddam Hussein, daily life in Iraq, and investigations into the reports of prisoner abuse.
http://news.bbc.co.uk/1/hi/in_depth/middle_east/2002/conflict_with_iraq/default.stm
- New York Times: Iraq, Afghanistan & The Reach of War** 
features news and updates on the conflicts in Iraq and Afghanistan, including interactive graphics and background information. Registration required.
<http://www.nytimes.com/pages/world/worldspecial/index.html>
- Yahoo! News Full Coverage: Iraq** 
ongoing, in-depth coverage of the conflict in Iraq, including photos, video, news, and analysis, as well as reports on the Iraqi parliament and attacks in the region.
<http://news.yahoo.com/fc/World/Iraq>
- Iraq Daily** 
news on Iraq's current affairs, sports, politics, business, and more. From Worldnews.com.
<http://www.iraqdaily.com/>
- CNN.com: Transition of Power** 
ongoing coverage of the reconstruction of Iraq and the transition of power.

Another problem: Link Analysis



Can we find patterns in the network/ graph structure or
Can we identify communities of Web users?

Goals of Data Mining – S.Weiss and U.Fayyad views

Prediction	Description
Classification Regression Time series	Deviation Detection Clustering Association Rules Summarization Visualization Text mining

- Two primary goals of data mining
 - **Prediction** : using some variables or fields in database to predict unknown of future value of other variables.
 - strongest and practical goal of data mining and other fields as pattern recognition, machine or statistical learning
 - **Knowledge description** : finding human-interpretable patterns describing data
 - Becoming more and more important for KDD
- Technical methods for solving these problems often overlap.

Components of Data Mining Algorithms

- Model or pattern structure → Knowledge representation; language determining the underlining structure forms that we seek from the data.
- Score function → estimates how well a particular patterns (or a model with its parameters) meet the criteria of KDD process.
- Optimization of parameters and search method:
 - Searching for the parameters which optimize the score function (quality of the model) → optimal estimation vs. heuristic or greedy iterative ..
 - Searching over different model or pattern spaces
- Data management strategy → handling data access efficiently during search/optimization

More in Hand et al. book.

Why Data Mining? -- Potential Applications

- Database analysis and decision support
 - Market and customer analysis; analytical CRM
 - target marketing and advertising, customer relation management, market basket analysis, cross selling, market segmentation.
 - Risk analysis and management
 - Forecasting, customer changes, quality control, loan approval.
 - Diagnostics (e.g. technical conditions of objects)
 - Fraud detection
- Other Applications:
 - Text mining (news group, email, documents) and Web analysis; Search engines; Adaptive Web servers and e-commerce systems



Customer Changes: Case Study

- Situation: Attrition rate at for mobile phone customers is around 25-30% a year (US)!

Task:

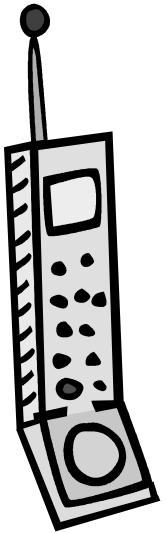
- Given customer information for the past N months, predict who is likely to attrite next month.
- Also, estimate customer value and what is the cost-effective offer to be made to this customer.



Customer Attrition Results

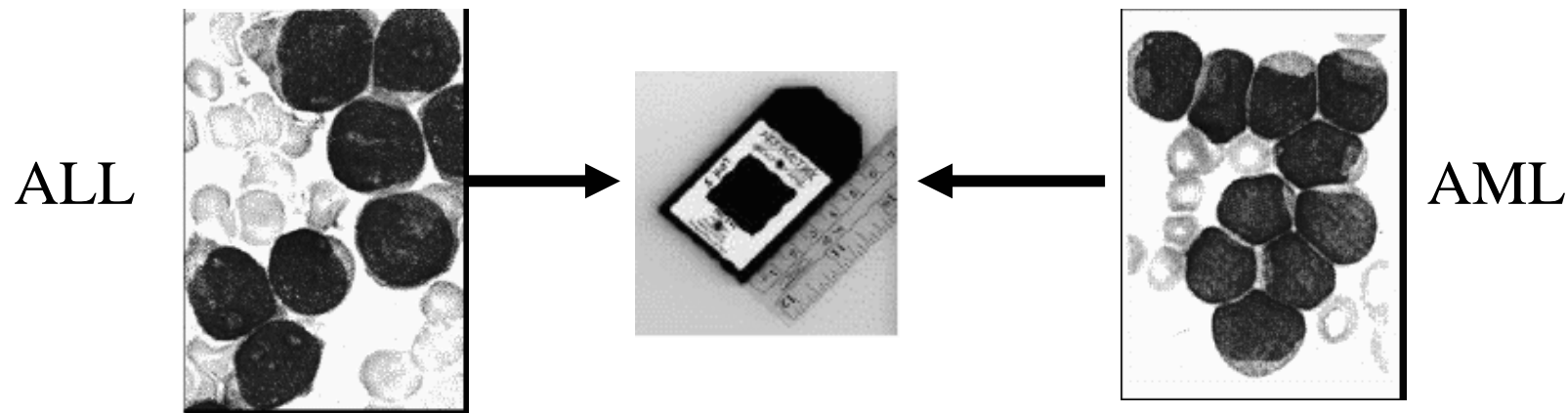
- Verizon Wireless (US) built a customer data warehouse
- Identified potential attriters
- Developed multiple, regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

(Reported in 2003)



Biology: Molecular Diagnostics

- Leukemia: Acute Lymphoblastic (ALL) vs Acute Myeloid (AML)
 - 72 samples, about 7,000 genes



Results: 33 correct (97% accuracy),
1 error (sample suspected mislabelled)

Outcome predictions?

Problems Suitable for Data-Mining

- Require knowledge-based decisions.
- Have a changing environment.
- Have accessible, sufficient, and relevant data
- Data are difficult or complex.
- Problems are not trivial, cannot be solved „manually” by humans.
- Provides high payoff for the right decisions!

Privacy considerations important if personal data is involved!

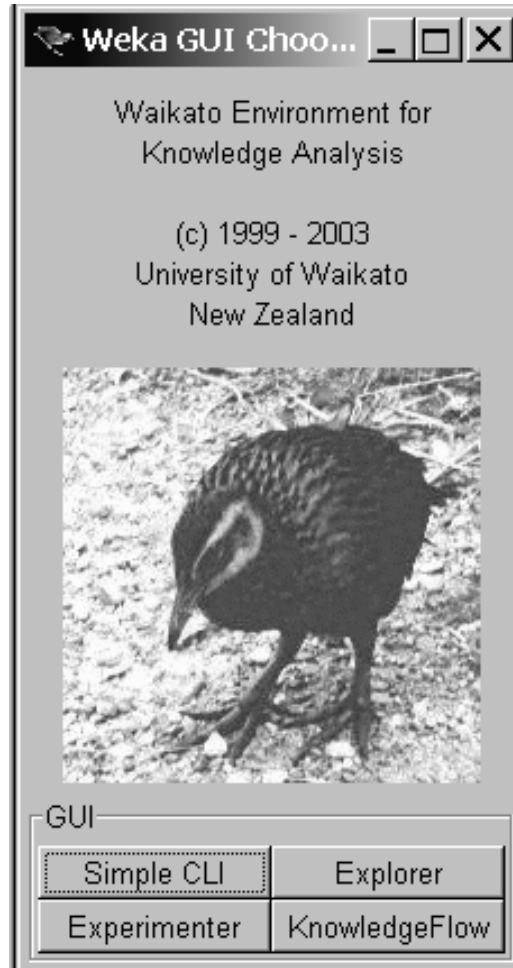
Warnings!

- Historically data mining was used in a pejorative sense by statisticians for the idea that, if you search long enough, you can always find some model to fit your data arbitrarily well.

Example:

- David Rhine, a "parapsychologist" at Duke in the 1950's tested students for "extrasensory perception", by asking them to guess 10 cards—red or black. He found about 1/1000 of them guessed all 10, and instead of realizing that that is what you would expect from random guessing, declared them to have ESP. When he retested them, he found they did no better than average. His conclusion: telling people they have ESP causes them to lose it! Quote from Jeffrey Ullman, Stanford

Weka – software for data mining



- Waikato Environment for Knowledge Analysis (WEKA); developed by the Department of Computer Science, University of Waikato, New Zealand
- Data mining / Machine learning software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications

<http://www.cs.waikato.ac.nz/ml/weka/>

- Ian Witten, Eibe Frank

WEKA - Tools

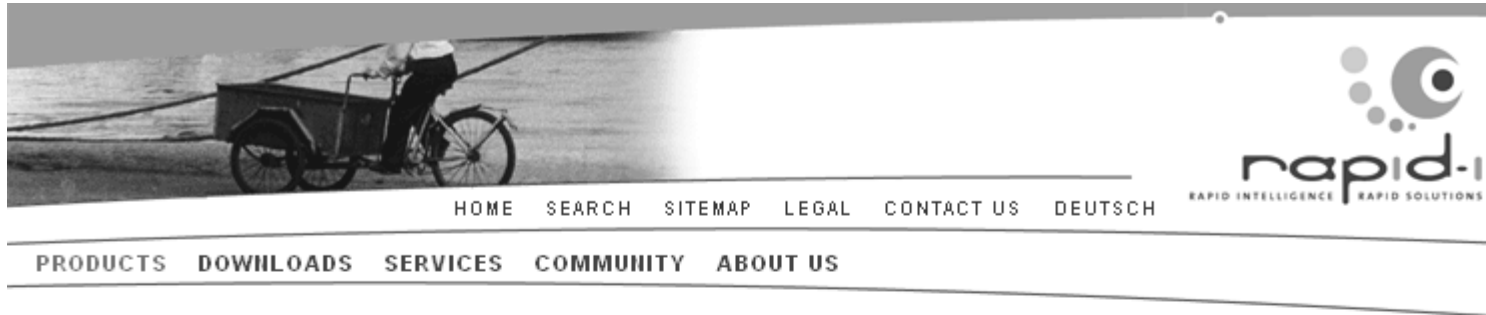
- Pre-processing Filters
- Attribute selection
- Classification/Regression
- Clustering
- Association discovery
- Visualization



Examples of Systems for Data Mining

- IBM: QUEST and Intelligent Miner
- Silicon Graphics: MineSet
- SAS Institute: Enterprise Miner
- SPSS / Integral Solutions Ltd.: Clementine
- Oracle Miner
- Rapid Miner (YALE)
- Orange
- Other systems
 - Information Discovery Inc.: Data Mining Suite
 - SFU: DBMiner, GeoMiner, MultiMediaMiner

RapidMiner (YALE)



[HOME](#) [SEARCH](#) [SITEMAP](#) [LEGAL](#) [CONTACT US](#) [DEUTSCH](#)

[PRODUCTS](#) [DOWNLOADS](#) [SERVICES](#) [COMMUNITY](#) [ABOUT US](#)

TESTIMONIALS

"I have encountered various learning environments, but none so broad, powerful, and easy-to-use as RapidMiner / YALE. Many of us who are not skilled in programming are thankful."

Roberto E. Ferrer, Venezuela

DOWNLOADS

[RapidMiner / YALE](#)

[RapidMiner / YALE Plugins](#)

[RapidMiner / YALE Documentation](#)

[RapidMiner / YALE Interactive Tour](#)

TRAINING SEMINARS

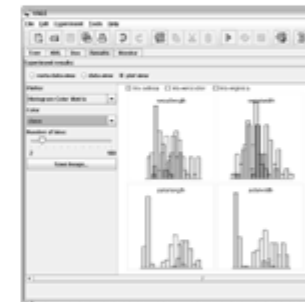
[Data Mining for Marketing and Customer Service](#)

[Data Mining Techniques: Theory and Practice](#)


[Extending RapidMiner and Integration as a Data](#)

[HOME](#) : [PRODUCTS](#) : [RAPIDMINER \(YALE\)](#) : [SCREENSHOTS](#)
RAPIDMINER / YALE SCREENSHOTS

This web page provides a selection of screenshots for RapidMiner (formerly YALE). These pictures might help you to get a first impression of the abilities of RapidMiner. This page contains a large number of images. Please be patient until all pictures were loaded.



Orange (Slovenia)



orange
DATA MINING
FRUITFUL&FUN

- Home
- Screenshots
- Contact & Support
- Acknowledgements
- Download
- Forum (RSS)
- Documentation
- Search
- Visual Programming
- Catalog of Widgets
- Scripting for Beginners
- Class Reference
- Modules
- Example Scripts
- Data Sets

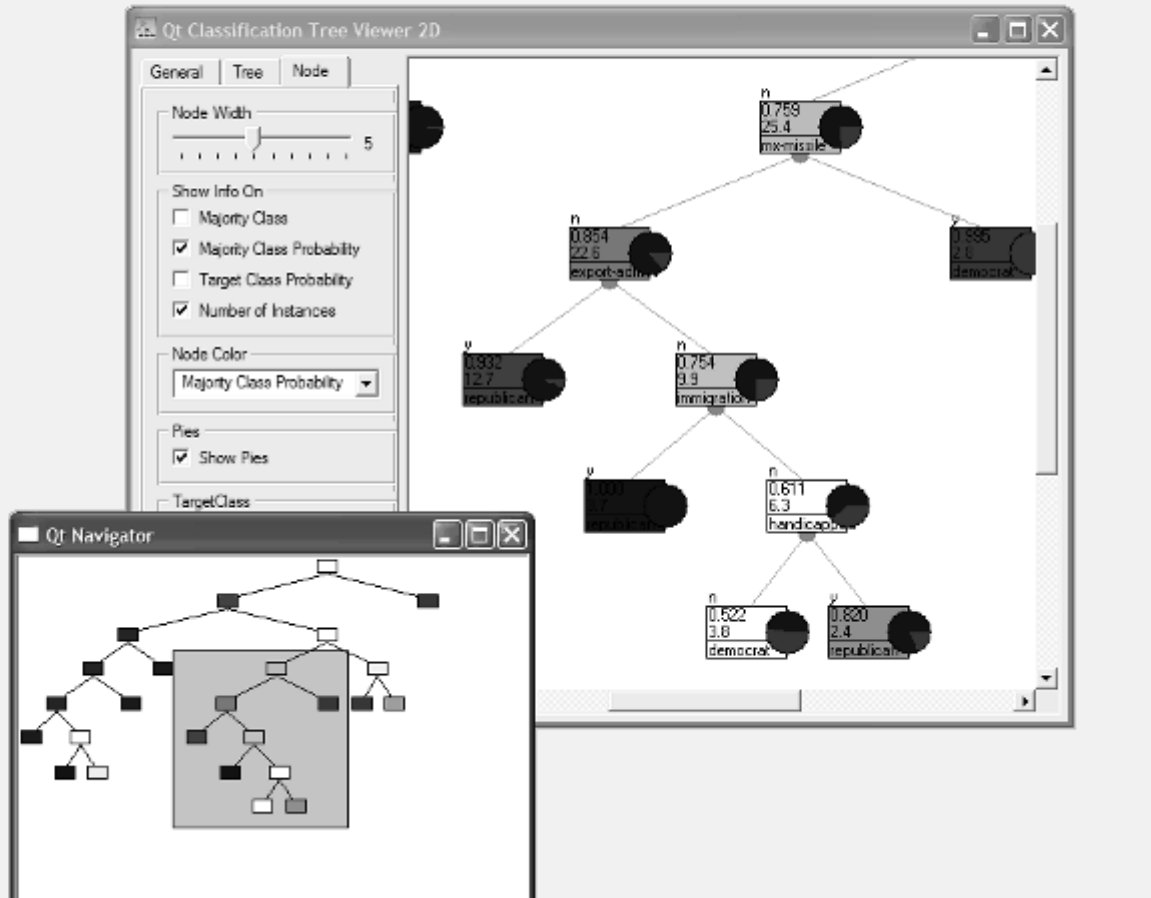
Latest News

Oct 31: The list of example scripts from documentation works again. For instance, you want to know how to induce random forests in

Orange Screenshots

Following are screenshots of Orange Widgets and Orange's visual programming interface for data mi

Classification tree viewer with a navigator.



Statsoft – Statistica Data Miner

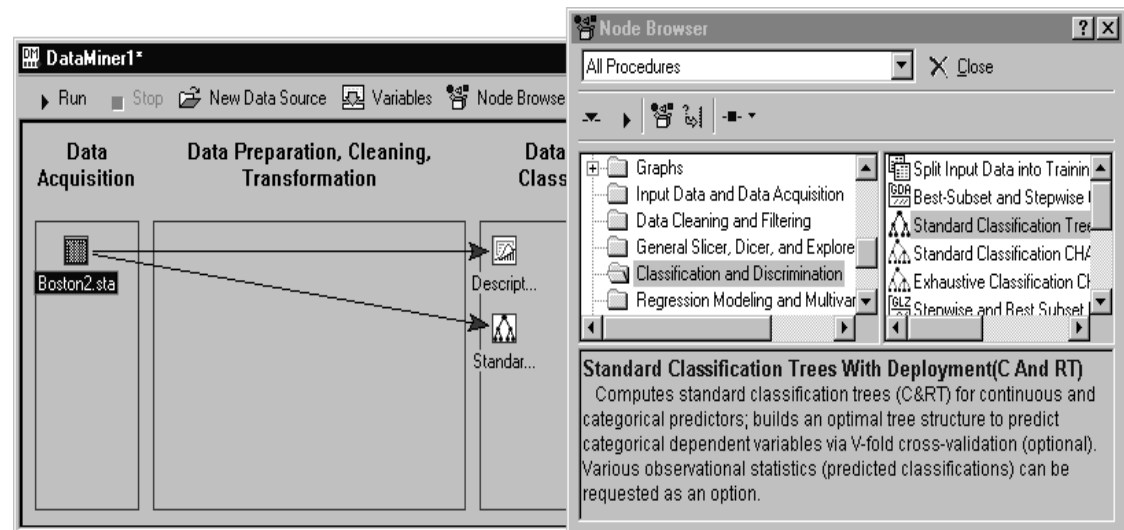
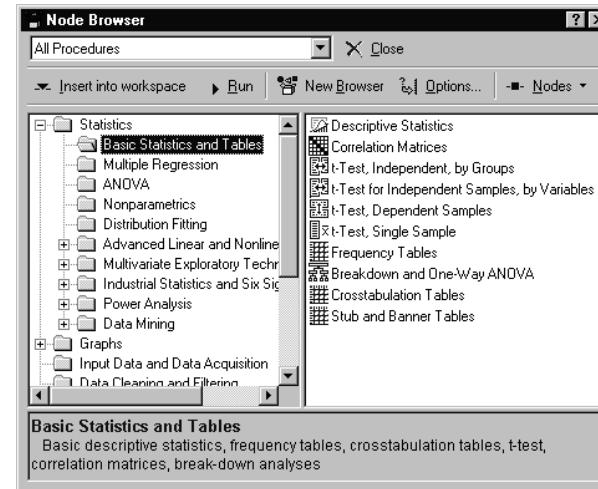
- Data Miner - My Procedures
- Data Miner - All Procedures

- Data Miner - Data Cleaning and Filtering

- Data Miner - General Slicer/Dicer Explorer with Drill-Down
- Data Miner - General Classifier (Trees and Clusters)
- Data Miner - General Modeler and Multivariate Explorer
- Data Miner - General Forecaster
- Data Miner - General Neural Network Explorer

- Neural Networks
- Generalized EM & k-Means Cluster Analysis
- Association Rules
- General Classification/Regression Tree Models
- General CHAID Models
- Interactive Trees (C&RT, CHAID)
- Boosted Tree Classifiers and Regression
- Generalized Additive Models
- MAR Splines (Multivariate Adaptive Regression Splines)

- Rapid Deployment of Predictive Models (PMML)
- Goodness of Fit, Classification, Prediction
- Feature Selection and Variable Screening



Data mining: what is on the market?

- Consultancy
 - There are some people which offer data mining consulting.
- Software
 - SAS, Statistica, Oracle, SGI Miner, IBM intelligent miner, ...

www.KDnuggets.com - various references

Address <http://www.kdnuggets.com>

KDnuggets™ Data Mining, Knowledge Discovery, Genomic Mining, Web Mining
[Data Mining Consulting](#) | [Data Mining Jobs](#) | [Advertising](#) | [Site Map](#)

CLEMENTINE 7.0 = POWER, PREDICTION, PRODUCTIVITY
SPSS Clementine 7.0 - The next generation of Data Mining

Free Webinar:
[Why Use Predictive Analytics?](#)

KDnuggets News, the Data Mining & Knowledge Discovery newsletter: data mining news, jobs, software, courses, ...
[2003 issues](#) | [Schedule](#) | [Archive](#) | [Submit](#) | [Subscribe!](#)

Current Issue: NEW! [03:19, Oct 14, 2003: Data preparation; NSF deadline; ICDM-2003, Nov 19-22 ... \(29 items\)](#)

Match in: [help](#)

Software: [Classification](#), [Suites](#), [Text](#)
Jobs: [Industry](#), [Academic](#)

Solutions: [Bioinformatics](#), [CRM](#), [Web](#)
Courses: [Oct](#), [Nov](#), [Dec](#)
[Education](#)

Companies: [...](#)
Meetings: [...](#)

Insightful Miner
Easy to Use & Extensible Data Mining

- Build predictive models easily
- Modern visual interface
- Advanced analytic methods
- Scalable capabilities

Free Webcast & Whitepaper!

Insightful Miner
Easy to Use & Extensible Data Mining

Poll

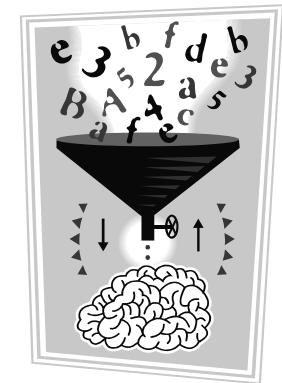
How frequently do you do a separate feature selection in classification (rather than have a learning algorithm do selection)

Always
 Most of the time
 Frequently
 Rarely
 Never

[View Results](#)

Summary:

- Technology trends lead to data flood
 - data mining is needed to make sense of data
- Data Mining has many applications, successful and not
- Knowledge Discovery Process
- Data Mining Tasks
 - classification, prediction, clustering, associations, ...



Short Break



Lecture 1b.

Few remarks on
the nature of data to be mined

Data Mining: On What Kind of Data?

- **Flat files** → **Attribute-value data tables**
- Relational databases and data warehouses → relations.
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, ...)
- Sequence data bases
- Other advanced data repositories
 - Object-oriented and object-relational databases
 - Spatial databases
 - Time-series data and temporal data
 - Text databases and multimedia databases
 - WWW resources
 - ...



Flat files

- Actually the most common data source for data mining, especially at the research level.
- Simple data files in text or binary format with a structure known by the data mining algorithm to be applied.
- The data in these files can be transactions, time-series data, scientific measurements, etc.
- Big data – efficiency of access and management.

Instance	f_1	...	f_k	Y
x_1	$V_{1,1}$...	$V_{1,k}$	$V_{1,k+1}$
...
x_i	$V_{i,1}$...	$V_{i,k}$	$V_{i,k+1}$
...
x_n	$V_{n,1}$...	$V_{n,k}$	$V_{n,k+1}$

Typical Input → data tables

concepts-classes, instances, attributes

- A data table is a set of measurements taken from some environment of process ($n \times k$ table)
- In general, the input takes the form of: concepts-classes, instances, attributes.
 - Target concepts (classes): kinds of things that can be discovered
 - Aim: intelligible and operational concept description.
 - Instances: the individual, independent examples of a concept
 - „Objects” of the same type.
 - Note: more complicated forms of input are possible.
 - Attributes: (features) measuring aspects of an instance
 - Could be defined on various measurement scales – however we will focus mainly on nominal and numeric ones.
 - In statistics → variables.

Again how to get a data table?

- Instance: specific type of example
 - Thing to be classified, associated, or clustered
 - Individual, independent example of target concept
 - Characterized by a predetermined set of attributes
- Input to learning scheme: set of instances / data table
 - Represented as a single relation/flat file
- Rather restricted form of input
 - No relationships between objects
- Most common form in practical data mining
- Process of flattening a file “denormalization”
 - Several relations are joined together to make one
- Possible with any finite set of finite relations

The weather problem (Quinaln's play sport)

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Given past data,
Can you come up
with the rules for
Play/Not Play ?

Classics in ML



Types of attributes

- The most common distinction comes from measurement scale and statistics:
 - Nominal (also binary)
 - Ordinal
 - Interval-scaled
 - Ratio-scaled.
- Other names:
 - Categorical vs. numeric/continuous ones.
- Other types:
 - Criteria (preference-ordered), hierarchical, ...

Transforming text documents into a standard form

- Transformation into Vector Representation

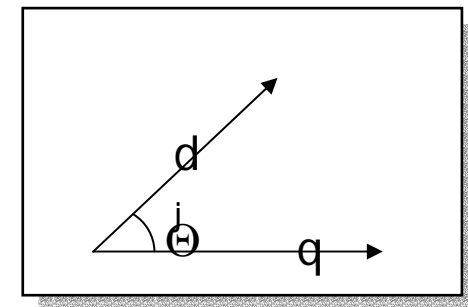
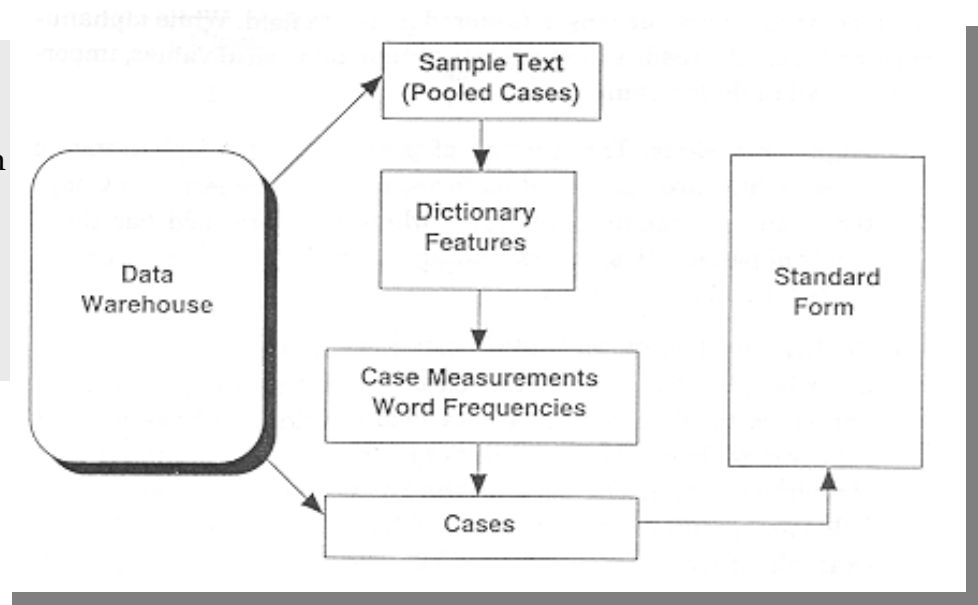
The $d=7$ documents:

- D1: Large Scale Singular Value Computations
- D2: Software for the Sparse Singular Value Decomposition
- D3: Introduction to Modern Information Retrieval
- D4: Linear Algebra for Intelligent Information Retrieval
- D5: Matrix Computations
- D6: Singular Value Analysis of Cryptograms
- D7: Automatic Information Organization

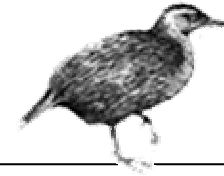
The $t=5$ terms:

- T1: Information
- T2: Singular
- T3: Value
- T4: Computations
- T5: Retrieval

$$A = \begin{pmatrix} 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 1.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$



WEKA - the ARFF format



```
%  
% ARFF file for weather data with some numeric features  
%  
@relation weather  
  
@attribute outlook {sunny, overcast, rainy}  
@attribute temperature numeric  
@attribute humidity numeric  
@attribute windy {true, false}  
@attribute play? {yes, no}  
  
@data  
sunny, 85, 85, false, no  
sunny, 80, 90, true, no  
overcast, 83, 86, false, yes  
...
```

Attribute types in WEKA

- ARFF supports numeric and nominal attributes.
- Interpretation depends on learning scheme:
 - Numeric attributes are interpreted as
 - interval scales if less-than and greater-than are used,
 - ratio scales if distance calculations are performed (normalization/standardization may be required).
 - Instance-based schemes define distance between nominal values (0 if values are equal, 1 otherwise)
- Integers: nominal, ordinal, or ratio scale?

Any questions, remarks?

