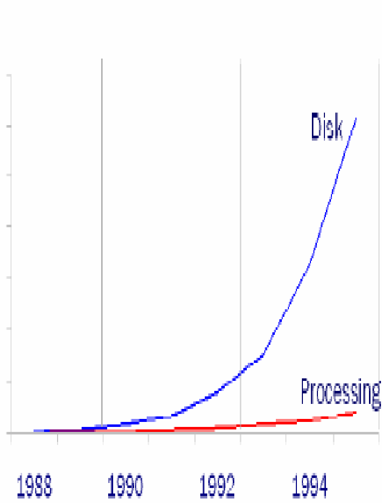
Knowledge Discovery Process and Data Mining - Final remarks

Lecturer: JERZY STEFANOWSKI
Institute of Computing Sciences
Poznan University of Technology
Poznan, Poland

Lecture 14
SE Master Course
2008/2009

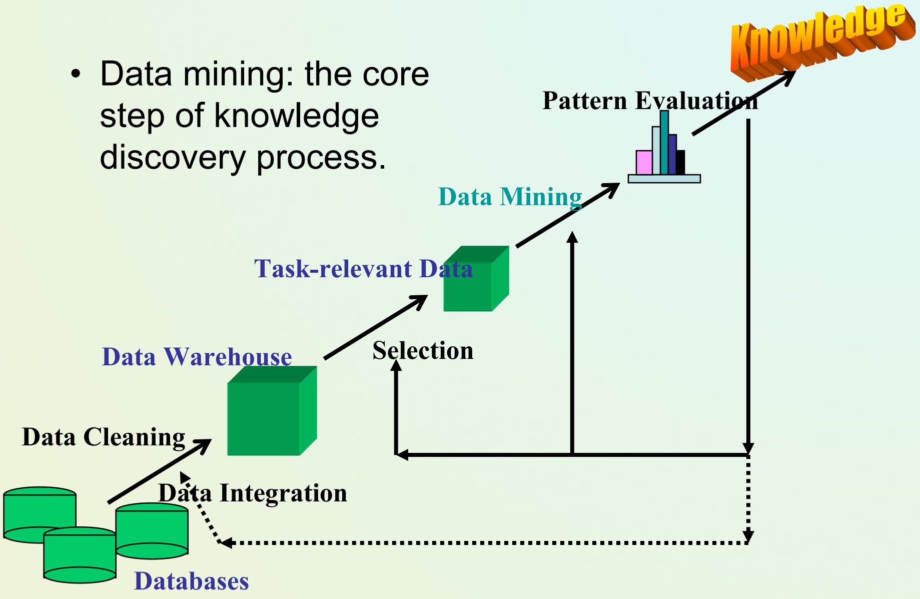
Growth Trends

- Moore’s law
  - Computer Speed doubles every 18 months
- Storage law
  - total storage doubles every 9 months
- Consequence
  - very little data will ever be looked at by a human
- Knowledge Discovery is **NEEDED** to make sense and use of data.



Data Mining a step in A KDD Process

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



Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleaning

Data Integration

Databases

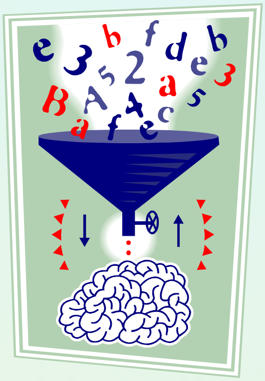
Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing
- Data reduction and projection:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Interpretation: analysis of results.
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Interacting with a user / expert in KDD

- KDD is not a fully automatically way of analysis.
- The user is an important element in KDD process.
- Should decide about, e.g.
  - Choosing task and algorithms, selection in preprocessing.
- Interpretation and evaluation of patterns
  - Objective interestingness measures,…
  - Subjective,…
- By definition, KDD may have several iterations.

Data Preparation for Knowledge Discovery



A crucial issue: The majority of time / effort is put there.

Data Understanding: Relevance

- What data is available for the task?
- Is this data relevant?
- Is additional relevant data available?
- How much historical data is available?
- Who is the data expert ?

Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
  - Object-relational database
  - Spatial and temporal data
  - Time-series data
  - Stream data
  - Multimedia database
  - Heterogeneous and legacy database
  - Text databases & WWW

Are All the “Discovered” Patterns Interesting?

- Data mining 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.

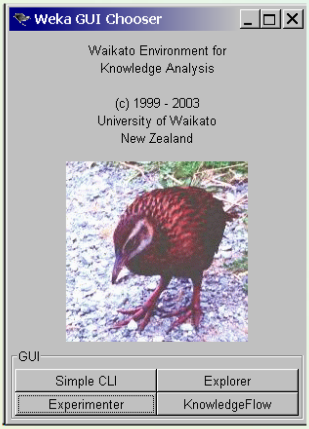
Can We Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering
- Search for only interesting patterns: An optimization problem
  - 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

Examples of Systems for Data Mining

- IBM: QUEST and Intelligent Miner
- Silicon Graphics: MineSet
- SAS Institute: Enterprise Miner
- Statistica Data Miner
- SPSS / Integral Solutions Ltd.: Clementine
- Oracle 9i Miner
- Rapid Miner (YALE)
- Orange
- Other systems
  - Information Discovery Inc.: Data Mining Suite
  - SFU: DBMiner, GeoMiner, MultiMediaMiner

WEKA – Machine Learning and Data Mining



Weka GUI Chooser

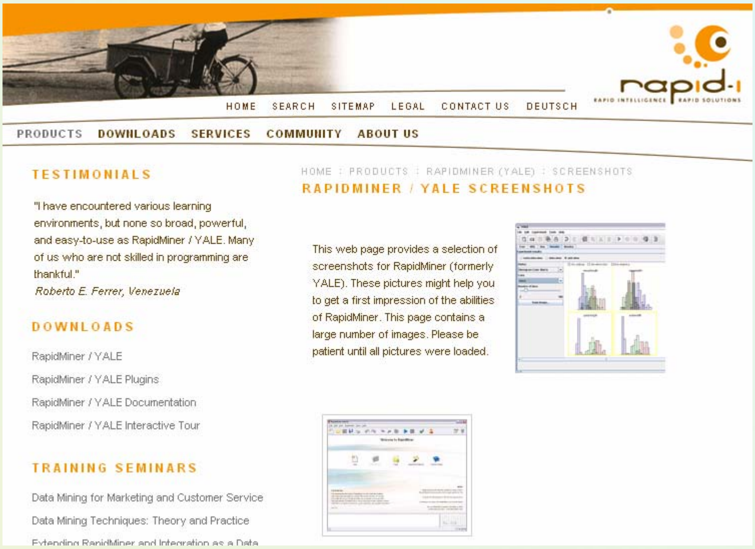
Waikato Environment for
Knowledge Analysis

(c) 1999 - 2003
University of Waikato
New Zealand

Java implementation
of many algorithms

No ideal solutions → but …

RapidMiner (YALE)



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.

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

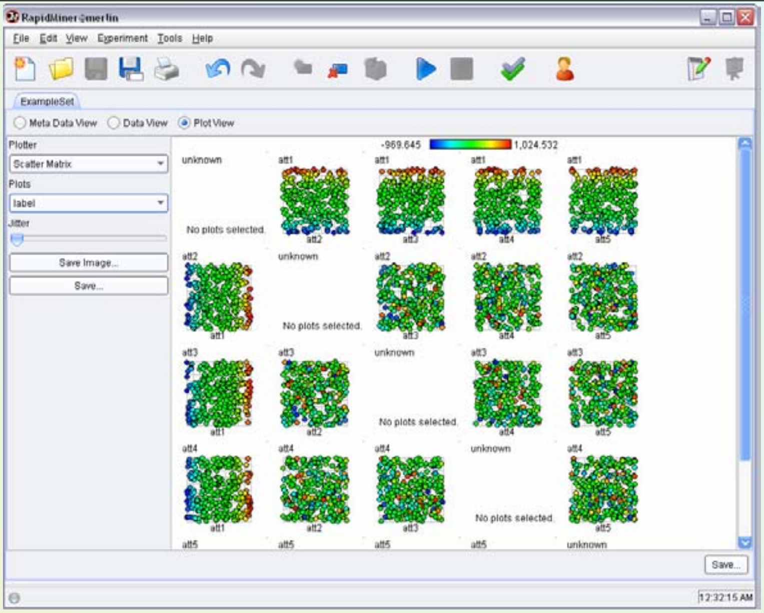
TRAINING SEMINARS

Data Mining for Marketing and Customer Service

Data Mining Techniques: Theory and Practice

Extending RapidMiner and Integration as a Data

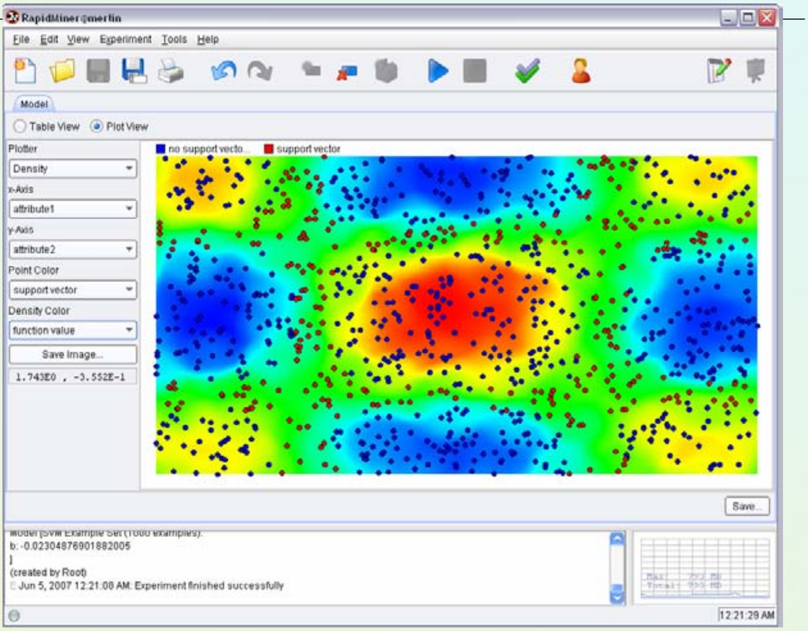
Some Rapidminer screenshots



Plotter: Scatter Matrix

Plots: label

SOM



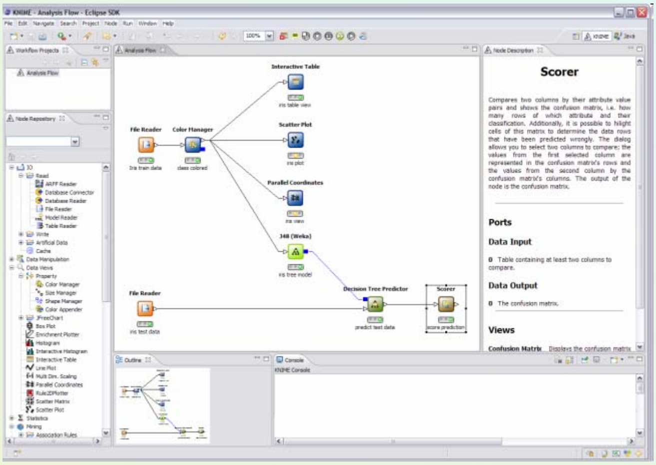
Plotter
Density
x-Axis
attribute1
y-Axis
attribute2
Point Color
support vector
Density Color
function value

1.743E0 , -3.552E-1

```
Model [SVM Example Set (1000 examples):
b: -0.02304876901882005
]
(created by Root)
Jun 5, 2007 12:21:08 AM: Experiment finished successfully
```

KNIME

- KNIME was developed (and will continue to be expanded) by the Chair for Bioinformatics and Information Mining at the University of Konstanz, Germany.
- It integrates all analysis modules of the well known Weka data mining environment and additional plugins allow R-scripts to be run, offering access to a vast library of statistical routines.



Scorer

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification. Additionally, it is possible to hilight cells of this matrix to determine the data rows that have been predicted wrongly. The dialog allows you to select two columns to compare; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix.

Ports

Data Input

0 Table containing at least two columns to compare.

Data Output

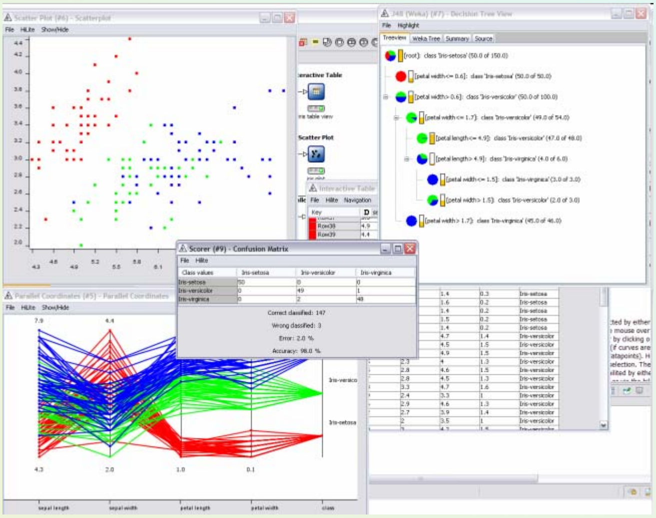
0 The confusion matrix.

Views

Confusion Matrix Displays the confusion matrix

KNIME - **An Example of Data Analysis Workflow**

- More http://www.knime.org/



J48 (Weka) (#7) - Decision Tree View

File Highlight

Treeview | Weka Tree | Summary | Source

```
[root]: class 'Iris-setosa' (50.0 of 150.0)
  [petal width<= 0.6]: class 'Iris-setosa' (50.0 of 50.0)
  [petal width> 0.6]: class 'Iris-versicolor' (50.0 of 100.0)
    [petal width<= 1.7]: class 'Iris-versicolor' (49.0 of 54.0)
      [petal length<= 4.9]: class 'Iris-versicolor' (47.0 of 48.0)
      [petal length> 4.9]: class 'Iris-virginica' (4.0 of 6.0)
        [petal width<= 1.5]: class 'Iris-virginica' (3.0 of 3.0)
        [petal width> 1.5]: class 'Iris-versicolor' (2.0 of 3.0)
    [petal width> 1.7]: class 'Iris-virginica' (45.0 of 46.0)
```

Scorer (#9) - Confusion Matrix

File Hilite

Class values	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	50	0	0
Iris-versicolor	0	49	1
Iris-virginica	0	2	48

Correct classified: 147

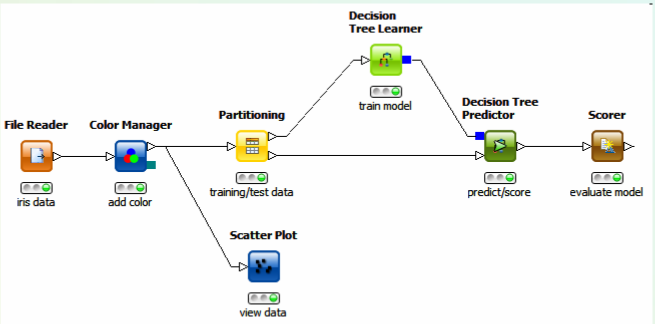
Wrong classified: 3

Error: 2.0 %

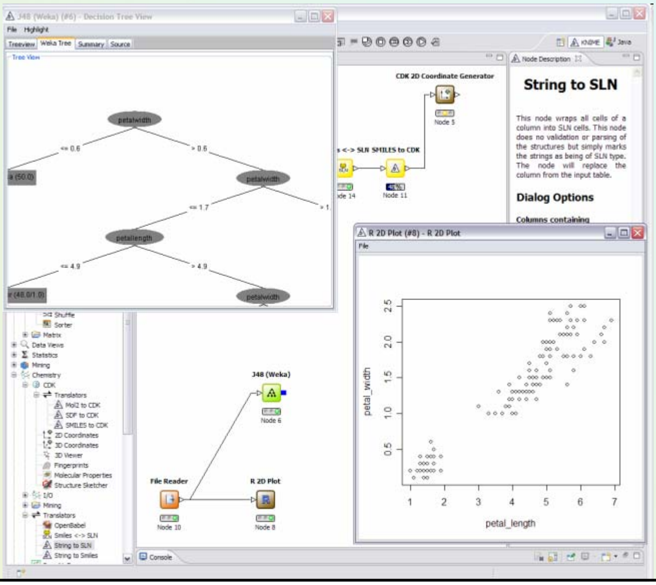
Accuracy: 98.0 %

Developing Trees

- Node flows



KNIME working with decision trees



String to SLN

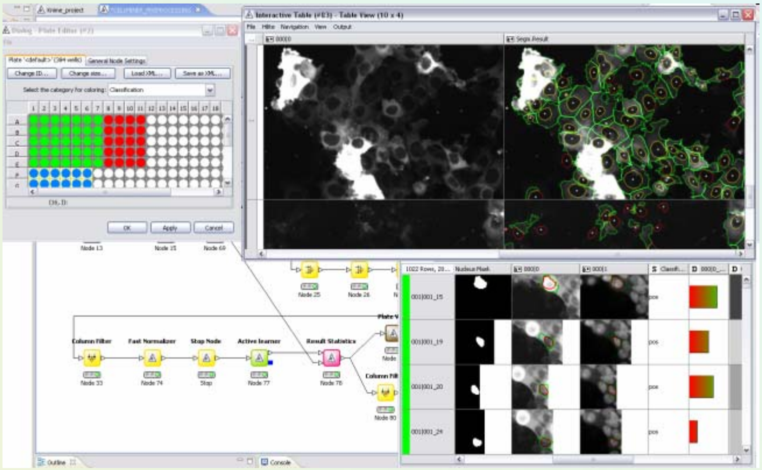
This node wraps all cells of a column into SLN cells. This node does no validation or parsing of the structures but simply marks the strings as being of SLN type. The node will replace the column from the input table.

Dialog Options

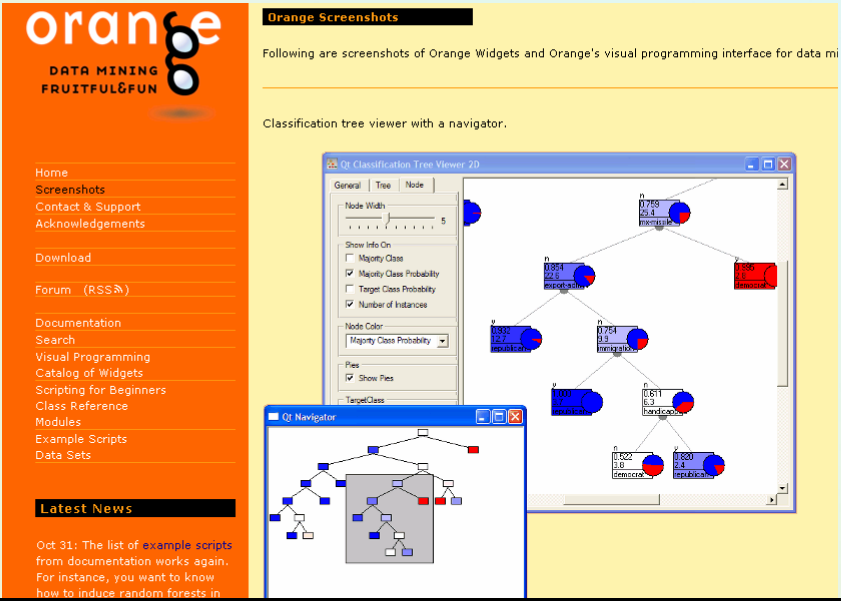
Columns containing

Cell Miner

- KNIME has been used to analyze cell images.



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.

Qt Classification Tree Viewer 2D

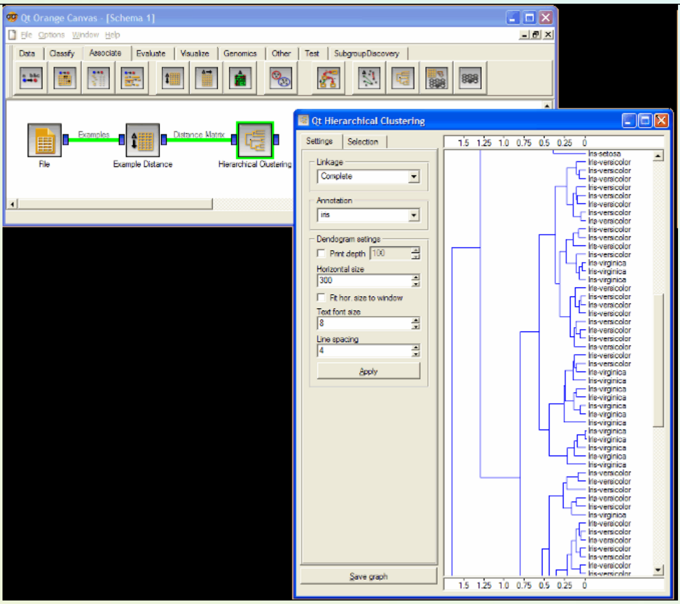
Node Width: 5

Show Info On:
- ☐ Majority Class
- ☑ Majority Class Probability
- ☐ Target Class Probability
- ☑ Number of Instances

Node Color: Majority Class Probability

Pies: ☑ Show Pies

Orange - clustering



Qt Orange Canvas - [Schema 1]

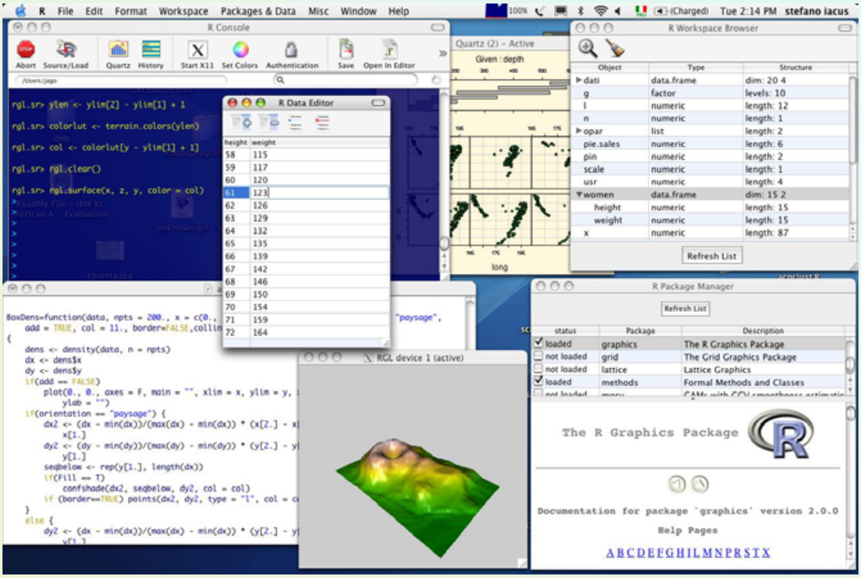
Qt Hierarchical Clustering

Linkage: Complete

Annotation: iris

Dendogram setings
- ☐ Print depth 100
- Horizontal size: 300
- ☐ Fit hor. size to window
- Text font size: 8
- Line spacing: 4

R project – statistical data exploration



```
rgl.sr> ylen <- ylim[2] - ylim[1] + 1

rgl.sr> colorlut <- terrain.colors(ylen)

rgl.sr> col <- colorlut[y - ylim[1] + 1]

rgl.sr> rgl.clear()

rgl.sr> rgl.surface(x, z, y, color = col)
>
>
>
>
>
>
>
>
```

```
BoxDens=function(data, npts = 200., x = c(0.,                    "paysage",
    add = TRUE, col = 11., border=FALSE,collin
{
    dens <- density(data, n = npts)
    dx <- dens$x
    dy <- dens$y
    if(add == FALSE)
        plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
            ylab = "")
    if(orientation == "paysage") {
        dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
            x[1.]
        dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
            y[1.]
        seqbelow <- rep(y[1.], length(dx))
        if(Fill == T)
            confshade(dx2, seqbelow, dy2, col = col)
        if (border==TRUE) points(dx2, dy2, type = "l", col = c
    }
    else {
        dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
            y[1.]
```

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

Object	Type	Structure
▶ dati	data.frame	dim: 20 4
g	factor	levels: 10
l	numeric	length: 12
n	numeric	length: 1
▶ opar	list	length: 2
pie.sales	numeric	length: 6
pin	numeric	length: 2
scale	numeric	length: 1
usr	numeric	length: 4
▼ women	data.frame	dim: 15 2
height	numeric	length: 15
weight	numeric	length: 15
x	numeric	length: 87

status	Package	Description
☑ loaded	graphics	The R Graphics Package
□ not loaded	grid	The Grid Graphics Package
□ not loaded	lattice	Lattice Graphics
☑ loaded	methods	Formal Methods and Classes

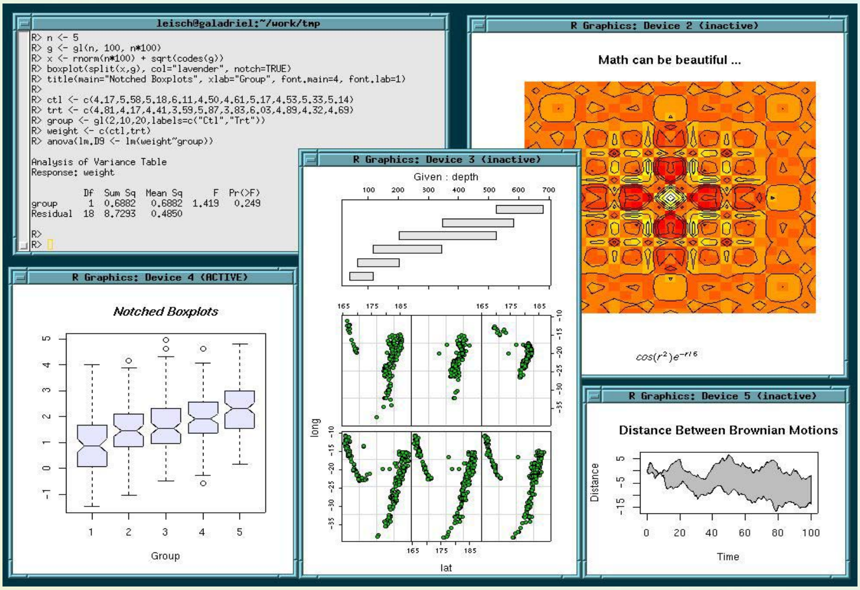
The R Graphics Package

Documentation for package `graphics' version 2.0.0

Help Pages

A B C D E F G H I L M N P R S T X

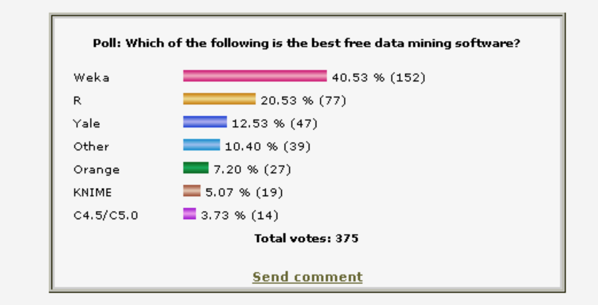
R project


leisch@galadriel:~/work/tmp
R Graphics: Device 2 (inactive)
Math can be beautiful ...
$cos(r^2)e^{-r/6}$
R Graphics: Device 3 (inactive)
Given : depth
100
200
300
400
500
600
700
165
175
185
-10
-15
-20
-25
-30
-35
long
lat
R Graphics: Device 4 (ACTIVE)
Notched Boxplots
5
4
3
2
1
0
-1
Group
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
5
-5
-15
0
20
40
60
80
100
Time

```
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))

Analysis of Variance Table
Response: weight

          Df  Sum Sq  Mean Sq      F  Pr(>F)
group      1  0.6882   0.6882  1.419   0.249
Residual  18  8.7293   0.4850

R>
R> 
```

2008 Pool on the popular free software

- Internet users - www.eruditionhome.com



Poll: Which of the following is the best free data mining software?

Weka 40.53 % (152)

R 20.53 % (77)

Yale 12.53 % (47)

Other 10.40 % (39)

Orange 7.20 % (27)

KNIME 5.07 % (19)

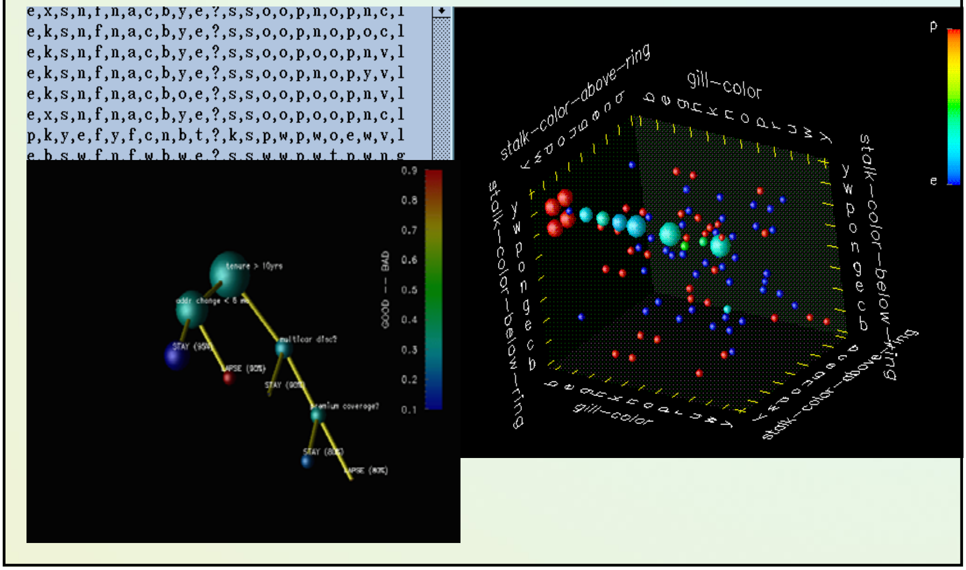
C4.5/C5.0 3.73 % (14)

Total votes: 375

IBM Intelligent Miner: Major Features

- Highly scalable, large database-oriented data mining algorithms
- Multiple data mining functions:
  - Association
  - Classification
  - Sequencing analysis
  - Clustering.
- Visual graphical display
- Influential in database and data mining research communities.

IBM Miner – example of visualisation



```
e,x,s,n,f,n,a,c,b,y,e,?,s,s,o,o,p,n,o,p,n,c,l
e,k,s,n,f,n,a,c,b,y,e,?,s,s,o,o,p,n,o,p,o,c,l
e,k,s,n,f,n,a,c,b,y,e,?,s,s,o,o,p,o,o,p,n,v,l
e,k,s,n,f,n,a,c,b,y,e,?,s,s,o,o,p,n,o,p,y,v,l
e,k,s,n,f,n,a,c,b,o,e,?,s,s,o,o,p,o,o,p,n,v,l
e,x,s,n,f,n,a,c,b,y,e,?,s,s,o,o,p,o,o,p,n,c,l
p,k,y,e,f,y,f,c,n,b,t,?,k,s,p,w,p,w,o,e,w,v,l
e,b,s,w,f,n,f,w,b,w,e,?,s,s,w,w,p,w,t,p,w,n,g
```

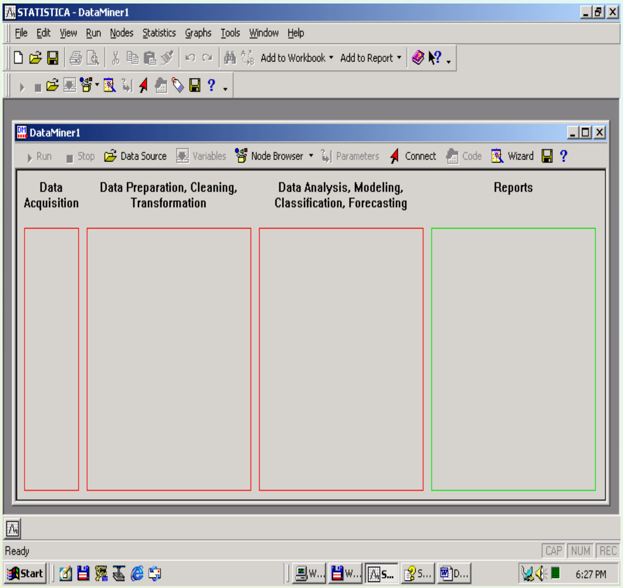
Statistica – Statsoft (www.statsoft.pl / *.com)

- User friendly for MS Windows; mainly based on statistical approaches.
- It contains numerous data analysis methods.
- Efficient calculations, good managing results and reports.
- Excellent graphical visualisation.
- Comprehensive help, documentations, supporting books and teaching materials.
- Drivers to data bases and other data sources

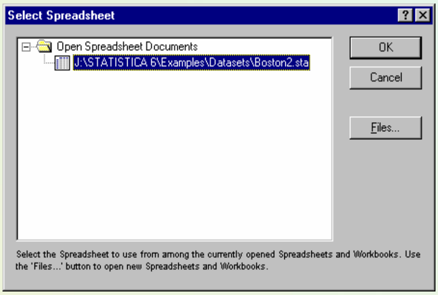
Main systems:

- Statistica 6.0 – mainly statistical software
- Statistica Data Miner – specific for DM / user friendly
- Specialized systems – Statistica Neural Networks.
- Quality and Control Cards
- Corporation Tools
- …

DataMiner – main panel

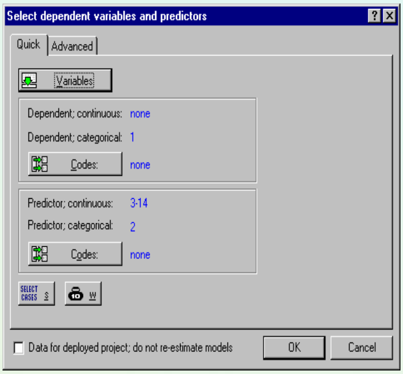


Data Miner – loading data and selecting attributes



J:\STATISTICA 6\Examples\Datasets\Boston2.sta

Select the Spreadsheet to use from among the currently opened Spreadsheets and Workbooks. Use the 'Files...' button to open new Spreadsheets and Workbooks.



Dependent; continuous: none

Dependent; categorical: 1

Codes: none

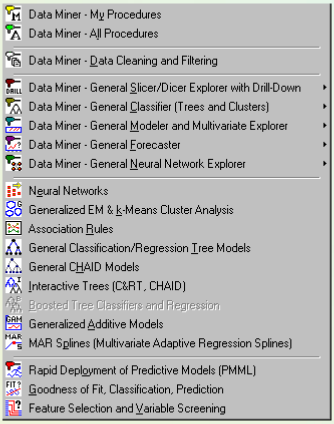
Predictor; continuous: 3-14

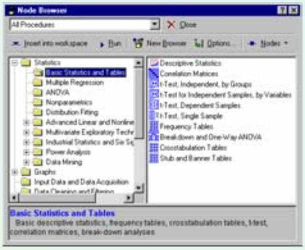
Predictor; categorical: 2

Codes: none

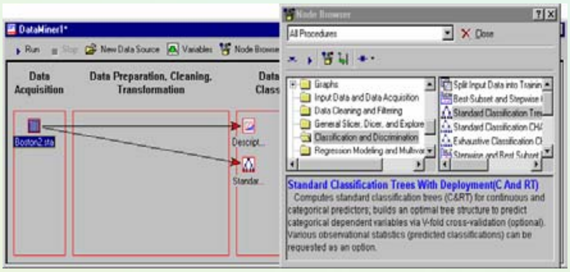
□ Data for deployed project; do not re-estimate models

Data Miner – choosing methods


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

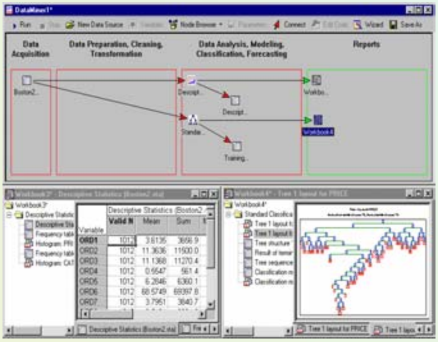


Basic Statistics and Tables
Basic descriptive statistics, frequency tables, crosstabulation tables, t-test, correlation matrices, break-down analyses

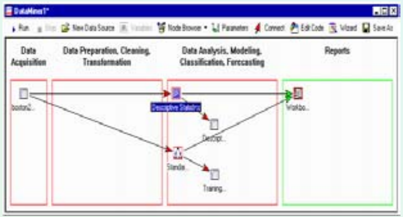

DataMiner1*
Run
Stop
New Data Source
Variables
Node Browser
Data Acquisition
Data Preparation, Cleaning, Transformation
Data Class
Boston2.sta
Descript...
Standar...
Node Browser
All Procedures
Close
Graphs
Input Data and Data Acquisition
Data Cleaning and Filtering
General Slicer, Dicer, and Explore
Classification and Discrimination
Regression Modeling and Multivar
Split Input Data into Trainin
Best Subset and Stepwise
Standard Classification Tre
Standard Classification CH
Exhaustive Classification C
Stepwise and Best Subset

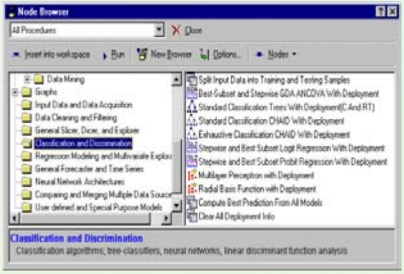
Standard Classification Trees With Deployment(C And RT)
Computes standard classification trees (C&RT) for continuous and categorical predictors; builds an optimal tree structure to predict categorical dependent variables via V-fold cross-validation (optional). Various observational statistics (predicted classifications) can be requested as an option.

Extra tools for defining projects


DataMiner1*
Run
Stop
New Data Source
Variables
Node Browser
Parameters
Connect
Edit Code
Wizard
Save As
Data Acquisition
Data Preparation, Cleaning, Transformation
Data Analysis, Modeling, Classification, Forecasting
Reports
Boston2...
Descript...
Standa...
Descript...
Training...
Workbo...
Workbook4
Workbook3* - Descriptive Statistics (Boston2.sta)
Descriptive Statistics (Boston2
Variable
Valid N
Mean
Sum
Workbook4* - Tree 1 layout for PRICE

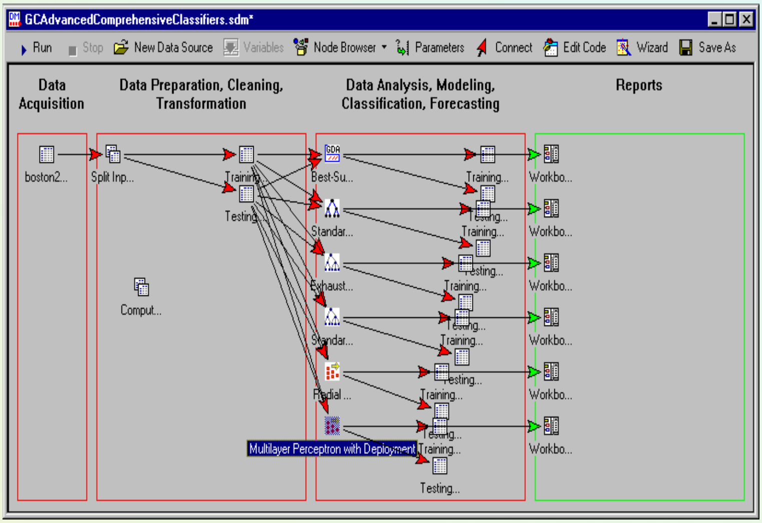
Variable	Valid N	Mean	Sum
ORD1	1012	3.6135	3656.9
ORD2	1012	11.3636	11500.0
ORD3	1012	11.1368	11270.4
ORD4	1012	0.5547	561.4
ORD5	1012	6.2846	6360.1
ORD6	1012	68.5749	69397.8
ORD7	1012	3.7951	3840.7



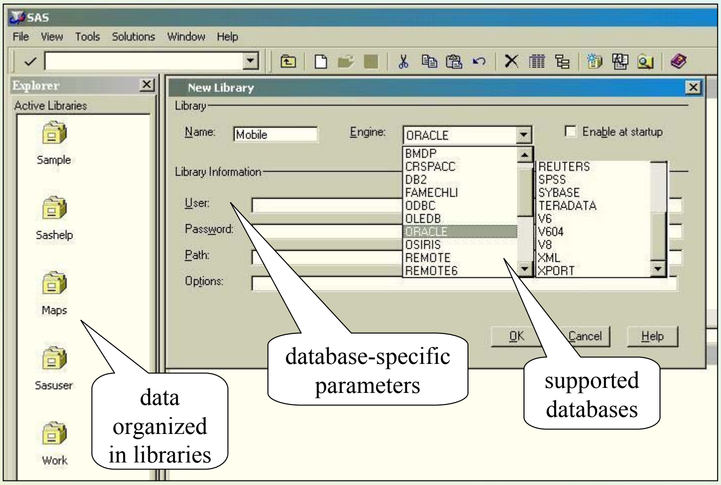

Node Browser
All Procedures
Close
Insert into workspace
Run
New Browser
Options...
Nodes
Data Mining
Graphs
Input Data and Data Acquisition
Data Cleaning and Filtering
General Slicer, Dicer, and Explorer
Classification and Discrimination
Regression Modeling and Multivariate Explor
General Forecaster and Time Series
Neural Network Architectures
Comparing and Merging Multiple Data Source
User defined and Special Purpose Models
Split Input Data into Training and Testing Samples
Best-Subset and Stepwise GDA ANCOVA With Deployment
Standard Classification Trees With Deployment(C And RT)
Standard Classification CHAID With Deployment
Exhaustive Classification CHAID With Deployment
Stepwise and Best Subset Logit Regression With Deployment
Stepwise and Best Subset Probit Regression With Deployment
Multilayer Perceptron with Deployment
Radial Basis Function with Deployment
Compute Best Prediction From All Models
Clear All Deployment Info

Classification and Discrimination
Classification algorithms, tree-classifiers, neural networks, linear discriminant function analysis

Using several methods on the same data



SAS Enterprise Miner



Name: Mobile

Engine: ORACLE

□ Enable at startup

User:

Password:

Path:

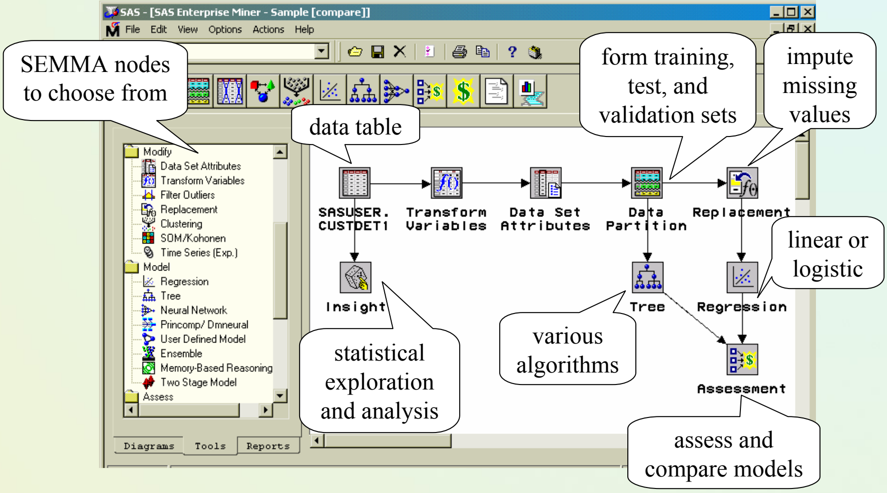
Options:

data organized in libraries

database-specific parameters

supported databases

Enterprise miner project



SEMMA nodes to choose from

data table

form training, test, and validation sets

impute missing values

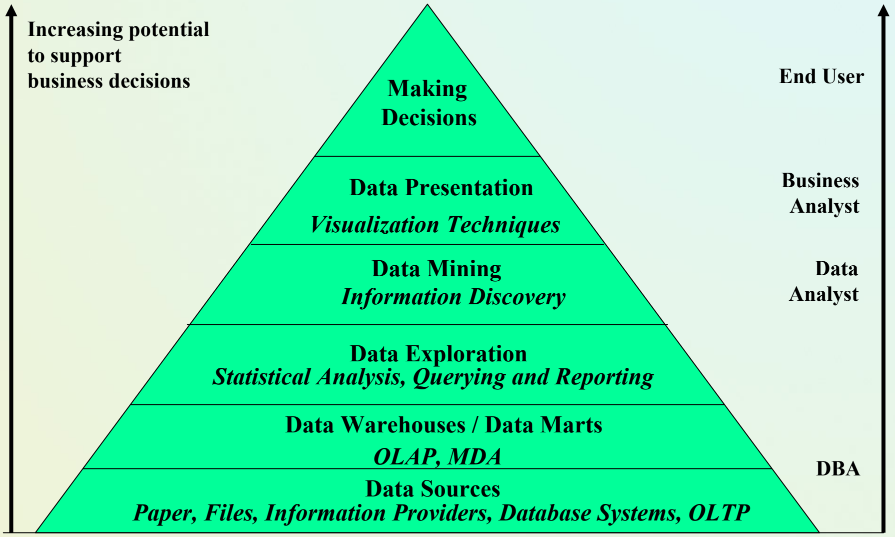
linear or logistic

statistical exploration and analysis

various algorithms

assess and compare models

Data Mining and Business Intelligence



Increasing potential to support business decisions

Making Decisions

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts
OLAP, MDA

Data Sources
Paper, Files, Information Providers, Database Systems, OLTP

End User

Business Analyst

Data Analyst

DBA

Industries/fields where you currently apply data mining [KDD Pool - 216 votes total]

Banking (29) 13%

Bioinformatics/Biotech (18) 8%

Direct Marketing/Fundraising (19) 9%

eCommerce/Web (12) 6%

Entertainment/News (1) 0%

Fraud Detection (19) 9%

Insurance (15) 7%

Investment/Stocks (9) 4%

Manufacturing (9) 4%

Medical/Pharma (15) 7%

Retail (9) 4%

Scientific data (20) 9%

Security (8) 4%

Telecommunications (12) 6%

Travel (2) 1%

Other (19) 9%

Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers
- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)

Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

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.

Controversial Issues: Society and Privacy

- Data mining (or simple analysis) on people may come with a profile that would raise controversial issues of
  - Discrimination
  - Privacy
  - Security
- Examples:
  - Should males between 18 and 35 from countries that produced terrorists be singled out for search before flight?
  - Can people be denied mortgage based on age, sex, race?
  - Women live longer. Should they pay less for life insurance?
- Can discrimination be based on features like sex, age, national origin?
- In some areas (e.g. mortgages, employment), some features cannot be used for decision making

Data Mining and Privacy

- Can information collected for one purpose be used for mining data for another purpose
  - In Europe, generally no, without explicit consent!
  - In US, generally yes,…
- Companies routinely collect information about customers and use it for marketing, etc.
- People may be willing to give up some of their privacy in exchange for some benefits

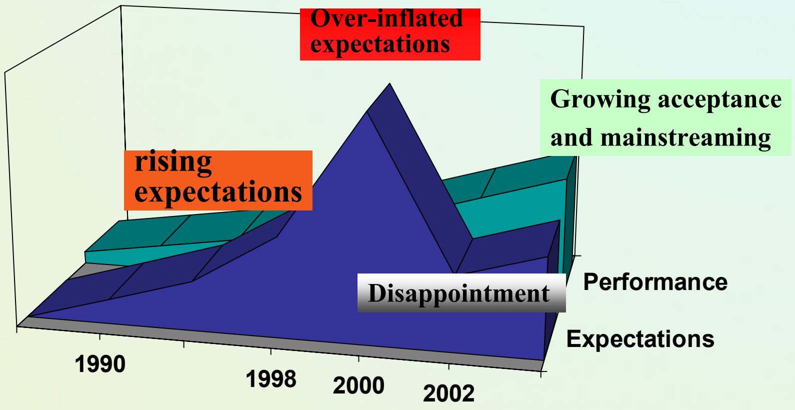
Data Mining Future Directions

- Currently, most data mining is on flat tables
- Richer data sources
  - text, links, web, images, multimedia, knowledge bases
- Advanced methods
  - Link mining, Stream mining, …
- Applications
  - Web, Bioinformatics, Customer modeling, …

Challenges for Data Mining

- Technical
  - tera-bytes and peta-bytes
  - complex, multi-media, structured data
  - integration with domain knowledge
- Business
  - finding good application areas
- Societal
  - Privacy issues

The Hype Curve for Data Mining and Knowledge Discovery



Data Mining Central Quest

Find true patterns
and avoid *overfitting*
(false patterns due
to randomness).

So, be lucky in using this course!

Background literature

- Witten Ian and Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.
- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001.
- Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001.
- Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press 1996.
- Mitchell T.M., Machine Learning, McGrawHill, 1997.
- Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, PP Press, 2003.

Any questions and remarks

- I prefer other questions than
- What about the final exam?

Thank you !