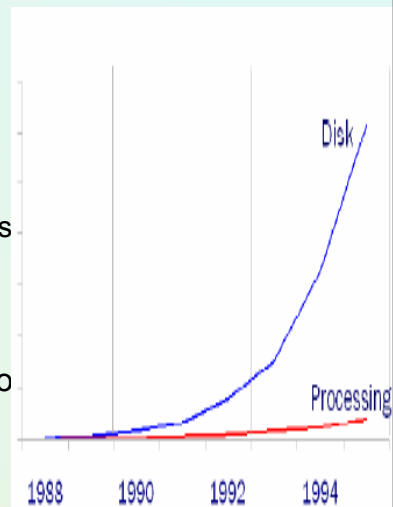

Knowledge Discovery Process and Data Mining - Final remarks



Lecturer: JERZY STEFANOWSKI
Institute of Computing Sciences
Poznan University of Technology
Poznan, Poland
Lecture 14
SE Master Course
2008/2009

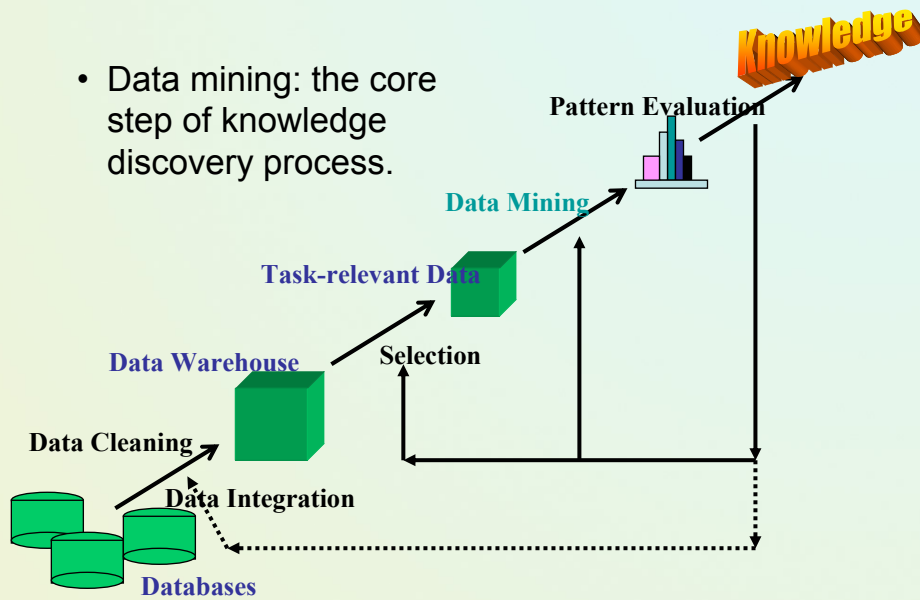
Growth Trends

- 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.



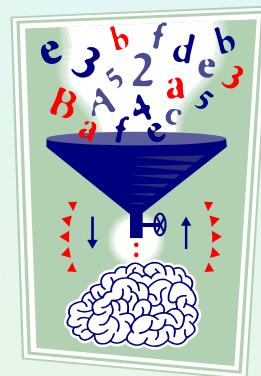
Steps of a KDD Process

- 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.

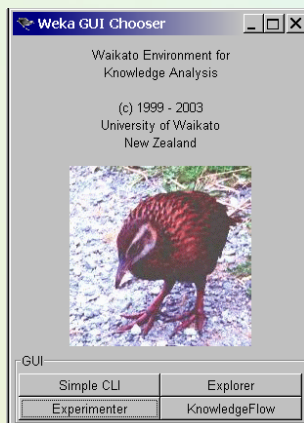
Can We Find All and Only Interesting Patterns?

- 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 generate 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

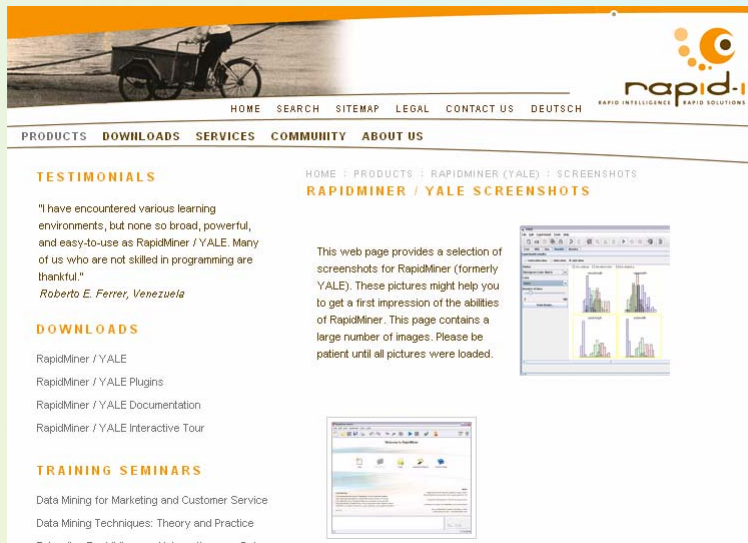
WEKA – Machine Learning and Data Mining



Java implementation
of many algorithms

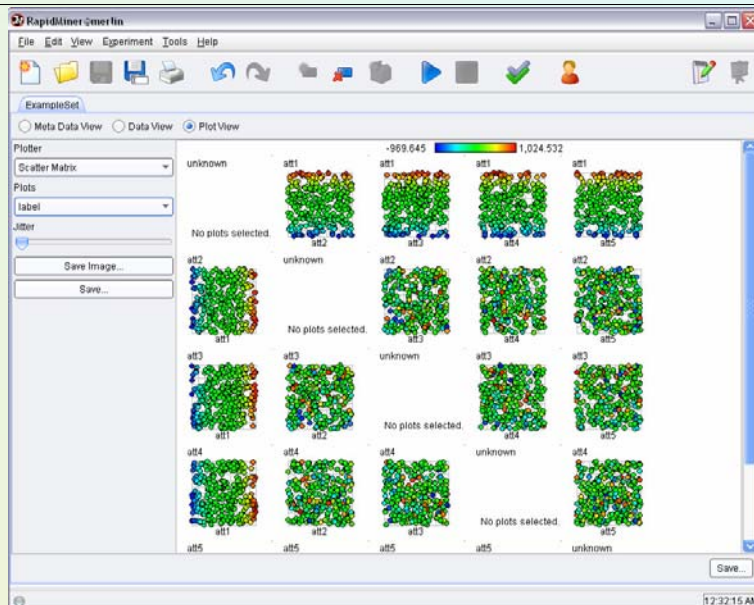
No ideal solutions → but ...

RapidMiner (YALE)

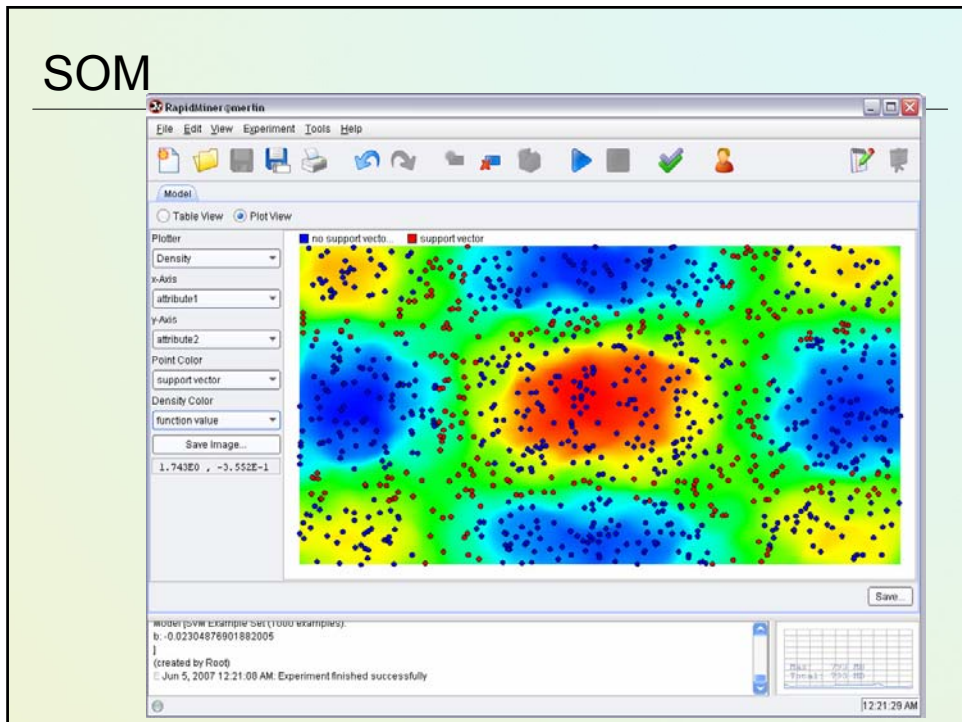


The screenshot shows the RapidMiner (YALE) website homepage. At the top, there is a navigation menu with links for HOME, SEARCH, SITEMAP, LEGAL, CONTACT US, and DEUTSCH. Below the menu, there are sections for TESTIMONIALS, DOWNLOADS, and TRAINING SEMINARS. The TESTIMONIALS section features a quote from Roberto E. Ferrer, Venezuela. The DOWNLOADS section lists various resources like RapidMiner / YALE, plugins, documentation, and an interactive tour. The TRAINING SEMINARS section lists courses such as "Data Mining for Marketing and Customer Service" and "Data Mining Techniques: Theory and Practice". A central section titled "RAPIDMINER / YALE SCREENSHOTS" provides a selection of screenshots from the software, with a note that the page contains a large number of images and users should be patient until they are fully loaded. The RapidMiner logo is visible in the top right corner.

Some Rapidminer screenshots

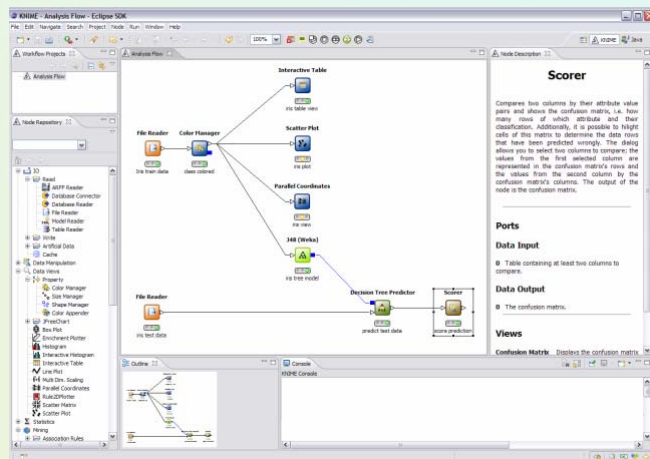


SOM



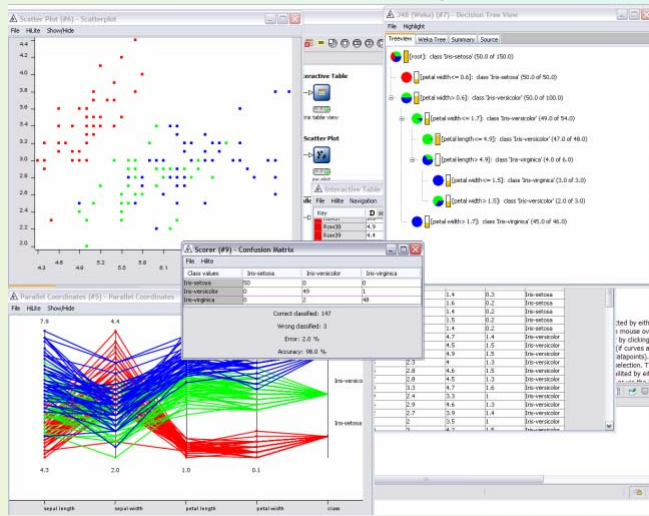
KNIME

- 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.



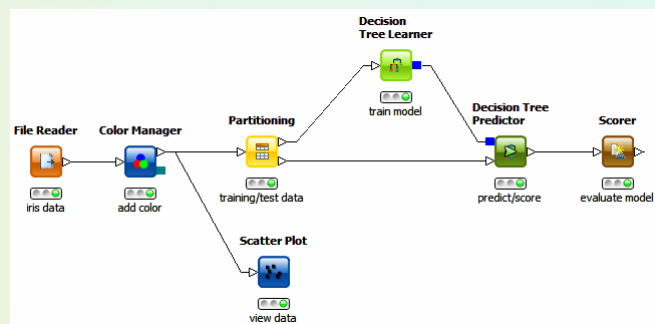
KNIME - An Example of Data Analysis Workflow

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

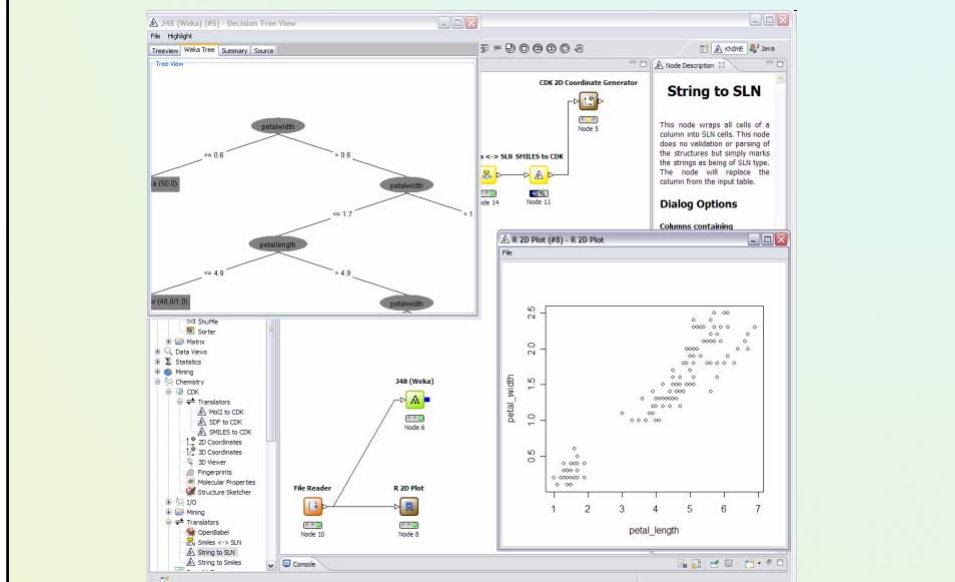


Developing Trees

- Node flows

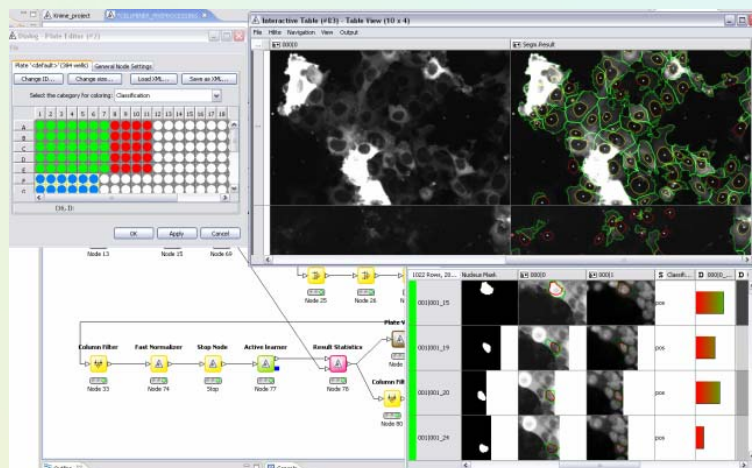


KNIME working with decision trees



Cell Miner

- KNIME has been used to analyze cell images.



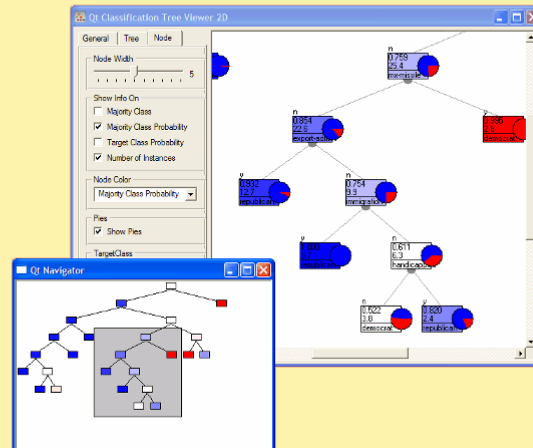
Orange (Slovenia)



Orange Screenshots

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

Classification tree viewer with a navigator.



Home
Screenshots
Contact & Support
Acknowledgements

Download

Forum (RSS)

Documentation

Search

Visual Programming

Catalog of Widgets

Scripting for Beginners

Class Reference

Modules

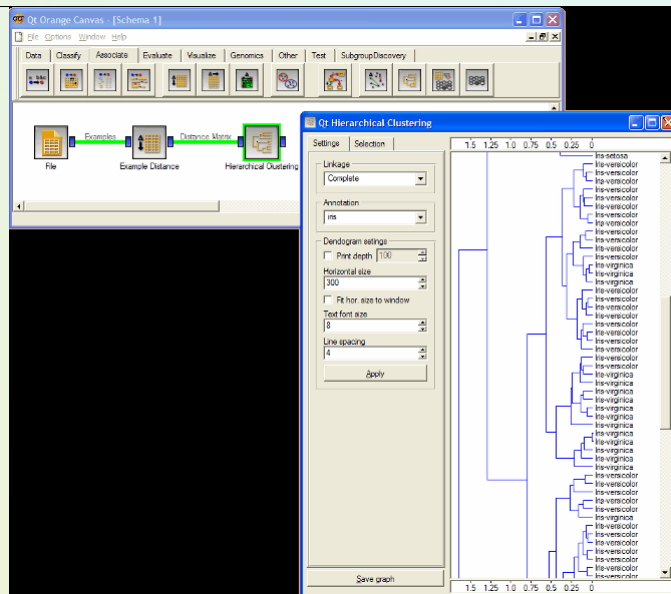
Example Scripts

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 - clustering



R project – statistical data exploration

The screenshot displays the RStudio interface with several windows open:

- R Console:** Contains R code for data manipulation and plotting.


```

rgl.srs.ylen <- ylen[2] - ylen[1] + 1
rgl.srs.colorlut <- terrain.colors(ylen)
rgl.srs.col <- colorlut[y - ylen[1] + 1]
rgl.srs rgl.clear()
rgl.srs rgl.surface(x, z, y, color = col)

dens <- density(data, n = npts)
dx <- densx
dy <- densy
!!(add == TRUE)
plot(0, 0, axes = F, main = "", xlim = x, ylim = y,
      ylab = "")
!!(orientation == "paysage") {
  dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * Cx[Z] - x
  dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * Cy[Z] - y
  seqlow <- rep(1, length(dx))
  !!(fill == T)
  contour(dx2, seqlow, dy2, col = col)
  !!(border == TRUE) points(dx2, dy2, type = "l", col = c
    } else {
  dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * Cy[Z] - y
  }

```
- R Data Editor:** Shows a table with columns 'height' and 'weight'.

| height | weight |
|--------|--------|
| 58 | 115 |
| 59 | 117 |
| 60 | 120 |
| 61 | 125 |
| 62 | 126 |
| 63 | 129 |
| 64 | 132 |
| 65 | 135 |
| 66 | 139 |
| 67 | 142 |
| 68 | 146 |
| 69 | 150 |
| 70 | 154 |
| 71 | 159 |
| 72 | 164 |
- Quartz (2) - Active:** Displays a 3D surface plot of the terrain data.
- R Workspace Browser:** Lists objects in the workspace:

| Object | Type | Structure |
|-----------|------------|------------|
| data | 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 |
- R Package Manager:** Shows installed packages:

| status | Package | Description |
|-------------------------------------|----------|----------------------------|
| <input checked="" type="checkbox"/> | graphics | The R Graphics Package |
| <input type="checkbox"/> | grid | The Grid Graphics Package |
| <input type="checkbox"/> | lattice | Lattice Graphics |
| <input checked="" type="checkbox"/> | methods | Formal Methods and Classes |

R project



The screenshot shows an RStudio workspace with several graphics devices:

- R Console:** Contains R code for generating data and performing an ANOVA.


```

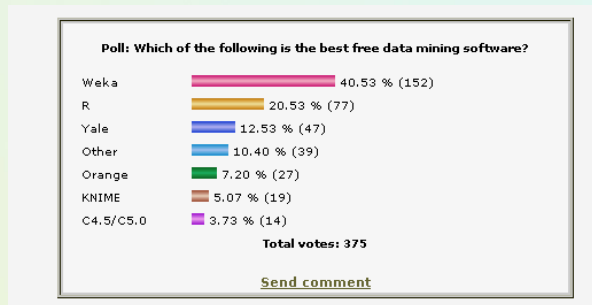
R> n <- 5
R> g <- gl(n, 100, n=100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(splitted(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", Font.main=4, Font.lab=1)
R>
R> c1 <- c(4,17,5,28,5,18,5,11,4,50,4,81,5,17,4,53,5,33,5,14)
R> trt <- c(4,31,4,17,4,41,2,28,5,87,2,83,6,105,4,88,4,32,4,68)
R> group <- gl(2,10,20,labels=c("C1","Trt"))
R> weight <- c(c1, trt)
R> aovwt(aov ~ weight ~ group)

```
- Analysis of Variance Table:**

| | 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 Graphics: Device 4 (ACTIVE):** Displays a Notched Boxplots plot with five groups on the x-axis and response values on the y-axis.
- R Graphics: Device 3 (inactive):** Displays a horizontal boxplot titled "Given: depth" with values on the x-axis ranging from 100 to 700.
- R Graphics: Device 2 (inactive):** Displays a colorful fractal-like pattern titled "Math can be beautiful ...". Below it is the equation $\cos(r^2) \cdot e^{-r^2}$.
- R Graphics: Device 5 (inactive):** Displays a line plot titled "Distance Between Brownian Motions" showing distance over time.

2008 Pool on the popular free software

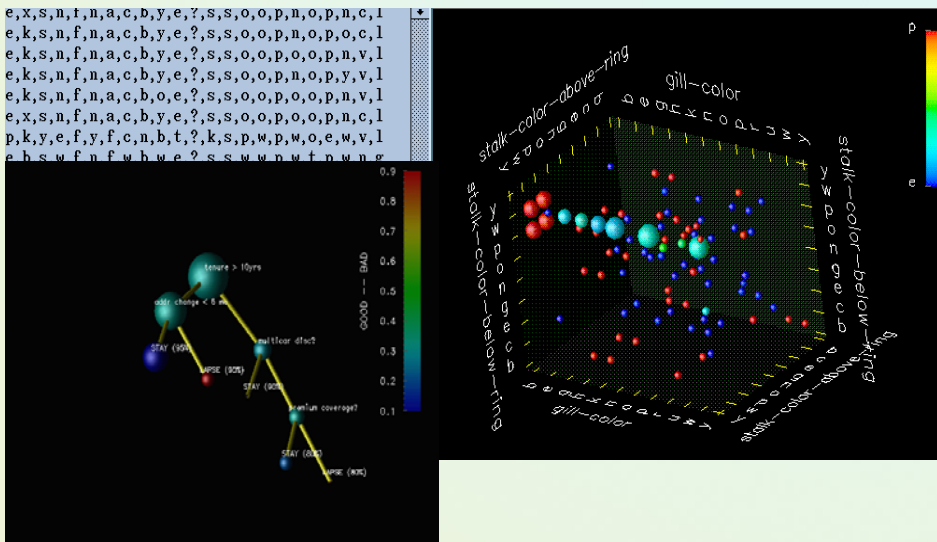
- Internet users - www.eruditionhome.com



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



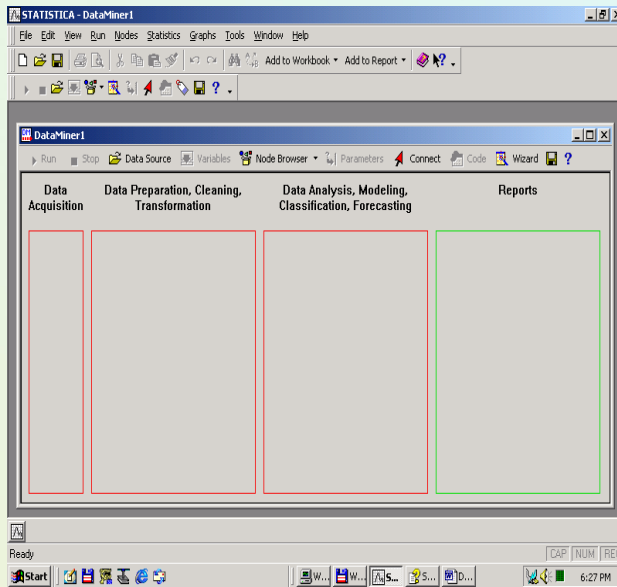
Statistica – Statsoft (www.statsoft.pl / *.com)

- 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

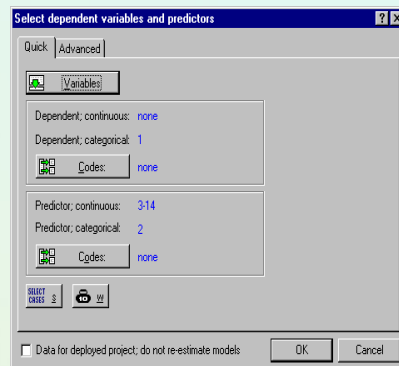
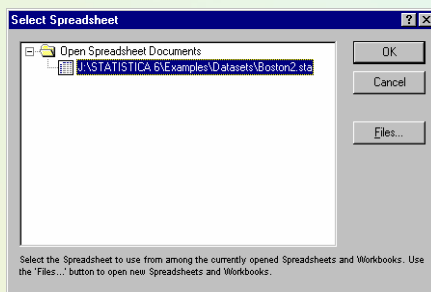
Main systems:

- 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

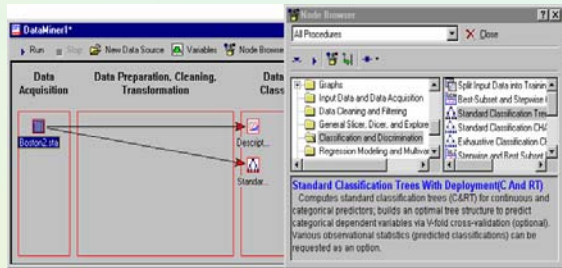
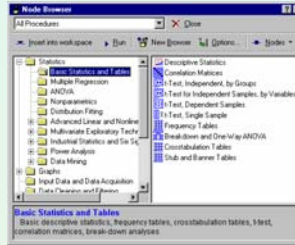


Data Miner – loading data and selecting attributes

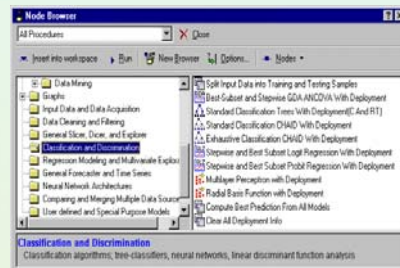
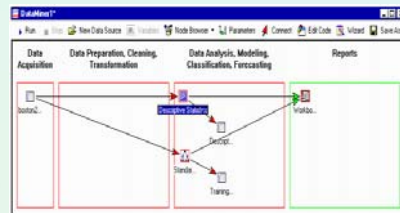


Data Miner – choosing methods

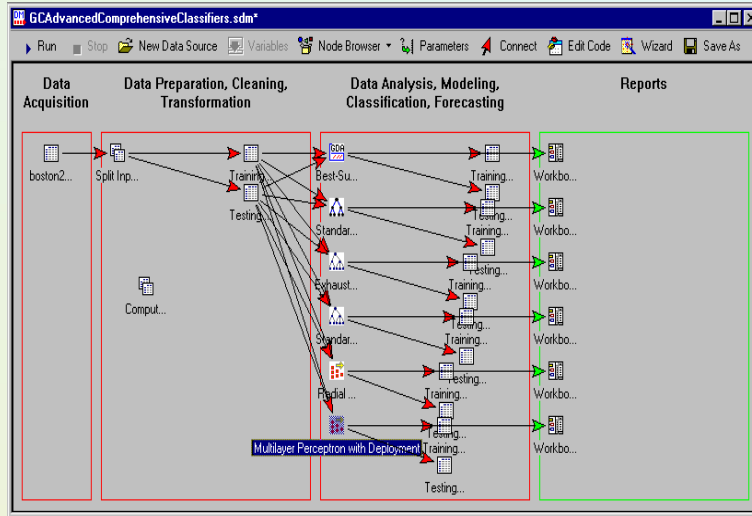
- Data Miner - My Procedures
- Data Miner - All Procedures
- Data Miner - Data Cleaning and Filtering
- Data Miner - General Slicer/Dicer Explorer with Drill-Down
- Data Miner - General Classifier (Trees and Clusters)
- Data Miner - General Modeler and Multivariate Explorer
- Data Miner - General Forecaster
- 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



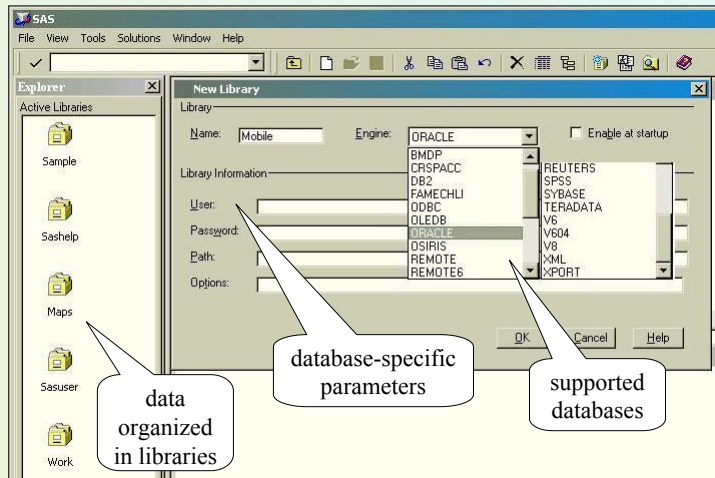
Extra tools for defining projects



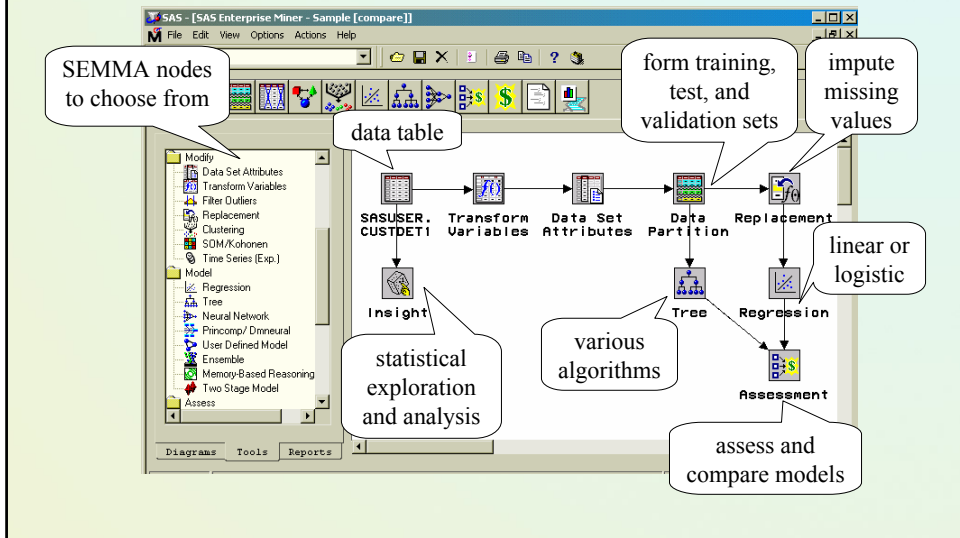
Using several methods on the same data



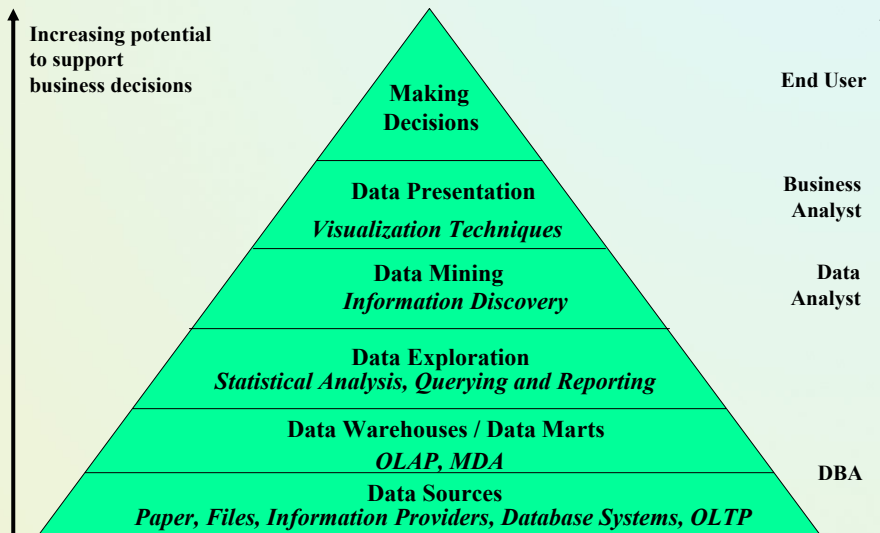
SAS Enterprise Miner



Enterprise miner project



Data Mining and Business Intelligence



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

| | |
|--------------------------------------|----------------------------|
| Banking (29) 13% | Manufacturing (9) 4% |
| Bioinformatics/Biotech (18) 8% | Medical/Pharma (15) 7% |
| Direct Marketing/Fundraising (19) 9% | Retail (9) 4% |
| eCommerce/Web (12) 6% | Scientific data (20) 9% |
| Entertainment/News (1) 0% | Security (8) 4% |
| Fraud Detection (19) 9% | Telecommunications (12) 6% |
| Insurance (15) 7% | Travel (2) 1% |
| Investment/Stocks (9) 4% | 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

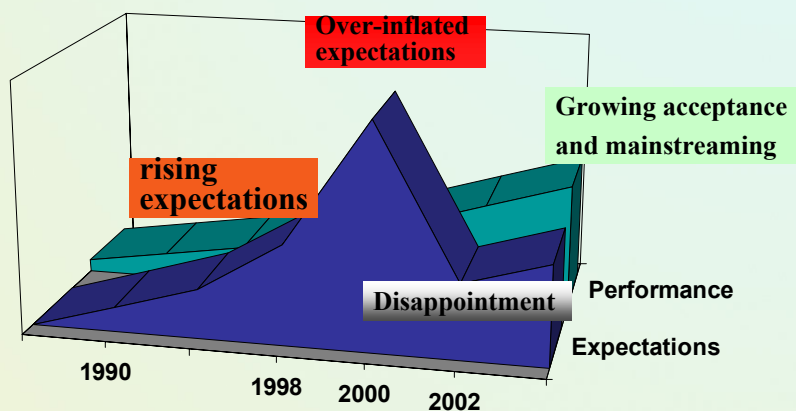
Data Mining Future Directions

- 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 !