

---

# Data Mining

## Knowledge Discovery, Data Warehousing and Machine Learning

### Final remarks

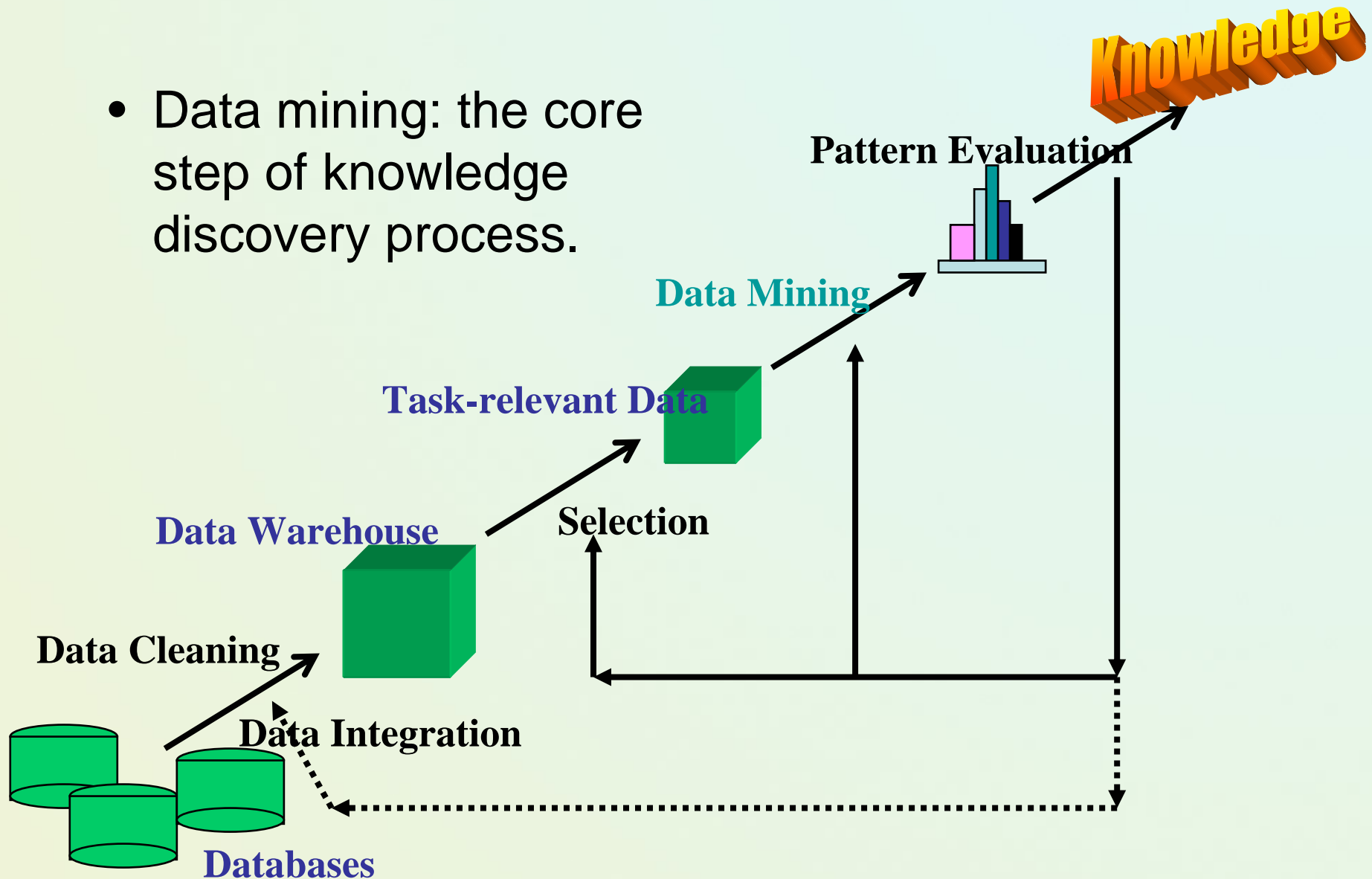


Lecturer: JERZY STEFANOWSKI

Email: [Jerzy.Stefanowski@cs.put.poznan.pl](mailto:Jerzy.Stefanowski@cs.put.poznan.pl)

# Data Mining a step in A KDD Process

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



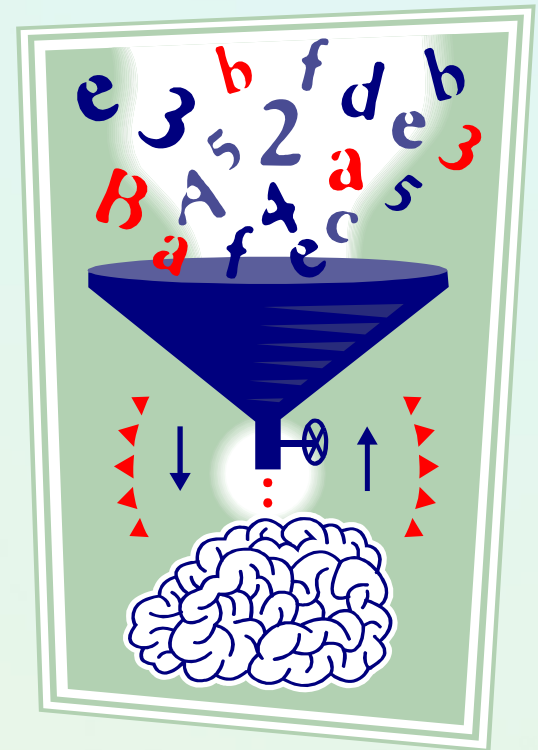
# Steps of a KDD Process

- 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 Understanding: Quantity

---

- Number of instances (records)
  - *Rule of thumb: 5,000 or more desired*
  - if less, results are less reliable; use special methods (boosting, ...)
- Number of attributes (fields)
  - *Rule of thumb: for each field, 10 or more instances*
  - If more fields, use feature reduction and selection
- Number of targets
  - *Rule of thumb: >100 for each class*
  - if very unbalanced, use stratified sampling

# Data Cleaning Steps

---

- Data acquisition and metadata
- Missing values
- Unified date format
- Converting nominal to numeric
- Discretization of numeric data
- Data validation and statistics



# Data Cleaning: Metadata

---

- **Field types:**
  - **binary, nominal (categorical), ordinal, numeric, ...**
  - **For nominal fields: tables translating codes to full descriptions**
- **Field role:**
  - input : inputs for modeling
  - target : output
  - id/auxiliary : keep, but not use for modeling
  - ignore : don't use for modeling
  - weight : instance weight
  - ...
- **Field descriptions**

# Data Cleaning: Unified Date Format

---

- We want to transform all dates to the same format internally
- Some systems accept dates in many formats
  - e.g. “Sep 24, 2003” , 9/24/03, 24.09.03, etc
  - dates are transformed internally to a standard value
- Frequently, just the year (YYYY) is sufficient
- For more details, we may need the month, the day, the hour, etc
- Representing date as YYYYMM or YYYYMMDD can be OK, but has problems
- **Q: *What are the problems with YYYYMMDD dates?***
  - A: Ignoring for now the Looming Y10K (year 10,000 crisis ...)
  - YYYYMMDD does not preserve intervals:
    - 20040201 - 20040131  $\neq$  20040131 – 20040130
  - This can introduce bias into models

# Data Cleaning: Missing Values

---

- Missing data can appear in several forms:
  - <empty field> “0” “.” “999” “NA” ...
- Standardize missing value code(s)
- Dealing with missing values:
  - ignore records with missing values  
(only if you have enough data)
  - treat missing value as a separate value
    - Not-recommended approach
  - Imputation / Substitution:
    - Fill in with mean or mode values
      - Several options (all examples vs. class)
    - Regression or dependency from other fields

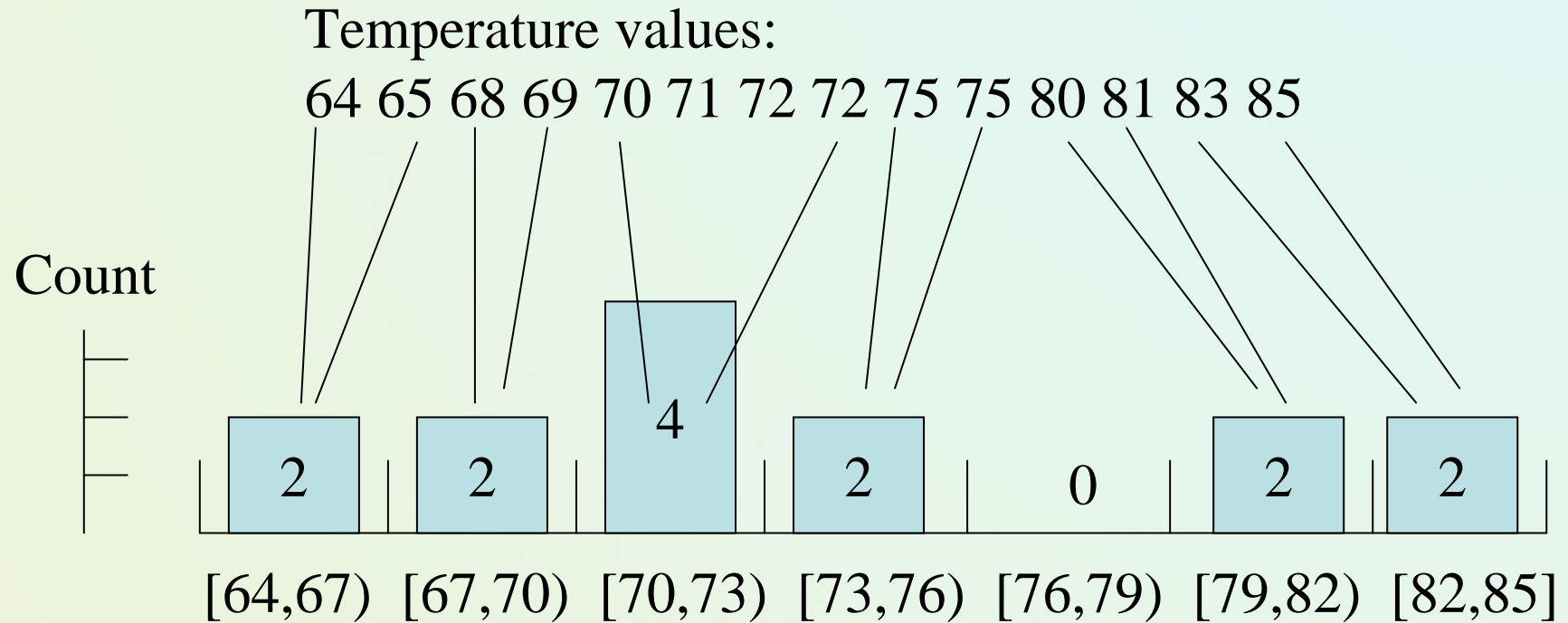
# Data Cleaning: Discretization

---

- Some methods require discrete values, e.g. most versions of Naïve Bayes, CHAID
- Discretization is very useful for generating a summary of data
- Also called “binning”
- Many approaches have been proposed:
  - Supervised vs. unsupervised,
  - Global vs. local (attribute point of view),
  - Dynamic vs. Statitic choice of paramteres

# Discretization: Equal-Width

---



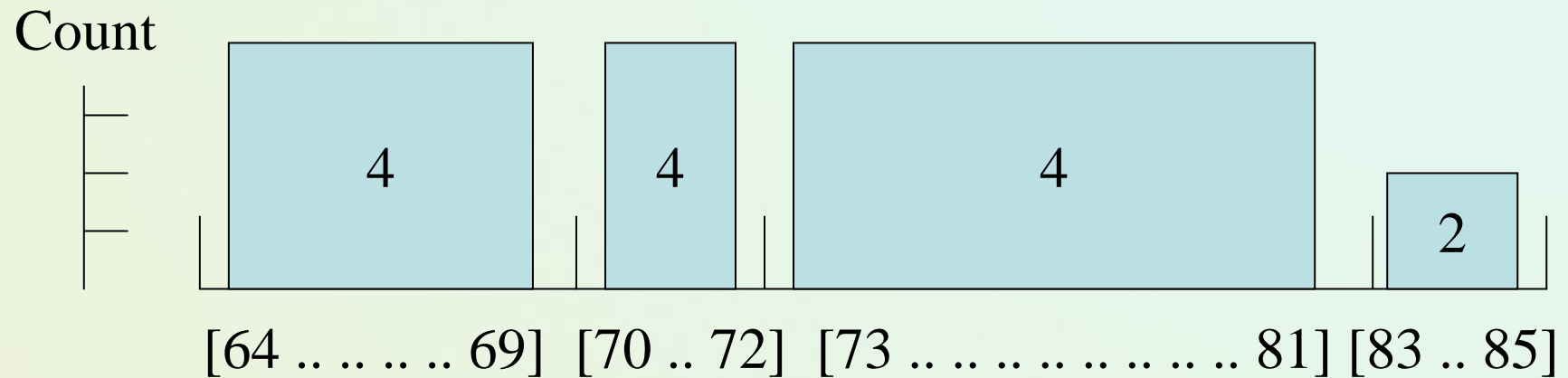
Equal Width, bins  $\text{Low} \leq \text{value} < \text{High}$

# Discretization: Equal-Height

---

Temperature values:

64 65 68 69 70 71 72 72 75 75 80 81 83 85

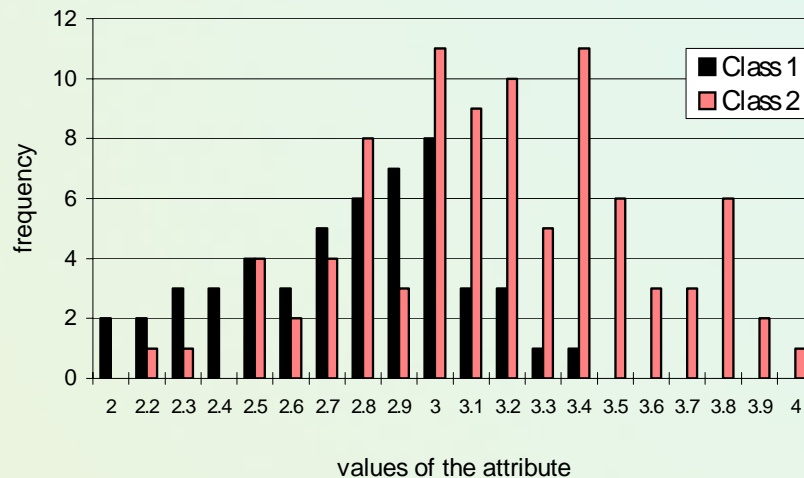


Equal Height = 4, except for the last bin

# Supervised discretization

---

- Use information about attribute value distribution + class assignment.



- Minimal entropy based approaches; Chi-Merge, others

# Data Cleaning: Attribute Selection

---

First: Remove fields with no or little variability

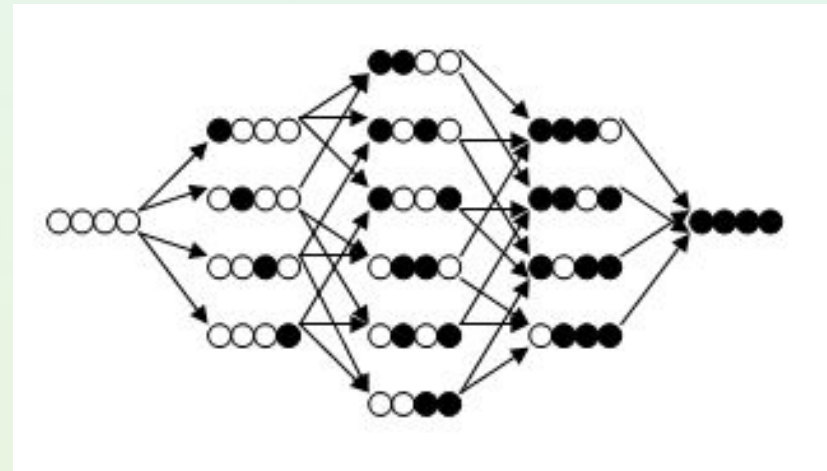
- Examine the number of distinct field values
  - *Rule of thumb: remove a field where almost all values are the same (e.g. null), except possibly in  $minp$  % or less of all records.*
  - $minp$  could be 0.5% or more generally less than 5% of the number of targets of the smallest class
- More sophisticated (statistical or ML) techniques specific for data mining tasks
  - In WEKA see attribute selection



## A few remarks on selecting attributes

---

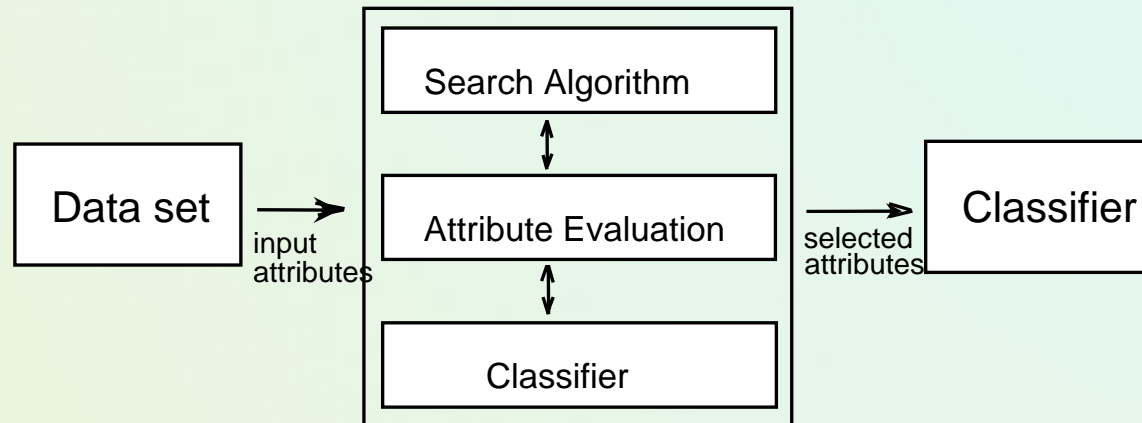
- Irrelevant attributes (features) in the input data may decrease the classification performance
- Attribute (feature) selection:
  - Find the smallest subset of attributes leading to a higher classification accuracy than all attributes
- Search problem in the space of attribute subsets
- Three components:
  - Search algorithm
  - Evaluation function
  - Classifier



# Wrapper approach

---

- Filter vs. Wrapper approach (Kohavi, and ...)



- The classifier is used by the evaluation function
- Search algorithms:
  - Forward selection
  - Backward elimination
  - Random search

# Different attribute selection methods

---

- Random selection.
- Correlation-based measure.
- Contextual-merit.
- Info-Gain.
  - Gain ratio
  - Chi-squared statistic
  - Liu Consistency measure
- and
  - Relief method
  - Wrapper model

# Conclusion

---

Good data preparation is  
key to producing valid and  
reliable models!

# 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


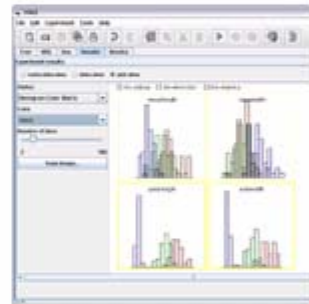
## TRAINING SEMINARS

- 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)



[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](#)

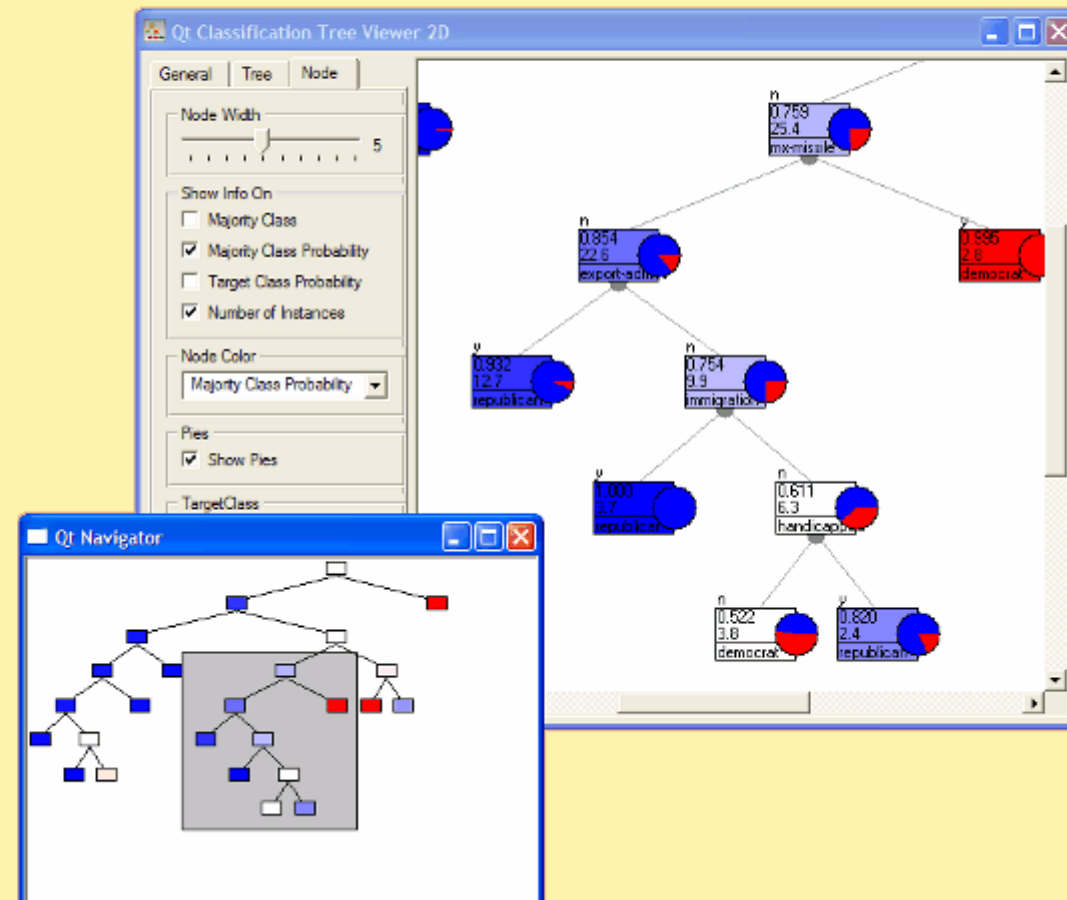
## Latest News

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 mining.

Classification tree viewer with a navigator.



# IBM Intelligent Miner: Major Features

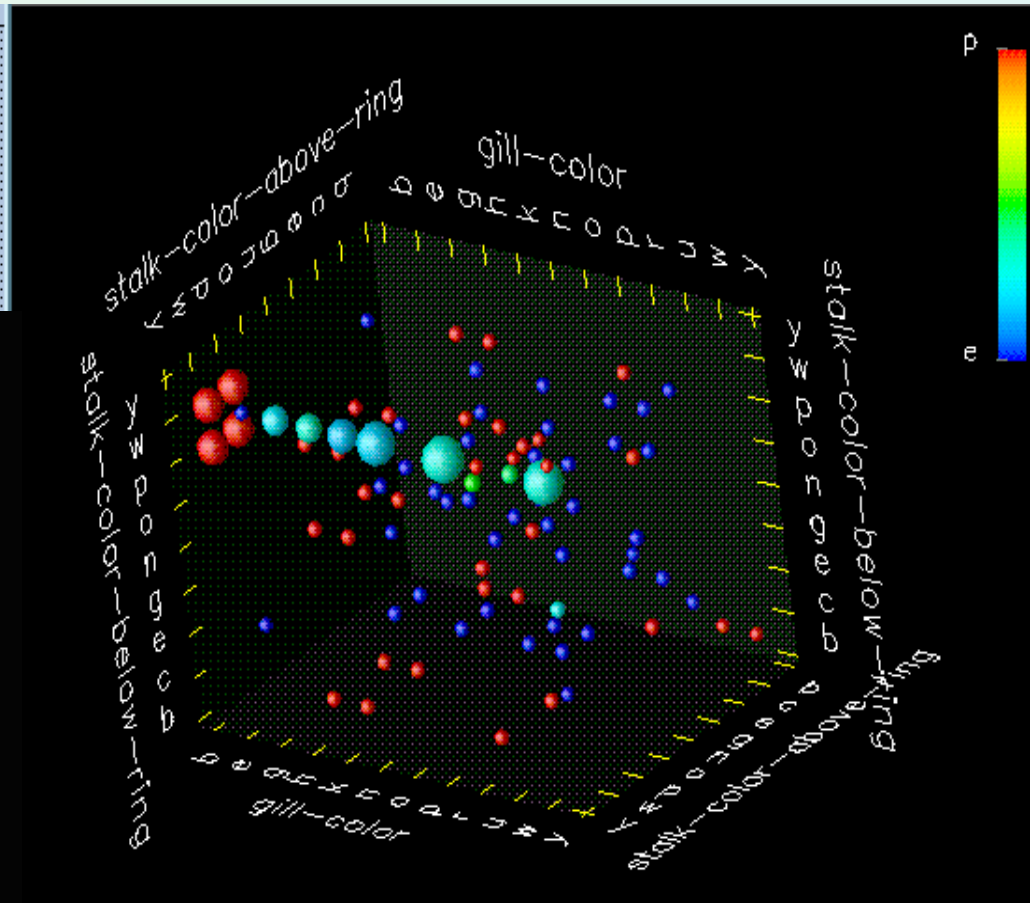
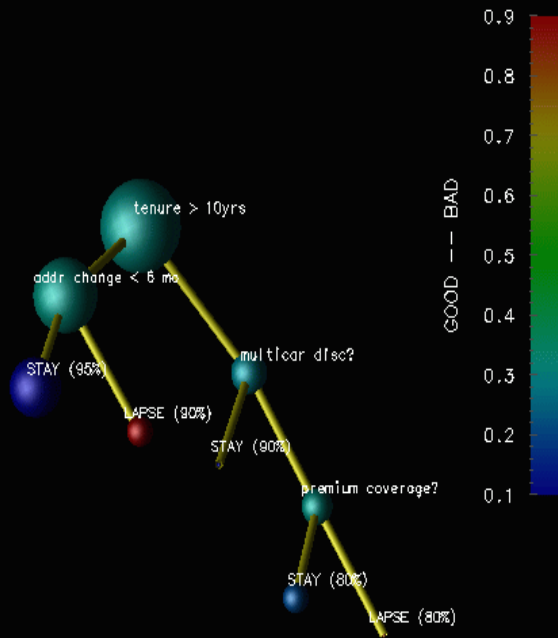
---

- 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](http://www.statsoft.pl) / \*.com)

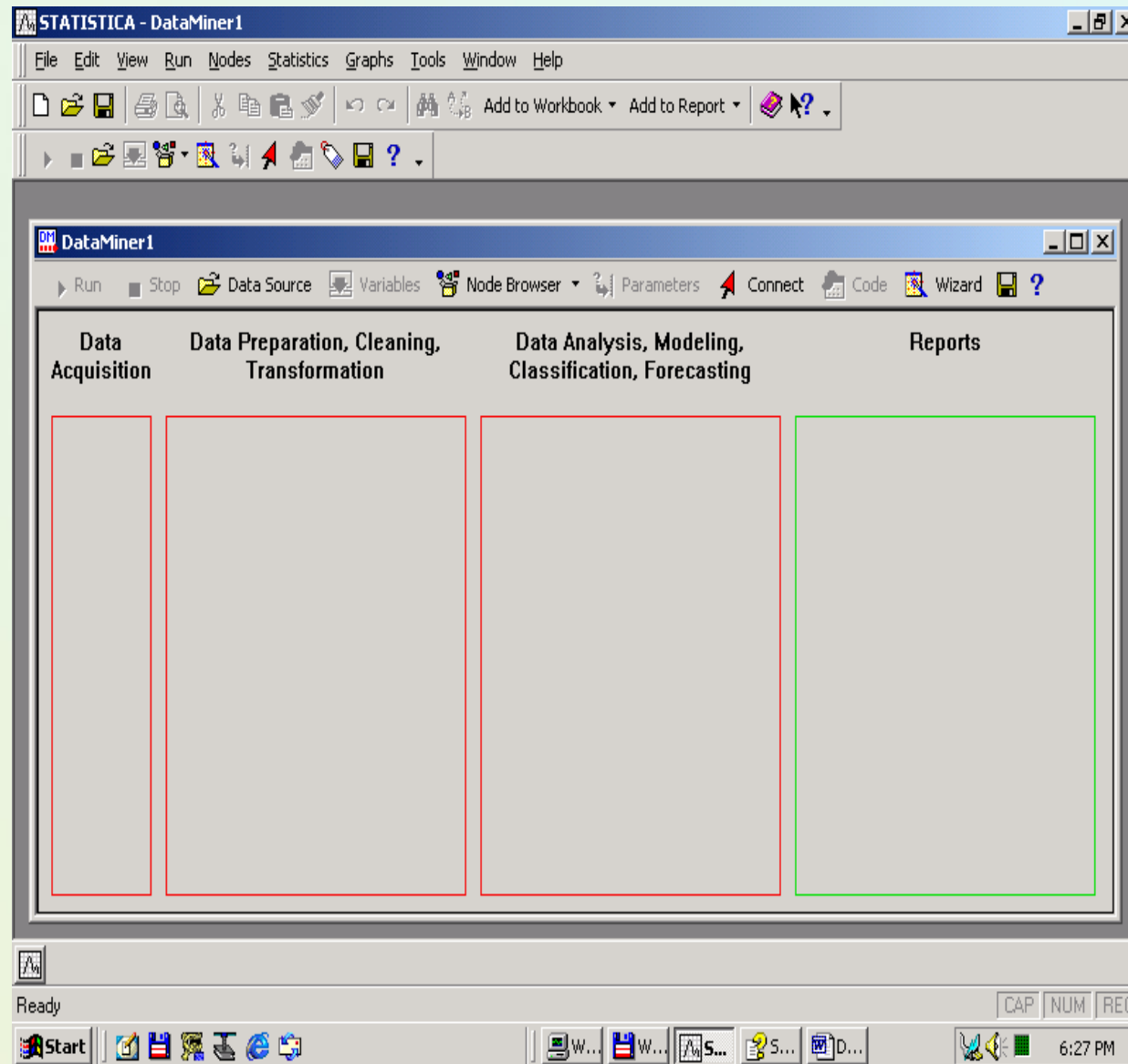
---

- 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

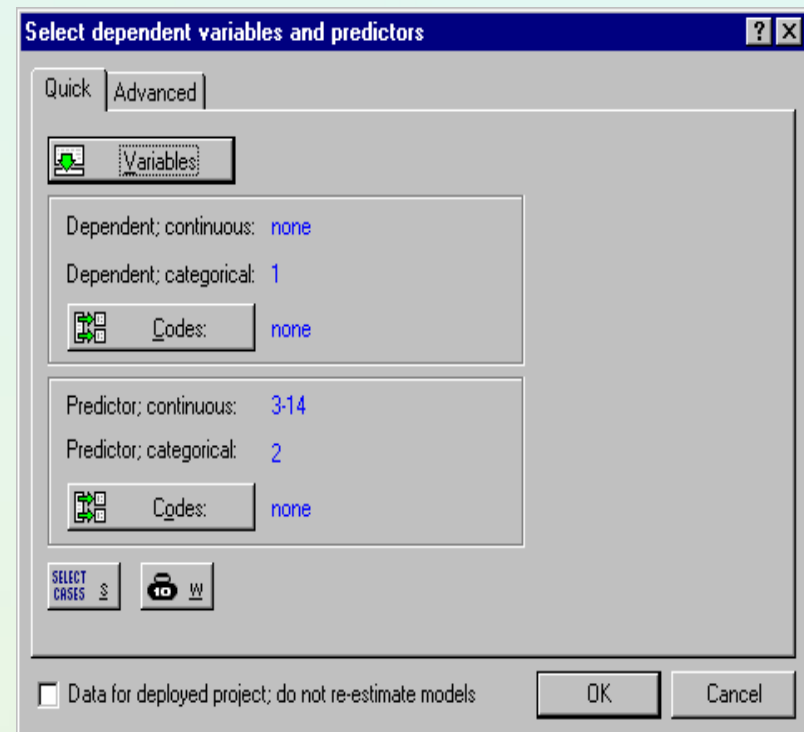
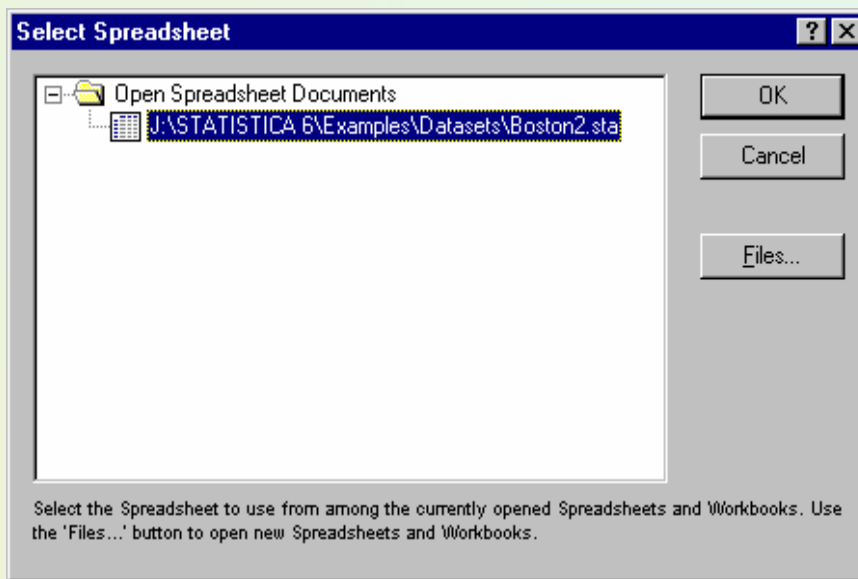
## Main systems:

- 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

---

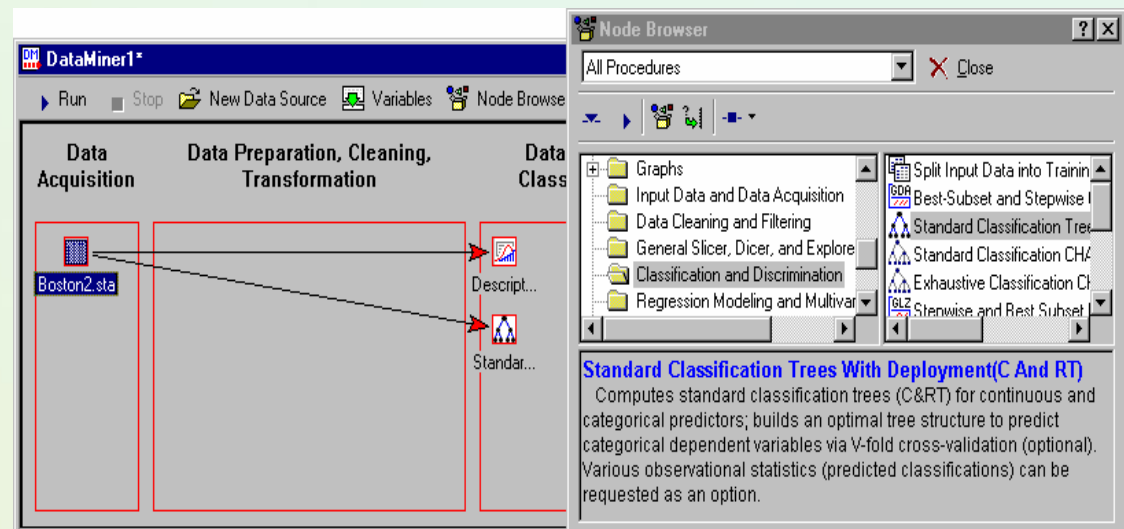
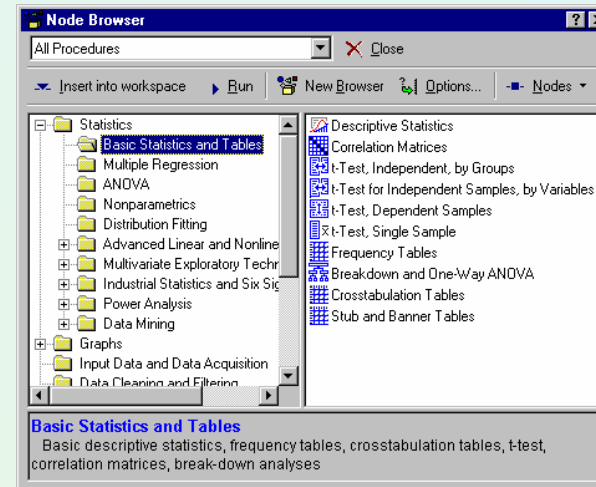
- Data Miner - General Slicer/Dicer Explorer with Drill-Down
- Data Miner - General Classifier (Trees and Clusters)
- Data Miner - General Modeler and Multivariate Explorer
- Data Miner - General Forecaster
- Data Miner - General Neural Network Explorer

---

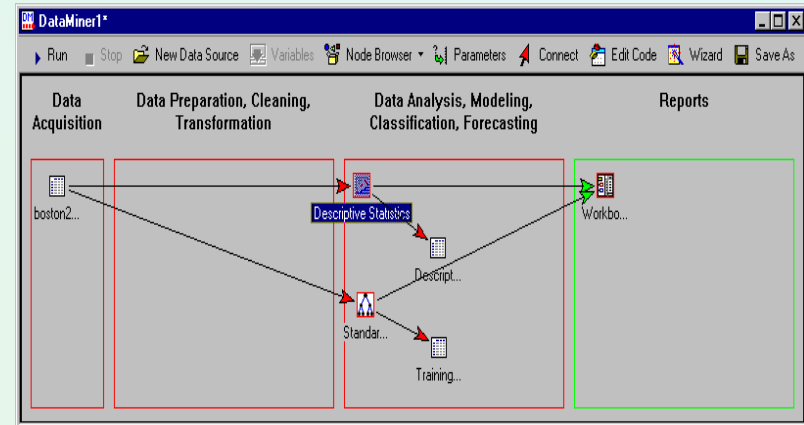
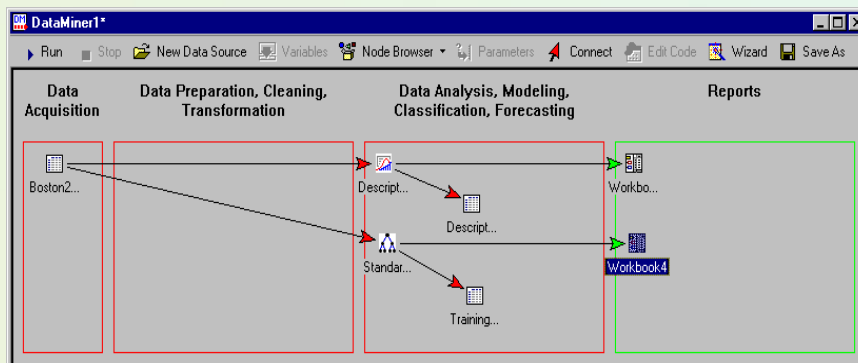
- Neural Networks
- Generalized EM & k-Means Cluster Analysis
- Association Rules
- General Classification/Regression Tree Models
- General CHAID Models
- Interactive Trees (C&RT, CHAID)
- Boosted Tree Classifiers and Regression
- Generalized Additive Models
- MAR Splines (Multivariate Adaptive Regression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

---

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



# Extra tools for defining projects



Workbook3\* - Descriptive Statistics (Boston2.sta)

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

Workbook4\* - Tree 1 layout for PRICE

The screenshot shows two workbooks. Workbook3\* displays a table of descriptive statistics for the variable 'ORD' (ORD1 to ORD7). Workbook4\* displays a decision tree diagram for the variable 'PRICE', showing a hierarchical structure of nodes and branches.

Node Browser

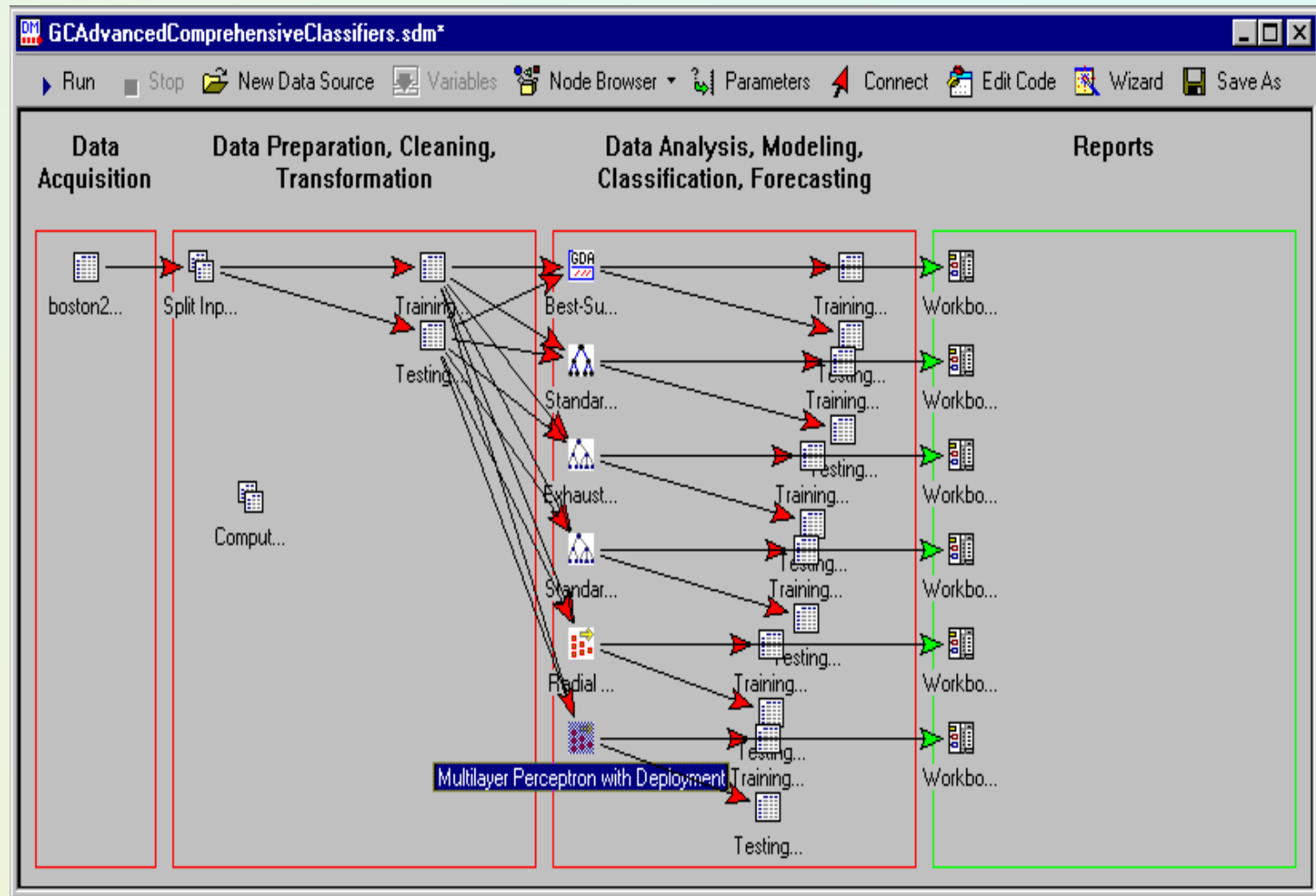
All Procedures

Insert into workspace Run New Browser Options... Nodes

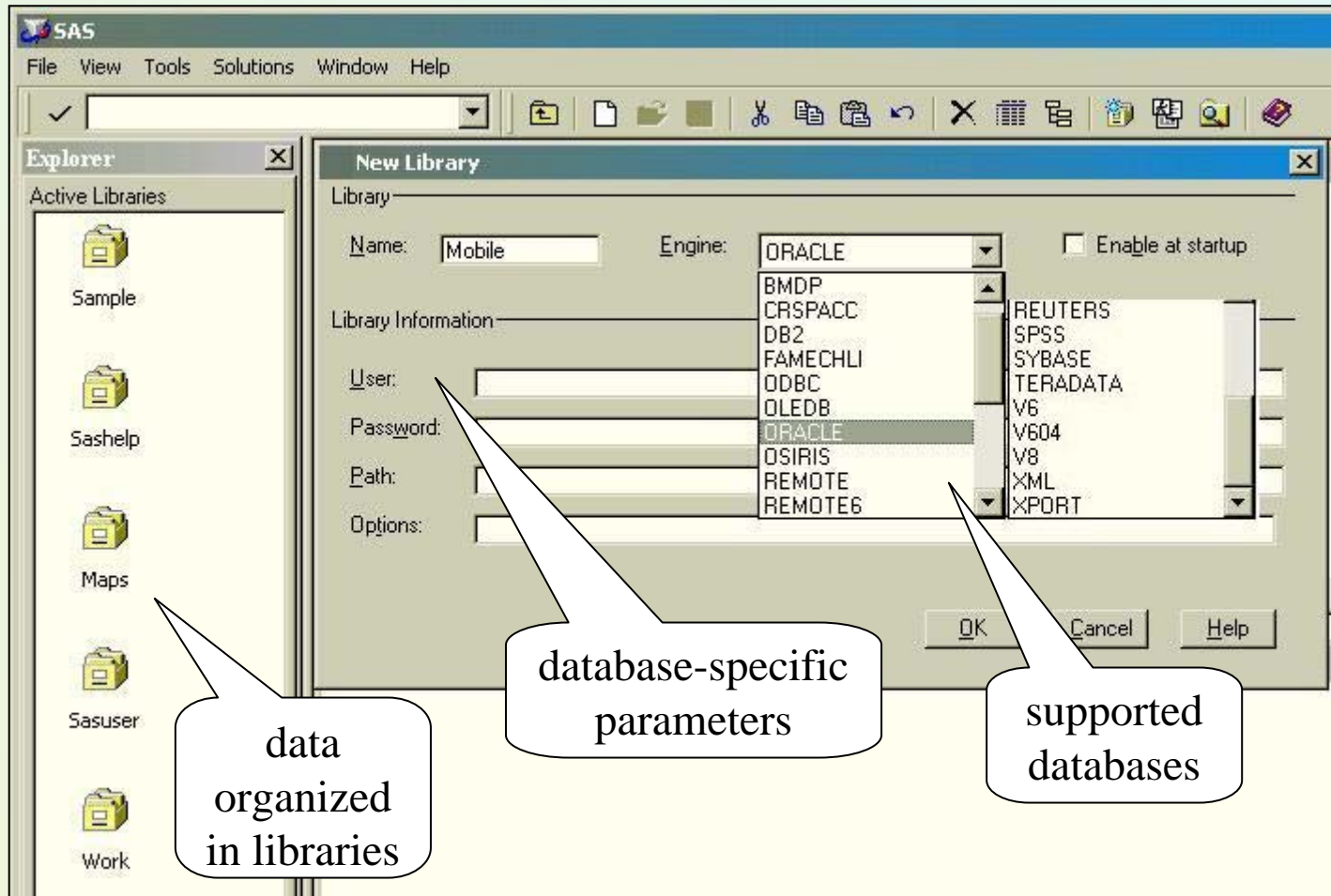
- Data Mining
  - Graphs
  - Input Data and Data Acquisition
  - Data Cleaning and Filtering
  - General Slicer, Dicer, and Explorer
  - Classification and Discrimination**
    - Split Input Data into Training and Testing Samples
    - Best-Subset and Stepwise GDA ANCOVA With Deployment
    - Standard 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
  - Regression Modeling and Multivariate Explore
  - General Forecaster and Time Series
  - Neural Network Architectures
  - Comparing and Merging Multiple Data Source
  - User defined and Special Purpose Models

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

# Using several methods on the same data

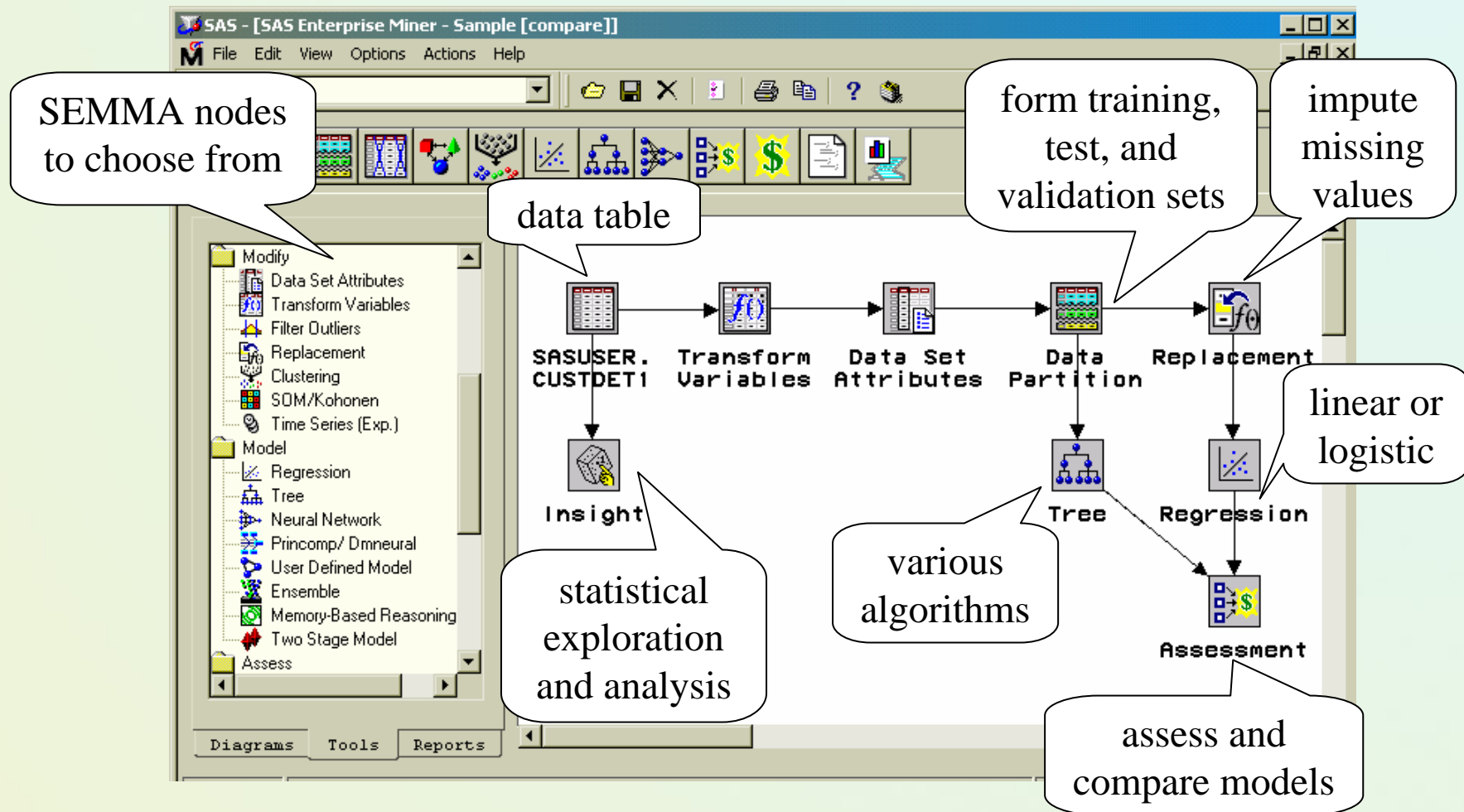


# SAS Enterprise Miner



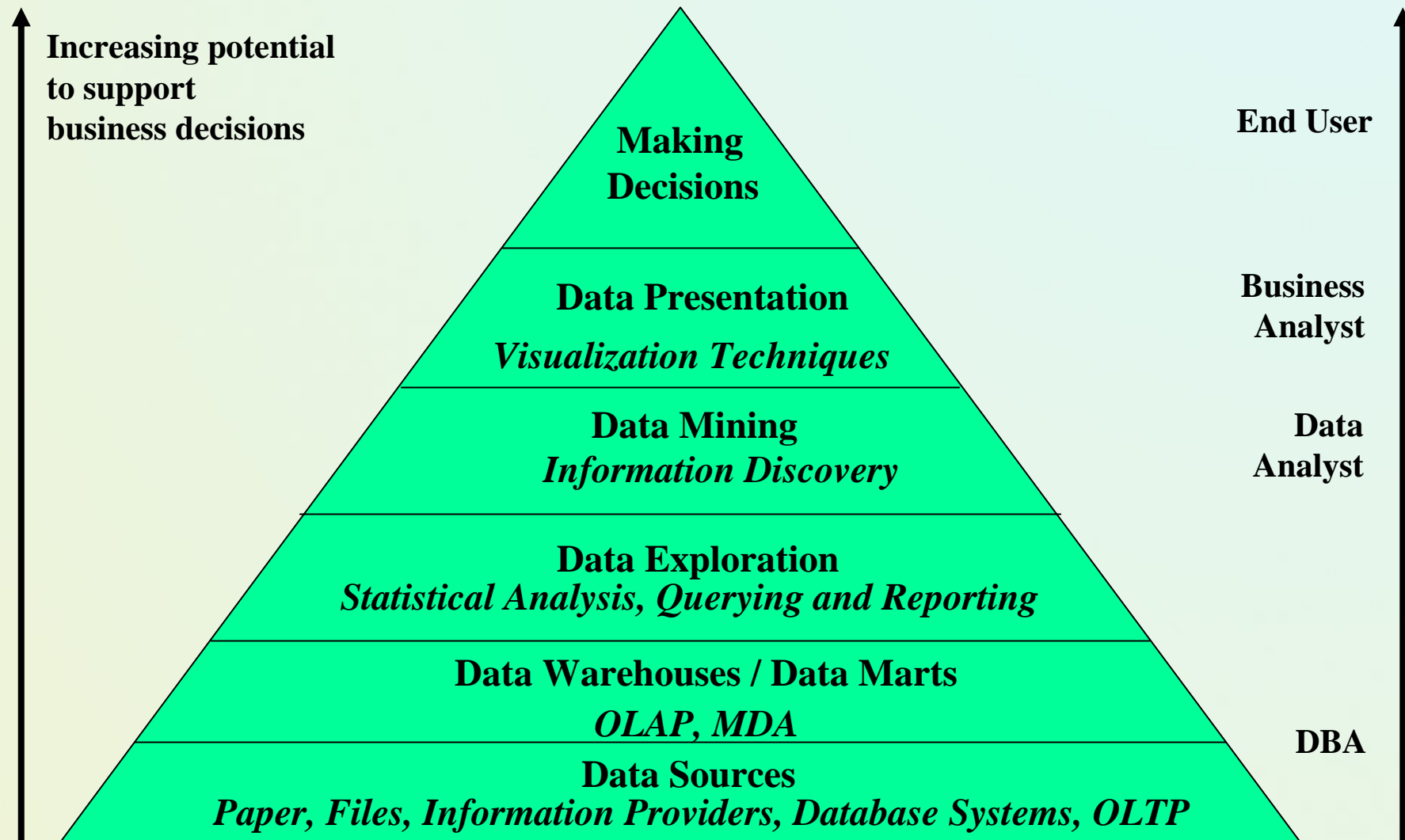


# Enterprise miner project



# Data Mining and Business Intelligence

---



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

---

**Banking** (29) 13%

**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%

# 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

# 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.





Thank you !