Data Mining in Ship Construction & Operation

A review of innovative methods and Open Software's

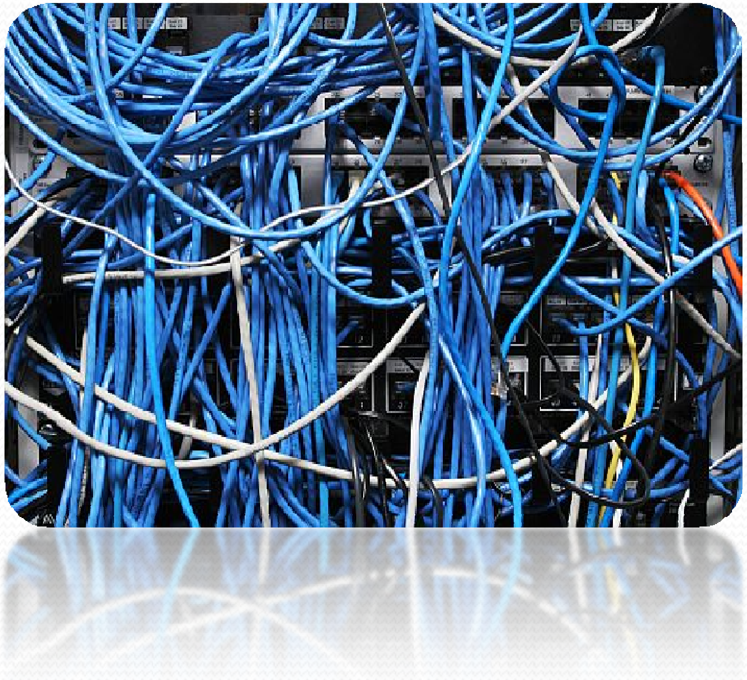
Dr. Jean-David Caprace
Post PhD at UFRJ

January 2011

Summary

Outline of the presentation

- Data Mining
- Mathematical models
- Applications
- Open software's comparison



Data Mining (DM)

Why?

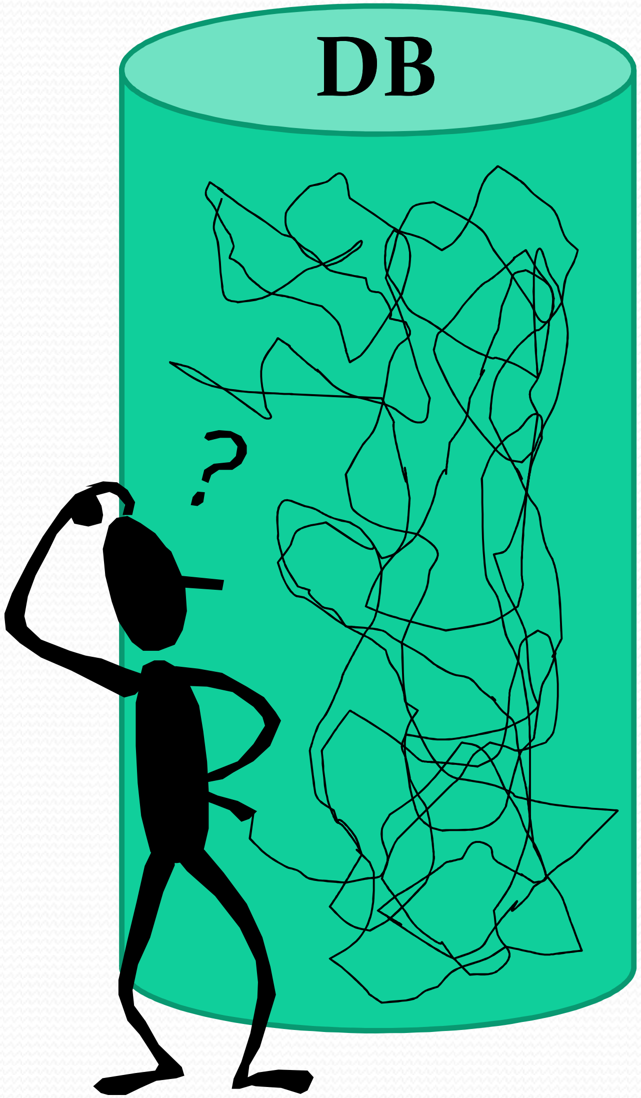
- Complexity of modern manufacturing processes
- Massive investment in automation & monitoring systems

➔ Generation of large DBs

➔ DBs underused

➔ Human analysts take weeks to discover useful information

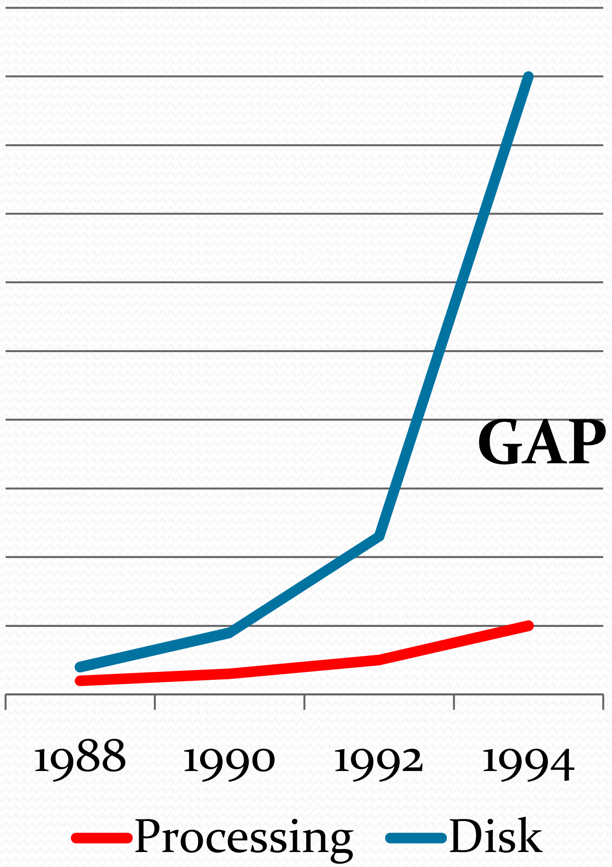
➔ **How to solve this problem?**



Data Mining (DM)

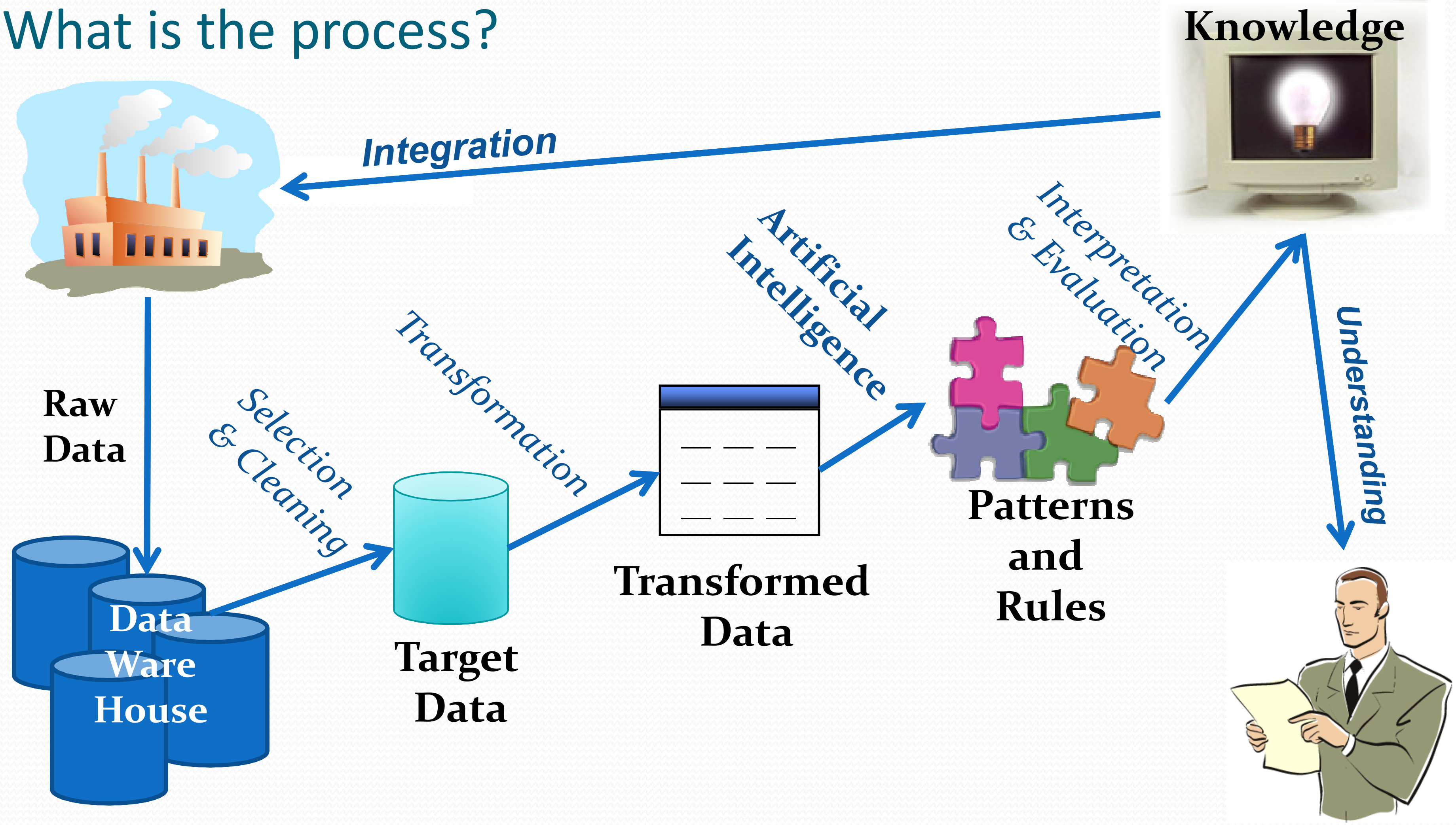
Why?

- Moore's law
  - Computer speed doubles every 18 months
- Storage law
  - Total storage doubles every 9 months
- Consequence
  - very little data will ever be analyzed by humans
- Knowledge discovery is **NEEDED** to make sense and use of data ➔ Data Mining



Data Mining (DM)

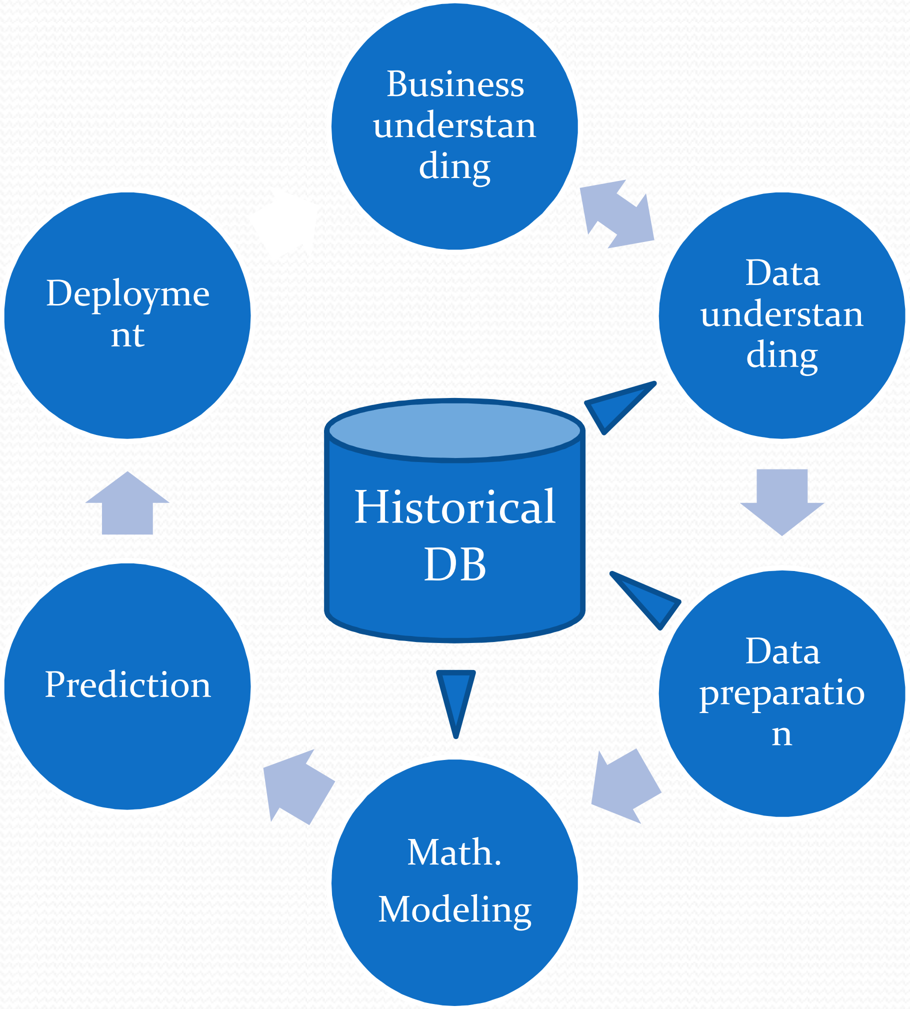
What is the process?



Data Mining (DM)

What is the methodology?

- Easiness to retrieve the knowledge
- Detect **hidden** and **complex** relationships
- The crisp DM process (www.crisp-dm.org)
  - CRoss Industry Standard Process for DM = **World Standard**
  - Step-by-step data mining guide



Data Mining (DM)

What is the methodology?

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Data Mining (DM)

Result of DM includes …

- **Forecasting** what may happen in the future
- **Classifying** objects into groups by recognizing patterns
- **Clustering** objects into groups based on their attributes
- **Associating** what events are likely to occur together
- **Sequencing** what events are likely to lead to later events

Data Mining (DM)

Is **not** …

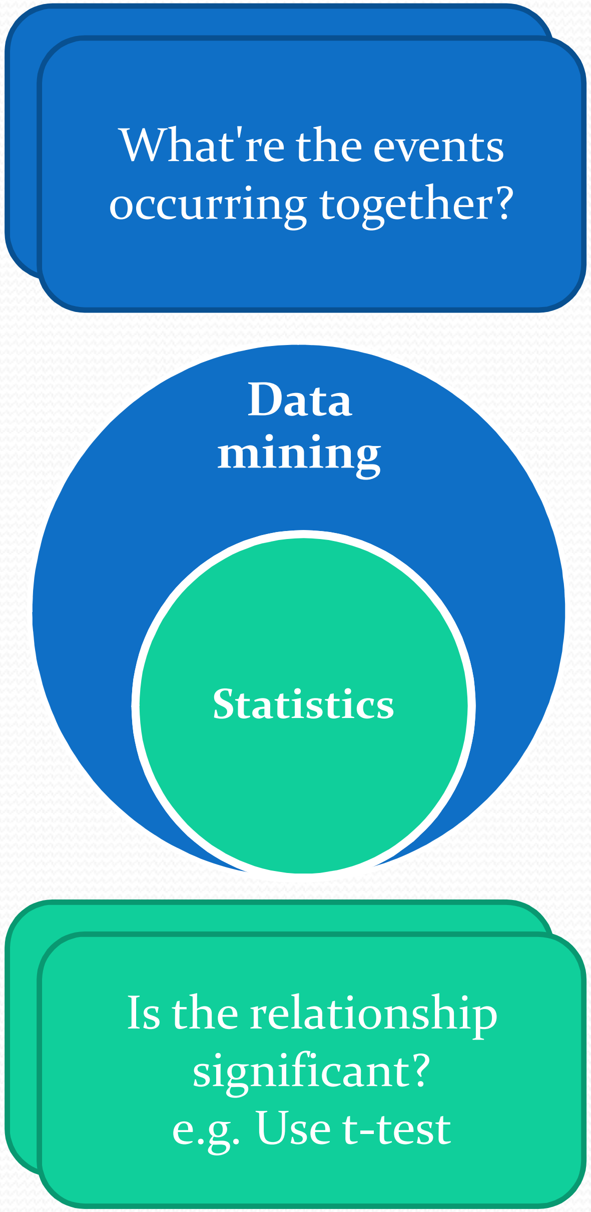
- Brute-force crunching of bulk data
- “Blind” application of algorithms
- Going to find relationships where none exist
- Presenting data in different ways
- A database intensive task
- A complex technology requiring an advanced degree in computer science

Data Mining (DM)

Versus statistical analysis

- **Statistics**
  - Core of Data Mining
  - Help to make the difference between noise and significant findings

- **Data Mining**
  - Cover the entire data analysis process
  - Knowledge extraction



What're the events occurring together?

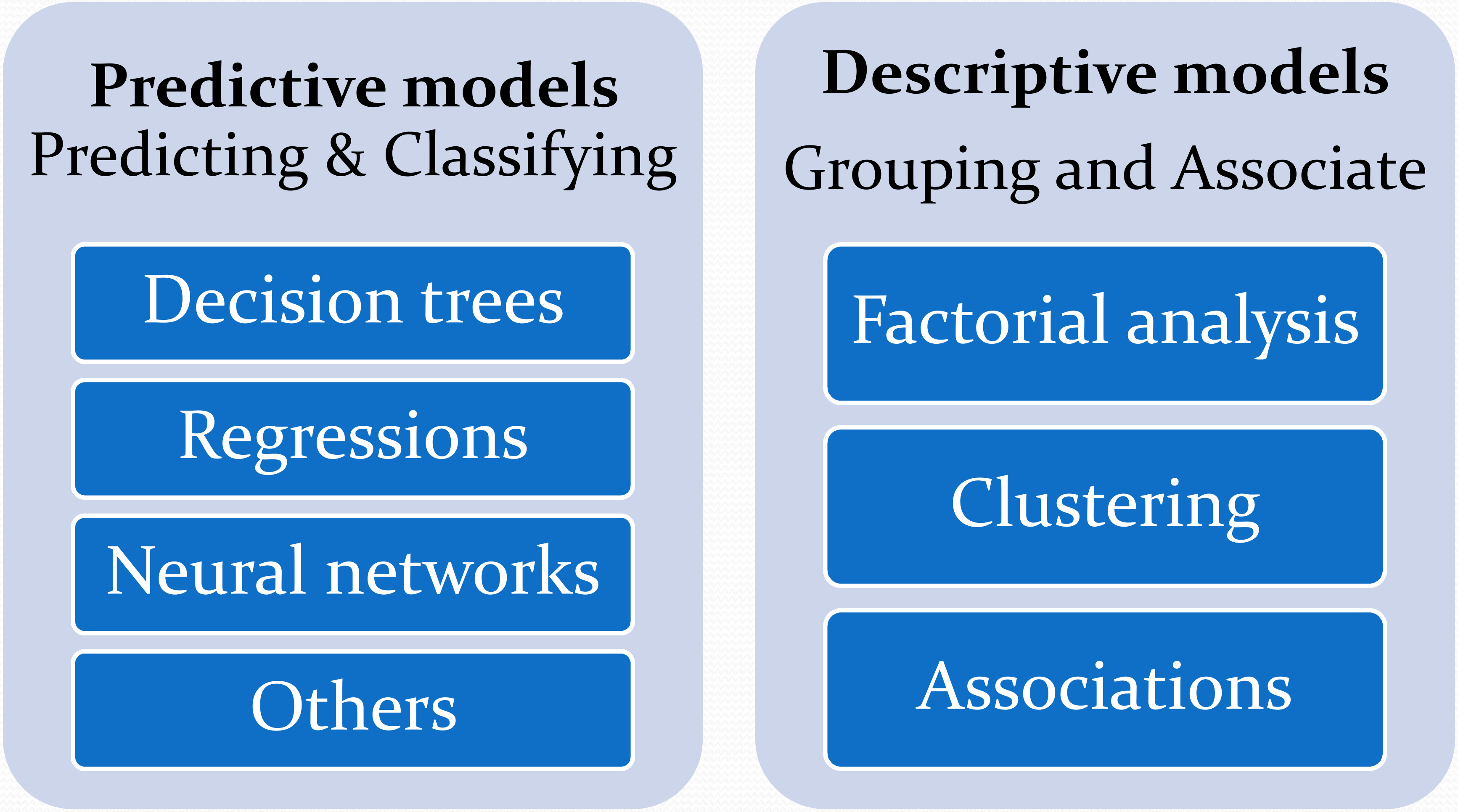
Data mining

Statistics

Is the relationship significant?
e.g. Use t-test

Mathematical models

Different types



Predictive models
Predicting & Classifying

- Decision trees
- Regressions
- Neural networks
- Others

Descriptive models
Grouping and Associate

- Factorial analysis
- Clustering
- Associations

Mathematical models

Predictive models

Decision trees

- Classification tree
- ID3
- **C4.5**
- CHAID
- ECHAID
- CART
- C5
- J48
- QUEST
- M5P

Regressions

- Multi linear regression
- Polynomial regression
- Logistic regression
- Proportional hazards models – Cox
- Partial Least Squares regression - PLS

Neural networks

- **Multi Layer Perceptron - MLP**
- Probabilistic Neural Network – PNN
- **Plenty ➔** AHC, TDNN, ARP, AMF, ALN, GRNN, BSB, FCM, BM, MFT, RCC, BPTT, RTRL, EKF, AG, BAM, TAM, etc.

Other supervised learning

- Bayesian networks
- Support Vector Machine – SVM
- SVM for Regression - SVR

Mathematical models

Descriptive models

Factorial analysis

- Principal Component analysis - PCA
- Independent Component Analysis - ICA
- Correspondence analysis
- Multiple correspondence analysis
- Multiple discriminant analysis

Clustering

- Hierarchical clustering – dendrograms
- **K-Means**
- X-Means
- K-Medoids
- Fuzzy c-Means
- Self Organizing Maps – SOM
- Nearest Neighbor Search – NNS
- Expectation-Maximization
- Optics

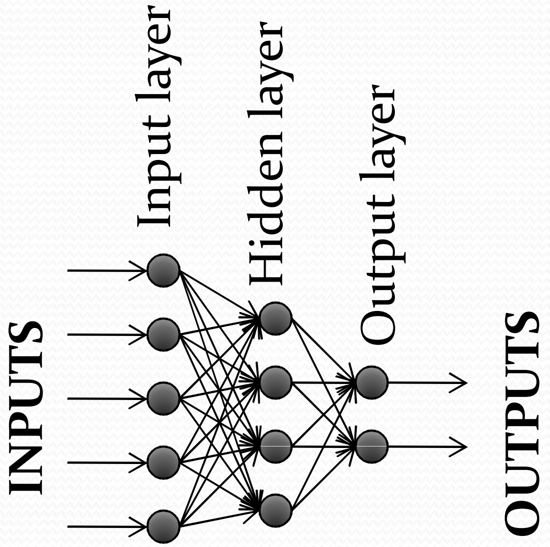
Association

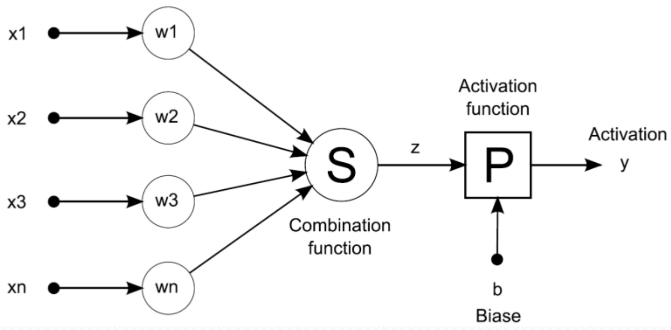
- **Apriori**
- Generalized Rule Induction – GRI
- Carma
- Tertius
- Generalized Sequential Patern - GSP

Artificial Neural Networks

What is it?

- Advantages
  - Learn from training experience
  - Extract non-linear relationships
  - Works with all data types
  - Accurate
- Drawbacks
  - Risk of 'overfit' the data
  - Extensive amount of training time
  - Black box → difficult interpretation





Artificial Neural Networks

Possible applications …

- Straightening cost/time assessment during ship design

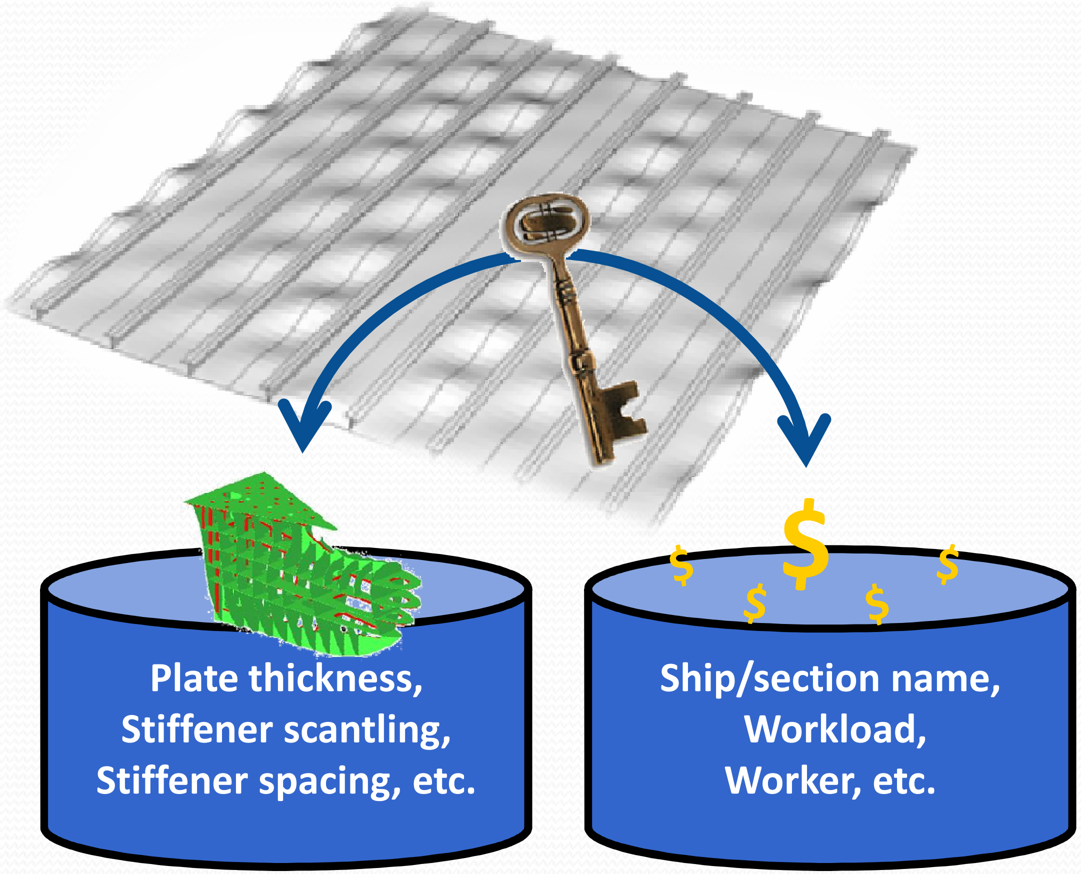
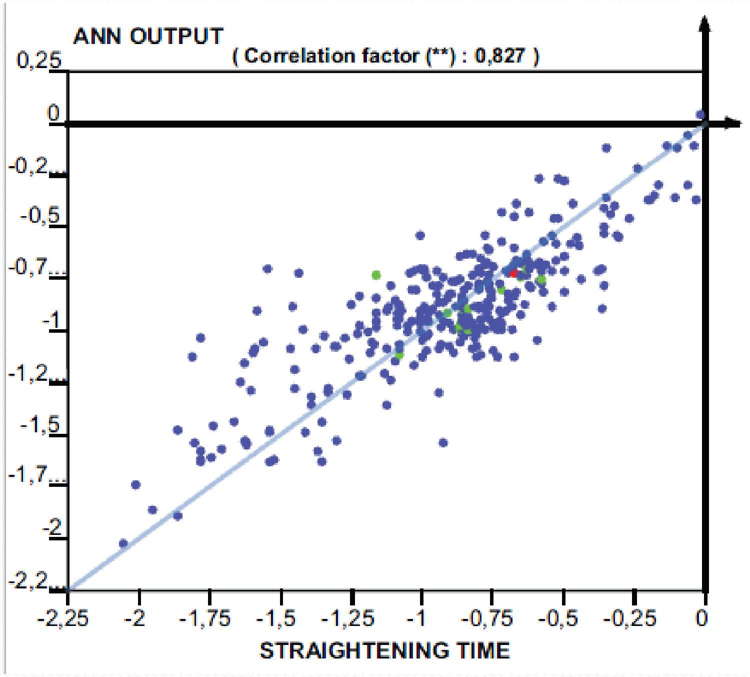


Plate thickness,
Stiffener scantling,
Stiffener spacing, etc.

Ship/section name,
Workload,
Worker, etc.



9 inputs parameters
1 output parameter
R2 = 0.827

Artificial Neural Networks

Possible applications …

- Blocks cost/time assessment during ship design

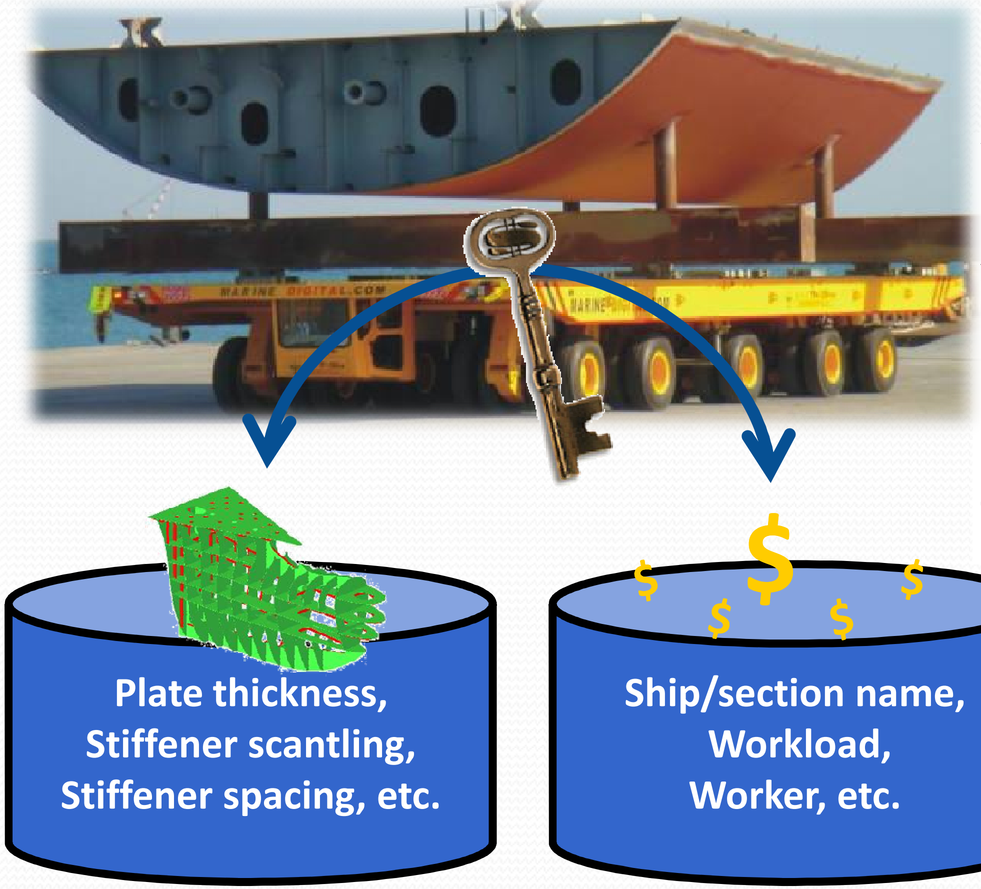
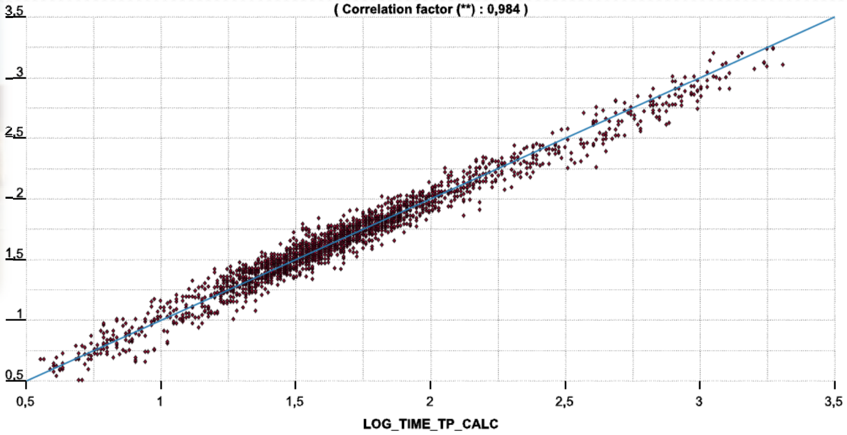


Plate thickness,
Stiffener scantling,
Stiffener spacing, etc.

Ship/section name,
Workload,
Worker, etc.



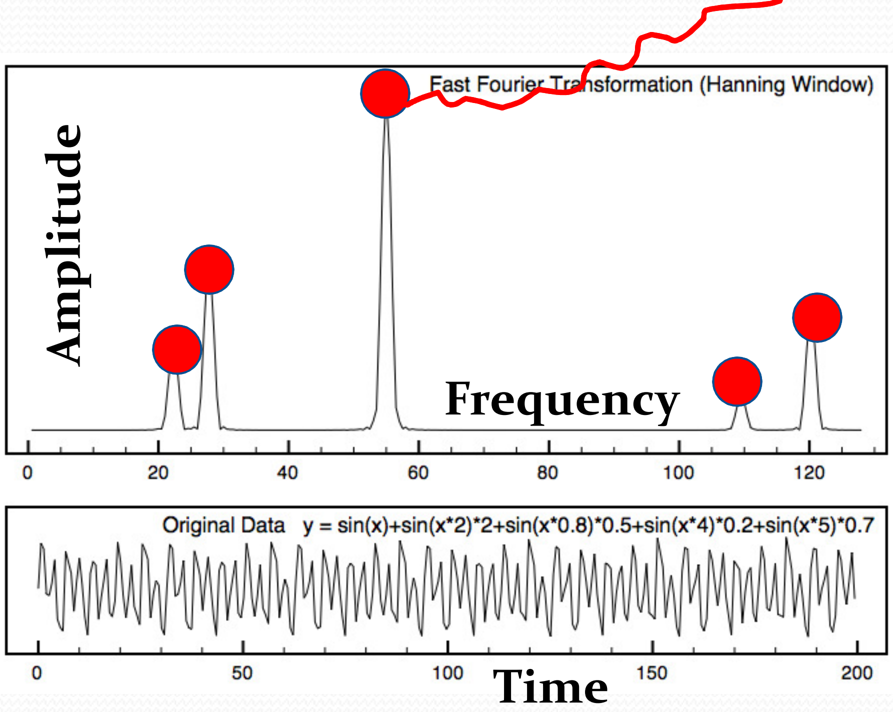
12 inputs parameters
1 output parameter
R2 = 0.984

Artificial Neural Networks

Possible applications …

- Place of corrosion prediction
- Part failure prediction

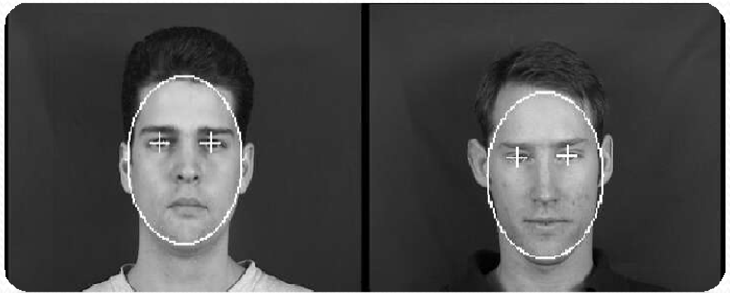
(Conditioned Based Maintenance)

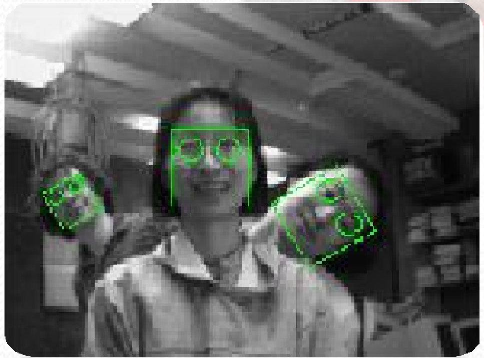


Artificial Neural Networks

Possible applications …

- Face detection
  - Count number of passenger in a cruise ship (evacuation)
  - Recognize passenger in a cruise ship
  - Detection of intrusions (terrorism)

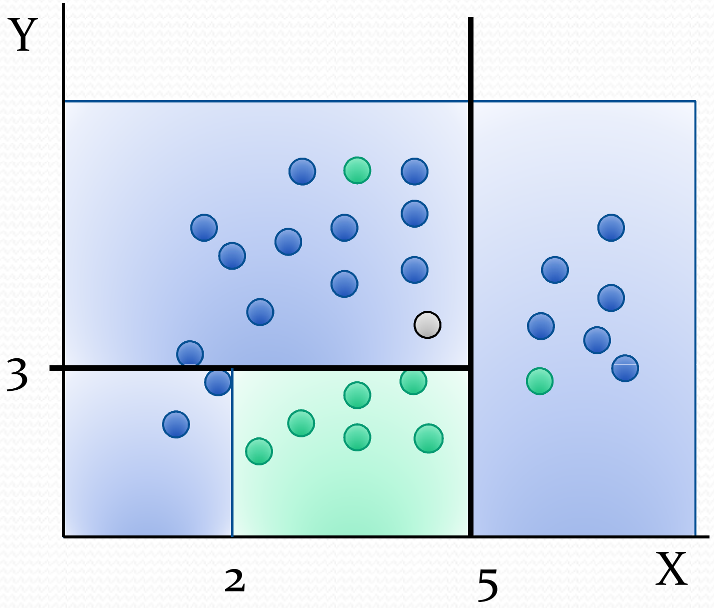




Decision trees

What is it?

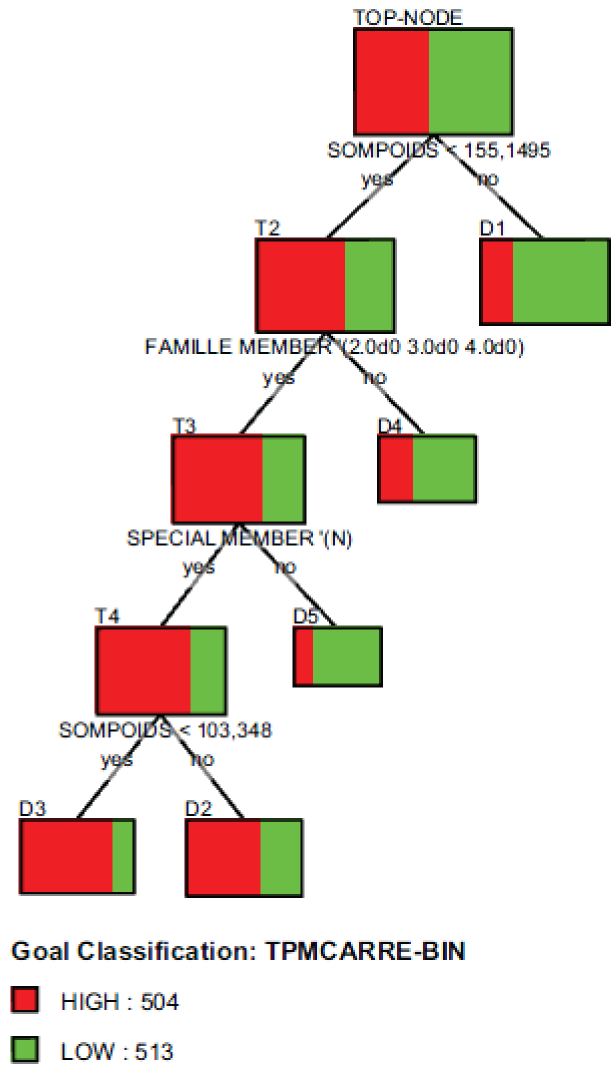
- Advantages
  - Intuitive outputs
  - Handle all types of attributes (numeric and symbolic)
  - Have value even with little hard data
  - Can be combined with other DM
- Drawbacks
  - Target must be symbolic



if X > 5 then blue
else if Y > 3 then blue
else if X > 2 then green
else blue

Decision trees
Possible applications…

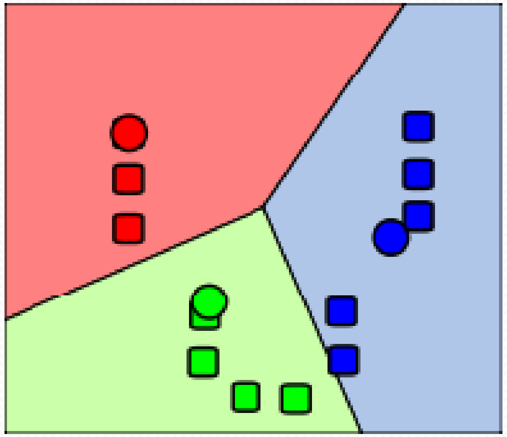
- Identify the rules that generate costs in the design or the operation of a ship
- Prediction of symbolic attributes
  - What is the risk of (high, …, low)?
  - What is the cost of (high, …, low)?
  - Characterization of the complexity of a system (for maintenance)



Clustering models

What is it?

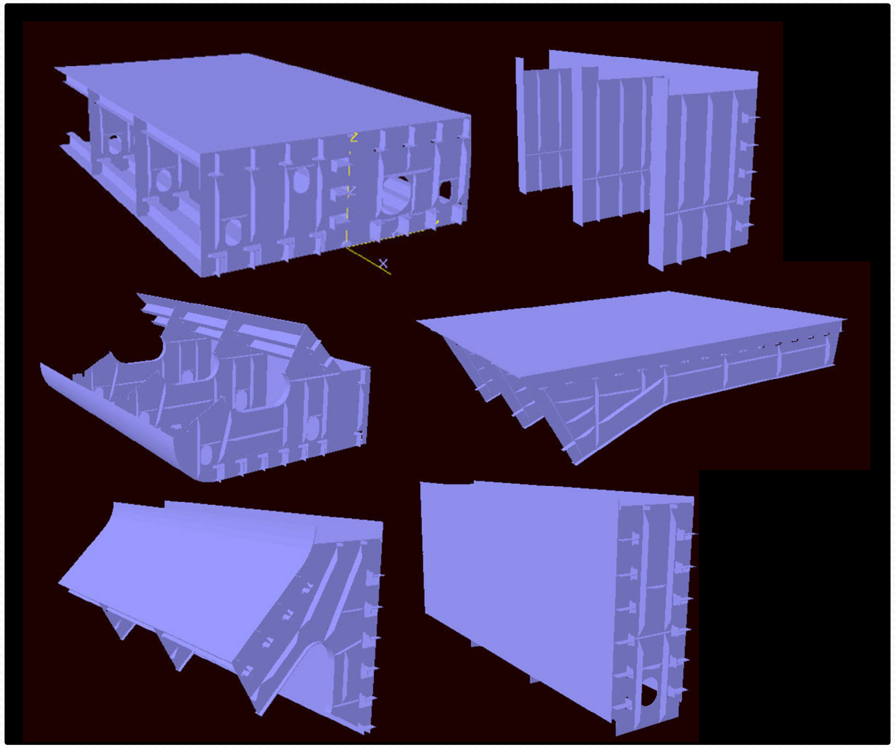
- Advantage
  - Data segmentation
    - Reduce the quantity of data for future analyze
  - Classification of the data in different groups
  - Extraction of knowledge about not known groups
- Drawbacks
  - Sometimes give different results for each run
  - Concept of mean is often required



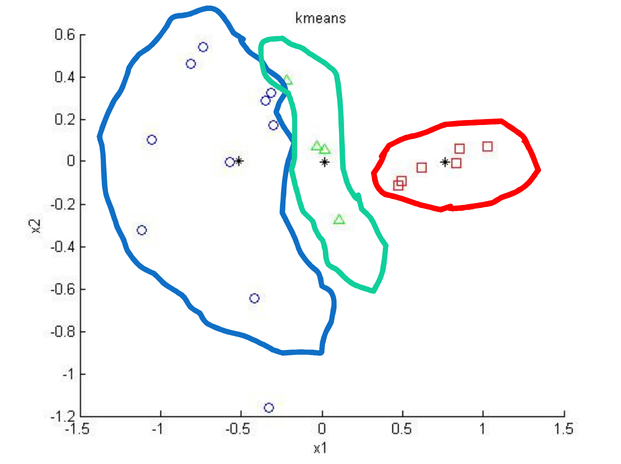
Clustering models

Possible applications …

- Classification of ship parts/section/blocks
- Identify different groups of ships in a fleet
- Gather different identical event sequences (maintenance/repair)



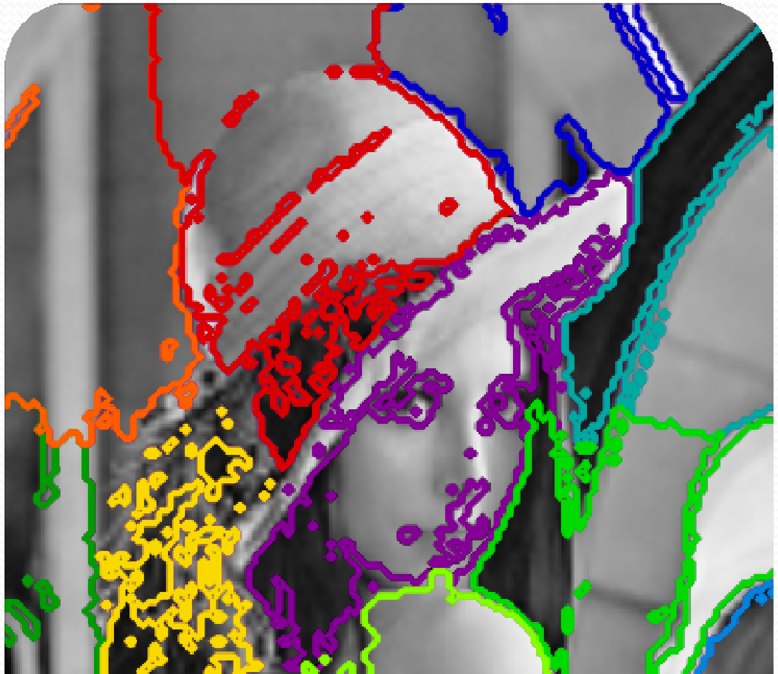
- Possibility to combine with another model for the prediction



Clustering models

Possible applications …

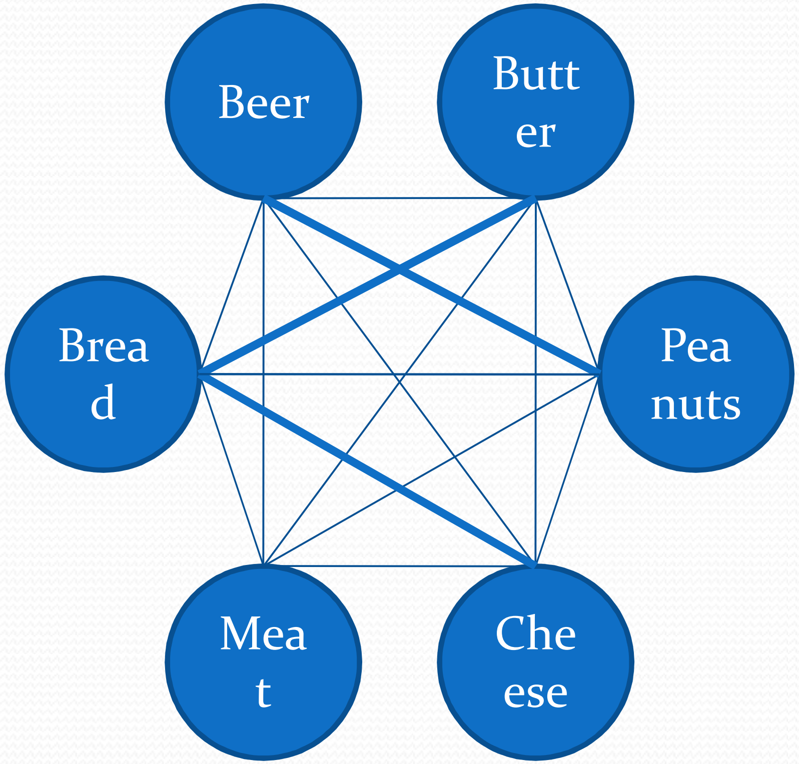
- Image segmentation
  - Contour detection
- Edge distortion measurement?
- Corrosion surface detection



Association rules

What is it?

- Advantages
  - Rules are intuitive
  - Works with huge DB
  - Can detect event sequences
- Drawbacks
  - Huge number of rules
    - Need to be filtered



Association rules

Possible applications …

- Ship maintenance and operation
  - Do certain faults/incidents lead to specific repairs?
  - Do certain repairs produce subsequent faults/incidents?
  - Are there repairs that lead to other repairs?

Data mining software's

Open source or commercial?

- Open source software's
  - Considerably improved
  - Integrates huge number of different algorithm's
  - Can manage huge DB
- Commercial software's
  - Better access to different DB
  - Better exploitation of models
  - Better reporting

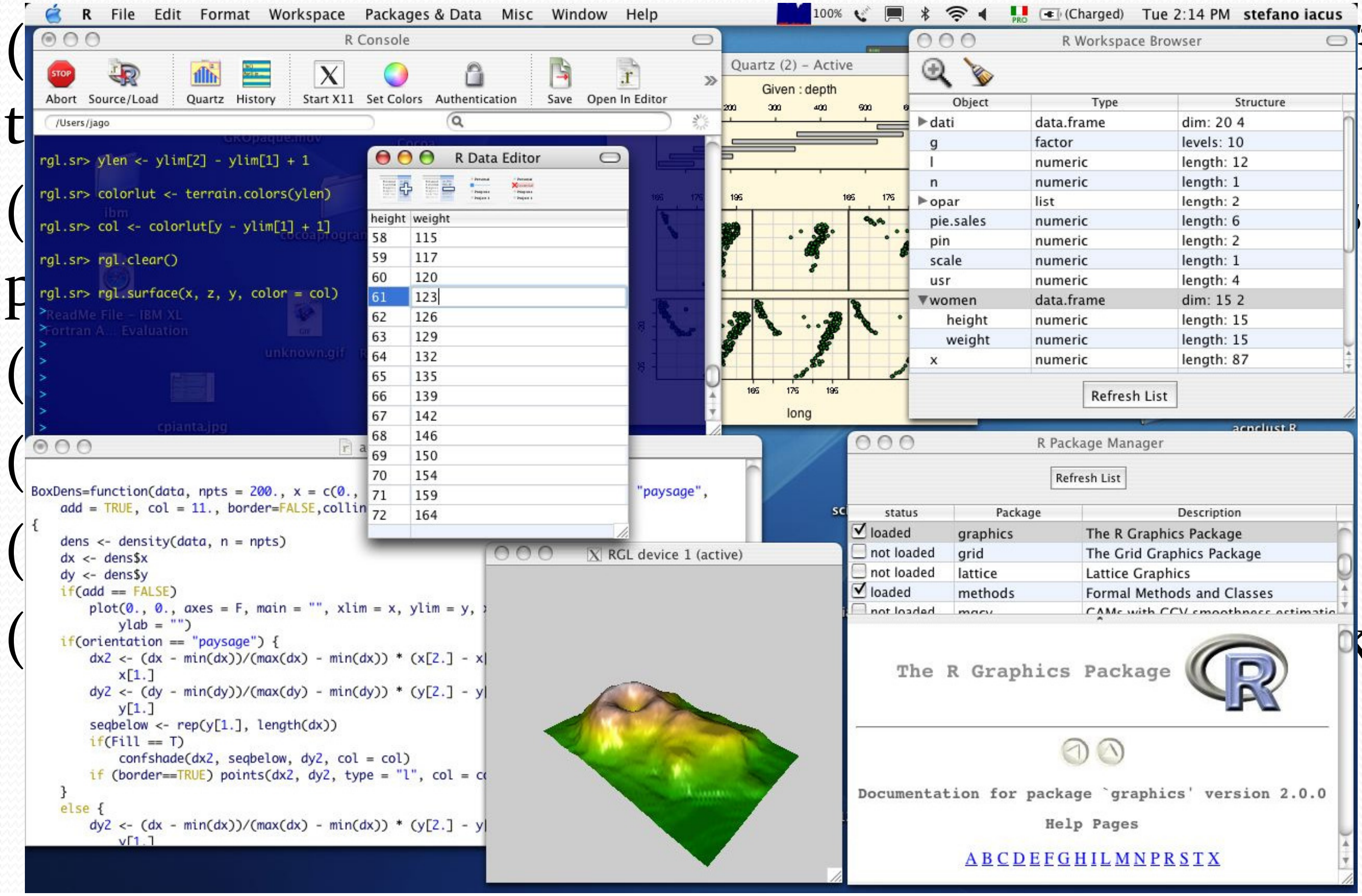
Knime
Tanagra
Weka
R
Orange
RapidMiner
...

SPAD
SAS
SPSS
STATISTICA
S-PLUS
...

Data mining software's

Open source

- R - http://www.r-project.org/
  - (icult
    t
  - (ting,
    p
  - (
  - (
  - (
  - (kages



R File Edit Format Workspace Packages & Data Misc Window Help
100% (Charged) Tue 2:14 PM stefano iacus
R Console
Abort Source/Load Quartz History Start X11 Set Colors Authentication Save Open In Editor
/Users/jago
Quartz (2) – Active
Given : depth
long
RGL device 1 (active)
R Data Editor
R Workspace Browser
R Package Manager

```
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)
>
>
>
>
>
>
>
>
```

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

Refresh List

```
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.]
```

Refresh List

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
□ not loaded	mgcv	GAMs with GCV smoothness estimatio

The R Graphics Package

Documentation for package `graphics' version 2.0.0

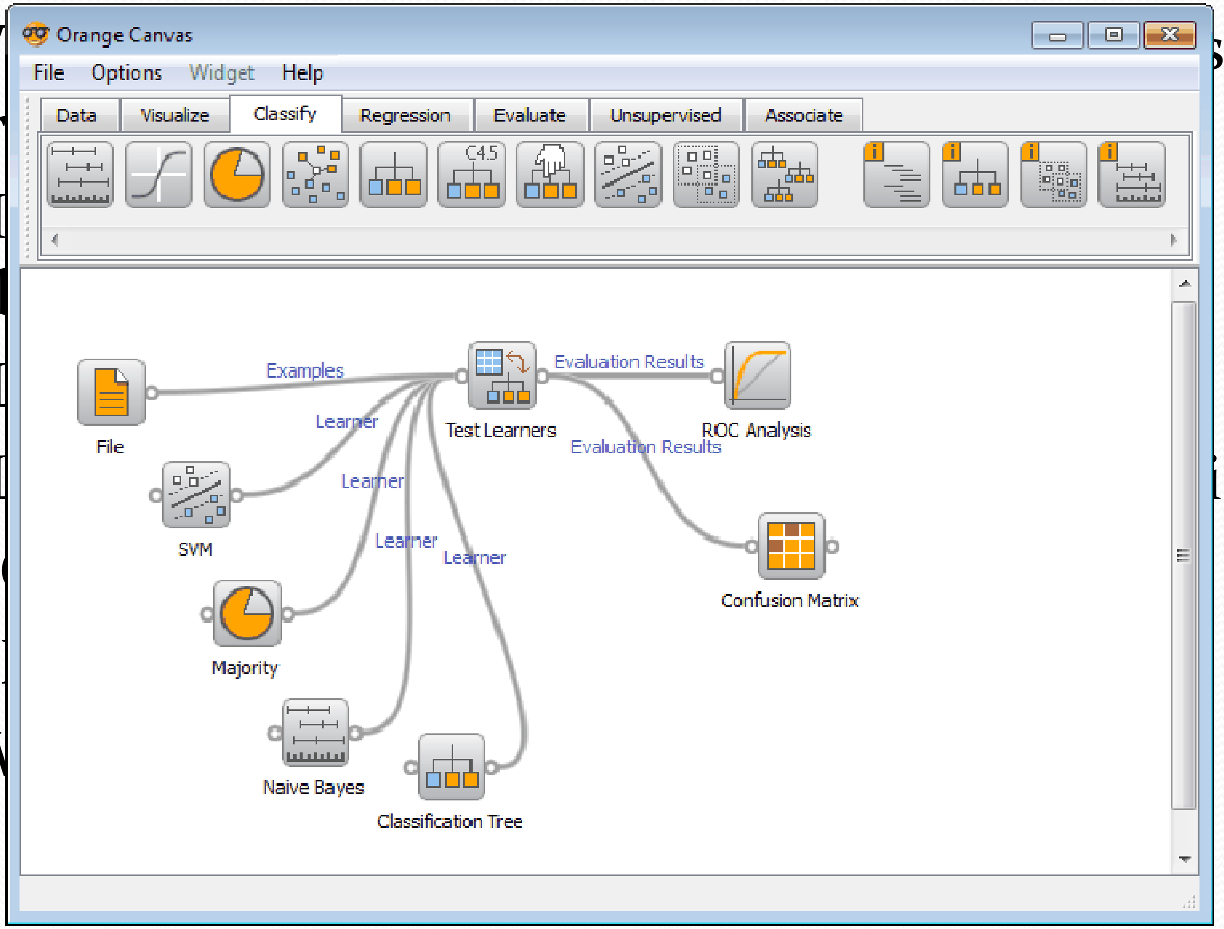
Help Pages

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

Data mining software's

Open source

- Orange - http://www.ailab.si/orange/
  - (+) T se, and stor
  - (-) I prol
  - (-) I
  - (-) I iles)
  - (+) (
  - (+) I
  - (-) V

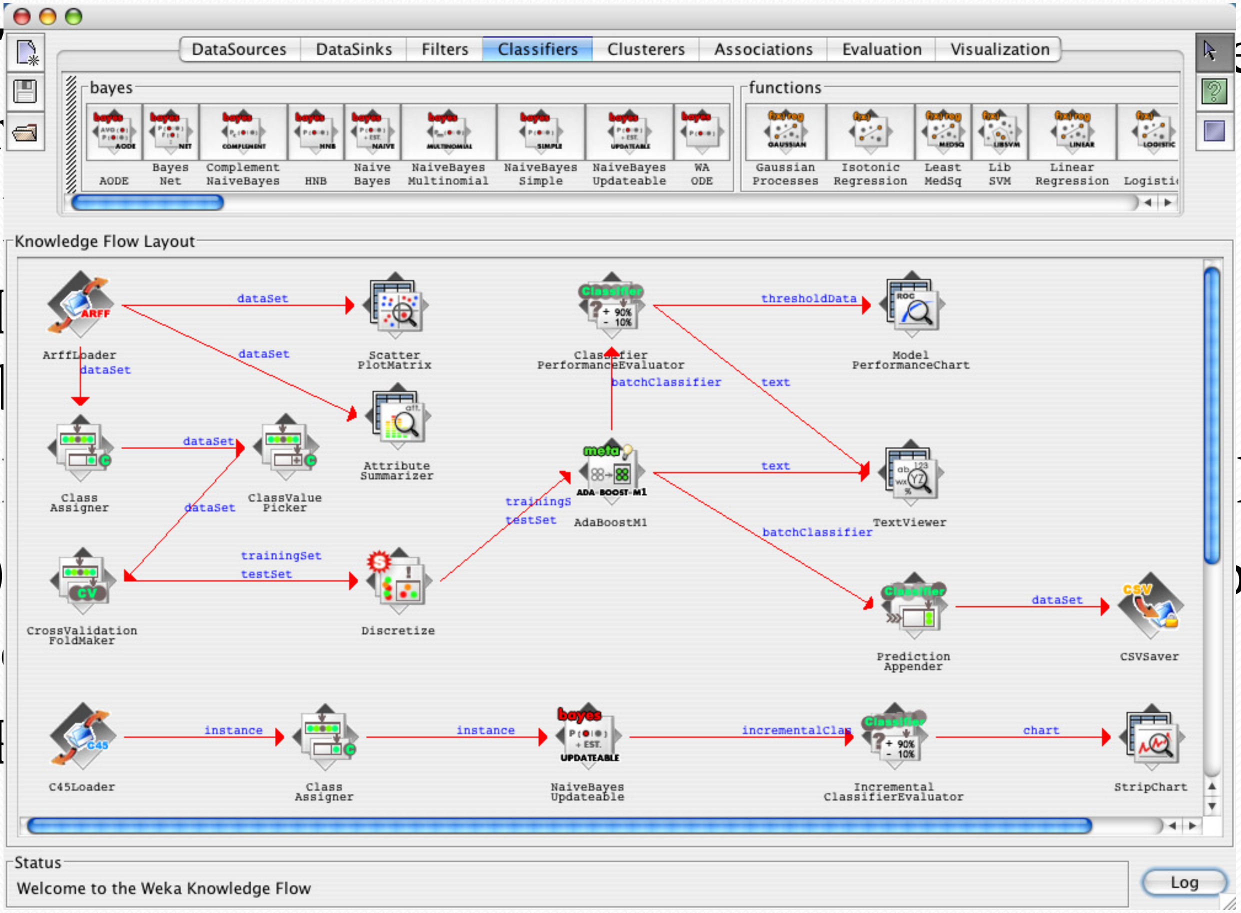


Data mining software's

Open source

WEKA
The University of Waikato

- Weka - http://www.cs.waikato.ac.nz/ml/weka/
  - (+) T… e and stor…
  - (+) …
  - (-) I… prob…
  - (+) … hm
  - (+-) … program (fre…
  - (-) I…

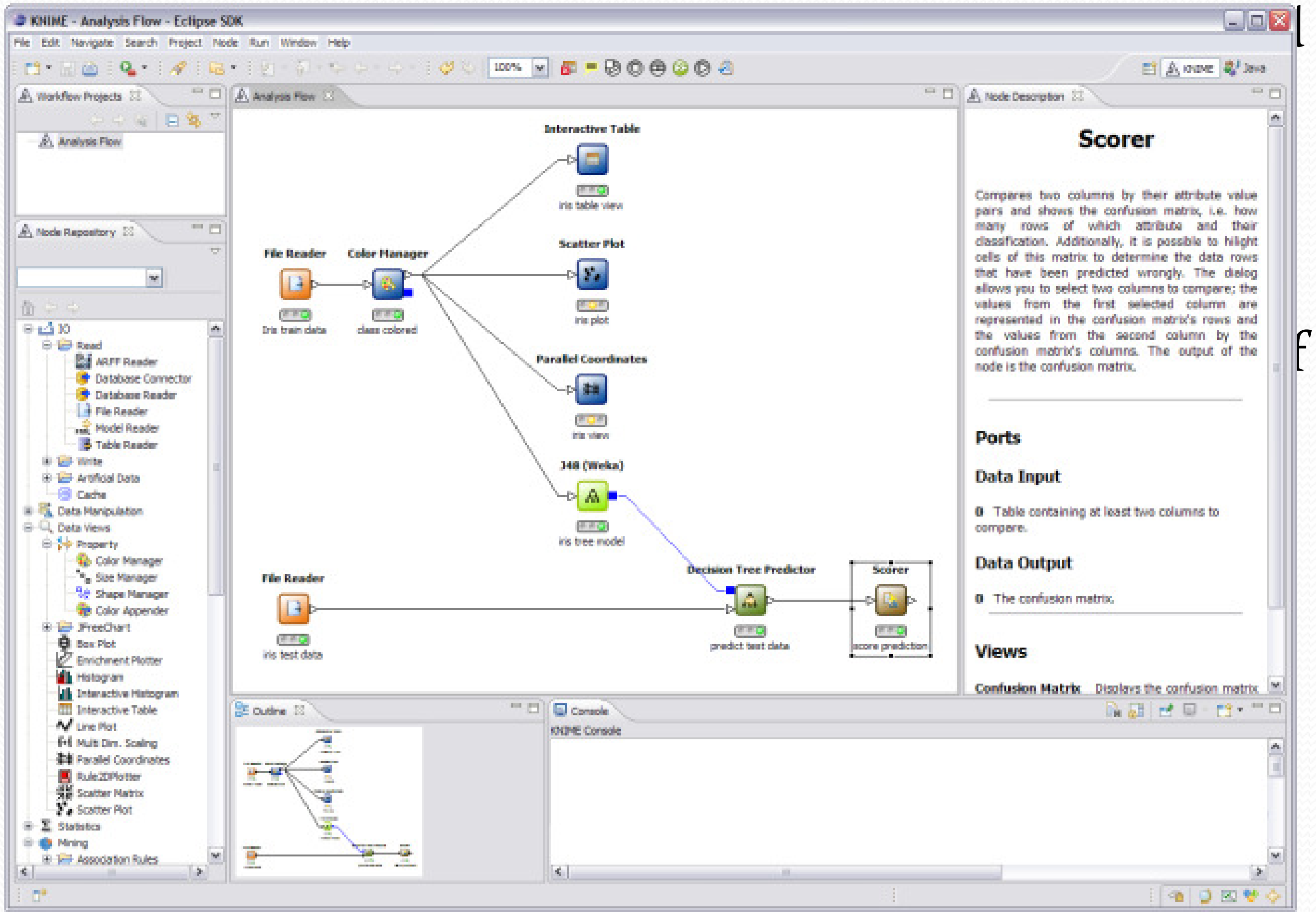


Welcome to the Weka Knowledge Flow

Data mining software's

Open source

- Knime - http://www.knime.org/
  - l store
  - f DB



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

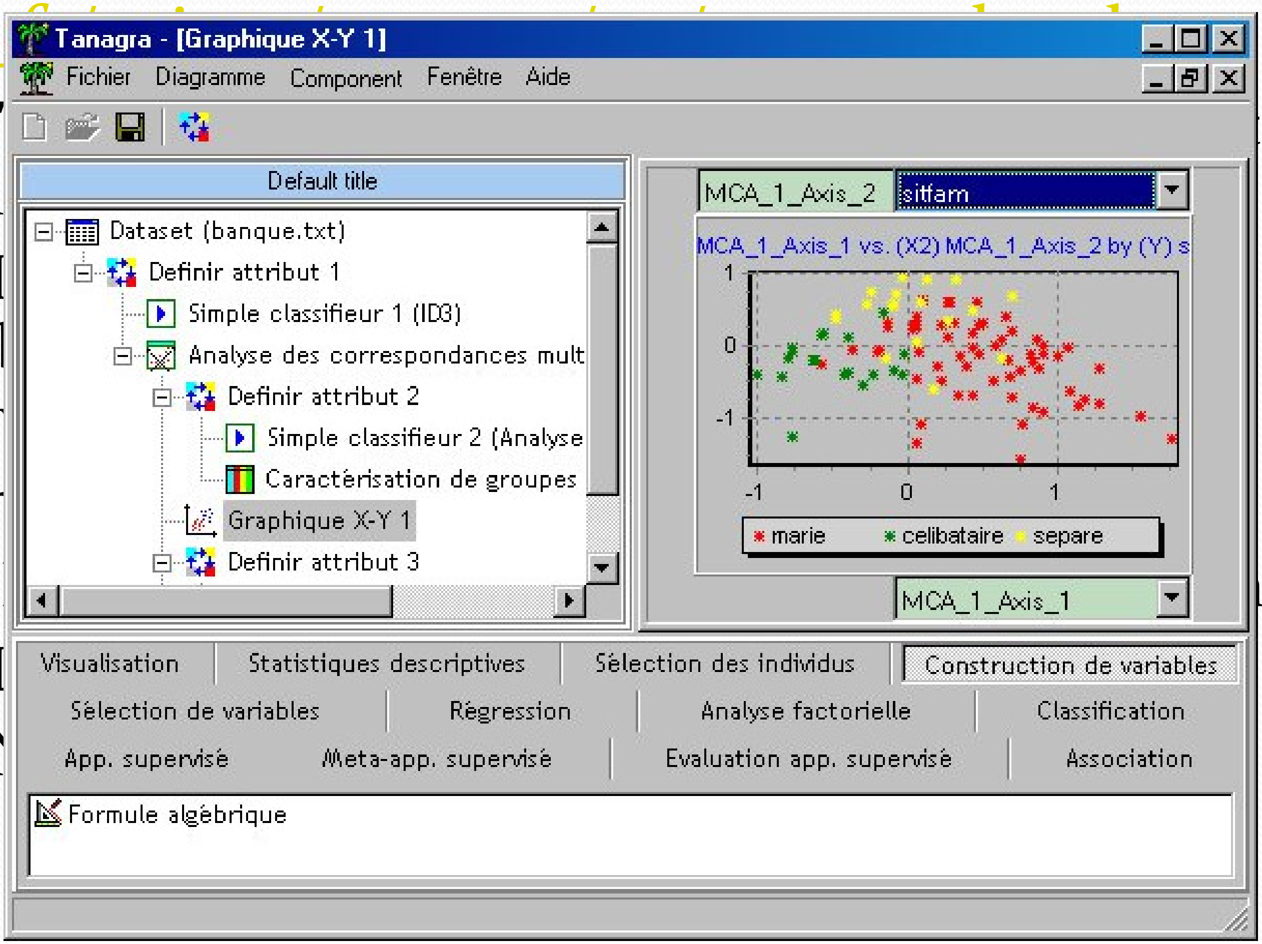
Confusion Matrix Displays the confusion matrix

Data mining software's

Open source

TANAGRA

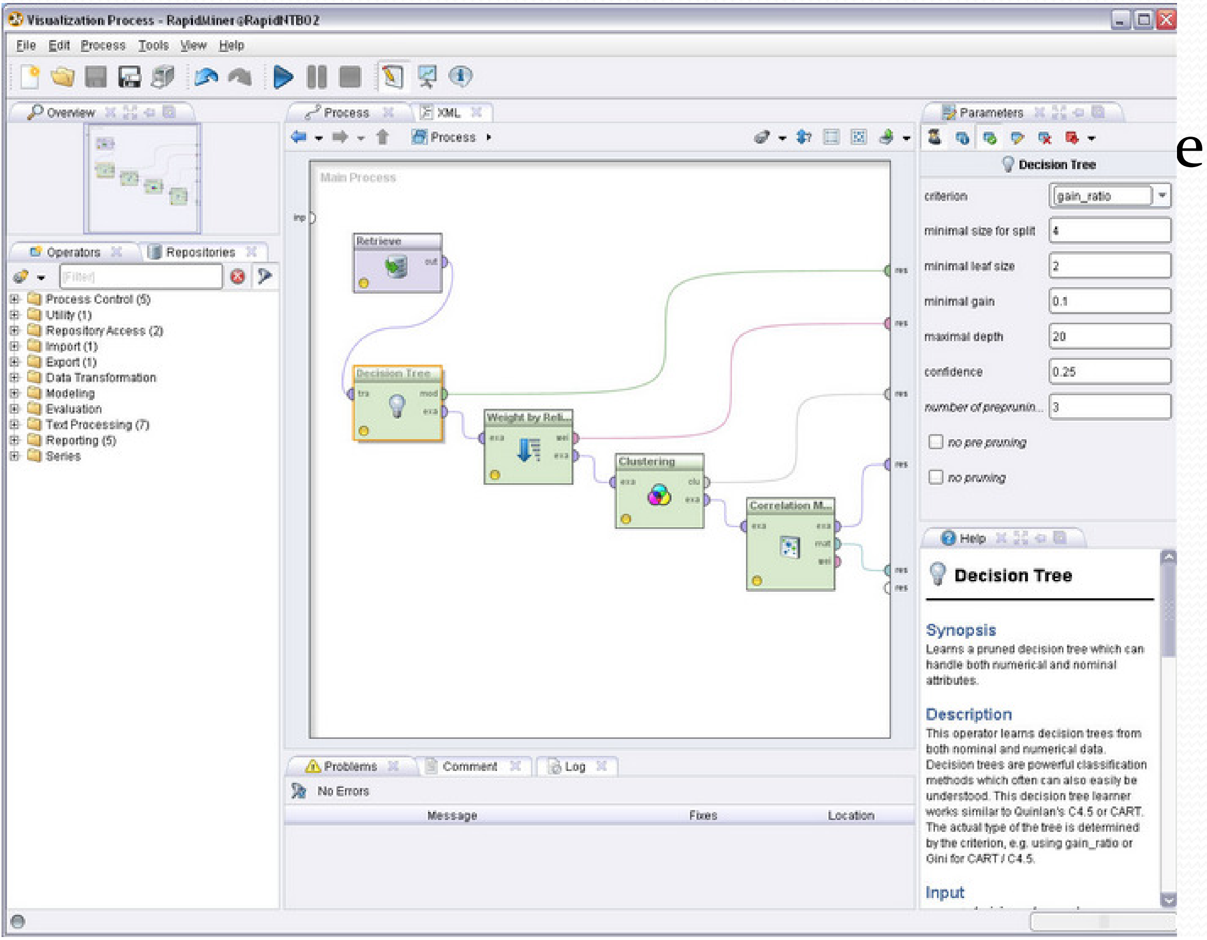
- Tanagra - http://eric.univ-lyon2
  - (+) ' he logi
  - (-) I prol
  - (+) \ com
  - (+) m
  - (-) I
  - (-) N



Data mining software’s

Open source

- Rapid Miner - http://rapid-i.com/c
  - (+) Tr e and store
  - (-) No exam
  - (+) Bi



Decision Tree

criterion: gain_ratio
minimal size for split: 4
minimal leaf size: 2
minimal gain: 0.1
maximal depth: 20
confidence: 0.25
number of preprunin...: 3
□ no pre pruning
□ no pruning

No Errors

Decision Tree

Synopsis

Learns a pruned decision tree which can handle both numerical and nominal attributes.

Description

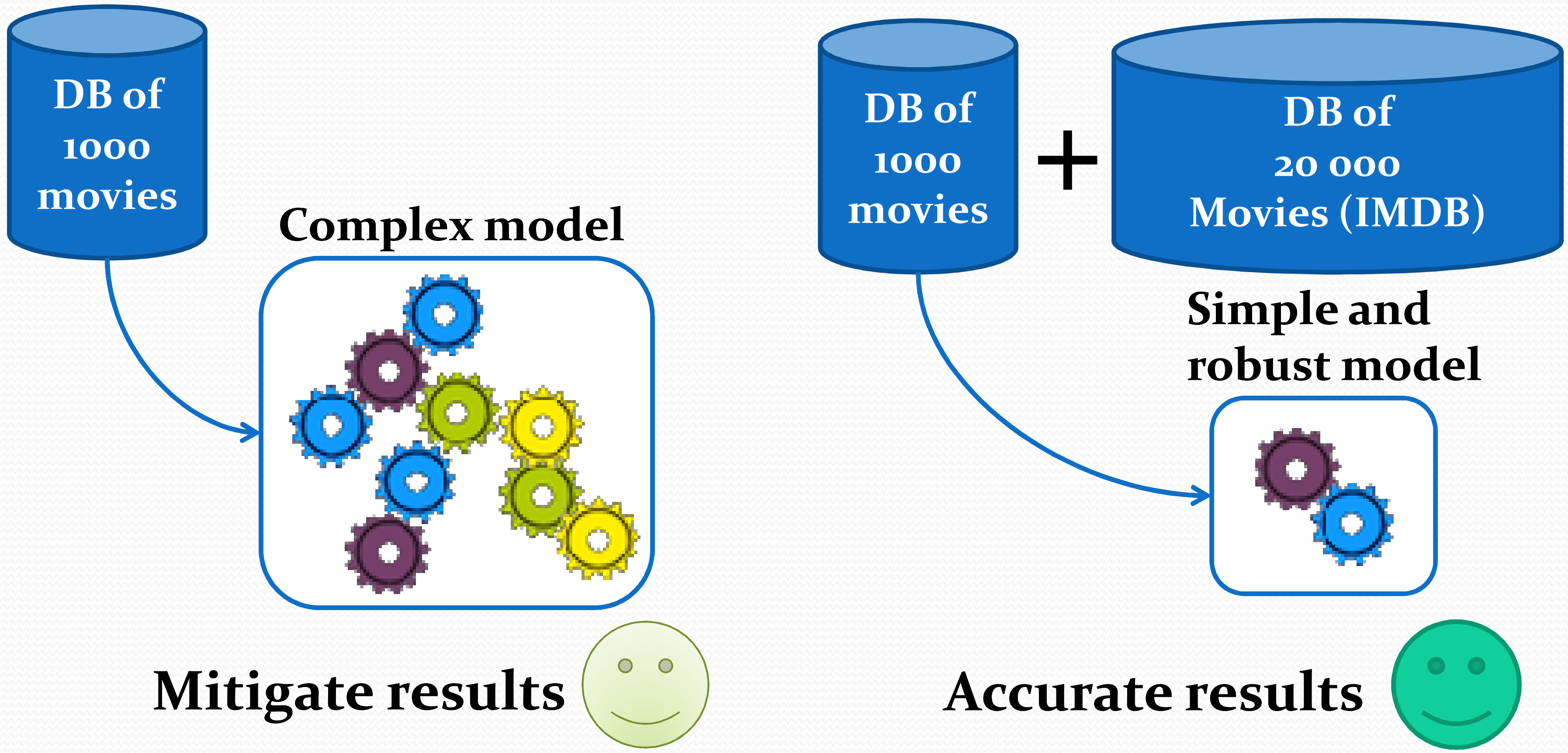
This operator learns decision trees from both nominal and numerical data. Decision trees are powerful classification methods which often can also easily be understood. This decision tree learner works similar to Quinlan's C4.5 or CART. The actual type of the tree is determined by the criterion, e.g. using gain_ratio or Gini for CART / C4.5.

Input

An interesting story …

Data mining to predict movie rating

- Researcher I
- Researcher II



DB of 1000 movies

Complex model

Mitigate results

DB of 1000 movies + DB of 20 000 Movies (IMDB)

Simple and robust model

Accurate results

Questions?