Data Mining

Motivation

Motivation

- Since the nineties, more and more data was stored because hard drivers became cheaper and with higher capacity
- Today: “90% of world's data generated over last two years”
  - http://www.sciencedaily.com/releases/2013/05/130522085217.htm (article from 22/5/2013)
- Question: how to extract useful information from all this data?

Motivation

Examples:

1. How to select customers with similar purchase profiles in order to adjust the promotion campaigns to these natural groups of customers?
   - It is necessary to know the history of purchases of these customers
2. How to suggest a new book to a client?
   - It is necessary to know the history of this and other shopping customers
3. How to detect breast cancer?
   - It is necessary to have measurable characteristics of the cyst/lump

Data Mining is …

Data Mining is ...

Sources of data are increasing everyday:

- Sensors
- Internet, e.g. social networks
- Data warehousing systems

Data Mining

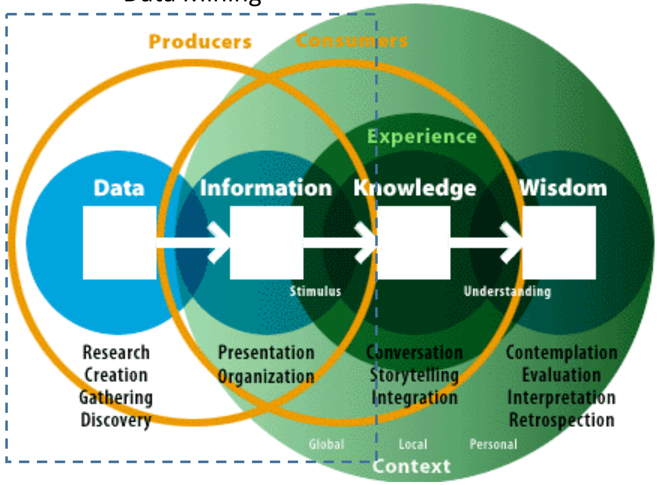


Image obtained in 31-5-2014 from
http://www.nathan.com/thoughts/course.html

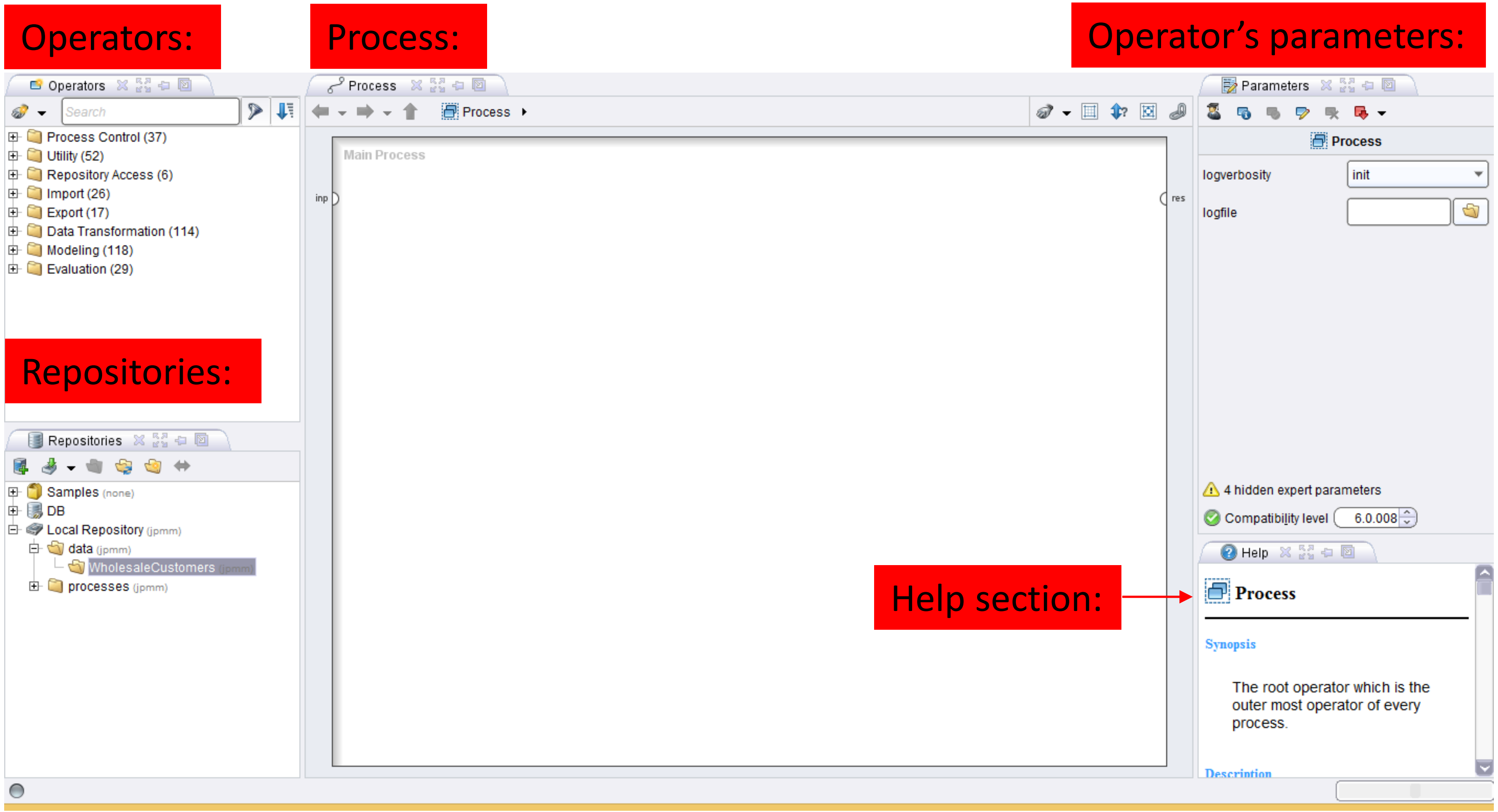
Extracting information from data:

- Without doing assumptions about data distribution
- Discovering unknown information
- Using computational resources

Rapid Miner

Two simple examples

Rapid Miner



Operators:

Process:

Operator's parameters:

Repositories:

Help section:

- Process Control (37)
- Utility (52)
- Repository Access (6)
- Import (26)
- Export (17)
- Data Transformation (114)
- Modeling (118)
- Evaluation (29)

- Samples (none)
- DB
- Local Repository (jpmm)
  - data (jpmm)
    - WholesaleCustomers (jpmm)
  - processes (jpmm)

Process

logverbosity	init
logfile	

4 hidden expert parameters

Compatibility level 6.0.008

Process

Synopsis

The root operator which is the outer most operator of every process.

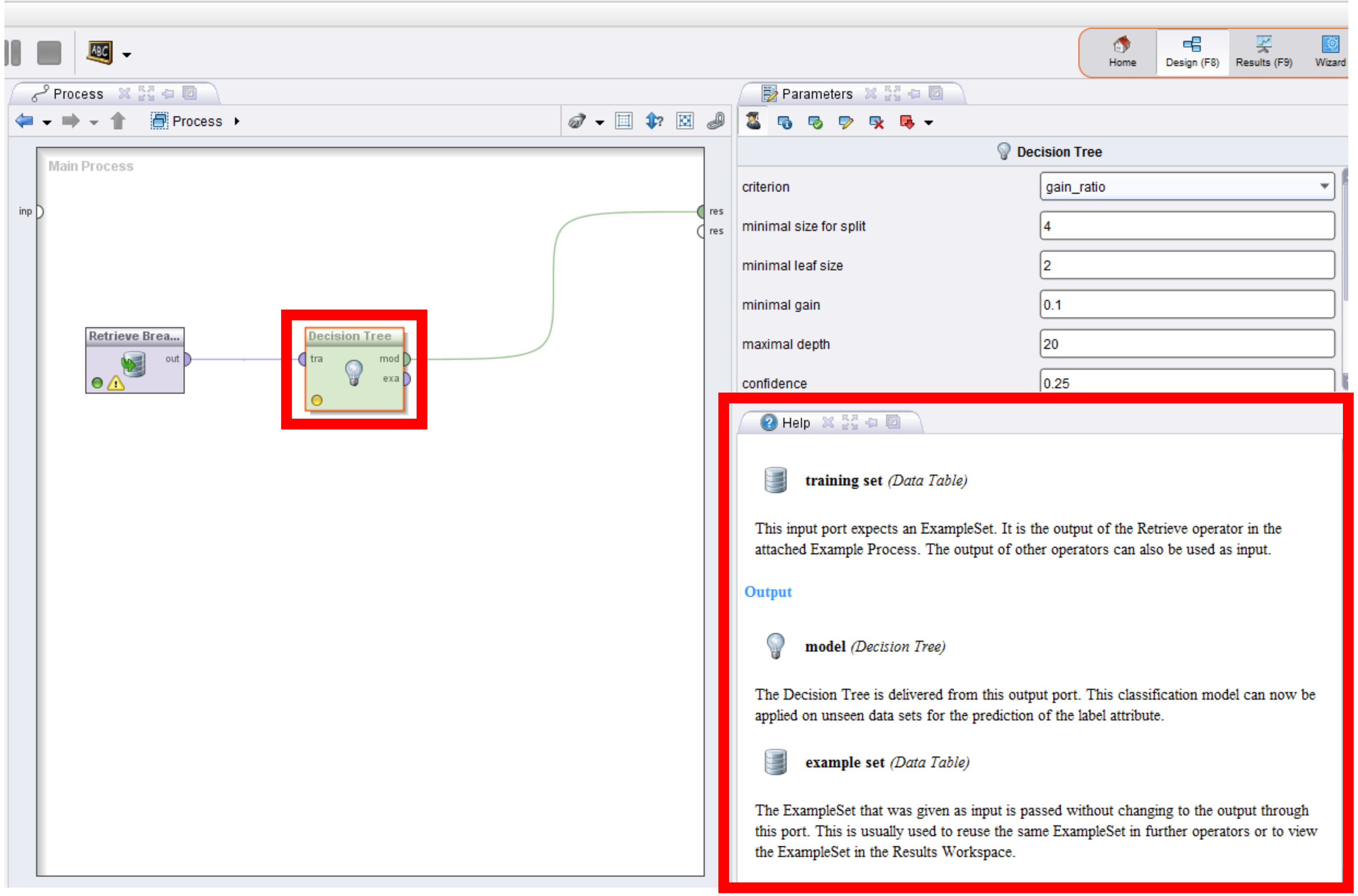
Description

Rapid Miner

The explanation of the operator’s inputs and outputs is given in the help section for the highlighted operator.

In the example:

- tra is the training set
- mod is the model
- exa is the example set



Decision Tree

criterion	gain_ratio
minimal size for split	4
minimal leaf size	2
minimal gain	0.1
maximal depth	20
confidence	0.25

training set *(Data Table)*

This input port expects an ExampleSet. It is the output of the Retrieve operator in the attached Example Process. The output of other operators can also be used as input.

Output

model *(Decision Tree)*

The Decision Tree is delivered from this output port. This classification model can now be applied on unseen data sets for the prediction of the label attribute.

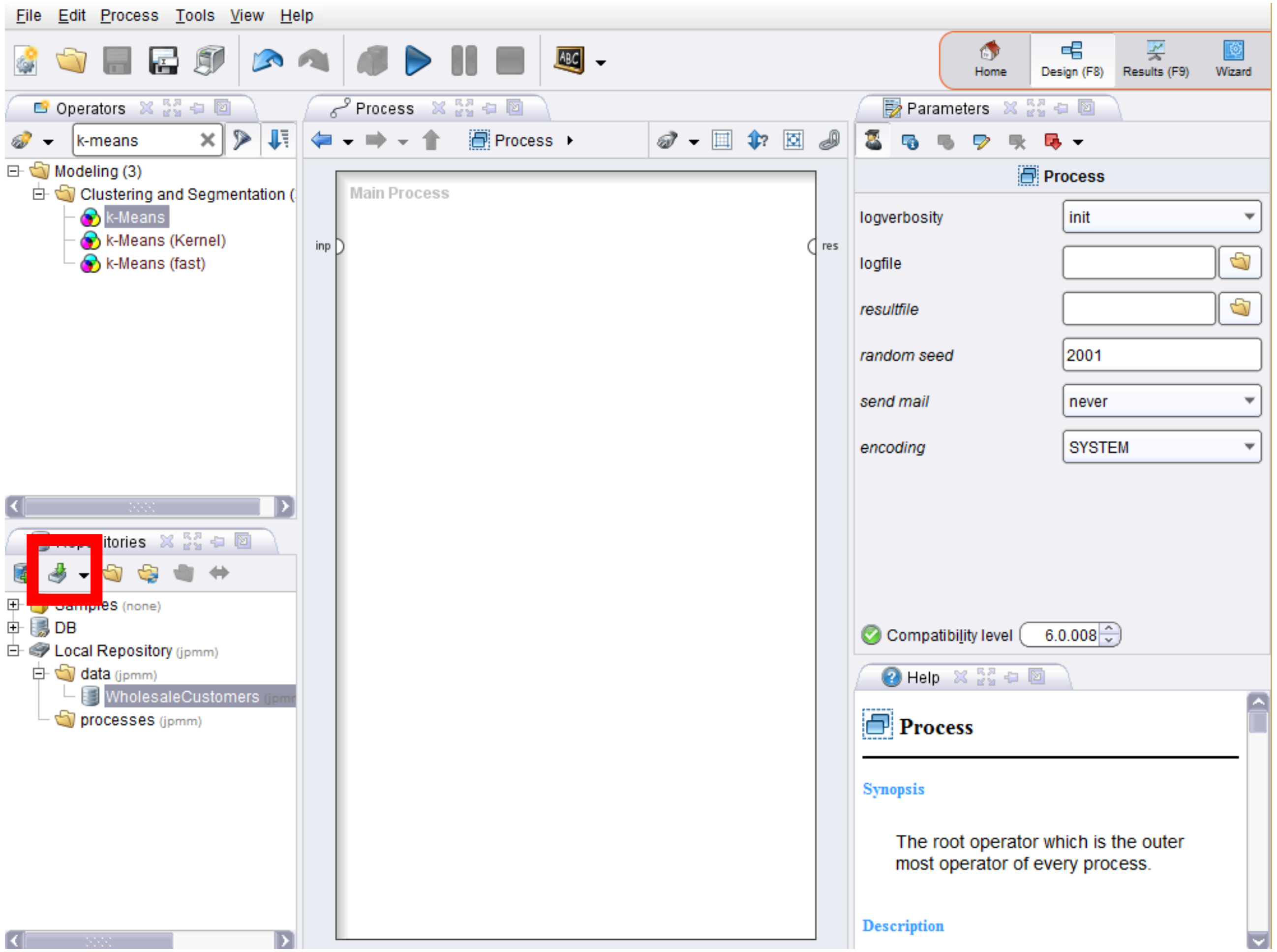
example set *(Data Table)*

The ExampleSet that was given as input is passed without changing to the output through this port. This is usually used to reuse the same ExampleSet in further operators or to view the ExampleSet in the Results Workspace.

Rapid Miner

1. How to select customers with similar purchase profiles in order to adjust the promotion campaigns to these natural groups of customers?
    - It is a clustering task
    - We are using the dataset *WholesaleCustomers*
        - Use only the quantitative variables

Rapid Miner



logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Compatibility level 6.0.008

Process

Synopsis

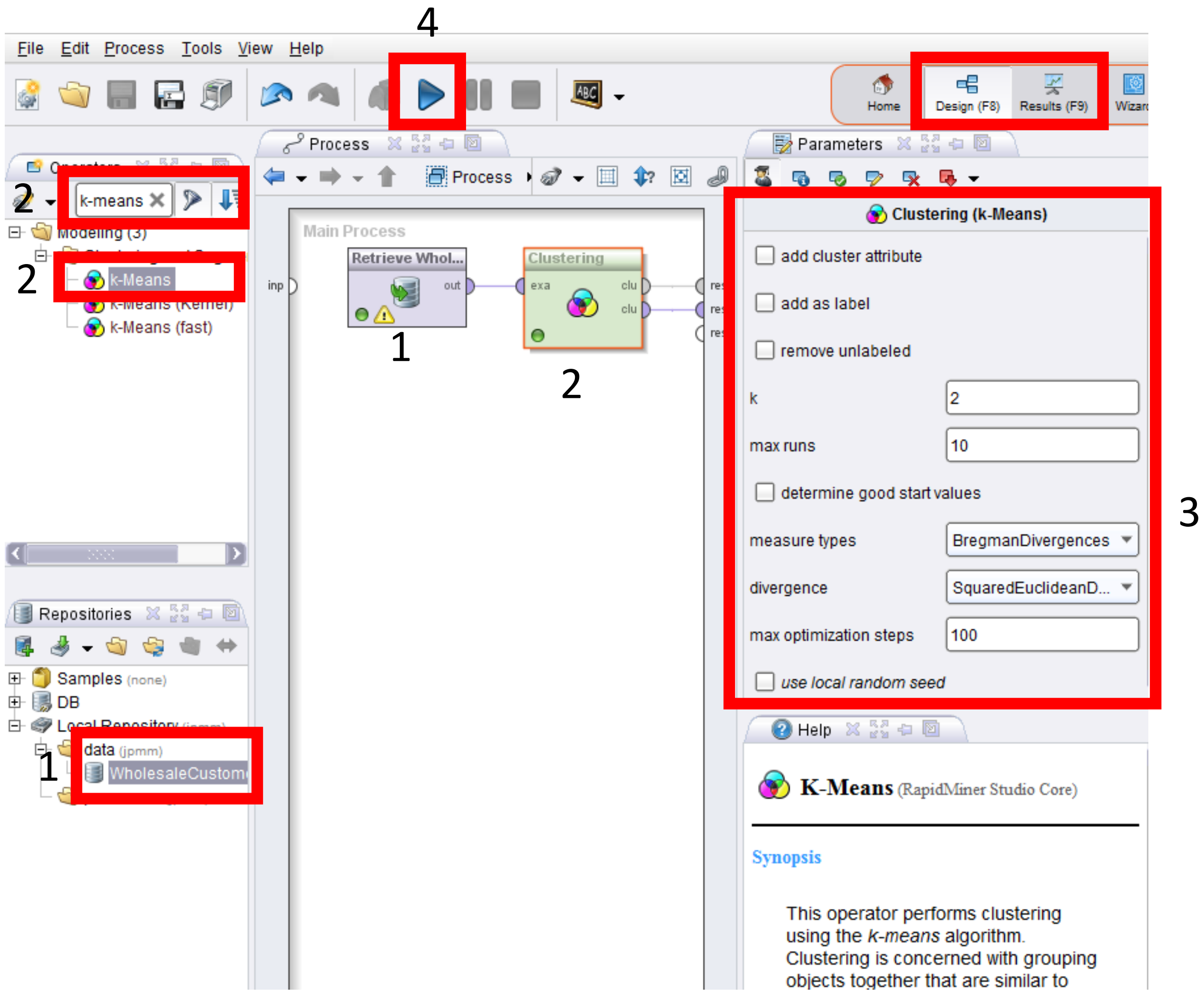
The root operator which is the outer most operator of every process.

Description

Rapid Miner

1. Loading data to the workspace
2. Insertion of the k-means operator
3. Definition of the parameters
4. Execution
5. Interpretation of the results

Rapid Miner



Clustering (k-Means)

□ add cluster attribute

□ add as label

□ remove unlabeled

k: 2

max runs: 10

□ determine good start values

measure types: BregmanDivergences

divergence: SquaredEuclideanD...

max optimization steps: 100

□ *use local random seed*

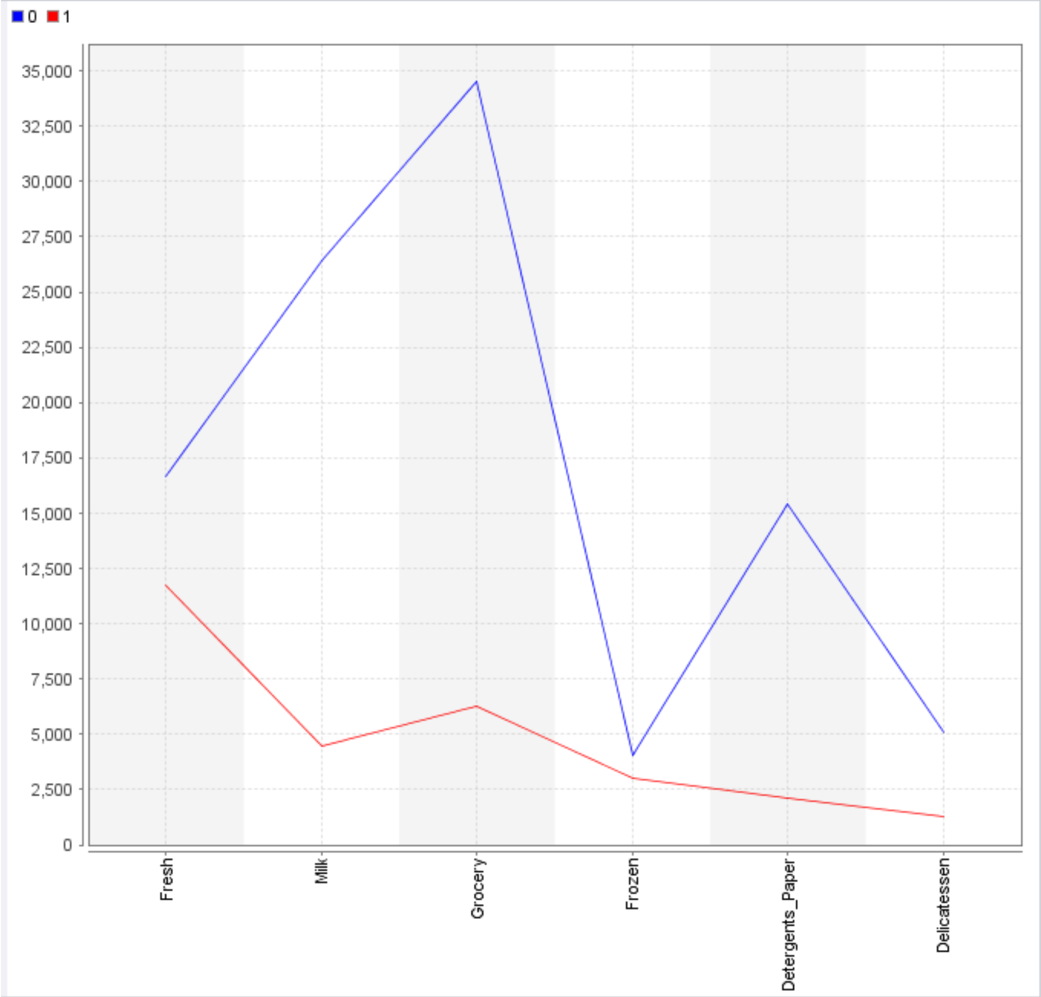
K-Means (RapidMiner Studio Core)

Synopsis

This operator performs clustering using the *k-means* algorithm. Clustering is concerned with grouping objects together that are similar to

Rapid Miner

5



Analising and interpreting the results (see the ones in the image).
Save the process.

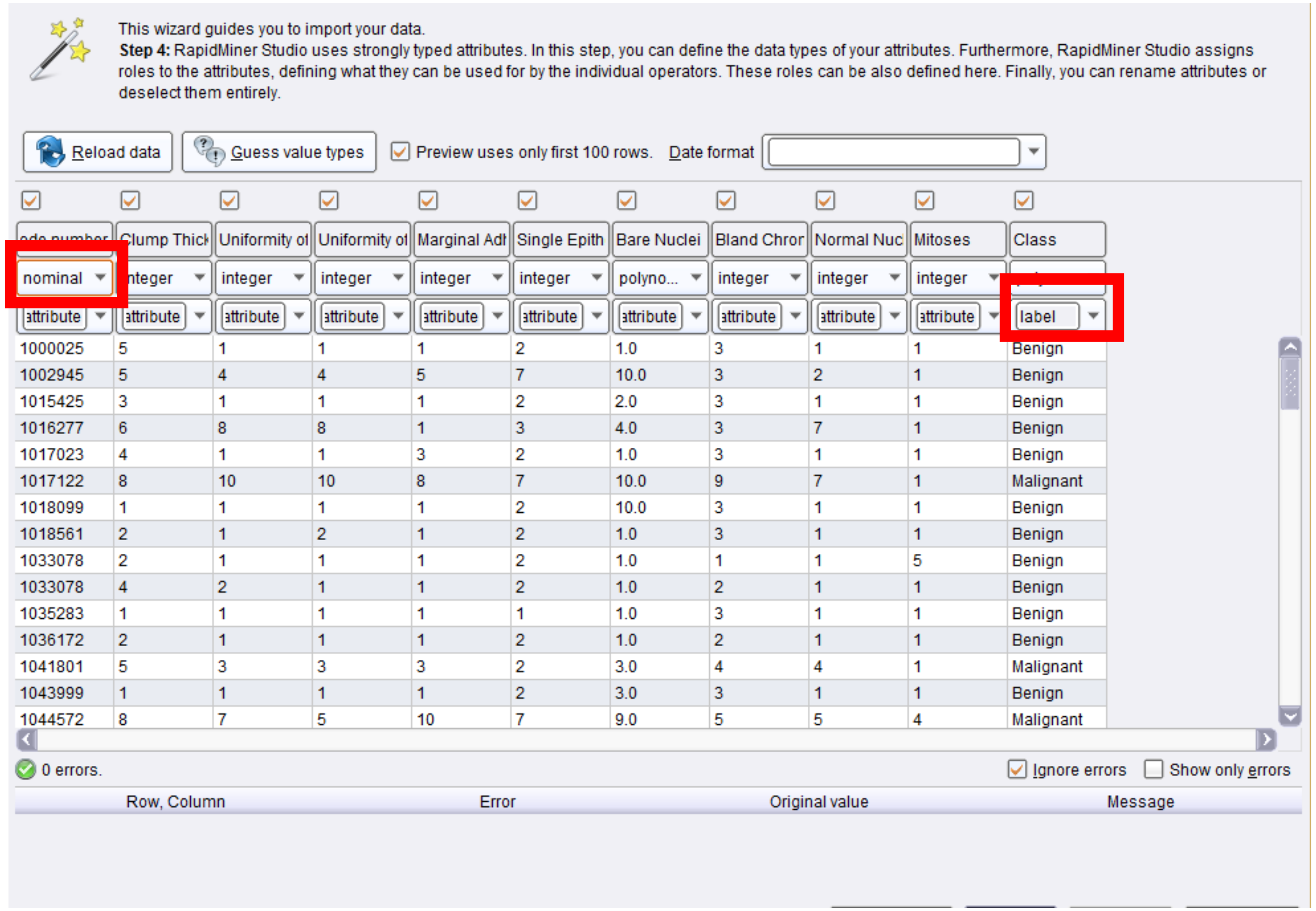
Rapid Miner

2. How to determine the benignity/malignancy of a breast cyst?
    - Is a prediction task, more precisely, a classification task
    - Let's use the dataset BreastCancerWisconsin

Rapid Miner

1. Loading data to the workspace
   - You should define the target variable
2. Insertion of the decision tree operator
3. Definition of the parameters
4. Execution
5. Interpretation of the results

Rapid Miner



This wizard guides you to import your data.
Step 4: RapidMiner Studio uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner Studio assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.

☑ Preview uses only first 100 rows. Date format

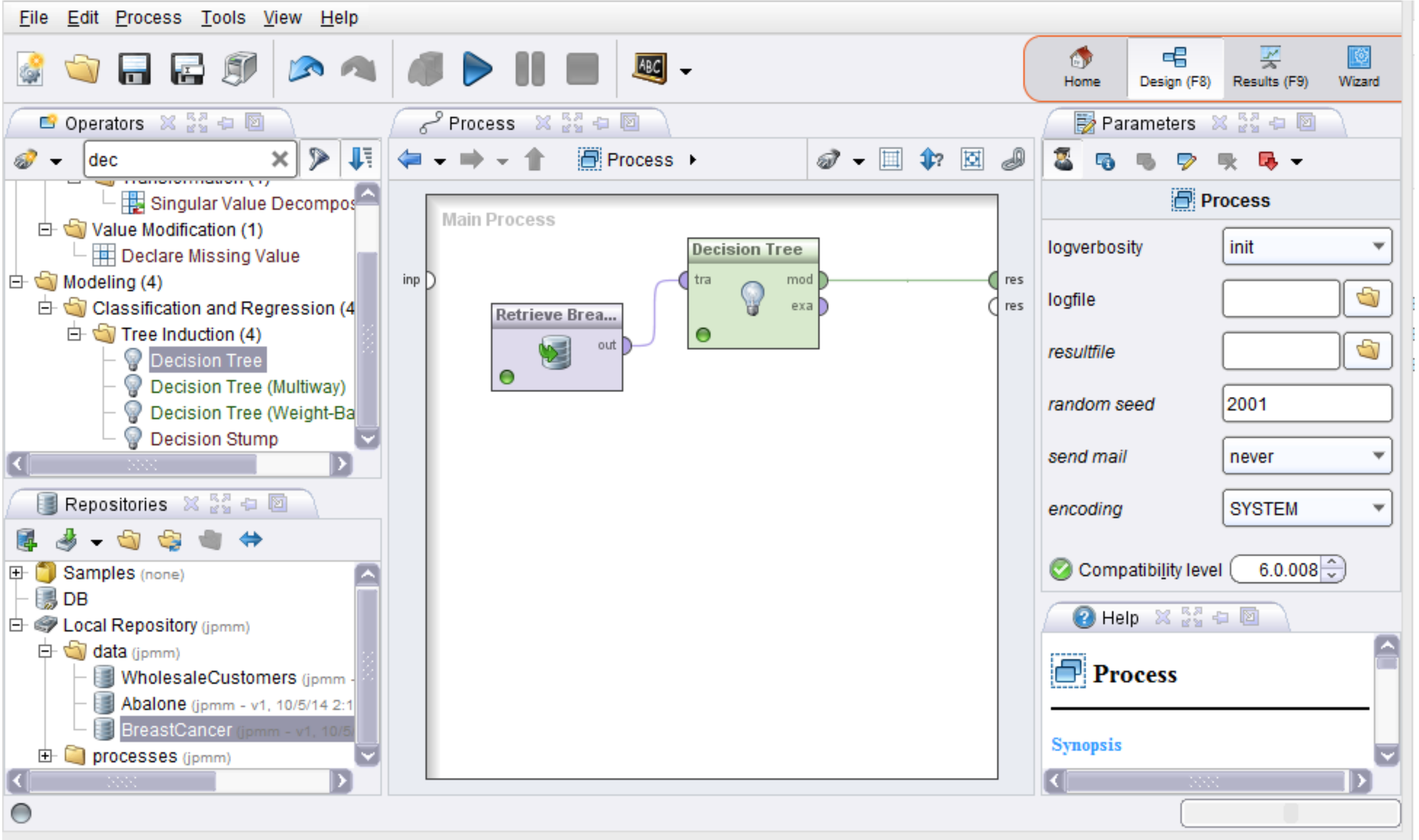
☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
ode number	Clump Thick	Uniformity of	Uniformity of	Marginal Adh	Single Epith	Bare Nuclei	Bland Chror	Normal Nuc	Mitoses	Class
nominal	nteger	integer	integer	integer	integer	polyno...	integer	integer	integer	[illegible]
attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	label
1000025	5	1	1	1	2	1.0	3	1	1	Benign
1002945	5	4	4	5	7	10.0	3	2	1	Benign
1015425	3	1	1	1	2	2.0	3	1	1	Benign
1016277	6	8	8	1	3	4.0	3	7	1	Benign
1017023	4	1	1	3	2	1.0	3	1	1	Benign
1017122	8	10	10	8	7	10.0	9	7	1	Malignant
1018099	1	1	1	1	2	10.0	3	1	1	Benign
1018561	2	1	2	1	2	1.0	3	1	1	Benign
1033078	2	1	1	1	2	1.0	1	1	5	Benign
1033078	4	2	1	1	2	1.0	2	1	1	Benign
1035283	1	1	1	1	1	1.0	3	1	1	Benign
1036172	2	1	1	1	2	1.0	2	1	1	Benign
1041801	5	3	3	3	2	3.0	4	4	1	Malignant
1043999	1	1	1	1	2	3.0	3	1	1	Benign
1044572	8	7	5	10	7	9.0	5	5	4	Malignant

0 errors.

☑ Ignore errors □ Show only errors

Row, Column	Error	Original value	Message

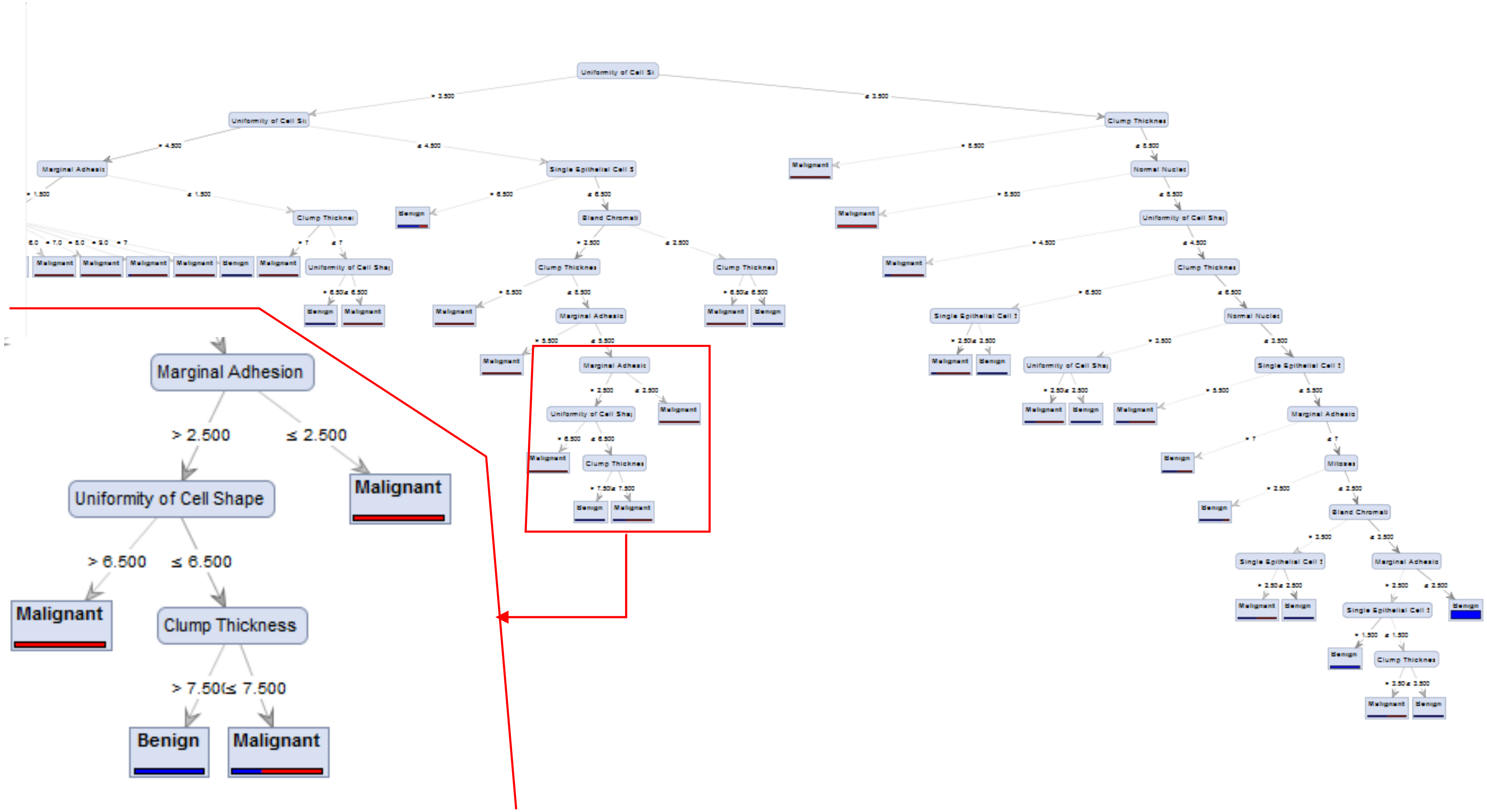
Rapid Miner



logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

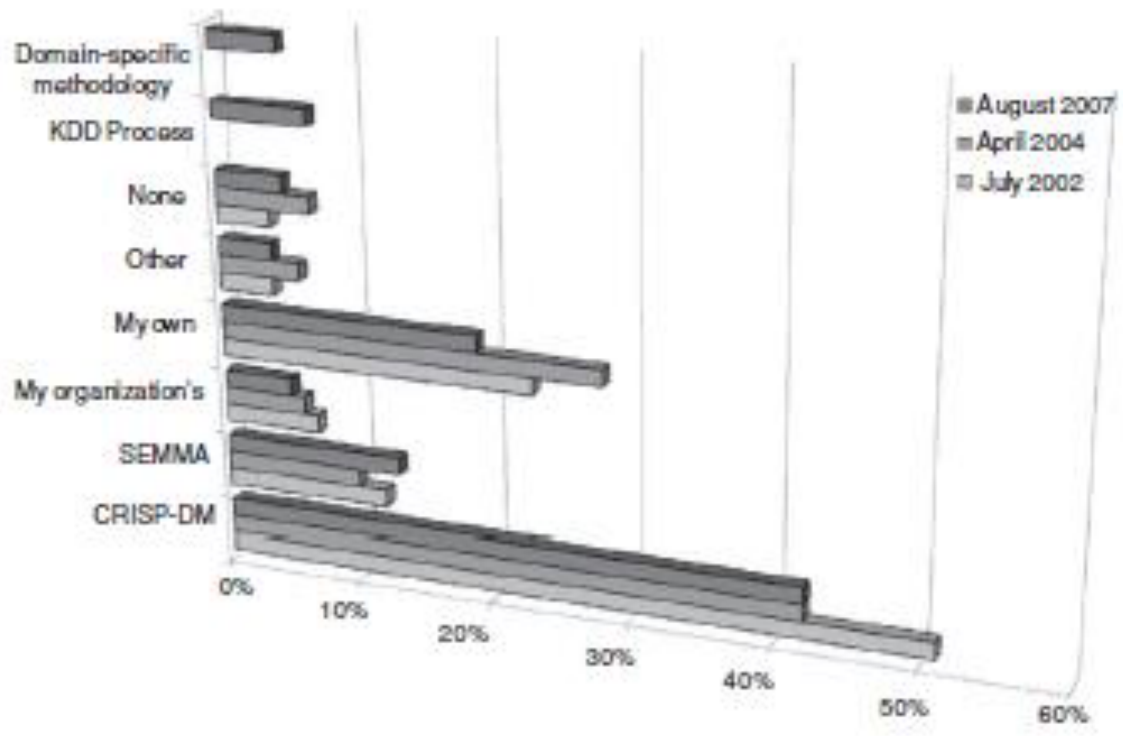
Compatibility level 6.0.008

Rapid Miner


Marginal Adhesion
> 2.500
≤ 2.500
Malignant
Uniformity of Cell Shape
> 6.500
≤ 6.500
Malignant
Clump Thickness
> 7.50(≤ 7.500
Benign
Malignant

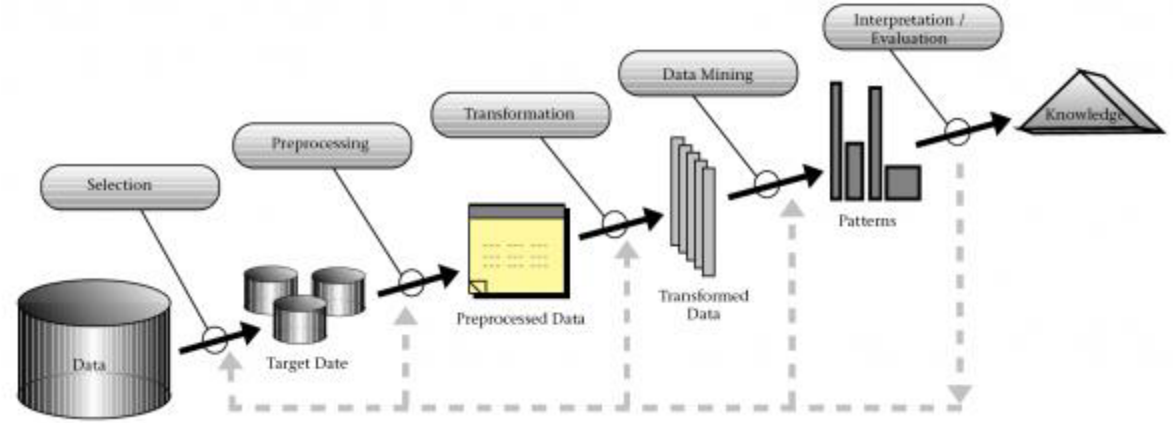
Process model

Process models



The KDD process model

1. Selection
2. Preprocessing
3. Transformation
4. *Data Mining*
5. Interpretation / Evaluation



Fayyad, U. M. et al. 1996. From data mining to knowledge discovery: an overview. In Fayyad, U. M. et al (Eds.), Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

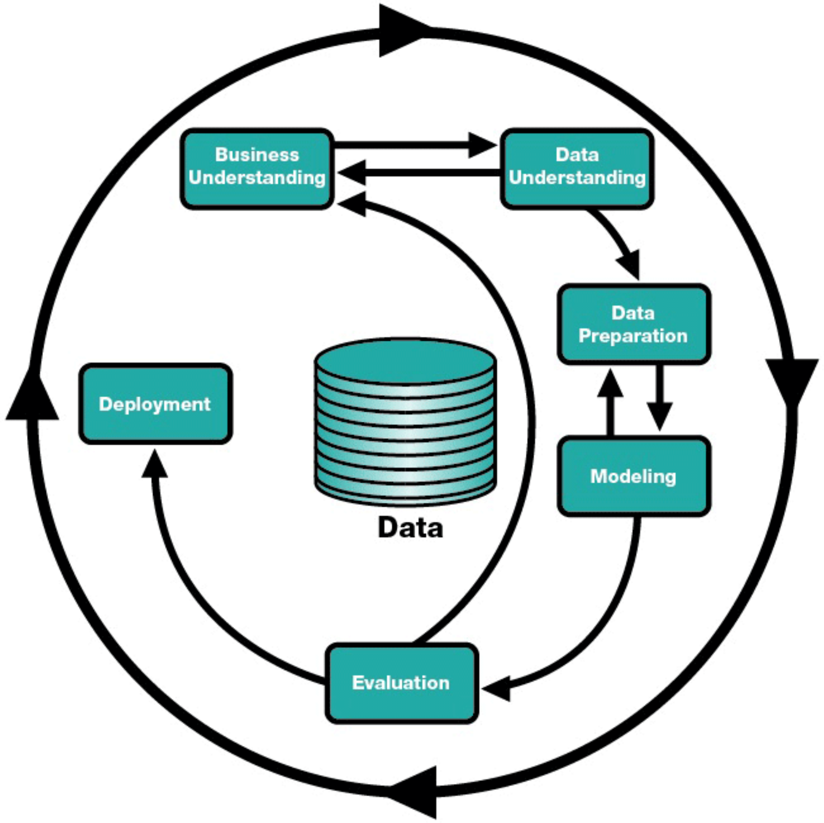
The KDD process model

- **Selection**: identification and selection of all external and internal sources of information and selection of the subset of data or variables needed for the KDD process.
- **Preprocessing**: includes the removal of data with extreme values (outliers), filling in missing values, etc.
- **Transformation**: converting data into a format suitable for Data Mining algorithms.
- **Data Mining**: in this step the specialized tools seek, through specialized algorithms, existing patterns in the data. This search can be performed automatically or interactively systems through the aid of the analyst responsible for the generation of the hypotheses. At the end of the process, the DM system should generate a report of the analysis carried out in order to enable analysts to verify the results obtained.
- **Interpretation/evaluation**: this step should be performed in conjunction with business analysts. If the knowledge generated is not satisfactory, analysts can form a new set of experiments giving rise to a new iteration of the process.

The CRISP-DM process model

CRISP-DM (CRoss Industry Standard Process for Data Mining)

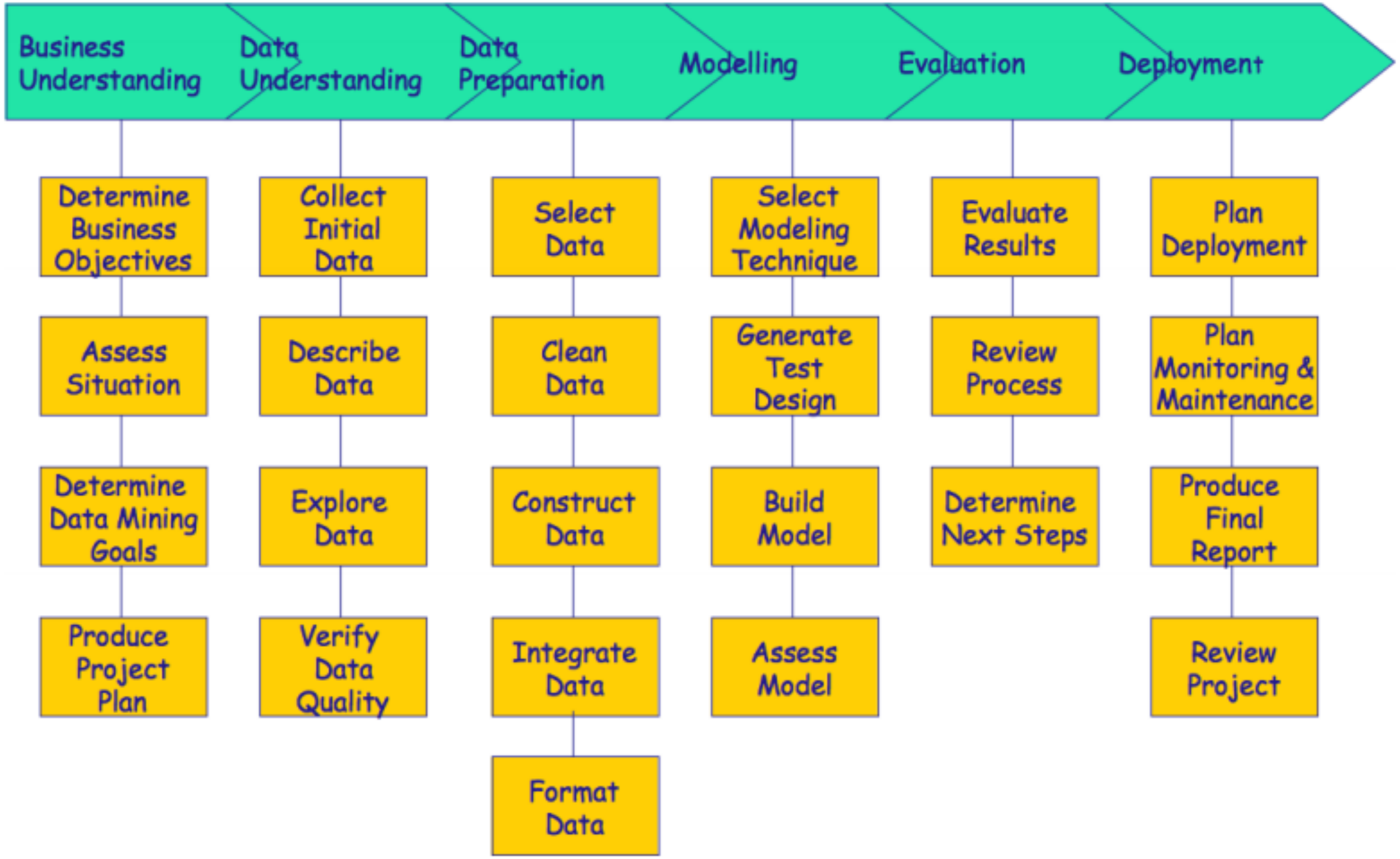
1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment



The CRISP-DM process model

1. **Understand the Business**: focuses on understanding the objective of the project from a business perspective, defining a preliminary plan to achieve the goals.
2. **Understand the data**: data gathering and early activities to better understand the data, identifying problems or interesting sets.
3. **Data preparation**: construction of the final data set from the initial one. Normally occurs several times in the process.
4. **Modeling**: several modeling techniques are applied, and its parameters calibrated for optimization. Thus, it is common to return to Data Preparation during this phase.
5. **Evaluation**: a model that seems to have great quality in a data analysis perspective was constructed. However, it is necessary to verify if the model reaches the goals of the business.
6. **Deployment**: the knowledge acquired by the model is organized and presented in a way that the customer can use.

The CRISP-DM process model

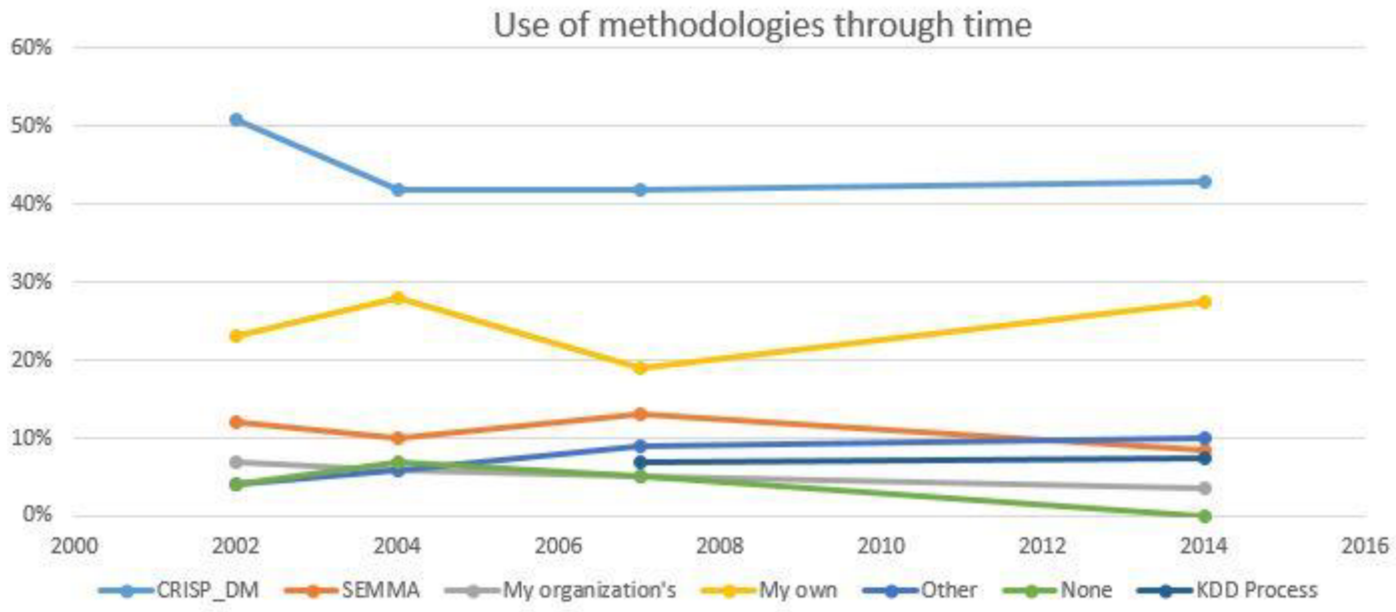


Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Determine Business Objectives	Collect Initial Data	Select Data	Select Modeling Technique	Evaluate Results	Plan Deployment
Assess Situation	Describe Data	Clean Data	Generate Test Design	Review Process	Plan Monitoring & Maintenance
Determine Data Mining Goals	Explore Data	Construct Data	Build Model	Determine Next Steps	Produce Final Report
Produce Project Plan	Verify Data Quality	Integrate Data	Assess Model		Review Project
		Format Data			

Process models

KDD	CRISP-DM
Pre-process KDD	Business understanding
Selection	Data understanding
Preprocessing	
Transformation	Data preparation
Data Mining	Modeling
Interpretation/Evaluation	Evaluation
Post-process KDD	Deployment

Process models


Use of methodologies through time
CRISP_DM
SEMMA
My organization's
My own
Other
None
KDD Process