

# Introduction to RapidMiner



# RapidMiner

- A very comprehensive open-source data mining tool
  - The data mining process is visually modeled as an operator chain
  - RapidMiner has over 400 built-in data mining operators
  - RapidMiner provides broad collection of charts for visualizing data
- Project started in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at University of Dortmund, Germany
- Today: Maintained by commercial company plus open-source developers
- RapidMiner Editions
  - Community Edition: Free
  - Educational Edition: Free for students and instructors
  - Enterprise Edition: Commercial



# Gartner: Data Science Platforms

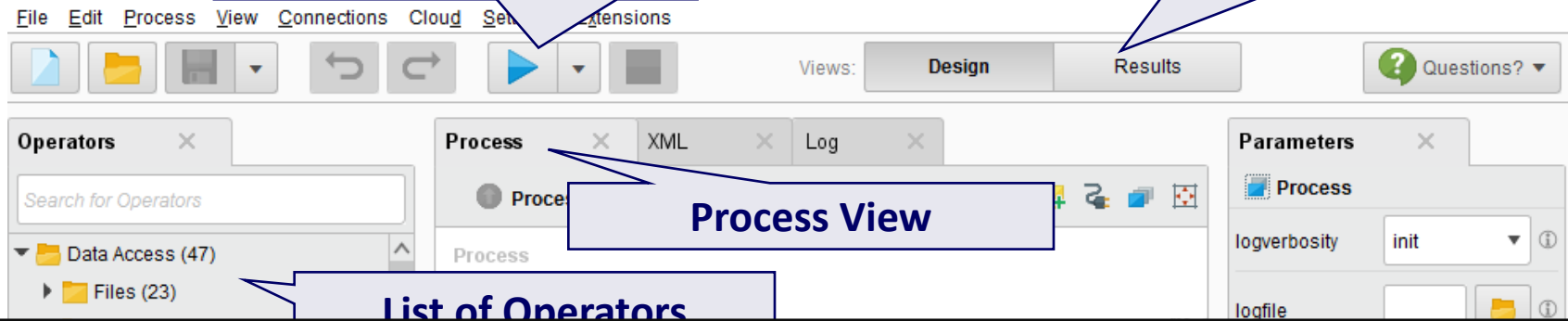
Figure 1. Magic Quadrant for Data Science and Machine Learning Platforms



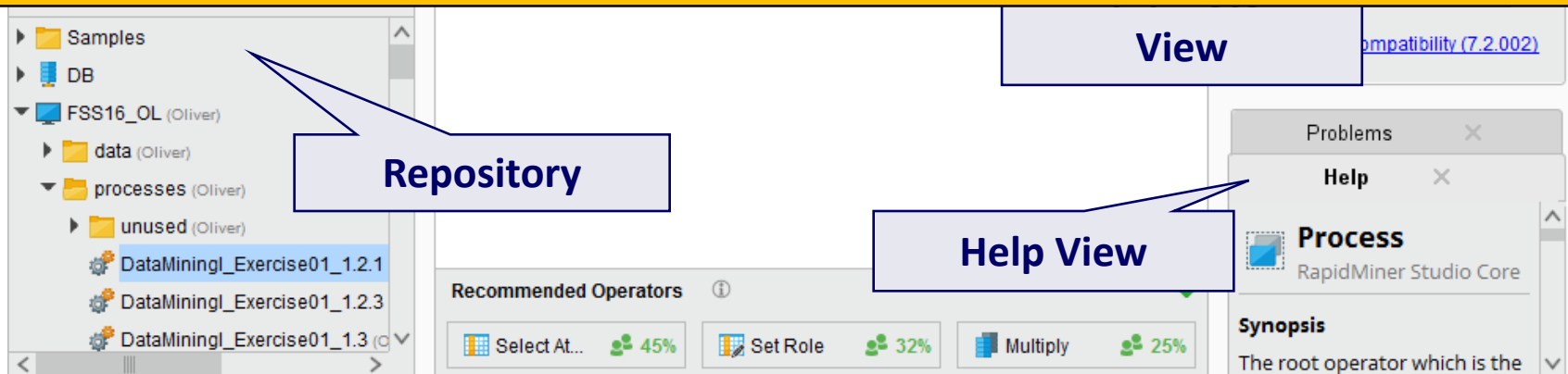
# Let's have a look at RapidMiner

Execute Process

Change Perspective

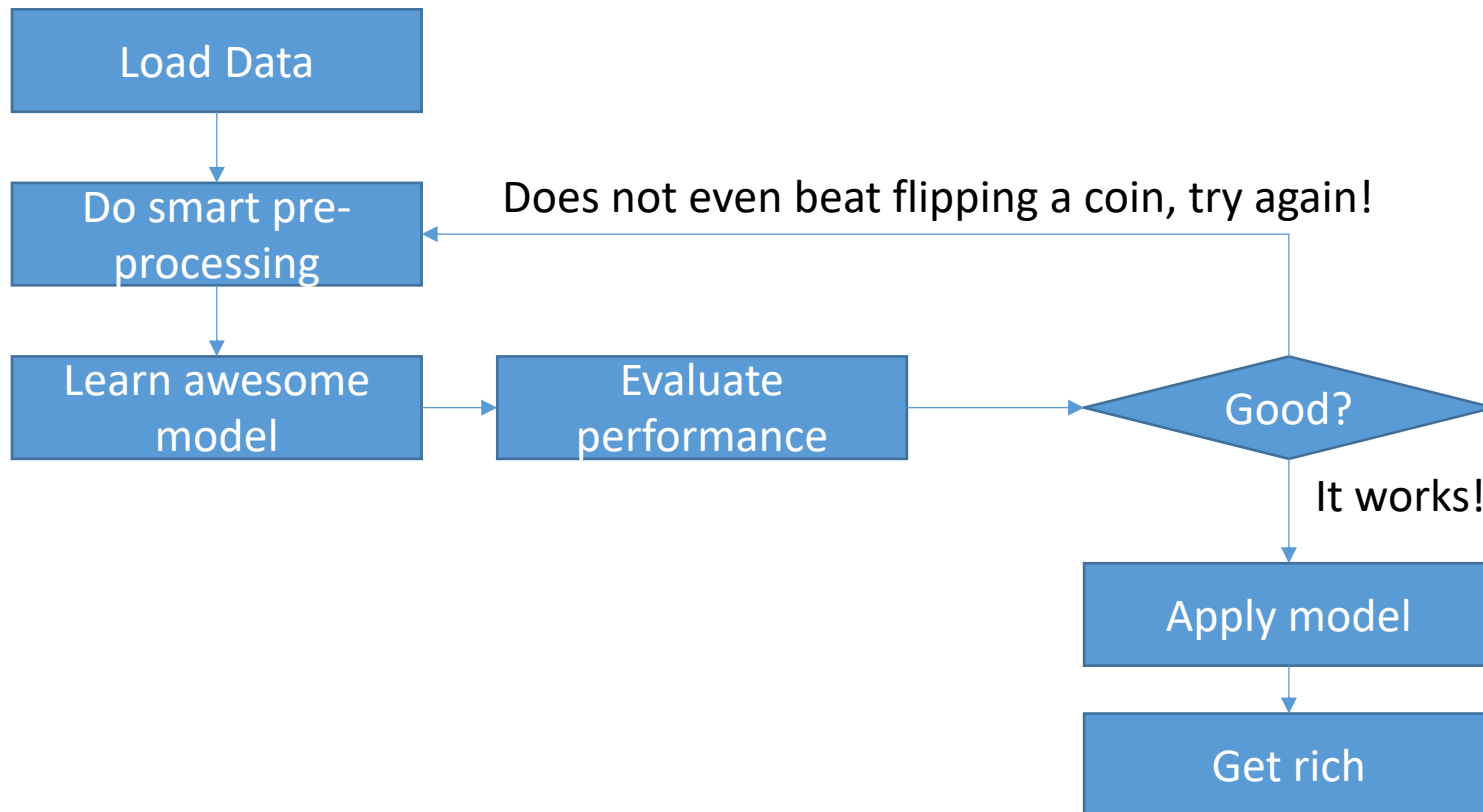


But let's take it step by step ...

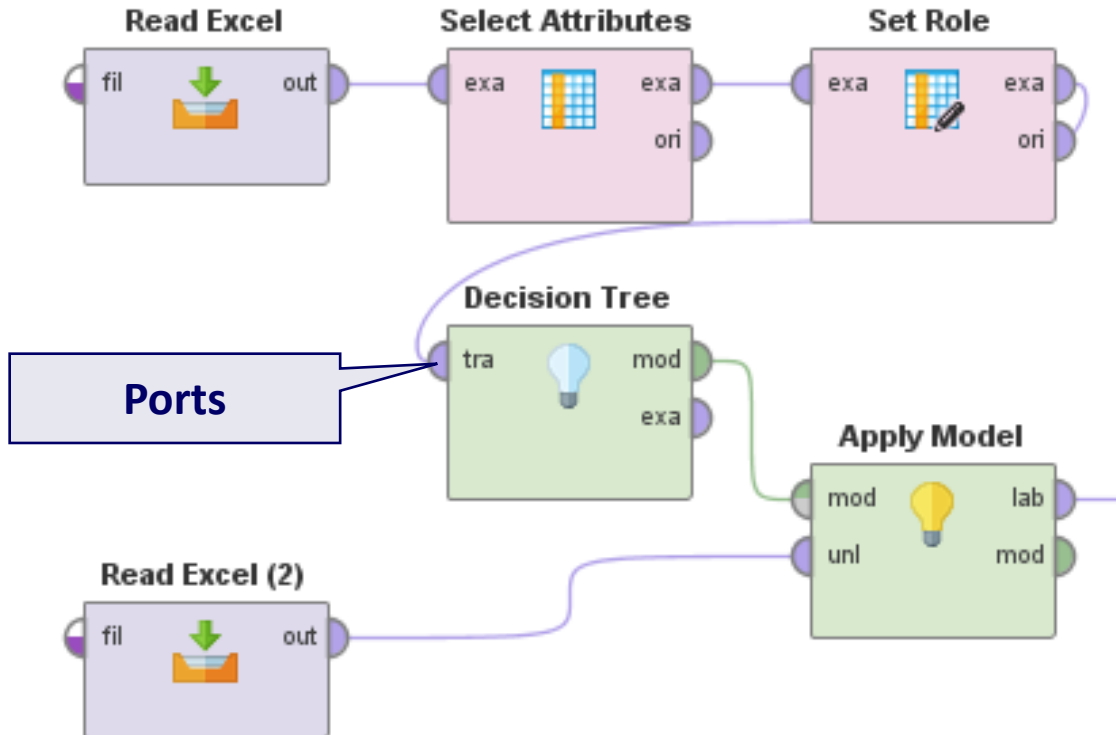


# How does it work?

- You visually design a data mining process
- A process is like a flow chart for mining operators



# Specifying a Process by Chaining Operators



Common Port Names


Name	Meaning
out	Output
exa	Example Set
ori	Original Input
tra	Training Data
mod	Model
unl	Unlabelled Data
lab	Labelled Data
per	Performance

# RapidMiner Operators: Loading Data


- Many operators to read data from files
- Output Port labelled “out”
  - Creates an **Example Set**
- An Example Set contains your data!
  - The records are called **Examples**



**Parameters** ✕

 **Read CSV**

[Import Configuration Wizard...](#) ⓘ

csv file   ⓘ

column separators  ⓘ

☐ trim lines ⓘ

☒ use quotes ⓘ


quotes character  ⓘ


escape character  ⓘ

☐ skip comments ⓘ

☒ parse numbers ⓘ

decimal character  ⓘ

 [Hide advanced parameters](#)

 [Change compatibility \(7.2.002\)](#)

# Data in RapidMiner

- All data that you load will be contained in an example set
- Each example is described by **Attributes** (a.k.a. features)
  - Attributes have **Value Types**
  - Attributes have **Roles**

Customer ID	ItemsBought	ItemsReturned	ZipCode	Product
polynomin... ▼	integer ▼	integer ▼	polynomial ▼	polynomial ▼
id ▼	attribute ▼	attribute ▼	attribute ▼	attribute ▼
4	45	10	2	1365
5	42	18	5	2764
6	50	0	1	1343
8	13	12	4	2435
9	10	7	3	2435
10	34	17	6	2896
11	40	20	8	2869
12	40	8	2	1236
14	9	9	8	2435
15	36	7	2	1764
16	42	1	1	1547

Attribute Names

Value Types

Roles

# Data in RapidMiner

- Value types define how data is treated
  - Numeric data has an order (2 is closer to 1 than to 5)
  - Nominal data has no order (red is as different from green as from blue)

Value Type	Description
binominal	Only two different values are permitted
polynominal	More than two different values are permitted
numeric	For numerical values in general
integer	Whole numbers, positive and negative
real	Real numbers, positive and negative
date_time	Date as well as time
date	Only date
time	Only time
text	Random free text without structure

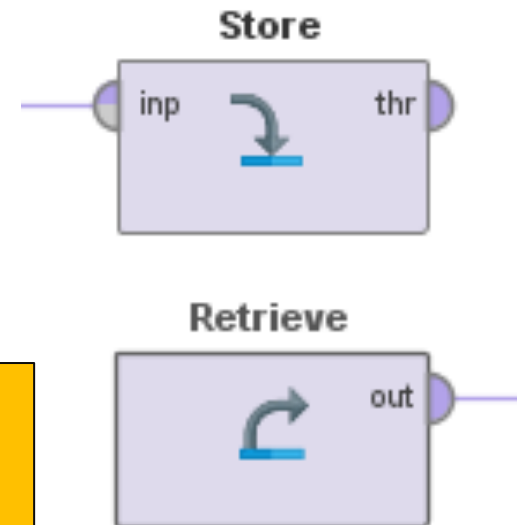
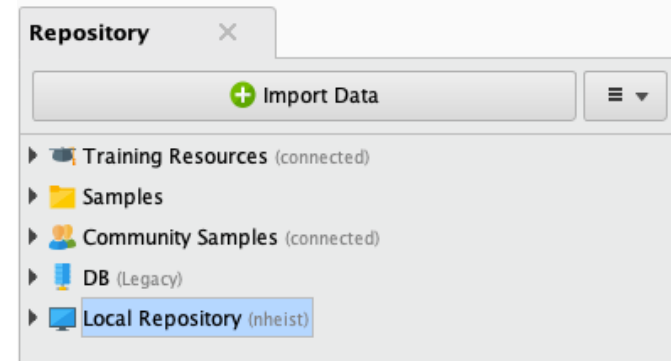
# Data in RapidMiner

- Roles define how the attribute is treated by the Operators

Role	Description
Id	A unique identifier, no two examples in an example set can have the same value
Attribute	Regular attribute that contains data
Label	The target attribute for classification tasks
Cluster	Created by RapidMiner as the result of a clustering task
Prediction	Created by RapidMiner as the result of a classification task

# The Repository

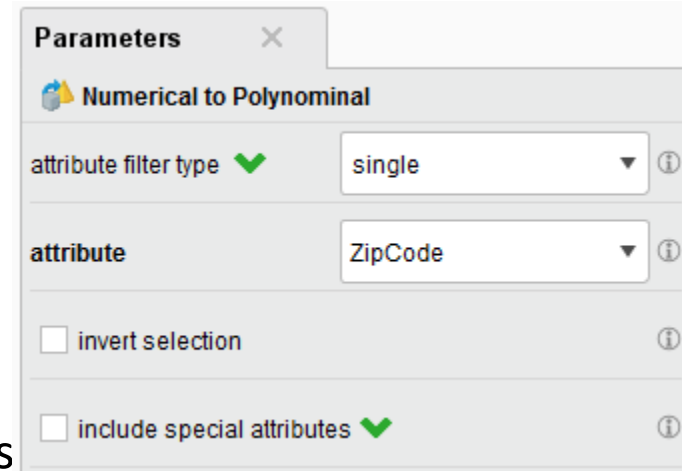
- This is where you store your data and processes
- Stores data and its meta data (!)
  - Only if you load data from the repository, RapidMiner can show you which attributes exist
- Add data via the “Import Data” button or the “Store” operator
- Load data via drag ‘n’ drop or the “Retrieve” operator



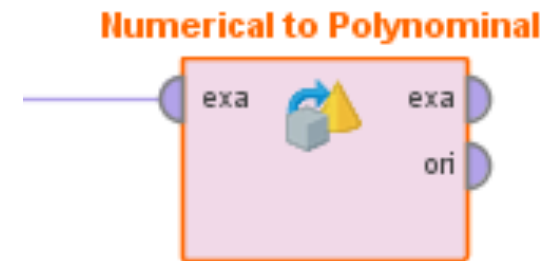
If you have a question starting with  
**“Why does RapidMiner not show me ...?”**  
Then the answer most likely is  
**“Because you did not load your data into the Repository!”**

# RapidMiner Operators: Pre-Processing

- Type and Role Conversions
  - “TypeA to TypeB”: Change the type
  - “Set Role”: Change the role
- Attribute Set Transformation
  - “Select Attributes”: Remove attributes
  - “Generate Attributes: Create new attributes
- Value Transformation
  - “Normalize”: transform all values to a certain range
- Filtering
  - “Filter examples”: Remove examples
- Aggregation
  - “Aggregate”: SQL-like aggregation (count, sum)

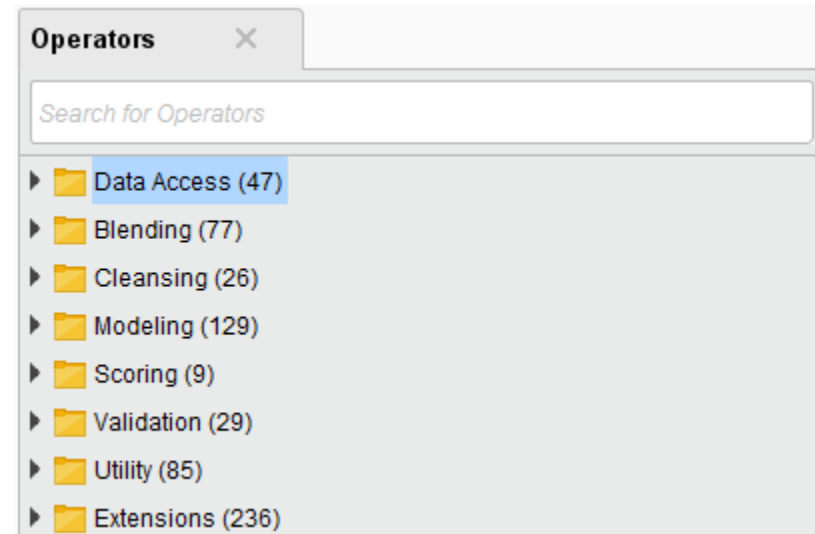


The screenshot shows the 'Parameters' window for the 'Numerical to Polynomial' operator. It includes a close button (X) and an information icon (i). The 'attribute filter type' is set to 'single' with a green checkmark and an information icon. The 'attribute' is set to 'ZipCode' with a dropdown arrow and an information icon. There are two checkboxes: 'invert selection' (unchecked) and 'include special attributes' (checked with a green checkmark), each with an information icon.



# How to find Operators

- The Operators Panel lets you browse all available operators
- You can search for operators by typing in the search bar
- You add operators by double clicking or by dragging them onto the process view



Frequently Asked Questions – And their surprising answers ...

How can I ...?	Type ... into the search bar!
Select which Attributes to use?	Select Attributes
Filter out examples?	Filter Examples
Read a CSV file	Read CSV
Learn a decision tree	Decision Tree

# How to use RapidMiner

- Use the “Design Perspective” to create your Process
  - See your current Process – “Process”
  - Access your data and processes – “Repository”
  - Add operators to the process – “Operators”
  - Configure the operators – “Parameters”
  - Learn about operators – “Help”
- Use the “Results Perspective” to inspect the output
  - The “Data View” shows your example set
  - The “Statistics View” contains meta data and statistics
  - The “Charts View” allows you to visualise the data

# The Design View

Execute Process

Change View

The screenshot displays the RapidMiner Design View interface, which is used for building and executing data mining workflows. The interface is divided into several panes and sections, each with a specific function:

- Process View:** The central workspace where the workflow is designed. It shows a sequence of operators connected by lines. In this example, a "Retrieve DataMining..." operator is connected to an "exa" (Example Set) operator, which is then connected to an "ori" (Output) operator. The "Process" tab is active, showing the workflow diagram.
- List of Operators:** A pane on the left side that lists all available operators. It is organized into categories: Data Access (47), Files (23), Database (4), Applications (9), and Cloud Storage (5). The "Retrieve" operator is highlighted under the "Data Access" category.
- Operators:** A pane on the left side that lists all available operators. It is organized into categories: Data Access (47), Files (23), Database (4), Applications (9), and Cloud Storage (5). The "Retrieve" operator is highlighted under the "Data Access" category.
- Parameter View:** A pane on the right side that displays the parameters for the selected operator. In this case, it shows the parameters for the "Process" operator, including "logverbosity" (set to "init"), "logfile", "resultfile", "random seed" (set to "2001"), "send mail" (set to "never"), and "encoding" (set to "SYSTEM").
- Repository:** A pane on the left side that displays the data sources available in the repository. It shows a tree structure with folders like "Samples", "DB", and "FSS16\_OL (Oliver)". Under "FSS16\_OL (Oliver)", there are folders for "data (Oliver)" and "processes (Oliver)". The "DataMiningI\_Exercise01\_1.2.1" operator is highlighted under the "processes (Oliver)" folder.
- Help View:** A pane on the bottom right that displays the help documentation for the selected operator. It shows the "Process" operator's synopsis, stating that it is the root operator which is the
- Recommended Operators:** A section at the bottom of the interface that suggests operators based on the current workflow. It lists "Select At..." (45%), "Set Role" (32%), and "Multiply" (25%).

# The Results View - Data



Data



Statistics



Charts



Advanced  
Charts



Annotations

ExampleSet (41 examples, 0 special attributes, 5 regular attributes)

Row No.	Semester	Name	Course	Mark	Attended
1	FSS2010	Alex Krausche	Database Sy...	1.300	13
2	FSS2010	Tanja Becker	Database Sy...	2	12
3	FSS2010	Mariano Selina	Database Sy...	1.700	5
4	FSS2010	Otto Blacher	Database Sy...	2.300	13
5	FSS2010	Frank Fester	Database Sy...	2	13
6	FSS2010	Susanne Müll...	Database Sy...	3	12
7	FSS2010	Avid Morvita	Database Sy...	4	13
8	FSS2010	Steve Queck	Database Sy...	2.700	8
9	FSS2010	Michaela Mart...	Database Sy...	5	5
10	FSS2010	Ulrich Gester	Database Sy...	5	7
11	HWS2010	Alex Krausche	Database Sy...	1	12
12	HWS2010	Tanja Becker	Database Sy...	1.700	13
13	HWS2010	Mariano Selina	Database Sy...	2	10
14	HWS2010	Otto Blacher	Database Sy...	2.300	10
15	HWS2010	Frank Fester	Database Sy...	2	9
16	HWS2010	Michaela Mart...	Database Sy...	3.700	8
17	HWS2010	Ulrich Gester	Database Sy...	5	9

# The Results View - Statistics



Data



Statistics



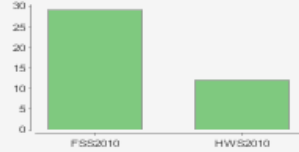
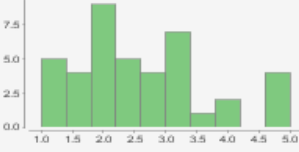
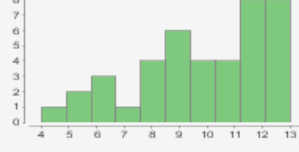
Charts



Advanced  
Charts



Annotations

Name	Type	Missing	Statistics
<a href="#">Semester</a>	Binominal	0	 <p>Least: HWS2010 (12)      Most: FSS2010 (29)</p> <p><a href="#">Open chart</a></p>
<a href="#">Name</a>	Polynomial	0	 <p>Least: Tanja Becker (3)      Most: Frank Fester (5)</p> <p><a href="#">Open chart</a></p>
<a href="#">Course</a>	Polynomial	0	 <p>Least: Algorithms I (5)      Most: Database Systems I (10)</p> <p><a href="#">Open chart</a></p>
<a href="#">Mark</a>	Numeric	0	 <p>Min: 1      Max: 5      Average: 2.593</p> <p><a href="#">Open chart</a></p>
<a href="#">Attended</a>	Integer	0	 <p>Min: 4      Max: 13      Average: 9.976</p> <p><a href="#">Open chart</a></p>

# The Results View - Charts



Data



Statistics



Charts

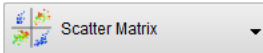


Advanced  
Charts

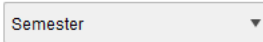


Annotations

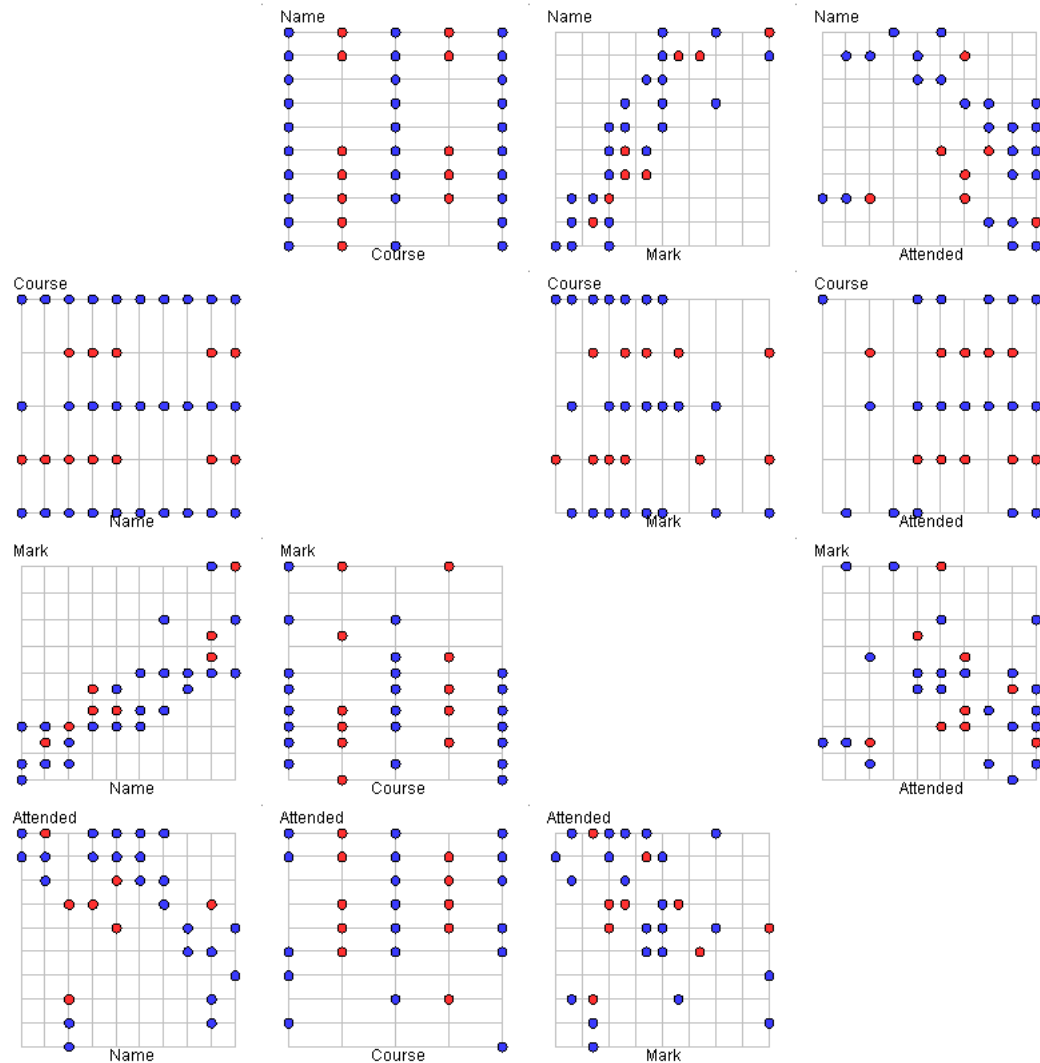
Chart style:



Plots:



Jitter:



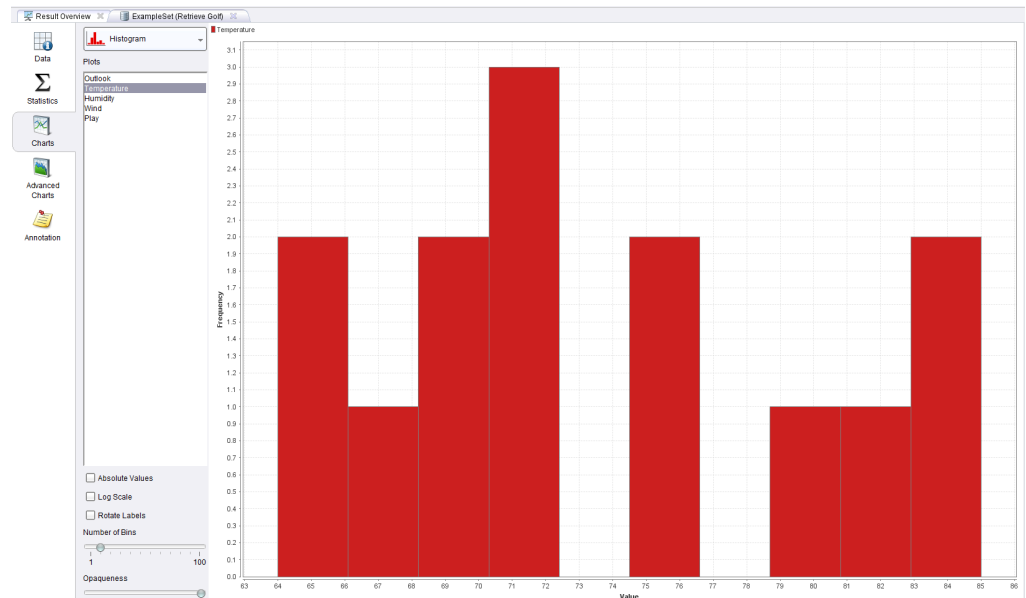
# Data Visualisation

- Visualisation of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyse large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

**Visualisation is the conversion of data into a visual format so that the characteristics of the data and the relationships among data items or attributes can be analysed.**

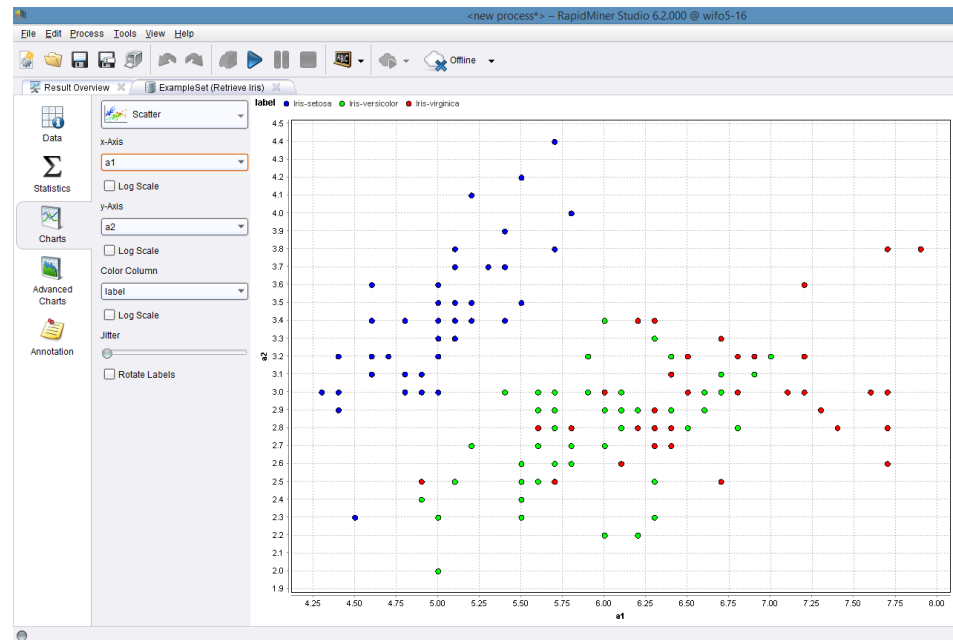
# Visualisation Techniques: Histogram

- Usually used to display the distribution of values of a **single attribute**
  - Divide the values into bins and show a bar plot of the number of objects in each bin
  - The height of each bar indicates the number of objects per bin
  - Shape of histogram depends on the number of bins



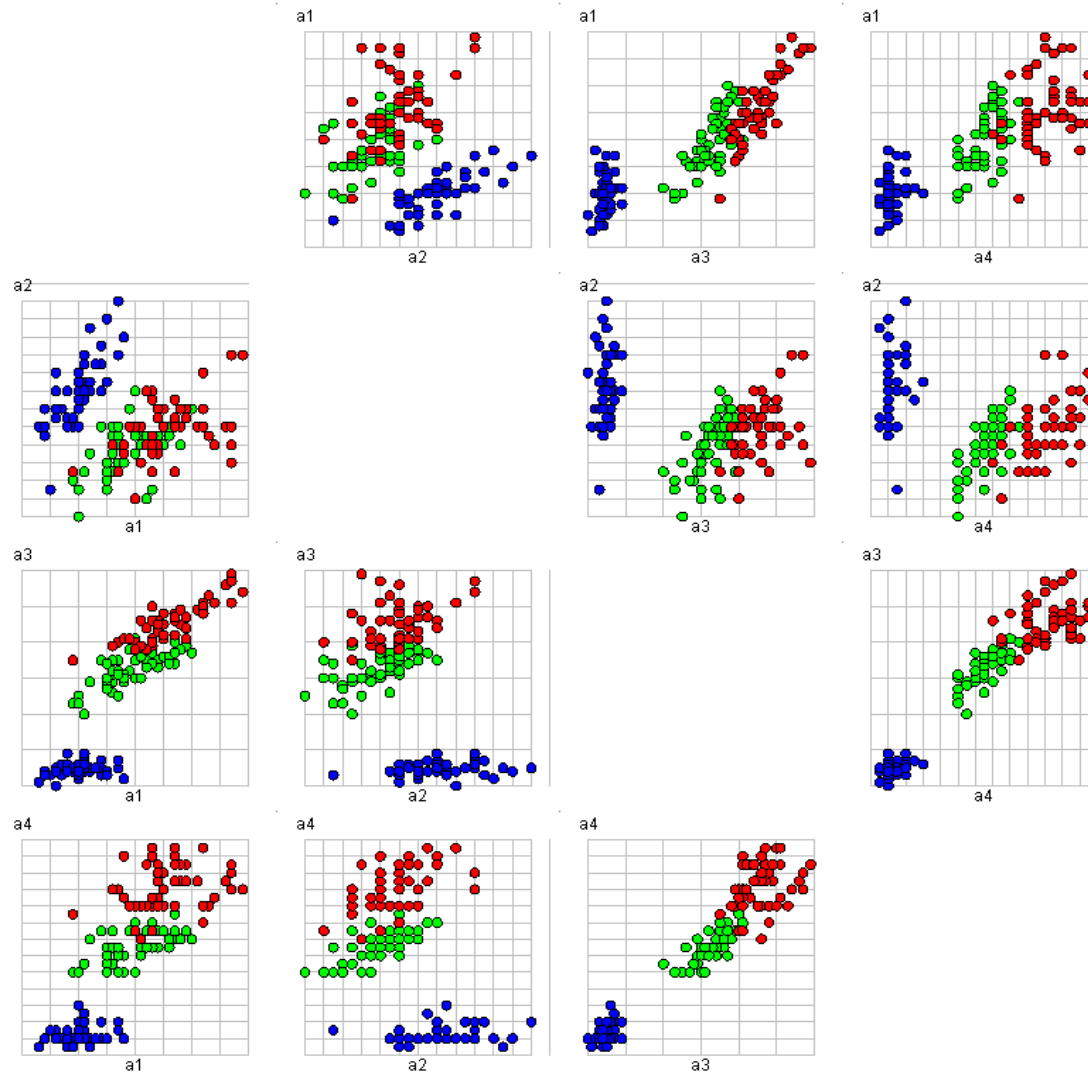
# Visualisation Techniques: Scatter Charts

- Two-dimensional scatter charts are most commonly used
- Often additional attributes/dimensions are displayed by using the size, shape, and color of the markers that represent the objects
- It is useful to have arrays of scatter charts that can compactly summarise the relationships of several pairs of attributes
- RapidMiner Scatter Charts
  - Scatter (single chart)
  - Scatter Multiple
  - Scatter Matrix
  - Scatter 3D



# RapidMiner Chart: Scatter Matrix

label ■ Iris-setosa ■ Iris-versicolor ■ Iris-virginica



# RapidMiner Resources

- RapidMiner 9.3:
  - <https://my.rapidminer.com/nexus/account/index.html#downloads>
- Rapidminer User Manuals: <http://rapidminer.com/documentation/>
- Open Access Book covering RapidMiner
  - Matthew North: Data Mining For The Masses:  
<https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf>
- RapidMiner Forum and Discussion Groups: <https://community.rapidminer.com/>
- Video Tutorials
  - by Rapid-I: <https://www.youtube.com/user/RapidIVideos>
  - by Neutral Market Trends: <http://www.neuralmarkettrends.com/tutorials/>
- MyExperiment: process repository: <http://www.myexperiment.org/>

# Hands-on!

- Now start RapidMiner
- Load your first dataset
- Start exploring the data!

# Examples for Data Profiling

- Students Data Set

Course	Taught in	# Students	Grade Range	Max. Attend
Algorithms I	HWS2010	5	1.7 – 5.0	12
Database Systems I	FSS2010	10	1.3 – 5.0	13
Database Systems II	HWS2010	7	1.0 – 5.0	13
Electronic Markets	FSS2010	10	1.0 – 3.0	13
Software Engineering	FSS2010	9	1.3 – 4.0	13

- Scatter Chart

- Y-Axis: Course
- X-Axis: try!