UNIVERSITÄT
MANNHEIM

Data Mining

Introduction to RapidMiner

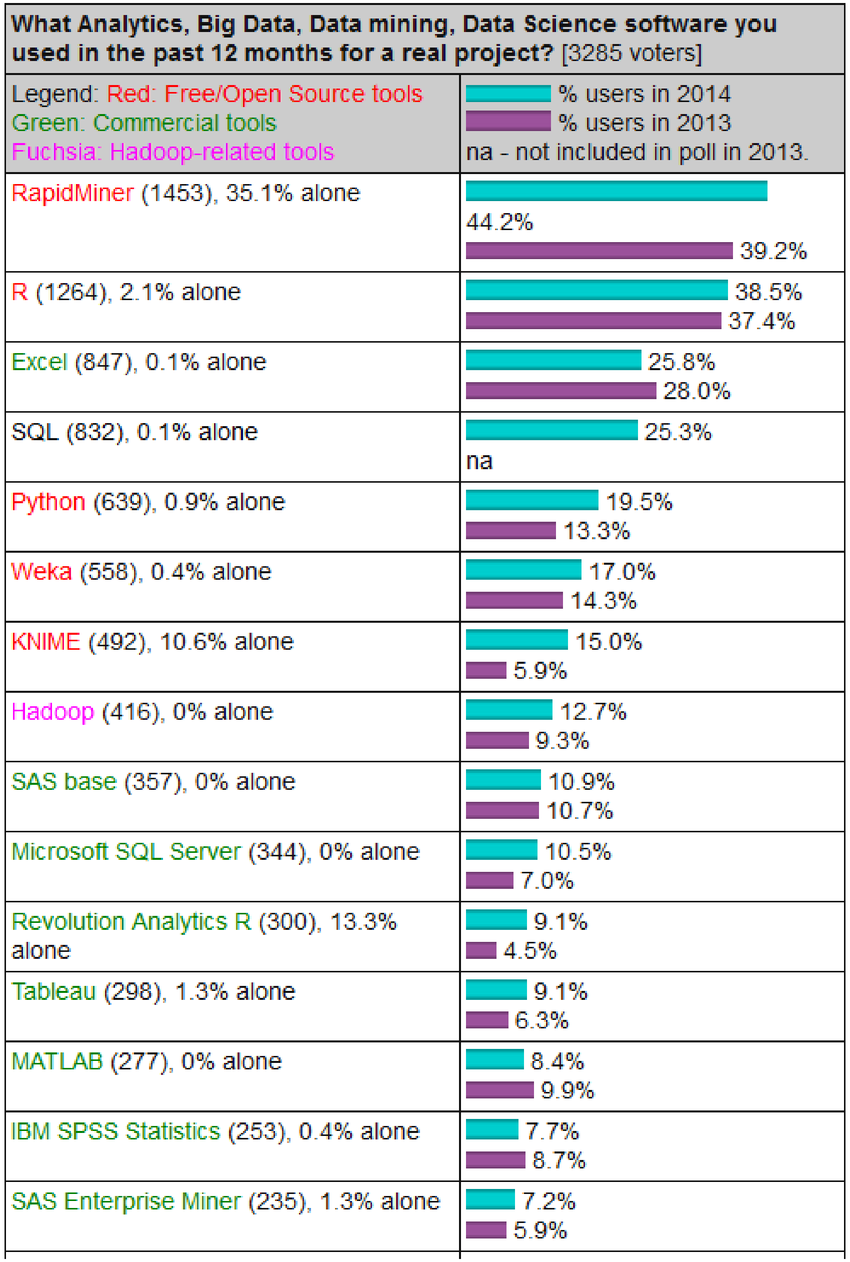
Outline

1. What is RapidMiner?
2. RapidMiner Terminology
3. RapidMiner User Interface
    1. The Design Perspective
    2. The Results Perspective
4. The RapidMiner Repository
5. Preprocessing Operators
6. Data Visualization
7. RapidMiner Resources

RapidMiner

- A very comprehensive open-source data mining tool
  - The data mining process is visually modeled as an operator chain
  - RapidMiner has over 400 build 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
    (= Second Last Edition)
  - Enterprise Edition: Commercial
    (= Last Edition plus professional support)

KDnuggets Poll 2014 (and 2013)



What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [3285 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop-related tools	🟦 % users in 2014 🟪 % users in 2013 na - not included in poll in 2013.
RapidMiner (1453), 35.1% alone	44.2% 39.2%
R (1264), 2.1% alone	38.5% 37.4%
Excel (847), 0.1% alone	25.8% 28.0%
SQL (832), 0.1% alone	25.3% na
Python (639), 0.9% alone	19.5% 13.3%
Weka (558), 0.4% alone	17.0% 14.3%
KNIME (492), 10.6% alone	15.0% 5.9%
Hadoop (416), 0% alone	12.7% 9.3%
SAS base (357), 0% alone	10.9% 10.7%
Microsoft SQL Server (344), 0% alone	10.5% 7.0%
Revolution Analytics R (300), 13.3% alone	9.1% 4.5%
Tableau (298), 1.3% alone	9.1% 6.3%
MATLAB (277), 0% alone	8.4% 9.9%
IBM SPSS Statistics (253), 0.4% alone	7.7% 8.7%
SAS Enterprise Miner (235), 1.3% alone	7.2% 5.9%

Source: http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html

RapidMiner Terminology

- Concept: Class of examples, e.g. *person*, *flower*
- Example: Instance of a class, e.g. *Franz Müller*
- Example Set: The collection of all examples, that are mined.
- Attributes: Properties, e.g. *name*, *address*, *age*, *size*
  - Attributes have Value Types (see next slides)
  - Attributes have Roles (see next slides)
- Process: The complete data transformation and mining process.
- Model: The result of applying a process, e.g. *a classification model or set of rules*.

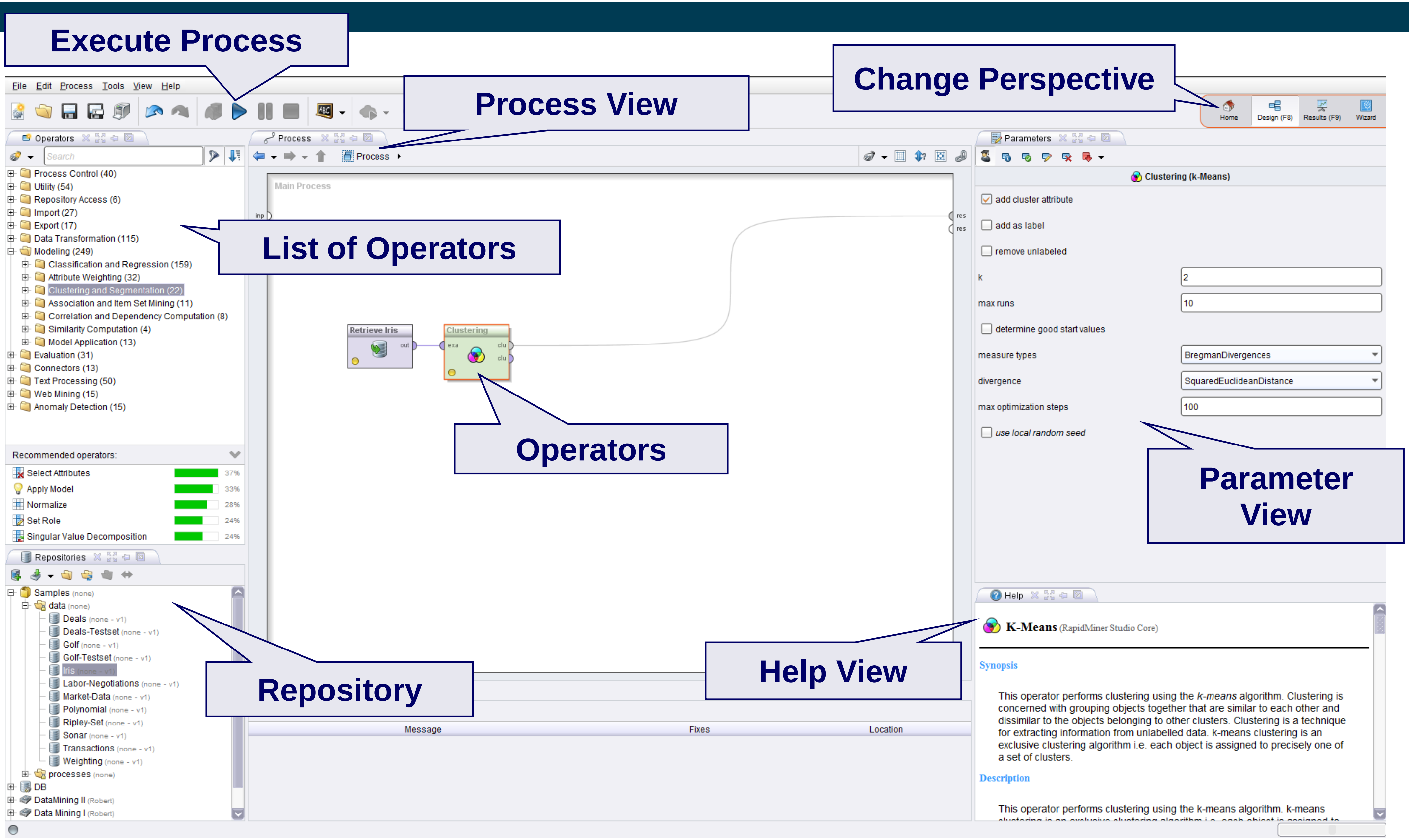
Value Types in RapidMiner

Value Type	RapidMiner Name	Description
Nominal	nominal	Categorical non-numerical values, usually used for finite quantities of different characteristics
2-value nominal	binominal	Special case of nominal, where only two different values are permitted
Multi-value nominal	polynominal	Special case of nominal, where more than two different values are permitted
Numerical values	numeric	For numerical values in general
Integers	integer	Whole numbers, positive and negative
Real numbers	real	Real numbers, positive and negative
Date time	date_time	Date as well as time
Date	date	Only date
Time	time	Only time
Text	text	Random free text without structure

Attribute Roles in RapidMiner

- Special attributes
  - Id: A unique identifier of examples, e.g. *matriculation number*
  - Cluster: The assigned cluster of an example, e.g. *after clustering*
  - Label: The class attribute, e.g. *for classification tasks*
  - Prediction: The predicted label, e.g. *after applying a classification model*
- Regular Attributes
  - Regular attributes are used to describe characteristics of an example

The Design Perspective



Execute Process

Process View

Change Perspective

List of Operators

Operators

Parameter View

Repository

Help View

Clustering (k-Means)

☑ add cluster attribute

□ add as label

□ remove unlabeled

k	2
max runs	10

□ determine good start values

measure types	BregmanDivergences
divergence	SquaredEuclideanDistance
max optimization steps	100

□ use local random seed

K-Means (RapidMiner Studio Core)

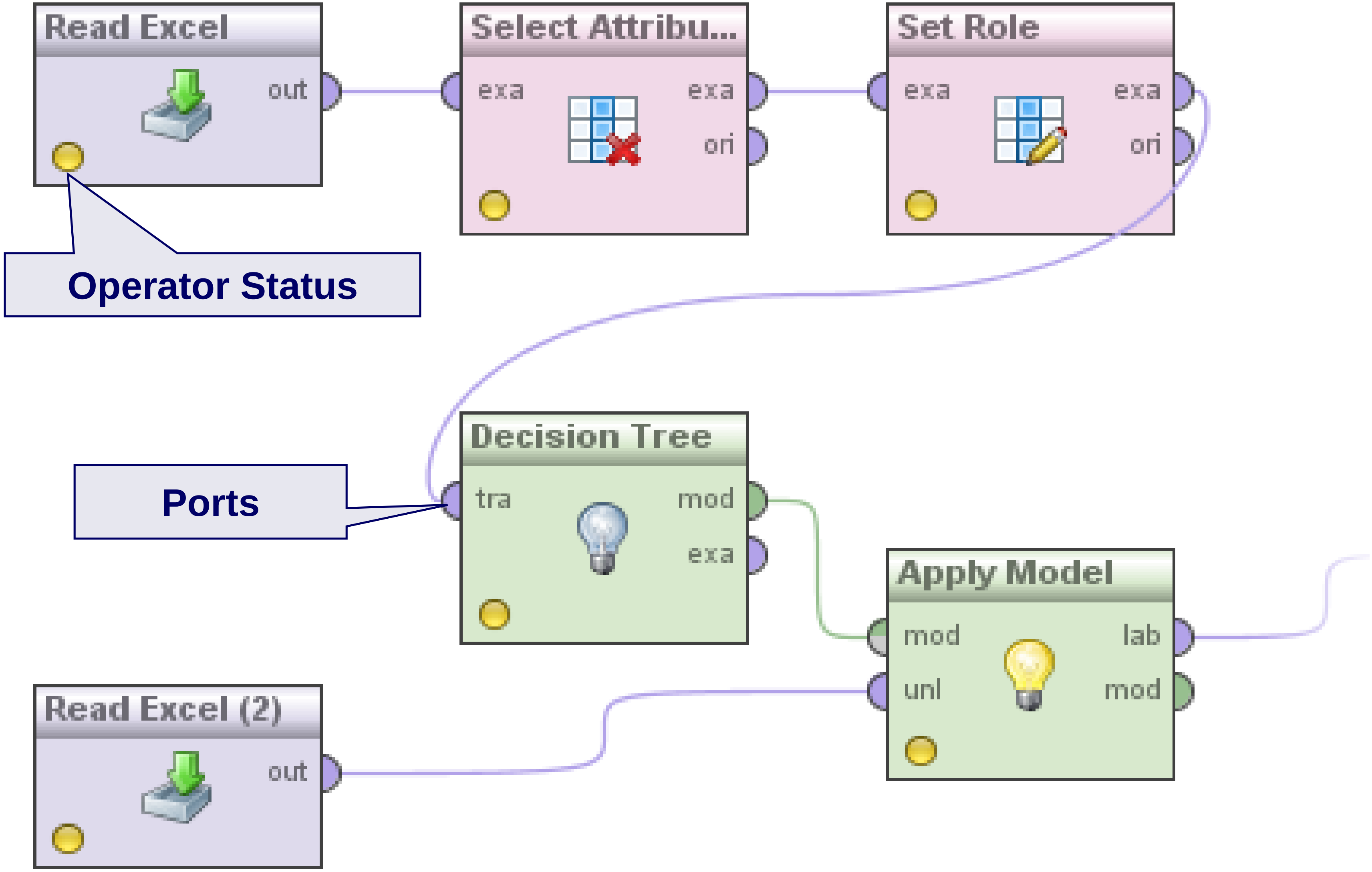
Synopsis

This operator performs clustering using the *k-means* algorithm. Clustering is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters. Clustering is a technique for extracting information from unlabelled data. k-means clustering is an exclusive clustering algorithm i.e. each object is assigned to precisely one of a set of clusters.

Description

This operator performs clustering using the k-means algorithm. k-means

Specifying a Process by Chaining Operators

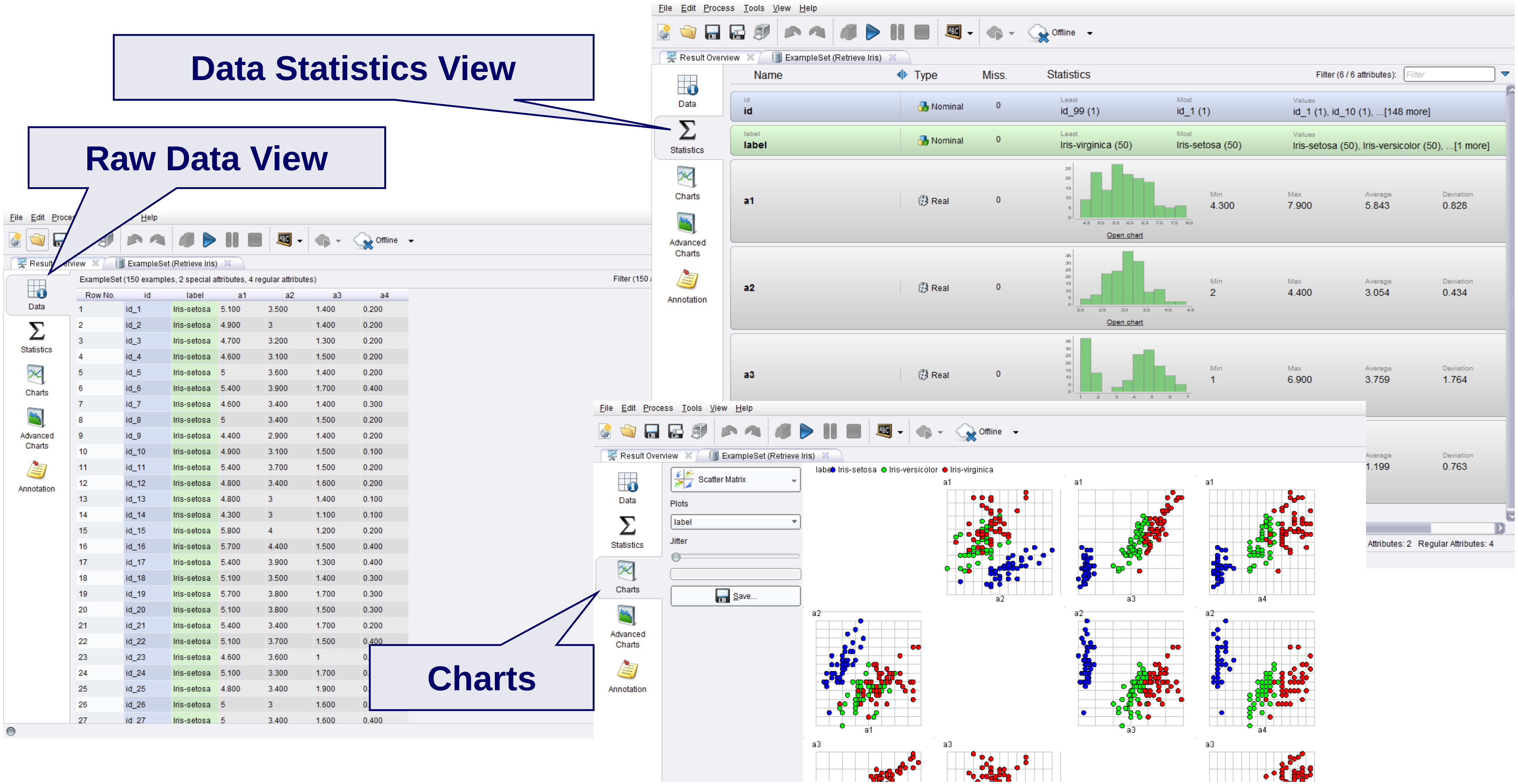


Operator Status

Ports

The Results Perspective

– Perspective for displaying and visualizing Example Sets and Models.



Data Statistics View

Raw Data View

Charts

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Filter (150 /

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	4.600	3.600	1	0.
24	id_24	Iris-setosa	5.100	3.300	1.700	0.
25	id_25	Iris-setosa	4.800	3.400	1.900	0.
26	id_26	Iris-setosa	5	3	1.600	0.
27	id_27	Iris-setosa	5	3.400	1.600	0.400

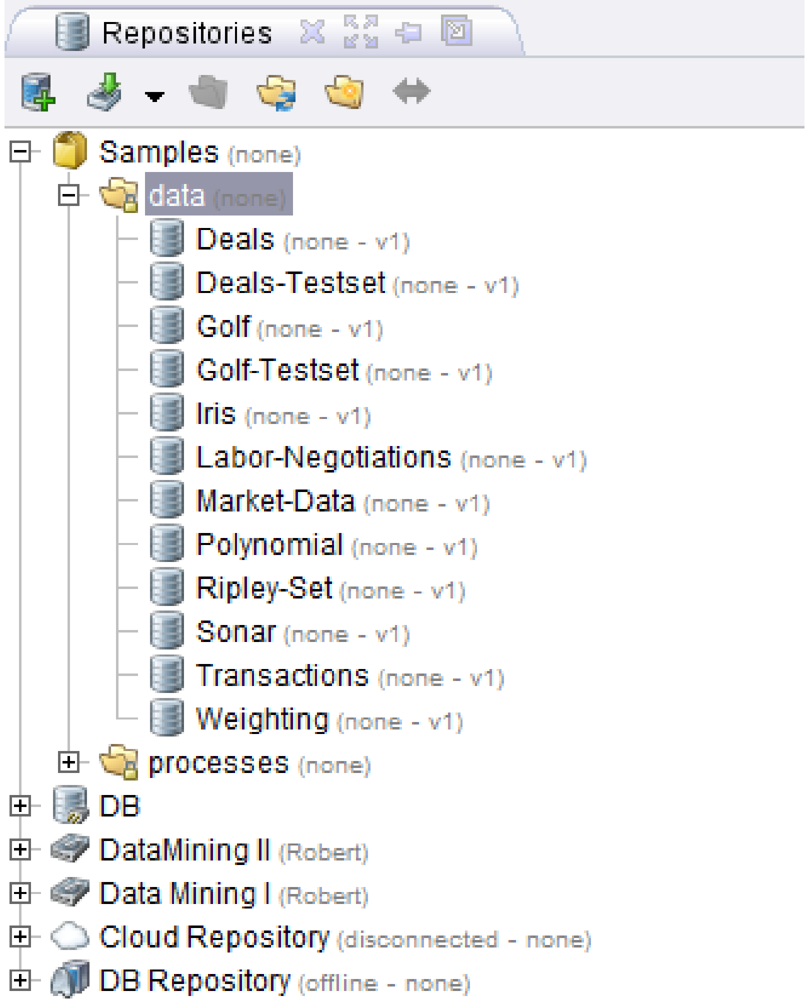
Filter (6 / 6 attributes): Filter

Name	Type	Miss.	Statistics			
id	Nominal	0	Least id_99 (1)	Most id_1 (1)	Values id_1 (1), id_10 (1), ...[148 more]	
label	Nominal	0	Least Iris-virginica (50)	Most Iris-setosa (50)	Values Iris-setosa (50), Iris-versicolor (50), ...[1 more]	
a1	Real	0	Min 4.300	Max 7.900	Average 5.843	Deviation 0.828
a2	Real	0	Min 2	Max 4.400	Average 3.054	Deviation 0.434
a3	Real	0	Min 1	Max 6.900	Average 3.759	Deviation 1.764
					Average 1.199	Deviation 0.763

Attributes: 2 Regular Attributes: 4

The Repository

- Repository types
  - Local (files on your machine)
  - Remote (in the Cloud or own RapidMiner server)
- Location to store:
  - Datasets and its meta-data
  - Processes
- Data Import/Export Formats
  - CSV and XML Files
  - Excel Spreadsheets
  - Microsoft Access Tables
  - ARFF Files (Weka file format)
  - ...



Repositories

Samples (none)
data (none)
Deals (none - v1)
Deals-Testset (none - v1)
Golf (none - v1)
Golf-Testset (none - v1)
Iris (none - v1)
Labor-Negotiations (none - v1)
Market-Data (none - v1)
Polynomial (none - v1)
Ripley-Set (none - v1)
Sonar (none - v1)
Transactions (none - v1)
Weighting (none - v1)
processes (none)
DB
DataMining II (Robert)
Data Mining I (Robert)
Cloud Repository (disconnected - none)
DB Repository (offline - none)

(Some common) Preprocessing Operators

- Type and Role Conversions
  - Type: Provides for converting between different attribute types
  - Role: Provides for setting attribute roles (e.g. label, id)
- Attribute Set Reduction and Transformation
  - Select Attributes: Selects a sub set of the available attributes
  - Generate Attribute: Generates new attribute from existing attributes
- Value Transformations
  - Normalize: Normalizes attribute to certain range
- Filtering
  - Filter Examples: Filters the examples for a give criterion
- Aggregation
  - Performs a SQL-like aggregation (count, sum, ...)

Data Visualization

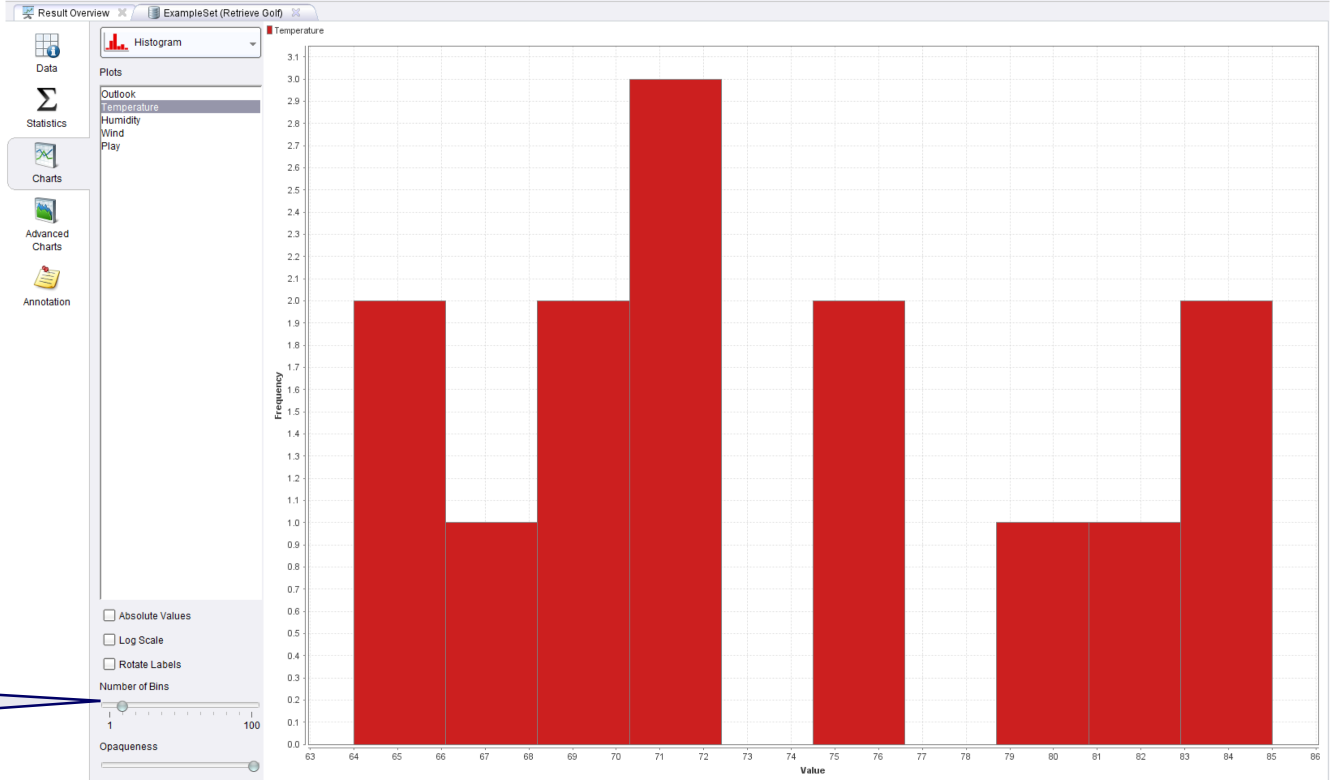
- Visualization of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Visualization 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 analyzed.

Visualization 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

RapidMiner Chart: Histogram



Histogram

Plots

Outlook
Temperature
Humidity
Wind
Play

□ Absolute Values

□ Log Scale

□ Rotate Labels

Number of Bins

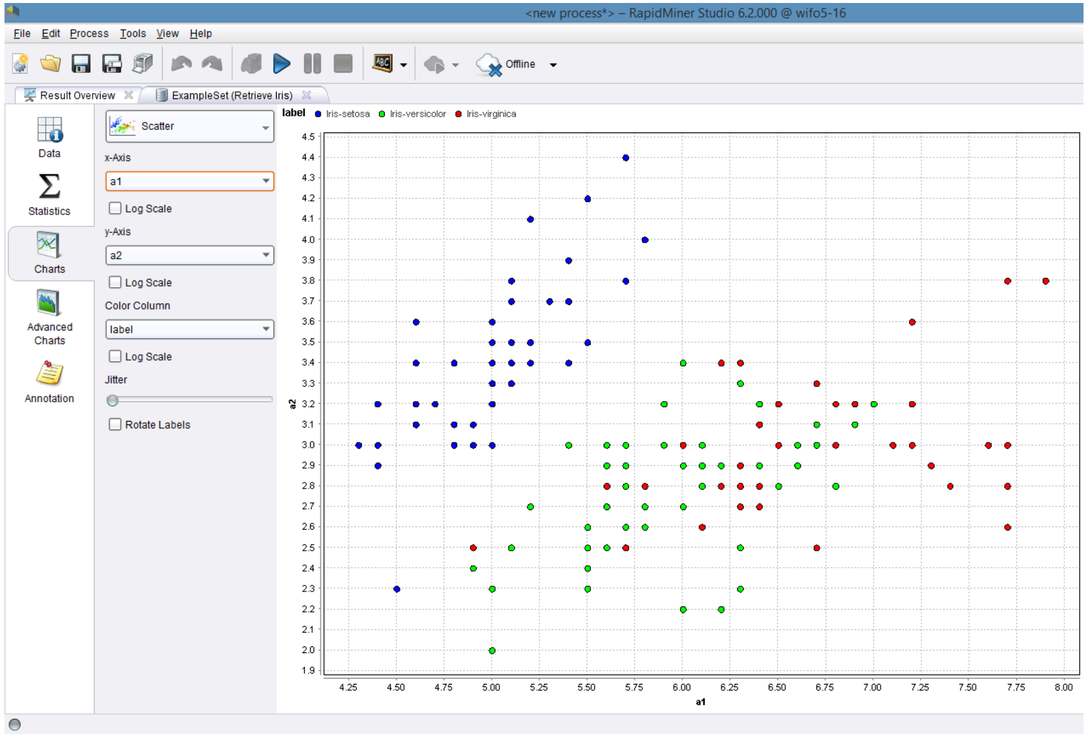
Opaqueness

Number of Bins

Visualization 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 summarize the relationships of several pairs of attributes
- RapidMiner Scatter Charts
  - Scatter (single chart)
  - Scatter Multiple
  - Scatter Matrix
  - Scatter 3D

RapidMiner Plotter: Scatter



<new process*> – RapidMiner Studio 6.2.000 @ wifo5-16

Scatter

x-Axis

a1

☐ Log Scale

y-Axis

a2

☐ Log Scale

Color Column

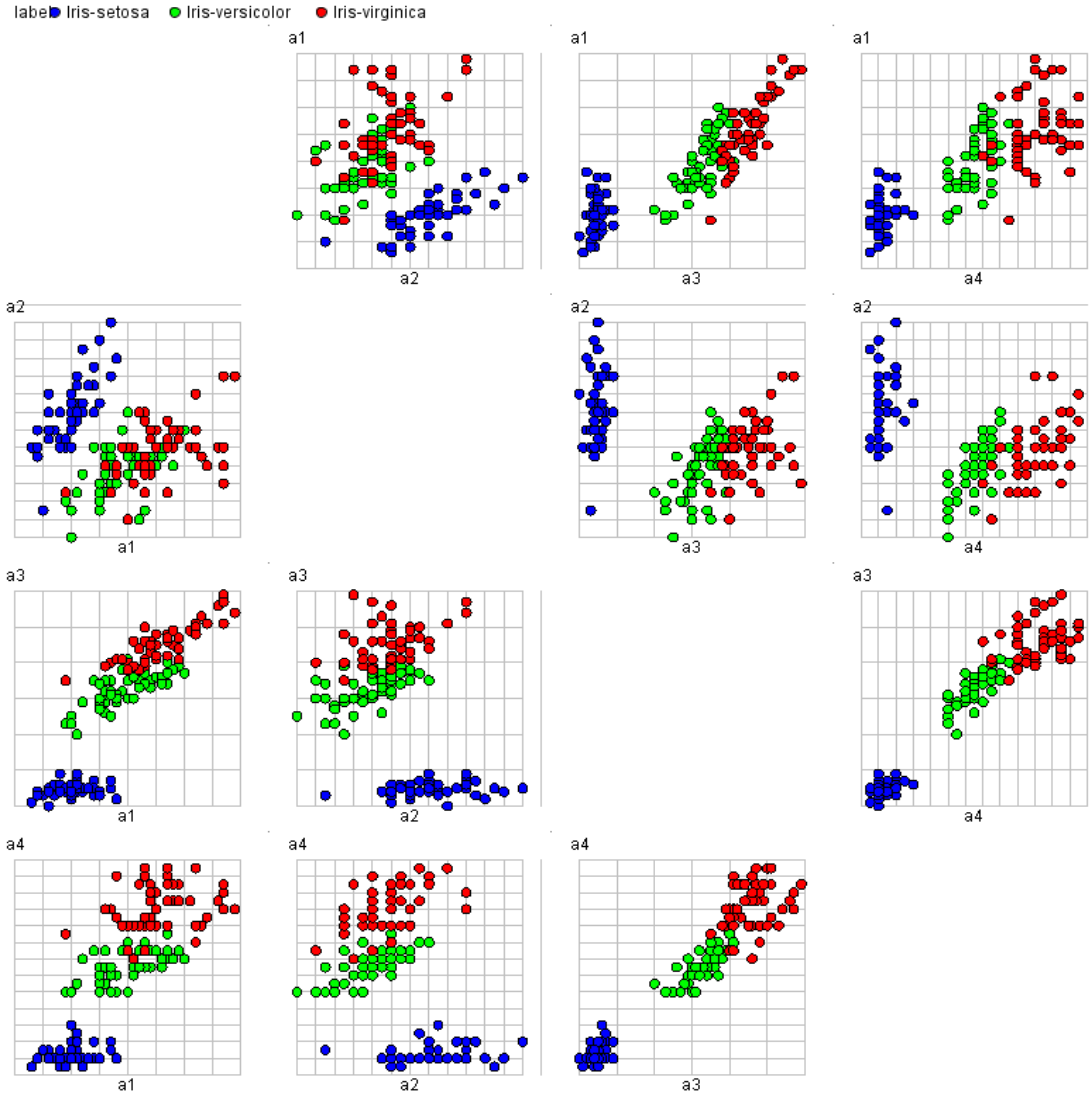
label

☐ Log Scale

Jitter

☐ Rotate Labels

RapidMiner Chart: Scatter Matrix



RapidMiner Resources

- RapidMiner 7:
  - https://my.rapidminer.com/nexus/account/index.html#downloads
  - **License Key** can be found in ILIAS
- Rapidminer User Manuals: http://rapidminer.com/documentation/
- Open Access Book covering RapidMiner
  - Matthew North: Data Mining For The Masses: http://dl.dropbox.com/u/31779972/DataMiningForTheMasses.pdf
- Operator Documentation: http://rapid-i.com/wiki/
- RapidMiner Forum and Discussion Groups: http://forum.rapid-i.com/
- Video Tutorials
  - by Rapid-I: https://www.youtube.com/user/RapidIVideos
  - by NDLR: https://dspace.ndlr.ie/jspui/handle/10633/2353
  - by Neutral Market Trends: http://www.neuralmarkettrends.com/tutorials/
- MyExperiment: process repository: http://www.myexperiment.org/