Data Mining

(Mining Knowledge from Data)

Statistics

Magda Friedjungová

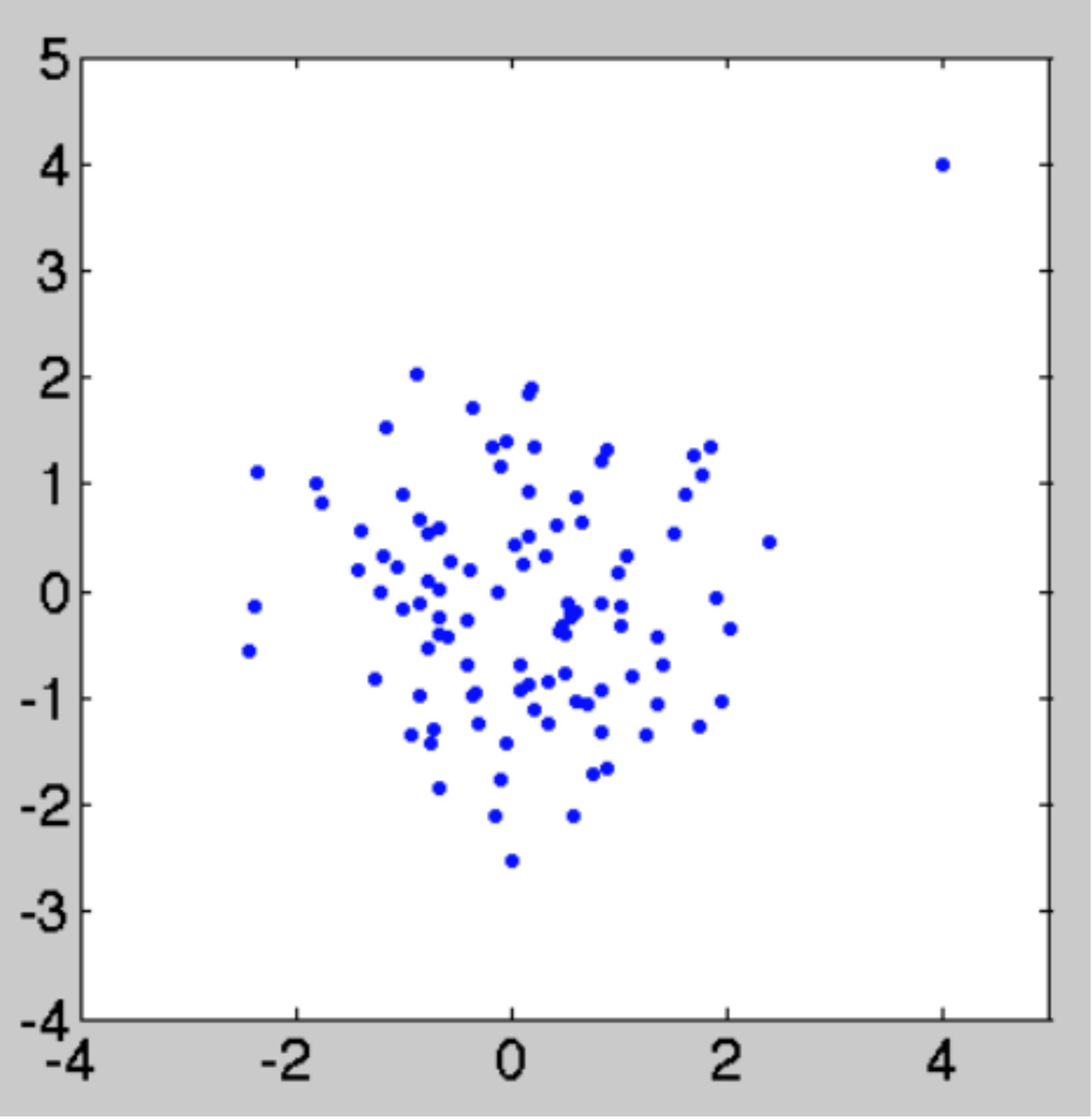
Outliers

Outliers

- An observation point that is distant from the other observations.
- Due to variability in the measurement or it may indicate an experimental error or any other error in the dataset.

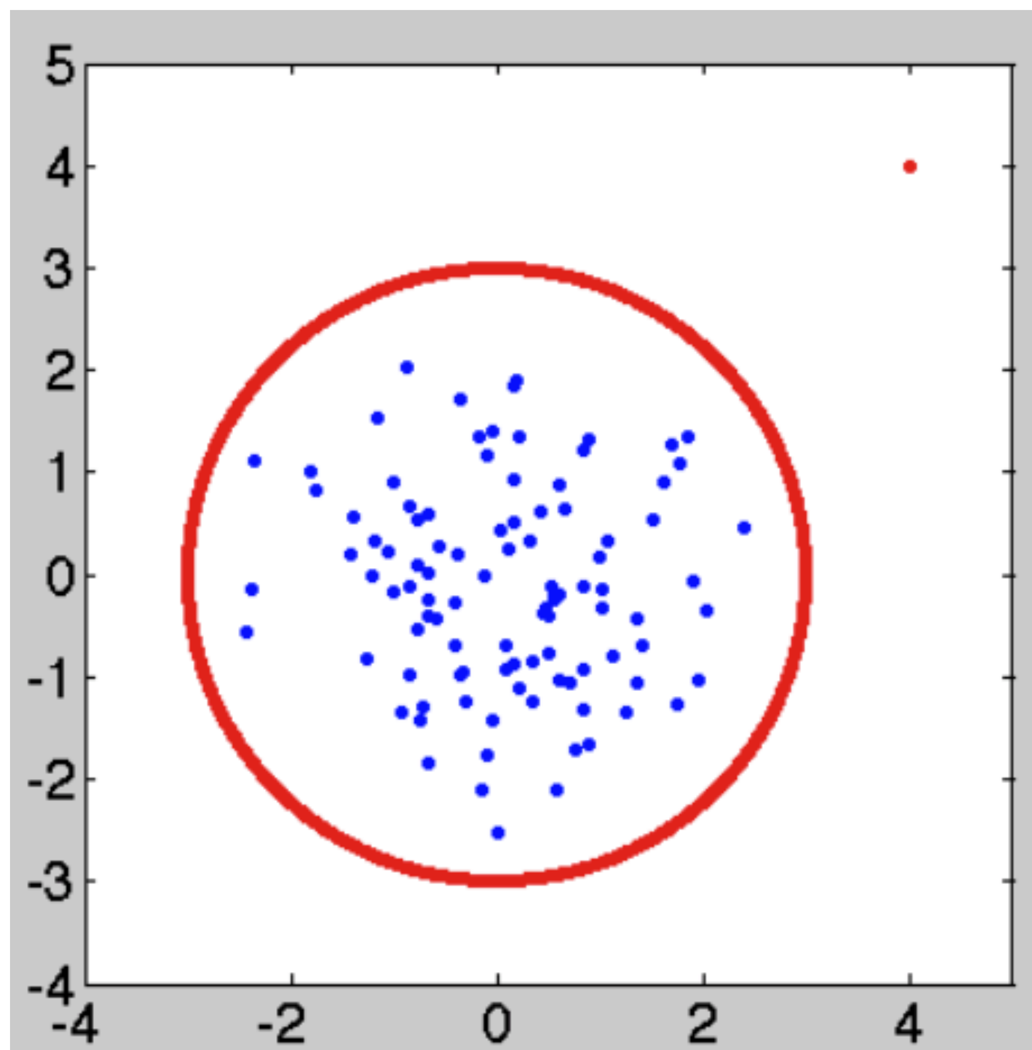
Outliers

- Which of the points is an outlier?



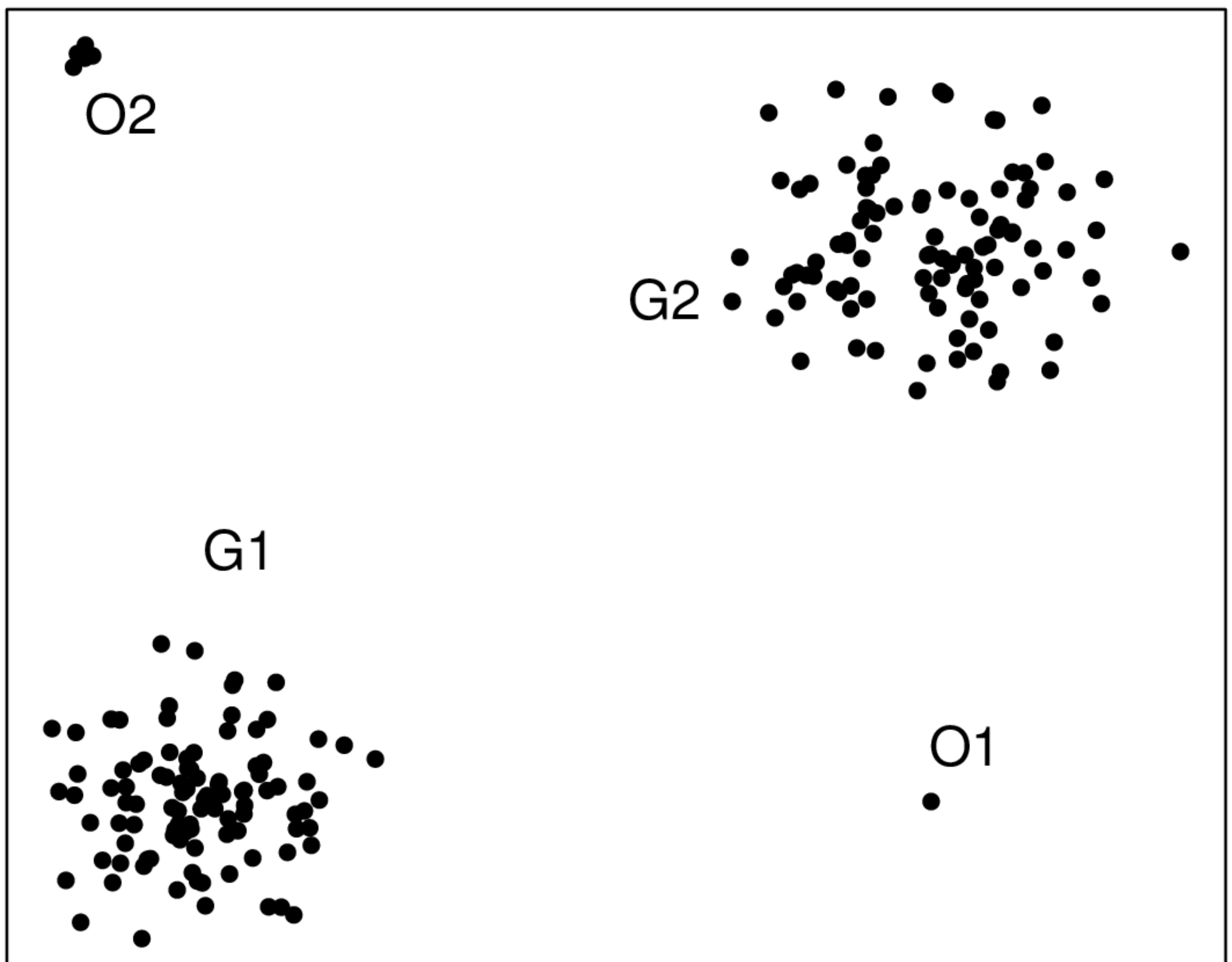
Outliers

- The blue data were generated from the Bell curve centered at [0, 0], the red dot was added later.



Outliers

- Two-dimensional outliers.

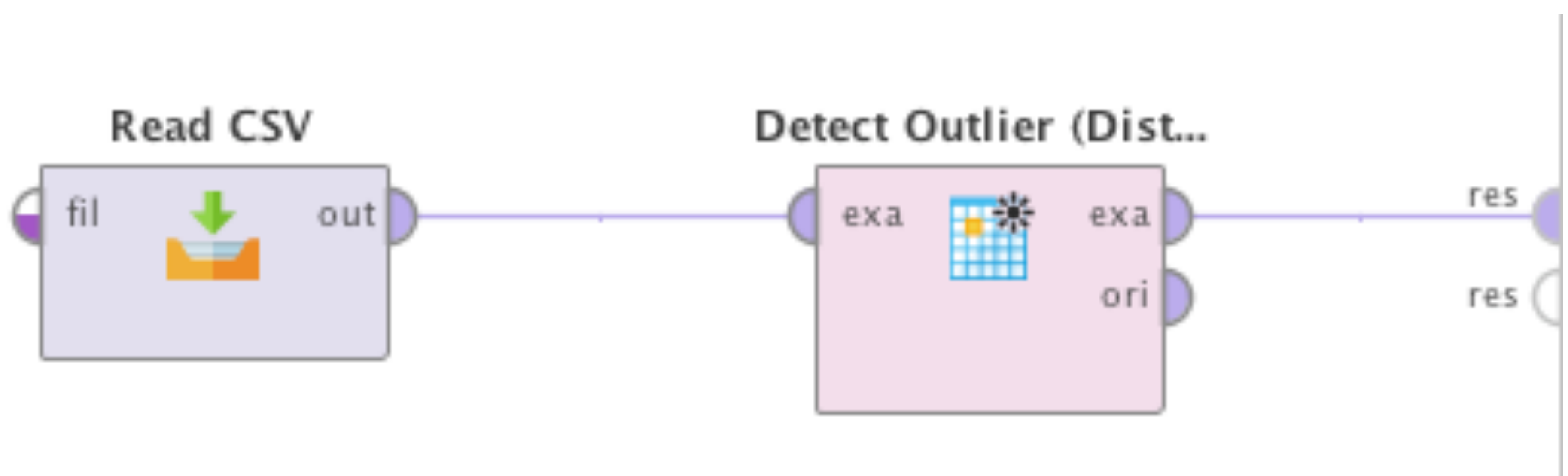


Outliers

- Download the dataset1.csv from EDUX.
- Start RapidMiner Studio.
- Load dataset1.csv to RapidMiner Studio.
- Find one outlier using the “Detect Outlier (Distances)” operator.

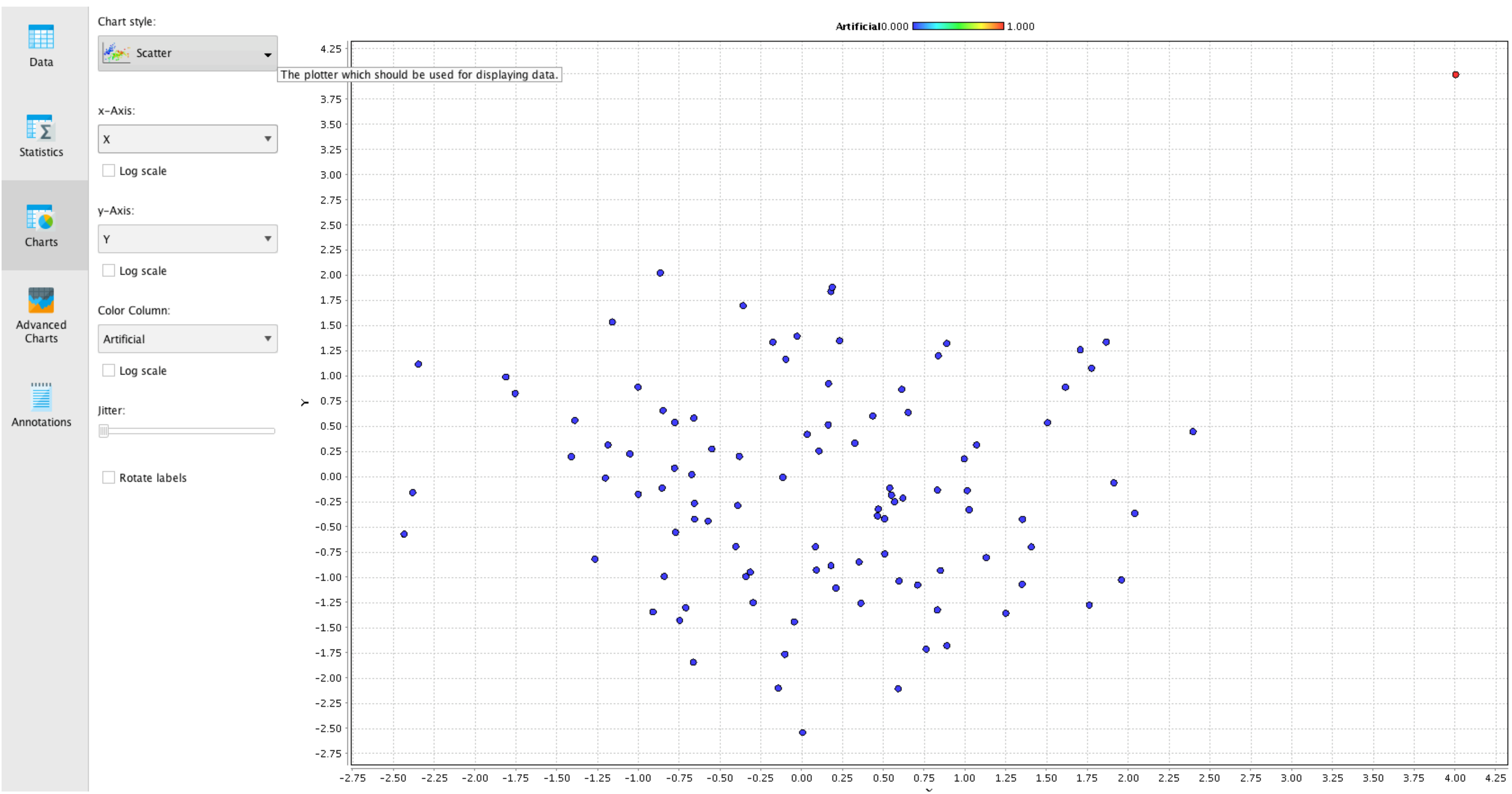
Outliers

- Set the delimiter to “Tab” in the “Read CSV” block.
- Set the number of outliers in the “Detect Outlier (Distances)” to one outlier.
- Press the Play button.



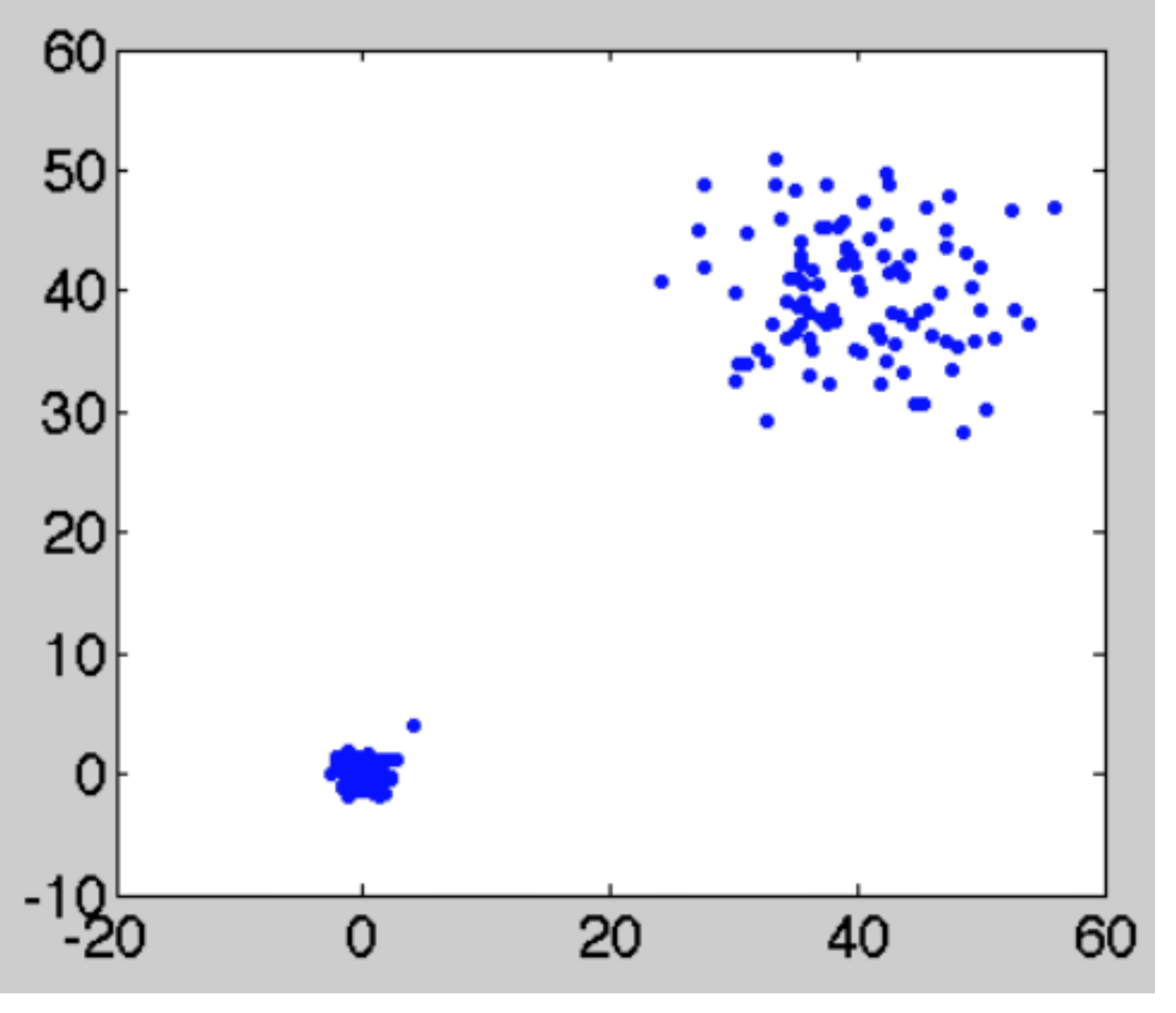
Outliers

- The result is an outlier in the upper right corner.



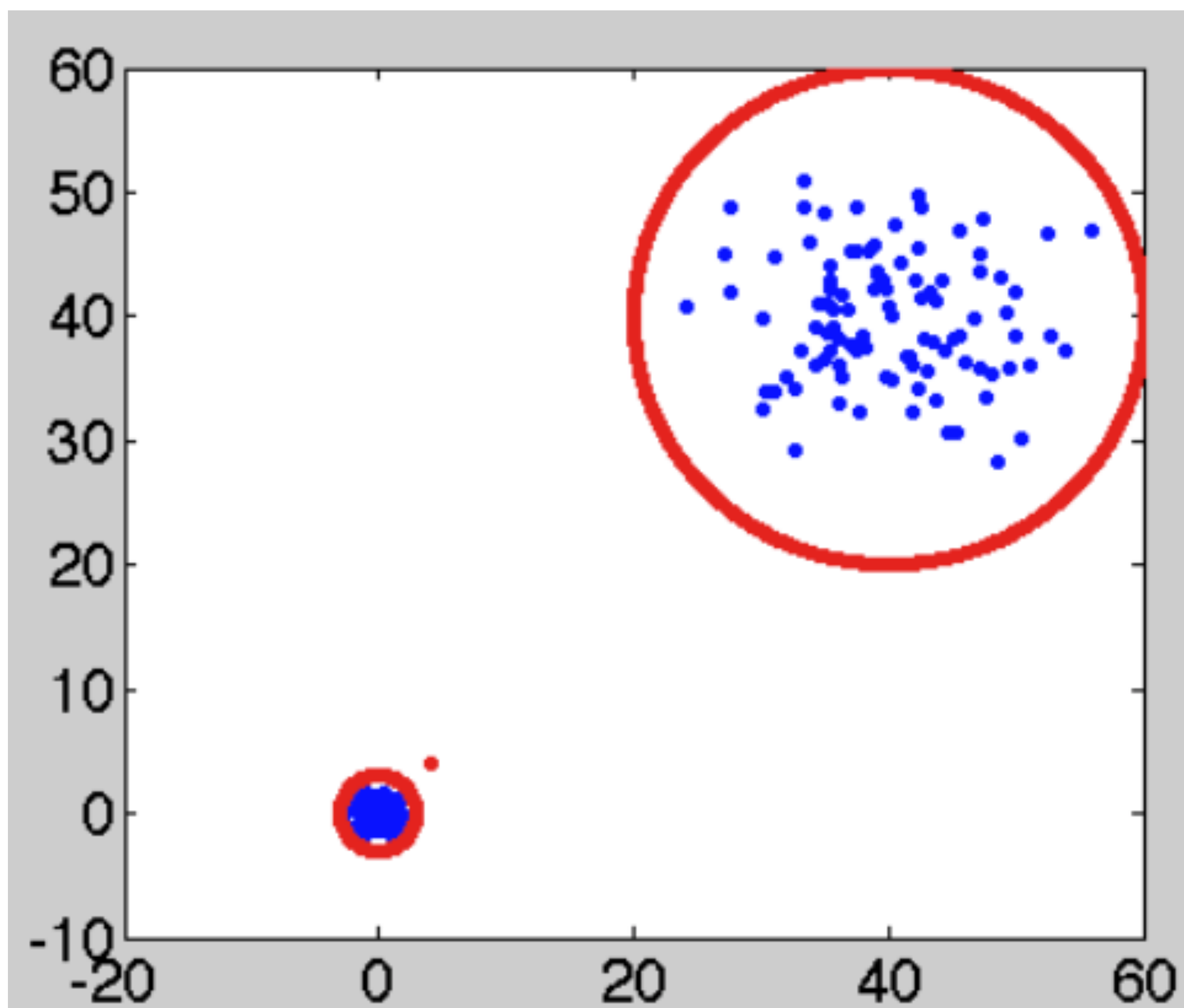
Outlier

- What is the outlier now?



Outliers

- The data were generated using two Bell curves positioned at [0, 0] and [40, 40]. The red dot was added later.

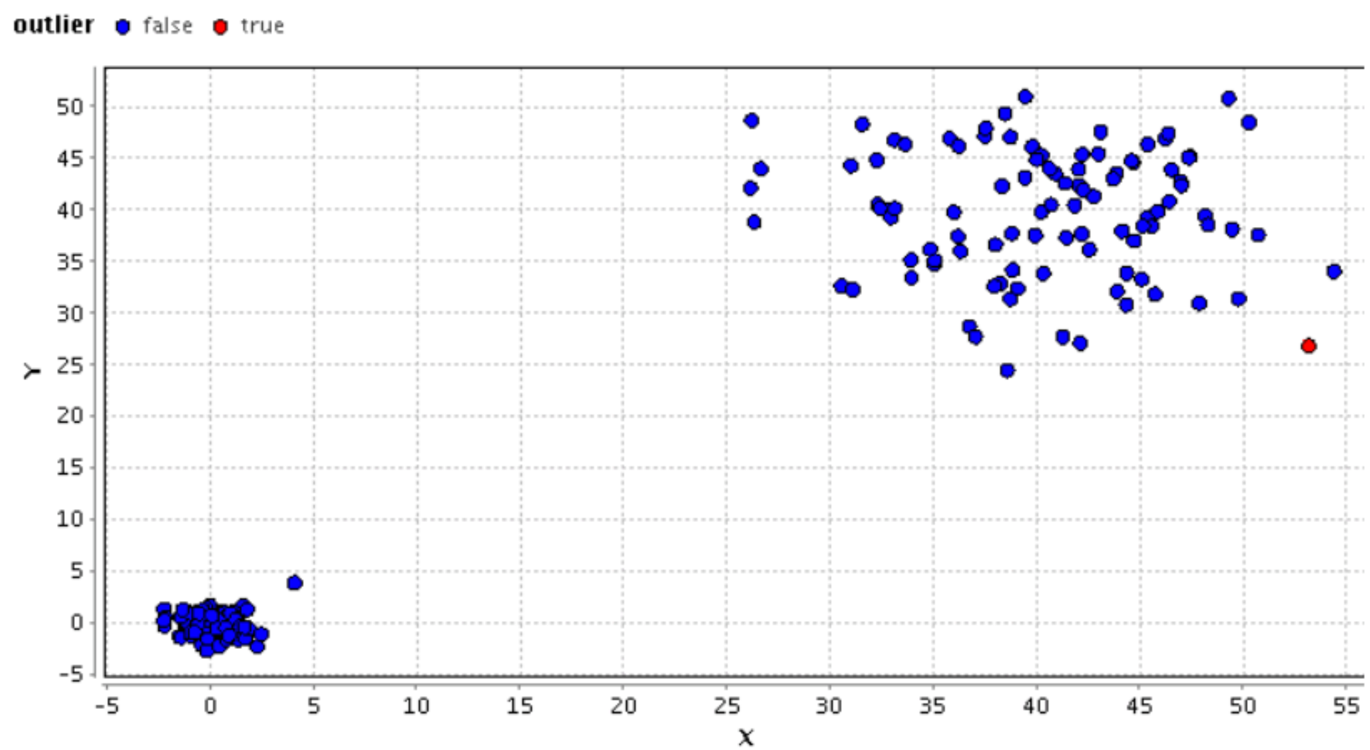


Outliers

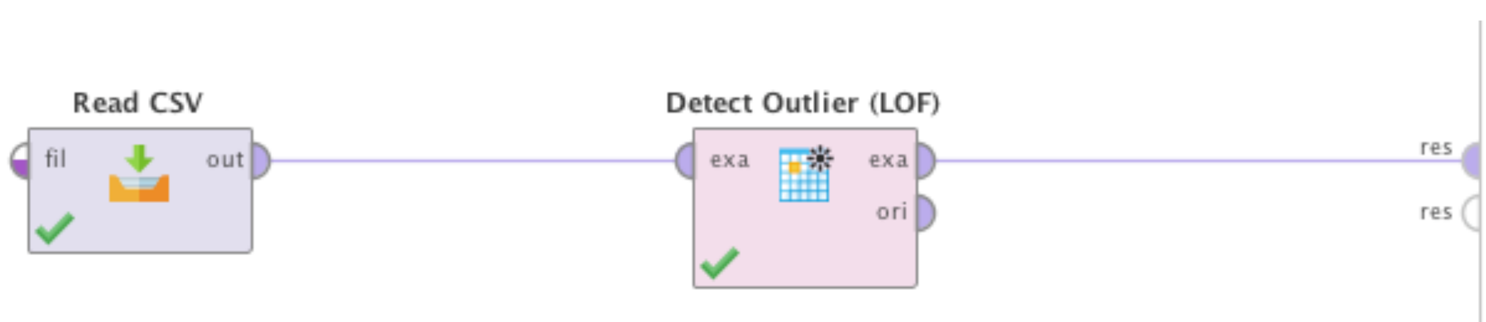
- Download the dataset2.csv from EDUX.
- Load dataset1.csv to RapidMiner Studio.
- Find one outlier by the “Detect Outlier (Distances)” operator.

Outliers

- This is not what we expected.

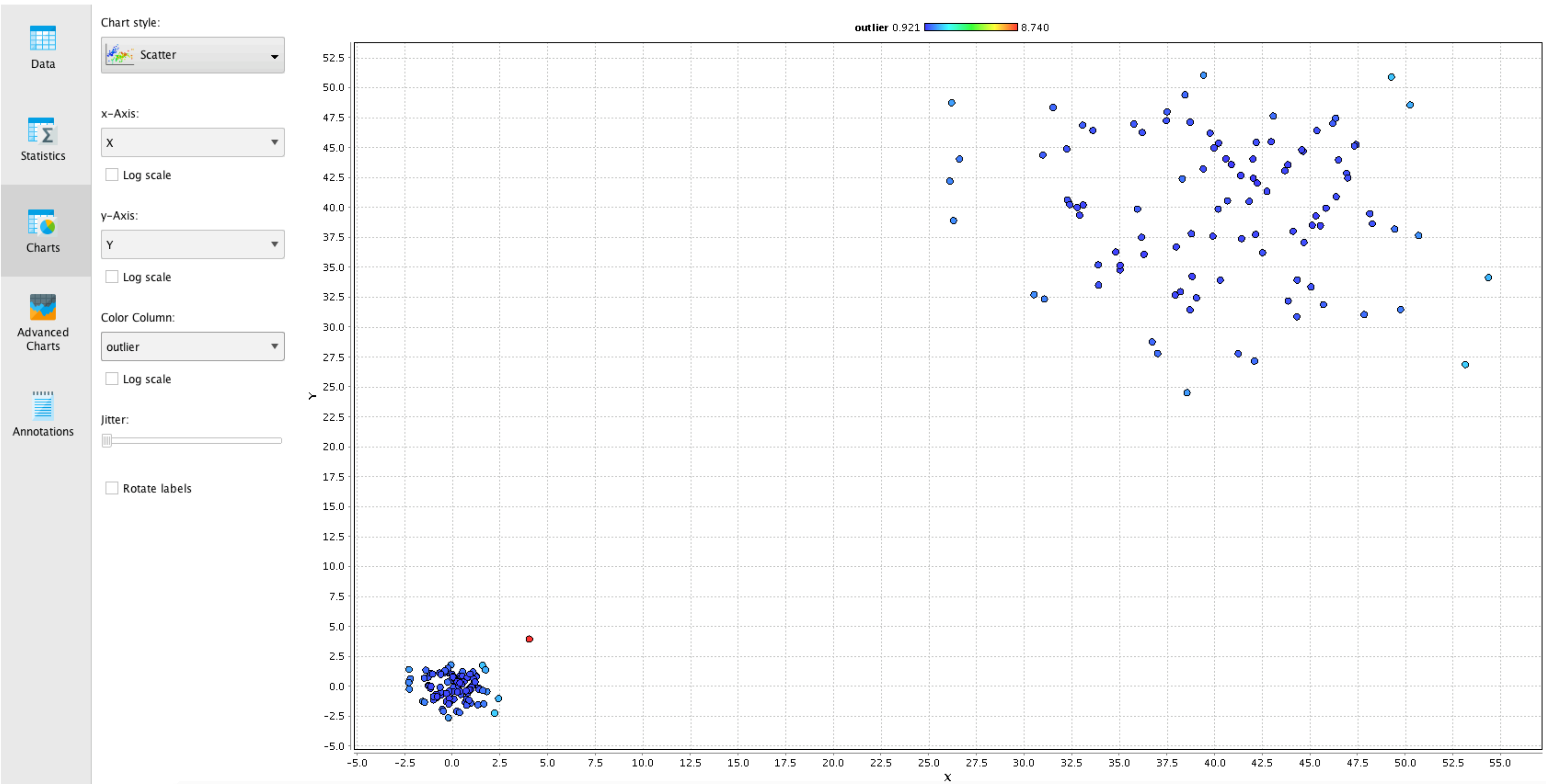


- Try to use “Detect Outliers (LOF)” block.



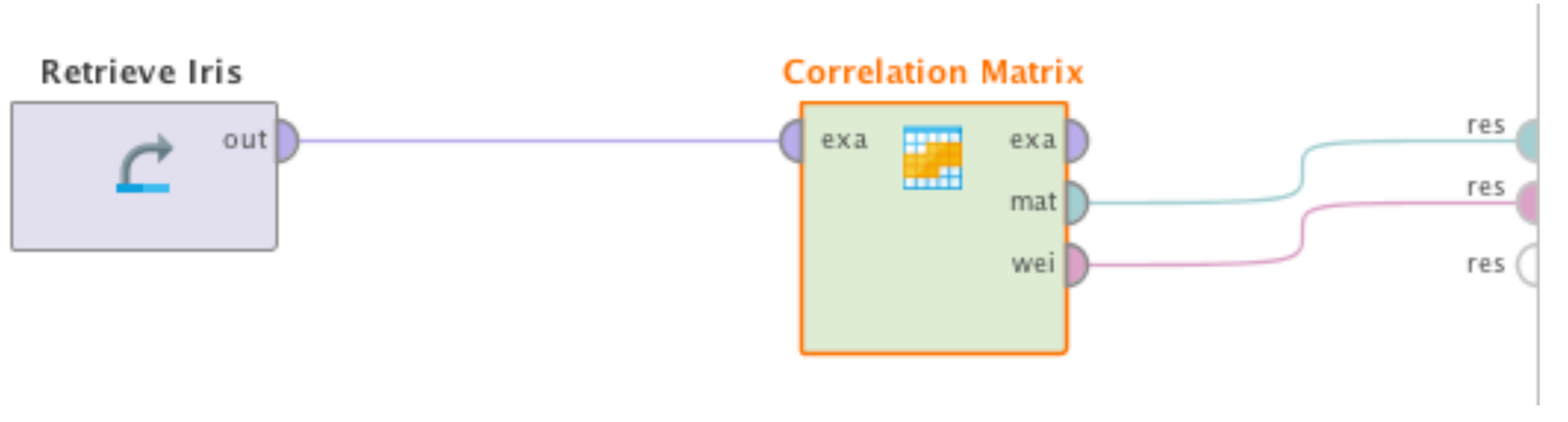
Outliers

- Finally, the desired result.



Correlation Matrix

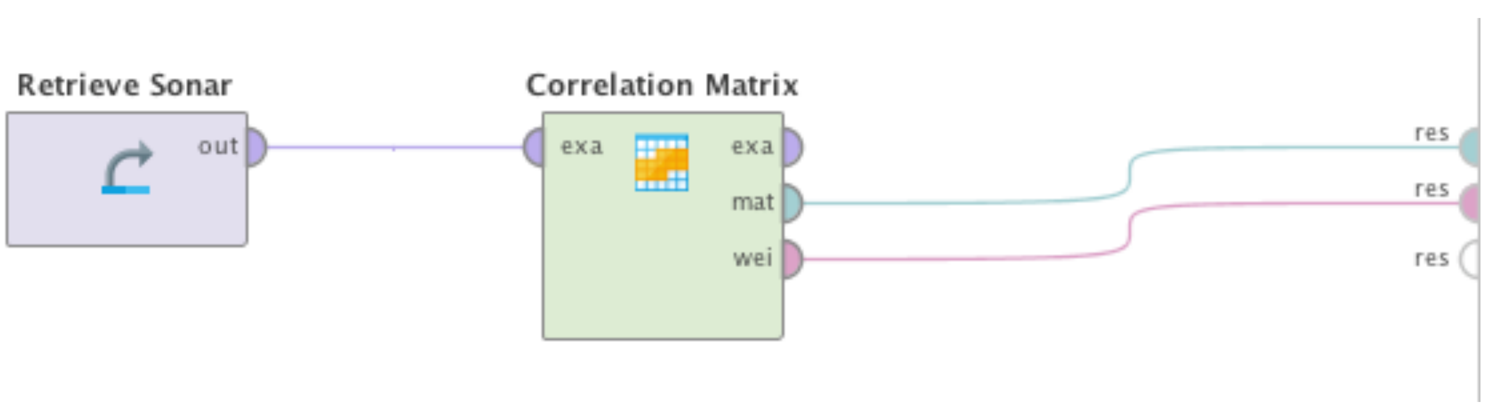
- Calculate the correlation matrix for the Iris dataset.



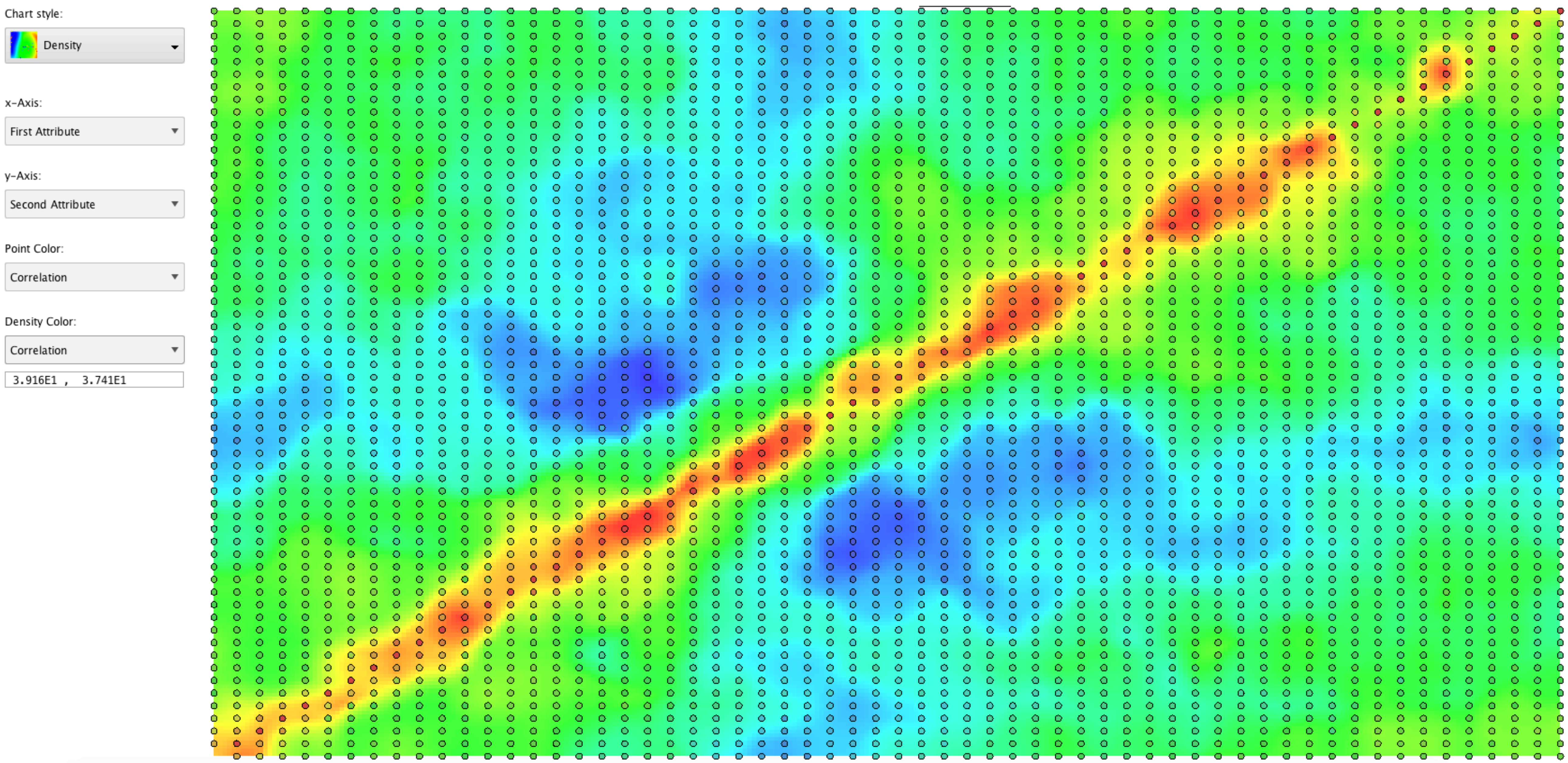
Correlation Matrix

Attributes	a1	a2	a3	a4
a1	1	-0.109	0.872	0.818
a2	-0.109	1	-0.421	-0.357
a3	0.872	-0.421	1	0.963
a4	0.818	-0.357	0.963	1

- How do you interpret it?
- Try to do the same for the Sonar dataset.

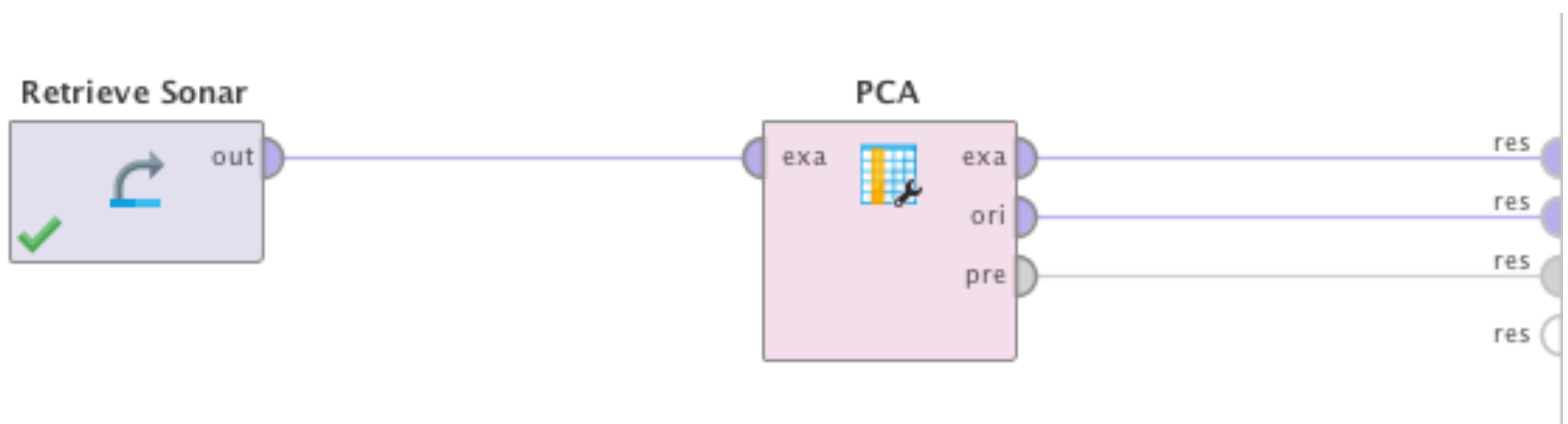


Sonar Correlation Matrix

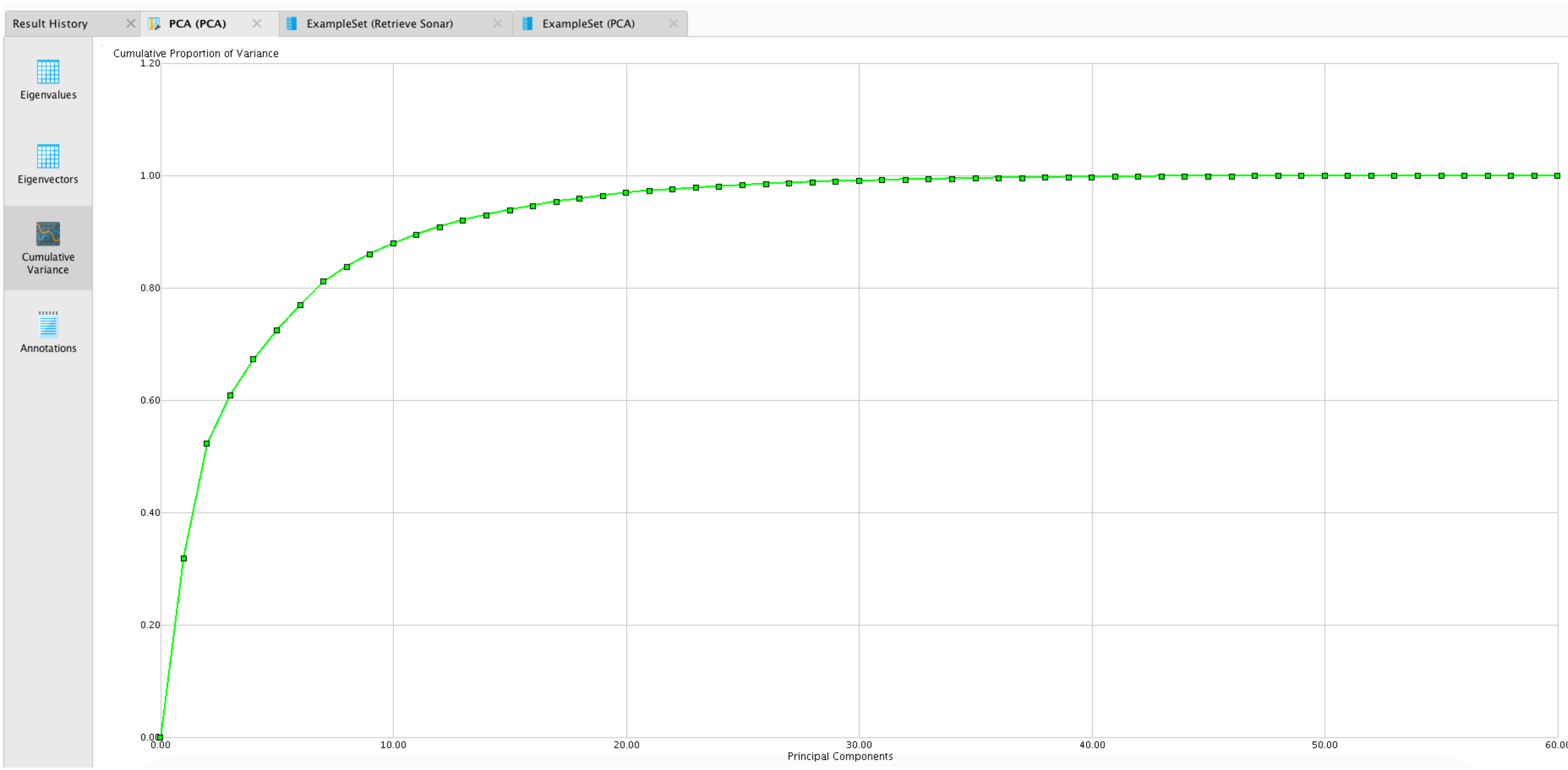


Linear Dependencies

- Principal Component Analysis (PCA)
- Use the Sonar dataset



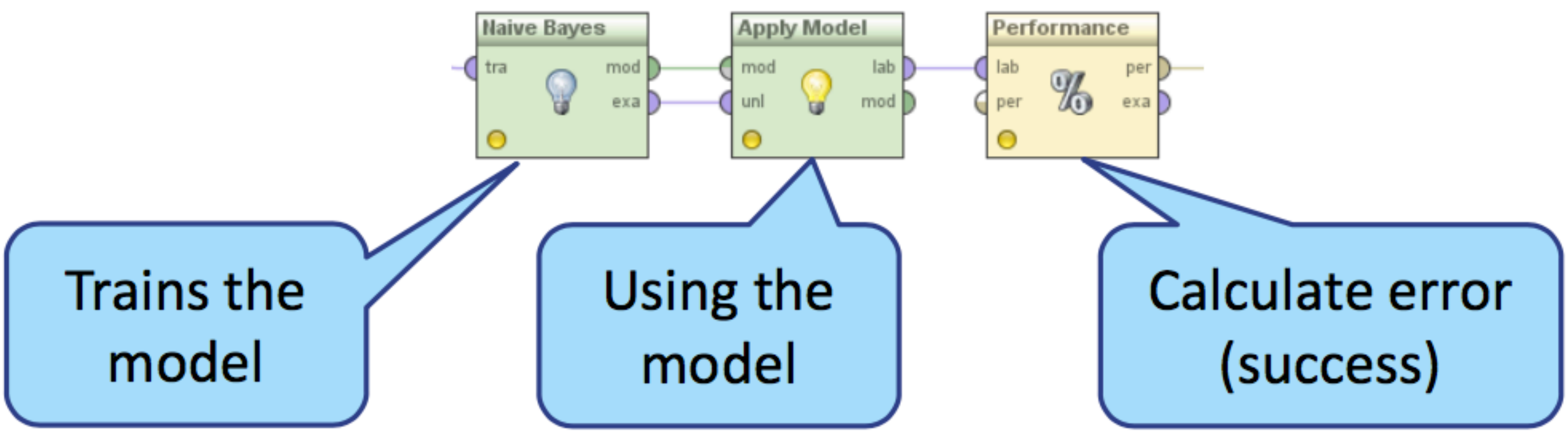
PCA - Cumulative Variance Plot



- How do you interpret the graph?

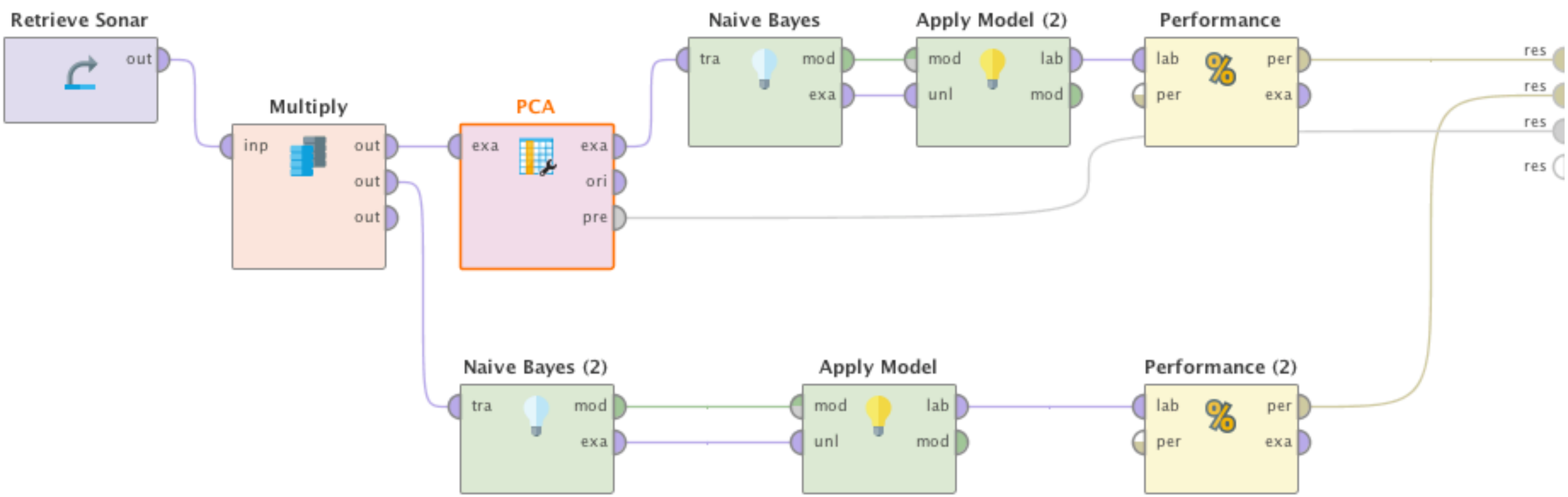
How much the PCA helps/hurts the models?

- View the model as a “black box” for now.
- Learn the model, use it on training data and detect the error.



PCA Experiment

- Build an experiment as shown.
- Which classifier (Naive Bayes) gives better results?



Accuracy

- PCA & Bayes

accuracy: 83.17%

	true Rock	true Mine	class precision
pred. Rock	77	15	83.70%
pred. Mine	20	96	82.76%
class recall	79.38%	86.49%	

- Bayes

accuracy: 73.08%

	true Rock	true Mine	class precision
pred. Rock	86	45	65.65%
pred. Mine	11	66	85.71%
class recall	88.66%	59.46%	

- Which one is more accurate and why?

Accuracy

- There are 60 attributes in the original dataset. When the PCA is used, 95% of the variance is captured in 17 attributes.
- This means that the model has simplified from 2^60 (considering binary attributes) to 2^17 possibilities.
- The dramatic simplification has led to classification accuracy increasing, even at the cost of losing 5% of the information.

Accuracy

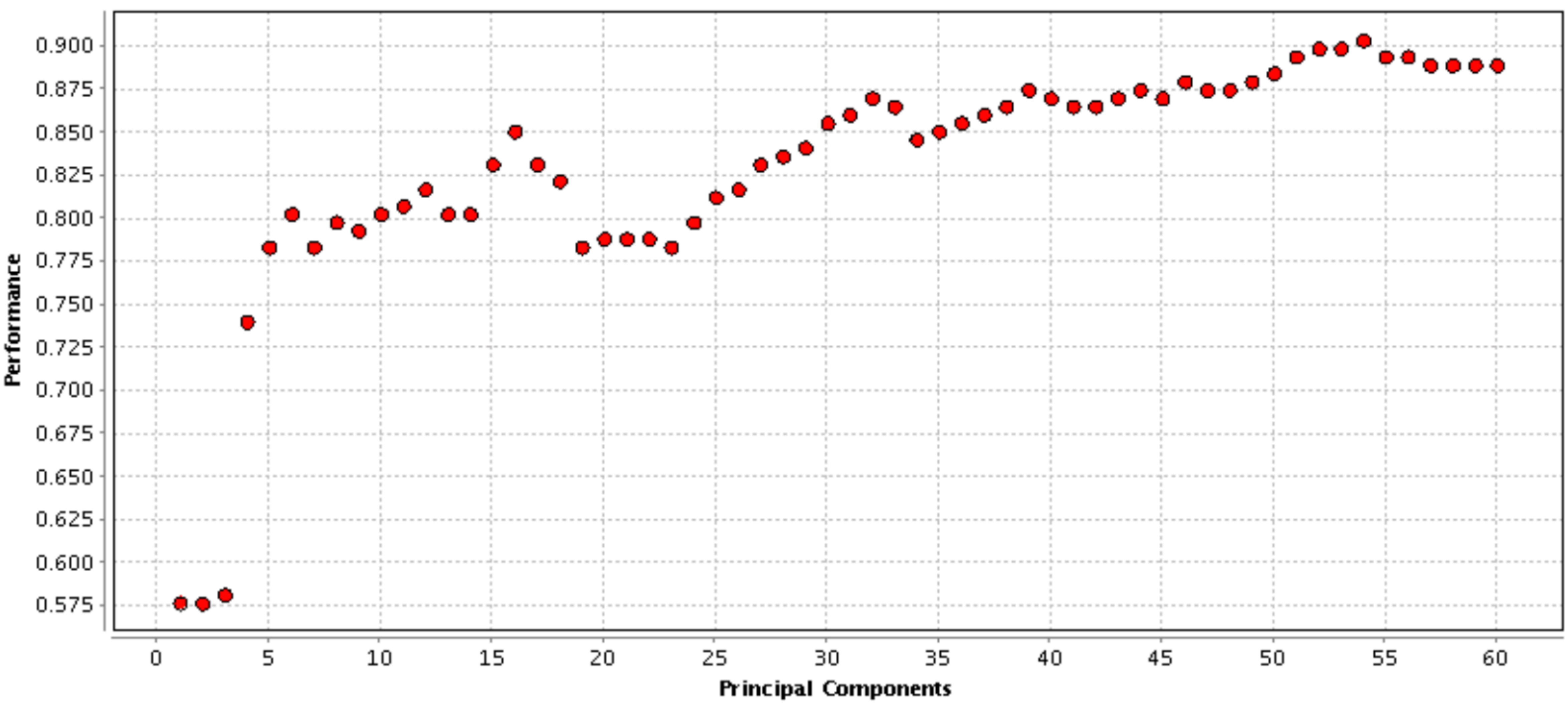
- We committed several inaccuracies and errors - for example assessing errors on the training data is not correct (as you will learn in later lecture).

- Also drawing conclusions from one value of success is quite risky (although it is a ten percent difference in our case).

- What is the **optimal** number of principal components?

Principal Components

- The ideal number is 16 or 54 attributes, depending on whether we want to save the computing time and memory, or we want to maximize accuracy.



Principal Components

- How do we create such a plot?