# Data Mining II - Exercise 3: Anomaly Detection

In this exercise we will focus on Anomaly Detection. Out of the box, RapidMiner offers four different operators for the detection of outliers. In addition, we will use the “Anomaly Detection” extension, which you can install via the RapidMiner Marketplace (it’s free).

## 1) Visually explore the output of the different methods!

Load the dataset “artificial.txt” in your repository and experiment with some of the outlier detection operators. This data set has only two attributes so it is easy to visualize. In addition to the “correct” data points, which form clusters that you can easily spot, 27 random data points were added.

Try to find these random points using “Detect Outlier (Distances)”, “Detect Outlier (Densities)”, “Detect Outlier (LOF)” and “Cluster-Based Local Outlier Factor (CBLOF)”. Try to figure out suitable parameters for these methods and check the result visually in a scatter plot.

## 2) Evaluate your results!

Load the dataset “breast_cancer_outliers.csv”. This dataset contains various features from cancer diagnostics. The majority of the dataset is data obtained from non-cancer patients (label “B”) and 20 examples are from samples where a cancer is present (label “M”). Because of their low frequency, the cancer examples can be treated as outliers and we can use the methods that you learned to detect them. To compare the different methods, use the “Generate ROC” operator and specify “M” for the parameter “label value for outliers”. The operator will generate data that you can use to plot an ROC curve. Compare the outlier detection methods that you have used in the previous task using ROC curves.

## 3) Apply it in classification!

Load the dataset “iris_shuffled_train.csv”. This is an alternative version of the Iris dataset, where some errors happened during data entry. Learn a decision tree classifier and check its performance on the test set “iris_shuffled_test.csv”. Try to improve the performance by removing the outliers when you train the model. (to simplify the task, outliers are marked in the training set)

## 4) Learn to recognise what you know!

Load the dataset “shuttle_train.csv” and learn a one-class SVM to recognise examples that are similar to those in the training set (use a sample of 5000 examples to reduce the runtime). The label “anomaly” tells you whether something is an outlier or not, and in this training set, nothing is an outlier. Apply your model to the test set “shuttle_test.csv”, which does contain outliers. Are you satisfied with the performance?