**Development of a Data Science study**

**on smart water metering data**

**Guidance document**

**Data4Waer Summer School**

**Bucharest, June 2 – June 27, 2017**

Antonio Candelieri, Francesco Archetti

University of Milano-Bicocca

Table of Contents

[1. Background of the summer school assignment 3](#_Toc484172170)

[2. Learning objectives of the summer school assignment 4](#_Toc484172171)

[3. Introduction to The Summer School Assignment 5](#_Toc484172172)

[4. Data Science tools 6](#_Toc484172173)

[5. Suggestions and Guidelines 10](#_Toc484172174)

[6. Most Relevant Resources 11](#_Toc484172175)

# Background of the summer school assignment

This second summer school of the DATA4WATER project follows the training course delivered at UPB on the topic of “Using Data Science for Urban Water Management”. Among other topics, this training course contained a short introduction to open source tools for developing Machine Learning based applications, in particular R (with a specific reference to the “mlr” and the “mlrMBO” packages) and Massive Online Analysis (MOA) for online machine learning algorithms

During the exercise sessions with R, a series of simple analytical tasks were developed consisting of loading a set of individual water consumption data acquired through a smart meter and analyzing the corresponding time series through clustering and regression models (in particular Support Vector Machine regression).

Next to the “mlr” tool, a set of additional R packages were briefly introduced such as “kernlab” and “e1071”, as well as “mlrMBO” relatively to the topic of Sequential Model Based Optimization for hyperparameter tuning and automatic algorithm configuration.

Further to the R suite, the training course also introduced the software MOA, which is a software environment providing several online machine learning algorithms. MOA can be used to evaluate the different pros and cons between a more traditional “batch” and a more innovative online learning schema.

While R is more oriented towards a “command-line” interaction, MOA offers the possibility to work directly through a GUI as well as to use its libraries, algorithms and functionalities from the code developed by the user.

The assignment for this summer school is focused on using one or both these suites in order to develop an validate a Data Science study on a set of individual water consumption data acquired – with a hourly resolution – through smart meters.

The basic examples developed during the exercise sessions is useful in order to deal with the summer school assignment.

*Schedule*

Summer school has three main components:

1. Training component: 31 May - 2 June 2017
2. Definition of tasks and development of work in groups: 2 June - 27 June 2017
3. Presentation of results and evaluation: 27/28 June 2017

# Learning objectives of the summer school assignment

As most of the summer school participants are either students of computer science or have that same background, the main focus of the assignment is software development.

In the assignment, these participants will learn how to work with R and/or MOA, and how to manipulate and analyse the available set of data in order to develop some analytical workflow which could be helpful for the implementation of data-driven decision support functionalities as well as predictive analytics functionalities.

Through this process they will also become familiar with generic concepts and procedures used in Data Science, in particular with respect to the topic of time series data analysis, both in the batch and online settings.

Several summer school participants who have background in computer science, statistical inference and machine learning, as well as other water related topics, will also benefit from the assignment because they will understand much better the potential of the Machine Learning tools and the role of suitable decision in the design of an analytical workflow.

Furthermore, they may also develop skills for manipulating and analysing data related to the monitoring and management of the water resource, which can be very useful for their future design and development practice.

Summarizing, the learning objectives of the summer school are:

* Understanding the main concepts and design choices in developing a machine learning based application by using real-world water-related data.
* Understanding the differences, as well as advantages and limitations, between batch and online machine learning in the time series data analysis domain.
* Enhancing software development skills through the development of code/scripts in R and MOA.
* Development of skills related to making design decisions and evaluating their impact on results of the data analysis task.

# Introduction to The Summer School Assignment

The main task in the summer school assignment is to develop a machine learning study by using individual water consumption data from smart meters.

The machine learning study should address one or more specific tasks of time series analysis, such as clustering and/or forecasting, as well as anomaly detection.

Specific design choices would be motivated and evaluation procedures (i.e. validity measures in the case of clustering and error measures in the case of forecasting) would be performed, reported and critically discussed.

The possibility to work with both batch and online machine learning tools would also allow for a comparison between these two paradigms in order to better understand – and quantify – their pro and cons.

Furthermore, the possibility to use the R package “mlrMBO” would also enable the possibility to identify the best hyperparameters tuning or algorithm configuration efficiently, through global optimization, in particular Sequential Model Based Optimization.

Some general guidelines about the development of the assignment follows:

* **Loading of the available set of data**

The set of data consists of 21 .csv files whose rows are related to a specific day and columns are the corresponding date and 24 numeric values, which are the hourly water consumption for that day. Each file is related to a specific smart meter (i.e. a customer) and the observation periods is about three months. In the case online machine learning is addressed, data have to be managed to simulate a streaming/online setting.

* **Definition of the time series data analysis task(s) to perform**

Several analyses are possible, such as clustering (among customers as well as independently for each customer), forecasting, anomaly detection. Each task can provide different insights and can be used to design and develop a possible decision support functionalities

* **Make design choices and perform analysis**

After the definition of the analytical task(s), some choices have to be taken, for instance according to representation of the data, preprocessing, distance/similarity to use, etc. design choices have to be motivated and then analysis has to be performed by using R and/or MOA. Any other library/suite can be used in case the attendant has already some specific skill. Some other possible suites, for instance are WEKA, KNIME, ScikitLearn. In any case tasks have to be compliant with the overall objectives of the summer school.

* **Summarize and discuss results**

At the end of the design and development activities, results should be presented and discussed, in particular with respect to the initial design choices as well as the choice of the tool to use.

# Data Science tools

The first of the two proposed tools is the language R. It can be downloaded and used with its traditional – and minimal GUI – or along with R-Studio, which is a more sophisticated IDE. Functionalities do not change, just the look-and-feel and the user experience.

The following figures show the R’s basic GUI and the R-Studio’s one, respectively.

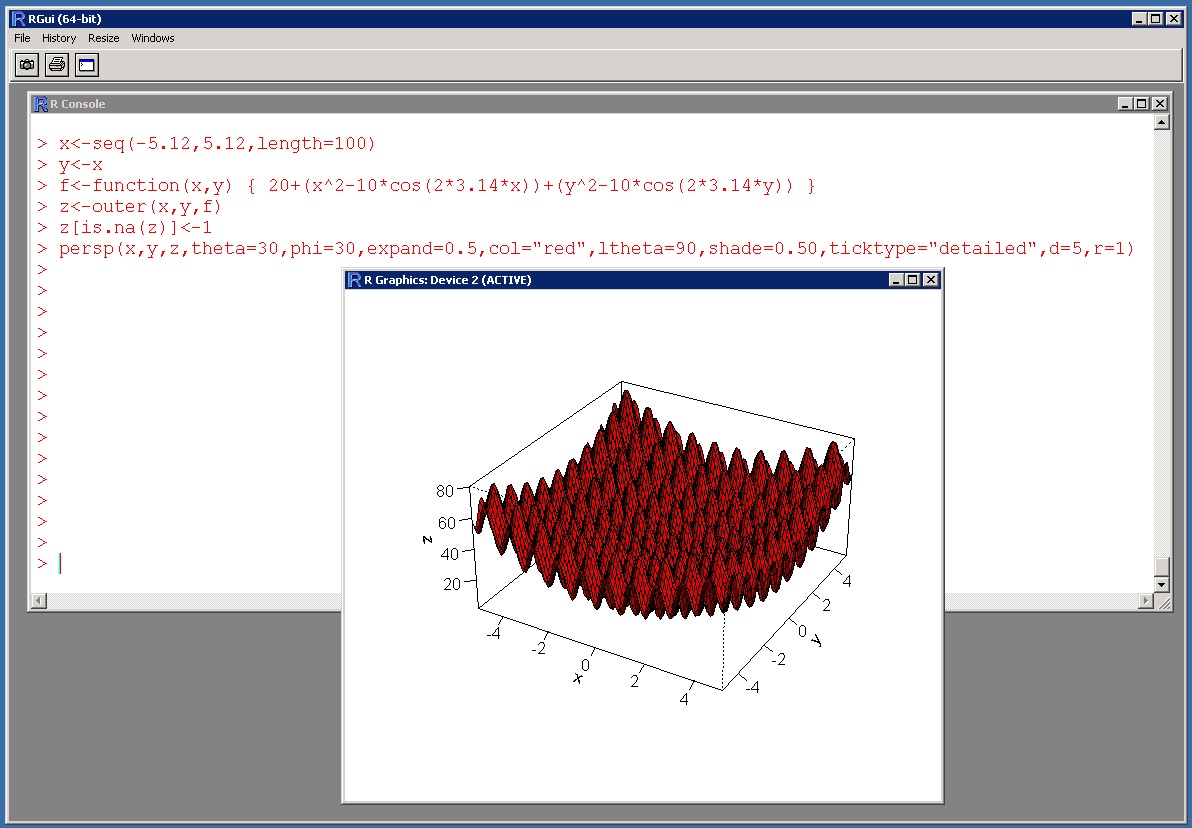
[](https://www.google.it/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjilqfByZ3UAhVCaVAKHUxxAoAQjRwIBw&url=https://jamesmccaffrey.wordpress.com/2015/02/02/graphing-rastrigins-function-using-the-r-language/&psig=AFQjCNGFxMofA34hTfnWHhUu1tJW3fPeVg&ust=1496438244402608)

Fig. 1 – The Graphical User Interface of R

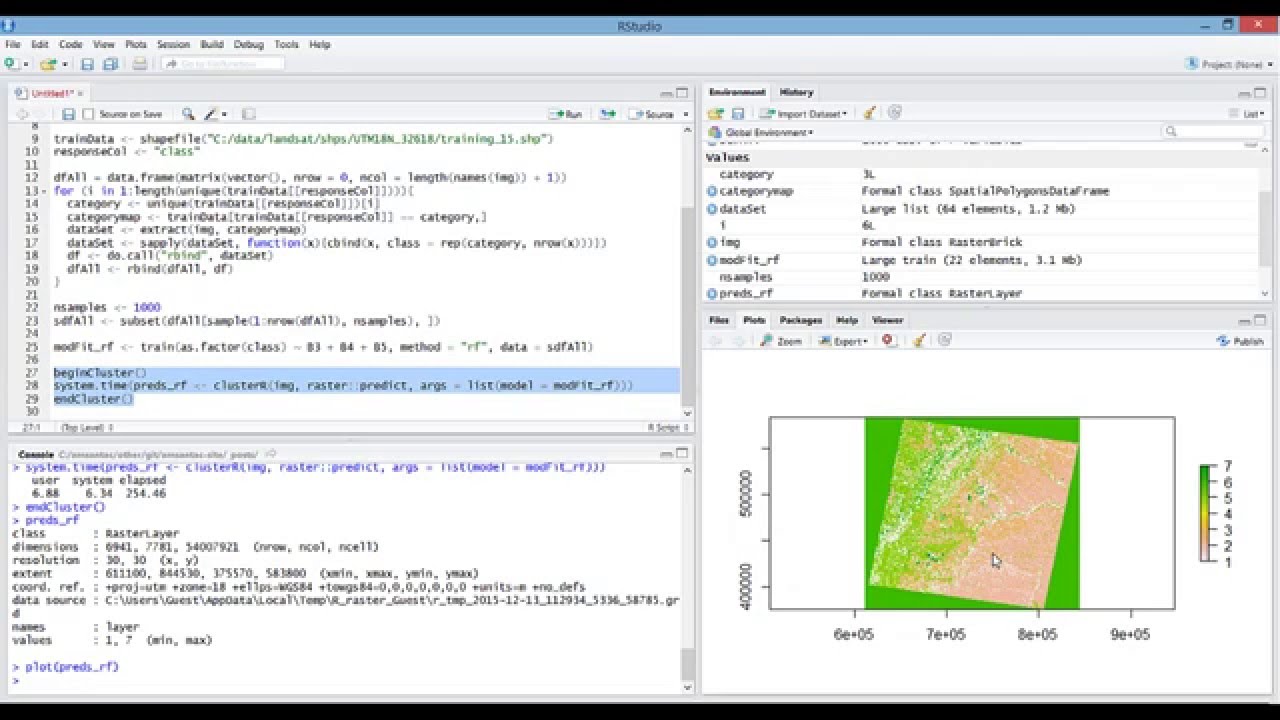
[](https://www.google.it/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwi8t4TKyZ3UAhVSLlAKHfX9AacQjRwIBw&url=https://www.youtube.com/watch?v%3Dfal4Jj81uMA&psig=AFQjCNGFxMofA34hTfnWHhUu1tJW3fPeVg&ust=1496438244402608)

Fig. 2 – The Graphical User Interface of R-Studio

Independently on the choice of the GUI, the packages “mlr” and “mlrMBO” can be installed and loaded to be used in the R environment.

The following two pictures are related to the documentation web site of these two packages.

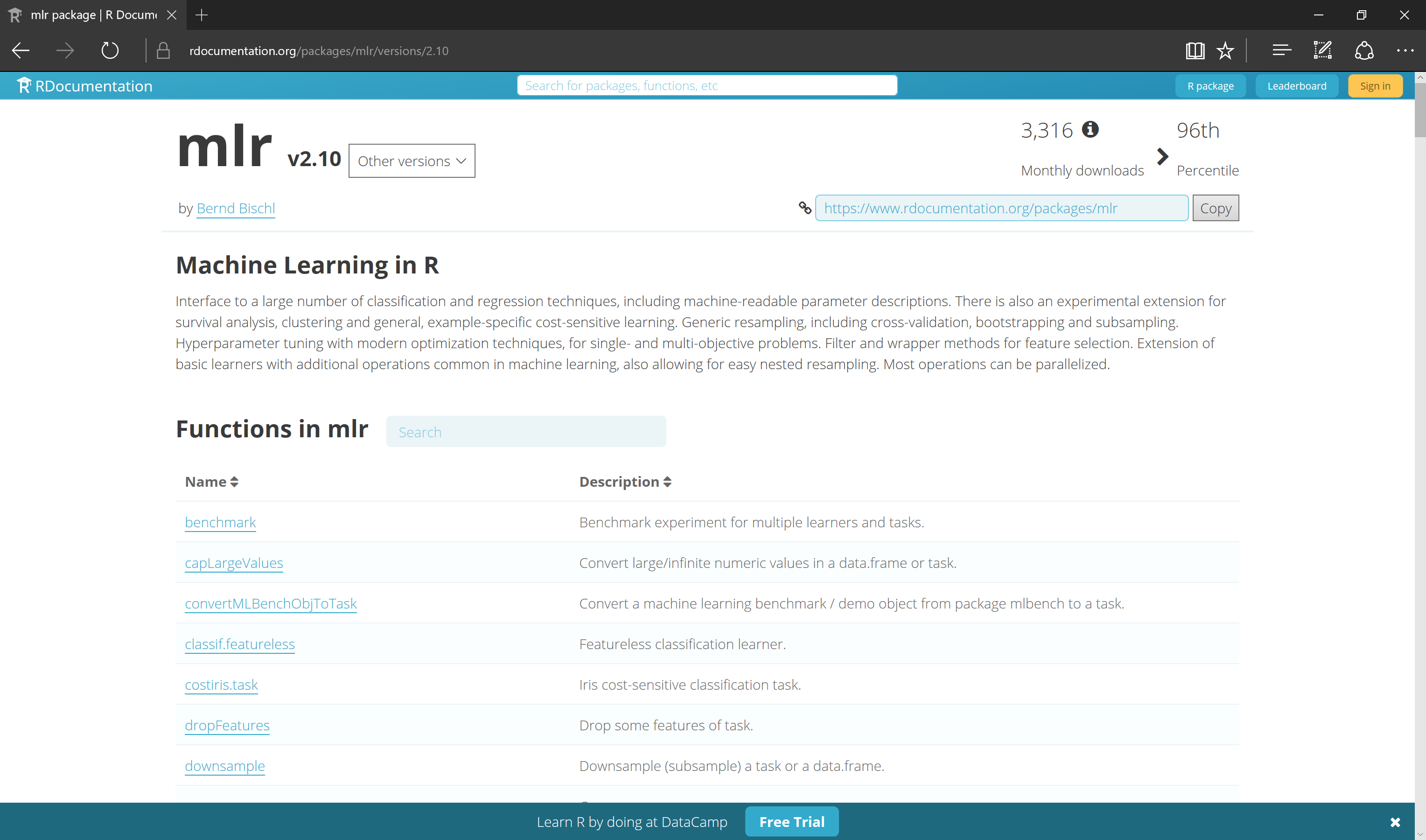


Fig. 3 – Documentation web site for “mlR package

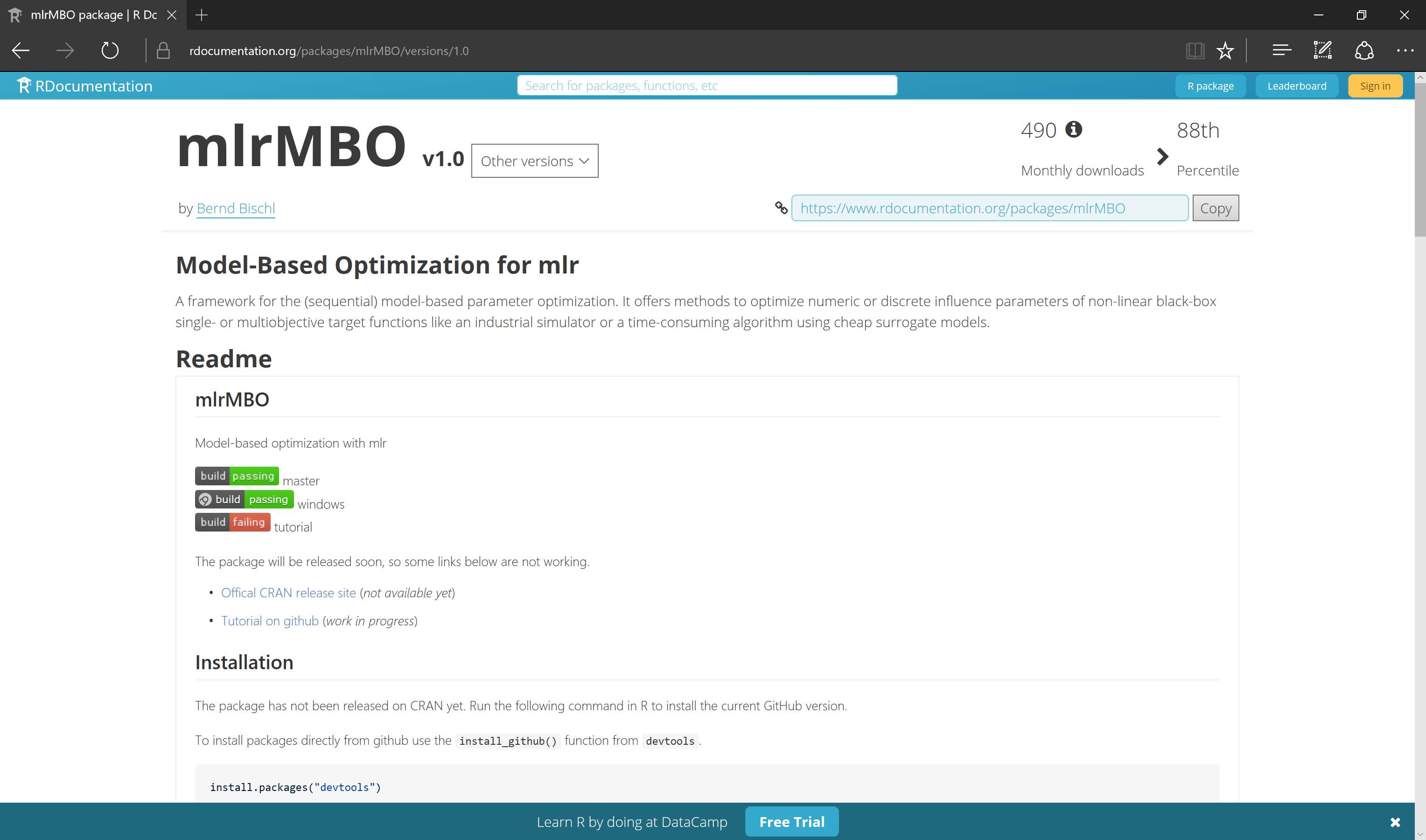


Fig. 4 – Documentation web site for “mlrMBO” package

Some other packages which are strongly suggested are:

* **skmeans** 🡪 a clustering package useful to perform clustering based on cosine similarity/distance
* **dtw** 🡪 a package implementing dynamic time warping
* **dtwclust** 🡪 a package to perform dtw-based clustering
* **e1071** 🡪 a package offering several Machine Learning algorithm, including artificial neural networks and support vector machines
* **kernlab** 🡪 a package offering kernel-based algorithms
* **Metrics** 🡪 a package providing easy computation of several error measures

With respect to the online machine learning tool, the suite MOA (Massive Online Analysis) of the WAIKATO University is suggested.

It provides a very user-friendly GUI with different tabs organized “per-task”: classification, regression, clustering, etc.

MOA is completely developed in Java, it is open source and it is really easy to be integrated in user-defined code.

Following figures are related to the web site of MOA and its GUI (i.e. comparison of two clustering algorithms on streaming data).

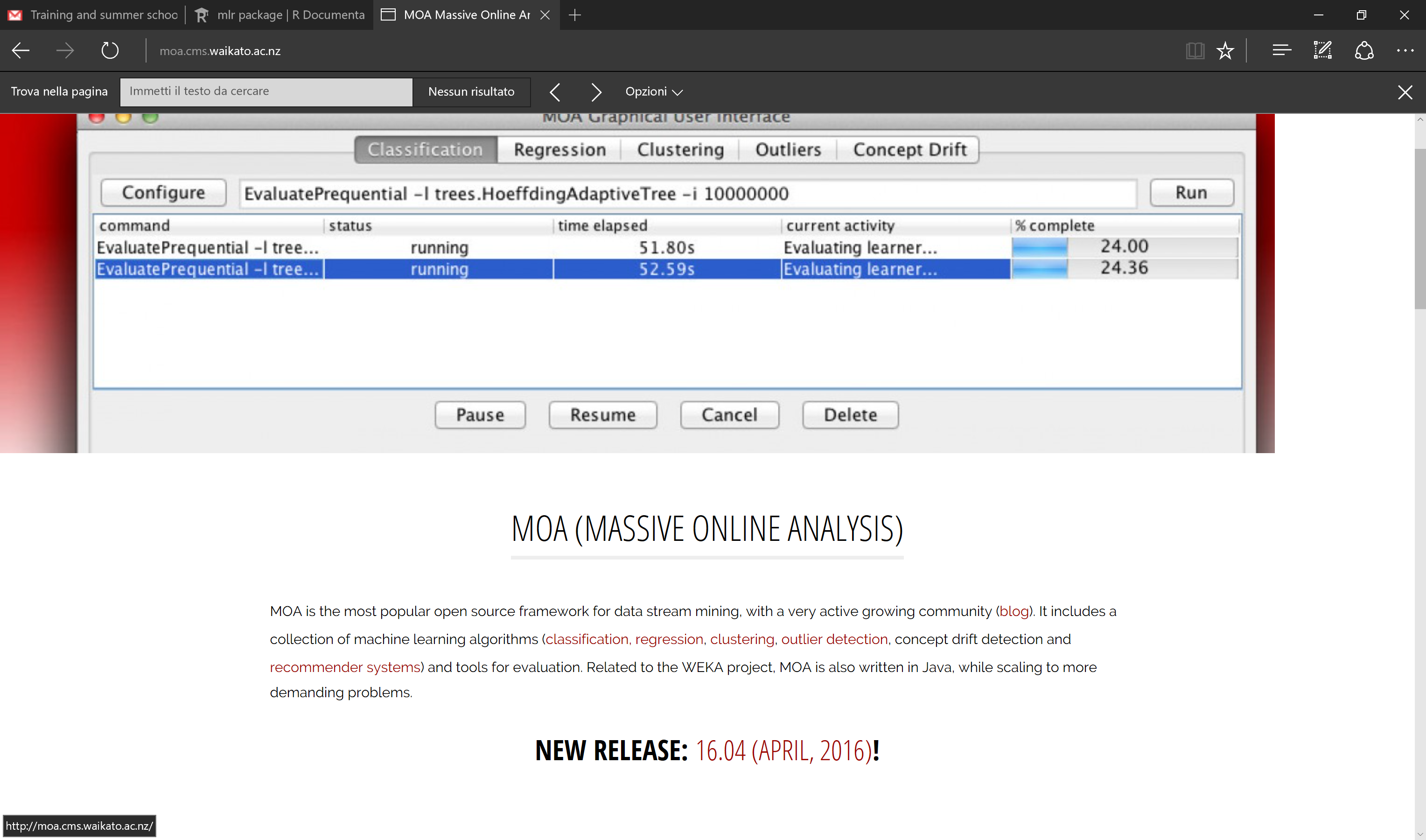


Fig. 5 – MOA’s web site

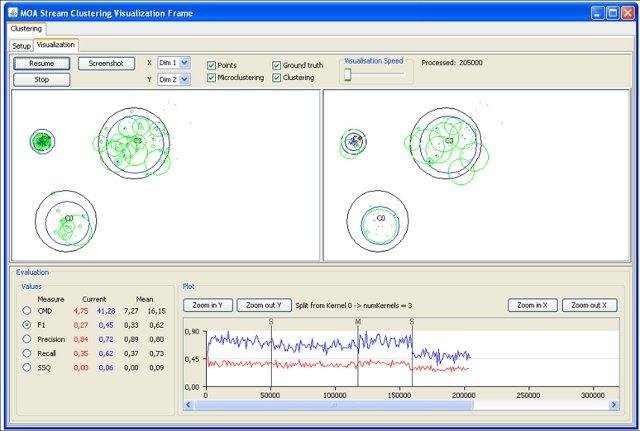
[](https://www.google.it/url?sa=i&rct=j&q=&esrc=s&source=imgres&cd=&cad=rja&uact=8&ved=0ahUKEwjE1PzhyZ3UAhVQU1AKHUOzDWgQjRwIBw&url=https://sourceforge.net/projects/moa-datastream/&psig=AFQjCNFvcGmFTtOyqqi6-1oWECfOR_m3GA&ust=1496438446846363)

Fig. 6 – MOA’s GUI: the comparison between two online clustering algorithms

# Suggestions and Guidelines

* The choice of the task to address as well as the tool to use is up to participant.
* It is not strictly required to address the assignment individually; however, in case the work is addressed by a team, the maxim number of components should be three.
* An overall team work is anyway strongly encouraged: all the participants may collaborate in order to select and perform specific tasks, combining results and approaches and discuss results.
* Basic instructions to load, manipulate and analyse data are provided in the material of the training session (PowerPoint slides).

# Most Relevant Resources

[1] R Project <https://www.r-project>.

[2] R package “mlr” <https://www.rdocumentation.org/packages/mlr/versions/2.10>

[3] R package “mlrMBO” <https://www.rdocumentation.org/packages/mlrMBO/versions/1.0>

[4] Massive Online Analysis (MOA) <http://moa.cms.waikato.ac.nz/>