# Data Analysis using Scala and Spark

[**Apache Spark™**](http://spark.apache.org) is a fast and general engine for large-scale data processing.

More information on Spark programming can be found [here](https://spark.apache.org/docs/0.9.1/scala-programming-guide.html).

The data set used has been taken from the [Linkage Data](https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns) (358 MB, consisting of 5.7 million records) of the UCI machine-learning repository.

The exercise is to identify which features in the data set reduce the number of False Positives in detecting a match/non-match.

To start a Spark shell, you need to access the hadoop cluster and type: spark-shell - -master yarn-client