# Machine Learning Course
# Project 1 Report

Thibaut Chamard - Priscille De Dumast - Mathilde Guillaumot

October 26, 2017

## Introduction

*This first Machine Learning project was meant to do exploratory data analysis, understand the given realwolrd dataset and its features as well as doing feature processing and engineering to clean the dataset and extract more meaningful information. In this report, we will explain the pre-processing choices we made and the methods we used to do our analysis on the given dataset and share our results in regard of the models used.*

## 1 Data Analysis

### 1.1 Exploring the data

When going through the different features of the train and test datasets, we can observe that 11 of them seems to hold a lot of -999 values which is meant to describe non-computable or meaningless values. When reading the Higgs boson machine learning challenge description of the features, we learn that 10 of these features are jet particules related. Depending on the number of jet particules created during the collision, some quantities are computable and some are not. The 11th feature, independent from jet particules but nonetheless containing a lot of -999 values, describes the estimated mass mH of the Higgs boson candidate. The documentation tells us that it may be undefined if the topology of the event is too far from the expected topology, thus explaining the meaningless values.

### 1.2 Pre-processing the data

**Estimated mass mH of the Higgs boson candidate feature:** In order to counteract the topology issue that leads to meaningless values in this features's column, we replace all meaningless values with the mean of the estimated mass values for all other values.

**Jet particules related features:** We observe that the uncomputable values of quantities for the jet particules related features depend on number of jet particules created during the collision as mentioned above. There are 3 different subsets of values of interest: 0, 1 and 2,3. According to these subsets of values, some features can be computed and some can not. Thus, we will drop some in regard of the number of jets values. Dealing with this issue implies the splitting of our dataset into 3 subsets in regard of the 3 subsets of values.