# MUD laboratories

## Session 2

Introduction

Through this part f the project we aimed to enhance the sequence tagging process. These improvements relayed in two essential parts: feature extraction and classifier. Therefore, we added features to the pool of the feature extractor and tried some models to achieve better classification rates.

Feature Extraction

Before adding any features the macro average percentage was of 55% as it was indicated in the project statement.

The first kind of features that we decided to add to the pool were some basic ones: presence or absence of upper cases, numbers, abbreviations, etc. Despite of resulting in an improvement of the macro average of success (both in CRF and NB), this advance was too mild since it oscillated in between 0.5% and 5% of success increasement, leading to a 60% of success.

The second kind of features resulted in more significant changes. These were inferred from the [Annotation Guidelines for DDI Corpus](https://hulat.inf.uc3m.es/DrugDDI/annotation_guidelines_ddi_corpus.pdf), through this document, the usual nomenclatures of drugs are explained. The abbreviations, numbers and words from usual chemistry formulas are exposed together with some examples, for most of the drugs included in this project. By using this information, we added some features, for example: numbers followed by commas and some dash are used in drugs (organic chemistry) like *1-methyl-4-phenyl-1,2,3,4-tetrahydropyridine*.

The greatest improvement was achieved by including vocabularies of words provided in the resources folder. This includes the *DrugBank.txt* and the *hsdb.txt* files, with already known drug names. We also added a custom drug list of 50 names. With these changes the macro average percentage scaled up to 73% with the CRF.

By looking at the confusion matrix, it became clear that the *drug\_n* entities were the ones offering resistance to the classifier, this was due to their low presence in the original dataset. Once again, we used the *Annotation Guidelines for DDI Corpus* to better understand these kinds of drugs. The annotation defined *drug\_n* as “*active substances altering drugs such as drugs, food, drinks or environmental chemicals*”. We also included some *drug\_n* vocabularies, but it didn´t impact classifier success. The last remaining option is to include all these entities present in the original dataset as a feature vocabulary, however we considered this out of the scope of this project.

Feature Extraction

The first classifier was the CRF (Conditional Random Field), was the one providing better results. And the one used for feature testing of the previous section. Some initial proves lead us to use this first classifier as the reference one.

Other used classifiers were the NB (Naïve Bayes), which obtained a final macro average of success of 58% over the testing set. Other considered model was the SGD (Stochastic Gradient Descent) providing a 58.7% of accuracy.

## Session 3

Introduction