# Task 1 Binary Classification with Sequence data

Objective: Train a classification models to make prediction on testing data set, using the data in the “Sequence model data.zip” file.

Note:

1. This is a sequence classification task, where the order of each feature matters. You could train a model without considering order as a baseline model, but must train a model addressing sequence because in real work, sequence analysis is part of the project.
2. Each row represents a training/testing sample containing a sequence, where the first element is “PPD\_197” and last element is “PPD\_0”.
3. All the sequences have been padding, which is the reason why lots of zeros show up in “PPD\_0”.
4. All the values in each entry is categorical variable. Imaging every value in the entry as an index of a word in natural language.
5. Y-variable is the first column called “LABEL”, in testing data, there is no label

What we expect,

1. A good prediction result, which we will compare with the hold out y variable in the testing data set
2. The process of how the prediction is made, including model comparison, hyper-parameter tuning, feature analysis, possibly, feature selection, create new features based on existing variables could be needed.
3. If you could use library such as Keras, Tensorflow to train a deep neural network (DNN) classifier, that will be a very good plus, even if neural networks might not be the best performed model.

You could use any tools available to you for this task. Ultimately, we will assess your work based on two criteria. 1. predictive accuracy on the test set using the PR-AUC metric, 2. model structure you finally applied, for example, we will consider how advance the model is, or if you could create additional meaningful features from the data we gave to you.

You should return to us the following:

- A 23,910 x 1 csv or txt file containing one prediction per line for each row in the test dataset.

- A brief report describing the techniques you used to obtain the predictions, that at least should include the following parts: a. why do you choose the model you use? b. your estimates of predictive performance on the test data set, c. some words telling us your understanding about the model you use.

- The code for building the model, or the saved model such as pickle file.