# Data Challenge - Explanation of Method

Nathan Rosenfeld

April 9, 2023

## 1 Data Importation

To construct the dataframe for the patient data, I did the following:

- Using a for-loop, the program read each patient file.
- From each file, it extracted:
  - Patient ID
  - Age
  - Unit
  - Gender
  - For each quantitative variable, it extracted:
    - The maximum value.
    - The minimum.
    - The mean.
    - The slope of the regression line (using `Hour` as the predictor.
  - Finally, it took the maximum recorded `Hour` and recorded it as `Duration`, representing the length of the patient's hospital stay.
- The patients whose `ID` values were listed in the `test_nolabel` file were separated into another dataframe.
- The remaining patient data was merged with the `Outcome` data.

## 2 Imputation of Missing Values

For missing predictor values, I used the iterative imputation method by chained equations. This was done separately for each data set.

## 3 Selecting the Model and Parameters

To select the model, I split the dataframe with known outcomes into two parts, so that I could test the accuracy of the model on the 30% of patient data with known outcome. Given the data, I chose the `Adaboost` classification model. To optimize the parameters, I did the following, using the `training` portion of the data:

- I separately imputed the missing values in each dataset.
- Using cross-validation, I looked for the the number of trees that got optimal `BER` and `AUC` values.

- I did likewise for search-depth.
- Then, I did the same for the learning-rate.
- Given the cross-validation results, I chose 100, 2, and 0.6 as the optimal values for `number of trees`, `tree-depth`, and `learning-rate` respectively.
- Using those values, I trained the model on the entire `training` portion (of the patients with known outcomes) and tested it on the `testing` portion.
- I measured the error rate on the `testing` portion.

# 4 Training and Applying the Final Model

After being satisfied with the results from the above steps, I did the following:

- I used the iterative imputation method on the entire training dataset (all patients with known outcomes).
- I trained the `Adaboost` model on this dataset, using the parameters I found to be optimal in the previous step.
- I used the iterative imputation on the `nolabels` dataset.
- I got the predicted probability and predicted outcome for this dataset from the model.
- The predictions were saved to the file `Predicted_outcomes.csv`