**CMPS 396X Project Phase 1**

**Team Members:**

* Hilal Breiss
* Johnny Bou Saab
* Othman Ibrahim
* Ahmad Nsouli
* Anis El Rabaa

Our team is composed of Hilal Breiss, Johnny Bou Saab, Othman Ibrahim, Ahmad Nsouli and Anis El Rabaa. We will be working on predicting the modes of action (MoA’s) of drugs based on their genetic data and their cell viability data. Modes of action of a drug represent on what proteins or substances the drug has effect, and the nature of the effect. For example, “acetylcholine inhibitor” is such a MoA. Our goal is to predict the modes of action of a new drug based on data we have about other drugs with known MoA’s. So basically, this is a probabilistic multi-label classification problem. When successful, our project could dramatically speed up the development of new medications, as the modes of action of a new drug could be conveniently inferred from those of already known drugs.

Our target customers would be medical and pharmaceutical researchers who need a convenient way of predicting whether a potential drug will indeed have the desired effect on humans, while also looking out for any potential side effects the drug may have. Our model will return for each drug the probability that it may have a specific MoA, so that these scientists can have concrete estimations and margins to base their decisions on.

Our competitors would be AI software companies working on a similar project for rival researchers. Time to market and accuracy of results are a key factor to our success over rivals.

In this project, we are going to get our data from Kaggle website under Mechanisms of Actions (MoA) competition. To get access to the data use this link: <https://www.kaggle.com/c/lish-moa/data>. Moreover, Broad Institute of MIT and Harvard, together with the Laboratory for Innovation Science at Harvard (LISH), presents this challenge with the goal of advancing drug development through improvements to MoA prediction algorithms. The data set folder contains 5 csv files: sample\_submission.csv (3.18 MB), test\_features.csv (24.93 MB), train\_features.csv (149.1 MB), train\_targets\_nonscored.csv (18.56 MB) and train\_targets\_scored.csv (9.66 MB). A total of 205.43 MB. The data combines gene expression and cell viability data, though it is made available when researchers in this fields do research and collect data about the drugs and their effects on the cells. Then data availability is research based. The data is well structured into training files and testing files. Moreover, data inside the files is organized into rows and columns and each row represents a unique drug given a unique id. Total columns in the dataset is 2569 column. In the training files the data is numeric. There are some columns in the testing file that are textual. no images, no audios and it is labeled. The data is easily readable and simple. The data is based on a new technology that measures simultaneously human cells’ responses to drugs in a pool of 100 different cell types. In addition, the data available made access to MoA annotations for more than 5,000 drugs in this dataset. Hence, since names or any specific attribute that indicate a human aren't mentioned, we shouldn't have restriction on using the available data.

Our data is complete in the sense that it does not contain errors or missing information. However, to make it more challenging, we’re planning to add gaussian to some of the data, around 15-20% of it. The data is not small, it can be representative; however, there is a lot of classes (between 200-300) to predict from, so it’s going to be a challenge to actually predict one class with higher confidence than others since the data is not big data and the number of features is around 800, while the number of train data is around 27k.

Our data measures simultaneously human cells’ responses to drugs in a pool of 100 different cell types. In addition, our data monitors 772 individual genes. Both of which reduce the bias and solve the problem of identifying ex-ante. However, one of biases that may affect our results is the use of only one new investigational tool to detect and measure a target molecule's presence or functional activity. The tool used is new that our data is only existing external datasets based on this tool.

Uncertainty quantification tries to determine how likely certain outcomes are if some aspects of the system are not exactly known. For example, Bayesian deep learning can estimate using a neural network, the output parameters of a certain probability distribution. Instead of finding one parameter vector, we approximate the posterior and sample many parameter vectors (this approximation tell us which parameter can be dropped) and then sample many parameter from this approximation, so that we can use it to get uncertainty. So mainly we infer distribution to each parameter so W can become a matrix of distribution for each parameter by applying a Gaussian approximation. As we previously mentioned this can be applied to our model where it will predict the probability of each drug having this specific MOA.

We also might integrate meta learning, the ability to learn, as much as possible in our project. Our data is not small per say, but meta learning is almost always a good idea, if not, not a bad one, to consider.