**Project 1 #Planning**

**Team**: Hongyan Chen, Shuyang Wu

**Date**: 2023-10-05

**Dataset: CHL5230 dataset**

**Research Question(s)**:

CVD dataset

1. Can various factors including sBP, dBP, FBS, HDL, LDL, and HBA1c be used to predict CVD risks?
2. Can we detect CVD risk using select variables such as medication usage, lifestyle and physical condition?
3. Among patients who have CVD risks, can various treatments or medicine usage affect CVD death?

diabetes dataset

1. Can we use date-related variables such as date of diagnosis, testing and diabetes onset to detect the trends in development of diabetes by time-series analysis?
2. Can we analyse the relationship between diabetes and CVD risks?

**Data Engineering Process:**

1. Data Preparation:

* Data cleaning and identity:Handling missing values, looking for influential outliers.performing EDA and data visualization, performing descriptive statistics, and feature scaling.
* Data exploring and visualization:performing EDA and data visualization(looking at distribution), perform descriptive statistics, and feature scaling.
* Normalization or standardization

1. Feature selection:

Based on EDA and descriptive statistics summary, select significant variables which strongly related with response variable for the following model fitting.

1. Model fitting:

* KNN for classification
* logistic regression model and generalized linear regression model for prediction
* survival model for time between treatments
* time-series model for trends detection

1. model evaluation and interpretation:

evaluating model, analyzing results, and identify relationship and associations by Maximum Likelihood Estimation or other statistical model.