### The Cutting Edge: Machine Learning for Predicting Diabetes

### Team 2A - Ruicheng Li, Chun-Hui Tsai, Lalitha Bhupalam, Hilary Nguyen, Abhinav Samavedula

**Introduction:**

Diabetes prediction models are crucial in the medical field for providing early warnings to those individuals at risk. According to the CDC, diabetes affects over 34.2 million Americans, with over 88 million pre-diabetics unaware of their condition. We will use a dataset from Kaggle (<https://shorturl.at/hjwB9>), which includes demographic information like gender and age, as well as their lifestyle and health indicators such as BMI and smoking. Our goal is to identify patterns behind all this data, enabling us to signify the potential risk of developing diabetes and allow for early precautions.

**Question:**

For our analysis, we refer to a dataset from Kaggle, which allows us to focus on the following questions:

1. Which factors most strongly influence the risk of diabetes?
2. What is the prevalence of diabetes and pre-diabetes among different demographic groups in the dataset?
3. How can we construct a risk profile for diabetes based on the lifestyle factors and lab test results data, and how can we build a machine learning model based on these?

**Potential Analytical Approach:**

Our approach starts by combining data, data cleaning and preparation, and then we do an exploratory analysis to identify the potential diabetes predictors. We would build and test various machine learning models to determine their predictive accuracy. Next, we emphasize the most predictive risk factors and propose a concise set of survey questions to effectively predict diabetes risk. The report will document all of our methods and discuss the practical applications for screening and preventive healthcare. If possible, we will find more data through web scraping.