**CS-GY 6513 Big Data**

PROJECT PROPOSAL

**Team members**

* Junyi He (jh7687)
* Sai Kiran Uriti (su2092)
* Tejasvi Dasari (td2317)
* Liu Yang (ly2062)]

**Introduction**

With the development of the economy and society and the improvement of people's consumption level, citizens are willing to spend more money on food. People prefer a better dining experience, such as dining environment, food quality, service, in a limited amount of money. Therefore, restaurant recommendations become increasingly vital. At the same time, the owner of the restaurant can also learn from it and better improve their own restaurant.

**Dataset description**

In recent years, Yelp, the largest comment website in the United States, has disclosed its internal data set. The dataset includes 4.7 million user evaluations, 150000 merchant information, 200000 pictures and 12 metropolises.

In addition, it also covers 1 million tips of 1.1 million users, and more than 1.2 million business attributes (such as business hours, whether there is a parking lot, whether it can be reserved, and the environment), and the total number of users who sign in at each business over time. This dataset provides a good sample for implementing restaurant recommendations.

**Project description**

This project aims to extract valuable parts and build the user's mental model by processing and analyzing the Yelp open dataset, especially the evaluation information, for example, to understand which part of the restaurant attracts users most, food quality or service.

Furthermore, according to the psychological model and evaluation information, we do restaurant recommendations to help users’ decision making and give advice to restaurant owners for their improvement.

**Project challenge**

The Yelp dataset has several GB, which is a good entry point for academic purposes. An element that can generate economic returns is the ability of the project to operate stably in a real production environment, where there is no known upper bound on the size of business attributes, so we must handle the situation when the data is unlimited. However, it is impossible for a single computer to store and process data of arbitrary size, so big data technologies are crucial to make our project efficient and scalable in a production environment. To achieve our goal, we must implement a big data-based computing program to run our algorithms for a fast recommendation on any scale, a distributed storage for our massive statistical insights, and an orchestration system to automate the cluster deployment and scaling.