# Lab 5: Final Project
# Project Proposal

**Bharat Suri ( suri.40@osu.edu )**
**Nithin Senthil Kumar ( senthilkumar.16@osu.edu )**

## Yelp Dataset Challenge

## Task

The dataset we have chosen is from the Yelp Dataset Challenge. Our task will be to use the public Yelp dataset which is available in this challenge and study the businesses listed on Yelp as part of this subset. The main objective will be to draw visualizations of businesses based on their ratings and plot the locations on a map, giving the user a better idea about the options they have in the vicinity. Using the features, we plan to perform location-based clustering on the businesses. Using these predictions from the system, we plan to design the pipeline in a way such that it can also be incorporated into a larger recommender system which would make use of these outputs as features and also use information from the user and review files to give the user location-based recommendations. We will also make use of some basic Bag-of-words techniques from Natural Language Processing for the reviews dataset.

## Dataset

As part of the challenge, Yelp released the Yelp Open Dataset which is a subset of their businesses, reviews, and user data. This dataset is available here.

Individual files available as part of the downloaded dataset are:
Business.json
Review.json
User.json
Checkin.json
Tip.json

## Data Mining Approach and Software

The approach we plan to take will be cluster analysis. We plan to study the dataset first, and after performing a thorough exploratory data analysis, we will utilize different feature subsets to draw clusters. These will be clusters based on ratings, clusters based on location, clusters based on food trends (for restaurants), clusters based on business category. Through this project, we think it will be beneficial in terms of learning more about complex visualizations in exploratory data analysis as well as learning about advanced clustering algorithms. We plan to use the same software we have been using in previous labs, i.e. Scikit-Learn, Seaborn, Pandas, Numpy, Jupyter.

Our main focus will be studying the data, engineering features for businesses, users, as well as reviews which can be used in the cluster analysis, and finally using a clustering technique to show the clusters based on location on a local map.