**Final Project**

Group members: *Gregg Maloy, Jacob Silver and Justin Williams*

## **Objective:**

Utilize Yelp API data to visualize the relationship between users and businesses.

**Updated Proposed Analysis**

1. Exploratory data analysis (EDA)
2. Calculate unweighted centrality measures of a bipartite network where reviewers/users representing nodes and edges represented businesses reviewed.
3. Graph the above network and utilize the island method to look for power users who may or may not be connected via the same restaurant.
4. Text analysis to find the most common sentiment for each user, restaurant and star rating.
5. Develop a machine learning algorithm to predict restaurant review rating based on text analysis of reviews.
6. Determine if power user reviews more accurately predict user star rating.

**Files**

This analysis was separated into three Jupiter notebooks:

1. Data acquisition and exploratory analysis  
   *(file name: 01\_data\_acquisition\_eda.ipynb)*
2. Network graphing and centrality measures   
   *(file name: 02\_Network\_Graphing\_and\_Centrality\_Measures.ipynb)*
3. Yelp review text modeling  
   *(file name: 03\_yelp\_review\_text\_modelling.ipynb)*

**High Level Overview**

The below provides a summary of key takeaways from each part followed by a conclusion.

*Data Acquisition & Exploratory Data Analysis (EDA)*

* Yelp Fusion API was utilized to secure the data
* Yelp API limits significantly reduced the size of the dataset
* Two data frames were created with data limited to reviews and business of restaurants in any of the NYC five boroughs

Businesses Data Frame EDA

11 variables including business\_id, city, average user rating

647 unique businesses

Average overall business rating determined to be 4.27

Review Data Frame EDA

5 variables including business\_id, review\_user\_id, review\_rating, review\_text

1,941 unique reviews

The dataset was significantly skewed in terms of favorable reviews with ~1,200 of the 1,941 reviews being rated 5 stars.

Some users were found to review restaurants in multiple cities

*Network graphing and centrality measures*

* Unweighted centrality measures were calculated on a bipartite network represented by nodes as users and businesses as edges.
* Degree centrality was decided upon as being the most appropriate centrality measure.
* 88.4% of users have a degree measurement of 2. This means a significant majority of reviewers (88.4%, count=1479 reviewers) have only reviewed two restaurants. The data dataset was significantly skewed to users with a degree measure of 2.
* The island method was utilized. It was determined that a degree centrality of 8 displayed a group of power users. However, due to low sample size, different water levels may be appropriate for future analysis of larger datasets.

*Review Text Modeling*

* A model was built to predict star rating based on the text of a review.
* Multinomial Naive Bayes model and XGBoost classifiers were utilized.
* XGBoost was determined to have greater accuracy in predicting 5 star reviews (.71 vs. .67).
* the words "comfort" and "welcome" appeared with greater frequency in 5 star reviews
* References to celebrations, parties and birthdays, as indicated by the stem "celebr", were highly predictive of 5-star reviews.
* Delivery, as indicated by the stem “deliveri”, appears to be a highly polarizing and "reviewable" experience, so if restaurants offer it, they should ensure the experience is as convenient and high quality as possible.

**Conclusion**

1. Due to the limited dataset/API limits and time constraints, the final part of the purposed analysis ‘Determine if power user reviews more accurately predict user star rating’ was not performed.
2. Next steps would be to run this analysis on a larger dataset and perform the ‘Determine if power user reviews more accurately predict user star rating’ which would essentially marry NLP and network analysis in a single analysis.