# **CS-5630 / CS-6630 Project - YelpHelper - Process Book**

Sateesh Tata - u0942976

Vairavan Sivaraman - u0942570

***Overview and Motivation***: This project is basically a simple visual representation of the yelp sample data set provided by YELP for their Yelp Dataset Challenge competition for 2016. The dataset is really vast with a numerous possible attributes and factors that could be considered to analyze and visualize the data and trends. We were mainly motivated by the variety of factors that could be considered to actually visualize such a large data. Although this might not be a standard bread and butter dataset to directly throw a bunch of numbers and statistics to the user to plug them in and use in the visualizations, we believe that the richness and vastness of the data would provide us with a challenging task to analyze, understand the visualize the data that might not be pretty straight in the most obvious ways. Some of the main ideas that were suggested with the data by YELP in their official challenge are the following,

* **Location Mining and Urban Planning:** How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?
* **Seasonal Trends:** What about seasonal effects: Are HVAC contractors being reviewed just at onset of winter, and manicure salons at onset of summer? Are there more reviews for sports bars on major game days and if so, could you predict that?
* **Infer Categories:** Do you see any non-intuitive correlations between business categories e.g., how many karaoke bars also offer Korean food, and vice versa? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text?
* **Natural Language Processing (NLP):** How well can you guess a review's rating from its text alone? What are the most common positive and negative words used in our reviews? Are Yelpers a sarcastic bunch? And what kinds of correlations do you see between tips and reviews: could you extract tips from reviews?
* **Changepoints and Events:** Can you detect when things change suddenly (i.e. a business coming under new management)? Can you see when a city starts going nuts over cronuts?
* **Social Graph Mining:** Can you figure out who the trend setters are and who found the best waffle joint before waffles were cool? How much influence does my social circle have on my business choices and my ratings?

These problems might not all be standard visualization techniques or provide straightforward data to plug in but are definitely interesting in many ways. We would like to push our understanding of our concepts and visualization techniques to bring out ways to be able to better show such subtle changes and properties and relations between various factors and trends.

***Related Work***: The websites which have quite a lot of user base always come up with some really nice and innovative approaches to visualize the information on their websites. This has primarily inspired us to pick up this particular set of data and come up with better and more cleaner visualizations. So we picked the YELP dataset challenge so that we could understand the data and represent it in such a way so that an end user could clearly see the data representation and could understand how the businesses are doing in and around a particular area and also some user related data.

With this kind of data and many of the visualizations that were discussed from the class, we think we can better represent the data. For example, the geo location of the maps and also the bar and area graphs with brushes included are a good way to analyze large amounts of data location wise, over a period of time or over a big group. Also, we feel that this could be a good exercise to kind of implement the standard visualization techniques on some real world data sets which are not just tailored for visualization demonstrations but are actually the kind of data we expect from an actual application.

***Questions***: We have considered a lot of possible questions which we would like to answer as a part of our analysis and visualization of the data. The data we have is a wholesome representation of business data from the application’s perspective of the various businesses in a lot of areas along with a lot of users who are interacting with these businesses and are rating them. We primarily can think of two kinds of analysis that could be drawn from the given dataset.

One approach is to show the users the trends of businesses in a particular area over a period of time or the overall ratings in every category. The data given is per business with a lot of factors to consider and analyse how good is a business doing based on user ratings and also how each business fares with a bunch of other businesses of the same category with respect to the overall ratings.

The other approach is to view the customer interactions with the businesses based on the overall ratings of every single customer and the number of reviews that are given. When both the business and customer interactions are taken into consideration, there are several other possible trends that could be visualized using the data.

***Data***: The data we have for the given task is massive. The following are the rough estimates of various parameters of data size we are going to deal with during the course of this project.

* **1.6M** reviews and **500K** tips by **366K** users for **61K** businesses.
* **481K** business attributes, e.g., hours, parking availability, ambience.
* Social network of **366K** users for a total of **2.9M** social edges.
* Aggregated check-ins over time for each of the **61K** businesses.

There are primarily 5 kinds of json files present in the entire dataset namely, *yelp\_academic\_dataset\_business.json, yelp\_academic\_dataset\_checkin.json, yelp\_academic\_dataset\_review.json, yelp\_academic\_dataset\_tip.json and yelp\_academic\_dataset\_user.json.*

These JSONs are quite large and so we have decided to process the data and clean it up in multiple steps so that the data would finally be reduced to a small enough size and also the format of the data would be in such a way that there would be not a lot of effort for the application to process and filter the data in the runtime. We have initially parsed the JSON objects and removed the set of features that we are not going to be using for our visualizations. This has drastically reduced the size of the files that are to be loaded (more details to be added on finalizing all of our data formats). In the second step, we have parsed the data from the reduced JSON files in order to change the format of the data so that it is readily available for the visualizations to use. This would reduce the effort and decrease the time for the data to be loaded in the visualizations and for faster updates. The data count is so huge that every simple modification takes us a few hours to convert using a standard python script.

Lastly, we are planning to organize the data from the formatted JSONs in such a way that the filtering of the data would be really fast so that the visualizations update with very less delay.

***Exploratory Data Analysis***: We initially had the following ideas for visualization in mind in order to represent the visual analysis of the Businesses in an area across a bunch of categories.

* *Map* - This visualization would plot the locations of all the businesses and also the filtered data based on the categories, regions and other factors on the map. This would help us visualize the trend of locations on certain kinds of businesses and also how well a particular category of businesses are doing in a certain area. This helps the user to quickly select and compare between multiple local businesses and view their ratings history and overall scores over a bunch of businesses in an area of interest. This task would be challenging because the sheer number of businesses that are covered as a part of YELP data are huge and could cause problems and cluttering in the data on the map if not filtered properly.
* *Bubble/Pie Diagram* - This visualization is to represent the percentage weightage of each category among the list of all businesses. This representation shows how dense is a particular category of a business over a certain region and count (and probably more data). This gives the user an overall understanding of how and area could influence any particular category of business and also how dense it is in general in terms of sheer number.
* *AreaChart* - This visualization helps us to analyse how the ratings for a certain business or a bunch of businesses have changed over a period of time. With the help of this the user can see how the ratings for that business have changed over a certain interval of time or till it went out of business. This we believe would help one to understand the general trends over a bunch of businesses in the same or different categories over time.
* BarGraph *-* This visualization helps to analyse the overall ratings of a business and also specifically the number of ratings across all the stars. This would help the user to see how well the business is doing not just based on the number of ratings in general but also based on the number of ratings per every single category. This graph would plot a lot of data and in order to be able to view the data clearly, we are implementing a brush on top of this so that the user gets to select a subsection of data based on his interest.
* *BipartiteGraph* - This visualization helps the user to better understand the overall data that is being represented by the visualizations. This represents a bipartite graph relation between the areas and different categories of businesses so that a user can easily check which area has what categories of businesses in majority and also which category of business is more well established in which area. This provides the user with better understanding and helps him better decide what kind of filters he needs to use over a certain area or for a particular category of business.

These are the primary ideas of various visualizations to represent the data we have to analyse the trends in local businesses from the YELP data. We may come up with more interesting and different ways of handling the data in time and also by considering other factors which we haven’t so far from the huge dataset.

***Design Evolution***: (What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?) The list of above visualizations are a part of our initial phase to represent the data. We have encountered a lot of problems while trying to represent the data in these visual charts. So there are eventually a few places where we had to change our approach a little and compromise on certain things due to the limited time resources to work on such vast data. Since there is no straight method to represent this data, there have been a few changes in the designs and representations as a whole and a few ideas that we stumbled upon later which we thought were more suitable for such data representation. More on this section to come up...

***Implementation***:(Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.) This section would be added soon once we have finalized all the designs and implemented the changes.

***Evaluation***: (What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?)