Process Book

Yelp Data Challenge

Marissa Stanvick and Maxwell Lloyd

CS171 Final Project

Spring 2015

[2) Visualization Idea Evolution: 3](#_Toc417056295)

[2.1) Visualization Idea 1: Visualization of Ice and Fire 3](#_Toc417056296)

[2.2) Visualization Idea 2: Yelp Data Challenge 3](#_Toc417056297)

[3) Visualization Overview and Motivation 3](#_Toc417056298)

[4) Data 3](#_Toc417056299)

[5) Exploratory Data Analysis 3](#_Toc417056300)

[5.1) Map Layout: 3](#_Toc417056301)

[5.2) Force Layout 4](#_Toc417056302)

[5.3) Area Graph 5](#_Toc417056303)

[5.4) Word Cloud 6](#_Toc417056304)

[6) Implementation Challenges: 7](#_Toc417056305)

[6.1) Mapping 7](#_Toc417056306)

[6.2) Data Wrangling 7](#_Toc417056307)

[7) Evaluation 7](#_Toc417056308)

[8) Appendix 7](#_Toc417056309)

[8.1) Visualization of Ice and Fire Initial Proposal & Sketches 7](#_Toc417056310)

[8.2) Yelp Data Challenge 1 Initial Proposal & Sketches 7](#_Toc417056311)

# Visualization Idea Evolution:

## Visualization Idea 1: Visualization of Ice and Fire

Our initial design proposal and sketches (Appendix 8.1) detailed a “Game of Thrones” based visualization. The Game of Thrones books and television shows are a complex story following dozens of characters in tens of locations. The complexity lends itself to confusion for the readers and viewers. To help clarify where a character spends their time, we aimed to create a map based visualization. One could select characters and move through time in the book or television show to see where they moved and how much of their scenes were spent in a specific location. The user could select more than one character to compare the path each has taken. The initial proposal in Appendix 1 has more detail on the initial visualization.

However, based on the feedback from our TF, we decided to move to a different plan. The major problem for this proposal was that the data would be difficult to acquire. Neither Max nor I have experience with Python or data scraping. Additionally the format of data isn’t straightforward to obtain. For example, just because a location is mentioned in a recap, does not mean any of the characters were actually in the location. We tried to contact the creator of <http://quartermaester.info/> to find out more about their data set, but we were not successful. Without data, our project would not have made it far! Using a fictional map would have also added a an extra layer of complexity to the mapping especially because we could not utilize D3 built in mapping functionalities. Because of these difficulties, we generated a new idea.

## Visualization Idea 2: Yelp Data Challenge

For our new project proposal (Appendix 8.2), we focused on an idea with easily accessible data. From Section 7, we learned of the Yelp academic data set. Using this data set we wanted to generate an interactive map that would help users decide where to eat. You could filter the restaurants based on location. Within a specific location, the restaurants could be filtered on restaurant category. Each restaurant would be represented on the map by symbols that could be colored by # of reviews or # of stars. Finally, if the user clicks restaurants, they could compare them based on the aggregated review data in a Word Cloud. This proposal can be found in Appendix 8.2 for more detail. Based on feedback from the TF, we amended our idea to be less application based and more data focused. These ideas will be explored in the rest of the process book.

# Visualization Overview and Motivation

Yelp is a popular website used to rate and review business of all types. The website stores a very large amount of data ranging from business information to review details to user profiles. All of this data could be explored together to garner insights and trends about businesses.

# Data

Our data set is from <https://www.yelp.com/academic_dataset>. Yelp provided data for academic purposes, that are a compilation of business, review, and user data for businesses near 30 different schools. The data set includes business near Harvard University and MIT. The data is provided as a tar file which contained all of the business, user, and review objects in a single json file. We then separated the business, user, and review data into three files for easier processing, both by us, and chrome.

# Exploratory Data Analysis

## Map Layout:

### Design:

The map layout (Figure 1) is the first step in the Yelp data visualization. On the map of Cambridge, we will plot a heat map of all the businesses in the data set. The user will have the option of choosing whether the heat map is populated by the number of reviews, the cost, or the rating of the business via a dropdown menu. The user can brush a certain area of the map to filter the force layout.

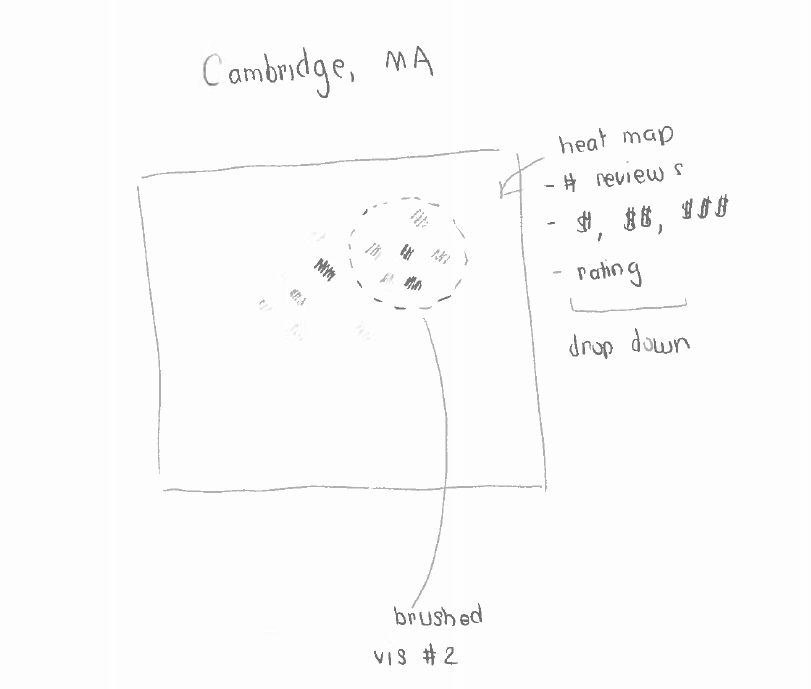


Figure 1: Map Layout

### Questions:

#### Where are businesses located near Harvard and MIT?

#### Are there clusters of businesses with a high number of reviews in a certain area?

#### Are similarly priced businesses located in the same area?

#### Are there differences in the rating of business between the Harvard and MIT areas? Price?

## Force Layout

### Design: The force layout (Figure 2) will have nodes represented by each business. The user can select whether to group the nodes by category. The user can select whether to color the nodes by # of reviews, # of stars, or category. A user can select a business to filter the area graph.

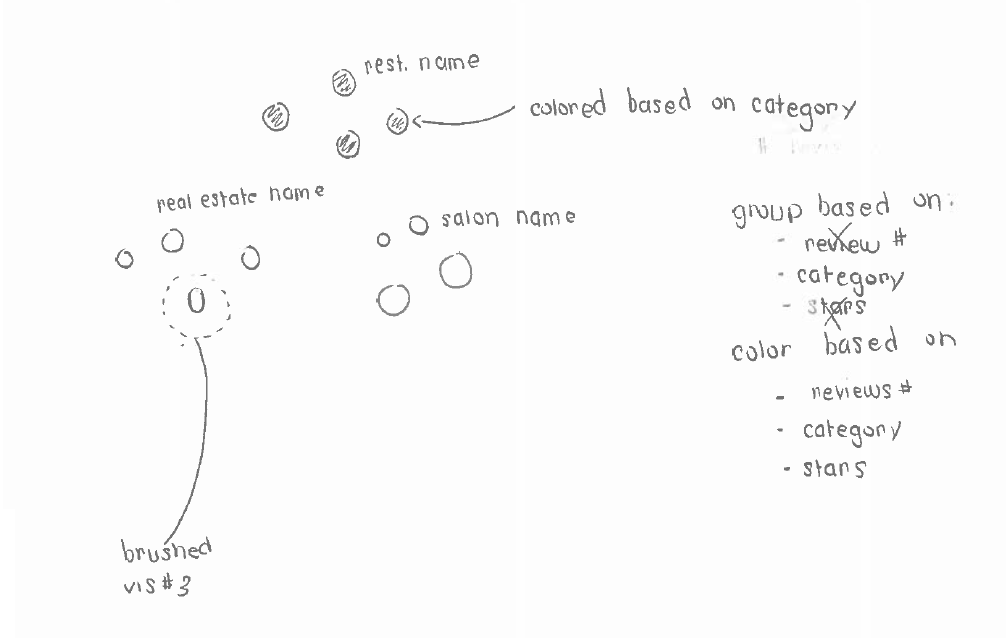


Figure 2: Force Layout

### Questions to Answer:

#### Are there a large number of restaurants in a certain brushed area? Are there a large number of bars near restaurants? What other types of businesses are in the brushed area?

#### Are the businesses with a lot of reviews all restaurants?

#### Are all business with a high # of stars located in the same area?

## Area Graph

### Design: The area graph (Figure 3) is specific to the selected business from the force layout. The area graph will plot the # of reviews over time. The user can brush a certain time period and receive the average # of stars and the total # of reviews in that time period.

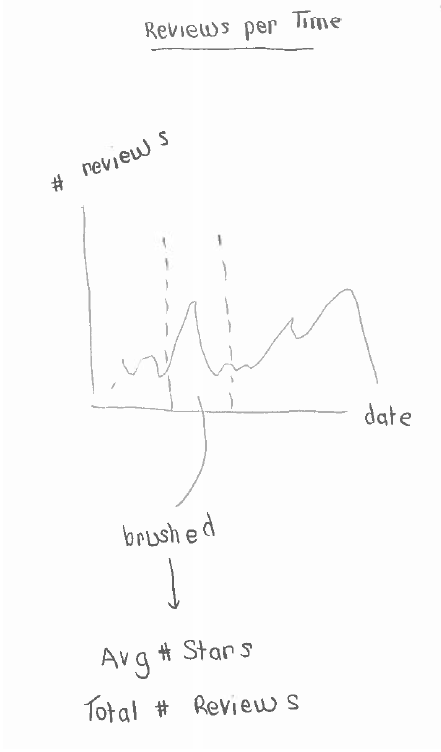


Figure 3: Area Graph

### Questions

#### What do the # of votes look like over time?

#### What is the average # of stars given in the summer?

## Word Cloud

### Design: The word cloud will be generated when a business is selected in the force layout. The word cloud will show the most representative words represented in the reviews for that business. Additionally, the words could be colored based on the average rating of reviews that included that word.

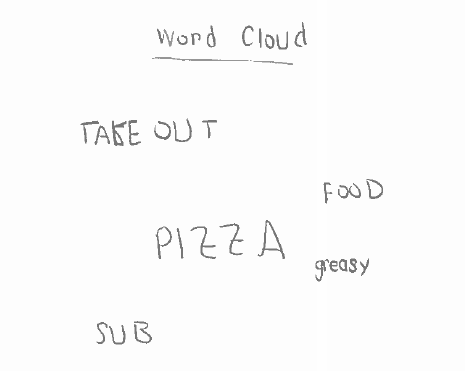


Figure 4: Word Cloud

### Questions:

#### What should I order from this restaurant?

#### Who is the best manicurist at this beauty salon?

#### What are the common words associated with very poor, or very good reviews?

#### Does this bar show a lot of games for a certain sports team?

# Implementation & Challenges:

## Mapping

The first map layout will be based on a map of Cambridge. The Cambridge GitHub webpage (<https://github.com/cambridgegis/cambridgegis_data>) provides topojson files of many different aspects of the city. We would like to use the boundary topojson to generate a map of the neighborhoods in Cambridge. To make a map, we followed Mike Bostock’s example (<http://bost.ocks.org/mike/map/>). We successfully generated a map of the US using a SHP file from the US census (<https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html>). Currently, we are in the process of generating a map of Cambridge. It is only generating a square and we are trying to troubleshoot this.

## Data Wrangling

The Yelp academic data set included all of the businesses, reviews, and users in one JSON file. We separated the data into three separate JSON files. This data was loaded into our code using queue, similar to HW3. First, we filtered the data to include only businesses associated with Harvard and MIT. Second, we filtered the reviews to only included reviews that were for businesses associated with Harvard and MIT. This data is what will be loaded into our visualizations

## Milestone 1 Timeline Update

Because we revised our idea twice, we are behind schedule for generating our visualization. Below is an updated schedule:

Week 1-2: April 4-17 – Done

* Acquired data
* Processed and cleaned up data
* Created initial SVG elements and JS files
* Outlined JS functions and general functions to do
* Sketched out visualizations

Week 3: April 18 – 24

* Generate Map Layout
* Generate Force Layout
* Generate Area Graph
* Interactivity:
  + - Brush map layout to update force layout
    - Click filter to update map layout and force layout
    - Click node grouping to update force layout
    - Click node colors to update force layout
    - Click node to update area layout and word cloud
    - Brush area layout to update data aggregation outputs

Week 4: April 25 - May 2

* Generate Word Cloud

Final Push: May 3 – May 5

# Evaluation

To be completed when project is complete

# Appendix

## Visualization of Ice and Fire Initial Proposal & Sketches

## Yelp Data Challenge 1 Initial Proposal & Sketches