Process Book

Yelp Data Challenge

Marissa Stanvick and Maxwell Lloyd

CS171 Final Project

Spring 2015

[2) Visualization Idea Evolution: 3](#_Toc418529054)

[2.1) Visualization Idea 1: Visualization of Ice and Fire 3](#_Toc418529055)

[2.2) Visualization Idea 2: Yelp Data Challenge 3](#_Toc418529056)

[3) Visualization Overview and Motivation 3](#_Toc418529057)

[4) Data & Data Processing 4](#_Toc418529058)

[4.1) Businesses 4](#_Toc418529059)

[4.2) Business Data Categories 4](#_Toc418529060)

[4.3) Review Data Size 4](#_Toc418529061)

[4.4) Review Data Counts Processing Speed 4](#_Toc418529062)

[4.5) Review Text for Word Cloud 5](#_Toc418529063)

[5) Exploratory Data Analysis 5](#_Toc418529064)

[5.1) Map Layout: 5](#_Toc418529065)

[5.2) Force Layout 6](#_Toc418529066)

[5.3) Area Graph 7](#_Toc418529067)

[5.4) Word Cloud 8](#_Toc418529068)

[6) Implementation & Challenges: 8](#_Toc418529069)

[6.1) Mapping 8](#_Toc418529070)

[6.2) Milestone 1 Timeline Update 9](#_Toc418529071)

[6.3) Force Layout Challenges 10](#_Toc418529072)

[6.4) Area Layout Challenges 10](#_Toc418529073)

[6.5) Visualization Changes: 11](#_Toc418529074)

[7) Evaluation 11](#_Toc418529075)

[8) Appendix 12](#_Toc418529076)

[8.1) Visualization of Ice and Fire Initial Proposal & Sketches 12](#_Toc418529077)

[8.2) Yelp Data Challenge 1 Initial Proposal & Sketches 12](#_Toc418529078)

# 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 & Data Processing

Our data set is from <https://www.yelp.com/academic_dataset>. Yelp provided data for academic purposes, which is a compilation of business, review, and user data for businesses near 30 different schools. The data set includes businesses 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. A significant amount of further data processing was implemented:

## Businesses

The initial businesses file was filtered to include on businesses associated with Harvard and MIT. This data set included approximately 1000 businesses.

## Business Data Categories­­­

Each business initially contained an array of multiple categories. In our visualization plan, we hoped to color code based on the category. This would be problematic because 1) each business included more than on category and 2) there were nearly 400 unique categories in the data set. It would be unreasonable to generate that many different colors. To solve this problem, every category from the entire data set was sent to an array. Each category was then counted for frequency. We then analyzed the top 20 most frequent categories. Overlapping categories such as restaurants and food or shopping and fashion were eliminated. We then updated the data set to only include one category if any of the categories matches one of the top 9 categories. If the business did not include one of the top nine categories, it was given a category of other.

## Review Data Size

The review data was initially 300mb large. It included data from 30 different universities and all of the text of the Yelp reviews. Because of the size of the file, we could not use chrome to load and manipulate the data. To minimize the data size, we used nodejs to run our javascript locally to create a new JSON file which only included reviews of businesses associated with Harvard and MIT. The smaller data file was only 50mb in size. This dataset could be loaded successfully with and Chrome

## Review Data Counts Processing Speed

The smaller review data set (50mb) was used to generate an area graph which plotted the number of reviews over time. Initially, the filter of this code was done in the visualization javascript (areavis.js). However, the load time was very slow. To make this process faster, the data was pre-processed with nodejs. A new JSON file was generated. The JSON file included each business. Within each business, the date and cumulative number of votes on that date were included. This file was passed into the area visualization. The area graph is then generated by summing the number of reviews for the brushed businesses­. Because this file was still slow to node, we limited the scope to a two year time frame.

## Review Text for Word Cloud

To generate the Word Cloud, the review data had to be pre-processed significantly. This data was processed in nodejs and written to a json file to increase the visualization speed. First, an empty array was created where for each business id, there was an empty array of text and an empty array of top words. We looped through every reviews and pushed all of the review text to the text array. To process the data, we removed any punctuation. We then split the large string of review text into an array of individual words which we sorted alphabetically. We then looped through every word in each business. We counted the number of times every word appeared and pushed this to a new array. We sorted this array by count and included only the top 50 words. TALK ABOUT AVERAGE OF STARS

# 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.

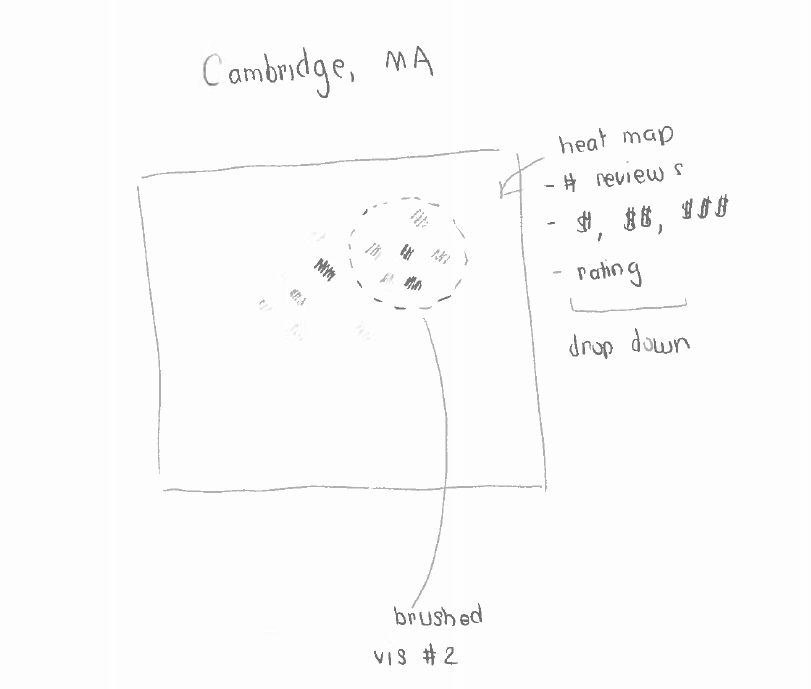


Figure : Map Layout

### 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.

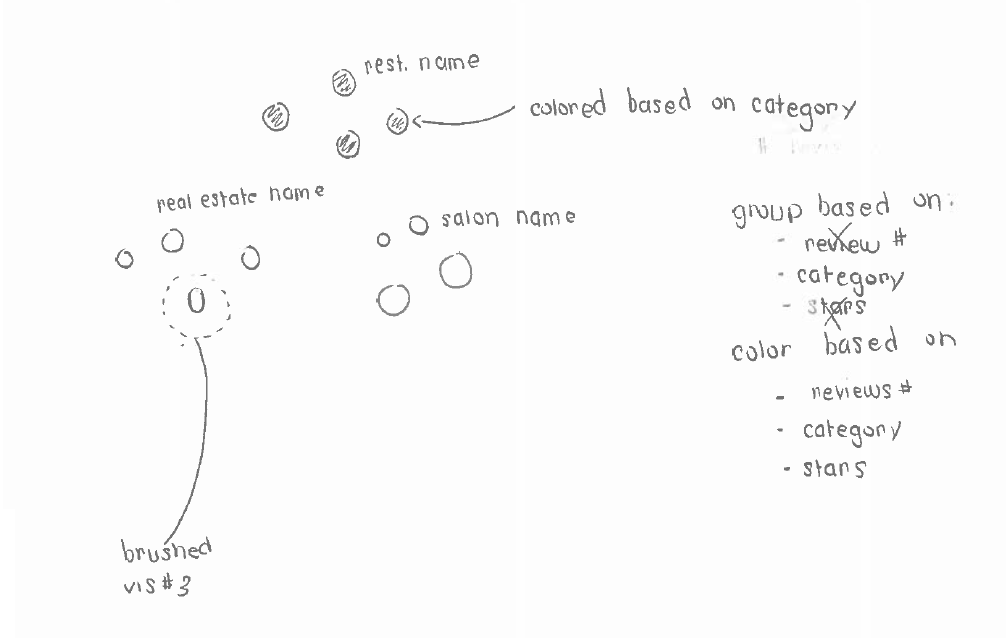


Figure : Force Layout

### 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.

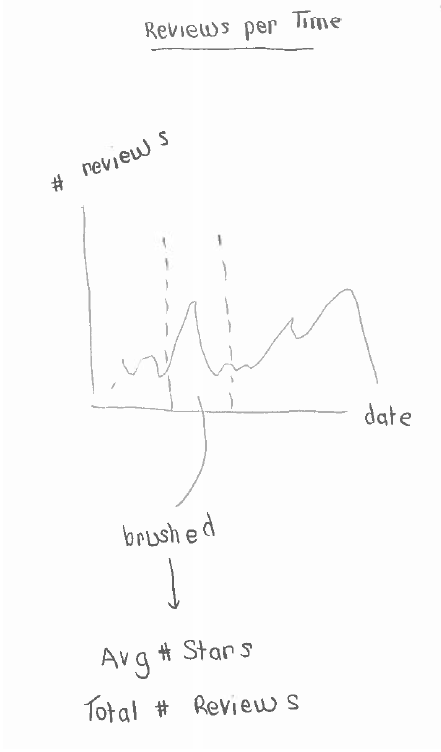


Figure : Area Graph

### 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.

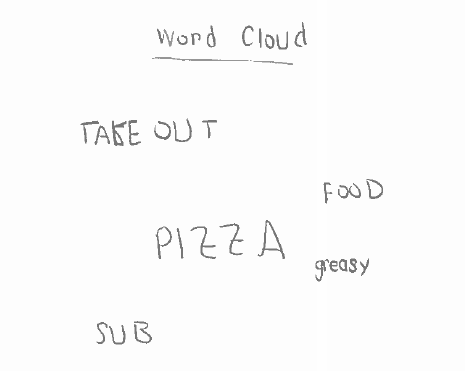


Figure : Word Cloud

### 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.

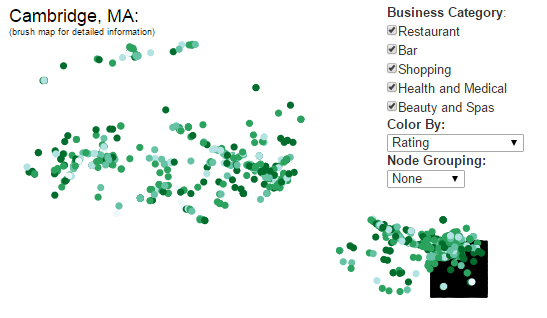


Figure : Map Vis Problem

## 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

## Force Layout Challenges

* When brushing, change the forces and attributes because a smaller number of data sets
* When trying to change radius, the old points with a different radius stayed
* When changing to non-grouped layout, it stays grouped because the points are moving from their grouped positions

## Area Layout Challenges

When our area graph was first generated, it looked like Figure 6. We determined that the dates were not sorted properly and thus caused the area graph to plot out of order. We fixed this by formatting the data in javscript using “new Date”. This ensured we were working with dates and not strings.

Our review count area graph is re-generated upon brushing the map. However, when we brushed we were getting the graph represented in Figure 7 below. The problem turned out to be from the code in Figure 8. Instead of using reviewsByDate we were using this.reviewsByDate. Therefore, each time a new selection was picked, this.reviewsByDate was getting longer and longer. To fix this we used reviewsByDate and reset the variable each time the function was entered. This fixed our graph and significantly increased the speed of our code.

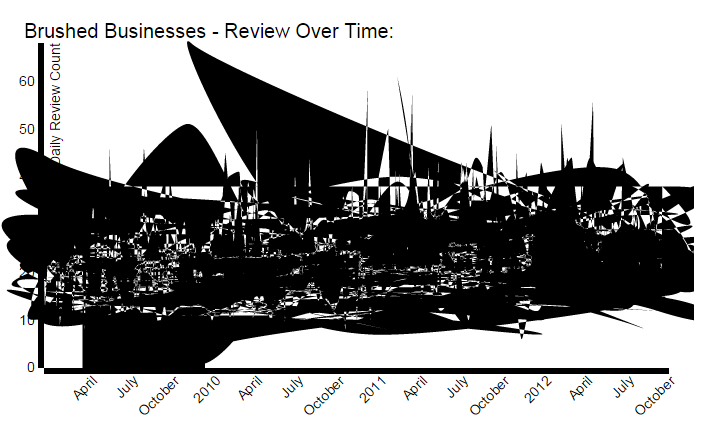


Figure : Area Vis Problem 1

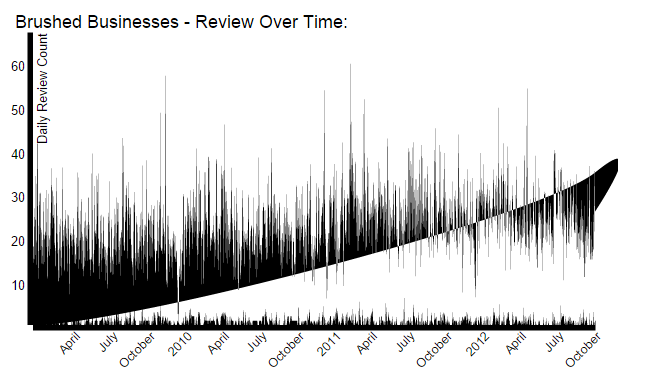


Figure : Area Vis Problem 2

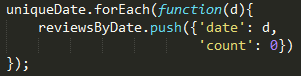


Figure : Area Vis Problem Code

## Visualization Changes:

Over the course of designing our visualization we changed different aspects of our design.

* Initially (Figure 1) we intended to create a heat map of Cambridge MA. However, to take advantage of the map functionality when plotting latitude and longitude, we altered our visualization slightly. We created circles for each business and colored similar to a heat map. Therefore, the user can see where an individual business is located and how it compared to others around it by the coloring.
* Initially (Figure 2) we intended to let the user select to color the nodes based on number of stars or number of reviews. Because this encoded the same information as the map visualization, we removed this from the options
* Initially (Figure 3) we intended to update the area graph (reviews per time) based on a user selecting a node in the force layout. However, many of the businesses had a small number of reviews which rendered the graph meaningless. Instead, we updated the area graph based on the brushed area so that it would be plotting a larger number of businesses.

# Evaluation

To be completed when project is complete

# Appendix

## Visualization of Ice and Fire Initial Proposal & Sketches

## Yelp Data Challenge 1 Initial Proposal & Sketches