CS171- Process Book

https://github.com/sbemagx/CS171-Metro-Boston-Food-Exploration

Project Proposal- Background and Motivation

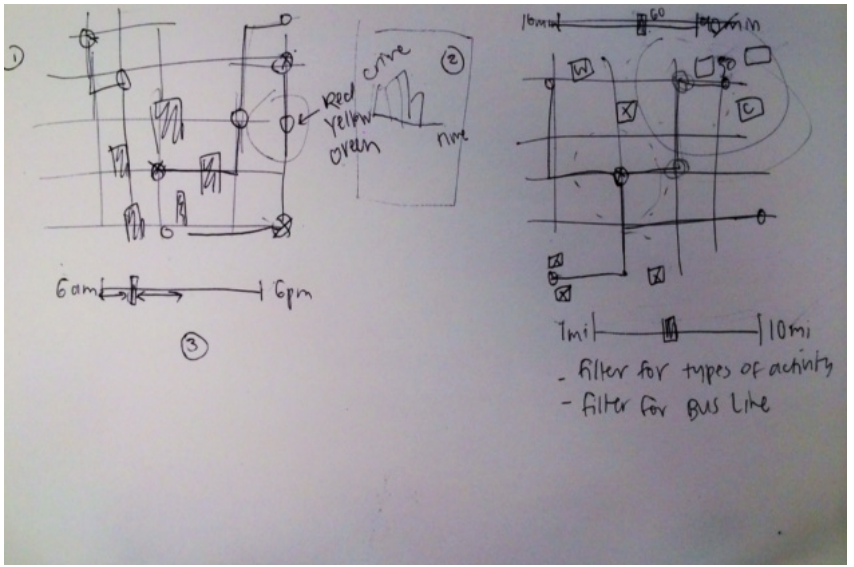
Early on we decided to do something related to transportation. Jack Birger works at a transportation consulting company and had the idea of using real-time bus data through the OneBusAway application and MTABustime. We explored ideas of combining MTA data with weather and crime data, but struggled to come up with specific questions that the data, and subsequent visualizations would answer. Therefore, we began to explore new datasets, looking for ways to combine interesting aspects of transportation with data from another source. We sought data that has both an excitement factor and that could be utilized to answer meaningful questions. Along the way we came across a very cool visualization that juxtaposes the London Tube system with second languages spoken, uncovering insights into the cultural fabric of the city. We think it would be interesting to apply a similar approach: using MBTA locations as a basis for Yelp data to explore metro Boston and its culture by mapping the constellation of restaurants within walking distance of train stops by ethnic category. Utilizing this data we aim to expose the clustering of restaurant categories (Italian, Vietnamese, etc) around particular MBTA stops.

Project Ideation

Initially, we were very interested in the benefits and drawbacks of public transportation. We considered a routing task dealing with peaks of public transport and how fast you could get to a certain location at a given time, utilizing some predictive analytics. Since we had just been dealing with Snowpocalypse, we were considering incorporating weather data as well, but there wasn't a clear story and outline of how we could tackle this task in a couple weeks time.

US Climate Data

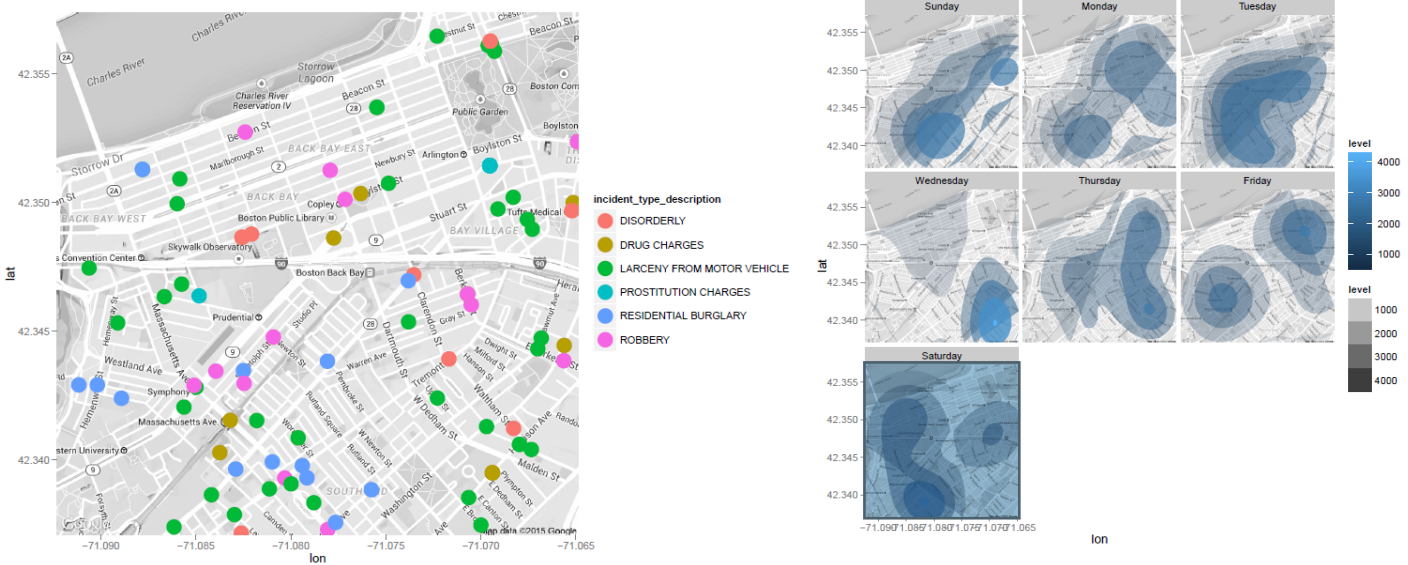
National Weather Service Forecast



Project Ideation

Sarah's coworker had demonstrated some simple ggplot features in R with mapping Boston crime data, and it led us to consider what factors persuade and dissuade us to go to a certain neighborhood for food/activities

[Boston Crime Data]



Project Ideation

We really wanted a solid data source that could tell us a story about Boston and its surrounding neighborhoods, and drew upon a lot of different sources and existing visualizations for inspiration.

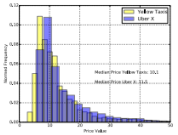


FIG. 2: Distribution of prices per journey for Uber X and Yellow Taxis in New York City.

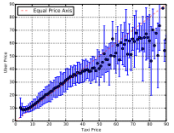
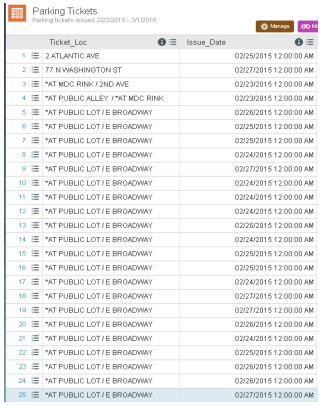
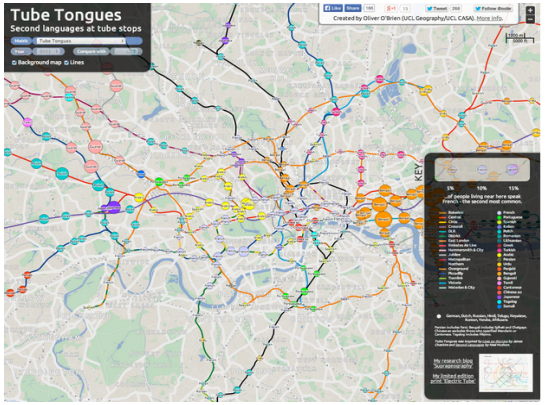
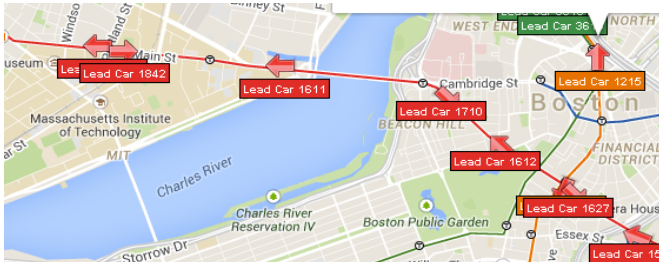
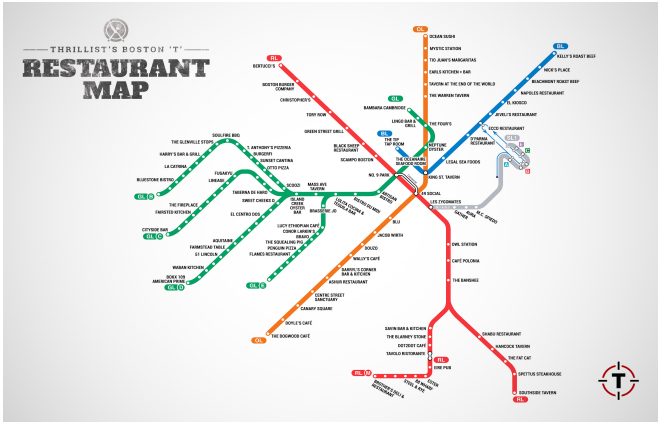


FIG. 3: Median Uber price for a given Yellow Taxi price.









Food Establishment Inspections

Health inspections of licensed food establishments

MASSACHUSETTS INCOME GROWTH BY ZIP CODE FROM 2001-2005

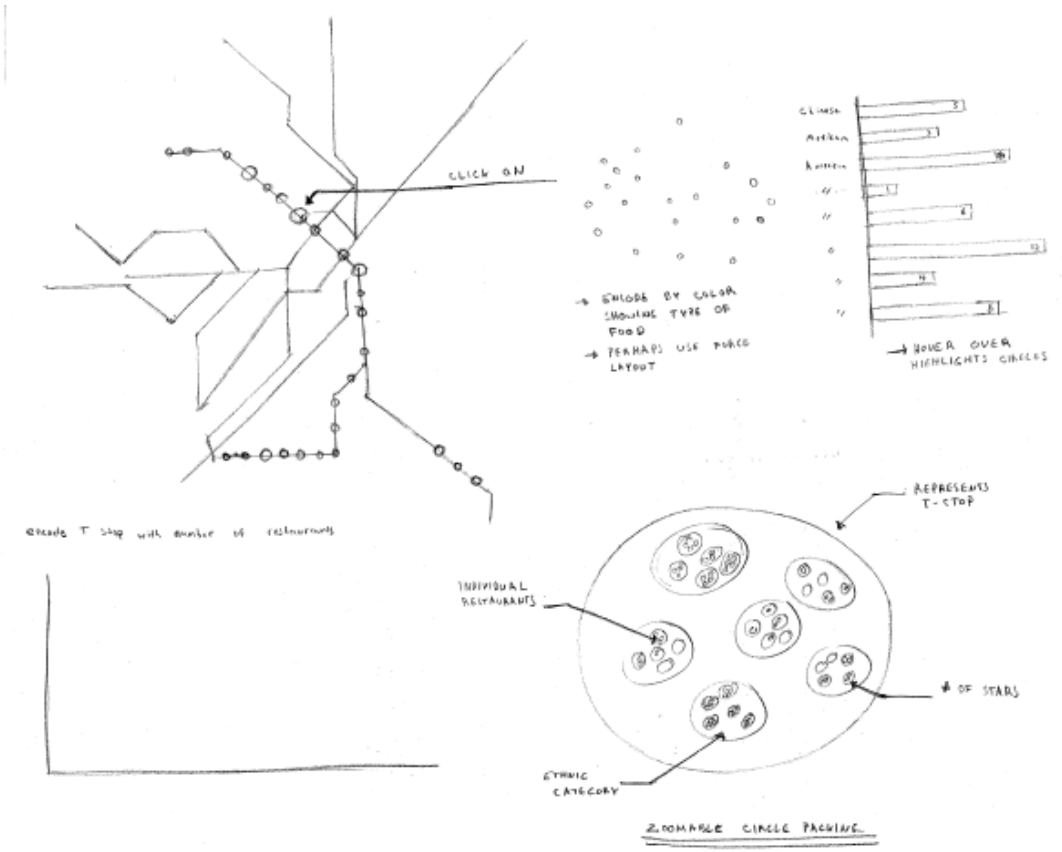
This chart shows the percentage change in average income by ZIP codes in Massachusetts from 2001 through 2005, the most recent figures available. Averages include both single and joint filers. ZIP codes are listed by community, but ZIP boundaries do not always align precisely with city and town borders. The chart can be sorted by clicking on column headers. Source: IRS.

Zip code	City or town	Number of returns, 2001	Average income, 2001	Number of returns, 2005	Average income, 2005	% change in avg. income, 2001-5
01001	Agawam	8,239	$43,292	8,390	$45,686	5.5
01002	Amherst	10,577	$51,729	10,135	$61,240	18.4
01003	Amherst	260	$8,377	194	$10,361	23.7
01004	Amherst	553	$33,203	474	$47,127	41.9
01005	Barre	1,978	$44,321	2,113	$49,272	11.2
01007	Belchertown	6,166	$51,179	6,709	$54,219	5.9
01008	Blandford	609	$43,125	617	$51,139	18.6
01009	Palmer (Bondsville)	672	$33,579	659	$35,757	6.5
01010	Brimfield	1,677	$52,016	1,768	$56,764	9.1
01011	Chester	863	$36,373	594	$41,944	15.3
01012	Chesterfield	393	$37,926	372	$46,476	22.5
01013	Chicopee	10,126	$31,485	9,872	$34,162	8.5

```
rate_marriage   : How rate marriage, 1 = very poor, 2 = poor, 3 = fair,
                  4 = good, 5 = very good
age             : Age
yrs_married     : No. years married. Interval approximations. See
                  original paper for detailed explanation.
children        : No. children
religious       : How religious, 1 = not, 2 = mildly, 3 = fairly,
                  4 = strongly
educ            : Level of education, 9 = grade school, 12 = high school,
                  14 = some college, 16 = college graduate, 17 = some
                  graduate school, 20 = advanced degree
occupation      : 1 = student, 2 = farming, agriculture; semi-skilled,
                  or unskilled worker; 3 = white-collar; 4 = teacher
                  counselor social worker, nurse; artist, writers;
                  technician, skilled worker, 5 = managerial,
                  administrative, business, 6 = professional with
                  advanced degree
occupation_husb : Husband's occupation. Same as occupation.
affairs         : measure of time spent in extramarital affairs
```

Project Ideation- round 474

Ultimately decided there are a lot of tools out there that exist to help plan and predict trips, and we wanted to understand more about the mapping of culture. We decided to utilize the existing tools and information that Yelp provides us to look into the makings and patterns of neighborhoods defined by the MBTA.



Project Proposal- Project Objectives

Primary question:

What cultural patterns exist around metro Boston that can be exposed through the visualization of restaurant clusters located near MBTA train stops?

Secondary questions:

Which filters can aid in seeing these patterns? Which filters are most accessible and necessary for restaurant goers?

At a practical level, Which filters can aid in selecting a restaurant near a particular stop?

Learn and Accomplish:

We are eager to provide a novel method for the exploration of food and culture around the city of Boston. We hope to allow users to gain new insights into cultural patterns as they relate to the combination of food and culture in relation to public transportation and settlement patterns in the metro area.

Project Proposal- Data

Primary data sources:

Yelp:

Link: https://www.yelp.com/developers

Format: JSON

Sample Data:


```
{ u'categories': [ [u'American (New)', u'newamerican'],[u'Pizza', u'pizza'],[u'Cocktail Bars', u'cocktailbars']],
u'display_phone': u'+1-617-500-3055',u'id': u'russell-house-tavern-cambridge',u'image_url': u'http://s3-media4.fl.yelpassets.
com/bphoto/M7YViqqBZM7Pl43JSocI1Q/ms.jpg',u'is_claimed': True,u'is_closed': False,u'location': { u'address': [u'14 JFK St'],
u'city': u'Cambridge',u'coordinate': { u'latitude': 42.373122,u'longitude': -71.119703},u'country_code': u'US',
u'display_address': [ u'14 JFK St',u'Harvard Square',u'Cambridge, MA 02138'],u'geo_accuracy': 9.5,u'neighborhoods': [u'Harvard
Square'],u'postal_code': u'02138',u'state_code': u'MA'},
```

Project Proposal- Data Processing

Yelp data vis will require multiple requests because Yelp's API limits search results to 20 per query. Also, requests of different types will need their data stitched together. Data will need to be collected and transformed into objects containing lat/long, categories, and ratings. We will be exploring two paths to deliver this data: (1) getting the data real-time from yelp via their API and (2) programatically collecting their data, decomposing, and storing it in a RDBMS so that it can then be recomposed and sent to the client via a lightweight REST server (such as Django REST Framework) as needed.

Project Proposal- Visualization

The primary display will be a stylized map of the MBTA train system in metro Boston. Users will be able to see the density of the categories of restaurants (American, Chinese, etc) that make up the majority at each stop (if there was a stop in the North End it would be Italian, for example) or filter to see the prevalence of a particular category across the system (show density of Vietnamese across the system).

Project Proposal- Features

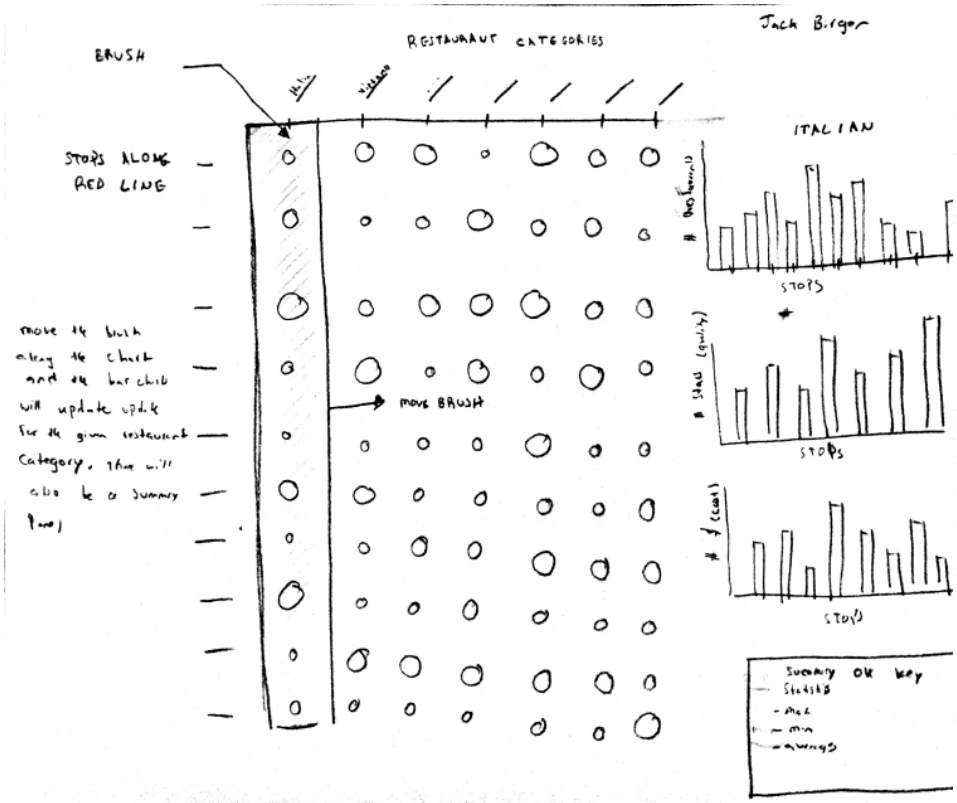
Must-Have Features

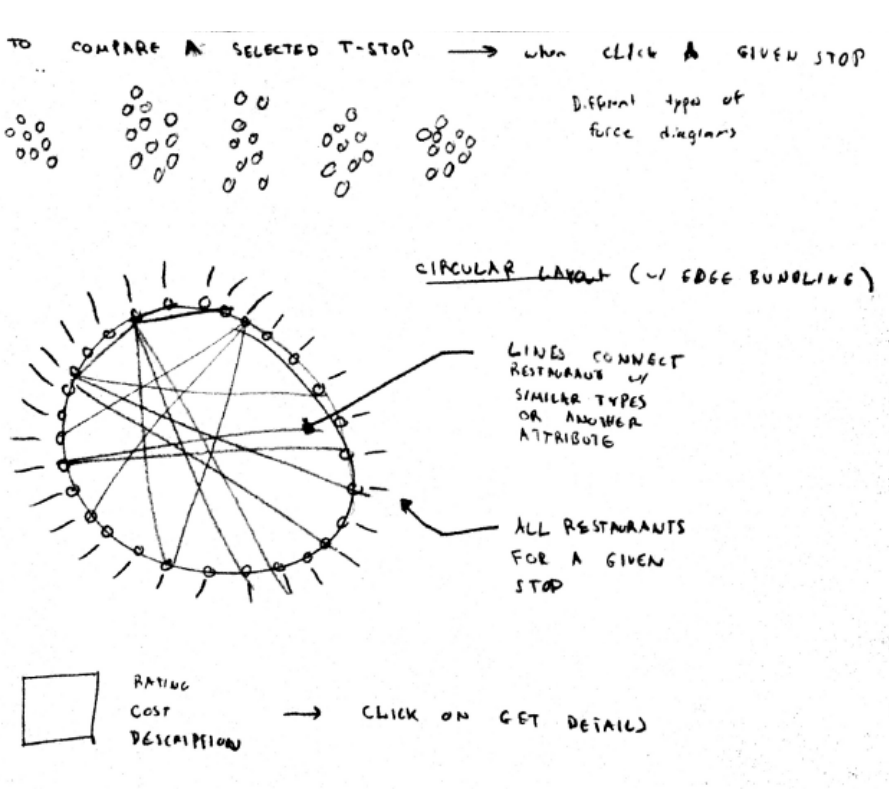
- Ability to layout stylized MBTA train map
- Ability to place relevant Yelp data around each of the T stops
- Ability to filter Yelp data by several criteria (such as category and rating)

Optional Features

- Ability to compare multiple T stops in different filters.
- Adjust size of the MBTA stops to reflect the total number of restaurants at a given stop.
- Adjust size of the MBTA stops to reflect average cost or ratings across restaurants at a given stop.
- Provide MBTA service alerts/status that pertain to the location.
- Ability to add MBTA real time data

Project Proposal- Sketches





Project Proposal- Project Schedule

April 3-6: To layout an outline/stub for the entire visualization. This will allow us to have a plan for how everything will interact and work together. This portion will be a collaborate effort by the team.

April 6-15: Once the outline is complete, we will divide different portions and views to different individuals. We will work on these primarily independently, but will use each other as resources if we get stuck or for any other purposes.

April 15-17: Combine individual portions to create a working prototype. Reformat process book if necessary.

April 17 Milestone 1: Complete data acquisition, have data structure ready. Create working prototype. Turn in process book.

April 17-30: Update user interface with additional filters and seamlessly combine each individual part.

April 30-May 3: Complete the process book, create screencast, put finishing touches on website.

TF Feedback

- Storytelling
  - make sure we have a clear question that we solve
  - Can either be done by a longer page with text between visualizations, or a walkthrough on one page
- Filters
  - think about search radius, how to implement
  - Maybe add a google map?
- Interaction + linking
  - finalize what other information we need in our visualization and MBTA map
  - Solidify a layout of website and how views interact with each other
- Data
  - No need for real time updates, use a static dataset and aggregate via D3/Javascript
- For milestone, possible to just start with 1 T stop for proof of concept

Data Setup

Christian secured data through both Yelp and MBTA APIs. In order to convert the data provided into a usable format, he had to use XHR, set up the database, and set up REST server with endpoints for accessing the data. He found that converting the data from a .txt to JSON format yields more than 1.5Gb per day sampled, and we needed to serve and load this data JIT to avoid massive lag. Another approach to create a manageable data set was to filter the data for specific elements as seen below.

```
data structure for top view (for each T stop):
[
  { x : int
    y : int
    line : attr
    id : int
    total_restaurants : int
  },
]
```

Transformed data:

```
{
        "rating": 5.0
        "review_count": 13
        "name": "The Table At Season
To Taste"
        "stop_id": 101
        "latitude": 42.3983409
        "longitude": -71.1310318
        "line": ["red"]
        "categories": ["American (New)","
Breakfast & Brunch"]
}
```

Data Setup

- **VPS created via digitalocean**
  - IP:45.55.178.178
  - URL: gaslight.grav.io
- **Configure DNS of new VPS**
  - $ping gaslight.grav.io
  - PING gaslight.grav.io (45.55.178.178): 56 data bytes
  - 64 bytes from 45.55.178.178: icmp_seq=0 ttl=56 time=19.530 ms
- **Install and configure tech stack**
  - Nginx
  - Postgresql
  - Bootstrap
  - Gunicorn
  - Virtualenv
  - Django
  - Supervisor

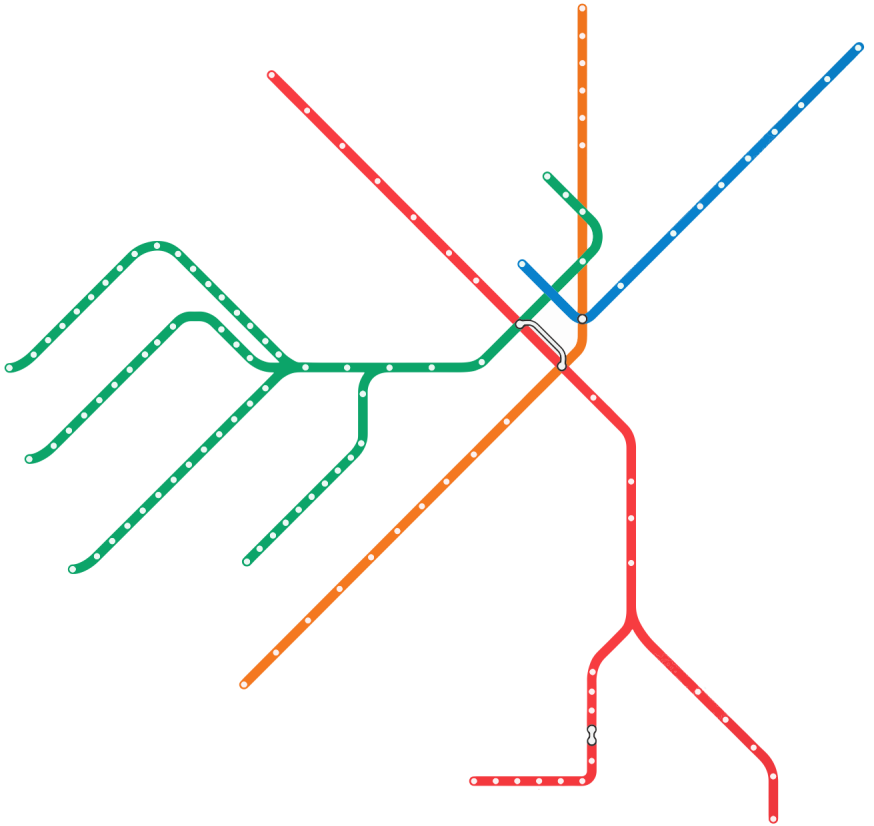
Front End Setup

- Loads data, SVG T Stops
- Filters
  - by MBTA line
  - by average rating
  - by number of restaurants

SVG Setup

```
stops = [
          { 'location': 'alewife station, cambrige, ma', 'line': 'red', 'stop_id': 1 },
          { 'location': 'davis square, cambrige, ma',    'line': 'red', 'stop_id': 2 },
          { 'location': 'porter square, cambrige, ma',   'line': 'red', 'stop_id': 3 },
          { 'location': 'harvard square, cambrige, ma',  'line': 'red', 'stop_id': 4 },
          { 'location': 'central square, cambrige, ma',  'line': 'red', 'stop_id': 5 },
          { 'location': 'kendall square, cambrige, ma',  'line': 'red', 'stop_id': 6 },
      ]
```


```
  {
          "rating": 5.0,
          "name": "The Table At Season To Taste",
          "longitude": -71.1310318,
          "stop_id": 1,
          "latitude": 42.3983409,
          "line": "red",
          "categories": ["American (New)","Breakfast & Brunch"]
   },
```

Metadata to Create Circles on the SVG

We decided the best way to visualize meaningful data on the svg mbta map was to create circles over the svg and then manipulate those circles as needed. In order to do this, we needed to manually collect data that related the locations of each stop in the svg to the yelp data. We did this by creating a metafile that had an array of objects. Each object represented key information to track the stop:

{"stop_id":104, "station":"Harvard Square Station", "line":["red"], "x":863, "y":418}

We were able to loop through this metadata file (mbta_metadata.json), creating circles over each stop and binding the following information:

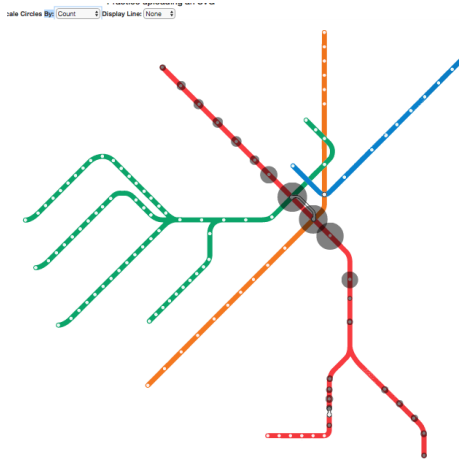
- Assigned the x/y location based on“x” and “y”
- Assigned an “id” to each circle to enable easy identification/selection of a given stop based on “station”
- Assigned a “class” to each circle to enable easy identification/selection of a given line “line”
- Assigned a stop_id that is used to map the Yelp data to circle elements.

Circles Plotted on SVG

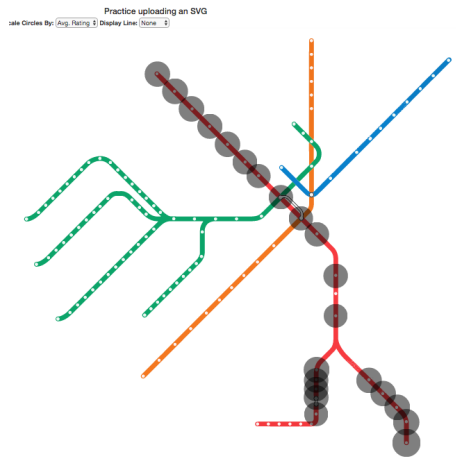
Creating Initial Filters and Views

To begin we only used the data for the redline to simplify our proof of concept - although we can easily and will expand to the full mbta dataset. The next goal was to prove that we could manipulate the data and create an initial visualization. To do this we created a function that would loop through our yelp dataset and calculate the total number of restaurants for each stop and the average rating across that total of restaurants per stop. These totals were stored in an array of objects that stored the aggregate information for each stop. From there we simply scaled the radius of the correlating circles based on these rankings.

STOPS



AVERAGE RATING

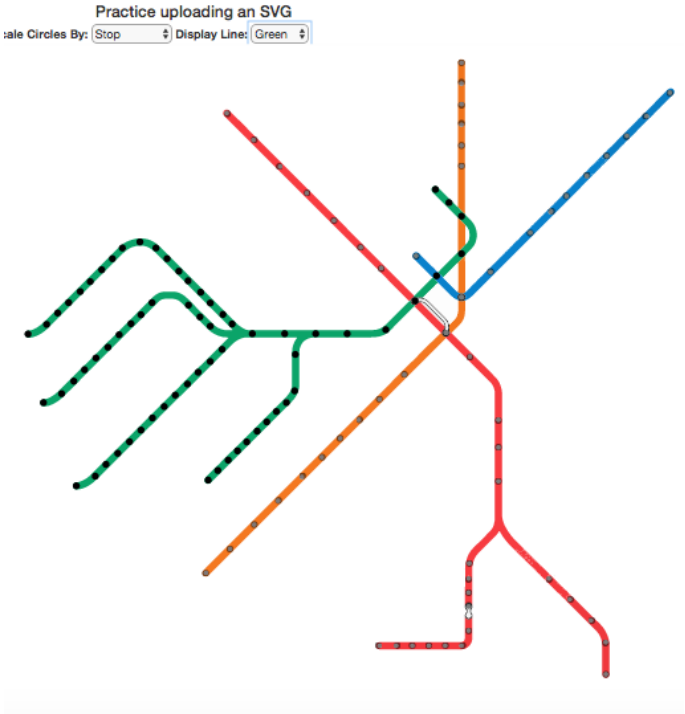


Highlight Individual Lines

A simple addition we wanted to add for this milestone was to show that we could highlight individual the view for individual lines. This ability enables the user to reduce the complexity of the map, dig into the visualization and make more interesting conclusions.

Image Highlights the

Green Line



Next Steps for Map View

For this map view there are some additions that we plan to incorporate in the upcoming days. We would like to add the ability to scale the stops based on category of restaurant. This gives the user the ability to make much greater conclusions.

The user will be able to answer theoretical questions such as:

> “What stop in the mbta has the most options for quality Chinese food?”
>
> “If I want to stay on redline what location has the most number of high ranking sushi restaurants?”

Additionally the user might be able to make conclusions about why these locations are ranked the way they are and if these patterns indicate any cultural relationships within the MBTA.

We also discussed adding slider that could change the walking distance length additionally filtering the data.

Next Steps

- Add at least 2 linked interactions
- get the main image situated in Bootstrap
- Create walkthrough feature
- Nail down the story we want to tell by iterating through data and filters

Next Steps

Circle Packing breakdown of stop data

Filter and represent size by:

Circle Level 1: Stop:

Circle Level 2: Category

Circle Level 3: Data Circle Size:

- # reviews
- distance
- ratings

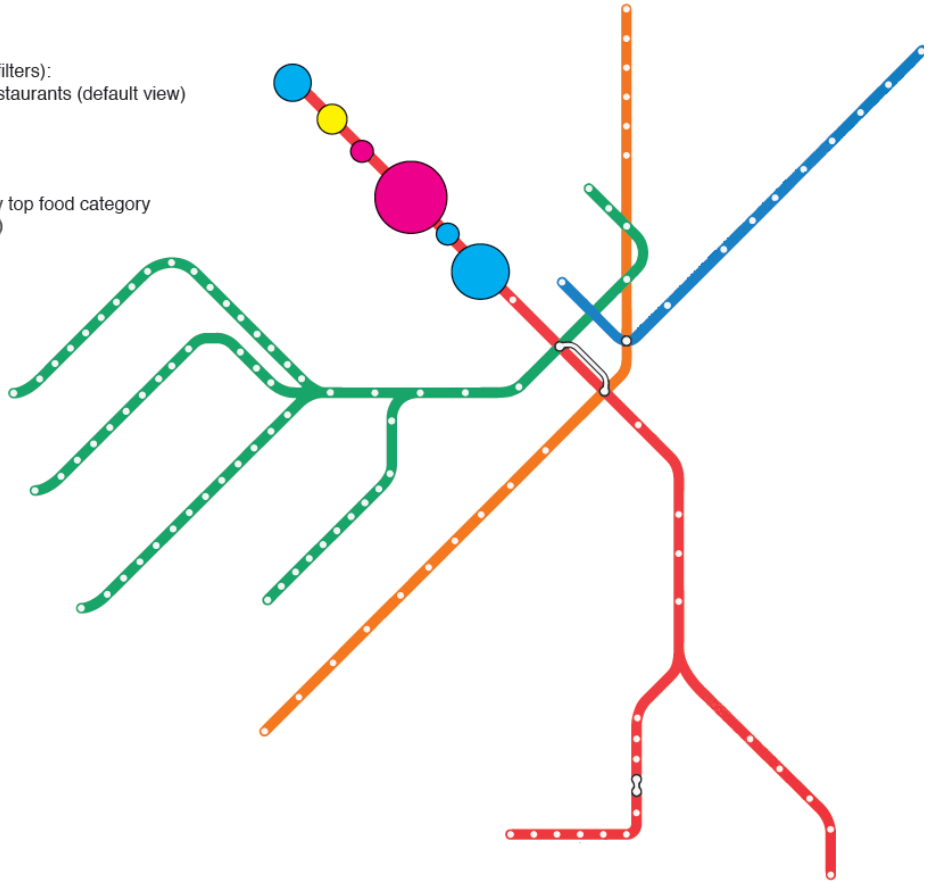
Circle Packing

Next Steps- potential view

Main View:

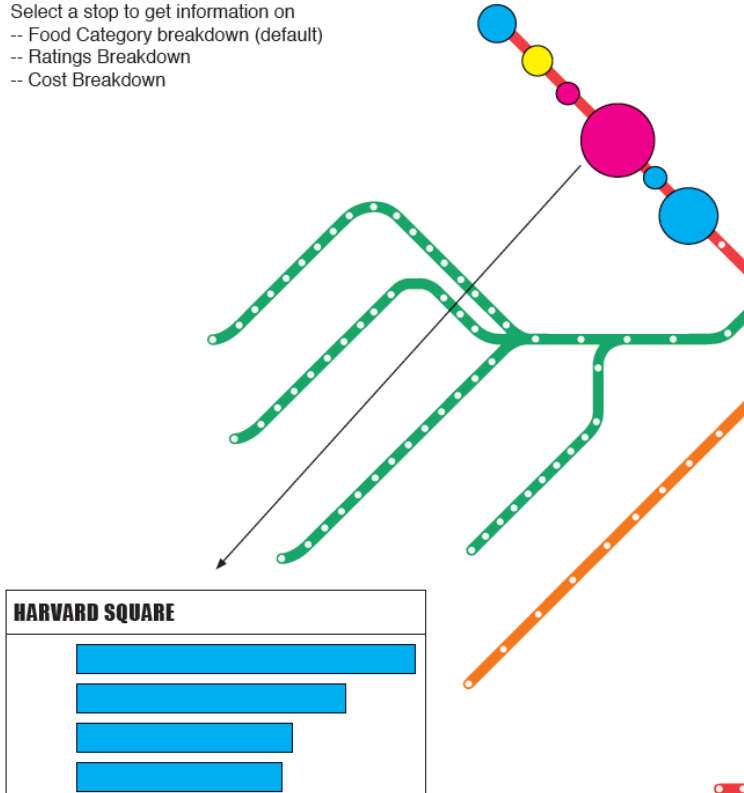
MBTA Stop sized by (filters):
-- Total number of restaurants (default view)
-- Restaurant cost
-- Category Type
-- Yelp Rating

MBTA stops colored by top food category (Italian, American, etc.)



Drill Down View:

Select a stop to get information on
-- Food Category breakdown (default)
-- Ratings Breakdown
-- Cost Breakdown



1.00 x 8.50 in