CS171 Final Project Process Book

Ina Chen and Fox Morone

April 2014

1 Overview and Motivation

Bike sharing is an increasingly popular mode of transportation within large cities such as Boston, New York, Chicago, D.C., and others. Unlike ride shares or car pooling where location and rides are relatively easily adjusted to rider volume and locations, bike sharing requires infrastructure carefully planned to accommodate riders at convenient locations and stocked to ensure availability of bikes. As the number of bike share users increases, it becomes useful for both city planners and bike program management as well as bike share users to look at ride patterns along with the flow of bikes at stations to complement program planning or program participation. For these reasons, a few bike share programs have released data on ride trips and some have even encouraged clear and creative visualization through visualization challenges. Here, we take advantage of these available bike share data in three cities, Boston, Chicago, and Washington, D.C.

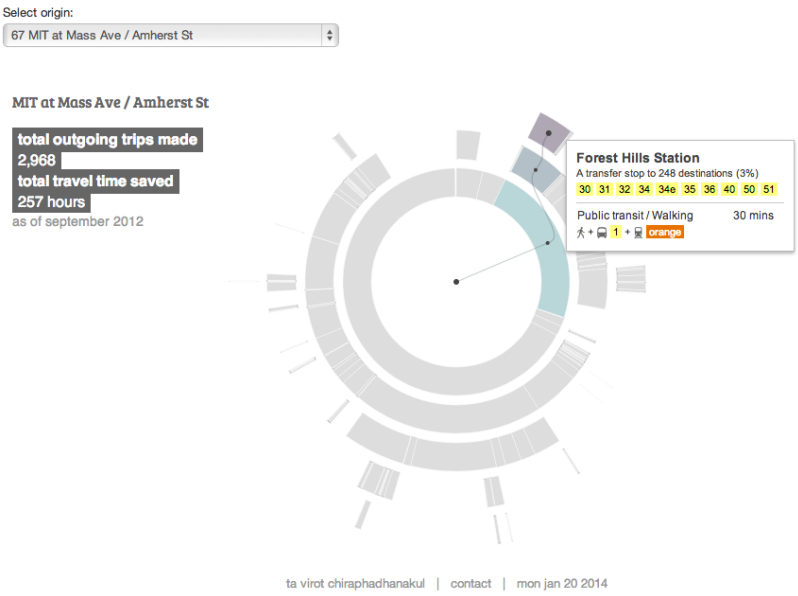
One focus of our visualization is to compare ride patterns between cities. There has been some visualizations of bike share data within individual cities , but we found none that compared bike share patterns between cities. Given the recent trend for programs release bike share data, we thought it would be interesting to compare the ride patterns across different cities. Since our data comes from bike share in large cities, we expect to find similar trends such as commute riders on weekends or tourist riders in summers but also differences that arise due to different geographies or climates of the cities. Comparisons between cities thus could provide both interesting patterns as well as open the way to looking at how bike share programs could improved in different cities.

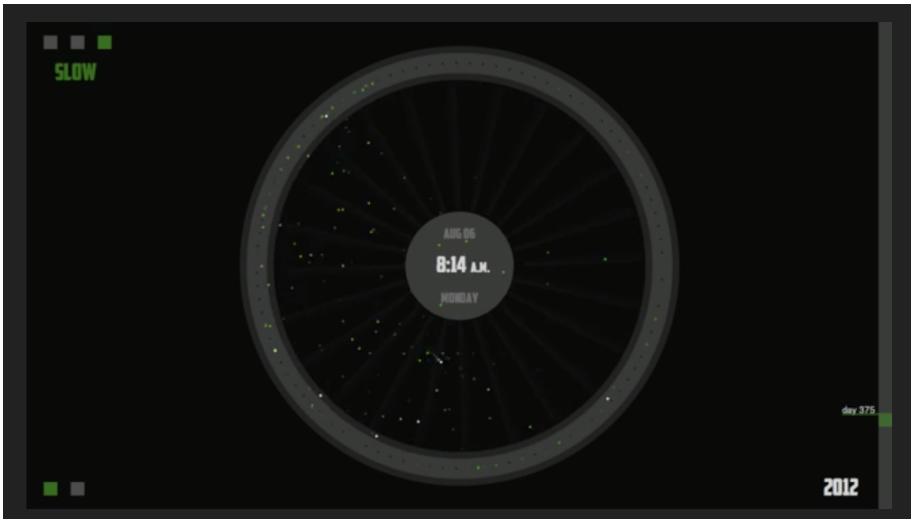
The second focus of our visualization is to look at bike rides data in its native geographical setting. Quite a few bike share data employs chord graphs or summary graphs that, while highlighting certain features, loses the geographical nature of bike rides (see Related Works below). In this focus, we want to complement inter-city comparison visualization with a detailed geographical-based visualization of ride patterns in each city. One particularly interesting feature we wanted to visualize was the flux of bikes at each station and ride routes in and out of each station. In this, we hope to visualize popular stations, trips, and the overall flux of bikes in both space and time.

2 Related Work

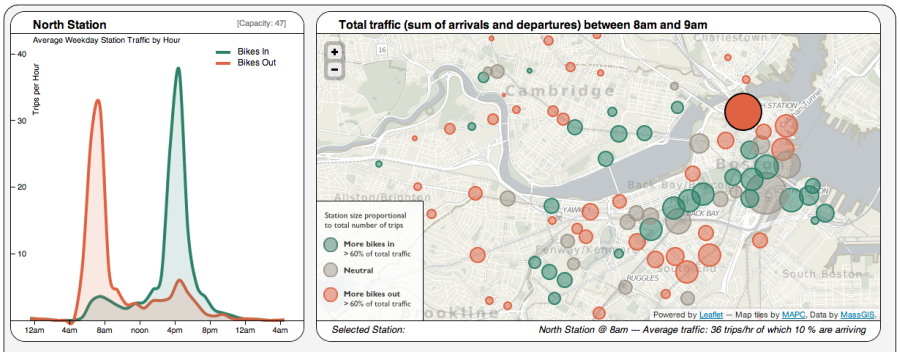
We found and drew inspirations (as well as rejections) from previous bike share visualizations.

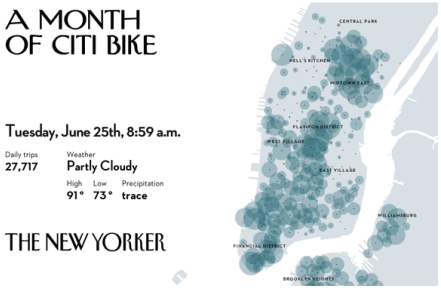
Visualization of rides between stations. We chose not to follow this style.

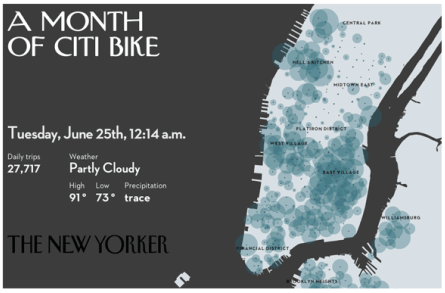


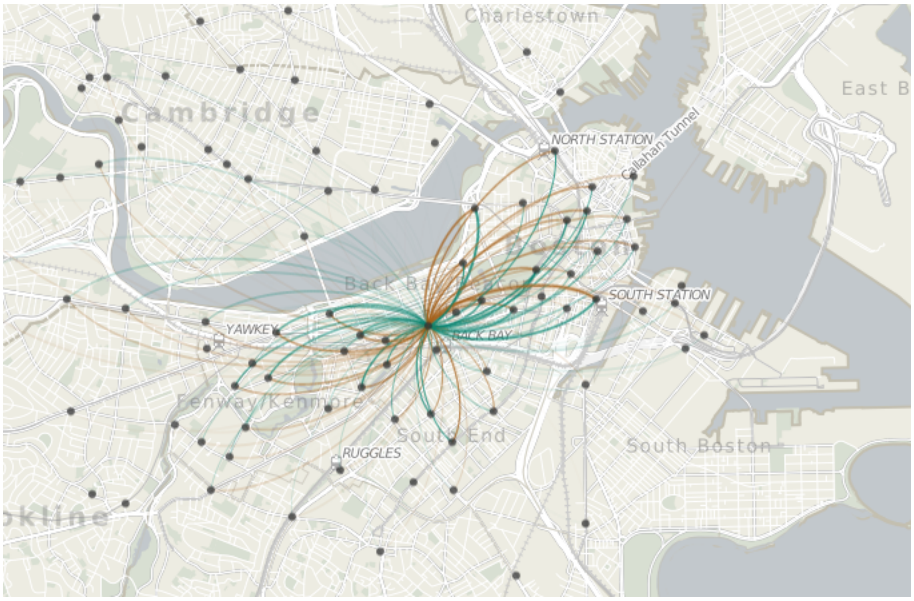


Visualization of station-related on a map. We particularly liked this type of visualization and drew much inspiration from these visualizations

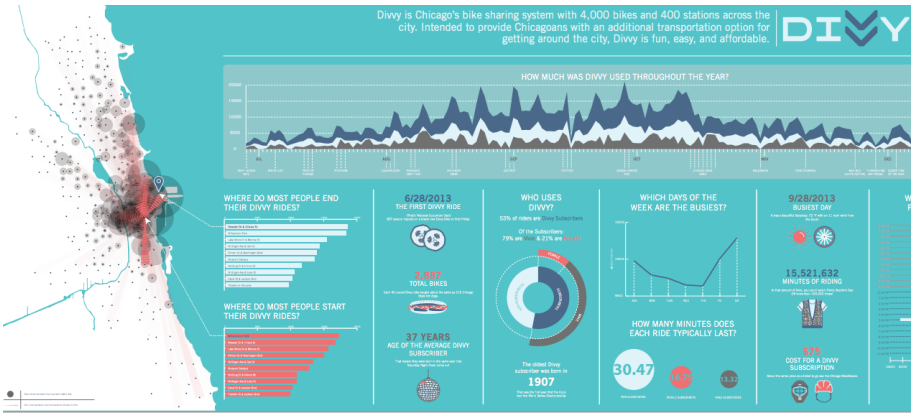


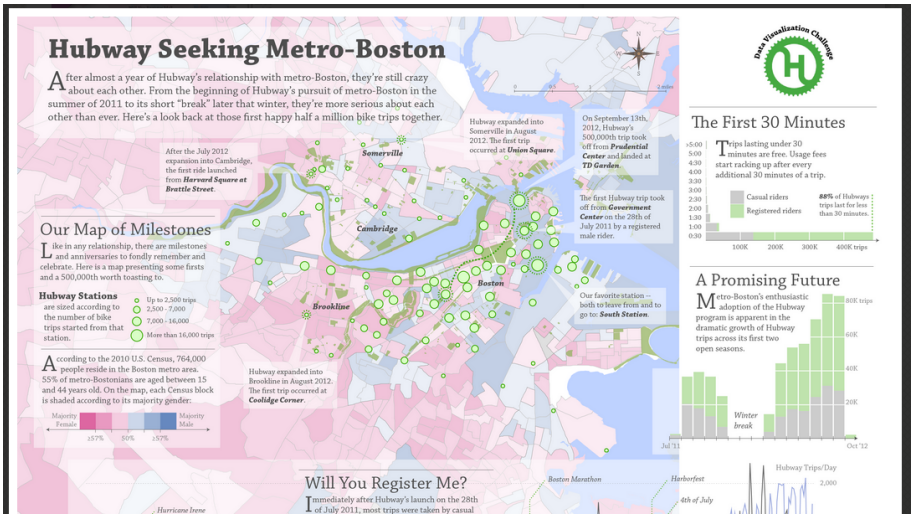






Static Visualizations that told stories. We hope to possibly incorporate some story highlighting feature but will focus on interaction so that the users may explore stories for themselves





3 Questions

Questions we attempt to answer with this visualization include how do bike ride patterns vary between subscriber and casual riders in different cities. Ride patterns is assessed through various metrics including average speed, number of trips, average duration per trip, average distance per trip, distance traveled which is calculated for each day over the course of a year. What are some interesting stories that the visualization say about deviations in bike rides over weeks, seasons, or for special events? What are patterns in trips within a city? What does the flux of bikes in and out of stations say about the commutes or travels of the riders?

As we worked on the project, questions were to adapted such that we can best visualize the data to maximize features that can be clearly presented while also maximizing smooth visualizations by reducing the data volume. This was particular important for the map visualization. While we initially asked if there are interesting stories from tracking bike data over the course of the year. When it became quickly obvious that this was somewhat unfeasible to tackle even the data size (and 300+ stations per city), we focused our attention on asking questions related to average patterns of bike share rides. On average over a month, how does the trip patterns change through the week? We are also considering further reducing the data to ask how bike patterns varied over a week when aggregated through a season.

4 Data

We pulled available CSVs (1000000 rows each) with ride data from each of the three cities. Generally, we have end and start station, end and start date/time, and casual user/subscriber data. we also had to get separate CSVs with stations' latitudes and longitudes.

The CSVs were several hundred megabytes and therefore unsuitable for a d3 project and also un-pushable to Github.

Cleaning was thorough and can be seen in our iPython notebook. Some of the ride data CSVs have strings of station names within their rows. These had to be removed to consolidate CSV size. Also, not all the cities provided distance data, so longitude and latitude values had to be used in conjunction with a global distance calculator to get that statistic.

As of now, we have cleaned all Washington D.C. data in order to start doing preliminary visualizations, but have not finished cleaning the Boston or Chicago files. This should be trivial after completing the D.C. set, however.

Concerning structure, we originally wished to use a CSV which essentially contained all the data we needed, but quickly saw that would also be too large of a file. We continued to restructure and aggregate data, and ultimately we have begun using a JSON file of average daily values of distance, ride duration, etc, which works well for our exploratory visualizations. On the other hand, our initial map visualizations were driven by CSV data which is separated by weekday, hour of the day, and station, but now we are using a JSON file for this as well. All of these files of tractable size for a D3 project.

We thought that it would be interesting to look at the ride share data in the context of neighborhoods in the city. To generate these maps, we first looked for files of city maps for each of the three cities. GeoJSON files were acquired for all three cities at https://github.com/codeforamerica/click_that_hood/ and were converted to TopoJSON files for use with d3 projections (We chose the mercator projection which visualizes city maps well). While not creatively nor intellectually difficult, finding, creating, and utilizing TopoJSON were not trivial tasks and took a fair amount of time.

We eliminated a few of the early data points in the Washington D.C. and Chicago datasets because the number of rides was very low, making certain measurments look like huge spikes and ultimately messing the y-scale of the graph. Data omission was not too absurd to do, however, because in Chicago's case, the bike system did not officially exist until 28 June, 2013. The days before this and some days afterward were eliminated from the dataset.

5 Exploratory Data Analysis

We first used rudimentary, simple visualizations to take a better look at our data. This included looking at the overall data over different metrics such as distance, ride duration, the resulting average speed, and number of rides per day.

We planned to look at $(rides_{in} - rides_{out})$ to display something of a 'flux' from each station. Hopefully depending on the hour and weekday, viewers would have been able to discern trends in tourist locations and residential locations. Implementations were not possible, though, because the size of the data shaped in such a way proved to be prohibitive for a D3 project.

6 Design Evolution

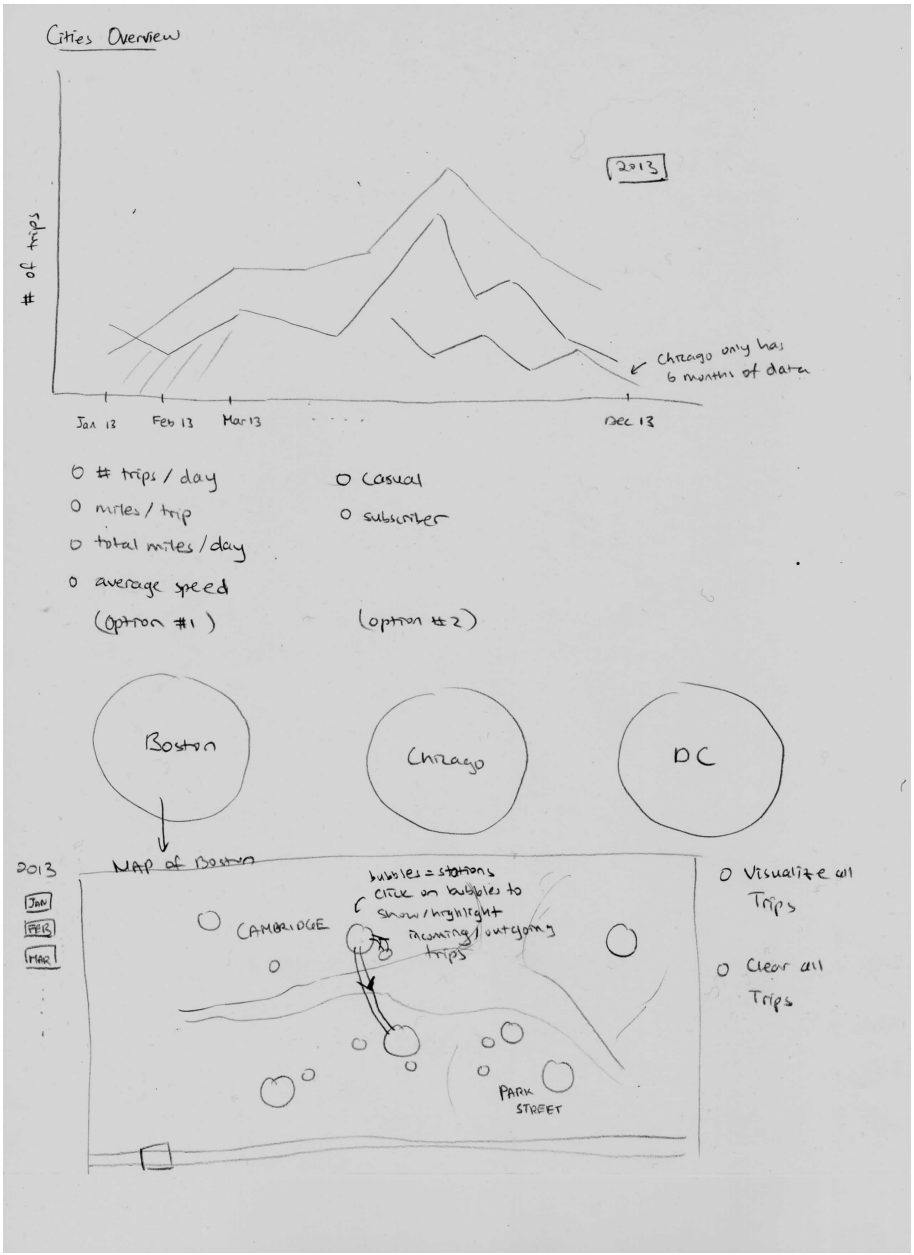
Our original map visualization idea involved getting to-from data for each station and drawing thicker or thinner chords between station nodes that were more or less 'connected' by bike usage. After much data cleaning, however, we realized two things: the number of rides—even over the entire year—for just about any pair of stations was generally minimal, and that storing so many small values in different JSON objects or CSV rows would result is data files that were far too large.

That idea eventually led to data aggregation which is discussed in the 'Data' Section above.

The line graph aspect of our visualization has been there since the beginning and to us seems like a very apt method of comparison between the cities.

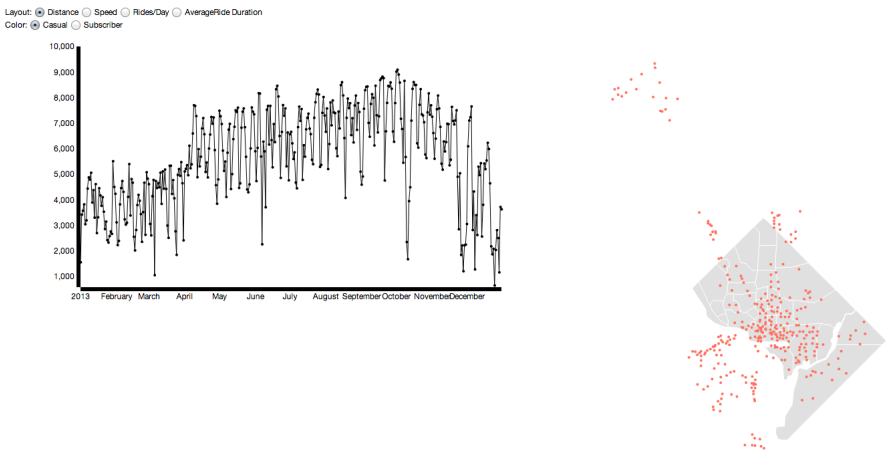
6.1 Proposed Design

We initially proposed a split presentation with the overview graph on top and the map visualization on the bottom. The top and bottom graphs cannot be show on the screen at the same time. The overall data visualization was to include interactions to display different features of the graph (two sets of features to choose from as shown in the image). The map visualization was to be an animation that displayed the ride patterns over the course of the year.



6.2 Milestone 1 Design

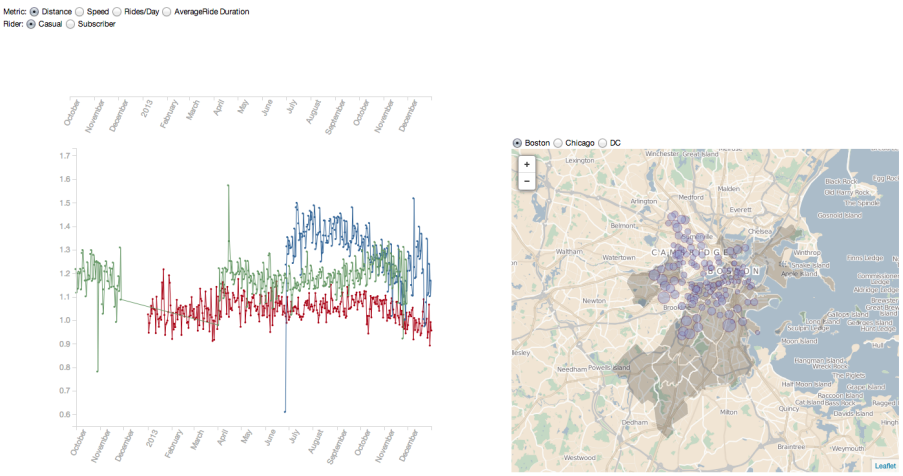
Visualization as of Milestone 1



6.3 Intermediate Design

Our intermediate design is very similar to our final implementation but left out some subtle adjustments. Of course, this intermediate design differs quite significantly from our milestone design.

Example of Intermediate formatting



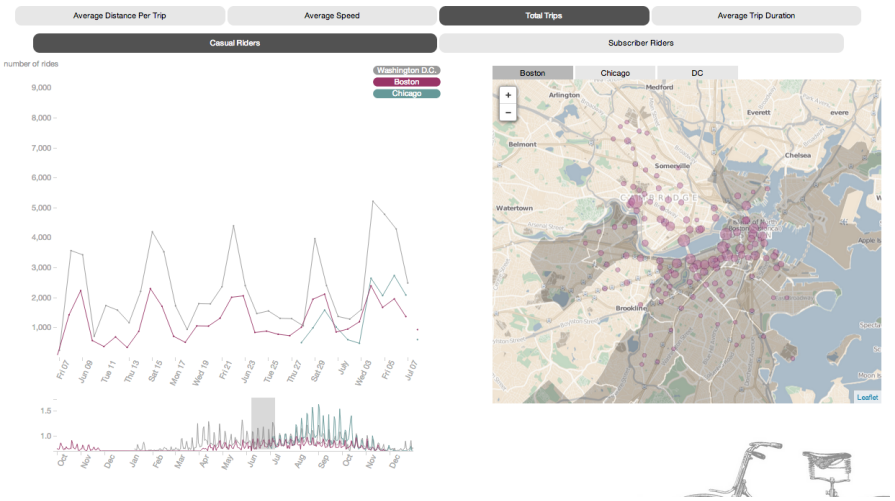
Obviously, the layout and formatting were improved and the map was laid over a photo of the surrounding area. This allows for more complete and intuitive exploration of the map, complete with street names and landmarks. Bubbles were placed on the map to reflect the number of outgoing rides and therefore how "popular" as site might be.

Moving to the other panel, all three cities data were plotted, addressing our goal of bike share system comparison. A brushable axis was incorporated, which allowed for focusing on select data, something that is indispensable when working with displays of the relatively small width we used to develop this visualization

7 Implementation

And now for the main event.

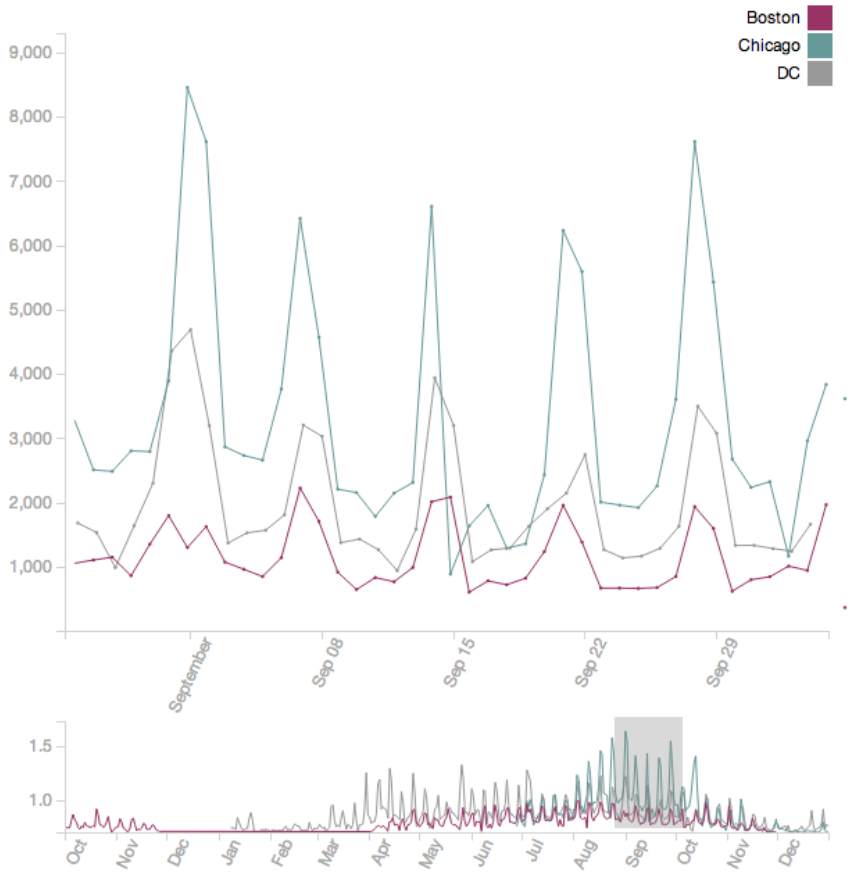
Final Implementation



Our final design differs most notably, if not most obviously, form our intermediate design due to the addition of panel-linking. In other words, our design updates the map as well as the line graph to display corresponding metrics on each. This allows for a more complete and relevant message to be brought to the viewer. In other words, all information about rider speed or average rider distance is shown simultaneously on the graph and on the map.

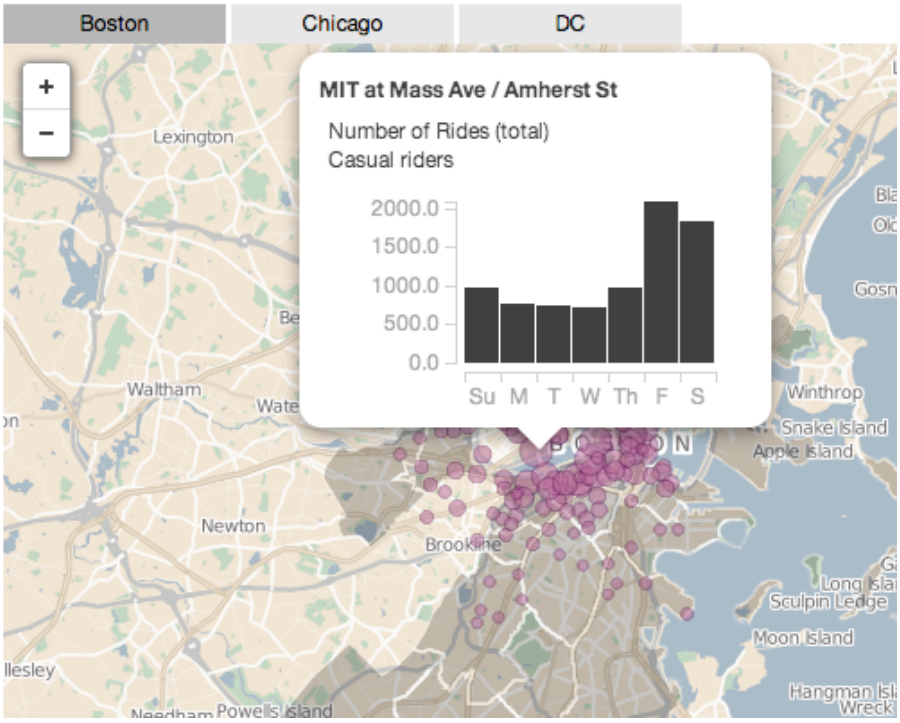
Also, useful is the addition of a brushable overview graph, which replaced the brushable axis exhibited by the intermediate step. This allows for more natural viewing of the data in that the viewer can get a "bigger picture" alongside the selected expansion.

Example of brushing expansion



The map panel carries with it quite a bit of data and insight. It is zoomable and draggable; it displays neighborhood names when the user hovers over each region; it visually encodes aggregate bike share data in bubble sizes; and subtly but most impressively, each bubble reveals tooltips upon being clicked and each tooltip shows a selected metric over the course of a week, which is bounds and leaps away from our intermediate design's map. That was a mouthful... or perhaps a "mapful."

Example of map tooltips



Lastly, the legend and the large toggles compared to the radio buttons we used originally, bring this final project together aesthetically and make our website much more welcoming than our intermediate implementation was.

8 Evaluation

Our map is really, really nice and displays data in several different ways and on several different spatial and temporal scales, which allows for a lot of exploration by the user. The line graph is quite nice for what it is. It does a bang-up job comparing the three cities in a concise way and even raises questions with a few anomalies we address in our section on telling "stories" on our website.

Before the negatives, we would like to say that we are thoroughly satisfied with our visualization's final manifestation, functionally and aesthetically.

There were a few things, however, that we really wished to do from the beginning but were unable to do properly given the time constraints. Two of these regrets are related to the timescale over which the data is averaged and displayed:

First off, we wanted to use the map to show the usage patterns at each hour of a given week. The original idea was to incorporate a slider which could show your metric of choice at, for example, 3:00 PM on Sundays. The ability to see the bubbles grow and shrink depending on day and night or morning and afternoon would give some insights into the way a city's bike share system is really used. We settled for displaying data averaged over each weekday, but to us this felt like a less satisfying and enlightening timescale.

Related to the previous issue, we realized that our data would prove difficult to work with due to its sheer size. We attempted to assemble a dataset that would allow for the creation of a visualization like the one just described, but to iterate through the larger dataset and produce a working dataset would have taken about 15 hours or more for each city. We decided that our time would be better spent focusing on other aspects.

Lastly, and perhaps most unfortunately, the maps do not compare different cities bike share systems at all, and therefore veer away from our original goal. Each map refers to just one city, even if the the information presented is extensive.

All the above are considered to be potential improvements.

What we learned about our data is largely summarized in the stories section on our website. But here are a few things: a really relevant factor that we did not anticipate was the age of each bike share system. Chicago's statistics were at times strange or unexplainable, but that was largely because of the small number of rides, especially immediately following opening. Further, we hoped that the speed metric would tell us something about how fast a traveller can get from point A to B in different cities. For example, we expected Chicago's average speeds to be on average higher than Boston's and Washington D.C.'s to be still higher on account of the curvy road plan hindering fast street travel in Boston and the more gridded street plans speeding things up.

Hopefully you click around our site and find something you find intriguing. So happy hunting... producing this was a reeeeal trip!