Lyla Fadden
Micah Lanier

CS 171: Final Project Proposal
MBTA GameDay

Background & Motivation

We are members of a four-person team of Harvard students working alongside MBTA staff to analyze and better predict demand for public transportation. The primary goal of that project is a machine learning model to predict subway demand, accounting for recent past ridership, seasonal/cyclical effects, and sporadic events like severe weather and major events. When building these models, we realized that sporting events in particular precipitate acute changes in demand that we can detect at key stations before and after games.

We wish to use our insight about sporting events (Bruins, Celtics, and Red Sox games in particular) to visualize MBTA traffic flow by line and station (for stations proximate to venues) before and after games. We will allow users to intuitively select sets of games—likely by brushing and providing filters for non-contiguous games (e.g., weekend games) or particularly interesting sets of games (playoff games in particular). Beyond visualizing the specific games selected, we will provide useful visual comparisons to other events (so that users can detect notable games) and to days without games.

Note that our work with the MBTA is part of our capstone course (AC 297r), but this project is *not* a required component of that work.

Project Objectives

As suggested above, our primary objective is to help users understand how MBTA traffic changes before, during, and after major sporting events. Fans travel to games from different parts of the city and surrounding area (in the latter case, they often travel from far-flung T stations like Riverside and Alewife). They might come early or stay late in order to spend time in the surrounding neighborhood, and they might change in number or behavior for special games—say, for Patriot's Day or a late-season series against the Yankees.

Beyond simple entertainment, planners might find this tool useful for anticipating large crowds or timing shifts and train runs to best accommodate swells of riders along specific lines. For this reason, we hope to add a measure of precision to our interface, reporting relevant statistical

observations with our pretty pictures. We have found these traffic patterns quite difficult to analyze, and we will consider our project a success if it is not only aesthetically pleasing, but useful as well.

Data

There are two important data sources that we will use: MBTA traffic data and sports schedules.

We have MBTA traffic data going back to 2013. It shows entries (and exits) for every subway station and above-ground Green Line route in 15-minute increments. Some example raw data for April 1 of last year:

locationid	servicedate	servicetime	entries	exits
1002	2014-04-01	500	5	1
1002	2014-04-01	515	81	8
1002	2014-04-01	530	60	15
1002	2014-04-01	545	68	13
1002	2014-04-01	600	90	15
1002	2014-04-01	615	135	23
1002	2014-04-01	630	121	17
1002	2014-04-01	645	90	26
1002	2014-04-01	700	117	11

We also have useful metadata with information about each station, including the lines that each station serves and latitude/longitude measurements for geographic plotting (should we decide to use it).

Additionally, we have sports schedules for Bruins, Celtics, and Red Sox home games since 2013. All were manually scraped from ESPN schedule listings. A few example rows for Boston Bruins games in 2014:

Datetime	Playoff	Opponent
2013-01-19 19:00	0	NY Rangers
2013-01-21 13:00	0	Winnipeg
2013-01-25 19:00	0	NY Islanders
2013-01-29 19:00	0	New Jersey
2013-01-31 19:00	0	Buffalo
2013-02-12 19:30	0	NY Rangers
2013-02-28 19:00	0	Ottawa
2013-03-02 13:00	0	Tampa Bay
2013-03-03 19:30	0	Montreal

We hope that opponents may prove to be an interesting dimension for analysis. Fan reactions to the Yankees' or Canadiens' arrivals may cause particularly interesting traffic patterns to emerge.

At this point, the only thing missing from our dataset is Red Sox playoff games. We expect that those will be easy to retrieve manually.

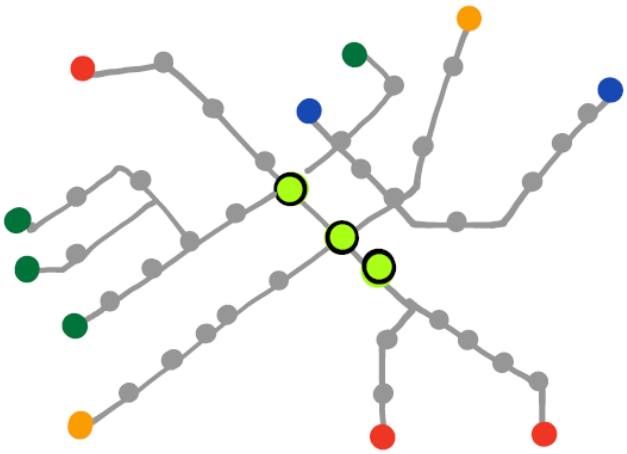
Data Processing

All of our data is very clean and low-dimensional, so we do not expect to have to contend with serious cleanliness or missing data issues. However, our MBTA is granular and voluminous—it spans several hundred megabytes and billions of individual records—it is wholly unsuitable for web visualization in its raw form. For visualization, we will have to summarize it into a compact form that still retains information pertinent to our display.

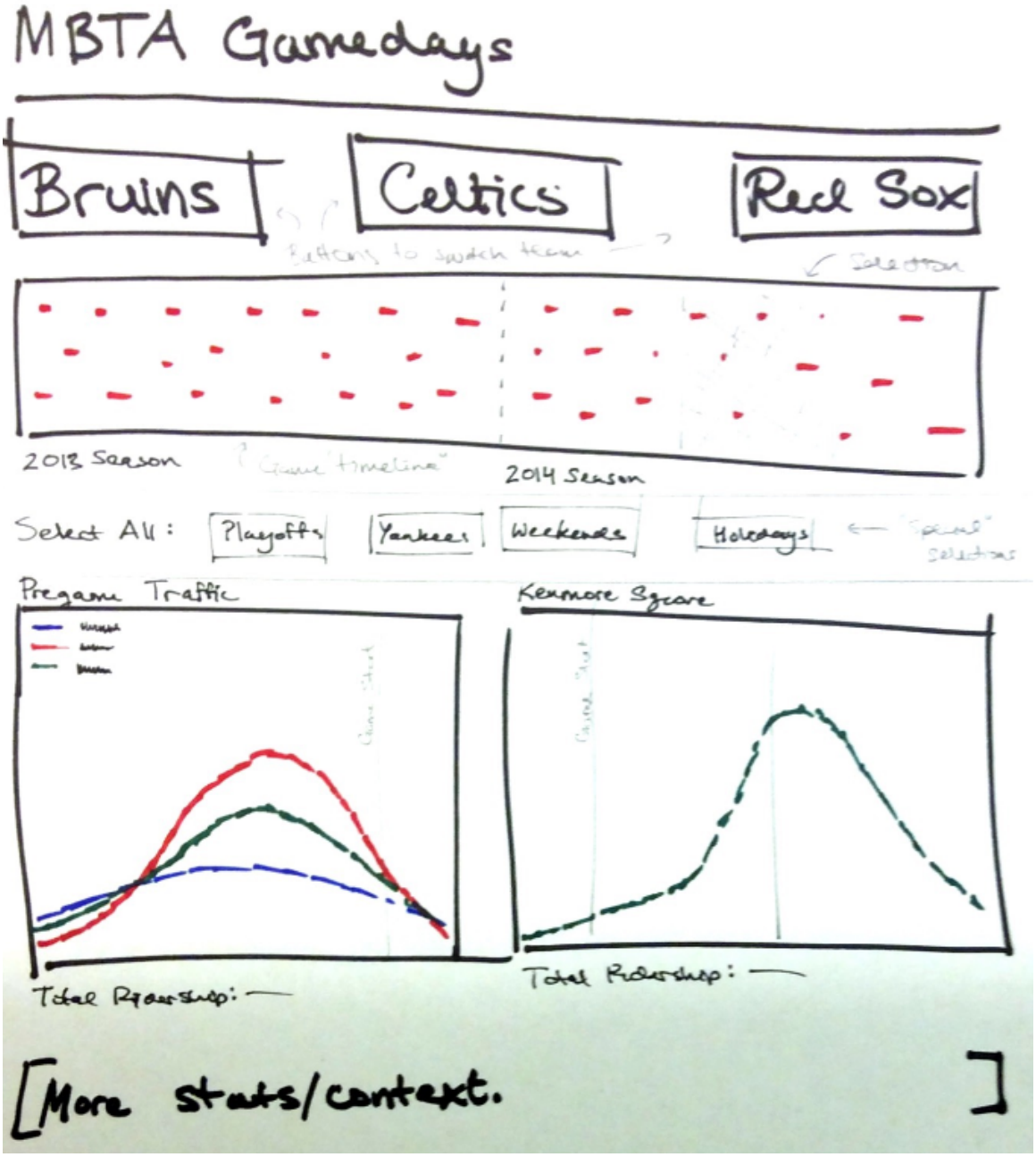
Having spent some time working with the data, we expect to be able to summarize game day traffic by line (for pre-game visualization) and by stations close to Fenway Park and TD Garden Arena (for post-game visualization). We can incorporate comparisons to average traffic in those calculations and avoid including that data in our final deliverable. This way, we can essentially "throw away" data irrelevant to gamedays themselves. Back-of-the-envelope calculations suggest that this will yield observations in the thousands—not a small amount, but an amount suitable for visualization.

Visualization

The selection of rail lines and stations could be done by providing a map representation of the MBTA rail system with the ability of selecting single or multiple stations or full lines:



The basic workflow will include a method for selecting the type of game (Bruins, Celtics, Red Sox) and the game(s) of interest. The game schedule can be displayed in a time frame indicating date/season on the horizontal axis and game start time on the vertical axis. The user can select a single game of interest or multiple ones by brushing. A different way of selecting groups of games for analysis is by selecting predefined filters including interesting sub-selections of game types such as playoffs, weekend games, and holidays. An additional grouping based on opponent will also be available. Visualizations of pre-game and post-game MBTA entries based on all selections and relevant statistics will provide a way of examining and understanding traffic trends.



Must-Have Features

In order to realize our project objective, several important features will need to be included in our final product as outlined below:

- Method for viewing and selecting specific rail station(s) of interest
- Method for viewing and selecting specific rail line(s) of interest
- View of game dates (schedule) for each type of sporting event (Bruins, Celtics, Red Sox games) including opponent information
- Method for selecting game or set of games in time range of interest
- Visualization of pre-game traffic by station or line for selected games
- Visualization of post-game traffic at stations near the sporting event for the selected games

Additional components of the traffic visualizations include:

  - Indicator of starting/ending time of the sporting event taking place where appropriate
  - Comparison of average traffic to game day traffic

Optional Features

After observing some final projects from last year, we realized that it might be interesting to add a "narrative" element to the visualization that highlights notable events that precipitated interesting demand patterns. But we do not yet know what those events might be (but using our visualization might help us find them!), so we do not know if such a feature would be valuable.

We also hope to make use of live updating as much as possible (e.g., charts will transform depending on brush extents). However, due to data volume this might not be possible to do elegantly. We will have to determine feasibility once we have a better handle on our data.

Project Schedule

Week of April 6:

- Gather final missing data (Red Sox playoff dates).
- Solidify main visualization page components.
- Begin page design and component layout.
- Summarize game day traffic for all games.

Week of April 13 (includes Milestone 1):

- Implement basic visualization framework, primary page elements (charts of traffic by line, traffic at nearby stations).
- Improve cosmetic appearance of visualization.

Week of April 20:

- Add ancillary interactions (hovering effects, additional contextual information).
- Add additional metrics and visual elements depending on fit with elements developed. These may include histograms, confidence intervals, summary statistics, etc.

Week of April 27:

- Polish visualization appearance (typography, artwork, colors, etc.).
- Potentially: add narrative page elements.

Week of May 4: By this point, we only anticipate minor polishing work before final submission.