Exploration of The 2016 RODS (Rolling Origin and Destination Survey) London Underground Data Sets

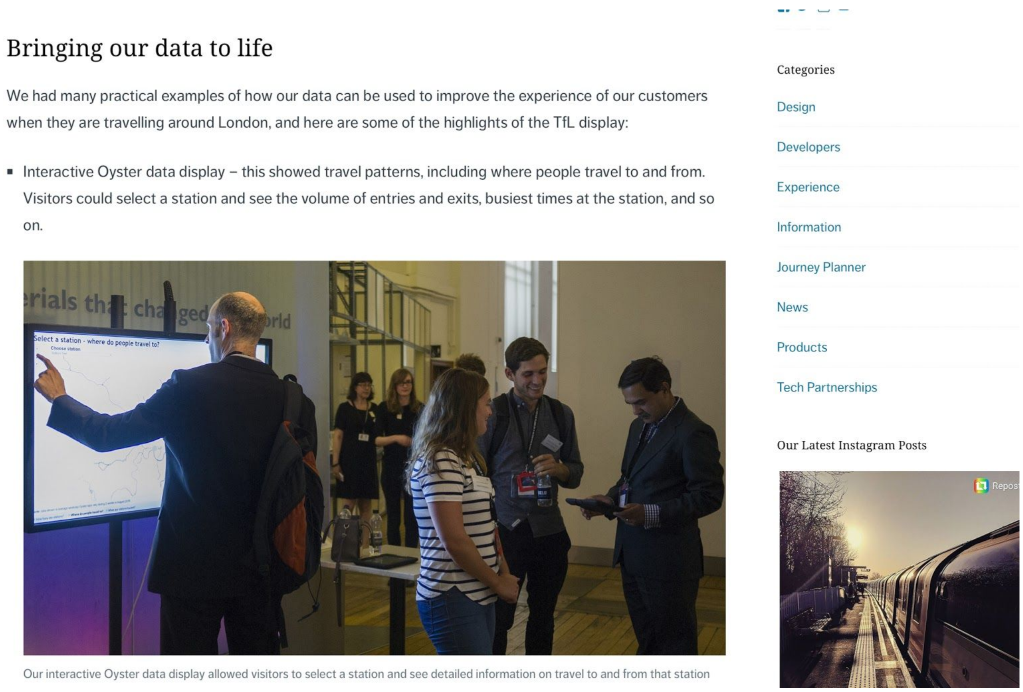
CS 5630 Final Project

By: Jonathan Bown, Andrew Yang and Pablo Napan

Inspiration:

The London Underground is a fascinating transportation system. London is one of the largest cities in the world yet, the TFL (Transport for London) manages to maintain a state of the art system that gets people where they need to go. Experiencing it for myself was somewhat of an inspiration for this project because having used other transportation systems such as the metro in Paris, France the Tube seems like a luxurious way to travel in comparison. But its not just the fact that its nice, the TFL is well known for being very data driven in their decision making and as a consequence of that philosophy they open a lot of their data to the public. The developers have access to the API for their apps but the London DataStore has a treasure trove of public transportation data waiting to be analyzed by anyone who has a peaked curiosity.

The TFL recently held an exhibit at the Science Museum in London which I (Jon) was able to visit during my time there. There was one visualization they had that blew me away. It showed something similar to what our visualization attempts to convey which is what does a day look like on the Tube? It was amazing to see the circles on each station grow and shrink as time progressed. The other display they focused on was the Oyster card structure. The Oyster card is similar to a UTA pass but a little more exclusive. The showed what the chip looks like in the card and what kind of data it collects. Just from that one card the TFL collects everything it needs to know about its passengers in order to improve the service. (Picture below of the interactive display)



Bringing our data to life

We had many practical examples of how our data can be used to improve the experience of our customers when they are travelling around London, and here are some of the highlights of the TfL display:

- Interactive Oyster data display – this showed travel patterns, including where people travel to and from. Visitors could select a station and see the volume of entries and exits, busiest times at the station, and so on.

Our interactive Oyster data display allowed visitors to select a station and see detailed information on travel to and from that station

Categories

Design

Developers

Experience

Information

Journey Planner

News

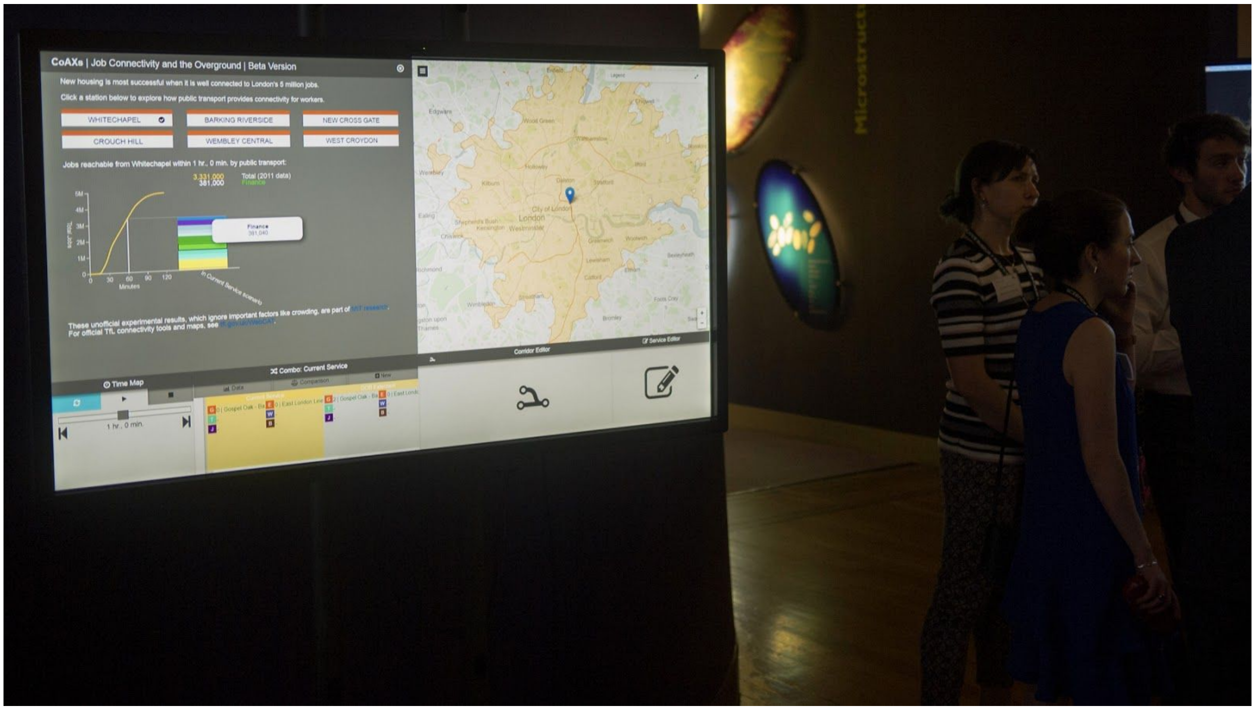
Products

Tech Partnerships

Our Latest Instagram Posts

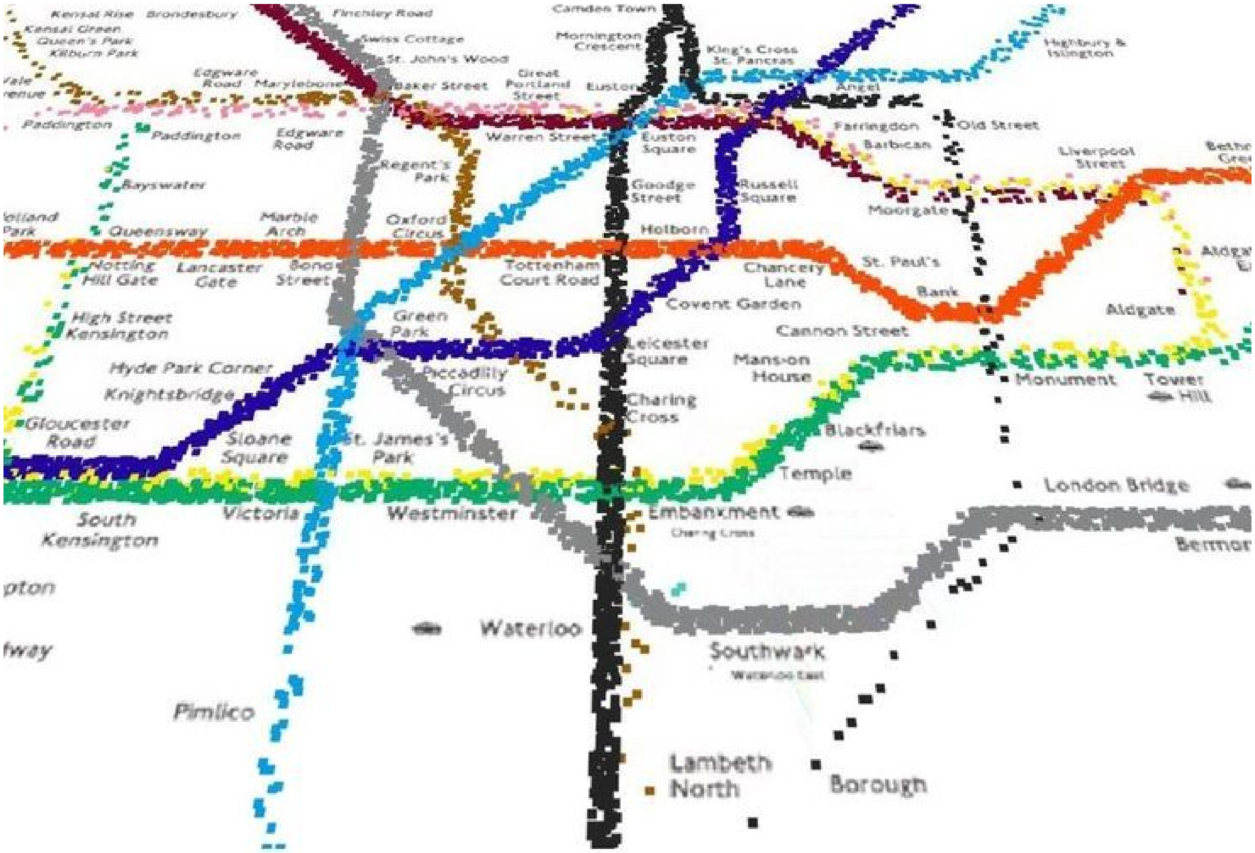
erials that changed world
Select a station - where do people travel to?
Choose station
Repos

One of the main visualizations at the exhibit.


CoAXs | Job Connectivity and the Overground | Beta Version
New housing is most successful when it is well connected to London's 5 million jobs.
Click a station below to explore how public transport provides connectivity for workers.
WHITECHAPEL
BARKING RIVERSIDE
NEW CROSS GATE
CROUCH HILL
WEMBLEY CENTRAL
WEST CROYDON
Jobs reachable from Whitechapel within 1 hr., 0 min. by public transport:
3,331,000
381,000
Total (2011 data)
Finance
Total Jobs
5M
4M
3M
2M
1M
0
0
30
60
90
120
Minutes
Finance
381,040
in Current Service scenario
These unofficial experimental results, which ignore important factors like crowding, are part of MIT research.
For official TfL connectivity tools and maps, see tfl.gov.uk/WebCAT.
Combo: Current Service
Time Map
Data
Comparison
New
1 hr., 0 min.
Corridor Editor
Service Editor
London
City of London

As we were looking for ideas of something to do my mind went back to that display and how powerful it was and we started delving into the London Datastore to see what's there. We also looked into see if anyone has done anything with D3 to visualize this data.

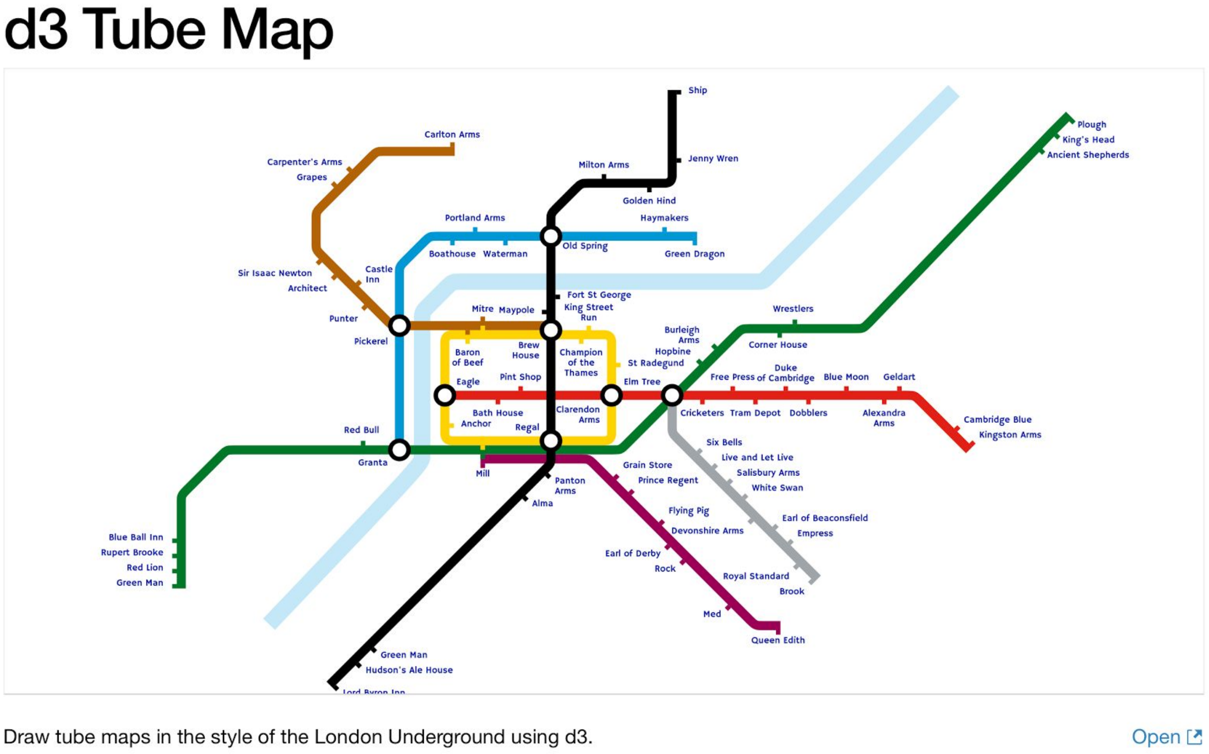
The example pictured below featured a dataset that contains oyster card data tied to time of entry and exit. This dataset was one of the first we downloaded and started looking at. So from that csv file someone made an animation of the dots representing the journeys and when they are aggregated they come together to form a nice map of the tube.



There was another unique one that we found that was a 3D real time tube tracker. Not only can you see the trains in the 3D london underground system as they are moving live. But the sound in the background almost takes you there if you have or haven't experienced the actual London Underground. This came in later when deciding to add sound.

After a lot of searching for what we could use as a structure for the data, we could not find anyone that had done a visualization using D3 and the tube map that the average, everyday user would see. There wasn't even an existing JSON file that held the coordinates for the stations. We did find someone that had constructed a package for drawing subway systems in New York and London called d3-tube-map. The user had constructed a map for the system at Cambrige (image below) but this was the perfect starting point for seeing if we could pull this off. This one is by John Wally.

Source: https://bl.ocks.org/johnwalley/9b6d8af7a209b95c5b9dff99073db420



d3 Tube Map

Draw tube maps in the style of the London Underground using d3.

The point here is that this style of visualizing the data hasn’t been done before using D3 or the subway-like map of the Tube that commuters see.

The TFL conducts what is called a Rolling Origin and Destination Survey which is just data collected on a rolling basis to get exit/entry counts reconciled every November and then pushed out to the public. In 2016, the TFL went above and beyond this normal data collection. They collected everything from how did you get to the station (car, bus, bike) to what route are you taking and why? This produced enormous amounts of interesting data that we could potentially use in the visualization. While there is still much more to uncover, we focused on a few key data sets.

Questions:

Overall we just wanted to create something unique, and that could serve a purpose for various types of interests in the data. The goal we were shooting for was basically how do we get a snapshot of what's happening on the tube everyday? What does the system flow look like? What kinds of people are riding the tube and from what stations? What is the average journey time and how does that compare with other categorical variables such as age and purpose of riding? What are the most common routes? Where do people spend most of their time? These were the questions we discussed and ultimately could have answered all of them (time-permitting) because the amount of data we obtained was enough to cover anything a researcher could think of.

But we decided to focus on a few:

- What does the demographic that use these stations look like?
- Where do people come from and go that use these stations?
- What is their average journey time?
- What does a day look like on the tube and how can we show that in an interactive way?

If we were to choose a target audience for the visualization, we thought first of a marketer interested in what stations to target with ads and when. Things like what demographics commonly enter each station, how long are they riding for, what is their reason for riding. Advertising in the London Underground stations is big business that needs lots of research.

Data: Source, scraping method and cleanup

Source: London DataStore https://data.london.gov.uk

Scraping method: Downloaded CSV

Cleanup: Every dataset described below.

RODS Rolling Origin and Destination Survey 2016.

This dataset was relatively well organized and fairly clean, however a big challenge was to uniform all the station names since many station names did not match with other datasets. The problem at first seemed to be the difference between the number of spaces between the name components but later we found that other group of names contained special characters and others were even hand typed with misspellings.

Counts

This dataset comprised the number of entries and exits of all stations registered during the period 2007-2016. It was fairly well organized and did not cause too much trouble except for the spacing between the stations name elements.

RODS_2016/Misc/Station Flows/AEI - Summary

This dataset has 15 minute intervals for in London talk 'A' for entry, 'E' for exit and 'I' for interchange. Every 15 minutes a station has a count for people entering leaving and interchanging. This worked perfectly for the main visualization.

Exploratory Data Analysis:

We knew the dataset well and it was structured in a manner that we did not need to make use of exploratory visualizations. However, getting the minimum and maximum values of the different attributes helped us decide the radius of the circles, the scale types we would use, container overflow, etc.

Design Evolution: What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

VIsualizations considered:

U-shaped tube: For the tube transformation, we first considered building a U-shaped tube because these can contain many stations and easily overflow its containers, however, we dropped this idea and decided to design a scrollable straight line which is visually less misleading, more accurate and coherent.

On-the-fly design changes:

Google maps: One of the biggest novelties in our design was the incorporation of google maps. Using the google map API we were able to include heatmaps and bubble maps, which helped to give a realistic element to the visualization.

Implementation:

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

1. *Map Window-Tube Map view:*

The visualization system starts with the a map of the UNDERGROUND system lines (fig. x) which represents pictorially every line. The hardest thing about this view was that the json file for the coordinates had to be built by hand. Nothing like it could be found at least in the research we did. But this gave us the flexibility to know exaclty where things can be appended, where information can be derived form, and where errors arise.

Functionality:

a) By selecting entries or exits in the drop down menu (upper-left corner) and clicking the ‘A day on the tube’ button, it will display the time evolution of the entries and or exits of all the stations in the transportation system (fig. x).
b) By clicking in any line, it will display the *Map window - Stations statistics.* For explanation, see item 2 below.

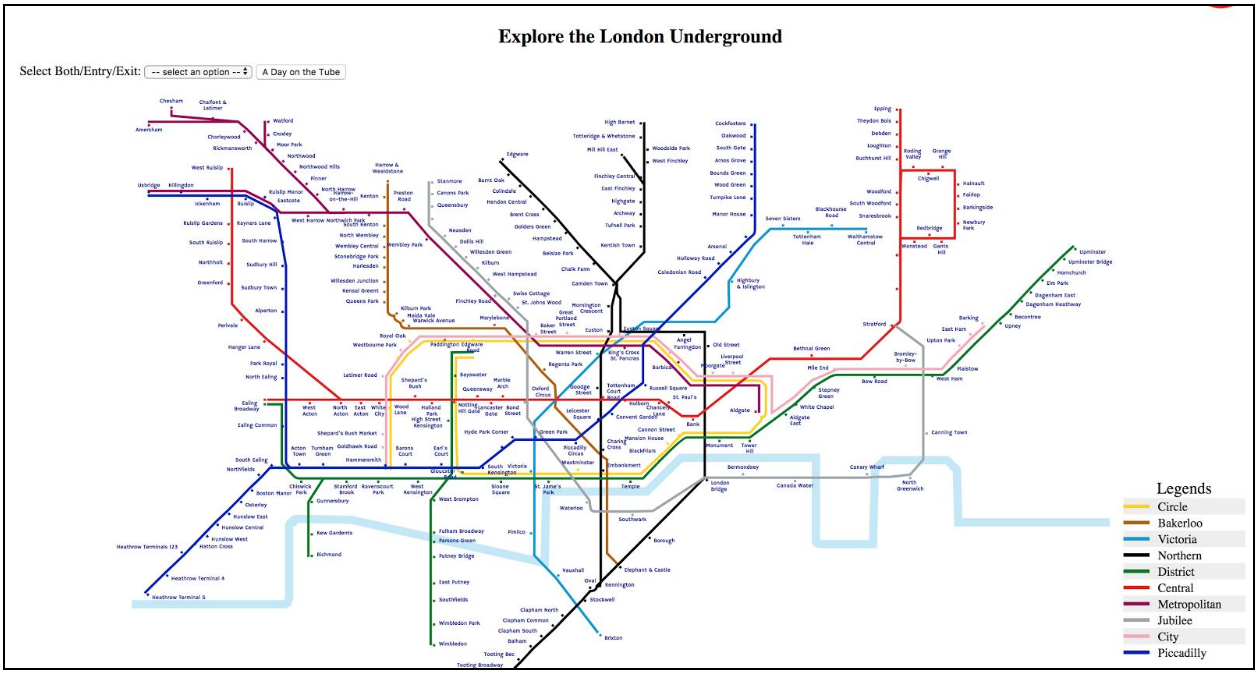
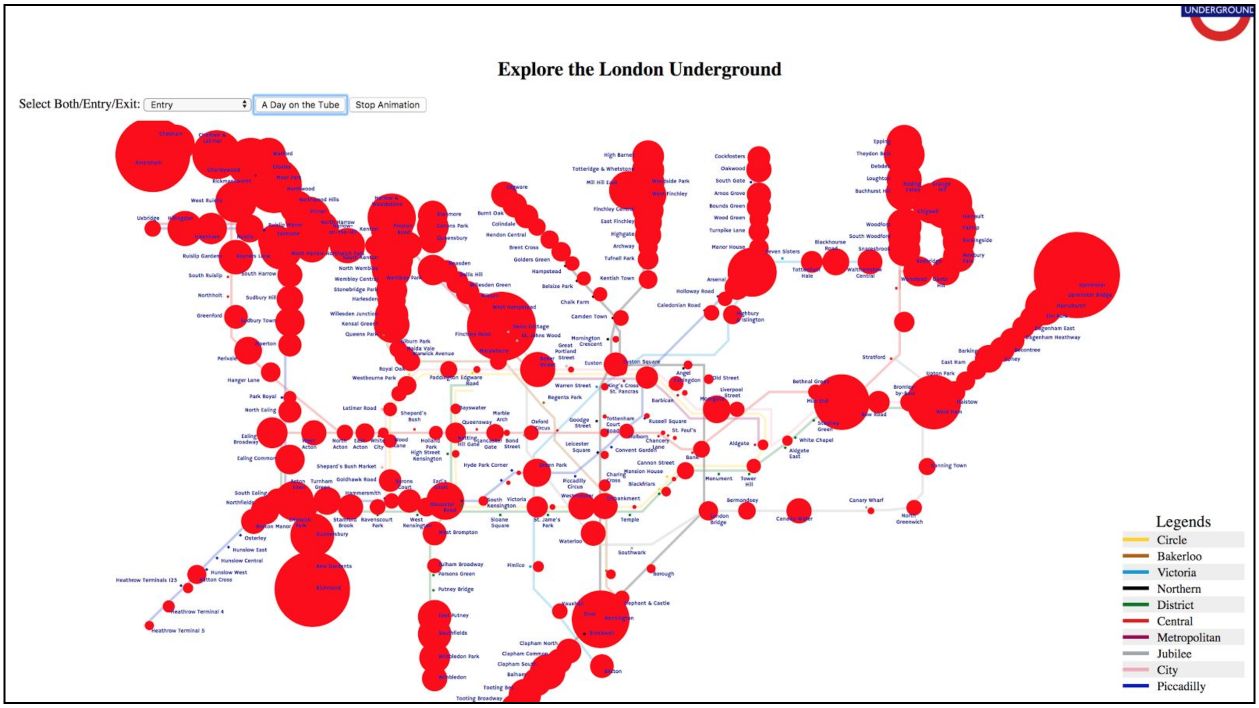


Fig 1. Tubemap



Explore the London Underground

Select Both/Entry/Exit: Entry

Fig 2. Time animation

2. *Map window - One-line-stats view*: Analysis of station statistics of a specific line

After clicking in one line/tube the line will get stretched into a flat line. In this view the data (age, sex, journey time, market segment) of the stations can be analyzed and compared.

Functionality:

a) By clicking a station (circle), 4 barcharts will be displayed, each one corresponding to the data categories listed above.
b) Next step, by selecting a station in the dropdown above, the barchart will turn from a single-bar chart to a grouped-bars chart. In case an error shows up, this means that there is no data for that specific station and hence cannot be displayed.

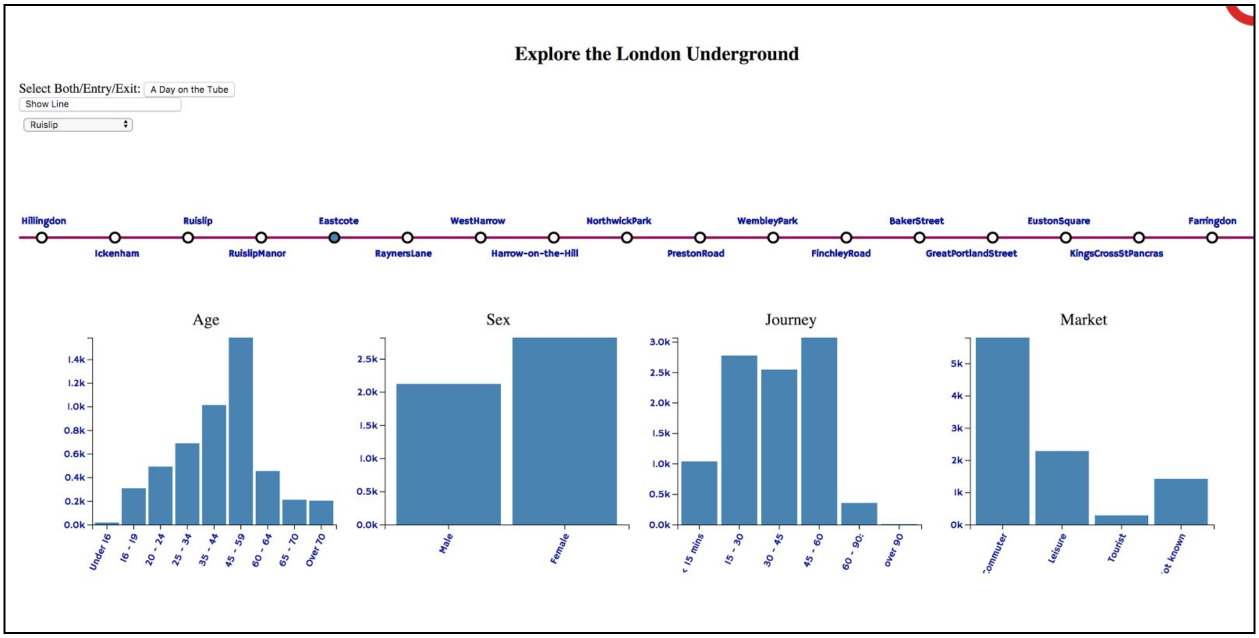


Fig 3. Statistics for one station

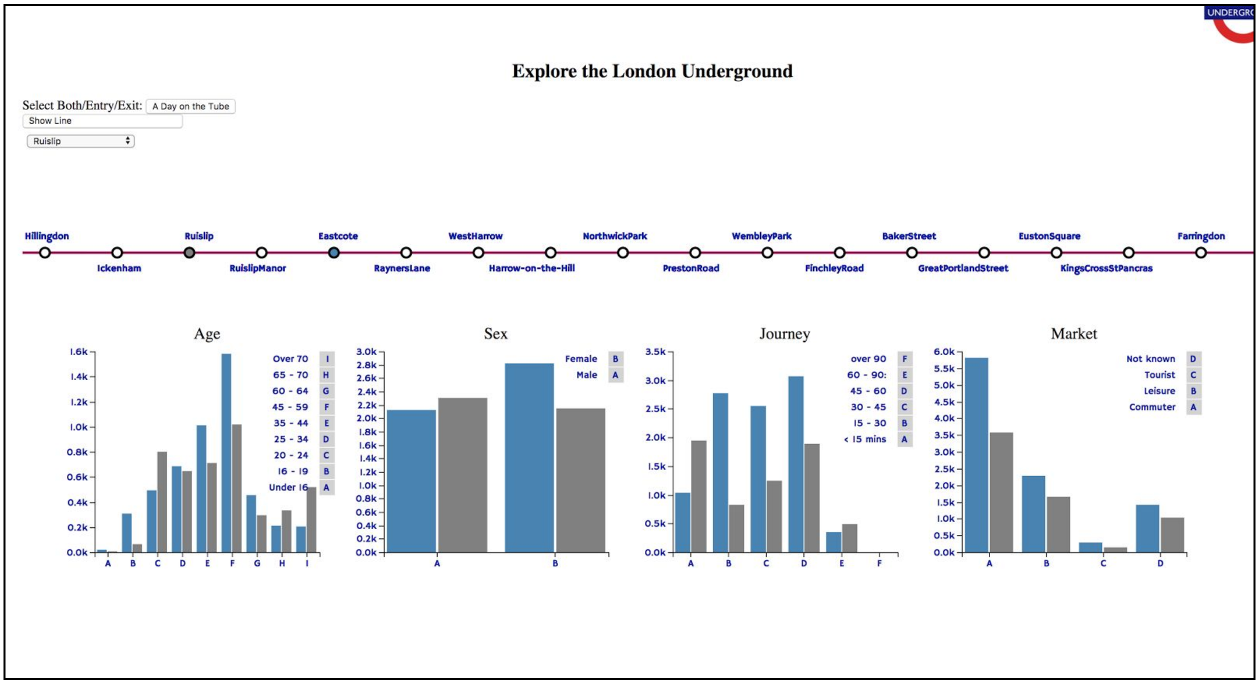


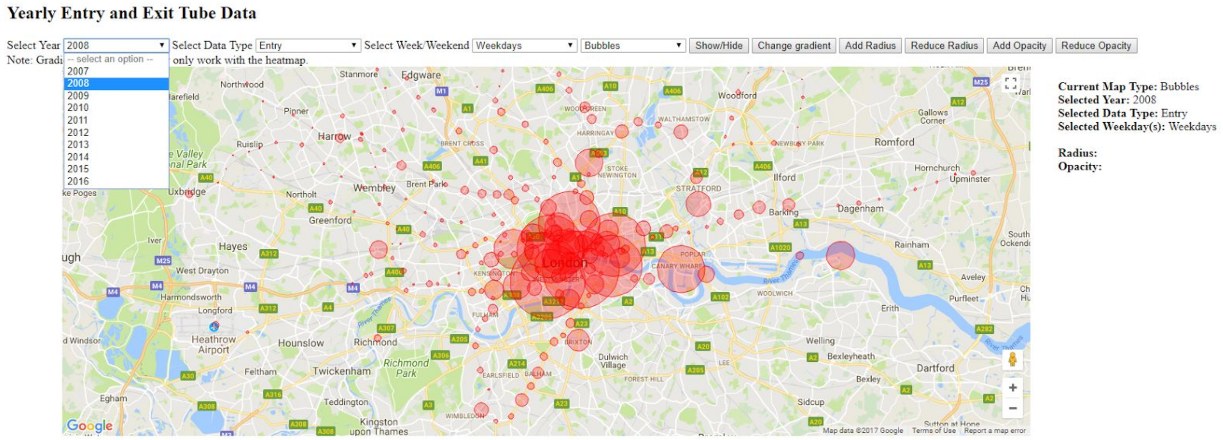
Fig. 4. Statistics for 2 stations

3. *Google-map Window:*

On a separate web page, we created two maps using Google Map's public API. These maps are centered over London, and allow us to create heatmaps and bubble maps that showcase certain statistics about the London Tube system. Above the maps, we have interactive drop down menus and buttons that lets one select what data set one wants to see in different visualization types. Additionally we have options to adjust the gradient, opacity and radius for the heatmap.

Functionality:

a) Select year,data set and visualization type using drop down menu..
b) Optionally change radius/opacity/gradient of heat map by clicking on buttons



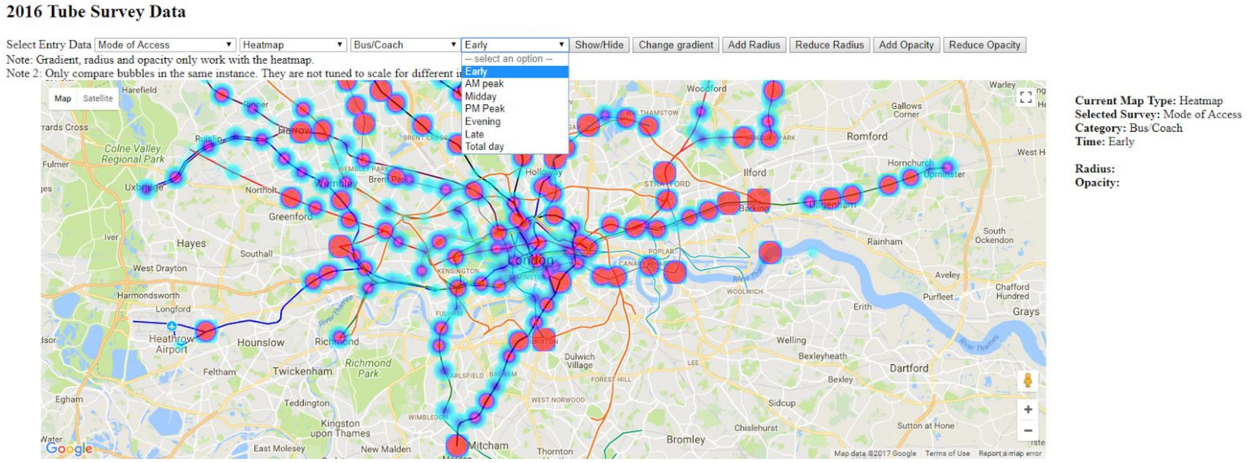
Yearly Entry and Exit Tube Data

Select Year 2008 Select Data Type Entry Select Week/Weekend Weekdays Bubbles

Note: Gradi… only work with the heatmap.

Current Map Type: Bubbles
Selected Year: 2008
Selected Data Type: Entry
Selected Weekday(s): Weekdays

Radius:
Opacity:



2016 Tube Survey Data

Select Entry Data Mode of Access Heatmap Bus/Coach Early

Note: Gradient, radius and opacity only work with the heatmap.
Note 2: Only compare bubbles in the same instance. They are not tuned to scale for different i…

Current Map Type: Heatmap
Selected Survey: Mode of Access
Category: Bus/Coach
Time: Early

Radius:
Opacity:

Evaluation:

What did you learn about the data by using your visualizations?

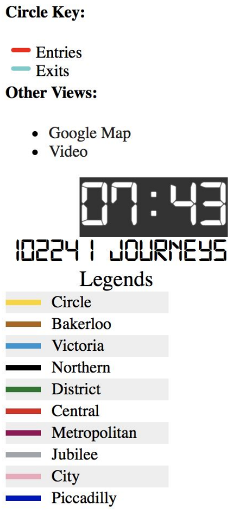
We learned that the London underground is one of the biggest and complex systems in the world. From the time animation we learned that it, unsurprisingly, obeys the schedule of most metropolis where peak times are around 8am and 5pm, the average travel time is 15-30 minutes and most people commute to the city center in the morning and to the suburbs in the afternoon. Analyzing the heat maps we learned a lot about the different modes of transport that are used and how often they are used in certain areas. It gives a good perspective of the change in entries and exits over the weekend. Heatmaps can be extremely useful for categorical data analysis which is what we were trying to go for there with our more categorical data sets.

The maybe surprising fact is that on average, clearly more women use this transportation system.

Our broad question was: What are the changes over time for London's underground tube stations?. We answered this by visualizing statically (maps and statistics) how the demography has changed over the years and dynamically (time animation) how popular the stations become over the day.

The visualization, except for some minor issues, works well. We would improve it by calculating statistics for the whole system

Overall we are happy with how it turned out. I think the main page with the clock turned out really well and gives a good look and feel to the visualization and not to mention the sound. The bar charts/comparison charts are also laid out nicely. Other than a few little csv bugs they work really well.



Circle Key:

- ▬ Entries
- ▬ Exits

Other Views:

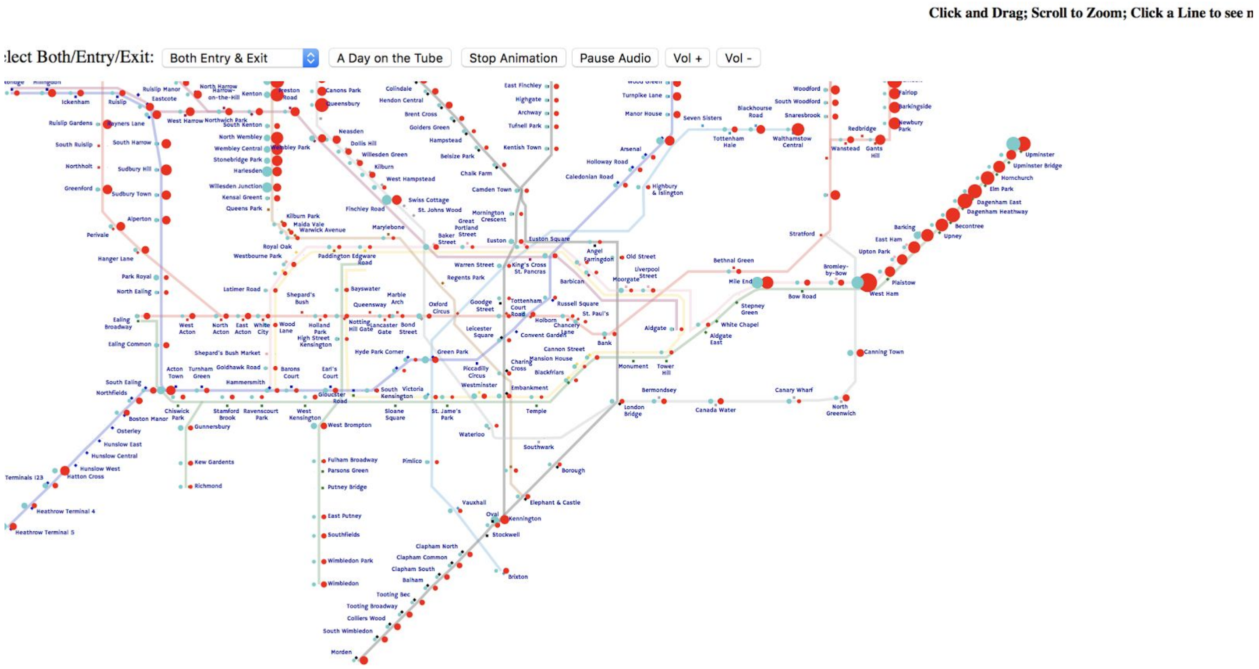
- Google Map
- Video

07:43

102241 JOURNEYS

Legends

▬	Circle
▬	Bakerloo
▬	Victoria
▬	Northern
▬	District
▬	Central
▬	Metropolitan
▬	Jubilee
▬	City
▬	Piccadilly



Click and Drag; Scroll to Zoom; Click a Line to see n

Going Forward:

We still have a lot of data that we didn’t get to but would have liked to work with more. The most clear next step would be to use the Origin-Destination Matrix data set in the RODS survey to build a matrix layout of where people most often travel to and when. We also considered a chord layout for this but it’s hard to make it really interactive with all the different lines present at once and we weren’t sure how to handle the fact that we want to see where people are going and interchanging but with several lines in the chord at once it can become extremely complicated. These data sets are also very large and building them in would take a lot of restructuring of the layout we already have. But this does provide a platform to continue to add information from the 2016 survey to continue to learn from.