Salman Arif
Andrew Chuang
Francesco Scivittaro
https://github.com/salman-arif/CS123-Project

CS 123 Project Proposal

Dataset: Taxi Cab trip data for both New York City and Chicago. These are publicly available via the NYC Taxi & Limousine Commission (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) and the Chicago City Data Portal (https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew).

The NYC data is comprised of individualized trip data for yellow and green taxi cabs. The data is divided into a single CSV file for each month for the years between 2009 and 2016. For an idea of size, each dataset for the months in 2015 is approximately 2gb. Because we have so much data, we plan to either subsample per month or select a specific several months we want to focus on for our project. Within a single month is usually around one million entries, each entry corresponding to a single ride. The information for each ride includes information such as start and end datetime, fare amount, distance, latitude and longitude of pickup and drop-off, and certain years include driver identifying information.

The Chicago data also contains individual trip data very similar to the NYC data. It contains 109M rows by 23 Columns and is nearly 10gb in size. Each row is a single taxi trip. Chicago's data is 2013-current, so we are able to select a matching time period with the NYC data.

Possible Topics/Hypotheses:

Modeling deviations in pricing
Comparisons between cities
Similarities between Chicago "community areas" and New York boroughs
Possibly information about drivers, related to pricing
Comparisons between yellow and green cab (restricted to area) behavior
Do green cabs cheat/bend rules about restricted pickup areas
Changes in yellow cab behavior after green cab service was initiated
Borough preference in payment type, destination, length of trip, fare amount
Tip amount by borough, possibly by driver as well, passenger count, fare rate type
Relationship between disputes and voided trips with other attributes
Differences in street-hail/dispatch trips for green cabs
Behavioral changes in response to events
How each city responds to changes (weather, pricing, public transit)
Comparisons in the behavior of cabs belonging to different Chicago cab companies

Further Possibilities:

Integrate data on weather, public transit (CTA?), Divvy

Explore possibility of obtaining data from Uber
Comparing taxi usage before and after specific events/festivals/MTA closures

Questions:

What is our final product supposed to look like? (report, platform, etc)
Is data cleaning expected to be a large portion of the project?
Can we look at the data from a stat/econ perspective? (make regression models, etc)
How do you recommend we deal with datasets being really large in size? (10gb+)