Analyzing NYC Taxi Data

CS 123 Final Project

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Research Questions

1) Has Uber affected yellow taxi demand after big concerts in New York City? *Uber launched in NYC in May 2011
2) What are the travel patterns of people in Manhattan? How likely is it for someone else to be taking a similar trip as you are?

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

The Data

Uber Taxi and Limousine Commission Freedom of Information Law response

- All trips for Apr-Sep 2014
- Size: 1 Gb, 28 million rows
- Columns used: Pickup date, Pickup latitude and longitude

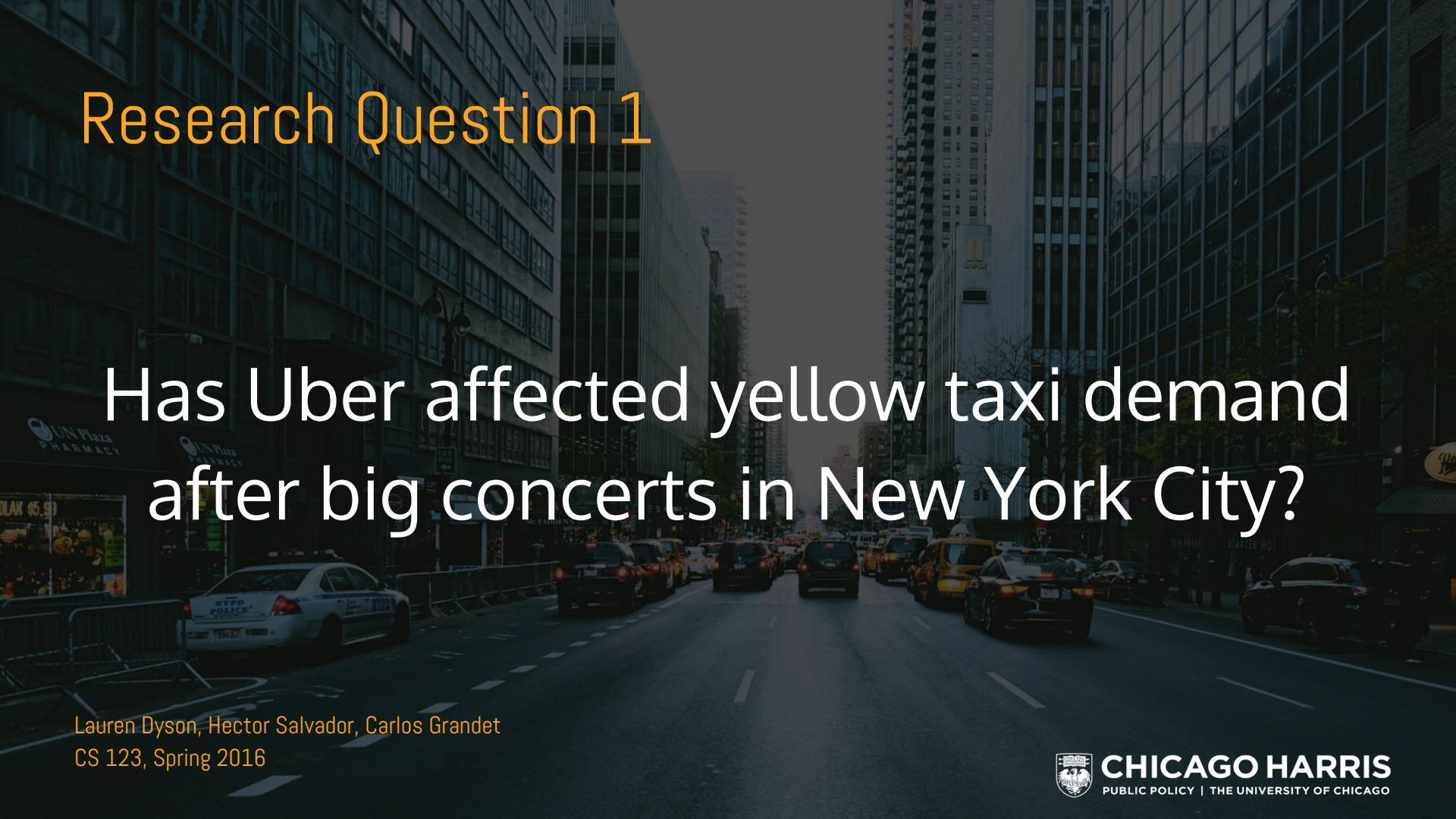
New York City Taxi and Limousine Commission Trip Record Data

- All trips for 2010, 2013, and 2014
- Size: ~65 Gb, ~530 million rows (1.8 Gb/14.7 m rows per month)
- Columns used: Pickup date and time, Dropoff date and time, Pickup longitude and latitude, Dropoff longitude and latitude

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Research Question 1



Has Uber affected yellow taxi demand after big concerts in New York City?

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

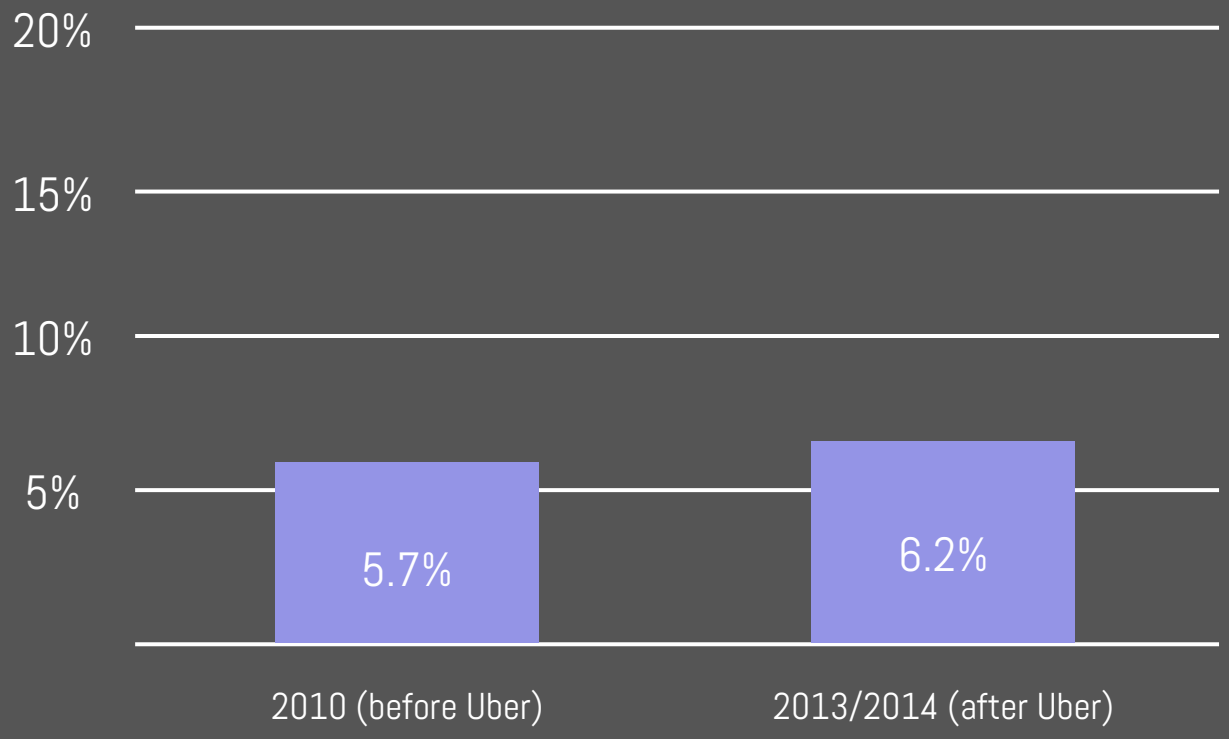
CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Approach

- Scraped data on NYC concerts for Billboard Top 100 artists (BandsInTown API and MusicBrainz API to get gigography)
- Defined time and location windows (within 3 hours after concert start and 0.2 km of the venue lat/long)
- Filtered taxi trips for 2010 and 2013/2014 to count trips within each window
- Compared volume of demand (as percentage of venue total capacity) before and after the establishment of Uber

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

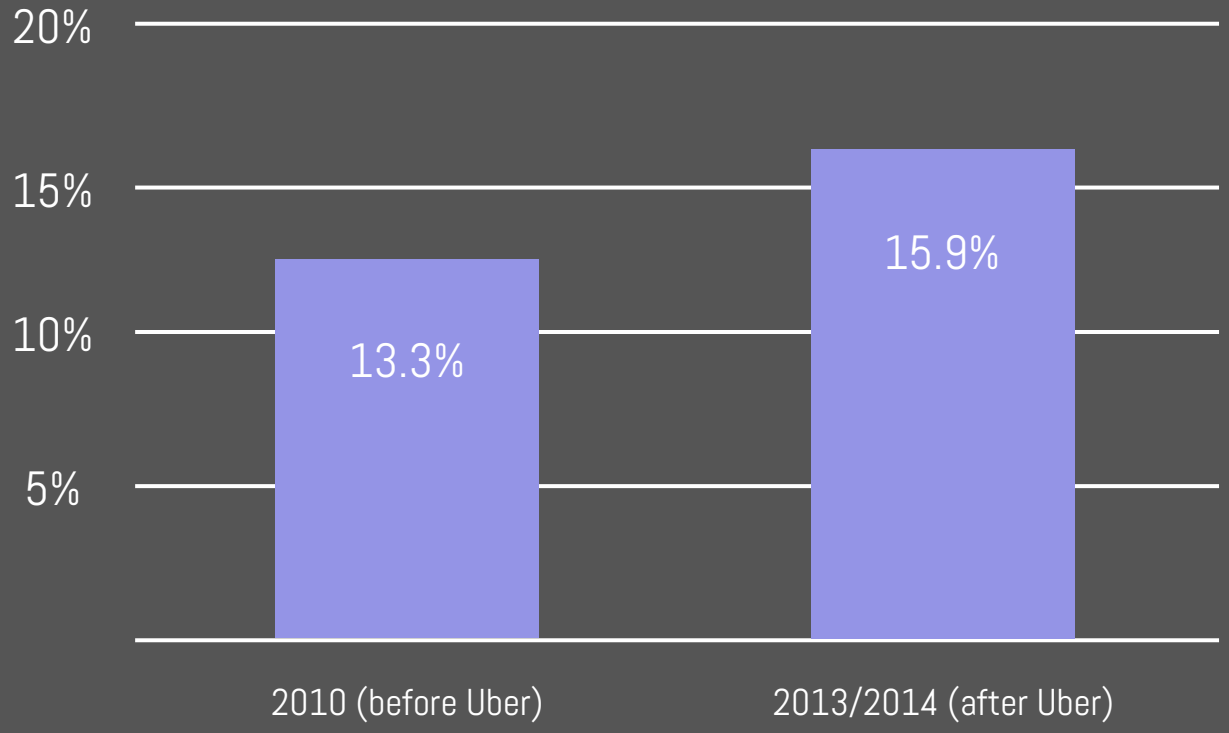
Taxi rides leaving Madison Square Garden within 3 hrs of major concert as % of total venue capacity



Within .2 km of venue lat/long and 3 hours of concert start time

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Taxi rides leaving all venues within 3 hrs of major concert as % of total venue capacity



Within .2 km of venue lat/long and 3 hours of concert start time

Results interpretation

- At Madison Square Garden, differences are rather small so it might be only the effect of picking more/less popular concerts on each year
- Not enough concert venues to make statistical significance tests
- We don't know trends on taxi rides volume over time, so we might be looking at the trend effect
- Next steps could be to run analysis on larger sample of events

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Research Question 2

What are travel patterns of people in Manhattan at different days (weekdays and weekends) and hours (day and night)?

How likely is it for someone else to be taking a similar trip as you are at the same time?

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Approach

- Reduce data to trips that start and end in Manhattan within the specified time window (weekdays, weeknights, weekends)
- For each combination of (pickup, dropoff) X (weekend, weeknight, weekday) run K-Means (using adapted Lab algorithm) to generate 50 clusters
- Analyze the number of trips done between clusters
- K-means is O(n,k,t) -- MapReduce allows us to accommodate large t
- Difficulty running multi-step job on EMR; for preliminary results, we drew a random sample and ran locally

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

Approach cont.

- For each trip starting and ending in Manhattan, determine to which pickup and dropoff cluster does it belong
- Reduce on pickup clusters and break this down into 30 minute increments
- Save into a dictionary the total trips by pickup clusters, time and day
- Calculate the probability (as a relative frequency) of going to any given dropoff cluster at that time from that pickup cluster
- Using MapReduce on EMR

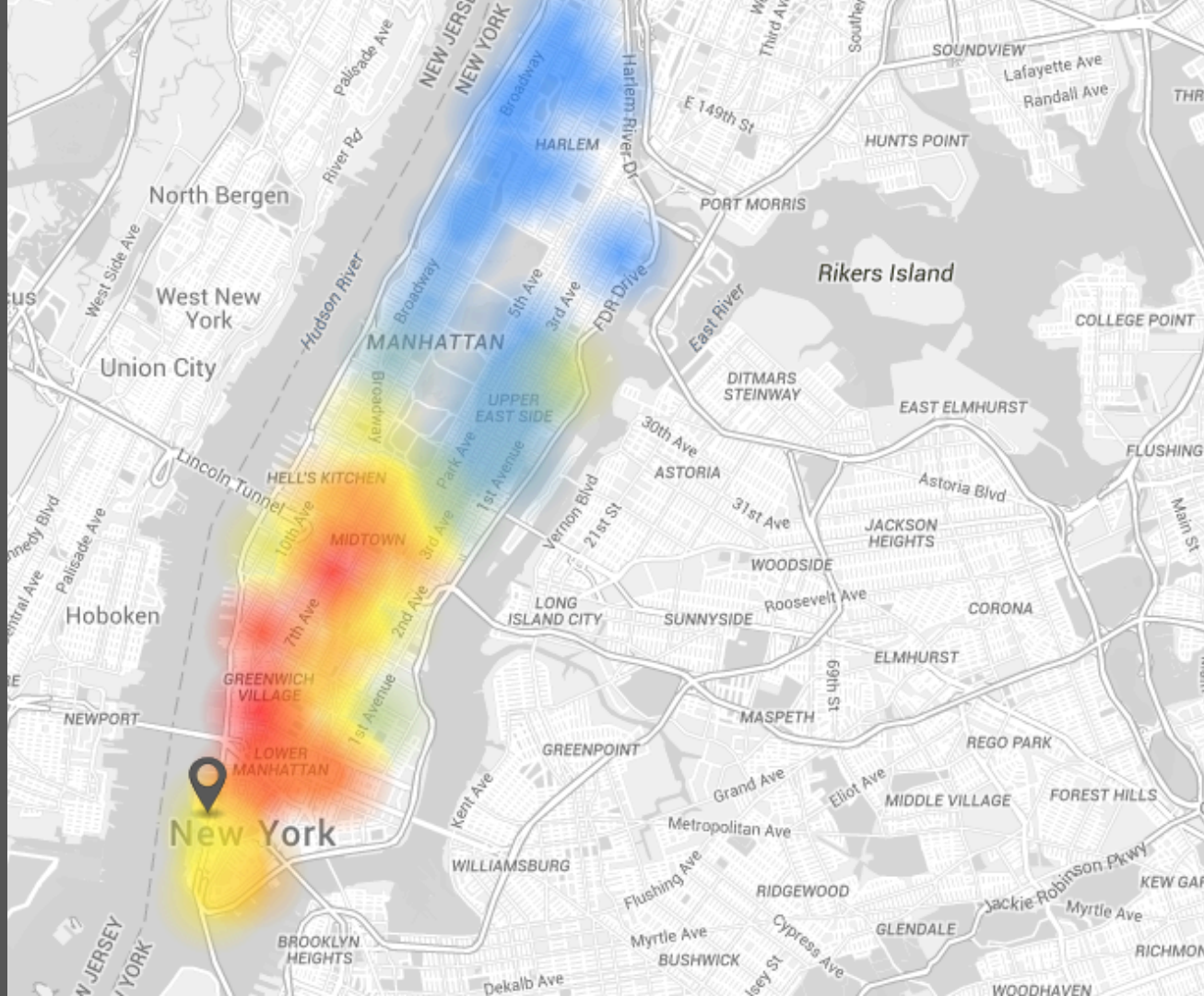
Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

Dropoffs from the Goldman Sachs Office Tower

6-7am on weekdays
Jan. 2015

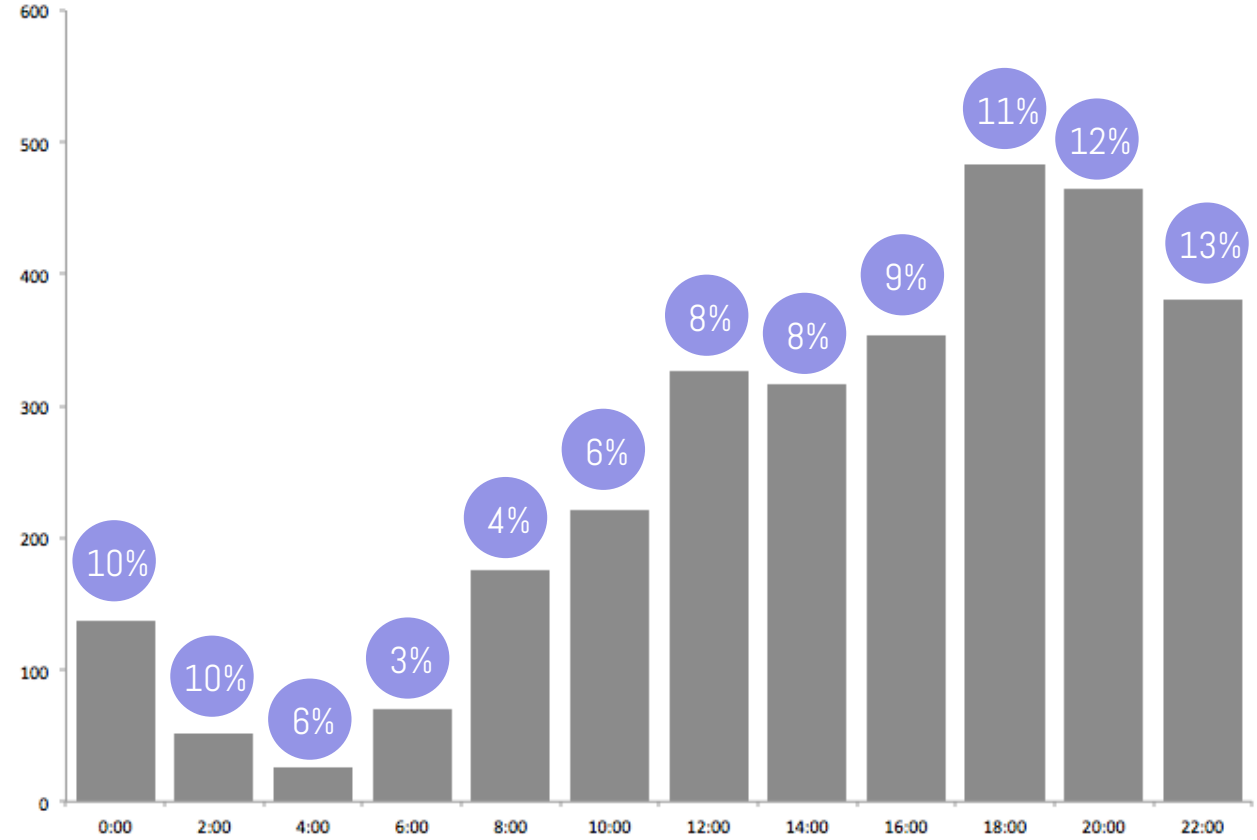
Dropoffs from the Goldman Sachs Office Tower

4-5pm on weekdays
Jan. 2015



Total trips

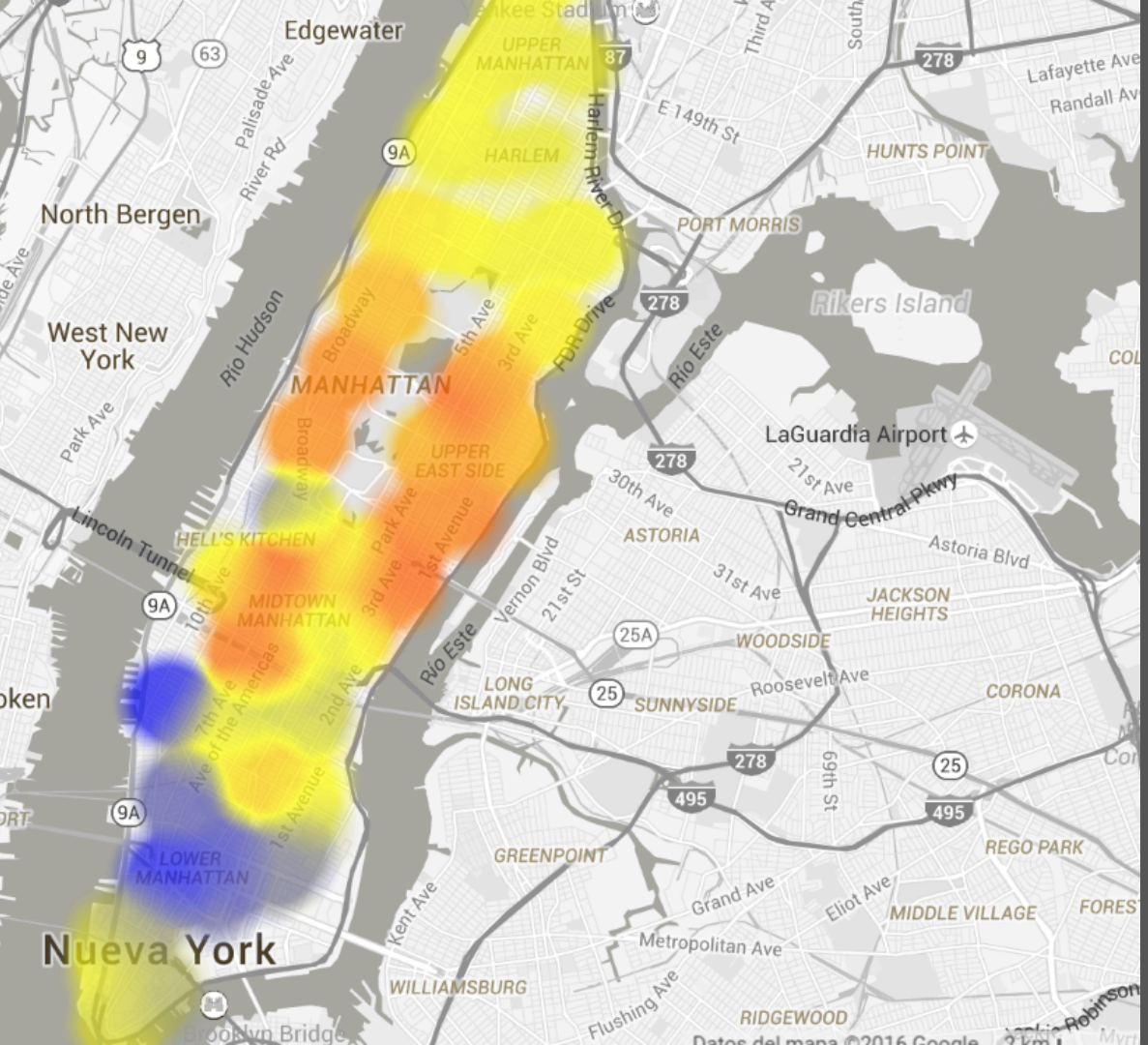
% Percentage of total trips originating from Wall Street that went to the LES at indicated hour



Taxi trips from Wall Street to the Lower East Side

Total rides every two hours

Jan. 2015



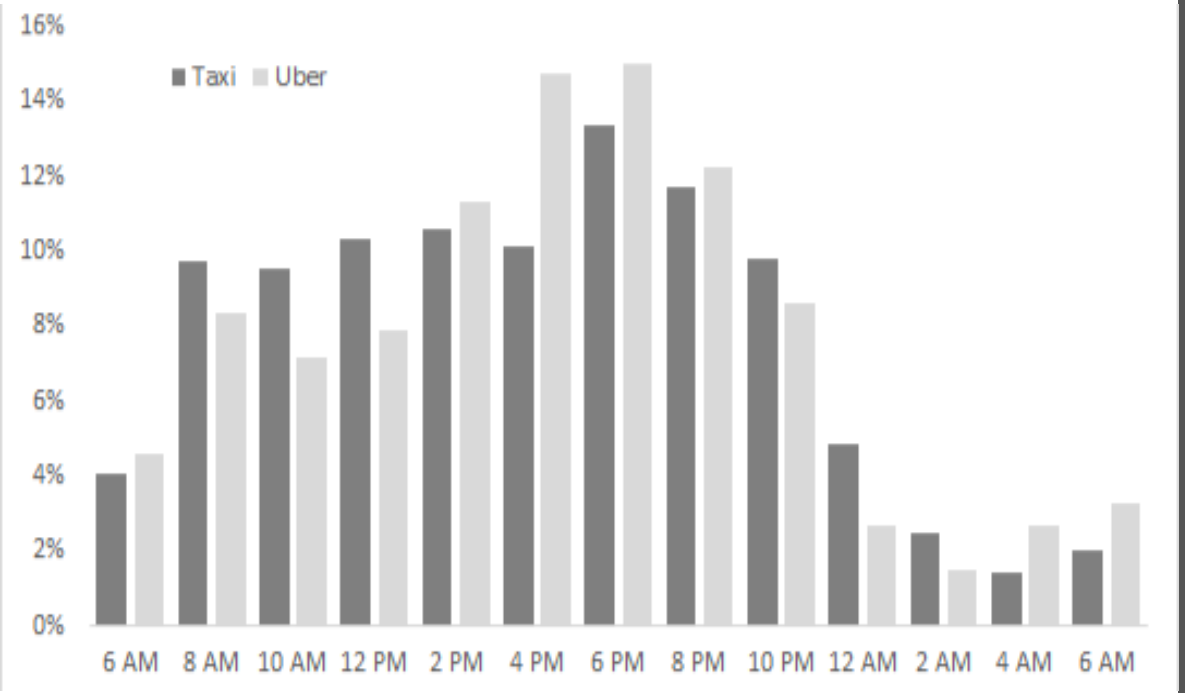
Uber vs Taxis
Pick-up
Spatial Concentration

Jan. 2015

Red: relatively more Taxis than Ubers

Blue: relatively more Ubers than Taxis

Percentage of trips during a weekday



Uber vs Taxis Pick-up Temporal Concentration

Jan. 2015

Policy Implications

- If there is no apparent change in taxi ride volumes after Uber entered the market, is Uber really competing against taxis? Or is it taking demand away from other methods of transport?
- Does available public transportation covers demand between areas with a high volume of trips?
- How do we design better routes for public transportation, adapting to traffic patterns at different times of the day?

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Big Data Considerations

Running times

- Counting taxi rides from concerts for one year (21 GB, 176 million rows):
  - Locally: impossible (aborted with one month)
  - 20 instances: 5 hours 19 minutes
  - 80 instances: 87 minutes
- Assigning trips into clusters and obtaining percentages (2.5 GB, 14 million rows):
  - Locally: aborted with sample size greater than 2 million rows
  - 2 instances: 2.5 hours
  - 20 instances: 25 minutes

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

Challenges

Computational analysis

- S3 buckets: names and upload times
- EMR
  - Long downtime for multistep jobs
  - Inability to run to completion
  - Random "sleeping" periods
  - Hard debugging
  - Saving intermediate results and dealing with generator type
- MRJob: passing in command line arguments when using AWS

Data Analysis

- Inaccurate coordinates from API
- Few concert venues to compare, lack of information about events in New York
- Lack of Uber data for drop-offs and for several years

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

Next Steps & Remaining Work

- Run Task 1 with larger set of comparable events at venues, for example, taking into account the number of passengers in each taxi
- Running the full set of available years (2008-2015)
- Fix k-means algorithm to run on EMR, and create clusters from full dataset
- Modify EMR to save intermediate results into a dictionary and access them
- Run on EMR the nodes clustering for a whole year (instead of only for one month)
- Process results; database with trip frequency every 10 minutes has 500,000 rows

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Lauren Dyson, Hector Salvador, Carlos Grandet
CS 123, Spring 2016

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO