Citadel – Correlation One Datathon

Team 5

Peter Adam – Masters of Business Analytics - UCD
Andy McSweeney – Masters of Business Analytics – UCD
Glenn Moynihan – PhD Theoretical Physics – Trinity

8 – April – 2017

Topic Question

Traditional cab drivers plan their day based on experience. That is, they have an understanding based on their many years of experience that certain areas of the city produce greater demand levels at different times in the day, and at different times of the year.

However, this ‘experienced-based’ planning is susceptible to cognitive bias, including recency (most recent success will be over-weighted), confirmation (luck will distort perceptions) and saliency (large fares will be over-weighted). This results in an imperfect human prediction of the answer to the question ‘where is my best chance of picking up a fare?’

Uber has separated itself from the Taxi industry with its use of technology to increase the user experience of hired transportation. We propose that Uber continue to use technology to optimize Time in Service for their drivers.

As such, the question we set out to answer is:

When a driver drops off a fare, which direction should they travel to maximize their chance of finding an additional fare and optimize Time in Service?

Answering this question means that drivers are idle less, and hence earn more money for themselves and for Uber. Additionally, the scheme will decrease waiting times in busy areas for consumers, widening the user experience moat that Uber has created.

The data used to answer this question was the Yellow Cabs pickup and drop-off locations. Due to this size of this data, analysis was conducted on April, 2014 data only, for the hour between 9am and 10am.

Non-Technical Executive Summary

When an Uber driver picks up a fare, both they and Uber are aware of the planned drop-off location, and the pickup location. Once the driver delivers the fare, they face two options to find the next fare:

- Stay in the drop-off area; or
- Move to another area, and if so, which area?

Based off an analysis of Yellow Cab transit around New York City, a directed flow graph identified the flow relationships between each of the 195 municipal zones. The primary drop-off location for pickups in any municipal zone for Yellow Cabs in Manhattan is shown below:

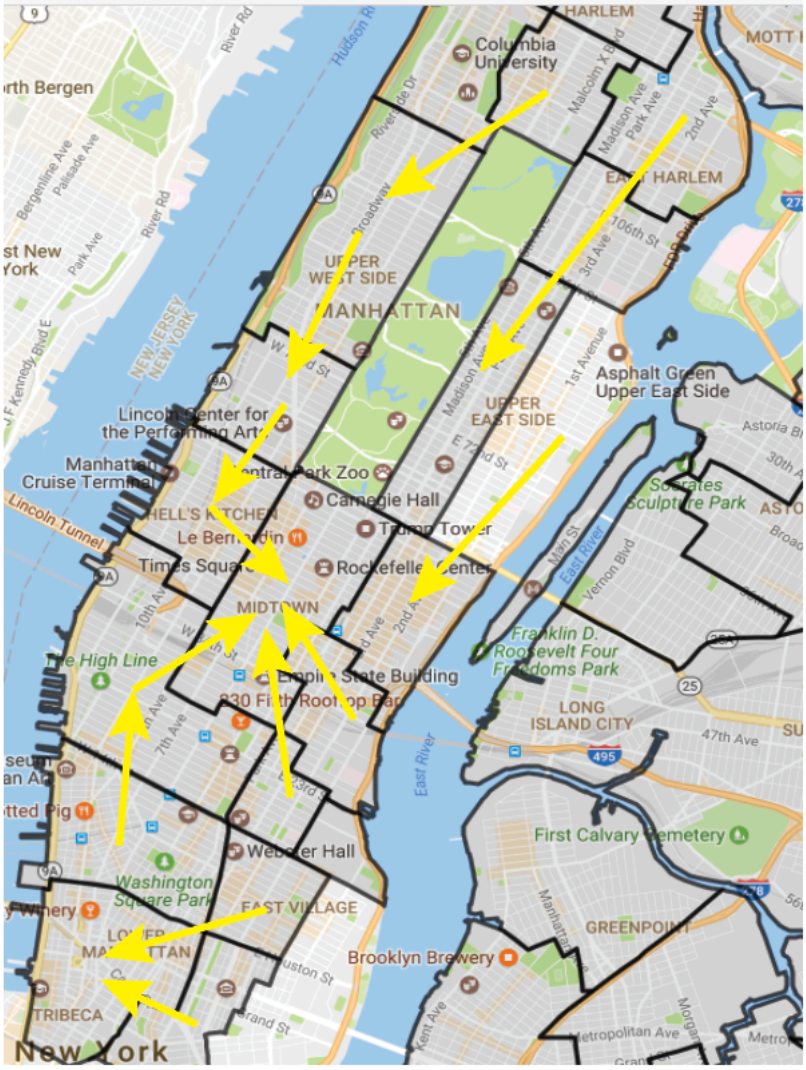

Figure 1: Main pickup – drop-off flows around Manhattan for Yellow Cabs.

Knowing how traffic moves around the city is vital for Uber, as it allows them to predict areas of demand throughout the day. Our solution utilizes this knowledge to assist drivers in real time.

While our solution is adaptable for all routes through the city, we focus on one route as a case study:

- Pickup Location: W 72nd St
- Drop-off Location: Washington Square Park
- Pickup Time: 09:00am

Based on the pickup and drop-off location, the expected drop-off time is 9:12am. By Analyzing the following map of all Yellow Cab pickup locations between 9 and 10am, Uber can then calculate where the driver should move to find their next fare.

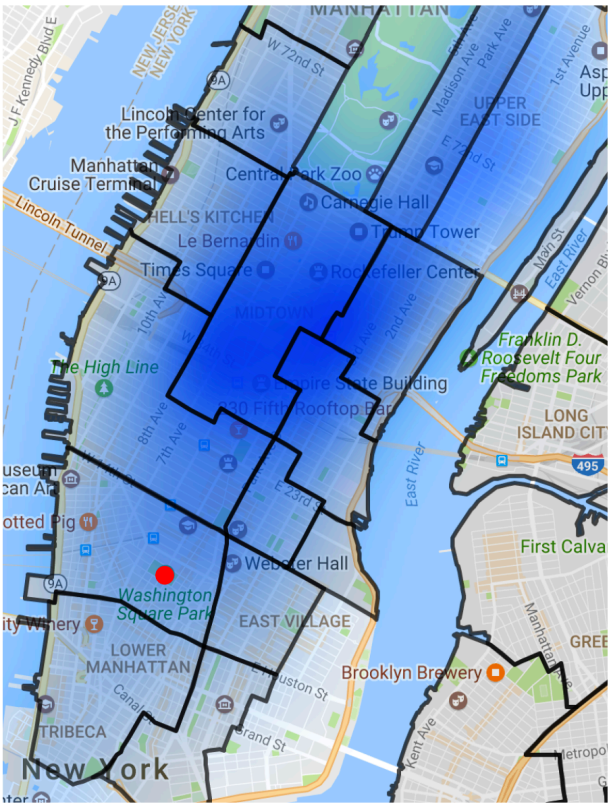

Figure 2: Yellow Cab pickup locations between 9 and 10am (Case Study Uber location in red).

The map shows most pickups during this time are in midtown, with pickups diminishing further south. Based off this and the driver's location, we can generate a likelihood of pickup for each of the municipal zones surrounding the driver.

Exact details of this likelihood function are explained in Section 3, but incorporates:

- Driver location;
- Pickup likelihood per zone;
- Distance from other zones; and
- Current Uber density in other zones.

In this theoretical example, the driver is currently in zone MN23. The best 5 chances of a pickup based on our likelihood function are:

- Stay in zone MN23: Pickup Likelihood 17%
- Move to zone MM21. Pickup Likelihood: 8%
- Move to zone MM20. Pickup Likelihood: 6%
- Move to zone MM13. Pickup Likelihood: 3%
- Move to zone MM17. Pickup Likelihood: 11%

This finding then suggests that the driver should stay in zone MN23 for the most likely pickup.

The best thing about this feature is that Uber knows as soon as the pickup is made and the drop-off location is made where the 'best next location' is for every driver. That means, approximately 10 minutes in advance where each driver will be, and where they will be going to next. This information is incredibly valuable, and effectively allows Uber to front-run demand it knows is coming based off historical data.

Furthermore, because Uber controls the recommendation it gives to each driver about their suggested next location, they can efficiently distribute driver density across the city as required. All of this combines to increase driver efficiency by maximising their Time in Fare, and reduce waiting times for consumers.

The benefits don't end there, but other Uber programs like Uber Pool can be optimized by knowing primary traffic flows around the city. Uber Pool, where passengers car pool for a lower fare, could benefit by drivers taking routes through zones where the primary flow is the same as their current destination. For example, if a driver makes an Uber Pool pickup at Grand Central heading towards Columbia, they have the choice of taking Madison through the East side, or up Broadway/Amsterdam up the west side. Based on the primary flows map, Uber can suggest that the driver take 1st avenue north on the east side into MN19, across Midtown through MN17, and then north on the west side through MN15, MN14 and MN12 as this route travels through 4 of the 5 zones which primarily feed zone MN09 (Columbia).

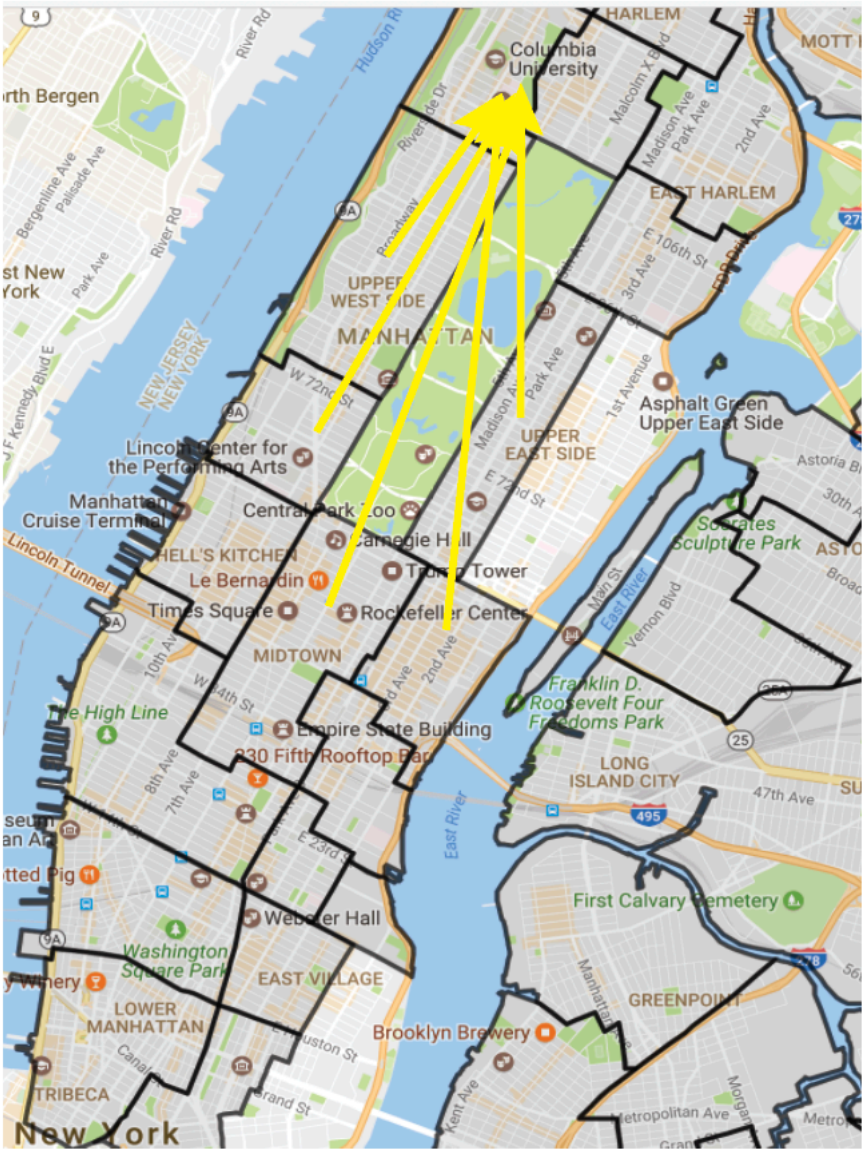

Figure 3: 5 Primary traffic flows into MN09 (Columbia University)

While we have provided a single case study to show the effectiveness of our method, it is easily adaptable to any other time period, and definitely customizable to efficiently predict primary traffic flows per hour

The answer to the question of ‘what is the best next location’ for an Uber driver after a drop-off adds value to Uber’s operations across the board. Drivers are more efficient, passengers are happier, and Uber can continue to outperform Taxis by efficiently distributing their fleet ahead of time to best serve consumers.

Technical Executive Summary.

The first step taking was to filter data and study a small time period. This was conducted using the grep functionality in bash. This was the most computationally efficient. Subsequently, all analysis was conducted in Python using Pandas DataFrames.

Once a specified time period of data was identified, pickup and drop-off locations were mapped to municipal zones. This was completed using a 'closest centroid' method, but computing the centroid of the zone and finding the closest centroid to a pickup / drop-off location, and assigning that zone to the pickup / drop off. With this completed, a directed graph matrix was constructed which showed the number of trips beginning in a zone and ending in all others, or ending in itself. This allowed the first criteria of the weighting equation to be calculated, as primary traffic flows throughout Manhattan were identified.

The weighting equation itself was constructed to prioritize pickup likelihood, while penalizing for distance travelled to the desired zone, and whether there is an oversaturation of Uber drivers in an area. This last criteria couldn't be calculated accurately due to limitations in data. An approximation would have been to use the number of Uber pickups in each zone during that time, but as that is likely correlated with the first criteria, it was decided to set this factor to 1 for all zones in the Case Study. It is suggested that this factor takes a value between 0.5 and 1.5 to add a multiplicative effect to the overall criteria.

With this in mind, the 'preferential zone' formula was constructed as:

$$\frac{P}{D} * U_d$$

where

- P is the proportion of all trips in Manhattan picked-up in that zone during the time period;
- D is this distance to the zone in km; and
- U_d is the current Uber density in this area, set to 1 for all zones in the case study.

With all the other factors able to be calculated from the given data, the results are below:

Zone	Distance (km)	Pickup Likelihood	Uber Density	Result
MN23	1 (current zone)	0.169455458	1	0.17
MN21	1.592	0.130960947	1	0.08
MN20	2.394	0.148026668	1	0.06
MN13	6.782	0.24601583	1	0.03
MN17	2.772	0.305541097	1	0.11

There is no part of this analysis that was specific to the case study chosen for presentation, and with some basic data engineering, a scalable solution could be relatively easily completed.

The primary road block was size of the datasets, and trying to manipulate those to work with a sensible subset. Building a weighted adjacency matrix for the municipal zones took a very long time.

Other avenues of investigation that were stopped due to time constraints were sub-graphing within municipal zones to further drill down into the data, or graph analysis to identify key nodes in the traffic flow graph which could then be avoided in the event of traffic difficulty.