UBER TAXI FARE ANALYSIS

Shaiwal Sachdev

Mentor: Sankarshan Mridha

Problem Definition

- Aim is to **compare and analyze fare data** of large number of trips for Taxi Companies mainly New York Yellow Cabs and Uber.
- We will also try **to predict the value of surcharge** at a given location and time.
- The analysis is focused on New York region only.

Dataset

For this work we used two datasets. NYC Yellow taxi data and Uber Data.

- ❑ **NYC Dataset**

Pickup (Lat,Lon)	Dropoff (Lat,Lon)	Distance (miles)	Duration (seconds)	Total Fare (USD)	Time stamp

- ❑ **Uber Dataset**
  - ❑ We selected popular origin, destination (OD) pairs from the NYC dataset for different hour bucket and get fare data from Uber API.
  - ❑ For comparison, we focused on **four** different one hour time buckets, 6 [5:30am-6:30am], 10 [9:30am-10:30am], 16 [3:30pm-4:30pm], 20 [7:30pm-8:30pm] and collected uber data for **425308 OD pairs**.

Min Fare (USD)	Low-High Estimated Fare (USD)	Distance (miles)	Duration (seconds)	Surcharge Multiplier

Pickup(lat,lon) = 399 W 15th St, New York
(40.7416561,-74.0048858)

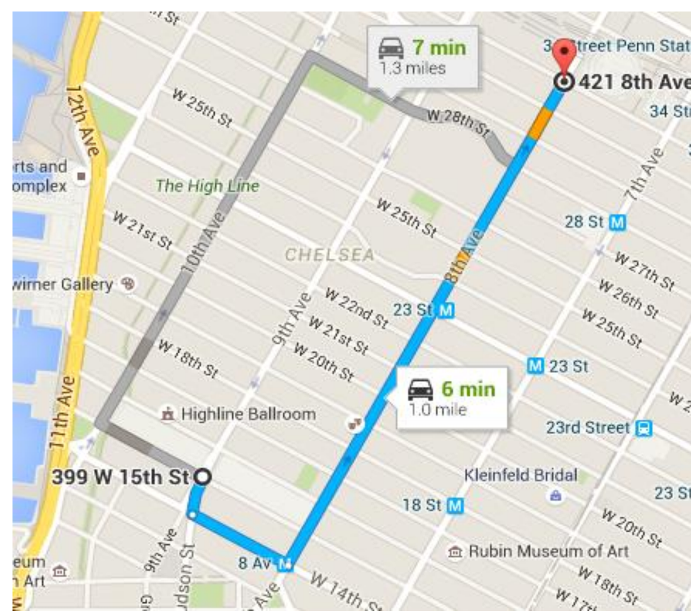
Dropoff(lat,lon) = 421 8th Ave, New York
(40.7502935,-73.9948451)

UberX

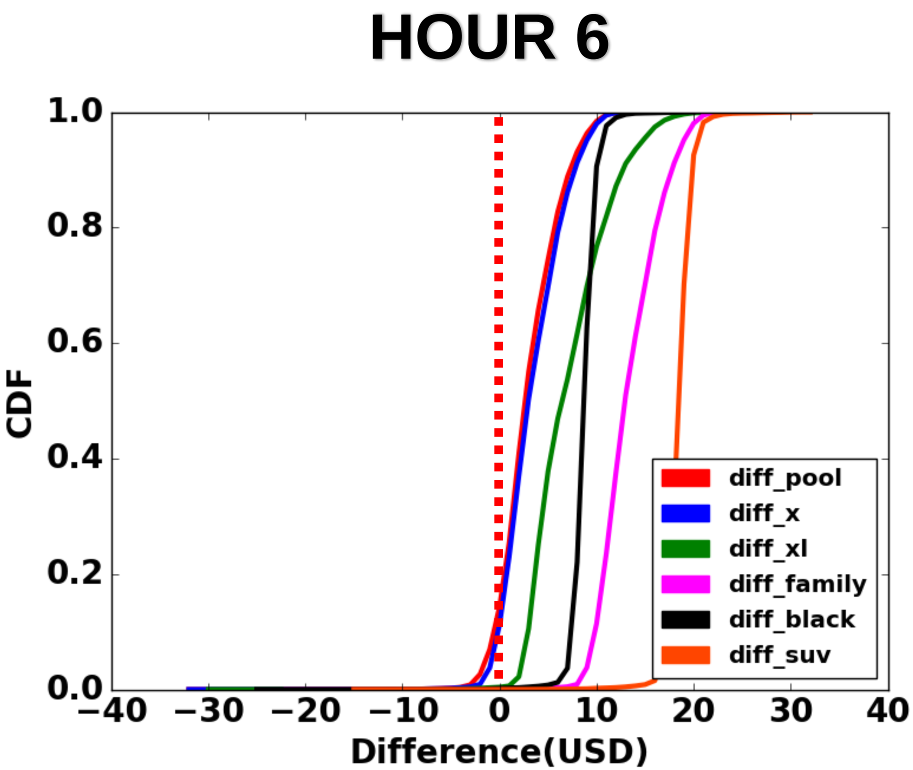
Hour	Surcharge	Min Fare($)	Low-High Estimate($)	Distance (miles)	Duration (minutes)
6	1.5	12	12-13	1.04	5
10	1.0	8	8-9	1.06	9
16	2.2	18	18-21	1.04	11
20	1.0	8	8-10	1.30	9

NYC

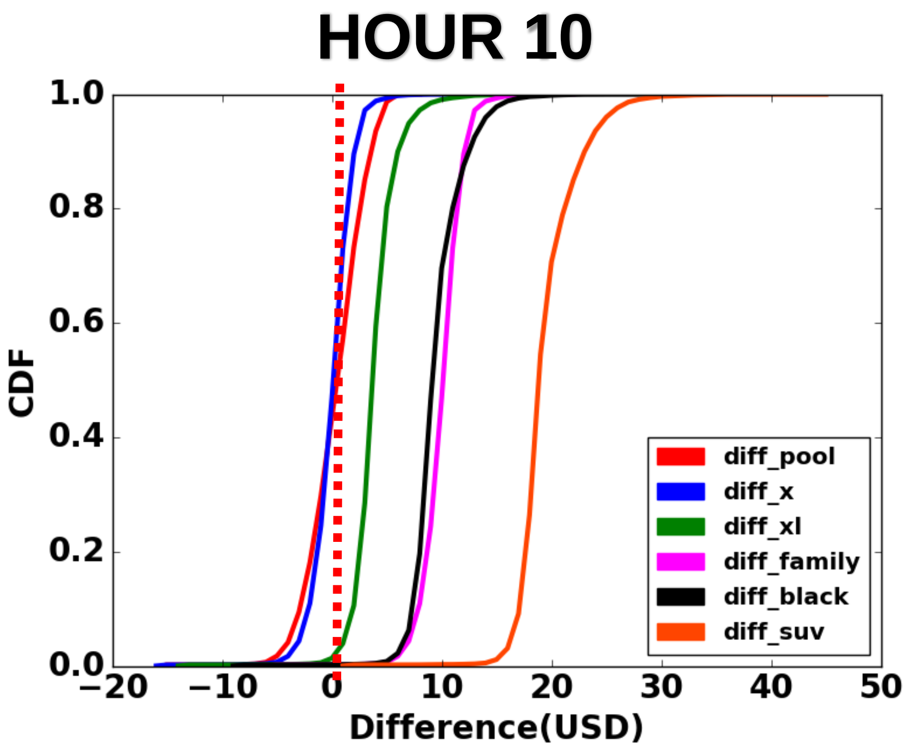
Hour	Trip Count	Avg. Fare($)	Avg. Dist (miles)	Avg. Duration (minutes)
6	31	6.2	0.95	5.5
10	105	8.5	1.02	8.2
16	121	8.9	1.05	7.0
20	239	8.1	1.06	6.2



Fare Comparison(Cumulative Distribution Function)

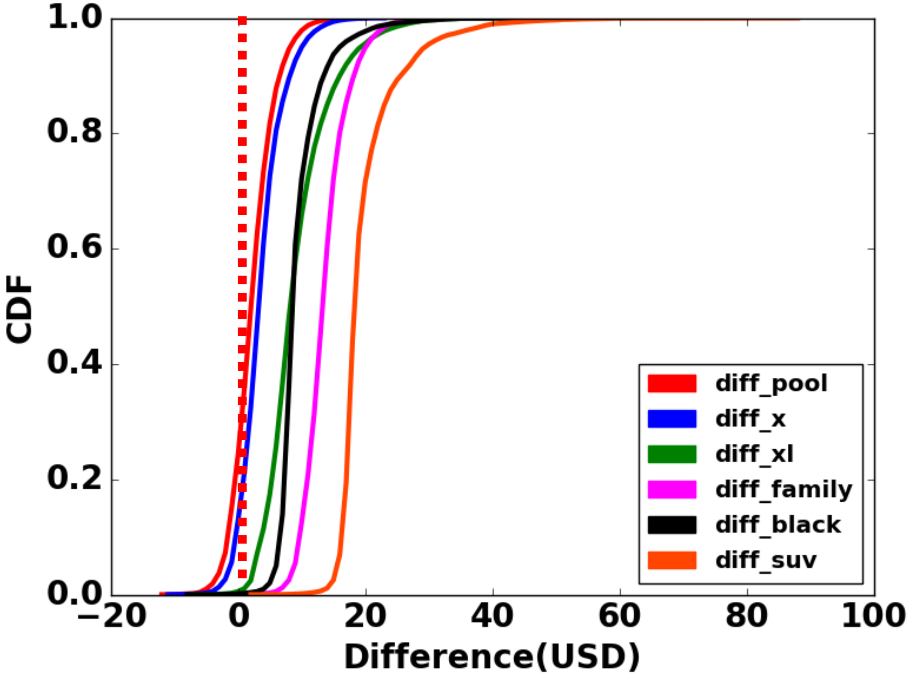


Cheapest UberX is costlier than NYC
For 90% trips UberX > NYC



UberX costs compared to NYC cabs
For ~50% trips UberX < NYC
For ~50% trips UberX > NYC

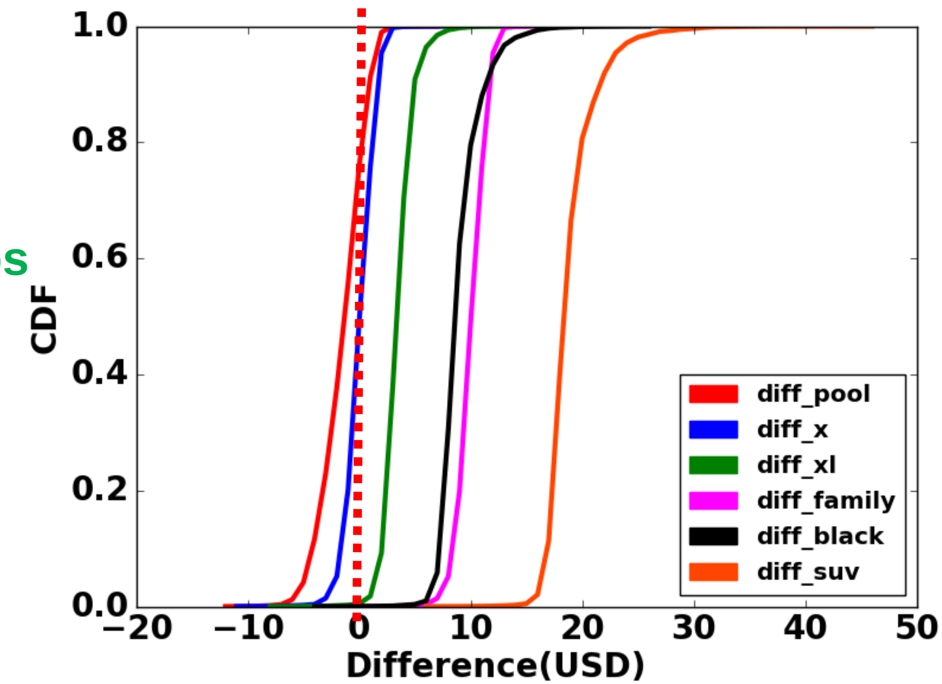
HOUR 16



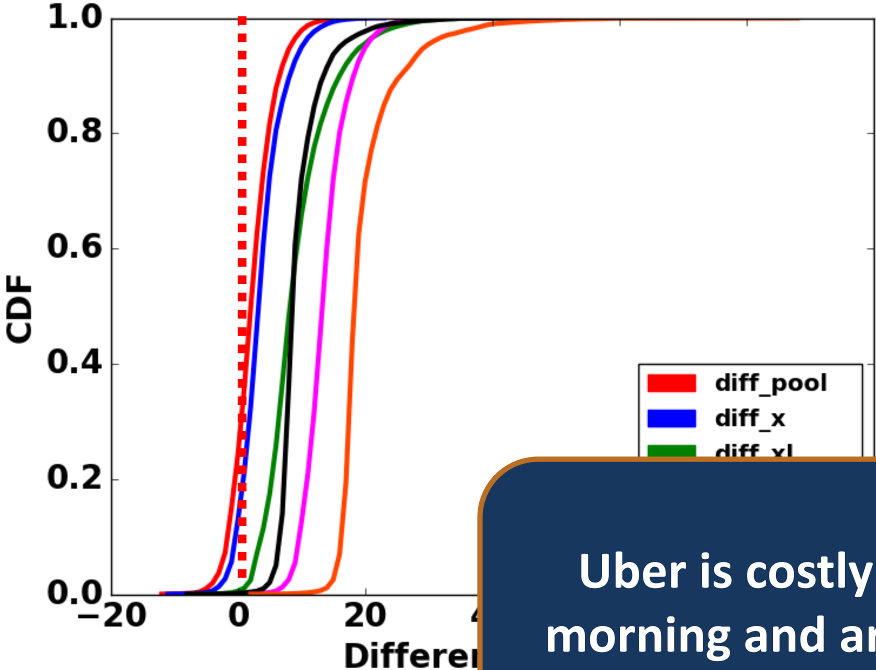
Cheapest UberX is costlier than NYC
For ~80% trips UberX > NYC

UberX costs almost equal to NYC cabs

HOUR 20



HOUR 16

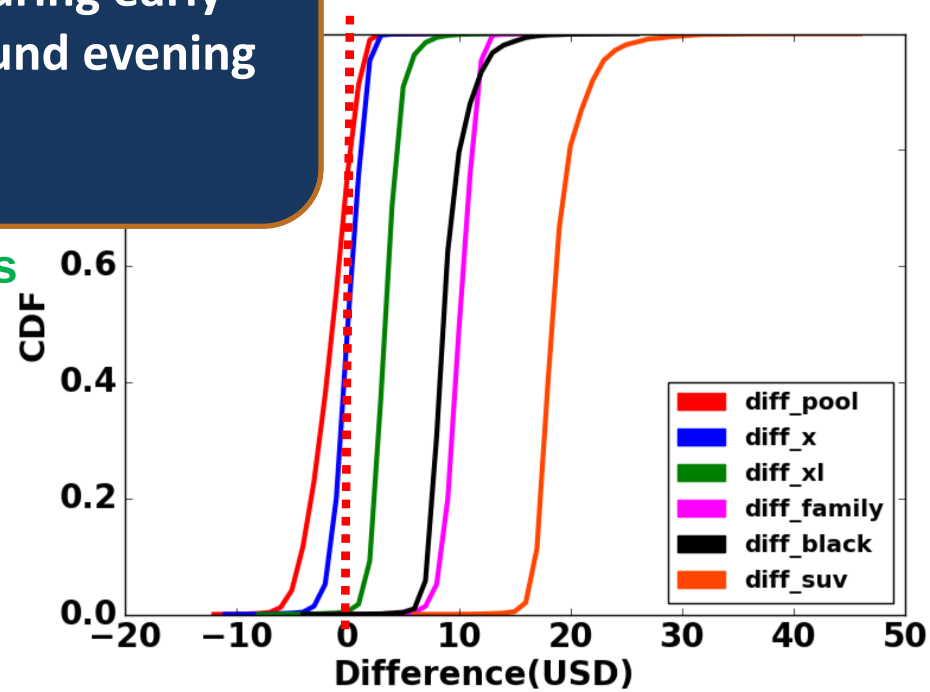


Cheapest UberX is costlier than NYC
For ~80% trips UberX > NYC

Uber is costly during early morning and around evening time

HOUR 20

UberX costs almost equal to NYC cabs



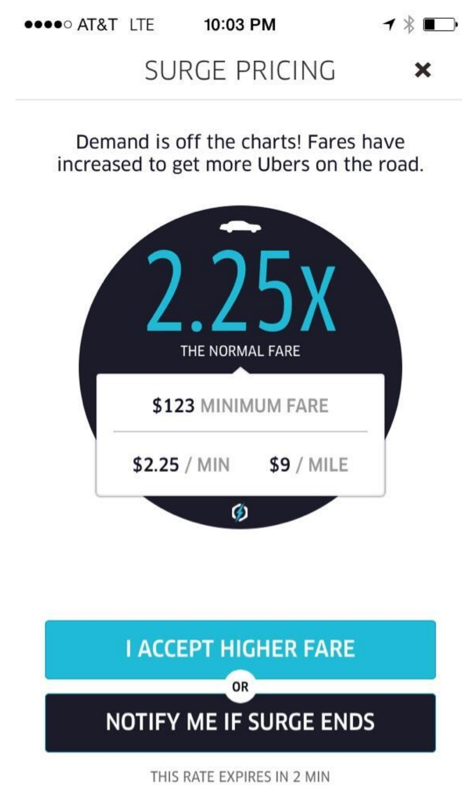
Surcharge Estimator

- After doing the analysis ,we found out that uber earns most of the its money by its dynamically changing Surcharge values.
- Surcharge Multiplier =

(Demand By Customers)/(Supply of Drivers).

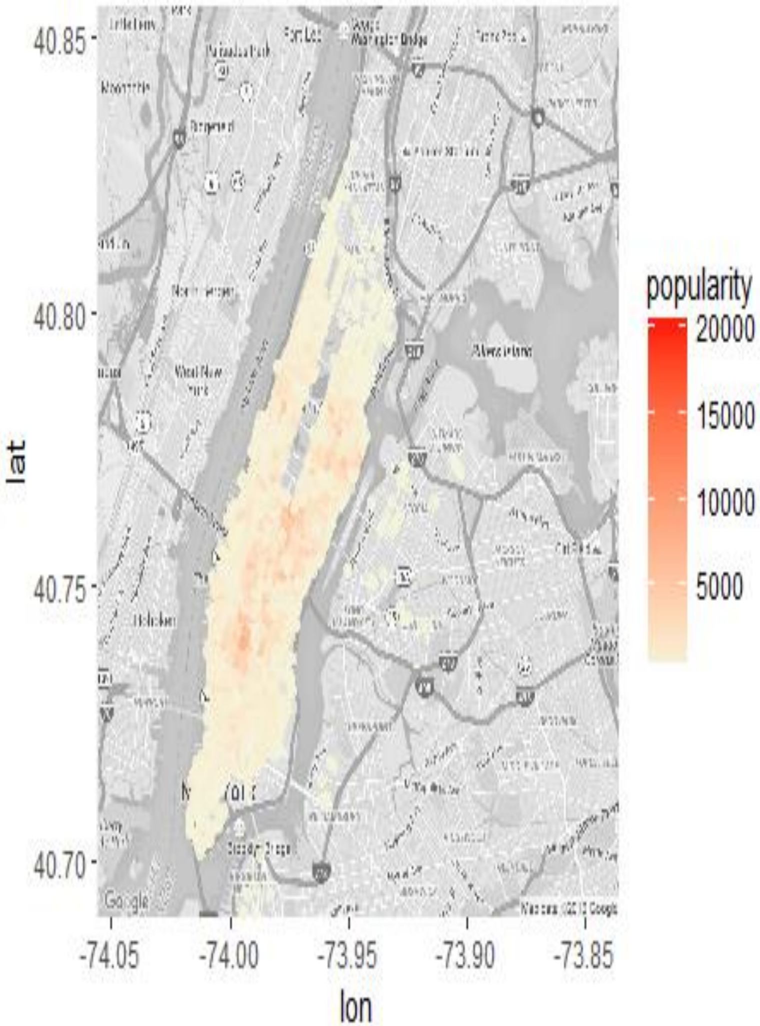
- **Can we make a model that will predict the surcharge value giving the following inputs**
  - 1. pickup location
  - 2. dropoff location
  - 3. Hour
  - 4. cab service

- Assuming uniform supply of drivers. We consider, according to *Anastasios et al* [1], that surcharge only depends on the Popularity of the pickup location.

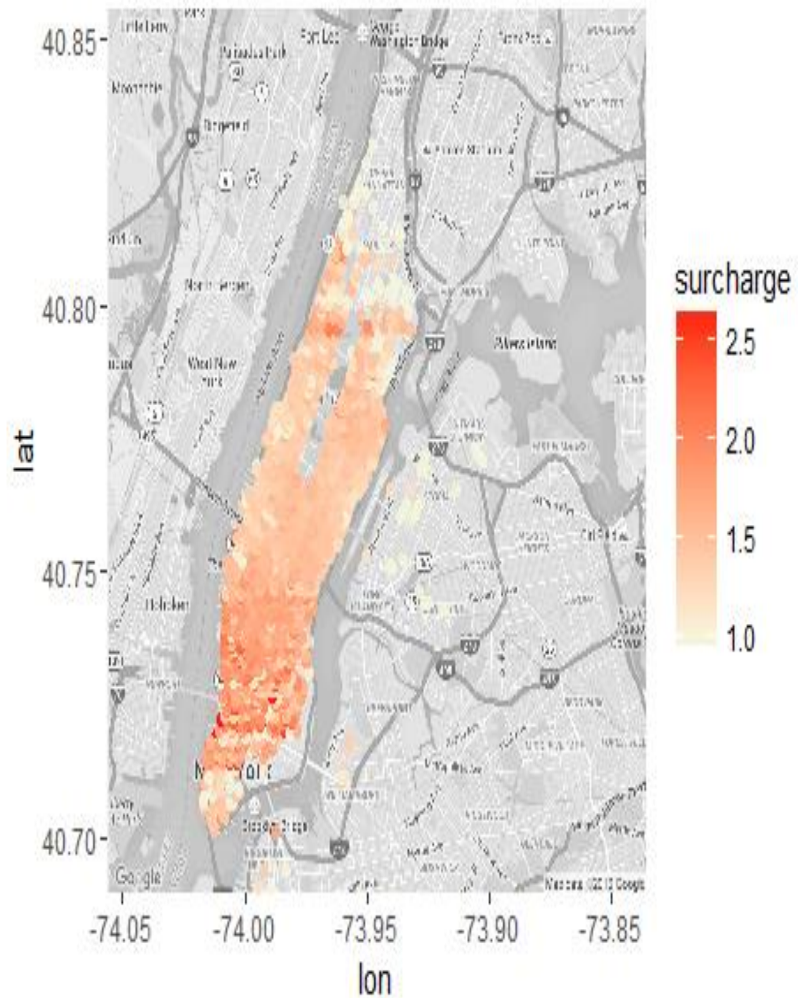


1. Mining open datasets for transparency in taxi transport in metropolitan environments Anastasios Noulas1,3*, Vsevolod Salnikov2, Renaud Lambiotte2 and Cecilia Mascolo1, EPJ Data Science 2015

How the data looks at hour 16 ?



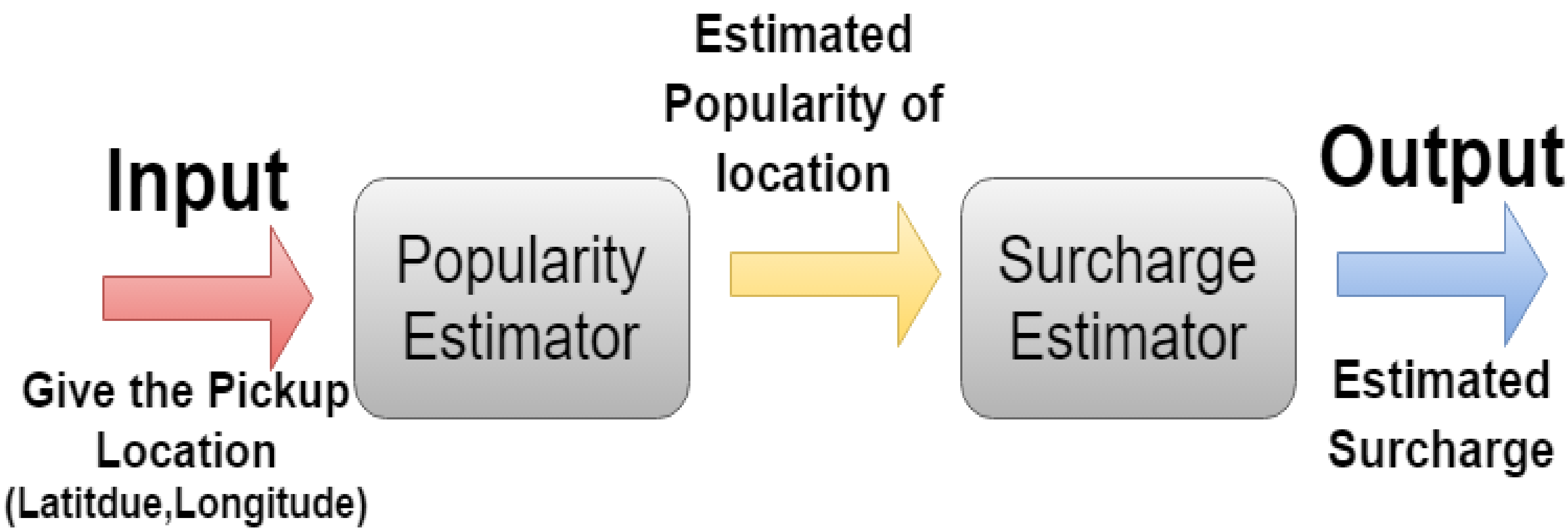
Original Popularity from NYC data



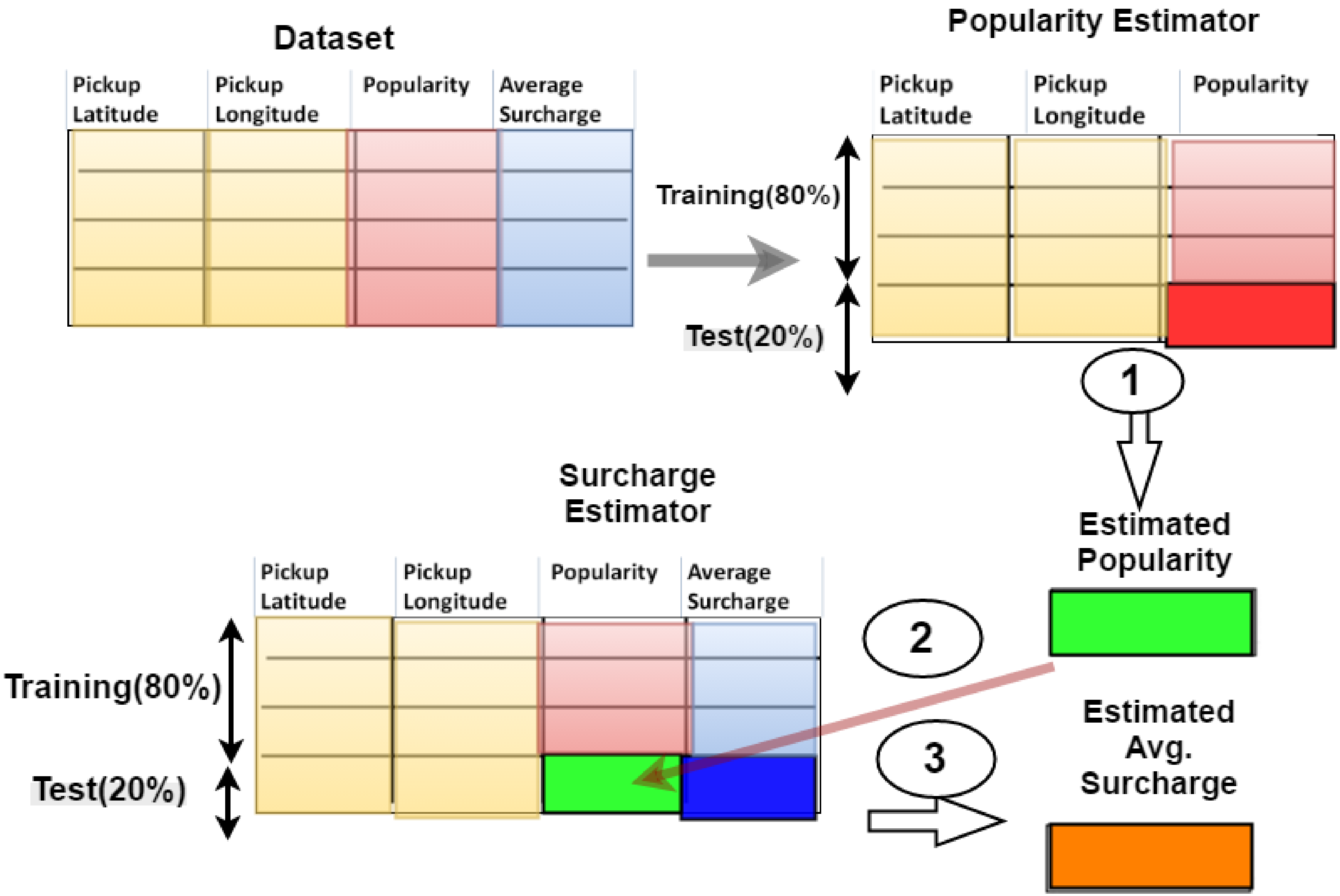
Original Surcharge from Uber

Proposed System

For a given Hour and Cab Service

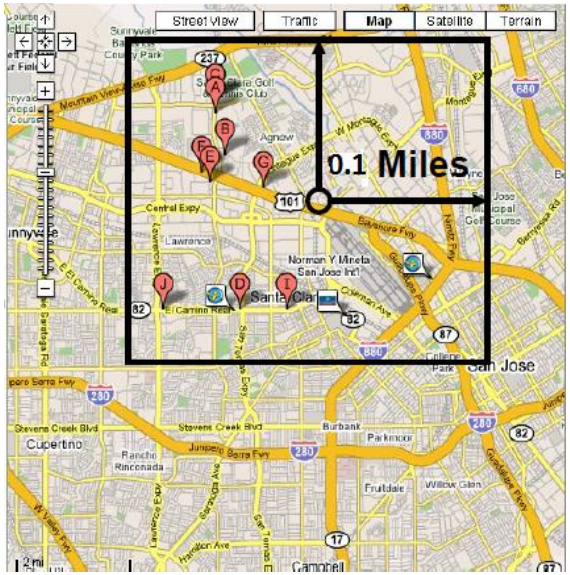


PROCESS FLOW



Popularity Estimator

- The demand (popularity) at a test location depends on the popularity of neighbour.
- Start with a distance of 0.1 mile and go on searching upto 1 mile till we find at least one neighbouring location.
- We can take the average of demand of all locations within the bounding box. *This method did not give us good results.*
- Even within the bounding box ,the location nearer should have more effect or more weight.



So we use Inverse Distance Weighting Average(IDW)

$$w_i(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x}_i)^2} \qquad y.(\mathbf{x}) = \sum_{i=1}^{d} w_{.i} x_i$$

Surcharge Estimator

We give popularity as of a location as input and get estimated surcharge as output for a fixed hour.

- **Locally Weighted Regression(LWR)**
- Estimated value of surcharge should depend on local regions more.
- Let X_0 be the popularity of the new location.
- Find the squared difference between X_0 and its neighbors. We will use this distance for the weight function.
- Find diagonal weight matrix.
- X is estimated popularity
- Y is surcharge

$$w_i(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x}_i)^2}$$

$$w(x) = a e^{-\frac{(x-b)^2}{2c^2}}$$

$$\beta = (X'WX)^{-1}X'WY$$

$$\hat{y} = \beta X_0.$$

MSE for Surcharge Estimation

Hour	6	10	16	20
Linear Regression	0.0287	0.0221	0.0422	0.00239
Decision Tree Regression	0.02713	0.01980	0.03089	0.00231
KNN Regression (k- nearest neighbor)	0.02806	0.02050	0.03680	0.002720
Locally Weighted Regression(LWR)	**0.0091**	**0.0081**	**0.0101**	**0.00168**

For the above MSE the original range of surcharge values lie from 1 to 2.0

Conclusion

- The proposed model estimates the average surcharge at a location with maximum Mean Square Error of 0.01.
- This model helps to predict the surcharge of a location for any hour bucket using history data. So the user can plan accordingly.
- This model also helps to compare the estimated surcharge given by Uber against the estimated surcharge of the location using history data.


Thank You