Taxi Usage Modeling with Spatio-Temporal Correlations

# Introduction

This project attempts to provide insights into how taxi usage in New York City is related with factors such as population, average income, land use, car ownership, road network and public transit accessibility. Additionally, we plan to investigate the spatial and temporal patterns of taxi usage which is potentially caused by the uneven distributions of taxi demand and supply across the city.

Census tracts are used as the basic geographic units to capture taxi usage, transportation, land use, demographic and socioeconomic data. A Hadoop mapreduce program will be run in Amazon Web Service to compute the monthly taxi pick-ups and drop-offs for each census tract. Taxi usage models will be developed to relate the taxi pick-ups and drop-offs with features collected. A conditional autoregressive (CAR) effect term will be used to capture the spatial correlation of neighboring areas. Meanwhile, an unstructured random effect term will be included into the model to account for the temporal variation.

# Datasets

* Taxi trip data
* Census data
* Demographic data
* Socioeconomic data
* Land use data
* Road network
* Subway data
* Bus data

# Timeline

* April 10
  + Write mapreduce program for taxi usage calculation and test it locally (Ya)
  + Collect demographic, socioeconomic and transportation data for each census tract (Yuzheng, Kun)
* April 17
  + Run mapreduce program on AWS (Ya)
  + Finalize the data preparation (Kun, Yuzheng)
* April 24
  + Model specification and development (Kun)
  + Begin writing report manuscript (All)
* May 1
  + Test model on AWS and obtain the final outcomes (Yuzheng, Kun)
  + Visualize the outcomes (All)
* May 11
  + Wrap up the project presentation and report (All)

# GitHub Repository

<https://github.com/coolshaker/taxi_project>