# Markov Chains to Analyze Taxicab Rides in New York City

Mitas Ray, Gary Cheng, Kaylee Burns

**Come up with a new application for PageRank or Markov chains.**
We will model the routes of taxicab drivers in New York City for a given time interval as a random walk among geographic regions (intervals of latitude and longitude). The probability of transitioning between states will be calculated by counting the number of rides to other states and dividing by the total number of rides out of that state. In addition to transitional probabilities, we will associate an average fare with each edge.

**What is the motivation behind your idea?**
We selected the taxi cab data set from Kaggle because we wanted to find an application of Markov chains to a domain we were all interested in. As users of common taxi services like Uber or Lyft, we're fascinated by the logistics of driving taxi cabs. We hope that the insights we gain from this project will allow us to connect and reflect on our past tourist experiences in New York.

Furthermore, Markov chains are well suited to trends in taxi cab drivers' routes and can enlighten us about the popularities of various pickup locations over time, opportunities for taxicab drivers to increase their revenue, and the trends in "migration" among customers. This investigation not only provides direct financial benefit to drivers but also satisfies our curiosity.

**What is the dataset you will be exploring/analyzing?**
We will be analyzing the NYC Yellow Taxi Trips (https://www.kaggle.com/nyctaxi/yellow-taxis) from January - June 2016. Our model will categorize latitude, longitudes, and time of day into states that represent geographic regions of New York City.

**How do you plan on applying properties of Markov Chains/PageRank?**
The next location of a taxi is strictly dependent on the immediately previous location. This observation embodies the markov property. As mentioned earlier, each state is characterized by a latitude, longitude interval. Then we can determine transition probabilities between states using our dataset. We don't plan on making Pagerank the focus of our project, but if we have time, we want to find the steady state of our transition matrix. This information would give us a popularity ranking of the geographic/temporal states of our Markov Chain.

**What do you expect to see from the results?**
We want to answer the as many of the following questions:

- Given some starting location, as the number of rides you give goes to infinity, do you converge to one location?
- Where should your start to make the most money in (some amount of time)?
- What locations makes the least amount of money in (some amount of time)?
- Expected profit of ride based on starting location
- Average duration of rides based on starting location