# Problem Statement

## Background

The first taxi in NYC was deployed in July 1897, when the Electric Carriage and Wagon Company (E.C.W.C.) began running 12 hansom cabs. Over the years, the taxicabs of NYC have become some of the mostly widely recognized icons in the world, with over 13,000 medallions issued as of 2014 (the right to run a yellow taxi). On top of that, numerous green taxis, community cars, and black cars roam the streets under the supervision of the NYC Taxi and Limousine Commission (TLC).

But in 2010, a small company then known as UberCab would make a splash in San Francisco and become one of the largest behemoths in the world, valued at over $60 billion as of August 2016. Uber, as it came to be known over the years, has disrupted and is now on the verge of overturning the traditional for-hire vehicle (FHV) industry as we know it.

Uber, and other FHV companies like it, have made life difficult for traditional taxi drivers, as the increasing competition has driven down costs and put many long-time drivers out of business. It has even been tough on Uber's own drivers, as the lower fares have been mostly passed down to them. With the advent of autonomous vehicles and California's recent approval to allow the testing of fully driverless cars on the road, it seems like the industry will soon look completely different than what we have been used to for over 100 years.

## Your Task

Your goal is to analyze the NYC Uber trip data (described below), potentially in combination with supplementary datasets, in order to increase understanding of how developments in the NYC for-hire transport industry relate to broader trends in the public and private transportation industries at large.

We have partially pre-cleaned several supplementary datasets for your use. Additional trip data is available, including data from green boro taxis, yellow medallion taxis, and NYC subway turnstiles. We also provide demographic info on age and income, geographic data for the Neighborhood Tabulation Areas (NTAs) of NYC, and daily weather data for NYC.

**You are asked to pose your own question and answer it using the available datasets**. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is more important over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to predict future trip trends. Submissions may also be illuminating, through the use of data visualizations or through sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: NYC has both green boro taxis and yellow medallion taxis. How have trip trends over time differed between these two classes of taxis in relation to Uber ridership?

Sample Question 2: Is there a connection between the demographics of particular NTAs in NYC and the dynamics of taxi or Uber rides involving those NTAs?

Sample Question 3: Investigate the details of pickup times and endpoints. Is there a relationship between these and the NYC weather?

Sample Question 4: Are there interesting similarities/differences in the servicing patterns of Uber, green taxis, and yellow taxis?

Sample Question 5: Explore the interplay between Uber and taxi trip trends and NYC public subway trip trends. What sorts of dynamics exist between these?

## Datasets

The provided datasets are spread across nine tables. Your team should only use the tables that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to "play nice" with each other.

***uber_trips_2014***
Trip data (pickup times, pickup coordinates, etc.) from Uber vehicles in 2014.
*~4.5 million rows & 4 columns.* Size: ~30MB zipped, ~200MB unzipped.

***uber_trips_2015***
Trip data (pickup times, pickup location IDs, etc.) from Uber vehicles in 2015.
*~14 million rows & 4 columns.* Size: ~65MB zipped, ~550MB unzipped.

***demographics***
Demographic data (population, age, income, etc.) organized alphabetically by NTA.
*188 rows & 33 columns.* Size: ~0.1MB.

***geographic***
Data about the shape of each NTA (latitude and longitude coordinates, in order).
*9,302 rows & 195 columns.* Size: ~4MB.

***green_trips***

Trip data (pickup/dropoff times, pickup/dropoff locations) from NYC green boro taxis. *Note: in order to keep the dataset size manageable, the provided data is a 20% unbiased sample of the raw data. If using trip count metrics, remember to multiply quantities by 5 to approximate the actual data.*
*~3.5 million rows & 9 columns.* Size: ~140MB zipped, ~400MB unzipped.

***mta_trips***
Trip data (time intervals, entries, exits, etc.) from NYC public subway turnstiles.
*~7.5 million rows & 10 columns.* Size: ~50MB zipped, ~700MB unzipped.

***weather***
Temperature and precipitation data for three areas in the NYC metropolitan area.
*2,190 rows & 10 columns.* Size: ~0.1MB.

***yellow_trips***
Trip data (pickup/dropoff times, pickup/dropoff locations) from NYC yellow medallion taxis. *Note*: *in order to keep the dataset size manageable, the provided data is a 5% unbiased sample of the raw data. If using trip count metrics, remember to multiply quantities by 20 to approximate the actual data.*
*~8 million rows & 9 columns.* Size: ~260MB zipped, ~800MB unzipped.

***zones***
Information about each ride pickup zone in the NYC metropolitan area.
*263 rows & 5 columns.* Size: ~0.1MB.

## Additional Datasets

You are welcome to scour the Web for custom datasets to supplement your analysis. All additional data used should be public and should not exceed 2GB unzipped (consult your TA if you believe your idea is worthy of an exception).

## Other Materials

We will provide you the schema for each of the data tables in another packet.

## Tips & Recommendations

We've compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not appertain to every team, so we recommend that you and your team apply any tips accordingly.

| Tips for Success | Try to Avoid |
|---|---|

| | |
|---|---|
| **1.** Focus on hypothesis testing when brainstorming your research question | **1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy |
| **2.** Spend ample time on your report to ensure strong communication through visualizations and writing | **2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient |
| **3.** Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality | **3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile |

**Ask for Help**

The TAs are here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.