# Big Data Project Proposal - By Luis Severien Marcilla and Jelena Stankovic

## High level requirements: Motivation and goal

In big cities such as New York one is overwhelmed by the traffic and just happy to get safely and fast from A to B. Especially when it comes to hailing a cab there are other worries than calculating prices. Still, there are big differences in fares which the consumer might want to take into account when deciding about if he is going to travel by a NYC Cab, Uber or just may take the metro.

The most prominent way to travel New York is in a typical yellow NYC Taxi. The fares of these taxis can easily be calculated retrospectively (the composition of the price will be presented in the analytics part and can be looked up here [wiki/New\_York\_City\_Taxi\_Cabs](https://de.wikipedia.org/wiki/New_York_City_Taxi_Cabs" \l "Boro-Taxis)), but a prediction turns out to be difficult due to uncertainty of travelling time through NYC traffic and the actual distance. Some tourists might also not have in mind that there are extra fees charged for rides at night and during rush hours.

To create some transparency, we are going to use the NYC taxi dataset and create an application for the consumer, which will enable him to foresee the NYC Cab price for a future ride. The application takes the time of the day, the pick-up and the drop-off coordinates and forecasts the prices of the NYC TLC (Taxi and Limousine Commission) company.

The publicly available dataset ([www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml](http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)) comprises data covering more than 1 billion taxi rides from 2009 to 2017. These huge amount of past data will enable us to create a predictive price model for future rides.

High Level requirements:

A dataset comprising travel data

## Implementation requirements:

### Analytics

Use scikit learn, Python’s machine learning library for clustering and regression analysis

Predicting the taxi price is a regression problem (supervised learning problem). The independent variables to take into account: pick up time, pick up neighborhood, drop off neighborhood + dummy variables for the optional premiums. The regression is most likely going to require previous clustering of drop-off and pick-up coordinates into neighborhoods (clustering algorithm, unsupervised learning problem).

The taxi price is calculated on the basis of several variables. Basis price 2,60 $ + 0,55$ NY state tax per ride + 0,80 $ for a night shift premium (20:00h-6:00h) + 1,00 $ rush hour premium (16:00h-20:00h on weekdays) which will be covered by dummy variables. Of course the price also depends on the miles driven (0,60$ per 1/5 mile) and in case of driving slow or standing in a traffic jamb 0,50$ per 60 seconds. To account for this we will need to regress on the pick-up time and the pick-up and drop-off neighborhoods as this will give us a clue about the typical traffic conditions between the neighborhoods at different times of the day.

Y = 3,15 + 0,8d1 + 1d2 + 0,6\*5x1 + 0,5\*x2

X1 : Average distance between two clusters in miles

X2: Average waiting time due to traffic between to clusters

**bX1X2 – x1 pick up cluster x2 drop off cluster – interaction effect based on distance and time**

Calculate parameters for the model which will account well for the distance and time travelled. For creating the model we will use the dataset 2010 – 2012, this will be our training data. Data from 2013 will serve as test data, so we see if our predictions by the first model were accurate. Include 2013 and 2014 data into the model.

Stream 2015-2017 data and adjust the parameters at the end of day as new data is created.

### Architecture

As the given data source is not a data stream we will have to simulate just this by ingesting it part by part within a prespecified time frame (maybe even simulate a real data stream with timestamps from the previous year) into our data pipeline. The ingestion will be done with flume.

* Use the datetime package to access the current time and create the timestamps for streaming simulation
* Ingest data with Flume
* Connect to Spark and process with an API for Python

As the given data source is not a streaming data source we will have to simulate a stream by ingesting it part by part within some time frame (maybe even simulate a real data stream with timestamps that match the previous year).

1. The ingestion will be done with Apache Flume, the distributed Hadoop service for ingesting streaming data. As we only have data from one source and in one single format this should be relatively easy, so we refrain from using more complicated services like Kafka.

* save files on a local HDFS directory according to the timestamp. This local directory will be used by flume as the source (spooldir).
* Ingest data with Flume and write it on to HDFS (sink another HDFS directory or HBase).
* Connect to Spark (pick data from HDFS) and process with an API for Python – use pyspark
* Create RDDs in Spark from data files on HDFS. From those extract the values we need for the analytics part. Perform transformations on the RDDs so only the needed values are left (filter, then action “collect” which will return an array – store these arrays in variables which will be used to improve the model)