What is the problem statement?

The time spent on road primarily depends on multiple perspectives: Regions, weather, accidents etc.

For taxi or Uber driver, **how to balance the traffic congestion and benefit maximization?** In details:

1. How could the drivers keep track of the traffic situation by visualization & prediction the trip duration?
2. Will the weather have influence on the traffic?
3. How could drivers acquire the distribution information of requests number?

For these questions, the focus of paper is to:

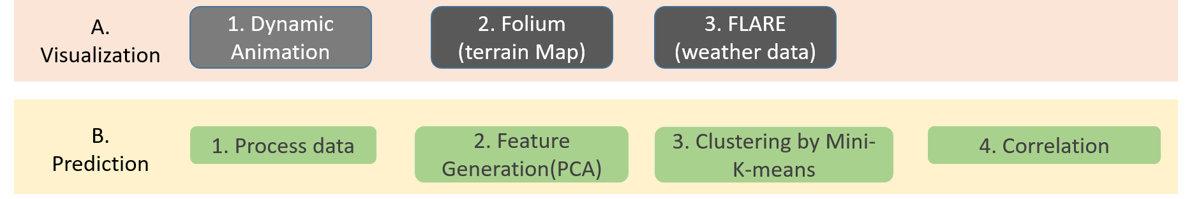
1. **Predict the total ride duration of taxi trips in New York based on TPOT.**
2. **Explore how weather data impacts the traffic with data & explore if it could be validated by sentiment analysis.**
3. **Visualize the traffic and requests number by geospatial information.**

Why have you chosen this particular problem?

I am curious about

1. How could we plot the geospatial data by real map, rather than the usual chart.
2. Will the weather exactly affect on the traffic trip duration?

How are you implementing your problem statement?



1. How could the drivers keep track of the traffic situation by visualization & prediction the trip duration?

A:

1. Visualization:

Plot **the black heatmap with hot area** highlighted, to check the changes with time hours.

1. Prediction:

**Mini batch K-means** is to predict the pickup & dropoff cluster based on geospatial data for getting the hot area.

**TPOT+K fold validation**. Selecting a pipeline is complicated and not very accurate. So I would like to try TPOT to help me. Then I will predict with the model generated by TPOT.

**XGBoostRegressor**: Validated by single XGBoostRegressor to check RMSE(Root mean squared error).

1. Will the weather have influence on the traffic?

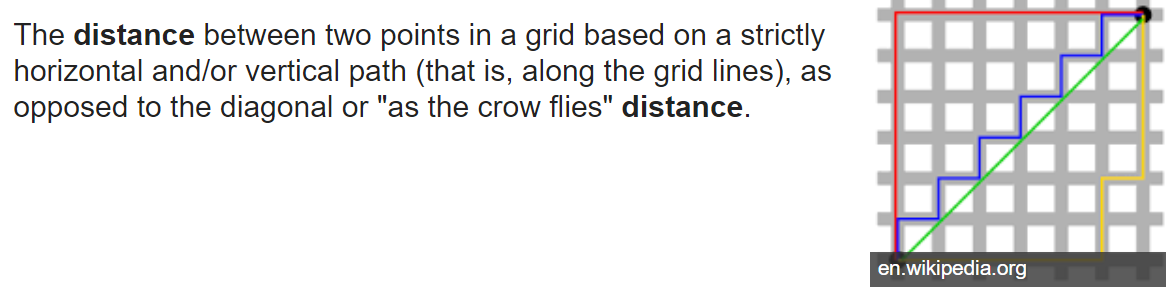
A: Calculate the correlation after preprocessing and feature extraction by **PCA**.

Besides, I’d like to check it out if the result has consistent conclusion with **twitter analysis**, so I apply the sentiment analysis based on twitter.

1. How could drivers acquire the distribution information of requests number?

A: Import **Folium**, an open source map kit and label the areas with more requests.

**Manhattan Distance:**



**Haversine Distance:**

The **haversine** formula determines the great-circle **distance**between two points on a sphere given their longitudes and latitudes.

