# **Machine Learning to Generate Activity Plan**

# Goals of the Simulation

To create a simulation model which visualizes the movement of people in Phnomh Penh, and the medium of travel they use.

# Goals of the Machine Learning Model

To predict the activity plan/movement pattern of 2,000,000 residents of Phnom Penh, by using the activity plan of 10,000 residents.

# Training the Machine Learning Model

## Data

* ‘***trip\_od.csv***’: Records the origin and destination of each trip, along with the path taken, date of trip, time taken, and average speed.
  + Id : id of the agent/traveller
  + start\_lat : starting latitude of the trip
  + start\_lon : starting longitude of the trip
  + end\_lat : ending latitude of the trip
  + end\_lon : ending longitude of the trip
  + path : a list of segment IDs to identify all the trip segments of the trip
  + time : the starting time of the trip
  + dt : the duration of the trip, in seconds
  + distance : the distance of the trip, in meters
  + aver\_speed : the average speed during the trip, in meters/second
* ‘***trip\_traj.csv***’: Records the trajectory in which each trip (as recorded in ‘trip\_od.csv’) was taken.
  + id : id of the agent/traveller
  + trip\_id : to identify which trip the path corresponds to
  + split\_path : segment id
  + start\_time : starting time of the segment
  + start\_lat : starting latitude of the segment
  + start\_lon : starting longitude of the segment
  + end\_lat : ending latitude of the segment
  + end\_lon : ending longitude of the segment

## Challenges

* Currently, we have no way to identify what medium of transportation is used in each trip (car, walking, tuktuk, motorcycle, etc.). However, this information is crucial for the ABM.
* We cannot simplistically classify the medium of transportation simply based on the average speed information (?). Some transportation mediums may travel at the same range of speed (e.g. a car and a motorcycle can travel at the same speed of 40 km/hour under certain circumstances; a tuktuk and a person can travel at the same speed of 4 km/hour during a traffic jam).
* So it seems that we need to collect more data before moving forward.

## Moving Forward: Some Ideas

* ***Can we use logistic regression to identify the medium of transportation used in each trip?*** This will probably require more data than just currently available. We can identify certain variables which can help us differentiate between a certain transportation medium and another. Then, we can implement a logistic regression ML model to predict the transportation medium of each trip.
* Some potential variables which may help:
  + **Relationship between trip duration & average speed**: It is extremely rare for pedestrians to travel 15 km in 30 minutes (average speed = 30 km/hour). So it is more likely that this trip was done using a car, motorcycle, or tuktuk. On the other hand, it seems more feasible if a person was traveling at 30km/hour only for 100m. In short, analyzing the relationship (using a certain metric/ratio) between the trip duration and average speed might help us predict the medium of transportation used.
  + **Geographic Feature**: Some terrains do not allow for travel using a certain medium of transportation (e.g. no tuktuks or pedestrians allowed in highways; no cars allowed in super small roads).
  + **Weather Condition**: During rain, it is less likely (although still possible) for pedestrians to walk.
  + **Demographic Information**: Different age groups travel differently. Younger people can walk a greater distance at a greater speed than elderly people.
  + **Medium of Transportation**: Ultimately, we will also need this information to train our ML model. We need to “feed” the model with information regarding the medium of transportation used in each trip (perhaps using a separate/different but still relevant dataset) for the model to be able to accurately identify the relationship between the independent variables mentioned above and the dependent variable which we are trying to predict (medium of transportation).
* The challenge would be to gather all this additional information.
* If the dataset could be retrieved, I am quite confident that we will be able to train a machine learning model to predict the medium of transportation. We can then use this model to predict the travel pattern in Phnom Penh and generate afterwards the activity plans of 2 million Phnom Penh residents.