# Data Science Capstone Project

## Title: finding the right neighborhood (The Battle of Neighborhoods)

### Introduction

In today's dynamic world, it is common that people find a new job and have to move to a new city or neighborhood. Let's say a person got a job offer from a big company with great career prospects in another city or maybe another country. If this person accepts the job offer, then he must move to a new location. Probably this person would prefer to move to a location that is similar to the place he lives currently in. In this way, he can continue to follow his hobbies and habits and can integrate easier and faster. He has access to venues of his interest in his current neighborhood like gym, swimming pool, cinema, theater, amusement park, restaurants, coffee shops, etc. in the new location, too. To this end, here we want to provide a possibility to find out what are the similar neighborhoods in the new city that are similar to the current neighborhood.

### Dataset

To solve this problem, we need the borough and neighborhood data of the current neighborhood and the destination. As an example, here we choose as the current city New York. We can get a list of New York City Neighborhood Names from the https://geo.nyu.edu/ website. The data is in JSON format and it can be very easily transformed into the Pandas data frame. In this project, we will use only the related information including borough, neighborhood, latitude, and longitude. For more information about the data, please visit the website. Once we have the list of neighborhoods and corresponding latitudinal and longitudinal information we will use Foursquare API to get the venues near each neighborhood.

| | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Bronx | Wakefield | 40.894705 | -73.847201 |
| 1 | Bronx | Co-op City | 40.874294 | -73.829939 |
| 2 | Bronx | Eastchester | 40.887556 | -73.827806 |
| 3 | Bronx | Fieldston | 40.895437 | -73.905643 |
| 4 | Bronx | Riverdale | 40.890834 | -73.912585 |

Figure 1. Sample data from New York dataset

We choose Toronto city as the destination. Because the data cannot be directly downloaded, Postal Code, borough, and neighborhood are scrapped from the Wikipedia website. The borough name is not available for some rows. We will not include these rows in our analysis. However, if the neighborhood is not assigned but the borough is assigned then we consider the corresponding borough as the neighborhood, too. We will merge the rows if Postal Code and borough of two or more rows are the same and the corresponding neighborhoods will be separated by a comma ",". Latitude and longitude information can be downloaded from this address: https://cocl.us/Geospatial_data. Next, this data can be merged with Toronto data together. Once we have the list of neighborhoods and corresponding latitudinal and longitudinal information we will use Foursquare API to get the venues near each neighborhood. To this end, we need to have a Foursquare account to get the required credentials.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Regent Park, Harbourfront | 43.65426 | -79.360636 | Roselle Desserts | 43.653447 | -79.362017 | Bakery |
| 1 | Regent Park, Harbourfront | 43.65426 | -79.360636 | Tandem Coffee | 43.653559 | -79.361809 | Coffee Shop |
| 2 | Regent Park, Harbourfront | 43.65426 | -79.360636 | Cooper Koo Family YMCA | 43.653249 | -79.358008 | Distribution Center |
| 3 | Regent Park, Harbourfront | 43.65426 | -79.360636 | Body Blitz Spa East | 43.654735 | -79.359874 | Spa |
| 4 | Regent Park, Harbourfront | 43.65426 | -79.360636 | Impact Kitchen | 43.656369 | -79.356980 | Restaurant |

Figure 2. Sample data from Toronto dataset

## Methodology

To find the matching neighborhoods, first, we will find the nearby venues to the current borough. Next, we will find the venues in the neighborhoods of the destination borough. Considering the fact that some of the venue categories may not exist in the other borough, we will only consider common venue categories. Now using a similarity measure, we can compare the similarity between the current neighborhood and the neighborhoods in the destination borough to find a neighborhood that has the highest similarity.