# Coursera

# IBM Data Science Capstone Project

## Problem Definition & Data

### Background

The Coursera IBM Data Science Professional Certification course consists of 9 online courses that covers topics including open-source tools and libraries, Python, databases, SQL, data visualization, data analysis, statistical analysis, predictive modeling, and machine learning algorithms. The course finishes with a Capstone project.

The purpose of this Capstone project is to demonstrate the use of the data science toolsets, methodologies, and skills that have been acquired during this course to help solve a business problem.

### Problem

In this project I am looking at the possibility of opening a new Martial Arts School in Toronto, Ontario. I am going to explore neighborhoods and boroughs of existing schools and try to determine a potential location for a new school.

## Data

For this project, the data that will be used:

- List of districts on Toronto
    - Data will be acquired through web scraping the Canadian Postal Code's Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Geospatial coordinates of the neighborhoods and boroughs of Toronto
    - Data will be captured using the Geocoder package and stored into a csv file for easy consumption – the use of the Geocoder package is no longer free
- Top venues of neighborhoods and boroughs of Toronto
    - Data will be acquired through the use of the Foursquare API

## Methodology

This project will compare suburbs and will determine similarities based on clustering techniques using location data services.

This project uses web scraping techniques to retrieve data from the Canadian Postal Code's Wikipedia page.

The data is then acquired and cleansed in preparation for clustering.

The geospatial locations data import will be merged with the post code data which will enable the data to be visualised over a map of the area.

The data will be clustered and plotted over the map.

The clustering is carried out by K Means and the clusters are plotted using the Folium Library.

The data will be mapped across Toronto and then focused/clustered in on boroughs containing the name 'Toronto'.