**Capstone Project #1 Proposal**

***Predict TalkingData User Demographics (Gender and Age) in School District***

***Problem Statement***

TalkingData, China’s largest third-party mobile data platform, understands that everyday choices and behaviors paint a picture of who we are and what we value. The company is seeking to leverage behavioral data from more than 70% of the 500 million mobile devices active daily in China to help its clients better understand and interact with their audiences.

TalkingData asked tobuild a model predicting users’ demographic characteristics based on their app usage, geolocation, and mobile device properties. Doing so will help millions of developers and brand advertisers around the world pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.

The goal of my project is to predict the demographics of a user (gender and age) at a certain geolocation (eg. School District) based on their app download and usage behaviors.

***Data Source and Schema:***

Since this project is inspired by a Kaggle Competition sponsored by TalkingData, the user behavioral data is collected from TalkingData SDK integrated within mobile apps TalkingData serves under the service term between TalkingData and mobile app developers. Full recognition and consent from individual user of those apps have been obtained, and appropriate data anonymization has been performed to protect privacy.

Another key data source in this project is OpenStreetMap database. Since TalkingData provides the geolocations with longitude and latitude details, I would like to further explore and categorize the locations to predict user demographics.

The data schema provided by TalkingData can be represented in the following table:

