**Battle of the Cities**

*Author: Sebastiaan Hemmer*

**Introduction/Business Problem**

My imaginary traveling agency ‘*Sacramentum’* has a unique different proposition; we offer mystery trips to different US cities without our clients knowing their destination in advance.

Our clients purchase tickets and are only provided information regarding the climate of a destination so they know what clothes they should pack. On the airport, our clients are given an envelope which contains the tickets to their destination. Their destination is drawn completely random at this moment.

Recently, we have been dealing with many customers that were unsatisfied with the mystery location they have traveled to. Because this is impacting our business, we are looking for solutions to improve our customer satisfaction. Therefore, we have decided to design a content-based recommender system that takes user input of 3-5 cities that they rate positively and then recommends a mystery location that will most likely be rated positively by our client.

In order to determine a profile of our clients, we will leverage the Foursquare API to cluster different US cities based on certain features. These features shall mainly consist of the amount of different types of venues and attractions in a city, as well as geographical and demographic data.

Because we will be using unlabeled data and do not have access to a training/testing dataset, we are limited to using unsupervised learning techniques such as clustering.

**Data**

As mentioned in the previous section, we will be leveraging the Foursquare API and combining it with both geographical and demographic data that are publicly available. In order to determine an accurate client profile, we will need enough features to make distinctions between different cities but not too many as that will overfit and overcomplicate our model. Therefore, we have selected the following data per city:

*Demographical data*

This will be our starting dataset, we will select the top 1000 largest US cities by population. We will use the available dataset on ODS, which will provide us information about the:

* Name of city
* Population
* Coordinates

(source: <https://public.opendatasoft.com/explore/embed/dataset/1000-largest-us-cities-by-population-with-geographic-coordinates/table/?sort=-rank>)

*Geographical data*

We will then combine our city demographical data with geographical data from open NOAA datasets, such as:

* Elevation
* Average annual temperature
* Average annual precipitation

*Foursquare API*

Lastly, we will include the Foursquare API to include venue data. We have selected the following features of importance:

* Top 5 most popular venue categories (in 5km radius from center)
* Amount of venues in a 5km radius from the center
* Average amount of likes per venue