# Discovering Health-Related Needs in the Community with Data Science and Open Data

Germán H. Alférez, Ph.D.

Global Software Lab, School of Engineering and Technology, Universidad de Montemorelos, Mexico

harveyalferez@um.edu.mx, www.harveyalferez.com

## Abstract

As Seventh-day Adventists, we have been called to minister people's needs. However, discovering the real needs of people in our communities is not an easy task. At the Global Software Lab, School of Engineering and Technology, Universidad de Montemorelos, we are currently working on several research projects to analyze open data by means of data science in order to discover health-related needs in Mexico and the US. This poster summarizes four of these projects.

## Current Projects

### 1. Application of Data Science to Classify Causes of Maternal Mortality in Mexico

✓ **Abstract:** In Mexico, the maternal mortality ratio is very high: 38 deaths per 100,000 live births in 2015 (WHO). The early discovery of possible causes of maternal mortality is key to reduce the number of deaths. To this end, we created a software to automatically classify patients in the following two causes of death by means of applying data science and machine learning to open data provided by the Mexican Ministry of Health: eclampsia during labor and postpartum hemorrhage [1].

✓ **Method:** IBM Methodology for Data Science [2].

✓ **Results:** The model generated with Naïve Bayes was chosen to carry out classifications within the software (precision = 0.75). The model was trained with 1,018 instances. In the case of eclampsia during labor: precision = 0.71, recall = 0.94, and $F_1$ score = 0.81. In the case of postpartum hemorrhage: precision = 0.81, recall = 0.43, and $F_1$ score = 0.56.

| Classifier | Accuracy | Precision | Recall | $F_1$ score |
|---|---|---|---|---|
| KNN | 1.00 | 0.60 | 0.59 | 0.60 |
| Logistic Regression | 0.73 | 0.73 | 0.73 | 0.73 |
| **Naïve Bayes** | **0.72** | **0.75** | **0.74** | **0.71** |
| SVM | 1.00 | 0.54 | 0.57 | 0.42 |

Table 1. Results with the generated models

### 2. Application of Data science to Confirm the Relationship between Dental Caries and Diabetes in Dental Records

✓ **Abstract:** In Mexico, there were over 11 million cases of diabetes in 2015 (IDF). Our contribution was to apply novel data science techniques to medical records at a dental clinic in Northeast Mexico to discover the relationship between diabetes and dental caries [3]. The analysis of data was carried out by means of unsupervised learning (K-Means). Experiments were performed on 193 dental records. Our findings corroborate the results in related work.

✓ **Method:** IBM Methodology for Data Science [2].

✓ **Results:**

1. Diabetic patients tend to present teeth loss and food accumulation in some zones.
2. Diabetic patients tend to present 9 to 17 teeth with caries, whereas healthy patients tend to present 1 to 9 teeth with caries.

### 3. Discovering Hidden Patterns in US Health-Related Open Data with Machine Learning

✓ **Abstract:** We applied unsupervised learning (K-Means) to discover hidden patterns in the Community Health Status Indicators (CHSI) open dataset provided by the Centers for Disease Control and Prevention (CDC) on Data.Gov [4]. It contains over 200 measures for each of the 3,141 US counties.

✓ **Method:** IBM Methodology for Data Science [2].

✓ **Results:**

1. An increasing number of primary care physicians/dentists per county correlates with a decreasing number of people with diabetes per county.
2. Increasing numbers of Medicare beneficiaries and community/migrant health centers per county correlate with a decreasing number of people with diabetes per county.

### 4. Discovering Mission-Oriented Patterns with Open Data in New York City

✓ **Abstract:** We analyzed an open dataset of motor vehicle collisions in NYC, which is freely provided by NYPD. This dataset registers vehicle collisions in Bronx, Brooklyn, Manhattan, Queens, and Staten Island from 2014 to 2016. It contains 932,904 registered incidents. Each registered incident has 30 variables. We applied the K-Means algorithm to this dataset [5]. We have a big interest on applying data science to understand the needs of people in NYC and take action. See our previous results on Adventist Review [6].

✓ **Method:** IBM Methodology for Data Science [2].

✓ **Results:**

1. On Thursdays, Fridays, and Saturdays, drivers tend to drive aggressively. This situation increases the number of accidents during those days.
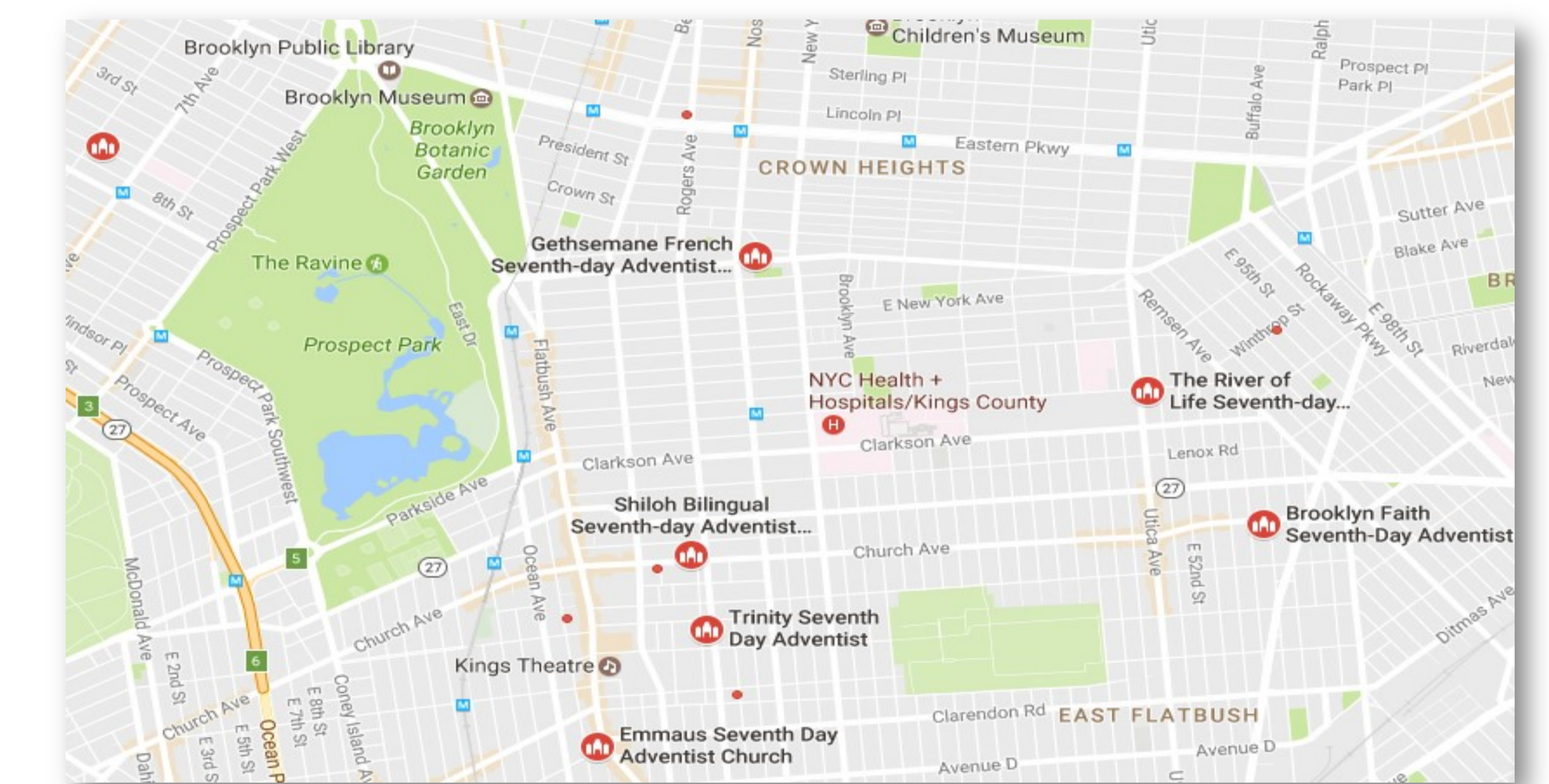2. On Fridays, around Prospect Park, Brooklyn, there were over 77,000 pedicab accidents.



Figure 1. SDA churches near Prospect Park

## Future Work

We are planning to apply data science to data from Adventist World Radio and the North American Division. We are currently finishing a project to understand the needs of people in the 10/40 Window.

1. R. Domínguez, "Aplicación de ciencia de datos para la creación de software predictivo de morbimortalidad materna en México," M.S. thesis, School of Eng. and Tech., Universidad de Montemorelos, Mexico, 2017.
2. *Foundational Methodology for Data Science*, IBM, Somers, NY, 2015.
3. G.H. Alférez et al. "Application of data science to discover the relationship between dental caries and diabetes in dental records," presented at the 2016 International Conference on Health Informatics and Medical Systems (HIMS 2016), Las Vegas, NV, 2016.
4. G.H. Alférez. (2016, November 11). *Discovering hidden patterns in US health-related open data with machine learning* [Online]. Available: http://www.sdadata.org/blog/discovering-hidden-patterns-in-us-health-related-open-data-with-machine-learning
5. G.H. Alférez. (2016, May 1). *Discovering Mission-Oriented Patterns with Open Data in New York City* [Online]. Available: http://www.sdadata.org/blog/discovering-mission-oriented-patterns-with-open-data-in-new-york-city
6. G.H. Alférez, "Tweeting in New York City - data science can teach us to sympathize," Adventist Review, vol. 193, no. 2, pp.47-49, Feb. 2016.