Bloomberg

Skills-based Volunteering: Providing Data Science Resources to Non-Profit Organizations

Michael Campellone
Bloomberg L.P. Global Data
120 Park Ave
New York, NY 10165
+1 646 324 2106
mcampellone2@bloomberg.net

Ikkei Itoku
Bloomberg L.P. Global Data
100 Business Park Dr
Skillman, NJ 08558
+1 609 279 4844
iitoku2@bloomberg.net

Mia Zhao
Bloomberg L.P. Global Data
100 Business Park Dr
Skillman, NJ 08558
+1 609 279 3338
czhao17@bloomberg.net

Anna Cianciara
Bloomberg L.P. Global Data
100 Business Park Dr
Skillman, NJ 08558
+1 609 279 4804
acianciara@bloomberg.net

Nicole Barberis
Bloomberg L.P. Global Data
100 Business Park Dr
Skillman, NJ 08558
+1 609 279 4629
nbarberis@bloomberg.net

ABSTRACT

In this paper, we describe one of the first skills-based data science volunteer initiatives at Bloomberg.

General Terms

Algorithms, Measurement, Economics, Experimentation, Human Factors, Standardization, Theory, Verification.

Keywords

Bloomberg, Clean Ocean Action, Data for Good Exchange, d4gx.

1. INTRODUCTION

As non-profit organizations continue to face challenges in backing their respective initiatives to drive policy change and receive the necessary investment from both government and private sector sources, the increasing reliance on data to support their claims becomes fundamentally critical. That being said, non-profit organizations often lack the necessary resources internally to perform the required data science approaches and techniques to fully leverage the power behind the data the organization is collecting. With Bloomberg's existing commitment to supporting philanthropic initiatives across the Public Health, Environment, Education, Arts and Government Innovation spaces, we as an organization want to ensure non-profits have access to the resources crucial to applying data science techniques to data that will ultimately drive policy and behavior change in the future.

In order to pilot this skills-based volunteer initiative, we partnered with Clean Ocean Action (COA), an organization Bloomberg already had an extensive relationship with. Bloomberg has had a long time history of performing Beach Sweeps with COA since 2011. As of July 2015, 185 unique employees have dedicated almost 800 hours across 17 volunteer events.

COA's mission is to "improve the degraded water quality of the marine waters off the New Jersey/New York coast".[1] The organization aims to identify the potential sources of pollution and deliver a solution in order to reduce the level of pollution by using a combination of research, public education, and citizen action.[1]

COA presented our Bloomberg team with Beach Sweep data collected by volunteers from 1993-2014. The data included type and count of garbage collected, site locations where the garbage was collected and the season and date when the garbage was collected. COA collected this data by having volunteers manually write and report the required information onto summary papers. This method of data collection not only created an unstandardized format but also required additional time and resources necessary to have the data inputted into an Excel spreadsheet, thereby increasing the probability of human error. This analysis of Beach Sweep data will be the first of its kind in COA history. The goal is to more accurately monitor trends throughout the years and eventually associate trends with various legislation that has been enacted, as well as consumerism, weather, and industry changes.[1] Ultimately, Bloomberg aims to use this pilot program with COA not only to provide resources and tools to the organization but to

create a model in which a similar approach can be used to help any organization leverage the power of data science and statistical analysis, particularly Bloomberg non-profit partners: American Littoral Society, Thames21, and Arakawa CleanAid.

2. DATA GOVERNANCE AND WEB APPLICATION

The key to utilizing data for operational excellence is the integrity of the underlying raw data. Without clean and standardized data, it is difficult to analyze and make critical business decisions based on it. Thus, our primary goal was to introduce a new data governance method by designing a database schema and developing a web-based application to maintain the data integrity.

2.1. Schema Architecture

The first task of introducing the new data governance system was to thoroughly review the current data retrieval process in order to fully understand the data flow from beginning to end, as data could be deteriorated at any point of the process. To begin, we interviewed the COA and American Littoral Society staff as well as internal Bloomberg employees who participated in the COA beach clean-ups to identify the steps volunteers take to report the collected trash items from beaches and the procedures the COA staff follow to compile the data. Based on the observations and analysis of the 1993-2014 datasets, database schemas were developed in MySQL, for both its reputation and significant presence in the data science community.

2.2. Historical Data Clean-up

After designing the schemas, data munging was required. Due to the data volume and inconsistency, this process demanded a significant amount of time. Each year contained two Excel files with multiple sheets of detailed data collected by COA. Furthermore, there was little format consistency between files and the classified categories had changed overtime. Thus, there was no way to programmatically process these files and manual examination of each category and file was needed. Utilizing volunteer time at Bloomberg, the most recent years of the data were standardized and successfully transferred to the database. The remaining files will undergo a similar process during an upcoming "Bloomberg Datathon" event.

2.3. Web Application for New Data Entry

After the schema design and significant data munging, a new data entry framework was needed for volunteers and COA to update collected items into the database (See Figure 1). To solve this, we created a custom web application based solely on open source solutions so that COA could invest the resources for its core operations. For readability and future maintenance purposes, Python was our language choice and we utilized a micro web framework, Flask, as the backend. The front end was built on top of a popular open source solution, Bootstrap, so the site could be optimized for desktop computers and smartphones. This application will not only solve data entry and integrity problems but will function as a real time data analytics dashboard as well (See Figure 2, 3 and 4). As of September 2015, the application was successfully deployed from an internal Bloomberg web server to Amazon Web Services, more precisely, Amazon Relational Database Service and Elastic Beanstalk.

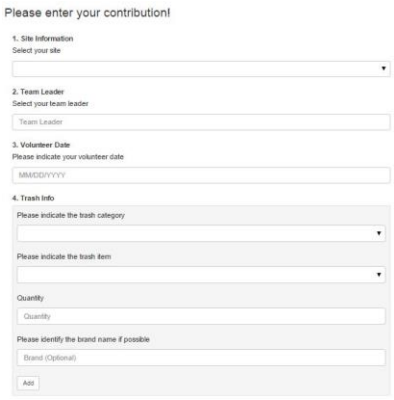

Figure 1: The data entry form on web application

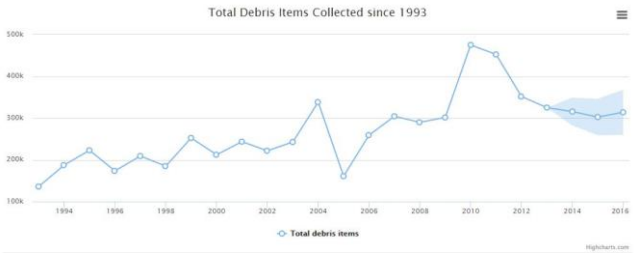

Figure 2: Data analytics dashboard of total debris collected since 1993

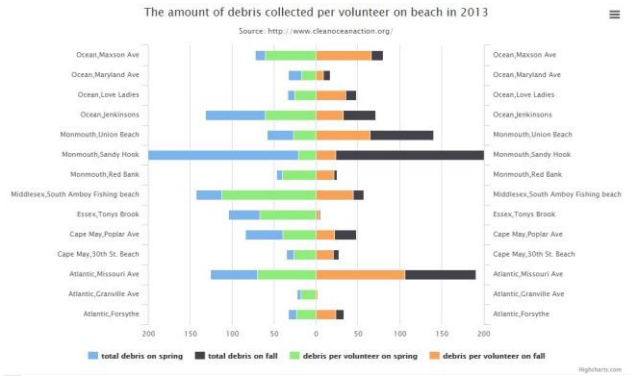

Figure 3: Data analytics dashboard of amount of debris collected per volunteer on beach in 2013

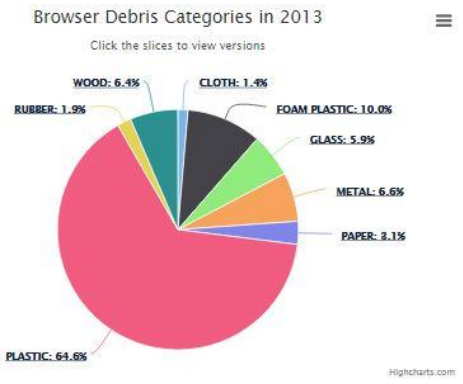

Figure 4: Data analytics dashboard of debris categories in 2013

3. DATA ANALYSIS METHODS

Using the semi-annual (Fall and Spring) trash data provided by COA, we analyzed the volume of trash collected against: macro indicators including GDP, unemployment rate, and New Jersey population; related industry indicators including aggregated inventory, sales, and revenue data; and finally, related company information such as Pepsi and Coca-Cola revenue and sales data from the US market.

We performed Granger causality test on the data series containing strong correlation, setting our constraints to $t<0.05$ as statistically significant and r^2 to 0.8.

Additionally, as part of our analysis we mapped trash categories from the COA data to the Bloomberg Industry Classification Standard (BICS). The Bloomberg Industry Classification Standard is an industry classification system developed and maintained by Bloomberg that classifies companies based on business, economic function, and other characteristics. By mapping the trash category into BICS codes we are able to identify companies in specific industries who could produce the type of trash found in our dataset from COA.

Our analysis utilized random forests to rank the companies and identify which ones might have more influence on trash volume. We then used a forecast skill calculation to measure the effectiveness of our predictions.[2]

4. CONCLUSIONS

It is important to note that our data analysis is not complete. With a majority of the historical data in need of munging, our upcoming "Bloomberg Datathon" will allow our team to complete the analysis outlined in Section 3. When complete, our analysis will allow COA to more accurately identify potential sources of specific types of trash found at various beaches.

The development and implementation of the web application, as mentioned above, will serve four purposes. The first will allow our Bloomberg team to standardize the remaining historical years of data. Second, the application will allow volunteers and COA to standardize the data they are collecting and directly load this data into the database. As a result, future data will no longer need to undergo the time intensive process that was necessary for the historical data we received. Third, we see the application serving as a tool to increase volunteer engagement. Analytics on the web application are updated real time and allow an individual or team leader to visualize the impact they are making. Lastly, the analytics made available to the organization via the web application will allow COA to plan more efficiently from an operations standpoint. We see this application being used to properly allocate resources at the various beach sites and also to allow the organization to reflect on prior Beach Sweeps and strategies that may or may not have worked.

5. ACKNOWLEDGMENTS

Our thanks to Chelsea Bentley, Monica Hilliard, Victoria Cerullo, Catie Tobin, Eduardo Hermesmeyer, Daniel Yielding, Christian Bellmann, John Mahoney, Wendy Huang, Alex Gorden, Elizabeth Campbell, Bloomberg Philanthropy & Engagement and of course, Clean Ocean Action.

6. REFERENCES

[1] http://www.cleanoceanaction.org/index.php?id=334

[2] Allan H. Murphy, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424. DOI= http://dx.doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2