Interactive Data Visualization for Public Health Applications: An Evaluation of Plotly's Capabilities

Summary

Leveraging multiple data variables is critical for effective infectious disease outbreak management. Contextual actionable data provided to decision makers is often best analyzed and conveyed visually. However, limited human resources and collaborative platforms create many challenges for the effective use of data. B.Next executed a project to explore, evaluate, and demonstrate how and to what extent information technology capabilities might empower public health analysts with limited or no coding experience to create enhanced data visualizations during an infectious disease outbreak. Dynamic visualizations with multiple variables have proven beneficial in past epidemic management operations for situational awareness[1]. The popular tools available to produce interactive visualizations require financial resources beyond what is often available to state and local public health organizations or coding expertise to utilize open source coding libraries such as JavaScript, R, and Python. This report outlines the methods and results of a B.Next project with Plotly[2], a company that creates open source tools for visualization, to further develop their existing web-based interface to create interactive cross-filtering visualizations with multivariate datasets for non-coders. Plotly enhanced their Chart Studio tool to enable cross-filtering functionalities with nine different charts in a dashboard display. An evaluation of Plotly's enhancement was completed by both public health subject matter experts and IQT Labs personnel with expertise in infectious disease and visualization tools. The evaluation suggests that the cross-filtering functionality within Plotly provides new capabilities accessible to non-coders. However, Plotly needs additional improvements to compete with more intuitive interfaces such as Tableau[3].

Background

A critical task during an outbreak is interpreting data in order to direct response actions. There are many correlated variables that must be considered to answer three important questions during an outbreak: (1) what is happening?; (2) where is it happening?; and (3) when do we anticipate impacts? Useful variables include, but are not limited to, location of outbreaks and transmissions, risk demographics, economics, case counts, transmission severity, personnel impacted and needed for the response, and infrastructure resource status. To coordinate an effective response, analysts must interpret these variables and provide decision makers with contextual actionable data, which is often best analyzed and conveyed visually.

[1] Findings from discussions at B.Next Roundtable – Data Analytic Technologies for Key Decisions; held August 2016
[2] https://plot.ly/
[3] https://www.tableau.com/

© 2018 In-Q-Tel, Inc.

The need for improved data visualizations was a recurring theme at our Key Decisions Workshop and Roundtable, held in August 2016. Understanding the demographics of individuals infected by an outbreak and the corresponding spatial variations were major challenges during the Ebola epidemic. Decision makers wanted to be able to correlate the status and movement of the outbreak with several other cultural and resource related variables, but there were many stakeholders managing their own data at various stages of the outbreak. Sharing and collaborating in one platform with various partners was challenging due to the limited human resources to develop and manage a central platform. Moreover, combining data from multiple sources into one visualization platform required a team of data analytics and visualization experts which are in low supply in resource-limited environments[4].

Most often, to create data visualizations for epidemic management within public health organizations, non-technical analysts rely on Microsoft Excel and, where resources are available, business intelligence dashboards, such as Tableau. Public health data scientists or academic groups with strong analytic skills and abilities utilize non-commercial grade, bespoke tools and open source libraries. For dynamic visualizations, public health analysts are at a disadvantage because many do not have the financial resources for expensive business intelligence "point and click" tools, nor do they have access to the required coding expertise to use popular visualization open source libraries. B.Next explored commercially-available visualization tools with the goal of demonstrating how such tools could provide the ability to create dynamic, multivariate visualizations in a user-friendly interface, without the need for coding expertise and at a more affordable cost for public health organizations.

B.Next executed a project with Plotly to leverage their existing, commercially available, and collaborative visualization tool that enables non-coders to quickly explore and display multivariate datasets in multiple views simultaneously. Plotly offers a suite of open source libraries for data scientists, as well as business intelligence tools and dashboards for non-technical analysts that can be used for free if the data is made publicly available, or used privately with a paid subscription[5]. Furthermore, Plotly's cloud platform hosts its graphics in a way that improves scientific reproducibility by bundling scientific figures with their underlying data, version history, and discussion among collaborators. Plotly integrates a wide variety of existing tools and work flows, and is language flexible, allowing coding analysts to use the most popular data science languages and libraries (Data-Driven-Documents (D3), R, Python®, Julia) to create, share and edit charts. For example, if an interactive chart was produced in R, Plotly allows other users to download and edit the same chart in Python®.

The technique of cross-filtering between plots and maps (which uses a shared data key but different axes) has become a popular means to quickly explore and display multivariate datasets. The most powerful and widely used cross-filtering programming language is

[4] Fast, Larissa and Adele Waugaman. 2016. Fighting Ebola with Information: Learning From Data and Information Flows in the West Africa Ebola Response. Washington, DC: USAID
[5] Learn more about Plotly's tools and services at https://plot.ly

© 2018 In-Q-Tel, Inc.

JavaScript due to its graphical capabilities that are not available in traditional scientific programming languages; however, JavaScript can be difficult to master. Before our project, R Shiny® Server (RStudio®) combined with Plotly.js (Plotly's JavaScript library) enabled powerful, open-source cross-filtering in R (a programming language widely used for its statistical computing and graphics); however, this approach is limited to the R language. Plotly.js is a free open source JavaScript charting library that can be downloaded to interface with other open source libraries. To enable cross-filtering capabilities across all programming languages that interface with Plotly, this project developed robust RStudio® and JavaScript cross-filtering capabilities into the core Plotly.js feature. For users who are unable or prefer not to use code, Plotly.js has a graphical user interface (GUI), Chart Studio, for data upload and chart creation. Prior to our project, Chart Studio did not include any cross-filtering capabilities. The new cross-filter interaction added to Plotly's dashboard environment was designed using datasets in the field of public health to ensure relevance to researchers in this field.

Explanation of the enhanced tool

Plotly enhanced their web-based interface to produce cross-filtering dashboards with charts in Cartesian axes and geographic maps. Using Plotly's Chart Studio, a non-technical analyst can create nine different charts compatible with cross-filtering: Mapbox maps, D3 maps, scatter plots, 3D scatter plots, box plots, histograms, choropleth maps, bar charts and bubble charts. The Chart Studio import tool supports .csv file uploads, URLs from GitHub, and SQL database connections for live updating. The data should be structured and prepared for analysis prior to importing. For geographic maps, Chart Studio supports latitude and longitude coordinates, GEOJSON files, state and country names, and abbreviations, but does not yet support zip or country codes. The original and manipulated datasets are version-tracked within each plot by a source tree. The source tree displays the root nodes that make up the plot and a link to the dataset within Plotly. The user has the option to include metadata and links to the original source.

The cross-filtering charts are capable of displaying up to six datasets and six variables via Plotly's Chart Studio and up to 30 datasets and 30 variables within Plotly.js (open source libraries). These charts can then be added to a dashboard that allows the user to enable cross-filtering via four filter options:

1. Rectangular selection by click-drag: default cross-filter selection for charts and maps. The selection makes a union of the selected x and y data and sends that selection to the cross-filter.
2. Non-rectangular selection by click-drag: Plotly's mode bar allows the user to change the cursor mode to lasso for non-rectangular data selections.
3. Drop-down menu: the user can enable and create drop-down filter options by selecting a data column, equality operator (>, <, or ==), or a single data value. The user can also enable a search bar feature to highlight textual data within the charts.
4. Pie chart selection: cross-filter selections in pie charts are triggered by clicking one of the "slices" of the pie.

© 2018 In-Q-Tel, Inc.

There are limitations in the tool that could affect the user experience to include the following:

1. Chart Studio was not designed as a "big data" tool. For an optimal user experience, the cross-filtering dashboard datasets should not exceed 5,000 rows[6].
2. The cross-filtering dashboards do not support common key identifiers. If dashboard charts are made from multiple datasets of different columns/variables, they must have the same number of rows, and the rows should be ordered in the same way. If multiple datasets are used with different numbers of rows, cross-filtering will automatically be disabled because it will be unclear how to link the datasets. Note: Plotly recommends combining data from various sources into one single dataset prior to importing into Chart Studio.
3. The drop-down menu filter only supports a specified value, instead of selecting a range of data values, and all filter values are limited to data values in the dataset. Dates are all also not supported as a data value.
4. The time to restyle dashboard plots after cross-filtering ranges from one to six seconds due to a performance bottleneck within Plotly.js.
5. Table charts only respond to cross-filtering and cannot be used to initiate cross-filtering selections.
6. The x-axis will not automatically resize after filtering. The user has to manually adjust and resize the x-axis.

The individual charts within the dashboards can be exported as an image in six different formats and embedded via HTML; the datasets can be downloaded in four different formats; and the code can be exported in six different languages[7]. The cross-filtering dashboards are shareable online via links; as such, viewers do not need to download any additional software or create accounts for viewing. Viewers also have the option to create free public accounts and fork and edit charts to apply changes to the cross-filtering analysis for further data exploration.

[6] From a local, 'in-memory'data source or csv file, Tableau Desktop is limited to 100,000 rows. However, most Tableau Desktop users query against a live, connected database (e.g. SQL database) that can allow billions of rows.

[7] Image export formats: .png, .jpeg, .svg, .pdf, .eps, .webp
Data export formats: .csv, .xlsx, .ppt, .json, .ipynb
Code export formats: Python, MATLAB, R, Julia, plotly.js, Node.js

© 2018 In-Q-Tel, Inc.

Evaluation Process and Results

To assess how Plotly's platform enhancements could benefit the public health community, we tested the overall functionality with various chart types and public health datasets and solicited qualitative feedback from public health subject matter experts. Publicly available datasets were downloaded from the University of Pittsburgh and the Humanitarian Data Exchange to evaluate the tool's ability to clarify relationships among various data dimensions and reproduce operationally relevant cross-filtering dashboards. Qualitative reviews from our public health evaluators consisted of several discussions and surveys on whether and how the tool could benefit their current workflow and the tool's ease of use. Below are a few examples of dashboards with different chart types for public health applications[8].

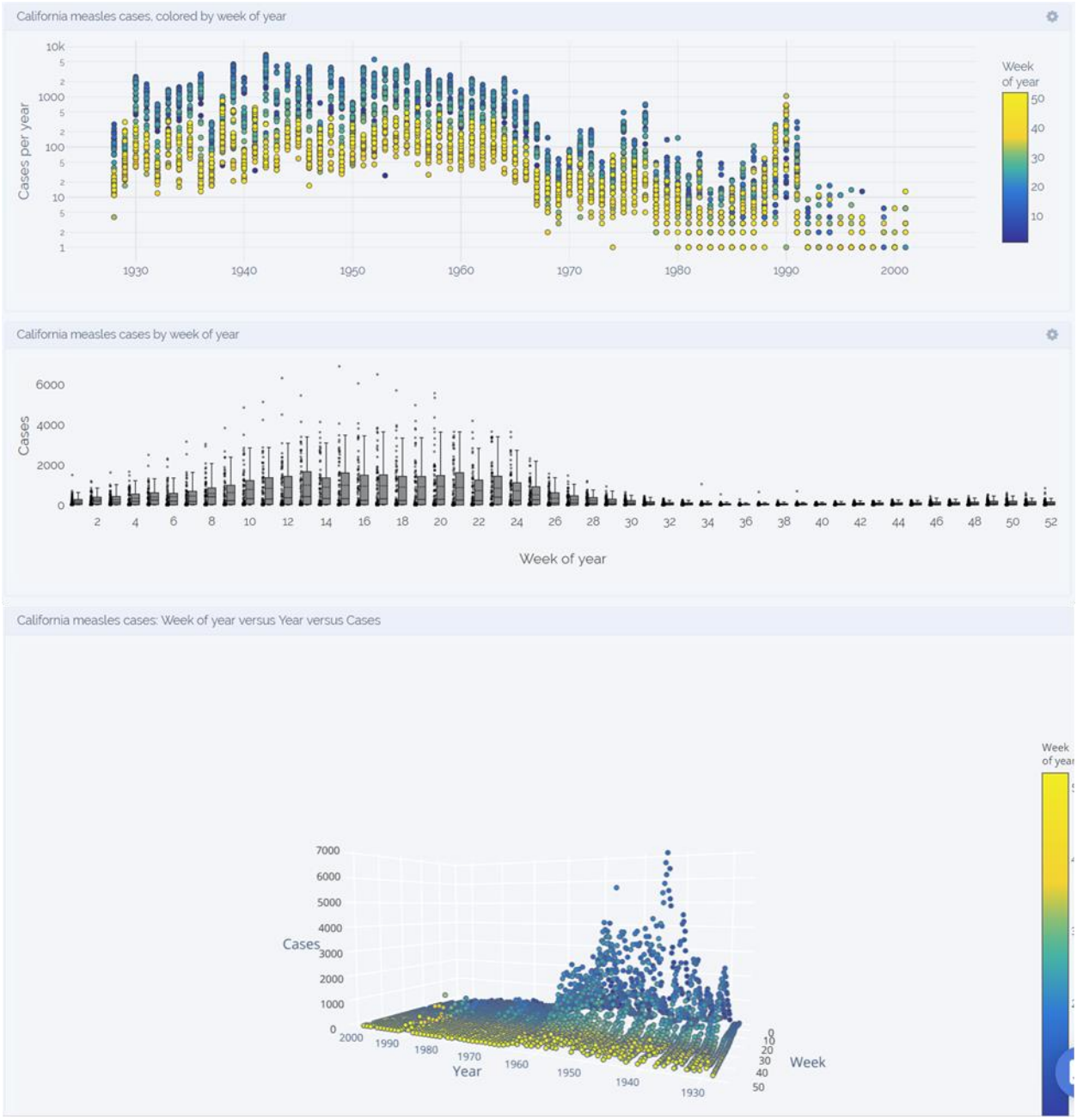
Case Counts Over Time

The University of Pittsburgh digitized the entire history of weekly National Notifiable Disease Surveillance System (NNDSS) reports for the U.S. from 1888 to 2013 as part of their Project Tycho database[9]. Measles data from the Project Tycho database were downloaded to demonstrate Plotly's cross-filtering tool with public health data. Figure 1 displays a 2D scatter chart, box plot, and 3D scatter plot of measles cases in California by year and week between 1928 and 2004. The measles vaccine was distributed in the U.S. in 1968 and the dashboard visualizations in Figures 1 and 2 highlight the effectiveness of the vaccine and the seasonal impact on measles outbreak.

[8] The public health examples in this report along with five others are documented with step-by-step instructions to create each visualization and the original datasets on GitHub.

[9] Willem G. van Panhuis, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Y Lee, Vladimir Zadorozhny, Shawn Brown, Derek Cummings, Donald S. Burke. Contagious Diseases in the United States from 1888 to the present. *NEJM* 2013; 369(22): 2152-2158.

© 2018 In-Q-Tel, Inc.



[Figure 1](#) – Unfiltered dashboard for California measles cases from 1928 to 2004.

© 2018 In-Q-Tel, Inc.

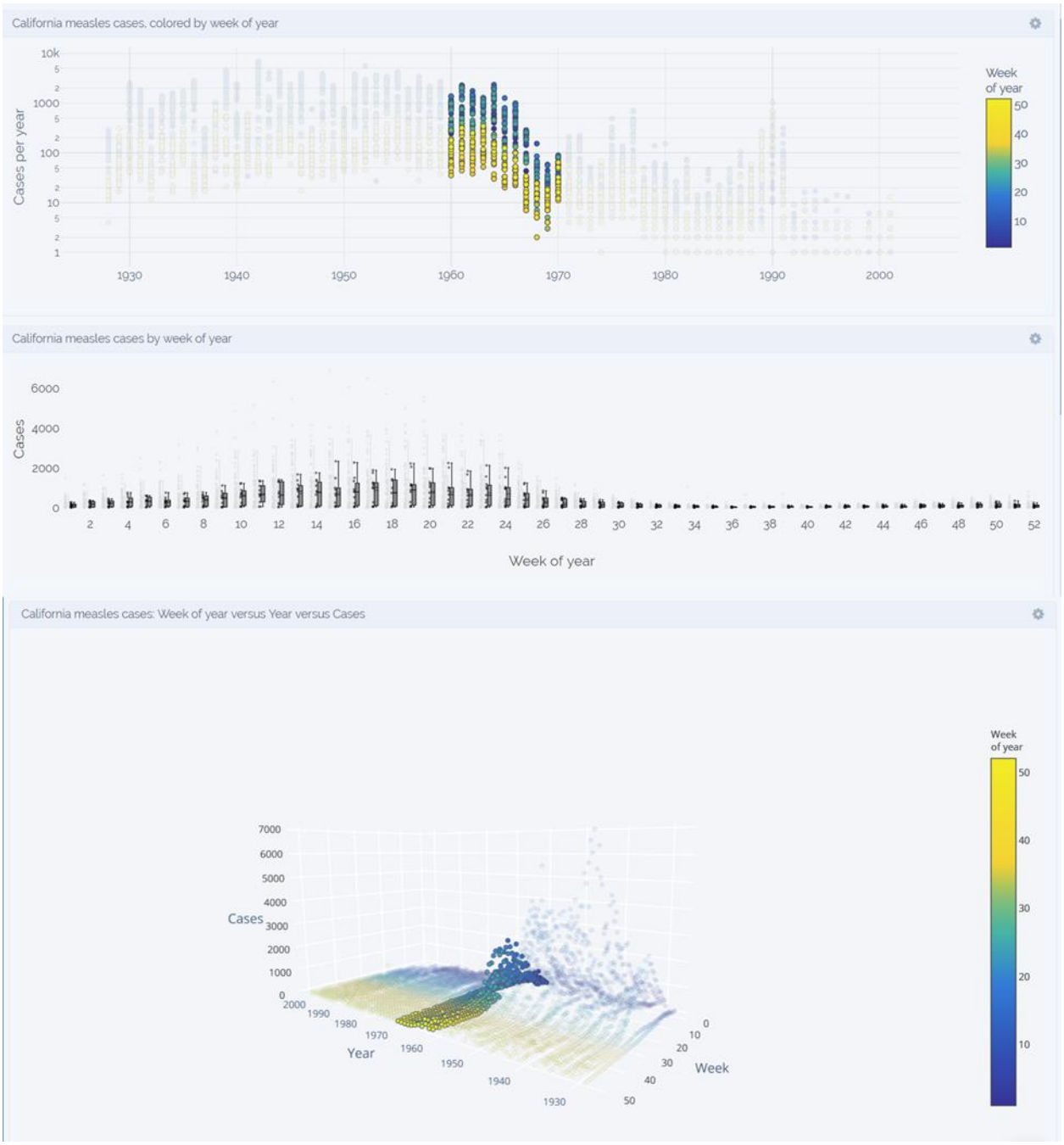


Figure 2 – Filtered dashboard for California measles cases between 1960 – 1970 via data point selection

Tracking Response Organization Activities

During the Ebola crisis, the Humanitarian Data Exchange (HDX) implemented an online tool which served as a central repository for public and restricted datasets. These datasets included information about the organizations involved in the response to the Ebola outbreak. Simon B. Johnson, a data manager with the British Red Cross, recognized the

© 2018 In-Q-Tel, Inc.

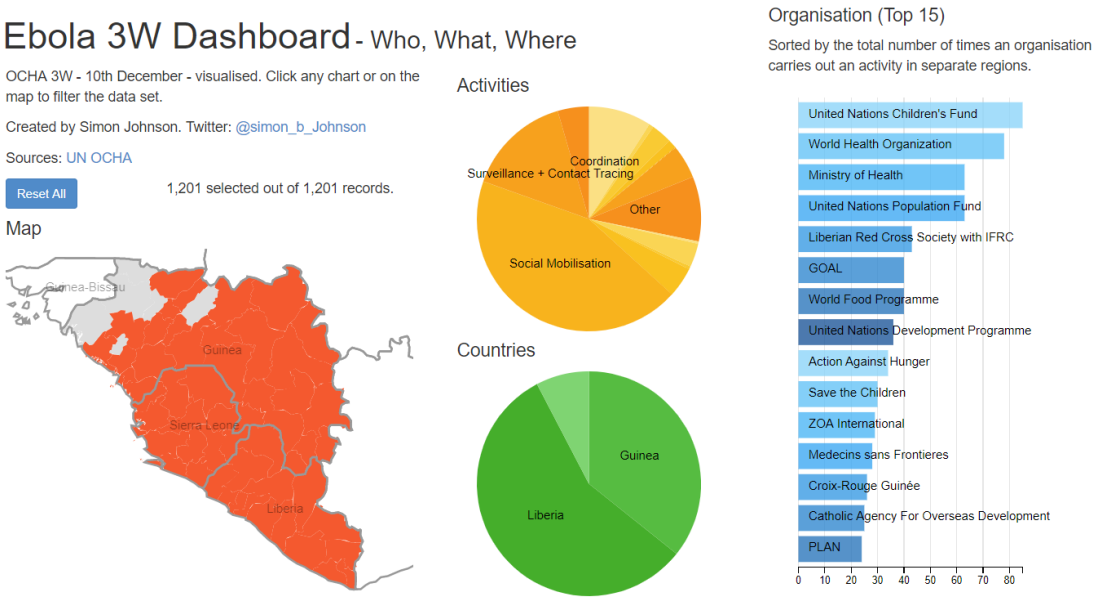
challenges with static maps to understand the gaps in response and the latest status of where organizations were working[10]. Because he was unfamiliar with JavaScript D3 libraries, it took him several days and iterations to learn JavaScript and create a semi-functional interactive cross-filtering dashboard. After Johnson's first iteration with the Ebola outbreak, he received several inquiries for interactive dashboards for other humanitarian efforts but was restricted to coding custom dashboards per use case. Eventually, together with a team of developers, he was able to create a reusable version of his dashboard, called the 3W Dashboard.

USAID wrote a report in 2016 called "Fighting Ebola with Information: Learning from the use of data, information, and digital technologies in the West Africa Ebola outbreak response"[11] that highlighted Johnson's efforts and the specific challenge with data demands and the need for data experts with visualization expertise. Although Johnson's code driven visualizations with the HDX data were successful, a tool with the ability to create interactive dashboards for non-coders would have been useful. Within Plotly's Cloud platform, coders and non-coders can collaborate in an interface that accommodates each user's technical background. The same data used in Johnson's dashboard (Figure 3) was imported into Plotly's Chart Studio to create a dashboard with similar functionality and without the need to write code. The first version of the original 3W Dashboard took several days and had limitations with the ability to pan and zoom on the map to explore the depth of the data, while Plotly's version of the dashboard was recreated in 3 to 4 hours by a first time user. Some of the major differences in functionality between Johnson's and Plotly's dashboard include the following:

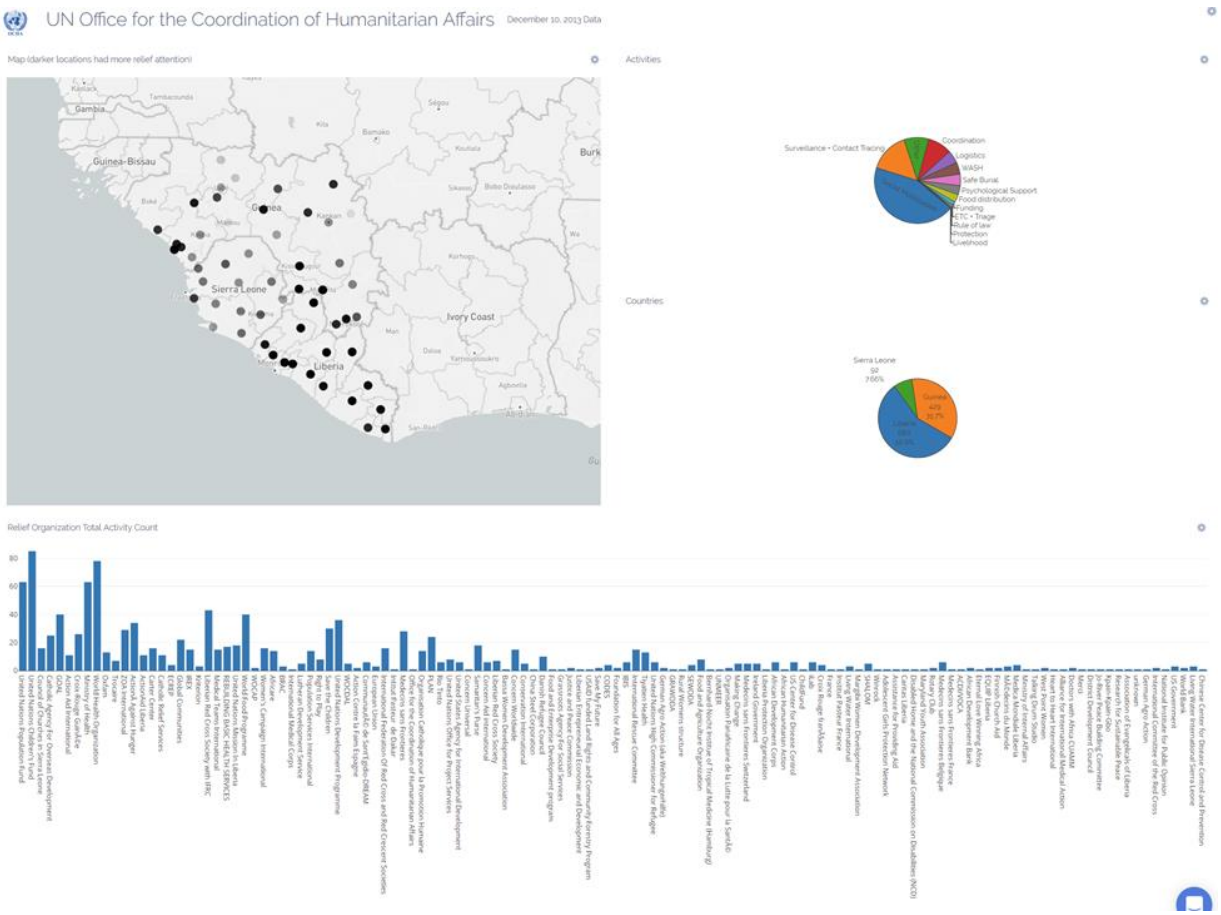
(1) Plotly users select areas on the map vs choropleth regions in the 3W Dashboard. Plotly's finest resolution in choropleths is limited to whole countries and US states.
(2) Instead of being limited to clicking single data points like in the 3W Dashboard, Plotly users are able to select groups of data points simultaneously by click-drag.
(3) Plotly users can enable a search bar feature to assist with filtering.
(4) In Plotly, a table showing what has been selected is optionally available.

[10] https://medium.com/@Simon_B_Johnson/who-is-doing-what-where-visualisation-from-concept-to-use-6aa0ae1c3cd5
[11] Fast, Larissa and Adele Waugaman. 2016. Fighting Ebola with Information: Learning From Data and Information Flows in the West Africa Ebola Response. Washington, DC: USAID. https://www.usaid.gov/sites/default/files/documents/15396/FightingEbolaWithInformation.pdf

© 2018 In-Q-Tel, Inc.



[Figure 3](#) – Simon B. Johnson's Ebola 3W Dashboard



[Figure 4](#) – Recreation of Simon B. Johnson's 3W Dashboard for Aid Agencies during the Ebola Outbreak (unfiltered)

© 2018 In-Q-Tel, Inc.

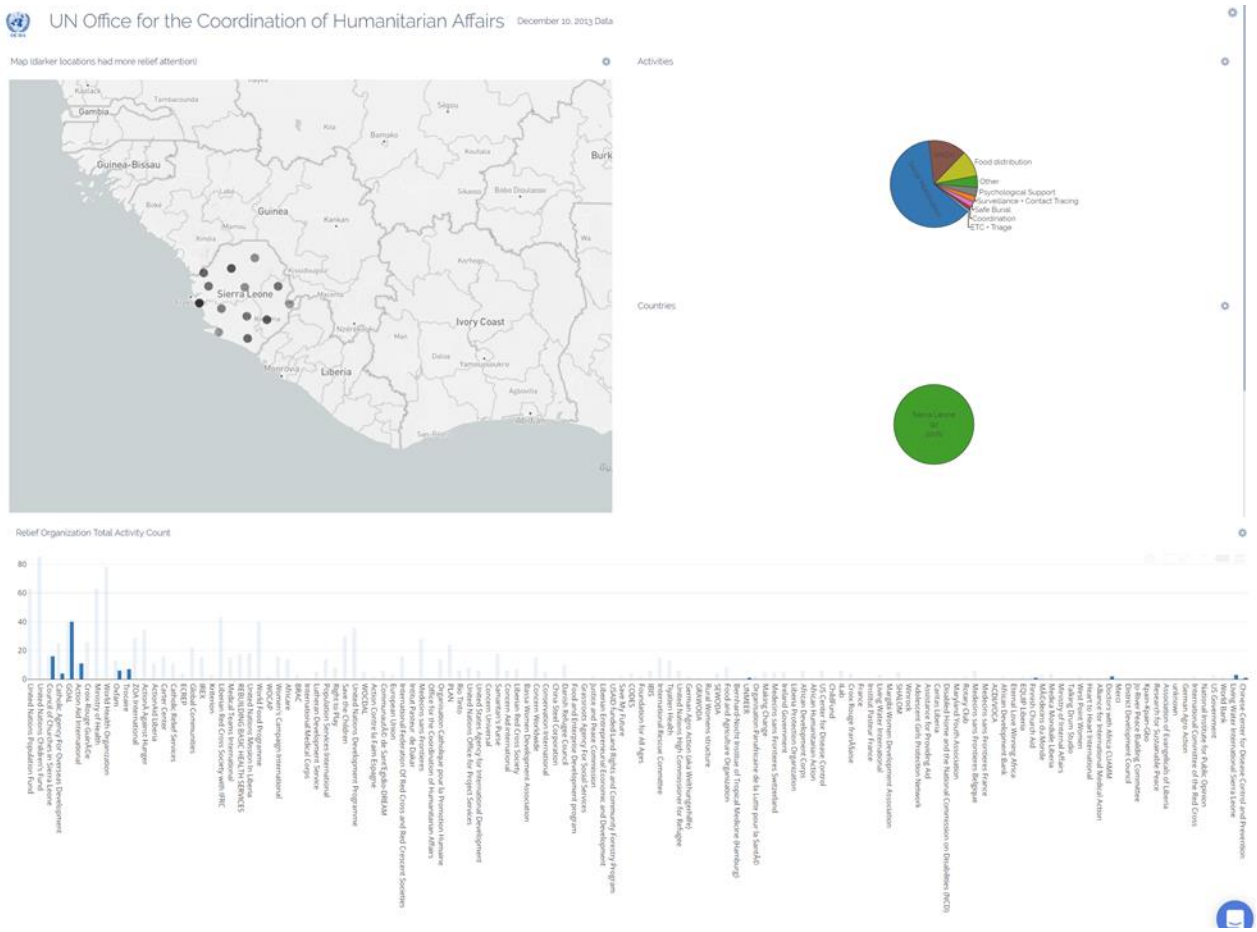


Figure 5 – Recreation of Simon B. Johnson's 3W Dashboard for Aid Agencies during the Ebola Outbreak (filtered on Sierra Leone)

Generating Reports

Plotly's dashboard also has a report mode that contains a rich-text editor to produce web-based reports with D3 interactive graphs for online sharing. Plotly's Cloud platform could be used as an end-to-end tool to upload structured data, create dynamic visualizations, and produce reports. Public health analysts often complain about one-off platforms that have to be integrated with other tools to produce a final product; however, Plotly has enabled a workflow to create visualizations, collaboratively explore data dynamically, and produce data driven reports within one platform.

© 2018 In-Q-Tel, Inc.

Public Health Subject Matter Experts Assessments

B.Next engaged subject matter experts from reputable public health organizations to assess how this tool could impact their workflow. These evaluators were chosen because of their experiences working with public health analysts, or as an analyst themselves, during an epidemic. Each public health entity is unique in their make-up of technical skills, infrastructure, and resources. Our evaluators provided their assessment with these diverse communities in mind, including (1) state and local public health agencies, that are typically limited in technology, resources, and technical skill, (2) a structured and specialized organization with less-limited resources and expertise, and (3) a large entity with diverse skills and resources.

The experts reiterated the status quo for public health broadly of resource limited, non-technical analysts, explaining the current process to generate static charts individually and copy and paste them into Microsoft Word to produce a report. When new data becomes available, such as case count updates during an epidemic, some analysts have to recreate a visual from scratch every time. With Plotly's enhancement, several charts within one dashboard can be updated simultaneously by manipulating the original datasets. The tool also provides the capability to explore data dynamically within a dashboard report view as an alternative to static Microsoft Word reports. Rather than spending time reproducing static reports and visuals, Plotly's tool allows the user to focus his time on data exploration. For traditional state and local health departments, our colleagues believe this tool could improve their workflow because it allows new capabilities to explore data dynamically and utilize interactive dashboards as their reoccurring reports.

Tableau and R are two popular tools used to clean and manipulate data and produce visualizations. Tableau Desktop provides users with an intuitive point and click interface to create dynamic visualizations, without the need for tutorials or instructions. Although Tableau Desktop is limited to Windows operating systems, it is capable of connecting to a wider suite of data sources and contains data quality and auditing functionalities[12]. Tableau Desktop also provides the functionality to develop code to enhance visualization and alerting capabilities. Compared to Tableau Desktop, some of our evaluators described Plotly's Chart Studio interface as a less intuitive tool that requires a demo or tutorial to navigate.

For our technical evaluators with coding expertise in dynamic visualizations, this project demonstrated approaches to reduce the amount of overhead to produce customizable dashboards given the flexibility to use familiar programming languages such as R and Python®. Currently there are analysts within federal organizations who use Plotly because it makes D3 visualizations more manageable. However, they are unable to use some of its features that allow simultaneous multi-user collaboration on visualizations because of internal policies governing software and data usage. The Plotly tool is still valuable for coders and non-coders, even if restricted to non-collaborative use. However, the full utility of the tool would provide the greatest impact.

[12] https://www.tableau.com/products/desktop#data-sources-professional

© 2018 In-Q-Tel, Inc.

Conclusion and Future Opportunities

Even though B.Next's project with Plotly is complete, the company plans to incorporate several additional cross-filtering enhancements into their commercial development path that could benefit epidemic management, based on feedback from the subject matter experts. Those under consideration include:

1. Choropleth maps with increased granular spatial capability via the display of aggregated data at the zip code and county level. This enhancement will be specifically useful for state and local health departments.
2. Performance optimization to reduce the bottlenecks in re-styling delays, and increase the dataset limit to 100,000 rows vs. 5,000.
3. Implementation of shared data keys and common identifiers to ensure proper filtering of various datasets with a different number of rows and columns.
4. Additional cross-filter chart types: violin plots, parallel coordinate plots, Sankey diagrams, heat maps, and clustering on maps.
5. Additional filter options in the dropdown menu for selecting time range, aggregation type and location.

Based on the feedback from our evaluators, we also advised Plotly to consider updating their Chart Studio with a more intuitive "point and click" interface to create and style visualizations and pre-built charts and dashboards. While some users will appreciate the options to custom build visualizations, others may become frustrated and overwhelmed with the number of options and features. That said, we recognize that the user experience and benefits of Plotly's cross-filtering tools are dependent on a user's appetite for creativity, their technical expertise, and their overall end goal.

This project illustrates how existing, commercially available tools can be modified to address data challenges for public health applications. Although some of our evaluator's current workflow is amenable to a more established and user-friendly platform, our evaluation demonstrated the enhanced tool in its current state could positively impact the public health community, especially the state and local public health departments.

© 2018 In-Q-Tel, Inc.