Paper SAS1972-2015

Social Media and Open Data Integration through SAS® Visual Analytics and SAS® Text Analytics for Public Health Surveillance

Manuel Figallo-Monge, SAS Institute Inc.

Emily McRae, Ph.D., SAS Institute Inc.

ABSTRACT

A leading killer in the United States is smoking and, moreover, over 8.6 million Americans live with a serious illness caused by smoking or second-hand smoking. Despite this, over 46.6 million U.S. adults smoke cigarettes, cigars, and pipes all of which have been linked to chronic diseases. What about e-cigarettes – how might they affect this public health situation, particularly among each of the 50 United States? In other words, can monitoring public opinions of e-cigarettes using open data (such as Twitter), SAS® Visual Analytics, and SAS® Text Analytics help inform insights on the potential dangers of these new products? To what extent can social media provide lead generation for anti-smoking campaigns, prioritization of surveys, and early detection of e-cigarette outcomes? These are some of many analytic questions which may arise in coming years, as more data pertaining to e-cigarettes becomes available. As such, this paper will not attempt to make any conclusive decisions on the potential role e-cigarettes may play in chronic health conditions. Rather, we will aim to provide a broad overview of a technological solution that can be utilized for the purpose of public health surveillance. In short, this paper will introduce the topics of open data, APIs, SAS® Intelligence Platform (especially, SAS® Metadata, stored processes, macros) as well as SAS® capabilities including text and statistical analysis and visualizations.

The research in this paper is conducted on thousands of tweets from September to October 2014 (a subset of a collection from April to August 2014). It includes API-sources beyond Twitter that are used to enrich Twitter data in order to implement a surveillance system developed by SAS® for the Centers for Disease Control and Prevention (CDC). The Health Indicators Warehouse (HIW), an on-line repository made available by Health and Human Services (HHS) of over 1,200 Federal health indicators accessible via API's, is one such example.

The analysis is especially important to The Office of Smoking and Health (OSH) at the CDC which is responsible for tobacco control initiatives to help states, for example, promote smoking cessation and prevent smoking initiation in young people.

INTRODUCTION

It is the aim of this paper to identify how federal agencies can use both SAS® and API-based data in order to understand trends and patterns through analytics for structured and unstructured data. Integrating multiple data sources is key for the betterment of government programs; e.g., to help identify U.S. states where there may be a strong need for surveys to better understand and fight diseases.

Although this paper provides some insights into public opinion and open data, it also provides some background on how to operationalize analytics using API-based data within the SAS® Intelligence Platform, which will be explained shortly, for public health surveillance.

Public health surveillance is the continuous, systematic collection, analysis and interpretation of health-related data for the planning, implementation, and evaluation of a public health practice.[1]

The diseases that Federal agencies, like the CDC, are seeking to control are ones such as heart disease and cancer. Although outside the scope of this paper, this can ultimately be done through a variety of measures that run the gamut from regulations (e.g., excise taxes) to public health campaigns.

In this paper, over 50,000 tweets have been collected through an automated process over several months for the public surveillance of e-cigarettes. The resulting output has been enhanced using SAS® technologies for text analytics and visualizations. Through the insights ultimately attained by this process, decisions can

be gained to better help Federal agencies, such as the CDC, prioritize programs, allocate resources, and validate investments made.

Key to acquiring or "systematically collecting" data for public health surveillance are API's or Application Programming Interfaces. API's are connectors (much like power outlets that connect buildings to a larger electrical grid) between two disparate systems. In the Federal government, as in the private sector, there has been a tremendous growth in API's in recent years.

Federal government leaders, such as Aneesh Chopra (The first Chief Technology Officer of the United States), have claimed that API's are one of the most important technologies that will have the "greatest impact on how the government functions", especially because they dramatically lower the barriers to information sharing. And because they enable and facilitate the distribution of data, they also promote the government values of transparency, collaboration and participation.

BACKGROUND

A chronic condition is a human disease that is persistent, otherwise long-lasting in its effects (i.e., more than three months), and typically preventable. Because they are not passed from person to person, they are also known as non-communicable diseases.

The four main types of chronic diseases are cardiovascular diseases (such as heart attacks and stroke), cancers, chronic respiratory diseases (such as chronic obstructed pulmonary disease and asthma) and diabetes. [2]

As a nation, the United States spends 86% of its health care dollars on the treatment of chronic diseases. These persistent conditions—the nation's leading causes of death and disability—leave in their wake deaths that could have been prevented, lifelong disability, compromised quality of life, and burgeoning health care costs. [3]

Moreover, as demonstrated in *Figure 1*, poor states with a median household income less than $50,000 typically have a higher lung, trachea, and bronchus cancer (diseases traditionally associated with tobacco usage) death rate per 100,000 population. In the future, these states may be the focus of particular interest by the CDC, if in fact they demonstrate a higher prevalence of e-cigarette usage.

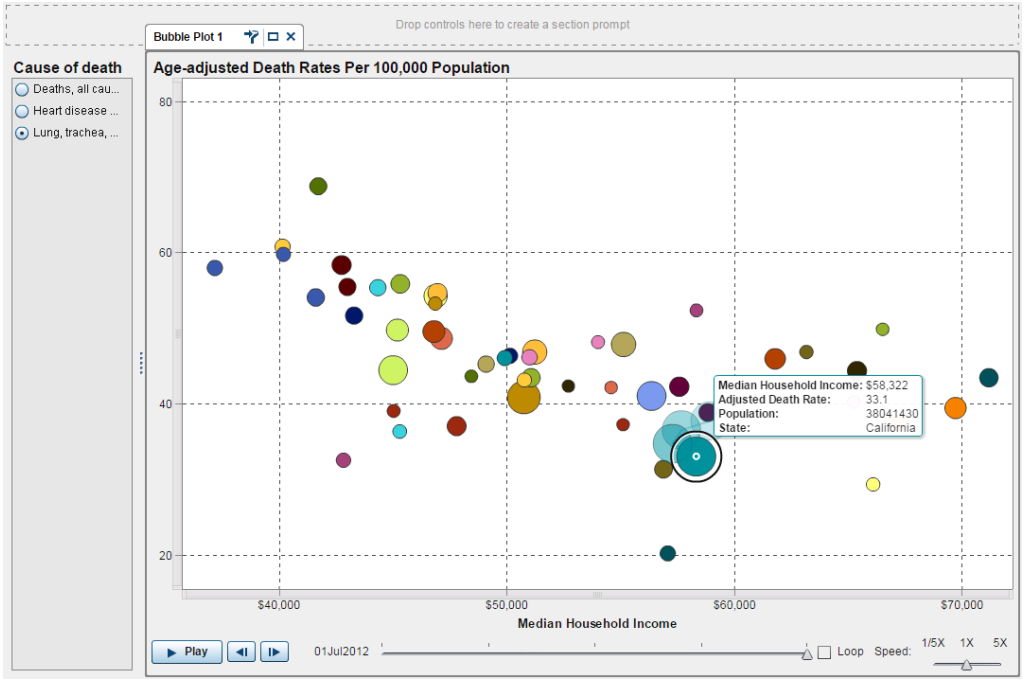


Figure 1 –Health Indicators Warehouse (HIW) data presented in a SAS® Visual Analytics Bubble Plot.

And, because e-cigarettes misuse has become more widespread nationwide, as shown in *Figure 2*, analysis to determine public opinion regarding them is important; this is so that public health surveys, policies and overall surveillance can be better informed.

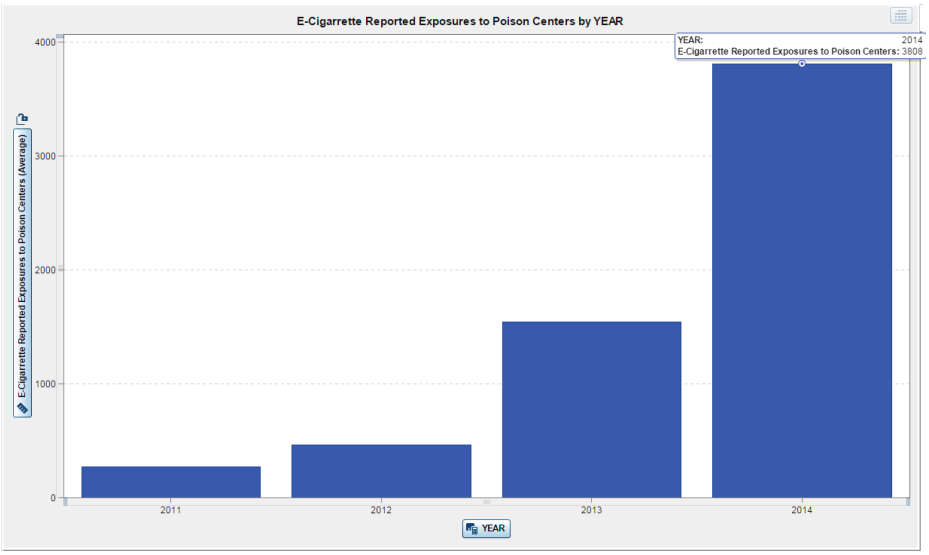


Figure 2 – Poison Center reports of e-cigarettes have experienced a 1305% uptick in calls from 2011 (271) to 2014 (3808).

As e-cigarettes are a relatively new phenomenon, however, data does not yet exist which relates their usage to any long-term effects on public health. As such, this paper does not try to link e-cigarettes use or misuse to chronic diseases, rather we aim to chart public opinion towards the more widespread use of e-cigarettes, particularly as it pertains to the on-going controversy towards the promotion of e-cigarettes as a smoking cessation tool. We are nonetheless able to establish a baseline understanding of chronic diseases by looking at data from the HIW; this adds some context to the social media data collected.

PROBLEM

As mentioned, public health surveillance must include the continuous and systematic collection of data. This, put another way, involves operationalizing analytics. *Figure 1* above is result of integrating SAS® Visual Analytics with the HIW.

To operationalize analytics, further, implies that data for analysis must be available on-demand through tightly coupled or integrated systems. This is because poor integration leads to slow analytic processing that's error prone, costly and time consuming. Integration between disparate systems is therefore key.

The first problem that any federal agency must grapple with, in other words, involves systems that utilize manual ETL (Extract, Transform, and Load) processes leading to poor integration. As shown in *Figure 3*, manual ETL processes are a bottleneck between source data and analytic reports and visualizations.

This introduces requirements risks as well as political risks. Reports, in other words, may not meet end-user specifications, and IT groups invariably will hear constant complaints from irate business end-users who expect more timely and accurate reporting.

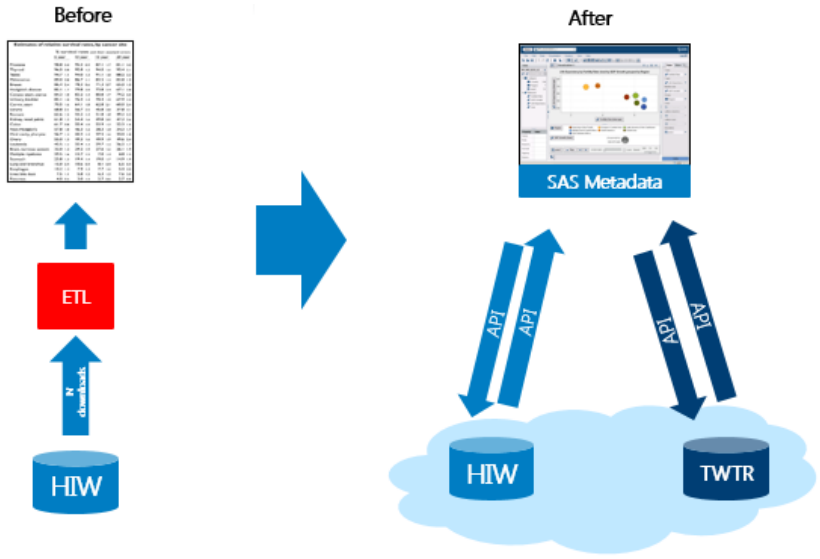


Figure 3 – Gap Analysis. Current state *without* the SAS® Intelligence Platform and SAS® Metadata (Before), and future state data processing for public health surveillance *with* the SAS® Intelligence Platform, SAS® Metadata, and SAS® Visual Analytics (After). The red box denotes a manual process.

Using SAS, particularly the SAS® Intelligence Platform and SAS® Metadata as demonstrated shortly, organizations can "connect" directly to disparate data sources made available via API's. This streamlines operations and enables "pedal to the metal" analytics—in other words, API's permit data feeds directly into SAS®' analytic systems, making processing automated, uniform and integrated.

The second problem occurs when reports and graphs are static and solely tabular because this hamstrings analytics (see *Figure 4*). Tables are useful to look up individual values or when individual values need to be compared. However, if the data is large and the message of the data is contained in "the shape of the values", [4] tables are relatively ineffective and may even thwart effective analysis. Use SAS® Visual Analytics to produce graphs that reveal the complex relationships among multiple values and to make an enormous amount of data easily digestible.

Data presentation, (whether it be through a table, chart or graphs), in other words, must establish a relationship between data values, represent quantities accurately and compare quantities for effective analytics.

Estimates of relative survival rates, by cancer site

	% survival rates and their standard errors			
	5 year	10 year	15 year	20 year
Prostate	98.8 0.4	95.2 0.9	87.1 1.7	81.1 3.0
Thyroid	96.0 0.8	95.8 1.2	94.0 1.6	95.4 2.1
Testis	94.7 1.1	94.0 1.3	91.1 1.8	88.2 2.3
Melanomas	89.0 0.8	86.7 1.1	83.5 1.5	82.8 1.9
Breast	86.4 0.4	78.3 0.6	71.3 0.7	65.0 1.0
Hodgkin's disease	85.1 1.7	79.8 2.0	73.8 2.4	67.1 2.8
Corpus uteri, uterus	84.3 1.0	83.2 1.3	80.8 1.7	79.2 2.0
Urinary, bladder	82.1 1.0	76.2 1.4	70.3 1.9	67.9 2.4
Cervix, uteri	70.5 1.6	64.1 1.8	62.8 2.1	60.0 2.4

Figure 4 – Tables vs Graphs. The table above is meant to highlight the limitations of tables.

SAS® visualizations allow us to not only represent data graphically, but to also interact with those visual representations to change the nature of the display, filter out what's not relevant, drill into lower levels of detail, and highlight subsets of data across multiple graphs simultaneously.

This makes good use of our eyes and assists our brains, resulting in insights that cannot be matched by traditional approaches. Static graphs delivered on paper or electronically on a computer screen help us communicate information in a clear and enlightening way, which is a benefit that should not be undervalued, but it is from SAS® visualizations that businesses will derive the greatest benefits.[4]

In sum, integrated systems with enhanced visualizations enable greater insights and at significantly lower "barriers to entry"; this is especially critical with the mounting pressure of federal health agencies to address public health concerns, such as e-cigarettes. Integration with visualizations that inform insights can ultimately do two things: 1) provide a starting point and underpinnings for a more formalized, long-term study; and, 2) identify the extent of the public health concern, in this case e-cigarettes.

This leads to the precipitating reason for the public health surveillance case study in the paper:

Realizing that formal surveys and data gathering are time-consuming and expensive, an integrated technical solution enabling the continuous, systematic collection and analysis of open data using visualizations is needed to understand public opinion about e-cigarettes, particularly in those U.S. states where there is a high-incidence of chronic diseases associated with smoking. How can this be done?

SOLUTIONS

SAS® technologies can be integrated in many different ways and, ultimately, the approach in this paper produced a hub and spoke architecture where SAS® Text Analytics and other components, such as SAS® Visual Analytics, can integrate. This is shown in *Figure 5*. This architecture can be leveraged, for example, to *write* stored processes in Enterprise Guide that are *run* in Visual Analytics.

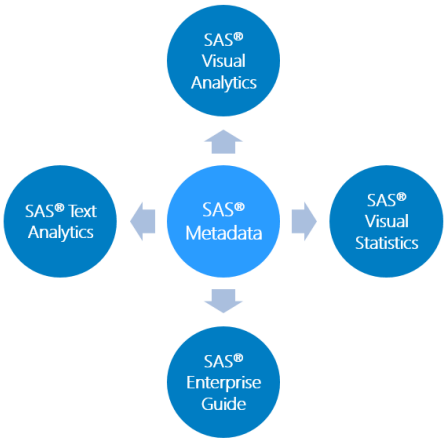


Figure 5 – Hub-and-Spoke Architecture with SAS® Metadata and the SAS® Intelligence Platform.

The hub is the SAS® Metadata server, a centralized resource for storing, managing, and delivering metadata for SAS® applications across an organization, project, or team.[5]

The metadata server enables: 1. Extensibility; 2. Scalability; 3. Centralized Management; and, of course, 4. Integration.

It's also important to note that Metadata can house SAS® code, such as SAS® Stored Processes which will be discussed later, to enable all components in a SAS® ecosystem to interact with one another. SAS® Enterprise Guide, furthermore, is an excellent development to *write* stored processes that can be *run* in SAS® Visual Analytics.

Using the SAS® Enterprise Guide "spoke", put another way, has several advantages for the traditional SAS® programmer--more efficiencies, auto-completion, syntax suggestions, etc. ---to provide a set of comprehensive facilities for SAS® development.[6]

SAS® Metadata is a chief component of the SAS® Intelligence Platform, a comprehensive, end-to-end infrastructure for creating, managing, and distributing analytic visualizations.

For the purposes of this paper, it includes tools and interfaces that enable the following:

- Centrally control the accuracy and consistency of analytic data which is processed by stored processes
- Extract data from a variety of cloud enabled data source made available through APIs, and build data marts and analytic data stores that result from integrating the extracted data
- Give business users at all levels the ability to explore data in a web browser using SAS® Visual Analytics, perform simple and complex query and reporting functions, and view up-to-date results of complex analyses
- Use high-end analytic techniques to provide capabilities such as correlation, clustering, regressions, geomaps, sentiment analysis of text, and categorization of content, as shown in *Figure 6*.

Figure 6 – A more integrated system leads to greater insights and more robust analytics.

In sum, once the components in the architecture are integrated, more complex analysis can be achieved while addressing users' needs: facilitating data collection, generating leads and anomalies, and enabling analytics (particularly for early detection and situational awareness).

With an architecture in place and analytic needs defined, data will be required. As mentioned, API's play an essential role as they facilitate the distribution of open data.

OPEN DATA AND API'S

In this paper, the term "open data" refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users.[7] Open data is distributed by way of API's, a term defined earlier. The rise in API's has been impressive in the last several years, as shown in *Figure 7*, and in 2014, Programmable Web recorded over 10,000 APIs.

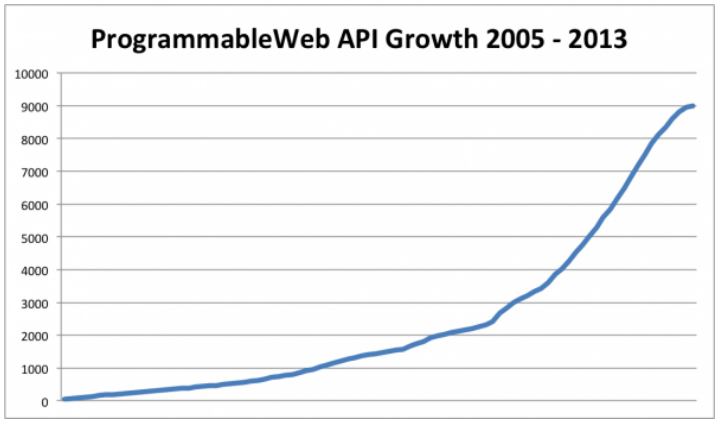


Figure 7 – The Rise of API's

Two data sources available through API's include the Health Indicators Warehouse (HIW) and Twitter. The HIW is the Department of Health and Human Services' (HHS) response to open government efforts to make federal data more accessible to all users. Through it, users can view and download data and metadata for more than 1,200 indicators on health status, outcomes and determinants from more than 180 federal and nonfederal sources. These sources include NCHS data systems, Centers for Disease Control and Prevention (CDC) surveillance data, census data, Medicare and Medicaid administrative data and much more. HIW provides access to data through an API.[8]

Advantages of HIW data:

- It is publicly available, and SAS® Visual Analytics can readily use it through stored processes.
- It includes diverse indicators for preliminary analysis on topics such as chronic diseases.
- It covers both state and county level data.

Disadvantages of HIW:

- It is historical and often annual. That is, the CDC may need closer to real-time measures of public health.
- Typically, the data does not include the current year.
- Some indicators are aggregated at only the national level.

To gain a better understand of the public health situation involving e-cigarettes, Twitter data can be used to understand public opinion and even sentiment regarding new products, such as e-cigarettes, across a variety of users. Because the messages are 140-characters, they are characterized as unstructured data suitable for SAS® Text Analytics.

Advantages of Twitter

- The user base is large. As of December 2014, Twitter has more than 500 million users, out of which more than 284 million are active users.[9]
- Twitter is intended to be used as a way of handling situation awareness. Situation awareness involves being aware of what is happening in one's environment in order to understand how information, events, and one's own actions will impact goals and objectives[10]; this is useful to prioritize epidemiological surveys, for example.
- It can also be used, at least in theory, be used to mobilize public sentiment and mobilize efforts. For this reason, OSH may consider using Twitter data as a means to guide campaigns

Disadvantages of Twitter

- The volume is high
- It's very noisy
- It needs to be enriched for any meaningful analysis.

The solution in this paper combines both HIW and Twitter data made available through APIs that are invoked through SAS® Stored Processes.

STORED PROCESSES, MACROS, AND BATCH PROGRAMS

A stored process is a SAS® program that is stored on a server and defined in metadata, and which can be executed as requested by client applications. For the system analysis in this paper, stored processes were used for building a web application that can access structured (HIW) and unstructured (Twitter) open data via API's.

Stored processes can be written by anyone who is familiar with the SAS® programming language or with the aid of a SAS® code generator such as SAS® Enterprise Guide.

The basic steps to creating a stored process are as follows:

1. Writing the SAS® Stored Process
2. Choosing or Defining a SAS® server to run it on
3. Registering stored process in the SAS® Metadata Server

The ability to store SAS® programs on the metadata server provides security, effective change control management but, more importantly, it produces the hub-spoke-architecture we described earlier.[11]

An optimal way of writing stored processes include the use of macros, or reusable SAS® components that are flexible and modular enough to be applied to a variety of use cases. In the use case of the HIW, several macros interact with one another inside a stored process to form a larger composite application that enables Visual Analytics to stream open data via API's on-demand into its in-memory server (SAS® LASR Analytic Server)—see *Figure 8*.

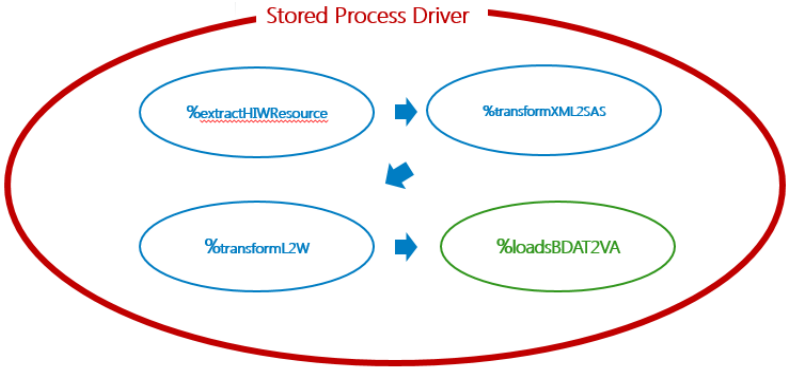


Figure 8 – Stored processes can be composed of several reusable SAS® components or macros that interact with one another in order to form a composite application. This design of a stored process re-uses the macro in green for the HIW use case and Twitter use case.

As demonstrated in *Figure 8*, they include the following:

%extractHIWResource – this macro will extract data from the HIW and produce an XML file as output

%transformXML2SAS – this macro will transform an XML file into a SAS® dataset

%transformL2W – this macro will transform a SAS® dataset long to wide

%loadBDAT2VA – this macro will load a SAS® dataset (BDAT file) into the Visual Analytics LASR (in-memory) server

Note that *Figure 8* shows a design of a stored process solution that would ultimately be implemented as SAS® macro code. *Figure 9* shows an example of a SAS® macro, %extractHIWResource.

```
%let MyHIWREST="http://services.healthindicators.gov/v5/REST.svc?indicator_id=71&num_pages=50000&time=2010&fips=0";
%let MyOutputPath="C:\SAS\Projects\CDC1\data\HIW_output.xml";
%extractHIWResource(RESTQuery=&MyREST, FSOutput=&MyOutPath)
```

Figure 9 – Moving from reusable component design to implementation involves macros. The macro above will make an API call to the HIW to produce an XML file as output with only three lines of code.

By creating a stored process and reusing the macros above, analysts will be able to select a health indicator from the HIW that can be loaded into SAS® Visual Analytics for analysis and visualization. This is also an excellent example of how SAS® code and SAS® Visual Analytics together can enhance an organization's daily workflow, efficiency, and effectiveness.

Macros can also interact with one another inside a scheduled batch program. We collected Twitter data as part of a batch process because the Twitter API places limits on the amount of data that can be pulled from one call. In addition, Twitter data is constantly refreshed with new tweets which we also wanted to capture.

Four additional macros were used in the batch program to implement the methodology in *Figure 10*:

%extractTweetsBySearchQuery – this macro will extract tweets based on the query provided as input to produce a csv file with associated metadata (e.g., user screen name)

%importData – this macro will import a csv into a SAS® BDAT file

%applyCCModel – this macro will apply a SAS® Enterprise Content Categorization model to tweets

%geocodeLocation – this macro will generate latitude and longitude coordinates as variables from the location details in a twitter user's profile

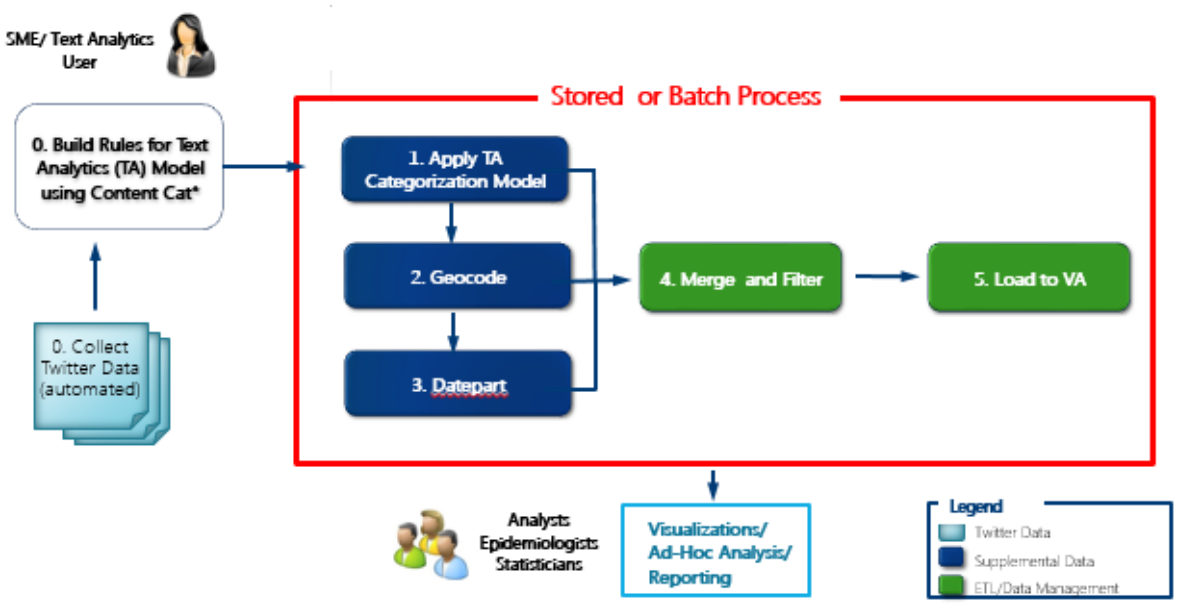


Figure 10 – A Methodology to operationalize analytics with Twitter Data. Note that %loadBDAT2VA macro from the HIW use cases was *reused* for this Twitter use case.

Once Twitter data is collected the data needs to be prepared for analysis and visualization. As part of this process, a Subject Matter Expert (SME) or Text Analytics user can build rules that can be applied to the Twitter data, as shown in *Figure 10*. This text model is applied to the data in order to categorize it (step 1); next the %geocodeLocation macro assigns latitude and longitude variables to the location data from the Twitter data (step 2); then, date fields are converted in a format suitable for analysis (step 3); the results are merged with previous collected tweets and filtered (step 4); and finally, loaded to the SAS® Visual Analytics LASR (in-memory) server using the %loadBDAT2VA macro (step 5). It's worth noting that the %loadBDAT2VA macro was reused for this purpose. Macros are designed to be modular and flexible components that can be reused in a variety of use cases, which leads to operational efficiencies.

Two of these macros are described in detail below.

The macros consist of inputs, outputs, and an interface as shown in the *Figure 11* example. An input is a variable with a value assigned to it. This is possible using the %let statement in SAS®. An interface specifies one or more variables in a parameter list separated by commas. Output refers to an object that is produced (for example, an XML file, a CSV file) or an action that is accomplished by invoking the macro (for example, in the case of the %loadBDAT2LASR macro, registering and lifting a BDAT file from the file system and into the Visual Analytics in-memory LASR server).

```
/*INPUTS*/
/*SAS Metadata source objects:*/
%let MySrcData=hiw_data;
%let MySrcDataPath="C:\SAS\Projects\CDC1\data";
%let MySrcLib=hiw1;
%let MySrcLibFolderPath="/Shared Data/CDC1";
/*SAS Metadata target objects:*/
%let MyTargetLib="Visual Analytics Public LASR";
%let MyTargetLibFolder="/Shared Data/SAS Visual Analytics/Public/LASR";
/*INTERFACE*/
%loadBDAT2LASR(SrcDS=&MySrcData, SrcPath=&MySrcDataPath, SrcLib=&MySrcLib, SrcLibFolderPath=&MySrcLibFolderPath,
               TargetLib=&MyTargetLib, TargetLibFolderPath=&MyTargetLibFolder)
/*OUTPUT. A SAS (sas7bdat) Dataset loaded in the SAS Visual Analytics in-memory LASR server.*/
```

Figure 11 – Inputs, outputs, and an interface are key components of the %loadBDAT2LASR macro. NB: input variables are followed by the %let statement.

The %loadBDAT2LASR macro in *Figure 11* is important because it gives the user an automated, fast and easy way of registering data in Metadata and lifting data into the VA in-memory LASR server in one fell swoop.

This macro and others are provided for download at this site:

https://github.com/ManuelSAS/SGF2015Macros

Documentation is provided with the macros to install and configure them along with the download.

Macros are also easy to use. The %extractTweetsBySearchQuery macro in *Figure 12* requires only two inputs to produce a CSV file containing over 20 variables with tweets related to the input keywords.

```
/*INPUTS*/
/*Twitter query keywords with or without boolean operators:*/
%let MyTwitterKeywords="cdc AND(e-cigarettes OR ecigarettes OR ecigs OR electronic cigarettes OR e-smokes)";
/*Filesystem location of CSV output:*/
%let MyOutputPath="C:\SAS\Projects\CDC1\data\ECigs_Tweets.csv";
/*INTERFACE*/
%extractTweetsBySearchQuery(keyword=&myKeyword, outputPath=&myOutputPath)
/*OUTPUT. A CSV file containing Tweets and associated metadata--e.g., screename, location, etc.*/
```

Figure 12 – Macro are not only reusable but are an easy to use method of acquiring API-related data.

In conclusion, as a result of using macros such as %extractTweetsBySearchQuery, data can be extracted from external remote sources (e.g., HIW and Twitter) for analysis. And, with macros such as %loadBDAT2VA, interaction and integration between the SAS® Metadata Server and the different "spokes" in a hub and spoke architecture is possible –e.g., SAS® Visual Analytics and SAS® Text Analytics.

TEXT ANALYTICS

As part of the data pre-processing for Visual Analytics, a SAS® Enterprise Content Categorization (ECC) model was developed and applied to Twitter data. SAS® EEC provides the means to automatically categorize large volumes of text in a context-dependent manner according to defined linguistic rules, as well as to integrate the information and insight offered by subject matter experts. The large volume of e-cigarette and tobacco-related content which was collected from Twitter for this project necessitated an approach which would allow us to view this content in a clearly defined and organized manner. To this end, we developed an ECC taxonomy which reflected the categories defined by the World Health Organization's MPOWER initiative, comprising the following categories:

- **Advertising**: content related to e-cigarette advertising, branding or marketing, particularly when pertaining to children
- **Cessation**: content related to the promotion of e-cigarettes as 'smoking cessation' devices, a claim currently viewed by some as contentious at best
- **Enforcement**: content related to the efforts to regulate, ban or restrict e-cigarette products
- **Poison**: content related to reports of poisonings caused by e-cigarette liquids
- **Prevention**: content related to preventative measures for smoking or e-cigarette use

An additional ECC category was developed- '**Noise'**, which allowed us to tag tweets which were not deemed useful for this analysis, in particular, those touting e-cigarette paraphernalia for sale.

Following the development of this ECC taxonomy, our corpus of tweets were scored for inclusion in

any of the above categories. Those tweets belonging to the 'Noise' category were removed from the final dataset. This method allowed us to 'tag' tweets as belonging to one or several categories, in a context-specific manner. As seen in *Figure 13*, treemaps utilizing variable hierarchy structures are a useful way of visualizing multiple levels of data.

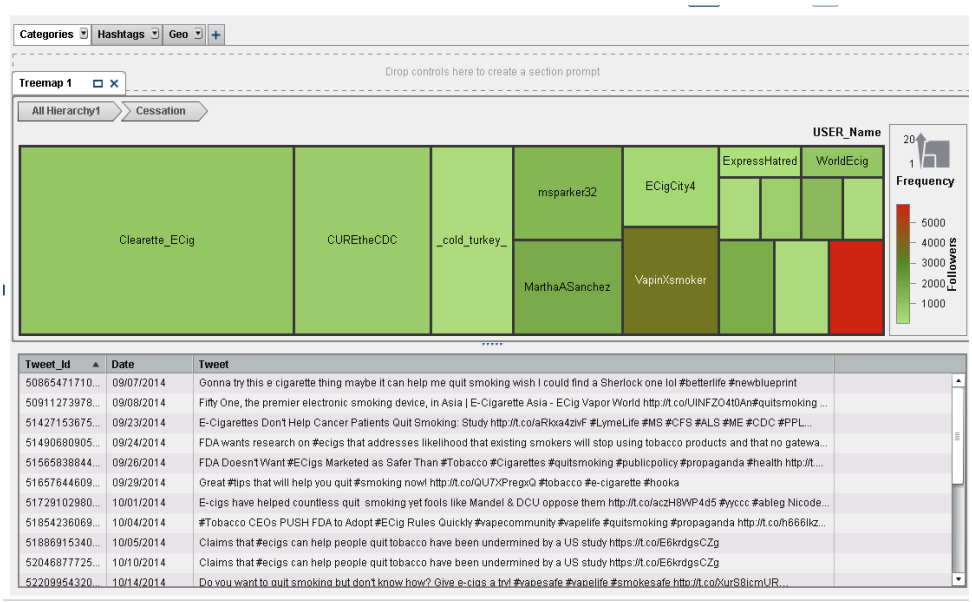


Figure 13 – In this instance, we are looking at the authors associated with tweets coded to the 'Cessation' category, where the size of the tile corresponds to their number of tweets and the color corresponds to their number of followers.

ECC also allows for the extraction of content through the use of regular expressions, classifier strings, grammar rules, and parts-of-speech tagging in concept rules. As Twitter data is often annotated with 'hashtags', in which keywords or phrases are prefixed with the '#' symbol, we applied the regular expression: \#[a-zA-Z1-9]* in a concept rule, which allowed us to extract hashtagged terms from each tweet. These terms allowed us to view individual tweets as they pertained to trending topics or stories, as seen in *Figure 14* in which we rank the top 20 hashtags associated with e-cigarette Twitter data.

Figure 14 –Highlighting the hashtags #Orwellian and #CDC reveals some of the context associated with these keywords. In addition, including a cross tabulation allows the user to associate hashtags with particular smoking categories.

By applying both of these methods to our twitter corpus, we were able to quickly and accurately summarize large volumes of e-cigarette twitter data, without the need to manually review individual tweets. Furthermore, this approach can be combined with the analysis of twitter metadata, such as time-stamps and geo-locations, to give an enhanced view of the public's opinions towards e-cigarettes and tobacco products.

VISUAL ANALYTICS

SAS® Visual Analytics is an easy-to-use, web-based product that leverages SAS® high-performance analytic technologies. SAS® Visual Analytics allows analysts to explore huge volumes of data very quickly to identify patterns and trends for further analysis

Users can even bring in data from the HIW using stored processes that are run in SAS® Visual Analytics as demonstrated earlier in order to explore and mine data quickly and easily.

Because of a highly visual, drag-and-drop interface and speed of an in-memory server (SAS® LASR Analytic Server), analysts can accelerate analytic computations to derive value from large amounts of diverse datasets; this creates an unprecedented ability to solve difficult problems.[12]

To further understand the costs of chronic diseases associated with smoking, for example, HIW data can be streamed into SAS® Visual Analytics for explorations and to see "the forest for the trees".[13]

A correlation matrix is a good place to kick-off exploring data and any two variables that exhibit a strong correlation is a good indication that this is a good place to focus the analysis. *Figure 15* represents this.

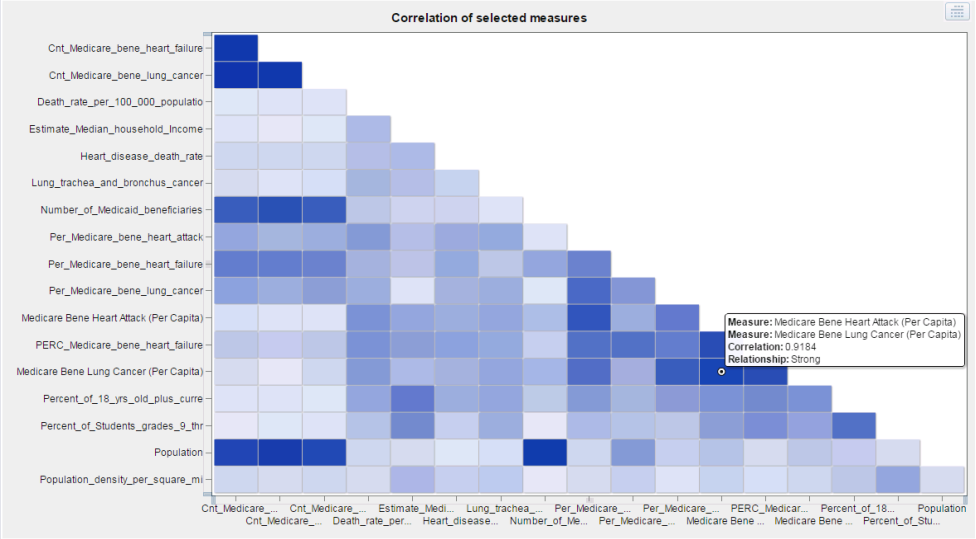


Figure 15 – Items in dark blue indicate a strong correlation, and in this case it is a per capita rate of Medicare Beneficiaries for Heart Attacks and Lung Cancer

To analyze the two variables from the correlation matrix further, a linear regression can be used to look at specific observations and how often (frequency) they fit a regression line as well as to determine the residuals (*Figure 16*). This may indicate candidate variables use in a more formalized linear regression project.

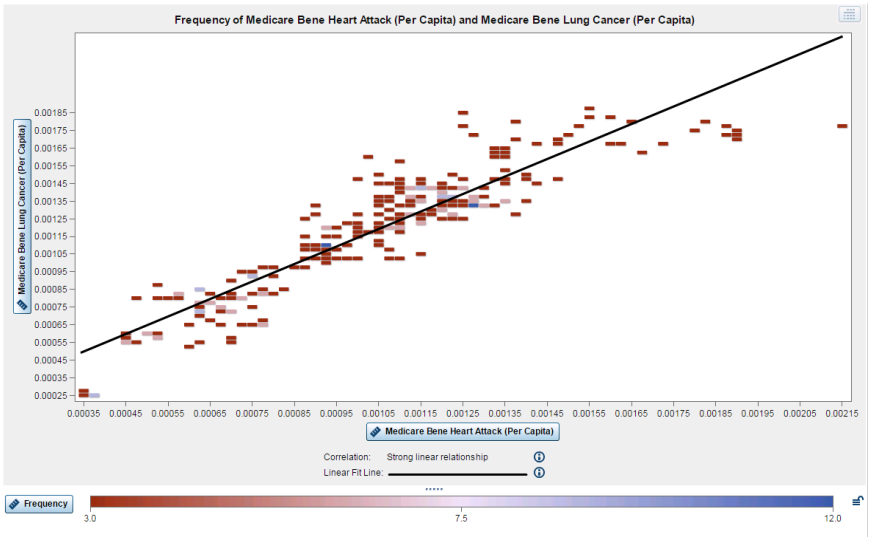


Figure 16 – A fit line can be produced from the correlation matrix to show a linear relationship.

Put another way, if the points on the scatter plot are tightly clustered around the line, then it likely provides a good approximation for the relationship. If not, another fit line should be considered to represent the relationship. If outliers (points which are distant from the rest of the data) are present, moreover, they can have a strong influence on the slope of the line, and those points should be examined more closely.

Whatever the case, the linear relationship in *Figure 16* can lead to several hypothesis that can be tested with additional visualizations, such as a cluster. Clustering is a method of data segmentation that puts

observations into groups that are suggested by the data. The observations in each cluster tend to be similar in some measurable way, and observations in different clusters tend to be dissimilar. Observations are assigned to at most one cluster. From the clustering analysis, you can generate a cluster ID variable to use in other tools.

The two strongly correlated variables in the analysis ("Medicare Beneficiaries for Heart Attacks and Lung Cancer") was supplemented with a third similar variable ("Medicare Beneficiaries for Heart Failure") and evaluated with Estimated Median Household Income to produce the visualization in *Figure 17*. Including additional variables from the correlation exercise and earlier analysis, demonstrate that observations with several Medicare beneficiaries who have chronic diseases linked to smoking also have an inverse relationship to "Estimated Median Household Income"; this is similar to the Bubble Plot shown earlier.

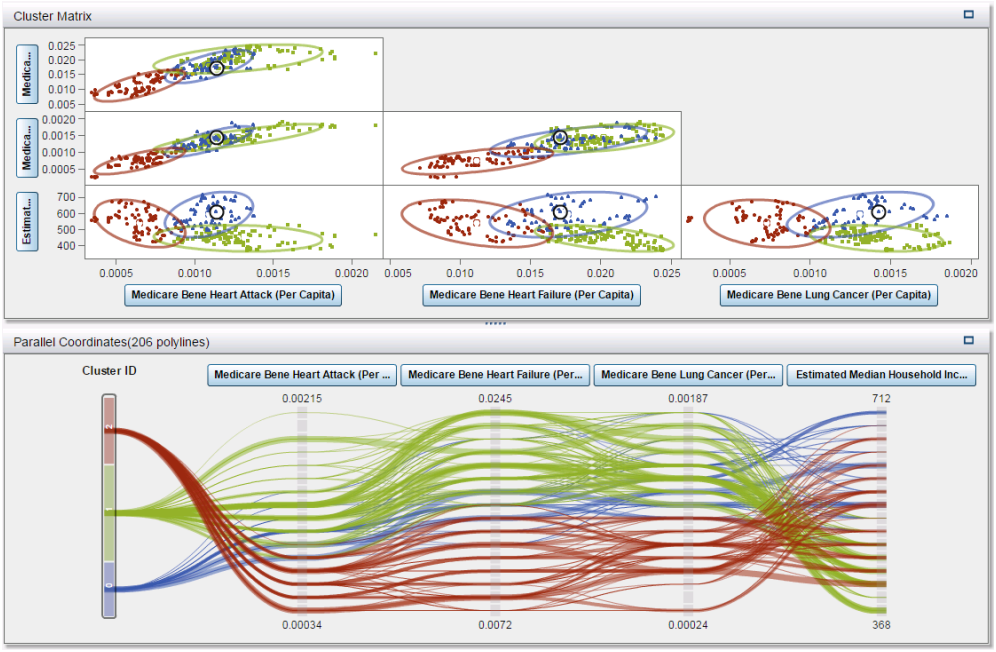


Figure 17 – Using clustering methods, SAS® Visual Analytics assign a "Cluster ID" to observations.

In summary, SAS® Visual Analytics tool provides many analytic capabilities that can be published as reports for widespread distribution as demonstrated in the next section; this moves the data collected from stored processes and API's from the x-ray to the cinema.

RESULTS AND HOW VISUALIZATIONS INFORM INSIGHTS

The first step in providing e-cigarette public health surveillance in this paper involved the "systematic and continuous collection of data"; this required API integration that streamlined the processes associated with acquiring data. This was made possible by stored processes as shown in *Figure 18*.

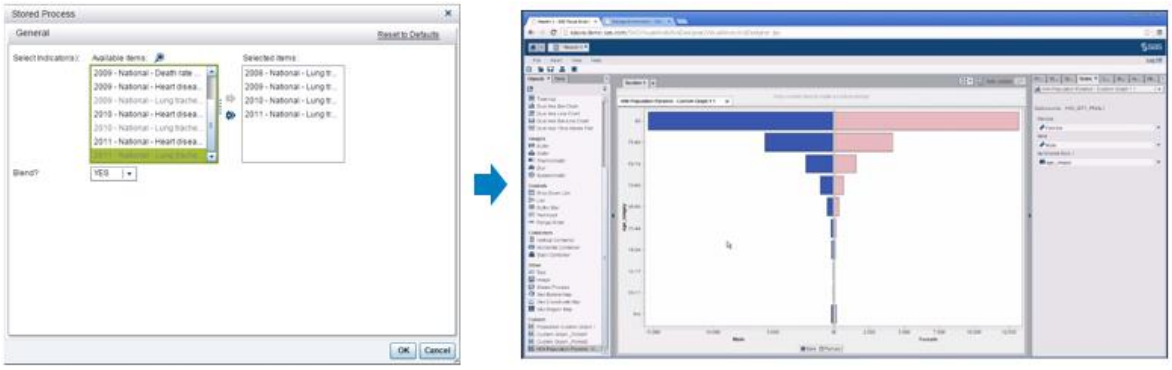
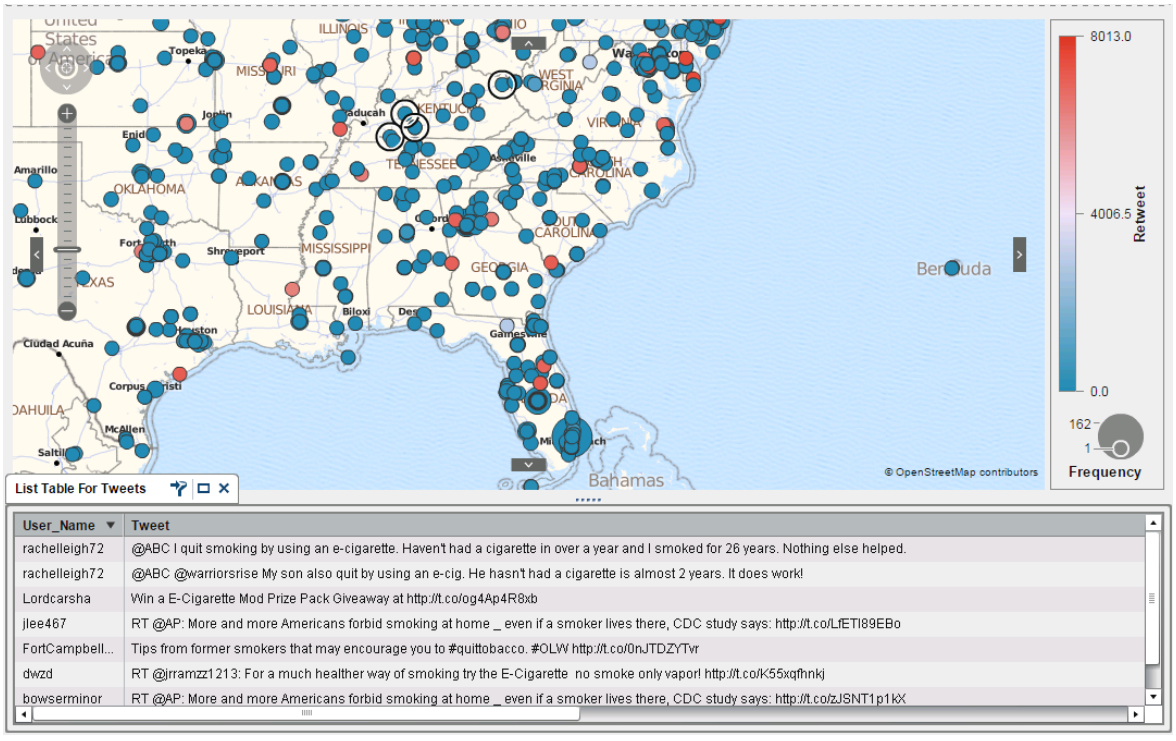


Figure 18 – A stored process (left) can include prompts or selection boxes to choose specific HIW indicators. The selected items can also be "blended" or integrated, as demonstrated by the "Blend?" drop-down box, to form a larger dataset for analysis and report creation—e.g., population pyramid (right).

As shown, stored processes and metadata produce a single integrated system which resulted in several efficiencies. This type of integration can lead to explorations of HIW data that provide context for the public health surveillance of e-cigarettes.

This analysis of structured and unstructured data form the second step for effective public health surveillance—i.e., after data has been collected.

The integration of data and capabilities as discussed up to this point, can lead to an understanding of public opinion about these e-cigarettes as shown in *Figure 19*. The Kentucky tweets below talk about using e-cigarettes to quit smoking and less toxic vapor.



User_Name	Tweet
rachelleigh72	@ABC I quit smoking by using an e-cigarette. Haven't had a cigarette in over a year and I smoked for 26 years. Nothing else helped.
rachelleigh72	@ABC @warriorsrise My son also quit by using an e-cig. He hasn't had a cigarette is almost 2 years. It does work!
Lordcarsha	Win a E-Cigarette Mod Prize Pack Giveaway at http://t.co/og4Ap4R8xb
jlee467	RT @AP: More and more Americans forbid smoking at home _ even if a smoker lives there, CDC study says: http://t.co/LfETI89EBo
FortCampbell...	Tips from former smokers that may encourage you to #quittobacco. #OLW http://t.co/0nJTDZYTvr
dwzd	RT @jrramzz1213: For a much healther way of smoking try the E-Cigarette no smoke only vapor! http://t.co/K55xqfhnkj
bowserminor	RT @AP: More and more Americans forbid smoking at home _ even if a smoker lives there, CDC study says: http://t.co/zJSNT1p1kX

Figure 19: Plotting Twitter user locations against a map will allow users to follow attitudes towards e-cigarettes by region or state.

Additional SAS® capabilities also uses text analytics to display whether the Tweets in a topic express positive, negative, or neutral sentiment as shown in *Figure 20*. With the topic **"cdc, +study, +smoke, +smoker"** selected, the word **"e-cigarettes"** demonstrate many negative sentiment tweets in California. Although the "Relevance" field indicates little relevance between the "e-cigarettes" and the topic, these results provide valuable leads (i.e., User Screename) and context (i.e., State).

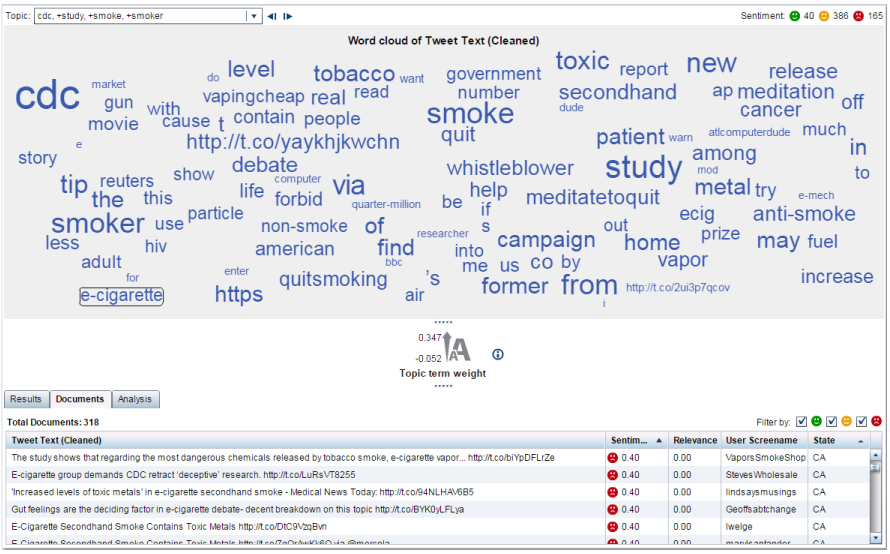


Figure 20: You can use text analytics in a word cloud to identify topics and terms that appear together in a tweet and to analyze the sentiment of the tweets in a topic.

These capabilities form the bedrock for effective public health surveillance that can inform insights and ultimately drive effective decision making.

CONCLUSION

This paper has shown how open data may help with the planning, implementation, and evaluation of a health practice for public health surveillance. In addition, the SAS® Intelligence Platform and Metadata especially can be used to heavily automate, make uniform and operationalize business processes.

Furthermore, some of the value with open data lies not only in providing answers but in raising questions that provide continual guidance towards more formal analysis, surveys, or campaigns.

These conclusions can be summarized as followed:

- API-based data sources and integration are key. Multiple data sources and technologies can be integrated through the SAS® Intelligence Platform, i.e., stored processes in Metadata
- Additional inclusion of SAS® macros and stored processes through SAS® Metadata allows users to customize and enrich data. For example, by including latitude and longitude coordinates with Twitter data for use in geomaps.
- Ongoing analysis of HIW data can be married to social media (for public opinions) in order to identify research opportunities. This is possible through system integration
- Visualizations are a boon towards enabling analysis and making diverse and even large datasets easily digestible.

Moving forward, the approach in this paper can be continued and extended to include more data over time and to possibly be repeated with other use cases involving Twitter and open data.

As a result of our initial foray into this topic, we have identified particular questions pertaining to e-cigarettes that may be of ongoing interest to the CDC:

- **E-cigarette usage**. What is e-cigarette usage by median household income or chronic disease on a state by state basis? To what degree is e-cigarette usage a leading indicator of chronic disease?[14,15]
- **Attitudes**. How can attitudes towards e-cigarettes, as reflected in social media, manifest themselves as regional differences?
- **Causal factors**. What factors are driving the increase in calls to poison centers about e-cigarettes? By state?
- **Individuals and screennames**. Are particular individuals or groups driving attitudes towards e-cigarettes on Twitter? Does this present an opportunity to utilize social network analysis?
- **Disease control**. Can e-cigarettes be viewed as a cessation tool and regulated as such? Or should they be regulated a tobacco product with the attendant health risks? Will rates of chronic diseases increase or decrease over time and can they be linked to e-cigarette usage?
- **Technical.** What other API-based data sources need to be included to enrich any analysis?

This analysis must, in the end, move beyond the use of output indicators associated with the counts or frequency of chronic diseases and any related tweets. Although very useful for the monitoring and evaluation framework described in this paper, ultimately, outcomes or the change that is expected as a result of e-cigarettes usage is of paramount importance. This leads to the final point in this paper that open government values of transparency, collaboration, and participation require more than data and must also include analytics.

REFERENCES

1. World Health Organization - Public Health Surveillance.

http://www.who.int/topics/public_health_surveillance/en/

2. World Health Organization - Noncommunicable diseases.

http://www.who.int/topics/noncommunicable_diseases/en/

3. Centers for Disease Control and Prevention -- Chronic Disease Prevention and Health Promotion

http://www.cdc.gov/chronicdisease/index.htm

4. Few, Stephen: Data Visualization Past, Present, and Future. (2007)

http://www.perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf

5. SAS Products -- SAS® Metadata Server

http://support.sas.com/software/products/metadatasrvr/index.html

6. Ravenna, Andy. SAS® Enterprise Guide® 4.2: Getting to Code You. (2010)

http://support.sas.com/resources/papers/proceedings10/145-2010.pdf

7. Office of Management and Budget -- Open Data Policy

https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

8. Health Indicators Warehouse

http://www.healthindicators.gov

9. Wikipedia. “Twitter”.

http://en.wikipedia.org/wiki/Twitter

10. Wikipedia. “Situation Awareness”.

http://en.wikipedia.org/wiki/Situation_awareness

11. SAS® 9.4 Stored Processes

http://support.sas.com/documentation/cdl/en/stpug/67499/HTML/default/viewer.htm#n180km1hadyuton13gx4bzlqhpwr.htm

12. SAS® Visual Analytics 7.1 -- User’s Guide
http://support.sas.com/documentation/cdl/en/vaug/67500/PDF/default/vaug.pdf

13. Abousalh-Neto, Nascif. The Forest and the Trees: See it All with SAS® Visual Analytics Explorer. (2013)

http://support.sas.com/resources/papers/proceedings13/058-2013.pdf

14. CDC Leading Health Indicators (2010)

http://www.cdc.gov/nchs/data/hpdata2010/hp2010_final_review_leading_health_indicators.pdf

15. CDC Leading Health Indicators – Tobacco (2014)

https://www.healthypeople.gov/sites/default/files/HP2020_LHI_Tobacco_0.pdf

ACKNOWLEDGMENTS

We would like to thank Lily Chen, Ph.D. at the CDC for her assistance in completing these visualizations. We would also like to thank all of our colleagues at the SAS Federal Government Business Unit, a uniquely congenial workplace. In particular we would like to thank Reuben Richards, Joseph Boland, Brian O’Mara, Theresa Do, and Tom Sabo, all of whom provided invaluable feedback.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Manuel Figallo
SAS Federal
1530 Wilson Blvd, Suite 800
Arlington, VA 22201
Manuel.Figallo@sas.com
http://www.linkedin.com/mfigallo

Emily McRae
SAS Federal
1530 Wilson Blvd, Suite 800
Arlington, VA 22201
Emily.McRae@sas.com