Oak Ridge Bio-surveillance Toolkit (ORBiT): Integrating Big-Data Analytics with Visual Analysis for Public Health Dynamics

Arvind Ramanathan, Laura L. Pullum, Chad A. Steed*
Computational Science and Engineering Division, Oak Ridge National Laboratory
Tara L. Parker †
Computer Science and Engineering, Texas Tech University
Shannon P. Quinn, Chakra S. Chennubhotla‡
Department of Computational and Systems Biology, University of Pittsburgh

ABSTRACT

In this position paper, we describe the design and implementation of the Oak Ridge Bio-surveillance Toolkit (ORBiT): a collection of novel statistical and machine learning tools implemented for (1) integrating heterogeneous traditional (e.g. emergency room visits, prescription sales data, etc.) and non-traditional (social media such as Twitter and Instagram) data sources, (2) analyzing large-scale datasets and (3) presenting the results from the analytics as a visual interface for the end-user to interact and provide feedback. We present examples of how ORBiT can be used to summarize extremely large-scale datasets effectively and how user interactions can translate into the data analytics process for bio-surveillance. We also present a strategy to estimate parameters relevant to disease spread models from near real time data feeds and show how these estimates can be integrated with disease spread models for large-scale populations. We conclude with a perspective on how integrating data and visual analytics could lead to better forecasting and prediction of disease spread as well as improved awareness of disease susceptible regions.

Index Terms: J.3 [Life and Medical Sciences]: Biology and genetics—Medical Information Systems;

1 INTRODUCTION

The imminent threats from novel and emerging air-, water- and food-borne diseases that can potentially have devastating social and economic impact on widespread geographic regions within a short period of time underscores the importance for developing effective early-warning/forecasting systems that can enable rapid identification, analysis and detection of these diseases [1]. While traditional indicators of public health related data sources, including the national and international surveillance systems and other data repositories, track and monitor constantly for emerging diseases, there is an emerging need to integrate information from heterogeneous data sources, including novel data streams arising from social media (voluntary information reported by citizens) and from prescription sales data. Citizen surveillance involves monitoring information from diverse, potentially high-volume, noisy data sources including social media and other data sources such as images on sites (e.g. Instagram) to identify emerging bio-threats [15, 5, 16, 26, 4]. The collective intelligence and social power of individuals augmented with information from traditional sources of public health, promises to empower analysts, decision-makers and the general public with actionable insights on emerging bio-threats. However, these non-traditional data-sources need to be filtered to reveal features that are relevant to public health, annotated with trust models and analyzed for gaining insights into emerging disease outbreaks. We hypothesize that analysis of big-data from social media, synergistically aggregated with trusted data-sources (e.g. emergency room visits at hospitals and clinics, prescription sales data, etc.), can provide an improved, effective and reliable early warning and situational awareness to characterize biological events of interest.

Current biosurveillance tools/systems include BioSense 2.0 [2], HealthMap [9], Google Flu Trends [10], EARS (Early Aberration Reporting System) [14], NEDSS (National Electronic Disease Surveillance System) [12], BCON (BioSurveillance Common Operating Network) [1], Indiana Public Health Emergency Surveillance System (PHESS) [11], Linked Animal-Human Health Visual Analytics (LAHVA) [17], pandemic visualization tools [13, 19, 18], ESSENCE [3], and GEIS (Global Emerging Infections Surveillance and Response System) [28]. We would like to note that we have only listed only a few of the many frameworks available for disease surveillance; for a survey of various techniques, readers are referred to a review by Shmueli and Burkom [25]. Most of these tools include some data analytics capabilities including natural language processing (NLP), data aggregation, basic statistical analyses, and time series counts/ratios. The primary visualization capability provided by the tools consists of a map of the area of interest, augmented by icons representing, or color-coded based on, the metric of interest (MOI), e.g., number of cases. Other visualizations include graphs of MOI varying over time, pie charts of MOI by demographic category, and bar charts comparing MOI amongst various entities. A more detailed survey of the different tools for bio-surveillance is presented elsewhere [21]. One of the main challenges within the bio-surveillance community is the inability to integrate information from diverse data sources (including both structured and unstructured data) and analyze vast datasets in a reliable and efficient way to forecast and warn public health officials about emerging epidemics, to improve situational awareness, and to predict the effects of disease spread and intervention strategies in widespread geographic areas.

In this paper, we describe our experience developing a novel and extensible data analytics platform for bio-surveillance, namely the Oak Ridge Bio-surveillance Toolkit (ORBiT). ORBiT is a component based system that integrates information from existing traditional sources such as clinical data including emergency room visits and prescription data from private and public entities, as well as non-traditional social media sources including Twitter and Instagram, environmental data feeds from the Environment Protection Agency (EPA), and climatological/weather related data. Unlike other bio-surveillance systems, where the primary emphasis is to drive information collection and support visualization of possible alerts about emerging diseases, ORBiT is largely focused on developing novel statistical and machine learning tools that can

*e-mail: ramanathana,pullumll, steedca@ornl.gov
†e-mail: tara.parker@ttu.edu
‡e-mail:spq1,chakracs@pitt.edu

provide insights from large-scale heterogeneous datasets. In addition, the machine learning tools (or analytics components) are tightly integrated with visualization tools in a web-based framework to aid the end-users (or analysts) to explore potential links between heterogeneous datasets, detect patterns/correlations across multiple data streams, identify emerging disease outbreaks, forecast emerging epidemics and monitor control strategies. ORBiT is implemented as a component-based plug-and-play toolkit that exploits existing distributed cloud-based analytics frameworks including Hadoop and Mahout.

The paper is organized as follows: in the next section, we provide an overview of ORBiT, in particular describing the various data aggregation and analytic tools. In section 3, we describe our preliminary results from ORBiT to extract, analyze and visualize information from a large-scale social media corpus to identify potentially interesting temporal patterns of communication. Finally, we conclude with a perspective on developing data analytics tools for bio-surveillance specifically targeting emerging air-, water- and food-borne infectious diseases.

2 ORBiT: Implementation

As illustrated in Figure 1, ORBiT is implemented as a distributed analytic platform: it consists of a software stack atop of Hadoop and makes use of Titan, a distributed graph database as a backend for data storage. The data from each of the traditional and non-traditional sources are hosted as a massive linked structure, with extensible interfaces provided for each data stream. The data from the linked structure is interfaced with streaming and graph-data analytic modules. The outputs from the analytic modules are interfaced with visualization tools that enable analysts to detect spatial and temporal patterns/correlations across multiple data sources. In the subsequent sections, we present an overview of the algorithms and tools that are included as part of ORBiT.

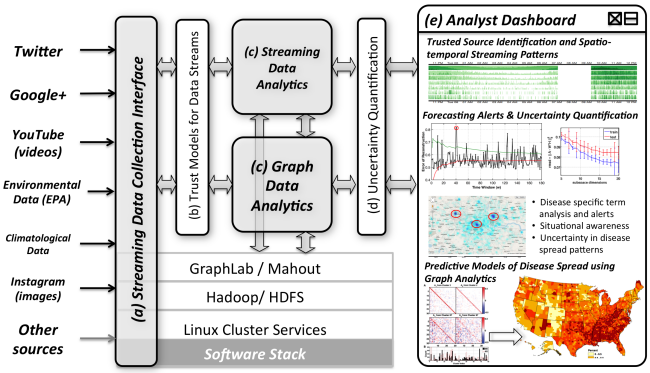

Figure 1: Architectural Overview of Oak Ridge Bio-surveillance Toolkit showing the various components and interfaces in ORBiT as well as the visualization interface.

2.1 Data Collection Interface

The data collection interface represents a collection of tools to handle multiple diverse/disparate, potentially high volume data streams including: (a) social media sites such as Twitter; (b) climatological data; (c) traditional structured data records of emergency room visits and prescription sales that include data regarding physician issued prescriptions for patients, and (d) extensions to accommodate other non-traditional multimedia data such as images from Instagram. Further, the collection interface can interact with existing reporting tools for bio-surveillance such as HealthMap and/or Google Flu Trends. Additionally, with minimal extensions to the framework, it is possible to integrate data from other data streams.

A secondary aspect of the data collection interface is the ability to stream the data to the Titan distributed graph database [24] to enable efficient storage and retrieval of large-scale datasets. The linked representation of the data provides additional services to query and search the data, while returning them as objects for further analysis. Titan provides a distributed environment to store the datasets; in addition, the graph structure provides an intuitive means to connect disparate data sources, in spite of not seeing 'obvious' connections between them. For example, when a patient reports to the emergency room with respiratory stress, the physician may also prescribe several medicines that target his/her symptoms. However, the information is spread across two separate sources, with the emergency room visit being one source and the prescription data being the other. By linking the information based on perhaps the patient (without violating any privacy rules/regulations), it becomes apparent how the data from the two sources are linked inherently.

2.2 Streaming and Graph Data Analytics Components

The central core of the analytic components consists of a powerful NLP (natural language processing) toolkit that can effectively build a statistically relevant vocabulary or bag-of-words model to process text-related data-streams such as Twitter [6]. The NLP tools build statistically principled models of disease associated terms from existing ontologies (e.g., BioCaster), PubMed literature and other textual data-sources (see Figure 2A). Additionally, extensions to the framework will also to accommodate analyst specified terms (from the user interface) for filtering these data-streams with the NLP toolkit. Similar annotation capabilities are also built for image data.

Once the data-streams are filtered using NLP, we used higher-order statistical tools to track/tag events of interest using multi-scale temporal windows (hours, days, weeks, months, years) that can be typically specified by the analyst/end-user [22]. Statistical feature-sets extracted from the filtered data allow one to quickly identify a baseline and tag events as outliers from these baselines. In order to track correlations across multiple data-streams and make predictions, we include several linear, non-linear and hybrid statistical inference tools that achieve good performance in terms of an applied loss function within ORBiT [8, 7].

For detecting spatio-temporal patterns from geo-tagged data-streams, we used tensor analysis to characterize emerging correlated behaviors in other domains [23]. As illustrated in Figure 2B, terms extracted from the symptom lists for various infectious diseases are first tagged for geo-locations and conveniently captured as a matrix. The temporal evolution of this information is tracked as a three-dimensional tensor. Similar tensor representations are also captured for other data sources including prescription data. Using tensor analysis, we capture a small subspace in the potentially high-dimensional space to identify which geographic regions show correlated behaviors in terms of patterns observed from the data. Additionally, tensor analysis tools can signal the time-points that show anomalous behavior observed, implying a potentially interesting event in the data streams being tracked. The machine learning tools themselves are available as individual components and the analyst can select/combine/define data-workflows that allow her/him to customize the analytic outputs.

2.3 Analyst Dashboard: Visual Analytic Interface for Bio-surveillance

The analysis modules closely interface with the visual front-end, which consists of a front-end that allows the analysts (or end-users) to interact with and provide feedback to the data analytics components in the toolkit. Visual analytics has recently emerged as an effective means to integrate information from diverse data sources and develop hypotheses about emergent behaviors in the data [21]. The front-end allows the end-user to: visualize data-streams, identify potentially interesting leads (from the different data sources) and tag them, visualize anomalous behaviors, and visualize spatio-temporal correlations across multiple data-streams. Similar to pre-

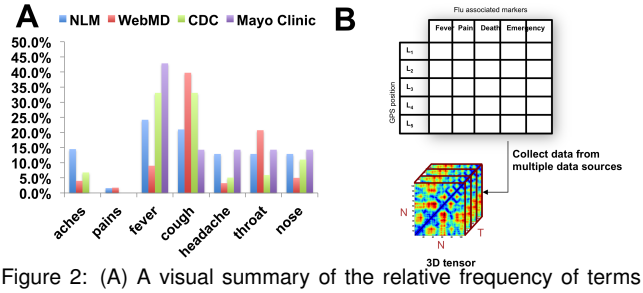

Figure 2: (A) A visual summary of the relative frequency of terms associated with the flu as catalogued from the National Library of Medicine (NLM), the website WebMD (http://www.webmd.com), Centers for Disease Control (CDC) and Mayo Clinic. Note the similar frequency of terms associated with the flu occur consistently, in spite of compiling the information from different websites. (B) A visual representation of the tensor data, where information from different locations associated with the flu are tracked over time.

vious work [27], the visual interfaces transform the outputs from the analysis modules into graphs- and geographic information system (GIS)-based outputs. These outputs can then be navigated by users to gain insights into bio-surveillance related questions.

For example, from social media feeds, we extract terms related to bio-surveillance, such as 'flu' or 'sore-throat' and summarize the frequency information of these terms as a visual graph where users can look for sudden spikes in the occurrences of these terms. Further, when terms such as the 'flu' and 'sore-throat' occur in the same stream of data, we can then examine that information as a graph, with the nodes representing the disease-related terms and the edges between the nodes representing their relative weights (i.e., how often the words co-occur within the same stream). This information can enable users to gather insights regarding time-evolving behaviors within complex data streams. In addition, users can also visualize related multimedia content, based on hashtags related to different disease-relevant topics to quickly correlate information across multiple data-streams. Finally, for data-streams where geographic locations are tagged, one can also visualize the information as a map [27] to enable end-users to quickly spot regions where unusual behaviors are observed.

For structured data records, such as prescription record data, ORBiT includes geo-location information and visual representation of time-evolving data streams such as number of prescriptions for specific disease/symptoms observed, number of patients with a particle disease/symptom and other epidemiological metrics commonly used by public health officials.

3 ORBiT: Applications

In this section, we present three examples of applying ORBiT (1) to visualize the time-evolution of co-referencing hashtags from a large Twitter corpus, (2) to cluster large-scale datasets automatically to identify emerging patterns of topics and (3) to integrate epidemiological models with ground-level laboratory test observations to project the total number of infections within different geographic regions. Although the applications are not specific to bio-surveillance, they demonstrate the utility of ORBiT in sifting potentially large-scale datasets and quickly summarizing for end-users tasks relevant to bio-surveillance.

3.1 Analyzing Social Media Data for Detecting Time-evolving Patterns

In this section, we present an overview of how we track the time-dependent changes in how users tag different topics of interest. A typical scenario of the same is illustrated in Figure 3A, where several tweets correspond to people experiencing the flu, as evidenced by the number of hashtags that correspond to '#fever'. However, some tweets (shown as a red rectangle in Figure 3A) can correspond to entertainment topics, such as '#beiber'. It is important to note that using only '#fever' can result in a large number of hashtags that do not have relevance to the flu. Therefore, it is important to also consider hashtags that co-occur with '#fever' (and hence are referred to as co-referenced hashtags). In order to track hashtags relevant to the flu, we observe from Figure 2A that there is a high probability of observing terms such as '#sorethroat', '#sick' or '#headache' along with '#fever'. Using these co-occurring terms as a filter, we can then summarize the co-occurring terms for the flu as shown in Figure 3B. The co-occurring terms are centered around '#fever' with other symptoms connected using edges that have been weighted using the relative occurrences of other hashtags, including '#sorethroat', '#sick' or '#headache'. In the next sub-section, we present an overview of how the co-occurrence of hashtags can be utilized to cluster the data and visualize different topics that emerge from the analysis.

3.2 Visual Topic Modeling

To demonstrate the efficacy of co-occurring hashtags for identifying events, we used a dataset spanning the four-day period from April 14 through April 17, 2013. This time period covered the Boston Marathon and the bombing tragedy. We sampled from the Twitter public stream. This results in capturing roughly 1% of all public tweets. For the sake of simplicity, we discarded all tweets whose language flag was set to something other than *en*, or English. This resulted in the retention of 6,482,226 tweets in a 96-hour period. From these tweets, we extracted 1,742,540 hashtags including repeats. Next, we examined all hashtags that co-occurred with other hashtags; that is, we identified those which appeared in the same tweet. From these co-occurrences, we constructed a graph of all hashtags, summarized in Figure 4A, in which an edge was defined between hashtags if they appeared together in a tweet. We discarded all hashtags that did not co-occur with any other hashtags, resulting in a final unique hashtag count of 182,580. Using the graph defined by co-occurring hashtags, we performed spectral clustering [7] to identify regions of tightly-coupled co-occurring hashtags. One of these clusters is shown in Figure 4B, where '#prayforboston' and '#boston' appear extremely often. The presence of other hashtags unrelated to the events in Boston can be attributed to the frequencies more common hashtags that pervade Twitter on a daily basis, such as '#ff', '#music', and '#rt'. These very high-frequency hashtags (Figure 4B) can essentially be regarded as "stop words" and discarded from consideration. This clustering technique is already effective enough to capture overarching topics; further filtering will improve its accuracy.

3.3 Integrating Epidemiological Models with Ground-level Observations

In order to further illustrate the utility of integrating heterogeneous datasets for public health dynamics within ORBiT, we present a strategy for modeling the spread of influenza within a specific geographic region based on data compiled from multiple data-sources including laboratory confirmed H1N1 diagnosis (based on the 2009 H1N1 outbreak in the state of TX) and other data derived from specific prescription records for patients. The prescription records are first processed to include location specific details and identifying patients that exhibit symptoms similar to the flu (based on the ICD9 codes 487.[xx] or 488.[xx]) and who have been prescribed drugs specific to the flu (based on the National Drug Codes (NDC) such as 49281-0392-15, 49281-0707-55, etc.). The challenge in analyzing the datasets is that both datasets include anonymized data (i.e., patient information is protected); however, we need to link the prescription records with laboratory confirmed tests to identify a portion of the population that is infected with the H1N1 flu.

For the laboratory diagnostics data obtained from the Texas State Department of Health Services, we observed that out of the 127 zip

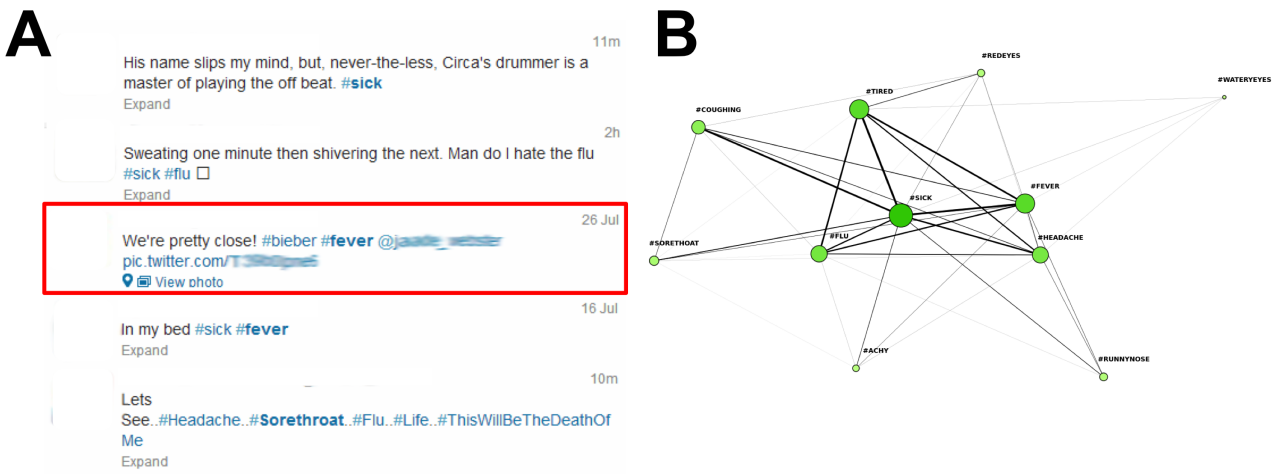

Figure 3: (A) Examples of tweets that contain #sick, #sorethroat, #flu or #fever relating to the flu symptoms versus tweet that is not related to the symptoms of flu (highlighted in a red rectangle). (B) Co-referenced hashtags summarized as a graph for the flu. The size of the nodes indicate how often the hashtags occur where as the size of the edges indicate how often the hashtags occur together in the tweets. Note that the weight on the edges may change as time progresses.

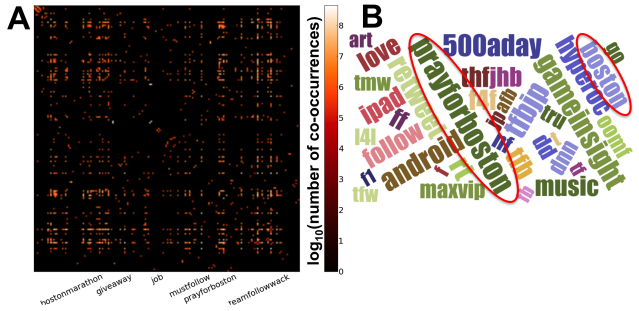

Figure 4: (A) An adjacency matrix representation of the graph of co-occurring hashtags extracted from Twitter during the Boston Marathon bombing tragedy. The adjacency matrix reveals an inherent structure in the co-occurrence of hashtags which can be discovered using spectral clustering techniques. (B) A word-cloud representation of the top cluster shows the presence of #prayforboston and #boston tags along with other tags that also occur on a common basis. These visual representations can guide the end-user to further filter the data and examine more relevant tweets.

codes for which test data was available, only 5 zip-codes exhibited statistically significant numbers of H1N1 infections. Hence, we restricted our analysis to only these locations even in the prescription record dataset. For the prescription record dataset, we used a surrogate measure of the number of patients with confirmed ICD9 codes for the H1N1 flu (487.[xx] or 488.[xx]) and compared this with the laboratory confirmed tests obtained from the Texas State Department of Health Services. The comparison enabled us to obtain a base estimate for the number of infected individuals with H1N1. We then tracked the time-evolution of the number of patients that presented symptoms for the H1N1 flu from the prescription record dataset and used it to estimate the rate of infection within the population. We used the census data to estimate the total number of susceptible people in the population.

The initial number of infected people and the rate of infection are used as input to the simulation framework based on our Susceptible (S), Infected (I), Deceased (D) and Recovered (R) model. The model is initialized with the parameters estimated from the real data (for the number of S, I and infection rate) as well as parameters calibrated from historical datasets [20]. Based on the inputs, the epidemiological model is then run beginning with day 0, which corresponds to the onset of the epidemic. We then obtain a projection of the time-course of the H1N1 based on the information within the sub-regions of interest (i.e., the 5 zip-codes within TX that exhibited significant H1N1 infections). The outputs are then streamed to the geographic location and augmented with an interactive visual aid to allow the end-user to gain insights into the disease spread process.

3.4 User Interface for ORBiT

A visual representation of the total number of influenza reports from the 2009 prescription data is summarized in Figure 5. ORBiT allows customizing the view for different users: Figure 5A shows the perspective from the viewpoint of the public health researcher/analyst and Figure 5B shows the user interface from the perspective of a layman user. Both users have different usage scenarios: in particular, public health researchers/analysts require access to data that provide a fine resolution of the data, including integration with epidemiological models (as discussed above). On the other hand, mobile clients such as regular users can have access to summary information about occurrences (of influenza or other infectious diseases) along with public warnings issued in and around their immediate vicinity/neighborhood (purple dot indicates where the user is located currently).

4 Conclusion

In this paper, we have described ORBiT, which emphasizes our novel statistical and machine learning tools to analyze potentially large datasets and provide a visual analytics front-end for biosurveillance related tasks. Users can (1) interact with the system to potentially integrate heterogeneous datasets, (2) visualize time-evolving patterns from large-scale datasets and (3) integrate epidemiological simulations by estimating their parameters directly from the data itself. In addition to these capabilities, ORBiT also includes mechanisms to integrate with existing systems to incorporate additional sources of data. The visual tools developed within the framework will enable users to provide feedback into the machine learning tools themselves so that the results and the workflows can be defined by the end-users or analysts to customize their tasks of interest. Future development on ORBiT will include an application programming interface (API) that can allow developers to contribute machine learning tools into the framework, while customizing them to the needs of the bio-surveillance community. In addition, streaming analysis tools will be integrated into the framework to facilitate near real-time analysis on very large-scale datasets. We

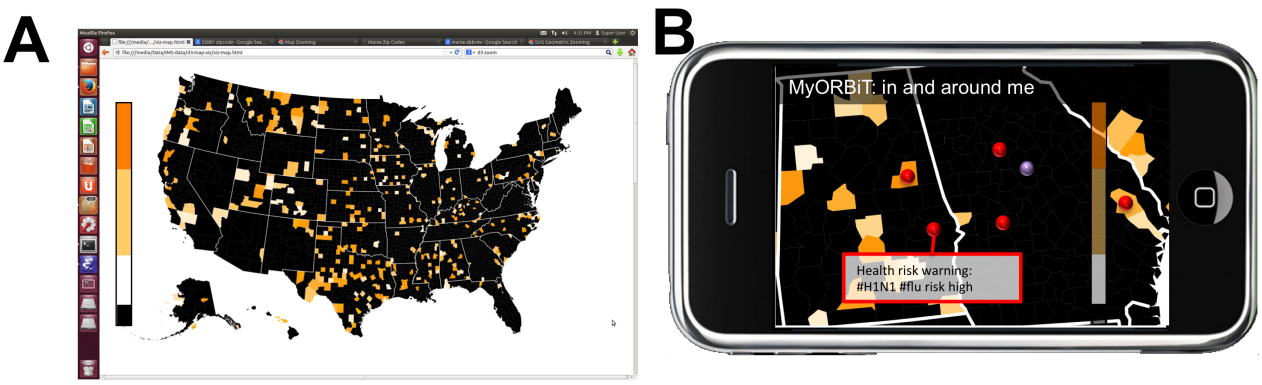

Figure 5: (A) User interface from ORBiT for public health researcher/analytst. (B) User interface from ORBiT for mobile devices.

hope that the availability of this platform will facilitate better forecasting and prediction capabilities for disease surveillance and enable real-time situational awareness that can guide public health officials to respond effectively to emerging bio-threats/diseases.

Acknowledgements

The authors wish to thank Ronita Adams (Tennessee State University), Vivek Datla (University of Memphis) and Eric Mitchell (Pelissippi State Community College) for participating in the design and implementation of the ORBiT framework as part of their summer internship program. The authors also wish to thank Dr. Grace Kuban from the Texas State Department of Health Services for providing us access to the laboratory test data for H1N1 (year 2009). ORNL is operated by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] Efforts to develop a national biosurveillance capability need a national strategy and a designated leader. GAO-10-645, Jun 2010.

[2] C. Bradley, H. Rolka, D. Walker, and J. Loonsk. Biosense: Implementation of a national early event detection and situational awareness system. *Morb Mor Wkly Rep*, 54 (Suppl):11–19, 2005.

[3] K. Brown, J. Pavlin, J. Mansfield, E. Elbert, V. Foster, and P. Kelley. Identification and investigation of disease outbreaks by essence. *J Urban Health*, 80(1):i119–i119, 2003.

[4] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection — harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009. PMID: 19423867.

[5] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*, 5(5):e1206, 05 2011.

[6] W. Chapman. *Natural language processing biosurveillance*. Handbook of Biosurveillance. Elsevier Inc., 2005.

[7] C. Chennubhotla and A. Jepson. Eigencuts: Half-lives of eigenflows for spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 689–696, 2003.

[8] C. S. Chennubhotla and A. Jepson. Sparse-pca: Extracting multi-scale structure from data. In *International Conference on Computer Vision*, pages 641–647, 2001.

[9] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, 2008.

[10] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 02 2009.

[11] S. Grannis, M. Wade, J. Gibson, and J. Overhange. The indiana public health emergency surveillance system: Ongoing progress, early findings, and future directions. In *AMIA Annual Symposium Proceedings 2006*, pages 304–308, 2006.

[12] N. W. Group. National electronic disease surveillance system (nedss): a standards-based approach to connect public health and clinical medicine. *J Public Health Manag Pract*, 7(6):43–50, 2011.

[13] R. Hafen, D. Anderson, W. Cleveland, R. Maciejewski, D. Ebert, A. Abusalah, M. Yakout, M. Ouzzani, and S. Grannis. Syndromic surveillance: Stl for modeling, visualizing and monitoring disease counts. *BMC Med Inform and Decision Making*, 9:21–32, 2009.

[14] L. Hutwagner, W. Thompson, G. Seeman, and T. Treadwell. The bioterrorism preparedness and response early aberration reporting system (ears). *Journal of Urban Health*, 80(1):i89–i96, 2003.

[15] M. Kamel Boulos, B. Resch, D. Crowley, J. Breslin, G. Sohn, R. Burtner, W. Pike, E. Jezierski, and K.-Y. Chuang. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, ogc standards and application examples. *International Journal of Health Geographics*, 10(1):67, 2011.

[16] J. Lombardo and D. Buckeridge, editors. *Disease Surveillance: A Public Health Informatics Approach*. John Wiley and Sons, 2006.

[17] R. Maciejewski, Y. Jang, D. Ebert, W. Cleveland, M. Ouzzani, S. Grannis, and L. Glickman. Lahva: Linked animal-human health visual analytics. *Advances in Disease Surveillance*, (4–11), 2007.

[18] R. Maciejewski, P. Livengood, S. Rudolph, T. Collins, D. Ebert, R. Brigantic, C. Corley, G. A. Muller, and S. Sanders. A pandemic influenza modeling and visualization tool. *J Vis Lang and Comput*, 22:268–278, 2011.

[19] J. Malone, R. Brigantic, G. A. Muller, A. Gadgil, W. Delp, B. McMahon, R. Lee, J. Kulesz, and F. Mihelic. U.s. airport entry screening in response to pandemic influenza: modeling and analysis. *Travel Med Infect Dis*, 7(4):181–191, 2009.

[20] O. Ozmen, J. Nutaro, A. Ramanathan, and L. Pullum. Comparing sir models of epidemiology. *BMC Infect Dis*, (submitted), 2013.

[21] A. Ramanathan, L. Pullum, C. Steed, S. Quinn, C. Chennubhotla, and T. Parker. Integrating heterogeneous healthcare datasets and visual analytics for disease bio-surveillance and dynamics. In *3rd IEEE Workshop on Visual Text Analytics*, 2013.

[22] A. Ramanathan, A. J. Savol, P. K. Agarwal, and C. S. Chennubhotla. Event detection and sub-state discovery from biomolecular simulations using higher-order statistics: Application to enzyme adenylate kinase. *Proteins: Structure, Function, and Bioinformatics*, 80(11):2536–2551, 2012.

[23] A. Ramanathan, J. O. Yoo, and C. J. Langmead. On-the-fly identification of conformational substates from molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 7(3):778–789, 2011.

[24] M. Rodriguez and J. Shinavier. Exposing multi-relational networks to single-relational network analysis algorithms. *J Infomet*, 4(1):29–41, 2009.

[25] G. Shmueli and H. Burkom. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, 52(1):39–51, 2010.

[26] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5):e19467, 05 2011.

[27] C. Steed, T. Potok, R. Patton, J. Goodall, C. Maness, and J. Senter. Interactive visual analysis of high throughput text streams. In *2nd Interactive Visual Text Analytics Workshop, Seattle, WA*, 2012.

[28] C. Witt, A. Richards, P. Masuoka, D. Foley, A. Buczak, L. Musila, J. Richardson, M. Colacicco-Mayhugh, L. Rueda, T. Klein, A. Anyamba, J. Small, J. Pavlin, M. Fukuda, J. Gaydos, K. Russell, and the AFHSC-GEIS Predictive Surveillance Writing Group. The afhsc-division of geis operations predictive surveillance program: a multidisciplinary approach for the early detection and response to disease outbreaks. *BMC Public Health*, 11(Suppl 2):S10, 2011.