Integrating Heterogeneous Healthcare Datasets and Visual Analytics for Disease Bio-surveillance and Dynamics

Arvind Ramanathan, Laura L. Pullum, Chad A. Steed
Computational Science and Engineering Division, Oak Ridge National Lab
{ramanathana, pullumll, steedca}@ornl.gov

Shannon S. Quinn, Chakra S. Chennubhotla
Department of Computational and Systems Biology
University of Pittsburgh
{spq1, chakracs}@pitt.edu

Tara Parker
Computer Science Department
Texas Tech University
tara.parker@ttu.edu

ABSTRACT
In this paper, we present an overview of the big data challenges in disease bio-surveillance and then discuss the use of visual analytics for integrating data and turning it into knowledge. We will explore two integration scenarios: (1) combining text and multimedia sources to improve situational awareness and (2) enhancing disease spread model data with real-time bio-surveillance data. Together, the proposed integration methodologies can improve awareness about when, where and how emerging diseases can affect wide geographic regions.

Author Keywords
Heterogeneous datasets; Big data; Bio-surveillance; Visual analytics; Public health dynamics

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION
According to the World Health Organization (WHO), public health surveillance is the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice. Such surveillance can: (a) serve as an early warning system for impending public health emergencies; (b) document the impact of an intervention, or track progress towards specified goals; and (c) monitor and clarify the epidemiology of health problems, to allow priorities to be set and to inform public health policy and strategies. Public health surveillance can include both communicable (e.g., syphilis, HIV, etc.) and non-communicable (e.g., cancer) diseases; however, infectious and communicable diseases receive the most attention within the surveillance community, because of their potential to affect a large proportion of the population within a short time-period [23]. Several events in the recent past, including the 2001 anthrax attack on the United States mail system (which signified an intentional release of anthrax bio-agents) [30], the 2009 influenza pandemic [17], and the recent Middle East Respiratory Syndrome events [19] underscore the importance of developing effective public health surveillance systems.

In this paper, we discuss big data challenges in public health surveillance, focusing on data visualization, integration and analytics. We present a survey of existing bio-surveillance systems and highlight the role of data integration from heterogeneous datasets including text, images and multimedia as well as visual analytics in gathering insights and guiding decision makers with possible intervention strategies in case of a disease outbreak. In addition, we highlight the outstanding challenges in public health surveillance and summarize how novel statistical techniques can aid data integration and analysis of diverse datasets. Finally, we conclude with a summary of use cases for visual analytics and its role in providing valuable insights for public health dynamics.

Public Health Surveillance as a Big Data Problem
Both traditional and non-traditional information sources are contributing to disease surveillance data for public health. Traditional information, such as health records (from, e.g., clinics or health care providers and hospitals) are increasingly delivered electronically. Other sources for traditional syndromic surveillance data include the Centers for Disease Control and Prevention (CDC), ambulance and emergency medical records, and laboratory records. Databases and sources of aggregate health information are available for query. The CDC releases data on reportable infectious diseases, which it collates given reports from state health agencies. WHO and CDC produce reports whose data can be used for baselining and tracking infectious diseases and associated environmental conditions. In addition to these traditional data sources, huge amounts of novel data are being generated that can make important contributions to public health and disease surveillance. [16, 4]

Disease surveillance can also be informed by indirect or non-traditional means such as understanding sales data (e.g., from over the counter medications, prescriptions), attendance records (school or work), news feeds, and other disease, economic, agricultural, and environmental monitoring appli-

CHI'13, Oct–13, 2013, Atlanta, Georgia, USA.

Source	Volume	Variety	Velocity	Value
Electronic Health Records (EHR)	O(PB[1])	Structured and unstructured text and multimedia	High throughput + High latency	Highly specific, description of patients, symptoms and treatments
Emergency call (911) and Poison control centers	O(TB[2])	Structured and unstructured multimedia (voice) and text	High throughput during epidemics	Specific descriptions + geo-tagged data; may have false positives
Prescription datasets	O(TB)	Structured and unstructured text data	Variable	Specific symptoms; drug prescriptions; geo-tagged data
Emergency Medical Services (EMS) data	O(TB)	Structured and unstructured text and multimedia	Variable	Specific symptoms, drugs administered; low noise
Sales and school attendance data	O(TB)	Mostly structured text data	Low throughput and high latency	Aggregate indicators may indicate widespread disease prevalence
Twitter (and other social media)	O(PB)	Unstructured text (and multimedia) data	High throughput and low latency	Highly noisy, unspecific indicators
News reports (and other media)	O(TB)	Unstructured text (and multimedia) data	High throughput and high latency	Need experts and/or NLP$^+$ tools to curate and filter relevant data

Table 1. Summary of direct and indirect sources relevant for public health surveillance. [1]PB: petabyte, 10^{15} bytes; [2]TB: terabyte, 10^{12} bytes. The direct and indirect sources are separated by a horizontal line.$^+$ NLP: Natural Language Processing.

cations. Novel information sources include passive search query, micro-blogging data, crowd-sourced data and other social media data. The use of this data for public health surveillance has been well documented and validated for major public health events, including influenza and dengue epidemics [4, 32, 6]. The time delay between incident and report of an outbreak is significantly less for non-traditional sources (e.g., from minutes to hours) than that for traditional sources (on the order of hours to weeks to months) depending on how the data is aggregated. Each data source may provide a distinct perspective on key indicators for public health; however, no single source is sufficient to provide guidance to public health officials regarding the disease spread process.

A summary of the various data sources and their usage in the context of public health surveillance is shown in Table 1. Commonly used attributes for data such as volume (size of the dataset), variety (including structured versus unstructured data, multimedia, text, etc.), velocity (throughput and latency in the data), and value (the overall utility of the dataset and a qualitative measure of how noisy the dataset can be) are shown for the different data sources. One can clearly observe that the variety of datasets used as well as the rich embedding of textual and multimedia data can make it quite challenging to develop analytic tools specific to public health surveillance. In addition, each data source can potentially carry information at various time-resolutions; for e.g., EHR and EMS datasets can be reported only after the incidences have been reported, thus inducing a delay of about an hour or more, versus Twitter (and other social media datasets) information is passed on almost instantaneously.

Given the diversity in the data sources, the different timescales at which they are reported, and quality of data, it is tremendously challenging to develop automated techniques to gain insights relevant to public health surveillance questions. In addition, collecting information across multiple (direct and indirect) sources and developing statistical techniques that detect emerging patterns across diverse datasets and/or correlate information across data sources related to disease surveillance present unique challenges to the community. Thus, public health surveillance is a typical big data problem, similar to other realms such as social media monitoring and financial data analytics.

Bio-surveillance monitoring systems

In this section we present a survey of current tools developed for public health surveillance and summarize their ability to handle diverse datasets and integrate multiple data sources. A summary of the capabilities as well as the diverse datasources that these programs use are summarized in Table 2. In each of the systems, we also summarize how visual analytics is utilized in gathering and providing insights into disease outbreaks across the US and the world. Popular bio-surveillance platforms include HealthMap [13], Global Public Health Intelligence Network (GPHIN) [26], ProMED-mail [36], BioCaster [10] and EpiSPIDER [34], which gather information from reports regarding disease outbreaks (including food-, water- or air-borne diseases) occurring throughout the world. These tools rely on open source data, i.e., information from indirect sources including news and media reports and alerts from global surveillance agencies. Other systems such as ESSENCE (Electronic Surveillance System for Early Notification of Community-based Epidemics) [3], RODS (Real-time Outbreak and Disease Surveillance) [12], GEIS (Global Emerging Infections surveillance and response System) [35] and PHIN (Public Health Information Network) [1], utilize information from traditional sources of syndromic surveillance. Note that almost all bio-surveillance platforms provide data ingestion from a variety of sources and possess only cursory analysis capabilities.

A majority of the tools described above predominantly make use of geographic information system (GIS) maps to effectively summarize alerts. GIS outputs are a natural frame of reference to convey effectively to the end-users the nature of outbreaks and provide an estimate of the number of affected people. Further, these results also provide novel insights into: (a) population movement and potential contacts between people across geographic locations [7, 20], (b) disease occurrences with projection of spatio-temporal changes [18], (c) detection of potential hotspots [25] and integrating computational simulations with real datasets [22].

Given the large amount of heterogeneous textual data, natural language processing (NLP) tools are also utilized in the systems described above to process the collected information and filter/cluster/classify the datasets relevant to specific diseases. For example, EpiCanvas [14] and CommonGround [21] are

Program	Data sources	Analytics	Visualization
ProMED-mail	Media reports, official reports, local observers, etc	Highly curated by manual filtering; global views	GIS reports and alerts prioritized by disease occurrence and scale; textual summary of data
HealthMap	Baidu, Eurosurveillance, Google, ProMED,WHO, etc.	Classification, Clustering, Filtering, integration; Mostly unstructured textual data analytics	GIS reports, timelines and tables/graphs for summarizing disease outbreaks
BioCaster	EurekAlert!, European Media Monitor Alerts, Google, CDC's Morbidity and Mortality Weekly Report, MeltWater, ProMED, WHO, etc.	Multi-lingual reporting; Ontology driven analysis; Classification, clustering, filtering, integration from various data sources; unstructured textual data analytics	GIS reports; Graphs and tables representing different diseases
EpiSPIDER	Daylife, Google, Humanitarian news, Moreover, ProMED, Twitter, WHO	Only English language articles; Mapping system to view and filter reports	Timeline visualization and ordering of events; word-cloud depiction of topics
GPHIN	Internet reports/news articles/ health hazards, etc.	Continuous acquisition and data analysis to aggregate news feeds; multi-lingual filtering; classification and automatic ranking	Alerts reported as text to users
ESSENCE	Hospital emergency departments; pharmacy reports; diagnostic laboratory tests	Multivariate data analysis; time series analysis; statistical process control	Web-based search; GIS reporting; alert lists; tables
RODS	Hospital emergency departments; chief complaints data	Multi-variate data analysis; Bayes classifiers	graphs and tables to indicate alerts
GEIS	NASA GIMMS (Global Inventory Modeling and Mapping System); ecologic niche modeling; arthropod vector (mosquito, sand fly, tick), animal disease-host/ reservoir	Time-series analysis on homogeneous data, thresholds, spatial co-occurrence of reported events	GIS visualization with layering of information
PHIN	Electronic health records and other civilian data sources (CDC)	Application models (externally developed analytics)	Web-based interfaces including GIS, tables and graphs

Table 2. Summary of common bio-surveillance systems. Systems using open-source or non-traditional data and traditional data are separated by a horizontal line.

two front ends for bio-surveillance tools where NLP outputs are effectively integrated with GIS and other graphical/tabular outputs to allow users to interact, visualize, explore and develop hypotheses about emerging diseases. Other tools that use interactive visual analytics of high-throughput streams (such as Twitter) can also be customized for bio-surveillance.

Graphical representation of data integrated with GIS and NLP tools is essential to guide end-users to answer questions related to bio-surveillance. For example, several recent studies have demonstrated that multi-panel graphs [8], visualization of outputs from clustering of high-dimensional data sets [33], tag interactions/formatting (i.e., identifying terms of interest and linking them based on statistical measures of importance within the data), and correlation lines (i.e., summarizing temporal correlations between co-occurrences of terms) [21] can provide better situational awareness from bio-surveillance. Additionally, several tools such as LeadLine [11] exist to develop a story-board description of events detected from social media sites, and can be customized for bio-surveillance; however the use of these visualization tools are largely dependent on the backend statistical algorithms used to characterize the data.

VISUAL ANALYTICS FOR BIO-SURVEILLANCE

An outstanding challenge within the public health community is the ability to identify the origin of novel and emerging (food-, water-, air-borne and/or other infectious) diseases within widespread, heterogeneous populations, both accurately and in a timely fashion from the aforementioned data sources. This would require integration of diverse sources of information, including direct and indirect sources (Table 1) as well as development of novel statistical analysis and inference tools to (a) better predict potential disease hotspots and sentinel points for occurrence of large-scale epidemics (i.e., situational awareness), (b) provide early warning of imminent health threats and emerging health-related events and (c) provide decision makers with insights on how to manage disease outbreaks and control strategies. Interfacing the back-end data analytic tools with the end-user would require the development of novel visual analytics and user interfaces that effectively allow flow of information between such complex systems and their respective users. In this section, we highlight two use cases that illustrate how visual analytics can play an important role in bio-surveillance applications:

Use case 1: Linking text and multimedia data sources to improve situational awareness

As the adage goes "a picture is worth a thousand words", viewing and collating information from diverse data streams that include both text and multi-media can be particularly relevant for bio-surveillance tasks involving situational awareness. For example, a cursory search on the web for the flu yields thousands of images (with annotated information such as #flu), with additional textual information that can be gathered from Twitter (with the same/similar annotations). However, using the information from these images (for e.g., their geo-location) and correlating them with Twitter streams (e.g., their geo-location, or reports of flu occurrences) to get a better perspective on which geographic regions are more affected by an emerging epidemic. Although Twitter data streams have been used to track and monitor epidemics [9], only recently images and location-based monitoring have been used to gather insights on disaster management in the context of the Boston bombing incident [5].

The Oak Ridge Bio-surveillance Toolkit (ORBiT) [27] we are developing, provides a user interface that enables users to link information from multiple datasources, as illustrated in Fig. 1.

Note that information from potential links within the textual streams can (potentially) navigate into multimedia streams (such as YouTube or Instagram) and in some cases, the textual data might itself contain self-reported images/videos. Linking the information, in particular, for identifying the geographic region that may be affected or a user (or groups of users), can be valuable to quickly analyze which regions may need more attention during an emerging epidemic. This motivates the need for intuitive user interfaces that can potentially bring up both GIS based maps and time-series analyses on individual datasources to facilitate a visual integration of the diverse data-streams and develop intuition regarding emergent behavior from current data sources. Similar to other domains such as analyses of molecular simulations [28] and crisis management [24], organizing the data as a storyboard can provide novel insights into emerging bio-events.

The ability to examine multiple data sources simultaneously with multi-media can provide additional inputs to end-users regarding which regions are susceptible to disease outbreaks and guiding decision makers to focus more resources to these regions. In addition, the ability to corroborate similar information across multiple sites can provide additional confidence in these indirect sources of data. Thus, the linking of information can potentially inform users about related topics, how information from one source connects to the other and lead to better clustering of datasets that can improve situational awareness in the context of bio-surveillance.

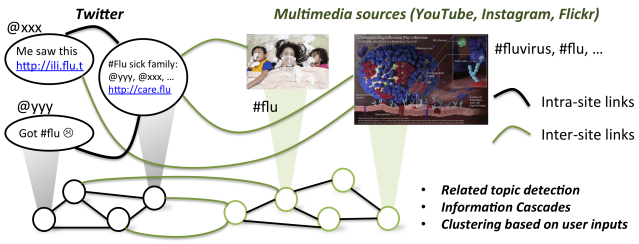

Figure 1. Linking between heterogeneous data-sources in ORBiT.

For traditional syndromic surveillance data sources, one can use similar annotations to potentially integrate data from these large datasets. For example, most EHR or prescription records provide descriptions of patient names or geographic locations, which can represent potential starting points to link information across these data sources. Many surveillance datasets often provide additional records on the patients' symptoms, medical procedures, and/or images of scans/X-rays. However, a significant challenge associated with these datasets is the ability to link information across the diverse data sources, especially when most records have privacy restrictions, which can pose challenges in allowing integration of data. One would either have to resort to privacy preserving data mining approaches [2] or use statistical techniques to potentially summarize the content of the data before allowing end-users such as analysts to visually interact with the data. However, more research in visual analytics and de-identification of data needs to be carried out to alleviate the issues of privacy and allow end-users to interactively link these diverse datasets [25, 15].

Use case 2: Integrating epidemiological models with real-time bio-surveillance data

Current bio-surveillance systems provide visual interpretation of results within the perspective of the data being analyzed. However, a significant challenge within the public health surveillance community is the ability to utilize the data and gather insights into the consequences of an emerging disease/epidemic. For example, public health officials are usually interested in gathering insights into the projected number of people infected during an epidemic and how fast the epidemic is spreading through the population. Therefore, there is a need to translate/transform the information obtained from analysis of both direct and indirect data sources into epidemiological models. Within ORBiT, we are implementing analytical tools that will enable the translation of direct and indirect data into parameters for epidemiological models. Although these tools need further research and development, the availability of tools to analyze and visualize the data and incorporate user-feedback to translate these observations into useful epidemiological models can have significant impact on improving situational awareness and develop better epidemiological models that better reflect ground-level observations.

CONCLUSIONS

From a visual analytics stand point, it is important to note that the current bio-surveillance tools need to implement the aforementioned use-cases to have a positive impact on enabling early warning and detection of emerging disease outbreaks, improving our ability to predict when, where and how novel/emerging diseases are spreading and incorporating user feedback into bio-surveillance work-flows. Enabling user interfaces to translate information from a data analytics perspective to a simulation perspective, where the objective is to forecast the impact of emerging disease outbreaks is key to having better situational awareness. In addition, there is a need to develop novel analytic tools that address anomaly detection from multi-modal (potentially high dimensional) datasets [29], and time-series analyses that respect the underlying statistical structure of the data [31] will also be key in developing automated tools for bio-surveillance. As ORBiT evolves, we will actively pursue incorporating such features into its user interface to improve situational awareness as well as guide epidemiological models to achieve high fidelity with ground-level observations.

ACKNOWLEDGMENTS

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. *Public Health Information Network*. Centers for Disease Control, 2013.
2. Agrawal, R., and Srikant, R. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, ACM (New York, NY, USA, 2000), 439–450.

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

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

5. Cassa, C., and Chunara, R. Twitter as a sentinel in emergency situations: Lessons from the boston marathon explosions. *PLoS Currents: Disasters 1*, doi:10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b (2013).

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

7. Choi, J., Lee, S.-j., Gigitashvilli, S., and Wilson, J. Situation awareness tool for global argus. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, VAST '07, IEEE Computer Society (Washington, DC, USA, 2007), 213–214.

8. Chui, K. K. H., Wenger, J. B., Cohen, S. A., and Naumova, E. N. Visual analytics for epidemiologists: Understanding the interactions between age, time, and disease with multi-panel graphs. *PLoS ONE 6*, 2 (02 2011), e14683.

9. Chunara, R., Andrews, J. R., and Brownstein, J. S. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene 86*, 1 (2012), 39–45.

10. Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics 24*, 24 (2008), 2940–2941.

11. Dou, W., Wang, X., Skau, D., Ribarsky, W., and Zhou, M. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the 2012 IEEE Visual Analytics Science and Technology* (2012).

12. Espino, J., Wagner, M., Szczepaniak, C., Tsui, F.-C., Su, H., Olszewski, R., Liu, Z., Chapman, W., Zeng, X., Ma, L., Lu, Z., and Dara, L. J. Removing a barrier to computer-based outbreak and disease surveillance — the rods open source project. *Morb Mor Wkly Rep 53*, Suppl (2004), 32–39.

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

14. Gesteland, P. H., Livnat, Y., Galli, N., Samore, M. H., and Gundlapalli, A. V. The epicanvas infectious disease weather map: an interactive visual exploration of temporal and spatial correlations. *Journal of the American Medical Informatics Association 19*, 6 (2012), 954–959.

15. Giannotti, F., Giannotti, G., and Pedreschi, D. *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. SpringerLink: Springer e-Books. Springer, 2008.

16. Hay, S. I., George, D. B., Moyes, C. L., and Brownstein, J. S. Big data opportunities for global infectious disease surveillance. *PLoS Med 10*, 4 (04 2013), e1001413.

17. JK, L., M, A., K, W., and et al. Factors associated with death or hospitalization due to pandemic 2009 influenza a(h1n1) infection in california. *JAMA 302*, 17 (2009), 1896–1902.

18. Kamel Boulos, M., Resch, B., Crowley, D., Breslin, J., Sohn, G., Burtner, R., Pike, W., Jezierski, E., and Chuang, K.-Y. 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 (2011), 67.

19. Khan, K., Sears, J., Hu, V. W., Brownstein, J. S., Hay, S., Kossowsky, D., Eckhardt, R., Chim, T., Berry, I., Bogoch, I., and Cetron, M. Potential for the international spread of middle east respiratory syndrome in association with mass gatherings in saudi arabia. *PLoS Currents: Outbreaks* (2013).

20. Livnat, Y., Agutter, J., Moon, S., and Foresti, S. Visual correlation for situational awareness. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, IEEE Computer Society (Washington, DC, USA, 2005), 13–.

21. Livnat, Y., Jurrus, E., Gundlapalli, A. V., and Gestland, P. The commonground visual paradigm for biosurveillance. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on* (2013), 352–357.

22. Livnat, Y., Rhyne, T., and Samore, M. Epinome: A visual-analytics workbench for epidemiology data. *Computer Graphics and Applications, IEEE 32*, 2 (2012), 89–95.

23. Lombardo, J., and Buckeridge, D., Eds. *Disease Surveillance: A Public Health Informatics Approach*. John Wiley and Sons, 2006.

24. Lupo, L., Malizia, A., Diaz, P., and Aedo, I. Socialstory: a social storyboard system for sharing experiences in emergencies. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, ACM (New York, NY, USA, 2012), 786–787.

25. Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., and Ebert, D. A visual analytics approach to understanding spatiotemporal hotspots. *Visualization and Computer Graphics, IEEE Transactions on 16*, 2 (2010), 205–220.

26. Mawudeku, A., Blench, M., Boily, L., St. John, R., Andraghetti, R., and Ruben, M. *The Global Public Health Intelligence Network*. John Wiley Sons Ltd, 2013, 457–469.

27. Ramanathan, A., Pullum, L., Steed, C., Quinn, S., and Chennubhotla, C. Oak ridge bio-surveillance toolkit:. In *IEEE VAST Workshop on Public Health's Wicked Problems: Can InfoVis Save Lives?* (2013).

28. Ramanathan, A., Savol, A. J., Agarwal, P. K., and Chennubhotla, C. S. 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 (2012), 2536–2551.

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

30. Rasko, D. A., Worsham, P. L., Abshire, T. G., Stanley, S. T., Bannan, J. D., Wilson, M. R., Langham, R. J., Decker, R. S., Jiang, L., Read, T. D., Phillippy, A. M., Salzberg, S. L., Pop, M., Van Ert, M. N., Kenefic, L. J., Keim, P. S., Fraser-Liggett, C. M., and Ravel, J. Bacillus anthracis comparative genome analysis in support of the amerithrax investigation. *Proceedings of the National Academy of Sciences 108*, 12 (2011), 5027–5032.

31. Savol, A. J., Burger, V. M., Agarwal, P. K., Ramanathan, A., and Chennubhotla, C. S. Qaarm: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin. *Bioinformatics 27*, 13 (2011), i52–i60.

32. Signorini, A., Segre, A. M., and Polgreen, P. M. 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 (05 2011), e19467.

33. Sims, J. N., Isokpehi, R. D., Cooper, G. A., Bass, M. P., Brown, S. D., John, A. L. S., Gulig, P. A., and Cohly, H. H. Visual analytics of surveillance data on foodborne vibriosis, united states, 1973–2010. *Environmental Health Insights 5* (11 2011), 71–85.

34. Tolentino, H., Kamadjeu, R., Fontelo, P., Liu, F., Matters, M., Pollack, M., and Madoff, L. Scanning the emerging infectious diseases horizon - visualizing ProMed emails using EpiSPIDER, Apr. 2007.

35. Witt, C., Richards, A., Masuoka, P., Foley, D., Buczak, A., Musila, L., Richardson, J., Colacicco-Mayhugh, M., Rueda, L., Klein, T., Anyamba, A., Small, J., Pavlin, J., Fukuda, M., Gaydos, J., Russell, K., 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 (2011), S10.

36. Yu, V. L., and Madoff, L. C. Promed-mail: An early warning system for emerging diseases. *Clinical Infectious Diseases 39*, 2 (2004), 227–232.