

# **Towards A National Infectious Disease Information Infrastructure**

## ***A Case Study in West Nile Virus and Botulism***

Daniel Zeng   Hsinchun Chen   Chunju Tseng   Catherine A. Larson

Department of Management Information Systems  
University of Arizona, Tucson, Arizona

Millicent Eidson   Ivan Gotham  
New York State Department of Health  
SUNY, Albany

Cecil Lynch  
California Department of Health Services  
UC Davis

Michael Ascher  
Lawrence Livermore National Laboratory

### ABSTRACT

*Information technologies are playing an increasingly important role in preventing, detecting, and managing infectious disease outbreaks. This paper presents a collaborative infectious disease informatics project led by an interdisciplinary team of information systems researchers and public health researchers and practitioners. This project has resulted in a research prototype called the WNV-BOT Portal system, which provides an integrated infectious disease information sharing, analysis, and visualization environment across jurisdictions.*

## **1. Introduction**

Infectious disease outbreaks are critical threats to public health and national security (Damianos et. al., 2002). With the greatly expanded trade and travel, infectious diseases, either naturally occurred or caused by biological terror attacks, can spread at a fast pace within and across country borders, resulting in potentially significant loss of life, major economic crises, and political instability.

Information systems play a central role in developing an effective comprehensive approach to prevent, detect, respond to, and manage infectious disease outbreaks of plants, animals, and humans (Damianos et. al., 2002). Currently, a large amount of infectious disease data is being collected by various laboratories, health care providers, and government agencies at local, state, national, and international levels (Pinner et. al., 2003). Furthermore, many agencies have developed information access, analysis, and reporting systems of varying degrees of sophistication in this important area of infectious disease informatics. For example, in its role as the key agency responsible for human reportable diseases in the U.S., the Centers for Disease Control and Prevention (CDC) has developed computerized reporting systems for local and state health departments. Similarly, the United States Department of Agriculture (USDA) is enhancing data systems for certain animal diseases, and the U.S. Geological Survey (USGS), through its National Wildlife Health Center (NWHC) and numerous partners, manages databases for

wildlife diseases. Databases may also be available at other federal and state or local health, agriculture, and environment/wildlife agencies and laboratories. However, access to some of these data sources and related search and reporting functionalities may be limited to the agencies that have developed such systems (Kay et. al., 1998), reducing the effective use of infectious disease data in the national and global contexts. In addition, real-time data sharing, especially of databases across species, could enhance expert scientific review and rapid response using input and action triggers provided by multiple government and university partners. The key technical and policy-related challenges that have to be addressed when developing an effective *national infectious disease information infrastructure* are summarized below.

- ***Existing infectious disease information systems do not fully interoperate.*** Most existing systems have been developed in isolation (Kay et. al., 1998), particularly for different species. As such, when disease-control agencies need to share information across systems, they may need to use e-mail attachments, mailing, faxing, telephone calls, or manual data (re)entry. In addition, much of the search and data analysis function is only accessible to internal users.
- ***An efficient reporting and alerting mechanism across organizational boundaries is lacking.*** For both national security and public health purposes, certain information regarding infectious diseases needs to be quickly propagated through the chain of public health agencies; such information may also need to be shared with law enforcement and national security agencies in a timely manner (Berndt et. al., 2003). Certain models exist within the human public health community (e.g., CDC's ArboNet and Epi-X) and within certain states (e.g., New York State's Health Information Network). However, in general the current reporting and alerting mechanism is far from complete and efficient, and may involve extensive and error-prone human interventions.
- ***The information management environment used to analyze large amounts of infectious disease data and develop predictive models needs major improvements.*** Current infectious disease information systems provide very limited support to professionals analyzing data and developing predictive models. An integrated analytical environment that offers functionalities such as geocoding, advanced data mining and summarization capabilities, and visualization support is critically needed.
- ***Data ownership, confidentiality, security, and other legal and policy-related issues need to be closely examined.*** When infectious disease related data is shared across jurisdictions, important access control and security issues need to be resolved between the involved data providers and users. Subsets of such data are also governed by relevant healthcare and patient-related laws and regulations. Special care has to be taken when dealing with issues arisen from data sharing and aggregation, especially in the cross-jurisdiction context.

This paper summarizes our ongoing research and system development effort motivated to address some of the above challenges. Funded by the National Science Foundation through its Digital Government program and the Intelligence Technology Innovation Center, our effort is aimed at developing scalable technologies and related standards and protocols needed by the full implementation of the national infectious disease information infrastructure and at studying related policy issues.

Our highly interdisciplinary research team consists of three groups: (1) the Artificial Intelligence Laboratory at the University of Arizona, (2) the New York State Department of Health and its partner Health Research, Inc., and (3) the California State Department of Health Services and its partner PHFE Management Solutions. Initiated in October 2003, our project has been focused on two prominent infectious diseases: *West Nile Virus* (WNV) and *Botulism*. These two diseases were chosen as our first target because of their significant public health and homeland security implications and the availability of related datasets in both New York and California states. After an intensive 4-month research and system development effort, we have completed a research prototype called the ***WNV-BOT Portal*** system. This system provides integrated, Web-enabled access to a variety of distributed data sources related to WNV and Botulism. It also provides advanced information visualization capabilities as well as predictive

modeling support. In this paper, we summarize the background and application context of our project and present the main technical components of WNV-BOT Portal. We also discuss broader technical and policy issues related to the design and development of a scalable national infectious disease infrastructure based on the lessons learned through our bi-state case study and prototyping effort in WNV and Botulism.

The rest of the paper is structured as follows. Section 2 presents general research issues in infectious disease informatics. It helps to set up a reference framework for our research activities. Section 3 discusses WNV and Botulism datasets and the related existing public health systems which WNV-BOT Portal is designed to integrate and interoperate. In Section 4, we present the overall system design and main technical components of WNV-BOT Portal. We conclude the paper in Section 5 by summarizing our research and discussing our ongoing activities and future plan.

## **2. Infectious Disease Informatics: A Research Framework**

In this section, we provide a brief overview of the field of infectious disease informatics (IDI) and the major requirements of a national infectious disease information infrastructure (NIDII). The objective of IDI research can be summarized as the development of the science and technologies needed for collecting, sharing, reporting, analyzing, and visualizing infectious disease data and for providing data and decision-making support for infectious disease prevention, detection, and management. IDI research directly benefits public health agencies in their infectious disease fighting activities at all levels of government and in the international context. It also has important applications in law enforcement and national security concerning biological terror attacks (Siegrist, 2002).

IDI research is inherently interdisciplinary, drawing expertise from a number of fields including but not limited to various branches of information technologies such as data sharing, security, Geographic Information Systems (GIS), data mining and visualization, and other fields such as bioinformatics and biostatistics. It also has a critical policy component dealing with issues such as data ownership and access control, privacy and data confidentiality, and legal requirements. Because of its broad coverage and potential impact, IDI research, along with related infrastructure development efforts, can be most successful through broad participation and partnership from various academic disciplines, public health and other disease surveillance, management, diagnostic, or research agencies at all levels, law enforcement and national security agencies, and related international organizations and government branches. In addition, with the majority of potential bioterrorism agents being zoonotic (in common between humans and non-human animal species), it is particularly critical to integrate and analyze disease information across different animal species.

In the short term, developing a technical approach to enable information sharing between agencies that have infectious disease-related datasets is critical for the development of the NIDII. This technical approach should include data sharing protocols based on interoperable standards such as XML and a Web-enabled distributed data store infrastructure to allow easy access. It also needs to provide a basic reporting and alerting mechanism across organizational boundaries and provide important geocoding and GIS-based visualization tools to facilitate infectious disease data analysis. In the mid to long term, significant advances in both technical and policy fronts need to be made to develop and realize the potential benefits of the NIDII. From a technological standpoint, a scalable and effective approach needs to be developed to facilitate data sharing across a large number of distributed data sources across jurisdictions for most infectious disease types. More advanced altering and reporting mechanisms are called for, especially for applications that involve public health agencies, other disease management and research agencies, and other government branches such as law enforcement. New “privacy-conscious” data mining techniques also need to be developed to better protect privacy and patient confidentiality (Kargupta et. al., 2003). Furthermore, infectious disease predictive models need to be augmented to take into consideration factors that have not been included in existing research (e.g., global temperature changes, bird migration patterns, major public gathering), along with more refined visualization and geocoding techniques.

From a policy perspective, there are mainly four sets of issues that need to be studied and related

guidelines developed. The first set is concerned with legal issues. There exist many laws, regulations, and agreements governing data collection, confidentiality, and reporting, which directly impact the design and operations of the NIDII. Confidentiality is a concern not only for human data, but also for non-human animal data, because institutions, animal owners, veterinarians, and even those simply reporting wildlife deaths have legitimate concerns about not being identified or associated with disease locations. The second set is mainly related to data ownership and access control issues. The key questions are: Who are the owner(s) of a particular data set? Who are allowed to access, aggregate, or input data? Who own the derivative data? For both original and derivative data, who are allowed to distribute them to whom? The third set concerns data dissemination and alerting. What alerts should be sent to whom under what circumstances? What summaries should be available as public information, using what mechanisms? The policy governing data dissemination and alerting needs to be made jointly by organizations across jurisdictions and has to carefully balance the needs for information and possibility of information overflow (Chen et. al., 2003). The fourth set is concerned with data sharing and possible incentive mechanisms. To facilitate fruitful sharing of infectious disease data on an ongoing basis, all contributing parties at all levels of jurisdictions inside and outside of the public health system need to have proper incentives and benefit from the collaboration.

The research reported in this paper is a response to the short- and mid-term needs of IDI research with a primary focus on the enabling information sharing infrastructure and data analysis and visualization support. Our future work will be more inclusive to address broader policy-related issues.

### **3. West Nile Virus and Botulism Datasets and State Public Health Systems**

The emergence of WNV in the Western Hemisphere was reported first in New York State in late summer 1999. This unprecedented event required rapid mobilization and coordination of hundreds of public health workers, expenditure of millions of dollars on an emergency basis, and immediate implementation of massive disease surveillance and vector control measures. The Health Information Network (HIN) system has been used by New York State to enable rapid and effective response to the WNV crisis. The HIN is an enterprise-wide information infrastructure for secure Web-based information interchange between the New York State Department of Health (NYSDH) and its public health information trading partners, including local health departments and the New York State Department of Agriculture and Markets, New York State Department of Environmental Conservation, and the United States Department of Agriculture's Wildlife Services New York office (Gotham et. al., 2001). This system currently supports 20,000 accounts and 100 mission critical applications, cross-cutting all key public health response partners in the state of New York. It implements sophisticated data access and security rules, allowing for real-time use of the data within the state while protecting confidentiality and scientific integrity of the data. The infrastructure is well suited to public health response, as illustrated by New York's ability to rapidly incorporate it into its plan to respond to the WNV outbreak in NY in 1999-2000 (Gotham et. al., 2001). The system has evolved into an integrated surveillance system containing large quantities of real-time data related to WNV including (a) detailed human cases, (b) dead bird surveillance data, (c) asymptomatic bird surveillance data, (d) detailed reports on veterinarian, owner/residence, necropsy, clinical course for mammals, and (e) mosquito surveillance data.

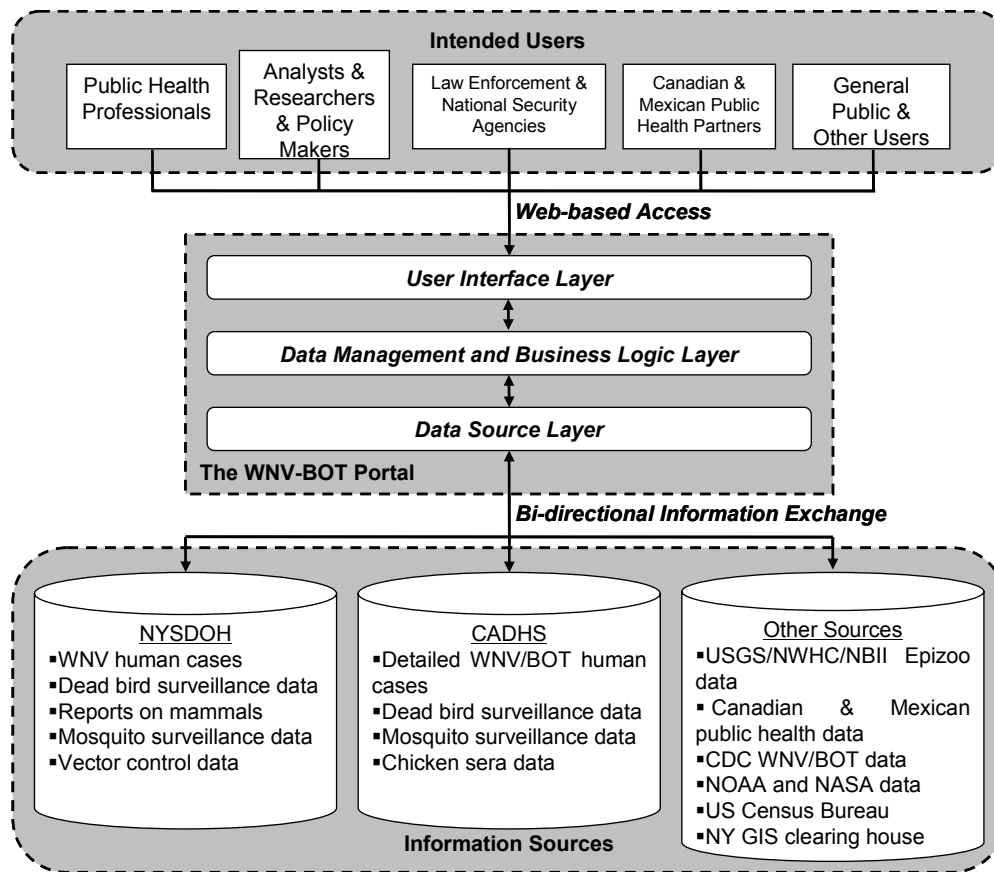
WNV has yet to manifest as an indigenous human disease in California, but the historical geographic spread of this disease and the fact that WNV has been detected in sentinel flocks of chickens and mosquito pools in California, would indicate the high likelihood that the state will have to deal with large numbers of cases later this year. Important as a cause of neurological morbidity and death, WNV is also a prototype of an emerging viral infection. The analysis of data collected regarding its occurrence and spread provides a basis for the development of predictive models for other emerging or as yet unidentified diseases. The California Department of Health Services (CADHS) has access to the detailed datasets from California's mosquito control districts and surveillance data on sentinel flock, dead bird, and equine specimens. In collaboration with USGS, we also have datasets concerning domestic and wild animal

populations that might be exposed to WNV; some related data is available through CDC.

Botulism is a disease rarely seen in the United States with fewer than 200 cases per year reported to the CDC. Despite the low volume of cases, because of the risks associated with the possibility of a terrorist event utilizing botulinum toxin, the importance of having a system in place to identify and manage larger numbers of cases of the disease cannot be overestimated. The transactional data generated in such a system must also be available for post-event analysis in order to improve public health methods and responses. Both New York and California represent a significant part of the world economy and are high risk targets for bioterrorism due to the high level of international traffic into the states. NYSDOH has an internal database for botulism cases that occurred in New York State. In California, no computerized system currently exists that is capable of handling the information gathering, retrieval, and dissemination needs for a bio-terrorist event involving botulinum toxin. However, detailed paper-based information is available on both Botulism cases and antitoxin inventory. In addition, nationwide avian botulism data is maintained and updated by the National Wildlife Health Center.

#### 4. WNV-BOT Portal System Development

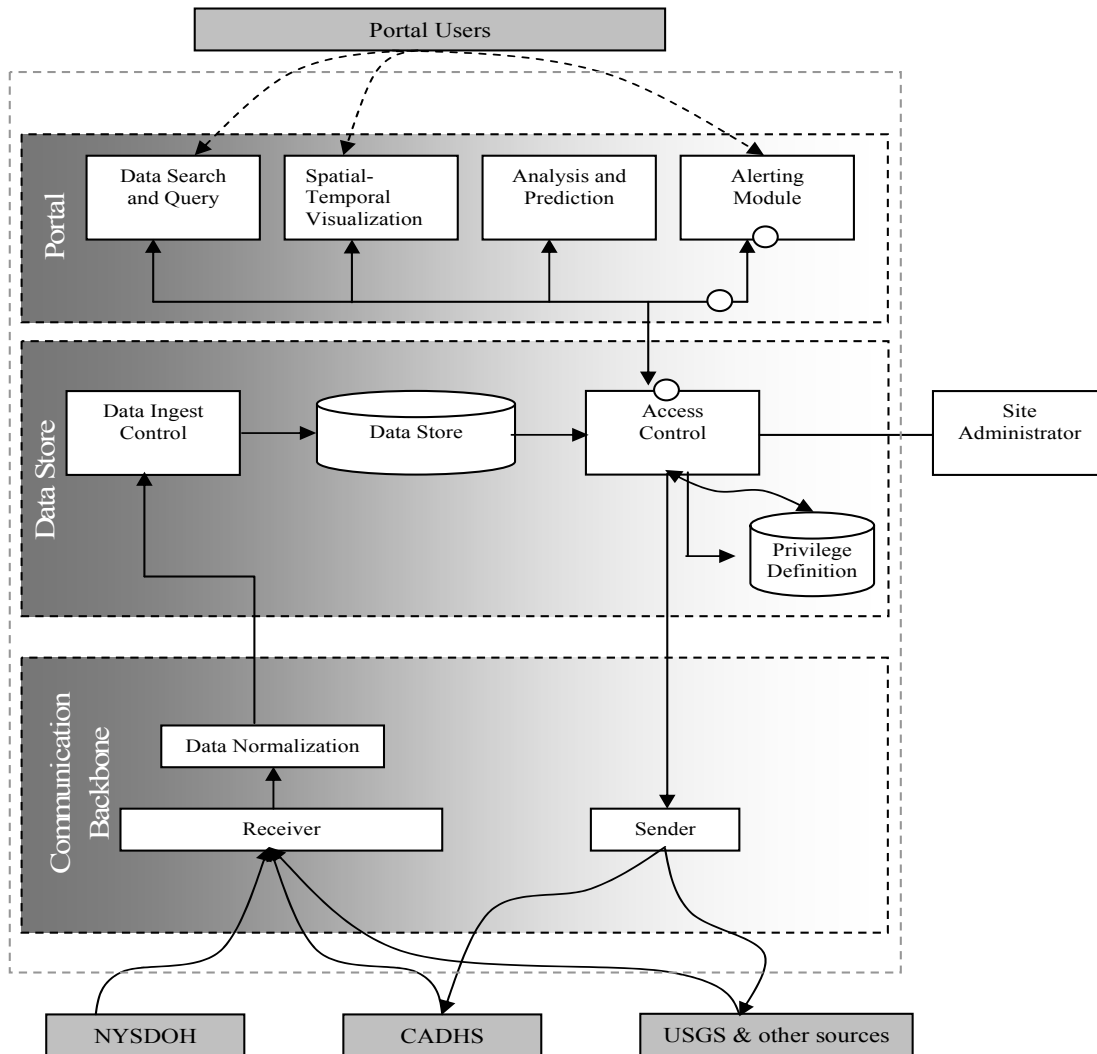
The WNV-BOT Portal system has been developed to integrate infectious disease datasets on WNV and Botulism from New York, California, and several federal data sources. It also provides a set of data analysis, predictive modeling, and information visualization tools tailored for these two diseases. Figure 1 summarizes these datasets and intended users of WNV-BOT Portal.



**Figure 1: Data Sources and Intended Users of the WNV-BOT Portal**

#### 4.1 WNV-BOT Portal System Design

As illustrated in Figure 2, from a systems perspective, WNV-BOT Portal is loosely-coupled with the state public health information systems in that the state systems will transmit WNV/BOT information through secure links to the portal system using mutually-agreed protocols. Such information, in turn, will be stored in the internal data store maintained by WNV-BOT Portal. The system also automatically retrieves data items from sources such as those from USGS and stores them in the internal data store.



**Figure 2. Overall Architecture of the WNV-BOT Portal System**

Architecturally, WNV-BOT Portal consists of three major components: a Web portal, a data store, and a communication backbone. Figure 2 illustrates these components and shows the main data flows between them and the underlying WNV/BOT data sources. The Web portal component implements the user interface and provides the following main functionalities: (a) searching and querying available WNV/BOT datasets, (b) visualizing WNV/BOT datasets using spatial-temporal visualization, (c) accessing analysis and prediction functions, and (d) accessing the alerting mechanism. The remainder of this section discusses the design philosophy and technical details of the data store layer, the communication backbone, and the visualization module. These system components have been fully

implemented. The next section will briefly discuss other components of WNV-BOT Portal which are being actively developed.

#### **4.2 Portal Data Store**

A main objective of WNV-BOT Portal is to enable users from partnering states and organizations to share data. Typically data from different organizations has different designs and stored in different formats. To enable data interoperability, we use Health Level Seven (HL7) standards (<http://www.hl7.org/>) as the main storage format. In our approach, contributing data providers transmit data to WNV-BOT Portal as HL7-compliant XML messages (through a secure network connection if necessary). After receiving these XML messages, WNV-BOT Portal will store them directly in its data store. This HL7 XML-based design provides a key advantage over an alternative design based on a consolidated database. In a consolidated database design, the portal data store has to consolidate and maintain all the data fields for all datasets. Whenever an underlying dataset changes its data structure, the portal data store needs to be redesigned and reloaded to reflect the changes. This severely limits system scalability and extensibility. Our HL7 XML-based approach does not have these limitations. To alleviate potential computational performance problems associated with this HL7 XML-based approach, we are identifying a core set of data fields based on which search will be done frequently and extracting these fields from all XML messages to be stored in a separate database table to enable fast retrieval.

An important function of the data store layer is data ingest and access control. The data ingest control module is responsible for checking the integrity and authenticity of data feeds from the underlying information sources. The access control module is responsible for granting and restricting user access to sensitive data.

#### **4.3 Communication Backbone**

The communication backbone component enables data exchanges between WNV-BOT Portal and the underlying WNV/BOT sources. Several federal programs have been recently created to promote data sharing and system interoperability in the healthcare domain. The CDC's Electronic Disease Surveillance System (NEDSS) initiative is particularly relevant to our research. It builds on a set of recognized national standards such as HL7 for data format and messaging protocols and provides basic modeling and ontological support for data models and vocabularies. NEDSS and HL7 standards are having a major impact on the development of disease information systems. Although these standards have not yet been tested in cross-state sharing scenarios, they provide a solid foundation for data exchange standards in the national and international contexts. WNV-BOT Portal heavily utilizes NEDSS/HL7 standards.

The communication backbone component uses a collection of source-specific "connectors" to communicate with underlying sources. We use the connector linking NYSDOH's HIN system and WNV-BOT Portal to illustrate a typical design of such connectors. The data from HIN to the portal system is transmitted in a "push" manner. HIN sends secure Public Health Information Network Messaging System (PHIN MS) messages to the portal at pre-specified time intervals. The connector at the portal side runs a data receiver daemon listening for incoming messages. After a message is received, the connector will check for data integrity syntactically and invoke the data normalization subroutine. Then the connector will store the verified message in the portal's internal data store through its data ingest control module. Other data sources (e.g., those from USGS) may have "pull"-type connectors which will periodically download information from the source Websites and examine and store data in the portal's internal data store. In general, the communication backbone component provides data receiving and sending functionalities, source-specific data normalization, as well as data encryption capabilities.

#### **4.4 Data Visualization**

The role of visualization techniques in the context of large and complex dataset exploration is to organize and characterize the data visually to assist users in overcoming the information overload problem (Zhu et al., 2000). WNV-BOT Portal makes available an advanced visualization module, called the Spatial

Temporal Visualizer (STV) to facilitate exploration of infectious disease case data and to summarize query results. STV is a generic visualization environment that can be used to visualize a number of spatial temporal datasets simultaneously. It allows the user to load and save spatial temporal data in a dynamic manner for exploration and dissemination. STV has three integrated and synchronized views: periodic, timeline, and GIS. The periodic view provides the user with an intuitive display to identify periodic temporal patterns. The timeline view provides a 2D timeline along with a hierarchical display of the data elements organized as a tree. The GIS view displays cases and sightings on a map. Figure 3 illustrates how these three views can be used to explore infectious disease dataset: The top left panel shows the GIS view. The user can select multiple datasets to be shown on the map in a layered manner using the checkboxes. The top right panel corresponds to the timeline view displaying the occurrences of various cases using a Gantt chart-like display. The user can also access case details easily using the tree display located left to the timeline display. Below the timeline view is the periodic view through which the user can identify periodic temporal patterns (e.g., which months have an unusually high number of cases). The bottom portion of the interface allows the user to specify subsets of data to be displayed and analyzed.

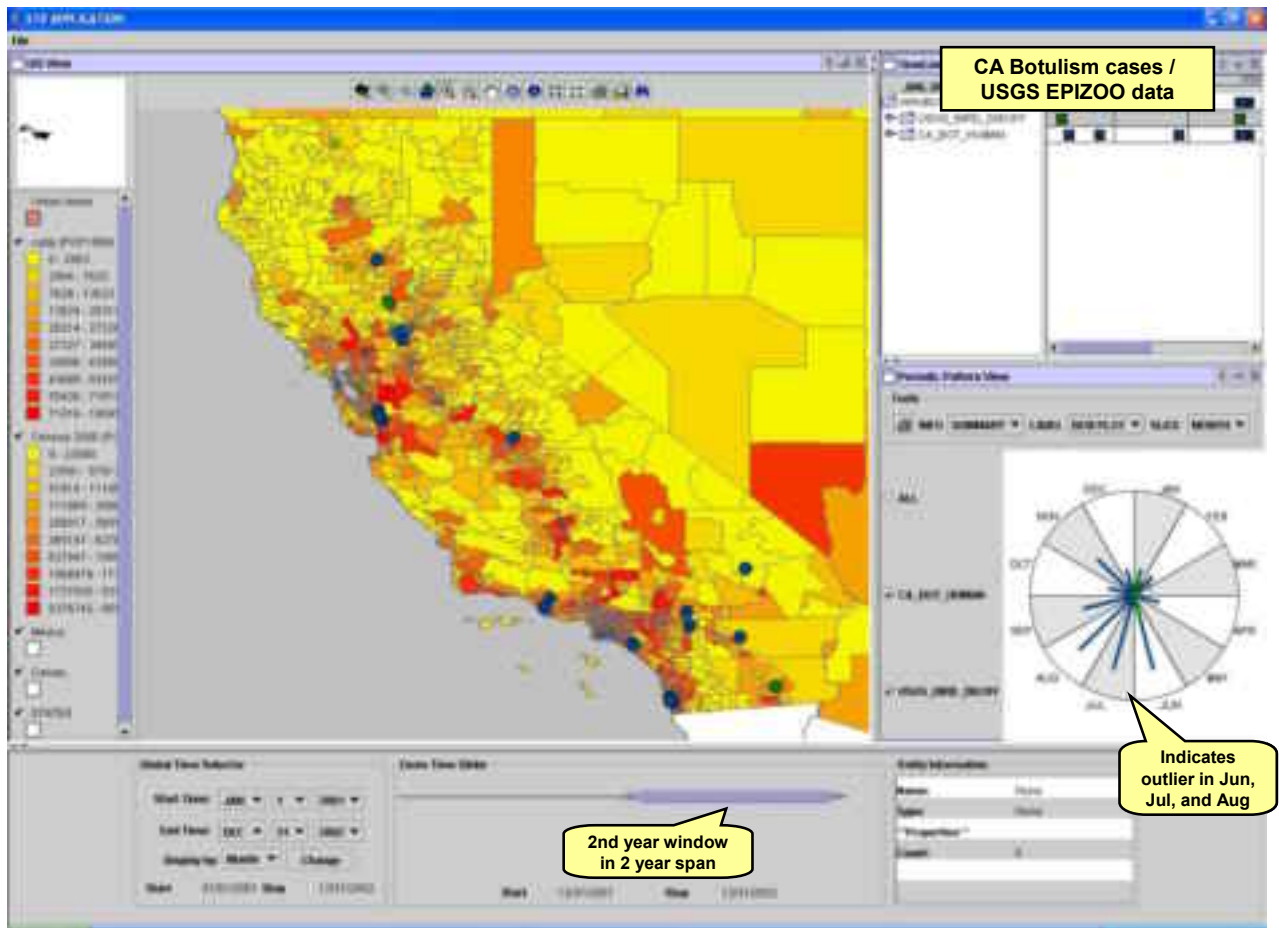


Figure 3. Using STV to Visualize Botulism Data

## 5. Summary and Ongoing Research

This paper presents a collaborative effort between IT researchers and public health agencies aimed at developing a scalable information sharing, analysis, and visualization environment in the domain of infectious diseases. The resulting prototype system, WNV-BOT Portal, focuses on two prominent disease types and has successfully demonstrated the technological feasibility of integrating and interoperating infectious disease datasets for multiple diseases and across jurisdictions. Our project has supported exploration of and experimentation with technological infrastructures needed for the full-fledged implementation of a national infectious disease information infrastructure and helped foster information sharing and collaboration among related government agencies at state and federal levels. In addition, we have obtained important insights and hands-on experience with various important policy-related challenges faced by developing a national infrastructure. For example, a nontrivial part of our project activity has been centered around developing data sharing agreements between project partners from different states.

Our ongoing technical research is focusing on two aspects of infectious disease informatics: hotspot analysis and efficient alerting and dissemination. Hotspot is a condition indicating some form of clustering in a spatial and temporal distribution. For WNV, localized clusters of dead birds typically identify high risk disease areas. Automatic detection of dead bird clusters using hotspot analysis can help predict disease outbreaks and allocate prevention/control resources effectively. Most of existing disease informatics research uses the spatial scan statistic techniques to perform hotspot analysis. We are currently applying other hotspot analysis techniques (e.g., Risk-Adjusted Nearest Neighbor Hierarchical Clustering) that have been developed and successfully applied in crime analysis to disease informatics (Levine, 2002; Jain et. al., 1999). Initial experimental results indicate that these techniques are complementary to the spatial scan techniques in many regards. In a broader context, we are pursuing research in vector borne emerging infection predictive modeling. In particular, we are (a) augmenting existing predictive models by taking additional factors (e.g., weather information, bird migration patterns) into consideration, and (b) tailoring data mining techniques for infectious disease datasets that have prominent temporal features.

There is a critical need to disseminate alert messages in a timely manner during suspected disease outbreaks or bio-terror attacks. In addition, in more routine situations, advisory or informative messages may need to be distributed among various interested parties. We are developing an advanced alerting module as part of WNV-BOT Portal to complement alerting and surveillance systems that already exist in various states. For instance, our alerting module is designed to send alert message across state boundaries or to agencies outside of the public health domain (e.g., law enforcement and homeland security). In our current design, alert messages can come from the following three sources. (a) The user can specify personalized triggering conditions (e.g., “notifying me if there are four Botulism cases within the past two days”). (b) The predictive models may suggest with high confidence that a disease outbreak is in progress. (c) Public health officials may want to send alerts across organizational and state boundaries. Depending on the applicable business rule, in some cases, alert messages will be automatically sent out without review (e.g., alerts triggered by personalized conditions and intended for individual users). In other cases, before alert messages are sent out, they will be reviewed by designated personnel. After approval, the alert dissemination module will deliver them to users or user roles.

Our ongoing and planned policy-related research is focused on (a) development of memoranda of understanding and data sharing agreements between information trading partners, and (b) a user evaluation study providing feedback on the design and functionality of WNV-BOT Portal.

We conclude this paper by discussing the pathway leading to the national infectious disease information infrastructure based on the lessons learned from our WNV-BOT project. Due to the complexity of such an infrastructure from both technical and policy standpoints, we envision that its development path will follow a bottom-up, evolutionary approach. Initially, each individual state will develop its own integrated infectious disease infrastructure for a limited number of diseases. Following successful deployment of such systems, regional nodes linking neighboring states can be established.

Such regional nodes will leverage both state sources and data from federal agencies such as CDC, USGS, and USDA. National and international infrastructures will then become a natural extension and integration of these regional nodes, covering most infectious disease types.

## References Cited

- Berndt, D., Hevner, A. and Studnicki, J., "Bioterrorism Surveillance with Real-Time Data Warehousing," *NSF/NIJ Symposium on Intelligence and Security Informatics*, 2003.
- Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W. and Schroeder, J., "COPLINK: manging law enforcement data and knowledge," *CACM* 46(1), pp. 28-34, 2003.
- Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G. and Hirschman, L., "MiTAP for Bio-Security: A Case Study," *AI Magazine*, 23(4), pp. 13-29, 2002.
- Gotham, I. J., Eidson, M., White, D. J., Wallace, B. J., Chang, H. G., Johnson, G. S., Napoli, J. P., Sottolano, D. L., Birkhead, G. S., Morse, D. L., and Smith, P. F., "West Nile virus: a case study in how NY State Health Information infrastructure facilitates preparation and response to disease outbreaks," *J Public Health Manag Pract*, 7(5), pp. 75-86, 2001.
- Hand, D. J., "*Discrimination and Classification*," Wiley, Chichester, U.K, 1981.
- Jain, A. K., Murty, M. N., and Flynn, P. J., "Data clustering: a review," *ACM Computing Surveys*, 31(3), pp. 264-323, 1999.
- Kay, B. A., Timperi, R. J., Morse, S. S., Forslund, D., McGowan, J. J. and O'Brien, T., "Innovative Information-Sharing Strategies," *Emerging Infectious Diseases*, 4(3), 1998.
- Kargupta, H., Liu, K. and Ryan, J., "Privacy Sensitive Distributed Data Mining from Multi-Party Data," *Proc. of ISI 2003*, pp. 336-342, 2003.
- Levine, N., *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 2.0)*. Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. May 2002.
- Pinner, R. W., Rebmann, C. A., Schuchat, A. and Hughes, J. M., "Disease Surveillance and the Academic, Clinical, and Public Health Communities", *Emerging Infectious Disease*, 9(7), 2003.
- Siegrist, D. W., "The Threat of Biological Attack: Why Concern Now?" *Emerging Infectious Diseases*, 5(4), 2002.
- Zhu, B., Ramsey, M. and Chen, H., "Creating a Large-scale Content-based Airphoto Image Digital Library," *IEEE Transactions on Image Processing, Special Issue on Image and Video Processing for Digital Libraries*, 9(1), pp. 163-167, 2000.