Activity Monitoring for Environmental and Public Health Events [1]

Chung-Sheng Li

IBM Thomas J. Watson Research Center

P O Box 704, Yorktown Heights, NY 10598, USA

csli@us.ibm.com

Received 9 July 2007; Accepted 9 July 2007

Abstract. Activity monitoring has received increasing attention due to the strong environmental and public health needs as well as the rapid advances in both hardware and software technologies. In this paper, we review a number of such activity monitoring scenarios and describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel health activity monitoring system that generates alerts from environmental, behavioral, and public health data sources. A model-based approach is used to develop environmental, disease and behavior models from multi-modal heterogeneous data sources. This system has been successfully applied to various genuine and simulated environmental events such as wild land fires and diseases outbreaks scenarios1.

Keywords: Health activity monitoring, Environmental activity monitoring, Event driven architecture

1. Introduction

Recent advances in both hardware and software technologies enable real-time or near real-time monitoring and alert generation for environmental and public health related activities. The availability of remotely sensed data together with the deployment of sensor networks facilitate environmental activity monitoring (EAM) such as global climate change (including the deterioration of atmosphere, ozone layer holes, temperature rise, glacial melting, sea level rise, vegetation response [1]), deforestation [2-5], flooding [6], earthquake [7], forest fire [8,9], and air pollution [10]. Monitoring of disease outbreaks for public health purposes based on environmental epidemiology has been demonstrated for a number of vector-born diseases such as Hantavirus Pulmonary Syndrome (HPS), malaria, and Denge fever [11-15]. Recently, health activity monitoring (HAM) concept has also been applied to the early detection of subtle human behavior changes due to disease outbreak to provide advanced warnings before significant casualties registered from clinical sources [16].

The alerts generated from either EAM or HAM systems are triggered through the fusion of multi-modal heterogeneous data sources. These data sources include data from remote sensing (such as satellite images), ground sensor network (such as ground sensors for moisture, rainfall, and air quality), retail transactions from supermarkets, phone records, web log, and various GIS related information. HAM will also leverage data generated from clinical sources such as in-patient/outpatient data and prescription data from the pharmacy.

In this paper, we describe the architecture and implementation of the Epi-SPIRE, which is a novel EAM/HAM system capable of generating early warning from monitoring environmental and public health activities. A model-based approach is used to develop the behavior models from multi-modal heterogeneous data sources. Furthermore, a model-based indexing technique has been developed to speed up the data access and retrieval. This system has been successfully applied to vector-born infectious disease such as HPS, pests in the agriculture area such as fire ants, and influenza. For HPS, the advanced warning for high risk regions by using a combination of satellite images and digital elevation map (DEM) can be as much as 9 months [15]. In the case of influenza, preliminary results indicate that early warnings can be generated by Epi-SPIRE using heterogeneous non-traditional data sources earlier than that can be achieved by using only traditional clinical data sources, thus demonstrating the potential benefit of such a system for public health applications.

[1] The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, Air Force Research Lab, NASA, or the United States Government.

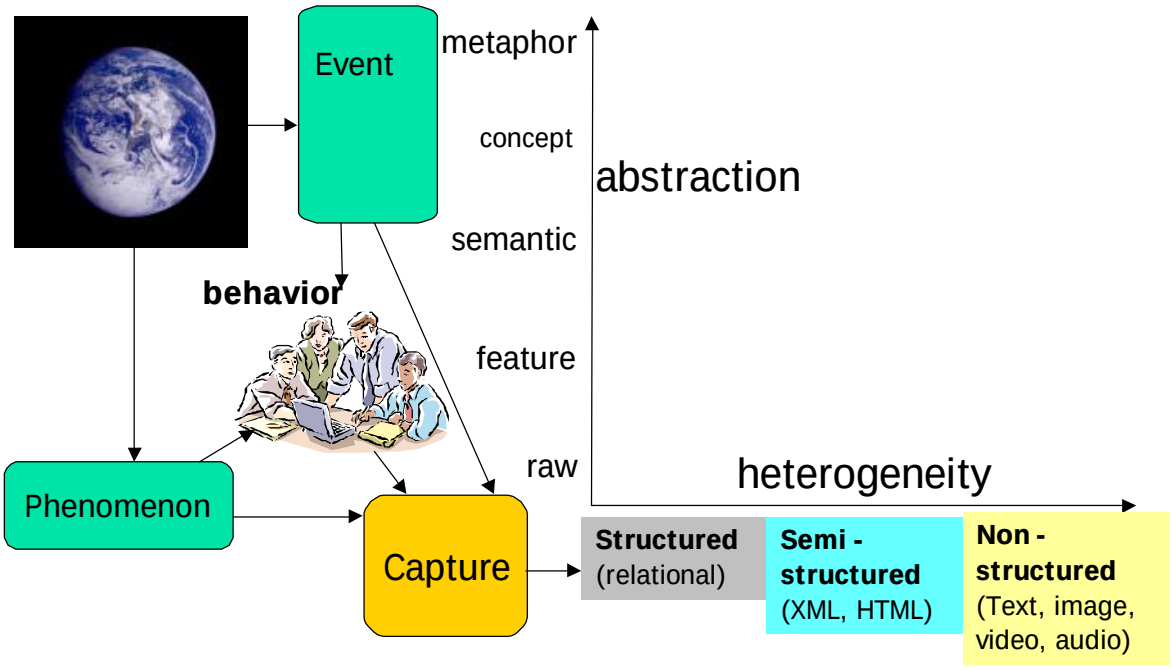

Fig.1. Process of generation of multi-modal heterogeneous data sources

2 Scenarios for Environmental and Health Activity Monitoring

The Environmental Activity Monitoring (EAM) scenarios include:

- **Deforestation**: Deforestation is known to be linked to the reduced biodiversity and increased carbon dioxide emission levels (which is one of the main causes for global warming). Deforestation is still occurring at a rapid pace in spite of the worldwide concern. For example, South America (mostly the Amazon rainforest in Brazil) is losing its rainforest at the rate of 4.3M hectares/year between 2000 and 2005, while Africa is losing 4M hectares/year. Remote sensing has been used for monitoring the rate of deforestation by analyzing the vegetation index and canopy closure of the forest from satellite instruments such as AVHRR.
- **Flooding**: Detecting flood from remotely sensed data often involves change detection of sequences of satellite images or aerial photos. Images taken at visible or infrared bands are constrained by the cloud cover. Recently, satellites with microwave sensors providing global coverage of the Earth's entire land surface on a near-daily basis without severe interference from cloud cover has become available. Using a strategy first developed for wide-area optical sensors, microwave data can be used to measure river discharge changes, river ice status, and watershed runoff. As rivers rise or fall, surface water areas will increase or decrease. Due to much lower microwave emission from water versus land surfaces, these changes can be readily measured in a consistent manner by comparing the newly measured radiance against a previously known target calibration location, and removes most other factors affecting the signal. The Dartmouth Flood Observatory is currently producing near real-time surface water watch [6].
- **Earthquake damage assessment**: Recently, a number of feasibility studies have been conducted based on performing change detection on high resolution remotely sensed data. Research has been mostly focused on investigating using remotely sensed data for various aspects of damage detection [7]. The change detection has been performed against optical images as well as Synthetic Aperture Radar (SAR) images. Real-time damage assessment through these mechanisms allows emergency response plans to be developed more precisely (based on the size of the damage) as well as more timely (based on the exact location of the damage).
- **Wild land fire** [8,9]: Research in wild land fire detection can be categorized as follows: fire risk assessment, fire detection, and fire damage assessment. Fire risk assessment is often based on the fuel load, ground moisture condition, wind direction, etc. Commercial solutions, such as Sanborn Wildland Fire risk Assessment System (WFRAS) [16], have been available for assessing the current fire risk as well as analyzing fire prevention and fuel treatment options for reducing future wildland fire risk. Near real-time wildland fire detection through satellites such as MODIS of the United States is available at [17]. The damage assessment can be done through the evaluation of the reduction of normalized difference vegetation index (NDVI) from the AVHRR instrument (available on a few satellite platforms), which is also provided by WFRAS.

- **Air pollution**: Since 2003, NOAA and EPA of the United States have begun their effort on developing systems for providing nationwide hourly air quality forecasts of ozone and PM2.5 (aerosol particle with diameter less than 2.5 μm). The ability to measure various tropospheric pollutants at the desired spatial and temporal resolution and accuracy still remains a challenge. Currently, NOAA has been using MODIS for demonstrating the feasibility of remote air quality monitoring.
- **Fire Ants**: The red imported fire ant, *Solenopsis invicta*, was introduced into the United States from South America about 50 years ago. Today this exotic pest infests close to 250 million acres in the southeastern United States. Imported fire ants have no natural enemies and adapt to changing climatic and environmental conditions. It is estimated they could infest almost a quarter of the land mass of the United States. High densities of *S. invicta* cause numerous environmental and economic problems. In urban areas ants infest yards, playgrounds and open fields, and as many as 200 mounds per acre have been documented. During flooding, water borne ants sting on contact. In the southeastern United States, fire ants are the number one cause of known stings and bites. Up to 2% of stings result in life threatening anaphylaxis reaction. Morbidity due to secondary bacterial infections is as high as 54%, because the sting site is intensely pruritic. In addition to the toll to human health from envenomization, anaphylaxis, and secondary infections, fire ants attack livestock and damage crops, infrastructure (roadway collapse), and telecommunication electric insulation. Risk model for fire ants based on surface moisture and surface temperature have been demonstrated recently. The spatial resolution of the surface temperature and moisture can be substantially improved by combining sensors on the ground and instrument data from the satellite images.

Health and infectious disease surveillance (a.k.a Health Activity Monitoring – HAM) for public health purposes have existed for many decades. The following is a subset of the 30+ national surveillance systems coordinated by the National Center of Infectious Disease (NCID) at CDC (http://www.cdc.gov/ncidod/osr/site/surv_resources/surv_sys.htm).

- Mortality Reporting System: from 122 cities & metropolitan areas, compiled by the CDC epidemiology program office;
- Electronic Foodborne Outbreak Investigation and Reporting System (EFORS): used by 50 states to report data about Foodborne Outbreaks on a daily basis;
- Foodborne Diseases Active Surveillance Network (FoodNet): consists of active surveillance for foodborne diseases and related epidemiologic studies. This is a collaboration among CDC, the ten emerging infectious program sites (EIPs), the USDA, and FDA;
- Global Emerging Infectious Sentinel Network (GeoSentinel): consists of travel/tropical medicine clinics around the world that monitor geographic and temporal trends in mobility;
- United States Influenza Sentinel Physicians Surveillance Network: 250+ physicians around the country report each week the number of patients seen and the total number with flu-like symptoms;
- The US DoD Global Emerging Infectious Surveillance System (DoD-GEIS): has its own set of surveillance network within the States as well as internationally (http://www.geis.fhp.osd.mil/).
- BioSense: a national initiative at US (coordinated by CDC) to establish near real-time electronic transmission of data to local, state, and federal public health agencies from national, regional, and local health data by accessing and analyzing diagnostic and prediagnostic health data.
- RODS (Real Time Outbreak and Disease Surveillance http://rods.health.pitt.edu/): a joint effort between University of Pittsburgh and CMU. This effort includes an open source version of RODS software (since 2003) and National Retail Data Monitor (NRDM) which monitors the sales of over-the-counter healthcare products in 20,000 participating pharmacies in the states.

3 Activity Monitoring

The multi-modal heterogeneous data sources collected by a EAM or HAM system can come from a wide variety of sources, including (1) sensors monitoring the environment either through *in situ* or remote sensing (such as satellites) to capture the events and phenomenon as they occur; (2) data already collected for other purposes, such as e-seminar, phone records, web log, newsgroup, sewage records; (3) data collected from clinical sources such as insurance claims, in-patient and outpatient data, lab tests, and Emergency Room records.

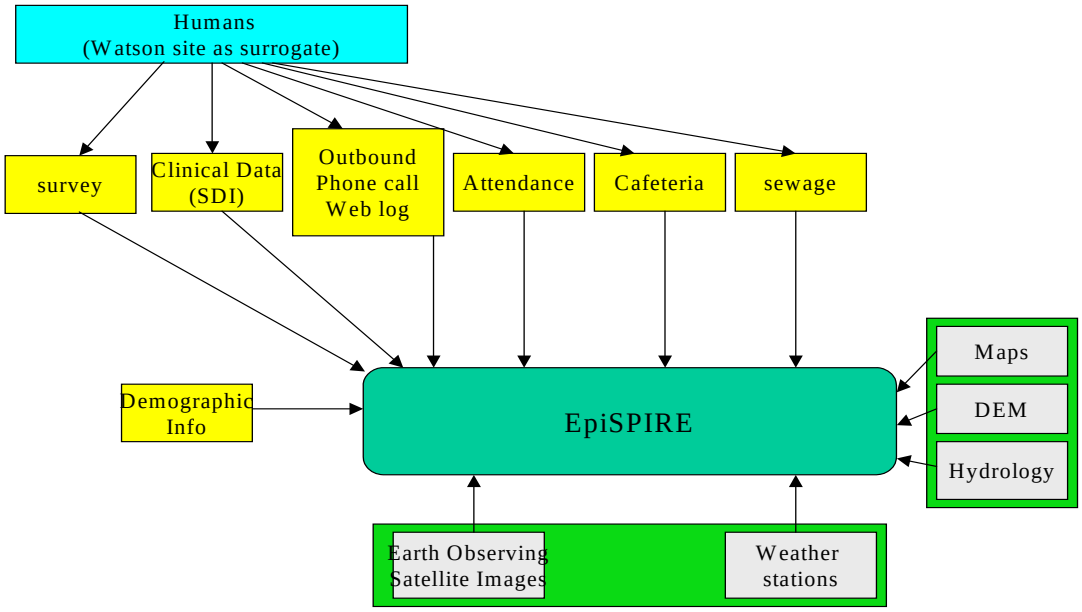

Fig.2. Epi-SPIRE environment

The data sources capturing events and phenomenon related to environments and human behavior, as shown in Fig.1, can be categorized as structured (parametric or relational), semi-structured (HTML or XML), and non-structured (text, image, audio, and video). The data can be potentially captured at various abstraction levels, including raw data (raw images or video), features extracted from the raw data (such as texture and spectral histogram from satellite images), semantic (road, houses), concepts (house surrounded by bushes), and metaphors.

The main challenge in EAM or HAM is to be able to fuse multi-modal heterogeneous information sources (based on models) at different abstraction levels, generate multiple hypothesis of the models for the events, phenomenon and behaviors, and test the validity of the hypothesis using the available data. The end objective of such a system is to predict or detect an upcoming event using the model derived from the fused heterogeneous data sources.

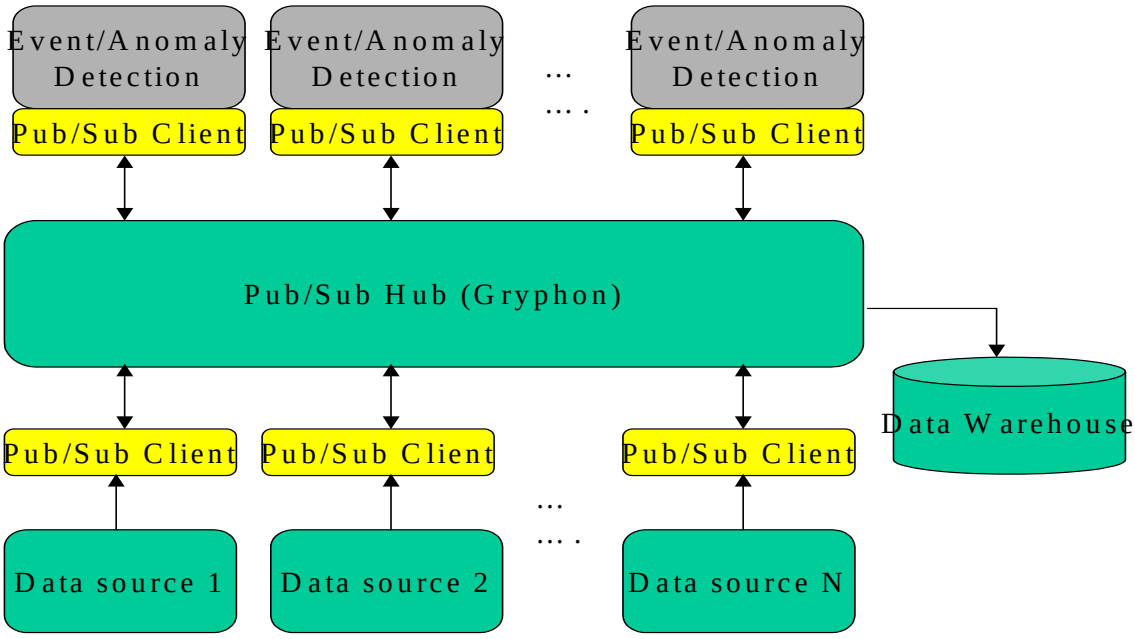

Fig.3. Epi-SPIRE architecture

4 Environments and Architecture

The system environment of Epi-SPIRE is shown in Fig.2. The Epi-SPIRE system uses (1) data collected from the natural environment (such as those collected by the satellites and weather stations), (2) data collected passively as a byproduct of human behavior (such as attendance at work or school, consumption records at cafeteria, sewage generation, web log and phone records), (3) data collected actively from probing the population that are being monitored, usually through periodic survey. In addition to the dynamic data that require real time processing, Epi-SPIRE also utilizes static data such as maps, digital elevation map, hydrology, and demographic information.

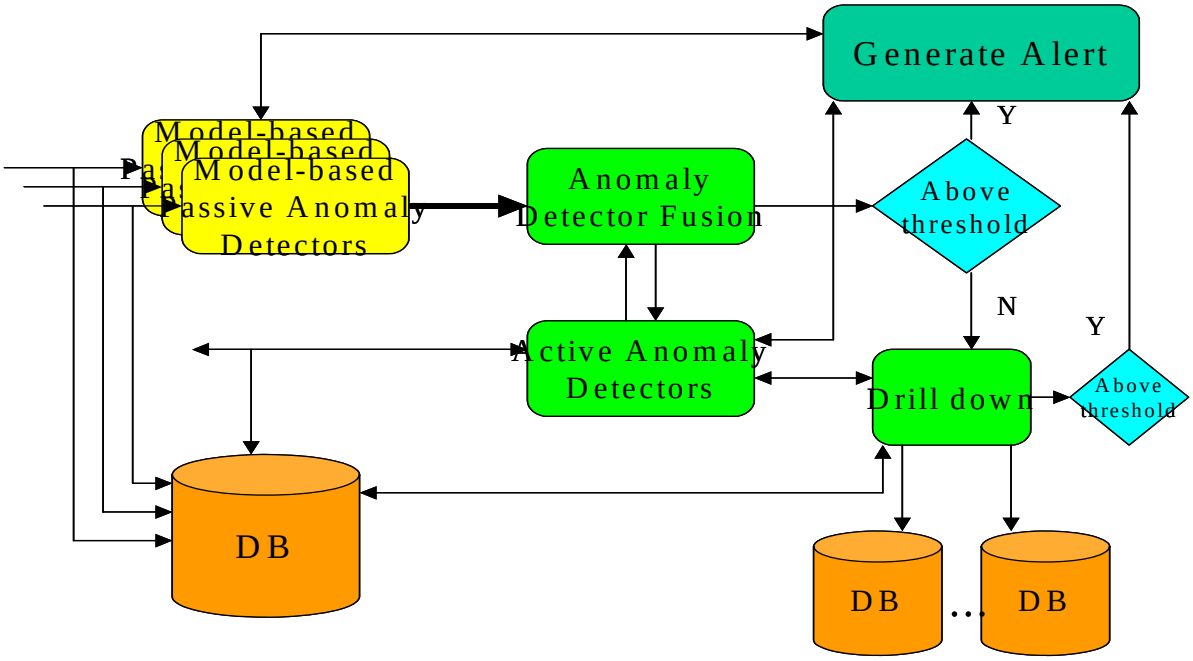

Fig.4. Alert generation process from active and passive detectors

The system architecture for Epi-SPIRE, which is based on the use of a content-based publisher/subscriber hub - Gryphon [18], is shown in Fig.3. All of the data sources are connected to the pub/sub hub as publisher so that the data (numeric message, text, audio, or video) from these sources can be routed through the hub to those subscribers that subscribe to these sources. All of the detectors are attached to the system as subscribers as well as publishers, so that they can subscribe to a number of data sources as well as the output from other detectors based on the topics of the data sources.

Note that each of the detectors within the system (as shown in Fig.3) may generate alerts based on the specific charter of the detector. There is also system level alert generation that fuses the alerts generated from other detectors. The system level alert generation uses alerts generated by both passive and active detectors, as shown in Fig.4.

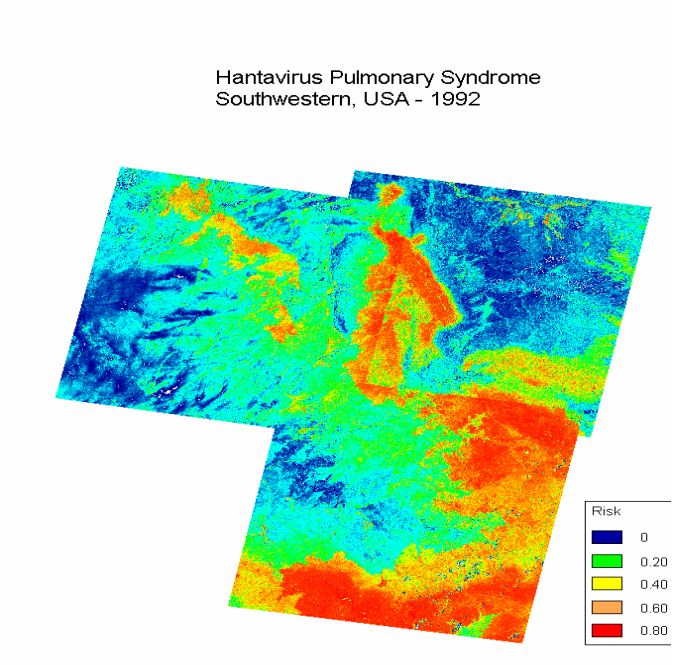

Fig.5. Risk map for Hantavirus Pulmonary Syndrome during 1992

5 Model-Based Data fusion and Detection

A number of modeling techniques have been developed in this system to model the spatio-temporal risk factor to certain infectious diseases (HPS, influenza, Denge fever, and anthrax). A linear time-invariant model, $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$, has been used to model the HPS, where each X_i represents the data itself or derived attributes/features from the multi-modal information sources, while the coefficient a_i represents the weights (relative contribution) of the attribute derived from the data. More specifically, the risk assessment model for the risk to HPS associated with a location (x,y) is:

$$R(x,y) = 0.443X_1 + 0.222X_2 + 0.153X_3 + 0.183\ X_4,$$

where X_1, X_2, and X_3 correspond to the pixel value of band 4, 5 and 7 of Landsat Thematic Mapper image at location (x,y), while X_4 corresponds to the elevation (in meters) from the corresponding DEM (digital elevation map). A risk map based on this model for the south western US during the summer of 1992 is shown in Fig.5. The actual HPS outbreak took place in 1993 with more than 85% of the cases occur within those highest risk areas. In addition to the linear model, finite state machine models have been successfully developed and applied to modeling the risk to fire ants (which are harmful to both crops and livestock of the southeast US), and Bayesian network models have been developed for other infectious diseases.

The same model for data fusion can also be used for indexing to facilitate model-based information retrieval. A model-based indexing technique, Onion [19], was developed for linear model based data fusion and retrieval and provide up to three order-of-magnitude speedups as compared to linear evaluation.

The risk map generated above provides the baseline for anomaly detection – as we are usually only interested in unexplainable anomalies. We have explored two general classes of model-based anomaly detectors (Fig.3 and 4) that have applicability to site surveillance. The first class, which we term differential detectors, is applicable in the case where there are two or more sites that have similar behaviors. A differential detector raises an alarm when the deviation between sites becomes sufficiently large. The second class of detectors is predictive, i.e., they predict “normal” site behavior and raise an alarm if a sufficiently large deviation from normal is detected.

6 Validation

The Epi-SPIRE system has been validated in a genuine environment between the fall 2001 and summer of 2002 to monitor the behavioral changes of a population caused by the earliest stages of illness. Examples of such behaviors include increased absenteeism, increased inquiries for medical information, changes in eating/drinking habits, increased coughing, increased traffic for leaving the building early, and increased sewage generation. IBM T. J. Watson Research Center, which consists two sites - Yorktown and Hawthorne, and is located in Westchester County, NY (50 km north of New York City), is used in this case study. The total population for the sites is approximately 2000. All of the data collected below have been properly anonymized so that the privacy of the population being investigated is not violated.

1) A weekly survey of self-reported health level was conducted from January 2002 through May 2002, during which an email-based survey of the population was run at the Watson site. About 400 IBM employees volunteered to participate. This survey had an excellent response rate: 92% of polled employees responded the same day, 73% by noon.
2) The IBM Watson worksite requires the swiping of a badge in order to gain entry. The badge number and time of entry are recorded in a database that is maintained for security purposes. We have been receiving an anonymized version of this information since 12/2001.
3) The IBM Watson site records, for billing purposes, all phone calls made outside the site. The calling number, called number, time of call, and duration of call are recorded in a database. A set of local medically related phone numbers was obtained from two main sources (scanned from yellow pages, internet directories). From an anonymized version of these data it is possible to count the number of calls made from Watson to medically related numbers, as well as the number of extensions that were used to place these calls.
4) The IBM Watson site records, for security purposes, all accesses to external websites at the firewall. The source IP, destination IP, and date/time of access are recorded in a database. Using an anonymized version of this database along with a manually generated list of medically related websites, it is possible to count the number of accesses to these medically related sites, as well as the number of computers from which these requests were made.
5) Consumption of cafeteria food and beverages at Hawthorne Cafeteria (one of the two sites for the IBM T. J. Watson Research Center) are recorded electronically. This cafeteria provides service to about 700 people.
6) A number of other potential data sources have been considered and undergone some preliminary evaluation. These include: site utility usage, site sewage generation, cough counting, and car counting (cars entering or leaving site).

The alerts generated from these data sources are compared to the insurance claims from the Westchester County. There is preliminary evidence that the warnings generated by some of the data sources (survey and phone in particular) lead the clinical sources.

We have also evaluated the Epi-SPIRE anomaly detection mechanisms in a synthetic environment in which site-specific or regional outbreaks are simulated. The results indicate that the pathogen release can be detected within 4 days for acceptable false alarm levels.

7 Summary

In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel environmental and health activity monitoring (EAM and HAM) system that generates alerts from environmental, behavioral, and public health data sources. A model-based approach is used to develop the disease and behavior models from multi-modal heterogeneous data sources. This system has been successfully validated in a number of scenarios involving infectious disease outbreak.

Acknowledgement

* This research is sponsored in part by the Defense Advanced Research Projects Agency and managed by Air Force Research Laboratory under contract F30602-01-C-0184 and NASA/IBM CAN NCC5-305.

** The author would like to acknowledge the entire EpiSPIRE team, which includes Charu Aggrarwal, Murray Campbell, Yuan-Chi Chang, Vijay Iyengar, Mahesh Joshi, Ching-Yung Lin, Milind Naphade, John R. Smith, Belle Tseng, Min Wang, Kung-Lung Wu, Philip Yu from IBM Research Division and Gregory Glass from Johns Hopkins University.

References

[1] http://rst.gsfc.nasa.gov/Sect16/Sect16_2.html

[2] http://www.eoearth.org/article/Deforestation_in_Amazonia

[3] http://en.wikipedia.org/wiki/Deforestation

[4] http://www.emporia.edu/earthsci/student/pepper1/rainradar.htm

[5] D. O. Fuller, "Tropical forest monitoring and remote sensing: a new era of transparency in forest governance?" http://www.as.miami.edu/geography/climatology/SJTG2.pdf

[6] http://www.dartmouth.edu/~floods/

[7] http://mceer.buffalo.edu/publications/resaccom/03-sp01/09eguchi.pdf

[8] http://dirs.cis.rit.edu/research/fires.html

[9] http://activefiremaps.fs.fed.us/

[10] http://www.orbit.nesdis.noaa.gov/star/AQ_AQM.php

[11] G. E. Glass, T. L. Yates, J. B. Fine, T. M. Shields, J. B. Kendall, A. G. Hope, C. A. Parmenter, C. J. Peters, T. G. Ksiazek, C.-S. Li, J. A. Patz, and J. N. Mills. "Satellite imagery characterizes local animal reservoir populations of Sin Nombre virus in the southwestern United States," *Proceedings of National Academy of Science*, Vol.99, pp.16817-16822, December 2002.

[12] G. E. Glass, "Public health applications of near real time weather data". *Proceedings of the 6th Earth Sciences Information Partnership Conf*erence, 2001.

[13] S. L. Klein, A. L. Marson, A. L. Scott, and G. E. Glass, "Sex differences in hantavirus infection are altered by neonatal hormone manipulation in Norway rats," *Soc Neuroscience*, 2001.

[14] S. L. Klein, A. L. Scott, and G. E. Glass, "Sex differences in hantavirus infection: interactions among hormones, genes, and immunity," *Am Physiol Soc*., 2001.

[15] G. E. Glass, "Hantaviruses. Climate Impacts and Integrated Assessment," Energy Modeling Forum 2001.

[16] http://www.sanborn.com/solutions/regional_fire_risk_assesment.asp

[17] http://activefiremaps.fs.fed.us/recent3.php

[18] S. Bhola, R. Strom, S. Bagchi, and Y. Zhao, "Exactly-once Delivery in a Content-based Publish-Subscribe System," *Dependable Systems and Networks*, 2002.

[19] Y.-C. Chang, L. D. Bergman, V. Castelli, C.-S. Li, M.-L. Lo, and J. R. Smith, ``The Onion Technique: Indexing for Linear Optimization Queries,'' *Proceedings of ACM SIGMOD 2000*, May 2000.

[20] B. L. Tseng, C.-Y. Lin, and J. R. Smith, "Real-Time Video Surveillance for Traffic Monitoring Using Virtual Line Analysis," *Proceedings of IEEE ICME*, Lausanne, Switzerland, August 2002.