BioSTORM: A System for Automated Surveillance of Diverse Data Sources

Martin J. O'Connor, M.Sc., David L. Buckeridge, M.D, M.Sc., Michael Choy,
Monica Crubezy, Ph.D., Zachary Pincus, Mark A. Musen, M.D., Ph.D.

Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA

Background. Heightened concerns about bioterrorism are forcing changes to the traditional biosurveillancemodel. Public health departments are under pressure to follow multiple, non-specific, pre-diagnostic indicators, often drawn from many data sources. As a result, there is a need for biosurveillance systems that can use a variety of analysis techniques to rapidly integrate and process multiple diverse data feeds using a variety of problem solving techniques to give timely analysis. To meet these requirements, we are developing a new system called BioSTORM (**Bio**logical **S**patio-**T**emporal **O**utbreak **R**easoning **M**odule).

System Description. The main goals of BioSTORM are: (1) integration of multiple data sources; (2) scalability; (3) responsiveness; (4) integral support of spatial and temporal analysis; (5) support for diverse problem solvers; and (6) flexible configuration support.

BioSTORM provides (1) a data broker, which integrate disparate data sources into a semantically uniform data stream; (2) a data mapper that tailors this data stream to the needs of individual problem solvers; and (3) a control structure, which manages the data flow in the system and governs the deployment of problem solvers; (4) a library of statistical and spatial problem solvers.

Data Broker Raw data are diverse and distributed in various databases and files with little common semantic structure. Thus, they can be difficult to integrate. The data broker uses two components to facilitate data integration (1) a data source ontology (Pincus, 2003); and (2) a software library that uses the ontology to access data. The data source ontology provides a semantic structure for raw data. It has an explicit vocabulary for describing attributes of data and data sources, providing a framework for relating data to shared semantics. The data broker queries the data source ontology to construct a stream of uniform data objects that conform to shared semantics.

Mapper The data mapper takes the data stream provided by the data broker and reduces, abstracts and transforms it to a format meaningful to each problem solver, allowing them to ignore the original sources or formats of data. Each problem solver must publish an input–output ontology describing the data that it wishes to receive. The mapper can then provide each problem solver with a customized set of data object by using mapping relations between a data stream and the problem solvers. For example, a relation can specify that data be aggregated at different spatial or temporal granularities.

Control Structure The control structure coordinates the flow of data from raw data, through the broker and mapper to problem solvers (O'Connor, 2003). It is responsible for invoking problem solvers when data arrive and for feeding them new arriving data through both the broker and the mapper. Its overall task is to unify the broker, mapper, problem solvers, and knowledge bases into a coherent, efficient runtime system. It provides a modular framework to enable concurrent application and structured evaluation of multiple analytic methods.

Problem Solvers The system employs a library of problem solvers, which are classified according to the type of analysis they perform (Buckeridge, 2003). This classification allows the control structure to identify appropriate problem solvers for a specific analysis task and then map the required data and knowledge to the problem solvers. This library currently contains statistical, spatial, temporal, and knowledge-based problem solvers.

Evaluation. We have evaluated the functionality of an end-to-end prototype with a number of statistical and spatial problem solvers using EMS data from San Francisco. This required (1) describing the EMS data in terms of our data source ontology, (2) using the data broker to convert the raw data into a semantically uniform data stream; (3) using the mapper to convert and abstract this data stream into the input required by the problem solvers; and (4) deploying the problem solvers to process their input. The problem solvers identified the annual influenza epidemic as validated by data from the California Influenza Surveillance Project.

Conclusion. The BioSTORM systems integrates disparate data sources into semantically uniform data streams, maps these streams to multiple problem solvers, and deploys these problem solvers to conduct surveillance.

References.

O'Connor *et al*. A Knowledge-Based Approach to Deploying Problem Solvers. *AMIA Annual Symposium* November, 2003.

Buckeridge *et al*. A Modular Framework for Automated Space–Time Surveillance Analysis of Public Health Data. *AMIA Annual Symposium* November, 2003.

Pincus *et al*. Contextualizing Heterogeneous Data for Integration and Inference. *AMIA Annual Symposium* November, 2003.