Summary

_Biosurveillance_ denotes the task of monitoring for the occurrence (outbreak) of diseases of public-health concern, particularly infectious diseases. The general goal of biosurveillance is to detect the outbreak of such diseases as early as possible, while minimizing the number of false positive alerts that are raised when there is no outbreak.

This project has concentrated on investigating Bayesian biosurveillance methods. This section briefly summarizes the contributions of this project within the discipline, which includes biosurveillance in particular and anomaly detection more generally. Sections 1 and 2 of this report (see below) describe our results in detail; the interested reader can see those sections for information that supports the summary statements made in the present summary section.

We can divide our contributions into two primary areas: new biosurveillance methods per se, and new developments related to biosurveillance. We discuss each area in turn, and then we close with an outline of our future directions.

_Biosurveillance Methods_

We developed a number of new algorithms for biosurveillance that performed well compared to state-of-the-art methods. Most of these evaluations where based on using semi-synthetic data, in which simulated outbreak cases (e.g., patient cases due to windborne anthrax) were added to real non-outbreak cases. We did, however, perform some evaluations using data on real outbreaks, including real, documented influenza and cryptosporidiosis outbreaks. Overall, we found that the new biosurveillance algorithms that we developed performed well, compared to several state-of-the-art algorithms, both in terms of detection performance and run time. In the remainder of this subsection, we summarize the main areas we investigated in developing the new biosurveillance methods.

Bayesian methods require prior probabilities that express prior belief about outbreaks before the data are considered. We developed new methods for specifying such prior probabilities. One method, which we call _semi-quantitative_, allows the model builder to express constraints on the
relationships between a disease and the appearance of one of its finding, without having to completely commit to the detailed quantitative relationship between them. Such assessments are sometimes much easier and practical to obtain.

Another complementary method for determining model parameters uses machine learning. We investigated the use of existing training data and the use of dynamic user feedback to develop models of the likely size, shape, and location of outbreak regions, as well as how outbreaks are likely to emerge over time.

We devoted a significant effort to developing new spatial, temporal, and spatio-temporal algorithms for biosurveillance. We investigated new methods for representing how disease outbreaks emerge over space, over time, and over space-time. We often found that historical data, during which there were no outbreaks, served as a very informative baseline with which to model and detect emerging outbreaks. We devoted considerable effort to the development of algorithms that were efficient enough to be practical in real time. As one example, we developed a Bayesian scan statistic method that is many times faster than the comparable frequentist algorithm, because the Bayesian algorithm avoids randomization testing.

Traditional approaches to biosurveillance are based on using as input the counts of some stream of data, such as the number of respiratory chief complaints of patients who visit a emergency department (ED) during a 24-hour period. We explored novel algorithms for count-based biosurveillance. We also investigated algorithms that are person-based in the sense that each person in the population is modeled to some degree of detail, depending on the data available. Such person-based models are particularly useful when we have rich data about at least some individuals in the population. We found that some of the efficiency techniques we used to speed-up count-based algorithms could be used to develop efficient person-based algorithms.

*Bayesian case detection* determines the disease probability for each patient who visits a healthcare facility; it then raises a public-health alert for those patients who have a high probability for an outbreak disease. Thus, it is closely related to person-based biosurveillance. We investigated new methods for case detection, with a focus on detecting influenza.

We developed some algorithms that located the single most probable model given data, as well as other algorithms that derived an outbreak assessment by averaging over a large set of models given the data. The latter are called model-averaging algorithms, and they can sometimes be more sensitive detectors of outbreaks, particular when no one outbreak model is dominant. We developed and evaluated efficient methods for performing Bayesian model averaging for biosurveillance.
We also developed and investigated algorithms that could use more than one data stream in performing biosurveillance, such as using both emergency department data and data on over-the-counter sales of medications and products (e.g., thermometers). We found that such multiple data streams can act synergistically in detecting outbreaks sooner than either alone.

A strength of Bayesian methods is the ability to monitor for specific diseases, based on their known characteristics. A strength of frequentist methods (e.g., methods that produce p values) is the ability to monitor for unknown diseases. We developed Bayesian algorithms that are able to monitor for both known and unknown diseases. An evaluation provided support for improved Bayesian outbreak detection when both types of diseases are modeled. In a similar spirit, we developed and investigated new non-parametric methods for outbreak detection, based on count data.

In terms of biosurveillance practice, we developed, implemented, and deployed a Bayesian system called BARD for detecting windborne release of anthrax, which remains one of the most concerning bioterrorists weapons. We ran BARD for several years to actively monitor in real time for anthrax-disease outbreaks in Pennsylvania.

**Additional Developments**

Evaluating biosurveillance systems is a challenging and evolving area of work. The AMOC curve is one of the most commonly used measures of biosurveillance-system performance. For a given algorithm, it plots the expected time to outbreak detection versus the expected false-alert rate. The AMOC-curve method is premised on the output of an algorithm (e.g., the posterior probability of an outbreak) crossing a given detection threshold. Such behavior, however, might not correspond to how a public-health epidemiologist would use the output in detecting an outbreak. He or she might require that the output signal be above a threshold for at least two consecutive days, for example, before seriously considering the possibility of an outbreak. We developed a generalization of the AMOC curve, called the G-AMOC curve, and explored its properties as a measure of disease outbreak detection.

Most biosurveillance algorithms output a signal that indicates the chance that there is an ongoing outbreak. What we really wish to detect, however, are outbreaks that are subtle enough to be missed by public-health officials and clinicians. To do so, we need to additionally model the extent to which a given outbreak is likely to be detected by humans. We developed and investigated new Bayesian methods that are focused specifically on searching for and detecting such subtle outbreaks.

Public health epidemiologists are unlikely to trust biosurveillance alerts blindly. They will want explanatory justifications. Producing such explanations could be important to the long-term
success of biosurveillance systems. Automated explanation of Bayesian inference remains a largely unexplored area of research. We have developed methods for explaining the posterior probabilities that are output by a Bayesian biosurveillance system. We plan to evaluate these methods in a randomized, controlled study in early 2010.

**Future Directions**

In this project, we developed a number of promising Bayesian methods for biosurveillance, as originally proposed. The project goals did not, however, include deploying those methods for actual day-to-day use by public-health epidemiologists. Nonetheless, many of the Bayesian methods we developed have formed the basis for a recently funded CDC Center of Excellence five-year project at the University of Pittsburgh (CDC grant 1P01-HK000086-01), which is specifically focused on refining, deploying, and extensively evaluating Bayesian biosurveillance methods that will be used in the field. Thus, the methods developed in this NSF project are being taken in a logical direction of further development. Our goal is for the Bayesian biosurveillance system that is developed and deployed in this CDC project to serve as a state-of-the-art prototype that will be adapted and deployed throughout the country.

Anomaly detection is a generalization of biosurveillance. In anomaly detection, the goal can include detecting many different types of unusual patterns in data. We have recently received NSF funding for a project that is focused on developing, implementing, and evaluating largely Bayesian methods for discovering complex anomalous patterns in data (NSF IIS-0911032 to CMU (primary) and University of Pittsburgh (subcontract)). The domains of application include, for example, the detection of anomalous patterns in the maintenance of a fleet of aircraft, which might indicate problems with the aircraft or the maintenance procedures. This project will leverage many of the insights and algorithms developed in this Bayesian biosurveillance project, and it represents another natural direction in which to take the research results of the Bayesian biosurveillance project. The goals of the new project are to develop novel methods that will be highly effective in discovering complex patterns from data in many different domains, not just in biosurveillance. We plan to make the algorithms available as open-source software with the intention of advancing the science and engineering of anomaly detection in many different fields of study and application.
1 Research and Education Activities

1.1 Research Activities

The goals of this project were to develop, implement, and evaluate new Bayesian methods for detecting outbreaks of disease. We developed and investigated a variety of novel methods for modeling, inference, explanation, and learning, as planned in the original project proposal. Section 1 of this final report summarizes the main methods that we developed in the project, as well as the primary educational accomplishments. Section 2 summarizes the main findings and results that we obtained from evaluating these methods. Since this has been a five-year project (with an additional one-year no-cost extension), many methods have been developed and results have been obtained, which accounts for the length of this document.

1.1.1 A Bayesian Network Model for Spatio-Temporal Event Surveillance

Event surveillance involves analyzing a region in order to detect patterns that are indicative of some event of interest. An example is the monitoring of information about emergency department visits to detect a disease outbreak. Spatial event surveillance involves analyzing spatial patterns of evidence that are indicative of the event of interest. A special case of spatial event surveillance is spatial cluster detection, which searches for subregions in which the count of an event of interest is higher than expected. Temporal event surveillance involves monitoring for emerging temporal patterns. Spatio-temporal event surveillance involves joint spatial and temporal monitoring.

When the events observed are of direct interest, then analyzing counts of those events is generally the preferred approach. However, in event surveillance we often only observe events that are indirectly related to the events of interest. For example, when monitoring for an influenza outbreak, it is an outbreak of influenza that is the event of interest. We generally cannot directly observe the influenza status of individuals, but must infer that status under uncertainty from evidence, such as the chief complaints of patients who have recently visited emergency departments. In this situation, a better surveillance approach may be to model the relationships among the events of interest (e.g., influenza) and those observed (e.g., chief complaints).

We developed a high-level Bayesian network architecture that represents a class of spatial event surveillance models. We also developed an architecture that represents a class of temporal event surveillance models. We combined these Bayesian network architectures into a single architecture that represents a class of spatio-temporal models. Using these architectures, it is often possible to construct a temporal, spatial, or spatio-temporal model from an existing Bayesian network event-surveillance model that is non-spatial and non-temporal. Our general hypothesis was that when a non-spatial and non-temporal surveillance model is extended to incorporate space and time, event surveillance will be improved significantly.
PANDA-CDCA (PC) [Cooper 2006; Jiang 2008a] is a non-temporal, non-spatial disease outbreak detection system that we previously developed. We extended PC both spatially and temporally. Our specific hypothesis was that each of the spatial and temporal extensions of PC will perform outbreak detection better than does PC, and that the combined use of the spatial and temporal extensions will perform better than either extension alone. We evaluated these hypotheses using semi-synthetic data and found that the experimental results supported the hypotheses, as reported in more detail in the Findings and Results section of this report.

1.1.2 A New Prior for Bayesian Anomaly Detection

Bayesian anomaly detection computes posterior probabilities of anomalous events by combining prior beliefs with evidence from data. However, the specification of prior probabilities can be challenging. We developed a new Bayesian prior in the context of disease outbreak detection. We wanted it to be intuitive, easy-to-use, and yield a posterior probability of an outbreak that performs at least as well as a standard frequentist approach. By doing so, the resulting posterior could be usefully incorporated into a decision analysis about how to act in light of a possible disease outbreak [Shen 2009a].

We used the new prior to develop a univariate outbreak detection algorithm called BU. The time complexity of BU is linear in the number of the observed events being monitored, due to a novel, closed-form derivation that we introduced. The Finding and Results section describes the positive results we obtained from evaluating this new algorithm.

1.1.3 Bayesian Modeling of Anomalies Due to Known and Unknown Causes

We investigated Bayesian modeling of known and unknown causes of events in the context of disease-outbreak detection. We developed a Bayesian approach that models and detects (1) known diseases (e.g., influenza and anthrax) by using informative prior probabilities, and (2) unknown diseases (e.g., a new, highly contagious respiratory virus that has never been seen before) by using relatively non-informative prior probabilities. We call this Bayesian approach the Bayesian Hybrid (BH) detection algorithm. A key contribution of this research is that it introduces a Bayesian approach for jointly modeling both known and unknown causes of events. Such modeling has broad applicability in artificial intelligence in general, where the space of known causes of outcomes of interest is seldom complete [Shen 2009b; Shen 2009c].

1.1.4 Explanation of Inference in Bayesian Biosurveillance

The importance of an explanation facility in intelligent systems was recognized early on. There are several studies that experimentally confirmed the positive impact of explanation on user perception of the system and on the accuracy of decision making. A Bayesian biosurveillance system outputs the probability of an outbreak and possibly additional probabilities for the type of outbreak and the size of the outbreak. Users of such a system, such as public health epidemiologists, are plausibly more likely to view such probabilistic output as credible if it is accompanied by an explanation. Explanation may help the users of such systems to make more informed decisions.
We developed a novel method for explaining Bayesian network (BN) inference on a network that represents a population of individuals, each of whom is modeled as a subnetwork [Sutovsky 2008]. The method complements previous approaches for explaining inference on BNs that model a single agent (e.g., explaining the diagnostic inference for a single patient using a BN that models just that patient).

An outcome of inference in a BN is a posterior probability distribution over variables, given observed evidence. Hence the explanation of inference in a Bayesian network is focused on the posterior distribution of a node of interest, such as a node representing whether an outbreak is present or not. Methods that explain BN inference try to clarify why and how a certain posterior probability was obtained given observed evidence. For example, consider evidence about patients, such as their symptoms, and suppose that a BN system infers that a respiratory anthrax outbreak is highly likely. A public-health epidemiologist who received such a report would generally want to know why anthrax is being given a high posterior probability.

The posterior probability that we want to explain results from a combination of several factors: the evidence, the BN structure (variables and arcs), and the BN parameters (local conditional probabilities), and inference. Information about these factors and how they combine could be included in an explanation. The calculation of a posterior probability, however, involves many calculations with many numbers. Moreover, the evidence can consist of many findings. An explanation that simply lists all this information is unlikely to be very useful. We developed methods for selecting and organizing information in order to provide an effective explanation of BN inference during outbreak detection. We have also developed a detailed experimental design for evaluating the proposed explanation method. We are currently finalizing the implementation of an explanation method which we plan to evaluate within the next three months.

1.1.5 Bayesian Aerosol Release Detector (BARD)

BARD is a Bayesian approach that we developed for detecting disease outbreaks that are due to windborne release of infectious agents [Hogan 2007]. Our progress on developing BARD is summarized in this section.

First, we have analyzed the results of a simulation study that compares the detection performance of BARD to that of other detection algorithms. Second, we have extended BARD to model multiple days of data to better exploit the temporal characteristics of anthrax outbreaks [Kong 2008]. Third, we completed the development of an expanded BARD model that accounts for commuting to and from work, and we have published two papers on the BARD commuting model. One paper develops the BARD commuting model for use in simulating aerosol releases of anthrax [Cami 2008a]. The second paper incorporates the commuting model into the BARD detection algorithm and evaluates the performances of the expanded detector [Cami 2008b]. Fourth, we have developed an adaptive threshold approach for improving the detection performance of algorithms, such as BARD, that repeatedly analyze time series data.

1.1.6 Extension of the Bayesian Aerosol Release Detector (BARD) Simulation Algorithm
BARD can be used for outbreak detection, as described above, as well as for simulating outbreaks due to an aerosol release, such as a release of anthrax spores. This section discusses progress with the latter use of BARD.

In collaboration with researchers at Penn Statute University Department of Meteorology, we extended the BARD anthrax-outbreak simulation algorithm to achieve more realistic anthrax release scenarios by replacing a previous Gaussian plume model with a state-of-the-art dispersion model, namely the Second Order Closure Integrated PUFF model (SCIPUFF), and by using weather data computed from a 5th generation mesoscale model (MM5) [Sahin 2009]. SCIPUFF is a sophisticated atmospheric dispersion model, currently used in the Hazard Prediction and Assessment Capability (HPAC) software tool (one of the Defense Threat Reduction Agency applications) for estimating anthrax dosage amounts at specified locations under given weather conditions.

1.1.7 Rank-Based Spatial Clustering (RSC) Framework

The Rank-based Spatial Clustering (RSC) framework [Que 2009] is a general framework using a non-parametric spatial clustering approach, which can employ either a time series algorithm computing Z-scores of a zip code or a Bayesian algorithm computing the probability of a zip code having an outbreak, to identify an outbreak region. This framework achieves rapid, early detection of disease outbreaks composed of a cluster of geographic regions such as zip codes or counties. The RSC framework is derived from the Multiscale Spatial Clustering (MSC) algorithm we developed previously [Que 2008a; Que 2008b].

We developed two approaches in MSC to identify the zip codes with higher risk of having an outbreak: the Bayesian method [Que 2009] and the Z-score method [Tsui 2009]. We also compared the detection performance of MSC using Bayesian and Z-score methods with the current state-of-the-art spatial cluster algorithms, namely, the Kulldorff Spatial Scan Statistics (KSSS) algorithm and Bayesian Spatial Scan Statistics (BSSS) algorithm.

We developed an RSC framework that automatically aggregates adjacent zip codes based on score rank, identifies clusters with various sizes at different geographic locations, and computes a posterior probability of a cluster having an outbreak [Que 2009]. The previous MSC approach requires the use of different thresholds for selecting zip codes whereas the RSC algorithm does not. We also compared the algorithms using the RSC framework (RSC\textsubscript{std\_score} and RSC\textsubscript{Bayesian}) with the KSSS and BSSS algorithms.

We are in the process of testing the scalability, computational efficiency, and detection performance of the RSC framework by processing over-the-counter sales in the entire nation, which is a major challenge given the amount of data and the large area. We currently are refactoring our RSC code to optimize its computational performance.
1.1.8 Bayesian Case Detector (BCD)

We developed a system called the Bayesian Case Detector (BCD) for detecting individual cases of an outbreak disease. We deployed BCD at the University of Pittsburgh Medical Center (UPMC) as a production service in response to recent H1N1 influenza (swine flu) concerns [Tsui 2009]. The BCD system automatically processes Emergency Department (ED) reports from ten UPMC hospitals on a daily basis, identifies fever and Influenza-Like-Illness (ILI) cases, and provides daily summary reports of fever and ILI cases to the Allegheny County Health Department through emails.

1.1.9 Combining Multiple Data Streams

Two data sources that are routinely monitored by syndromic surveillance systems are over-the-counter (OTC) medication sales and Emergency Department (ED) chief complaint records. One way to combine the advantages of both types of data is to develop a detection algorithm that integrates the data sources. If both data sources are being jointly monitored, then the combined information could reinforce our belief that an outbreak is happening. The key difficulty with this data fusion approach is in measuring the relationship between the data sources when an outbreak occurs. The correlation between OTC and ED data during outbreak conditions cannot be learned because no training data exists that captures the effects of a large-scale epidemic on these data sources. Although training data do not exist, we do have some background knowledge about the plausible relationship between OTC and ED data for a given disease. Our approach to combining multiple data streams relies on using this background knowledge to model the actions of individuals that result in possible OTC medication purchases and ED admissions. Details about this method are described in [Wong, 2005].

1.1.10 Evaluation of Joint Clinician and Algorithm Detection of Disease Outbreaks

The goal of automated biosurveillance systems is to detect disease outbreaks early, while exhibiting few false positives. Evaluation measures currently exist to estimate the expected detection time of biosurveillance systems. Researchers also have developed models that estimate clinician detection of cases of outbreak diseases, which is a process known as clinical case finding. However, little research has been done on estimating how well biosurveillance systems augment traditional outbreak detection that is carried out by clinicians. We developed a general approach for doing so for non-contagious disease outbreaks, which are characteristic of many possible bioterrorist induced diseases, such as respiratory anthrax [Shen 2008]. Our results support that such analyses are useful in assessing the extent to which computer-based outbreak detection systems are expected to augment traditional clinician outbreak detection.

1.1.11 Complete Bayesian Model Averaging Over Spatial Distributions of Disease

We developed a Bayesian disease-outbreak detection algorithm for monitoring a wide variety of spatial distributions of a disease outbreak [Shen 2007]. The algorithm performs complete
Bayesian model averaging (BMA) over all possible spatial distributions of disease in polynomial time (relative to a finite set of spatial primitives). We compared the detection performance of this BMA algorithm to an existing outbreak-detection algorithm on two different spatial patterns of disease outbreaks. Preliminary results support that the two algorithms are complementary in the types of disease outbreaks that they detect.

1.1.12 A Recursive Bayesian Spatial Outbreak Detection Algorithm

When doing spatial event surveillance, we must first have a representation of the subregions of the region being analyzed. A simple approach is to represent the entire region by an $n \times m$ grid. Then we let every subset of grid elements represent a subregion. Unless $n$ and $m$ are small, we cannot check every possible subregion. Thus, many researchers perform spatial event surveillance by only checking every subregion which is a rectangle, ellipse, or some other simple shape. We developed and implemented a recursive spatial outbreak detection algorithm that searches a much richer set of subregions than such simple shapes [Jiang 2007].

1.1.13 Generalized AMOC Curves for Evaluation and Improvement of Event Surveillance

We developed a new method for evaluating outbreak detection systems, which we call Generalized Activity Monitoring Operating Characteristic (G-AMOC) curves [Jiang 2009d]. G-AMOC curves provide a new approach to evaluating and improving the timeliness of disease outbreak detection by taking the user’s response protocol into account and considering when the user will initiate an investigation in response to the system’s alerts. The standard AMOC curve is a special case of G-AMOC curves that assumes a trivial response protocol (initiating a new and separate investigation in response to each alert signal). Practical application of a surveillance system is often improved, however, by using more elaborate response protocols, such as grouping alerts or ignoring isolated signals. Our experiments demonstrate that we can use G-AMOC curves as 1) a descriptive tool, to provide a more insightful comparison of systems than the standard AMOC curve, and 2) as a prescriptive tool, to choose appropriate response protocols for a detection system, and thus improve its performance [Jiang 2009d].

1.1.14 Detection of Anomalous Records and Patterns in General Datasets

We have developed a method for finding individually anomalous records [Das 2007]. The key insight in this work is that we can achieve better accuracy by abandoning the "whole-record-likelihood" approach and instead considering Bayesian models independently for all subsets (up to some size) of the variables. This frees us from having to do structure learning (e.g., as done with traditional Bayesian networks) and thus frees us from the mistakes that can result from it. The new algorithm uses several computational tricks including AD-Trees to make it efficient.

We have developed two methods for detecting anomalous patterns (i.e., sets containing more than one record). The first method, Anomalous Group Detection (AGD) [Neill 2009d; Das 2007] is a generalization of the spatial scan statistic, a commonly used method for detecting clusters of increased counts in spatial data. We extended this framework to non-spatial datasets with categorical attributes, where the degree of anomalousness of each record depends on its attribute
values, and we wish to find groups of similarly anomalous records. We model the relationship between the attributes using a Bayesian network probability distribution model, define a likelihood ratio statistic in terms of the pseudo-likelihoods for the null and alternative hypotheses, and maximize this statistic over all subsets of records. Since an exhaustive search over all such groups is computationally infeasible, we propose an efficient (but approximate) search heuristic.

Our second method, Anomaly Pattern Detection (APD) [Das 2009; Das 2008a; Das 2008b] assumes that anomalies are generated by some underlying process which affects only a particular subset of the data. Our method consists of two steps: we first use existing anomaly detection methods to identify individual records with anomalous attribute values, and then detect patterns where the number of anomalous records is higher than expected. Given the set of anomalies flagged by the local anomaly detector, we search over all subsets of the data defined by any set of fixed values of a subset of the attributes, in order to detect self-similar patterns of anomalies. We wish to detect any such subset of the test data that displays a significant increase in anomalous activity as compared to the normal behavior of the system (as indicated by the training data). We perform significance testing to determine if the number of anomalies in any subset of the test data is significantly higher than expected, and propose an efficient algorithm to perform this test over all such subsets of the data.

1.1.15 Rapid Detection of Multidimensional Rectangular Clusters

Assume a uniform, multidimensional grid of bivariate data, where each cell of the grid has a count $c_i$ (e.g., number of disease cases of a given syndrome type in a cell subregion) and a baseline $b_i$ (e.g., population at-risk, or expected count estimated from historical data). Our goal is to find spatial regions (d-dimensional rectangles, representing a contiguous range of values for each dimensions) where the counts are significantly higher than expected given the baselines. This problem can be solved using a spatial scan statistic (Kulldorff, 1997), where we compute the maximum of a likelihood ratio statistic over all spatial regions, and find the significance of this region by randomization. However, computing the scan statistic for all spatial regions is generally computationally infeasible, so we introduced a novel fast spatial scan algorithm [Neill 2004a; Neill 2004b; Neill 2005a] for two-dimensional data and generalized this approach to multi-dimensional data [Neill 2005b]. The fast spatial scan approach incorporates a branch-and-bound search using a new multi-resolution “overlap-kd tree” data structure. Our multidimensional, multi-resolution algorithm allowed us to find spatial clusters up to 1400x faster than the naive spatial scan, without any loss of accuracy, making large-scale disease surveillance computationally feasible.

1.1.16 Expectation Scan Statistic and Detection of Emerging Space-Time Clusters

Our expectation-based scan statistic [Neill 2005c; Neill 2006a; Neill 2007a; Neill 2009a; Neill 2009b] is a method for spatio-temporal event detection using a single data stream. It consists of two main steps: first computing the expected counts for each spatial location and each time step by time series analysis, then finding spatial regions (sets of locations) where the observed counts for recent time steps are significantly higher than expected. Spatial regions are detected by scanning over a huge set of potential regions, maximizing a likelihood ratio statistic, and
computing statistical significance by randomization testing. We have developed a number of variants of this method, which differ both in the underlying statistical model (e.g., Poisson [Neill 2005c], Gaussian [Neill 2006a], and robust Bernoulli-Poisson [Neill 2007a]), as well as the method of time series analysis used to predict the expected counts. We demonstrated that the robust scan statistic outperforms non-robust alternatives in datasets where many outlier counts are present [Neill 2007a]. We have also developed an “emerging cluster” variant of the Poisson expectation-based scan statistic [Neill 2005c], which achieves higher detection power when an outbreak emerges slowly over time. We have demonstrated [Neill 2009a] that this method of aggregating information across multiple time series, rather than monitoring each series separately, improves the accuracy, timeliness, and spatial resolution of detection. We compared multiple variants of the expectation-based method on the disease surveillance task (detecting simulated disease outbreaks injected into real hospital Emergency Department data), answering the questions of which set of space-time regions to search, which time series analysis method to use for computing expectations, and which model to use for detection. We also compared the standard expectation-based scan to more complex expectation-based models (including Gaussian statistics that learn the expected counts and variances from historical data, and robust statistics which avoid detecting irrelevant outliers) over a wide range of datasets and outbreak types [Neill 2009a; Neill 2009b].

1.1.17 Nonparametric Scan Statistics

We developed a general methodology for spatial event detection, which (unlike standard model-based event detection approaches) makes no assumptions about the underlying parametric distribution of the data [Neill 2007b]. This nonparametric scan statistic approach, because of the generality of its model assumptions, has significant potential to be applied to a wide variety of pattern detection problems. As discussed above, our expectation-based scan statistics attempt to model the expected distribution of counts for each spatial location under the null hypothesis of no clusters, then find regions where the counts are higher than expected. The simple expectation-based statistics method assumes that the counts are generated by some parametric model, and then it learns the parameters of this model, typically from historical data. For example, the Poisson statistic learns the baseline (expected count) $b_i$ for each spatial location, while the Gaussian statistic learns both the expected count and the expected variance. The disadvantage of these model-based approaches is that they rely heavily on our distributional assumptions: for example, the Poisson statistic cannot account for overdispersion or underdispersion of counts, and neither Poisson nor Gaussian statistics can account for heavy-tailed count distributions.

Our nonparametric scan statistic approach makes no model assumptions on the distribution of counts, but instead uses the empirical distribution of historical counts for each spatial location. Let us assume that we have a count $c_i$ and a time series of historical counts for each spatial location $s_i$. We also make three simplifying assumptions: that the historical data contains no relevant clusters, the time series of counts for each location is stationary, and that counts are uncorrelated. Then under the null hypothesis of no clusters, we expect that the current count $c_i$ for each location will be drawn from the same distribution as the historical counts for that location. Thus we can define the empirical p-value $p_i$ for each spatial location to be the ratio $(T_{\text{beat}} + 1) / (T + 1)$, where $T_{\text{beat}}$ is the number of historical counts larger than $c_i$, and $T$ is the total number of historical counts. Under the null hypothesis, and given the assumptions above, each
empirical p-value will be asymptotically uniformly distributed on the interval \([0, 1]\). We wish to
detect regions \(S\) where the counts \(c_i\) are higher than expected, and thus where the \(p_i\) are lower
than expected. The general form of the nonparametric scan statistic is \(\max_\alpha F(N_\alpha, N, \alpha)\), where \(F\)
is a function comparing the observed number \((N_\alpha)\) to the expected number \((N \times \alpha)\) of p-values
less than \(\alpha\).

### 1.1.18 Multivariate Bayesian Scan Statistics

The *Bayesian spatial scan statistic* [Neill 2006b] is an extension of the spatial scan statistic
(Kulldorff, 1997) that combines observations and prior knowledge in a principled Bayesian
framework. We have demonstrated that the Bayesian method has several advantages over the
standard frequentist approach, including higher detection power, faster computation (since
computationally expensive randomization testing is unnecessary in the Bayesian framework),
and easier visualization and calibration. In [Neill 2007c; Neill 2009c] we extended the Bayesian
scan statistic to a multivariate pattern detection methodology, the *multivariate Bayesian scan
statistic* (MBSS). MBSS allows us to achieve faster and more accurate detection of emerging
patterns by combining multiple streams of spatio-temporal data (for disease surveillance, these
could include ED visits with different chief complaint categories and sales of different types of
OTC medications), as well as modeling and distinguishing between different types of patterns
(e.g., distinguishing a bioterrorist anthrax attack from seasonal influenza). We demonstrated that
it has many advantages over traditional event detection approaches, including accurate
characterization and differentiation of multiple event types, and comparable detection power to
the current state of the art, even when uninformative priors are used [Neill 2009c]. Much higher
detection power can be achieved for specific event types (e.g., an anthrax attack) by providing an
informative prior describing the spatial and temporal characteristics of the event and its relative
effects on the various data streams.

### 1.1.19 Incorporating Learning into Spatial Pattern Detection

One of the major focus areas of our research was the development of a multivariate Bayesian
framework for pattern detection, the “multivariate Bayesian scan statistic” (MBSS). This method
allows faster and more accurate detection of emerging patterns by combining multiple streams of
spatio-temporal data (for disease surveillance, these could include ED visits with different chief
complaint categories and sales of different types of OTC medications), as well as modeling and
distinguishing between different types of patterns (e.g., distinguishing a bioterrorist anthrax
attack from seasonal influenza). Our ongoing work on MBSS is primarily focused on
incorporating *learning* into the event detection framework in various ways, including learning
from fully and partially labeled training data, and also learning from user feedback [Neill 2007d].
We developed a new generative model for outbreak regions, incorporated this model into the
MBSS framework, and demonstrated that the model parameters can be learned efficiently from a
small number of labeled outbreaks [Makatchev 2008a; Makatchev 2008b]. This enables learning
of priors for region size, shape, and also which locations are more or less likely to be affected
(e.g., certain outbreaks may be more likely to affect urban than rural areas, or vice-versa). In
other recent work [Das 2009], we focused on learning temporal patterns for the effects of
different event types. One novel aspect of this work is that the models are learned from partially
labeled (rather than fully labeled) data, in which the user specifies a small number of affected zip
codes and the region of effect is determined automatically. Additionally, we can learn models which incorporate time lags between affected regions. We are in the process of extending this work to learn joint models of the spatial and temporal effects of an outbreak, allowing more timely and accurate detection of clusters of events that can grow, shrink, or move over time.

1.1.20 Fast Subset Scanning for Pattern Detection

We have recently discovered [Neill 2008] that many pattern detection methods, such as the original and nonparametric spatial scan statistics, share the property of linear-time subset scanning (LTSS): the highest-scoring (most anomalous or most relevant) of all the exponentially many subsets of the data can be found in linear time, by sorting the data records according to some function and searching only over regions containing the $k$ highest-scoring records (letting $k$ vary from 1 to the total number of records $N$). This LTSS property can be applied directly to make certain detection tasks more efficient (for example, network intrusion detection tasks where we want to consider all subsets of one or more subnets), but only provides a solution to the unconstrained (all subsets) optimization problem if we assume a uniform prior over all subsets of the data. Most detection problems have either hard constraints (only some subsets can be considered) or soft constraints (some subsets have higher priors than others), preventing us from applying LTSS directly. However, LTSS can be used to make constrained searches more efficient in several ways. Some problems can be partitioned into multiple subset scan problems (e.g., “localized scans” search over each location and all subsets of its k-nearest neighbors), each of which can be solved rapidly using LTSS. Alternatively, the unconstrained maximum score is an upper bound on the constrained maximum score, and thus LTSS can be used as part of a branch and bound algorithm (performing an exact search, but pruning many irrelevant subsets) or can be used to provide direction for a heuristic search algorithm. These developments have open up many ways in which LTSS can enable computationally efficient event detection, removing some of the major computational barriers faced by subset scanning methods. As examples, (1) proximity-constrained LTSS performs fast spatial scanning, (2) connectivity constrained LTSS performs fast graph scanning, and similarity-constrained LTSS performs rapid detection of the most anomalous groups of records and attributes in general multivariate datasets.

1.1.21 Smart Regularization for Learning

A recurring challenge in machine learning problems is the need to regularize parameter estimates when the amount of labeled data available is limited compared to the number of parameters we want to estimate. This is a typical feature of many datasets encountered in the practice of biosurveillance. One approach to addressing this issue is to impose an L1 or L2 norm penalty on the parameters being estimated. This tends to make the resulting estimates sparse and small. A big opportunity exists to incorporate additional knowledge into the regularization structure. We have developed and investigated several methods for doing this in classification tasks including:

1. A semi-supervised approach where the correlation structure of the classification problem is learned from other, unlabeled records that need not even belong to the classes under consideration by the classifier. This results in a large improvement over current methods [Zhang 2008].
2. An approach based on exploiting known smoothness properties in the regularization. We propose various methods of penalizing large temporal and spatial changes.

3. A multi-task learning approach that penalizes parameters as in the previous three methods, but additionally penalizes parameter differences between tasks.

1.1.22 Scalability and Massive Screening of Multidimensional Surveillance Data

We have developed highly efficient surveillance algorithms by taking advantage of smart data representations, in particular our cached sufficient statistics data structure called T-Cube. We developed massive detection algorithms that are applicable to temporal transactional data in which each record is labeled with multiple descriptors. Such data is commonly encountered in public health surveillance. Examples include records of patients reporting to Emergency Departments, where each transaction (a visit) is annotated with signs, symptoms, chief complaints, demographics of the patient, and similar categorical entries. Finding the most interesting combinations of descriptors in such data is in general an extraordinarily time consuming task, given the large number of all possible combinations of descriptors and their values. We focused our attention on the temporal scan algorithm [Dubrawski 2009a] as an example event detector, and we developed its new variants allowing for the following:

1. Automated selection of the most appropriate scanning window size.
2. Hierarchical scanning: if a descriptor attribute value is used to denote an entry in a hierarchy, the scan process can be implemented to find the nodes in the hierarchy tree such that the data aggregates corresponding to them stand out the most from the background of similar aggregates obtained for their siblings or their parents in the hierarchy.
3. Retrospective and prospective scanning for anomalies. These algorithms have been successfully tested in public health, food safety, and equipment health domains [Ray 2008; Chen 2008; Dubrawski 2008a; Dubrawski 2007a; Dubrawski 2009b; Roure 2007; Roure 2008].

We developed and applied a lift statistic to predict the risk of microbial contamination of food at USDA-regulated establishments. Lift measures the increase in probability of failing a microbial test (e.g., Salmonella) given recently observed performance indicators (e.g., recorded regulatory non-compliances) relative to a prior probability of such a test failure. It is a simple but expressive statistic which has been accepted by the food safety domain practitioners [USDA 2008]. The original formulation of the lift estimation algorithm assumed binary input and output events. We have investigated multivariate alternatives, which can process and predict rates of events of interest.

1.1.23 Graph-Based Representations and Algorithms

We have developed methods for risk assessment, such as for assessing a positive outcome of a microbial test for Salmonella in light of any recent history of regulatory non-compliances at a particular USDA-regulated establishment. Those methods assumed probabilistic independence of data-generating entities (e.g., of food establishments) which allowed for the creation of universal
models built for the whole population of entities. That approach, however, ignored potentially informative characteristics of the individual entities in making predictions. An obvious alternative is to build models for each entity. However, in that case, the statistical reliability of the practically attainable models often suffers from the limited amount of data available per entity. An intuitively plausible mid-way solution is to group similar entities and infer separate models for each group. We took that approach in the above mentioned multi-task learning via regularization idea, in which building models for each cluster of entities is treated as a separate learning task, and the tasks are enabled to inform each other in order to improve the overall performance. Another approach is to represent the modeled process as a graph of entities connected via links corresponding to observed similarities, interactions and/or co-occurring events at pairs of entities.

In one of the proposed frameworks, the historical record of positive results of microbial tests for Salmonella, conducted at individual food production facilities, is used to construct a bipartite graph. Nodes of this graph represent respectively the individual facilities and the specific Salmonella serotypes. A facility node is linked with a serotype node if the particular type of Salmonella was isolated at the given plant during the period of observation. We used this model to leverage co-occurrences of the events of the same types at different establishments to benefit risk prediction accuracy. Any gain of accuracy of the proposed approach over the baseline which assumes independence of entities would indicate the potential utility of the network based approach. We implemented this method using our computationally efficient Dynamic Social Network in Latent space algorithm (DSNL) [Sarkar 2005]. It embeds the underlying graph of entities into a low dimensional Euclidean space, so that projections of the closely probabilistically related nodes in the graph are near each other in the embedded space. The evolution of the projections of the graph from one time step to the next is balanced to take into account the most recently observed graph topology and to prevent improbable large changes of mutual positioning of the entities in the embedded space. The proximity of entities in the embedded space can then be used to make predictions of potential linkages over the next time step, that is, DSNL can predict probability of occurrence of an individual Salmonella serotype at specific food factory given the microbiological history of this establishment and its peers [Sarkar 2008].

The DSNL method implemented in the way described above takes into account observed phenomenological linkages between entities. It does not take into account potentially informative similarities between them, such as -- in the case of food establishments -- their production profiles, corporate membership, geographical location, and so on. Another researched graph-based approach is able to consider such demographic properties, as well as observed linkages due to microbial test result co-occurrences. It also allows for some of the entities to be “of interest” (active), as for example when they had a recent record of a specific microbial test failure. What is the relative likelihood of each of the other entities in the graph being active? Our AFDL (Activity from Demographics and Links) algorithm is designed to answer this question in a computationally-efficient manner. AFDL is able to work with demographic data, link data (including noisy links), or both. It is also able to process very large datasets quickly [Dubrawski 2007b]. To evaluate its utility in bio-surveillance applications we used the food safety data mentioned above [Dubrawski 2008b].
1.1.24 Learning Specific Detectors of Events in Multiple Streams of Data

We developed a method for constructing detectors of events that simultaneously manifest in multiple streams of data which may carry complementary information [Roure 2007]. We investigated how specific detectors can be learned from a limited number of user-labeled events of interest (such as disease outbreaks) and how the learned detectors compare against manually designed alternatives in terms of detection power [Roure 2008]. The positive results we obtained suggest that this approach may help reduce the costs of development and maintenance of event detection systems.

1.2 Education Activities

Education of students and postdocs has been a major focus of the project. The following students have been funded to work on the project and have completed one or more graduate degrees at Carnegie Mellon University (CMU) or the University of Pittsburgh (Pitt):

Daniel B. Neill, obtained a Ph.D. in Computer Science (CMU) in 2006. He is now an Assistant Professor of Information Systems at CMU.

Anna Goldenberg, obtained a Ph.D. in the Machine Learning Department (CMU) in 2007. She is now a Postdoctoral Fellow in the Penn Center for Bioinformatics at the University of Pennsylvania.

Kaustav Das, obtained a Ph.D. in the Machine Learning Department (CMU) in 2009. He is now at Microsoft Corporation.

Xia Jiang, obtained a Ph.D. in Biomedical Informatics (Pitt) in 2008. She is now a Postdoctoral Fellow in Biomedical Informatics at Pitt.

Yanna Shen, obtained a Ph.D. in Intelligent Systems (Pitt) in 2009. She is now a Postdoctoral Fellow at the National Institutes of Health.

Sajid Siddiqi, obtained a Ph.D. in Computer Science (CMU) in 2009. He is now at Google Inc.

The following students have also been funded to work on the project and are in the process of completing their graduate degrees:

Purnamrita Sarkar, Ph.D. in Machine Learning (CMU) expected in 2010.

Peter Sutovsky, Ph.D. in Information Science (Pitt) expected in 2010.
Jeff Lingwall, expects to complete a Ph.D. in Economics and Public Policy (CMU) in 2011.

Maxim Makatchev, Ph.D. in Computer Science (CMU) expected in 2011.

Robin Sabhnani, Ph.D. in Machine Learning (CMU) expected in 2011.

Yi Zhang, Ph.D. in Machine Learning (CMU) expected in 2012.

The following postdocs were funded by and contributed to the project:

Denver Dash (Pitt). Now a Research Scientist at Intel Labs.

Weng-Keen Wong (Pitt). Now an Assistant Professor of Computer Science at Oregon State University.

Faculty on the project at CMU and Pitt have included material based on the project in a number of courses, including courses on (1) Machine Learning and Policy, (2) Data Mining and Business Intelligence, (3) Health Information Systems, and (4) Probabilistic Modeling. The faculty also developed and offered a course on Realtime Outbreak and Disease Surveillance and a graduate seminar on Biosurveillance. We offered a five-day intensive course in biosurveillance, which attracted attendees from throughout the U.S. The course provided an overview of the state-of-the-art issues in biosurveillance, including discussions about organizations and the data they collect, biosurveillance algorithms, building biosurveillance systems, and biosurveillance-related decision making. We published a textbook on biosurveillance that is appropriate for use in a graduate course [Wagner 2006]; members of this project contributed to 21 of the 25 chapters in the book. Project faculty also organized the Donald A.B. Lindberg Symposium at the University of Pittsburgh in 2008, which was devoted to the topic of Informatics in Public Health, with an emphasis on biosurveillance. The Symposium was jointly sponsored by the Department of Biomedical Informatics and the Graduate School of Public Health at the University of Pittsburgh. Approximately 100 people, including over 40 students, attended this all-day symposium in which more than half of the talks were about biosurveillance.
2 Results and Findings

2.1. A Bayesian Network Model for Spatio-Temporal Event Surveillance

We evaluated an approach to Bayesian spatial event surveillance modeling [Jiang 2009a]. Using semi-synthetic data, we compared the performance of our system and the well-known spatial scan statistic. The system we introduced outperformed the spatial scan statistic both in terms of outbreak detection and subregion detection.

We also proposed a Bayesian network architecture for a class of temporal event surveillance models [Jiang 2008a; Jiang 2008b]. Using this Bayesian network architecture, we can convert certain non-temporal surveillance systems to temporal ones. We applied this architecture to PC (see Section 1.1.1 in Research and Education Activities) and created a temporal system called PCT. PCT takes Emergency Department (ED) patient chief complaint data as its input. We compared PCT to PC using a real influenza outbreak. Furthermore, we compared PCT to both PC and a classic statistical method called CUSUM using a total of 240 influenza and Cryptosporidium disease outbreaks created by injecting stochastically simulated outbreak cases into real ED admission data. Our results indicate that overall PCT had better performance than PC in outbreak detection, and that PCT is more stable than PC in that once an outbreak is detected, PCT is better at maintaining the detection signal on future days.

We also developed a spatio-temporal version of PC called PCTS [Jiang 2008a]. We examined differences in the detection performance of PC and PCTS in terms of detection time, accuracy, and reliability. A well-known multivariate space-time scan statistic developed by Kulldorff was also included in the evaluation [Kulldorff 2007]. The study was based on 240 semi-synthetic outbreaks of influenza and Cryptosporidium. The results of these experiments show that the spatio-temporal Bayesian approach performs well, relative to the non-spatial, non-temporal approach. The comparison of the spatio-temporal Bayesian approach to the multivariate space-time scan statistic is more complex, with neither approach completely dominating the performance of the other. A paper on this research is under review [Jiang 2009c].

2.2. A New Prior for Bayesian Anomaly Detection

We used semi-synthetic data to evaluate the detection performance of the BU Bayesian algorithm and a control chart method (CC), which is a standard frequentist algorithm that is closest to BU in terms of the type of data it uses [Shen 2009a].
Let $w$ represent the baseline period (in weeks) that we use to estimate the baseline behavior. Figure 1 shows a set of AMOC curves for BU and CC, in which the baseline period $w$ is 2 to 22 weeks. Overall, BU demonstrated relatively better detection performance than CC, and BU’s performance was relatively strongest when using baseline periods in which $w = 6$ to 10 weeks.

**Figure 1.** AMOC curves of BU and CC from using $w = 2, 6, 10, 14, 18,$ and $22$ weeks, respectively. In each curve, the marker on the curve shows the expected detection time at a given false positive rate, and the bar shows the 95% confidence interval of the expected detection time.
At zero and one false positive per month, BU has an expected detection time that is less than that of CC over all values of $w$. The maximum detection-time gain of BU over CC is approximately 6.7 hours under zero false positives per month when using a baseline period $w = 2$ weeks. At two to five false positives per month for all values of $w$, CC performs about as well as BU, as shown in Figure 1.

We performed one-sided paired t-test to evaluate the detection times of BU and CC. The statistical analyses show that the disease-outbreak detection performance of BU is statistically significantly better than that of CC when shorter baseline periods are used to estimate the baseline behavior in order to avoid seasonal effects. When using longer baseline periods, BU performs as well as CC.

2.3 Bayesian Modeling of Anomalies Due to Known and Unknown Causes

We report here the results of simulation experiments of the univariate version of the BH algorithm that takes as input emergency department patient symptoms, such as cough, fever, and diarrhea, of the most recent 24 hours [Shen 2009c].

Let $d_u$ and $d_v$ be two distinct outbreak diseases. Table 1 shows our experiments for one such pair of $d_u$ and $d_v$. In this table, both experiments have simulated outbreaks due to disease $d_u$. However, disease $d_u$ is modeled by the system evaluated in Exp. 1 but not modeled (e.g., unknown) in Exp. 2. Let DSM be a detection system that includes a disease-specific model, which models non-outbreak diseases $d_0$ and a known outbreak disease $d_u$ (in Exp. 1) or $d_v$ (in Exp. 2). Let UDM be a detection system that includes an unknown-disease model, which models non-outbreak diseases $d_0$, a known outbreak disease $d_u$ (in Exp. 1) or $d_v$ (in Exp. 2), and an unknown outbreak disease $d^*$.  

Table 1. A $2 \times 2$ table that summarizes the two experiments that each involves the use of a disease-specific model (DSM) and an unknown disease model (UDM).

<table>
<thead>
<tr>
<th></th>
<th>DSM</th>
<th>UDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>Model $d_0, d_u$.</td>
<td>Model $d_0, d_u, d^*$</td>
</tr>
<tr>
<td></td>
<td>Simulate outbreak cases from $d_u$.</td>
<td>Simulate outbreak cases from $d_u$.</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>Model $d_0, d_v$.</td>
<td>Model $d_0, d_v, d^*$</td>
</tr>
<tr>
<td></td>
<td>Simulate outbreak cases from $d_u$.</td>
<td>Simulate outbreak cases from $d_u$.</td>
</tr>
</tbody>
</table>

Let event $G$ denote the following event: Given that an outbreak is occurring, it is due to a disease that is not being explicitly modeled in the detection system. According to Table 1, given that an outbreak is occurring, $G$ is true in Exp. 2 and is false in Exp. 1. Let $q$ be the probability that $G$ is
true. Recall that we wish to evaluate whether modeling the possibility of an unknown disease occurring is a net positive in detecting disease outbreaks. If \( q = 1 \), then modeling \( d_* \) will likely be helpful. If \( q = 0 \), however, modeling \( d_* \) will be useless and possibly harmful by allowing more chances for a false positive alert.

We represent models DSM and UDM in Exp. 1 as DSM1 and UDM1, respectively, and likewise represent models DSM and UDM in Exp. 2 as DSM2 and UDM2, respectively. Let \( E_{DSM1} \) be the average detection time of DSM1 over all the experiments described above at a false positive rate of one per month, since one false positive per month is frequently cited as an upper bound on a tolerable rate. Let \( E_{DSM2} \) be the average detection time of DSM2 over all the experiments described above at a false positive rate of one per month. Let \( E_{DSM} = (1 – q) \times E_{DSM1} + q \times E_{DSM2} \). Define \( E_{UDM} \) analogously.

Table 2 shows the mean outbreak detection time (in days) of all four disease models over all the experiments. By solving \((1 - q) \times E_{UDM1} + q \times E_{UDM2} < (1 - q) \times E_{DSM1} + q \times E_{DSM2}\), we obtain \( q > 0.14 \). Under the assumptions introduced, this result indicates that if the probability is greater than 0.14 of an outbreak being due to an unknown disease, then including \( d_* \) in the model is expected to decrease the detection time at a false-positive rate of one alert per month.

Table 2. Mean detection time (in days) at a false alert rate of one per month of all four models over all the experiments.

<table>
<thead>
<tr>
<th></th>
<th>DSM</th>
<th>UDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>6.05</td>
<td>6.18</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>7.14</td>
<td>6.38</td>
</tr>
</tbody>
</table>

It seems plausible that there are disease-outbreak-monitoring situations in which if there is an outbreak then the probability exceeds 0.14 of it being due to an unknown disease. The Olympics provide one possible scenario, where a bioterrorist might attempt to use a new infectious disease agent to maximize terror. In such situations, the methods we have developed for detecting unknown diseases could be beneficial.

2.4 Bayesian Aerosol Release Detector (BARD)

The original version of BARD assumes that exposure to anthrax would occur at one’s residential unit. We developed a refined version of BARD, called BARD-C, which accounts for the fact that if a release occurred during the working hours, a worker would be exposed at the exposure level of her work unit [Cami 2008b]. We conducted an experimental study to compare the
performances of BARD and BARD-C on semi-synthetic outbreaks generated with an algorithm that also accounted for commuting.

A total of 100 semi-synthetic outbreaks were generated using the BARD-C simulator. For all simulations, the release quantity was fixed at 0.5 kg. The release time was chosen uniformly at random from the test period of the year 2005. The remaining release parameters were drawn from their respective prior distributions. Each semi-synthetic outbreak was supplied as input to the BARD and BARD-C detectors.

Figure 2 shows the AMOC analysis for timeliness. As seen, BARD-C’s timeliness is nearly two hours less than BARD’s timeliness at every false alarm rate. A statistical test showed that, for each false alarm rate, the mean timeliness of BARD-C was statistically different from the mean timeliness of BARD, at the 0.05 level.

![Figure 2. AMOC curves for the timeliness of BARD and BARD-C detectors.](image)

We also compared the release localization performance of BARD and BARD-C. Figure 3 shows box plots and plots of the median absolute error in downwind distance for BARD and BARD-C. In each plot the horizontal axis shows the number of hours from the simulated release. The samples from which the box plots were constructed correspond to the different simulations. Several key conclusions can be derived from Figure 3. First, the absolute error in downwind distance of both BARD and BARD-C converges to relatively small values as the time from release (of simulated anthrax) increases. Second, the absolute error rates of BARD and BARD-C appear to converge around the same time. Third, after the convergence, the median absolute error in downwind distance for BARD-C is nearly 2 kms smaller than the median absolute error for BARD. Similar results were found for the crosswind distance.

![Figure 3. Box plots and plots of the median absolute error in downwind distance for BARD and BARD-C.](image)
2.5 Extension of the Bayesian Aerosol Release Detector (BARD) Simulation Algorithm

To evaluate the BARD detection algorithm, we used the extended BARD simulation algorithm to create 54 release scenarios. Our evaluation of the extended BARD simulation algorithm used the SCIPUFF model to read 6 different weather files generated by the MM5 model, and it assumed 3 release locations and 3 release quantities (0.5 kg, 1 kg and 1.5 kg).

The table below summaries the evaluation metrics including the detection time (the time from release to the time an alert is triggered) and the distance error, which is the distance between the simulated point-release location and the inferred point-release location.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection Time (hours)</td>
<td>60.0</td>
<td>59.48</td>
<td>48.00</td>
<td>80.00</td>
</tr>
<tr>
<td>Distance Error (Km)</td>
<td>8.8</td>
<td>16.6</td>
<td>19.1</td>
<td>43.0</td>
</tr>
</tbody>
</table>

2.6 Rank-Based Spatial Clustering (RSC) Framework

We used two metrics (ROC curves and AMOC curves) to evaluate the RSC framework using both Z-score and Bayesian approaches (RSC_{std_score} and RSC_{Bayesian} algorithms) and three other algorithms that we applied to semi-synthetic data. In particular, we injected artificial outbreaks into real over-the-counter (OTC) baseline data. The three other algorithms included two state-of-the-art spatial detection algorithms, namely, the Kulldorff spatial scan statistic algorithm (KSSS) and the Bayesian spatial scan statistic (BSSS) algorithm (using 16x16 and 24x24 grids), as well as one time series algorithm called the Wavelet Anomaly Detector (WAD) algorithm. In this study, all the evaluated algorithms used daily expected counts of monitored variables (e.g., the daily count of ED patients with respiratory chief complaints) that were modeled using a wavelet algorithm.

We constructed artificial outbreaks using two parameters: 1) outbreak size ($K$), which designates the number of connected zip codes over which the simulated outbreak takes place, and (2) the outbreak strength ($\delta$), which is the fraction of increase in counts from a mean baseline of no outbreaks. In this experiment, we set $K = 4$ and $\delta = 0.2$ or $0.3$. Based on the plots below, the RSC framework algorithms performed well compared to the other algorithms.
We used BCD to monitor for influenza cases in real time by processing real ED reports from ten UPMC hospitals on a daily basis. As an example, the figure below shows a Google Earth view of the counts of fever cases aggregated by zip codes in Allegheny County for patient visits on May 12, 2009. A zip code with ten fever cases, which is suspicious for a disease outbreak, is colored in red in the onscreen map of the production system; that suspicious zip code is labeled with a white X in the black and white version shown below. Such maps are being made available to public health epidemiologists in Allegheny County. We are in the process of gathering data longitudinally in order to evaluate how well the system assists the epidemiologists in detecting influenza outbreaks in the county.
A Google Earth view of fever cases (darker regions) aggregated by zip codes in Allegheny County for emergency department patient visits from 10 University of Pittsburgh Medical Center (UPMC) hospitals on May 12, 2009.

2.8 Complete Bayesian Model Averaging Over Spatial Distributions of Disease

To evaluate the Bayesian model averaging algorithm, we used 96 simulated anthrax outbreaks [Shen 2007]. Each outbreak consisted of a simulated time series of patient cases that each presented to an ED with a respiratory chief complaint and a home zip code. In the simulation, the probability that a case was assigned to live in a zip code was proportional to the population of that zip code. We call this a scattered outbreak. We then overlaid (injected) these simulated outbreak cases onto real ED cases to create a combined dataset. We ran PANDA [Cooper 2004] and the spatial Bayesian model-averaging algorithm (SBMA) on the combined datasets. The SBMA algorithm showed a better detection performance than PANDA, as expected due to the injections being widely scattered spatially rather than having a plume-like pattern of an outdoor release of anthrax, which PANDA monitors for. SBMA appears well-suited to detect scattered outbreaks. Based on the results of previous experiments, PANDA seems well-suited to detect plume-like patterns of disease outbreaks. Thus, the two algorithms appear to be complementary, which suggests that it would be useful to apply them both in performing outbreak detection.
2.9 A Recursive Bayesian Spatial Outbreak Detection Algorithm

Using synthetic data, the results of applying the recursive Bayesian spatial outbreak detection algorithm indicate that on the average we can detect a 30-day moderate outbreak in a complex region at approximately 2.2 days into the outbreak [Jiang 2007]. Furthermore, there is about a 50% overlap on average between the detected region and the outbreak region by day 2 of the outbreak, which increases to 80% by day 4, and to almost 100% by day 7.

2.10 Detection of Anomalous Patterns in General Datasets

The two anomalous pattern detection methods that we introduce in [Das 2009; Neill 2009d; Das 2007; Das 2008a; Das 2008b] were tested on real-world data sets for emergency department admissions, container shipping, and network intrusions. The emergency department dataset was tested using various methods of synthetic anomaly injections including the BARD simulator for aerial anthrax releases. In our tests we achieved statistically significant improvements in precision/recall curves up to a quadrupling of area under the curve measures in some cases. A detailed comparison of the two methods [Das 2009] demonstrated that the AGD method outperforms the APD method in cases when there is a high degree of similarity between records and the individual records are less anomalous; APD outperforms AGD in cases where the records are less similar but more highly anomalous. Moreover, both methods outperform previously proposed detection methods, such as individual record anomaly detection and the WSARE algorithm.

2.11 Rapid Detection of Multidimensional Rectangular Clusters

We applied the fast spatial scan method on real-world epidemiological data including Emergency Department visits and over-the-counter drug sales. The fast spatial scan gained huge performance improvements over the naïve approach, making the use of the scan statistic feasible in these large real-world datasets, and the extension of efficient scan statistics from two-dimensional to d-dimensional data dramatically increased the scope of problems for which these techniques can be used. The fast algorithm achieved speedups of 20-2000x over the naïve approach on two-dimensional ED and OTC datasets [Neill 2004b], 235-325x speedups on a four-dimensional (longitude, latitude, gender age) ED dataset, and 48-1400x speedups on three-dimensional, space-time OTC datasets [Neill 2005b].

2.12 An Expectation Scan Statistic and Detection of Emerging Space-Time Clusters
In [Neill 2009a; Neill 2009b], we examine the expectation-based scan statistic approach in detail, and compare the detection performance of this approach to other spatial scan methods. We demonstrated that expectation-based scan statistics, which aggregate information across multiple time series rather than monitoring each series separately, improves the accuracy, timeliness, and spatial resolution of detection [Neill 2009a]. Additionally, we compared twelve variants of the spatial scan statistic on the outbreak detection task over a wide range of datasets and outbreak types [Neill 2009b]. Two interesting conclusions of this work were (1) that our expectation-based Poisson (EBP) statistic achieves significantly better detection performance than Kulldorff’s widely used spatial scan statistic, especially for outbreaks with a large area of effect, and (2) that miscalibrated randomization testing can harm detection performance and lead to high false positive rates, thus suggesting the use of an approach in which we calibrate detection thresholds using the empirical distribution of maximum scores from historical data.

2.13 Nonparametric Scan Statistics

We have compared the detection performance of the nonparametric scan statistic to parametric expectation-based scan statistic approaches on univariate and multivariate health data including Emergency Department visits and over-the-counter medication sales, using a semi-synthetic evaluation framework. As compared to typical univariate scan approaches, the nonparametric multivariate scan statistic method demonstrated increased detection power by combining evidence from multiple disparate data streams. It also enabled characterization of events by identifying which data streams are affected, and outperformed other scan statistic approaches on datasets (such as binary sensor data) where typical parametric models are inadequate [Neill 2007b]. Our results demonstrate that the nonparametric method achieves higher detection power than multivariate parametric methods in cases where the parametric assumptions are incorrect (as is typical for OTC data), but has somewhat lower detection power when the parametric methods make reasonably correct assumptions (as is typical for ED data). Additionally, the nonparametric method is correctly calibrated (e.g., 5% of baseline days are significant at 0.05), while parametric methods have much higher false positive rates than desired. Finally, the nonparametric method can be used successfully in a variety of scenarios where the parametric approaches are inappropriate, such as combining alerts from multiple detectors or combining multiple disparate data streams.

2.14 Multivariate Bayesian Scan Statistics (MBSS)

We have demonstrated [Neill 2007c; Neill 2009c] that MBSS achieves faster and more accurate detection of emerging patterns by combining multiple streams of spatio-temporal data (for disease surveillance, these could include ED visits with different chief complaint categories and sales of different types of OTC medications), as well as modeling and distinguishing between
different types of patterns (e.g., distinguishing a bioterrorist anthrax attack from seasonal influenza). Many advantages over traditional event detection approaches include accurate characterization and differentiation of multiple event types, and comparable detection power to the current state of the art, even when uninformative priors are used. In [Neill 2009c] we show that much higher detection power can be achieved for specific event types (e.g., an anthrax attack) by providing an informative prior describing the spatial and temporal characteristics of the event and its relative effects on the various data streams. We demonstrated that these event models can be learned accurately from labeled training data, expert knowledge, or a combination of the two.

2.15 Incorporating Learning into Spatial Pattern Detection

We have shown that the spatial distribution of outbreaks can be learned efficiently from a small number of labeled outbreaks using a “latent center” model [Makatchev 2008a; Makatchev 2008b] and the effects on each data stream can be learned by a smoothed maximum likelihood approach [Neill 2009c; Neill 2007d]. The resulting models improved detection time by two full days. In [Das 2009] we demonstrated that learning temporal models can enable MBSS to automatically distinguish true disease outbreaks from increases in health-related behaviors caused by inclement weather. For example, an approaching hurricane causes a sizeable increase in over-the-counter medication sales before the hurricane (anticipatory effect), a very dramatic drop in sales during the hurricane, and often a smaller increase in sales immediately after the hurricane (restocking effect).

2.16 Fast Subset Scanning for Pattern Detection

Our preliminary results on linear-time subset scanning (LTSS) [Neill 2008] demonstrate that substantial speedups can be obtained by applying LTSS to both constrained and unconstrained subset scan problems. For example, we were able to complete one surveillance task on a dataset of Emergency Department visits from 88 Allegheny County zip codes in less than five seconds using LTSS, while a naïve approach would have required on the order of $10^{26}$ years. Similarly, proximity-constrained subset scans require time that is linear rather than exponential in the neighborhood size when LTSS is applied, and LTSS can be used with branch and bound to speed up a search over distinct rectangular regions by three orders of magnitude.

2.17 Smart Regularization for Learning

Of the four methods of regularization we studied, the multi-task learning approach bears great promise for immediate practical applications. We are in the process of testing it against multi-stream, multi-task food-safety data in which information extracted from a variety of sources can be used to simultaneously address multiple distinct surveillance tasks. In the presence of
redundancy of both information available in the input streams and the objectives of analysis, the multi-task approach appears to be perfectly fitting. We also tested the method against a specific subtask of the above large problem in which the multi-task approach was utilized to exploit groupings of food processing establishments. We have also tested our regularization methods using text, images, brain waves, and food safety data.

2.18 Scalability and Massive Screening of Multidimensional Surveillance Data

We have extensively tested the T-Cube data structure and surveillance algorithms. It is being deployed as a part of the Public Health Information System developed by the USDA. It is also being used as a part of the CMU Auton Lab’s T-Cube Web Interface in our ongoing applied research for USDA, CDC, and the US Air Force [Dubrawski 2008a; Dubrawski 2007a; Dubrawski 2009b]. A variant of it is being deployed as a part of the Real-Time Bio-surveillance Project fielded in Sri Lanka and India [Dubrawski 2009b].

2.19 Graph-Based Representations and Algorithms

The hypothesis being tested is that the observed networking of facilities through bacterial serotypes found in food can be predictive of the future microbial safety performance. To test it, we split the available historical data into independent training and test subsets in order to objectively measure prediction accuracies. We also compared our results with a baseline that only considers histories of individual establishments, independently of each other, to make predictions.

In [Sarkar 2008] we compare the DSNL network model with a simple, but usually hard-to-beat baseline algorithm that does assume independence of entities, using food safety data (historical record of Salmonella serotypes isolated at food establishments). The network model outperformed the alternative on the task of ranking salmonella serotypes according to their likelihood of occurrence at a given establishment. In [Dubrawski 2009c] we discuss trade-offs between agility and reliability of predictions arising due to sparseness of data. On one hand, the more frequent the prediction cycles, the less latent (more agile) the surveillance process can be. On the other hand, the narrower the window width, the lower the expected count of occurrences of specific microbial findings per food factory, hence the greater sparseness of data per each training time step and, consequently, the harder it is to obtain reliable predictions using DSNL. We identified that exploiting recentness of data and using the known patterns in it, such as seasonality, as plausible means of battling the challenge of sparseness. Another observed utility of the DSNL is that of a powerful visualization tool which can be used to support understanding of the evolution of the underlying network over time.
To evaluate the utility of the AFDL algorithm in biosurveillance applications, we used the food safety data mentioned above. It consisted of the record of positive results of tests for salmonella conducted at USDA regulated establishments over a period of three consecutive years. We used the two and a half initial years of data to derive connectivity features, and the last six months were used for evaluation of accuracy of predicting whether a given food-processing establishment would have a positive test for a particular serotype of *Salmonella Montevideo* (one of over a hundred known serotypes of Salmonella), given the demographic information of that establishment and its connectivity pattern. The data consisted of over 3,500 establishments. Their demographic features include three groups of characterizations, such as the production profile (a list of types of food produced, such as meat, poultry, and eggs), the corporate membership of the establishment, and the geographic region in which it is located.

An establishment is considered active if over a specific period of time it records a positive case of *Salmonella Montevideo*. Two or more establishments are considered linked if they record positives of identical serotype over the period of time under consideration. Such definition of a link produces inherently uncertain data, since the co-occurrences of specific test results at different establishments may be purely coincidental and our analysis does not attempt to distinguish them from truly correlated instances. Nonetheless, we observed that using this inherently noisy data in combination with demographic features of establishments, leads to improvements in predictive accuracy of AFDL.

We tested AFDL’s ability to predict which of the monitored establishments would record at least one positive for *Salmonella Montevideo* over six months following two years worth of training data. Using only demographic features of the establishment to make predictions does not help as much as the result is barely distinct from what can be obtained by randomly sampling the entities. But adding noisy linkage information data improves recall of active establishments. Using only *Montevideo* co-occurrences to create links does not in general surpass the results obtained previously, but it does help considerably for shorter subsets of the top ranked establishments. That actually brings about its practical utility, as the investigators would typically begin their work by inspecting the most likely active entities first, and that is where *Montevideo*-based AFDL provides an immediate improvement. However, using all types of matching serotypes to estimate connectivity patterns substantially boosts the characteristic, both in the context of a short list, as well as overall. These results can be explained by the benefits of the higher density of graphs created using more observed connections when using all known serotypes rather than just one to model potential relationships between establishments [Dubrawski 2008b].
2.20 Learning Specific Detectors of Events in Multiple Streams of Data

We showed that detectors of events of specific types trained using a large number of known historical events can substantially outperform non-specific detectors in terms of detection speed and false positive rate [Roure 2008]. Additionally, because it is very difficult in practice to acquire a large number of labeled events, we developed another kind of specific detector which can be trained to perform well using only very few labeled data. These detectors combine the well-known Fisher's p-value aggregation method to signal alerts and a filter trained to discern likely false alarms. We observed that it is possible to have effective filters using very few training examples and that adding more labeled training cases leads to positive marginal effects.
References


