Research and Educational Activities

The four main research thrusts of this award are:

1. Incorporation of incremental model learning into event detection, creating an integrated framework for learning and detection which allows continuous discovery and learning of new event models from user feedback, and enabling continual improvement in detection performance.

2. New methodological contributions for fast subset scanning which allow event detection methods to scale up to large and high-dimensional data without sacrificing accuracy.

3. End-to-end methods that augment the automatic detection of events by providing methodological tools for event characterization, explanation, visualization, investigation, and response.

4. New methods which incorporate data from emerging, transformative technologies and address the fundamental scalability challenges.

These research thrusts are integrated with a multi-pronged educational and curriculum development program, the Machine Learning and Policy (MLP) initiative, which focuses on incorporating machine learning into public policy research and education. Our progress toward achieving the goals of each research thrust, as well as our educational activities and results, are discussed in detail below.

Task 1: Incorporating Learning into Event Detection

Our first research thrust focuses on improving the timeliness, accuracy, and utility of event detection through the incorporation of incremental model learning. Current state-of-the-art detection systems combine spatio-temporal information from multiple data streams to detect emerging events. However, these methods rely on fixed, pre-specified models, and cannot improve performance over time. This creates a practical problem when they detect many patterns which are anomalous but irrelevant to the user, greatly diminishing the utility of the system. For example, even state-of-the-art disease surveillance systems produce a huge number of false positives due to non-outbreak causes, ranging from inclement weather to tourism to promotional sales. Worse yet, even if the user manages to find a relevant pattern (or wishes to rule out some irrelevant pattern type), he is unable to convey this knowledge to the system. These challenges suggest the need for detection methods which can discover new event types, and improve existing models, by learning from data or user interaction. Incorporation of learning into the event detection process will achieve three main benefits: more timely and accurate detection of events, ability to model and distinguish between different event types requiring different responses, and dramatically reducing false positives by learning the relevance of each event type and reporting only the most relevant detected patterns.

The past three years' work on Task 1 has primarily focused on three areas: incorporating learning into our Bayesian framework ("Fast Subset Sums") for scalable event detection and visualization, learning an underlying graph structure from observations, and incorporating multiple known models into the Fast Generalized Subset Scan framework for anomalous pattern detection. Our work on learning for Fast Subset Sums is described under Task 3b (scalable event detection and visualization) below, but can also

be considered part of Task 1a (learning complex models). As discussed below, in our Bayesian event detection framework, we can learn multiple event models which differ in their spatial extent, density or sparsity, and their relative effects on the different monitored data streams, and use these models to more accurately detect and characterize emerging events. Second, we have investigated how the use of graph-based event detection methods (typically used to detect the most anomalous connected subgraph for a known graph structure) can also be used to learn the underlying graph structure. This is discussed under Task 1a (learning complex models) below. Finally, we have begun to consider cases where the user does not have sufficient time and resources to examine the entirety of the monitored data, but can respond to only a limited number of potential events identified by the system. In this case, the system must rapidly focus the user's attention on the most relevant patterns, as well as modeling the different event types and the relevance of each event type to the user. One key component is to detect anomalous patterns which do not fit any of the multiple known models, as discussed under Task 1d (learning new event models) below. These anomalies can then be presented to the user, and may be labeled as examples of a new event type which can then be modeled (e.g. learning a Bayesian network from the labeled data).

Task 1a: Learning complex models

We propose to learn more complex model specifications and incorporate these models into the detection process. In addition to incorporating model learning into our Generalized Fast Subset Sums framework (described as part of Task 3b below), a second main focus has been learning the underlying graph structure for events that spread along a graph or network. While graph-based event detection algorithms (such as our GraphScan approach described below) typically assume a known graph structure, we have developed a new, general framework for **learning an unknown graph structure** from data, and demonstrated that the graph learning framework enables more timely and more accurate event detection (Somanchi and Neill, 2013, submitted for publication).

Processes such as disease propagation or information diffusion often spread over some latent network structure (e.g. social networks or person-to-person contacts) which must be learned from our observations of the nodes in the network. For example, in disease surveillance, we might observe the time series of case counts for each of a set of spatial locations, but not know which locations are likely to spread disease to which other locations (via spatial adjacency, travel patterns, common food or water sources, etc.). Thus we attempt to reconstruct the underlying network along which a disease outbreak or other event might spread, and use the learned network to improve the timeliness and accuracy of event detection. However, in many cases labeled data may not be available: for example, public health officials might be aware that an outbreak has occurred, but may not have detailed information about which areas were affected and when. Hence we focus on learning graph structure from **unlabeled** data, given only a time series of observed counts (such as hospital visits or medication sales) at each node.

Our solution is to compare the most anomalous subsets detected with and without the graph constraints: we score each of a set of potential graph structures $G_1 \ldots G_M$ for each training example $D_1 \ldots D_J$, finding the most anomalous (highest scoring) connected subset and its score using an efficient graph-based event detection algorithm (our GraphScan algorithm, or the fast but approximate ULS approach)

for each combination of graph structure and dataset. These scores are normalized by dividing by the score of the most anomalous unconstrained subset for that training example, which can be efficiently computed using LTSS, and the normalized scores for a potential graph structure G_m are averaged over all of the J training examples. The idea is that, if the given graph structure is close to the true underlying graph structure, then the maximum constrained score will be close to the maximum unconstrained score for many of the training examples, while if the graph structure is missing essential connections, then the maximum constrained score given that graph structure will be much lower than the maximum unconstrained score for many examples. However, any graph with a very large number of edges will also score very close to the maximum unconstrained score, and thus we compare the score of the given graph structure to the distribution of scores of random graphs with the same number of edges, and choose the graph structure with the most statistically significant score given this score distribution.

Within our general framework for graph structure learning, we compared five approaches which differed both in the underlying detection method (BestSubgraph) and the method used to choose the next edge for removal (BestEdge), incorporated into a provably efficient greedy search procedure. We demonstrated both theoretically and empirically that our framework requires fewer calls to BestSubgraph than a naive greedy approach. We also evaluated the scalability, detection power, and accuracy of these approaches on various types of simulated disease outbreaks, including outbreaks which spread according to spatial adjacency, adjacency plus simulated travel patterns, and random graphs (Erdos-Renyi and preferential attachment), as discussed in the "Results and Findings" section.

This work was presented at the 2011 International Society for Disease Surveillance Annual Conference (abstract published in the *Emerging Health Threats Journal*) and the 2012 INFORMS Annual Conference. A related book chapter will be published in the *Encyclopedia of Social Network Analysis and Mining* (Springer, 2013, in press). The full paper is currently under revision for re-submission to a computational statistics journal. We are currently working to scale up the approach to much larger graphs, e.g. for use in analysis of massive social networks. Our ongoing work also focuses on extending the graph structure learning framework in several directions, including learning graph structures with directed rather than undirected edges, learning graphs with weighted edges, and learning dynamic graphs where the edge structure can change over time.

Task 1d: Learning new event models

In recent work, we have developed several "model-based" methods for detection of previously known and modeled patterns (the multivariate Bayesian scan statistic, fast subset sums, and generalized fast subset sums methods), as well as numerous "anomaly-based" methods for detection of previously unknown pattern types (e.g. the fast generalized subset scan and other fast subset scan methods). One important goal of our ongoing work is to integrate detection of known and unknown patterns, reporting to the user both a) patterns corresponding to known and relevant pattern types, and b) patterns which are sufficiently anomalous to be potential examples of a new and previously unknown pattern type. By incorporating user feedback on both known and previously unknown patterns, the set of known patterns and the accuracy of the models will continue to grow over time.

Currently, we are working to extend both the fast generalized subset scan (FGSS) and generalized fast subset sums (GFSS) methods to the "known and unknown patterns" case. The idea is that FGSS can be used to detect anomalous patterns not matching the expected data distribution (modeled by a Bayesian network learned from training data), while GFSS can be used to distinguish between multiple known and modeled patterns. To integrate the two methods, our first step is to extend FGSS to **multiple known model types**, each modeled by a Bayesian network, thus identifying **novel anomalous patterns** that do not fit any of these models. Our current approach (McFowland and Neill, 2013, in preparation) iterates between three optimization steps: choosing a "best fit" model for each record given the current set of attributes; choosing the most anomalous subset of records given the current subset of attributes and the "best fit" models; and choosing the most anomalous subset of attributes given the current subset of records and the "best fit" models. We then detect subsets of records and attributes that are unlikely given each known pattern model as well as the null model, thus enabling FGSS to discover previously unknown pattern types given the current set of known patterns.

We compare this "Fast Generalized Subset Scan with Multiple Models" (FGSS-MM) approach to a simpler extension of FGSS, which we term FGSS-Mixture (FGSS-MIX). FGSS-MIX simply replaces the null model with a mixture model, representing the background data distribution and other known data patterns. The optimization procedure for FGSS-MIX then follows the same procedure as FGSS: iterate between choosing the most anomalous subset of records for the given set of attributes and vice-versa.

As described in the "Results and Findings" section below, we evaluate these two extensions of FGSS on two application domains, network intrusion detection and masquerade detection (i.e., rapidly identifying individuals using a computer system who have legitimate credentials but are not who they claim to be and are likely involved in harmful activities, based on their observed actions). The first draft of our "FGSS with multiple known models" paper is complete; we are currently extending the evaluation results, and plan to submit this work to ICDM 2013 this summer.

Task 2: Fast Subset Scanning for Scalable Event Detection

In the subset scan framework, our primary goal is to find the subsets of the data which are most anomalous (or that best match some known and relevant pattern) by maximizing the score function F(S). Since an exhaustive search over subsets is computationally infeasible, typical spatial scan methods either restrict the search space, e.g. by searching over circular or rectangular regions, or perform a heuristic search. The former approach has low detection power for regions outside the search space (e.g. elongated or irregular clusters), while the latter does not guarantee that an optimal or near-optimal region will be found. However, we have discovered that many pattern detection methods satisfy a property (**linear-time subset scanning**, or LTSS) which allows efficient optimization over all subsets of the data: 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 approach enables us to optimize F(S) by evaluating only N of the 2^N possible subsets. We are in the process of investigating many ways in which LTSS will enable efficient event detection, removing some of the computational barriers faced by subset scanning methods.

Our first paper on linear-time subset scanning (Neill, 2012) has recently been published in the *Journal of the Royal Statistical Society Part B (Statistical Methodology)*, the #1 statistics journal as ranked by impact factor. This paper included theoretical results on LTSS, the univariate fast subset scan framework, and incorporation of hard constraints on spatial proximity, as discussed in last year's report.

This year's work on Task 2 focused on six main areas: extension of fast subset scanning to general datasets through continued development of the Fast Generalized Subset Scan (FGSS) algorithm (Task 2a), incorporating soft constraints on spatial proximity (Task 2b), extending the fast subset scan framework to multivariate and tensor datasets (Task 2b), incorporating connectivity constraints through continued development of the GraphScan algorithm (Task 2c), detecting events in heterogeneous graphs derived from social media such as Twitter (Task 2c), and incorporating temporal consistency constraints (discussed as part of Task 3c, automated event detection and tracking, below).

Task 2a: Extend LTSS to general multivariate datasets

In McFowland, Speakman, and Neill (2013), we present Fast Generalized Subset Scan (FGSS), an extension of the LTSS approach which enables efficient pattern detection in general multivariate datasets. In this case, we do not have space-time data, but instead have an arbitrary set of attributes measured for each of a large set of data records. In this problem setting, our goal is to detect self-similar subsets of data records for which some subset of attributes are anomalous. Our approach consists of four steps: 1) efficiently learning a Bayesian network which represents the assumed null distribution of the data; 2) computing the conditional probability of each attribute value in the dataset given the Bayes Net, conditioned on the other attribute values for that record; 3) computing an empirical p-value range corresponding to each attribute value by ranking the conditional probabilities, where under the null hypothesis we expect empirical p-values to be uniformly distributed on [0,1]; and 4) using a nonparametric scan statistic to find subsets of records and attributes with an unexpectedly large number of low (significant) empirical p-values. The final step is computationally expensive (exponential in the numbers of records and attributes for a naïve search), but LTSS can be used to speed up this search, converging to a local maximum of the score function and ensuring that each iteration step is linear (not exponential) in the number of records or attributes.

The FGSS framework has been evaluated on several application domains, including early detection of simulated anthrax bio-attacks, discovery of patterns of illicit container shipments for customs monitoring, and network intrusion detection, demonstrating improved detection accuracy, efficient runtime, and ability to correctly characterize the affected subset of attributes in all three domains. Results of these evaluations are presented below.

In the past year, our first paper on FGSS (McFowland, Speakman, and Neill, 2013) has been accepted for publication in the *Journal of Machine Learning Research*. We also have presented talks on FGSS at the *International Workshop on Applied Probability (IWAP 2012)* and the *2011 INFORMS Annual Conference*.

We have also extended the FGSS approach to the case of **mixed real- and categorical-valued datasets**, augmenting the Bayesian network with a regression tree for each real-valued attribute. Then the probability density corresponding to each attribute value (conditioned on the other attribute values for

that data record) is computed by performing kernel density estimation using only the appropriate leaf of the regression tree. We can then compute the empirical p-values for that attribute by computing and ranking the kernel density estimates corresponding to each attribute value, and perform the fast LTSS-enabled nonparametric scan as before. We are also investigating an alternative model which uses a **dependency network** instead of a Bayesian network to model the distribution of the data when no events are occurring; this avoids the computationally expensive Bayes Net structure learning step, and enables the algorithm to scale to datasets with hundreds or thousands of attributes, making it usable for our work on discovering anomalous patterns of patient care discussed below. Additional improvements may be achieved by directly using the learned Bayes Net or dependency network, rather than the entire training dataset, to compute the empirical p-values. Finally, we are currently working to extend FGSS to **multiple known model types**, as discussed in Task 1d above.

Task 2b: Extend LTSS to constrained subset scans

Since LTSS only guarantees a solution to the unconstrained (all-subsets) optimization problem, the biggest challenge is to incorporate constraints such as spatial proximity, graph connectedness, or temporal consistency to ensure that relevant and useful subsets are detected. In recent work, we have developed a number of novel and powerful methods for *constrained* optimization, using the unconstrained LTSS method as a building block. In this section we focus entirely on the incorporation of spatial proximity constraints; fast graph scanning is discussed in Task 2c below, and our extensions to detection and tracking of dynamic events (by incorporating temporal consistency constraints) are discussed in Task 3c below.

In Neill (2012), in addition to presenting the basic theoretical framework for LTSS described above, we focus on the application of LTSS to univariate spatial data, and consider how spatial proximity constraints can be incorporated. We often want to use spatial information to constrain our search by penalizing or excluding unlikely subsets (e.g. spatially dispersed or highly irregular regions). Thus we propose "fast localized scan" approaches which incorporate spatial proximity constraints into the LTSS framework. For example, we can constrain our search to regions consisting of a center location and some subset of its k-nearest neighbors, using LTSS to reduce the complexity from exponential to linear in k. These efficient, spatially-constrained LTSS searches allow us to perform spatial detection tasks in milliseconds that would require years for exhaustive search, and substantially improve detection power and spatial accuracy as compared to the traditional spatial scan approach (searching over circular regions).

In the past two years, we have developed a new framework which allows us to incorporate "soft" constraints on spatial proximity, rewarding compact regions and penalizing sparse regions, and thus enabling efficient and accurate detection of irregularly-shaped spatial clusters (Speakman, McFowland, Somanchi, and Neill, 2012). Our previous fast subset scan approach can incorporate proximity constraints using a fixed neighborhood size k; however, each of the 2^k subsets are considered equally likely, and thus the fast localized scan does not take into account the spatial attributes of a subset. Thus we extended the fast localized scan by giving preference to spatially compact clusters while still considering all subsets within a given neighborhood. For a given local neighborhood with center

location s_c and size k, we place a bonus or penalty $\Delta_i = h(1 - 2d_i/r)$ on each location s_i, where d_i is that location's distance from the center, r is the neighborhood radius, and h is a constant representing the strength of the compactness constraint. Each Δ_i can be interpreted as the prior log-odds that s_i will be affected, and thus the center location is e^h times as likely as its (k-1)th nearest neighbor.

In work presented at the 2011 International Society for Disease Surveillance Annual conference (abstract published in the *Emerging Health Threats Journal*), we first demonstrated that this approach can efficiently and accurately detect irregularly-shaped outbreaks. This work was also presented at the International Workshop on Applied Probability in June 2012, and will be presented at the 6th International Conference on Computational and Methodological Statistics in December 2013. Our full paper is under revision, and will be re-submitted to the top-tier data mining conference ICDM 2013 this summer.

Task 2b, continued: Extension of LTSS to multivariate and tensor datasets

While Neill (2012) focuses on univariate event detection (in which we monitor a single spatio-temporal data stream), linear-time subset scanning can also be extended to multivariate event detection, in which we integrate information from multiple spatio-temporal data streams. In Neill, McFowland, and Zheng (2013), we demonstrate how LTSS can be used to speed up two different multivariate scan statistic methods, Subset Aggregation (an extension of the method proposed by Burkom et al., 2005) and the multivariate spatial scan proposed by Kulldorff et al. (2007). This work was recently published in the journal, *Statistics in Medicine*, and was presented as part of our invited talk on multivariate surveillance at the *2011 Joint Statistical Meetings*.

The key insight behind this paper is that LTSS can either be used to efficiently optimize a score function over subsets of attributes (e.g. monitored data streams) for a given subset of data records (e.g. monitored spatial locations), or to optimize over records for a given subset of attributes. Thus we can *iterate* between optimizing over records and attributes until the algorithm converges to a (local) maximum of the score function over all subsets of records and attributes, and use multiple randomized restarts to approach the global maximum. The above discussion assumes one particular formulation of the multivariate scan statistic, in which we add counts across the monitored subset of data streams. An alternative formulation by Kulldorff et al. (2007) proposes adding log-likelihood ratios across streams (e.g., assuming that the data streams are conditionally independent). We demonstrate that the Kulldorff multivariate scan can also be made efficient using LTSS, by iterating between two steps: optimizing over subsets of records (for given values of the multiplicative effect of the event on each data stream), and re-calculating the maximum likelihood values of the event's effects for the given subset of records. We then evaluated the detection performance of both variants of the multivariate spatial scan for synthetic and real-world disease surveillance datasets, demonstrating that our LTSS-based approach significantly improved detection power and spatial accuracy for both methods, while maintaining efficient and scalable computation. Results for this paper were described in previous annual reports, and are also available in the published paper (Neill, McFowland, and Zheng, 2013).

We have recently extended the multivariate LTSS approach from matrix data (records x attributes) to **multi-dimensional tensor data** with an arbitrary number of modes (Neill and Kumar, 2013). Our previous work allows efficient optimization over subsets of records and attributes, which can be thought of as the rows and columns of a matrix; the current work allows joint optimization over subsets of each mode of a tensor with three or more modes. Our approach is a natural generalization of our multivariate fast subset scan algorithm: we randomly initialize the algorithm, then iteratively optimize over subsets of each tensor mode given the other modes. Each such conditional optimization can be performed efficiently using the LTSS property; the iterative process converges to a local maximum of the score function, and then multiple randomized restarts can be used to approach the global maximum.

In recent work presented at the 2012 International Society for Disease Surveillance Annual Conference (abstract published in the *Online Journal of Public Health Informatics*, 2013), we applied this technique to the disease surveillance domain, using multivariate case data from individuals in a population. In this setting, we have not only multivariate spatio-temporal count information (the number of cases for each location, time step, and data stream), but also additional categorical attributes for each affected individual (such as age group and gender). Each such attribute is represented by a tensor mode, and location and data stream are represented by two additional modes. Our Multi-Dimensional Subset Scan (MD-Scan) approach identifies not only the affected spatial locations and data streams, but also the characteristics of the affected subpopulation, as represented by a subset of values for each monitored attribute (e.g. "males under 30 who use intravenous drugs"). We demonstrate that this approach enables accurate and timely detection of emerging events, while maintaining computational tractability for massive datasets, as discussed in the "Results and Findings" section below.

Our ongoing work involves applying MD-Scan to many other application domains, including detection of anomalous patterns of care in a hospital setting. One challenge in non-spatiotemporal domains is estimating the expected count corresponding to each combination of attributes in the tensor; our ongoing work learns the structure and parameters of a Bayesian network from the case data and uses the learned model to predict counts. We believe that the ability of MD-Scan to accurately characterize the areas and subpopulations affected by an event will have important applications in other areas, such as disaster and crisis response, where it is essential for an intervention to be both rapid and precisely targeted. We are also working with the Chicago Department of Public Health to obtain data about the prevalence of sexually transmitted illnesses, risk factors, treatments, and preventive measures, in order to detect emerging patterns of STIs which may differentially affect a particular subpopulation or a particular neighborhood.

Task 2c: Fast graph scanning

Another extension of linear-time subset scanning focuses on graph and network data, where we monitor one or more data streams at each node of the graph, and wish to detect the most anomalous subset of nodes subject to the graph connectivity constraints (i.e. the given subset of nodes must form a connected subgraph of the original graph). If the score function satisfies LTSS, we can prove the following rule: "If subset S_{in} is included in the highest-scoring connected subset S, and removing S_{in} would not disconnect S, then no connected subset S_{out} adjacent to S can have higher priority than S_{in}."

This rule was incorporated into a depth-first search procedure which enables us to rule out many subsets which are provably suboptimal, reducing the search space and resulting in huge speed improvements. Additional speed improvements can be obtained by branch and bounding, applying the unconstrained LTSS property to quickly compute an upper bound on scores and ruling out provably suboptimal subgraphs. Although the detected subgraphs are similar to the previously proposed FlexScan algorithm (Tango and Takahashi, 2005), GraphScan is able to scale to much larger graphs consisting of several hundred nodes, with a **450,000-fold increase** in speed compared to FlexScan for neighborhoods of size k = 30.

GraphScan can be used for spatial data (searching for the most anomalous connected cluster of zip codes, with edges defined by spatial adjacency, travel patterns, etc.), and can also be used for non-spatial data with an underlying graph structure (including cell phone call graphs, social networks, and the Enron e-mail dataset). We believe that this approach will be particularly useful for our future work on detecting and preventing hospital-acquired illness, monitoring the spread of nosocomial infections between hospitals and between rooms within a hospital based on the movement of patients and hospital staff.

Our GraphScan algorithm was recently presented at the Quality and Productivity Research Conference (QPRC 2012), and a related book chapter will be published in the *Encyclopedia of Social Network Analysis and Mining* (Springer, 2013, in press). Finally, the full paper is currently under review for the *Journal of Computational and Graphical Statistics*. The GraphScan optimization step was also embedded into our multidimensional subset scan approach described in Task 2b above, thus allowing us to place a connectivity constraint (assuming a pre-specified graph structure) on our search over subsets for any or all modes of the tensor. GraphScan has also been extended to incorporate temporal consistency constraints, as described in Task 3c below, and to learn the underlying graph structure, as described in Task 1a above.

Task 2c, continued: Event detection in heterogeneous social media graphs

Event detection in social media is an important but challenging problem, with applications to conflict prediction, outbreak detection, and many others. Most existing approaches are based on burst detection, topic modeling, or clustering techniques, which cannot naturally model the implicit heterogeneous network structure in social media. As a result, only limited information, such as terms and geographic locations, can be used. We have recently developed a novel, non-parametric scan statistic approach that considers the entire heterogeneous network for event detection: we first model the network as a "sensor" network, in which each node (including Twitter users, keywords, locations, hashtags, tweets, etc.) senses its "neighborhood environment" and reports an empirical p-value measuring its current level of anomalousness for each time interval (e.g., hour or day). Then, we efficiently maximize a nonparametric scan statistic over connected subgraphs (using a very fast, but approximate, variant of our GraphScan approach) to identify the most anomalous network clusters. Finally, the event represented by each cluster is summarized with information such as type of event, geographical locations, time, and participants. This work (Chen and Neill, 2013) is in progress and will be submitted to the top-tier data mining conference ICDM 2013 this summer.

Task 3: End-to-End Methods for Event Surveillance

While most current surveillance systems focus on the problem of early detection of events, detection alone is not sufficient to enable a timely and effective response by the system's users. Successful event surveillance requires careful consideration of every step in the end-to-end process of data collection, automated detection and characterization, and user investigation and response. Our proposed work will augment event detection methods with novel methodological contributions and deployable tools which public health, law enforcement, and health care organizations can use to understand, visualize, investigate, and respond to emerging events. The past three years' work on Task 3 has focused on the development of Generalized Fast Subset Sums, our Bayesian framework for scalable event detection and visualization (Task 3b), and the incorporation of temporal consistency constraints to enable detection and tracking of events that change dynamically over time (Task 3c).

Task 3b: Scalable event detection and visualization

The multivariate Bayesian scan statistic (MBSS) is a powerful detection method which can integrate information from multiple data streams and can model and distinguish between multiple event types (Neill and Cooper, 2010). The output of the MBSS method can be easily visualized by computing the posterior probabilities that each event type E_k has affected each spatial location s_i, summing the posterior probabilities for all regions S containing s_i. Unlike standard spatial scan visualizations, which do not compute probabilities but instead show the most likely cluster, this method is able to quantify the system's uncertainty about the spatial extent and type of events. However, our LTSS method cannot be used to efficiently generate this visualization, since we need to sum over probabilities rather than just finding the highest-scoring region.

Thus we developed an efficient **Fast Subset Sums** (FSS) method which computes the summed posterior probability over all subsets containing location s_i, without computing the posterior probability of each individual subset. This work extends the MBSS framework to enable detection and visualization of irregularly-shaped clusters in multivariate data, by defining a hierarchical prior over all subsets of locations. While a naive search over the exponentially many subsets would be computationally infeasible, we demonstrate that the total posterior probability that each location has been affected can be efficiently computed, enabling rapid detection and visualization of irregular clusters. We compared the run time and detection power of this "fast subset sums" method to our original MBSS approach (assuming a uniform prior over circular regions) on semi-synthetic outbreaks injected into real-world Emergency Department data from Allegheny County, PA. Our results (presented in previous annual reports) demonstrated substantial improvements in spatial accuracy and timeliness of detection, while maintaining the scalability and fast run time of the original MBSS method. The full paper was published in the journal *Statistics in Medicine* (2011).

We have recently developed a generalization of the fast subset sums method which allows the sparsity of the detected region to be controlled (Shao, Liu, and Neill, 2011). More precisely, we propose a hierarchical probabilistic model with three steps: first, choosing the center location s_c from a multinomial distribution; second, choosing the neighborhood size k from a multinomial distribution; and

third, independently choosing whether to include (with probability p) or exclude (with probability 1-p) each location in the k-neighborhood of the center. We demonstrate that our previously proposed MBSS and FSS methods correspond to special cases of this **Generalized Fast Subset Sums** (GFSS) method, with p = 1 and p = 0.5 respectively, and show that appropriate choice of the sparsity parameter p enables much faster detection and higher spatial accuracy than either MBSS or FSS. Moreover, we demonstrate that the distribution of the sparsity parameter can be accurately **learned** from a small amount of labeled training data, and that the resulting GFSS method with learned p distribution outperforms MBSS, FSS, and GFSS with a uniform p distribution. We also show that two otherwise identical event types with different sparsities can be reliably distinguished by learning each event's p distribution, and that learning both an event's sparsity distribution and its relative effects on different data streams leads to more timely detection and better characterization than learning either parameter on its own. These results were presented in a previous annual report, and the full paper was published in the proceedings of the top data mining conference, the IEEE International Conference on Data Mining (ICDM), in 2011.

Most recently, we have developed a new expectation-maximization (EM)-based method which enables simultaneous learning of the distributions for the sparsity parameter, neighborhood size, and center location. Our preliminary experiments suggest that this approach can accurately and efficiently learn these distributions, dramatically improving detection power; we plan to follow up with a more detailed set of experiments for possible submission to ICDM 2013. We are also working on extensions of the EM-based learning approach to partially labeled data, where only a small subset of the affected locations is provided.

Task 3c: Automated event investigation and tracking

Once a potentially relevant event is detected by a surveillance system, the user must often perform a detailed investigation in order to understand its source, extent, and potential impact, enabling an appropriate and effective response. We propose novel methods to assist public health users in two distinct types of post-detection investigation: contact tracing (identification of individuals who may have been exposed to a contagious disease by contact with an infected person), and back-tracing of food-borne outbreaks (identifying the source of contamination by investigating links back from affected consumers to distributors, suppliers, and producers). These problems are graph-based in nature, and thus we can use our GraphScan method to efficiently find the most anomalous connected subset of nodes. However, we must also take the problem's temporal constraints into account, e.g. a person cannot infect others until some time period after they have been infected. Thus we have extended fast graph scanning to the dynamic case, allowing the affected subset of data records to change over time.

In recent work (Speakman and Neill, 2013, in preparation), we have developed a novel method for incorporating **soft constraints** into our linear-time subset scanning framework. Unlike many of the LTSS approaches describe above, we do not restrict the search space, but instead consider all subsets of the data while rewarding subsets that are more likely or penalizing subsets that are less likely to be affected. Incorporating soft constraints into the LTSS framework is challenging because, for an arbitrary score function F(S) that satisfies the linear-time subset scanning property, a penalized version of that function is not guaranteed to satisfy LTSS.

However, we have shown that this problem can be circumvented by conditioning on the event's severity (or relative risk), denoted as q. For a given value of q, and assuming any expectation-based scan statistic in the separable exponential family, we have shown that the score function $F(S \mid q)$ can be written as an additive function, $F(S \mid q) = \sum_{s_i \in S} G_q(s_i)$. For such functions satisfying the **additive LTSS property**, we can write the penalized form $F_{pen}(S \mid q) = \sum_{s_i \in S} (G_q(s_i) + \Delta_i) = \sum_{s_i \in S} H_q(s_i)$, where $H_q(s_i)$ is the total contribution of location s_i to the penalized scan statistic for the given value of q. Thus, for a given severity value q, we can easily maximize $F_{pen}(S \mid q)$ over all subsets of the data, by choosing all and only those locations with positive values of $H_q(s_i)$. Connectivity constraints can also be incorporated into this framework: this becomes an NP-hard problem (minimum Steiner tree), but is still feasible for graphs with several hundred nodes. For the unconstrained case, we have recently shown that only a small (linear) number of distinct q values must be considered, thus enabling a polynomial-time, exact solution which is nearly as fast as the unconstrained LTSS algorithm (running in 40-50 ms per day of data for our experiments on Allegheny County Emergency Department data). Connectivity constraints can also be incorporated into this framework: this becomes an NP-hard problem (related to the minimum Steiner tree problem), but is still feasible for graphs with several hundred nodes, and we have recently developed a fast approximate algorithm which can scale to 10,000 nodes or more.

By applying the additive LTSS property, we can enforce soft constraints on **temporal consistency** by considering the patterns detected at adjacent time steps, and rewarding patterns that are not dramatically different between time steps t and t+1. This allows us to extend our detection methods from detecting static patterns (which affect a fixed set of locations for some time duration) to **dynamic patterns** (which can grow or spread over time) while still maintaining efficient computation. Additionally, by using temporal consistency constraints to share information between multiple time steps, we can allow patterns to evolve smoothly over time while penalizing patterns which display unrealistic temporal trends (e.g. affecting the east side of the city on day 1, the west side on day 2, and back to the east side on day 3).

A preliminary version of this approach, presented at the 2011 INFORMS Annual Conference, applies the temporal consistency constraints moving forward in time, rewarding locations which were present for each of the past two time steps and also the neighbors of these locations. This algorithm was used to detect dynamic patterns in graph data with connectivity and temporal consistency constraints, applied to the detection of **spreading contaminants in a water distribution network**. Our preliminary results (presented in a previous year's annual report) show that incorporating simple size and temporal consistency constraints in a penalized, expectation-based binomial scoring function allows GraphScan to detect the contaminants earlier and to more accurately identify which nodes are affected as the contamination spreads through the network.

Our current extensions of the algorithm, which have been presented at the 2012 International Society for Disease Surveillance Annual Conference (abstract published in the *Online Journal of Public Health Informatics*, 2013) and will be incorporated into a journal paper submission in the next few months, are threefold. First, instead of only propagating our beliefs about the affected subset forward through time, we have developed an iterative approach which enables propagation of information both backward and forward in time; the detected subset for each time step is optimized given the detected subsets for the

previous and next time steps. Second, we have developed new methods for choosing the deltas (bonuses or penalties) for including each location on a given time step, based on a simple (log-linear) generative model of event propagation: log odds (affected on time step t) = b_0 + b_1 (affected on time step t-1) + b_2 (proportion of neighbors affected on time step t-1). Third, we incorporate a novel, approximate Steiner tree optimization into the inner loop of our algorithm (optimizing a given time step given the neighboring time steps), allowing the algorithm to scale to much larger graphs with tens of thousands of nodes. Our ongoing evaluation results suggest dramatically improved detection performance for emerging dynamic events which may grow, shrink, or move over time, as well as improved scalability and efficiency.

Task 4: Novel Data Sources for Event Surveillance

The rapid growth and widespread adoption of new technologies such as electronic record systems, mobile phones, sensor networks, Internet search, and user-generated Web content, and the huge amount of data generated by these technologies, present limitless opportunities to apply event detection for the public good. Electronic health records and crime reports are the primary technologies facilitating our health and crime surveillance systems respectively; mobile phones have great potential as an enabling technology for health surveillance in the developing world; and Internet search queries have been used for early detection of influenza. These novel data sources could radically transform the field of event detection, but each also presents new methodological challenges, requiring us to "scale up" detection algorithms to huge numbers of data sources, data aggregations, sensor configurations, and data records respectively, as well as incorporating crowdsourced data from many human users. The past three years' work on Task 4 has primarily focused on Task 4a (prediction using leading indicator data) and Task 4c (incorporating rich text data), as described below. Most recently, we have also worked on detecting anomalous patterns in massive, hierarchical data (for example, multi-resolution digital pathology images), as described in Task 4d below, and on mining Twitter data to detect local-level conflict events (using a novel non-parametric scan statistic for heterogeneous graphs, as described in Task 2c above, but also relevant to Task 4d).

Task 4a: Prediction using leading indicator data (for law enforcement and urban analytics)

While most of our previous work has focused on event detection, we have recently extended this work to **event prediction**, detecting emerging spatial clusters of various types of leading indicators and using the detected clusters to predict that an event is likely to occur in that geographic area. For example, in the law enforcement domain, we have shown that detected clusters of certain minor crimes, or certain types of 911 calls, significantly increase the likelihood that a violent crime cluster will emerge in that area. In past work, we demonstrated that this approach can be used to accurately predict clusters of violent crime between 1 and 3 weeks in advance, by detecting clusters of less serious "leading indicator" crimes. This early warning has the potential to enable police to reduce crime through reallocation of patrols and other targeted interventions, and has been incorporated into our CrimeScan software, which was in day-to-day operational use by the Chicago Police Department (CPD) from 2009-2011.

In the past year, we have developed a new methodological approach and deployable software package, which we call CityScan. CityScan builds on our previous CrimeScan approach, which detected clusters of leading indicators and used these to predict that violent crime will occur nearby, by learning a sparse logistic regression model to predict the probability that a violent crime cluster will occur, as a function of location, the presence of various types of leading indicator clusters nearby in space and time, and other covariates. We are in the process of performing a comprehensive empirical evaluation of CityScan (and comparison to existing prediction methods in the literature) for both crime prediction and 311 call prediction tasks. For the former evaluation, we are using crime offense report and 911 call data supplied by the CPD; for the latter, we will use either publicly available 311 call data from the City of Chicago Data Portal, or a dataset with a much larger variety of call types made available by our collaborators in the City of Chicago Mayor's Office. We anticipate that our results will be complete, and a journal paper submitted for publication, this summer or early fall.

Additionally, we have successfully re-established our collaboration with the CPD (which was on hold for about a year due to a major personnel shake-up on their side), and after successful initial evaluations, they are in the process of rolling out CityScan for day-to-day operational use. Moreover, in collaboration with Brett Goldstein (Chief Data Officer and IT Commissioner of the City of Chicago), we are working to extend our crime prediction work in order to predict many other quantities that are relevant to the city. Our initial focus is on predicting emerging patterns of citizen needs (as measured by clusters of 311 calls for service, such as rodent removal, sanitation complaints, pot holes, graffiti cleanup, and abandoned buildings), and our preliminary results suggest that many types of 311 calls can be predicted accurately one week in advance, using other 311 call types as leading indicators. Future uses of CityScan (and potentially, our other approaches for pattern detection and event prediction) for urban analytics include identifying trends in Twitter data related to the city's public transportation, and check-ins on Foursquare, a location-based social network.

Our future plans (supported by Commissioner Goldstein and the city leadership) are to have our software running in real time to detect relevant trends and patterns which will be directly used by city services. We are very excited about this ongoing collaboration, which has the potential to make a significant, data-driven contribution to city management in practice. We recently collaborated with the City of Chicago Mayor's Office to submit a proposal to the Bloomberg Mayor's Challenge. Chicago was a runner-up in this competition, receiving $1M for "The Chicago SmartData Platform", which proposes to "Partner with leaders in data and computer science to build the first open-source, predictive analytics platform… to harness the power of data to understand underlying trends and better direct limited resources." A major piece of this platform will be the City's use of our CityScan software to predict emerging patterns of violent crime, citizen needs (as measured by 311 calls for service), and other patterns relevant to the city's operations.

One remaining challenge in the CityScan framework is to decide which of the many possible leading indicators are most relevant for predicting a given type of event. Last year, we developed a novel method to **identify spatially localized subsets of leading indicators** for event prediction (Flaxman and Neill, 2012). Given a spatially localized time series to be predicted (e.g. daily counts of violent crime for each census tract) and multiple potential predictors (e.g. daily counts of various types of calls for service

for each tract), our approach maximizes the cross-correlation between the predictor variable and an aggregated subset of leading indicators across a range of time lags, all subsets of potential predictors, and all proximity-constrained subsets of locations. This captures the fact that different subsets of leading indicators may be relevant in different areas of the city. However, even for relatively small numbers of locations and leading indicators, optimization over all such subsets is computationally infeasible, and unfortunately the function we wish to optimize (Pearson correlation between the independent and dependent time series) does not satisfy our linear-time subset scanning property. Instead, we propose a novel "iterative average dot product" method, with the key insight that both spatial subset search and feature selection can be performed by approximating the correlation with a function that can be efficiently maximized over subsets of locations or streams. We then iterate between conditionally optimizing the approximate correlation over subsets of locations (for a given subset of streams) and optimizing over subsets of streams (for a given subset of locations), until convergence to a joint local optimum. Our iterative procedure refines the quality of this approximation over time, approaching the true best correlation. The approach was tested on 311 service calls from Chicago, and compared both to ground truth (for small problem sizes) and existing feature selection methods (such as lasso regression). Our method found near-optimal correlations while scaling to large numbers of locations and data streams, and demonstrated significant improvements in the correlation of detected subsets as compared to existing methods. This work was presented at the International Symposium on Forecasting in June 2012, and at the CMU Workshop on Machine Learning and Social Sciences in October 2012. The full paper is currently under revision, for re-submission to one of the top machine learning/data mining conferences.

In the process of developing this work, however, we realized that cross-correlation is not necessarily the most relevant quantity to optimize for identification of leading indicators, since two event types could have high cross-correlation due to purely spatial and purely temporal correlations (e.g., the rates of fires and violent crimes both tend to be higher in neighborhoods with higher population density and in the summer, but seeing a fire at a given time and place is not necessarily predictive of future, nearby violent crime). Thus we have recently developed new kernel-based tests for space-time interaction in spatio-temporal point processes. Space-time interaction can be thought of as residual space-time dependence, after controlling for purely spatial and purely temporal dependence: if two events are close in space, they are likely to be close in time. We demonstrate that the recently proposed Hilbert-Schmidt Independence Criterion can be used to test whether the joint distribution of points in space and time P(S,T) is separable as the product of two distributions P(S)*P(T), the distributions of points in space and in time, respectively. Unlike previously proposed space-time interaction tests such as Mantel (1967), this non-parametric approach can distinguish arbitrary, rather than just linear, dependencies. We provide a new test for space-time interaction based on HSIC, draw connections to previous tests such as Mantel's, and compare the power of our test to the current state of the art (Diggle's test) on synthetic and real-world data, demonstrating both high power and robustness to parameter specifications. To use this approach for identification of leading indicators, we extended the HSIC-based test to the bivariate case (where we have two separate types of points, such as thefts and homicides) and to only predict forward in time. Finally, we used our directional, bivariate space-time interaction test to identify statistically significant leading indicators for shootings and homicides in Chicago, choosing a small set of

21 out of the 271 different types of emergency calls to 911. This work will be presented at the 2nd Spatial Statistics Conference this summer; we also hope to complete and submit the full paper to a top machine learning conference this summer.

Task 4c: Incorporating rich text data

Typical event detection systems aggregate data records into counts and then detect spatial areas with anomalous recent counts. For example, in disease surveillance, we count the number of disease cases with each of a small set of general symptom categories (such as respiratory, gastrointestinal, and fever) in each zip code for each day. This approach works reasonably well given limited data about each patient, but we believe that outbreak detection can be dramatically improved by incorporating rich text data from electronic health records, e.g. patient histories and chief complaints. Typical disease surveillance systems have difficulty detecting new emerging infections with unknown symptom patterns, or other diseases that do not correspond to the existing symptom categories.

Thus we have developed a new "semantic scan statistic" approach, which uses rich text data to detect previously unknown event types, forming and searching a huge number of aggregated count datasets on the fly (Liu and Neill, 2011; Murray, Liu, and Neill, in preparation). Each count represents the number of records in a given spatial area and time interval which match some set of keywords; different keyword sets are used for each aggregation. Since the number of possible keywords is huge, our challenge is to find the most interesting aggregations and anomalous subsets without an exhaustive search. Our approach uses topic models (created by Latent Dirichlet Allocation) to automatically discover possibly relevant subsets of keywords. We then form a separate count dataset from the case data for each topic, and find the maximum region score over all of the topics considered. Thus the semantic scan statistic provides information not only about whether an event has occurred and which space-time region has been affected, but also which set of keywords (topic) occurs with surprisingly high count in this region.

Our approach assumes that there is a latent "topic" in each case report, and thus we can utilize widely used topic models to extract those "topics" from text and then apply existing spatial scan techniques to them. For example, we might have one patient with "abdominal pain and nausea" symptoms, and another patient who has exhibited "vomiting", but both sets of symptoms might correspond to the same disease category (GI illness). Our first approach used topic models to extract some number of "static" topics from the entire training dataset, and additional "dynamic" topics learned from the current two weeks of test data, in order to capture both broad, typical syndrome categories and newly emerging trends in the recent data. For each day of data, for each of the extracted topics, we formed a count dataset from the case data by computing the number of cases in each zip code for each day which are most likely to correspond to the given topic. We then apply our novel spatial scan methods to the resulting count data, and report the maximum value of the spatial scan statistic over all topics.

Our preliminary results demonstrate that, for disease outbreaks with very specific sets of symptoms, or with novel combinations of symptoms that have not previously been seen in the data, the text-based analysis will enable earlier and more accurate detection than traditional count-based detection approaches. Using a combination of "static" and "dynamic" topics does reasonably well for picking up

patterns of symptoms corresponding to both typical and (simulated) newly emerging illnesses, but some detection power is lost because many of the dynamic topics do not capture sufficiently different syndrome groupings from those represented by the static topics. Thus our most recent approach learns a set of "incremental" topics that represent those trends in the current data which are not well captured by the "static" topics, and we demonstrate that the resulting incremental Latent Dirichlet Allocation approach shows substantially improved detection power for newly emerging illnesses.

Our preliminary work was presented at the 2011 International Society for Disease Surveillance Annual Conference (abstract published in the *Emerging Health Threats Journal*) and at the 2012 International Conference on Digital Disease Detection. Most recently, we have developed a new, EM-based method for classifying cases into topics which improves detection performance as compared to the standard, Gibbs sampling-based method, and are performing an in-depth evaluation and comparison of the different variants of the method. We anticipate submitting the full paper to the IEEE International Conference on Data Mining (ICDM 2013) this summer.

We have also begun collaborating with the NC Detect group at North Carolina's Department of Public Health, who are interested in detecting and characterizing small clusters of related cases (e.g., a cluster of patients coming into a single hospital Emergency during a short time frame) that do not fit existing syndrome groupings. This could include clusters of signs or symptoms, clusters of place names (e.g. mentioning a specific restaurant), clusters of events (e.g. mentioning a specific fair, concert, etc.). We are in the process of adapting semantic scan for this setting: the approach is very similar to our "incremental" method, but makes slightly different assumptions (in particular, that only one novel topic is emerging, and that topic may represent a substantial proportion of cases in the cluster of interest). Thus we perform a separate "incremental" topic modeling for each cluster under consideration, learning a single new topic (which was not well captured by the static topics) from the cases contained in that cluster. We then classify which cases belong to that topic (thus refining the cluster membership), and score the refined cluster by comparing actual and expected counts. We are in the process of implementing this variant of semantic scan and applying it to the NC Detect data.

We are also working to further improve detection power by incorporating spatial information into the topic modeling step (rather than just into the subsequent spatial scan), and are performing an in-depth evaluation and comparison of the different variants of the method. Future work will also apply the semantic scan method to informal, online data sources such as Twitter and evaluate the utility of such sources for disease surveillance.

Task 4d: Incorporating society-scale data

In addition to our work on mining Twitter data to detect civil unrest (described in Section 2c above), we are also working to develop event detection algorithms which can scale to datasets so large that even our linear-time subset scanning methods are computationally expensive or infeasible (Somanchi and Neill, 2013). Such datasets require approximations based on sampling or data summarization. We are currently working to combine LTSS with hierarchical sampling and search techniques, with the goal of

scaling up our current pattern detection techniques to datasets consisting of billions or trillions of records, with provable guarantees on the optimality and/or accuracy of detection.

As an initial step, we are working with UPMC pathologist Dr. Anil Parwani to develop and evaluate a novel Hierarchical Linear Time Subset Scanning (HLTSS) method for detecting regions of interest in massive, multi-scale digital pathology slides. Such images typically consist of 10B pixels or more at the finest resolution, but are stored as multiple "layers" each representing a hierarchical aggregation of data from the previous layer (as a first-order approximation, each "pixel" at layer L can be thought of as averaging the red, green, and blue components from a 4x4 pixel square at layer L-1). HLTSS exploits this hierarchical structure inherent in data produced through virtual microscopy in order to accurately and quickly identify regions of interest for pathologists to review.

For hematoxylin and eosin (HE) stained slide images, a region is interesting when it contains a higher than expected concentration of violet pixels (hematoxylin dye, which stains nuclear material) as compared to pink pixels (eosin dye, which stains cytoplasmic material). There are various applications where identifying this discoloration pattern is useful: identifying regions of inflammation in gastrointestinal tracts for Crohn's disease; finding regions of inflammation (gastritis) in the stomach, which may be indicative of colonization by Helicobacter pylori; and diagnosis of prostatic intraepithelial neoplasia, which may lead to prostate cancer.

The HLTSS method works by first performing a proximity-constrained fast subset scan at a relatively coarse level of the hierarchy, then iteratively refining the detected subset by "expanding" one of the coarse-level pixels in layer L to its component pixels at layer L-1 and re-running the fast subset scan. This fast procedure is then repeated to identify the top-k clusters, and a post-processing step is used to enforce connectivity of the resulting clusters and to merge clusters across the fixed partitions imposed by the hierarchical decomposition of the slide. One additional complication is that the anomalousness of the cluster is determined by applying a likelihood ratio statistic in a linearly transformed (white, pink, violet) color space, in order to identify areas with a higher than expected proportion of violet color. This procedure could be generalized to any choice of three RGB colors, representing background, uninteresting, and interesting pixels respectively, enabling use in many other applications involving event detection in massive multi-scale image data. The full paper is in progress, and we hope to complete this work by late summer or early fall.

Applications of Event Detection

Our work has primarily focused on applications of event detection to three areas (disease surveillance, crime prediction, and detection of anomalous patterns of patient care), but has also been applied to a variety of other domains, including human rights, customs monitoring, network intrusion detection, and infrastructure monitoring. Here we discuss new developments in our work in each of these domains.

The **disease surveillance** domain has served as our primary testbed for the development of new event detection methods; we have obtained real public health data from a number of sources, and our work has been incorporated into multiple deployed biosurveillance systems. The project PI, Daniel Neill, recently published an article on "New directions in artificial intelligence for public health surveillance" in

IEEE Intelligent Systems. This article, based in part on the semantic scan statistic and fast subset scan methods described above, focused on the need for methods that can incorporate rich, unstructured text data and can scale to huge numbers of records and data streams. In addition to continuing to work with our existing Allegheny County Emergency Department data as a testbed for development of new algorithms, and working with the ECADS/ASSET/Data Fusion project team to develop and deploy methods for monitoring of multiple public health data sources in Canada, we are currently working to acquire new data and establish broader collaborations with the Chicago Department of Public Health. As noted above, our first project will involve analysis of sexually transmitted illness data using our new multidimensional subset scan methods. Data acquisition has been delayed due to legal/IRB challenges on the DPH's side, but we are optimistic that these issues can be resolved in the near future. Through the Data Fusion project, we have also received new data from the Ottawa Heart Institute which includes information on patient location and movements within the hospital, along with patient-level information about various hospital acquired illnesses (ventilator-acquired pneumonia, central line infections, Clostridium difficile, and MRSA). We are currently building predictive models of disease transmission, and will soon be applying our multidimensional subset scan and GraphScan methods to detect and track illnesses in this setting. Most interestingly, we hope to apply our graph learning methods to compare the graphs formed by varying mechanisms of disease transmission (direct person-to-person, spatial spread, spread via common caregiver, and spread via shared equipment) for each of these diseases, with the hope of providing actionable information for prevention of the spread of hospital-acquired illness. Finally, we are excited by our new collaboration with the NC Department of Public Health, which will give us the opportunity to develop, evaluate, and deploy our semantic scan methods for analysis of free-text Emergency Department chief complaint data.

As discussed above, we are continuing to work with the City of Chicago (Mayor's Office) and Chicago Police Department in order to predict geographic hot-spots of violent crime, to identify emerging clusters of citizen needs (as measured by 311 calls for service), to monitor various other relevant quantities in real time, and to continue developing the underlying state-of-the-art prediction methods. These projects were discussed as part of Task 4a, prediction using leading indicator data, above. The latest version of our CityScan software is now being used operationally by the CPD (as was its predecessor, CrimeScan), and CityScan will also be deployed by the City of Chicago as part of its "Chicago SmartData Platform", with funding provided by the Bloomberg Mayor's Challenge.

Additionally, with the recent addition of postdoctoral fellow Feng Chen to our Event and Pattern Detection Laboratory, we are embarking on a number of projects related to event detection for conflict prediction and human rights, with initial funding from the MacArthur Foundation. Our first project, described in detail in Task 2c above, focuses on monitoring Twitter data for local-level conflict prediction in multiple Latin American countries. We will also be submitting a concept paper (which proposes to use very similar techniques for early detection of patterns of human rights atrocities) to the "Tech Challenge for Atrocity Prevention" (http://thetechchallenge.org), a contest sponsored by Humanity United and USAID.

For the next year, we are particularly excited about applying and extending our pattern detection approaches (including, but not limited to, the fast generalized subset scan and multidimensional subset

scan methods described above) to discovering anomalous patterns of patient care. We are working create a widely applicable methodological and implementation framework for using massive quantities of healthcare data (including electronic health records and health insurance claims data) to discover patterns of care with significant potential impacts on patient outcomes (e.g., mortality, length of stay, and readmissions) and healthcare costs. The goal is to enhance the current practice of evidence-based medicine by supplying a systematic basis for the initial steps of pattern discovery and hypothesis generation: identifying currently widespread care practices that are potentially suboptimal, or alternative care practices with potential to improve health outcomes and reduce costs, that can then be rigorously evaluated for potential use by the medical community. The proposed system will evaluate, extend and deploy our state-of-the-art pattern detection methods, in order to automatically detect substantial variations in care between groups which have significant impacts on patient outcomes. These impacts can either be negative (e.g. systematic errors), in which case we can detect and correct these sub-optimal patterns of care, or positive. In the latter case, our system will have discovered a new potential best practice, which can then be investigated further, and if appropriate, shared with other groups.

The first stage of the project will focus on the inpatient setting, applying our novel pattern detection methods to electronic medical records (EMR) data in order to discover potentially relevant patterns of care with significant positive or negative impacts on patient outcomes. This stage has already received seed funding from a UPMC Healthcare Technology Innovation grant, and UPMC has committed to providing data consisting of thousands of patients with an entering diagnosis of chronic obstructive pulmonary disorder (COPD). We will analyze this data to discover anomalous patterns of care for these patients that influence outcomes such as length of stay, readmission, mortality, and nosocomial illness (e.g. ventilator-acquired pneumonia and multidrug-resistant infections). Due to personnel turnover on the UPMC side, as well as delays with the data use agreement between CMU and UPMC, our data acquisition has been continually delayed, but we hope to receive the data in the near future. We are also planning to partner with Presence Healthcare (a major health provider in the Chicago area, with12 hospitals, 29 long term care and senior residential facilities, numerous outpatient services and clinics, etc.) on a related project to identify and investigate anomalous patterns of care for patients with congestive heart failure. (A data sharing agreement is in progress.)

The second stage of the project will focus on identifying care practices which are cost-effective as well as improving outcomes, by integrating claims data with EMR, and treating cost of care as an additional outcome to be optimized. Claims data, while providing somewhat more limited information than the EMR data, has the advantages of having much more structure (making it easier to use within our existing methodological framework without needing to analyze large quantities of free text) and providing detailed cost information.

The third stage of the project will focus on extending the work in three main directions: First, we will attempt to understand the spread of hospital-acquired infections by using the methods above to investigate the effect of exposure to potential sources of illness in the hospital environment. Second, we will scale up our methods to monitor massive quantities of patient data within a hospital system in real time, enabling rapid responses to emerging trends and patterns in the inpatient setting. Here, the

discovered patterns could include emerging outbreaks of hospital-acquired illness, systematic errors in care (e.g. poor hand-washing practices), patterns of adverse events, or new trends in care practices. Third, we will evaluate the feasibility of extending health pattern discovery beyond the inpatient setting to outpatient clinics and preventive care.

This project is currently in progress, and we are continuing to advance the underlying pattern detection methodology while waiting for data from our potential partners. In the meantime, we have written an overview paper on "Using artificial intelligence to improve hospital inpatient care", which describes both past related work and our ongoing work on the project described above, and this paper has been accepted for publication in *IEEE Intelligent Systems* (Neill, 2013, in press)

Educational Activities: The Machine Learning and Policy (MLP) Initiative

With the critical importance of addressing global policy problems ranging from disease pandemics to crime and terrorism, and the continuously increasing size and complexity of policy data, the use of machine learning has become increasingly essential for data-driven policy analysis and for development of new, practical information technologies that can be directly applied for the public good. The numerous challenges facing our world will require broad, successful innovations at the intersection of machine learning and public policy. This endeavor will require widespread collaboration between machine learning and policy researchers, increased emphasis on the education of future researchers with in-depth knowledge of both disciplines, and a broadly shared research focus on developing novel machine learning methods which directly address critical policy challenges. Thus this project's main educational focus is on a multi-pronged curricular program, the **Machine Learning and Policy (MLP) initiative**. This program will facilitate the widespread use of machine learning methods for the public good by incorporating machine learning throughout the public policy curriculum. The components of this program currently include the Joint Ph.D. Program in Machine Learning and Public Policy, a master's level course on Large Scale Data Analysis for Public Policy, a Ph.D.-level Research Seminar in Machine Learning and Policy, and a series of courses, Special Topics in Machine Learning and Policy, all discussed in detail in last year's report, as well as a workshop and seminar series in Machine Learning and Social Sciences.

Our first student in the Joint Ph.D. program in Machine Learning and Public Policy is now completing his second year of doctoral study, and has been extremely productive as a member of the PI's Event and Pattern Detection Laboratory, recently winning Heinz College's Suresh Konda Best Paper Award. Additionally, the PI has been actively involved in coordinating the joint Ph.D. program admissions and in leading a faculty search in "Societal-Scale Data Analysis" which successfully recruited an additional tenure-track faculty member in this area. MLP courses taught by the PI in the past year included Large Scale Data Analysis for Public Policy, the MLP Research Seminar, and a new Special Topics in MLP course focusing on "Mining Massive Datasets", i.e., applying machine learning to Web- and societal-scale datasets that are too large for standard, basic machine learning techniques to be applied.

In the near future, in addition to continuing with course and program development, directing the joint Ph.D. program, and being involved with additional faculty hiring, we intend to create an integrated web

resource for Machine Learning and Policy at CMU, and to develop a workshop series which will spread these ideas outside the university. As a first step toward wider dissemination of these ideas, the PI collaborated with several other CMU faculty members to put together an internal workshop on "Machine Learning for the Social Sciences" which was held in October 2012, with invitees across many CMU departments. The workshop was limited to CMU faculty and PhD students, and included talks and brainstorming on cross-disciplinary collaborations between machine learning and social science faculty. (The PI also gave an invited talk at this workshop, on "Predicting and Preventing Emerging Outbreaks of Crime", along with leading that afternoon's strategy discussion). The success of and enthusiasm generated by this workshop has led to our creation of a monthly series of research talks by CMU faculty and students, with the goals of information sharing and facilitating collaborations between ML and social science faculty and students. Finally, the PI was a member of the Heinz College committee which developed a new "Policy Analytics" track for our Master of Science in Public Policy and Management Program, which will help to get more master's students involved in applied research at the intersection of machine learning and policy.

Educational Activities: The Event and Pattern Detection Laboratory

Over the past three years, thanks in large part to the generosity of the National Science Foundation, the PI has founded and directs the **Event and Pattern Detection Laboratory** (EPD Lab) at CMU. The Lab currently consists of Dr. Neill and six fellows/students for whom he is the principal research advisor: one post-doctoral fellow (Dr. Feng Chen), three Heinz College Ph.D. students (Skyler Speakman, Edward McFowland III, and Sriram Somanchi), one student in the Joint Ph.D. Program in Machine Learning and Policy (Seth Flaxman), and one SCS master's student (Kenton Murray). All of these students are working on research highly relevant to the goals of this project, as described in the sections above, and at least two additional students will join the project in the coming year. Strong emphasis is placed on student mentoring and group cohesion: Dr. Neill meets individually with each fellow/student on a weekly basis, along with bi-weekly lab meetings and occasional social events. Eight additional master's students involved with this project have successfully completed their studies in the past two years, including one LTI student (Yandong Liu), four MSIT-VLIS students (Amrut Nagasunder, Tarun Kumar, Xin Wu, and Kai Liu), one secondary MS in Machine Learning student (Kan Shao), and two MISM students (Rajas Lonkar and Yating Zhang). All of these students were very successful in their job search and now have positions in industry, government, or Ph.D. programs. Notable accomplishments by EPD Lab students related to this project include:

- EPD Lab students have **won three of the last five "Best Paper Awards"** given each year by Heinz College to the best "First Heinz Research Paper" (presented by 2nd year students) and "Second Heinz Research Paper" (presented by 3rd year students):
    - Seth Flaxman won the 2013 Suresh Konda Award ("Best First Heinz Research Paper") for his paper, "Correlates of homicide: new space/time interaction tests for spatiotemporal point processes" (Task 4a).
    - Sriram Somanchi won the 2013 George Duncan Award ("Best Second Heinz Research Paper") for his paper, "Discovering Anomalous Patterns in Large Digital Pathology Images" (Task 4d).

    - Ed McFowland won the 2012 Suresh Konda Award ("Best First Heinz Research Paper") for his paper, "Fast Generalized Subset Scan for Anomalous Pattern Detection" (Task 2a).
    - Also, Skyler Speakman was the runner-up for the 2011 George Duncan Award ("Best Second Heinz Research Paper", for his paper, "Pattern Detection with Temporal Consistency and Connectivity Constraints" (Task 3c).
- Research related to this project has led to the completion of seven First and Second Heinz Research Papers (by Skyler Speakman, Ed McFowland, Sriram Somanchi, and Seth Flaxman), three MSIT Capstone Projects (by Amrut Nagasunder, Tarun Kumar, and Xin Wu), and Kan Shao's Data Analysis Project (equivalent to a master's thesis for the MS in Machine Learning).
- Two EPD Lab students (Skyler Speakman and Ed McFowland), currently pursuing their PhDs in Information Systems at Heinz College, were admitted into CMU's secondary MS program in Machine Learning. One additional student (Sriram Somanchi) will be applying this year.
- Two EPD Lab students (Ed McFowland and Sriram Somanchi) have been awarded fellowships to participate in this year's "Data Science for the Social Good" summer program.

While we have presented multiple papers and talks under the EPD Lab banner in the past year, in the next year we hope to give the lab greater visibility by putting together a high-quality website (to replace our temporary website, currently linked off of the PI's home page), making more of our software publicly available, and inviting our faculty research collaborators (and the students they are advising) to become affiliated with the lab.

Outreach and Professional Activities:

Dr. Neill has given multiple guest lectures related to his work on this project in the Heinz Ph.D. seminar, Heinz Faculty Research Seminar, SCS Immigration Course, and Workshop on Machine Learning and Social Science at CMU, as well as guest lecturing in multiple CMU courses. He also was the PI of a $10M, university-wide grant proposal effort (leading a team of 37 CMU faculty members across 21 university departments) to create a Center for Development Data Analytics, focused on applying state-of-the-art machine learning methods to various problems in international development. This proposal was well received by the U.S. Agency for International Development (USAID), receiving very positive feedback and reaching the final round of selection. Though it was not funded at that time, USAID is now reconsidering some proposals that reached the final round, including ours. Also, one piece of this work (focusing on Dr. Neill's event detection and prediction work, as applied to detection of patterns of human rights abuses, violence, and conflict) has received initial funding from the MacArthur Foundation.

As noted above, Dr. Neill helped to organize a CMU workshop on "Machine Learning for the Social Sciences" in October 2012, along with a follow-up lecture series that started in Spring 2013 and will continue in the 2013-14 academic year. Additionally, he served as Scientific Program Chair for the 2011 International Society for Disease Surveillance Annual Conference. As program chair, his responsibilities included choosing and directing the Scientific Program Committee, coordinating the abstract submission and review process, and inviting speakers, with principal responsibility for the scientific content of the conference. The conference was an outstanding success, with 321 registered participants, 96 contributed talks, 65 posters, and 10 invited talks. Dr. Neill has also been serving as the AI and Health

Department Editor of *IEEE Intelligent Systems*, with the goal of presenting novel research and exciting applications at the intersection of artificial intelligence, machine learning, and health. In the past year, he also served on two expert panels: "Urban Analytics and Neighborhood Health: Bridging Social and Computer Science Perspectives to Inform Research, Policy, and Practice", organized and hosted by the MacArthur Foundation (Chicago, IL, May 2012), as well as the Subcommittee on Youth Violence for the Advisory Committee to the Social, Behavioral and Economic Sciences Directorate, National Science Foundation (Washington, DC, February 2013). The subcommittee's report, commissioned by Congressman Frank Wolf in response to the shooting at Sandy Hook Elementary, and focusing primarily on risk factors associated with mass shootings, was presented to Congress by Dr. Brad Bushman and NSF Director Dr. Subra Suresh, and was discussed at a hearing before the House Appropriations Commerce-Justice-Science (CJS) subcommittee. Dr. Neill's role focused primarily on identifying potential uses, challenges, and open questions for future research in event prediction, and the potential roles of machine learning and data mining in understanding the signaling behavior that precedes rare events such as mass shootings. The full report, "Youth Violence: What We Need To Know", is available at: http://wolf.house.gov/uploads/Violence_Report_Long_v4.pdf.