

---

## 1. Research Strategy

The major theme of my current research is “Machine Learning and Event Detection for the Public Good.” This research agenda is focused on the development of new statistical and computational methods for the discovery of emerging events and other relevant patterns in complex and massive datasets, as well as the application of these methods to a variety of policy problems ranging from medicine and public health to law enforcement and security. My pattern detection work primarily focuses on three main application areas: **disease surveillance** (e.g., using electronically available public health data such as hospital visits and medication sales to automatically identify and characterize emerging outbreaks), **law enforcement** (e.g., detection and prediction of crime patterns using offense reports and 911 calls), and **health care** (e.g., detecting anomalous patterns of care which significantly impact patient outcomes). I have also applied this work to numerous other areas, including network intrusion detection, customs monitoring, infrastructure monitoring, and economic growth.

Most of these applications fall into the general paradigm of **event detection**: monitoring multiple streams of spatially localized time series data and searching for anomalous patterns that are indicative of emerging (and relevant) events. In addition to detecting such events, we wish to characterize these events by identifying the type of event (for example, distinguishing an influenza outbreak from a bio-terrorist anthrax attack) and also identifying the affected subset of the data (pinpointing the spatial region affected by the event, its time duration, and which data streams were impacted). I have also extended these methodologies to general **pattern detection** approaches which can be applied not only to event detection, but to the more general question of finding any anomalous, interesting, or relevant patterns in massive datasets, including application areas such as fraud detection, scientific discovery, and the detection of suspicious patterns in container shipping data. The key concept for these approaches is **subset scanning**: the pattern detection problem can be treated as a search over subsets of the data and over possible pattern types, evaluating the score of each subset according to how likely it is to correspond to the given pattern, and finding the highest scoring combinations of subset and pattern. While there are often exponentially many subsets of the data to consider, in many cases this search can be performed efficiently by employing efficient data structures, as in our “fast spatial scan” algorithm, or using novel properties such as “linear-time subset scanning” (both discussed below) to find the most likely patterns without searching over all possible subsets. In past work, my development of subset scan methods has resulted in five major methodological advances:

- 1) The expectation-based scan statistic enables more timely and accurate detection of events through better use of spatial and temporal information.
- 2) The nonparametric and Bayesian scan statistics further improve detection power by integrating information from multiple data streams.
- 3) The multivariate Bayesian scan statistic (MBSS) incorporates prior information and historical data to accurately model and differentiate between multiple types of events.
- 4) New pattern detection methods such as Anomalous Group Detection (AGD) and Anomaly Pattern Detection (APD) enable accurate detection of patterns in general datasets.
- 5) New methods for “linear-time subset scanning” (LTSS) and “fast subset sums” (FSS) enable very fast detection of the most anomalous patterns, either in spatial datasets or in general multivariate data.

---

These methods have been applied to various domains including disease surveillance, crime prediction, network intrusion detection, and customs monitoring, and we have demonstrated substantial improvements in the timeliness, accuracy, and specificity of pattern detection and characterization in each domain.

I have recently received three major funding awards from the National Science Foundation, including an NSF CAREER award. These grants include: “Fast Subset Scan for Pattern Detection” (three years, \$500K), “Discovering Complex Anomalous Patterns” (four years, \$2.6M, in collaboration with four other CMU and Pitt faculty), and “CAREER: Machine Learning and Event Detection for the Public Good” (five years, \$530K). Additionally, I have received additional funding from the University of Pittsburgh Medical Center for two additional projects, “Detecting Anomalous Patterns in Health Care Data Streams” (as PI) and “Information Visualization for Cognitively Guided Decision Making for Diabetes Risk Assessment and Guideline Compliance” (as co-PI). These grants have helped to shape my research and educational agenda, as well as providing me with the financial resources to hire new students and personnel to work on these research projects.

The six main focus areas of my ongoing research are:

1. New methodological contributions for fast subset scanning which allow event detection methods to scale up to large and high-dimensional data without sacrificing accuracy or flexibility (funded by FSS, CAREER; described in Section 1.1.1).
2. 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 (funded by DCAP, CAREER; described in Section 1.1.2).
3. End-to-end methods that augment the automatic detection of events by providing methodological tools for event characterization, explanation, visualization, investigation, and response (funded by CAREER; described in Section 1.1.3).
4. New methods which incorporate data from emerging, transformative technologies (e.g. mobile phone and Internet search data) and address the fundamental scalability challenges (funded by CAREER; described in Section 1.1.4).
5. Continued development of the underlying statistical and computational methods for event detection (described in Section 1.1.5).
6. Deployment and evaluation of methods to enhance public health, safety, and security (described in Sections 1.1.6-1.1.9).

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. Additionally, I have

---

founded and am directing the **Event and Pattern Detection Laboratory** (EPD Lab) at CMU to pursue these research goals.

An additional thrust of my research is the development of information systems to improve health care policy and management. As co-director of the Healthcare Information Technology (HIT) research thrust of Heinz College's iLab, I am currently working with other researchers and public health departments on the development and deployment of practical systems for disease surveillance, as discussed in detail in Section 1.1.6. However, many other questions in health care policy, rather than focusing on pattern detection, require the application of other machine learning paradigms such as prediction and modeling, as well as statistical tools for data analysis and visualization. These projects are described in Section 1.1.10. Finally, some of the other research projects I have been involved with, including contributions to the fields of machine learning, game theory, and cancer prevention, are described in Sections 1.1.11 through 1.1.13. All of these methods and applications are discussed in detail below, along with our major results and publications.

## **1.1. Descriptions of Specific Research Projects**

The following thirteen subsections describe some of the specific contributions of my past, present, and planned future research. The first four subsections focus on my current research thrusts: fast subset scanning for scalable event detection (1.1.1), incorporating learning into detection (1.1.2), end-to-end surveillance (1.1.3), and methodologies to incorporate novel data sources (1.1.4). Each of these sections is further subdivided into specific research projects ranging from detection and learning on graphs (1.1.1.5, 1.1.2.3) to scalable visualization (1.1.3.1), text-based event detection (1.1.4.1), and many more. Section 1.1.5 discusses my past and ongoing contributions to the development of new statistical and computational methods for event detection, ranging from fast spatial scan algorithms (1.1.5.1) to parametric (1.1.5.2), nonparametric (1.1.5.3), Bayesian (1.1.5.4), Bayesian network (1.1.5.5), and rule-based (1.1.5.6) methodologies. The next four subsections focus on applications of my event detection methods to disease surveillance (1.1.6), law enforcement (1.1.7), patient care (1.1.8) and other domains (1.1.9). While these first nine subsections discuss work in my core research area of event and pattern detection, the last four subsections discuss my projects in other areas, including health care information systems (1.1.10), game theory and the evolution of behavior (1.1.12), and other topics (1.1.11 and 1.1.13).

### **1.1.1. Fast Subset Scanning for Scalable Event Detection**

*This research thrust is the major focus of my recently awarded NSF grant, “Fast Subset Scan for Anomalous Pattern Detection,” [95] and is also one of the four proposed research tasks for my NSF CAREER award [93]. We propose to develop new event and pattern detection methods which can “scale up” to massive datasets.*

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, I have recently discovered [42, 62] 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 computationally efficient event detection, removing some of the major computational barriers faced by subset scanning methods. For example, we were able to find the most anomalous subset of 97 Allegheny County zip codes for a disease surveillance task (monitoring hospital Emergency Department visits with respiratory chief complaints) in approximately 40 milliseconds using LTSS, while a naïve approach (exhaustive search over all  $2^{97}$  subsets of locations) would require over  $10^{20}$  years.

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. These methods are discussed in detail in the following subsections.

#### 1.1.1.1. Univariate LTSS with Proximity Constraints

My initial work on linear-time subset scanning [42, 62] focused on the univariate LTSS framework, proving that many commonly used scan statistics satisfy the LTSS property (as discussed in Section 1.1.1.2 below), and incorporating **spatial proximity constraints**. 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, either restricting or penalizing the neighborhood size. 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  $k$ . In practice, this allows us to perform spatial detection tasks in milliseconds that would require thousands of years for an exhaustive search. Finally, we examine the detection power and spatial accuracy of our fast subset scan approaches as compared to the traditional Kulldorff scan statistic (searching over circular regions), using simulated disease outbreaks injected into real-world hospital Emergency Department data. We demonstrate that proximity-constrained subset scans substantially improve the timeliness and accuracy of detection, detecting two days faster with fewer than half as many missed outbreaks. This work was presented at the *International Workshop on Applied Probability* [85] and the *International Society for Disease Surveillance Annual Conference* [42]; an abstract was published in the journal *Advances in Disease Surveillance* [42], and the full paper has been accepted to the *Journal of the Royal Statistical Society (Series B: Statistical*

---

*Methodology*) [62]. Future extensions of this work will investigate other spatial constraints (e.g. shape and convexity), consider “soft” constraints which allow but penalize irregular shapes, and incorporate additional constraints (such as temporal consistency) made possible by stronger but more restrictive versions of the LTSS property (“strong LTSS” and “additive LTSS”) discussed below.

### 1.1.1.2. Theory of Linear-Time Subset Scanning

In the recent paper [62] described above (accepted to the *Journal of the Royal Statistical Society*), I present various proofs of the linear-time subset scanning property, thus demonstrating that a wide range of scan statistics satisfy LTSS and can be used efficiently for scalable event detection. Some of the **sufficient conditions** for LTSS to hold include:

- 1) Let  $F(S) = F(X, Y)$  be a quasi-convex function of two additive, non-negative sufficient statistics of subset  $S$ . If  $F(S)$  is monotonically increasing or decreasing with either  $X(S)$  or  $Y(S)$ , then  $F(S)$  satisfies the LTSS property. As a corollary, we can show that Kulldorff’s original spatial scan statistic satisfies LTSS.
- 2) Let  $F(S)$  be an expectation-based scan statistic for any distribution in the separable exponential family (Poisson, Gaussian, exponential, variance of a Gaussian with known mean, etc.) Then  $F(S)$  satisfies LTSS. As a corollary, the commonly used expectation-based Poisson, Gaussian, exponential, and Gaussian-variance scan statistics satisfy LTSS.
- 3) Let  $F(S) = F(X, |S|)$  be a function of one additive sufficient statistic of subset  $S$  and the cardinality of  $S$ . If  $F(S)$  is monotonically increasing with  $X(S)$ , then  $F(S)$  satisfies the LTSS property. Moreover,  $F(S)$  also satisfies the **strong LTSS** property, which allows us to find the optimal subset  $S$  among those subsets with any given cardinality. As a corollary, we can show that our nonparametric scan statistics [44, 103] satisfy LTSS and strong LTSS, as do the expectation-based exponential and Gaussian-variance scan statistics. Functions satisfying strong LTSS allow some useful optimization approaches that functions which only satisfy weak LTSS do not, such as efficient constrained optimization over subsets with hard constraints on region density. Interestingly, however, most commonly used scan statistics only satisfy the weak but not strong LTSS property; nevertheless, the weak property is sufficient for very efficient optimization over subsets of the data.

### 1.1.1.3. LTSS for Multivariate Spatial Datasets

With Heinz Ph.D. student Edward McFowland III and former Heinz master’s student Huanian Zheng, I have extended the linear-time subset scanning approach from univariate to **multivariate spatial datasets** [63]. The key insight behind this work 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 due to

---

Burkom (2003), 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 compared the detection performance of the Burkom and Kulldorff variants of the multivariate spatial scan for synthetic and real-world disease surveillance datasets. For both methods, we compared our “fast localized scan” approach (searching over all subsets of locations constrained by spatial proximity) to the traditional spatial scan approach (searching over circular regions). We demonstrated that the fast localized scan significantly improved detection power and spatial accuracy for both Burkom and Kulldorff methods, while maintaining efficient and scalable computation. Comparing the Burkom and Kulldorff methods, we demonstrated that Kulldorff's method tends to achieve somewhat higher detection power and spatial accuracy when data streams are affected to differing extents. However, when only a subset of streams is affected, Burkom's method more accurately characterizes events by identifying the affected streams. Thus Kulldorff's method may be preferable when event detection is the primary goal, while Burkom's method may be preferable when event characterization is paramount. Our fast, scalable algorithms enable either method to be effectively and efficiently applied to massive, high-dimensional datasets. This work was presented at the ENAR 2010 [80] and ISDS 2010 [34] conferences, and was also included as part of my invited talks on multivariate surveillance at the Data Fusion Research Meeting [74] and the 2011 Joint Statistical Meetings [72]. The full journal paper [63] was submitted to the journal *Statistics in Medicine*, and is currently under review.

With MSIT-VLIS student Tarun Kumar, I have recently extended the Burkom multivariate LTSS approach from matrix data (records x attributes) to **tensor data** with an arbitrary number of modes [107]. The approach is a natural generalization of our previous algorithm, in which we randomly initialize the algorithm then iteratively optimize over subsets of each tensor mode given the other modes. This process converges to a local maximum of the score function, and then multiple randomized restarts can be used to approach the global maximum. We are in the process of evaluating the utility of this approach: in the disease surveillance domain, we believe that it will improve detection power by enabling us to simultaneously search over space-time regions, subsets of the monitored data streams, and subpopulations with different demographic or behavioral characteristics (e.g. age groups, gender, socioeconomic status, and race/ethnicity), thus increasing our ability to detect disease outbreaks which have different impacts on different subpopulations. This will be extremely useful in our future work on detecting patterns of chronic illness as well as anomalous patterns of patient care in a hospital setting, described below.

---

#### 1.1.1.4. LTSS for General Multivariate Datasets

In joint work with Heinz Ph.D. student Edward McFowland III, I have also developed an extension of the LTSS approach which enables efficient pattern detection in general multivariate datasets [71]. 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 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 our nonparametric scan statistic (described in Section 1.1.5.3 below) to detect 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 as above, 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 resulting “Fast Generalized Subset Scan” (FGSS) method has been evaluated on multiple 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. FGSS was shown to consistently outperform our previously proposed Anomaly Pattern Detection (APD) method [20], and scales to much larger datasets than our previously proposed Anomalous Group Detection (AGD) method [68], thus providing a novel and useful approach for accurate and efficient pattern detection in massive, high-dimensional data. The FGSS approach was presented at the INFORMS 2010 [77] and CAARMS 2010 conferences; the full journal paper is nearly complete and will be submitted shortly to the *Journal of Machine Learning Research* [71].

While our first paper on LTSS for general data focuses on the case of categorical datasets, we are also working to extend this approach to the case of mixed real- and categorical-valued datasets [110]. In the simplest case, we assume that each real-valued attribute is independent of the other attributes, 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. A more challenging case occurs when we wish to model the conditional dependencies between the real-valued and categorical-valued attributes. In this case, we learn a regression tree to estimate the value of each real-valued attribute conditioned on the other attributes. Then the empirical p-value 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. Other ongoing work on FGSS includes the extensions to multiple models and multiple local anomaly detectors [111] described in Section 1.1.2.4 below.

---

### 1.1.1.5. LTSS with Connectivity Constraints for Graph and Network Data

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 a group of nodes is contained in the optimal subset  $S$ , and if removing that group does not disconnect the subgraph, any neighboring group of nodes with higher priority will be contained in  $S$  as well. We first developed a preliminary method which uses this rule to scale up to graphs with 100 nodes, requiring approximately two minutes of run time per day of data; for graphs with high connectivity (where we can perform a nearly unconstrained scan) or low connectivity (where there are few subsets to search), the algorithm scales approximately linearly, but its complexity is still exponential in the worst case. This preliminary GraphScan work was presented at the 2009 ISDS Annual Conference [39], in collaboration with Heinz Ph.D. student Skyler Speakman, and was successfully presented as Skyler's First Heinz Research Paper.

Since the ISDS presentation, we have further improved GraphScan by integrating branch-and-bound pruning techniques, leading to an approximately 100x speedup as compared to the original GraphScan approach (i.e. only one second is required to process each day of data). Additionally, we developed a new variant of the GraphScan algorithm which relies on a depth-first backtracking search rather than evaluating nodes in priority order. This latest version is able to scale to graphs of several hundred nodes, and we have demonstrated that both detection power and spatial accuracy are substantially improved when incorporating both connectivity and proximity constraints as compared to proximity constraints alone. 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. In addition to the ISDS 2009 presentation [39] mentioned above, our GraphScan algorithm was also presented at INFORMS 2010 [78]; the full paper is nearly complete, and will shortly be submitted to the *Journal of Machine Learning Research* [70].

In collaboration with MISM student Rajas Lonkar, Skyler and I have also extended GraphScan to the multivariate case [109], by embedding the univariate constrained optimization step (maximizing over connected subsets) within the multivariate Burkomp linear-time subset scanning algorithm described above. In addition to using GraphScan to monitor multivariate spatial data with connectivity constraints, e.g. for multivariate disease surveillance, we can also enforce connectivity constraints on the attribute space, for example, incorporating prior knowledge as to the hierarchy of International Classification of Disease (ICD)-9 codes to infer which patterns of symptoms are likely to co-occur. The GraphScan optimization step can also be embedded into our tensor scan approach [107] described above, thus allowing us to place a connectivity constraint (with a given 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 [108],

---

as described in Section 1.1.1.6, and to learn the underlying graph structure [64, 104], as described in Section 1.1.2.3.

### 1.1.1.6. Incorporating Temporal Consistency and Other Soft Constraints

In recent work with Heinz Ph.D. student Skyler Speakman, I have developed a novel method for incorporating **soft constraints** into our linear-time subset scanning framework [108]. 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.

For example, consider the penalized function  $F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$ , where each bonus or penalty  $\Delta_i$  represents an arbitrary bonus or penalty for including location  $s_i$  in the detected subset. If  $F(S)$  is a log-likelihood ratio scan statistic, then each value  $\Delta_i$  can be thought of as the prior log-odds of including that location in the detected subset. This allows multiple soft constraints to be included: for example, a size penalty can be used by choosing  $\Delta_i = \Delta < 0$  for all locations  $s_i$ . Alternatively, conditioning on the center of the spatial region, each  $\Delta_i$  could be a function of that location's distance from the center, thus enforcing a soft constraint on spatial proximity (as opposed to arbitrarily restricting the neighborhood size, as in the fast localized scan method described above).

Most importantly, 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). Our current approach [108] 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; we are currently working to extend this to an iterative approach which enables propagation of information both backward and forward in time, and believe that this extension will dramatically improve detection performance.

As noted above, we consider the penalized function  $F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$ . Unfortunately, however, this penalized function does not satisfy the LTSS property for most commonly used spatial scan statistics  $F(S)$ . 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  $F(S | q)$  can be written as an additive function,  $F(S | q) = \sum_{s_i \in S} G_q(s_i)$ . For such functions satisfying the **additive LTSS property**, we can write the penalized form  $F_{\text{pen}}(S | q) = \sum_{s_i \in S} (G_q(s_i) + \Delta_i) = \sum_{s_i \in S} H_q(s_i)$ , 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  $F_{\text{pen}}(S | q)$  over all subsets of the data, by choosing all and only those locations with positive values of  $H_q(s_i)$ . However, we wish not only to maximize  $F_{\text{pen}}(S | q)$  for a given  $q$ , but to jointly maximize  $F_{\text{pen}}(S | q)$  and  $q$ , thus maximizing the original penalized function  $F_{\text{pen}}(S)$ . We have developed two distinct approaches: for the univariate case, we can condition on the range in which  $q$  falls, exactly optimizing  $F_{\text{pen}}(S)$  for each such range. We show that only a linear number of ranges must be considered, thus enabling us to exactly optimize  $F_{\text{pen}}(S)$ . For the multivariate case, we must iterate between computing the optimal penalized subset  $S$  given the values of  $q$  for each monitored data stream, and computing the optimal values of  $q$  given the current subset, as in the Kulldorff multivariate linear-time subset scanning approach described above. This latter approach is only guaranteed to converge to a local, rather than global, maximum, but enables incorporation of multiple data streams and fits easily into our GraphScan approach, thus incorporating (hard) connectivity constraints as well as (soft) temporal consistency constraints into the detection framework.

Our first use of the additive LTSS property was 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** [108]. This work was successfully presented as Skyler’s Second Heinz Research Paper, and will be submitted for publication within the next few months. Our results 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.

### 1.1.2. Incorporating Learning into Event Detection

*This research thrust is the major focus of our recently awarded NSF grant, “Discovering Complex Anomalous Patterns,” [97] and is also one of the four proposed research tasks for my NSF CAREER award [93]. We propose to develop new methods for detecting, characterizing, explaining, and learning models for both known and previously unknown events, creating an integrated system for pattern detection, learning, and discovery.*

Our second 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.

In past work, we have developed several new methods which incorporate supervised and semi-supervised learning into our multivariate Bayesian scan statistic (MBSS) framework [8]. In [45], we show that the spatial distribution of each event type, and its average effects on each monitored data stream, can be learned accurately from a small number of fully labeled training examples by maximum likelihood estimation, substantially improving the timeliness and accuracy of detection. In [22, 41], we learned a “latent center” model for the spatial distribution of outbreaks. Our model representation has only a small number of parameters (and thus can be learned from a small amount of data) but is expressive enough to model many possible distributions of outbreak sizes, shapes, and commonly affected areas.

These results are promising and demonstrate the potential benefits of incorporating learning into detection. However, additional work is needed to develop and incorporate learning methods when examples are partially labeled or when complex models must be learned. For example, in our generalized fast subset sums framework for Bayesian event detection (see Section 1.1.3.1 below), we can learn event models which differ in their spatial extent, density or sparsity, and their relative effects on the different monitored data streams. In other work (discussed in Section 1.1.2.3), we have investigated how the use of our GraphScan method described above (typically used to detect the most anomalous connected subgraph for a known graph structure) can also be used to learn the underlying graph for detection. Additionally, it is essential 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. My current research thrusts for incorporating learning into event detection include the following:

#### **1.1.2.1. Learning Complex Models**

We propose to learn more complex model specifications and incorporate these models into the detection process, including dynamic models which allow regions to evolve over time and synergistic models which consider the combined effects when multiple events simultaneously affect the same locations. In recent work with former MLD graduate student Kaustav Das and Prof. Jeff Schneider, we focus on learning temporal patterns for the effects of different event types. We demonstrate that learning temporal models can enable our multivariate Bayesian scan statistic method (described in Section 1.1.5.4 below) 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). This work was part of Kaustav’s doctoral thesis, and 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 and characterization of events that can grow, shrink, or move over time.

---

### 1.1.2.2. Learning from Partially Labeled Data

Much of our past event detection work focuses on learning from fully labeled data, where the event type  $E_k$  and affected space-time region  $S$  are completely and correctly specified by the user. More commonly, we must deal with the case where data is only partially labeled: the user might fail to specify the event type or affected region, or might label the events and locations presented by the system without examining the rest of the data. We propose an iterative approach based on expectation-maximization (EM), where we alternately compute the posterior distribution of labels given our models, and compute the maximum likelihood values of the model parameters. In our work described in Section 1.1.2.1 above, we present preliminary results on learning event models from partially labeled data using EM, showing that we are able to accurately detect the full spatial extent of a disease cluster when the user labels a small number of affected zip codes. We are currently working to extend the learning procedure for our computationally efficient, Bayesian “Generalized Fast Subset Sums” approach [65] (discussed in Section 1.1.3.1) from fully to partially labeled training examples using similar techniques [105].

### 1.1.2.3. Learning the Underlying Graph Structure for Event Detection

As noted in Section 1.1.1.5 above, we have developed a new, fast algorithm (“GraphScan”) for exactly and efficiently detecting the most anomalous connected subgraph in graph or network data [39, 70]. While the GraphScan algorithm typically assumes a known graph structure, in joint work with Heinz Ph.D. student Sriram Somanchi, I am currently investigating the use of GraphScan for **learning an unknown graph structure** from data [64, 104]. 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.

In a recent paper submitted to the IEEE International Conference on Data Mining [64], and currently under review, we proposed a method to learn the underlying graph structure that best represents a given time series of data observed at each node of the graph. For each of a large set of training examples, the observed data is assumed to be generated by a mixture distribution such that some subset of nodes have been affected by an anomalous process, spreading over the (unknown) underlying graph structure, while the remaining nodes are generated according to their usual (background) distribution. However, we assume that the data is **unlabeled**, and thus the affected subset of nodes for each training example is not provided to the learning algorithm.

Our solution is to score each of a set of potential graph structures  $G_1 \dots G_M$  for each training example  $D_1 \dots D_J$ , finding the most anomalous (highest scoring) connected subset and its score using our GraphScan algorithm for each combination of graph structure and dataset. (While this approach might require  $O(MJ)$  runs of the GraphScan algorithm, which would be computationally infeasible for large values of  $M$  and  $J$ , we demonstrate that a much smaller

---

number of runs is necessary in certain cases.) These scores are normalized by dividing by the score of the most anomalous unconstrained subset for that training example, 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.

In [64], we demonstrated that this learning approach can accurately learn a graph structure which can then be used by graph-based event detection methods such as GraphScan (discussed in Section 1.1.1.5 above), enabling more timely and more accurate detection of events (such as disease outbreaks) which spread based on that latent structure. Our results demonstrate that the learned graph structure is similar to the true underlying graph structure, capturing nearly all of the true edges but also adding some additional edges. Interestingly, the resulting graph often enables better detection power than the true underlying graph, enabling more timely detection of outbreaks, while achieving similar spatial accuracy to the true graph, as measured by the overlap coefficient between true and detected clusters. Finally, we demonstrated that the method is computationally tractable for learning the structure of graphs with approximately 100 nodes, typically requiring less than an hour of computation time despite searching over a large number of graph structures. We are in the process of extending this conference paper [64] into a submission to one of the top machine learning journals [104]. The key extension in the journal version (in progress) is the use of a more scalable (but approximate) event detection method as a component in the structure learning algorithm, thus enabling us to learn the structures of graphs with tens or hundreds of thousands of nodes.

#### **1.1.2.4. Combining Methods for Detection of Known and Unknown Patterns**

In past 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, described in Sections 1.1.3.1 and 1.1.5.4 below), 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, described in Section 1.1.1 above). 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 [112]. 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, I am working (with Heinz Ph.D. student Edward McFowland III) on extending both the fast generalized subset scan (FGSS) [71] and generalized fast subset sums (GFSS) [65, 105]

---

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 detect known and modeled patterns. To integrate the two methods, our first step (in progress) is to extend FGSS to **multiple known model types**, each modeled by a Bayesian network, thus identifying patterns that do not fit any of these models. Our current approach [111] 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. While this is a natural generalization of the original FGSS approach (which iterates between optimization over subsets of records and attributes), it is more complicated because it is not simply an iterative ascent algorithm and thus the convergence properties are more difficult to establish: while optimizing over records and attributes maximizes the score for the given mapping of records to models (finding the most anomalous subset of the data), re-fitting the models minimizes the score for the given records and attributes by making them less anomalous.

Our second step (planned future work) is to extend GFSS from spatial to general datasets. The hierarchical prior approach of identifying a “center” and “neighborhood size”, and then sampling probabilistically from the local neighborhood of the center, generalizes naturally to the non-spatial setting, but rather than using a Gamma-Poisson count model (as in the spatial setting) we must incorporate Bayesian network models like those used in FGSS. Finally, in addition to extending FGSS to multiple background models, it can also be extended to combine information from multiple local anomaly detectors, each of which gives a different view of the anomalousness of a given attribute value in the data. As for the “FGSS with multiple models” scenario discussed above, we now have multiple values to consider rather than a single value for each record and attribute, but this case requires a different iterative optimization procedure. The “FGSS with multiple local anomaly detectors” work will be the subject of Ed’s internship at AT&T Labs this summer, and has the potential to expand into a broader collaboration between our lab and the Machine Learning group at AT&T Labs.

In our future work [112], we propose to iteratively learn models and detect events through a “human-in-loop” process in which the user can provide four types of feedback on each reported pattern: labeling the pattern as an instance of one of the currently known event types  $E_k$ , labeling the pattern as an instance of a new event type  $E_{\text{new}}$ , rating the relevance of the discovered pattern, and correcting the region  $S$  by adding or removing records. Based on the user’s feedback, the system will update its models for each known event type  $E_k$ , including its spatial prior, effects on each data stream, and relevance to the user. At each stage we report patterns which are most likely to be relevant to the user. This will allow the system to continually expand its set of “known” pattern types and to incrementally improve the quality of its models. Two unique machine learning challenges in this setting, which our future work will address, are extending “active learning” approaches (methods for choosing the most informative examples for the user to label) to the case where the user labels a subset of the data as belonging to a particular pattern type, and learning the model for a newly identified pattern type starting from a single labeled example.

---

### 1.1.3. 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. My ongoing 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.

Our past work has addressed many aspects of this **end-to-end surveillance** problem. We have developed new multivariate methods which not only detect events but also characterize them by pinpointing the affected locations, temporal duration, and data streams. Our multivariate Bayesian scan statistic (discussed in Section 1.1.5.4 below) can further enhance situational awareness by modeling and distinguishing between multiple event types. We have explored methods for distinguishing between known and unknown events, explaining detected events (by comparison of the most likely and alternative models), and automating drill-down investigations (by identifying a small set of highly indicative records). As part of my work on the National Retail Data Monitor project, I have also developed and deployed GIS tools for outbreak visualization and drill-down investigation, and a Web-based interface which allows users to coordinate outbreak investigation and communicate their findings [27]. My current research thrusts for the end-to-end surveillance problem include the following:

#### 1.1.3.1. Scalable Event Detection and Visualization

The multivariate Bayesian scan statistic (MBSS) [8], described in Section 1.1.5.4 below, is a powerful detection method which can integrate information from multiple data streams and can model and distinguish between multiple event types. The output of MBSS 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 linear-time subset scanning methods cannot be used to efficiently generate this visualization, since we need to sum over probabilities rather than just finding the highest-scoring region.

More recently, we have 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 compare 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. We demonstrate substantial improvements in spatial accuracy and timeliness of detection, while maintaining the scalability and fast run time of the original MBSS method. This work has been presented at the International Society for Disease Surveillance Annual Conference [38] and the International Workshop on Applied Probability [79]. Additionally, the full paper was recently published in the journal *Statistics in Medicine* [5].

In collaboration with LTI doctoral student Yandong Liu and EPP doctoral student Kan Shao, I have recently developed a generalization of the fast subset sums method which allows the sparsity of the detected region to be controlled. 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, our work demonstrates 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. This work was presented at the ISDS Annual Conference (abstract published in the *Emerging Health Threats Journal* [35]), and as part of my invited talks at the Data Fusion Research Meeting [74] and the 2011 Joint Statistical Meetings [72]. Kan's work on learning the sparsity parameter for the Generalized Fast Subset Sums model was successfully presented for his Data Analysis Project for the M.S. in Machine Learning, and our conference paper submission is currently under review for the IEEE International Conference on Data Mining [65]. For the journal paper version of this work [105], we intend to jointly optimize the distributions for the sparsity parameter, neighborhood size, and center location using expectation-maximization (EM). We are also working on extensions to partially labeled data, where only a small subset of the affected locations is provided. We plan to submit this work this summer to one of the top machine learning journals (JMLR or MLJ).

### 1.1.3.2. Evaluation and Improvement of Event Surveillance Systems

This work focuses on the development of evaluation metrics that accurately reflect the performance of a system in real-world public health practice. In joint work with Dr. Xia Jiang and Prof. Greg Cooper, I have recently developed a generalization of the Activity Monitoring Operating Curve (AMOC) framework for evaluating the tradeoffs between false positive rate and timeliness of detection. Our approach explicitly accounts for the response protocol of the public health user: we demonstrate that the relative performance of systems may vary when a non-trivial response protocol is used, and also demonstrate that the framework can be used to select

---

appropriate response protocols for a given system. This work was published in the proceedings of the *American Medical Informatics Association Annual Symposium* [18].

### **1.1.3.3. Prioritization of Sources for Data Collection (planned future work)**

In the developing world, the amount of data available for electronic surveillance systems may be severely limited, and collection of new data sources may be difficult and expensive due to lack of existing infrastructure. Thus we propose novel methods to prioritize data sources for collection [119]. Our planned future work will focus on learning a predictive model which estimates the marginal improvement in timeliness of detection as a function of which sources are collected, the coverage and lag time of each source, and other area-specific covariates. We will explore a semi-synthetic testing approach, in which we inject a huge number of simulated outbreaks into real background health data from multiple areas and compute the tradeoff between detection time and false positives. We then “hold out” pieces of the data (e.g. removing a data source, or downsampling to simulate lower coverage) and compute the differences in detection time. Our predictive model will be trained by conducting this hold-out analysis across multiple areas with different existing data resources. Finally, the estimated benefits of additional data will be combined with cost estimates to prioritize collection efforts.

### **1.1.3.4. Automated Event Investigation and Tracking (planned future work)**

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 will use our new GraphScan method [39, 70] (discussed in Section 1.1.1.5 above) 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. We will extend fast graph scanning to the dynamic case, allowing the affected subset of data records to change over time. I believe that our recent advances in incorporating temporal consistency constraints into our fast subset scanning framework [108] (discussed in Section 1.1.1.6 above) will provide a methodological basis for accurate and computationally efficient detection of dynamic patterns. In future work, we propose to use LTSS to find the optimal subset subject to multiple constraints on consistency determined by the graph structure, the event dynamics, and the subset of records currently known to be affected. One major application of this work will be conducted through the recently funded Living Analytics Research Centre (a collaboration between CMU and the Singapore Management University), using cell phone location and proximity data to determine potentially infected individuals during an epidemic [117].

---

#### 1.1.4. Novel Data Sources for Event Surveillance

*This research thrust is the fourth of the four proposed research tasks for my NSF CAREER award [93], and will be an increasingly important focus of my work in the next 3-5 years.*

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.

##### 1.1.4.1. 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 are exploring a new **semantic scan statistic** approach [106], which uses rich text data to detect previously unknown event types, forming and searching a huge number of aggregated count datasets on the fly. 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 current 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 goal is to show 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.

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 initial 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.

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 current approach [106] focuses on learning “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. 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. We anticipate that a conference paper version will be ready within the next month for submission to the International Society for Disease Surveillance Annual Conference (ISDS 2011), with the full (journal paper) version to follow in the fall.

#### **1.1.4.2. Prediction using Leading Indicator Data**

We will improve the timeliness of event response by incorporating novel data sources which can predict that an event is likely to occur. For example, in the law enforcement domain, we have shown that detected clusters of certain minor crimes significantly increase the likelihood that a violent crime cluster will emerge in that area [43]. We will explore the utility of these and other potential predictors of crime, and will incorporate prediction into population health surveillance [127], for example, using veterinary health data to predict outbreaks of zoonotic disease, and environmental sensor data to predict airborne and waterborne outbreaks. Our approach reduces the prediction problem to a detection problem using the leading indicator data, but a fundamental challenge is to determine which combination of variables should be used as predictors when there are a huge number of variables to consider.

In the past year, we have made progress identifying the most useful leading indicators for the Chicago crime data discussed below, using a combination of cross-correlation analysis (for preliminary analysis, reducing the hundreds of different “calls for service” types down to a more manageable 40-50), univariate spatial and subset scanning (using our CrimeScan software described below), and multivariate spatial and subset scanning (evaluating which combinations of leading indicators are most useful to monitor). We have identified a working set of 22 leading indicator types which have promising detection results both on their own and in combination with other leading indicators, and are working to refine our analysis to identify the optimal

---

subset for inclusion into the deployed version of CrimeScan used by the Chicago Police Department.

#### **1.1.4.3. Integrating Sensor Placement and Sensor Fusion (planned future work)**

Sensor network data is an increasingly important component of disease surveillance: for example, CDC's BioWatch system has been deployed to detect an airborne pathogen release. We are also extremely interested in using location and communication data from mobile phones as a massive sensor network to detect anomalous behavior patterns caused by external events. In the developing world, disease surveillance could be enhanced by providing micro-payments to "sentinel" individuals in potentially affected areas, in exchange for health reporting via SMS text messages. Since in this case our "sensors" are both noisy and costly, we have a challenging active sensing problem. Our linear-time subset scanning approach can be thought of as a method for integrating data from multiple noisy sensors, while past work by Carlos Guestrin, Andreas Krause, Jure Leskovec, et al. has resulted in novel methods for efficient approximate sensor placement. Thus we propose to combine these two streams of work into an efficient iterative approach for noisy active sensing [120] which alternates between optimal sensor fusion using LTSS and near-optimal sensor placement using submodular function optimization. For each iteration of the algorithm, we use LTSS to efficiently detect potentially affected areas given the current sensors, and then obtain additional sensor readings which are likely to be useful given the current detection results.

#### **1.1.4.4. Incorporating Society-Scale Data**

Many online data sources, ranging from search queries to user-generated content such as blogs and status updates, can be incorporated into detection systems. However, the huge quantity of online data makes even linear-time search algorithms infeasible, requiring approximations based on sampling or on streaming data. In a recently initiated project with Heinz Ph.D. student Sriram Somanchi [118], we are working to combine LTSS with approaches from combinatorial property detection (CPD), which focus on determining (with high probability) whether a set  $S$  either has some property  $P$ , or is "far" from having property  $P$ , using a sample size that is sub-linear in  $S$ . In this case, property  $P$  corresponds to the "business as usual" case when no relevant events are occurring. We will extend CPD sampling methods and proof techniques to determine whether any subset of set  $S$  is far from  $P$ , and to approximate the affected subset, thus both detecting and localizing emerging events. LTSS may allow tighter bounds by efficiently considering all subsets of each sample; additional benefits may be gained from successive sampling techniques, allowing iterative refinement of our search. Our primary goal is to scale 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.

#### **1.1.4.5. Incorporating Crowdsourced Data**

Our ongoing work on incorporating learning into detection (discussed in Section 1.1.2 above) will transform the event detection task by rapidly honing in on patterns which are most relevant to the user. However, the need for an expert "in the loop" to provide feedback on detected patterns can be a serious bottleneck when investigation is time-consuming or the number of

---

patterns is large. This may be true both in disease surveillance and in health care, where time-constrained public health practitioners and physicians respectively may be unable to spend sufficient time on the identification and diagnosis of disease patterns. One possible solution is “crowdsourcing”: we farm out the detection task to hundreds of (non-expert) users and aggregate their feedback. Crowdsourced data creates numerous challenges: evaluating user skill levels and combining responses, training users to perform challenging tasks (as compared to tasks such as image labeling, which require minimal training), and providing monetary, altruistic, or entertainment-based incentives to attract and retain users. While many of these questions have been addressed in the prediction paradigm (where users provide a single label), they become much harder for detection (where users label the affected subset of the data). We propose to explore how detection problems can be effectively partitioned into smaller “chunks” so that a) each user can make a quick and accurate decision about their chunk, and b) if an event occurs in any subset of the data, sufficiently many users are able to detect it. This problem is similar to the distributed sensing case described above: each user can be viewed as a “sensor” that aggregates information about their chunk and outputs a noisy detection signal. We will combine LTSS, submodularity, and multi-resolution partitioning to simultaneously optimize “placement” of users (assignment of chunks) and “fusion” of their responses. Heterogeneity of user skill levels, and variations in skill depending on chunk characteristics, can be learned from data and incorporated into this framework. While I am interested in the theoretical basis of crowdsourcing more generally, my ongoing work will focus on the use of crowdsourced data as part of the HealthMap project for event-based disease surveillance [113] (described in Section 1.1.6) and for medical diagnosis from pathology slides [124] (described in Section 1.1.11 below).

## **1.1.5. New Statistical and Computational Methods for Event Detection**

### **1.1.5.1. Fast Spatial Scan Statistics**

My initial contribution to the event detection literature was the development of the “fast spatial scan” algorithm [29, 30, 31], which incorporates new multi-resolution search methods and a novel spatial data structure (the “overlap-kd tree”) to make cluster detection 100-1000x faster with no loss of accuracy. This enables us to perform the cluster detection task in under an hour for massive datasets which would otherwise require weeks of computation. While our first version of fast spatial scan [31, 53] performed efficient optimization over the set of  $O(N^3)$  square regions on an  $N \times N$  grid, our extensions of the algorithm to elongated clusters [3, 30, 52] and multi-dimensional datasets [29, 92] vastly increased the set of application domains to which our cluster detection methods could be applied, as well as allowing us to perform fast space-time cluster detection [29] and to use non-spatial attributes (such as patient age and gender) as additional search dimensions. However, the original implementation of fast spatial scan could only be used to search over rectangular-shaped clusters for data aggregated to a uniform grid, and could only be used to maximize the original (Kulldorff) spatial scan statistic. However, our new approaches based on linear-time subset scanning (LTSS) [42, 62] enable us to extend the fast spatial scan to non-gridded data and to any statistic satisfying the LTSS property, as well as producing even larger speedups than the original fast spatial scan method.

---

### 1.1.5.2. Expectation-Based Scan Statistics

The **expectation-based scan statistic** [2, 12, 13] is a new statistical method for spatio-temporal event detection. 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. Anomalous areas are detected by scanning over a huge set of potential regions, maximizing a likelihood ratio statistic, and computing statistical significance by randomization testing. In a recent paper published in the *International Journal of Forecasting* [12], I demonstrated 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. In this work, I also compared multiple variants of the expectation-based method on the disease surveillance task (detecting simulated disease outbreaks injected into real hospital Emergency Department data from Allegheny County), answering the questions of which set of space-time regions to search, which time series analysis method to use for computing expectations, and which statistical model to use for detection.

Within the expectation-based scan statistic framework, I have developed a number of new statistics that can be used for event detection. Here I focus on the parametric scan statistics, in which the score function is the log-likelihood ratio  $\Pr(\text{Data} \mid H_1(S)) / \Pr(\text{Data} \mid H_0)$ , where  $H_0$  is the null hypothesis that no events are occurring, and  $H_1(S)$  is the alternative hypothesis that an event is occurring in region  $S$ . Nonparametric and Bayesian methods will be discussed below. My first contribution to this area [2, 26] was the development of an “expectation-based” variant of the original “population-based” Poisson spatial scan statistic proposed by Kulldorff (1997). The expectation-based statistic searches for spatial areas where the count inside the area is higher than expected, as opposed to Kulldorff’s statistic, which searches for areas where the ratio of count to population is higher inside than outside. The expectation-based statistic has higher detection power than Kulldorff’s statistic when the event affects a large region (relative to the total area being monitored), and similar detection power otherwise [13]. I have also extended this framework to a Gaussian scan statistic model [13, 56], learning both the expectation and variance from historical data, and to a novel space-time scan statistic [25, 50, 57] for detection of emerging clusters, where the relative risk in the affected region increases over time.

A related stream of my research involves making traditional approaches to event detection more robust by extending the underlying statistical models. Traditional detection approaches typically suffer from high false positive rates: they signal alarms in response to anomalies in the data that do not correspond to the true patterns of interest. For example, when monitoring over-the-counter medication sales for disease outbreaks, many false positives are caused by single “outlier” stores with anomalous sales counts (due to data irregularities, inventory movements, bulk purchases, promotions, or other unmodeled events). In work presented at the *International Society for Disease Surveillance Annual Conference* [48], I developed a robust expectation-based scan statistic approach which substantially reduces the number of false positive alerts due to outliers, enabling faster and more accurate detection of outbreaks and other events.

In work presented at the ISDS Annual Conference [46] and a paper published in the *International Journal of Health Geographics* [13], I compared twelve variants of the spatial scan

---

statistic on the outbreak detection task over a wide range of datasets and outbreak types. One interesting conclusion of this work was that the expectation-based Poisson statistic can achieve significantly better detection performance than Kulldorff's widely used spatial scan statistic. Finally, in my recent work on linear-time subset scanning [62], I show that not only the Poisson and Gaussian expectation-based scan statistics, but also the expectation-based scan statistic corresponding to any distribution in the separable exponential family, can be efficiently maximized over subsets of the data. My current work on expectation-based scan statistics involves improving the prediction of expected counts for multivariate data, both by incorporating recent work in linear dynamical systems and (in the crime hot-spot detection domain) incorporating state-of-the-art crime forecasting methods developed by Prof. Wil Gorr.

### 1.1.5.3. Nonparametric Scan Statistics

I am currently working on a general, non-parametric methodology for multivariate event detection, which (unlike standard model-based event detection approaches) makes no assumptions about the underlying parametric distribution of the data, thus providing a principled way of combining information from multiple disparate data streams. My recent work on this **nonparametric scan statistic** (NPSS) methodology, presented at the International Society for Disease Surveillance Annual Conference [44] and the Joint Statistical Meetings [84], has demonstrated that the nonparametric method a) increases detection power for multivariate detection tasks, particularly when typical parametric models are incorrect, b) allows accurate identification of the affected spatial region, and c) enables us to characterize events by accurately identifying the affected subset of data streams. I have recently shown that our linear-time subset scanning approaches [42, 62] can be used to make the run time of the NPSS method scale linearly rather than exponentially with the number of data streams, without any loss of accuracy. I have also performed detailed follow-up experiments to determine which variants of the nonparametric scan achieve highest detection power across a range of outbreak detection scenarios, and I am currently working on a journal paper on this topic [103]. The nonparametric scan statistic also forms the basis for our Fast Generalized Subset Scan method [71] discussed in Section 1.1.1.4 above. Finally, I am currently working on a different nonparametric method which I call the "empirical scan statistic", which uses the entire distribution of residuals from a large amount of historical data to more accurately detect patterns. Our preliminary results suggest that this method can achieve faster detection using multivariate data as compared to typical parametric approaches and our previous nonparametric method.

### 1.1.5.4. Bayesian Scan Statistics

One major accomplishment of my past research was the development of a new multivariate Bayesian framework for pattern detection, the **Multivariate Bayesian Scan Statistic** (MBSS) [8]. This method 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). In a paper published in the journal *Machine Learning* [8], my co-author (Prof. Greg Cooper) and I describe the MBSS framework in detail and demonstrate that it has many advantages over traditional event detection

---

approaches, including higher detection power through incorporation of prior information, accurate characterization and differentiation of multiple event types, and the ability to accurately learn event models from labeled training data, expert knowledge, or a combination of the two.

Our work on MBSS has been widely disseminated through my invited talks at the *Donald A.B. Lindberg Lecture and Symposium* [87], *Washington Statistical Society Seminar* [86], *Joint Statistical Meetings* [88], *International Symposium on Forecasting* [89], and *Twelfth Biennial CDC/ATSDR Symposium on Statistical Methods* [83], and was also presented at the *International Society for Disease Surveillance Annual Conference* [45, 47]. Several of these talks also presented my work on incorporating learning into the MBSS framework, as discussed below. MBSS has also been a major component of four funded grant proposals [93, 97, 99, 100] which have helped to fund my work and support my students. The company Health Monitoring Systems has expressed interest in licensing my MBSS software. Additionally, MBSS builds on our previous work on the Bayesian Scan Statistic, extending this work to monitor data from multiple streams and to model and differentiate between multiple event types. The original, univariate BSS work was published in the top tier computer-science conference *Neural Information Processing Systems* (NIPS 2006) [23] and presented at the *ECADS Syndromic Surveillance Conference* [90], and the *International Workshop on Applied Probability* [91]. Additionally, my presentation on the Bayesian Scan Statistic at the *International Society for Disease Surveillance Annual Conference* [49] received the conference's Best Research Presentation award.

In collaboration with Prof. Greg Cooper and Dr. Xia Jiang, I have developed a new Bayesian scan statistic approach which combines spatial and population-based approaches to detection. This **entity-based** event detection approach is similar to the MBSS method in that it uses a Bayesian model to differentiate between multiple event types, but here we model the effects of the event on each individual in a population rather than on a set of monitored data streams. This approach is preferable to MBSS when we have detailed individual-level data but may be less useful when we have only aggregate count data. We have successfully applied this method to disease surveillance using Emergency Department data, resulting in a journal paper published in the *International Journal of Approximate Reasoning* [9].

My current 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. In [8, 45], we demonstrated how to learn the relative effects of an event on the monitored data streams from labeled training data, and in [22, 41], 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. This enables learning of priors for region size, shape, and also which locations are more or less likely to be affected. In future work, as part of the NSF-funded "Discovering Complex Anomalous Patterns" project [97], we will also investigate broader questions of optimal query selection for learning complex and dynamic event models from user feedback. This work is described in more detail in Section 1.1.2 above. Additionally, I have recently developed an extension of MBSS which enables scalable event detection and visualization, and incorporated learning into this framework as well. These "Fast Subset Sums" [5] and "Generalized Fast Subset Sums" [65] methods are described in Section 1.1.3.1 above.

---

### 1.1.5.5. Bayesian Network Scan Statistics

One interesting feature of both the MBSS and entity-based scan statistic approaches is that they rely on an underlying Bayesian network representation to model and detect events. In collaboration with Prof. Jeff Schneider and former MLD graduate student Kaustav Das, I extended these event detection methods to pattern detection in more general categorical datasets using efficient methods for Bayesian network learning. Our Anomalous Group Detection (AGD) method scans over related subsets of the data, computes a likelihood ratio statistic for each subset (comparing the “local Bayesian network” learned from a given subset of the data to the “global Bayesian network” learned from the entire dataset), and reports the highest scoring subsets. In a recent paper [68], currently under revision, we demonstrated that this method can accurately detect anomalous groups in disease surveillance and container shipping datasets. This work was also a major part of Kaustav’s doctoral thesis, as well as a component of our book chapter [1] on Bayesian network-based scan statistic approaches (including AGD, MBSS, and the entity-based scan statistic) published as part of the book *Scan Statistics – Methods and Applications*.

### 1.1.5.6. Rule-Based Detection of Patterns of Anomalies

In collaboration with Prof. Jeff Schneider and former MLD graduate student Kaustav Das, I developed a rule-based method of anomalous pattern detection which searches for related groups of individually anomalous records. This method is distinct from approaches such as our Bayesian network scan statistics, in which a group of records could be anomalous even if none of the individual records are anomalous, and thus is useful for a different class of applications (e.g. finding patterns of suspicious behaviors for fraud detection applications). Our “Anomaly Pattern Detection” (APD) method works by first applying any “local anomaly detector” to determine the anomalousness of each individual record, then detecting subsets of the data (defined by one- or two-component conjunctive rules such as SYNDROME = RESPIRATORY AND AGE < 10) with a higher than expected number of anomalous records. This work was published in the proceedings of the top-tier computer science conference Knowledge Discovery and Data Mining (KDD 2008) [20], and was also presented at the International Society for Disease Surveillance Annual Conference, with an abstract published in the journal *Advances in Disease Surveillance* [40]. We are currently investigating several other promising methods for detection of anomalous patterns and groups in massive datasets, including the Fast Generalized Subset Scan [71] and tensor scan [107] methods described in Section 1.1.1 above.

### 1.1.6. Applications to Disease Surveillance

As can be seen from Sections 1.1.1-1.1.5, the **disease surveillance** domain has served as my primary testbed for the development of new event and pattern detection methods. My first experience developing a real-world disease surveillance system was on the National Retail Data Monitor project [16], a research collaboration between the University of Pittsburgh and CMU for disease surveillance using over-the-counter medication sales. In joint work with Robin Sabhnani and Andrew Moore, I published several papers [16, 27, 51] describing the implementation of spatial cluster detection methods to monitor daily, nationwide data feeds from the NRDM. Additionally, some of my early work was developed and funded through CDC’s BioSense project for nationwide disease surveillance [99], and through the NSF-funded Bayesian Biosurveillance Project [100].

---

More recently, I have obtained real public health data from a number of sources, including colleagues at the University of Pittsburgh, the Ottawa Heart Institute, and the company Health Monitoring Systems. I am currently collaborating with a team led by Dr. Rick Davies of the Ottawa Heart Institute, developing and deploying systems for real-time disease surveillance in the Grey Bruce region of Ontario (ECADS project), Ottawa (ASSET project [101]), and several other Canadian cities (Data Fusion project [98]). The ECADS and ASSET systems are complete, and the deployed systems are in regular use by Grey Bruce Public Health and Ottawa Public Health respectively. The ECADS system has enabled Grey Bruce to rapidly detect several emerging outbreaks (*E. coli*, scarlet fever, cryptosporidiosis, etc.) and our work has given both health departments powerful tools for outbreak investigation and response. As part of the ECADS project, we performed a retrospective analysis of the severe outbreak of gastroenteritis in Walkerton, Ontario, in 2000, and demonstrated that our novel disease surveillance methods could have detected the outbreak two full days before the initial public health response [69]. We are currently working to obtain the retrospective data from both ECADS and ASSET, and will use the labeled disease outbreak data to analyze the performance of current and future detection methods. The Data Fusion project [98] will focus on development and deployment of methods for fusion of multiple health data sources, with applications to the monitoring of **hospital-acquired infections** and detection of patterns of events related to **drug abuse**. Data for hospital-acquired infections has been collected from two major Canadian hospitals by our Data Fusion team [126]. I am very excited about both of these new application domains, as I am working to extend the scope of my work from disease surveillance to the more general goal of **population health surveillance** [127], as discussed in Section 1.2 below. Additional biosurveillance system deployments (in collaboration with CMU's Auton Lab) are in progress in Sri Lanka and in Tamil Nadu, India.

My student Skyler Speakman and I have recently begun to collaborate with Prof. John Brownstein and the HealthMap project team, focusing on the use of **event-based disease surveillance** from news reports, eyewitness accounts, and individual self-reports [113]. These non-traditional data sources have great potential for early outbreak detection, but novel analytical methods are needed to detect emerging patterns using this data. We are also considering the role of **crowdsourcing** (see Section 1.1.4.5 above) both in data collection and in validation of the submitted reports, which presents a number of challenges including collection of data (active-survey-based vs. passive-clickstream-based), active learning (to present the best queries to the appropriate members of the "crowd"), and combining noisy data from many expert and non-expert users. Skyler will be interning with HealthMap this summer, and we expect this collaboration to result in a major ongoing project which will form a large part of Skyler's doctoral thesis. We have also started discussions with Prof. Dennis Israelski, President and CEO of InSTEDD, about applying our detection work to their event-based disease surveillance tools.

### 1.1.7. Applications to Crime Hot-Spot Detection and Prediction

A second major application domain for my event detection research is the **detection and prediction of patterns of crime**. In recent work with Prof. Wil Gorr, I applied the expectation-based scan statistic approach to crime hot-spot detection using offense report data. In addition to successfully detecting spatial clusters of violent crime, we demonstrated that the scan statistic can also 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. Our preliminary results have been published in *Advances in Disease Surveillance* [43] and presented at the 2009 *Crime Mapping Research Conference* [81], and we are currently working on a journal paper on this topic [102]. We have also been asked to develop an event detection and forecasting module for the widely used CrimeStat software, and (with the Washington, D.C. police department) have proposed a novel, controlled experiment to deploy early warning systems for crime and evaluate the effects of the resulting targeted interventions on crime rates.

In 2009 we began collaborating with the Chicago Police Department, and they have recently provided us with a huge amount of data including crime offense reports, calls for service, web-based citizen complaints, curfew violations, contact cards, gang information, and deployment operations areas (representing targeted interventions by the CPD to prevent crime). We are in the process of analyzing this data to determine which crime types and other indicators are the best predictors of emerging hot-spots of violent crime, both individually and when integrated with other data sources, to evaluate and extend our methods for prediction of violent crime patterns, and to develop and compare other methods for crime prediction [102]. Moreover, we have developed and provided a software package (which we call “CrimeScan”) to the CPD, which uses our multivariate spatial and subset scan methods to monitor multiple leading indicators for crime prediction, and their use of this tool has grown substantially over the past year. Our CrimeScan software is now in day-to-day use (actually being run twice a day) by the CPD for crime prediction and tactical deployments, and we now receive continual feedback on their use of this tool, allowing us to continually improve both the interface and the underlying detection methods. CPD has been an enthusiastic partner for this work: they recently established a new Predictive Analytics Group, and firmly believe that predictive policing will be an essential component of their current and future law enforcement strategies. In 2010, Chicago had their lowest murder rate since 1965, and credited smarter policing, including the use of predictive analytics for real-time deployment of patrols, with the overall decline. Our CrimeScan software was featured in a recent Chicago Sun-Times article about the CPD (1/22/2011):

*“It was a bit like a scene from “Minority Report,” the 2002 Tom Cruise movie that featured genetically altered humans with special powers to predict crime. In October, the Chicago Police Department’s new crime-forecasting unit was analyzing 911 calls for service and produced an intelligence report predicting a shooting would happen soon on a particular block on the South Side. Three minutes later, it did, police officials say...”*

I am currently working with MSIT-VLIS student Amrut Nagasunder, MISM student YongJei Lee, and Prof. Wil Gorr to continue developing advanced methods for crime hot-spot detection and prediction, hopefully leading to continued improvement in the Chicago Police Department’s predictive policing strategies and corresponding reductions in the level of violent crime. Amrut’s current work focuses on evaluating and extending our CrimeScan software and the underlying spatial scan methods, as well as developing and comparing new crime prediction methods based on kernel density estimation and prediction. YongJei’s current work focuses on crime mapping, enhancing our understanding of the different leading indicator types and their ability to predict violent crime, and eliciting knowledge from the CPD to create a useful, historical “gold standard” for violent crime prediction. We believe that this collaboration with

---

CPD will continue to expand over the next 3-5 years (for example, colleagues from IBM T.J. Watson Laboratories have expressed interest in becoming involved), and the Chicago data will be very valuable in our development of new detection and prediction methods as well as our understanding of the spatial dynamics of crime.

### **1.1.8. Applications to Detecting Anomalous Patterns of Patient Care**

As part of our DCAP project [97], we intend to apply pattern detection methods to analyze data from intensive care units in Boston (collected by the MIT MIMIC II project), and UPMC hospitals (collected by our University of Pittsburgh collaborators). One aspect of this project includes predicting hospital re-admissions, and data recently made available through the Heritage Health Prize competition might be a useful testbed for our detection and prediction methods.

Additionally, I was recently awarded a Healthcare Technology Innovation grant from the University of Pittsburgh Medical Center (UPMC) Technology Development Center [94], with the goal of “Anomalous Pattern Detection in Healthcare Data Streams.” As part of this project, with my co-PIs Artur Dubrawski, Jeff Schneider, and Rema Padman, I plan to apply novel pattern detection methods to healthcare data flowing through UPMC’s Message Routing System, thus enabling the ongoing extraction of diagnostic and business intelligence from real-time healthcare data streams. We propose to develop a system which observes the massive quantity of data passing through the Message Router, detects and characterizes anomalous patterns in the data, and reports the most relevant patterns to human expert users.

While the proposed system is general enough to detect many different types of patterns, our initial focus will be **detecting anomalous patterns of care with significant impacts on patient outcomes**. Consider the natural variation in care practices between different groups and different clinicians within the UPMC systems. For example, when presented with a patient with severe breathing difficulties, different clinicians may choose to administer different types and dosages of medications, use different criteria to decide whether or not to place the patient on a ventilator, etc. Similarly, different hospital staff may exhibit variability in their care practices (such as hand-washing and isolation precautions) and adherence to physician orders. This variation in type and quality of care can have huge impacts on patient outcomes, such as mortality and morbidity rates, hospital re-admissions, and nosocomial infections. 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. Other potential uses of the system include early detection of emerging outbreaks of disease in the patient population, monitoring and prevention of nosocomial illness, and identification and correction of discrepancies between the standard of care for a patient, the actual care that the patient receives, and the billing records for that patient. An MSIT-VLIS student, Tarun Kumar, will be working with me, my faculty colleagues, and the UPMC Technology Development Center to implement, evaluate, and operationalize this system. Our first stage is to analyze existing data from UPMC consisting of thousands of patients with a discharge diagnosis of pneumonia (either community-

---

acquired or hospital-acquired), discover anomalous patterns of care for these patients, and then extend our detection methods from static to streaming data. Data acquisition is in progress.

### 1.1.9. Applications of Event Detection to Other Domains

As discussed above, two other highly relevant application domains for my event and pattern detection work are **customs monitoring** (discovery of patterns of illicit container shipments) and **network intrusion detection**. These security domains have been used as testbeds for our recently proposed Anomalous Group Detection [68], Anomaly Pattern Detection [20], and Fast Generalized Subset Scan [71] methods, using a novel container shipment dataset (“PIERS”) and the publicly available “KDD Cup 1999” network intrusion dataset. All three methods have been demonstrated to achieve high detection performance for identifying suspicious patterns of shipments and for most types of network intrusions. Additionally, in collaboration with colleagues from Intel Research Pittsburgh, my former student Maxim Makatchev and I have developed and implemented algorithms to detect “worms” propagating through a computer network. Our preliminary results, obtained during Maxim’s summer internship at Intel, indicate potential improvements over Intel’s current worm detection approaches. Finally, I am planning to work with Prof. Nicholas Christin on a project involving the detection of cybercrime networks using data collected from “honeypots” (systems used by information security researchers to collect and identify hacking attempts) [125].

With former CMU Civil and Environmental Engineering graduate student Daniel Oliveira, and CEE faculty members Lucio Soibelman and Jim Garrett, we applied the novel spatial cluster detection methods described above to **detect and explain clusters of pipe breaks** in a water distribution system. This work was part of Daniel’s doctoral thesis, and our paper describing this work was recently published in the *Journal of Computing in Civil Engineering* [7]. Similarly, my work with Heinz Ph.D. student Skyler Speakman on detecting dynamic patterns with temporal consistency constraints [108], described in Section 1.1.1.6 above, was applied to the detection of spreading patterns of environmental contaminants in a water distribution network.

Finally, with University of the Andes economics faculty member Prof. Samuel Malone, several of my graduate students and I have recently started investigating what we call “**economic growth outbreaks**” [116], using our subset scan methods to detect clusters of countries which exhibit sharp, sustained economic growth, and attempting to model how these “outbreaks” spread via bilateral trade (and other links) and common structural factors.

### 1.1.10. Health Care Information Systems

In addition to the disease surveillance work discussed in detail above, I am currently working with Prof. Rema Padman and other colleagues on three additional projects related to health care information systems. First, with University of Florida faculty member (and former Heinz Ph.D. student) Chris Harle, we have developed novel methods for **classification and visualization of high-dimensional health data**, applied to assessment of the risk of diabetes and related complications. This work was presented at the *2008 INFORMS Workshop on Data Mining and Health Informatics* [21], and became Chris’s First Heinz Research Paper; we have recently submitted this work to the *Journal of Biomedical Informatics* [66]. Additionally, Prof. Padman and I were recently awarded a Healthcare Technology Innovation grant from the University of

---

Pittsburgh Medical Center (UPMC) Technology Development Center [96]. As part of this grant, and in collaboration with Prof. Harle and UPMC physicians Dr. Fran Solano and Dr. Janice Zgibor, we plan to develop and evaluate novel information visualization tools and methods for improving diabetes care using UPMC ambulatory care settings and data. The primary goal is to develop intelligent visual data analysis tools that can be integrated with existing approaches to clinical data management and evaluation, in order to provide practitioners with usable systems that deliver critical information and new insights for responding to chronic disease risk among their patients. The proposed tools will offer an interactive interface through which clinicians can visually access, explore and compare risk predictions for a large cohort of patients in the context of many risk factors. Our initial results on evaluating clinician information needs for this data visualization-based diabetes risk assessment task have been accepted for poster presentation at the *American Medical Informatics Association Annual Symposium* [33].

Second, with Stanford faculty member (and former Heinz Ph.D. student) Sharique Hasan, master's student Huanian Zheng, and Prof. George Duncan, we have developed a collaborative filtering approach to **medication reconciliation**, predicting and correcting omissions from a patient's medication list in order to reduce the risk of adverse drug events. Our results demonstrate that collaborative filtering identifies the missing drug in the top-10 list about 40-50% of the time and the therapeutic class of the missing drug 50-65% of the time at the three clinics in this study. This work was presented at the 2008 *AMIA Annual Symposium* [19] and the *13<sup>th</sup> International Conference on Medical Informatics* [36], and became Sharique's Second Heinz Research Paper. Most recently, our full paper describing this work was published in the *Journal of the American Medical Informatics Association* [4].

Finally, with colleagues at the Technical University of Munich, we are working on **hospital length-of-stay management**, and have developed a method to predict a hospital inpatient's Diagnosis Related Group (DRG) at the beginning of their stay. This technology will enable hospitals to better predict costs, reimbursement, and patient flow, enabling better scheduling of patients to increase hospital profit and improve quality of care. This work was presented at the *35<sup>th</sup> Conference on Operational Research Applied to Health Services* [37]; our full paper [67] is under revision and will be re-submitted shortly to the journal *Artificial Intelligence in Medicine*.

#### 1.1.11. Miscellaneous Projects (ongoing and planned future work)

I have assisted the Grandis Lab at the University of Pittsburgh Medical Center with their work on **cancer chemoprevention**, performing the statistical analysis for several of their experiments. This collaboration resulted in three co-authored journal papers in the journals *Carcinogenesis* [11], *Clinical Cancer Research* [10], and *Cancer Prevention Research* [6]. In our most recent paper [6], we demonstrated (using a novel carcinogen-induced mouse model of oral cancer) that mice placed on a diet supplemented with the epidermal growth factor receptor erlotinib exhibited a 69% decrease in incidence of pre-cancerous and cancerous lesions compared with mice on the control diet, thus presenting evidence for testing the efficacy of erlotinib in a clinical trial setting.

In past work, I have advised former EPP graduate student Sean Green on his research on **predicting the incidence of diarrheal illness** using decision trees, and we have developed and applied new machine learning methods for computing **variable importance** (e.g. for setting

---

health spending priorities). This work was part of Sean’s doctoral thesis, and a journal paper describing this work is in progress [114]. I have also worked with Prof. Karen Clay, Prof. Joel Tarr, and Heinz graduate student Jeff Lingwall in an attempt to draw connections between heavy rainfall, sewage outflows, and rates of gastrointestinal illness. While no significant correlations were found, we believe that follow-up work using higher quality data might lead to interesting new understandings of the effects of sewage outflows.

I am working with Heinz Ph.D. student Skyler Speakman and Prof. Laura Dabbish on methods to harness the **wisdom of crowds** for medical diagnosis, including crowdsourcing the discovery of suspicious patterns (such as Reed-Sternberg cells, typically indicative of Hodgkin’s lymphoma) in pathology slides [124]. While this work is in the very early stages, it has potential to reduce the frequency of medical errors as well as contributing to the theory of human computation.

In collaboration with Prof. Alessandro Acquisti, I plan to investigate the use of Internet search patterns (e.g. Google queries) and user-generated Web content (Twitter feeds, Facebook status updates, etc.) for event detection, and believe that these informal sources will be important early indicators of disease outbreaks and other health events (e.g. Google Flu Trends). We are also considering a number of other ways in which patterns of online activity can be used for the public good [123], including predicting (and reducing) traffic congestion for special events, identifying areas with inadequate access to health clinics and other resources, and serving as early indicators of a variety of potential hazards. One challenge in this domain is preserving individual privacy while obtaining useful information in the aggregate, and I believe that this project will benefit greatly from the expertise of Alessandro and other Heinz faculty on privacy issues. Additionally, I am currently working with Alessandro and postdoctoral fellow Fred Stutzmann on a project which focuses on **predicting individuals’ credit scores** from their Facebook profiles [115], demonstrating the risks to personal privacy from information disclosures on online social networks, and plan to work with Prof. Robert Hampshire and Dr. Lavanya Marla to use Twitter feeds for traffic modeling and optimization [122].

In recent work, I have begun to investigate other core machine learning techniques with a wide variety of potential policy applications, including clustering (useful for customer database segmentation and for modeling heterogeneous datasets) and active learning (useful for scientific discovery, product development, and medical diagnosis). In particular, I am currently working to develop the theory and methodology of “competitive active learning” [121], a paradigm in which agents must choose queries that are likely to maximize their information gain while minimizing the information provided to an opponent, with applications to information exchange and collaborative exploration between competing firms.

#### **1.1.12. Game Theory and the Evolution of Behavior (past work)**

In the past, a major focus of my research has been the study of dynamic, multi-agent processes such as evolution, learning, and imitation, in order to understand and control the behavioral patterns that emerge in multi-agent systems. Within this research domain, one of my main goals is to understand how agents can develop and maintain cooperative behavior, both in situations requiring simple choices between cooperation and selfish behaviors (modeled by the well-known “Prisoner’s Dilemma”) and in more complex situations (such as turn-taking [32]) where players must learn to coordinate their cooperative behaviors in a noisy environment. A main

---

contribution of my work, published in the *Journal of Theoretical Biology*, is the identification of cooperative strategies that are robust to “mistakes” resulting from imperfect communication, as well as demonstration of how these behaviors can emerge by learning or evolution in dynamic systems [17, 32, 61]. A second contribution is the development of a novel model of evolutionary and learning dynamics with “large aggregate shocks” [58, 59]: this is the first dynamical model in which cooperative behaviors can remain dominant in the long run, even in certain games where the Nash equilibrium is non-cooperative. Another goal of my work on dynamic processes has been drawing connections between agents’ behavior on the “micro-scale” and “macro-scale”: that is, examining how forces influencing the choices of individual agents can lead to large-scale behavioral trends in a population. My work has developed various methods for the large-scale statistical analysis of populations, and one of my long-term research goals is to integrate these tools with core machine learning techniques, developing methods for optimal control of very large numbers of interacting learning agents.

Two other contributions of my research in dynamic systems are a better theoretical understanding of mathematical population dynamics, and a better understanding of observed trends in human behavior. In another paper published in the *Journal of Theoretical Biology* [15], I proposed a revision to a fundamental theoretical concept of population dynamics: John Maynard Smith’s conception of “evolutionary stability,” which explains when one type of behavior can supplant another in a dynamically changing (learning or evolving) population. Finally, in a paper published in the journal *Rationality and Society* [14], I presented a model of how humans combine personal preferences with “norm-following” imitative behavior; this model can be used to explain trends in product choice and voting, as well as sociological phenomena such as fashions, fads, and the well-known “bandwagon effect.”

### **1.1.13. Natural Language Processing and Other Topics (past work)**

My past work also includes research in natural language processing, link and group detection, and mathematical modeling. As a Churchill Scholar at Cambridge University, I designed a system for automatic word sense disambiguation [60], using machine learning techniques to automatically identify the different senses of a word (for example, the word “plant” could correspond to “industrial production facility” or “natural fauna”) and discriminate between these senses in context. At Carnegie Mellon, I used similar techniques to investigate the problems of alias detection in link data sets [28]: deciding whether the same name corresponds to multiple entities (e.g. “George Bush”) and deciding whether different names correspond to the same entity (e.g. “George W. Bush” and “the President”). Finally, as an undergraduate at Duke University, I co-authored two papers investigating mathematical models of air traffic control and disaster evacuation respectively [54-55]; each of these was selected as a winning paper in the nationwide MCM (Mathematical Contest in Modeling) competition.

## **1.2. Research Trajectory**

In the future, I will continue to pursue a research agenda at the intersection of machine learning, computer science, statistics, and public policy. I will further develop methods for pattern detection and characterization (including applications to health surveillance, detection and prediction of crime patterns, and many other application domains) and apply a variety of new machine learning methods to health care policy and management.

---

Having successfully tackled the three main event detection questions of integrating spatial and temporal information from multiple data streams, incorporating prior information, and modeling and differentiating between multiple event types, I intend to focus on four additional questions: developing efficient and scalable pattern detection methods (Section 1.1.1), incorporating learning into the multivariate event detection framework (Section 1.1.2), developing new “end-to-end” methods for investigation of discovered patterns (Section 1.1.3), and incorporating new data sources (Section 1.1.4). Additional objectives of this ongoing work include: developing computationally efficient detection algorithms for massive, high-dimensional datasets; learning complex models and structures; integrating detection of known and previously unknown pattern types; incorporating incremental model learning and user feedback to enable continual improvement in the quality of detected patterns; generalizing event detection methods to non-geographic pattern discovery; and applying these techniques to a wide variety of real-world datasets. One goal of this work is to create a general and flexible framework for pattern detection which can be applied to any type of data, enabling users to define and discover new pattern types “on the fly”, and quickly focusing their attention on the most relevant aspects of a complex and massive dataset.

I also intend to continue advancing the core statistical and computational methodologies for event detection (Section 1.1.5) and to focus on applications to disease surveillance, crime monitoring, and patient care (Sections 1.1.6-1.1.8), particularly in the contexts of my ongoing work with public health surveillance in Canada (ECADS, ASSET, and Data Fusion projects), the Chicago Police Department, and the UPMC hospital system, as well as generalizing to many other applications of pattern detection. I am particularly interested in using LTSS to efficiently identify patterns in dynamic graphs, and am currently working to acquire relevant datasets (e.g. social networks, cell phone calls and location data, and blog data) on which these methods will be useful. I believe that one of the “killer apps” for the next decade of disease surveillance will be the use of location and proximity data from cellular telephones, in combination with various health data sources, to perform automatic outbreak detection and epidemiological contact tracing [117], thus identifying outbreaks and providing accurate information as to not only where, but **who**, is likely to be affected.

More generally, I hope to broaden my disease surveillance work into **integrated population health surveillance** [127], focusing not only on infectious disease outbreaks but also the many other factors which influence population health, including prevention and monitoring of chronic illness, hospital-acquired infections, drug abuse, injury, crime and violence, poverty, and patient care. Many of these individual factors will be examined through my collaborations with the Chicago Police Department (crime and violence), UPMC (patient care, chronic illness), and the Data Fusion project (hospital-acquired infections and drug abuse). In the longer term, I would like to develop surveillance systems which incorporate and integrate many of these factors to provide a more complete view of population health, emerging trends, interrelationships between variables, and the necessary interventions that should be made in order to maintain and improve the health of populations. My primary contact at the Chicago Police Department has recently been named Chief Data Officer of the City of Chicago, reporting directly to the Mayor, and we intend to use Chicago as a testbed for integrated population health surveillance, hopefully resulting in significant benefits to the city’s administration and to its citizens’ quality of life.

---

## 2. List of Research Papers and Projects

### **Book Chapters (3)**

- 1) D. B. Neill, G. F. Cooper, K. Das, X. Jiang, and J. Schneider. Bayesian network scan statistics for multivariate pattern detection. In J. Glaz, V. Pozdnyakov, and S. Wallenstein, eds., *Scan Statistics: Methods and Applications*, 221-250, 2009.
- 2) D. B. Neill and A. W. Moore. Methods for detecting spatial and spatio-temporal clusters. In M. Wagner, A. Moore, and R. Aryel, eds., *Handbook of Biosurveillance*, 243-254, 2006.
- 3) D. B. Neill and A. W. Moore. Efficient scan statistic computations. In A. Lawson and K. Kleinman, eds., *Spatial and Syndromic Surveillance for Public Health*. Chichester, UK: Wiley, 189-202, 2005.

### **Refereed Journal Papers (15)**

- 62) D. B. Neill. "Fast subset scan for spatial pattern detection," *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, accepted for publication, 2011.
- 4) S. Hasan, G. T. Duncan, D. B. Neill, and R. Padman. Automatic detection of omissions in medication lists. *Journal of the American Medical Informatics Association*, 18(4): 449-458, 2011.
- 5) D. B. Neill. Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine*, 30: 455-469, 2011.
- 6) R. J. Leeman-Neill, R. R. Seethala, S. V. Singh, M. L. Freilino, J. S. Bednash, S. M. Thomas, M. C. Panahandeh, W. E. Gooding, S. C. Joyce, M. W. Lingen, D. B. Neill, and J. R. Grandis. Inhibition of EGFR-STAT3 signaling with erlotinib prevents carcinogenesis in a chemically induced mouse model of oral squamous cell carcinoma. *Cancer Prevention Research*, 4(2): 230-237, 2011.
- 7) D. Oliveira, D. B. Neill, J. H. Garrett Jr., and L. Soibelman. Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network. *Journal of Computing in Civil Engineering*, 25(1): 21-30, 2011.
- 8) D. B. Neill and G. F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 79: 261-282, 2010.
- 9) X. Jiang, D. B. Neill, and G. F. Cooper. A Bayesian network model for spatial event surveillance. *International Journal of Approximate Reasoning* 51: 224-239, 2010.
- 10) R. J. Leeman-Neill, Q. Cai, S. C. Joyce, S. M. Thomas, N. E. Bhola, D. B. Neill, J. L. Arbiser, and J. R. Grandis. Honokiol inhibits epidermal growth factor receptor signaling and

---

enhances the antitumor effects of epidermal growth factor receptor inhibitors. *Clinical Cancer Research* 16(9): 2571-2579, 2010.

11) R. J. Leeman-Neill, S. E. Wheeler, S. V. Singh, S. M. Thomas, R. R. Seethala, D. B. Neill, M. C. Panahandeh, E.-R. Hahm, S. C. Joyce, M. Sen, Q. Cai, M. L. Freilino, C. Li, D. E. Johnson, and J. R. Grandis. Guggulsterone enhances head and neck cancer therapies via inhibition of signal transducer and activator of transcription-3. *Carcinogenesis* 30(11): 1848-1856, 2009.

12) D. B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting* 25: 498-517, 2009.

13) D. B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics* 8: 20, 2009.

14) D. B. Neill. Cascade effects in heterogeneous populations. *Rationality and Society* 17(2): 191-241, 2005.

15) D. B. Neill. Evolutionary stability for large populations. *Journal of Theoretical Biology* 227(3): 397-401, 2004.

16) M. M. Wagner, F.-C. Tsui, J. Espino, W. Hogan, J. Hutman, J. Hersh, D. Neill, A. Moore, G. Parks, C. Lewis, and R. Aller. A national retail data monitor for public health surveillance. *Morbidity and Mortality Weekly Report, Supplement on Syndromic Surveillance* 53: 40-42, 2004.

17) D. B. Neill. Optimality under noise: higher memory strategies for the Alternating Prisoner's Dilemma. *Journal of Theoretical Biology* 211(2): 159-180, 2001.

### **Refereed Conference Proceedings (15)**

18) X. Jiang, D. B. Neill, and G. F. Cooper. Generalized AMOC curves for evaluation and improvement of event surveillance. *Proceedings of the American Medical Informatics Association Annual Symposium*, 281-285, 2009.

19) S. Hasan, G. T. Duncan, D. B. Neill, and R. Padman. Towards a collaborative filtering approach to medication reconciliation. *Proceedings of the American Medical Informatics Association Annual Symposium*, 288-292, 2008.

20) K. Das, J. Schneider, and D. B. Neill. Anomaly pattern detection in categorical datasets. *Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 169-176, 2008.

21) C. A. Harle, D. B. Neill, and R. Padman. An information visualization approach to classification and assessment of diabetes risk in primary care. *Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics*, 2008.

- 
- 22) M. Makatchev and D. B. Neill. Learning outbreak regions in Bayesian spatial scan statistics. *Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health Care Applications*, 2008.
- 23) D. B. Neill, A. W. Moore, and G. F. Cooper. A Bayesian spatial scan statistic. In Y. Weiss, et al., eds. *Advances in Neural Information Processing Systems 18*, 1003-1010, 2006.
- 24) M. R. Sabhnani, D. B. Neill, A. W. Moore, A. Dubrawski, and W.-K. Wong. Efficient analytics for effective monitoring of biomedical security. *Proceedings of the International Conference on Information and Automation*, 2005.
- 25) D. B. Neill, A. W. Moore, M. R. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. *Proceedings of the 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 218-227, 2005.
- 26) D. B. Neill and A. W. Moore. Anomalous spatial cluster detection. *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.
- 27) M. R. Sabhnani, D. B. Neill, A. W. Moore, F.-C. Tsui, M. M. Wagner, and J. U. Espino. Detecting anomalous patterns in pharmacy retail data. *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.
- 28) P. Hsiung, A. Moore, D. Neill, and J. Schneider. Alias detection in link data sets. *Proceedings of the First International Conference on Intelligence Analysis*, 2005.
- 29) D. B. Neill, A. W. Moore, F. Pereira, and T. Mitchell. Detecting significant multidimensional spatial clusters. In L. K. Saul, et al., eds., *Advances in Neural Information Processing Systems 17*, 969-976, 2005.
- 30) D. B. Neill and A. W. Moore. Rapid detection of significant spatial clusters. *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 256-265, 2004.
- 31) D. B. Neill and A. W. Moore. A fast multi-resolution method for detection of significant spatial disease clusters. In S. Thrun, et al., eds., *Advances in Neural Information Processing Systems 16*, 651-658, 2004.
- 32) D. B. Neill. Cooperation and coordination in the Turn-Taking Dilemma. *Proceedings of the Ninth Conference on Theoretical Aspects of Rationality and Knowledge*, 231-244, 2003.

### **Refereed Conference and Journal Abstracts (21)**

- 33) D. B. Neill, R. Padman, F. Solano, J. Zgibor, and C. Harle. Clinician information needs for data visualization based diabetes risk assessment and guideline compliance. Accepted to *American Medical Informatics Association Annual Symposium*, 2011.
- 34) D. B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate spatial biosurveillance. *Emerging Health Threats Journal*, 4:s42, 2011.

- 
- 35) D. B. Neill and Y. Liu. Generalized fast subset sums for Bayesian detection and visualization. *Emerging Health Threats Journal*, 4:s43, 2011.
- 36) H. Zheng, R. Padman, D. B. Neill, and S. Hasan. A comparison of collaborative filtering methods for medication reconciliation. *Proceedings of the 13th International Congress on Medical Informatics*, 2010.
- 37) D. Gartner, R. Kolisch, R. Padman, and D. B. Neill. Early DRG classification of inpatients in hospitals. *Proceedings of the 35th Conference on Operational Research Applied to Health Services*, 2009.
- 38) D. B. Neill. Fast subset sums for multivariate Bayesian scan statistics. *Proceedings of the International Society for Disease Surveillance Annual Conference*, 2009. Available online at [www.syndromic.org](http://www.syndromic.org).
- 39) S. Speakman and D. B. Neill. Fast graph scan for scalable detection of arbitrary connected clusters. *Proceedings of the International Society for Disease Surveillance Annual Conference*, 2009. Available online at [www.syndromic.org](http://www.syndromic.org).
- 40) K. Das, J. Schneider, and D. B. Neill. Anomaly pattern detection for biosurveillance. *Advances in Disease Surveillance* 5: 19, 2008.
- 41) M. Makatchev and D. B. Neill. Learning outbreak regions for Bayesian spatial biosurveillance. *Advances in Disease Surveillance* 5: 45, 2008.
- 42) D. B. Neill. Fast and flexible outbreak detection by linear-time subset scanning. *Advances in Disease Surveillance* 5: 48, 2008.
- 43) D. B. Neill and W. L. Gorr. Detecting and preventing emerging epidemics of crime. *Advances in Disease Surveillance* 4: 13, 2007.
- 44) D. B. Neill and J. Lingwall. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* 4: 106, 2007.
- 45) D. B. Neill. Incorporating learning into disease surveillance systems. *Advances in Disease Surveillance* 4: 107, 2007.
- 46) D. B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *Advances in Disease Surveillance* 4: 259, 2007.
- 47) D. B. Neill, A.W. Moore, and G. F. Cooper. A multivariate Bayesian scan statistic. *Advances in Disease Surveillance* 2: 60, 2007.
- 48) D. B. Neill and M. R. Sabhnani. A robust expectation-based spatial scan statistic. *Advances in Disease Surveillance* 2: 61, 2007.

---

49) D. B. Neill, A. W. Moore, and G. F. Cooper. A Bayesian scan statistic for spatial cluster detection. *Advances in Disease Surveillance* 1: 55, 2006.

50) D. B. Neill, A.W. Moore, M. R. Sabhnani, and K. Daniel. An expectation-based scan statistic for detection of space-time clusters. *Advances in Disease Surveillance* 1: 56, 2006.

51) M. R. Sabhnani, D. B. Neill, A. W. Moore, F.-C. Tsui, M. M. Wagner, and J. U. Espino. Monitoring pharmacy retail data for anomalous space-time clusters. *Advances in Disease Surveillance* 1: 62, 2006.

52) D. B. Neill, A. W. Moore, and M. R. Sabhnani. Detecting elongated disease clusters. *Morbidity and Mortality Weekly Report, Supplement on Syndromic Surveillance* 54: 197, 2005.

53) D. B. Neill and A. W. Moore. A fast grid-based scan statistic for detection of significant spatial disease clusters. *Morbidity and Mortality Weekly Report, Supplement on Syndromic Surveillance* 53: 255, 2004.

### **Other Journal Papers (2)**

Note: These two papers were not peer-reviewed in the typical fashion, but both papers were selected as winners of the Mathematical Contest in Modeling (in 2000 and 2001 respectively). The 4-6 winning papers each year were published in the *UMAP Journal* (citations given below).

54) S. W. Malone, C. A. Miller, and D. B. Neill. Traffic flow models and the evacuation problem. *Undergraduate Journal of Mathematics and its Applications* 22(3): 273-292, 2001.

55) S. W. Malone, J. A. Mermin, and D. B. Neill. Air traffic control. *Undergraduate Journal of Mathematics and its Applications* 21(3): 227-241, 2000.

### **Technical Reports and Theses (6)**

56) D. B. Neill. Detection of spatial and spatio-temporal clusters. Ph.D. thesis, Carnegie Mellon University, Department of Computer Science, Technical Report CMU-CS-06-142, 2006.

57) D. B. Neill and A. W. Moore. Detecting space-time clusters: prior work and new directions. Carnegie Mellon University, Department of Computer Science, Technical Report CMU-CS-05-115, 2005.

58) D. B. Neill. Evolutionary dynamics with large aggregate shocks. Carnegie Mellon University, Department of Computer Science, Technical Report CMU-CS-03-197, 2003.

59) D. B. Neill. An evolutionary resolution to the Finitely Repeated Prisoner's Dilemma paradox. Carnegie Mellon University, Department of Computer Science, Technical Report CMU-CS-03-155, 2003.

60) D. B. Neill. Fully automatic word sense induction by semantic clustering. Cambridge University, M.Phil. thesis, 2002.

---

61) D. B. Neill. Optimality under noise: higher memory strategies for the Alternating Prisoner's Dilemma. Duke University, undergraduate honors thesis, 2000.

### **Papers Under Review (3)**

[Paper #62 was accepted for publication, and moved to the "Refereed Journal Papers" section.]

63) D. B. Neill, E. McFowland III, and H. Zheng. "Fast subset scan for multivariate event detection," journal paper submitted to *Statistics in Medicine*, 2011.

64) S. Somanchi and D. B. Neill. "Fast graph structure learning from unlabeled data for event detection," conference paper submitted to *IEEE International Conference on Data Mining*, 2011.

65) K. Shao, Y. Liu, and D. B. Neill. "A generalized fast subset sums framework for Bayesian event detection," conference paper submitted to *IEEE International Conference on Data Mining*, 2011.

### **Papers in Draft Form (6)**

66) C. Harle, D. B. Neill, and R. Padman. "An information visualization approach to type 2 diabetes risk assessment in primary care," journal paper submitted to *Journal of Biomedical Informatics*, currently under revision.

67) D. Gartner, R. Kolisch, D. B. Neill, and R. Padman. "Machine learning approaches for early DRG classification of inpatient data in hospitals," journal paper submitted to *Artificial Intelligence in Medicine*, currently under revision.

68) K. Das, J. Schneider, and D. B. Neill. "Detecting anomalous groups in categorical datasets," conference paper, currently under revision.

69) R. F. Davies, D. B. Neill, et al. Detection of the Walkerton gastroenteritis outbreak by text mining of emergency room health records. Journal paper, currently under revision.

70) S. Speakman, E. McFowland III, and D. B. Neill. Scalable detection of anomalous patterns with connectivity constraints. Journal paper, to be submitted shortly to *Journal of Machine Learning Research*.

71) E. McFowland III, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. Journal paper, to be submitted shortly to *Journal of Machine Learning Research*.

### **Invited Talks and Tutorials (21)**

Most of these invited talks, presented at conferences such as the *International Workshop on Applied Probability* and the *Joint Statistical Meetings*, describe published research work or work in progress. Broader overview talks include: [73], in which I describe multiple uses of machine learning for population health surveillance; [75], in which I provide hints and tips for practitioners who wish to use spatial event detection methodologies in practice; [76], an invited

---

plenary talk which presents my vision of the next decade of disease surveillance research; and [82], my half-day tutorial on event detection at the KDD 2009 conference.

72) “Fast multivariate subset scanning for scalable cluster detection,” *Joint Statistical Meetings 2011*, Miami, FL, August 2011.

73) “Machine learning for population health and disease surveillance,” *Advanced Analytics Workshop*, Washington, DC, April 2011.

74) “Spatial and subset scanning for multivariate health surveillance,” *Data Fusion Research Meeting*, Ottawa, ON, March 2011.

75) “Spatial scanning tips and tricks for practical outbreak detection,” invited webinar for the International Society for Disease Surveillance, January 2011.

76) “Research challenges for biosurveillance: the next ten years” (invited plenary), *International Society for Disease Surveillance Annual Conference*, Park City, UT, December 2010.

77) Fast generalized subset scan for anomalous pattern detection,” *INFORMS Annual Conference*, Austin, TX, November 2010. Joint work presented by Edward McFowland III.

78) “Scalable detection of anomalous patterns with connectivity constraints,” *INFORMS Annual Conference*, Austin, TX, November 2010. Joint work presented by Skyler Speakman.

79) “Fast subset sums for scalable Bayesian detection and visualization,” *Fifth International Workshop on Applied Probability*, Madrid, Spain, July 2010.

80) “Fast subset scanning for multivariate event detection,” *ENAR 2010 Annual Meeting*, New Orleans, LA, March 2010.

81) “Application of spatial scan statistic methods to crime hot spot analysis,” *Tenth Crime Mapping Research Conference*, New Orleans, LA, August 2009.

82) “Event detection,” half-day tutorial (with Weng-Keen Wong). *15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Paris, France, June 2009.

83) “Multivariate Bayesian scan statistics for event detection and characterization,” *Twelfth Biennial CDC/ATSDR Symposium on Statistical Methods*, Decatur, GA, April 2009.

84) “A nonparametric scan statistic for multivariate spatial biosurveillance,” *Joint Statistical Meetings 2008*, Denver, CO, August 2008.

85) “Linear-time subset scanning,” *Fourth International Workshop on Applied Probability*, Compiègne, France, July 2008.

86) “Multivariate event detection and characterization,” *Washington Statistical Society Seminar*, Washington, DC, May 2008.

---

87) “Multivariate outbreak detection and characterization,” *Donald A. B. Lindberg Lecture and Symposium*, Pittsburgh, PA, May 2008.

88) “A multivariate Bayesian method for spatial biosurveillance,” *Joint Statistical Meetings 2007*, Salt Lake City, UT, July 2007.

89) “Monitoring multivariate spatial time series data for disease outbreak detection,” *27th Annual International Symposium on Forecasting*, New York, NY, June 2007.

90) “Bayesian disease surveillance by detection of anomalous clusters,” *Third ECADS Syndromic Surveillance Conference*, Ottawa, ON, October 2006.

91) “Bayesian disease surveillance by detection of anomalous clusters,” *Third International Workshop on Applied Probability*, Storrs, CT, May 2006.

92) “Scaling up geographic disease surveillance,” *Second ECADS Syndromic Surveillance Conference*, Ottawa, ON, June 2005.

### **Funded Projects (as Principal Investigator) (3)**

93) NSF IIS-0953330, Neill (PI), 7/1/2010-6/30-2015, funded by National Science Foundation. “CAREER: Machine Learning and Event Detection for the Public Good.” This project will create and explore novel methods for detection of emerging events in massive, complex, real-world datasets. This research will be integrated with a multi-pronged educational initiative to incorporate machine learning into the public policy curriculum. Total award: \$529,962.

94) UPMC Healthcare Innovation Grant, Neill (PI), unrestricted gift awarded 11/8/2010, funded by University of Pittsburgh Medical Center- Technology Development Center. “Anomalous Pattern Detection from Healthcare Data Streams.” This project will apply novel pattern detection methods to detect anomalous patterns of patient care. Total award: \$121,503.

95) NSF IIS-0916345, Neill (PI), 8/1/2009-7/31/2012, funded by National Science Foundation. “III: Small: Fast Subset Scan for Anomalous Pattern Detection.” This project will develop new, general subset scan methods for efficient pattern detection in massive datasets. Total award: \$499,991.

### **Funded Projects (as co-PI or other roles) (6)**

96) UPMC Healthcare Innovation Grant, Padman (PI), unrestricted gift awarded 11/8/2010, funded by University of Pittsburgh Medical Center- Technology Development Center. “Information Visualization for Cognitively Guided Decision Making for Diabetes Risk Assessment and Guideline Compliance.” This project will develop and evaluate novel information visualization tools and methods for improving diabetes care. Total award: \$110,120. Role: co-PI.

---

97) NSF IIS-0911032, Dubrawski (PI), 9/1/2009-8/31/2013, funded by National Science Foundation. “III: Large: Discovering Complex Anomalous Patterns.” This project will develop an integrated probabilistic framework for pattern discovery, incorporating detection, characterization, explanation, and learning from user feedback. Total award: \$2,598,153. Heinz portion: \$572,569. Role: co-PI.

98) CRTI-08-190RD, Davies (PI), 7/2009-6/2013, funded by CRTI. “Data Fusion Solutions for Monitoring CBRNE Threats.” This project focuses on general solutions for integrating multiple data sources for public health surveillance and integrates these solutions into two specific applications, detection of severe outbreaks in hospitalized patients and surveillance of events related to illicit substance abuse. Total award: \$3,000,000. CMU subcontract will be approximately \$25,000. Role: Technical team, expert in statistical detection methods and data mining.

99) CDC 8-R01-HK000020, Dubrawski (PI), 9/30/2006-9/29/2008, funded by Centers for Disease Control and Prevention. “Efficient, Scalable, Multisource Surveillance Algorithms for BioSense”. This project will develop multivariate Bayesian biosurveillance methods for inclusion in the BioSense system. Total award: \$1,198,409. Role: Co-PI.

100) NSF IIS-0325581, Cooper (PI), 9/1/2003-8/31/2008, funded by National Science Foundation. “ITR: Bayesian Modeling for Biosurveillance”. This project will develop novel Bayesian methodologies for the detection of disease outbreaks. CMU award: \$1,246,800. Role: senior personnel.

101) CRTI-06-0234TA, Davies (PI), 7/2007-7/2010, funded by CRTI. “Advanced Syndromic Surveillance and Emergency Triage (ASSET)”. This project will develop and deploy a system for syndromic surveillance of Emergency Department data in Ottawa, Ontario, for earlier detection of disease outbreaks and bioterrorist attacks. Total award: \$2,000,000. CMU subcontract: \$25,475. Role: Technical team, expert in statistical detection methods and data mining.

### **Other Work in Progress (15)**

102) “Detecting and preventing emerging epidemics of crime.” Joint work with Prof. Wil Gorr, students Amrut Nagasunder and YongJei Lee, and the Chicago Police Department (CPD). Preliminary results on Pittsburgh crime data were presented at the 2007 *National Syndromic Surveillance Conference* [43] and 2009 *Crime Mapping Research Conference* [81]. A full paper is in progress; our CrimeScan software based on this work is currently in day-to-day use by the CPD for crime prediction and tactical deployment.

103) “Nonparametric scan statistics for multivariate event detection and characterization.” Preliminary results were presented at the 2007 *National Syndromic Surveillance Conference* [44] and the 2008 *Joint Statistical Meetings* [84]. A full paper is in progress; the nonparametric scan is also an important component of our Fast Generalized Subset Scan framework [71].

---

104) “Fast graph structure learning from unlabeled data for event detection,” with Sriram Somanchi. Extended journal version of [64]. Our recent extensions of this work have the potential to scale up our graph learning algorithm to graphs with thousands of nodes.

105) “A generalized fast subset sums framework for Bayesian event detection,” with Kan Shao and Yandong Liu. Extended journal version of [65]. Our current extensions of this work will enable joint learning of an event’s center location, neighborhood size, sparsity, and effects on the monitored data stream from fully labeled data; we are also working on extensions to partially labeled data, where only a small subset of the affected locations are provided.

106) “A semantic scan statistic for detection of novel disease outbreaks,” with Yandong Liu. We combine new text mining techniques (based on incremental topic modeling) with multivariate spatial scan statistics, in order to detect emerging patterns of keywords in patient chief complaints. This can be used for automatic detection of never-before-seen outbreaks with surprising patterns of symptoms, such as “patient’s skin turned green and his nose fell off,” as well as enhancing detection power for any outbreak with a more specific pattern of symptoms (e.g. bloody stools), as compared to the typical approach based on lumping cases into broad, predefined symptom categories (e.g. respiratory). Empirical evaluation on Emergency Department data and journal paper writeup are in progress.

107) “Linear-time subset scanning for tensor data,” with Tarun Kumar. We have extended the multivariate linear-time subset scanning approach of [63] to tensor data. The work in [63] 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. In disease surveillance, this could be used to scan over subpopulations (e.g. age groups, gender, socioeconomic status, and race/ethnicity) in addition to scanning over space, time, and subsets of the monitored data streams.

108) “Fast penalized scan statistics by additive linear-time subset scanning,” with Skyler Speakman. We have discovered a new method of incorporating penalty terms into the efficient event detection methods facilitated by the linear-time subset scanning framework. This allows us to incorporate penalties on region size, soft proximity constraints (based on the distance of each location from the region center), and most interestingly, we can use this to detect dynamic patterns which change over time but satisfy soft constraints on temporal consistency (e.g. the affected subset must be similar from one time step to the next). This work has many applications including tracking of detected patterns, epidemiological contact tracing, back-tracing the source of food-borne illnesses, etc. In preliminary work presented as Skyler’s Second Heinz Paper, we incorporated simple temporal consistency constraints to detect spreading contamination in water network, and demonstrated that the new penalized scan statistics substantially improve the timeliness and accuracy of detection.

109) “Multivariate GraphScan with connectivity constraints on locations and streams,” with Skyler Speakman and Rajas Lonkar. By combining the techniques described in our multivariate fast subset scan [63] and GraphScan [70] papers, we can jointly optimize over subsets of locations and streams with connectivity constraints on both the locations (to represent spatial adjacency and travel patterns) and the streams (to represent common co-occurrences, hierarchies of disease symptoms, etc.)

---

110) “Fast Generalized Subset Scan with real-valued attributes,” with Edward McFowland III. We are in the process of extending our Fast Generalized Subset Scan work [71] from categorical datasets to mixed datasets with both real- and categorical-valued attributes. Preliminary results were presented by Ed at this year’s CAARMS conference.

111) “Fast Generalized Subset Scan with multiple models,” with Edward McFowland III. This extension of the FGSS framework [71] assumes multiple known models (each represented by a Bayesian network) and detects anomalous subsets which do not fit any of the known models. Preliminary results will be presented by Ed at this year’s INFORMS conference.

112) “Integrating detection of known and unknown events,” with Edward McFowland III, Prof. Artur Dubrawski, and Prof. Jeff Schneider. As part of the Discovering Complex Anomalous Patterns project [97], we are working to combine detection of previously known and identified event types with detection of previously unknown (anomalous) events, by integrating the Fast Generalized Subset Scan [71] and Generalized Fast Subset Sums [65, 105] frameworks. Our eventual goal is to maintain a continually growing set of known event types and to learn new event types via active learning from user feedback.

113) “Global event-based disease surveillance,” with Skyler Speakman, Prof. John Brownstein (MIT), and the HealthMap project team. This project was facilitated by Skyler’s internship with the HealthMap group (Summer 2011). Our goals are to adapt count-based event detection methods to their event-based data, and to incorporate clickstream data from HealthMap users to automatically learn which events are most relevant. We have also been in discussions with the InSTEDD team about a similar event-based disease surveillance project, and are interested more broadly in crowdsourcing disease outbreak detection to reduce the burden on public health users.

114) “Variable importance methods to prioritize development spending and reduce the global health impact of diarrheal illness,” with Sean Green. A preliminary version of this work was included as part of Sean’s doctoral thesis; journal paper in progress.

115) “Predicting credit histories from Facebook profiles,” with Fred Stutzmann and Prof. Alessandro Acquisti.

116) “Predicting economic growth outbreaks,” with Prof. Samuel Malone (University of the Andes), Sriram Somanchi, and Amrut Nagasunder. The goal of this project is to use our novel event detection methodologies to analyze which factors (structural similarities, geography, bilateral trade, etc.) cause clusters of economic growth to emerge and spread.

### **Planned Future Work (11)**

117) “Automatic epidemiological contact tracing using cell phone location and proximity data,” based on extensions of our GraphScan approach [70, 108] for detecting dynamic events in networks. We hope to use data from the new Living Analytics Research Centre for this project.

118) “Extending event detection to Web- and society-scale data,” with Sriram Somanchi. We propose to use multi-resolution sampling methods based on the fields of Combinatorial Property Detection and Combinatorial Group Detection to develop sub-linear time approximation algorithms with provable performance guarantees, thus enabling us to scale up our current event

---

detection approaches to massive datasets consisting of billions or even trillions of data records. This work will also address the related question of event detection based on huge quantities of streaming data, which is an important component of our UPMC-funded work on detecting anomalous patterns of patient care [94].

119) “Automatic prioritization of health data sources for optimal and cost-effective data collection,” combining outbreak simulation with supervised learning to predict the marginal improvement in detection time that would be gained by adding a new data source or improving the coverage of an existing data source.

120) “Integrating sensor placement and sensor fusion in noisy sensor networks,” combining our linear-time subset scanning approaches for event detection from multiple sources [63] with efficient sensor placement approaches based on submodularity (Guestrin et al.).

121) “Competitive active learning.” Joint work with SCS faculty member Jeff Schneider. Active learning is a field of machine learning which considers optimal query selection; we will consider the case where there are multiple learners in competition with each other, and the goal is to choose queries which are optimally more informative to the learner than to the competitors. Example applications include sharing of intellectual property between competitors and collaborative exploration projects such as oil drilling.

122) “Using Twitter feeds for traffic modeling and optimization,” with Prof. Robert Hampshire and Lavanya Marla.

123) “Using Internet search patterns for the public good,” with Prof. Alessandro Acquisti.

124) “Using the wisdom of crowds for medical image labeling,” with Skyler Speakman and Prof. Laura Dabbish.

125) “Detecting cybercrime networks,” with Nicholas Christin.

126) “Tracking patterns of hospital-acquired infections,” with Dr. Rick Davies (Ottawa Heart Institute) and the Data Fusion project team. Data acquisition is in progress.

127) “Integrated population health surveillance,” with Brett Goldstein (Chief Data Officer, City of Chicago). The goal of this proposed work, using Chicago as a testbed, will be simultaneous and integrated monitoring of multiple data sources indicative of population health to detect emerging events, predict future trends, and identify unexpected relationships between different factors influencing the health of individuals and communities. Some of these factors include: infectious disease outbreaks; chronic illnesses (asthma, diabetes, etc.); accident and injury (e.g. traffic accidents); crime and violence; quality of air, water, food, and sanitation; wealth and poverty; sufficient access to necessary resources (hospitals, pharmacies, groceries, child care, public transportation, etc.); levels of stress, depression, and overall happiness.