RESEARCH INTERESTS- DANIEL B. NEILL

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 techniques for discovery of emerging events and other relevant patterns in complex, massive, and high-dimensional data. I apply these novel methods to create, develop, and deploy systems that directly enhance the public good, in domains ranging from public health and patient care, to law enforcement and urban analytics, to human rights and conflict. I work directly with a variety of organizations in the public and private sectors, including public health practitioners, hospitals, police departments, and city leaders, to develop data-driven decision support systems that can improve public health, safety, and security.

Much of my pattern detection work has focused 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[1-2], **law enforcement and urban analytics**, e.g., prediction of crime patterns using offense reports and 911 calls[3-4], and identifying emerging citizen needs using 311 calls for service, and **health care**, e.g., discovering anomalous patterns of care with significant impacts on patient outcomes[5], and detecting prostate cancer in digital pathology slides[6]. I have also applied my work to numerous other areas, including prediction of civil unrest[7], early detection of emerging patterns of human rights events[8], network intrusion detection[9-10], customs monitoring of container shipments[9-10], physical infrastructure monitoring[11-12], classification and visualization of chronic disease risk[13], detection of omissions in patients' medication lists[14], and hospital length of stay management[15].

Many 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, 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 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 and scientific discovery.

One key methodological idea of this work is **subset scanning**: we frame the pattern detection problem as a search over subsets of the data, in which we define a measure of the "interestingness" or "anomalousness" of a subset, and maximize this "score function" over all potentially relevant subsets. Subset scanning often improves detection power as compared to heuristic methods, which are not guaranteed to find optimal subsets, top-down detection methods, which fail to detect small-scale patterns that are not evident from global aggregates, and bottom-up detection methods, which fail to detect subtle patterns that are only evident when a group of data records are considered collectively. Of course, subset scanning creates both statistical and computational challenges, the most serious of which is the computational infeasibility of exhaustively searching over the exponentially many subsets.

A key breakthrough of my recent work was the **fast subset scan**[16], which can efficiently identify the most interesting, anomalous, or relevant subsets of data records without an exhaustive search. This enables us to solve detection problems in milliseconds that would previously have been computationally infeasible, requiring millions of years to solve. However, fast subset scan only solves the unconstrained best subset problem, thus creating additional challenges as to how we can incorporate real-world constraints. Our recently developed fast subset scan approaches can find optimal subsets subject to constraints on spatial proximity[16], graph connectivity[17], group self-similarity[10], or temporal consistency[12]. They can be applied to univariate[16], multivariate[18], or multidimensional tensor[19] datasets, spatial[16] or non-spatial[10] data, including complex data such as text[20-21], images[6], and social media[7-8], and can track and source-trace dynamically spreading patterns[12]. These methods have been applied to various domains including disease surveillance, patient care, crime prediction and urban analytics, demonstrating substantial improvements in the timeliness, accuracy, and specificity of pattern detection compared to the previous state of the art. Our ongoing work extends these novel detection approaches to address multiple other problem formulations, including learning graph structure[22], predicting future spread of events[23], identifying heterogeneous treatment effects in randomized controlled trials[24], continual pattern discovery[24], and classifier model validation and refinement[25].

My **past work** on event and pattern detection has advanced the state of the art in multiple ways. For example, the expectation-based scan statistics[26-27] enable more timely and accurate detection of events through better use of **spatial** and **temporal** information; the nonparametric[28], Bayesian[29], and subset aggregation[18] multivariate scan statistics improve detection power by integrating information from **multiple data streams**; and the Multivariate Bayesian Scan Statistic[30-32] incorporates **prior information** and historical data to accurately model and differentiate between **multiple types of events**. New pattern detection methods such as Anomalous Group Detection[33], Anomaly Pattern Detection[9], and Fast Generalized Subset Scan[10] enable accurate and computationally efficient detection of patterns in **general datasets**, while new methods for Linear-Time Subset Scanning[16], Additive Linear-Time Subset Scanning[34], and Fast Subset Sums[31-32] enable **scalable detection** of the most anomalous patterns.

My **recent work** has mainly focused on three areas: first, we have developed novel subset scan methods such as the semantic scan statistic[21], hierarchical linear-time subset scanning[6], and non-parametric heterogeneous graph scan[7], that can incorporate massive, complex, heterogeneous, and unstructured data from multiple sources, including **rich text data** such as Emergency Department complaints and electronic health records[35], **massive image data** such as digital pathology slides[6], and **heterogeneous social media data** such as Twitter[7-8]. Second, we have developed novel Gaussian process inference and kernel methods, for scalable **event prediction**[4], **leading indicator selection**[36], **causal inference**[37], and **changepoint detection**[38]. Third, we are extending our detection approaches to many other problem settings, ranging from **graph structure learning**[22] to **improving classifier performance** through discovery and correction of systematic errors[25]. This methodological work provides a general and flexible basis for efficiently solving a vast array of real-world pattern detection problems.

One of my primary research goals has been to translate our methodological advances into **real-world systems** that can be deployed and used to benefit public health, safety, and security. For example, my disease surveillance methods have been in use by multiple state and local public health departments in the U.S., Canada, and Sri Lanka, for early detection of emerging disease outbreaks. My CityScan methodology and software were incorporated into the Chicago Police Department's day-to-day policing operations for crime prevention through targeted deployment of patrols, and have provided them with substantial value in their day to day operations: “based upon deployment suggestions indicated in the CityScan intelligence reports, important arrests were affected, weapons were seized, and crimes were prevented.”[39] Working with Chicago city leaders, we have applied CityScan to predict and prevent rodent complaints. Through advance prediction of locations where rodents are likely to occur, CityScan enables cities to more precisely target their proactive rodent baiting crews and other prevention measures. We are currently conducting a randomized, controlled experiment to determine whether we can reduce rodent complaints by predicting, targeting, and preventing rat infestations before they occur. Additional deployments are in progress in Pittsburgh, Chicago, and Baltimore.

Additional papers, presentations, and more detailed project descriptions are available on the Event and Pattern Detection Laboratory web page (http://epdlab.heinz.cmu.edu).

References:

[1]D. B. Neill. New directions in artificial intelligence for public health surveillance. *IEEE Intelligent Systems* 27(1): 56-59, 2012.

[2]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

[3]D. B. Neill and W. L. Gorr. Detecting and preventing emerging epidemics of crime. *Advances in Disease Surveillance* 4: 13, 2007.

[4]S. Flaxman, A. G. Wilson, D. B. Neill, H. Nickisch, and A. Smola. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. *Proc. 32nd Intl. Conf. on Machine Learning, JMLR: W&CP* 37, 2015.

[5]D. B. Neill. Using artificial intelligence to improve hospital inpatient care. *IEEE Intelligent Systems*, 28(2): 92-95, 2013.

[6]S. Somanchi and D. B. Neill. Discovering anomalous patterns in large digital pathology images. *Proceedings of the 8th INFORMS Workshop on Data Mining and Health Informatics*, 2013.

[7]F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. *Proc. 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1166-1175, 2014.

[8]F. Chen and D. B. Neill. Human rights event detection from heterogeneous social media graphs. *Big Data* 3(1): 34-40, 2015.

[9]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.

[10]E. McFowland III, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14: 1533-1561, 2013.

[11]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.

[12]S. Speakman, Y. Zhang, and D. B. Neill. Dynamic pattern detection with temporal consistency and connectivity constraints. *13th IEEE International Conference on Data Mining*, 697-706, 2013.

[13]C. A. Harle, D. B. Neill, and R. Padman. Information visualization for chronic disease risk assessment. *IEEE Intelligent Systems* 27(6): 81-85, 2012.

[14]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.

[15]D. Gartner, R. Kolisch, D. B. Neill, and R. Padman. Machine learning approaches for early DRG classification and resource allocation. *INFORMS Journal on Computing*, 27(4): 718-734, 2015.

[16]D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.

[17]S. Speakman, E. McFowland III, and D. B. Neill. Scalable detection of anomalous patterns with connectivity constraints. *Journal of Computational and Graphical Statistics*, 24(4): 1014-1033, 2015.

[18]D. B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. *Statistics in Medicine,* 32: 2185-2208, 2013.

[19]D. B. Neill and T. Kumar. Fast multidimensional subset scan for outbreak detection and characterization. *Online Journal of Public Health Informatics* 5(1), 2013.

[20]Y. Liu and D. B. Neill. Detecting previously unseen outbreaks with novel symptom patterns. *Emerging Health Threats Journal* 4: 11074, 2011.

[21]K. Murray, C. Dyer, Y. Liu, and D. B. Neill. A semantic scan statistic for novel disease outbreak detection. Paper under revision.

[22]S. Somanchi and D. B. Neill. Fast graph structure learning from unlabeled data for event detection. Submitted for publication.

[23]S. Speakman. *Fast Constrained Subset Scanning for Pattern Detection*. Ph.D. thesis, H.J. Heinz III College, Carnegie Mellon University, 2014. Paper in preparation.

[24]E. McFowland III. *Efficient Methods for Anomalous Pattern Detection and Discovery*. Ph.D. thesis, H.J. Heinz III College, Carnegie Mellon University, 2015. Papers in preparation.

[25]Z. Zhang and D. B. Neill. Detecting systematically poor fit in binary classifiers. Paper in preparation.

[26]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.

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

[28]D. B. Neill and J. Lingwall. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* 4: 106, 2007.

[29]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.

[30]D. B. Neill and G. F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 79: 261-282, 2010.

[31]D. B. Neill. Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine*, 30: 455-469, 2011.

[32]K. Shao, Y. Liu, and D. B. Neill. A generalized fast subset sums framework for Bayesian event detection. *Proceedings of the 11th IEEE International Conference on Data Mining*, 617-625, 2011.

[33]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.

[34]S. Speakman, S. Somanchi, E. McFowland III, and D. B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 2016, in press.

[35]M. Nobles, L. Deyneka, A. Ising, and D. B. Neill. Identifying emerging novel outbreaks in textual emergency department data. *Online Journal of Public Health Informatics* 7(1): e45, 2015.

[36]S. Flaxman, D. B. Neill, and A. Smola. Correlates of homicide: new space/time interaction tests for spatiotemporal point processes. Paper in preparation.

[37]S. Flaxman, D. B. Neill, and A. Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology*, 2016, in press.

[38]William Herlands, Andrew Gordon Wilson, Hannes Nickisch, Seth Flaxman, Daniel B. Neill, Willem van Panhuis, and Eric P. Xing. Scalable Gaussian processes for characterizing multidimensional change surfaces. *Proc. 19th International Conference on Artificial Intelligence and Statistics*, 2016, in press.

[39]Written communication from Officer Joseph Candella, Predictive Analytics Project Manager, Chicago Police Department. Contact information for Officer Candella and Deputy Chief Jonathan Lewin, CPD, can be provided upon request.