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.

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 violations[8], drug overdose surveillance[9], algorithmic biases in criminal justice risk assessment[10], network intrusion detection[11-12], customs monitoring of container shipments[11-12], physical infrastructure monitoring[13-14], classification and visualization of chronic disease risk[15], detection of omissions in patients' medication lists[16], and hospital length of stay management[17].

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.

One key methodological idea of this work is **subset scanning**[18]: 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**[19], 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[19], graph connectivity[20], group self-similarity[21], or temporal consistency[14]. They can be applied to univariate[19], multivariate[21], or multidimensional tensor[22] datasets, spatial[19] or non-spatial[12] data, including correlated data[23] and complex data such as text[24-25], images[6], and social media[7-8], and can track and source-trace dynamically spreading patterns[14]. 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 detection compared to the previous state of the art.

My **past work** on event 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 parametric[21], nonparametric[28], and Bayesian[29] multivariate scan statistics improve detection power by integrating information from **multiple data streams**; and the Bayesian scan statistics[29-32] integrate **prior information** and historical data to model and differentiate between **multiple event types**. Finally, our new methods[31-33] can efficiently and accurately detect **irregularly-shaped spatial clusters** rather than the fixed shapes used by traditional spatial scan approaches, improving detection power.

My **most recent methodological work** has mainly focused on three areas. First, we have developed novel subset scan methods such as the semantic scan statistic[25], 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[34], **massive image data** such as digital pathology slides, and **online social media data** such as Twitter. Second, our ongoing work extends these novel detection approaches to address many other problem settings, including learning graph structure[35], predicting future spread of events, continual pattern discovery, identifying heterogeneous treatment effects in both randomized controlled trials and observational data, and improving classifier performance through discovery and correction of systematic errors[10]. Finally, we have developed novel Gaussian process inference and kernel methods, for scalable **event prediction**[4], **leading indicator selection**[36], **causal inference**[37], and **change point detection**[38]. Most recently, we have effectively combined Gaussian processes with subset scan methods to analyze multiple, correlated streams of spatio-temporal data from urban settings[23], in order to model the continuous "pulse" of a city and detect anomalous events corresponding to population-wide changes in location, movement, or behavior.

Societal Impacts

The methodological work described above provides a general and flexible basis for efficiently solving a vast array of detection problems. One of my primary research goals has been to translate these methodological advances into **real-world systems** that can be used operationally to enhance the public good. 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 solutions** that can improve public health, safety, and security.

For example, my CrimeScan methodology and software for **crime prediction** has been used operationally by both the Chicago Police Department and the Pittsburgh Bureau of Police. CrimeScan predicts where geographic hot-spots of violent crime will occur, by detecting clusters of more minor crimes and other leading indicators and incorporating these clusters into a predictive model. Both departments have used CrimeScan to guide their day-to-day policing operations for crime prevention through targeted deployment of patrols. The Chicago PD has noted that CrimeScan provided them with substantial value in their day-to-day operations:

"CrimeScan was set up to run daily, completely autonomously, and predictions were sent via system-generated messages to police analysts within the Predictive Analytics Group. These messages were compiled into detailed intelligence reports which were disseminated through the chain of command. (…) Citywide response teams made routine use of these intelligence packages when making deployment decisions for both daily and long-term operations. (…) Based upon deployment suggestions indicated in the CrimeScan intelligence reports, important arrests were affected, weapons were seized, and crimes were prevented."[39]

With support from the R. K. Mellon Foundation, the Pittsburgh Bureau of Police has recently deployed our predictive model for crime prevention. We are in the process of performing a city-wide **randomized field trial** in Pittsburgh in order to quantify its impact on both violent and property crimes. Our team worked closely with PBP to make crime report data available on a real-time basis, improve geocoding, and provide crime maps that can be accessed by officers from mobile data terminals in their patrol cars. Predicted violent and property crime hot spots are overlaid on the map so that police can not only target these areas but have access to the context needed for effective prevention. Thus our work in Pittsburgh has helped both to enable PBP crime analysts, and to put real-time crime data and maps in the hands of PBP officers in the field.

Working with Chicago city leaders, we have developed and deployed CityScan, an extension of CrimeScan, 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, preventing rat infestations before they occur. The city of Chicago continues to use CityScan and claims that it is "20 percent more effective than the traditional method of baiting rats after they've been discovered."[40] We have also evaluated CityScan for the cities of Pittsburgh and Baltimore: our results suggest that substantial public health benefits could be gained through proactive rodent baiting in each city.

In the **disease surveillance** domain, my methodological approaches 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. Much of my early work along these lines was through large-scale funded collaborations (e.g., CDC BioSense and the National Biosurveillance Integration System) where I developed and contributed advanced detection methods but was not directly involved in system-building. Recently I have taken a more hands-on approach, working directly with three state and local public health departments. With the North Carolina DOH and New York City DOHMH, I am currently working to develop and deploy early warning systems for emerging "novel outbreaks" with previously unseen patterns of symptoms, as well as other patterns of public health interest that do not correspond to pre-defined syndrome groupings. I am also working with the Kansas DOH to analyze data from their Prescription Drug Monitoring Program and identify emerging trends that are predictive of future spikes in drug overdose deaths.

Additionally, I have been engaging with partners in the **healthcare industry**, both payers and providers, to develop and deploy approaches for discovering patterns of patient care that positively or negatively impact outcomes. Both UPMC and Highmark have funded our research and are looking for opportunities to deploy it in their healthcare systems, with potential benefits including improved patient outcomes, reduced costs, and new standards of care.

In addition to deploying my methods in information systems to directly impact the public good, I also hope to impact society through influencing policy at the federal, state, and local levels. For example, I served as a member of the **NSF Subcommittee on Youth Violence** commissioned by Congressman Frank Wolf in response to the shooting at Sandy Hook Elementary. Our report[41], focusing on risk factors associated with mass shootings and street violence, was presented to Congress and discussed at a hearing before the House Appropriations Commerce-Justice-Science (CJS) subcommittee. Several of my current projects, such as analyzing the heterogeneous causal impacts of building and neighborhood factors such as overcrowding and crime on individual-level health outcomes, also have potential to influence urban policy and planning in NYC.

Reputation and Recognition

My work and external reputation have been recognized in various ways. I was the recipient of a **National Science Foundation CAREER Award**, was named one of the **top ten AI researchers to watch**, and am serving as **Associate Editor** of four journals (*IEEE Intelligent Systems, Security Informatics, Decision Sciences, and ACM Transactions on Management Information Systems)*. I have been appointed as **AI and Health Department Editor** of *IEEE Intelligent Systems* and co-chair of two International Conferences on Smart Health. I served as expert panelist for the MacArthur Foundation's workshop on urban analytics and as a member of the NSF Subcommittee on Youth Violence described above. In public health, I serve as **advisor to the Board of Directors** for the International Society for Disease Surveillance (ISDS), have served as **Scientific Program Chair** of the ISDS Annual Conference, and have given several invited plenary talks and webinars on the future of the public health surveillance field. I have also given dozens of ISDS Annual Conference talks on new methodological approaches for surveillance, receiving the conference's **Best Research Presentation** award for my Bayesian spatial scanning work, and have helped ISDS to develop consultancies on various problems of critical importance to public health practice. On the methodological side, our work on Penalized Fast Subset Scanning[33] was recently selected as a **Best Paper** of the *Journal of Computational and Graphical Statistics* by the journal's editor in chief. At the university level, I have been awarded the **H. John Heinz III College Dean's Career Development Professorship**, which "recognizes junior and mid-level faculty members who have shown great achievement in their field of study", have been nominated for a NASPAA Spotlight Award for contributions to the public sector, and since 2012 my doctoral student advisees have earned six best student paper and dissertation awards.

Additional papers and more detailed project descriptions are available on my personal web page (http://www.cs.nyu.edu/~neill) and my (old) Event and Pattern Detection Laboratory web page (http://epdlab.heinz.cmu.edu). Please feel free to contact me at firstname.lastname@nyu.edu.

Daniel B. Neill, Ph.D.
Last updated: October 2017

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: 607-616, 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]D. B. Neill and W. Herlands. Machine learning for drug overdose surveillance. *Proceedings of the Bloomberg Data for Good Exchange Conference*, 2017.

[10]Z. Zhang and D. B. Neill. Identifying significant predictive bias in classifiers. *Proceedings of the NIPS Workshop on Interpretable Machine Learning for Complex Systems*, 2016.

[11]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.

[12]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.

[13]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.

[14]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.

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

[16]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.

[17]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.

[18]D. B. Neill. Subset scanning for event and pattern detection. In S. Shekhar and H. Xiong, eds., *Encyclopedia of GIS, 2nd ed.*, 2218-2228, 2017.

[19]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.

[20]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.

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

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

[23]W. Herlands, E. McFowland III, A. G. Wilson, and D. B. Neill. Gaussian process subset scanning for anomalous pattern detection in non-iid data. Submitted for publication.

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

[25]A. Maurya, K. Murray, Y. Liu, C. Dyer, W. W. Cohen, and D. B. Neill. Semantic scan: detecting subtle, spatially localized events in text streams. Submitted for publication.

[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]S. Speakman, S. Somanchi, E. McFowland III, and D. B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2): 382-404, 2016.

[34]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.

[35]S. Somanchi and D. B. Neill. Graph structure learning from unlabeled data for early outbreak detection. *IEEE Intelligent Systems*, 32(2): 80-84, 2017.

[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*, 7(2): 22:1-22:23, 2015.

[38]W. Herlands, A. G. Wilson, H. Nickisch, S. Flaxman, D. B. Neill, W. van Panhuis, and E. P. Xing. Scalable Gaussian processes for characterizing multidimensional change surfaces. *Proc. 19th International Conference on Artificial Intelligence and Statistics, JMLR: W&CP*, 51: 1013-1021, 2016.

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

[40]L. Poon. "Will Cities Ever Outsmart Rats?" Available at: https://www.citylab.com/solutions/2017/08/smart-cities-fight-rat-infestations-big-data/535407/.

[41]A substantially revised and extended version of our report was published as: B. J. Bushman, K. Newman, S. L. Calvert, G. Downey, M. Dredze, M. Gottfredson, N. G. Jablonski, A. Masten, C. Morrill, D. B. Neill, D. Romer, and D. Webster. Youth violence: what we know and what we need to know. *American Psychologist*, 71(1): 17-39, 2016.