IC Talk 2011: Machine Learning and Event Detection for the Public Good

Daniel B. Neill
Carnegie Mellon University
H.J. Heinz III College
neill@cs.cmu.edu

Daniel B. Neill (neill@cs.cmu.edu)
Assistant Professor of Information Systems, Heinz College
Courtesy Assistant Professor of Machine Learning and Robotics, SCS

My research is focused at the intersection of **machine learning** and **public policy**.

Increasingly critical importance of addressing global policy problems (disease pandemics, crime, terrorism…)

Continuously increasing size and complexity of policy data, and rapid growth of new and transformative technologies.

Machine learning has become increasingly essential for data-driven policy analysis and for the development of new, practical information technologies that can be directly applied **for the public good** (e.g. public health, safety, and security)

A quick advertisement for the new and growing MLP curriculum at CMU:

Joint Ph.D. Program in Machine Learning and Public Policy (MLD & Heinz)
Large Scale Data Analysis for Policy (introduction to ML for PPM students)
→ 10-830, Research Seminar in ML and Policy (ML & Heinz PhD students)
→ 10-831, Special Topics in Machine Learning and Policy

Event and Pattern Detection

Harnessing the Wisdom of Crowds

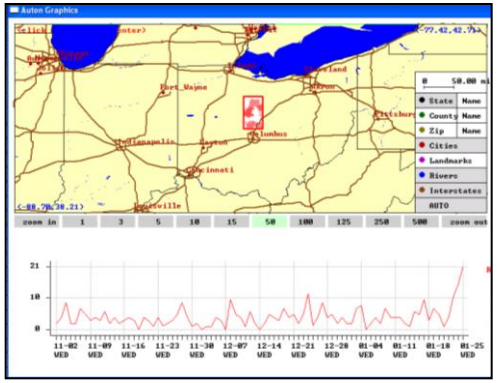
ML for the Developing World

Machine Learning and Health

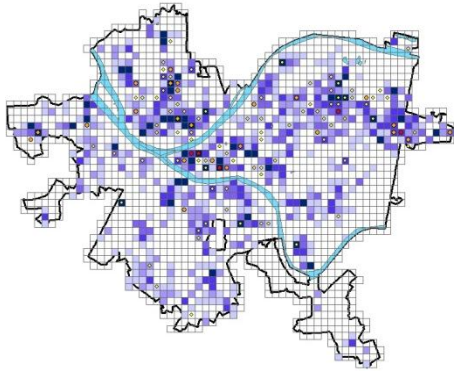
Daniel B. Neill (neill@cs.cmu.edu)
Assistant Professor of Information Systems, Heinz College
Courtesy Assistant Professor of Machine Learning and Robotics, SCS

My research has two main goals: to develop new machine learning methods for automatic **detection** of **events** and other **patterns** in massive datasets, and to apply these methods to improve the quality of public health, safety, and security.

<u>Customs monitoring</u>: detecting patterns of illicit container shipments



<u>Biosurveillance</u>: early detection of emerging outbreaks of disease



<u>Law enforcement</u>: detection and prediction of crime hot-spots

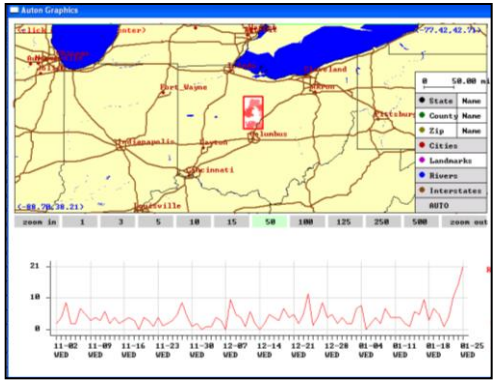
Our methods could have detected the May 2000 Walkerton *E. coli* outbreak two days earlier than the first public health response.

We are able to accurately predict emerging clusters of violent crime 1-3 weeks in advance by detecting clusters of more minor “leading indicator” crimes.

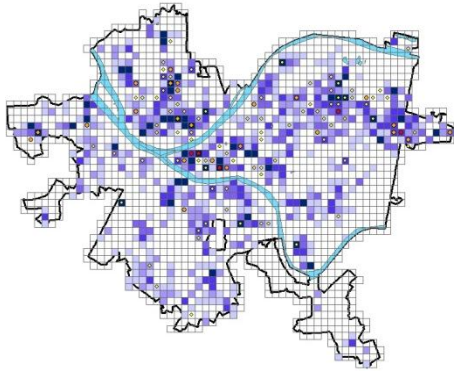
Daniel B. Neill (neill@cs.cmu.edu)
Assistant Professor of Information Systems, Heinz College
Courtesy Assistant Professor of Machine Learning and Robotics, SCS

My research has two main goals: to develop new machine learning methods for automatic **detection** of **events** and other **patterns** in massive datasets, and to apply these methods to improve the quality of public health, safety, and security.

Customs monitoring: detecting patterns of illicit container shipments



Biosurveillance: early detection of emerging outbreaks of disease



Law enforcement: detection and prediction of crime hot-spots

Our methods are currently in use for deployed biosurveillance systems in Ottawa and Grey-Bruce, Ontario; several other projects are underway.

We collaborate directly with the Chicago Police Department, and our “CrimeScan” software is already in day-to-day operational use for predictive policing.

Pattern detection by subset scan

One key insight that underlies much of my work is that pattern detection can be viewed as a **search** over subsets of the data.

Statistical challenges:
Which subsets to search?
Is a given subset anomalous?
Which anomalies are relevant?

Computational challenge:
How to make this search over subsets efficient for massive, complex, high-dimensional data?

New statistical methods enable more timely and more accurate detection by integrating **multiple data sources**, incorporating **spatial** and **temporal** information, and using **prior knowledge** of a domain.

New algorithms and data structures make previously impossible detection tasks computationally feasible and fast.

New machine learning methods enable our systems to learn from user feedback, modeling and distinguishing between relevant and irrelevant types of anomaly.

Disease surveillance

Bioterrorist attacks, e.g. release of anthrax spores

Avian influenza and other emerging infectious diseases

Early detection of disease outbreaks can save thousands of lives, and can significantly reduce the human and economic impacts of an outbreak.

Our solution: automatic surveillance

Automatic monitoring of electronically available public health data sources in near real-time.

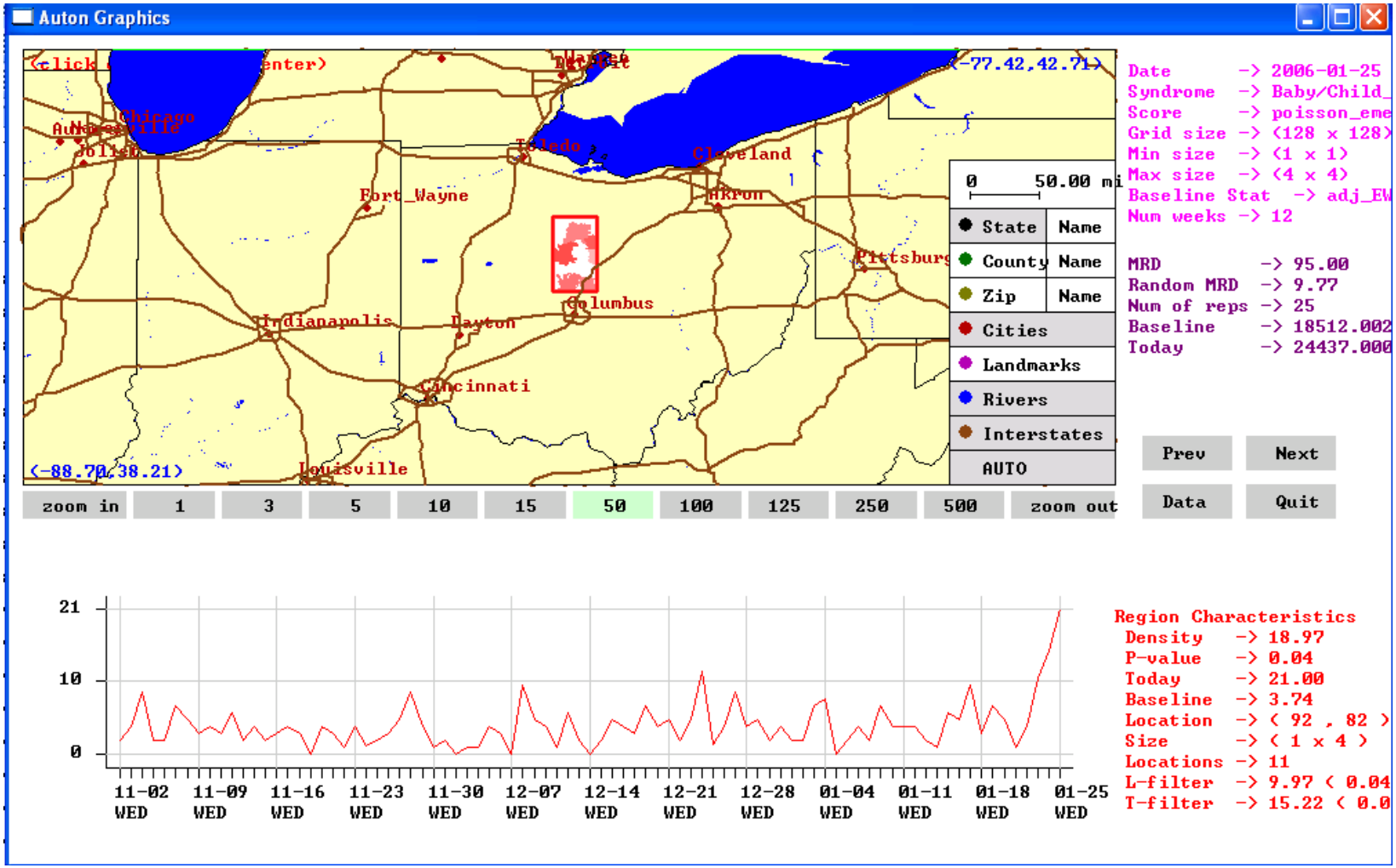
(Hospital ED visits, OTC medication sales, 911 calls, etc.)

Automatic detection of anomalous patterns that are indicative of an outbreak, using a variety of new statistical and ML methods.

Automatic reporting of alerts to public health.

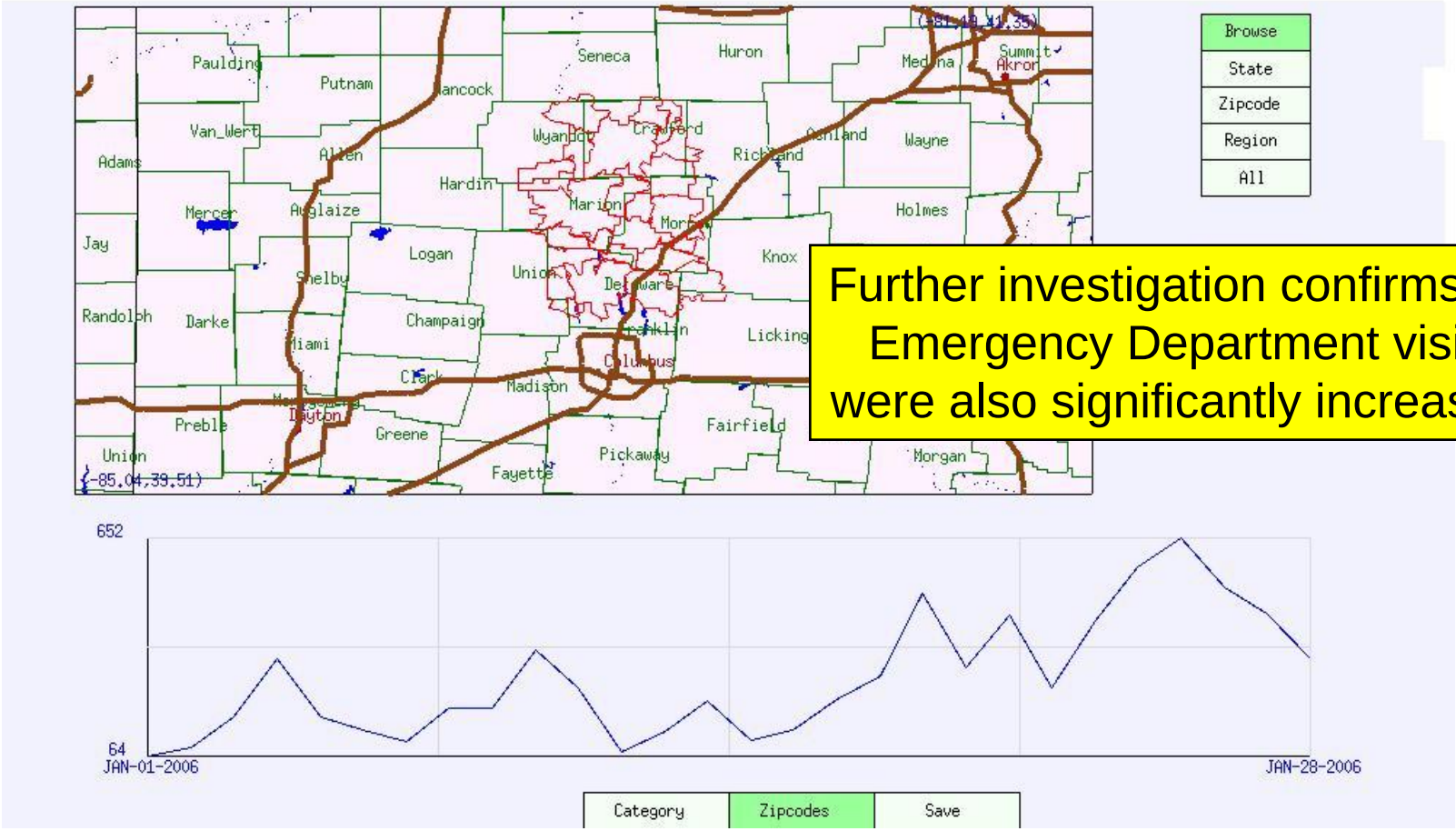
A recent potential outbreak

Spike in sales of pediatric electrolytes near Columbus, Ohio



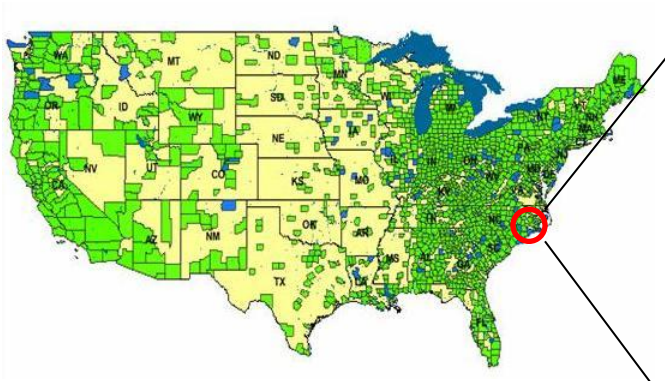
A recent potential outbreak

Spike in sales of pediatric electrolytes near Columbus, Ohio

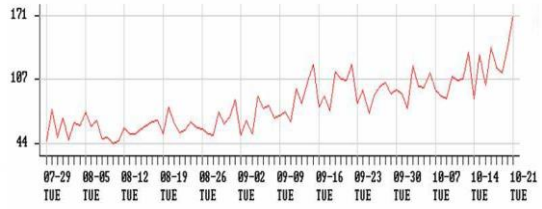


Further investigation confirms: GI Emergency Department visits were also significantly increased.

Under the hood: how does it work?



Daily health data from thousands of hospitals and pharmacies nationwide.



Time series of counts $c_{i,m}{}^t$ for each zip code s_i for each data stream d_m.

d_1 = respiratory ED

d_2 = constitutional ED

d_3 = OTC cough/cold

d_4 = OTC anti-fever

etc.

We want to obtain a complete **situational awareness** by integrating information from the multiple streams:

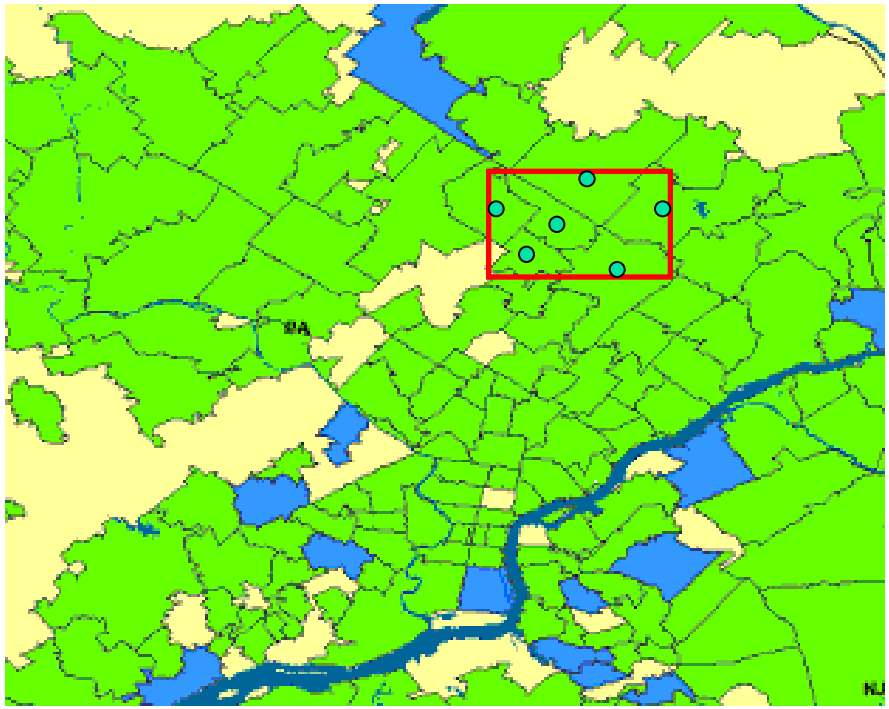
Detect emerging outbreaks.
Characterize the outbreak type.
Pinpoint affected areas.

<u>Expectation-based scan statistic</u>

1. Infer the expected count for each zip code, for each data stream, for each recent day.
2. Find regions where the recent counts for a subset of streams are higher than expected.

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

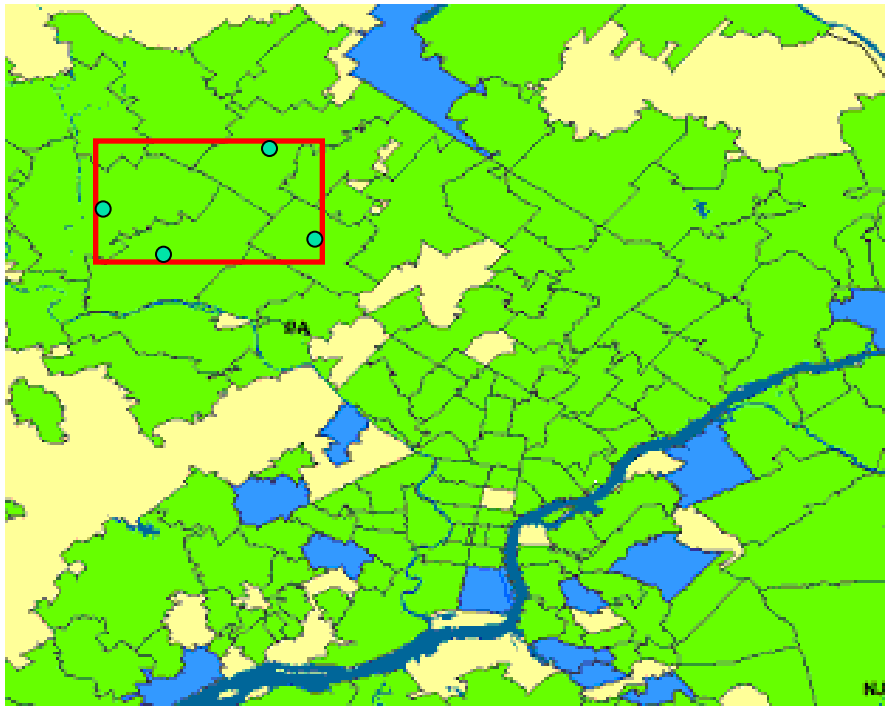


Rather than searching for single locations that are anomalous, we search for regions containing a group of anomalous locations.

Imagine moving a window around the scan area, allowing the window size, shape, and duration to vary.

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

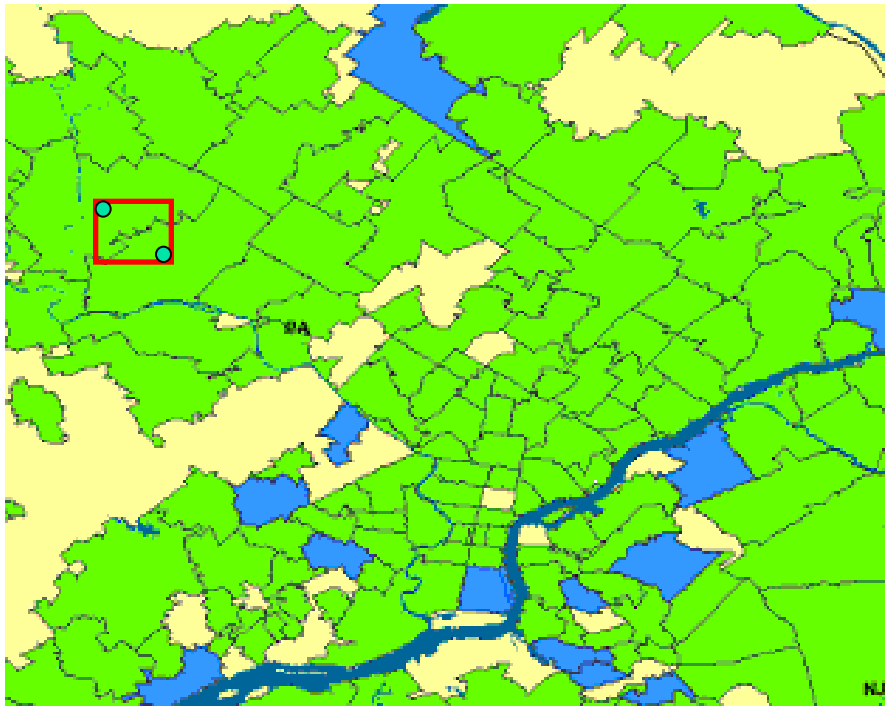


Rather than searching for single locations that are anomalous, we search for regions containing a group of anomalous locations.

Imagine moving a window around the scan area, allowing the window size, shape, and duration to vary.

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

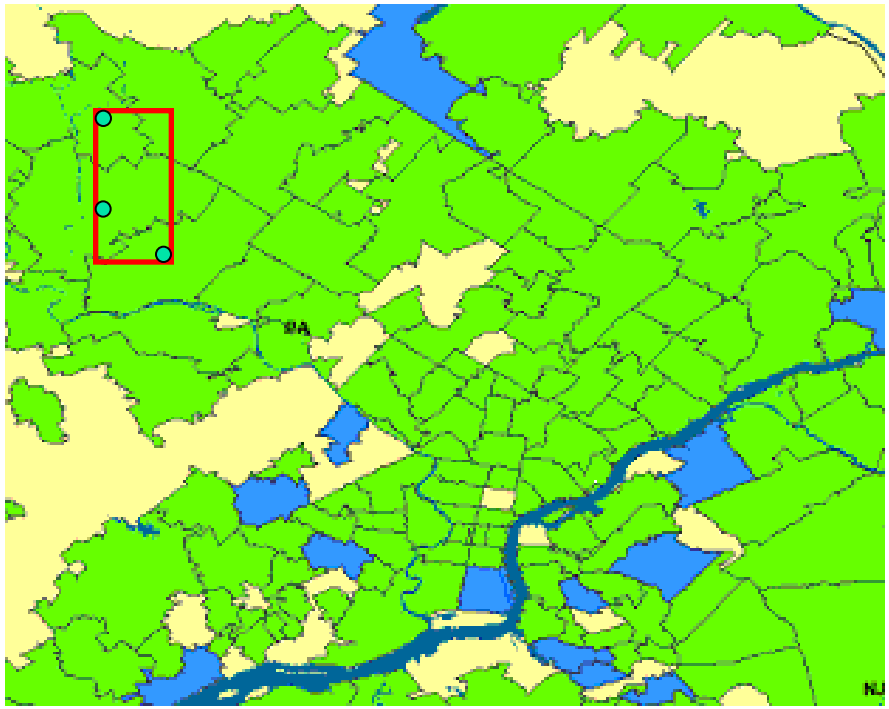


Rather than searching for single locations that are anomalous, we search for regions containing a group of anomalous locations.

Imagine moving a window around the scan area, allowing the window size, shape, and duration to vary.

The space-time scan statistic

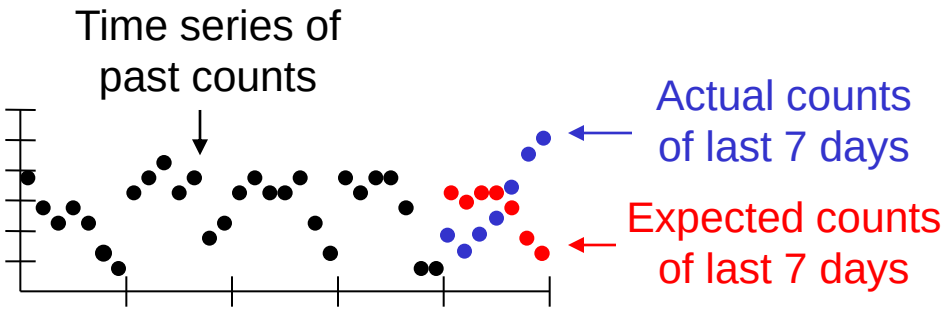
(Kulldorff, 2001; Neill & Moore, 2005)



Rather than searching for single locations that are anomalous, we search for regions containing a group of anomalous locations.

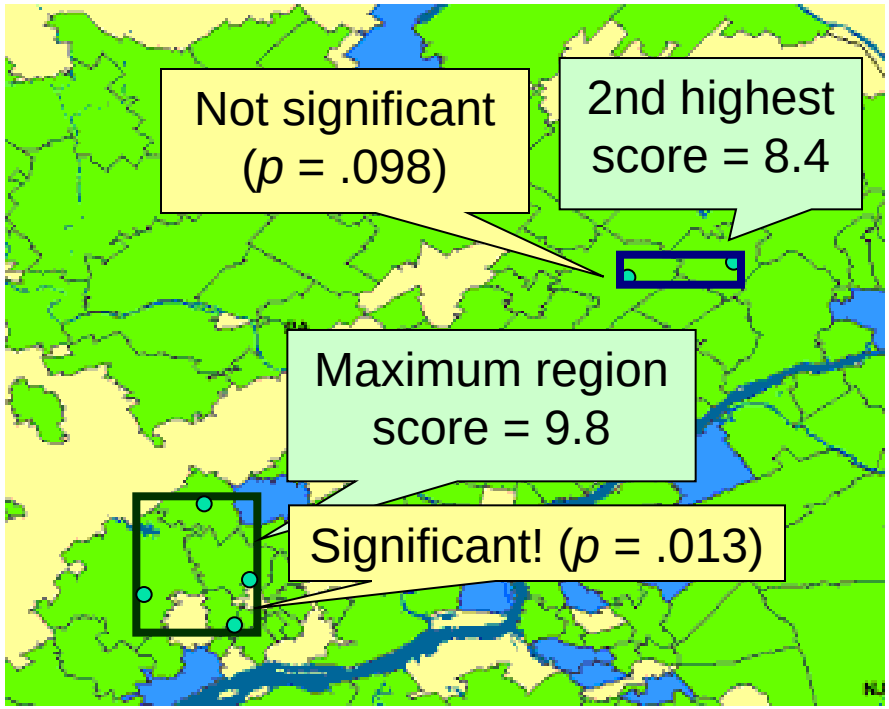
Imagine moving a window around the scan area, allowing the window size, shape, and duration to vary.

For each space-time region, we examine the aggregated time series, and compare actual to expected counts.



The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)



We find the highest-scoring space-time regions, where the score of a region is computed by the likelihood ratio statistic.

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

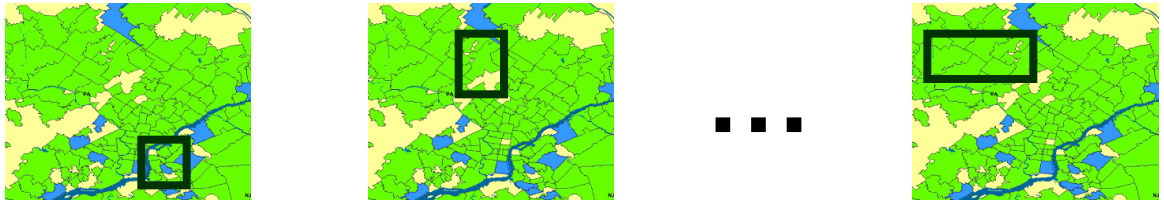
Alternative hypothesis: event in region S

Null hypothesis: no events

These are the most likely clusters… but how can we tell whether they are significant?

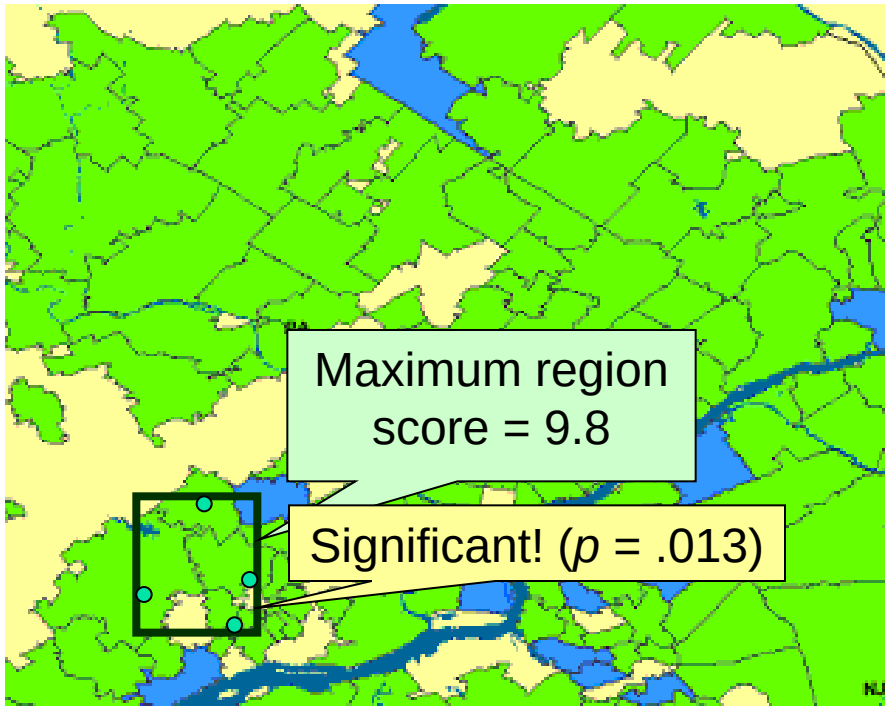
Answer: compare to the maximum region scores of simulated datasets under H_0.

$F_1^* = 2.4$ $F_2^* = 9.1$ … $F_{999}^* = 7.0$



The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)



Thus we end up with a list of the significant spatial clusters for each data stream.

Each cluster represents a potential event (e.g. an outbreak of disease) which may be of interest to the user.

- How do we:
  - accurately detect **irregularly-shaped** and **dynamic** clusters?
  - **efficiently** search over the billions of possible subsets?
  - improve detection by integrating **multiple data streams**?
  - reduce the number of **false positives** that aren't due to real events?
  - **generalize** these methods from spatial to other types of data?

Current Projects

- Fast Subset Scan for Pattern Detection (NSF IIS-0916345)
  - New, computationally efficient detection methods which scale to very large, high-dimensional datasets.
- Discovering Complex Anomalous Patterns (NSF IIS-0911032)
  - Integrating **detection**, **characterization**, **explanation**, and **learning**, in order to rapidly identify the most relevant patterns in complex data.
  - One application: detecting anomalous patterns of **patient care** in hospitals (with additional funding from University of Pittsburgh Medical Center).
- Machine Learning and Event Detection for the Public Good (CAREER)
  - Incorporating new and emerging **data sources** (cellular phones, sensor networks, Internet search queries, user-generated web content, etc.)
  - Developing **practical biosurveillance systems** for early outbreak detection (current deployments: U.S., Canada, Sri Lanka, India).
  - Predicting anomalous patterns of **violent crime** using leading indicators (in collaboration with the City of Chicago and Chicago Police Department).
  - Additional applications to food safety monitoring, fleet management, hospital-acquired illness, drug abuse, customs monitoring, many others.

Fast algorithms for subset scanning

Since there are exponentially many subsets of the data, it is often computationally infeasible to search all of them.

The most common approach is to use domain knowledge to restrict our search space: for example, we assume that an event will affect a contiguous spatial region, and often further restrict the region size and shape.

e.g. "search over circular regions centered at a data point" → only N^2 regions instead of 2^N.

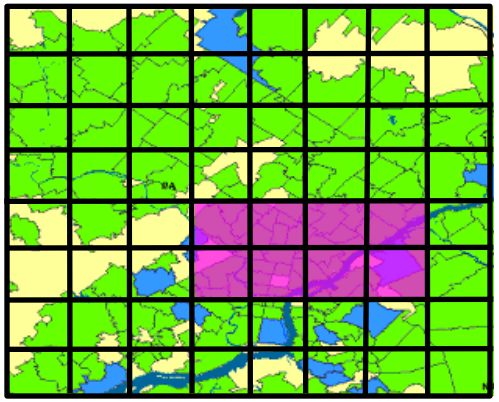
Another common approach is to perform a heuristic search. For example, we can greedily grow subsets starting from each data record, repeatedly adding the additional record that gives the highest scoring subset.

Tradeoff: much more efficient than naïve search, but not guaranteed to find highest scoring region.

In some cases, we can find the highest-scoring subsets **without** computing the scores of all possible subsets!

Fast spatial scan over rectangles

Consider searching over all rectangular regions for data aggregated to a N x N grid.

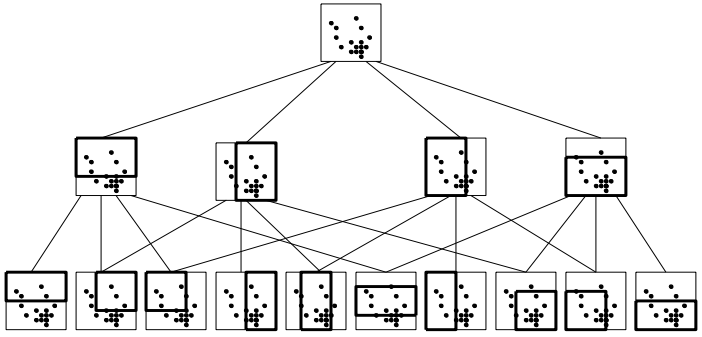


The number of search regions scales as $O(N^4)$, making an exhaustive search computationally infeasible for large N.

We can find the highest scoring clusters without an exhaustive search using **branch and bound**: we keep track of the highest region score found so far, and prune sets of regions with provably lower scores.

A new multi-resolution data structure, the **overlap-kd tree**, enables us to make this search efficient.

We can now monitor nationwide health data in 20 minutes vs. 1 week.



Linear-time subset scanning

- In certain cases, we can optimize F(S) over the exponentially many subsets of locations, while evaluating only O(N) regions.
- Many commonly used scan statistics have the property of linear-time subset scanning (LTSS):
  - Just sort the locations from highest to lowest priority according to some function…
  - … then search over groups consisting of the top-k highest priority locations, for k = 1..N.

The highest scoring subset is guaranteed to be one of these!

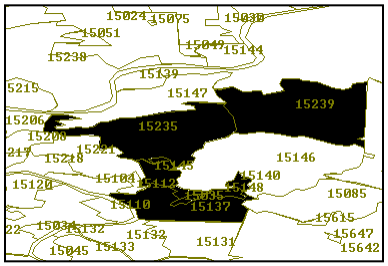
Sample result: we can find the **most anomalous** subset of Allegheny County zip codes in 0.03 sec vs. 10^{24} years.

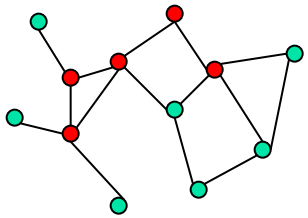
Fast Subset Scan for Pattern Detection

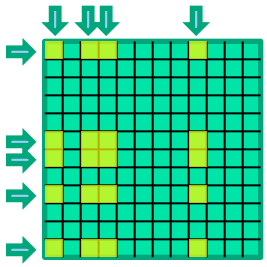
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the "best unconstrained subset" problem, and cannot be used directly for constrained optimization.

We are currently investigating how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

Proximity constraints	→	Fast spatial scan (irregular regions)
Multiple data streams	→	Fast multivariate scan
Connectivity constraints	→	Fast graph scan
Group self-similarity	→	Fast generalized subset scan





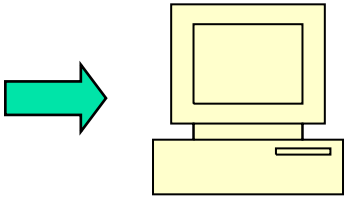


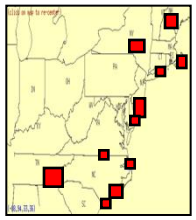
Other constraints? Shape, convexity, temporal consistency…
Other data types? Text, tensor data, dynamic graphs, etc.

Incorporating learning into detection

We have made major advances in detecting anomalous patterns, but not in determining which of these anomalies are relevant.

9/20, 15213, cough/cold, …
9/21, 15207, antifever, …
9/22, 15213, CC = cough, ...
1,000,000 more records…





Huge mass of data | Detection algorithm | Too many alerts | "What am I supposed to do with this?"

We must model and differentiate between multiple causes of a detected pattern, and provide only the relevant patterns to the user.

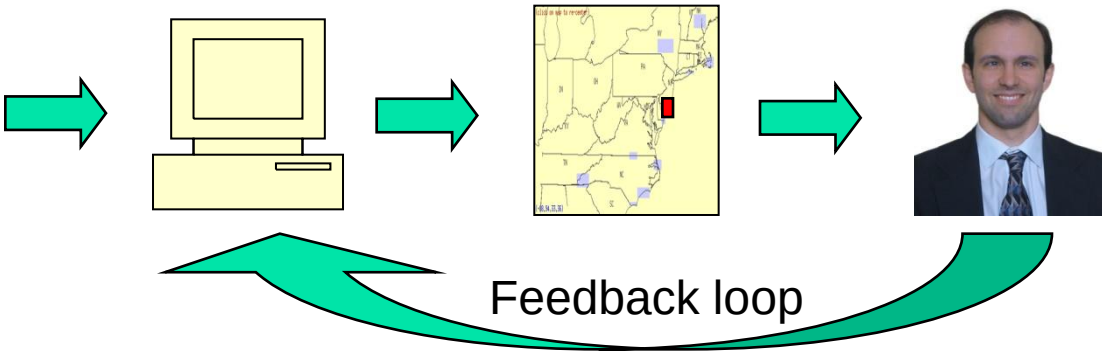
How can we classify patterns, and determine which ones are relevant to a given user at a given time?

Incorporating learning into detection

We have made major advances in detecting anomalous patterns, but not in determining which of these anomalies are relevant.

9/20, 15213, cough/cold, …
9/21, 15207, antifever, …
9/22, 15213, CC = cough, ...
1,000,000 more records…

Huge mass of data



We must model and differentiate between multiple causes of a detected pattern, and provide only the relevant patterns to the user.

How can we classify patterns, and determine which ones are relevant to a given user at a given time?

Incorporate user feedback into the detection process, and use it to learn models!

Active learning of new event types

In our proposed pattern discovery system, the user can define new classes "on the fly", by assigning a new label type to an example. The system can then find other potential examples of the new class in historical data, ask the user to label these, and learn a model for the new class.

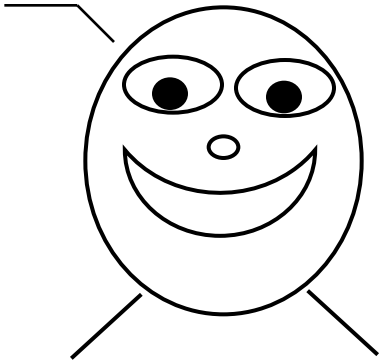
"Any clusters of interest today?"

"Yes, this appears to be an anthrax attack, based on increased OTC cough and fever."

"No, we don't see a corresponding increase in ED visits. I think this cluster is just a promotional sale."

"OK. Can you identify which of these historical clusters also correspond to promotional sales?"

"Yes, all of these. Any other clusters of interest?"



"Not really, just more seasonal flu and another promotional sale."

Challenges: Extending active learning from individual records to subsets
Learning complex models from partially labeled data
Learning a new model from a single labeled example

Discovering Complex Anomalous Patterns

- Our primary goal is to develop an integrated probabilistic framework for **pattern discovery** that allows the system and user to work interactively to discover and investigate anomalous patterns over groups of records.
- The framework will integrate four main components:
  - **Detection** of both known and previously unknown patterns, thus incorporating both model-based and anomaly-based detection methods.
  - **Characterization** of the pattern type, the affected subsets of the data records and attributes, and other parameters.
  - **Explanation** of why the pattern belongs to a particular pattern type, and why the system believes that it is relevant to the user.
  - **Active learning** of models from user feedback and/or partially labeled training data**,** with the goal of rapidly focusing the user's attention on the most relevant patterns.

While these methods will have wide applicability, our initial application will be detection of anomalous **patterns of care** in data from UPMC hospitals.

We want to find atypical treatment conditions that either improve patient outcomes ("best practices") or harm patient outcomes (e.g. systematic errors, improper hygiene, etc.)

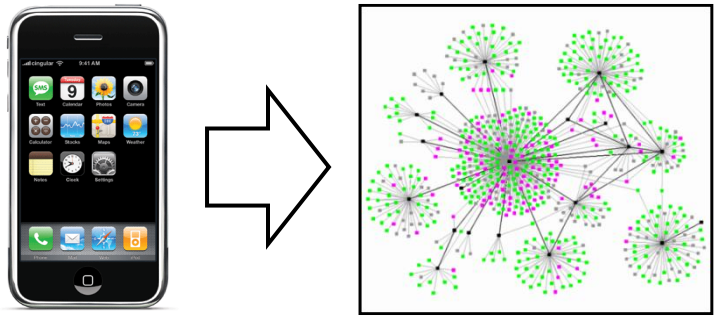
Novel data sources for event detection

The growing use of emerging technologies such as cellular phones, sensor networks, Internet search queries, and user-generated web content will open up a huge realm of possibilities for event detection.

Cell phone technologies are particularly exciting due to their rapidly increasing prevalence in the developing world and their potential to collect massive amounts of data from a huge population (location, proximity, call/SMS, built-in sensors, direct user queries…)

Automatic Contact Tracing

Use cell phone location and proximity data (possibly combined with other data) to detect disease outbreaks in their early stages, identify where and **who** is likely to be affected, and notify affected individuals (via “Reverse 911”).

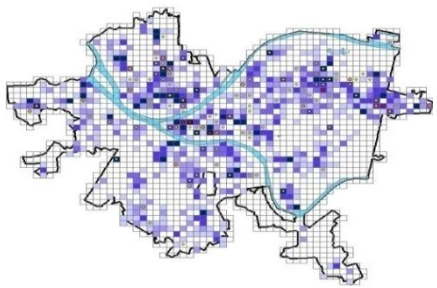


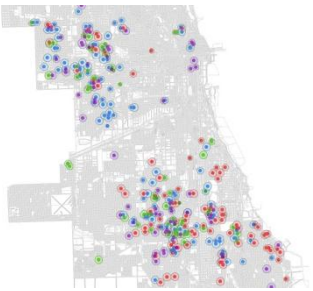
This work will use a novel dataset from Singapore, made available by the Living Analytics Research Centre (LARC).

Integrated Population Health Surveillance

Monitoring overall population health

Move beyond outbreak detection, to monitor chronic disease, injury, crime, poverty, drug abuse, patient care, etc.





Can we derive a useful aggregate indicator of population health, accurately measure it at a fine-grained spatial and temporal resolution, and predict the effects of various policy interventions?

Can we learn the (causal?) relationships and interactions between the various components of population health, use them to make fine-grained predictions, and optimize public decision-making?

Extension of our Chicago crime prediction project, in collaboration with the city's Chief Data Officer.

- Huge variety of potential data sources
- Quick translation to real-world practice
- Direct feedback from practitioners
- Possibility of controlled experiments

More topics in event and pattern detection

Currently unfunded projects, but I anticipate that you will have the opportunity to work on many of these research challenges in the next few years…

Privacy-preserving event detection (e.g. for local health departments with shared borders).

Combining **sensor placement** and **sensor fusion** (submodularity + LTSS + iterate = ???)

Developing and applying new methodologies for active learning, e.g. **competitive active learning**.

Using the **wisdom of crowds** to assist in outbreak detection, medical diagnosis, etc.

Applications to areas including disease surveillance, water quality, food safety, crime prediction, health care, fleet maintenance, network intrusion detection, fraud detection, customs monitoring, infrastructure management, scientific discovery, and many others…

Deployment of large-scale systems for health and crime surveillance will provide exciting opportunities to work with real world data, collaborate with law enforcement and health officials, and directly contribute to the public good by improving health, safety, and security.

Interested?

More details on my web page:
http://www.cs.cmu.edu/~neill

Or e-mail me at:
neill@cs.cmu.edu