Machine Learning and Event Detection for the Public Good

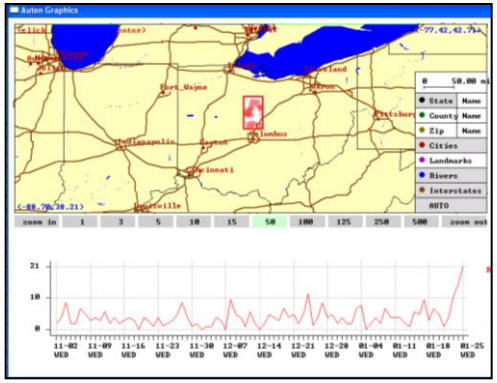
Daniel B. Neill, Ph.D.
H.J. Heinz III College
Carnegie Mellon University
E-mail: neill@cs.cmu.edu

We gratefully acknowledge funding support from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330.

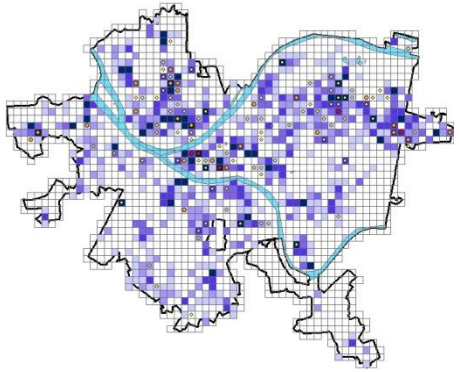
Daniel B. Neill (neill@cs.cmu.edu)
Assistant Professor of Information Systems, Heinz College
Courtesy Assistant Professor of Machine Learning and Robotics, SCS

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

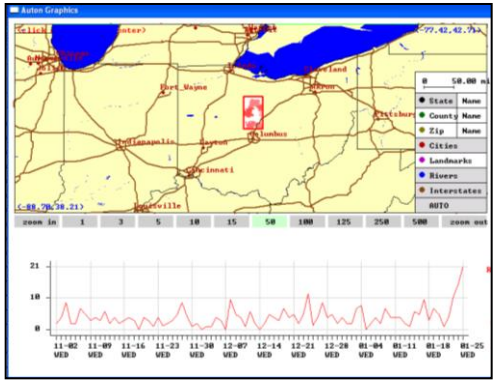
Our methods could have detected the May 2000 Walkerton *E. coli* outbreak two days earlier than the first public health response.

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.

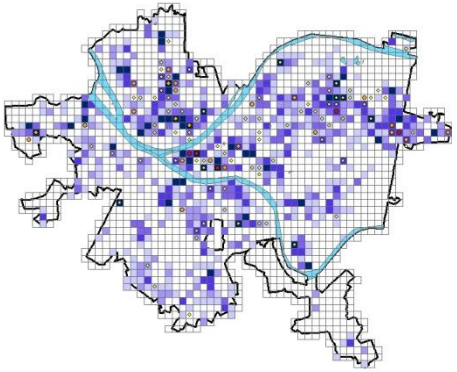
Daniel B. Neill (neill@cs.cmu.edu)
Assistant Professor of Information Systems, Heinz College
Courtesy Assistant Professor of Machine Learning and Robotics, SCS

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.

Why study machine learning?

Critical importance of addressing global policy problems: disease pandemics, crime, terrorism, poverty, environment…

Increasing size and complexity of available data, thanks to the rapid growth of new and transformative technologies.

Much more computing power, and scalable data analysis methods, enable us to extract actionable information from all of this data.

Machine learning techniques have become increasingly essential for 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)

Some definitions

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

Artificial Intelligence (AI) is the science of automating complex behaviors such as learning, problem solving, and decision making.

Data Mining (DM) is the process of extracting useful information from massive quantities of complex data.

I would argue that these are not three distinct fields of study! While each has a slightly different emphasis, there is a tremendous amount of overlap in the problems they are trying to solve and the techniques used to solve them.

Many of the techniques we will learn are **statistical** in nature, but are very different from classical statistics.

ML/AI/DM systems and methods:

Scale up to large, complex data
Learn and **improve** from experience
Perceive and **change** the environment
Interact with humans or other agents
Explain inferences and decisions
Discover new and useful patterns

How is ML relevant for policy?

ML provides a powerful set of **tools** for intelligent problem-solving.

Building sophisticated models that combine data and prior knowledge to enable intelligent decisions.

Automating tasks such as prediction and detection to reduce human effort.

Scaling up to large, complex problems by focusing user attention on relevant aspects.

Using ML to analyze data and guide policy decisions.

Proposing policy initiatives to reduce the amount and impact of violent crime in urban areas.

Predicting the adoption rate of new technology in developing countries.

Analyzing which factors influence congressional votes or court decisions

Using ML in information systems to improve public services

Health care: diagnosis, drug prescribing
Law enforcement: crime forecasting
Public health: epidemic detection/response
Urban planning: optimizing facility location
Homeland security: detecting terrorism

Analyzing impacts of ML technology adoption on society

Internet search and e-commerce
Data mining (security vs. privacy)
Automated drug discovery
Industrial and companion robots
Ethical and legal issues

Advertisement: MLP@CMU

We are working to build a comprehensive curriculum in **machine learning and policy (MLP)** here at CMU.

Goals of the MLP initiative: increase collaboration between ML and PP researchers, train new researchers with deep knowledge of both areas, and encourage a widely shared focus on using ML to benefit the public good.

Here are some of the many ways you can get involved:

Joint Ph.D. Program in Machine Learning and Public Policy (MLD & Heinz)
Ph.D. in Information Systems + M.S. in Machine Learning

Large Scale Data Analysis for Policy: introduction to ML for PPM students.
Research Seminar in Machine Learning & Policy: for ML/Heinz Ph.D. students.
Special Topics in Machine Learning and Policy: Event and Pattern Detection, ML for Developing World, Harnessing the Wisdom of Crowds

Workshop on Machine Learning and Policy Research & Education
Research Labs: Event and Pattern Detection Lab, Auton Laboratory, iLab
Center for Science and Technology in Human Rights, many others…

LSDA course description

This course will focus on applying large scale data analysis methods from the closely related fields of **machine learning**, **data mining**, and **artificial intelligence** to develop tools for intelligent problem solving in real-world policy applications.

We will emphasize tools that can “scale up” to real-world problems with huge amounts of high-dimensional and multivariate data.

Mountain of policy data
Huge, unstructured, hard to interpret or use for decisions

1. Translate policy questions into ML paradigms.
2. Choose and apply appropriate methods.
3. Interpret, evaluate, and use results.

Actionable knowledge of policy domain

Predict & **explain** unknown values
Model structures, relations
Detect relevant patterns

Use for decision-making, policy prescriptions, improved services

LSDA course syllabus

- Introduction to Large Scale Data Analysis
  - Incorporates methods from **machine learning**, data mining, artificial intelligence.
  - Goals, problem paradigms, and software tools (e.g. Weka)
- Module I (**Prediction**)
  - Classification and regression (making, explaining predictions)
  - Rule-based, case-based, and model-based learning.
- Module II (**Modeling**)
  - Representation and heuristic search
  - Clustering (modeling group structure of data)
  - Bayesian networks (modeling probabilistic relationships)
- Module III (**Detection**)
  - Anomaly Detection (detecting outliers, novelties, etc.)
  - Pattern Detection (e.g. event surveillance, anomalous patterns)
  - Applications to biosurveillance, crime prevention, etc.
  - Guest “mini-lectures” from the Event and Pattern Detection Lab.

Common ML paradigms: prediction

In **prediction**, we are interested in explaining a specific attribute of the data in terms of the other attributes.

Classification: predict a discrete value

"What disease does this patient have, given his symptoms?"

Regression: estimate a numeric value

"How is a country's literacy rate affected by various social programs?"

Two main goals of prediction

Guessing unknown values for specific instances (e.g. diagnosing a given patient)

Explaining predictions of both known and unknown instances (providing relevant examples, a set of decision rules, or class-specific models).

Example 1: What socio-economic factors lead to increased prevalence of diarrheal illness in a developing country?

Example 2: Developing a system to diagnose a patient's risk of diabetes and related complications, for improved medical decision-making.

Common ML paradigms: modeling

In **modeling**, we are interested in describing the underlying relationships between many attributes and many entities.

Our goal is to produce models of the "entire data" (not just specific attributes or examples) that accurately reflect underlying complexity, yet are simple, understandable by humans, and usable for decision-making.

Relations between entities	Relations between variables
Identifying link, group, and network structures	Identifying significant positive and negative correlations
Partitioning or "clustering" data into subgroups	Visualizing dependence structure between multiple variables

Example 1: Can we visualize the dependencies between various diet-related risk factors and health outcomes?

Example 2: Can we better explain consumer purchasing behavior by identifying subgroups and incorporating social network ties?

Common ML paradigms: detection

In **detection**, we are interested in identifying relevant patterns in massive, complex datasets.

Main goal: focus the user's attention on a potentially relevant subset of the data.

a) Automatically detect relevant individual records, or groups of records.

b) Characterize and explain the pattern (type of pattern, H_0 and H_1 models, etc.)

c) Present the pattern to the user.

Some common detection tasks

Detecting **anomalous** records or groups

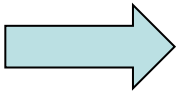
Discovering **novelties** (e.g. new drugs)

Detecting **clusters** in space or time

Removing **noise** or **errors** in data

Detecting **specific patterns** (e.g. fraud)

Detecting emerging **events** which may require rapid responses.



Example 1: Detect emerging outbreaks of disease using electronic public health data from hospitals and pharmacies.

Example 2: How common are patterns of fraudulent behavior on various e-commerce sites, and how can we deal with online fraud?

What is disease surveillance?

- The systematic collection and analysis of data for the purpose of detecting outbreaks of disease in people, plants, or animals.
  - Primary goal: timely and accurate detection and characterization of an outbreak.
    - Is there an outbreak?
    - If so, what type of outbreak, and where/who is affected?
    - End goal: enable public health to make rapid and informed decisions to prevent and control outbreaks.

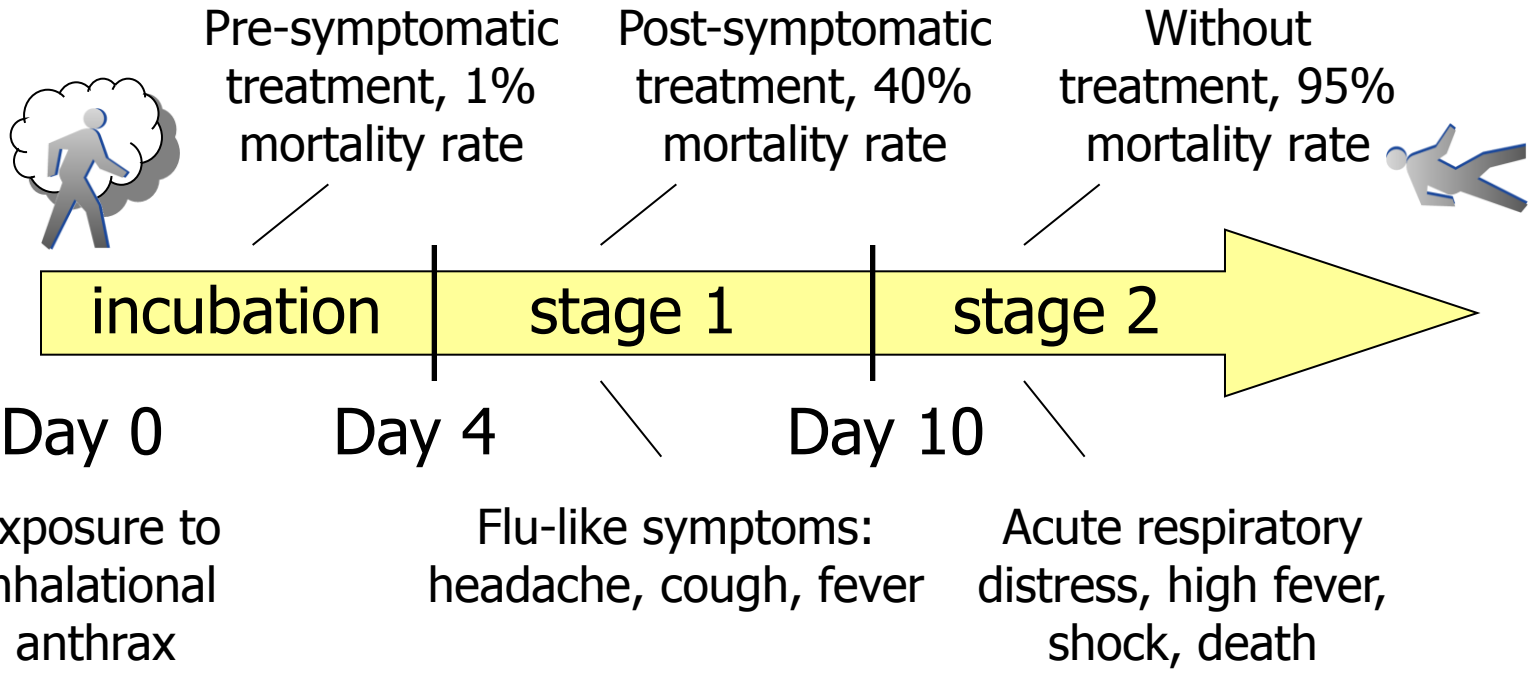
Treatment	Vaccination	Health advisories
Cleanup	Quarantines	Travel restrictions

Why worry about disease outbreaks?

- Bioterrorist attacks are a very real, and scary, possibility
  - 100 kg anthrax, released over D.C., could kill 1-3 million and hospitalize millions more.
- Emerging infectious diseases
  - “Conservative estimate” of 2-7 million deaths from pandemic avian influenza.
- Better response to common outbreaks (seasonal flu, GI)

Benefits of early detection

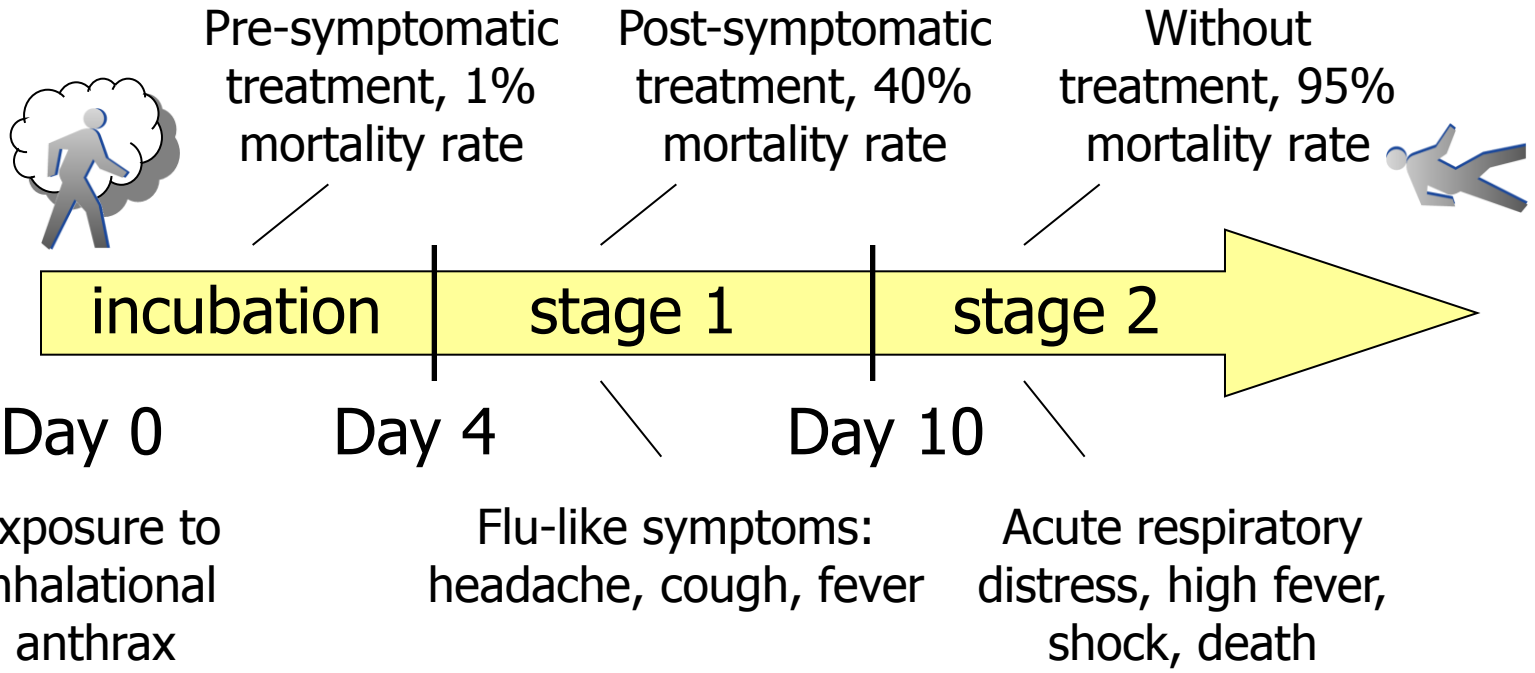
Reduces cost to society, both in lives and in dollars!



DARPA estimate: a two-day gain in detection time and public health response could reduce fatalities by a factor of six.

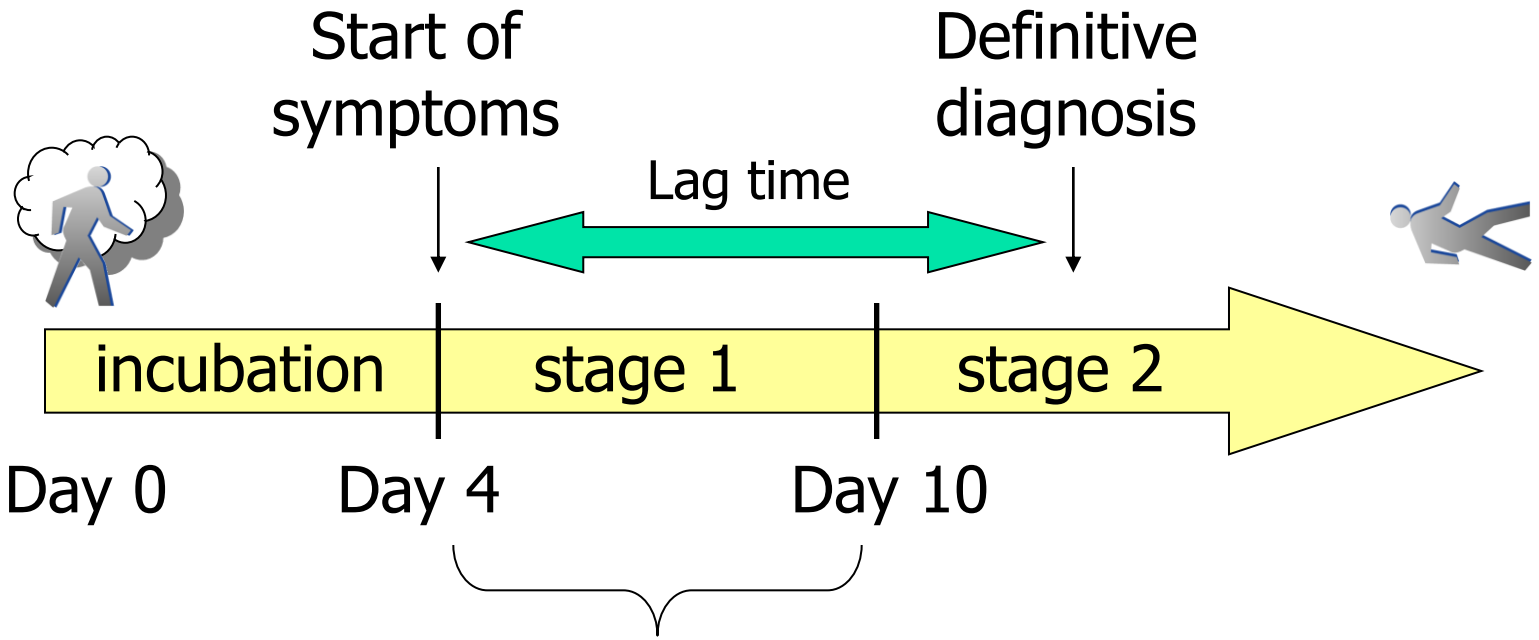
Benefits of early detection

Reduces cost to society, both in lives and in dollars!



"Improvements of even an hour over current detection capabilities could reduce economic impact of a bioterrorist anthrax attack by hundreds of millions of dollars."

Early detection is hard



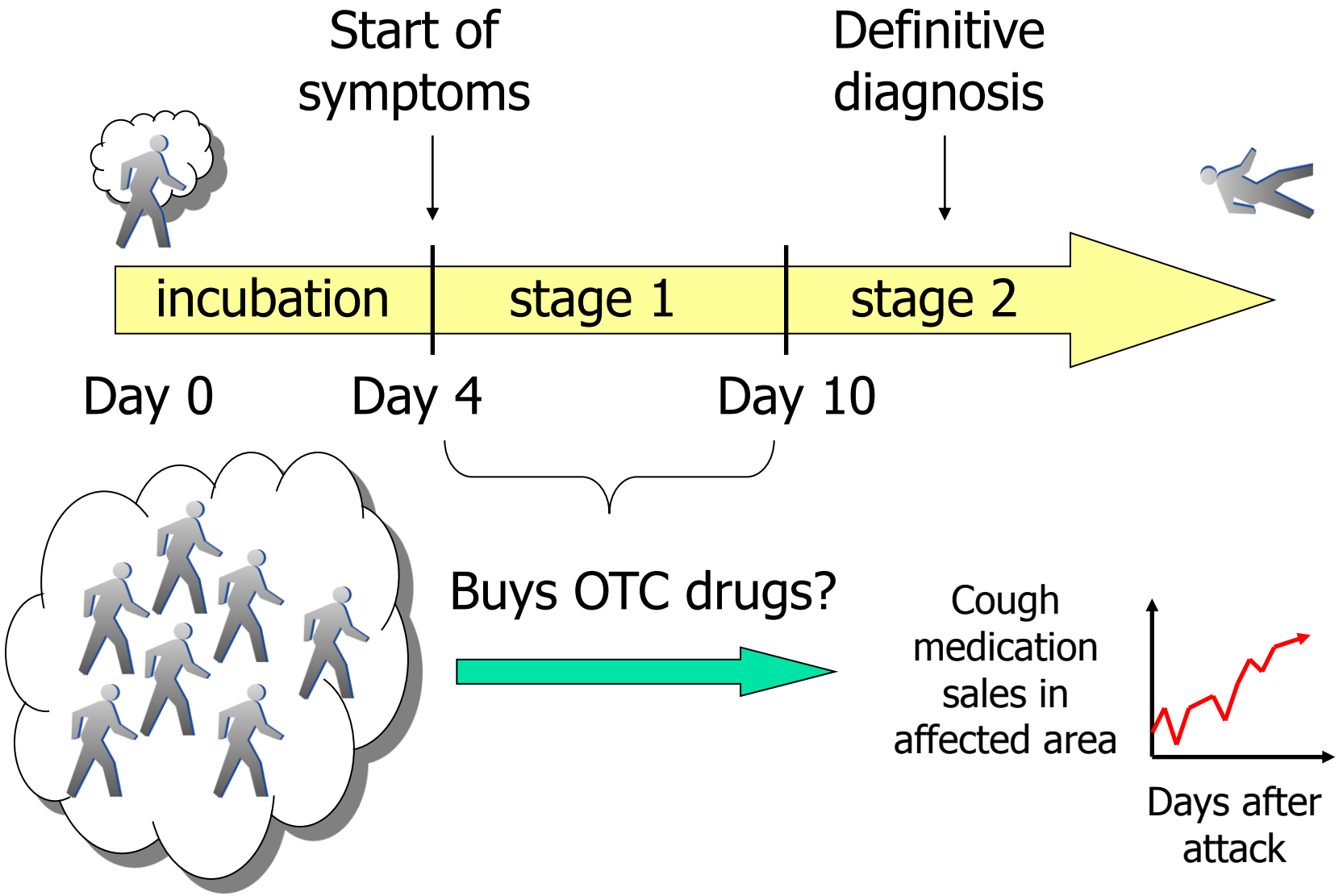
Buys OTC drugs

Skips work/school

Uses Google, Facebook, Twitter

Visits doctor/hospital/ED

Syndromic surveillance



Syndromic surveillance

Start of symptoms

Definitive diagnosis

We can achieve very early detection of outbreaks by gathering syndromic data, and identifying emerging spatial clusters of symptoms.

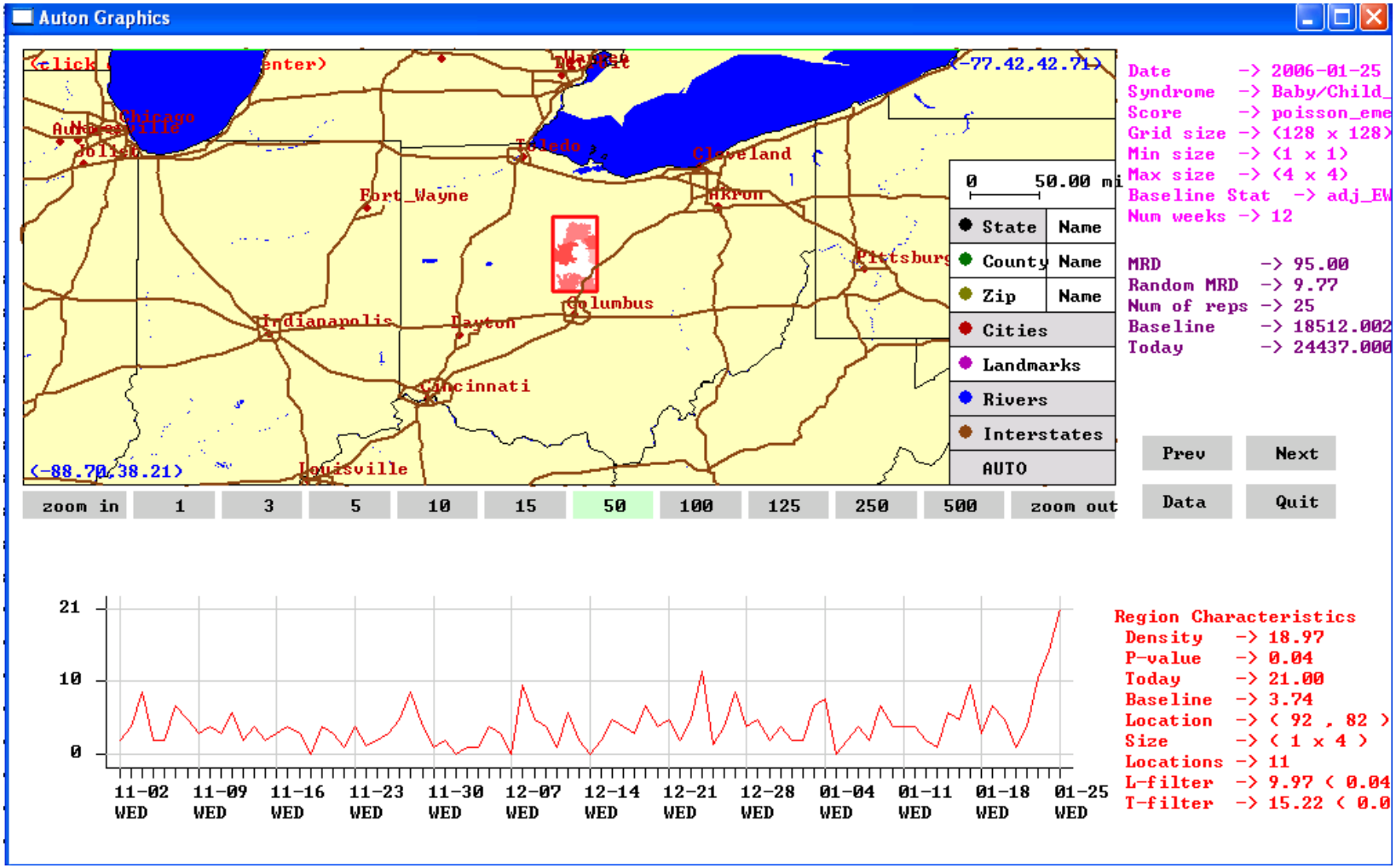
Buys OTC drugs?

Cough medication sales in affected area

Days after attack

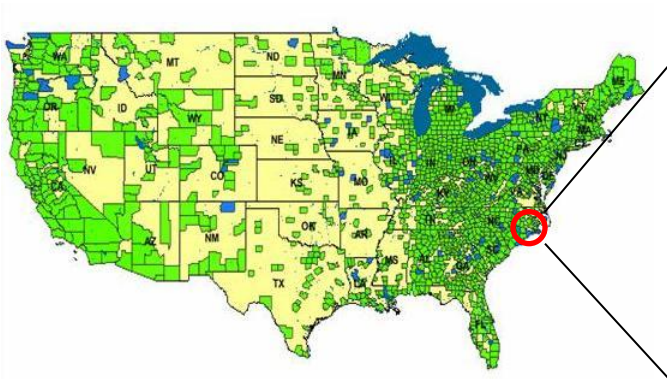
A recent potential outbreak

Spike in sales of pediatric electrolytes near Columbus, Ohio

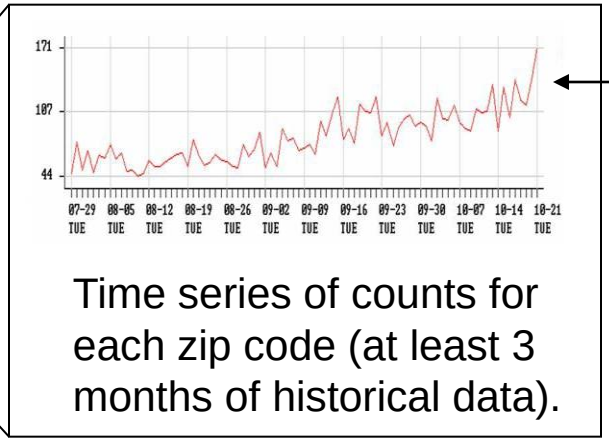


Under the hood: how does it work?

Finding emerging spatial clusters in a health data stream.



Daily over-the-counter sales of cough/cold medication, for each of over 20,000 zip codes nationwide.



Time series of counts for each zip code (at least 3 months of historical data).

This increase could be due to an outbreak, or due to chance.

Which increases are <u>significant</u>?

<u>Our solution</u>

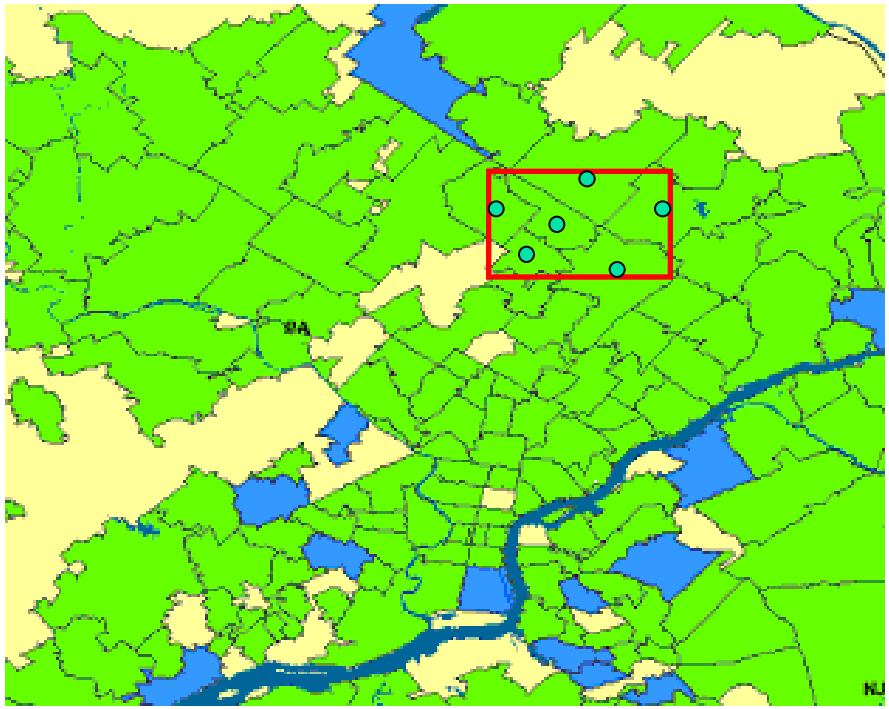
1. Infer the expected count for each zip code for each recent day.
2. Find regions where the recent counts are higher than expected.

We want to be able to detect outbreaks whether they affect a small or large region, and whether they emerge quickly or gradually.

<u>Solution</u>: the space-time scan statistic.

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

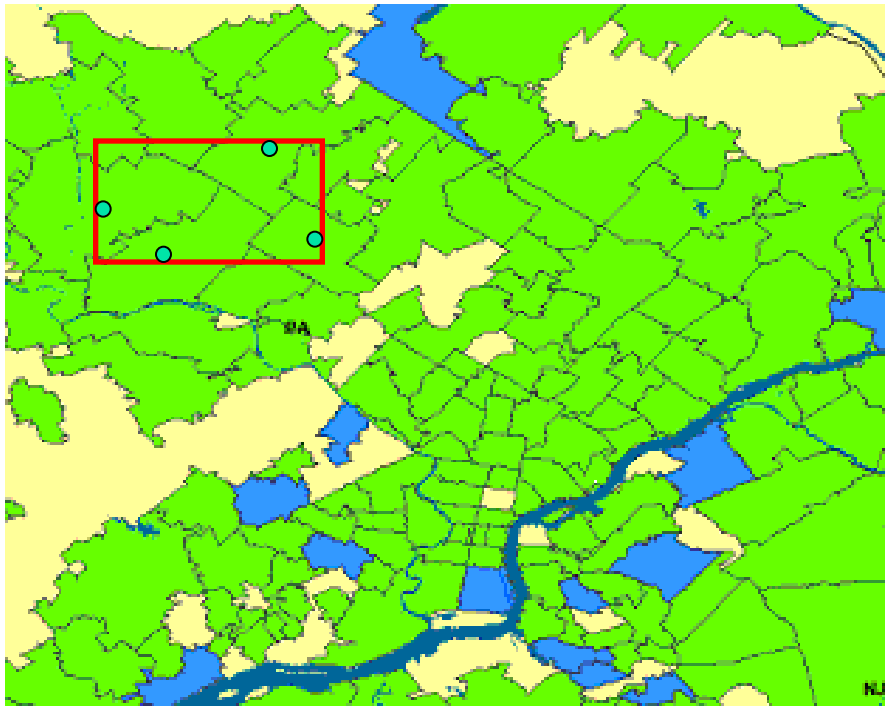


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

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

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

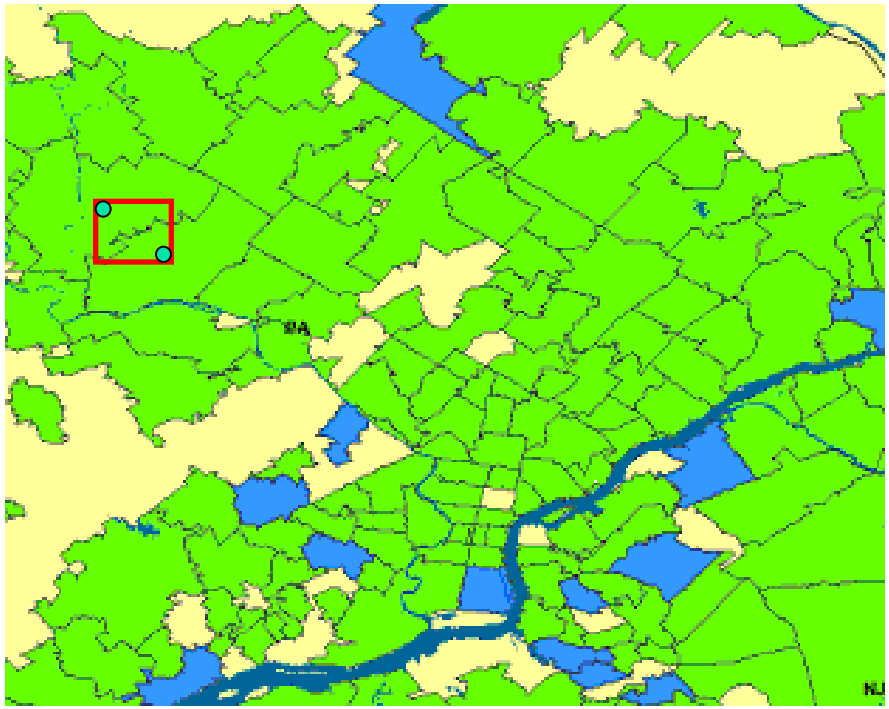


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

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

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

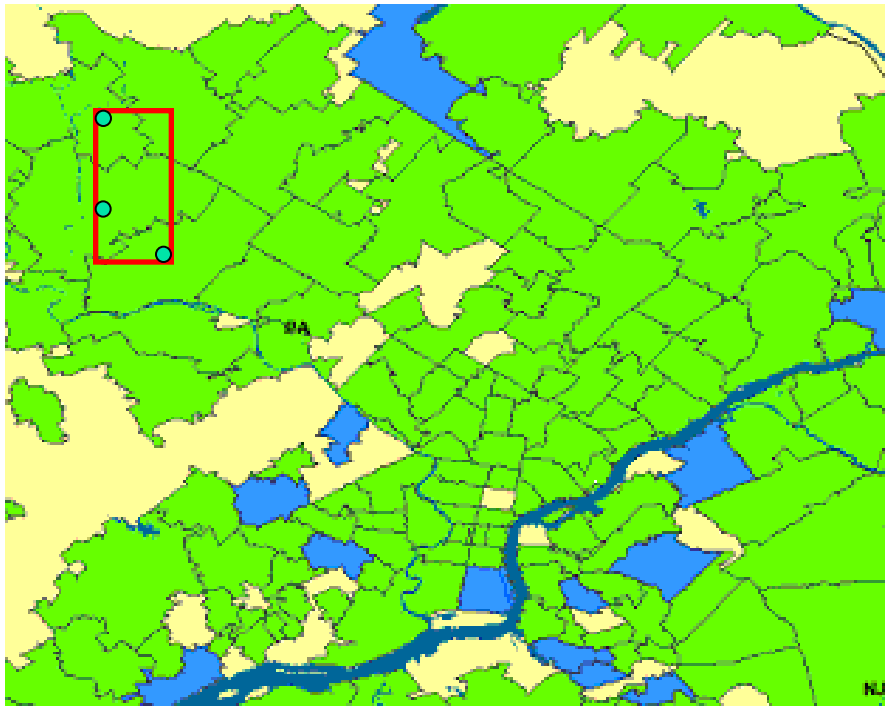


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

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

The space-time scan statistic

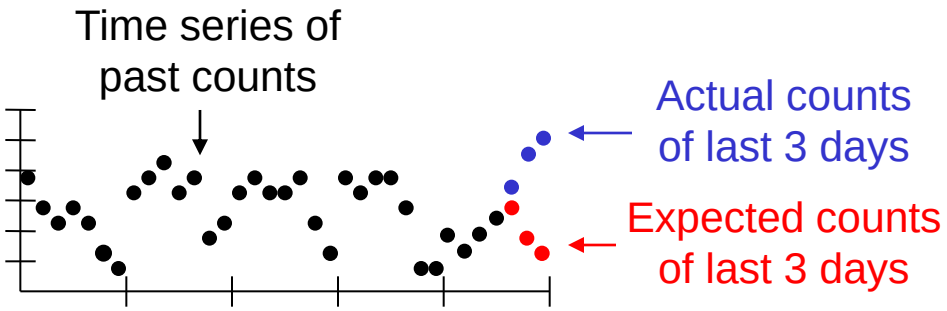
(Kulldorff, 2001; Neill & Moore, 2005)



To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

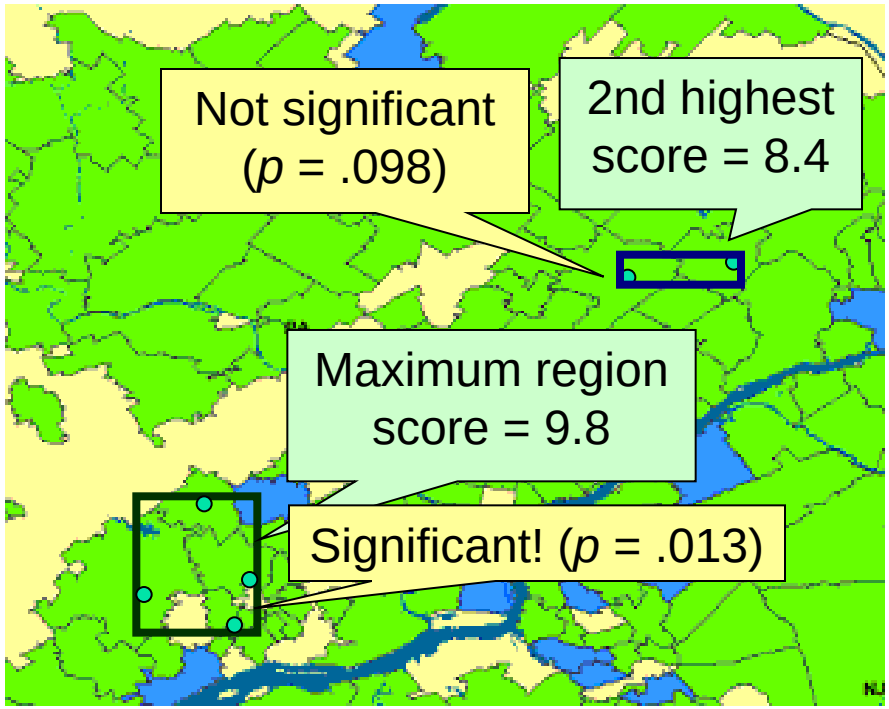
Imagine moving a window around the scan area, allowing the window size, shape, and temporal duration to vary.

For each of these regions, 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.

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

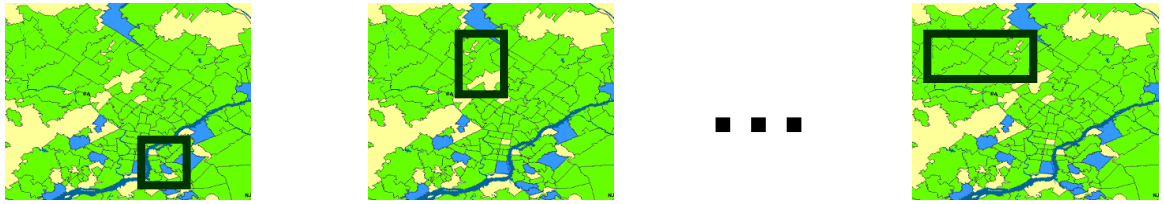
Alternative hypothesis: outbreak in region S

Null hypothesis: no outbreak

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

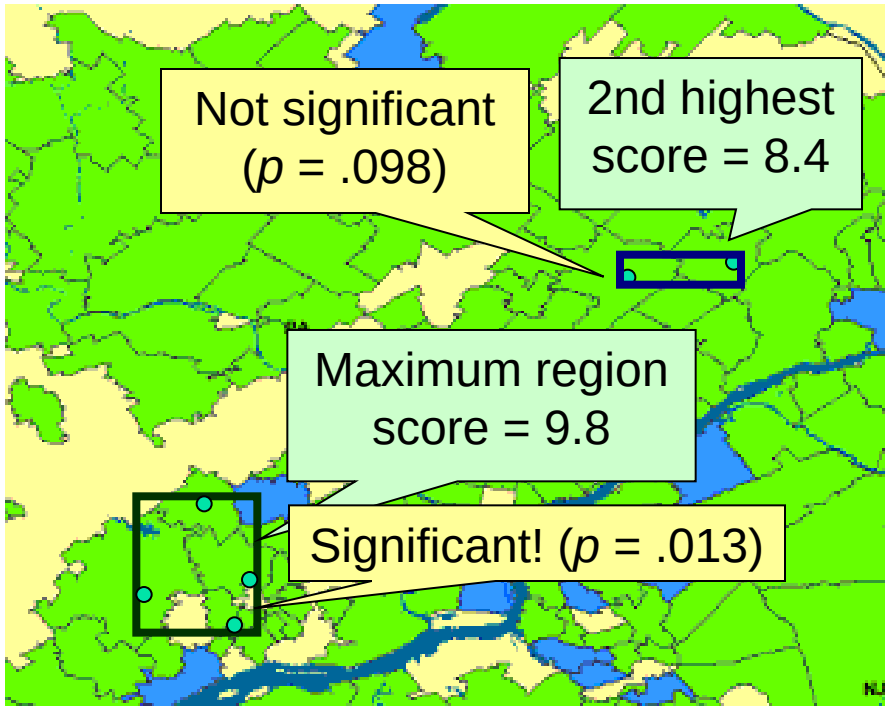
Answer: compare to the maximum region scores of simulated datasets under H_0.

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



The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)



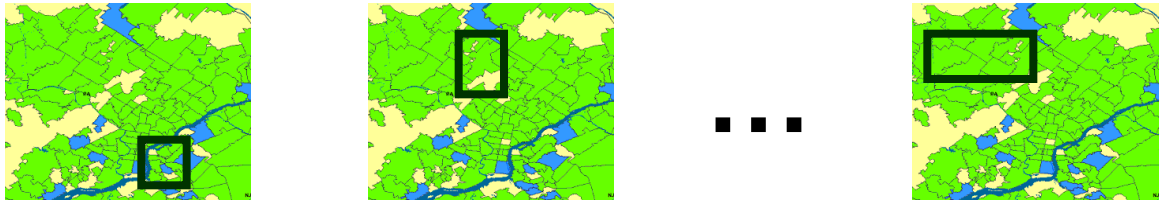
Recent advances in analytical methods for event detection enable us to:

- Integrate information from multiple streams
- Distinguish between multiple event types
- Scale up to many locations and streams
- Search over irregularly-shaped clusters
- Consider graph and non-spatial constraints

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.

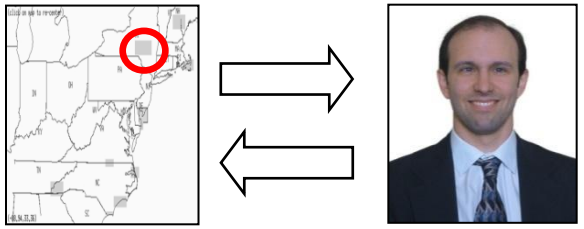
$F_1^* = 2.4$ $F_2^* = 9.1$ ... $F_{999}^* = 7.0$



A sampling of current projects…

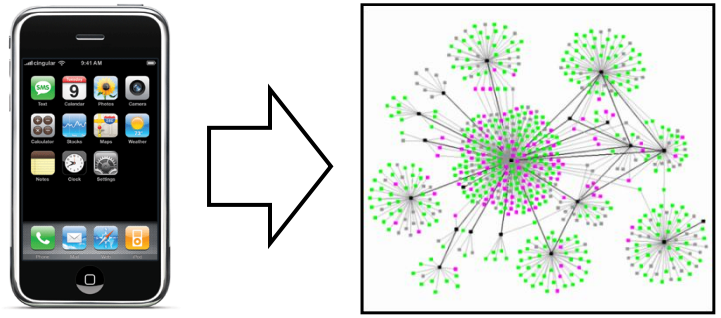
Integrating Learning and Detection

Incorporate user feedback, distinguish relevant from irrelevant anomalies



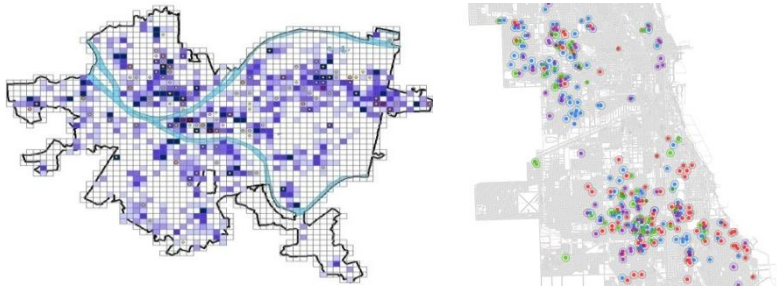
Automatic Contact Tracing

Use cell phone location and proximity data to detect outbreaks and identify where and **who** is affected.



Population Health Surveillance

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



Interested?

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

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