Machine Learning for Population Health and Disease Surveillance

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

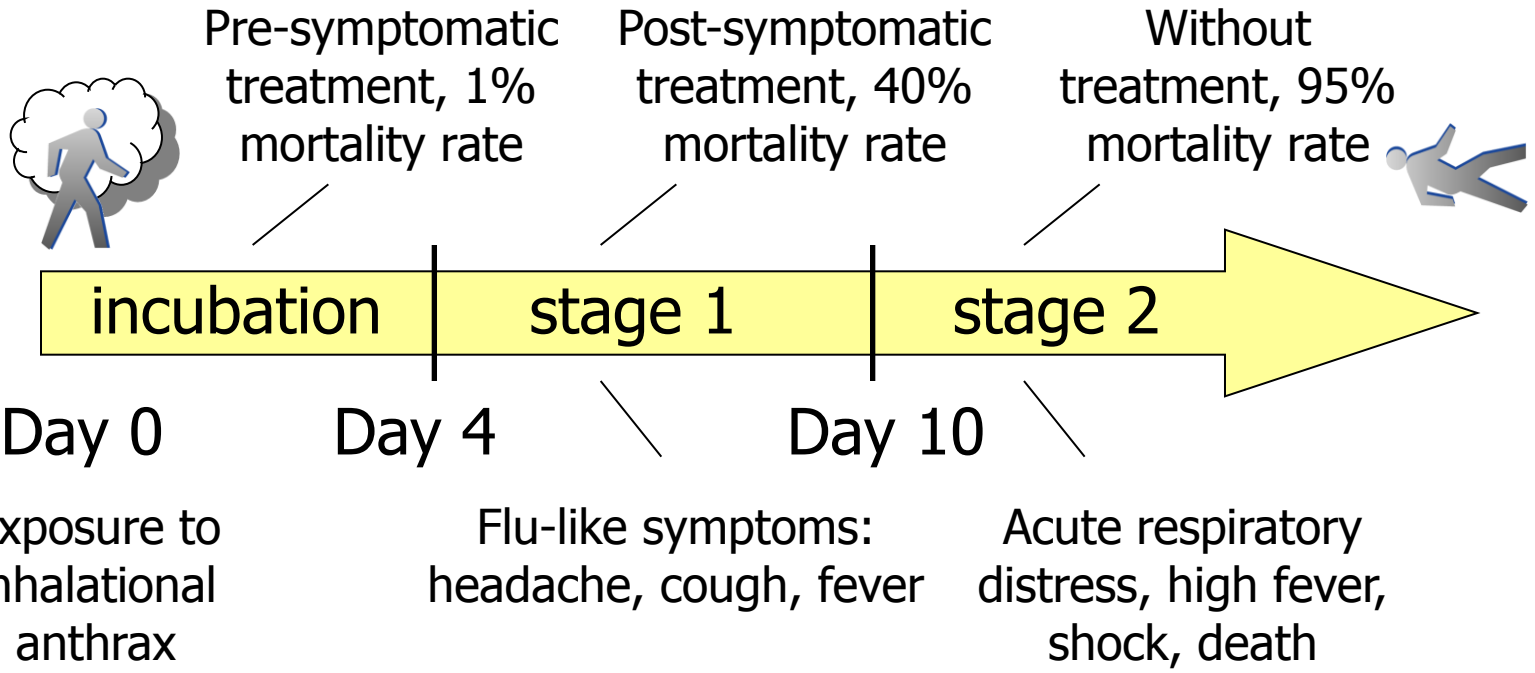
We gratefully acknowledge funding support from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330.

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

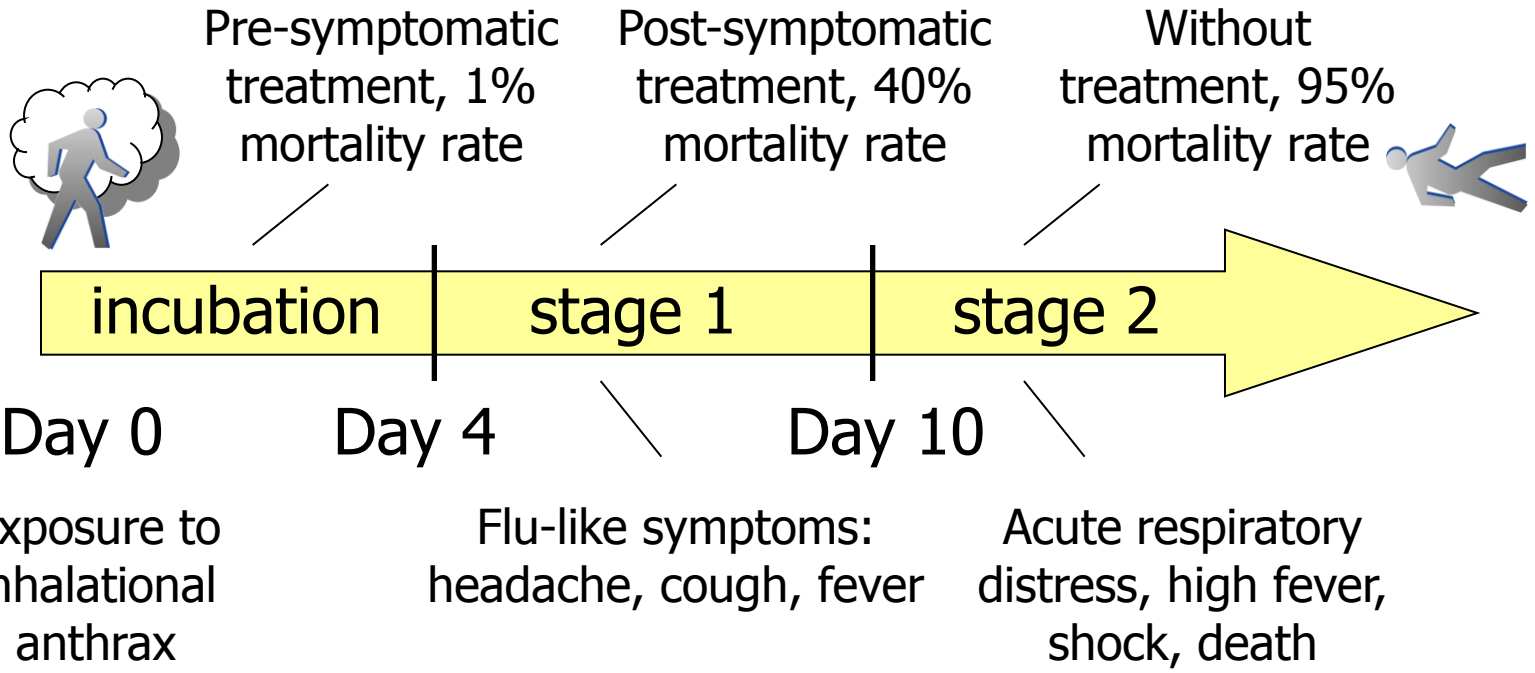
Reduces cost to society, both in lives and in dollars!



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

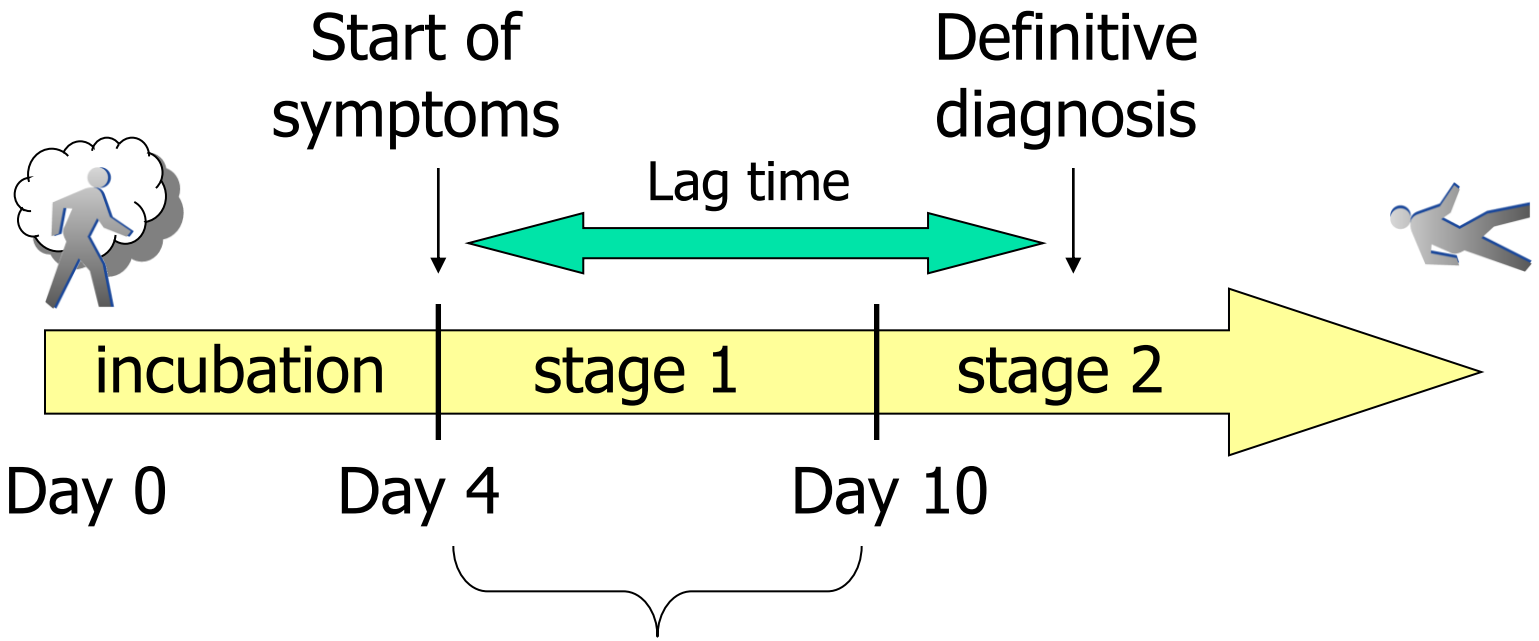
Benefits of early detection

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



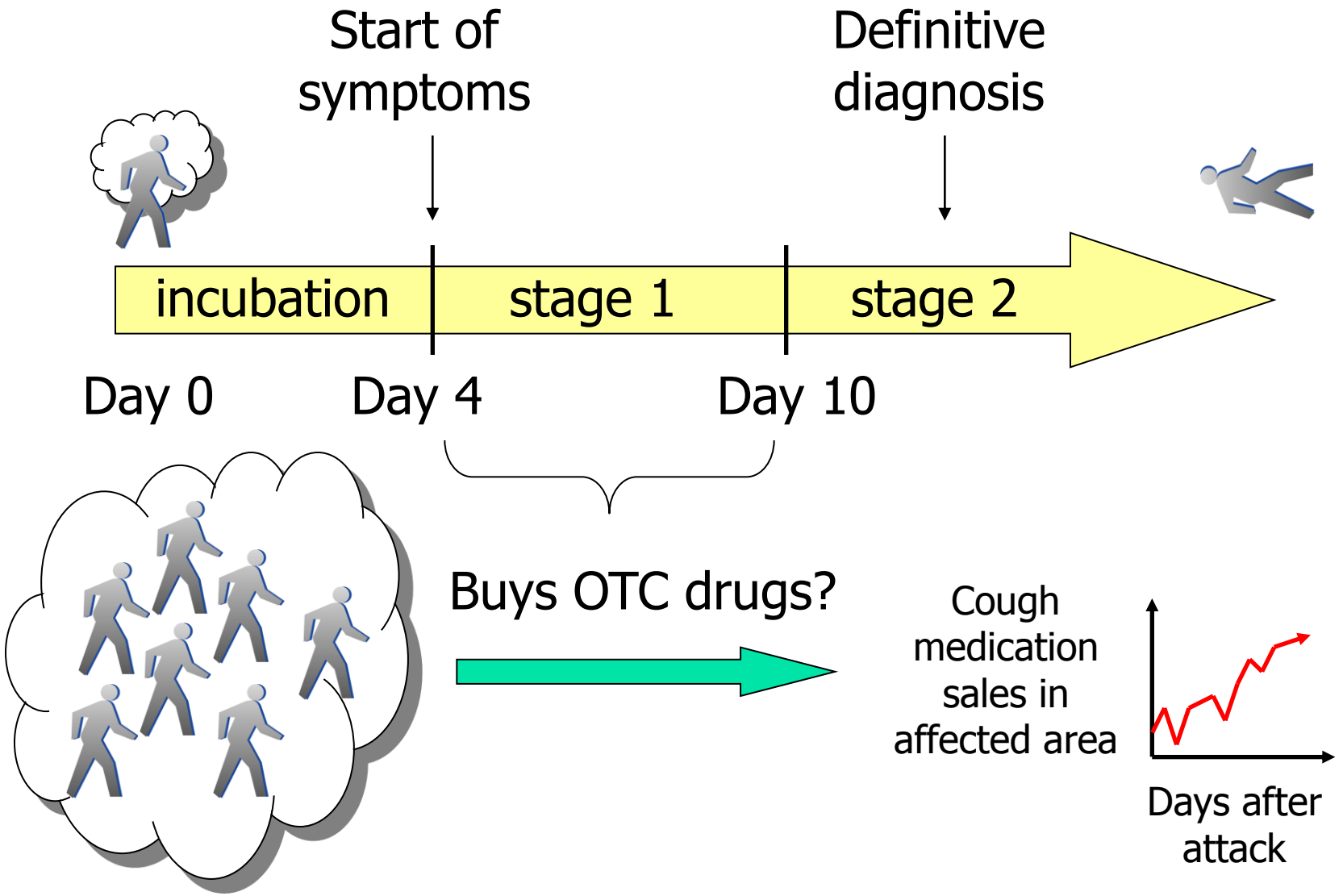
Buys OTC drugs

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.

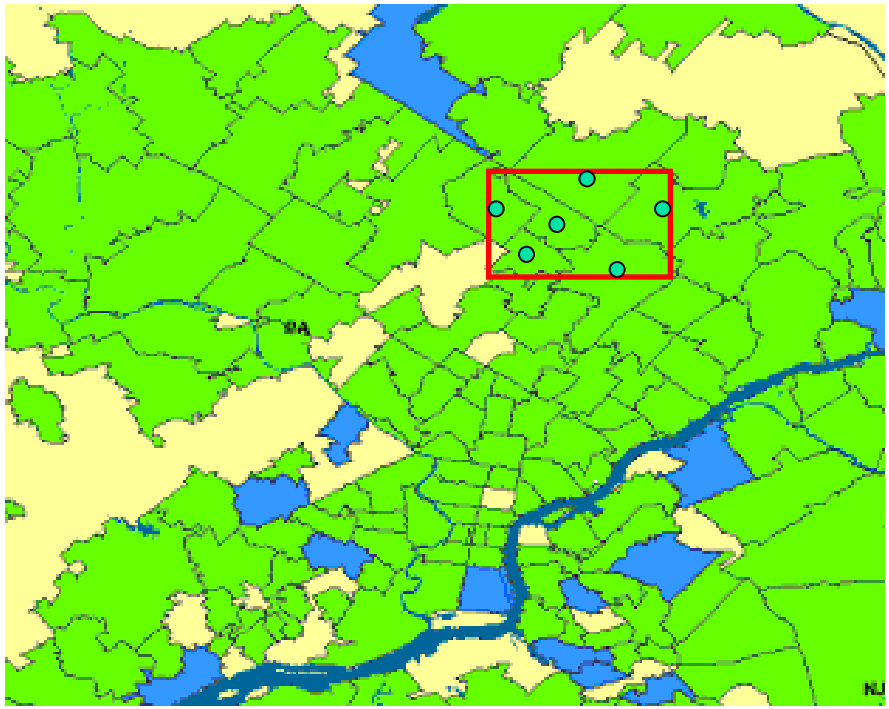
Buys OTC drugs?

Cough medication sales in affected area

Days after attack

The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)

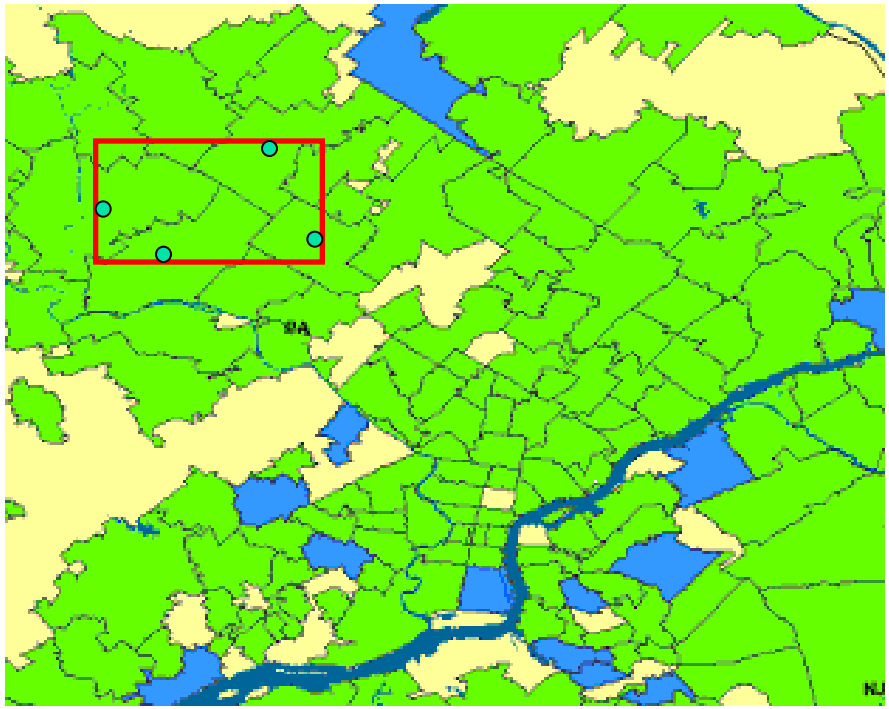


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

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)

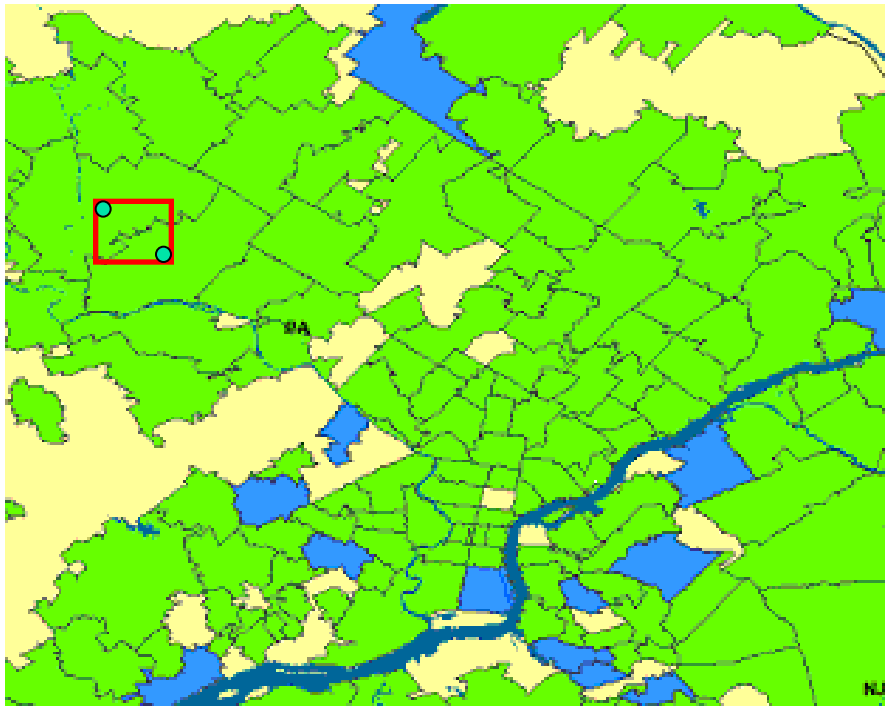


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

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)

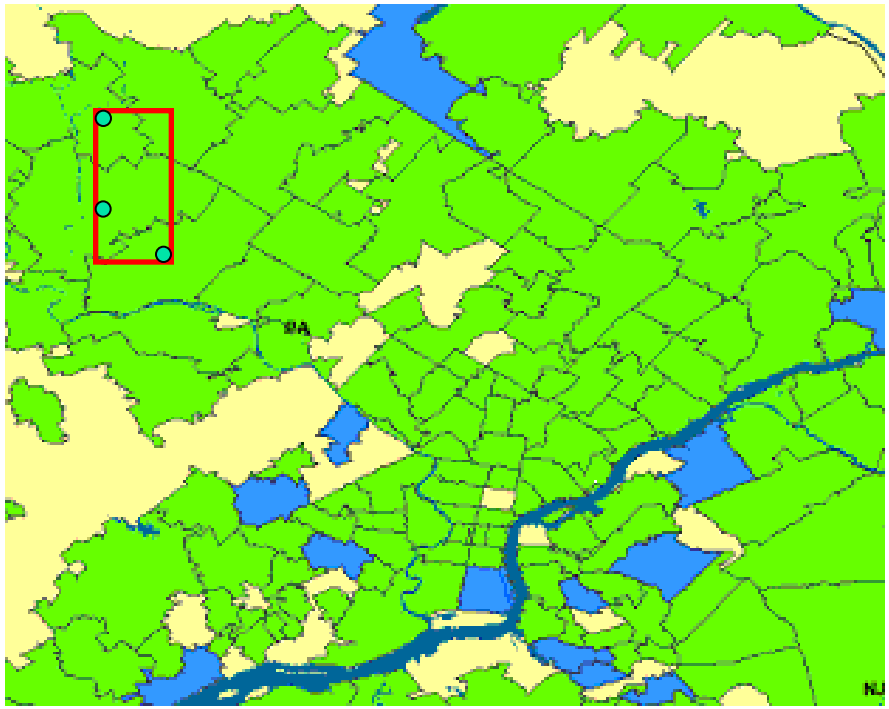


To detect and localize events, we can search for space-time regions where the number of cases is higher than expected.

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

The space-time scan statistic

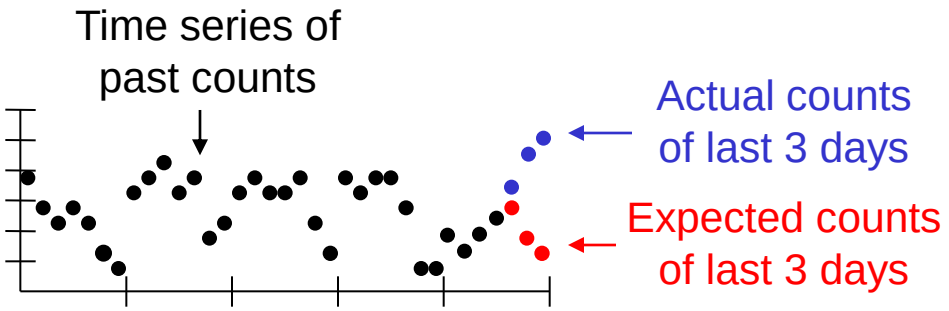
(Kulldorff, 2001; Neill & Moore, 2005)



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

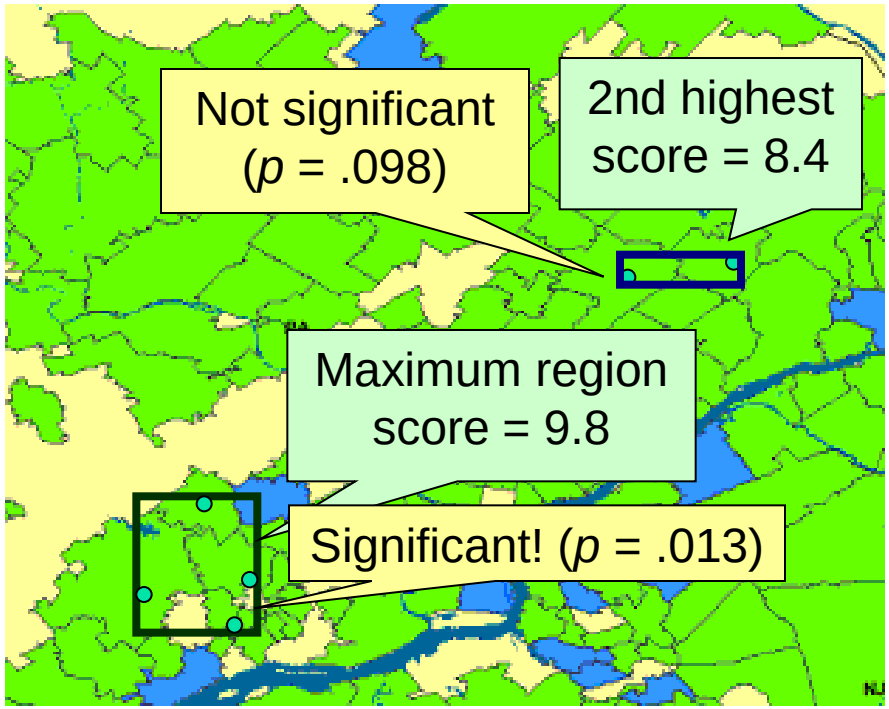
Imagine moving a window around the scan area, allowing the window size, shape, and temporal duration to vary.

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.

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

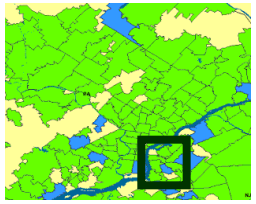
Alternative hypothesis: outbreak in region S

Null hypothesis: no outbreak

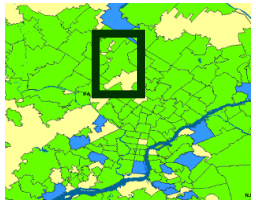
These are the most likely clusters… but how can we tell whether they are significant?

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

$F_1^* = 2.4$

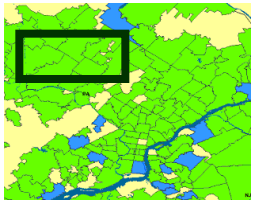


$F_2^* = 9.1$



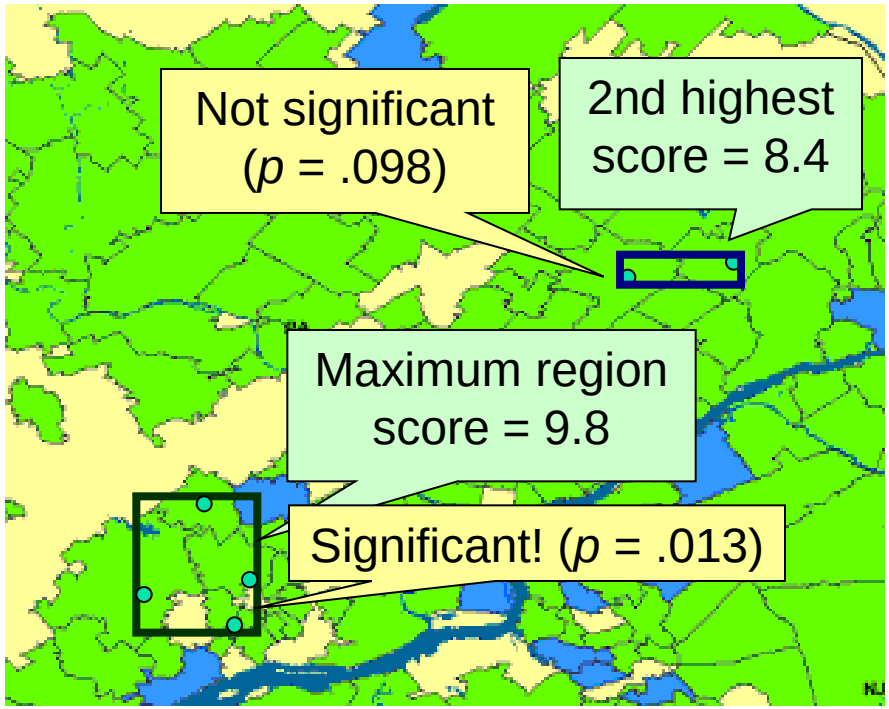
…

$F_{999}^* = 7.0$



The space-time scan statistic

(Kulldorff, 2001; Neill & Moore, 2005)



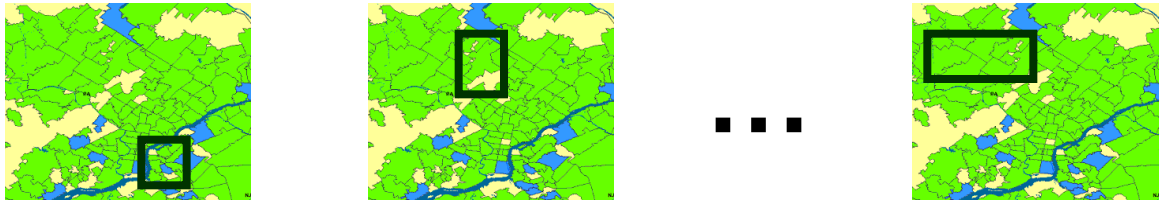
Recent advances in analytical methods for event detection enable us to:

- 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.

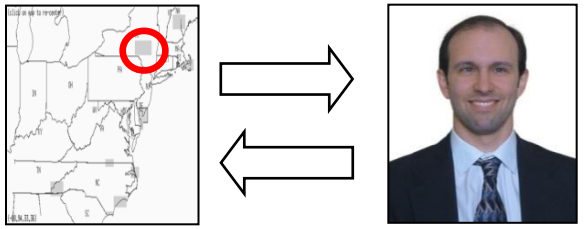
$F_1^* = 2.4$ $F_2^* = 9.1$ … $F_{999}^* = 7.0$



Current Projects

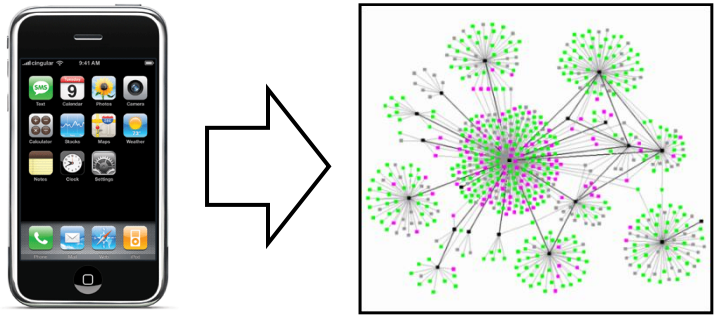
Integrating Learning and Detection

Incorporate user feedback, distinguish relevant from irrelevant anomalies



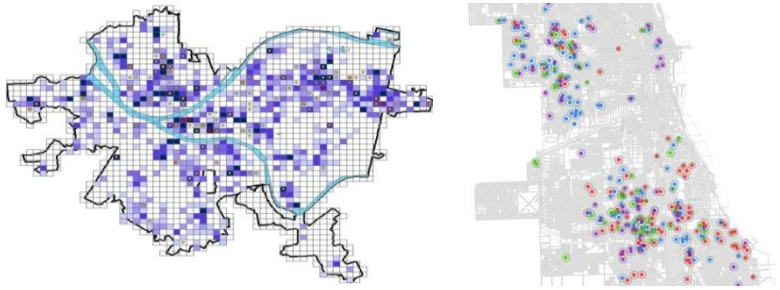
Automatic Contact Tracing

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