Machine Learning and Event Detection for the Public Good

Daniel B. Neill
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)
Associate Professor of Information Systems
Director, Event and Pattern Detection Laboratory
Courtesy Associate Professor of Machine Learning and Robotics

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…)

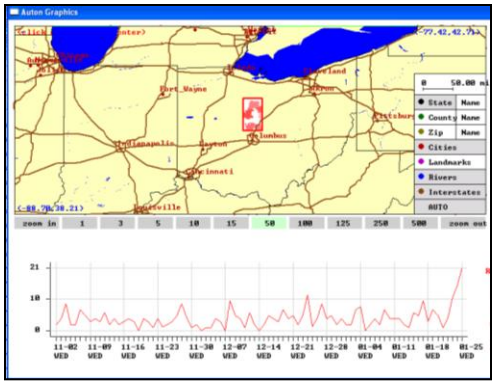
Continuously increasing size and complexity of policy data, and rapid growth of new and transformative technologies.

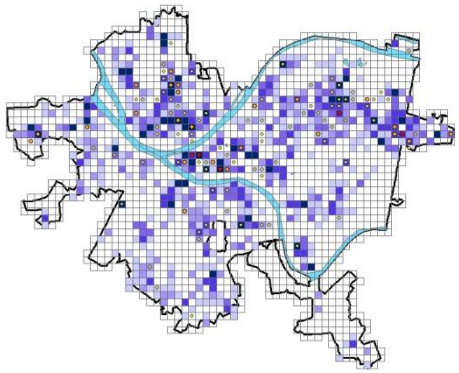
Machine learning has become increasingly essential for data-driven policy analysis and for the development of new, practical information technologies that can be directly applied **for the public good** (e.g. public health, safety, and security)

My research in this area has two main goals:

1) Develop new machine learning methods for better (more scalable and accurate) **detection** and **prediction** of events and other patterns in massive datasets.
2) Apply these methods to improve the quality of public health, safety, and security.

Daniel B. Neill (neill@cs.cmu.edu)
Associate Professor of Information Systems
Director, Event and Pattern Detection Laboratory
Courtesy Associate Professor of Machine Learning and Robotics





Disease Surveillance: Very early and accurate detection of emerging outbreaks.

Law Enforcement: Detection, prediction, and prevention of "hot-spots" of violent crime.

Medicine: Discovering new "best practices" of patient care, to improve outcomes and reduce costs.

Our disease surveillance methods are currently in use for deployed systems in the U.S., Canada, India, and Sri Lanka.

Our "CrimeScan" software has been in day-to-day operational use for predictive policing by the Chicago PD.

"CityScan" is being tested on 311 calls for anticipating citizen needs.

Advertisement: MLP@CMU

We have built 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
Ph.D. in Information Systems + M.S. in Machine Learning
M.S. in Public Policy and Management- Policy Analytics track

Large Scale Data Analysis for Policy; MLP Research Seminar; Special Topics in MLP: Event and Pattern Detection; ML for the Developing World; Harnessing the Wisdom of Crowds; Mining Massive Datasets…

Event and Pattern Detection Laboratory; Auton Laboratory; Heinz iLab; Center for Human Rights Science; Living Analytics Research Center…

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?

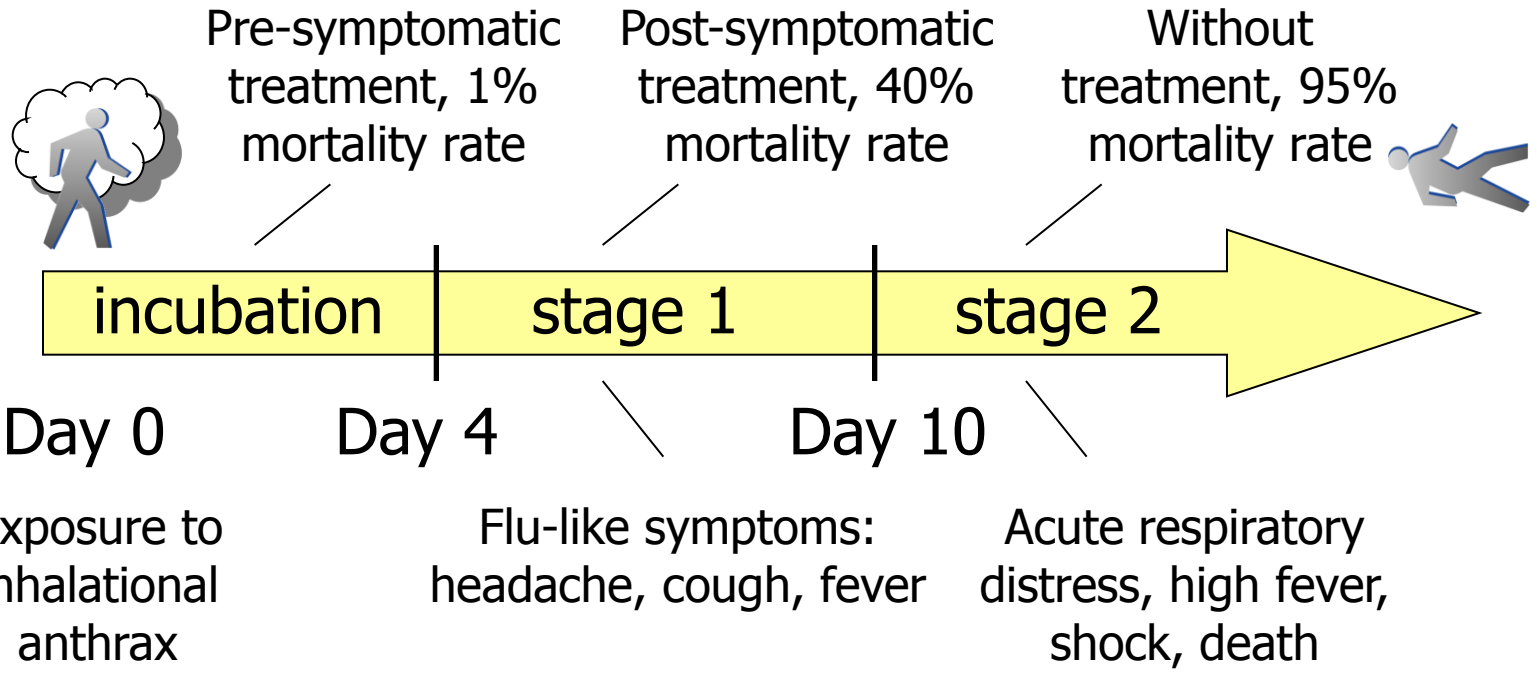
New statistical methods enable more timely and more accurate detection by integrating **multiple data sources**, incorporating **spatial** and **temporal** information, and using **prior knowledge** of a domain.

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.

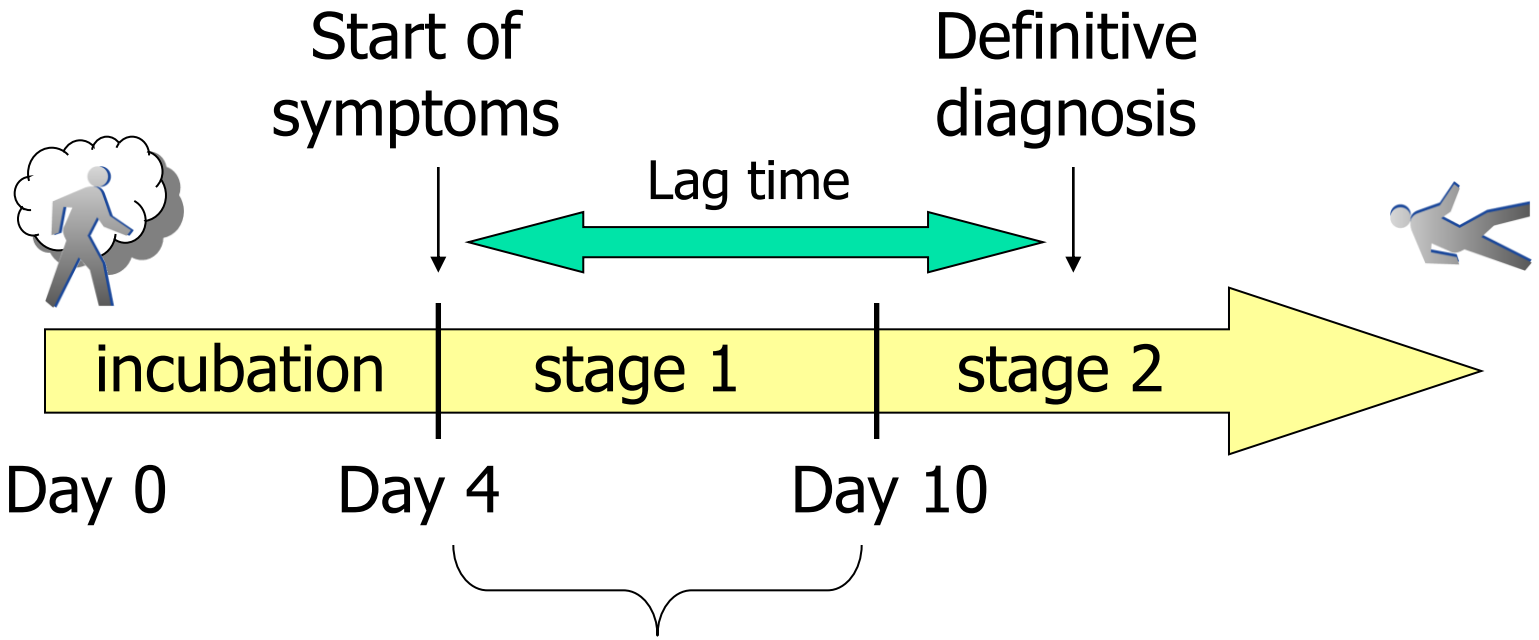
Case study 1: disease surveillance

Early detection reduces cost to society, 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.

Early detection is hard



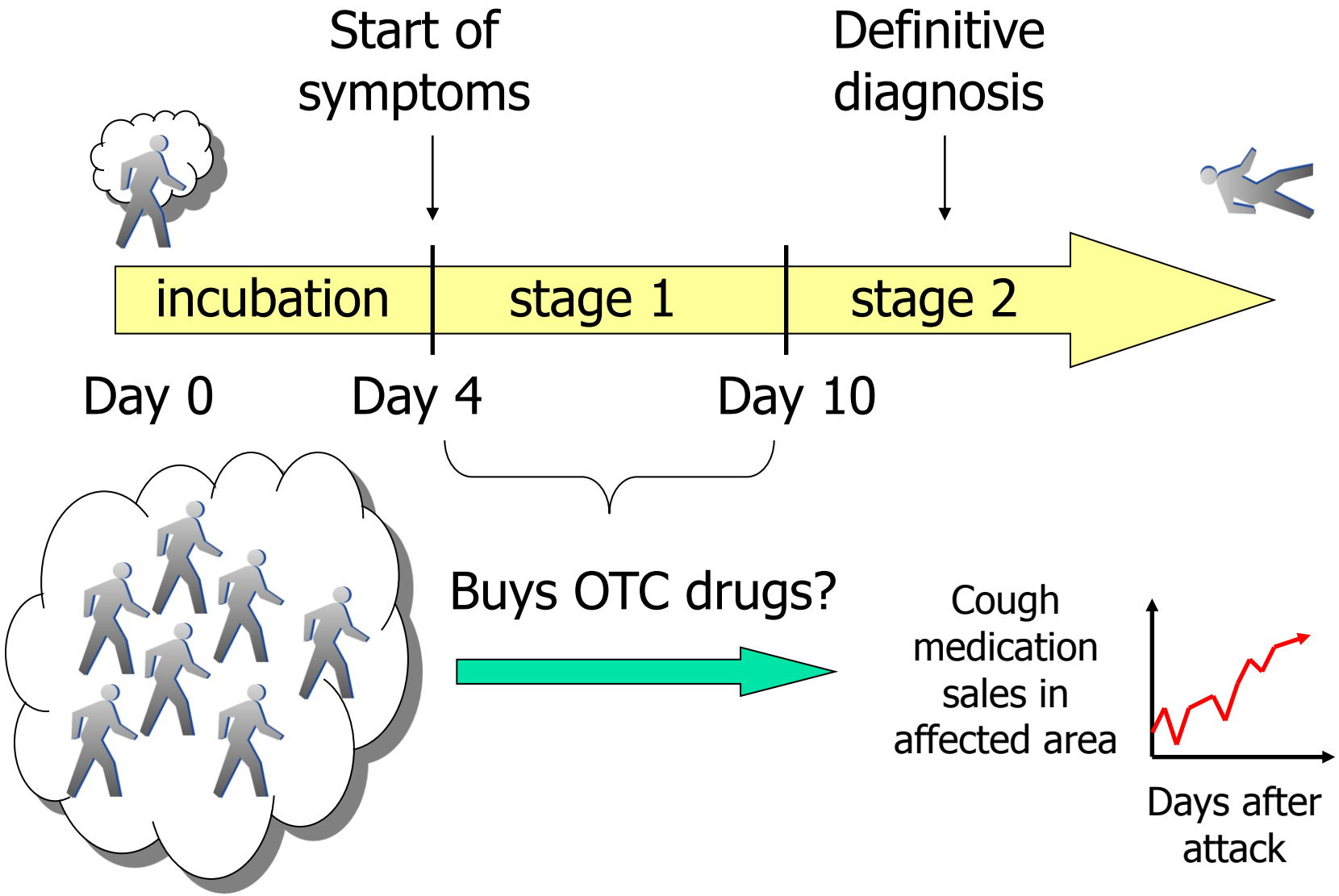
Buys OTC drugs

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.

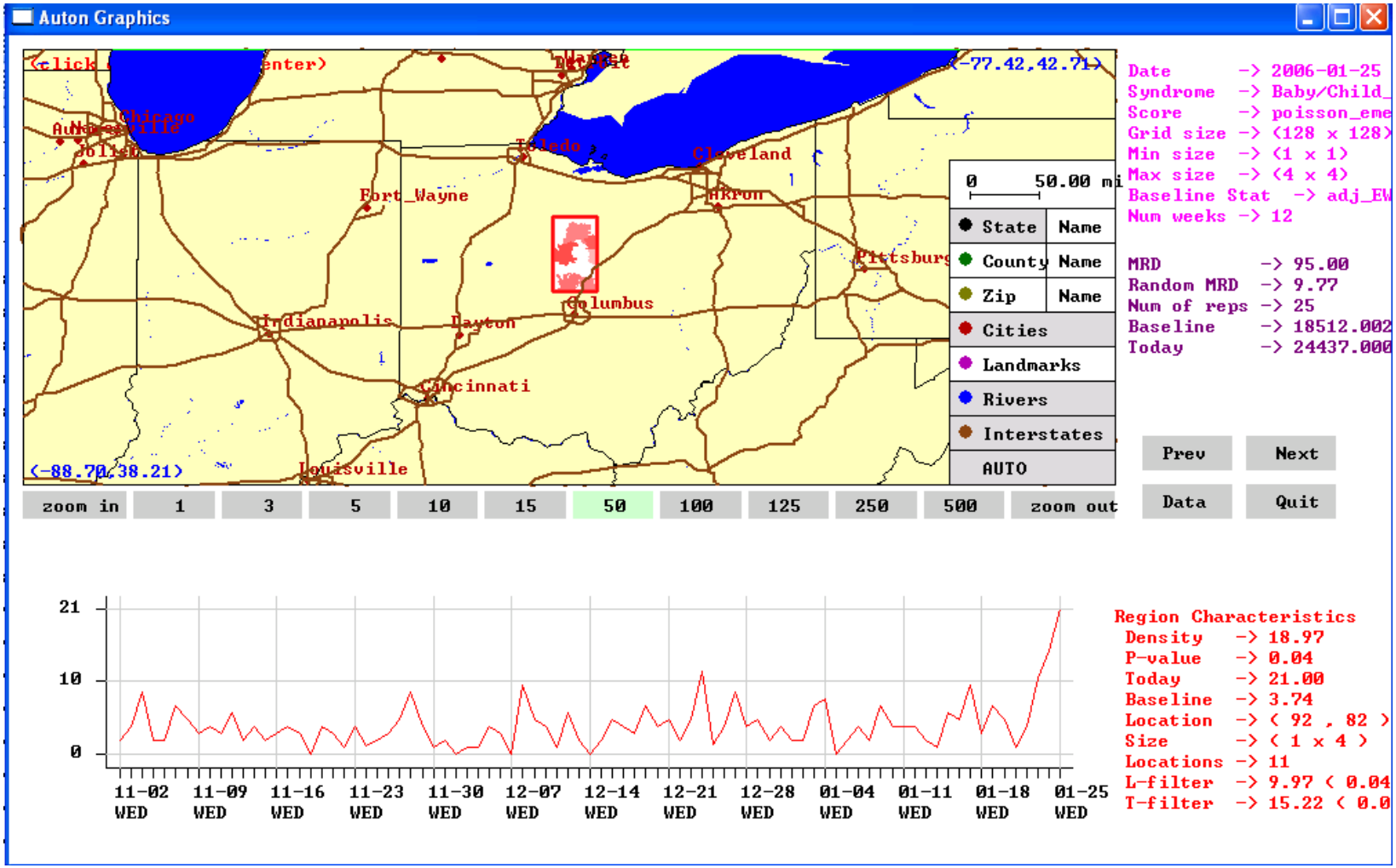
Buys OTC drugs?

Cough medication sales in affected area

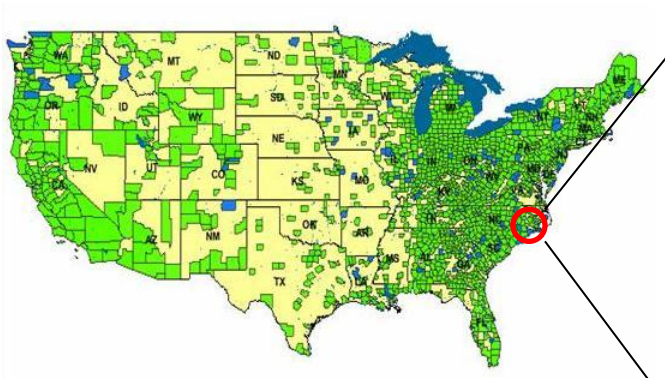
Days after attack

Univariate outbreak detection

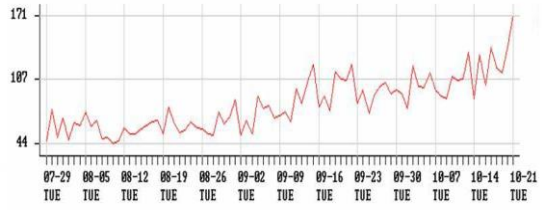
Spike in sales of pediatric electrolytes near Columbus, Ohio



Multivariate event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



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.)

Main goals:

Detect any emerging events.

Pinpoint the affected subset of locations and time duration.

Characterize the event by identifying the affected streams.

Compare hypotheses:

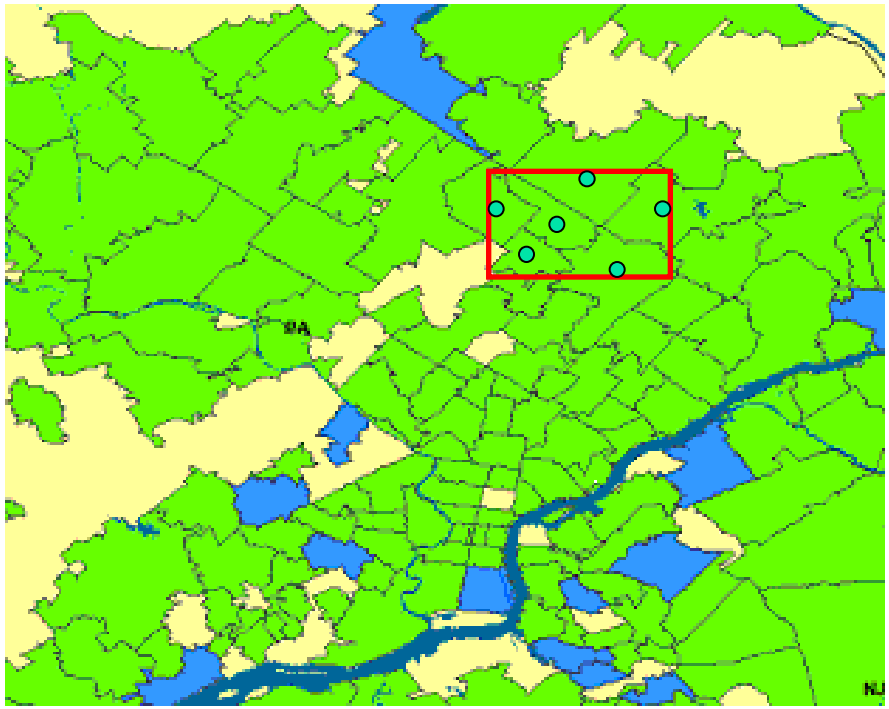
$H_1(D, S, W)$

D = subset of streams
S = subset of locations
W = time duration

vs. H_0: no events occurring

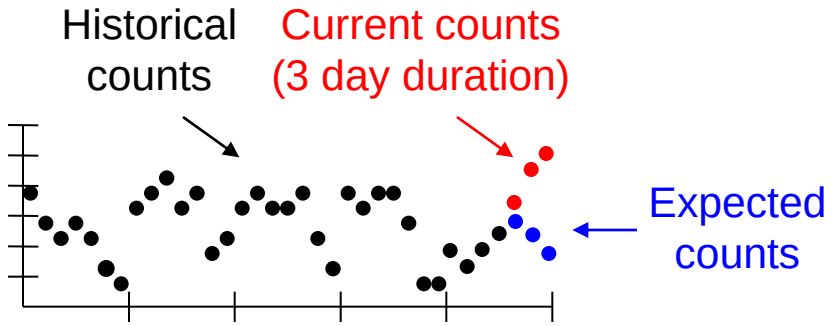
Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)



We search for spatial regions (subsets of locations) where the recently observed counts for some subset of streams are significantly higher than expected.

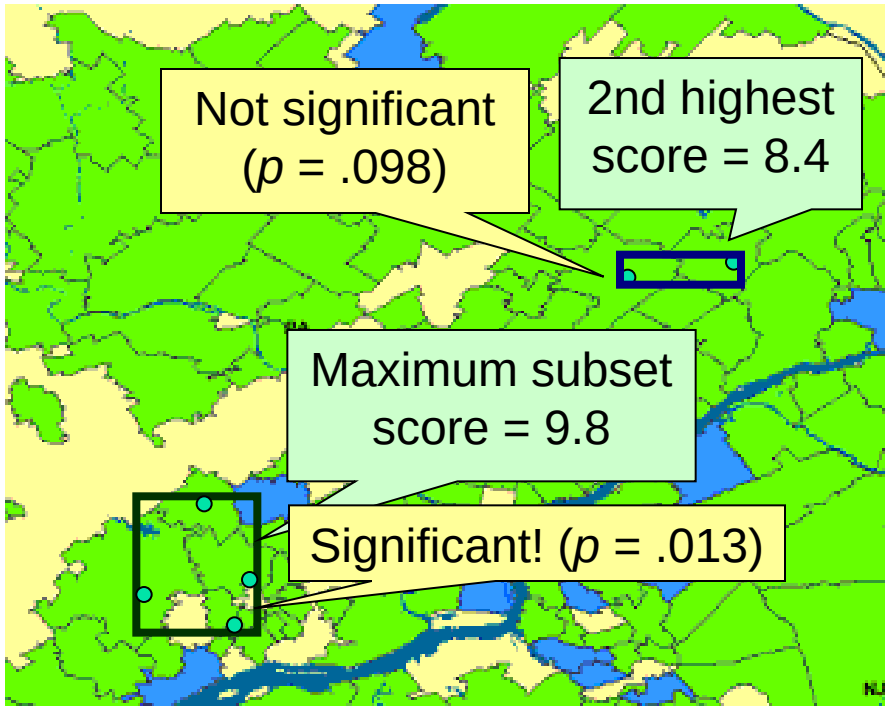
We perform **time series analysis** to compute expected counts ("baselines") for each location and stream for each recent day.



We then compare the actual and expected counts for each subset (D, S, W) under consideration.

Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)

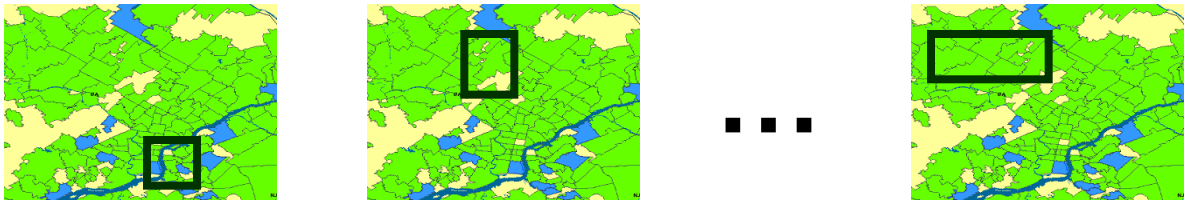


We find the subsets with highest values of a likelihood ratio statistic, and compute the *p*-value of each subset by randomization testing.

$$\mathrm{F}(D,S,W) = \frac{\Pr(\mathrm{Data} \mid H_1(D,S,W))}{\Pr(\mathrm{Data} \mid H_0)}$$

To compute p-value
Compare subset score to maximum subset scores of simulated datasets under H_0.

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



Scaling up surveillance

The landscape of surveillance is changing rapidly, due to increased availability of huge amounts of data at the societal scale.

Increasing use of detailed **electronic medical records** for patient data.

Informal, Web-based data sources such as Internet search queries and Twitter feeds.

New data sources have enormous **potential** for enabling more timely and accurate outbreak detection, but also pose many **challenges**.

Massive amounts of data…

Integrating many data sources…

Data mostly exists as **unstructured free text**!

Scaling up surveillance

The landscape of surve e is rapidly, du to increased a lability of hu t the etal scale.

Key message: New, cool data sources are not enough!

New methods are needed to deal with the **scale** and **complexity** of the new data.

sed data
search
ls.

New nabling more timely also pose many **challenges**.

Massive amount ta…

Integrating any data sources…

Data mostly exists as **unstructured free text**!

Where do existing methods fail?

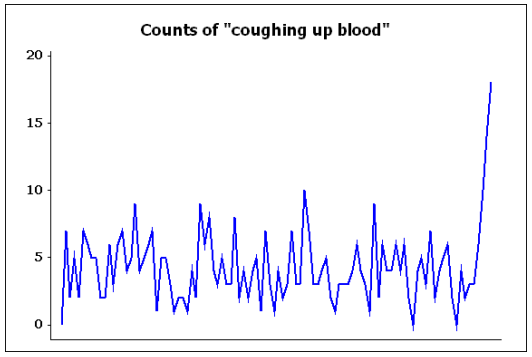
The typical, prodrome-based scan statistic approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

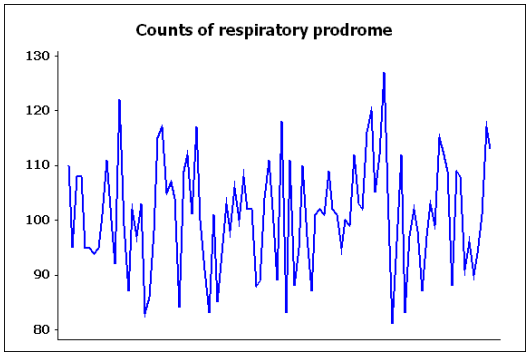
What happens when something new and scary comes along?

- **More specific symptoms** ("coughing up blood")
- **Previously unseen symptoms** ("nose falls off")

If we were monitoring these particular symptoms, it would only take a few such cases to realize that an outbreak is occurring!

Mapping specific chief complaints to a broader symptom category can dilute the outbreak signal, delaying or preventing detection.





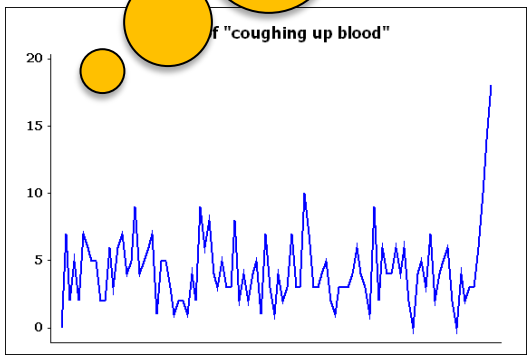
Where do existing methods fail?

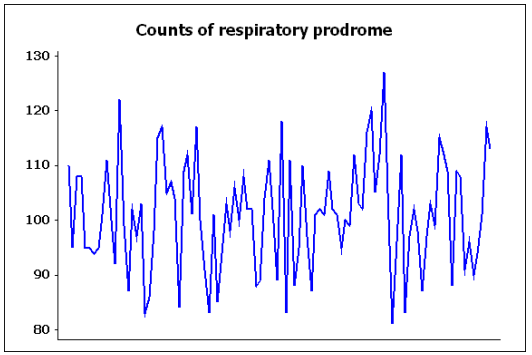
The typical, prod… scan statisti… effectively … outbreak… seen… sy…

…en something …s along? **…toms** …d") **…n** …off")

Our solution is to combine text-based (topic modeling) and spatial event detection (scan statistic) approaches, to detect **emerging spatial patterns of keywords**.

If we w… …ief complaints particular sy… …ptom category take a few such … …nate the outbreak signal, that an outb… …s occurring! …ying or preventing detection.





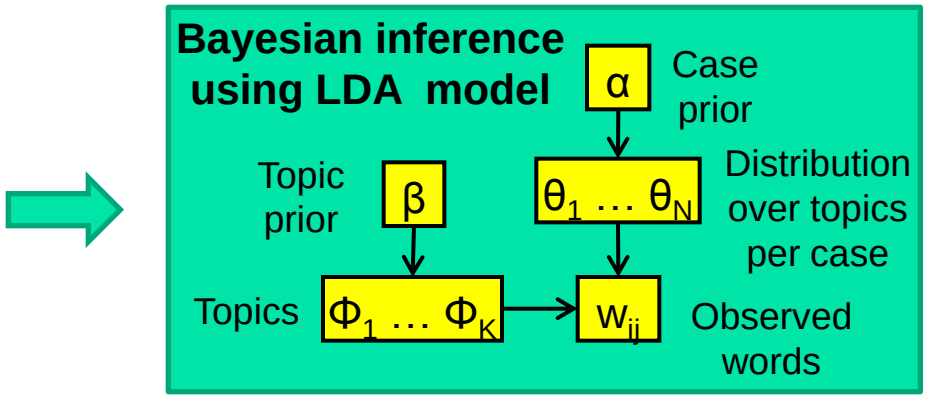
The semantic scan statistic

Date	Location	Complaint
1/1/11	15213	runny nose
1/1/11	15217	fever, chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp

2 years of free-text ED chief complaint data from 10 hospitals in Allegheny County.

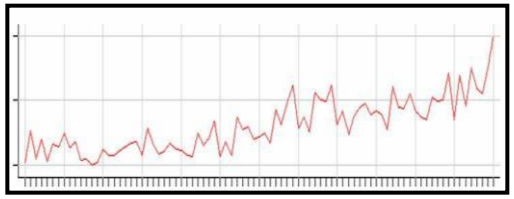
The semantic scan statistic

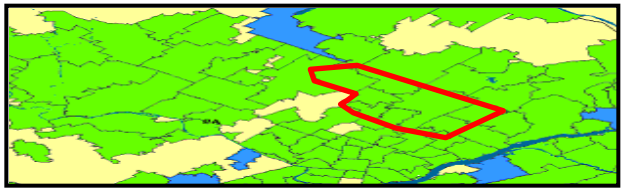
Date	Location	Complaint
1/1/11	15213	runny nose
1/1/11	15217	fever, chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp



Classify cases to topics

φ_1: vomiting, nausea, diarrhea, …
φ_2: dizzy, lightheaded, weak, …
φ_3: cough, throat, sore, …





Time series of counts for each location, for each topic T

Find topic T and region S maximizing the likelihood ratio statistic, F(S, T)

Results

Semantic scan achieved detected emerging, novel outbreaks **more than twice as fast** as the standard prodrome-based method (5.3 days vs. 10.9 days to detect)

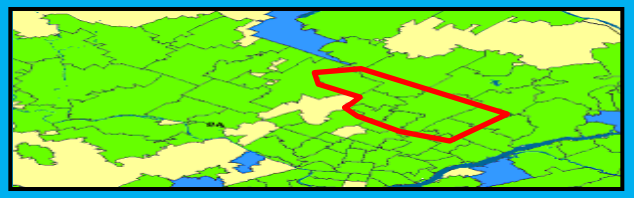
Simulated novel outbreak: “green nose”

green
nose
possible
color
greenish
nasal
…

Top words from detected topic

Fast subset scanning

We want to perform a constrained search over **subsets** of locations and data streams. but it is computationally infeasible to perform an exhaustive search over all subsets.



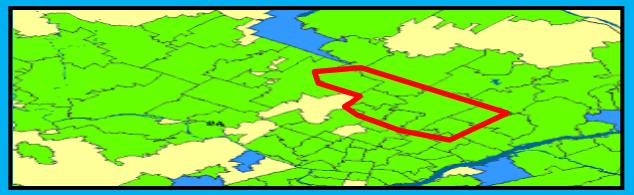
Find topic T and region S maximizing the likelihood ratio statistic, F(S, T)

Fast subset scanning

We show that it is possible to scan over the exponentially many subsets of the data in linear time. This approach reduces run time from **years** to **milliseconds** in practice!

Many likelihood ratio statistics satisfy the **linear-time subset scanning** (LTSS) property:

Sort the data from highest to lowest priority, then search over data records consisting of the top-k highest priority locations.



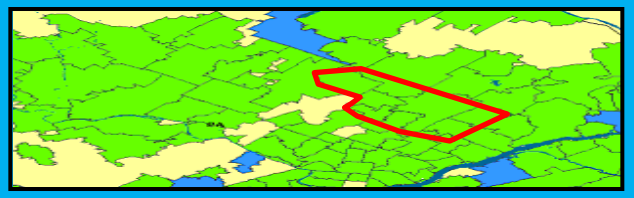
Find topic T and region S maximizing the likelihood ratio statistic, F(S, T)

Fast subset scanning

We show that it is possible to scan over the exponentially many subsets of the data in linear time. This approach reduces run time from **years** to **milliseconds** in practice!

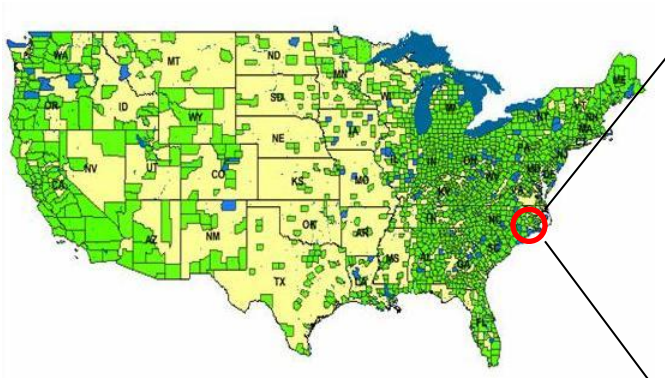
Many likelihood ratio statistics satisfy the **linear-time subset scanning** (LTSS) property:

The highest scoring of all 2^N subsets is **guaranteed** to be one of these N subsets!

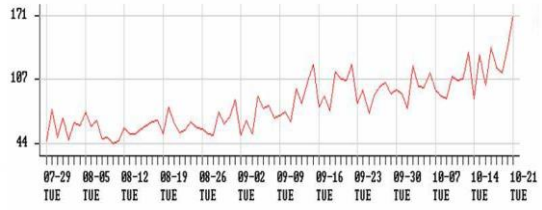


Find topic T and region S maximizing the likelihood ratio statistic, F(S, T)

Multivariate event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



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.)

Main goals:

Detect any emerging events.

Pinpoint the affected subset of locations and time duration.

Characterize the event by identifying the affected streams.

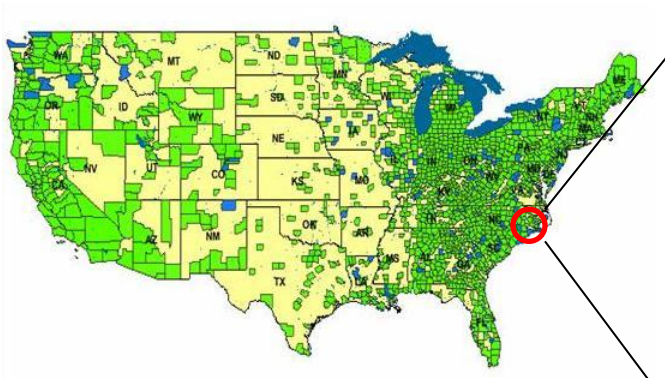
Compare hypotheses:

$H_1(D, S, W)$

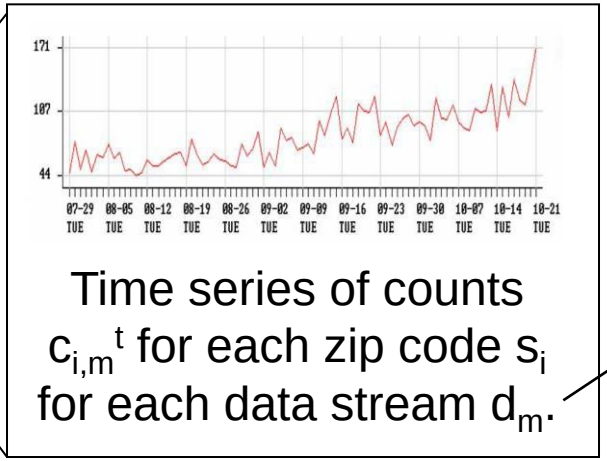
D = subset of streams
S = subset of locations
W = time duration

vs. H_0: no events occurring

Multidimensional event detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



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.)

Additional goal: identify any differentially affected **subpopulations** P of the monitored population.

Gender (male, female, both)

Age groups (children, adults, elderly)

Ethnic or socio-economic groups

Risk behaviors: e.g. intravenous drug use, multiple sexual partners

More generally, assume that we have a set of additional discrete-valued attributes $A_1..A_J$ observed for each individual case.

We identify not only the affected streams, locations, and time window, but also a **subset** of values for each attribute.

Multidimensional LTSS

- Our **MD-Scan** approach (Neill and Kumar, 2013) extends LTSS to the multidimensional case:
  - For each time window and spatial neighborhood (center + k-nearest neighbors), we do the following:

1. Start with randomly chosen subsets of **locations** S, **streams** D, and **values** V_j for each attribute A_j (j=1..J).

2. Choose an attribute (randomly or sequentially) and use LTSS to find the highest scoring subset of values, locations, or streams, conditioned on all other attributes.

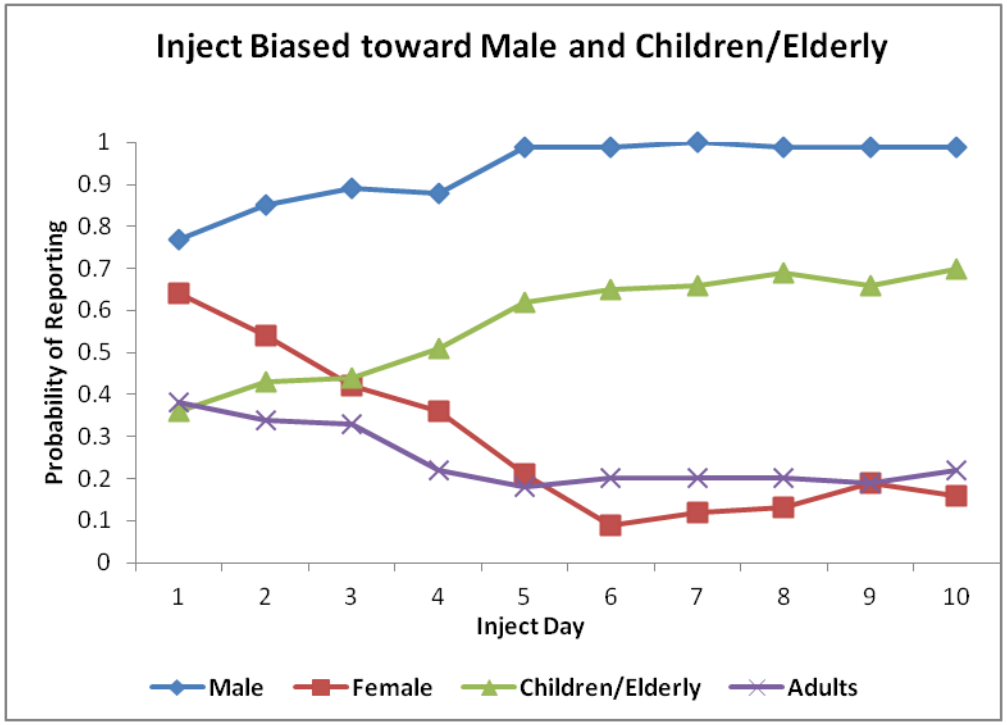
3. Iterate step 2 until convergence to a local maximum of the score function F(D,S,W, $\{V_j\}$), and use multiple restarts to approach the global maximum.

Evaluation

- We evaluated the detection performance of MD-Scan for detecting disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- We considered outbreaks with various types and amounts of age and gender bias.
  - Shown here: biased toward males, biased toward children and the elderly.

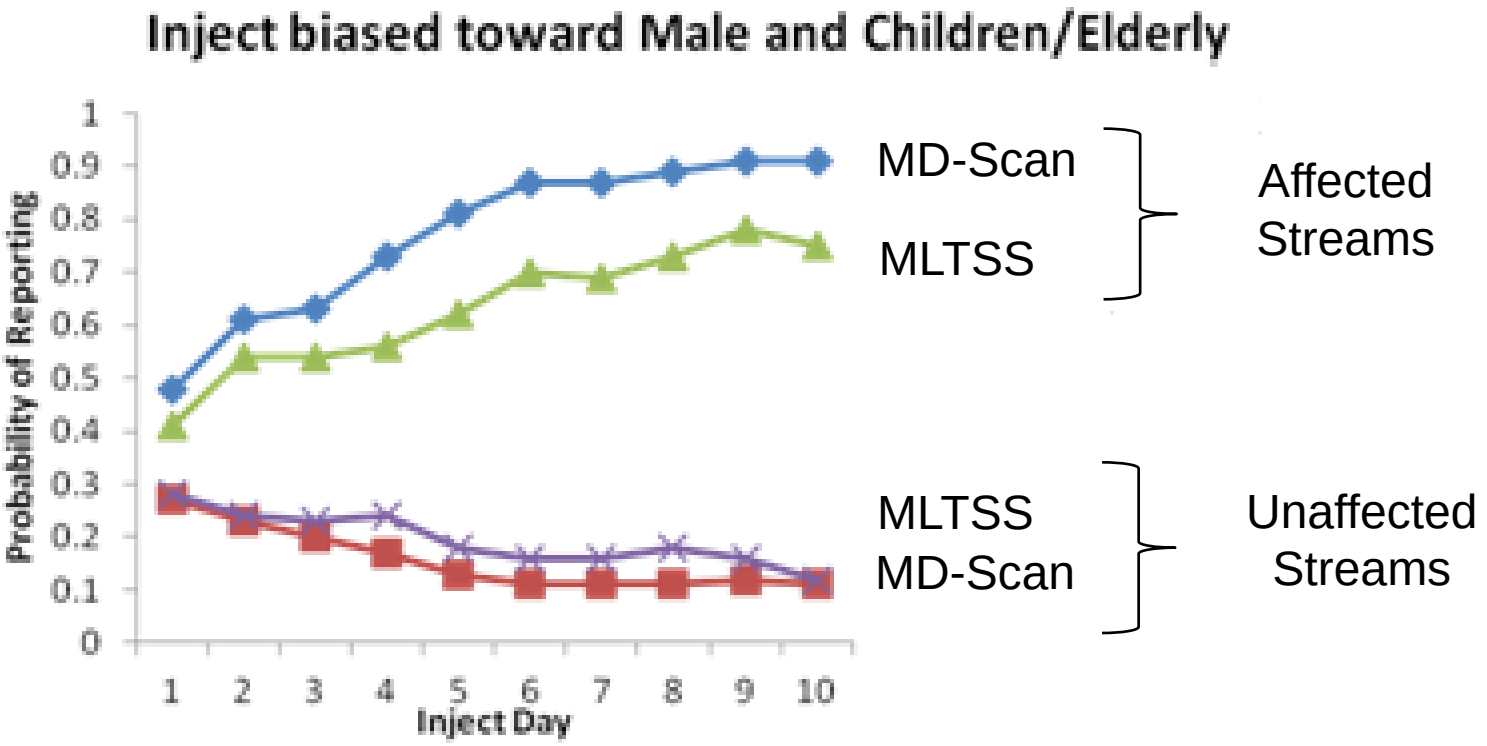
1) Identifying affected subpopulations

By the midpoint of the outbreak, MD-Scan is able to correctly identify the affected gender and age deciles with high probability, without reporting unaffected subpopulations.



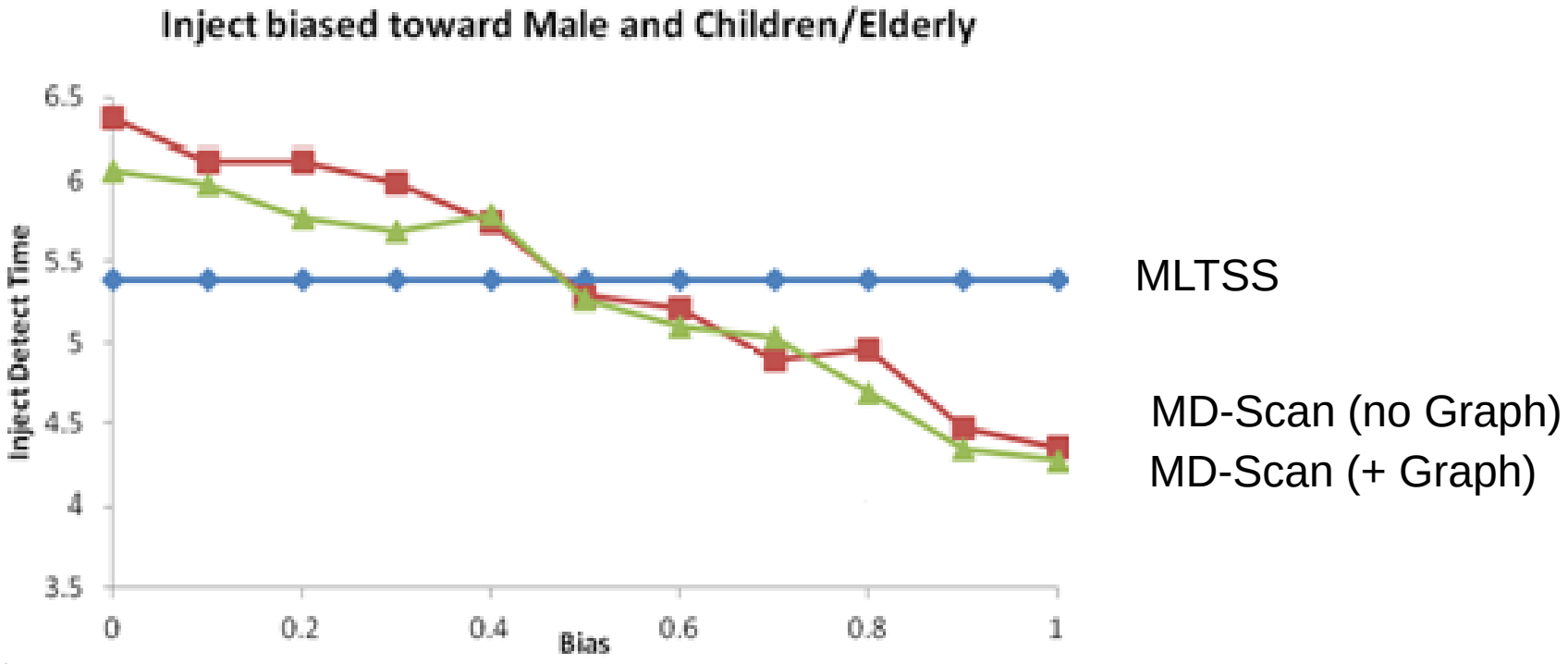
2) Characterizing affected streams

As compared to the current state of the art (multivariate linear-time subset scanning, or MLTSS), MD-Scan is better able to characterize the affected subset of the monitored streams.

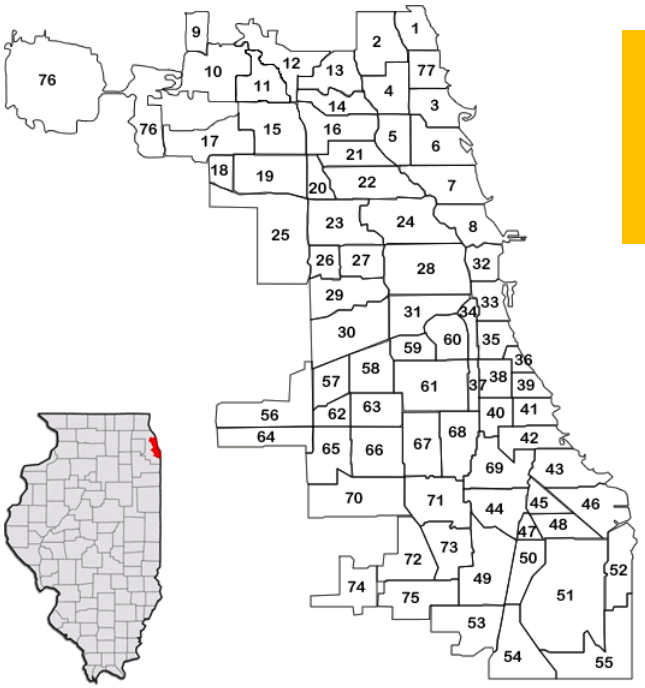


3) Timeliness of outbreak detection

At a fixed false positive rate of 1 per month, MD-Scan achieved faster detection for outbreaks which were sufficiently biased by age and/or gender.



Case study 2: Crime prediction in Chicago



Since 2009, we have been working with the Chicago Police Department (CPD) to predict and prevent emerging clusters of violent crime.

Our new crime prediction methods have been incorporated into our **CrimeScan** software, which has been used operationally by CPD for deployment of patrols.

From the Chicago Sun-Times, February 22, 2011:

"It was a bit like "Minority Report," the 2002 movie that featured genetically altered humans with special powers to predict crime. The CPD's new crime-forecasting unit was analyzing 911 calls and produced an intelligence report predicting a shooting would happen soon on a particular block on the South Side. Three minutes later, it did…"

CrimeScan

The key insight of our method is to **use detection for prediction**:

We can **detect emerging clusters** of various leading indicators (minor crimes, 911 calls, etc.) and use these to **predict** that a cluster of violent crime is likely to occur nearby.

Some advantages of the CrimeScan approach:

- Advance prediction (up to 1 week) with high accuracy.
- High spatial and temporal resolution (block x day).
- Predicting **emerging hot spots** of violence (as opposed to just identifying bad neighborhoods).

How to detect leading indicator clusters?
How to use these for prediction?
Which leading indicators to use?

CrimeScan: Prediction

We are currently investigating two different prediction methods, both of which use the detected leading indicator clusters as features of a predictive model.

Density-based prediction:

Areas which are closer to a significant cluster of any of the monitored LI are assumed more likely to have a spike in violence in the near future.

Total proximity to leading indicator clusters is computed by kernel density estimation:

$$\text{score} = \sum \exp(-d_i^2/2)$$

(where d_i is distance to the i^{th} leading indicator cluster)

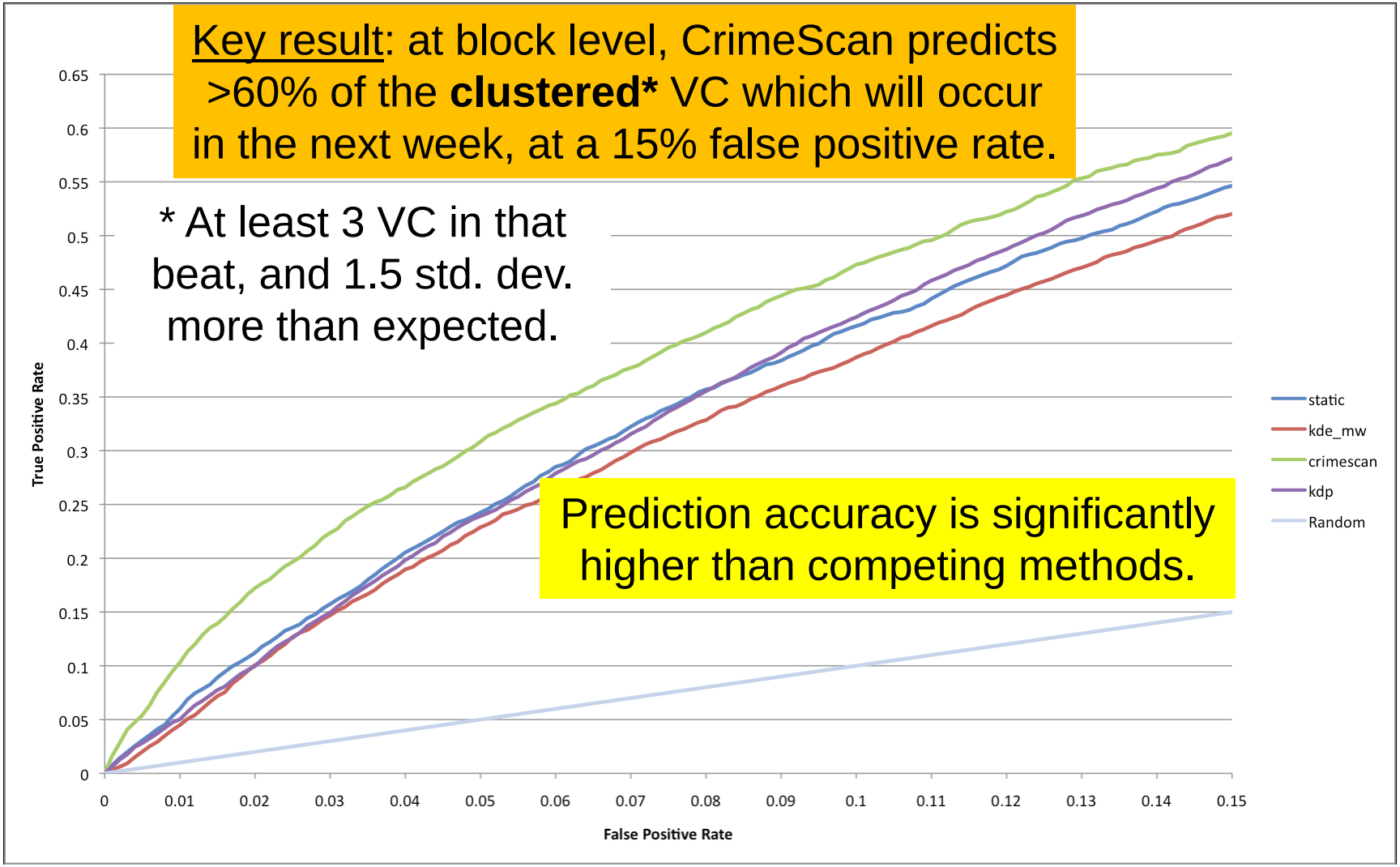
Model-based prediction:

We learn a sparse (penalized) logistic regression model, with binary features including the presence of each type of LI cluster within some radius.

$$\log(p / (1-p)) = \beta_0 + \beta_1 x_1 + \ldots$$

Advantages: can learn which LI types are most relevant for prediction, and can include various additional features (month, day of week, weather...)

CrimeScan: Preliminary Results



From CrimeScan to CityScan…

Working with the City of Chicago's Department of Innovation and Technology, we are currently using our new methods to analyze other data relevant to the city.

Most interestingly, we have very promising initial results for prediction of emerging patterns of 311 calls.

Examples: abandoned buildings, graffiti, sanitation complaints, rodent removal, garbage carts…

Our CrimeScan software has been renamed "CityScan" and will be an essential component of the city's new **Chicago SmartData platform** for real-time predictive analytics and decision making, with applications including rodent control, preventing STIs, and emergency response.

Interested?

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

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