

# Fast Subset Scan for Population Health and Disease Surveillance

**Daniel B. Neill**  
**H.J. Heinz III College**  
**Carnegie Mellon University**  
**E-mail: [neill@cs.cmu.edu](mailto:neill@cs.cmu.edu)**

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

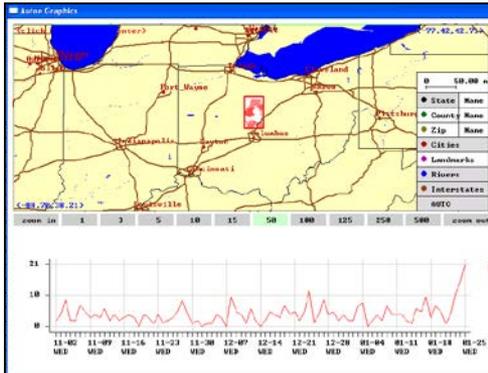
Carnegie Mellon University

**EPD Lab**

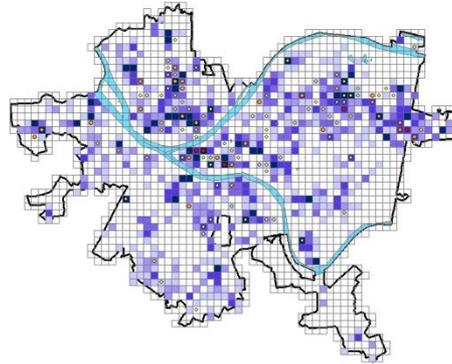
EVENT AND PATTERN DETECTION LABORATORY



Daniel B. Neill (neill@cs.cmu.edu)  
Associate Professor of Information Systems, Heinz College, CMU  
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.



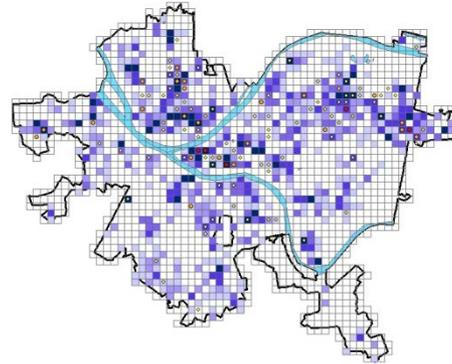
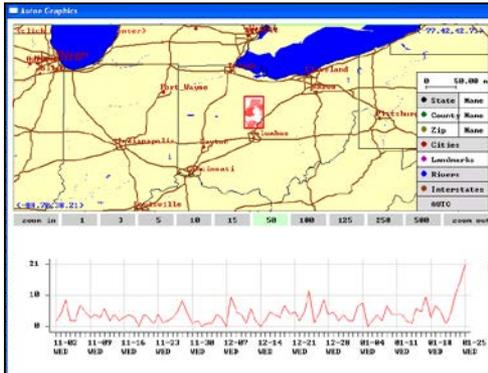
Medicine: Discovering new “best practices” of patient care, to improve outcomes and reduce costs.

My research is focused at the intersection of **machine learning** and **public policy**, with 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, Heinz College, CMU  
 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 Police Dept. “CityScan” has been used by Chicago city leaders for prediction and prevention of rodent infestations using 311 call data.

## Today's talk:

- Public health surveillance
  - Early outbreak detection (fast subset scan)
  - Accidental drug overdose surveillance (multidimensional scan)
  - “Novel” outbreak detection (semantic scan)



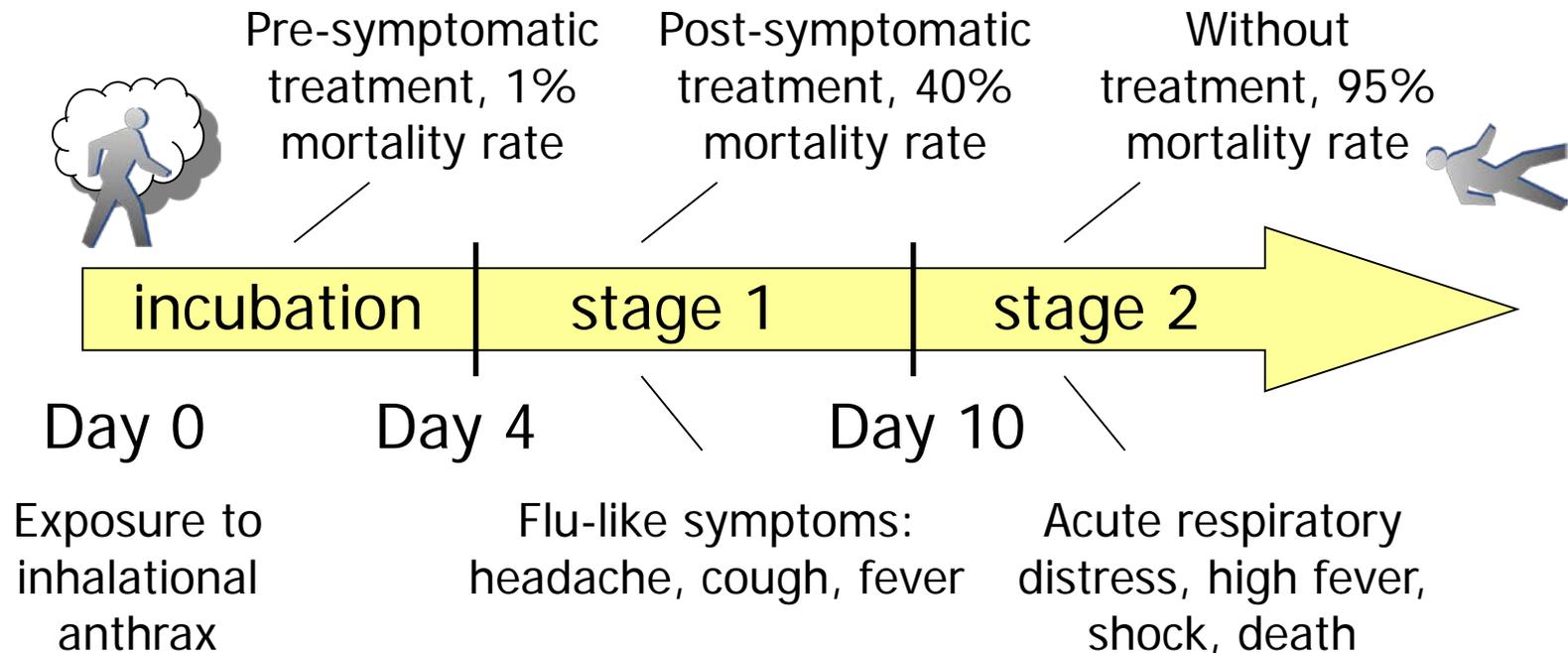
# Why worry about disease outbreaks?

- Bioterrorist attacks are a very real, and scary, possibility
  - Large anthrax release over a major city 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 and emerging public health trends.



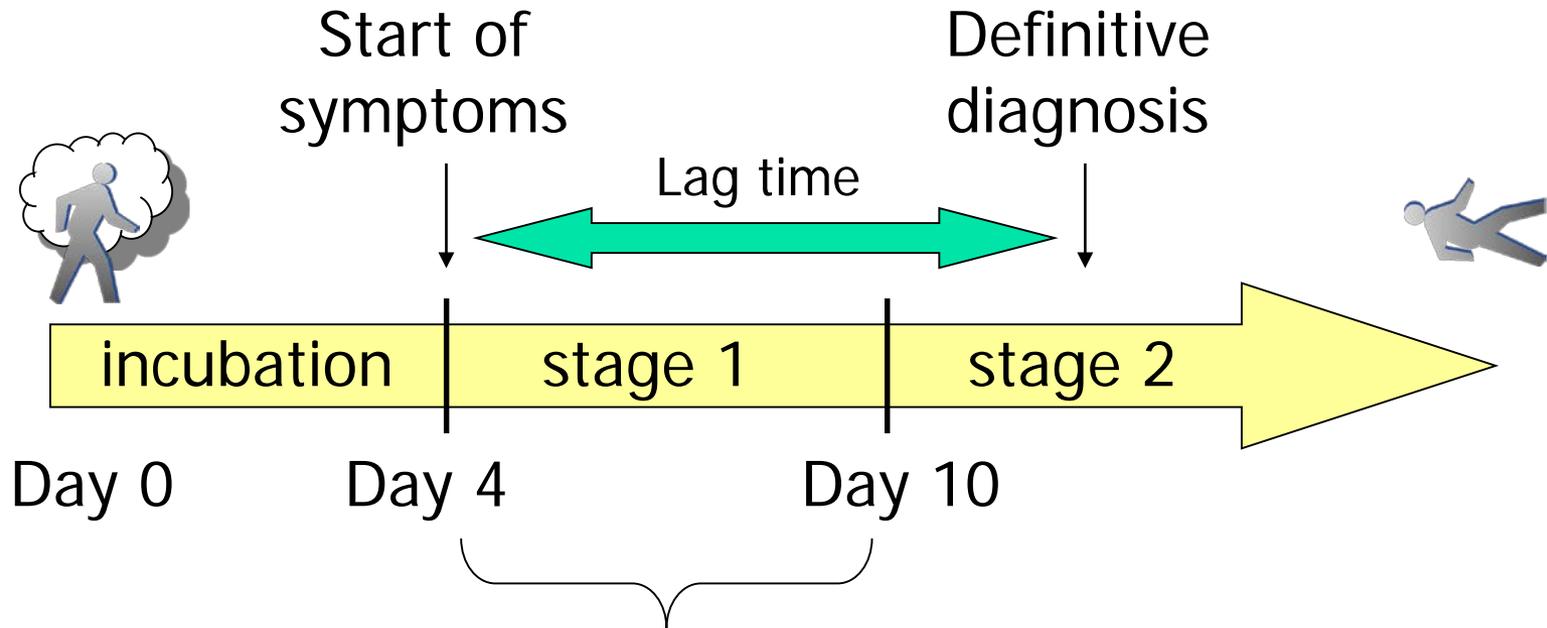
# Benefits of early detection

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.**

# Early detection is hard



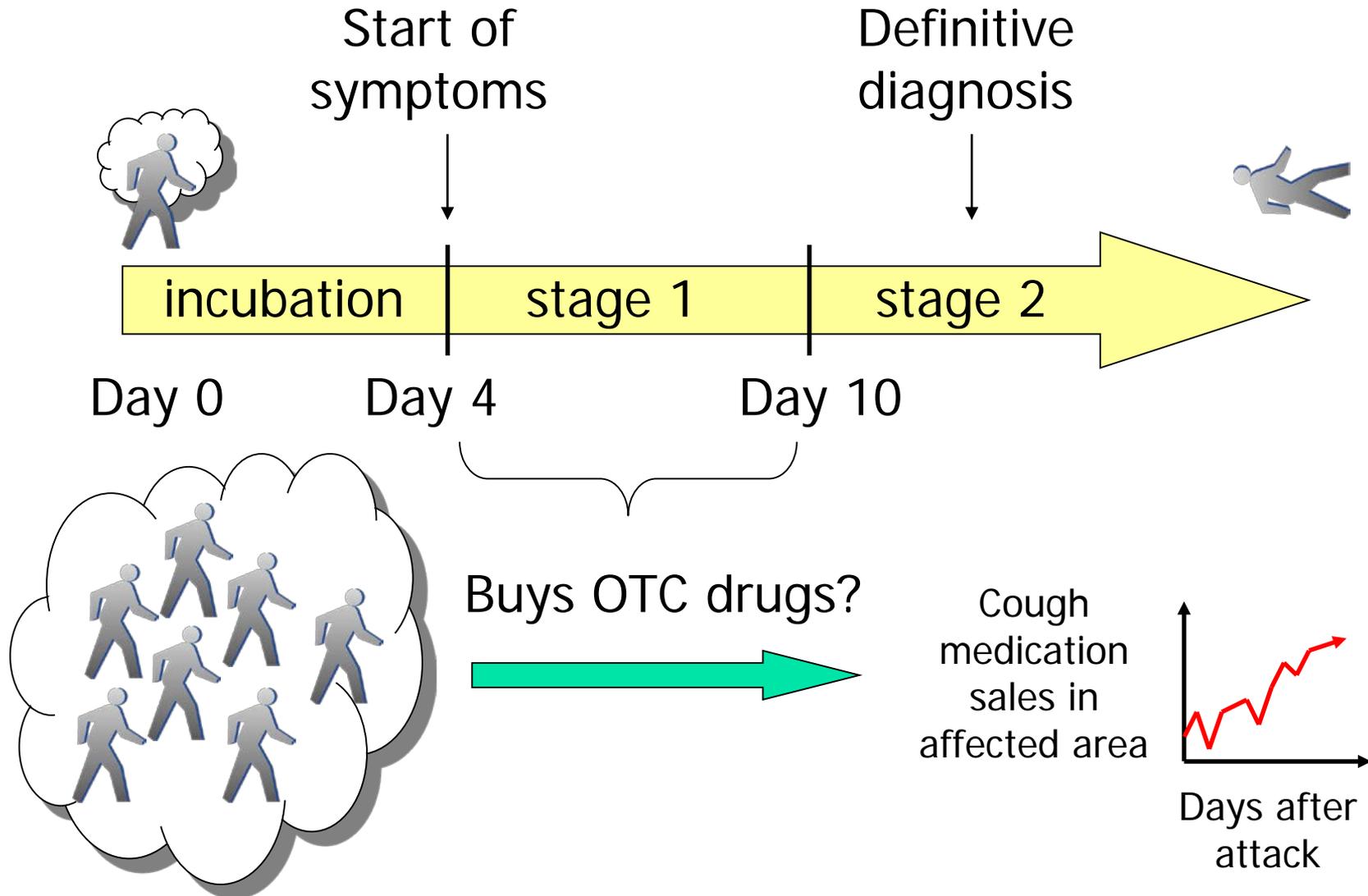
Buys OTC drugs

Skips work/school

Uses Google, Facebook, Twitter

Visits doctor/hospital/ED

# Syndromic surveillance



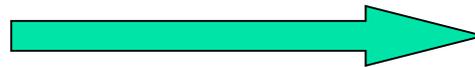
# Syndromic surveillance

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?



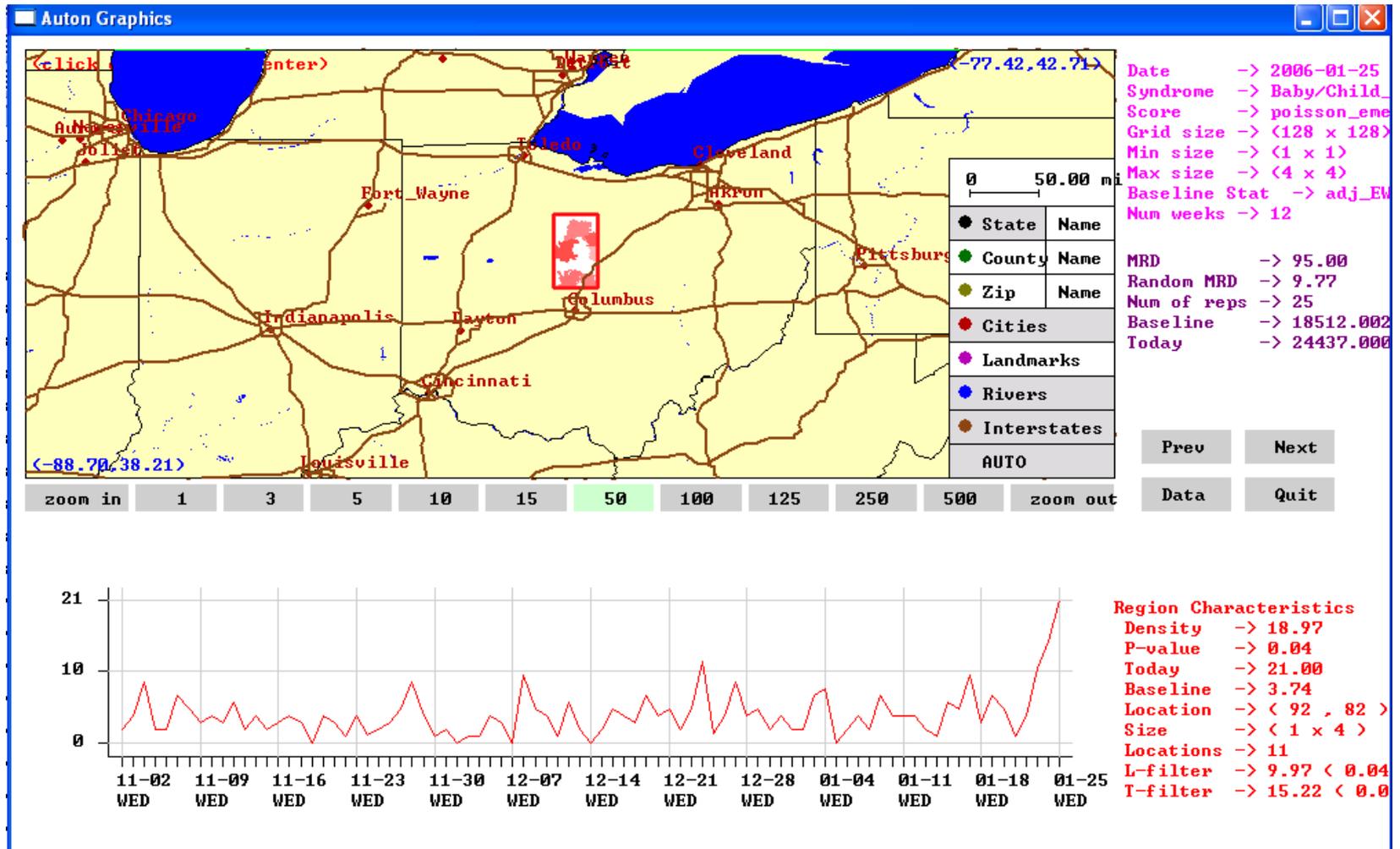
Cough  
medication  
sales in  
affected area



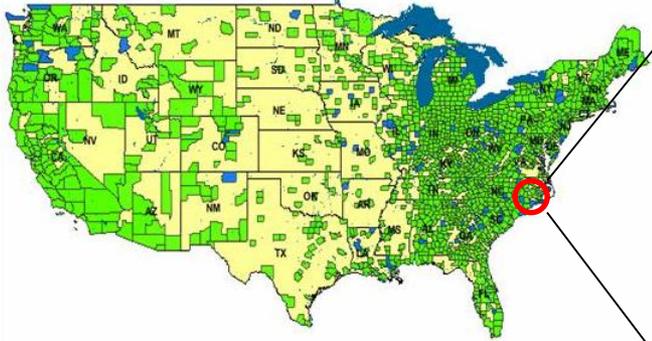
Days after  
attack

# Outbreak detection example

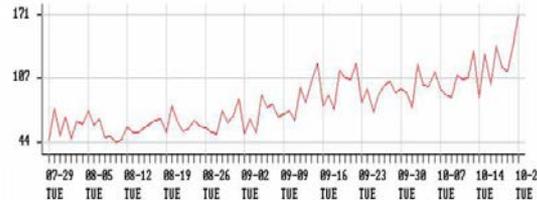
## Spike in sales of pediatric electrolytes near Columbus, Ohio



# 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  $d_m$ .

## Outbreak detection

- $d_1$  = 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, e.g., by identifying the affected streams.

## Compare hypotheses:

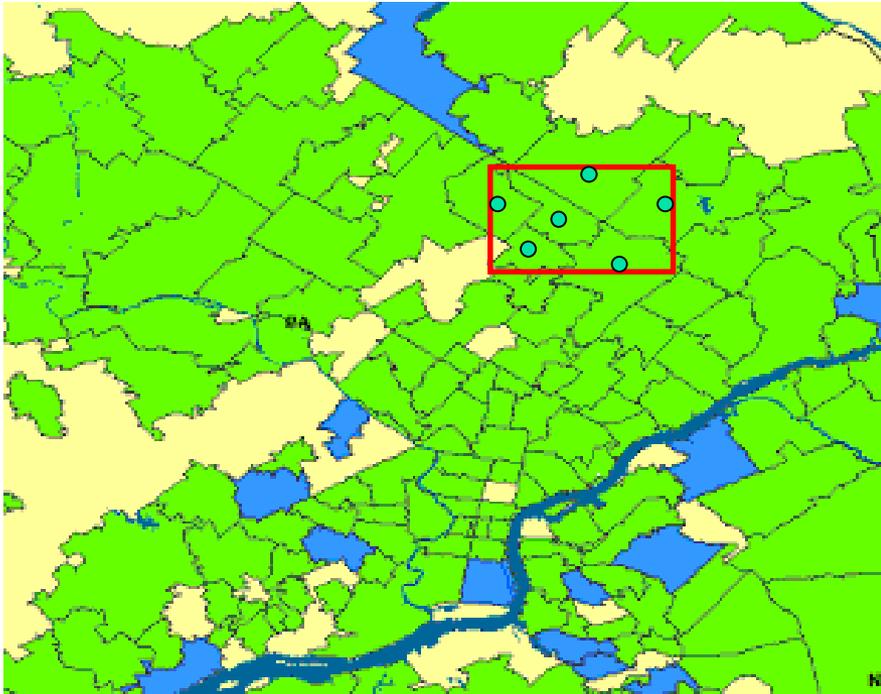
$$H_1(D, S, W)$$

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

vs.  $H_0$ : no events occurring

# Expectation-based scan statistics

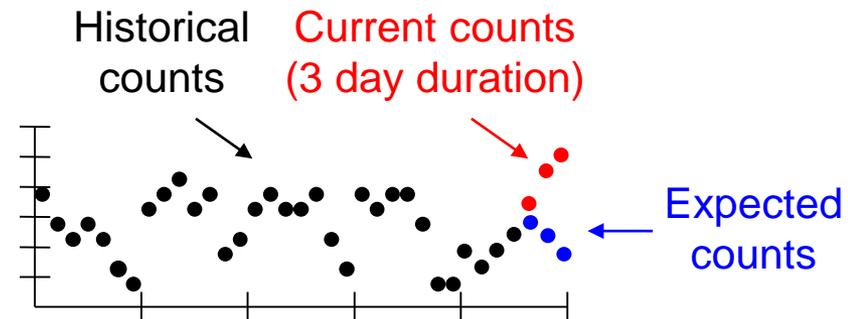
(Kulldorff, 1997; Neill and Moore, 2005)



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

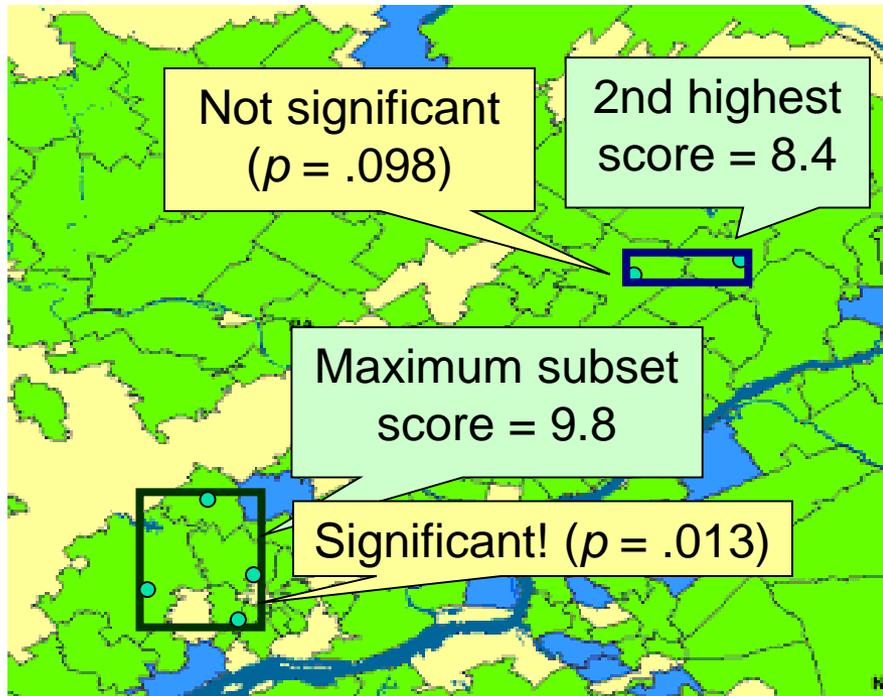
We perform **time series analysis** to compute expected counts (“baselines”) for each location and stream for each recent day.

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.

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

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

$F_1^* = 2.4$



$F_2^* = 9.1$



...

$F_{999}^* = 7.0$



# Likelihood ratio statistics

For our expectation-based scan statistics, the null hypothesis  $H_0$  assumes “business as usual”: each count  $c_{i,m}^t$  is drawn from some parametric distribution with mean  $b_{i,m}^t$ .  $H_1(S)$  assumes a multiplicative increase for the affected subset  $S$ .

## Expectation-based Poisson

$$H_0: c_{i,m}^t \sim \text{Poisson}(b_{i,m}^t)$$

$$H_1(S): c_{i,m}^t \sim \text{Poisson}(qb_{i,m}^t)$$

$$\text{Let } C = \sum_S c_{i,m}^t \text{ and } B = \sum_S b_{i,m}^t.$$

$$\text{Maximum likelihood: } q = C / B.$$

$$F(S) = C \log (C/B) + B - C$$

## Expectation-based Gaussian

$$H_0: c_{i,m}^t \sim \text{Gaussian}(b_{i,m}^t, \sigma_{i,m}^t)$$

$$H_1(S): c_{i,m}^t \sim \text{Gaussian}(qb_{i,m}^t, \sigma_{i,m}^t)$$

$$\text{Let } C' = \sum_S c_{i,m}^t b_{i,m}^t / (\sigma_{i,m}^t)^2 \\ \text{and } B' = \sum_S (b_{i,m}^t)^2 / (\sigma_{i,m}^t)^2.$$

$$\text{Maximum likelihood: } q = C' / B'.$$

$$F(S) = (C')^2 / 2B' + B'/2 - C'$$

Many possibilities: exponential family, nonparametric, Bayesian...

# Which regions to search?

Typical approach: “spatial scan” (Kulldorff, 1997)

Each search region  $S$  is a **sub-region** of space.

- Choose some region shape (e.g. circles, rectangles) and consider all regions of that shape and varying size.
- Low power for true events that do not correspond well to the chosen set of search regions (e.g. irregular shapes).

Our approach: “subset scan” (Neill, 2012)

Each search region  $S$  is a **subset** of locations.

- Find the highest scoring subset, subject to some constraints (e.g. spatial proximity, connectivity).
- For multivariate, also optimize over subsets of streams.
- Exponentially many possible subsets,  $O(2^N \times 2^M)$ : computationally infeasible for naïve search.

# Fast subset scan (Neill, 2012)

- In certain cases, we can optimize  $F(S)$  over the exponentially many subsets of the data, while evaluating only  $O(N)$  rather than  $O(2^N)$  subsets.
- Many commonly used scan statistics have the property of linear-time subset scanning:
  - Just sort the data records (or spatial locations, etc.) from highest to lowest priority according to some function...
  - ... then search over groups consisting of the top-k highest priority records, 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**.

# Linear-time subset scanning

- Example: Expectation-Based Poisson statistic
  - Sort data locations  $s_i$  by the ratio of observed to expected count,  $c_i / b_i$ .
  - Given the ordering  $s_{(1)} \dots s_{(N)}$ , we can **prove** that the top-scoring subset  $F(S)$  consists of the locations  $s_{(1)} \dots s_{(k)}$  for some  $k$ ,  $1 \leq k \leq N$ .
  - Key step: if there exists some location  $s_{\text{out}} \notin S$  with higher priority than some location  $s_{\text{in}} \in S$ , then we can show that  $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$ .
- Theorem: LTSS holds for expectation-based scan statistics in any exponential family. (Speakman et al., 2015)

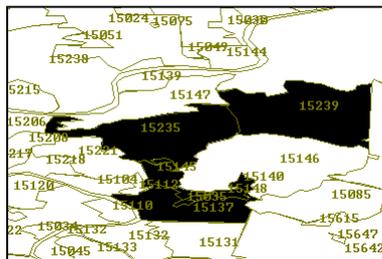
$$F(S) = \max_{q>1} \log \frac{P(\text{Data} \mid H_1(S))}{P(\text{Data} \mid H_0)} \quad \begin{array}{l} H_0 : x_i \sim \text{Dist}(\mu_i) \\ H_1 : x_i \sim \text{Dist}(q\mu_i) \end{array}$$

# Constrained fast subset scanning

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.

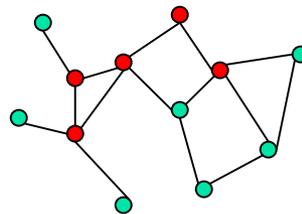
Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + 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

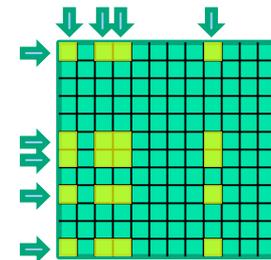


(Neill, *JRSS-B*, 2012)

(Neill et al., *Stat. Med.*, 2013)



(Speakman et al., *JCGS*, 2015)



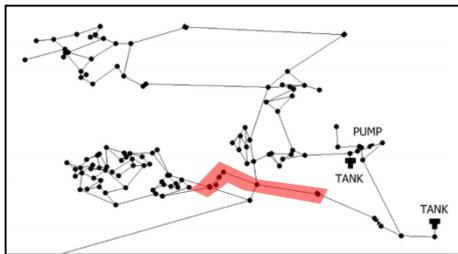
(McFowland et al., *JMLR*, 2013)

# Constrained fast subset scanning

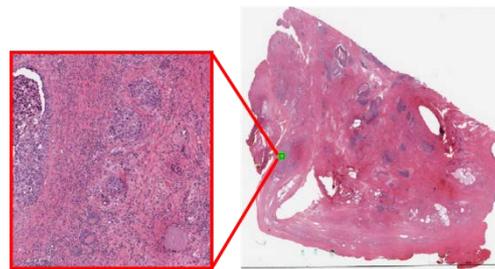
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + Temporal dynamics → Spreading contamination in water supply
- + Hierarchical scanning → Prostate cancer in digital pathology slides
- + Scalable GP regression → Predicting and preventing rat infestations



(Speakman et al., ICDM 2013)



(Somanchi & Neill, DMHI 2013)



(Flaxman et al., 2015;  
Neill et al., in preparation)

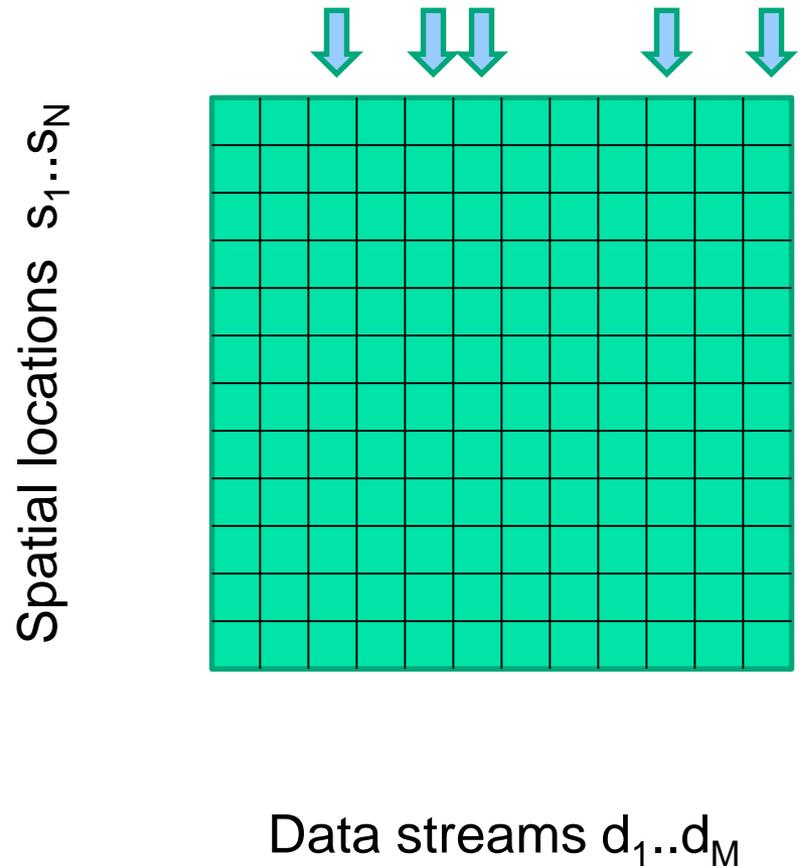
# Fast subset scan with spatial proximity constraints

- Maximize a likelihood ratio statistic over all subsets of the “local neighborhoods” consisting of a center location  $s_i$  and its  $k-1$  nearest neighbors, for a fixed neighborhood size  $k$ .
- Naïve search requires  $O(N \cdot 2^k)$  time and is computationally infeasible for  $k > 25$ .
- For each center, we can search over all subsets of its local neighborhood in  $O(k)$  time using LTSS, thus requiring a total time complexity of  $O(Nk) + O(N \log N)$  for sorting the locations.
- In Neill (2012), we show that this approach dramatically improves the timeliness and accuracy of outbreak detection for irregularly-shaped disease clusters.

# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

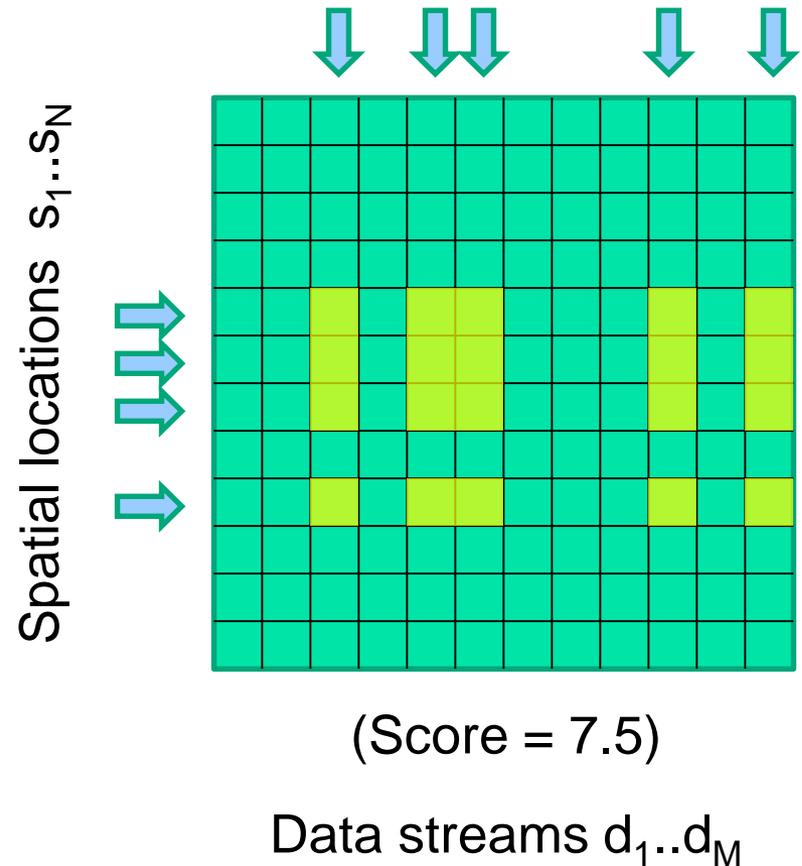
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

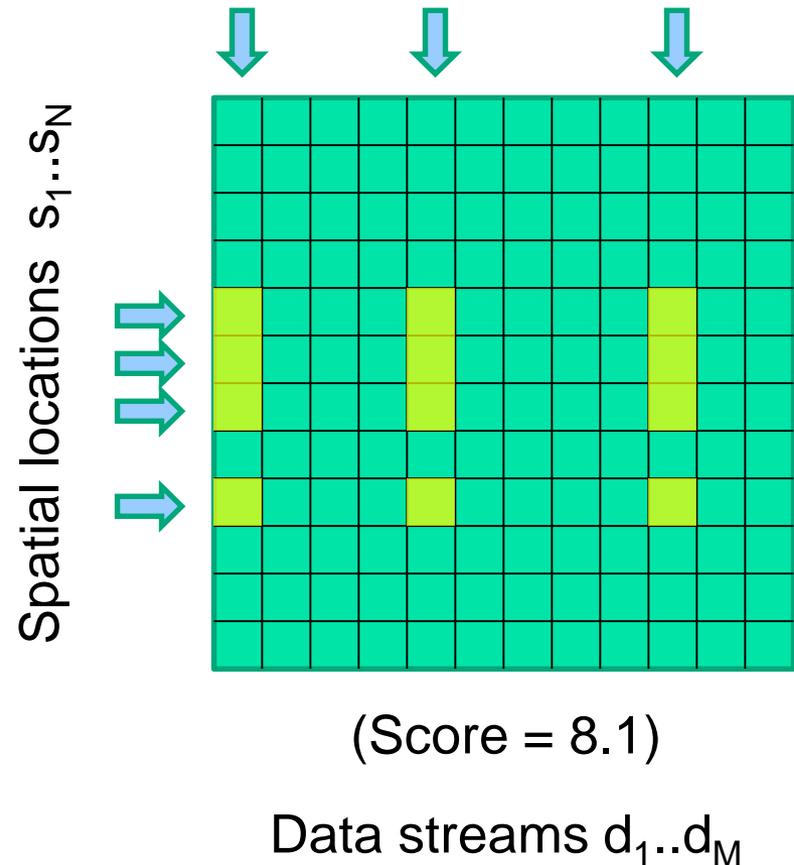
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

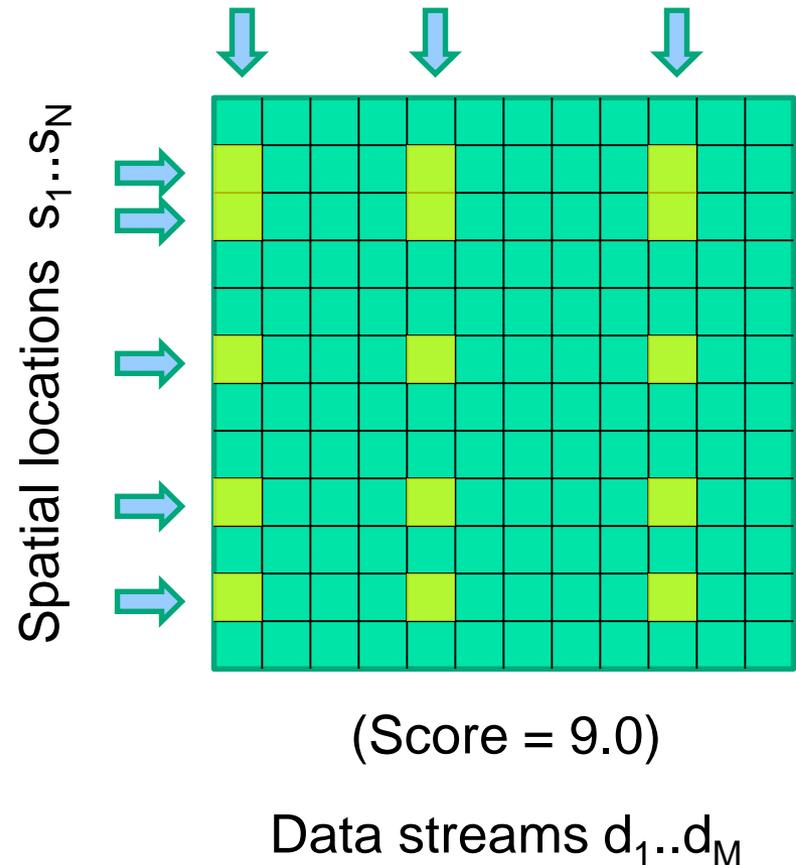
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

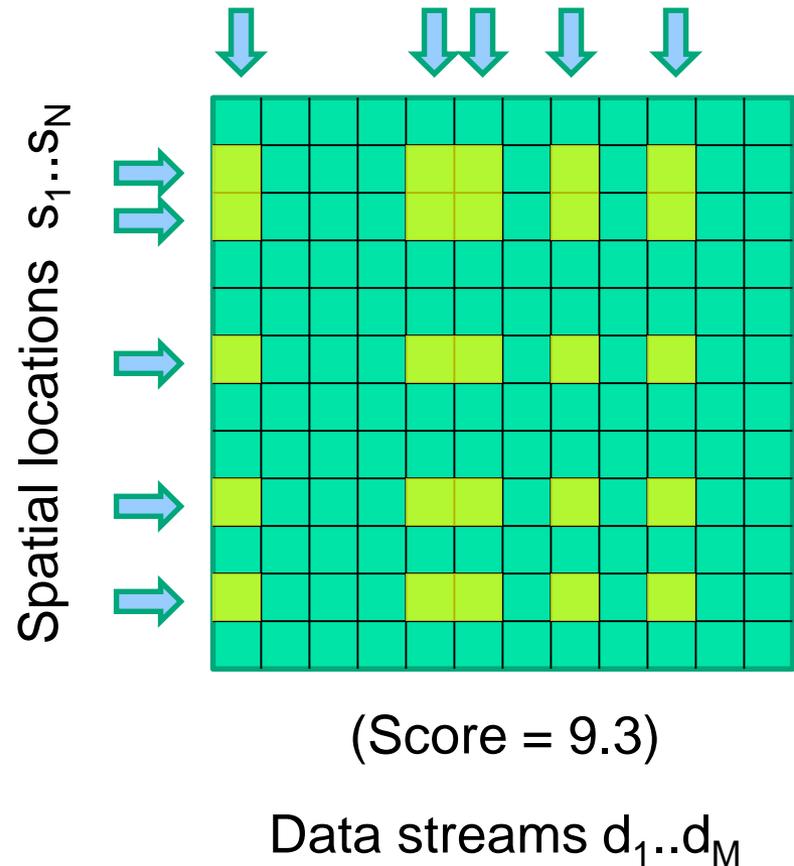
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

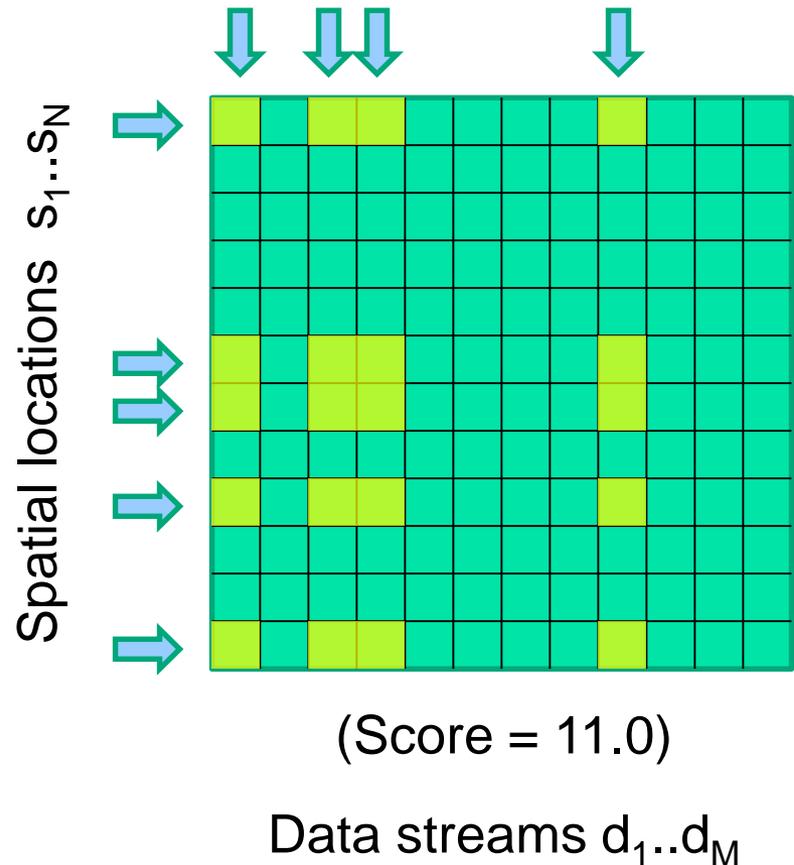
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!



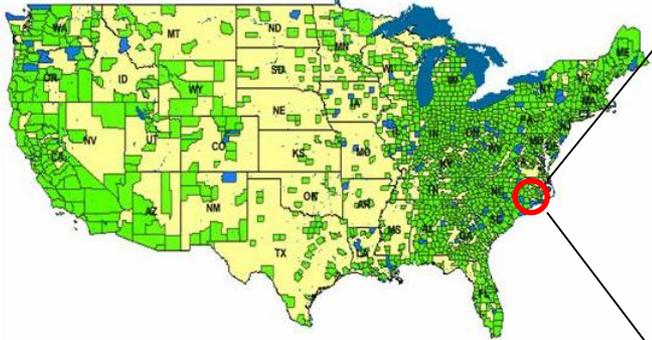
# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

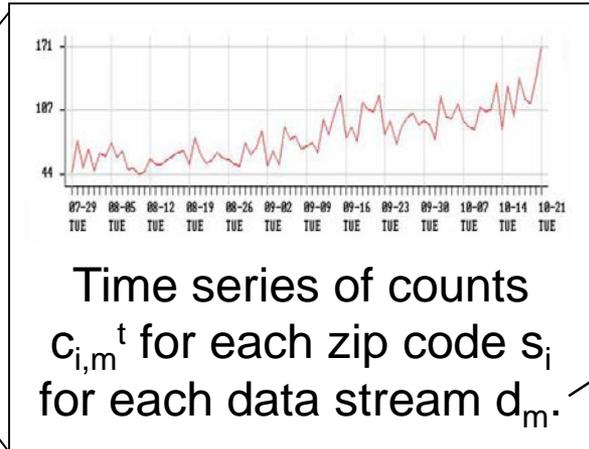
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!
- Converges to local maximum: we do multiple random restarts to approach the global maximum.



# Multivariate event detection



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



## Outbreak detection

- $d_1$  = 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, e.g., by identifying the affected streams.

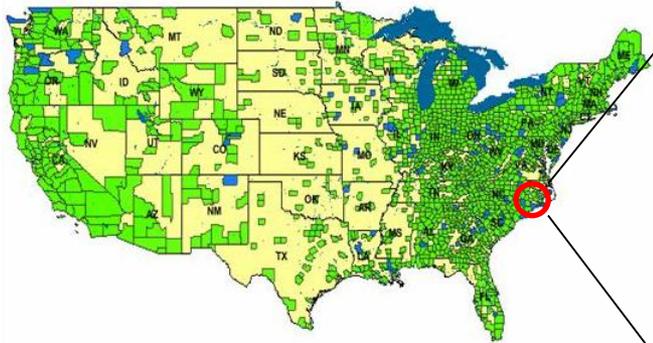
## Compare hypotheses:

$$H_1(D, S, W)$$

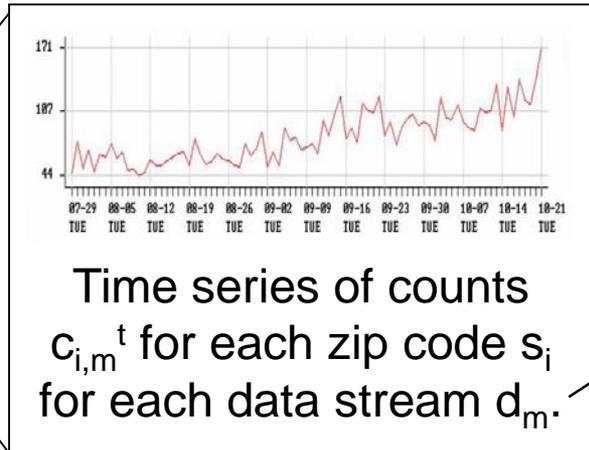
- $D$  = subset of streams
- $S$  = subset of locations
- $W$  = time duration

vs.  $H_0$ : no events occurring

# Multidimensional event detection



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



## Outbreak detection

- $d_1$  = 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

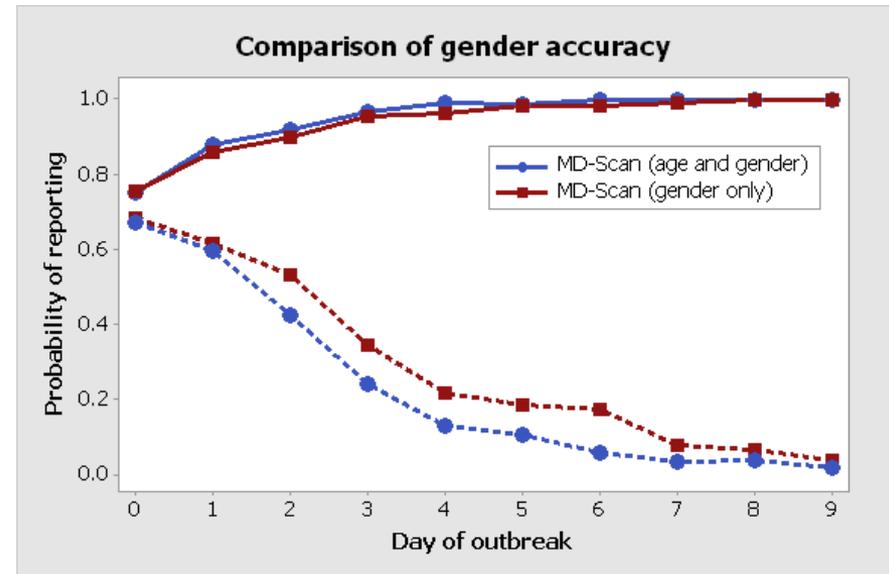
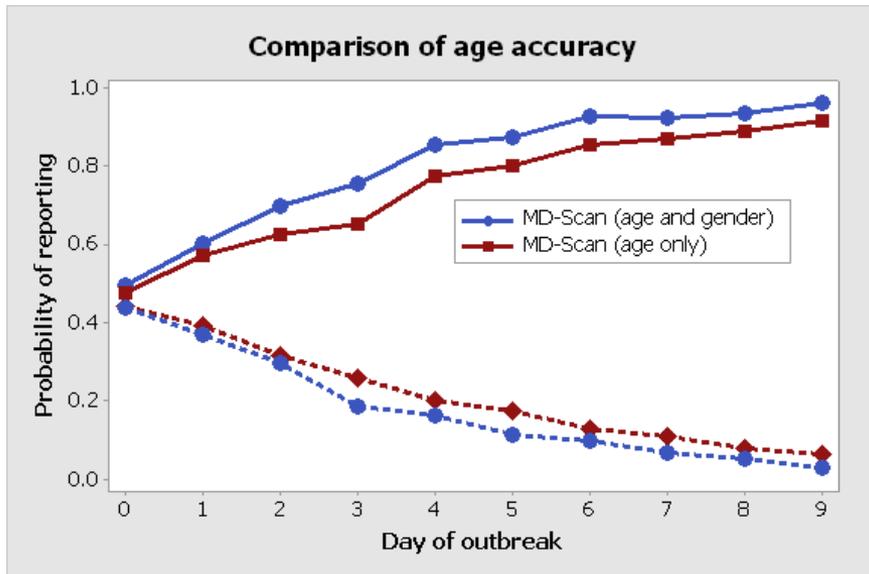
- Our **MD-Scan** approach (Neill and Kumar, 2013) extends LTSS to the multidimensional case:
  - For each time window and spatial neighborhood (center + k-nearest neighbors), we do the following:
    1. Start with randomly chosen subsets of **locations**  $S$ , **streams**  $D$ , and **values**  $V_j$  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.

# Empirical evaluation

- We first 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.

# 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.



**Proportions of correct and incorrect groups reported vs. time since start of outbreak.**

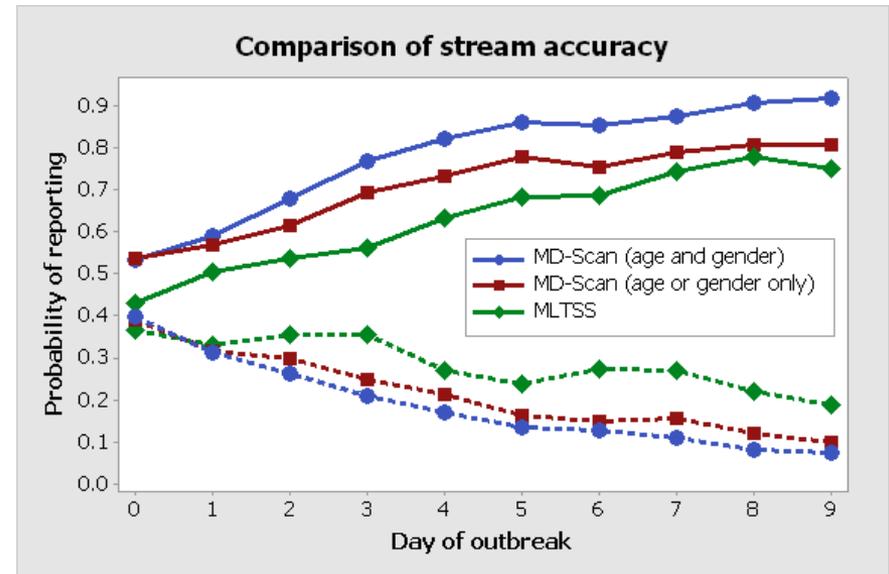
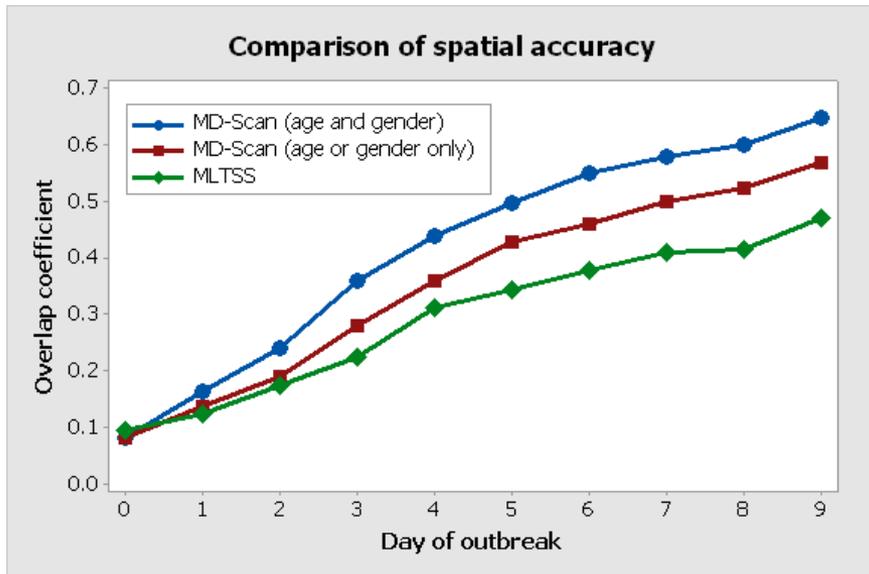
Solid lines: affected gender and/or age deciles. Dashed lines: unaffected.

Blue lines: outbreaks with differential effects by both age and gender (easier).

Red lines: outbreaks with differential effects by age or gender only (harder).

## 2) Characterizing affected streams

As compared to the previous state of the art (multivariate linear-time subset scanning), MD-Scan is better able to characterize the affected spatial locations and subset of the monitored streams.



**Left: overlap coefficient between true and detected subsets of spatial locations.**  
**Right: Proportions of correct and incorrect streams reported vs. day of outbreak.**

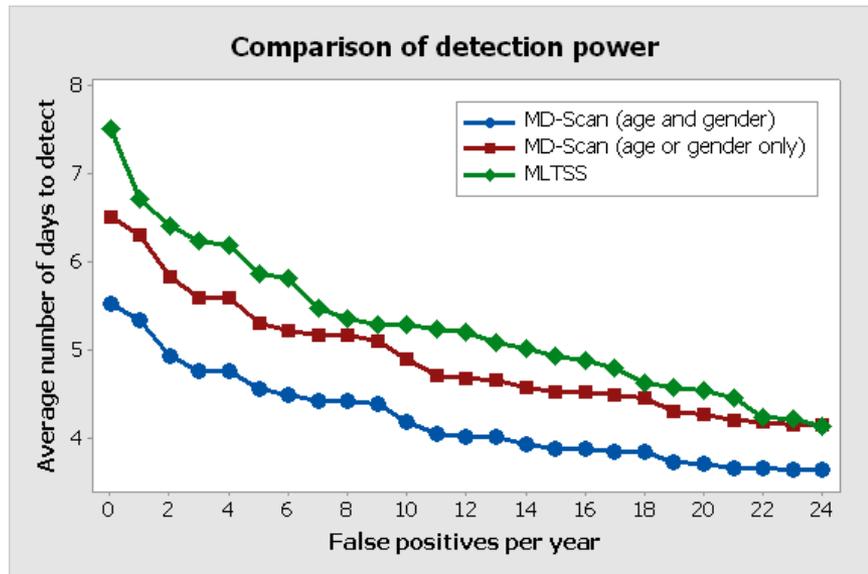
Blue lines: outbreaks with differential effects by both age and gender (easier).

Red lines: outbreaks with differential effects by age or gender only (harder).

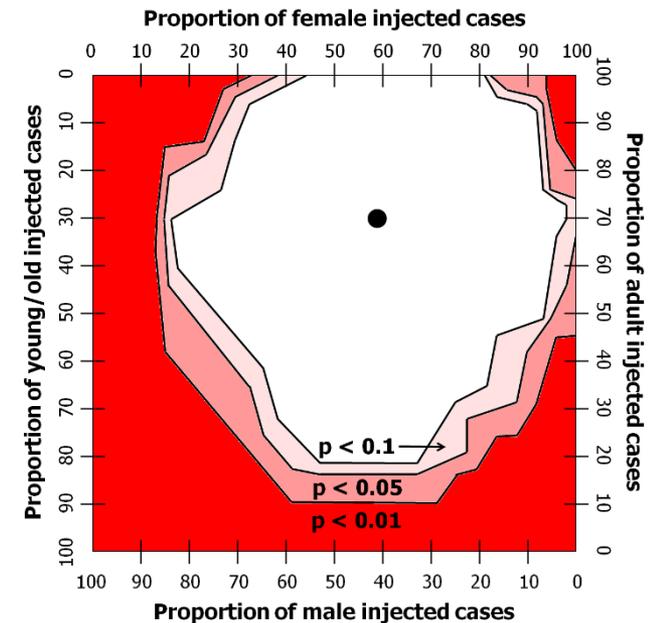
Green lines: MLTSS, ignoring age and gender information

# 3) Timeliness of outbreak detection

MD-Scan achieved significantly more timely detection for outbreaks that were sufficiently biased by age and/or gender.



For outbreaks with strong age and gender biases, time to detection improved from 5.2 to 4.0 days at a fixed false positive rate of 1/month.



Smaller biases in age or gender were sufficient for significant improvements; even when no age/gender signal is present, MD-Scan performs comparably to MLTSS.

# Allegheny County Overdose Data

- We analyzed county medical examiner data for fatal accidental drug overdoses, 2008-2015.
- ~2000 cases: for each overdose victim, we have date, location (zip), age, gender, race, and the set of drugs present in their system.
- Reduced to 30 dimensions (age decile, gender, race, presence/absence of 27 common drugs) plus space and time.
- Clusters discovered by MD-Scan were shared with Allegheny County's Department of Human Services; planned collaboration to build a prospective overdose surveillance tool.

# MD-Scan Overdose Results (1)



**Fentanyl** is a dangerous drug which has been a huge problem in western PA.

It is often mixed with white powder heroin, or sold disguised as heroin.

January 16-25, 2014:  
14 deaths county-wide  
from fentanyl-laced heroin.

March 27 to April 21, 2015:  
26 deaths county-wide from  
fentanyl, heroin only present in 11.

January 10 to February 7, 2015:

Cluster of 11 fentanyl-related deaths, mainly black males over 58 years of age, centered in Pittsburgh's downtown Hill District.

Very unusual demographic:  
common dealer / shooting gallery?

Started in the SE suburbs of Pittsburgh, including a cluster of 5 cases around McKeesport between March 27 and April 8.

Cluster score became significant March 29<sup>th</sup> (4 nearby cases, white males ages 20-49) and continued to increase through April 20<sup>th</sup>.

Fentanyl, heroin, and combined deaths remained high through end of June (>100).

# MD-Scan Overdose Results (2)

Another set of discovered overdose clusters each involved a combination of Methadone and Xanax.



The combination produces a strong high but can be deadly (~30% of methadone fatal ODs).

From 2008-2012: multiple M&X OD clusters, 3-7 cases each, localized in space and time.

Methadone: an opioid used for chronic pain relief and to treat heroin addiction, but also addictive and risk of OD.

From 2013-2015: no M&X overdose clusters; 33% and 47% drops in yearly methadone and M&X deaths respectively.



Xanax (alprazolam): a benzodiazepine prescribed for panic and anxiety disorders.

Why did these deaths cluster, when methadone and methadone + other benzo deaths did not?

What factors could explain the dramatic reduction in M&X overdose clusters?

# MD-Scan Overdose Results (2)

Another set of discovered overdose clusters each involved a combination of Methadone and Xanax.



Methadone: an opioid used for chronic pain relief and to treat heroin addiction, but also addictive and risk of OD.



Xanax (alprazolam): a benzodiazepine prescribed for panic and anxiety disorders.

Increased state oversight of methadone clinics and prescribing physicians after passage of the Methadone Death and Incident Review Act (Oct 2012).

Approval of generic suboxone (buprenorphine + naloxone) in early 2013 lowered cost of suboxone treatment as an alternative to methadone clinics.

Why did these deaths cluster, when methadone and methadone + other benzo deaths did not?

What factors could explain the dramatic reduction in M&X overdose clusters?

## Today's talk:

- Public health surveillance
  - Early outbreak detection (fast subset scan)
  - Accidental drug overdose surveillance (multidimensional scan)
  - “Novel” outbreak detection (semantic scan)



# Asyndromic surveillance

<u>Date/time</u>	<u>Hosp.</u>	<u>Age</u>	<u>Complaint</u>
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp

1 year of free-text ED chief complaint data from 3 hospitals in North Carolina.

Key challenge: A syndrome cannot be created to identify every possible cluster of potential public health significance.

A method is needed to identify relevant clusters of disease cases **without** pre-classification into syndromes.

Use case proposed by NC DOH and NYC DOHMH, solution requirements developed through a public health consultancy at the International Society for Disease Surveillance.

# From structured to unstructured...

nose caught in door

nausea  
vomiting

rabies shot

Each ED case does not just contain structured information, but also free text: the patient's **chief complaint**.

Q: How can we use this **unstructured** data to enhance detection?

n v d

Possible approach: map ED cases to broad syndrome categories ("prodromes") and do a **multidimensional scan**.

tired weak

food  
poisoning

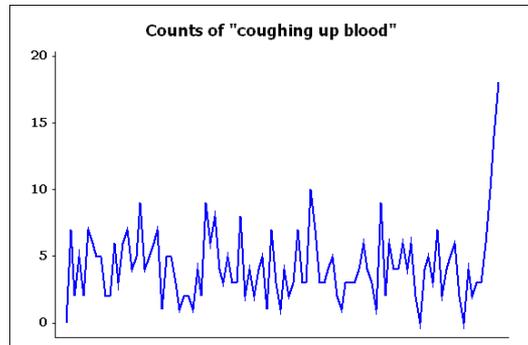
diarrhea

fever

# 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).

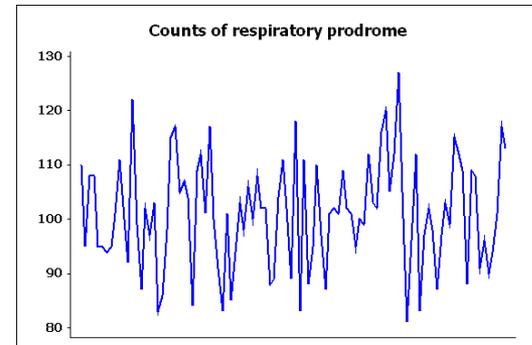
If we were monitoring these particular symptoms, it would only take a few such cases to realize that an outbreak is occurring!



What happens when something new and scary comes along?

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

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

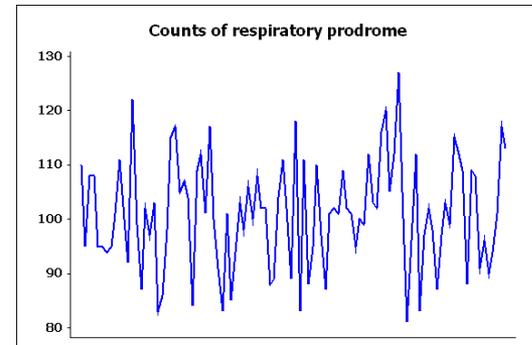
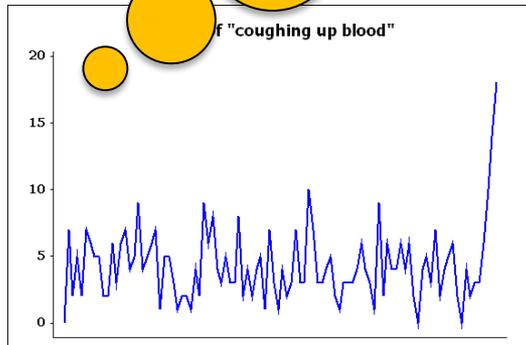


# Where do existing methods fail?

The typical, prodromal phase of an outbreak is often something that is not statistically significant. How can we detect something along? Existing methods (e.g., topic modeling) often fail to detect emerging patterns of keywords.

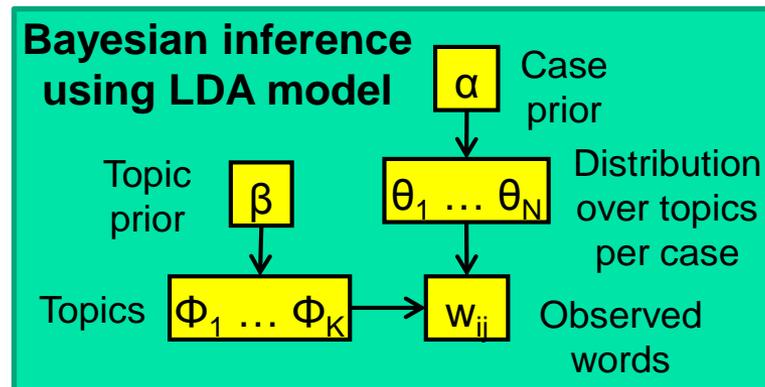
Our solution is to combine text-based (topic modeling) and event detection (multidimensional scan) approaches, to detect **emerging patterns of keywords.**

If we were to monitor a particular symptom category, we might take a few such as "coughing up blood" to estimate the outbreak signal, that an outbreak is occurring! Existing methods are often failing or preventing detection.



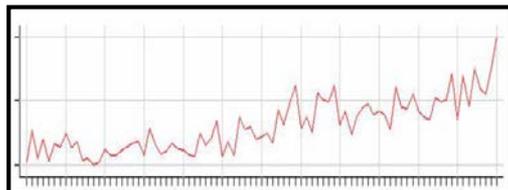
# The semantic scan statistic

<u>Date/time</u>	<u>Hosp.</u>	<u>Age</u>	<u>Complaint</u>
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp



$\phi_1$ : vomiting, nausea, diarrhea, ...  
 $\phi_2$ : dizzy, lightheaded, weak, ...  
 $\phi_3$ : cough, throat, sore, ...

Classify cases to topics



Time series of hourly counts for each combination of hospital and age group, for each topic  $\phi_j$ .

Now we can do a multidimensional scan, using the learned topics instead of pre-specified prodromes!

# Multidimensional scanning

(for learned topics)

For each hour of data (~8K):

For each combination  $S$  of:

- Hospital
- Time duration (1-3 hours)
- Age range
- **Topic**

**Count:**  $C(S)$  = # of cases in that time interval matching on hospital, age range, topic.

**Baseline:**  $B(S)$  = expected count (28-day moving average).

**Score:**  $F(S) = C \log (C/B) + B - C$ , if  $C > B$ , and 0 otherwise (using the expectation-based Poisson likelihood ratio statistic)

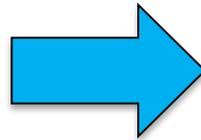
We return cases corresponding to each top-scoring subset  $S$ .

# Semantic scan results (1)

Semantic scan detected simulated novel outbreaks **more than twice as quickly** as the standard prodrome-based method: 5.3 days vs. 10.9 days to detect at 1 false positive per month.



Simulated novel outbreak: “green nose”



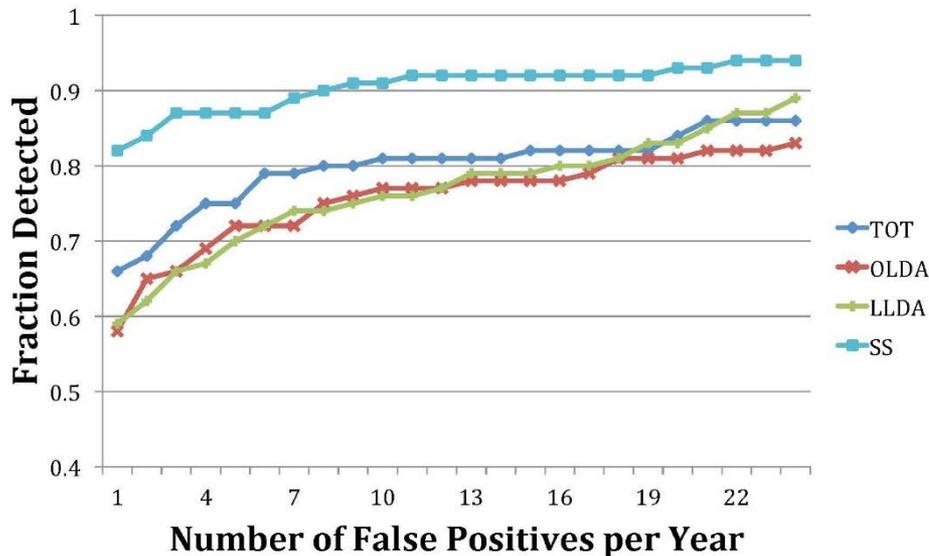
green  
nose  
possible  
color  
greenish  
nasal  
...

Top words from detected topic

# Semantic scan results (2)

Using a “leave one out” approach in which we hold out one International Classification of Diseases (ICD) code and inject cases as if from a novel outbreak, we observe huge improvements in detection power and accuracy vs. competing methods (Online LDA, Topic Over Time, Labeled LDA).

These gains resulted from development of a new **contrastive topic modeling** approach with higher power to detect newly emerging topics.



- 1) Learning a set of “background” topics from historical data.
- 2) Learning a set of “foreground” topics from recent data.
- 3) Combined LDA inference, holding the background topics constant, leads to discovery of foreground topics that are maximally different.

# Semantic scan results (3)

We used a combined multidimensional and semantic scan on datasets provided by the North Carolina DOH, with simulated novel outbreaks of interest injected by the NC DETECT group, and New York City's DOHMH.

We identified clusters of cases referring to specific locations, unusual sets of symptoms, or affected subpopulations. Here are some highlights:

Location and symptoms:

“sudden onset of rashes  
at the beach”

Ten cases that mentioned  
a local middle school  
within a four-hour span

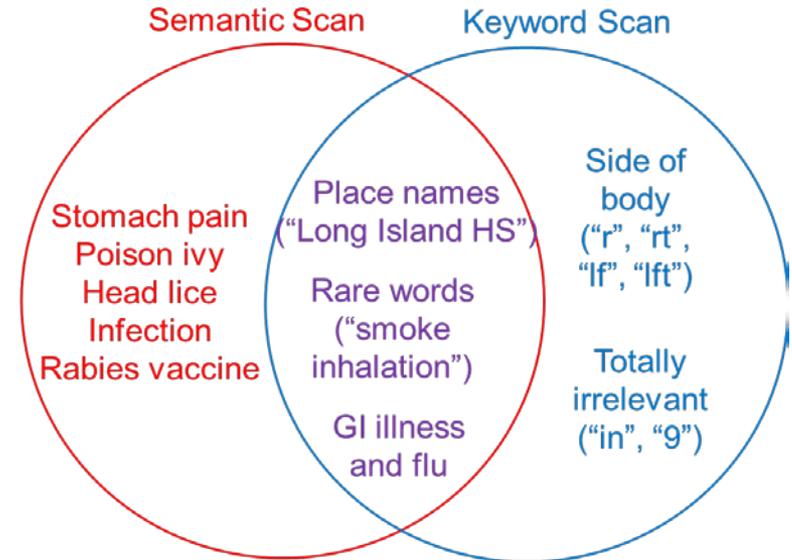
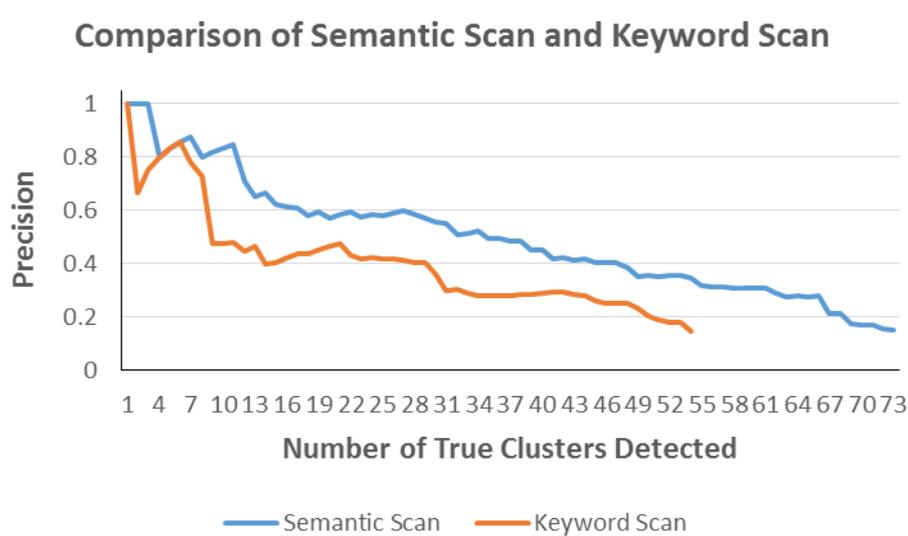
Clusters with related chief complaints:  
chemical spill, motor vehicle accidents,  
contagious diseases (head lice, scabies)

Specific subpopulations:

Seven young adults  
suffering from smoke  
inhalation

# Semantic scan results (4)

We compared the top 500 clusters found by the semantic scan and a keyword-based scan on the NC DOH data in a blinded evaluation, with public health officials labeling each cluster as “relevant” or “not relevant”.



Semantic scan: for 10 true clusters, had to report 12;  
for 30 true clusters, had to report 54.

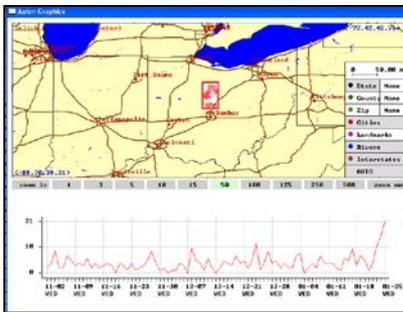
Keyword scan: for 10 true clusters, had to report 21;  
for 30 true clusters, had to report 83.

# Conclusions

The continually increasing **size**, **variety**, and **complexity** of data available for population health and disease surveillance necessitate the development of new detection methods to make use of these data.

**Fast subset scanning** (with constraints) can serve as a fundamental building block for efficient, scalable pattern detection in massive data.

Extensions to the **multivariate** and **multidimensional** settings, and incorporation of novel **topic modeling** approaches to handle free text data, enable us to address a variety of public health challenges.



Early outbreak detection



Drug overdose surveillance



Discovery of novel outbreaks

# Acknowledgements

- Event and Pattern Detection Laboratory (current members and alumni):  
<http://epdlab.heinz.cmu.edu/people>
- Students, postdocs, and collaborators:  
Abhinav Maurya, Skyler Speakman, Ed McFowland, Sriram Somanchi, Tarun Kumar, Kenton Murray, Yandong Liu, Chris Dyer.
- Funding support: NSF, MacArthur Foundation, Richard King Mellon Foundation.

# References

- 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.
- D.B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. *Statistics in Medicine* 32: 2185-2208, 2013.
- 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.
- 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*, 2014.
- S. Speakman, S. Somanchi, E. McFowland III, D.B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 2016, in press.
- T. Kumar and D.B. Neill. Fast multidimensional subset scan for event detection and characterization. Revised version in preparation.
- A. Maurya, D.B. Neill, et al., Semantic scan: detecting subtle, spatially localized events in text streams. Under review for KDD 2016.



**Thanks for listening!**

**More details on our web site:**

**<http://epdlab.heinz.cmu.edu>**

**Or e-mail me at:**

**[neill@cs.cmu.edu](mailto:neill@cs.cmu.edu)**