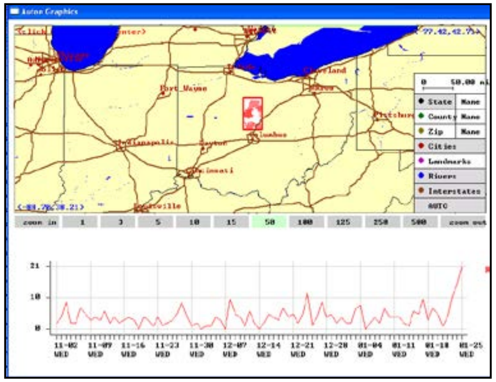
Machine Learning for Population Health and Disease Surveillance

Daniel B. Neill, Ph.D.
Center for Urban Science and Progress
New York University
E-mail: daniel.neill@nyu.edu

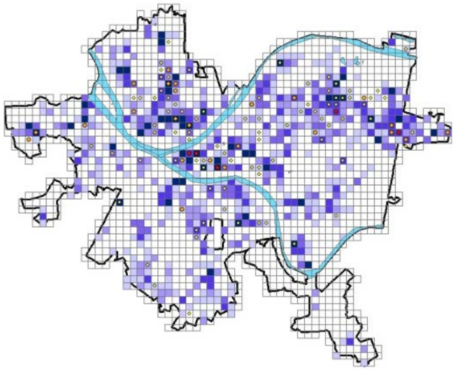
We gratefully acknowledge funding support from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330, MacArthur Foundation, and Richard King Mellon Foundation.

Daniel B. Neill, Ph.D.
Associate Professor of Computer Science (NYU Courant),
Public Service (NYU Wagner), and Urban Analytics (NYU CUSP).
E-mail: daniel.neill@nyu.edu



Population Health: Very early and accurate detection of emerging outbreaks.

Individual Health: Discovering novel "best practices" for patient care, to improve outcomes and reduce costs.

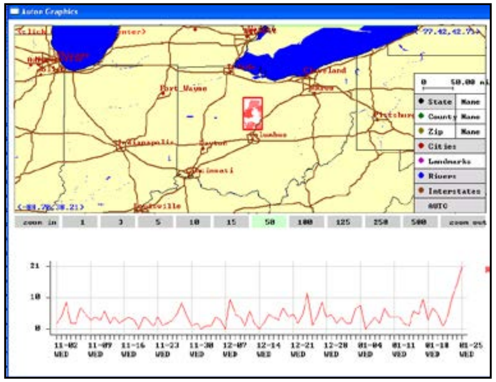


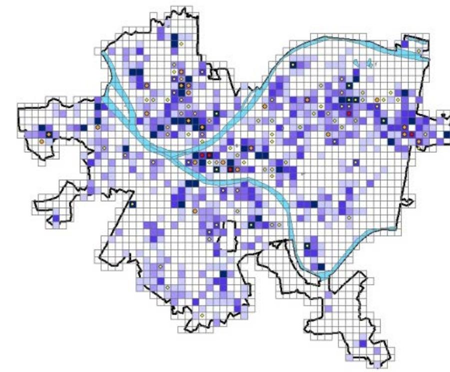
Community Health: Detection, prediction, and prevention of "hot-spots" of violent crime.

My research is focused at the intersection of **machine learning (ML)** and **public policy**, with two main goals:

1) Develop new ML methods for better (more scalable and accurate) **detection** and **prediction** of events and other patterns in massive datasets.
2) Apply these methods for the public good, particularly, to improve the quality of **individual**, **population**, and **community health**.

Daniel B. Neill, Ph.D.
Associate Professor of Computer Science (NYU Courant),
Public Service (NYU Wagner), and Urban Analytics (NYU CUSP).
E-mail: daniel.neill@nyu.edu





Population Health: Very early and accurate detection of emerging outbreaks.

Individual Health: Discovering novel "best practices" for patient care, to improve outcomes and reduce costs.

Community Health: Detection, prediction, and prevention of "hot-spots" of violent crime.

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

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.

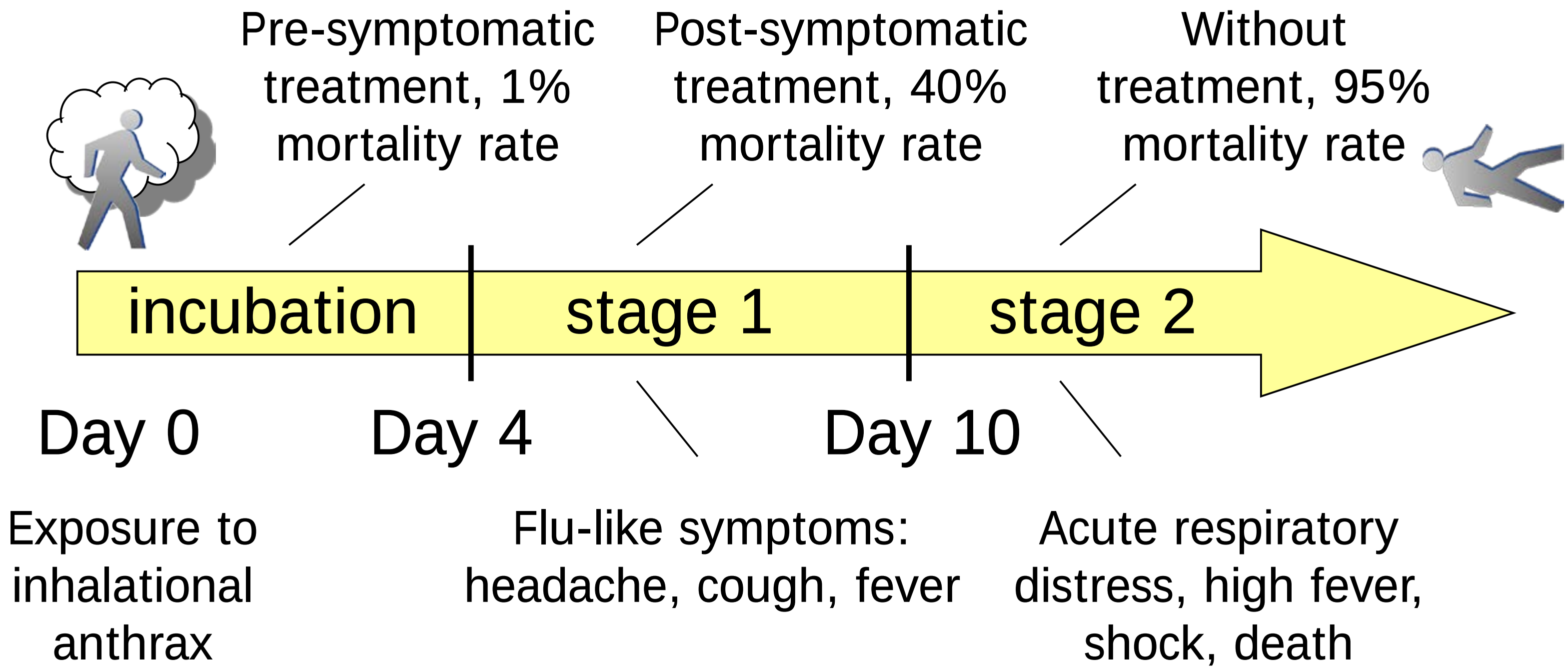
- Emerging infectious diseases

  "Conservative estimate" of 2-7 million deaths from pandemic avian influenza.

- Better response to common outbreaks and emerging public health trends.

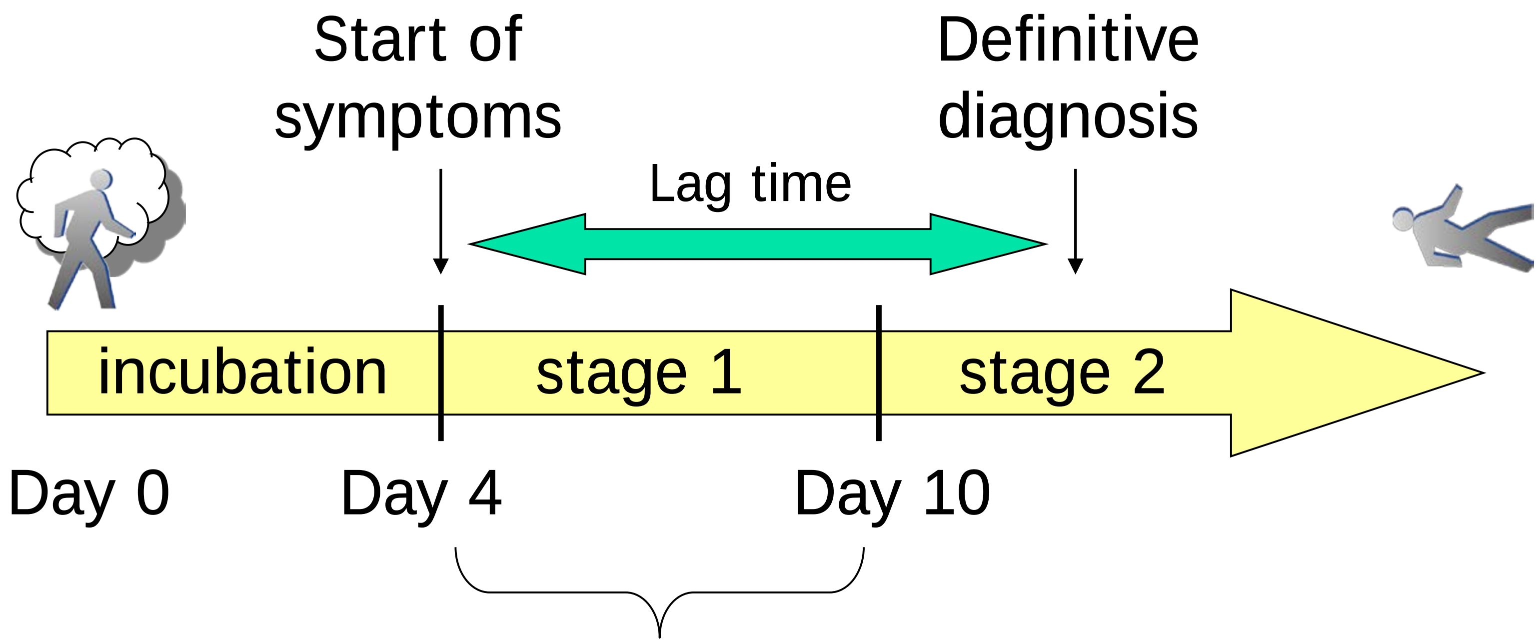
Benefits of early detection

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



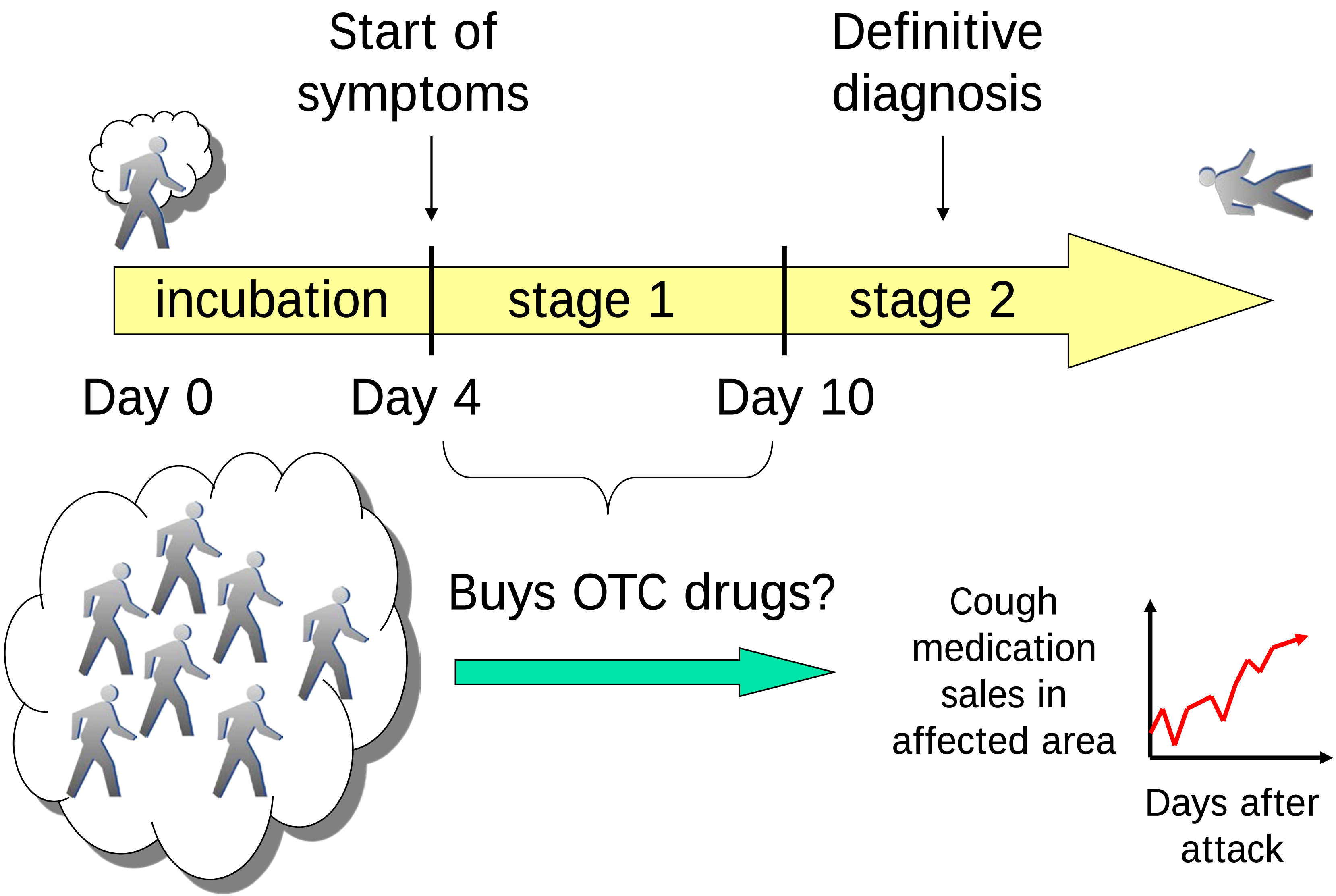
Buys OTC drugs

Skips work/school

Uses Google, Facebook, Twitter

Visits doctor/hospital/ED

Syndromic surveillance



Syndromic surveillance

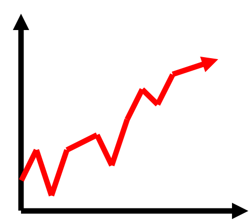
Start of symptoms

Definitive diagnosis

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

Buys OTC drugs?

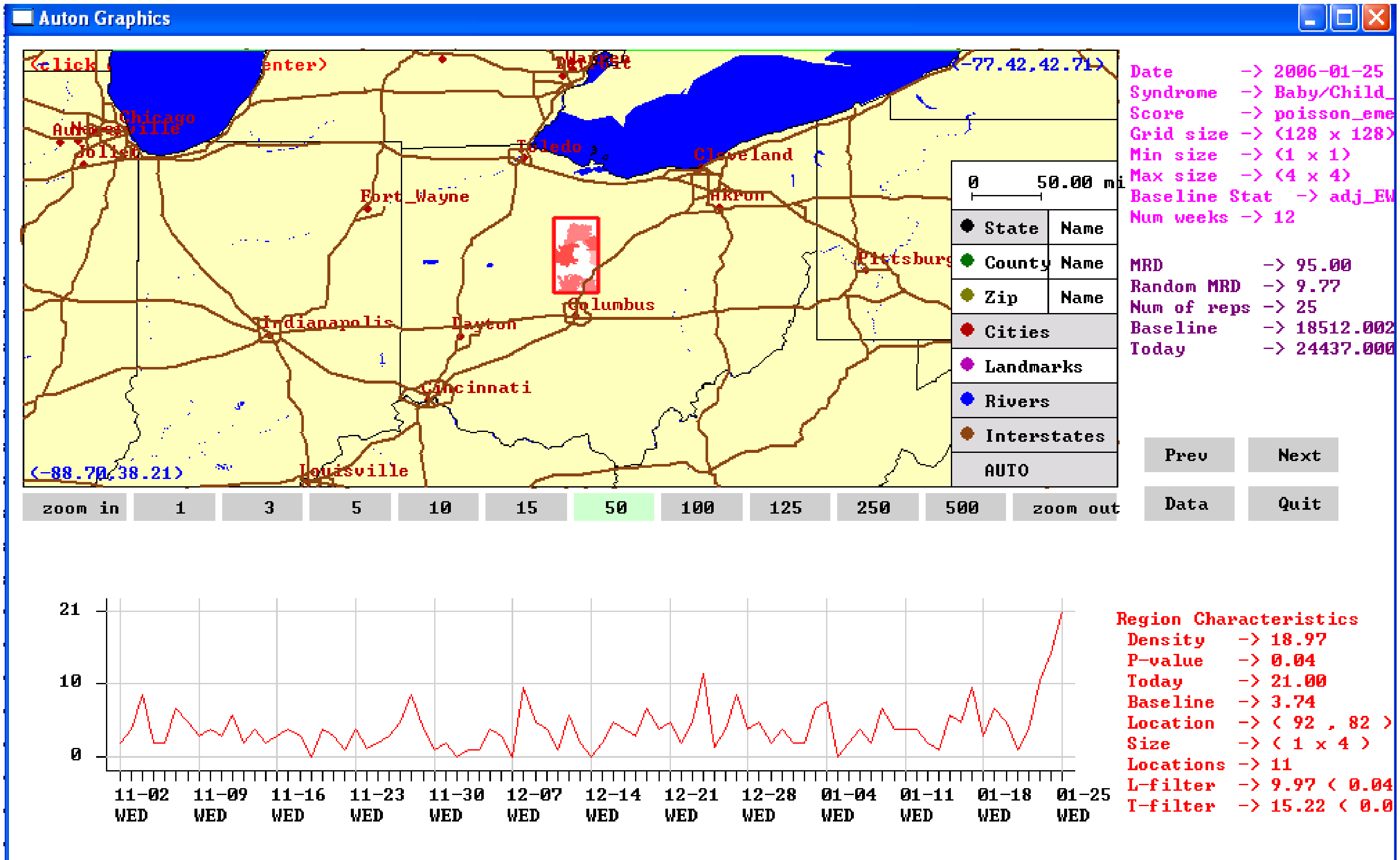
Cough medication sales in affected area



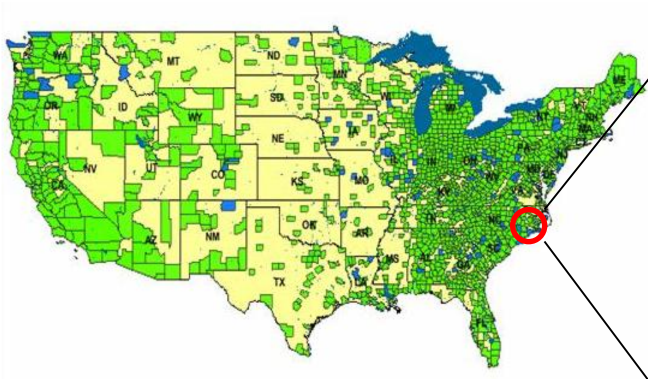
Days after attack

Outbreak detection example

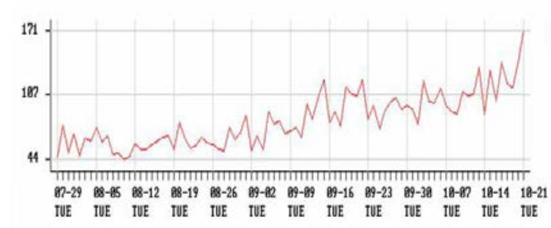
Spike in sales of pediatric electrolytes near Columbus, Ohio



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:

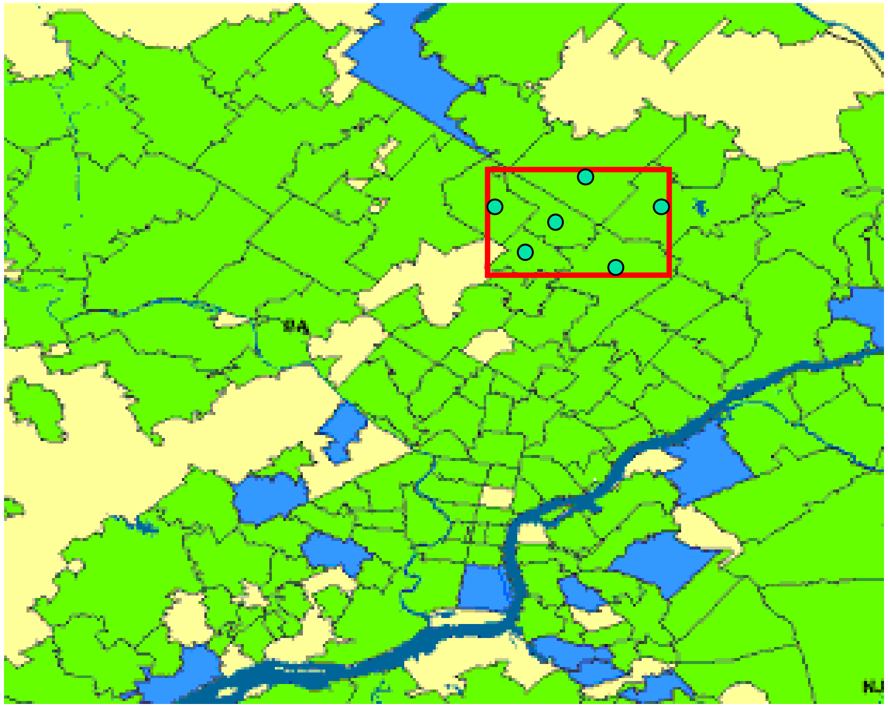
$H_1(D, S, W)$

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

vs. H_0: no events occurring

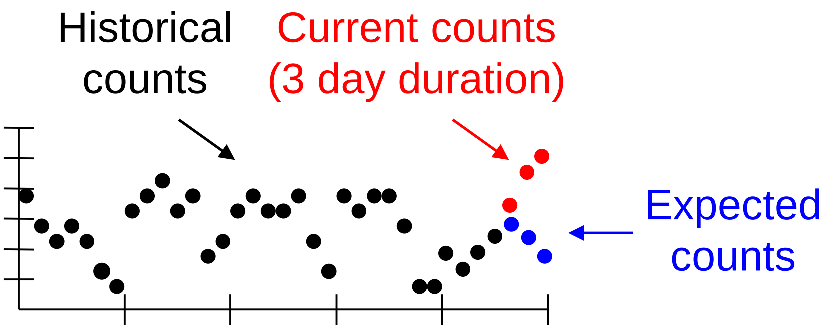
Expectation-based scan statistics

(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.

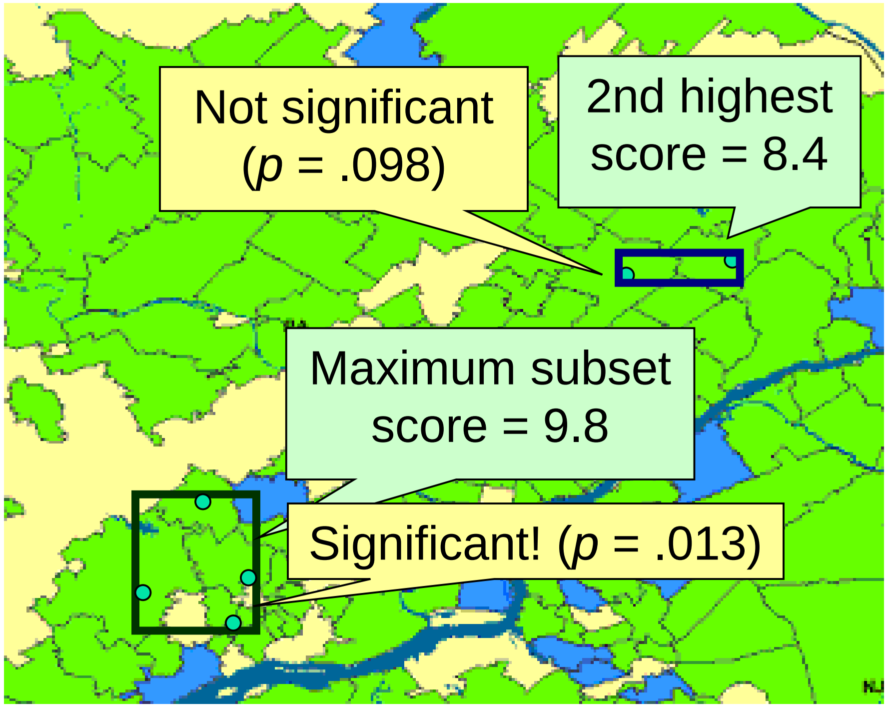
We perform **time series analysis** to compute expected counts ("baselines") for each location and stream for each recent day.



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)

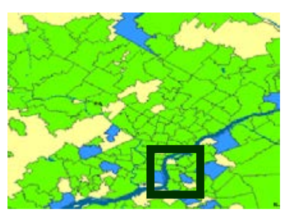


We find the subsets with highest values of a likelihood ratio statistic, and compute the *p*-value of each subset by randomization testing.

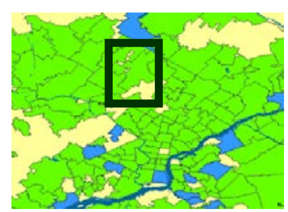
$$F(D,S,W) = \frac{\Pr(\text{Data} \mid H_1(D,S,W))}{\Pr(\text{Data} \mid H_0)}$$

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

$F_1^* = 2.4$

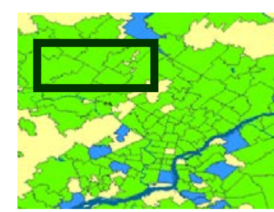


$F_2^* = 9.1$



…

$F_{999}^* = 7.0$



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.

<u>Expectation-based Poisson</u>

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

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

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

Maximum likelihood: $q = C / B$.

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

<u>Expectation-based Gaussian</u>

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

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

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

Maximum likelihood: $q = C' / B'$.

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

<u>Many possibilities</u>: 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 $s_{(1)} \ldots s_{(N)}$, we can **prove** that the top-scoring subset F(S) consists of the locations $s_{(1)} \ldots s_{(k)}$ for some k, $1 \leq k \leq N$.
  - <u>Key step</u>: if there exists some location $s_{out} \notin S$ with higher priority than some location $s_{in} \in S$, then we can show that $F(S) \leq \max(F(S \cup \{s_{out}\}), F(S \setminus \{s_{in}\}))$.
- <u>Theorem</u>: LTSS holds for expectation-based scan statistics in any exponential family. (Speakman et al., 2015)

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$

$$H_0 : x_i \sim Dist(\mu_i)$$

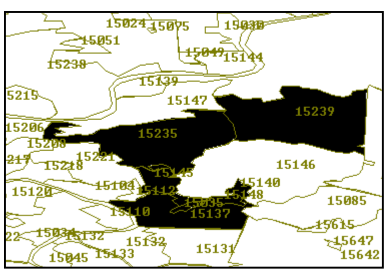
$$H_1 : x_i \sim Dist(q\mu_i)$$

Constrained fast subset scanning

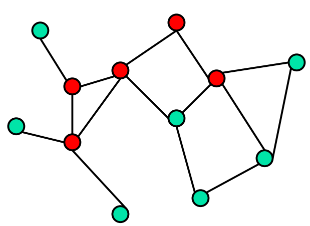
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the "best unconstrained subset" problem, and cannot be used directly for <u>constrained</u> optimization.

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.

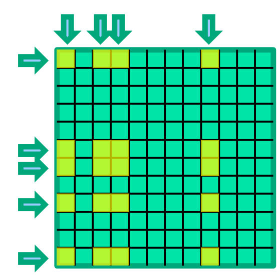
+ Proximity constraints → Fast spatial scan (irregular regions)
+ Multiple data streams → Fast multivariate scan
+ Connectivity constraints → Fast graph scan
+ Group self-similarity → Fast generalized subset scan



(Neill, *JRSS-B,* 2012)
(Neill et al., *Stat. Med.,* 2013)



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



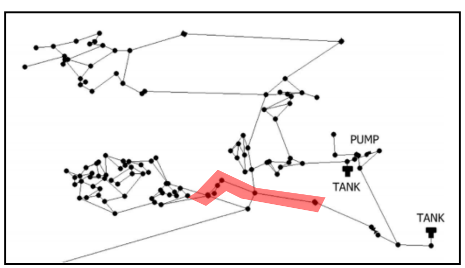
(McFowland et al., *JMLR*, 2013)

Constrained fast subset scanning

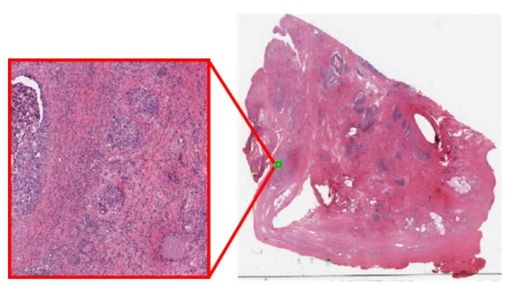
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

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)



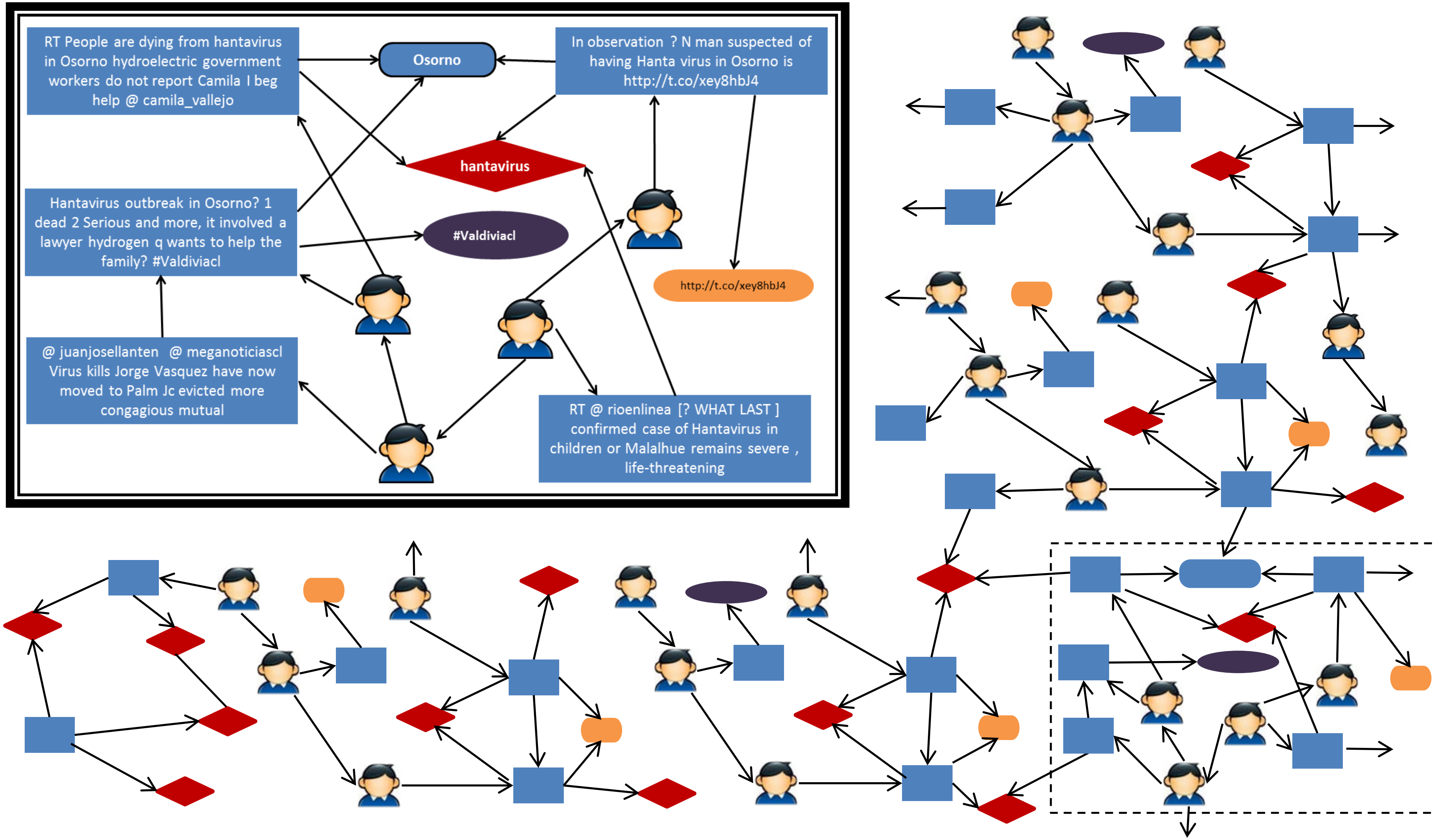
(Somanchi & Neill, DMHI 2013)

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

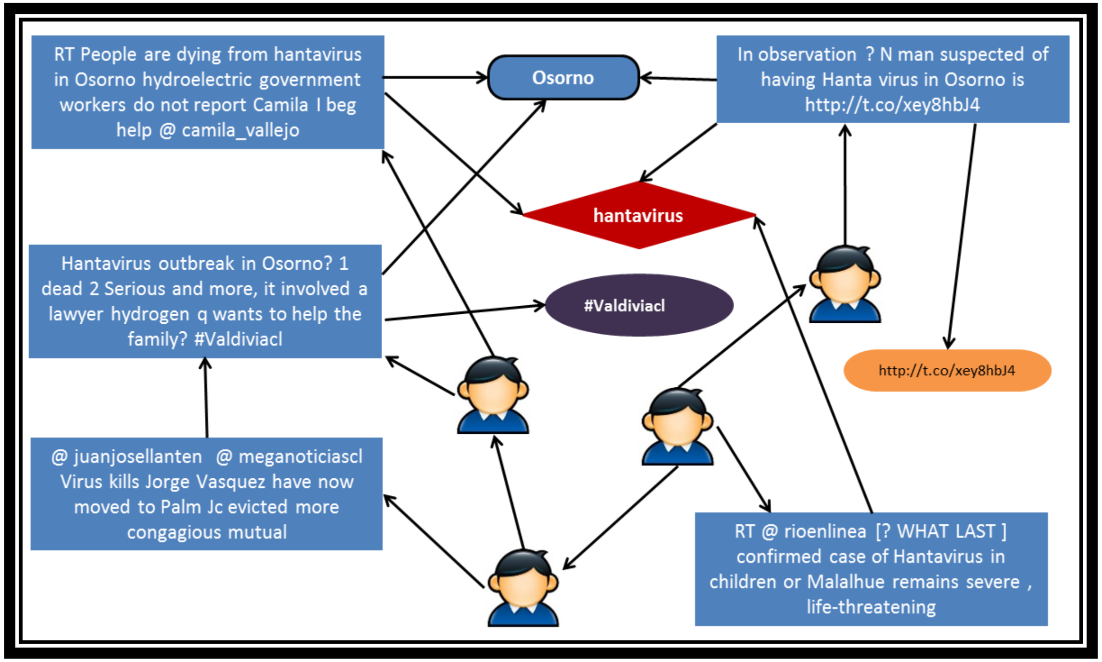
Fast subset scan with spatial proximity constraints

- 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.

Teaser #1: Detecting Rare Disease Outbreaks Using Twitter



Teaser #1: Detecting Rare Disease Outbreaks Using Twitter



Technical contributions:

- Modeling of Twitter data as a heterogeneous sensor network.
- Nonparametric subset scan to integrate data from multiple node types.
- Fast search algorithm scales to massive data.

Evaluation: 17 gold standard hantavirus outbreaks in Chile. Also applied to community health: predicting civil unrest and detecting emerging patterns of human rights violations.

Results: outperforms existing state-of-the-art methods with respect to timeliness of detection, detection power, and accuracy of outbreak characterization.

Detected Hantavirus outbreak, 10 Jan 2013

First news report: 11 Jan 2013

Temuco and Villarrica, Chile

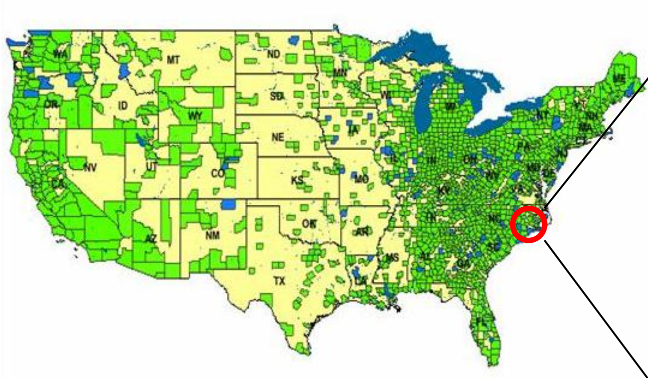
Locations
Users
Keywords
Hashtags
Links
Videos

F. Chen and D.B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. Proc. 20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 1166-1175, 2014.

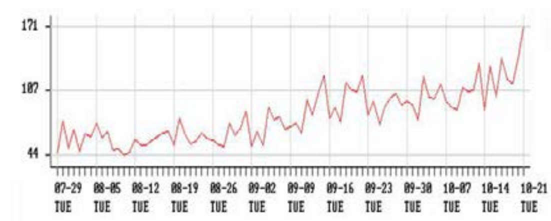
Today's talk:

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

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.

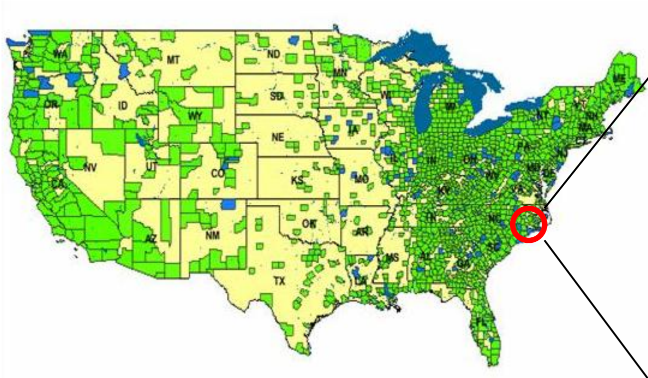
Compare hypotheses:

$H_1(D, S, W)$

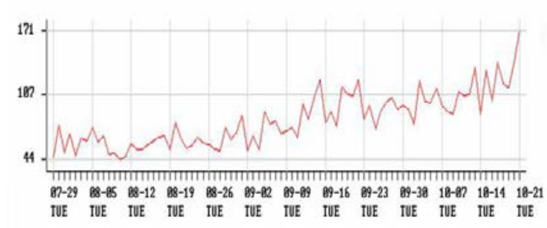
D = subset of streams
S = subset of locations
W = time duration

vs. H_0: no events occurring

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.

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.

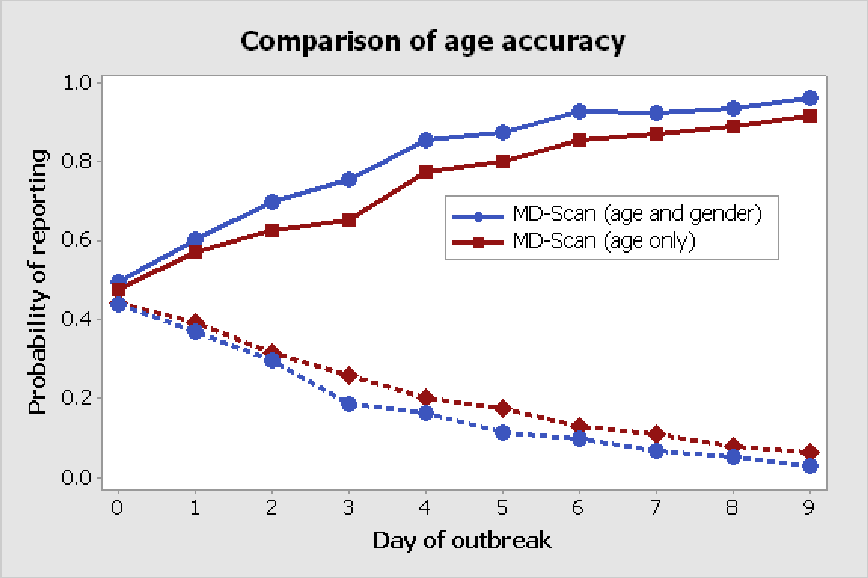
3. Iterate step 2 until convergence to a local maximum of the score function $F(D,S,W,\{V_j\})$, and use multiple restarts to approach the global maximum.

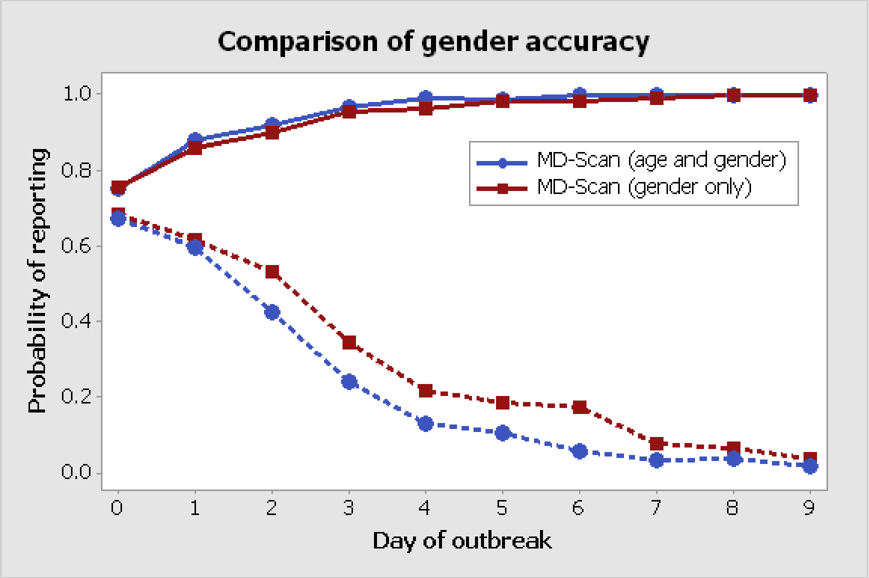
Empirical evaluation

- 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.

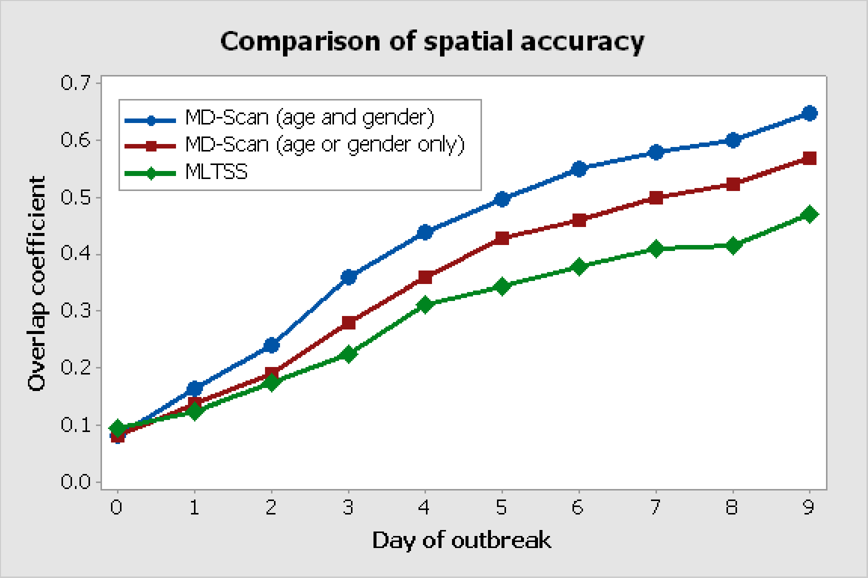


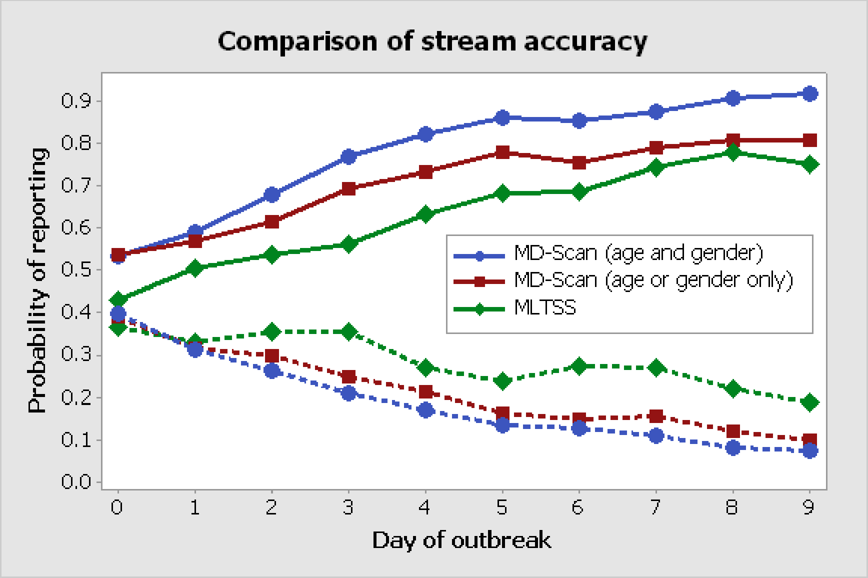


Proportions of correct and incorrect groups reported vs. time since start of outbreak.
Solid lines: affected gender and/or age deciles. Dashed lines: unaffected.
Blue lines: outbreaks with differential effects by both age and gender (easier).
Red lines: outbreaks with differential effects by age or gender 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.

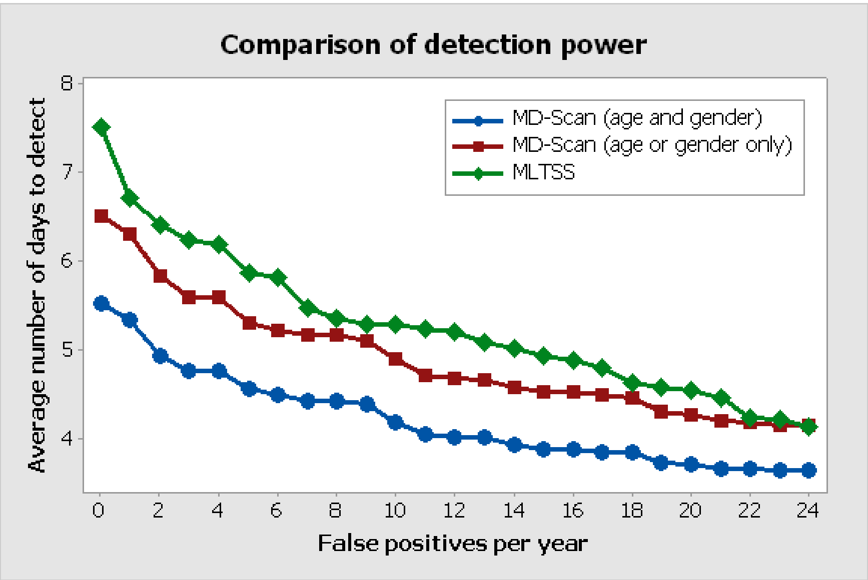


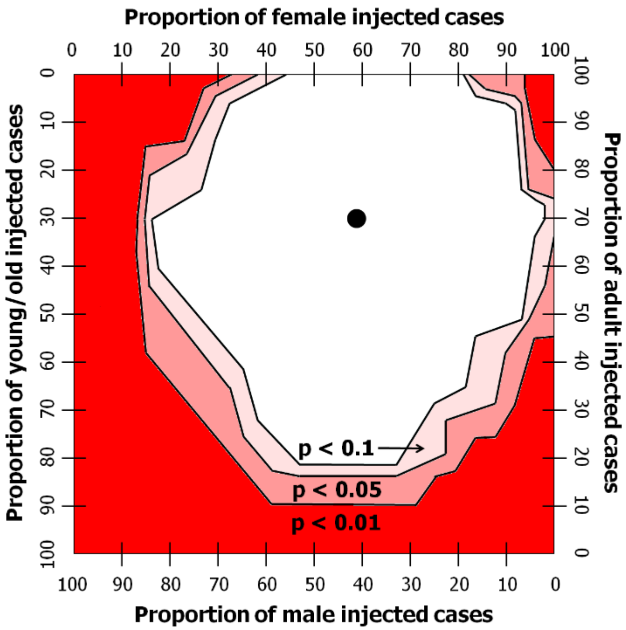


Left: overlap coefficient between true and detected subsets of spatial locations.
Right: Proportions of correct and incorrect streams reported vs. day of outbreak.
Blue lines: outbreaks with differential effects by both age and gender (easier).
Red lines: outbreaks with differential effects by age or 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

- Collaboration with Allegheny County's Department of Human Services, including retrospective analysis (2008-2015) and plans to build a prospective surveillance tool.
- ~2000 cases: for each overdose victim, we have date, zip code, age, gender, race, and which drugs present.
- We used an extension of MD-Scan, the **multi-dimensional tensor scan (MDTS)**, to detect emerging geographic, demographic, and behavioral patterns, many of which DHS had not previously identified.
  - Earlier detection of emerging overdose clusters.
  - Better characterization of **who** and **where** is affected by identifying affected subset in each dimension.
  - Quantifying the effects of drug legislation and other policy changes.

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 29th (4 nearby cases, white males ages 20-49) and continued to increase through April 20th.

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.

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.

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.

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

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?

Teaser #2: Discovering Anomalous Patterns of Care

Extensions of the multidimensional subset scan can be used to discover **heterogeneous treatment effects** in both experimental and observational data.

Using Highmark claims data from ~125K patients with diseases of the circulatory system, we discovered patterns including the following:

Glucocorticoids significantly increase mean number of hospitalizations following treatment in the subpopulation of hypertensive, overweight/obese males with endocrine disorders.

Regression on held-out data, controlling for observed covariates:

Glucocorticoids are associated with:
10.6% increase in hospitalizations across the entire population.
50.6% increase in hospitalizations for this subpopulation.

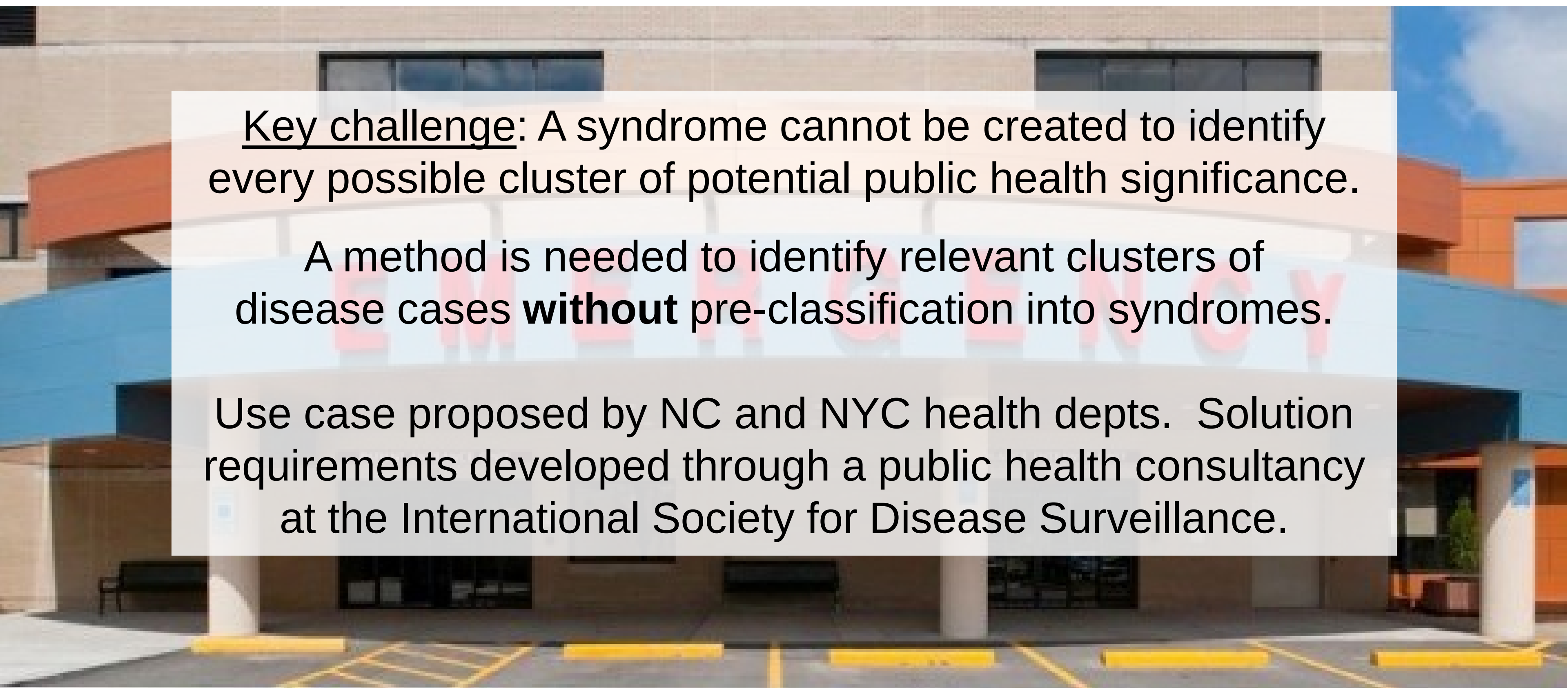
Today's talk:

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

Asyndromic surveillance

Date/time	Hosp.	Age	Complaint
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

Free-text ED chief complaint data from hospitals in New York City, North Carolina, and Allegheny County, Pennsylvania.



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 and NYC health depts. Solution requirements developed through a public health consultancy at the International Society for Disease Surveillance.

From structured to unstructured…

nose caught in door

rabies shot

a fib

tired weak

diarrhea

nausea
vomiting

n v d

food
poisoning

fever

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?

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

Where do existing methods fail?

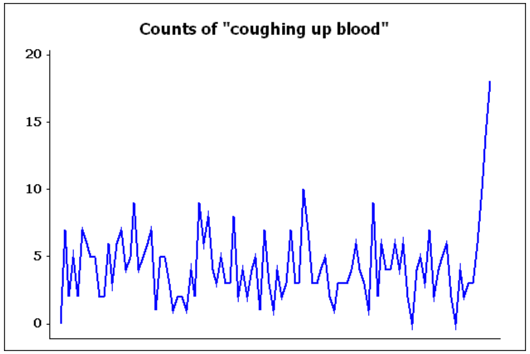
The typical, prodrome-based scan statistic approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

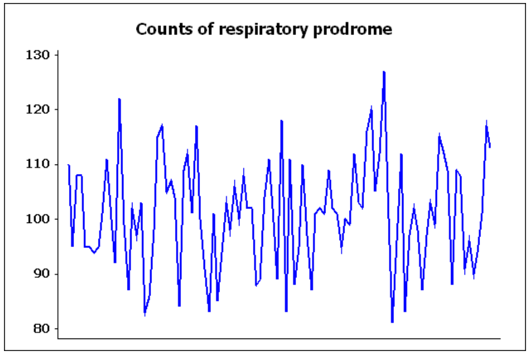
What happens when something new and scary comes along?

- **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.





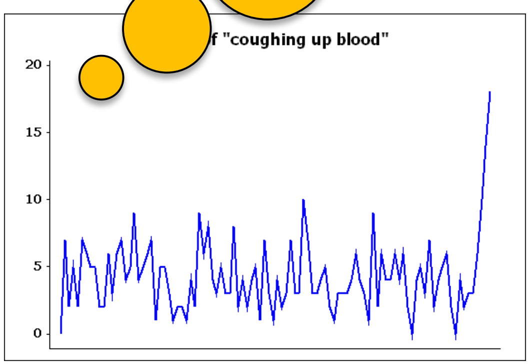
Where do existing methods fail?

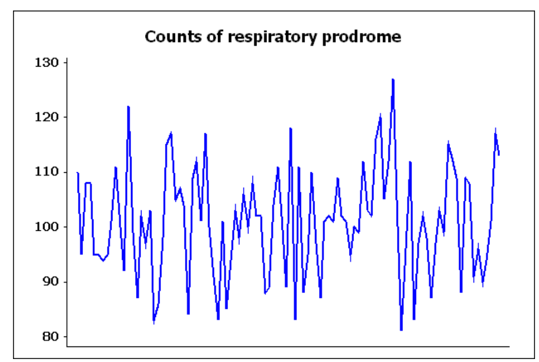
The typical, prodr... a...
scan statisti...
effectively...
outbrea...
seen...
sy...

...en something
...along?
...toms
...d”)
...n
...off”)

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

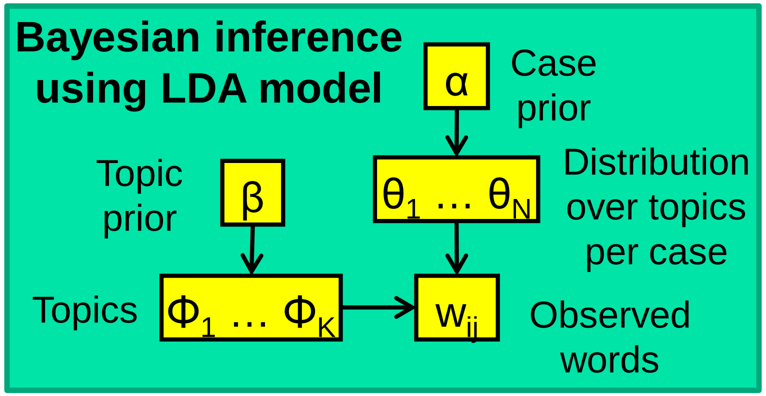
If we w... ...ief complaints
particular sy... ...ptom category
take a few such ... ...lute the outbreak signal,
that an outb...s occurring! ...ying or preventing detection.





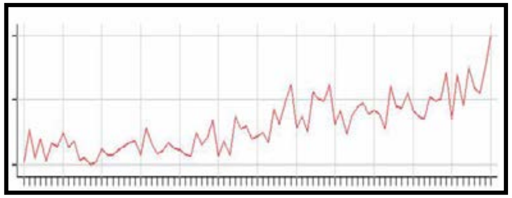
The semantic scan statistic

Date/time	Hosp.	Age	Complaint
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



Classify cases to topics

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



Time series of hourly counts for each combination of hospital and age group, for each topic φ_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)

(3 yrs. of data from 13 Allegheny County, PA hospitals)

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
…

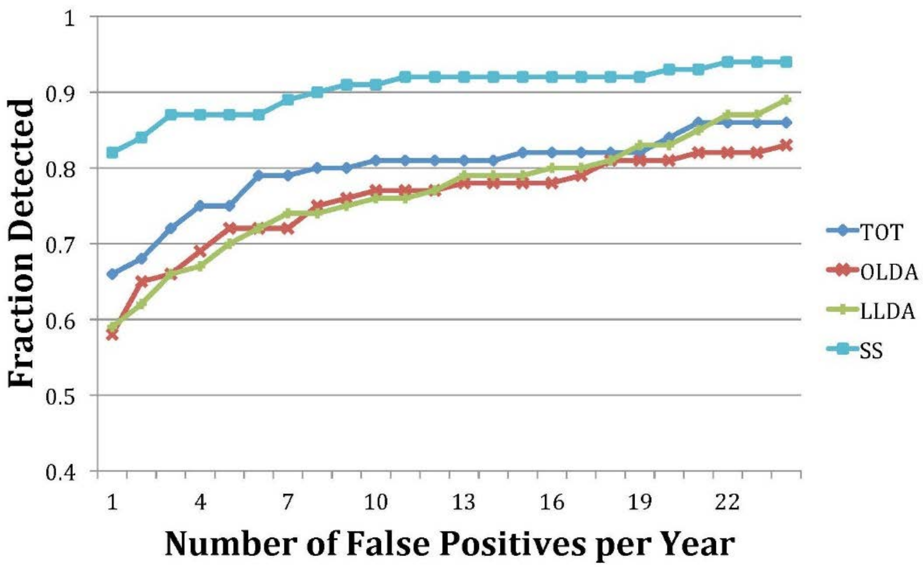
Top words from detected topic

Semantic scan results (2)

(3 yrs. of data from 13 Allegheny County, PA hospitals)

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.

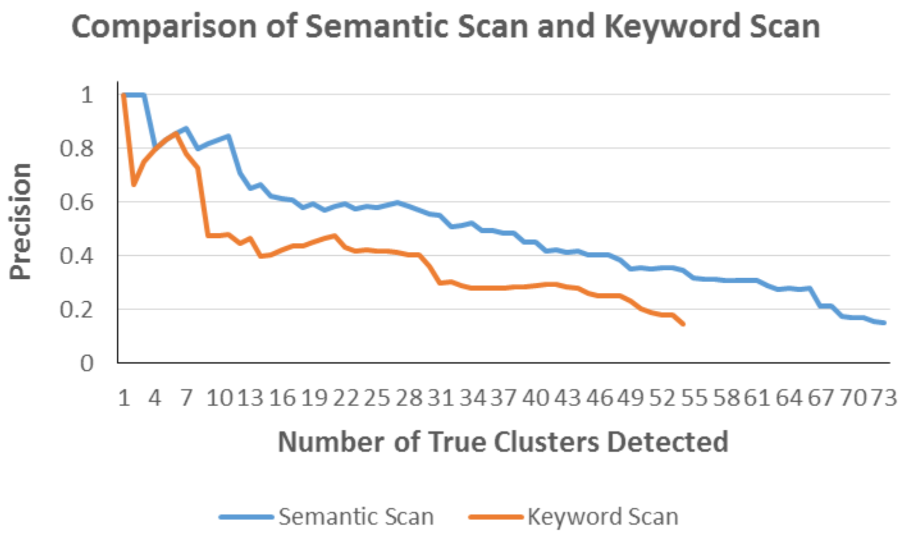


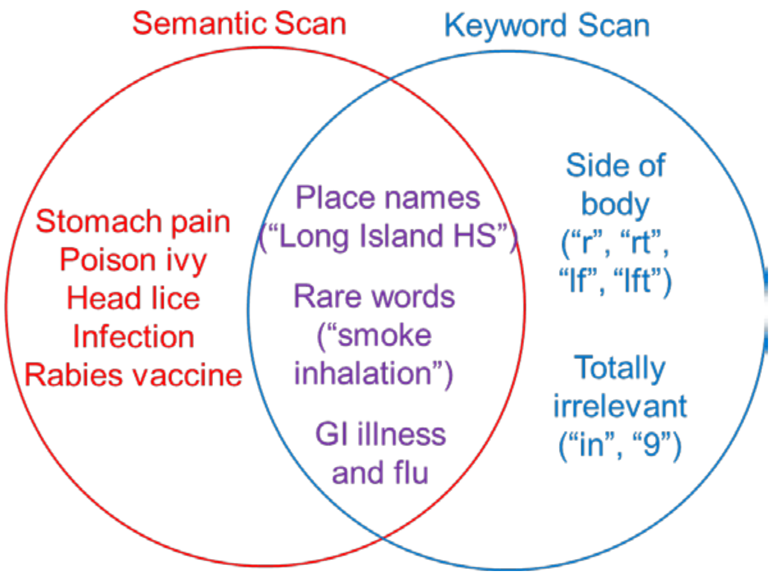
1) Learning a set of "background" topics from historical data.
2) Learning a set of "foreground" topics from recent data.
3) Combined LDA inference, holding the background topics constant, leads to discovery of foreground topics that are maximally different.

Semantic scan results (3)

(1 year of data from 3 hospitals in North Carolina)

We compared the top 500 clusters found by the semantic scan and a keyword-based scan in a blinded evaluation, with NC DOH 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.

Semantic scan results (4)

(5 yrs. of data from 10+ New York City hospitals)

Arrival Date	Arrival Time	Hospital ID	Chief Complaint	Patient Sex	Patient Age
11/28/2014	7:52:00	HOSP5	EVAUATION, DRANK COFFEE WITH CRUS	M	45-49
11/28/2014	7:53:00	HOSP5	DRANK TAINTED COFFEE	M	65-69
11/28/2014	7:57:00	HOSP5	DRANK TAINTED COFFEE	F	20-24
11/28/2014	7:59:00	HOSP5	INGESTED TAINTED COFFEE	M	35-39
11/28/2014	8:01:00	HOSP5	DRANK TAINTED COFFEE	M	45-49
11/28/2014	8:03:00	HOSP5	DRANK TAINTED COFFEE	M	40-44
11/28/2014	8:04:00	HOSP5	DRANK TAINTED COFFEE	M	30-34
11/28/2014	8:06:00	HOSP5	DRANK TAINTED COFFEE	M	35-39
11/28/2014	8:09:00	HOSP5	INGESTED TAINTED COFFEE	M	25-29

This detected cluster (in data from NYC DOHMH) represents patients complaining of ingesting tainted coffee, and demonstrates Semantic Scan's ability to automatically detect rare and novel events.

To extend our approach to the NYC data, we had to deal with additional questions of scale (~20M records) and complexity; much noisier text data required additional pre-processing steps.

Other NYC events identified by Semantic Scan include:

Accidents	Contagious Diseases	Other
Motor vehicle	Meningitis	Drug overdoses
Ferry	Scabies	Smoke inhalation
School bus	Ringworm	Carbon monoxide poisoning
Elevator		Crime related, e.g., pepper spray attacks

The progression of detected clusters immediately following Hurricane Sandy highlights the variety of strains placed on hospital emergency departments following a natural disaster:

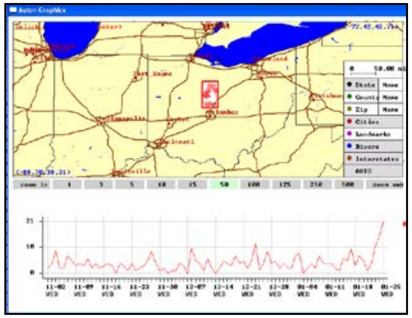
Acute cases: falls, SOB, leg Injuries → Mental health disturbances: depression, anxiety → Burden on medical infrastructure: methadone, dialysis

Conclusions

The continually increasing **size**, **variety**, and **complexity** of data available for population health and disease surveillance necessitate 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 (CMU): 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, Mallory Nobles, William Herlands.

- 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* 25: 382-404, 2016.
- T. Kumar and D.B. Neill. Fast multidimensional subset scan for event detection and characterization. Revised version in preparation.
- D.B. Neill and W. Herlands. Machine learning for drug overdose surveillance. *Journal of Technology in Human Services* 36(1): 8-14, 2018.
- A. Maurya, D.B. Neill, et al., Semantic scan: detecting subtle, spatially localized events in text streams. Submitted for publication.

Thanks for listening!

More details on my web site:
http://www.cs.nyu.edu/~neill

Or e-mail me at:
daniel.neill@nyu.edu