Event and Pattern Detection for Urban Systems

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, MacArthur Foundation, and Richard King Mellon Foundation.

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

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.

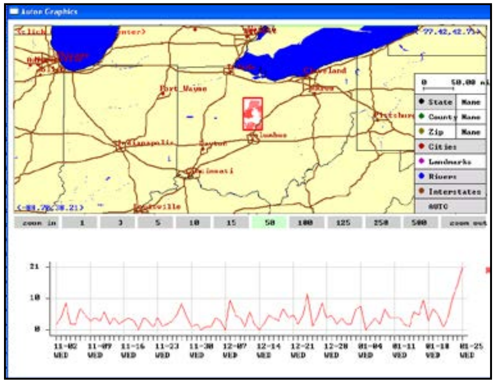
Machine learning has become increasingly essential for data-driven policy analysis and for the development of new, practical information technologies that can be directly applied **for the public good** (e.g. public health, safety, and security)

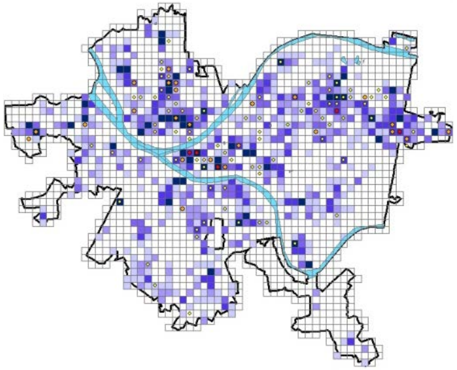
My research in this area has two main goals:

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, Heinz College, CMU
Director, Event and Pattern Detection Laboratory
Courtesy Associate Professor of Machine Learning and Robotics





<u>Disease Surveillance</u>: Very early and accurate detection of emerging outbreaks.

<u>Law Enforcement</u>: Detection, prediction, and prevention of “hot-spots” of violent crime.

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

Our disease surveillance methods are in use for deployed systems in the U.S., Canada, India, and Sri Lanka; currently collaborating 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)
- CrimeScan and CityScan
  - Crime prediction and prevention
  - Improving city services by predicting emerging patterns of citizen needs.

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.

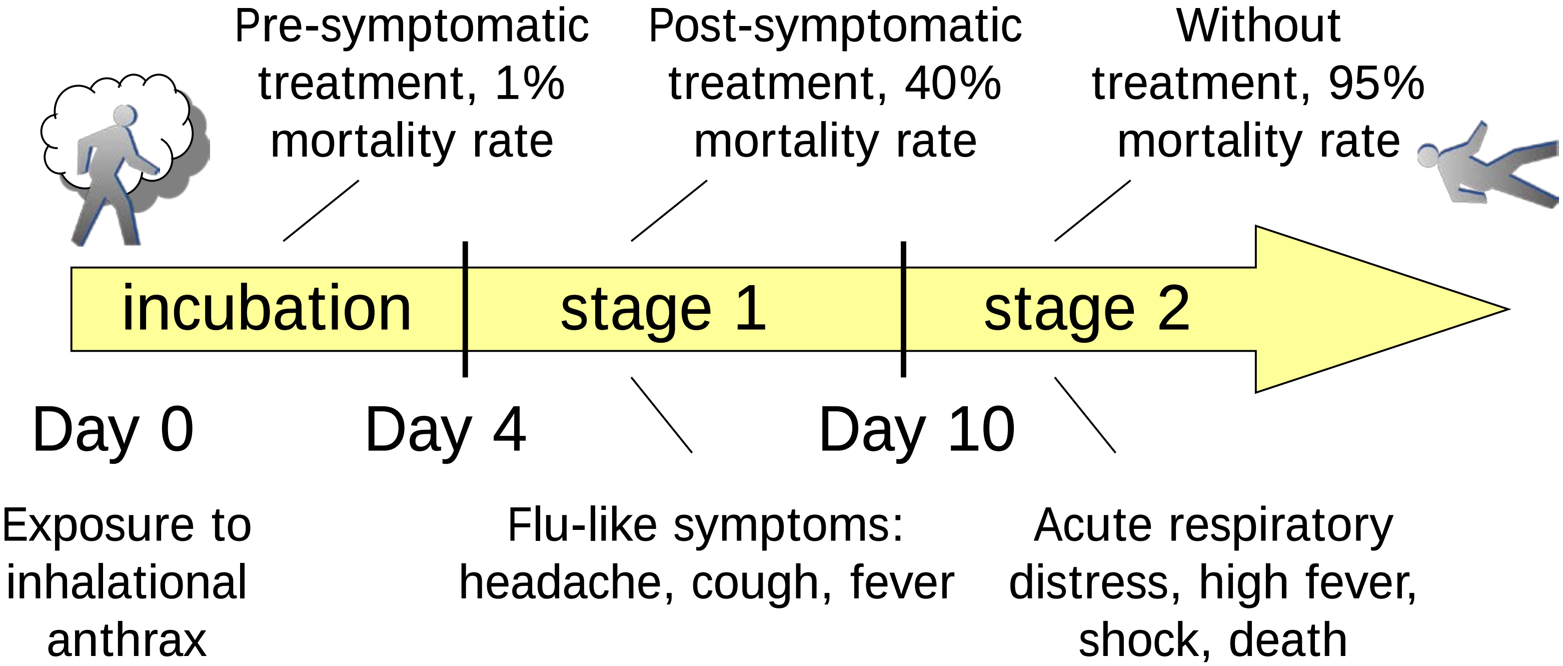
- Emerging infectious diseases

  “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!



Pre-symptomatic treatment, 1% mortality rate

Post-symptomatic treatment, 40% mortality rate

Without treatment, 95% mortality rate

incubation

stage 1

stage 2

Day 0

Day 4

Day 10

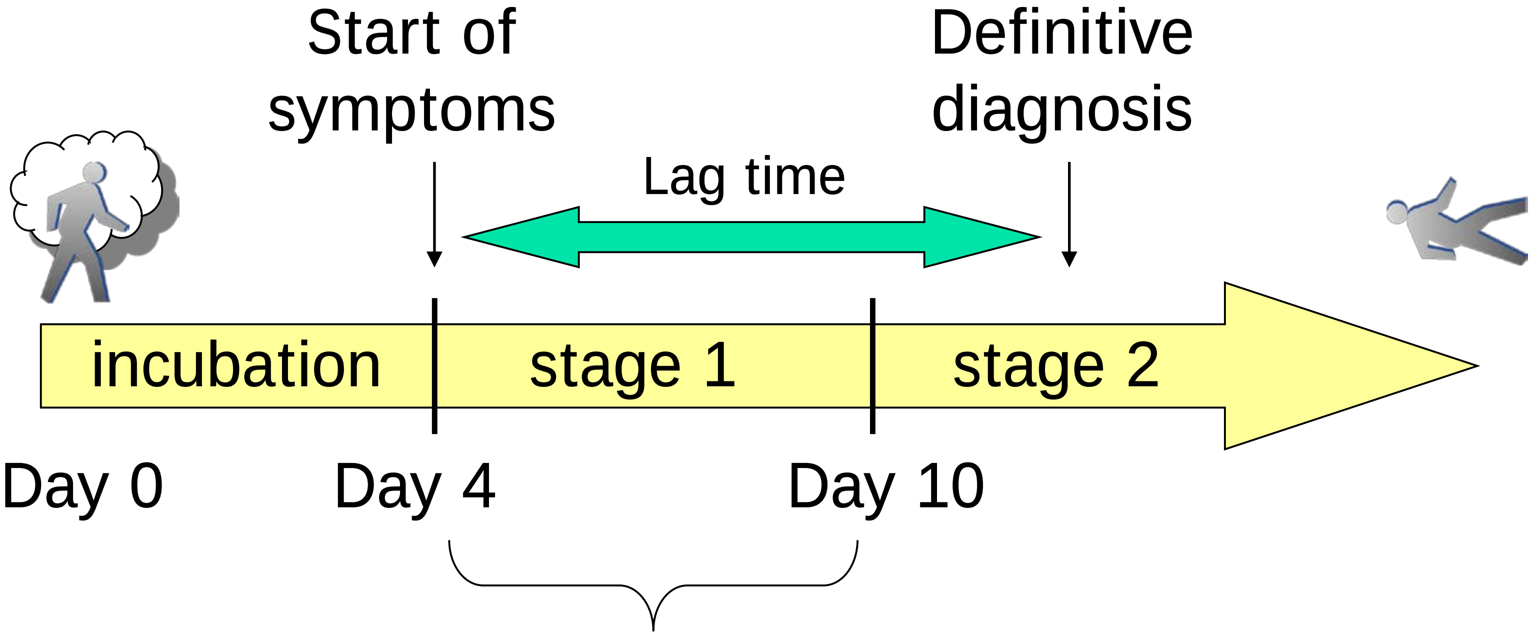
Exposure to inhalational anthrax

Flu-like symptoms: headache, cough, fever

Acute respiratory distress, high fever, shock, death

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



Start of symptoms

Definitive diagnosis

Lag time

incubation

stage 1

stage 2

Day 0

Day 4

Day 10

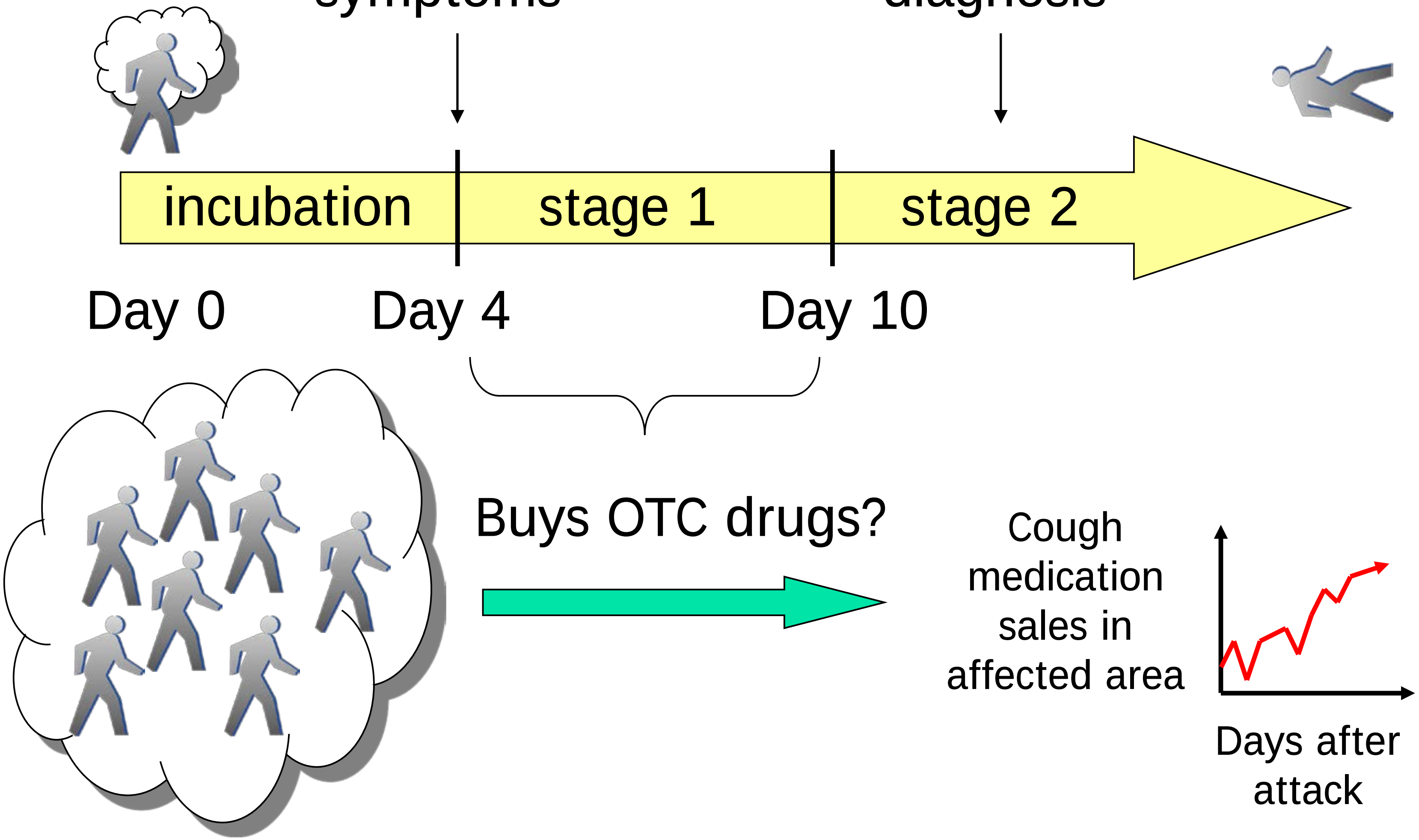
Buys OTC drugs

Skips work/school

Uses Google, Facebook, Twitter

Visits doctor/hospital/ED

Syndromic surveillance



Start of symptoms

Definitive diagnosis

incubation

stage 1

stage 2

Day 0

Day 4

Day 10

Buys OTC drugs?

Cough medication sales in affected area

Days after attack

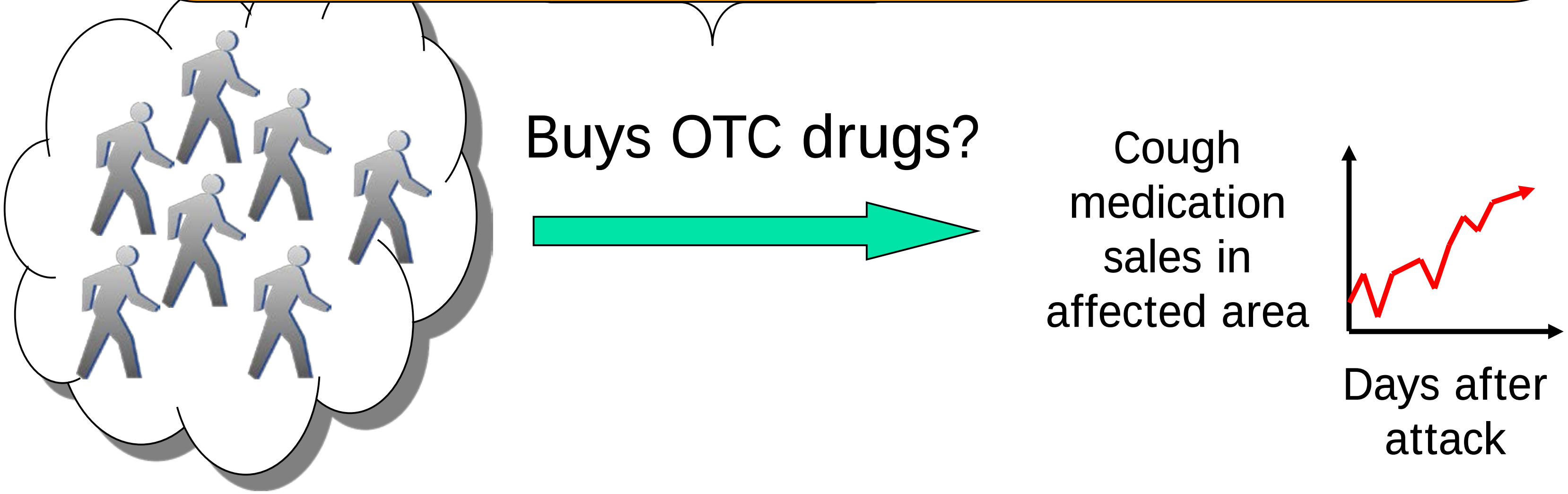
Syndromic surveillance

Start of symptoms

Definitive diagnosis

D

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



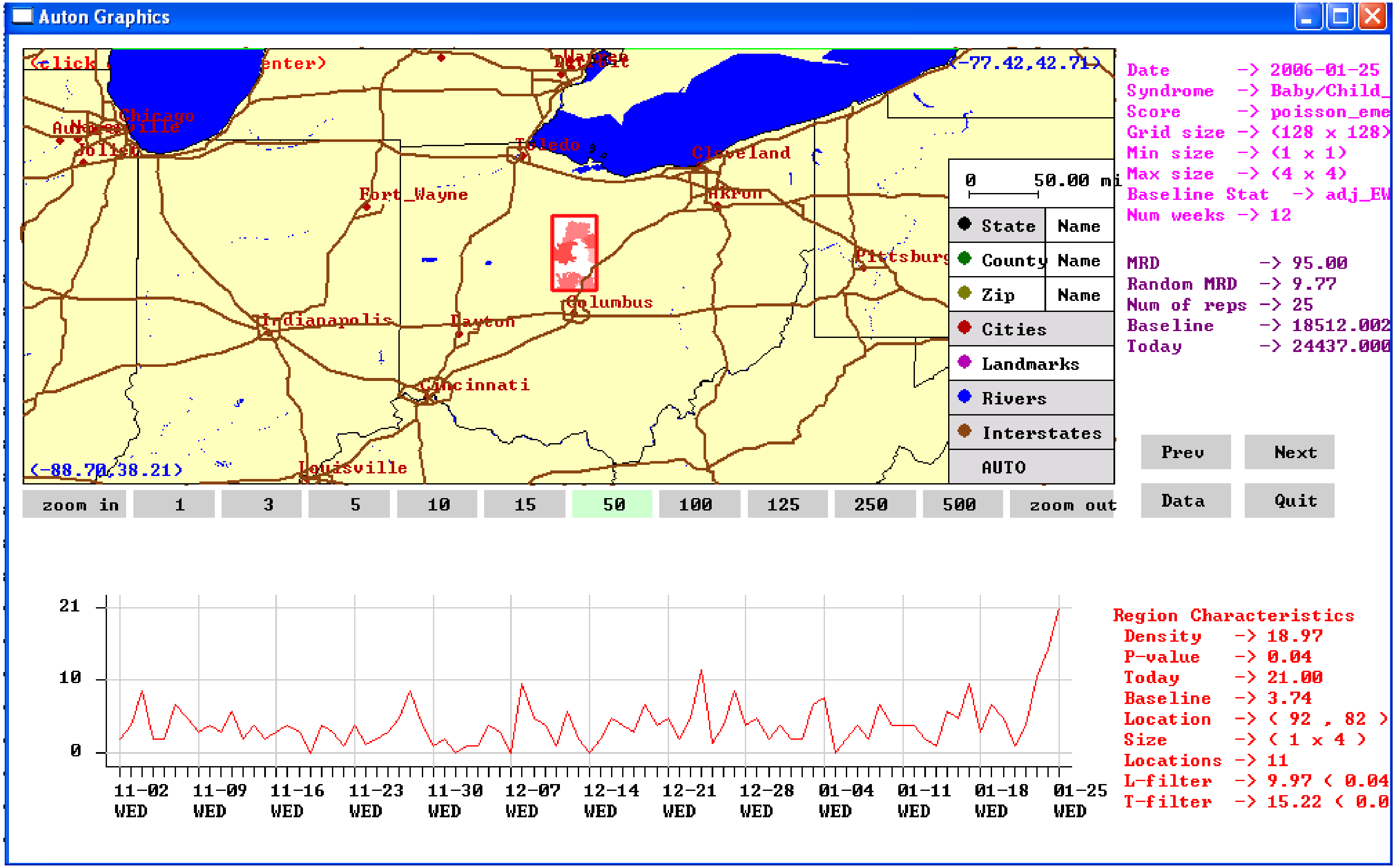
Buys OTC drugs?

Cough medication sales in affected area

Days after attack

Outbreak detection example

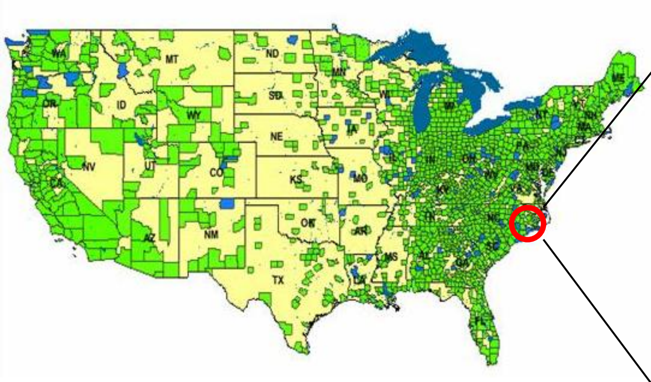
Spike in sales of pediatric electrolytes near Columbus, Ohio



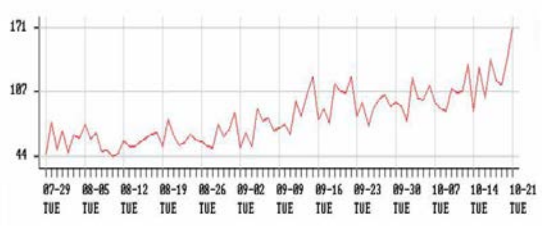
```
Date        -> 2006-01-25
Syndrome    -> Baby/Child_
Score       -> poisson_eme
Grid size   -> (128 x 128)
Min size    -> (1 x 1)
Max size    -> (4 x 4)
Baseline Stat   -> adj_EW
Num weeks   -> 12

MRD            -> 95.00
Random MRD     -> 9.77
Num of reps    -> 25
Baseline       -> 18512.002
Today          -> 24437.000
```

```
Region Characteristics
 Density     -> 18.97
 P-value     -> 0.04
 Today       -> 21.00
 Baseline    -> 3.74
 Location    -> ( 92 , 82 )
 Size        -> ( 1 x 4 )
 Locations   -> 11
 L-filter    -> 9.97 ( 0.04
 T-filter    -> 15.22 ( 0.0
```

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:

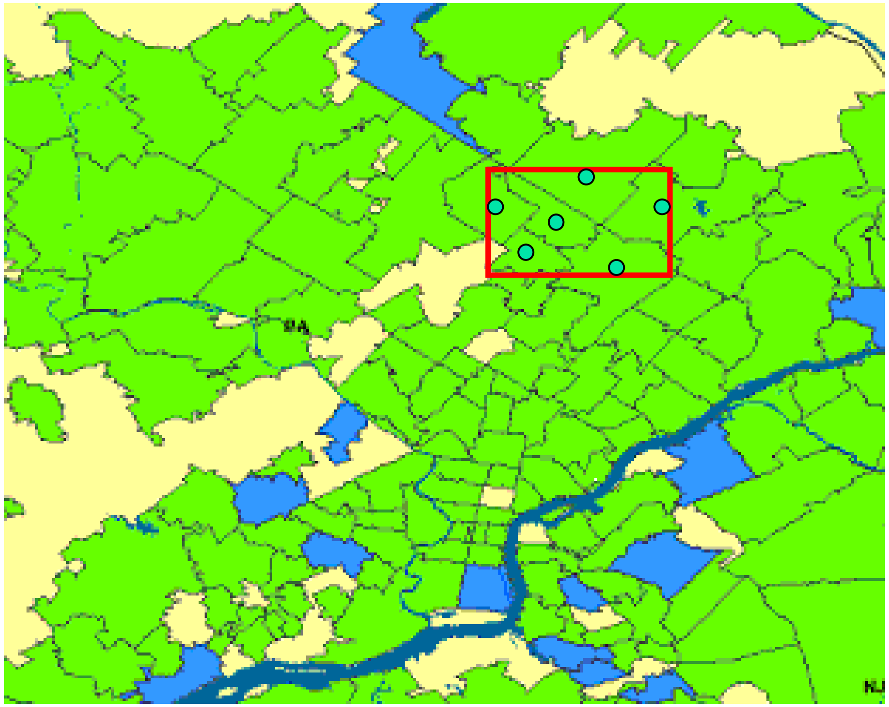
$H_1(D, S, W)$

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

vs. H_0: no events occurring

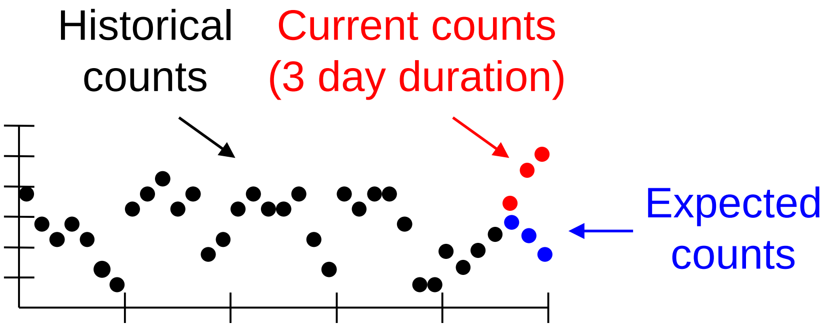
Expectation-based scan statistics

(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.

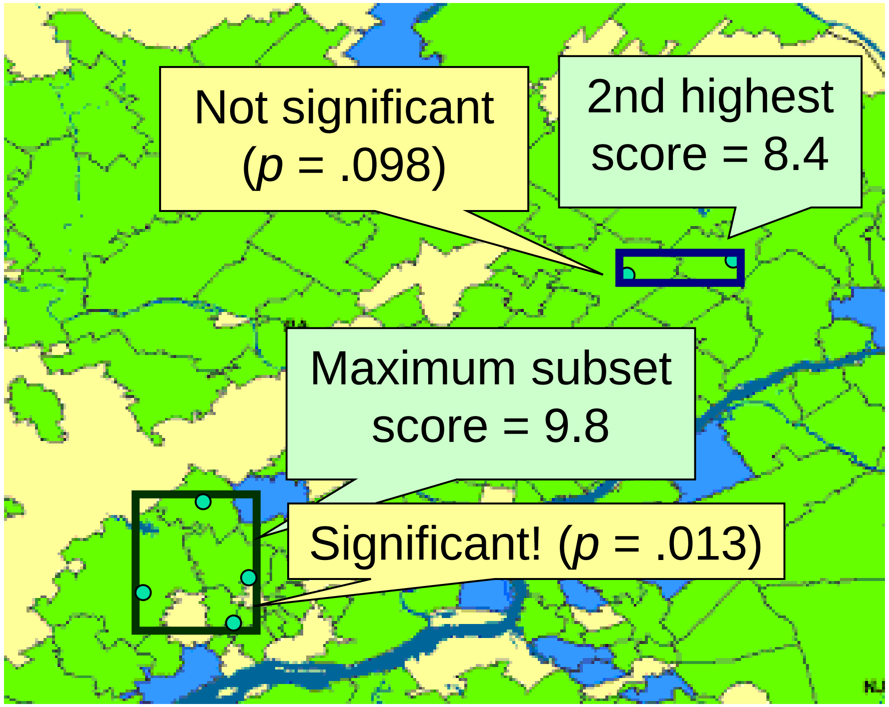
We perform **time series analysis** to compute expected counts ("baselines") for each location and stream for each recent day.



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)



Not significant (p = .098)

2nd highest score = 8.4

Maximum subset score = 9.8

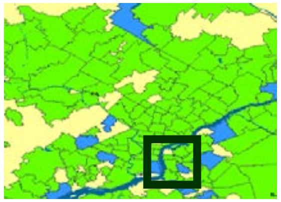
Significant! (p = .013)

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

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

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

Maximum likelihood: $q = C / B$.

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

Expectation-based Gaussian

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

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

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: $q = C' / B'$.

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

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

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

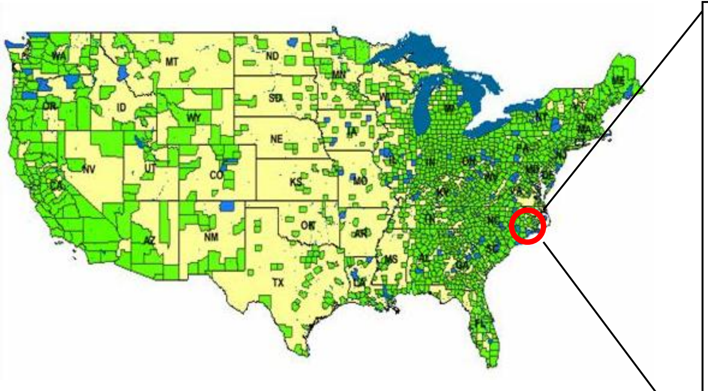
$$H_0 : x_i \sim Dist(\mu_i)$$

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

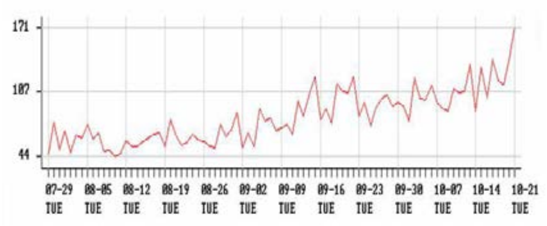
LTSS is a new and powerful tool for **exact** combinatorial optimization. But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

- To incorporate **spatial proximity** constraints, we maximize the likelihood ratio 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).
- This approach dramatically improves the timeliness and accuracy of outbreak detection for irregularly-shaped disease clusters. (Neill, *JRSS-B*, 2012)

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.

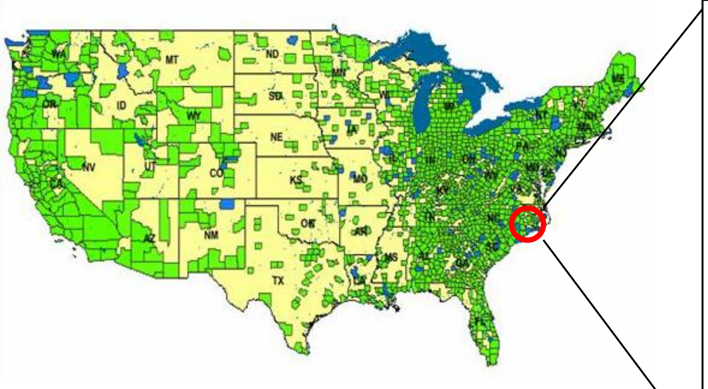
Compare hypotheses:

$H_1(D, S, W)$

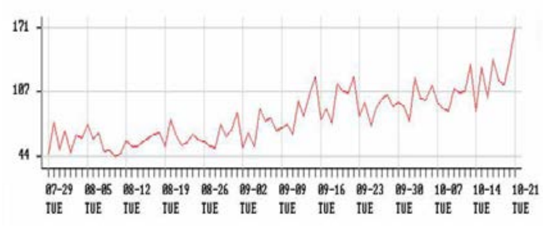
D = subset of streams
S = subset of locations
W = time duration

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:

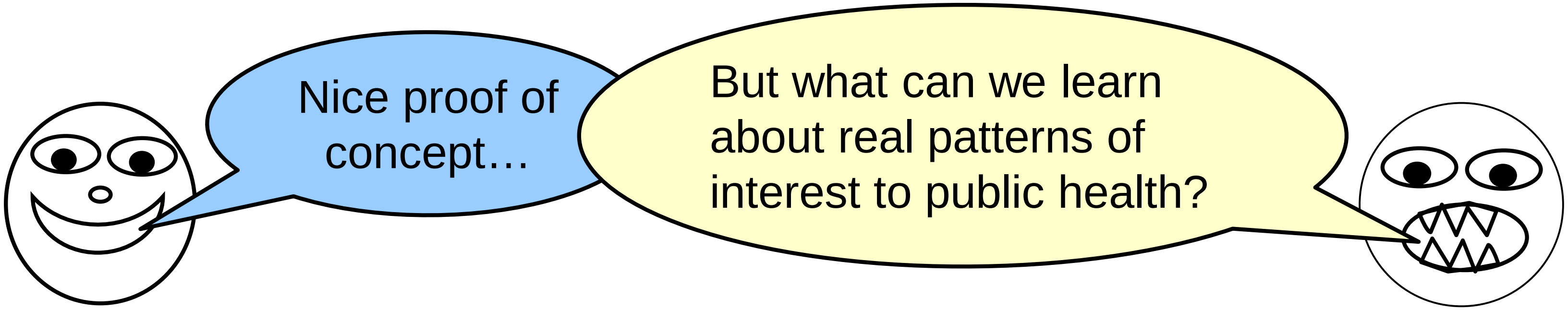
1. Start with randomly chosen subsets of **locations** S, **streams** D, and **values** V_j for each attribute A_j (j=1..J).

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.

Evaluation of MD-Scan

- We first evaluated the detection performance of MD-Scan for detecting simulated disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- For outbreaks with differential effects by age and gender, MD-Scan demonstrated **more timely** and **more accurate** detection, and accurately **characterized** the affected subpopulations.


Nice proof of concept…
But what can we learn about real patterns of interest to public health?

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)

40-100x more potent than heroin or morphine!

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?

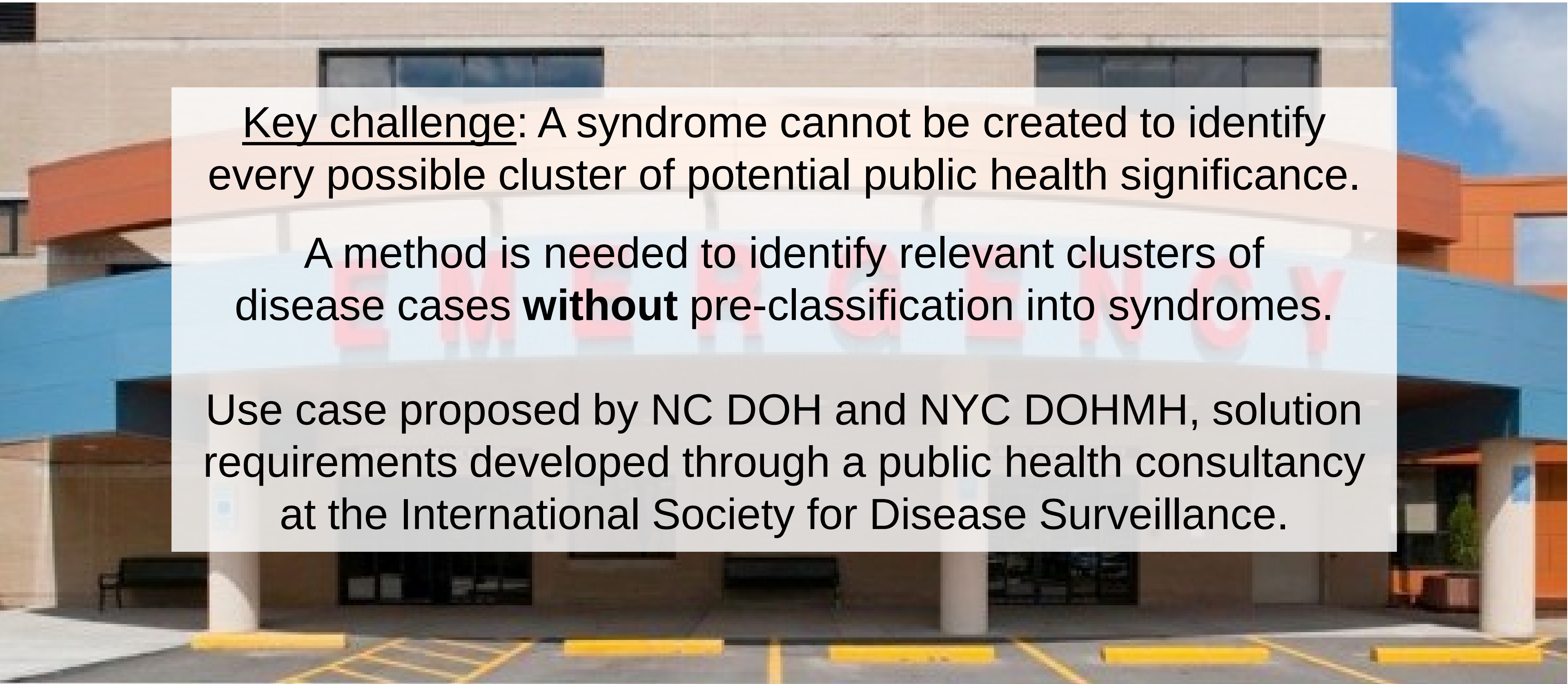
Today's talk:

- Public health surveillance
  - Early outbreak detection (fast subset scan)
  - Accidental drug overdose surveillance (multidimensional scan)
  - "Novel" outbreak detection (semantic scan)
- CrimeScan and CityScan
  - Crime prediction and prevention
  - Improving city services by predicting emerging patterns of citizen needs.

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

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…

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?

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

nose caught in door

rabies shot

a fib

tired weak

diarrhea

nausea
vomiting

n v d

food
poisoning

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).

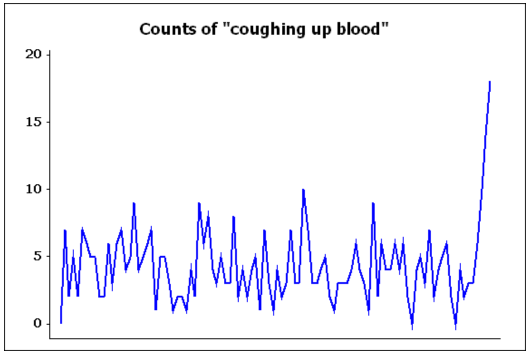
What happens when something new and scary comes along?

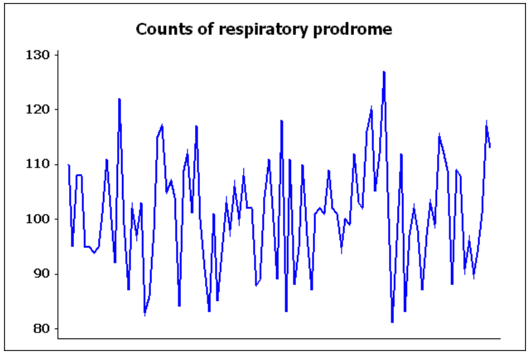
- **More specific symptoms** ("coughing up blood")

- **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?

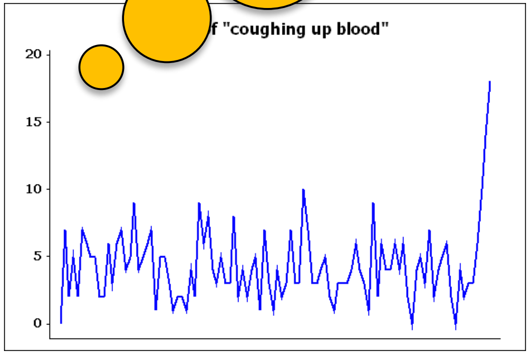
The typical, prodr
scan statisti
effectively
outbrea
seen
sy

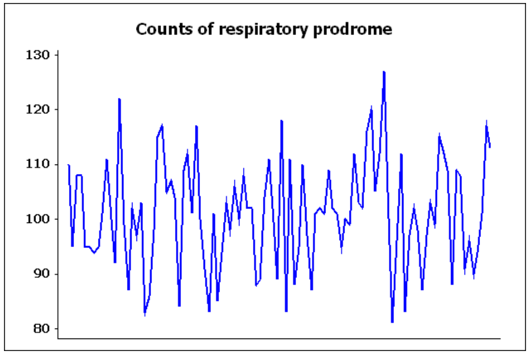
Our solution is to combine text-based (topic modeling) and event detection (multidimensional scan) approaches, to detect **emerging patterns of keywords**.

en something
along?
toms
d")
n
off")

If we w
particular sy
take a few such
that an outb
s occurring!

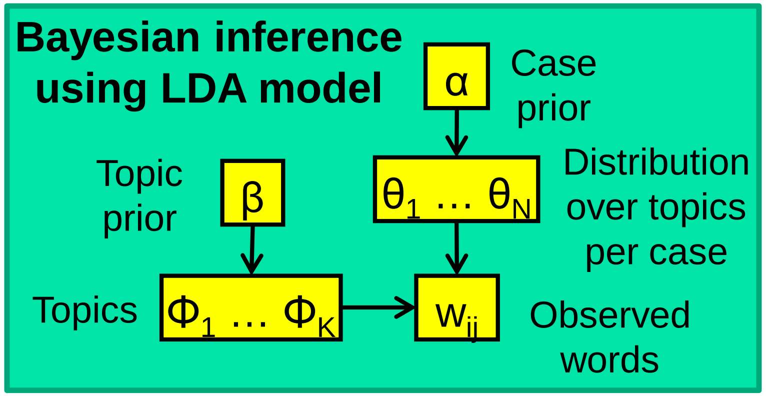
ief complaints
ptom category
ilute the outbreak signal,
ying or preventing detection.





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



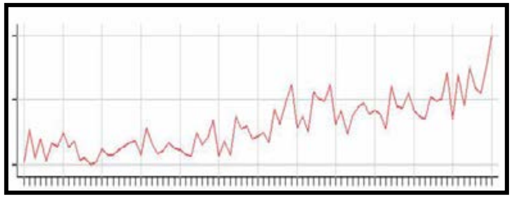
Bayesian inference using LDA model

φ_1: vomiting, nausea, diarrhea, …

φ_2: dizzy, lightheaded, weak, …

φ_3: cough, throat, sore, …

Classify cases to topics



Time series of hourly counts for each combination of hospital and age group, for each topic $\varphi_{j.}$

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”

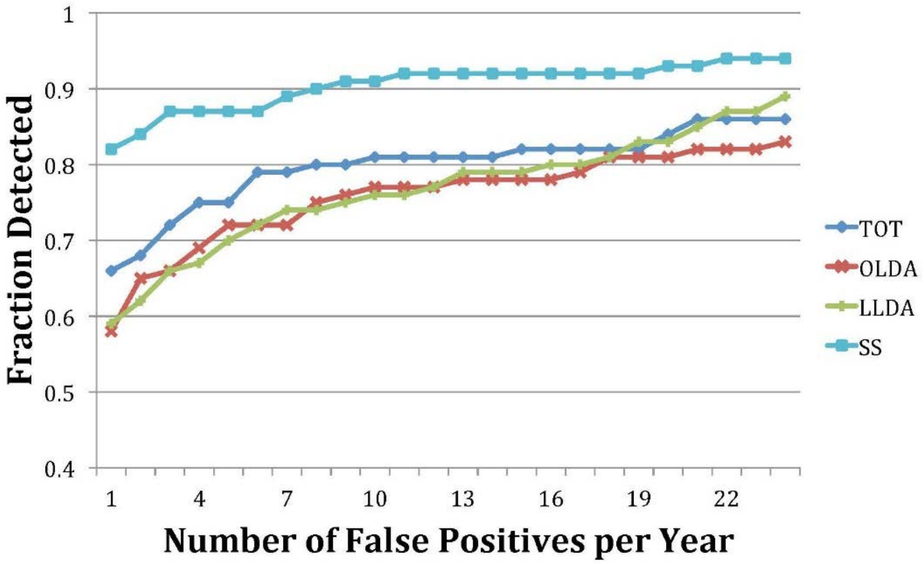
green
nose
possible
color
greenish
nasal
…

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"

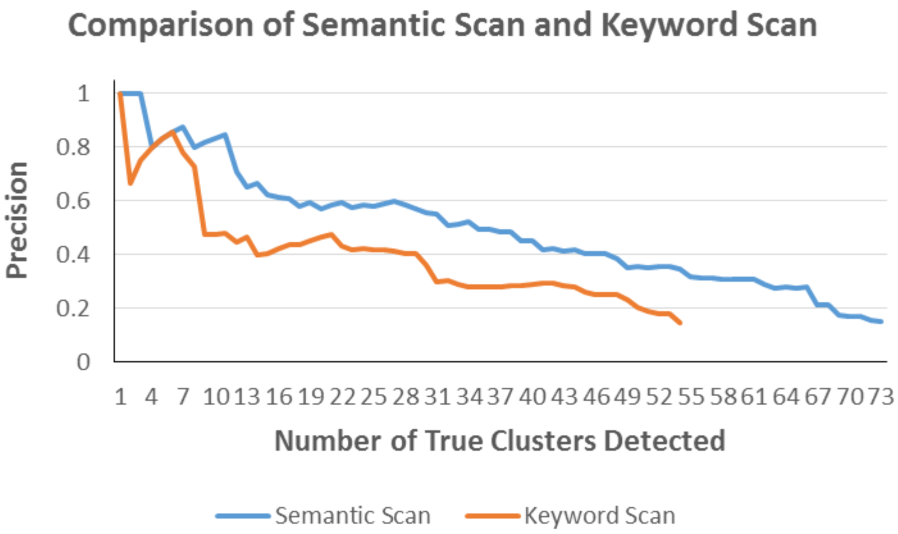
Clusters with related chief complaints:
chemical spill, motor vehicle accidents,
contagious diseases (head lice, scabies)

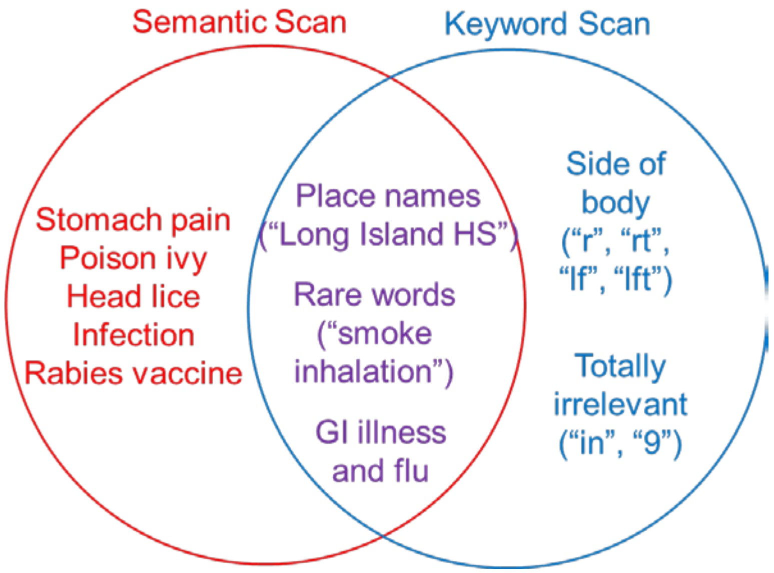
Ten cases that mentioned
a local middle school
within a four-hour span

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”.





Stomach pain
Poison ivy
Head lice
Infection
Rabies vaccine

Place names
(“Long Island HS”)

Rare words
(“smoke
inhalation”)

GI illness
and flu

Side of
body
(“r”, “rt”,
“lf”, “lft”)

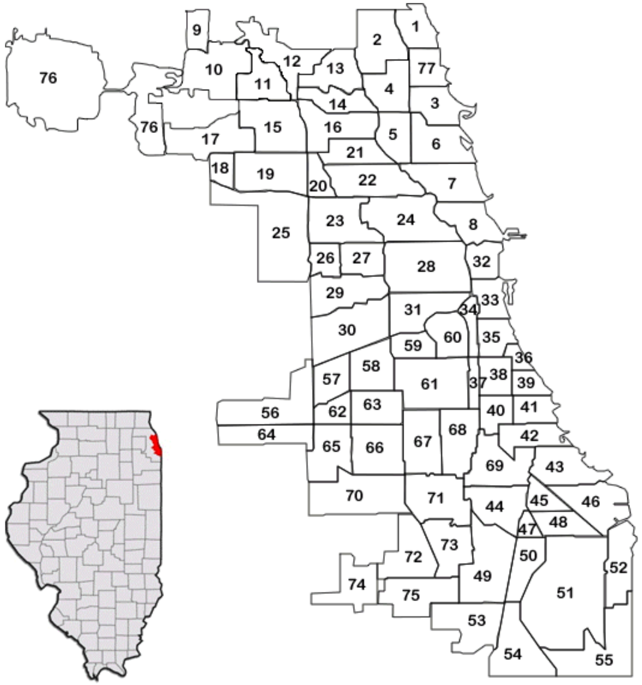
Totally
irrelevant
(“in”, “9”)

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.

Today's talk:

- Public health surveillance
  - Early outbreak detection (fast subset scan)
  - Accidental drug overdose surveillance (multidimensional scan)
  - "Novel" outbreak detection (semantic scan)
- CrimeScan and CityScan
  - Crime prediction and prevention
  - Improving city services by predicting emerging patterns of citizen needs.

Case study: Crime prediction in Chicago



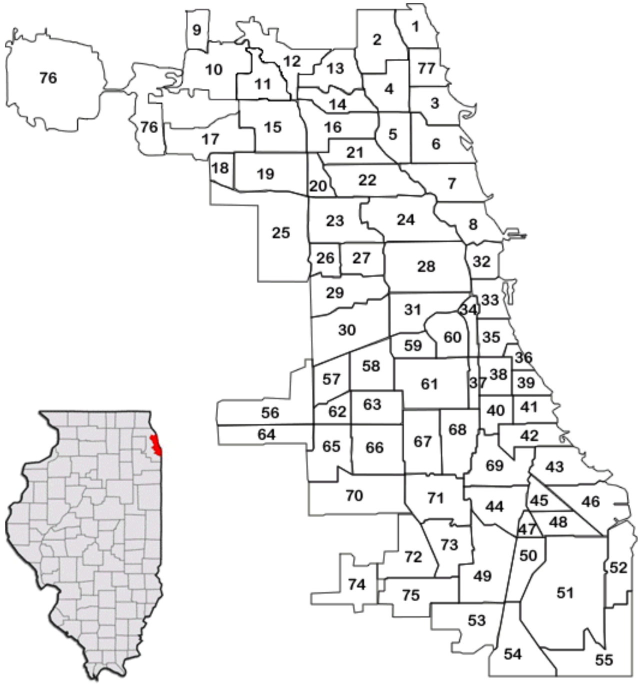
Since 2009, we have been working with the Chicago Police Department (CPD) to predict and prevent emerging clusters of violent crime.

Our new crime prediction methods have been incorporated into our **CrimeScan** software, run twice a day by CPD and used operationally 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…"

Case study: 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, run twice a day by CPD and used operationally for deployment of patrols.

“CrimeScan was set up to run daily, completely autonomously. Predictions were sent to police analysts, and messages were compiled into detailed intelligence reports disseminated through the chain of command. Based upon deployment suggestions indicated in the reports, important arrests were affected, weapons were seized, and crimes were prevented.”

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

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.

The fast subset scanning approaches described above enable **early and accurate detection** of emerging clusters.

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

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.

The fast subset scanning approaches described above enable **early and accurate detection** of emerging clusters.

Proximity to detected clusters → features in a predictive model.

We use **scalable Gaussian process regression** to model spatial correlation and improve prediction accuracy.

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

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.

"Kitchen sink" penalized regression does not work so well.

Correlation-based LI selection is confounded by purely spatial and purely temporal correlations.

Our solution is a new bivariate "kernel space-time independence" test that identifies space-time interactions between LI types while controlling for space and time.

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

From CrimeScan to CityScan…

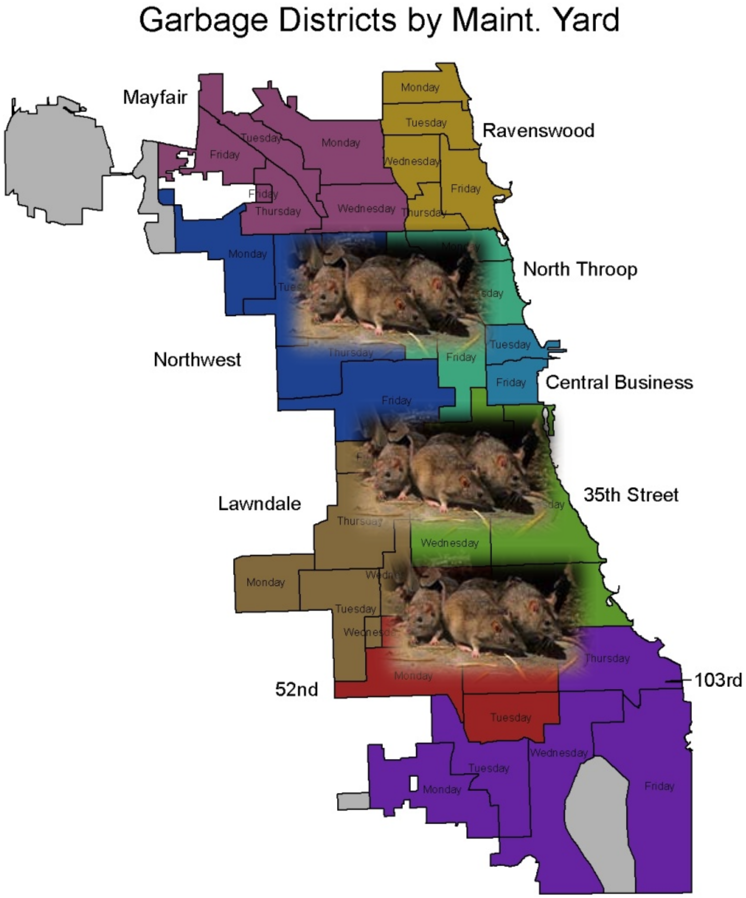
We have been working with city leaders in Chicago, Pittsburgh, and Baltimore to predict emerging spatial patterns of **311 calls** (non-emergency service requests).

By providing support for precisely targeted interventions, we will enable cities to respond **proactively** and **effectively** to emerging challenges and citizen needs.

Indicators of neighborhood decay (graffiti, abandoned buildings, etc.)

Health and sanitation issues, particularly focusing on rodent prevention.

CityScan: Preventing rat infestations



We are currently performing a controlled experiment with Chicago's Dept. of Streets and Sanitation, with the goal of predicting and preventing rodent infestations.

- Measured by "rodent complaint" 311 calls.
- Other 311 call types as leading indicators.

<u>"Treatment" garbage districts</u>:
We predict rodent complaints using CityScan and use predictions to direct the city's preventative rat baiting crews.

<u>"Control" garbage districts</u>:
Preventative baiting performed as usual.

Featured in Chicago Business Journal and Baltimore Sun-Times:
"Carnegie Mellon smells a rat, and Chicago is grateful"

Crime Prevention in Pittsburgh

Richard King Mellon Foundation

PITTSBURGH POLICE
ACCOUNTABILITY INTEGRITY RESPECT

Allegheny County · Department of Human Services ·

CITY OF PITTSBURGH
OFFICE OF THE MAYOR

Carnegie Mellon University

Integrating geographic crime prediction with subgroup and individual-level predictions.

Analyzing social media to identify causal mechanisms leading to outbreaks of violence.

Incorporating many data sources: 911 and 311 calls, incident reports, criminal justice, human services…

Integrating precisely targeted policing with non-punitive interventions by city and county (e.g., targeted clean-up efforts).

Our goal is to make the city’s crime prevention efforts both more effective and less intrusive (“data-driven community policing”).

Conclusions

Urban systems present unique data analysis challenges that cannot be solved by off-the-shelf methods, requiring new innovations in machine learning methodology.

Our work in event and pattern detection, applied to domains like **disease outbreak detection** and **crime prevention**, can effectively address many of these challenges.

We continue to work closely with public health departments, police, and city leaders to develop, deploy, and evaluate approaches that address critical urban problems.

Safer Cities

Cleaner Cities

Healthier Cities

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.
- S. Flaxman, D.B. Neill, et al., Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. *Proc. Intl. Conf. on Machine Learning*, 2015.

Thanks for listening!

More details on our web site:
http://epdlab.heinz.cmu.edu

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