

# Event and Pattern Detection at the Societal Scale

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, and the John D. and Catherine T. MacArthur Foundation.

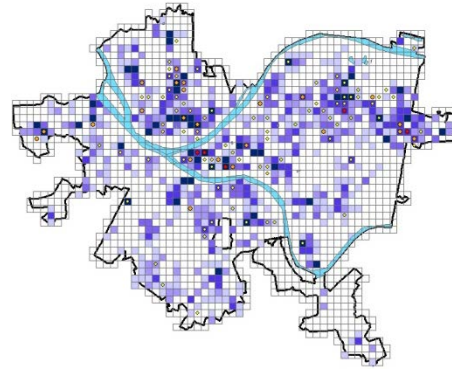
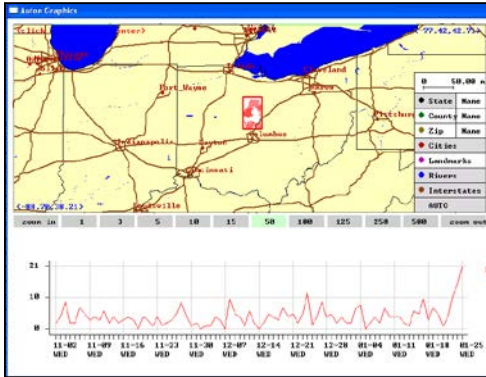
Carnegie Mellon University

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.

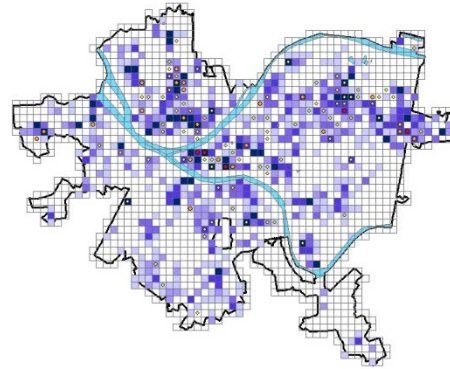
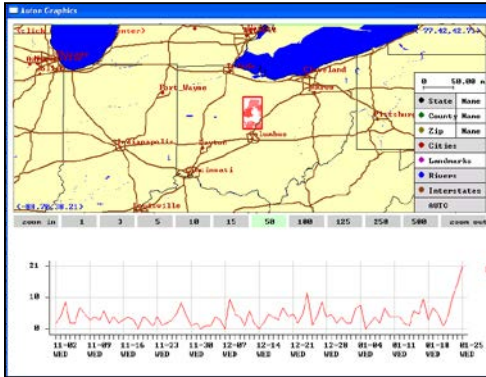
Medicine: Discovering new “best practices” of patient care, to improve outcomes and reduce costs.

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.

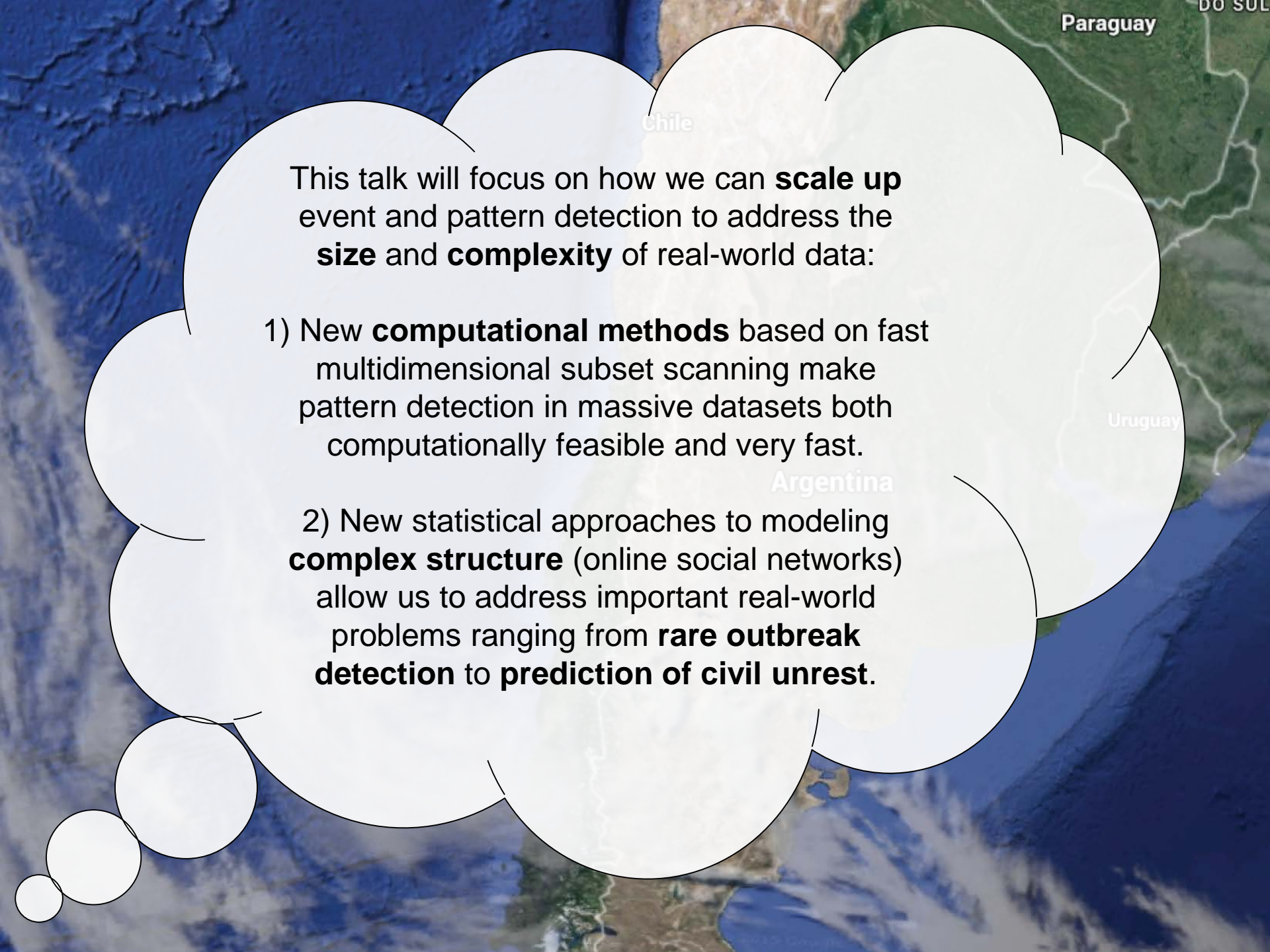
Law Enforcement:  
 Detection, prediction, and prevention of “hot-spots” of violent crime.

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

Our disease surveillance methods have been 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 Chicago and Pittsburgh PDs. “CityScan” has been used by Chicago city leaders for prediction and prevention of rodent infestations using 311 call data.



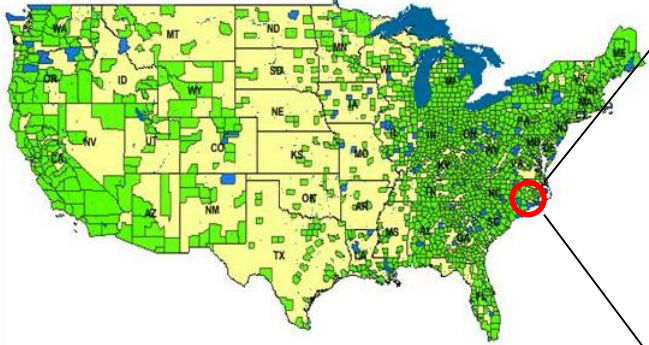


This talk will focus on how we can **scale up** event and pattern detection to address the **size** and **complexity** of real-world data:

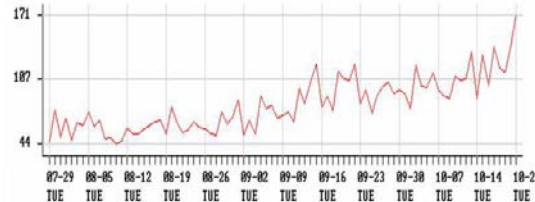
1) New **computational methods** based on fast multidimensional subset scanning make pattern detection in massive datasets both computationally feasible and very fast.

2) New statistical approaches to modeling **complex structure** (online social networks) allow us to address important real-world problems ranging from **rare outbreak detection** to **prediction of civil unrest**.

# 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:

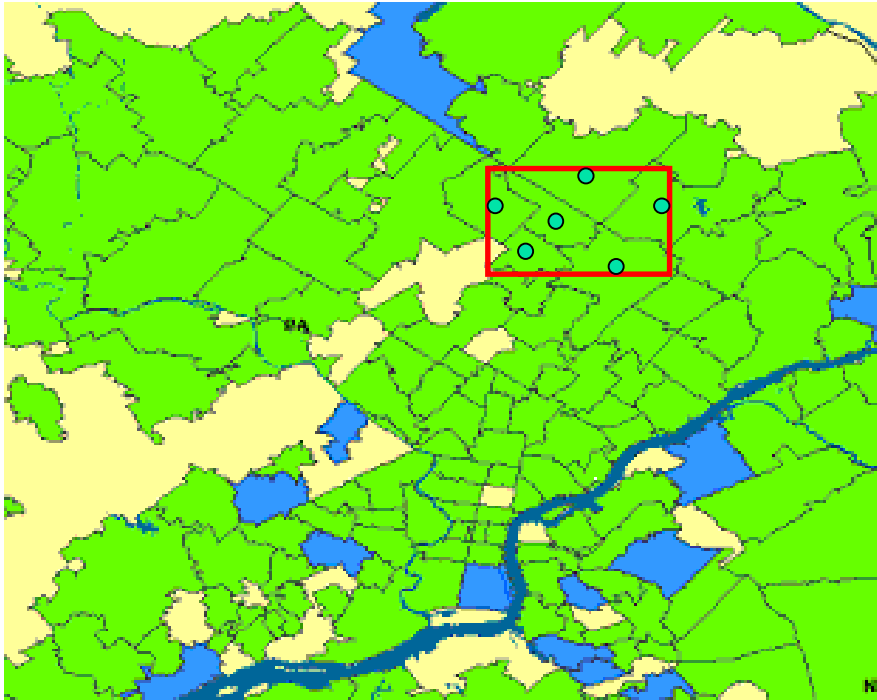
$$H_1(D, S, W)$$

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

vs.  $H_0$ : no events occurring

# Expectation-based scan statistics

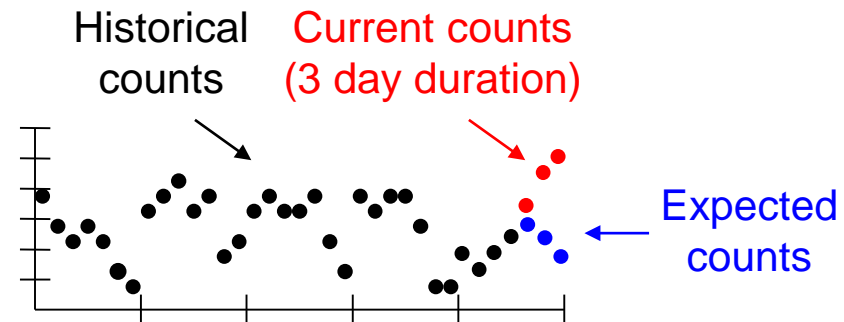
(Kulldorff, 1997; Neill and Moore, 2005)



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

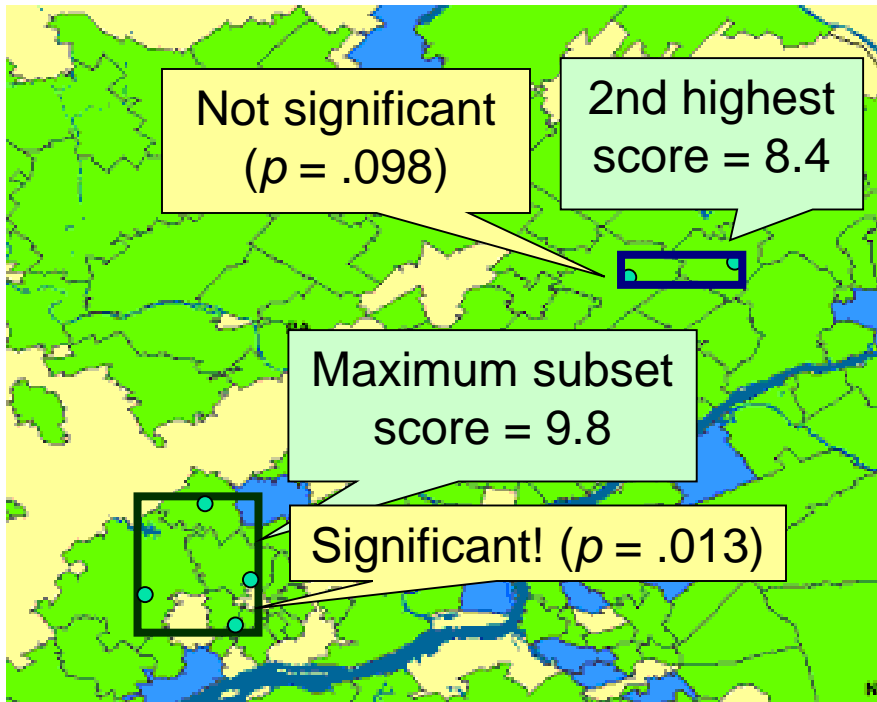
We perform **time series analysis** to compute expected counts (“baselines”) for each location and stream for each recent day.

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)

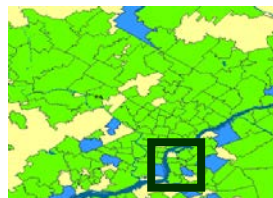


We find the subsets with highest values of a **likelihood ratio statistic**, and compute the  $p$ -value of each subset by randomization testing.

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

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.

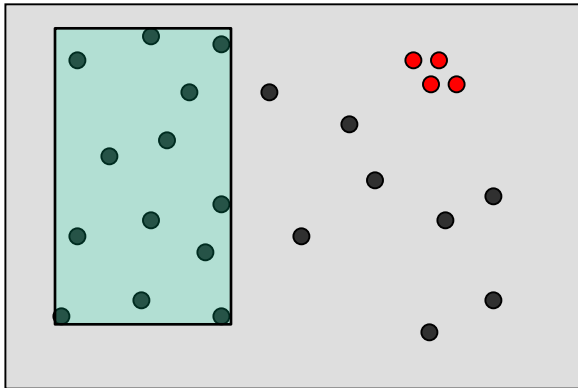
# Question: Why search over subsets?

## Answer: Simpler approaches can fail.

### Top-down detection approaches

Are there any globally interesting patterns? If so, recursively search the most interesting sub-partition.

Two examples: bump hunting;  
“cluster then detect”.

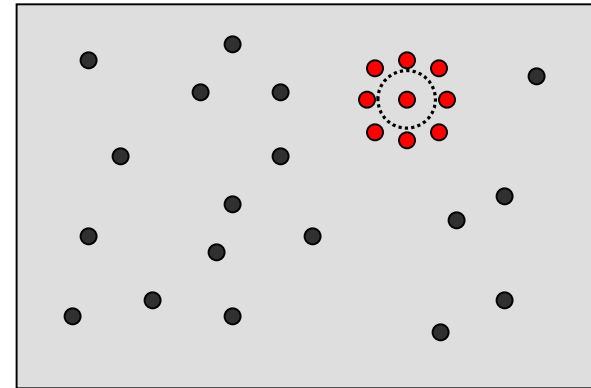


Top-down fails for **small-scale patterns** that are not evident from the global aggregates.

### Bottom-up detection approaches

Find individually (or locally) anomalous data points, and optionally, aggregate into clusters.

Two examples: anomaly/outlier detection;  
density-based clustering.



Bottom-up fails for **subtle patterns** that are only evident when a group of data records are considered collectively.

# Question: Why search over subsets? Answer: Simpler approaches can fail.

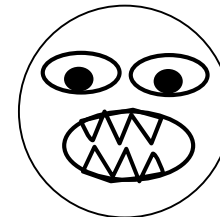
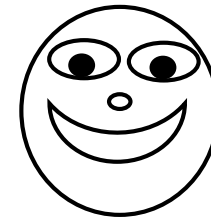
Top-down detection approaches

Are there any patterns?  
the most interesting

So here's where we are so far:

Treating pattern detection as a subset scan problem is statistically desirable for maximizing detection power...

but computationally infeasible (for exhaustive search at least).



Top-down fails to find **subtle patterns** that are not evident when a group of data words are considered collectively.

# 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., 2016)

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

# 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  $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$ .
- Even better theorem: We can also maximize the **penalized** scan statistic  $F(S) + \sum_{s_i \in S} \Delta_i$  in  $O(N \log N)$  time, evaluating only  $2N$  of the  $2^N$  subsets.

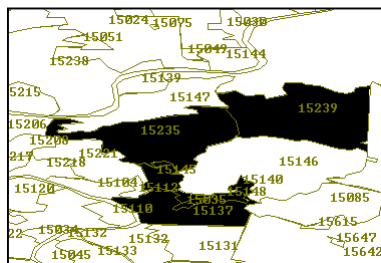
(Speakman et al., 2016)

# 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.

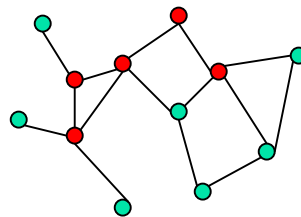
Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + 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

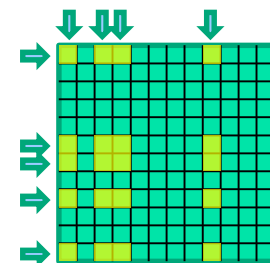


(Neill, *JRSS-B*, 2012)

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



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



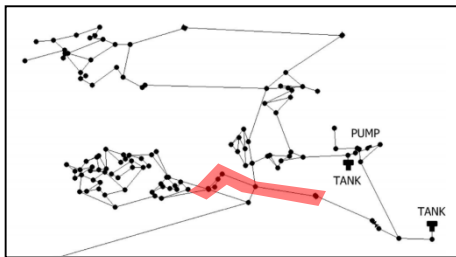
(McFowland et al., *JMLR*, 2013)

# Constrained fast subset scanning

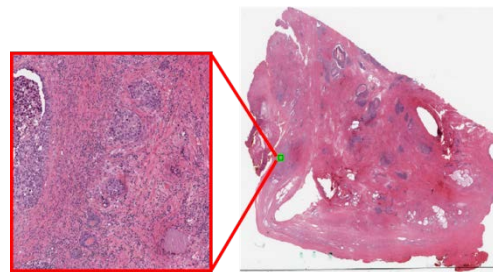
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

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)



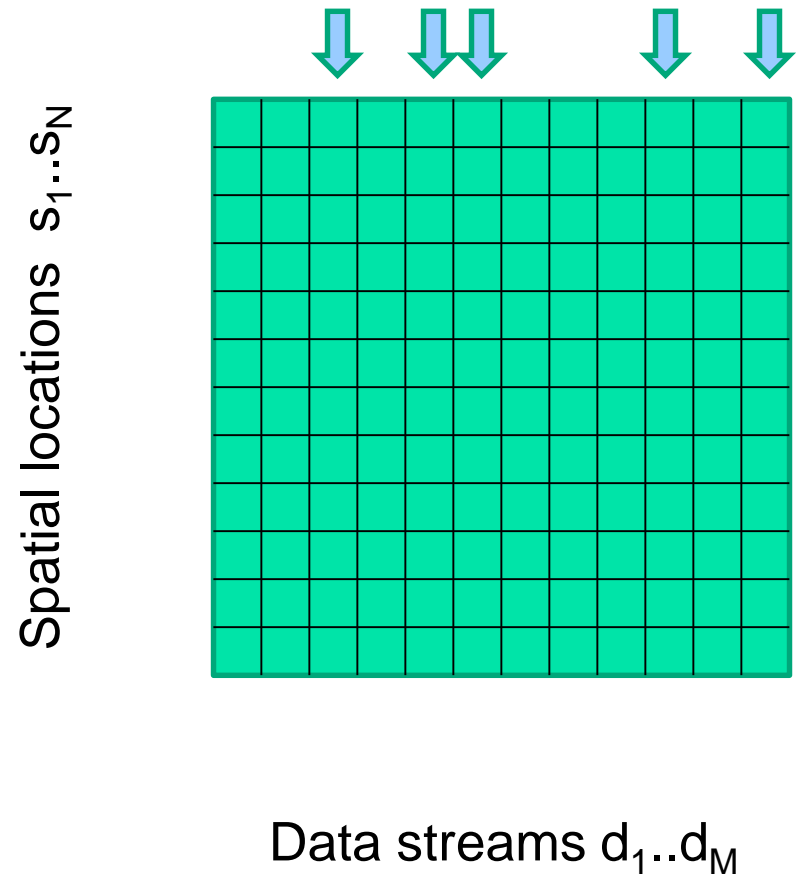
# Fast subset scan with spatial proximity constraints

- 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)

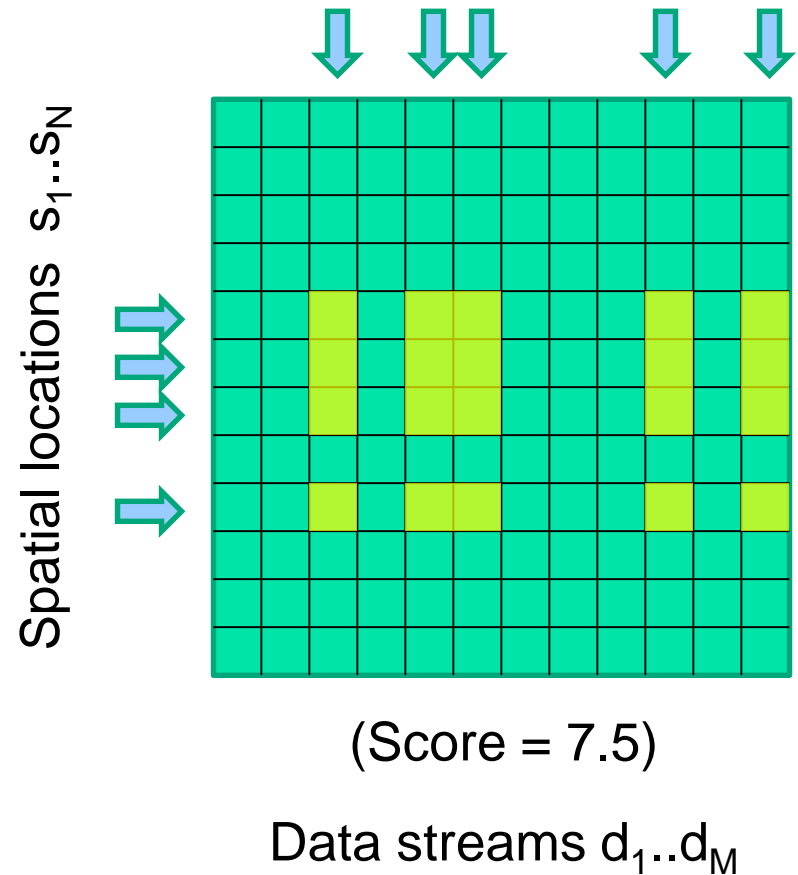
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

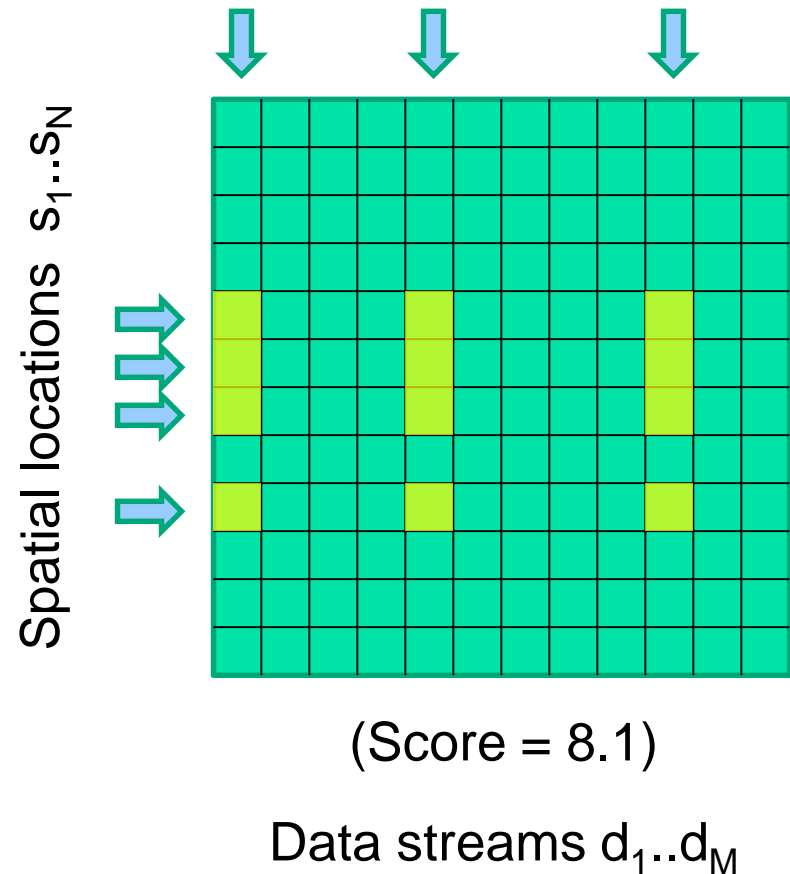
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



# 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**...

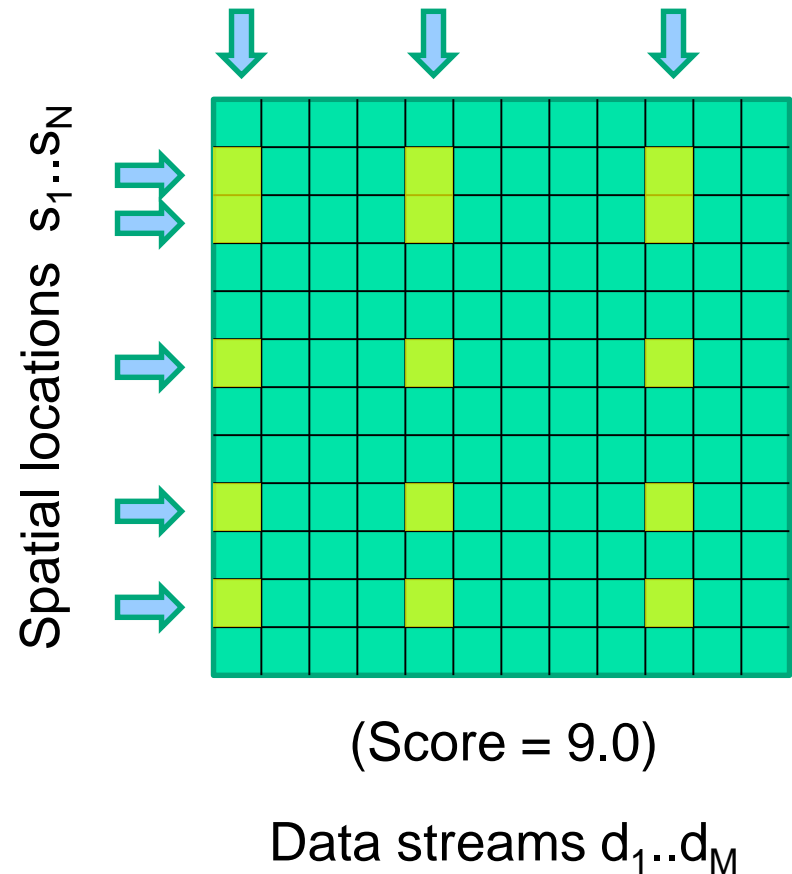




# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

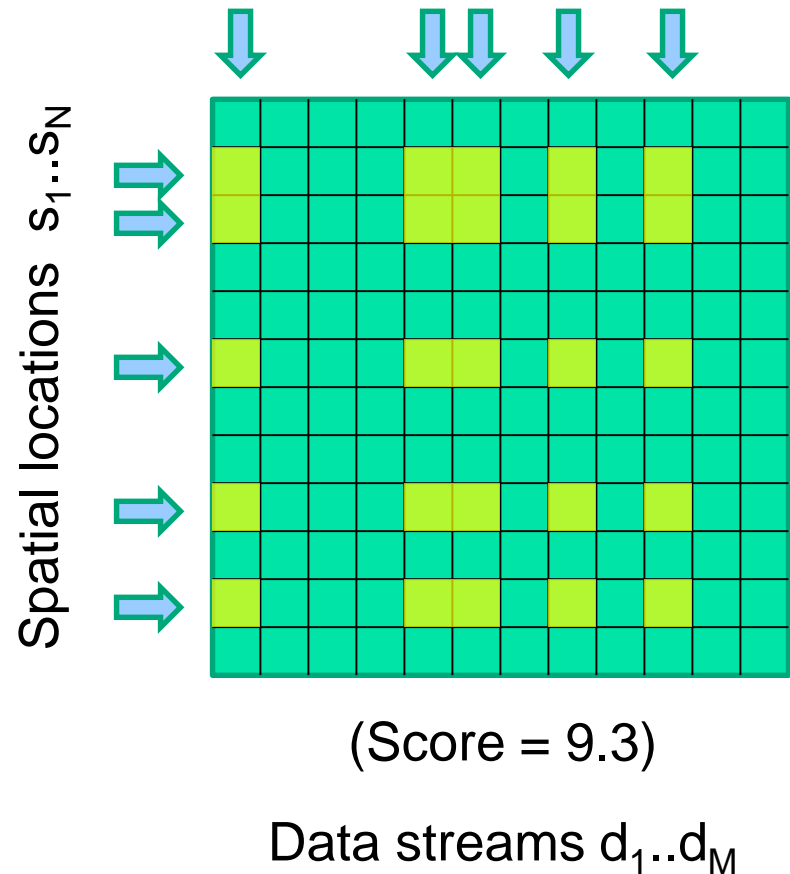
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations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



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

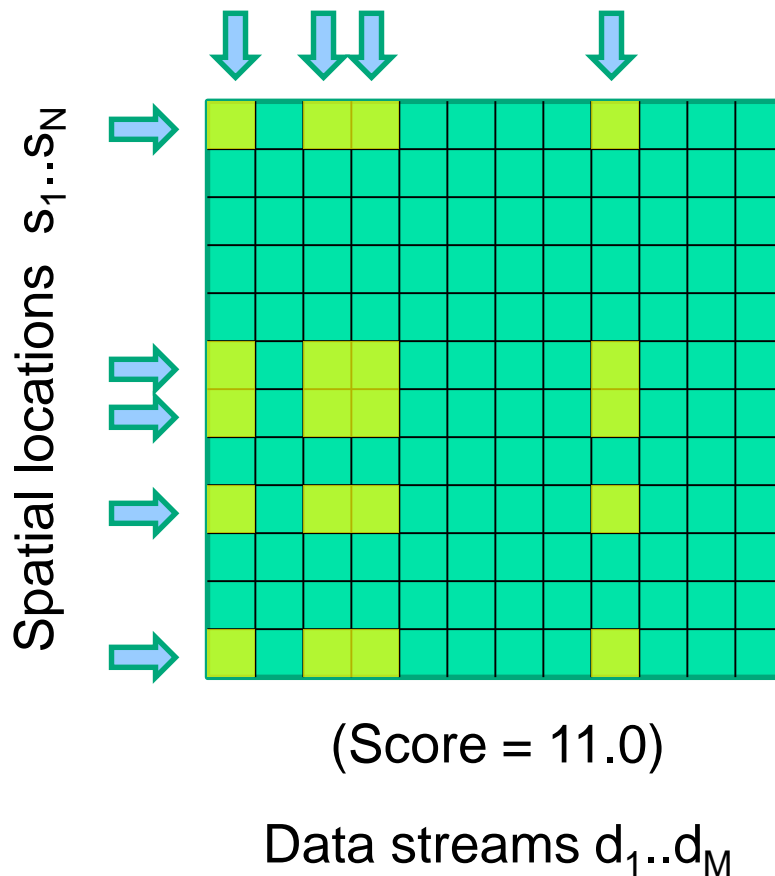
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations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



# Multivariate fast subset scan

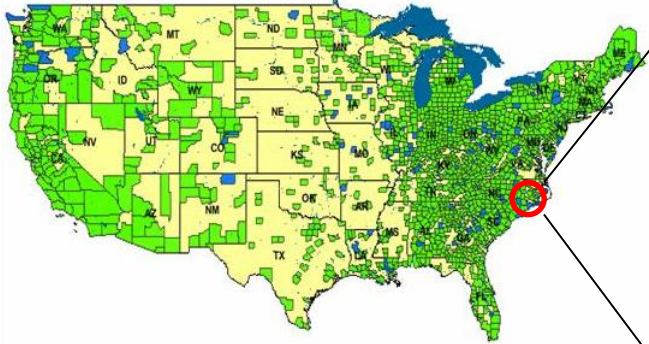
(Neill, McFowland, and Zheng, 2013)

- 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.
- For general datasets, a similar approach\* can be used to jointly optimize over subsets of data records and attributes.

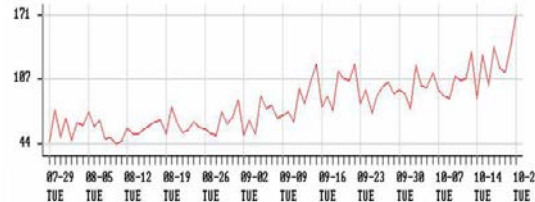


\*McFowland, Speakman, and Neill, *JMLR*, 2013

# 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.

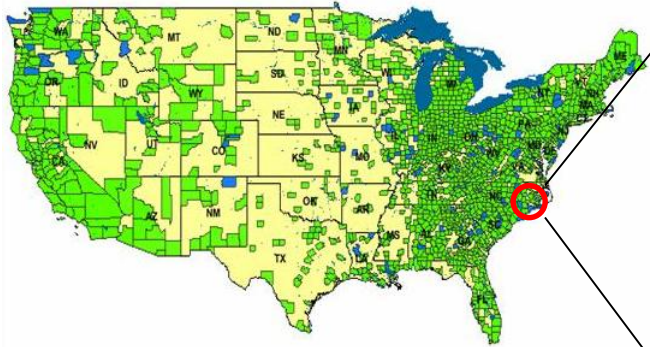
## Compare hypotheses:

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

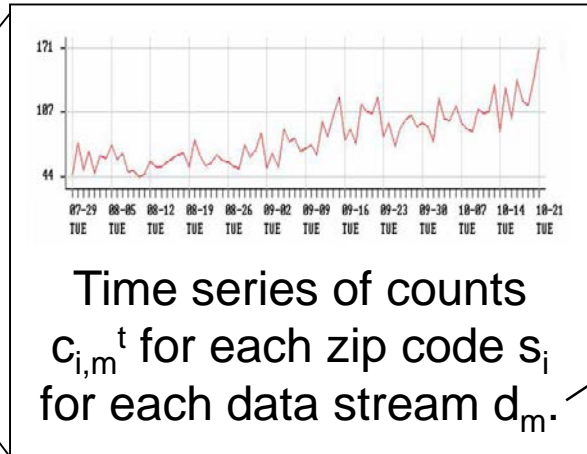
- $D$  = subset of streams
- $S$  = subset of locations
- $W$  = time duration

vs.  $H_0$ : no events occurring

# Multidimensional event detection



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



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 subset scan

- Our **MD-Scan** framework (Neill & 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  $A$  (randomly or sequentially) and use LTSS to find the highest scoring subset of values, locations, or streams, conditioned on all other attributes.  
**\*\*\* Linear rather than exponential in arity of A \*\*\***
    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.

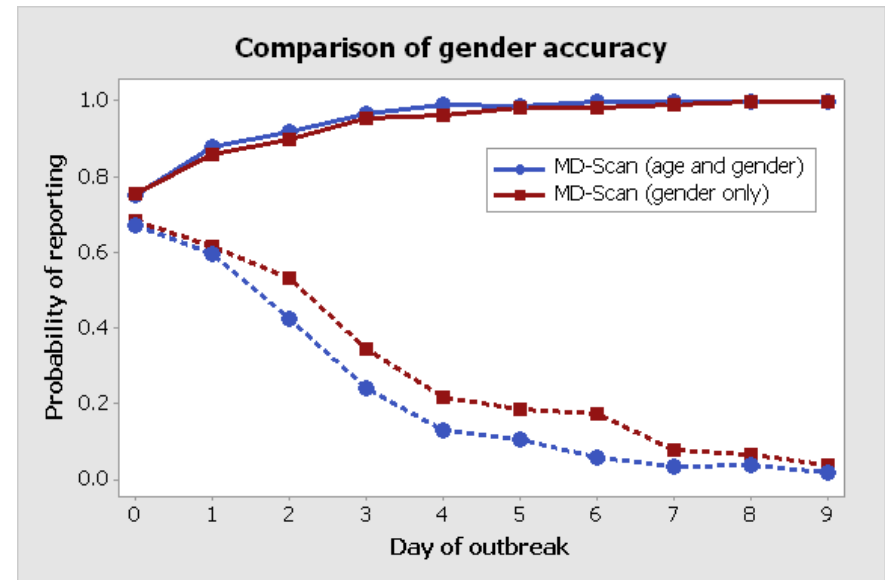
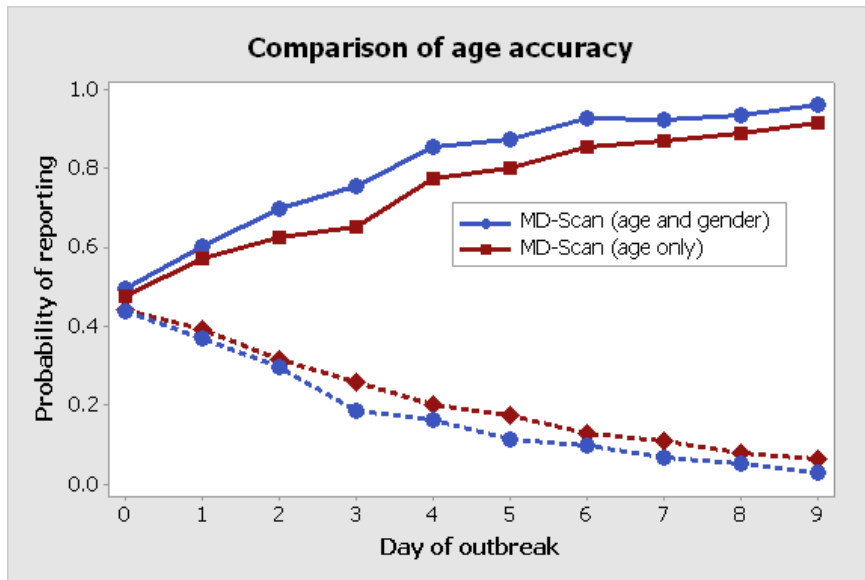
# Evaluation of MD-Scan

- 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.



# 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.**

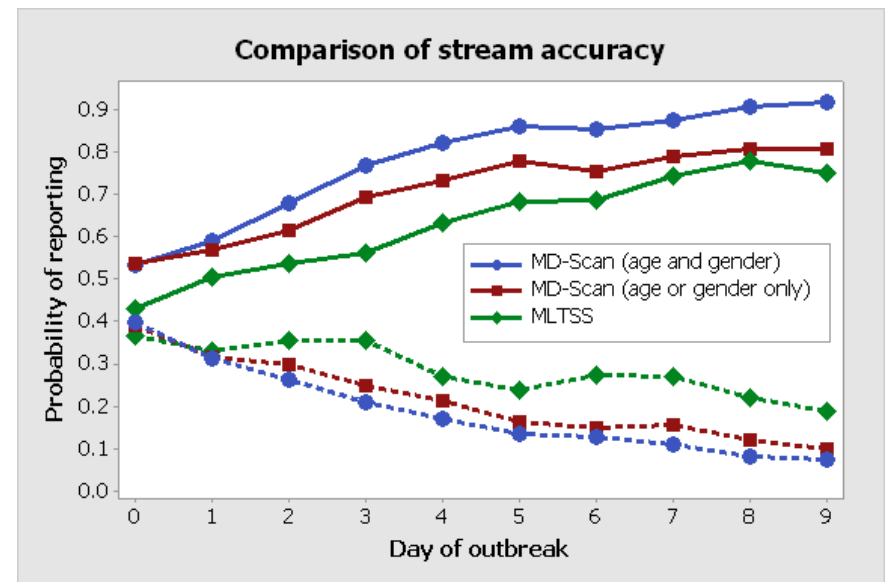
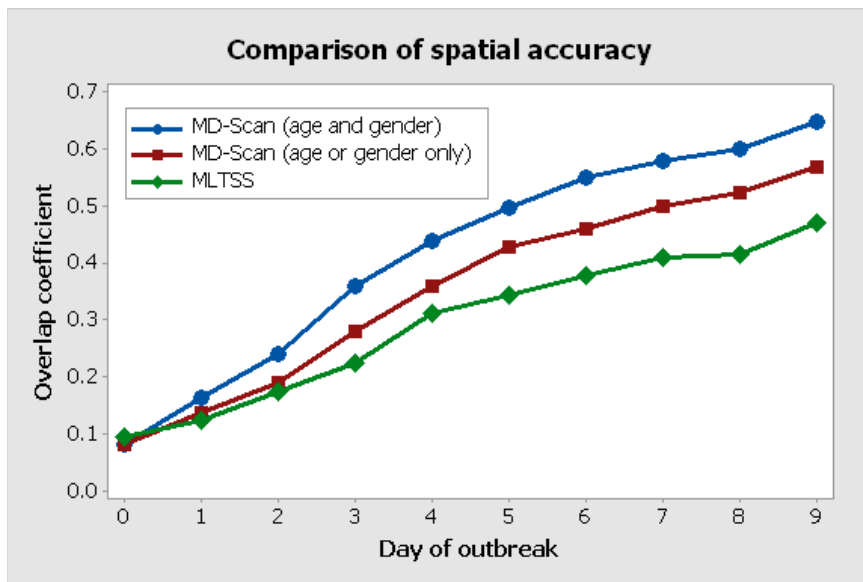
Solid lines: affected gender and/or age deciles. Dashed lines: unaffected.

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.**

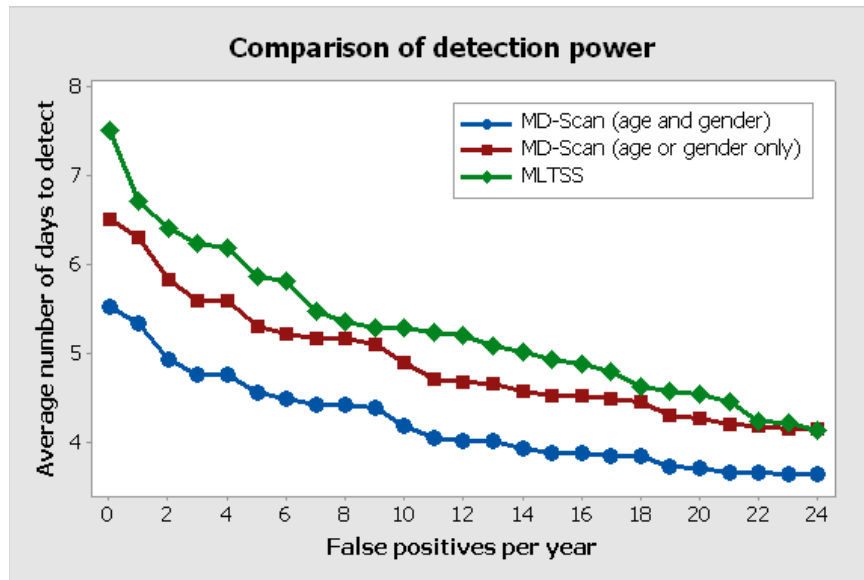
Blue lines: outbreaks with differential effects by both age and gender (easier).

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

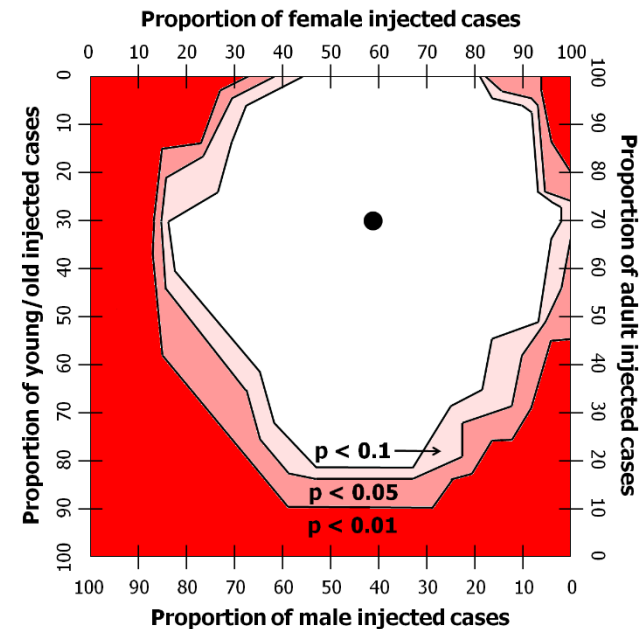
Green lines: MLTSS, ignoring age and gender information

# 3) Timeliness of outbreak detection

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



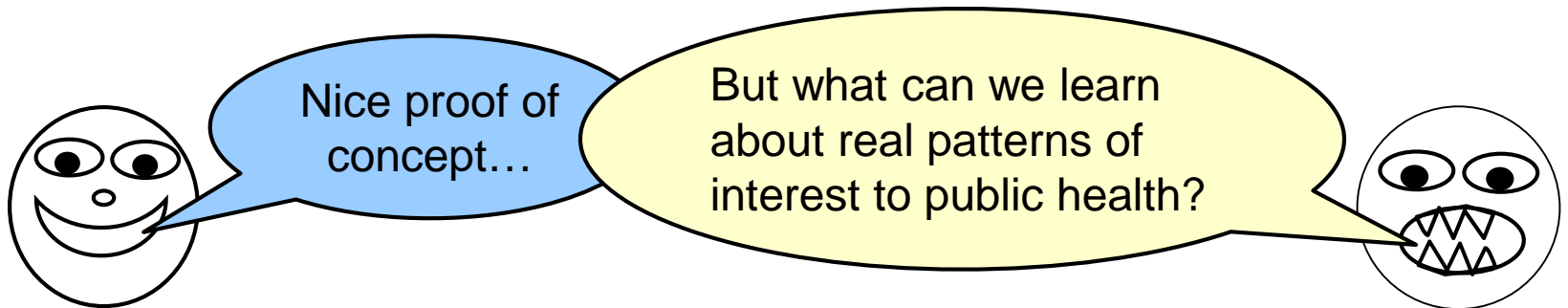
For outbreaks with strong age and gender biases, time to detection improved from 5.2 to 4.0 days at a fixed false positive rate of 1/month.



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.

# 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.



# 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 Dept. of Human Services.

# 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?

# Incorporating unstructured data

<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

Free-text ED chief complaint data from hospitals in NYC and North Carolina.

Key challenge: public health agencies must be able to identify relevant clusters of disease cases that may not correspond to known syndromes (e.g., rare or novel outbreaks)



# 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

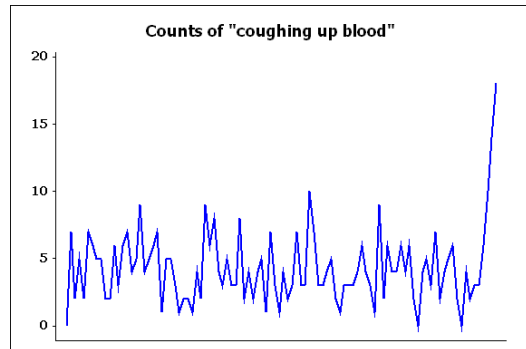
diarrhea

fever

# Where do existing methods fail?

The typical syndromic surveillance approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

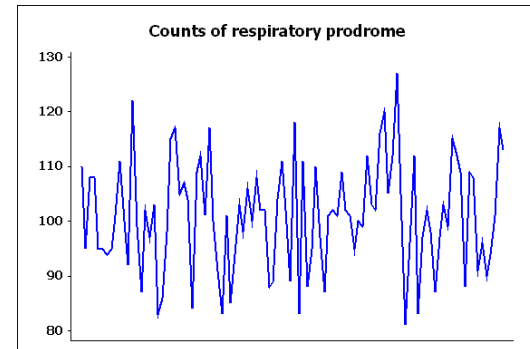
If we were monitoring these particular symptoms, it would only take a few such cases to realize that an outbreak is occurring!



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.

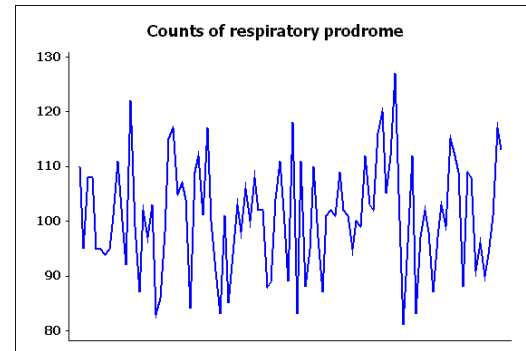
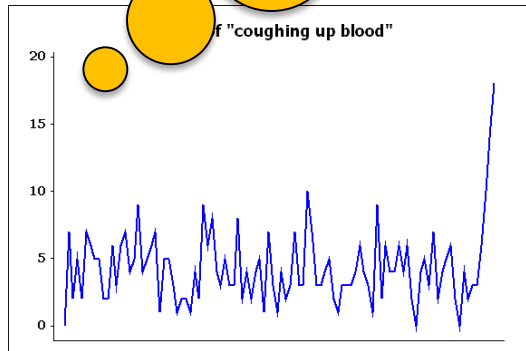


# Where do existing methods fail?

The typical surveillance system is designed to detect when something is going on along with the symptoms (e.g., "coughing up blood" or "shortness of breath") and then to alert the system (e.g., "off").

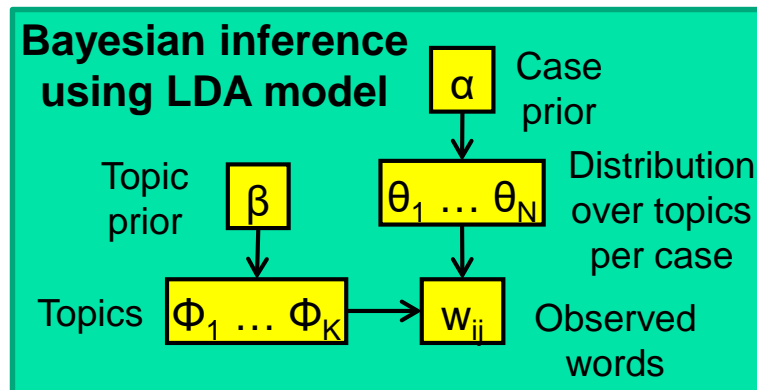
Our solution is to combine text-based (topic modeling) and event detection (multidimensional scan) approaches, to detect **emerging patterns of keywords.**

If we were to monitor a particular symptom category, we would take a few such symptoms to estimate the outbreak signal, that an outbreak is occurring! This is a challenging task, requiring a way of preventing detection.



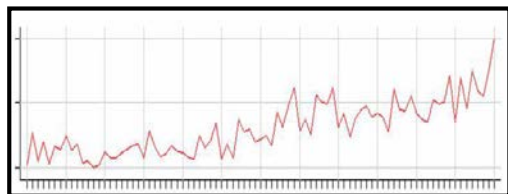
# The semantic scan statistic

<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 syndromes!

# Multidimensional scanning

For each hour of data:

For each combination  $S$  of:

- Hospital
- Time duration
- 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).

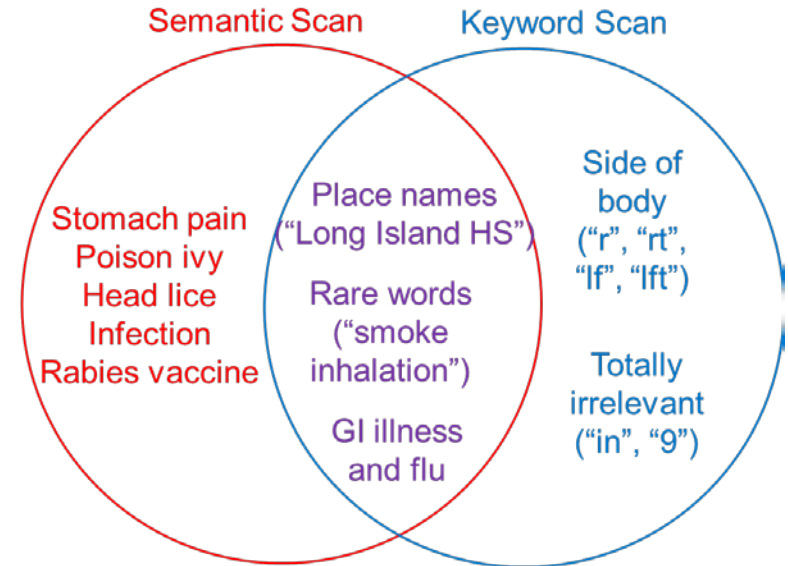
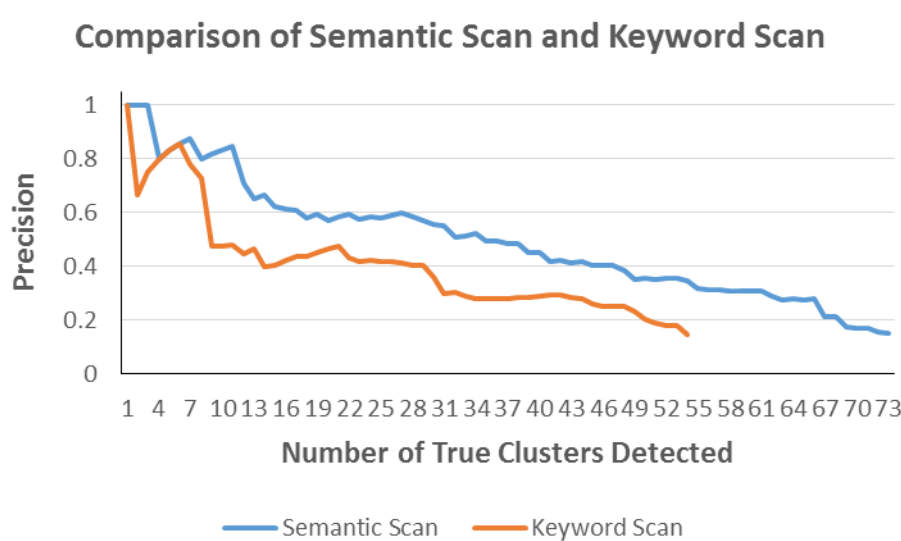
**Score:**  $F(S) = C \log (C/B) + B - C$ , if  $C > B$ , and 0 otherwise (using the expectation-based Poisson likelihood ratio statistic)

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



# NC DOH evaluation results

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

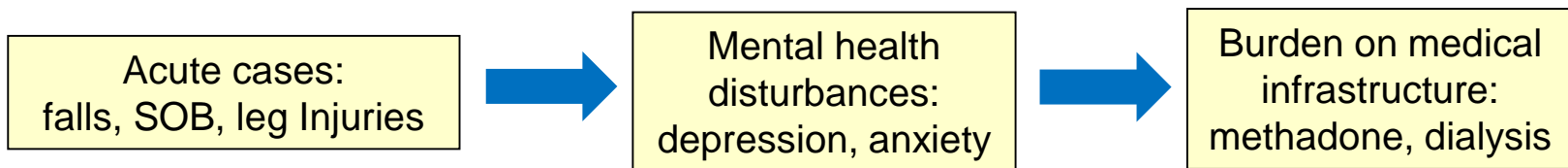
Keyword scan: for 10 true clusters, had to report 21;  
for 30 true clusters, had to report 83.

# NYC DOHMH dataset

- New York City's Department of Health and Mental Hygiene provided us with 5 years of data (2010-2014) consisting of ~20M chief complaint cases from 50 hospitals in NYC.
- For each case, we have data on the patient's chief complaint (free text), date and time of arrival, age group, gender, and discharge ICD-9 code.
- Substantial pre-processing of the chief complaint field was necessary because of size and messiness of data (typos, abbreviations, etc.).
  - Standardized using the Emergency Medical Text Processor (EMTP) developed by Debbie Travers and colleagues at UNC.
  - Spell checker for typo correction.
  - If ICD-9 code in chief complaint field, convert to corresponding text.

# Events identified by semantic scan

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

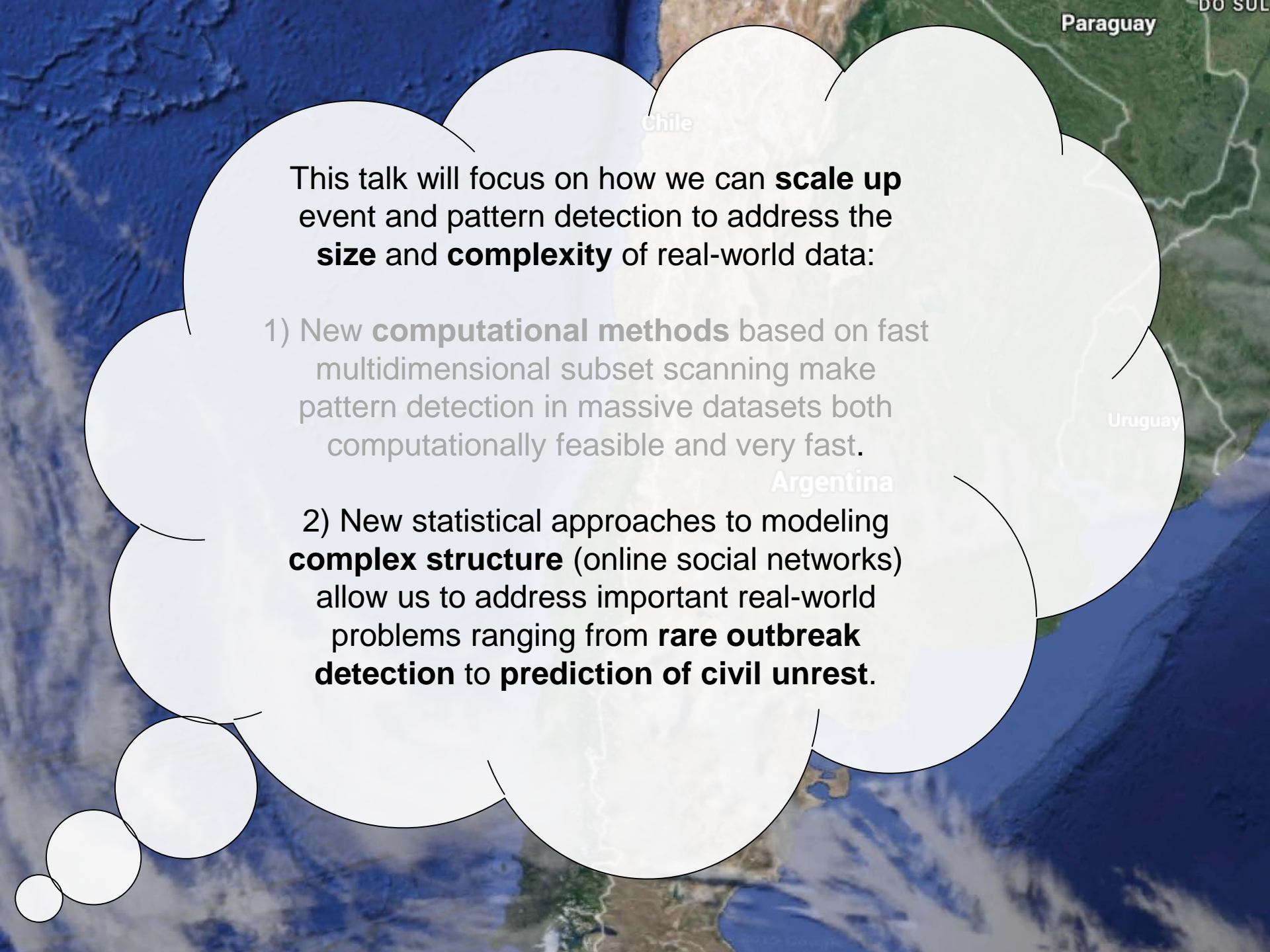


Many other events of public health interest were identified:

<b>Accidents</b>
Motor vehicle
Ferry
School bus
Elevator

<b>Contagious Diseases</b>
Meningitis
Scabies
Ringworm

<b>Other</b>
Drug overdoses
Smoke inhalation
Carbon monoxide poisoning
Crime related, e.g., pepper spray attacks



This talk will focus on how we can **scale up** event and pattern detection to address the **size** and **complexity** of real-world data:

1) New **computational methods** based on fast multidimensional subset scanning make pattern detection in massive datasets both computationally feasible and very fast.

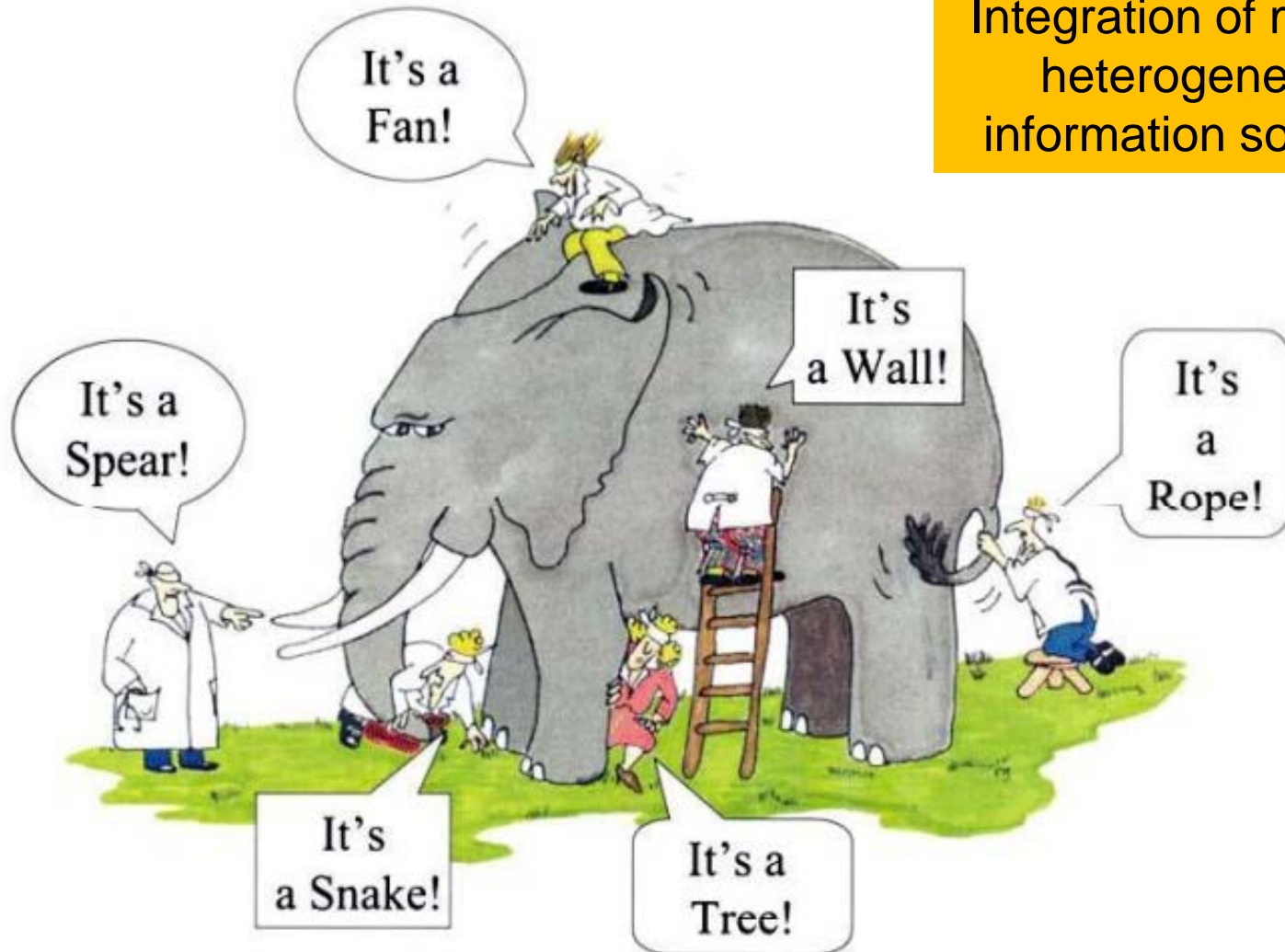
2) New statistical approaches to modeling **complex structure** (online social networks) allow us to address important real-world problems ranging from **rare outbreak detection** to **prediction of civil unrest**.





# Technical Challenges

Integration of multiple heterogeneous information sources!



# Technical Challenges

One week before Mexico's 2012 presidential election:

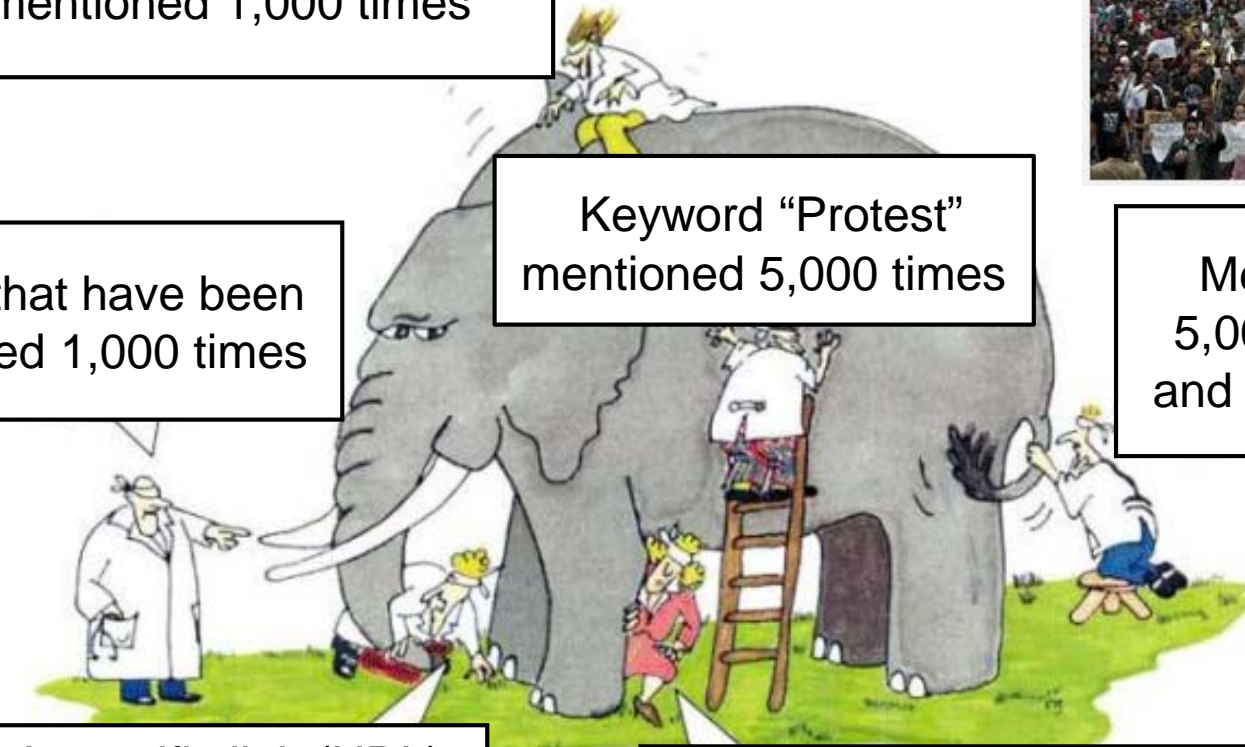
Hashtag "#Megamarch"  
mentioned 1,000 times



Tweets that have been  
re-tweeted 1,000 times

Keyword "Protest"  
mentioned 5,000 times

Mexico City has  
5,000 active users  
and 100,000 tweets



A specific link (URL)  
was mentioned  
866 times

Influential user "Zeka"  
posted 10 tweets



# Technical Challenges

One week before Mexico's 2012 presidential election:

Hashtag "#Megamarch"  
mentioned 1,000 times



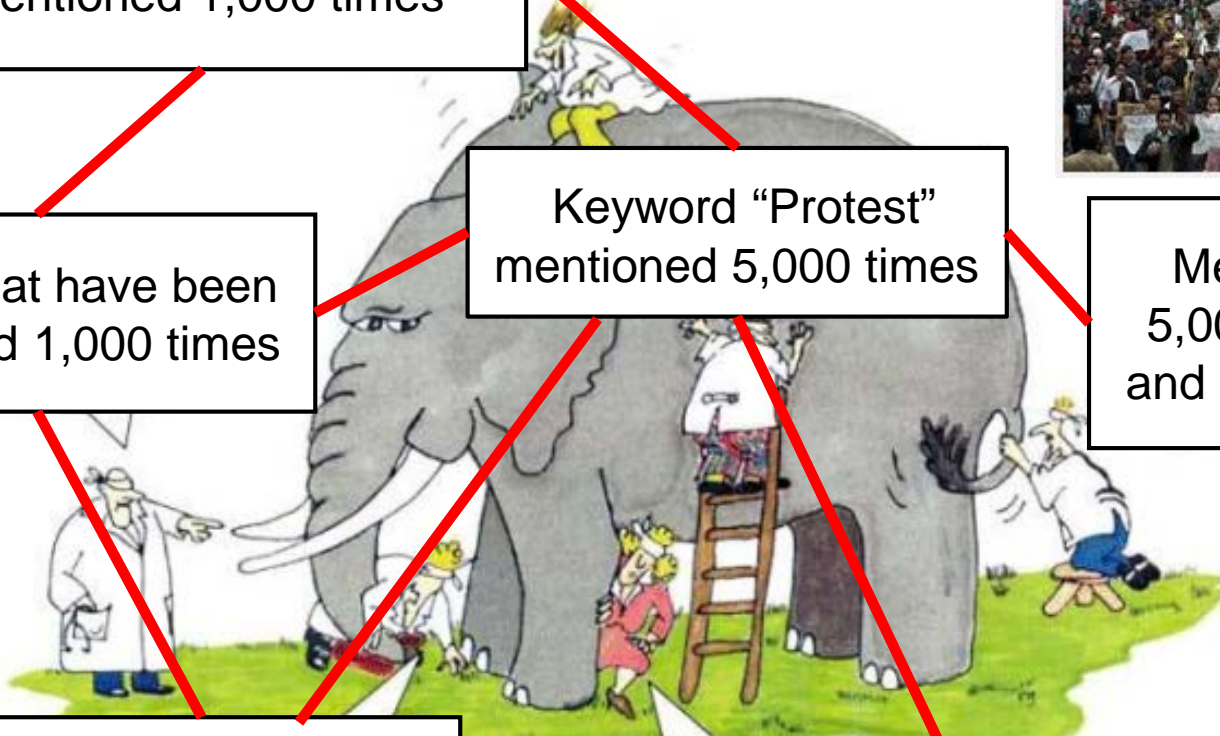
Tweets that have been  
re-tweeted 1,000 times

Keyword "Protest"  
mentioned 5,000 times

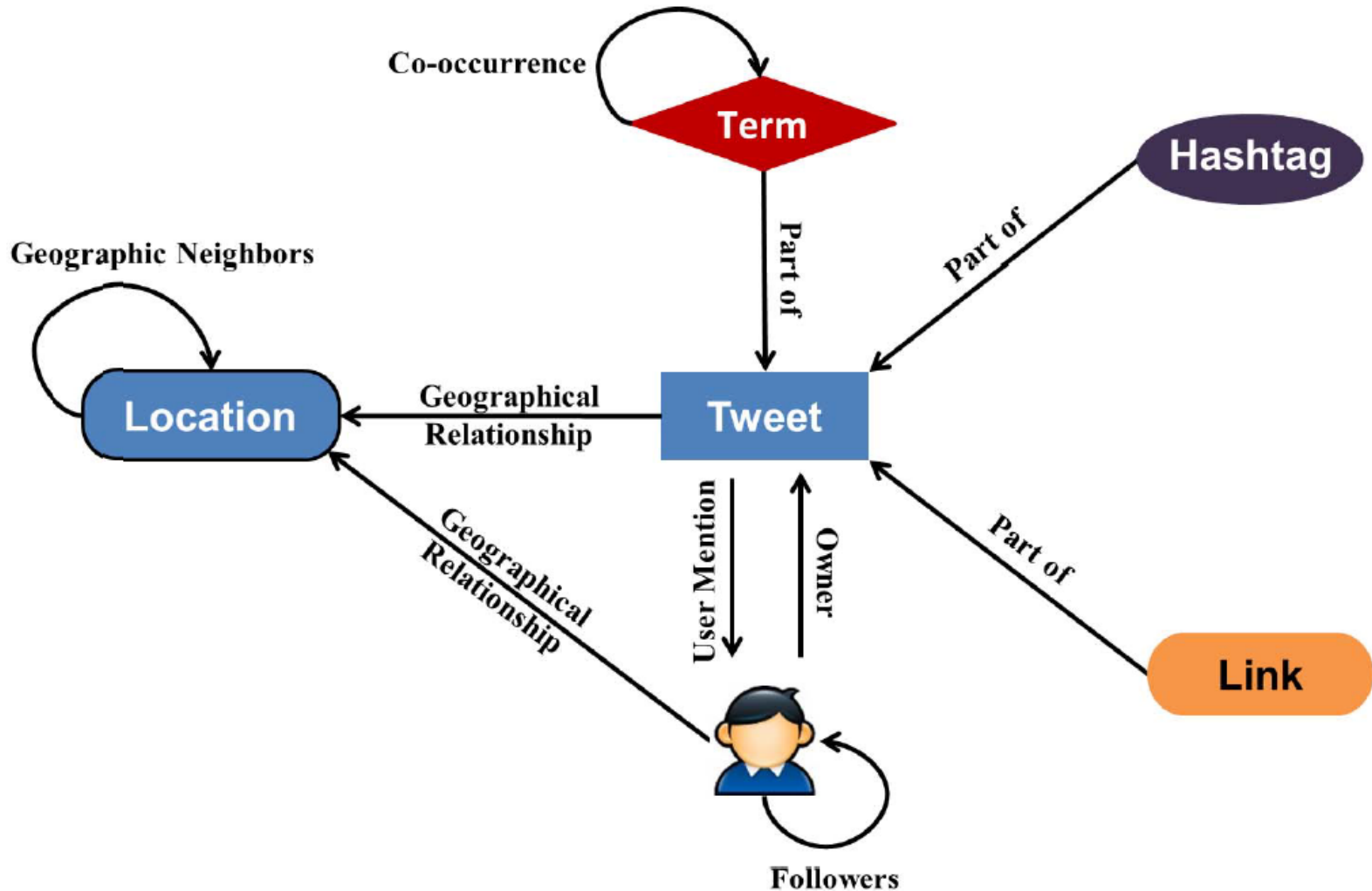
Mexico City has  
5,000 active users  
and 100,000 tweets

A specific link (URL)  
was mentioned  
866 times

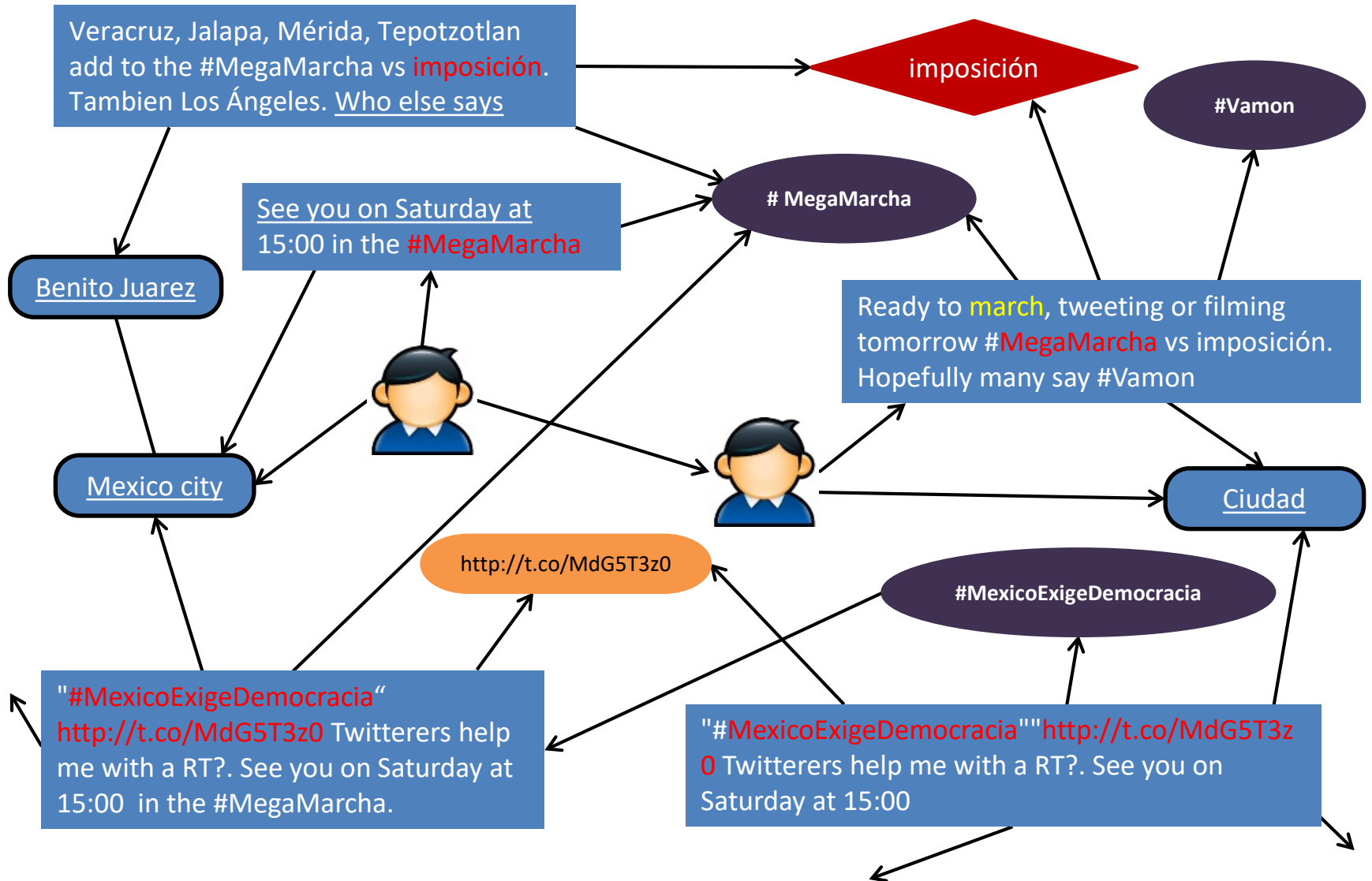
Influential user "Zeka"  
posted 10 tweets



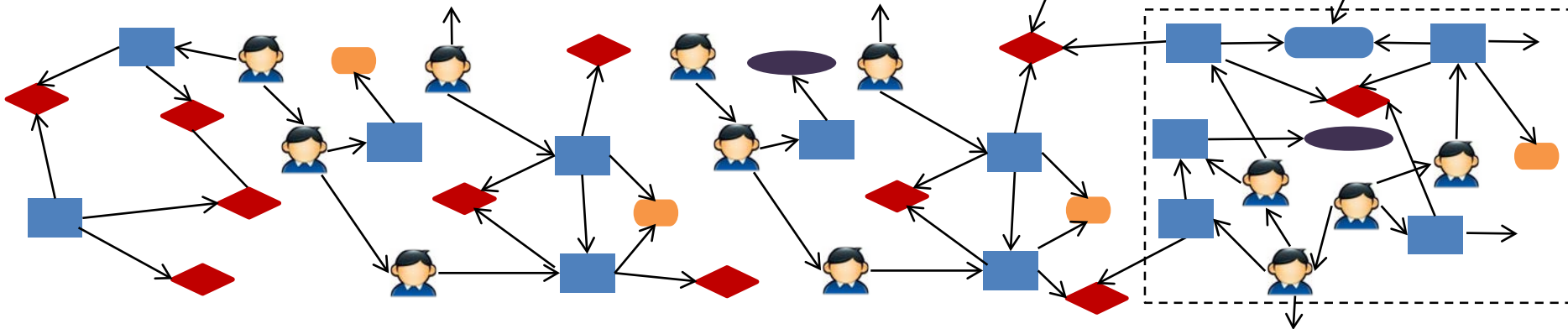
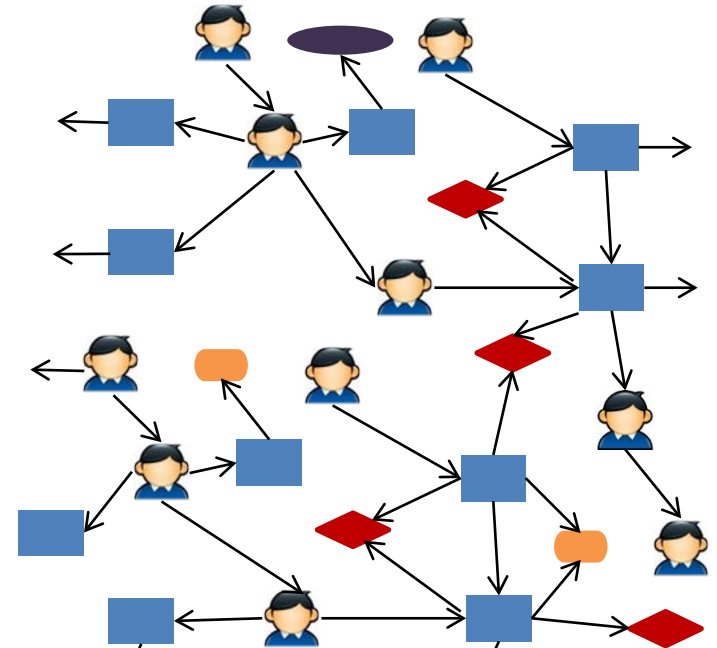
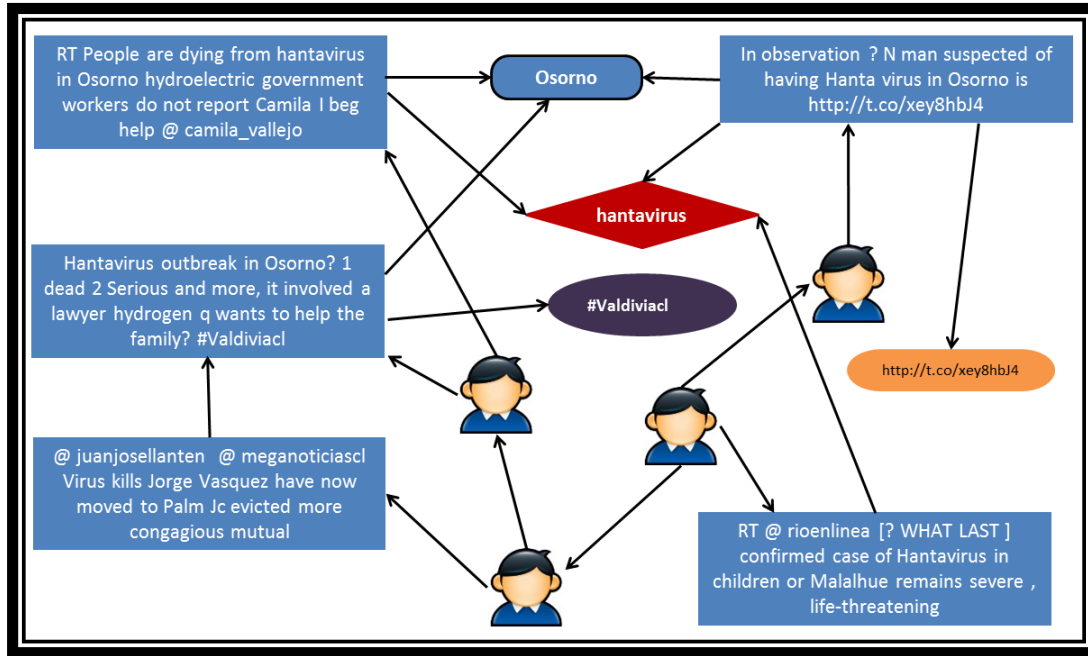
# Twitter Heterogeneous Network



# Twitter Heterogeneous Network



# Twitter Heterogeneous Network



# Nonparametric Heterogeneous Graph Scan

(Chen and Neill, KDD 2014)

1) We model the heterogeneous social network as a **sensor network**.

Each node senses its local neighborhood, computes multiple features, and reports the overall degree of anomalousness.

2) We compute an **empirical p-value** for each node:

- Uniform on  $[0,1]$  under the null hypothesis of no events.
- We search for subgraphs of the network with a higher than expected number of low (significant) empirical p-values.

3) We can scale up to very large heterogeneous networks:

- Heuristic approach: **iterative subgraph expansion** (“greedy growth” to subset of neighbors on each iteration).
- LTSS can efficiently find the best subset of neighbors, ensuring that the subset remains connected, at each step.

# Sensor network modeling

Each node reports an empirical p-value measuring the current level of anomalousness for each time interval (hour or day).

Object Type	Features
User	# tweets, # retweets, # followers, #followees, #mentioned_by, #replied_by, diffusion graph depth, diffusion graph size
Tweet	Klout, sentiment, replied_by_graph_size, reply_graph_size, retweet_graph_size, retweet_graph_depth
City, State, Country	# tweets, # active users
Term	# tweets
Link	# tweets
Hashtag	# tweets

Features

empirical  
calibration

Individual p-value  
for each feature

min

Minimum  
empirical p-  
value for  
each node

empirical  
calibration

Overall p-value  
for each node

# Nonparametric scan statistics

Number of nodes in  $S$  with p-values  $\leq \alpha$ .

Subgraph

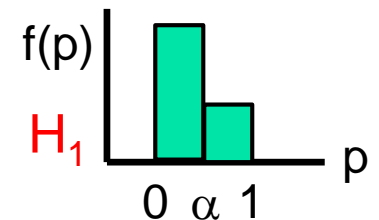
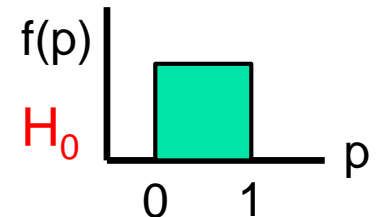
$$F(S) = \max_{\alpha \leq \alpha_{max}} F_\alpha(S) = \max_{\alpha \leq \alpha_{max}} \phi(\alpha, N_\alpha(S), N(S))$$

Significance level

Number of nodes in  $S$

Berk-Jones (BJ) statistic:

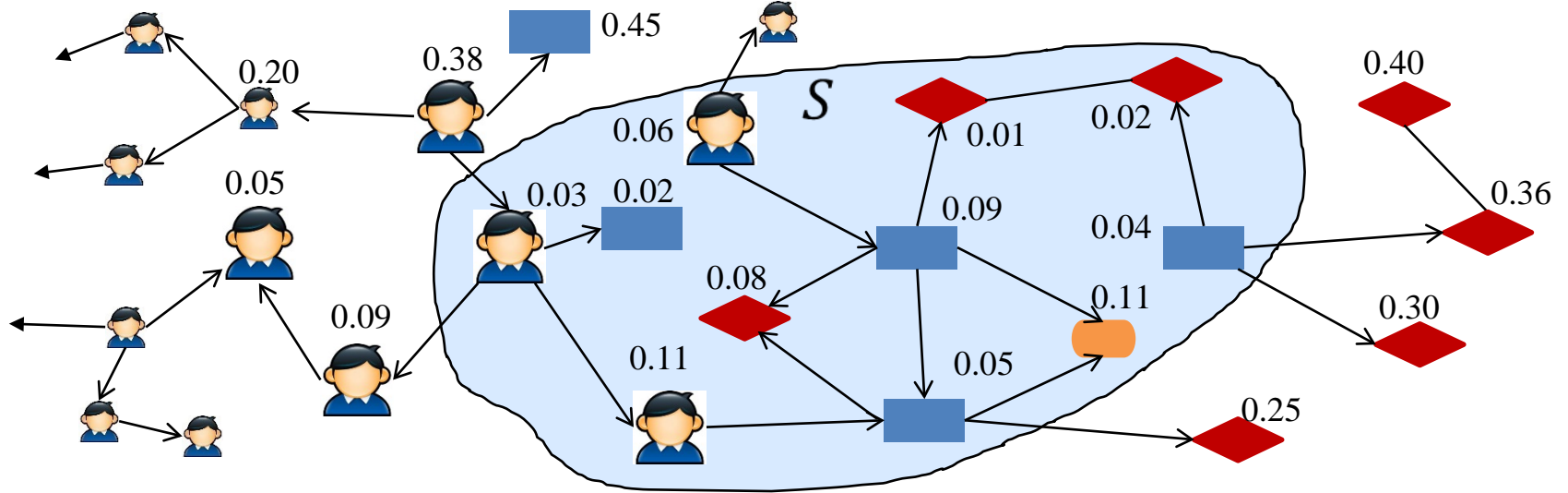
$$\phi_{BJ}(\alpha, N_\alpha(S), N(S)) = N(S)K\left(\frac{N_\alpha}{N}, \alpha\right)$$



Kullback-Liebler divergence:

$$K(x, y) = x \log\left(\frac{x}{y}\right) + (1 - x) \log\left(\frac{1 - x}{1 - y}\right)$$

# Nonparametric graph scanning



$$S^* = \operatorname{argmax}_{S \in V: S \text{ is connected}} F(S)$$

We propose an approximate algorithm with time cost  $O(|V| \log |V|)$ .



# NPHGS evaluation- civil unrest

Country	# of tweets	News source*
Argentina	29,000,000	Clarín; La Nación; Infobae
Chile	14,000,000	La Tercera; Las Últimas Noticias; El Mercurio
Colombia	22,000,000	El Espectador; El Tiempo; El Colombiano
Ecuador	6,900,000	El Universo; El Comercio; Hoy

**Gold standard dataset:** 918 civil unrest events between July and December 2012.

Example of a gold standard event label:

PROVINCE = “El Loa”

COUNTRY = “Chile”

DATE = “2012-05-18”

LINK = “<http://www.pressenza.com/2012/05/...>”

DESCRIPTION = “A large-scale march was staged by inhabitants of the northern city of Calama, considered the mining capital of Chile, who demanded the allocation of more resources to copper mining cities”

We compared the detection performance of our NPHGS approach to homogeneous graph scan methods and to a variety of state-of-the-art methods previously proposed for Twitter event detection.

# NPHGS results- civil unrest

Method	FPR (FP/Day)	TPR (Forecasting)	TPR (Forecasting & Detection)	Lead Time (Days)	Lag Time (Days)	Run Time (Hours)
ST Burst Detection	0.65	0.07	0.42	1.10	4.57	30.1
Graph Partition	0.29	0.03	0.15	0.59	6.13	18.9
Earthquake	0.04	0.06	0.17	0.49	5.95	18.9
RW Event	0.10	0.22	0.25	0.93	5.83	16.3
Geo Topic Modeling	0.09	0.06	0.08	0.01	6.94	9.7
NPHGS (FPR=.05)	0.05	0.15	0.23	0.65	5.65	38.4
NPHGS (FPR=.10)	0.10	0.31	0.38	1.94	4.49	38.4
NPHGS (FPR= .15)	0.15	0.37	0.42	2.28	4.17	38.4
NPHGS (FPR=.20)	0.20	0.39	0.46	2.36	3.98	38.4

Table 3: Comparison between NPHGS and Existing Methods on the civil unrest datasets

NPHGS outperforms existing representative techniques for both event detection and forecasting, increasing **detection power**, **forecasting accuracy**, and **forecasting lead time** while reducing **time to detection**.

Similar improvements in performance were observed on a second task:

Early detection of rare disease outbreaks, using gold standard data about 17 hantavirus outbreaks from the Chilean Ministry of Health.

Detected Hantavirus outbreak, 10 Jan 2013

First news report:  
11 Jan 2013



Temuco and Villarrica, Chile

- Locations
- Users
- Keywords
- Hashtags
- Links
- Videos

# NPHGS results- human rights

We performed an exploratory analysis of human rights-related events in Mexico from January 2013 to June 2014, using Twitter data (10% sample, filtered using relevant keywords).

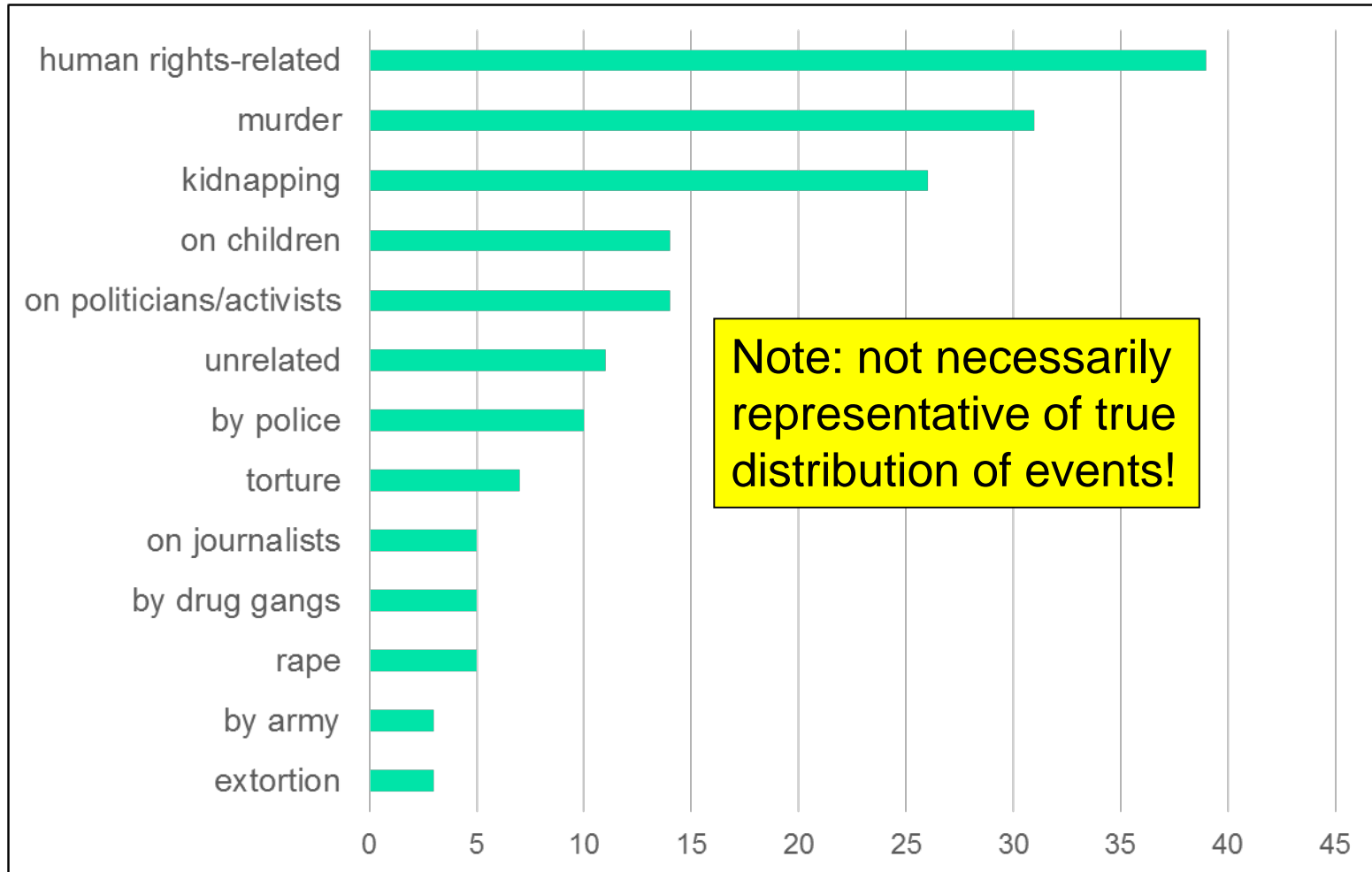
The top 50 identified clusters over the entire study period were analyzed manually to identify:

- (1) whether the cluster was human rights related
- (2) the types of human rights violations
- (3) the victims of the violations
- (4) the alleged perpetrators.

NPHGS was able to identify some human rights events of interest before international news sources...  
... and in some cases, before local news sources.

# Cluster characteristics

(top-50 detected clusters)





# Conclusions

Real-world problems at the societal scale require new computational methods to deal with both the **size** and the **complexity** of data.



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

Practical solutions to societal challenges also require an understanding of complex data (text, networks, images, streams, ...), leading to **new statistical and algorithmic tools** for extracting relevant patterns.

# Acknowledgements

- Event and Pattern Detection Laboratory (current members and alumni):  
<http://epdlab.heinz.cmu.edu/people>
- Students, postdocs, and collaborators:  
Feng Chen, Skyler Speakman, Ed McFowland, Sriram Somanchi, Tarun Kumar, Kenton Murray, Yandong Liu, Chris Dyer, Jay Aronson.
- Funding support: NSF and MacArthur 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.
- S. Speakman, S. Somanchi, E. McFowland III, D.B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics* 25: 382-404, 2016.
- 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.
- F. Chen and D.B. Neill. Human rights event detection from heterogeneous social media graphs. *Big Data* 3(1): 34-40, 2015.
- T. Kumar and D.B. Neill. Fast tensor scan for event detection and characterization. Revised version in preparation.
- A. Maurya, K. Murray, C. Dyer, Y. Liu, and D.B. Neill. A semantic scan statistic for novel disease outbreak detection. Submitted for publication.



**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)