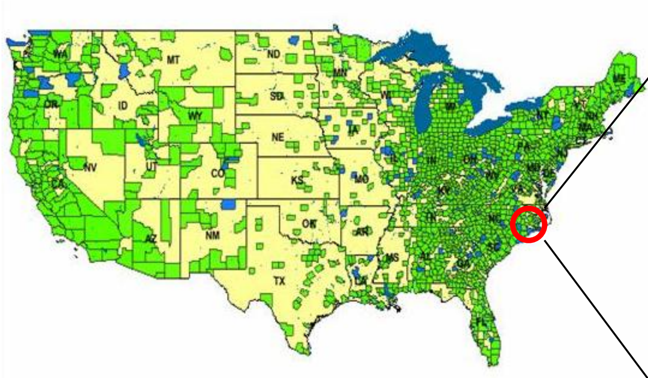
A Pre-Syndromic Surveillance Approach for Early Detection of Novel and Rare Disease Outbreaks

Daniel B. Neill[1,2,*] and Mallory Nobles[1]

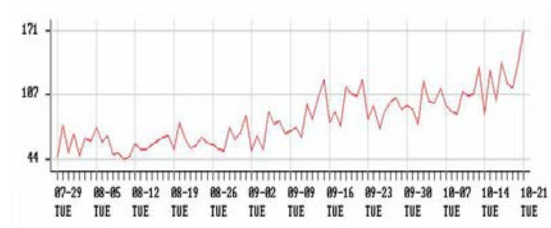
[1]Carnegie Mellon University
[2]New York University
***E-mail: neill@cs.cmu.edu**

This work was partially supported by NSF grant IIS-0953330. Data was provided by the NC DHHS/DPH NC DETECT system and the NYC Department of Health and Mental Hygiene. The NC DETECT Data Oversight Committee and NYC DOHMH do not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

Early outbreak detection (syndromic)



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.

<u>Outbreak detection</u>

d_1 = respiratory ED

d_2 = constitutional ED

d_3 = OTC cough/cold

d_4 = OTC anti-fever

(etc.)

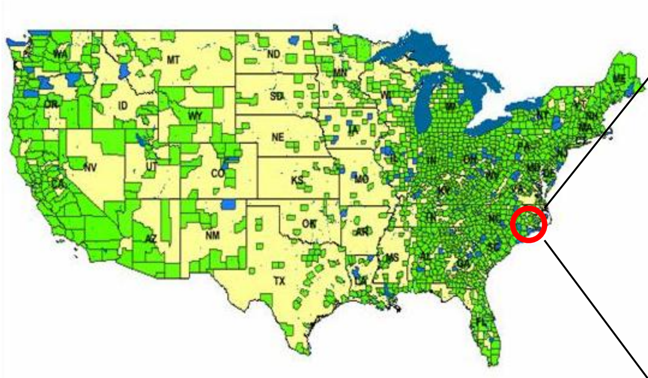
<u>Three main goals of syndromic surveillance</u>

Detect any emerging events

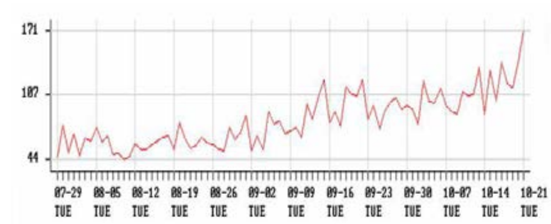
Pinpoint the affected subset of locations and time duration

Characterize the event by identifying the affected subpopulation

Early outbreak detection (syndromic)



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.)

Recent **spatial and subset scanning** approaches can accurately and efficiently find the most anomalous clusters of disease, by maximizing a likelihood ratio statistic over subsets.

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

Compare hypotheses:

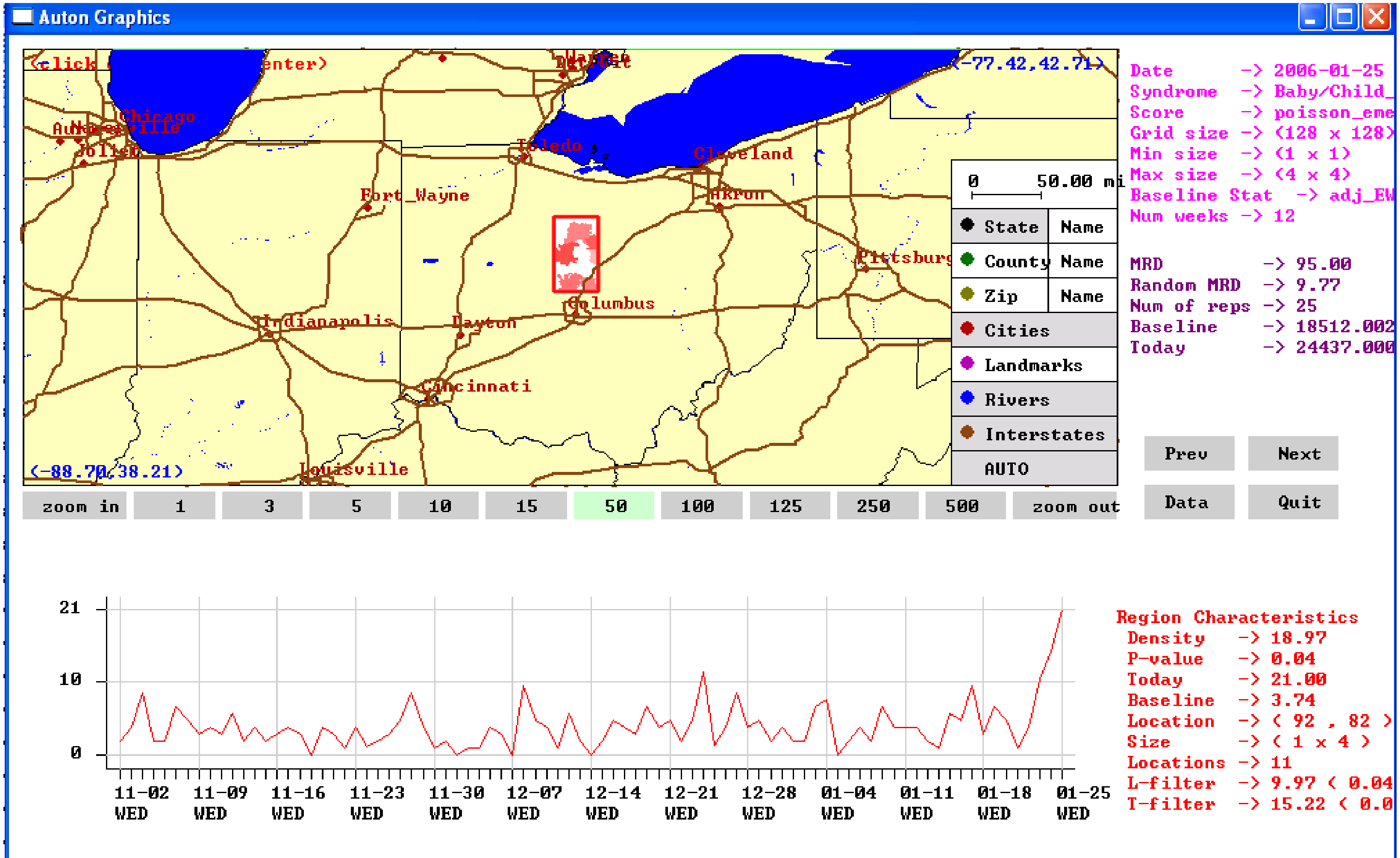
$H_1(D, S, P, W)$

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

vs. H_0: no events occurring

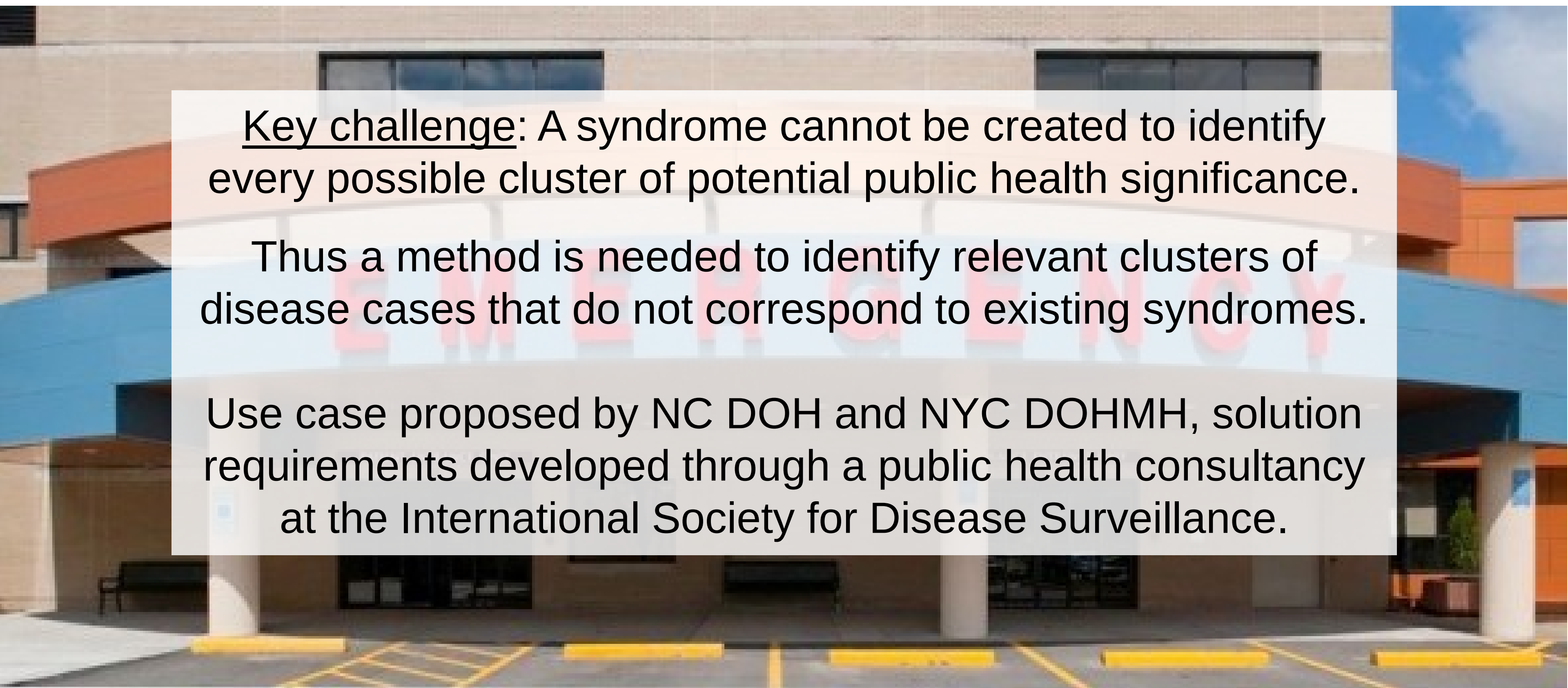
Syndromic surveillance example

Spike in sales of pediatric electrolytes near Columbus, Ohio



Pre-syndromic 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



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

Thus a method is needed to identify relevant clusters of disease cases that do not correspond to existing 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.

Where do existing methods fail?

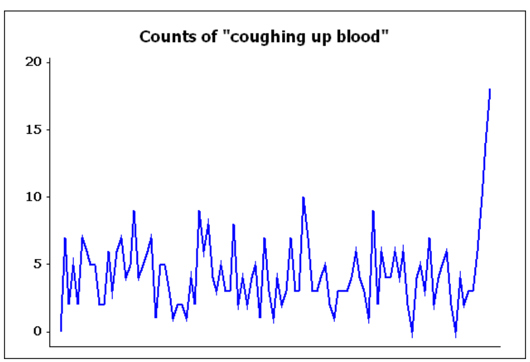
The typical syndromic surveillance approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

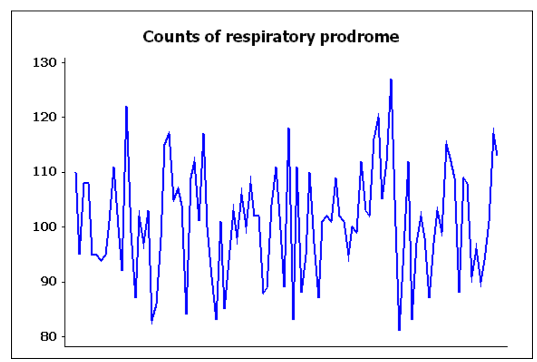
What happens when something new and scary comes along?

- **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.





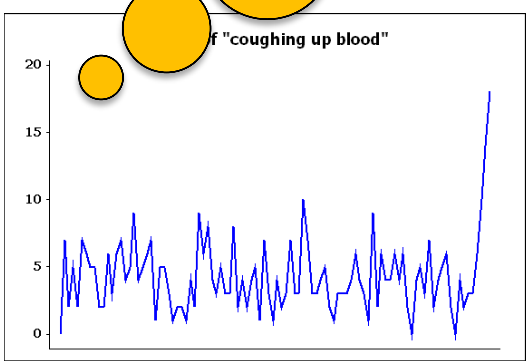
Where do existing methods fail?

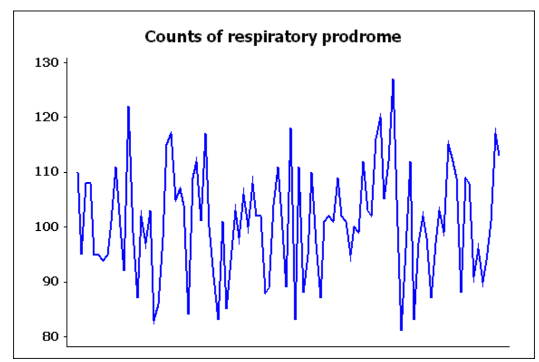
The typical sy... ic
surveillance
effectively
outbrea...
seen...
sy...

...en something
...s 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**.

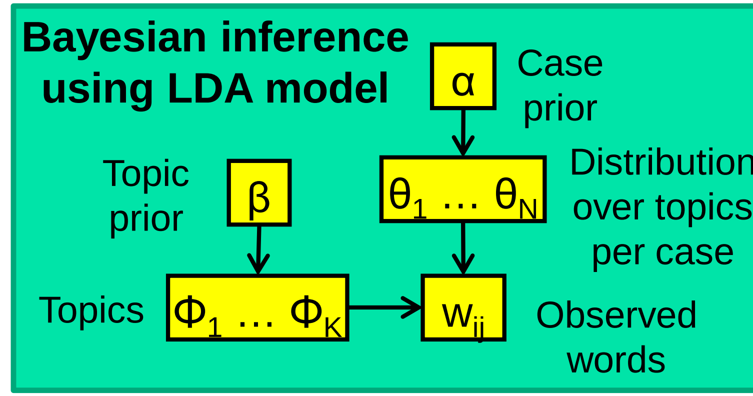
If we w... ...ief complaints
particular sy... ...ptom category
take a few such ... ...ilute the outbreak signal,
that an outb... ...s occurring! ...ying or preventing detection.





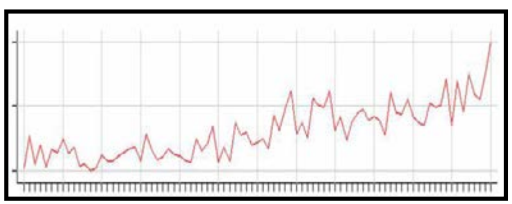
The semantic scan statistic

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 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.

Simulation results

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

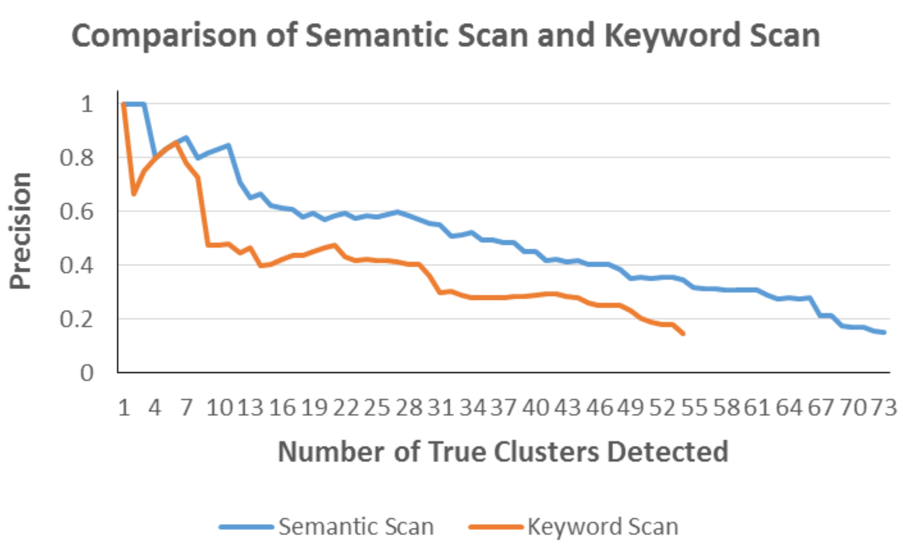
Simulated novel outbreak: “green nose”

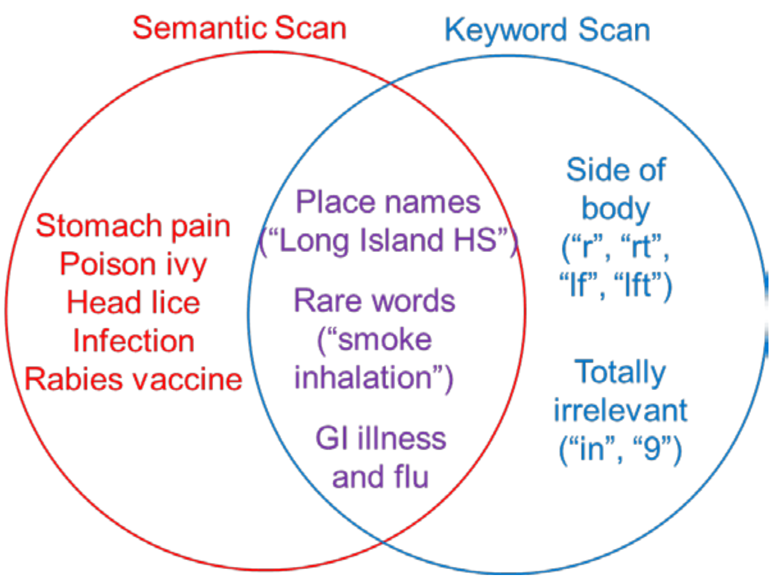
green
nose
possible
color
greenish
nasal
…

Top words from detected topic

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".





<u>Semantic scan</u>: for 10 true clusters, had to report 12;
for 30 true clusters, had to report 54.
<u>Keyword scan</u>: 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:

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

Many other events of public health interest were identified:

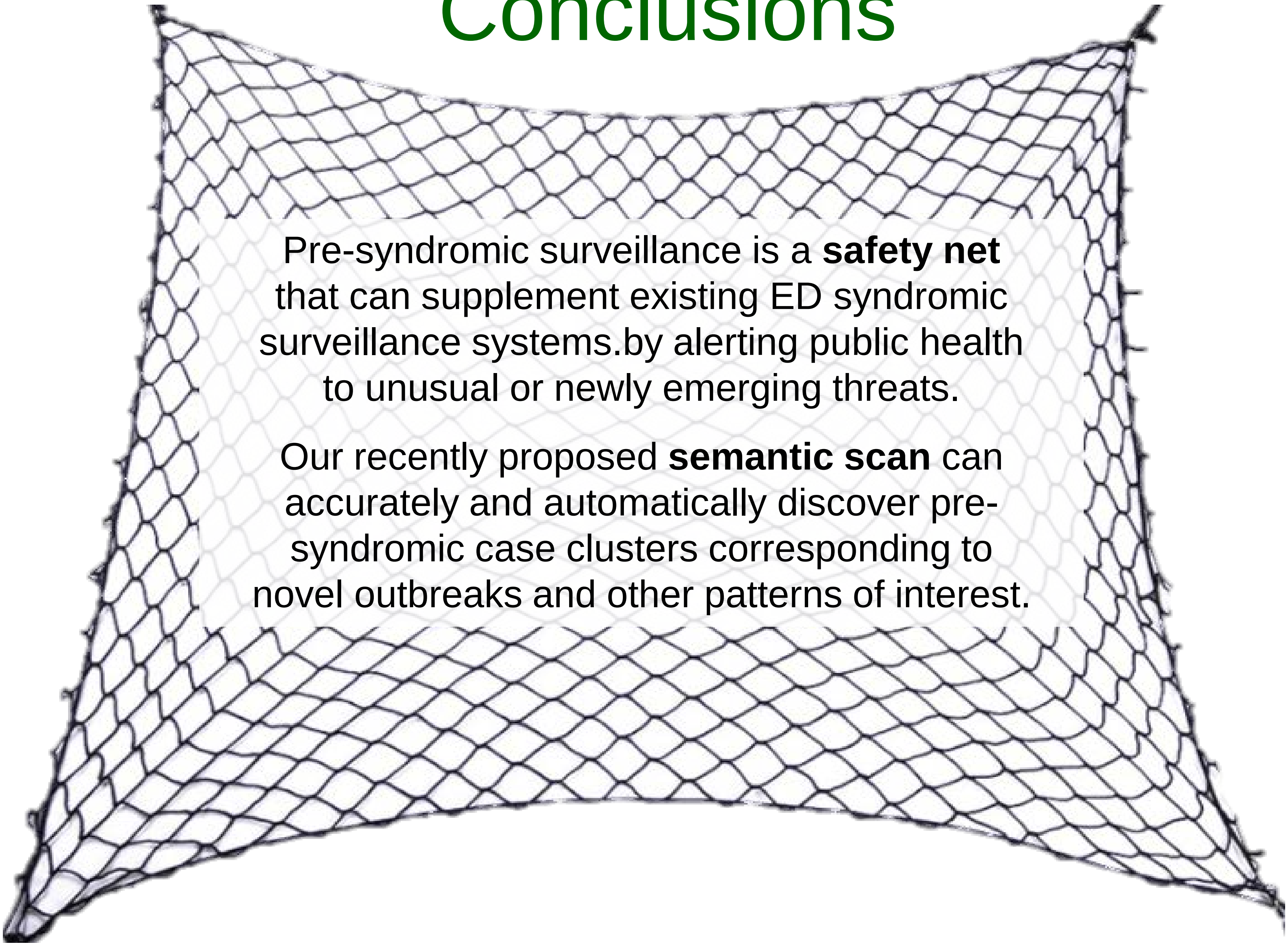
Accidents	Contagious Diseases	Other
Motor vehicle	Meningitis	Drug overdoses
Ferry	Scabies	Smoke inhalation
School bus	Ringworm	Carbon monoxide poisoning
Elevator		Crime related, e.g., pepper spray attacks

Example of a detected cluster

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 represents 9 patients complaining of ingesting tainted coffee, and demonstrates Semantic Scan’s ability to detect rare and novel events.

Conclusions



Pre-syndromic surveillance is a **safety net** that can supplement existing ED syndromic surveillance systems.by alerting public health to unusual or newly emerging threats.

Our recently proposed **semantic scan** can accurately and automatically discover pre-syndromic case clusters corresponding to novel outbreaks and other patterns of interest.

Thanks for listening!

More details on our web site:

http://epdlab.heinz.cmu.edu

Or e-mail me at:

neill@cs.cmu.edu