Novel Machine Learning Methods for Public Health and Disease Surveillance

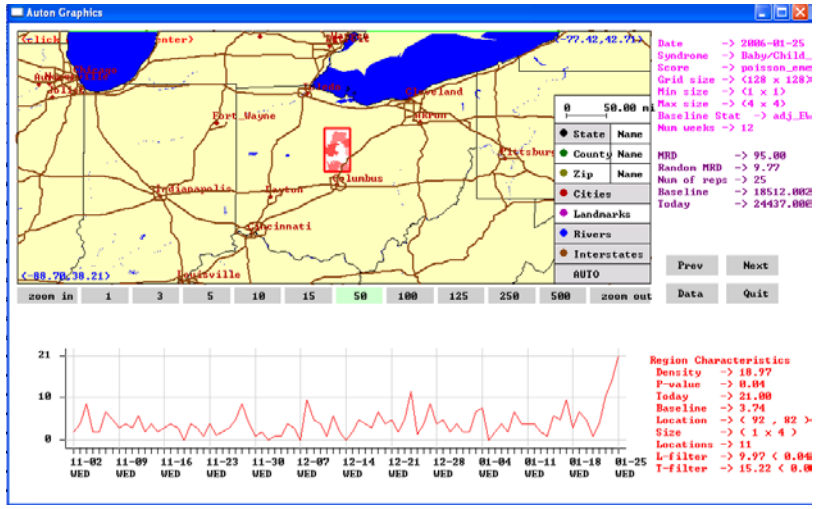
Daniel B. Neill, Ph.D.
Carnegie Mellon University (H.J. Heinz III College)
New York University (Center for Urban Science & Progress)
E-mail: neill@cs.cmu.edu

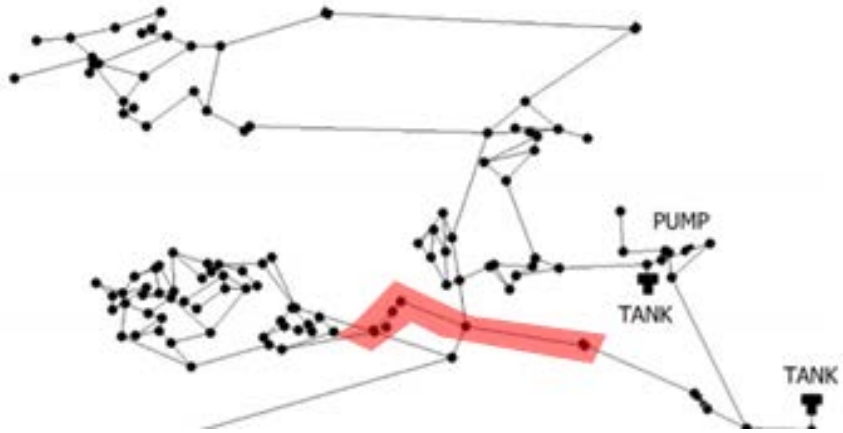
We gratefully acknowledge funding support from the National Science Foundation, grant IIS-0953330, and MacArthur Foundation.

Carnegie Mellon University
EPD Lab
EVENT AND PATTERN DETECTION LABORATORY

What can Machine Learning do for public health surveillance?

- Earlier, more accurate detection of emerging patterns in space and time.
  - Better **estimation** of what we expect to see, and better **detection** of deviations from “expected” (more power, fewer false positives).
  - Integrating **multiple data streams**; incorporating prior information.
  - **Tracking** and **source-tracing** outbreak spread.

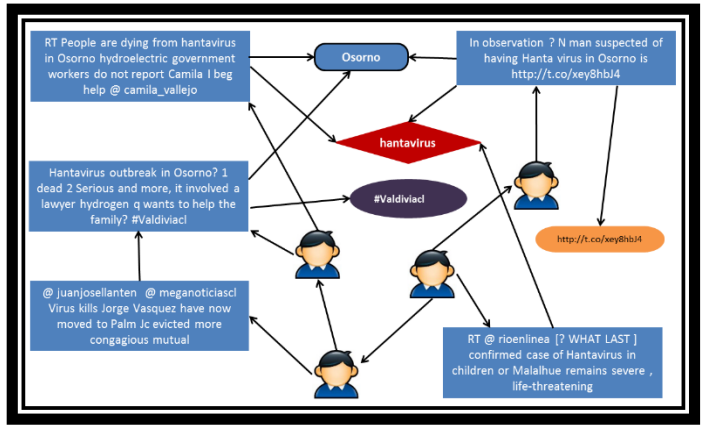




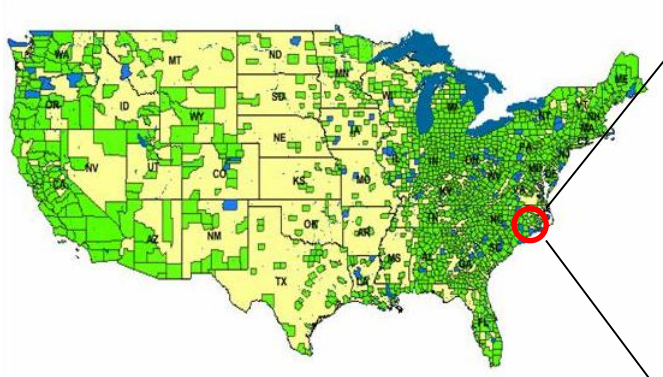
What can Machine Learning do for public health surveillance?

Today's talk will focus on two specific examples of how ML expands the scope and scale of data sources that can be used for surveillance:

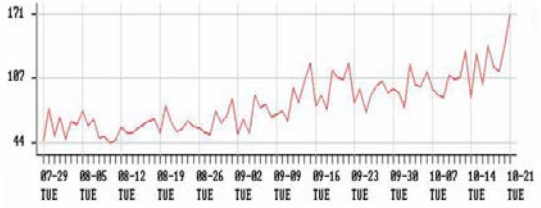
- Structuring **unstructured data**, e.g., free text from ED chief complaints.
- Incorporating complex data sources such as **online social media.**



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.)

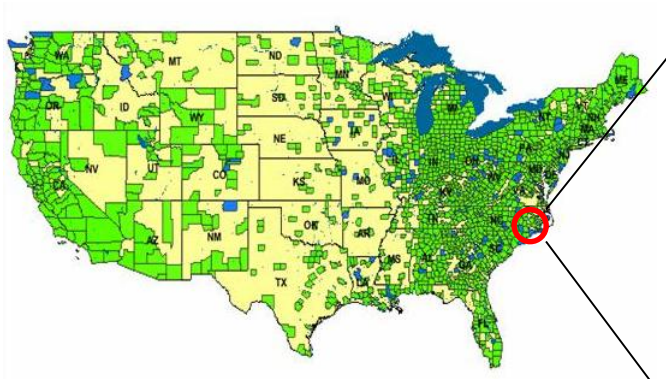
<u>Three main goals of syndromic surveillance</u>

Detect any emerging events

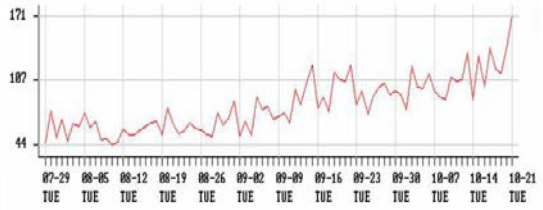
Pinpoint the affected subset of locations and time duration

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.

<u>Outbreak detection</u>

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.

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

<u>Compare hypotheses</u>:

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

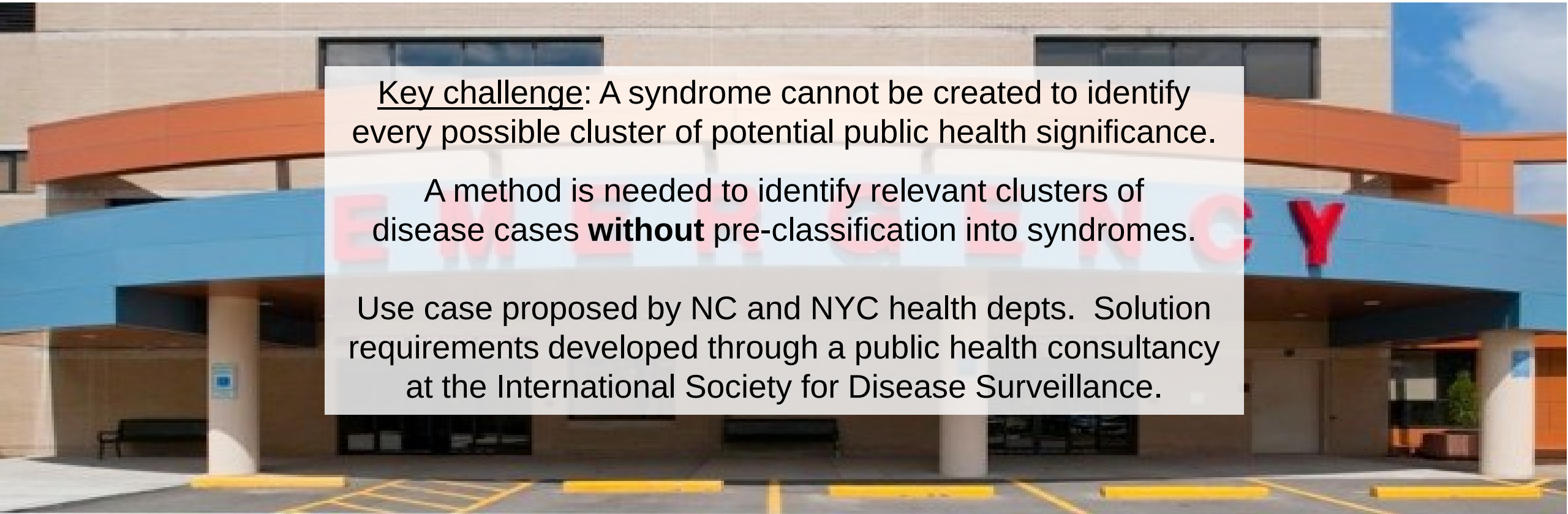
D = subset of streams
S = subset of locations
P = subpopulation
W = time duration

vs. H_0: no events occurring

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

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



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

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

Use case proposed by NC and NYC health depts. Solution requirements developed through a public health consultancy at the International Society for Disease Surveillance.

Where do existing methods fail?

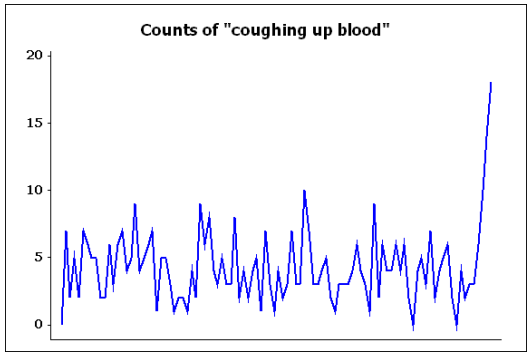
The typical syndromic surveillance approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

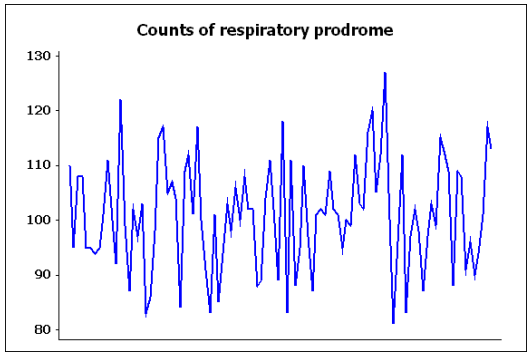
What happens when something new and scary comes along?

- **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.



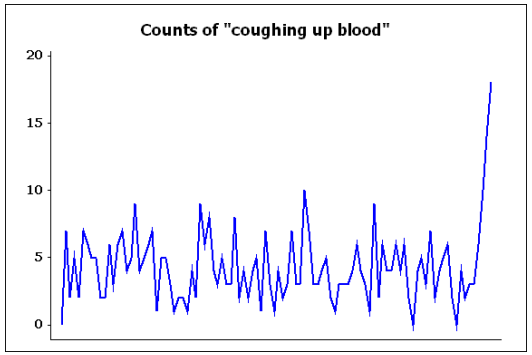


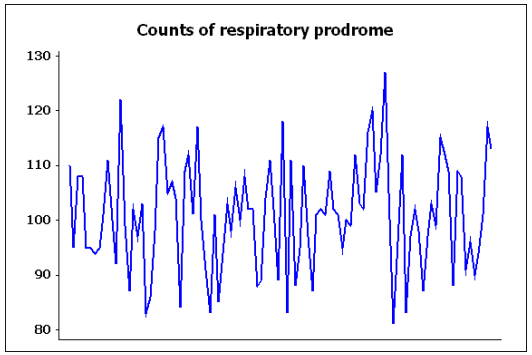
Where do existing methods fail?

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

particular s… take a few such cases to realize th… outbreak is occurring!

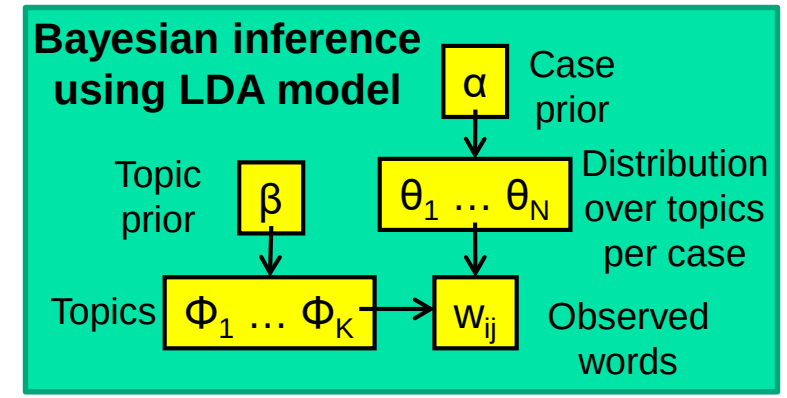
…plaints …der symptom category can dilute the outbreak signal, delaying or preventing detection.





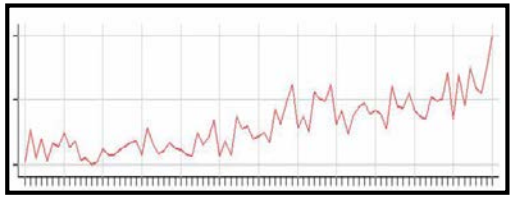
The semantic scan statistic

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



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 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, 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 log-likelihood ratio statistic)

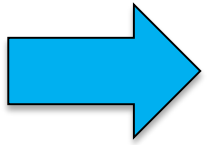
We return cases corresponding to each top-scoring subset S.

Simulation results

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

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”

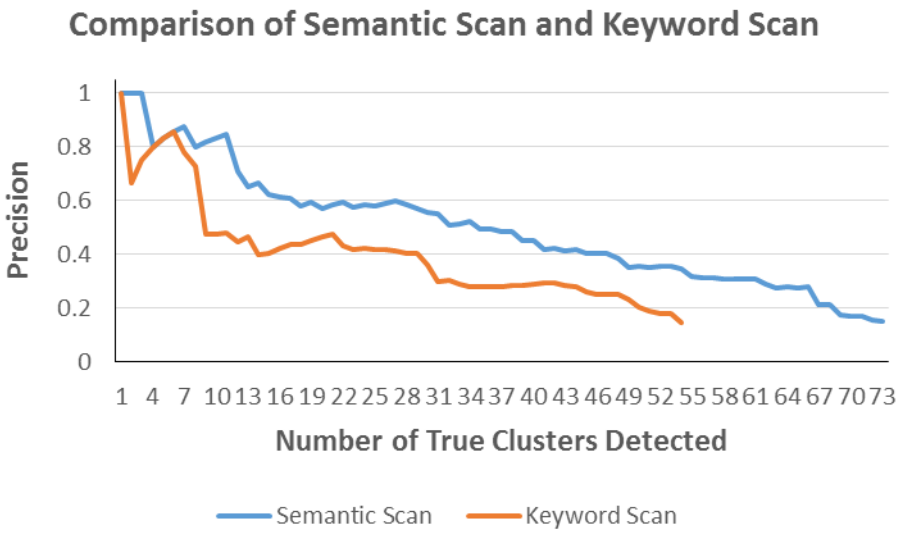


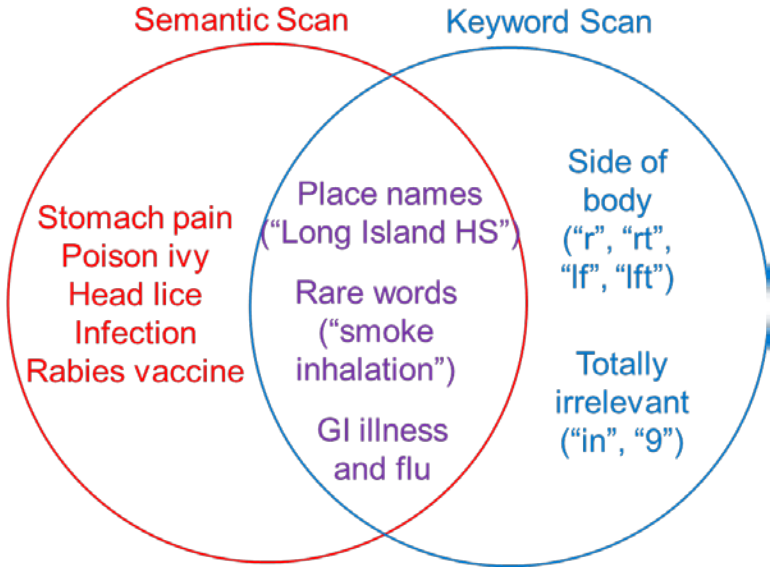
green
nose
possible
color
greenish
nasal
...

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”.





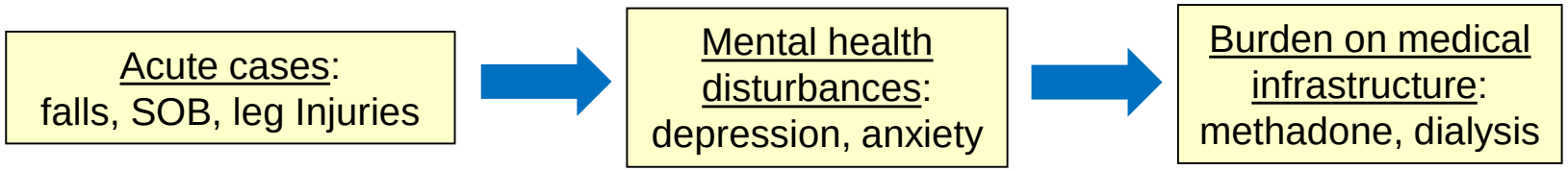
Semantic scan: for 10 true clusters, had to report 12;
for 30 true clusters, had to report 54.
Keyword scan: for 10 true clusters, had to report 21;
for 30 true clusters, had to report 83.

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:

Accidents
Motor vehicle
Ferry
School bus
Elevator

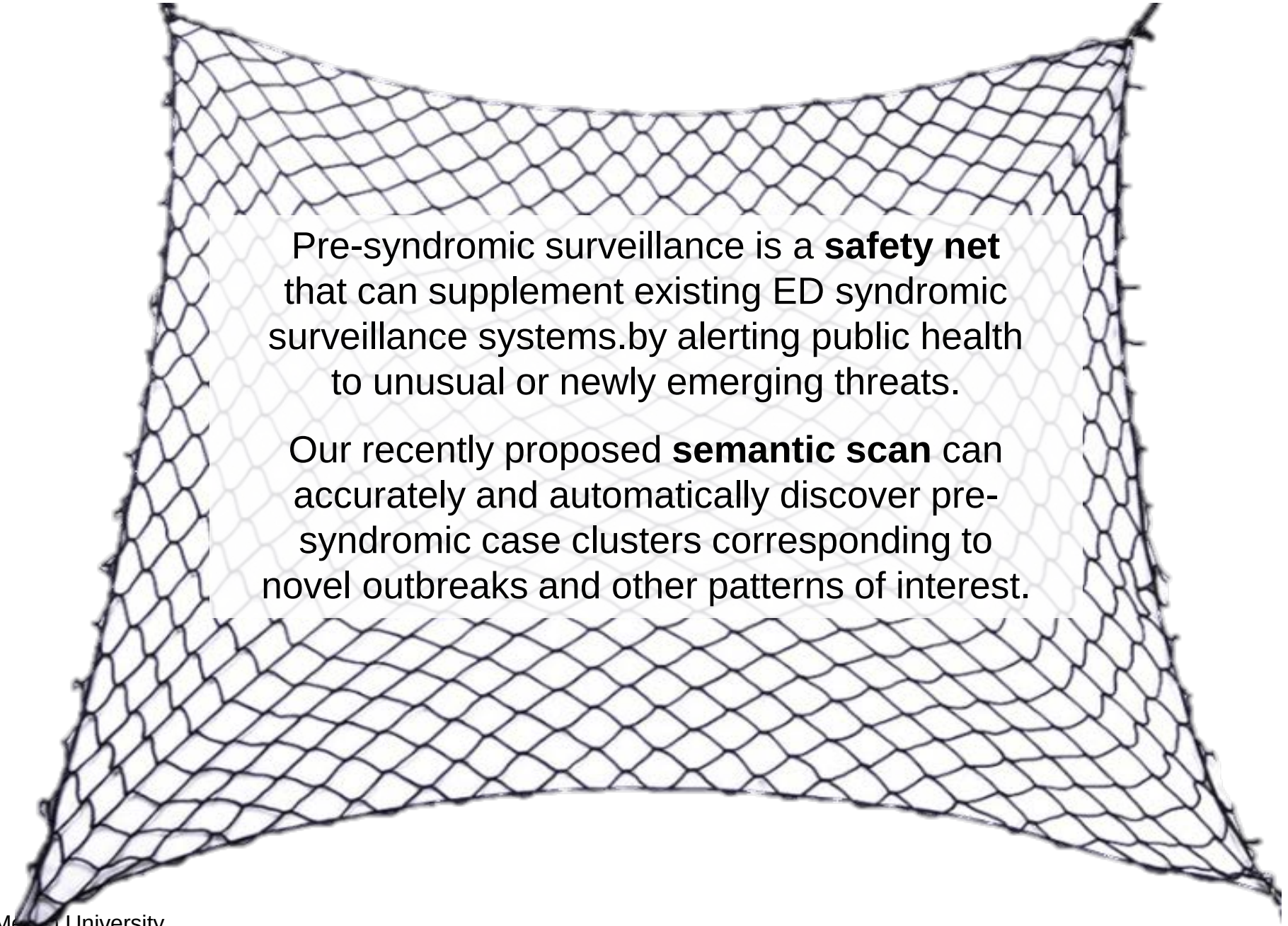
Contagious Diseases
Meningitis
Scabies
Ringworm

Other
Drug overdoses
Smoke inhalation
Carbon monoxide poisoning
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.



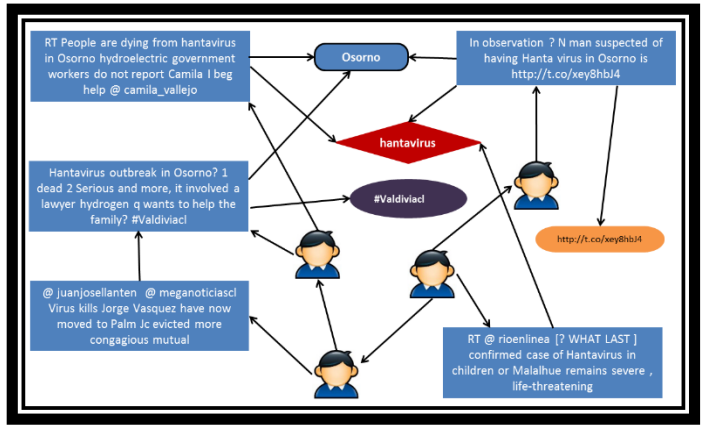
Pre-syndromic surveillance is a **safety net** that can supplement existing ED syndromic surveillance systems.by alerting public health to unusual or newly emerging threats.

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

What can Machine Learning do for public health surveillance?

Today's talk will focus on two specific examples of how ML expands the scope and scale of data sources that can be used for surveillance:

- Structuring **unstructured data**, e.g., free text from ED chief complaints.
- Incorporating complex data sources such as **online social media.**

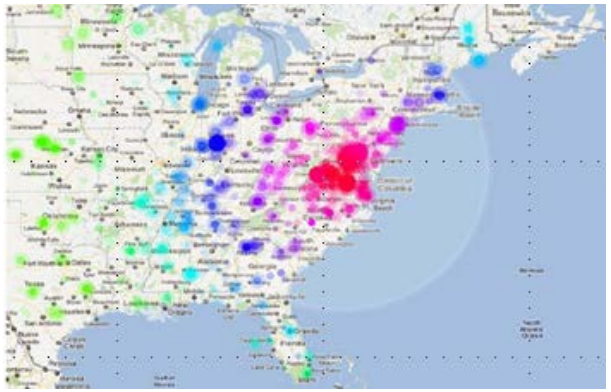


Event Detection from Social Media

Protest in Mexico, 7/14/2012

2012 Washington D.C. Traffic

Tweet Map for 2011 VA Earthquake



Social media is a real-time “sensor” of large-scale population behavior, and can be used for early detection of emerging events.

… but it is very complex, noisy, and subject to biases.

We have developed a new event detection methodology:

“Non-Parametric Heterogeneous Graph Scan” (NPHGS)

Applied to: **rare disease outbreak detection** (hantavirus in Chile).

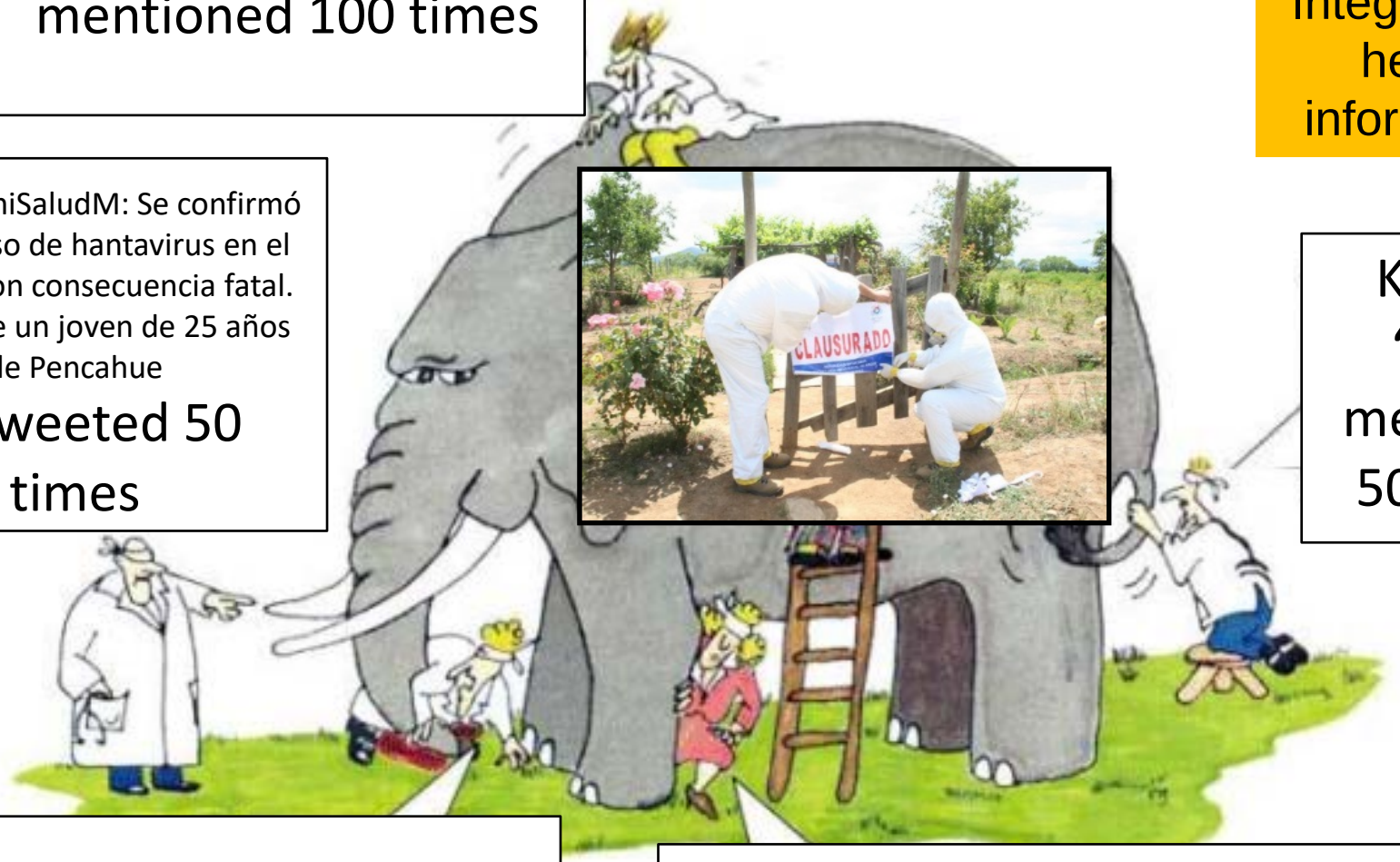
Technical challenge: Integration of multiple heterogeneous information sources!

"#VIRUSHANTA" mentioned 100 times

Technical challenge: Integration of multiple heterogeneous information sources!

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue

re-tweeted 50 times



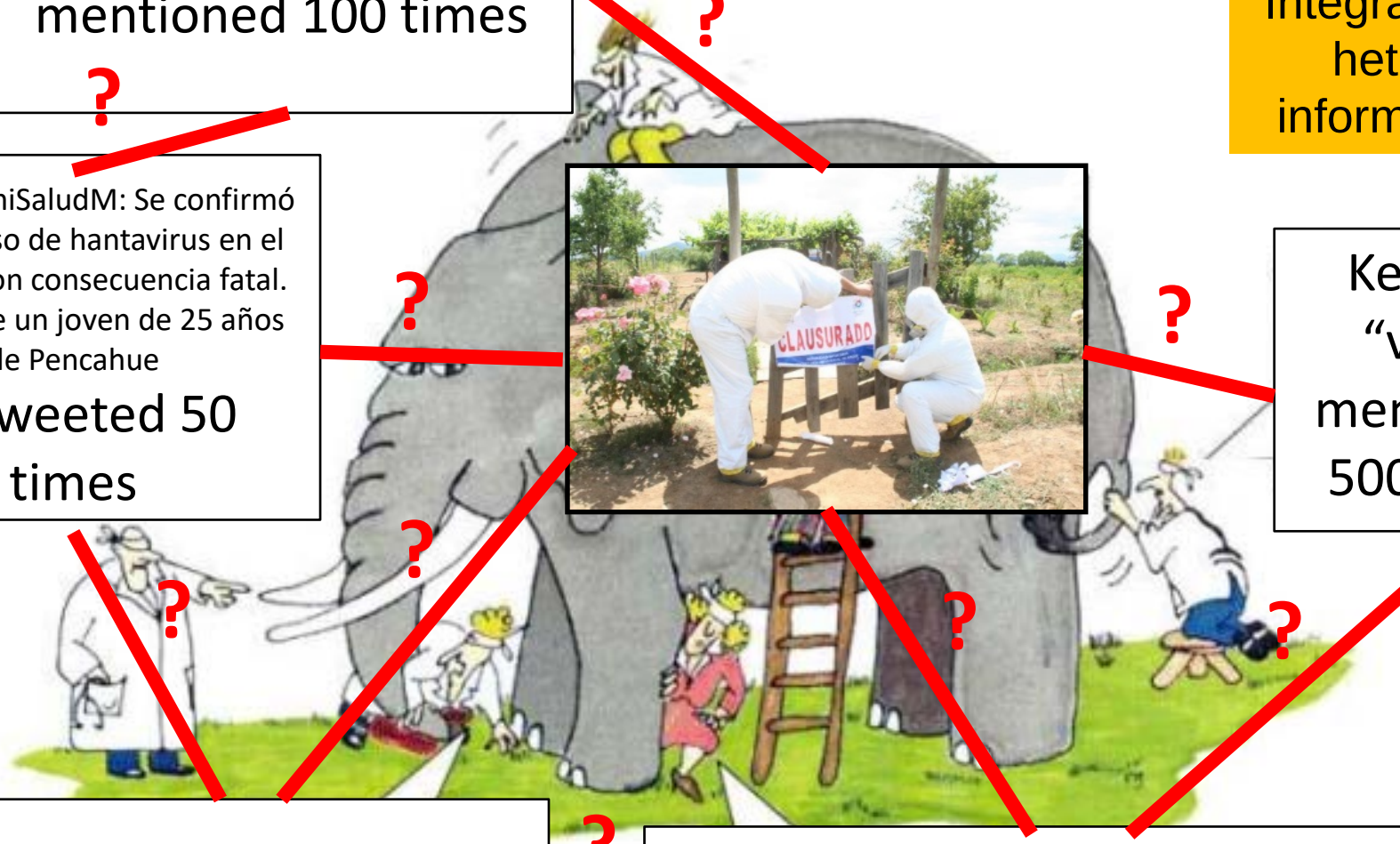
Keyword "virus" mentioned 500 times

http://t.co/5lkD0CZDmf mentioned 10 times

Influential User "SeremiSaludM" (1497 followers) posted 8 tweets

“#VIRUSHANTA” mentioned 100 times

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue re-tweeted 50 times



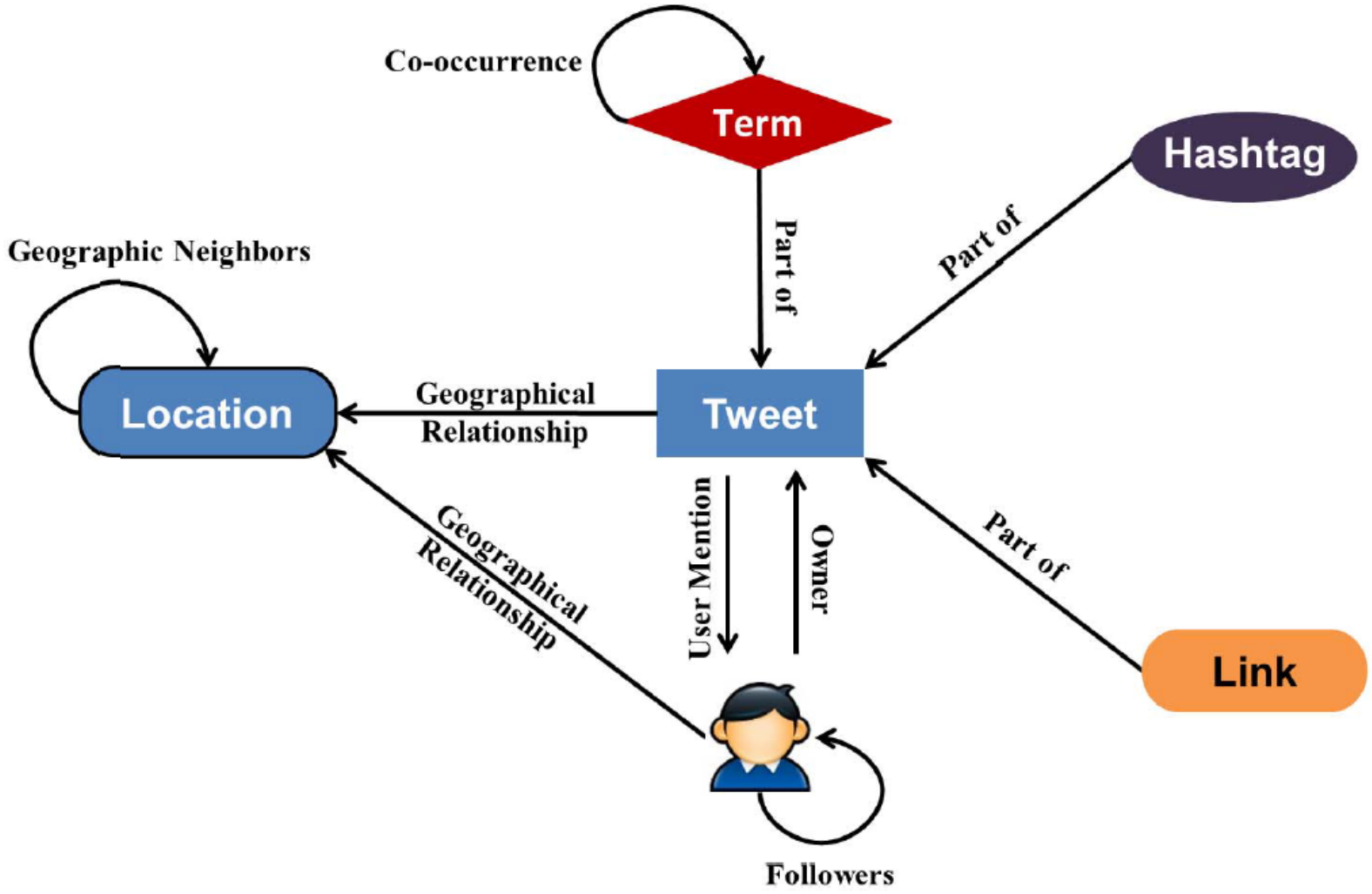
Keyword “virus” mentioned 500 times

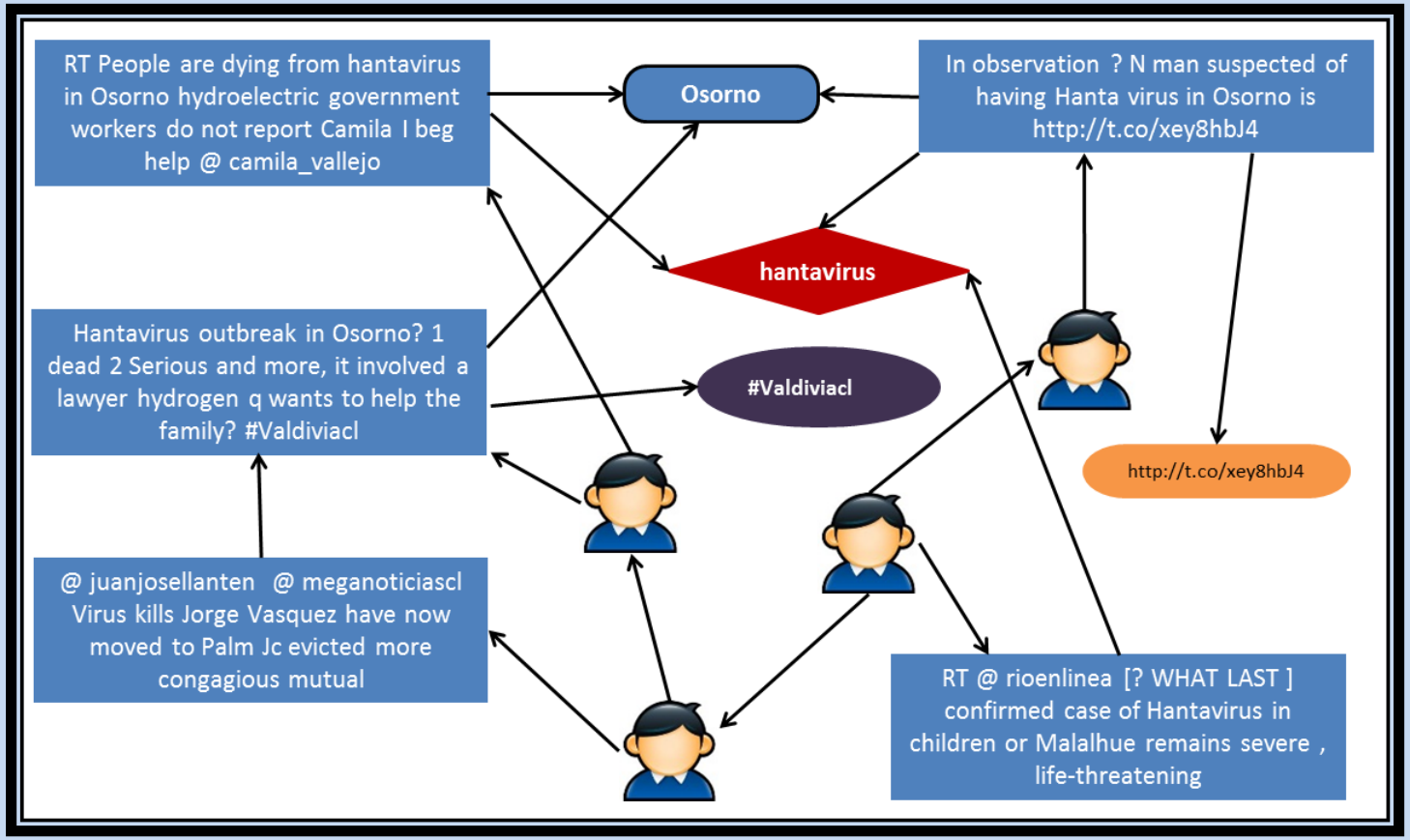
http://t.co/5lkD0CZDmf mentioned 10 times

Influential User “SeremiSaludM” (1497 followers) posted 8 tweets

Technical challenge: Integration of multiple heterogeneous information sources!

Twitter Heterogeneous Network

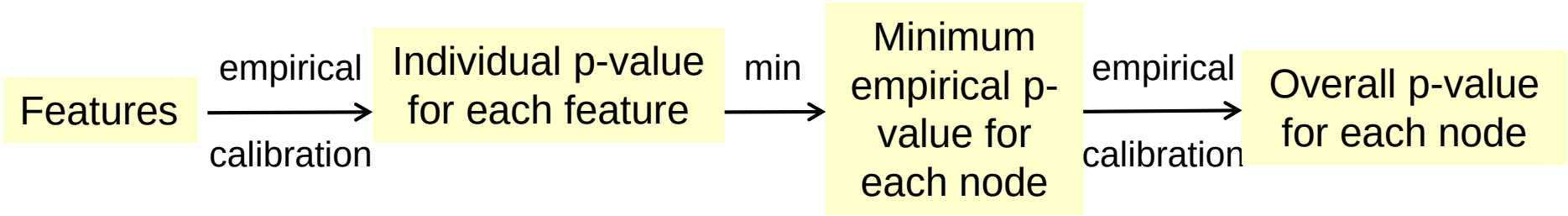



RT People are dying from hantavirus in Osorno hydroelectric government workers do not report Camila I beg help @ camila_vallejo
Osorno
In observation ? N man suspected of having Hanta virus in Osorno is http://t.co/xey8hbJ4
hantavirus
Hantavirus outbreak in Osorno? 1 dead 2 Serious and more, it involved a lawyer hydrogen q wants to help the family? #Valdiviacl
#Valdiviacl
http://t.co/xey8hbJ4
@ juanjosellanten @ meganoticiascl Virus kills Jorge Vasquez have now moved to Palm Jc evicted more congagious mutual
RT @ rioenlinea [? WHAT LAST] confirmed case of Hantavirus in children or Malalhue remains severe , life-threatening

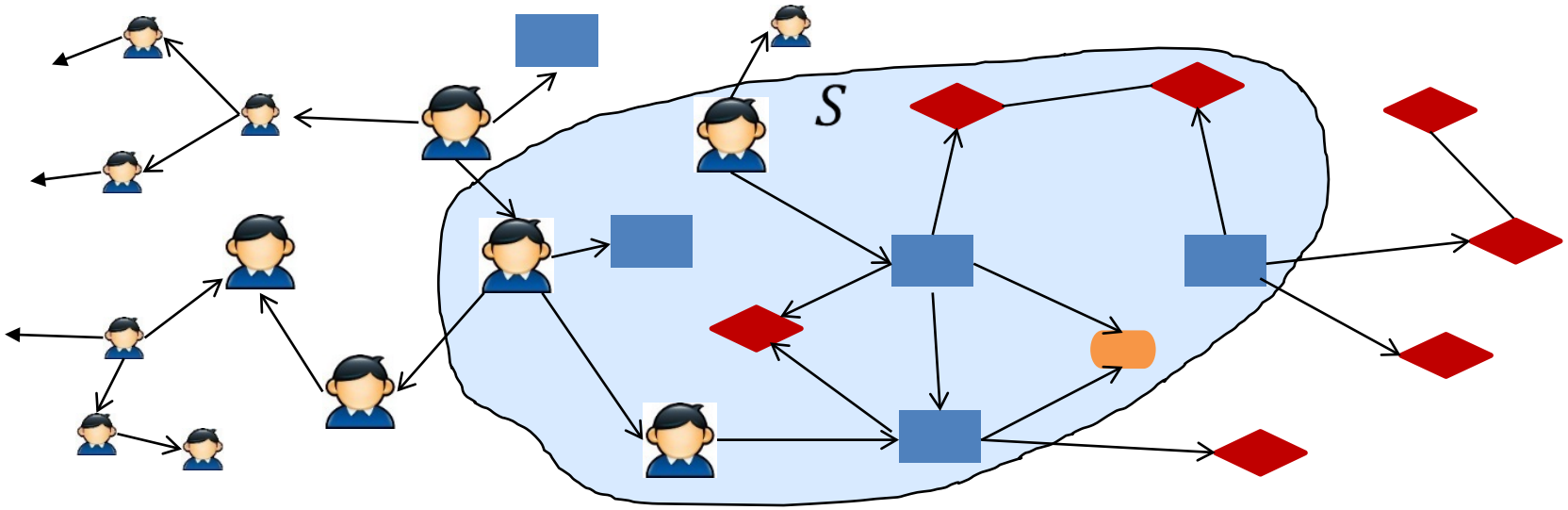
Step 1: Evaluate each node in the Twitter graph

Each node (user, tweet, location, etc.) reports a value measuring how **anomalous** it is for the current 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



Step 2: Find the most anomalous subgraphs



$$S^{\star} = \underset{S \in V:\, S \text{ is connected}}{\operatorname{argmax}} F(S)$$

This step allows us to find groups of nodes (users, keywords, tweets, hashtags, etc.), that are most anomalous when considered collectively.


Using gold standard data from 17 hantavirus outbreaks in Chile, we demonstrated that this approach outperforms existing state-of-the-art methods with respect to timeliness of detection, detection power, and accuracy.
Paraguay
STATE OF SÃO PAULO
Chile
STATE OF PARANÁ
STATE OF SANTA CATARINA
Uruguay
Argentina
Photo credits: Claudio Co-Bar, Martin Garrido

Detected Hantavirus outbreak, 10 Jan 2013

Locations
Users
Keywords
Hashtags
Links
Videos

Temuco and Villarrica, Chile

First news report: 11 Jan 2013

Photo credits: Claudio Co-Bar, Martin Garrido

“Word cloud”
from tweets

Conclusions

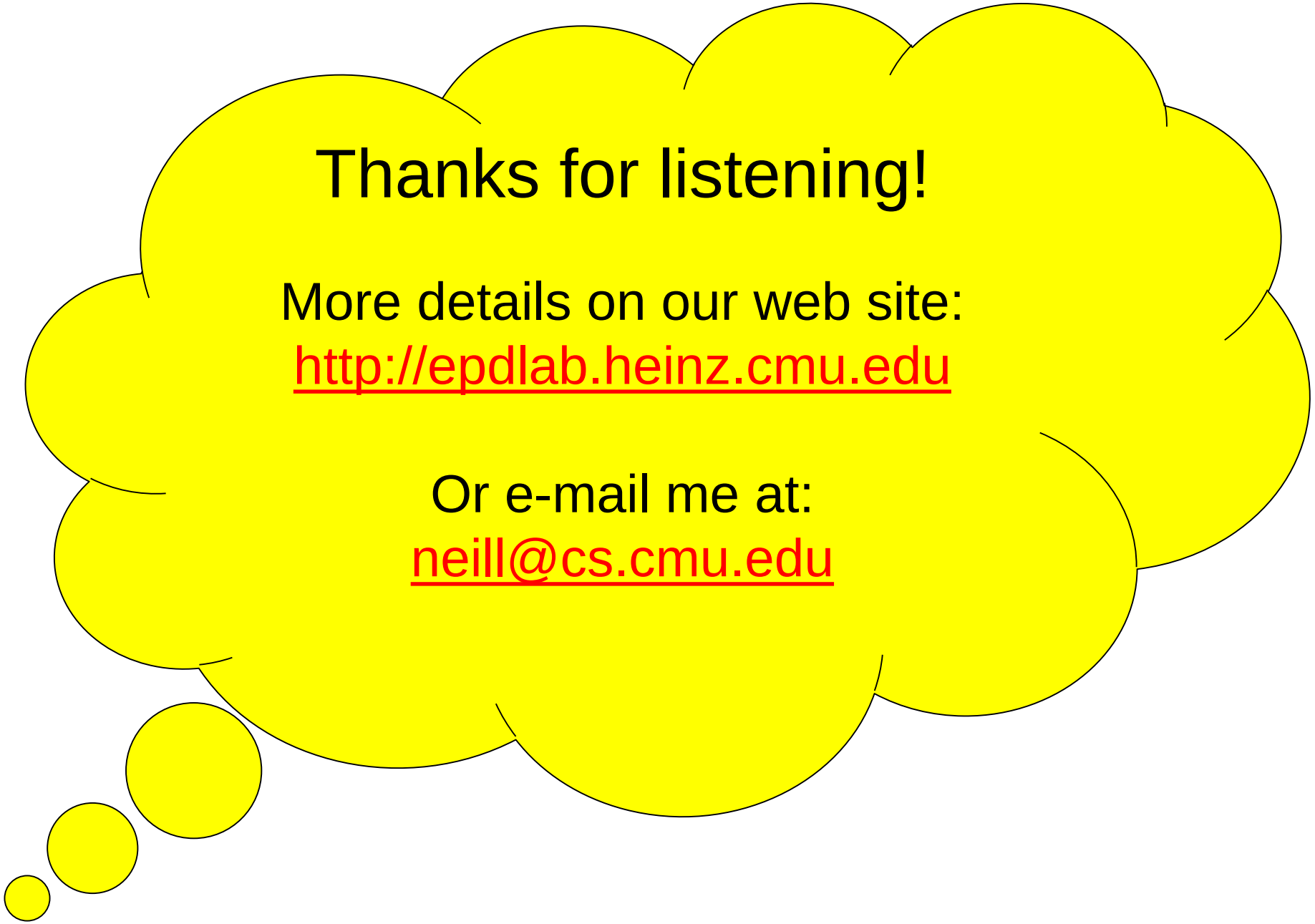
The continually increasing **size**, **variety**, and **complexity** of data available for population health and disease surveillance necessitate development of new detection methods to make use of these data.

We have developed new machine learning approaches for extracting emerging patterns from **free text** data, and for analyzing the heterogeneous network structure of **online social media**.

These methods enable us to detect rare or previously unseen outbreaks that would be missed by typical syndromic surveillance approaches.

References

- D.B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.
- D.B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. *Statistics in Medicine* 32: 2185-2208, 2013.
- E. McFowland III, S. Speakman, and D.B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14: 1533-1561, 2013.
- F. Chen and D.B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. *Proc. 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.
- S. Speakman, S. Somanchi, E. McFowland III, D.B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics* 25: 382-404, 2016.
- D.B. Neill and W. Herlands. Machine learning for drug overdose surveillance. *Journal of Technology in Human Services, 2*018, in press.
- T. Kumar and D.B. Neill. Fast multidimensional subset scan for event detection and characterization. Submitted for publication.
- A. Maurya, D.B. Neill, et al., Semantic scan: detecting subtle, spatially localized events in text streams. Submitted for publication.


Thanks for listening!
More details on our web site:
http://epdlab.heinz.cmu.edu
Or e-mail me at:
neill@cs.cmu.edu