Multidimensional Semantic Scan for Pre-Syndromic Disease Surveillance

Mallory Nobles[1], Ramona Lall[2], Robert Mathes[2], and Daniel B. Neill[1,3,*]

[1]Event and Pattern Detection Laboratory, Carnegie Mellon University
[2]NYC Department of Health and Mental Hygiene
[3]Machine Learning for Good Laboratory, New York University
[*]Corresponding author, e-mail: daniel.neill@nyu.edu

We wish to thank the BCD Syndromic Surveillance Unit at NYC DOHMH for providing retrospective data and participating in the blinded evaluation, and the DHS Hidden Signals Challenge for providing funding support.

Traditional Syndromic Surveillance Classifies Free Text Data into Known Disease Categories

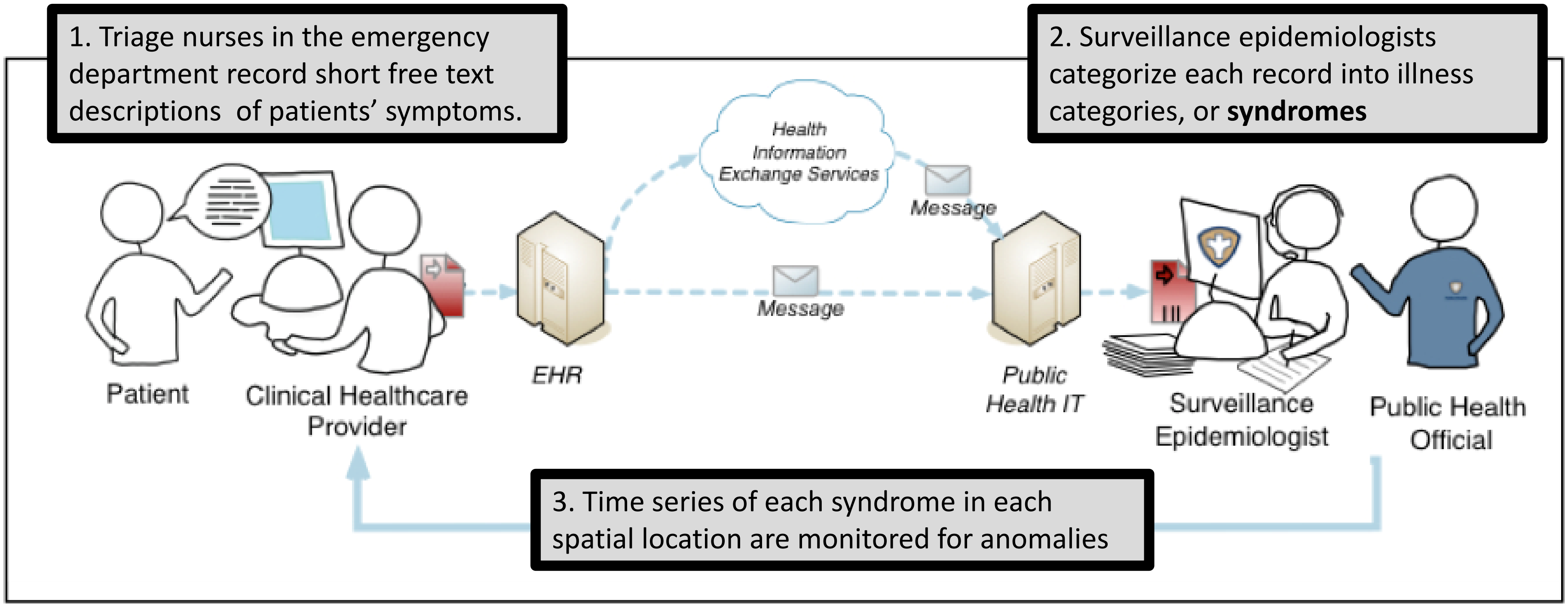
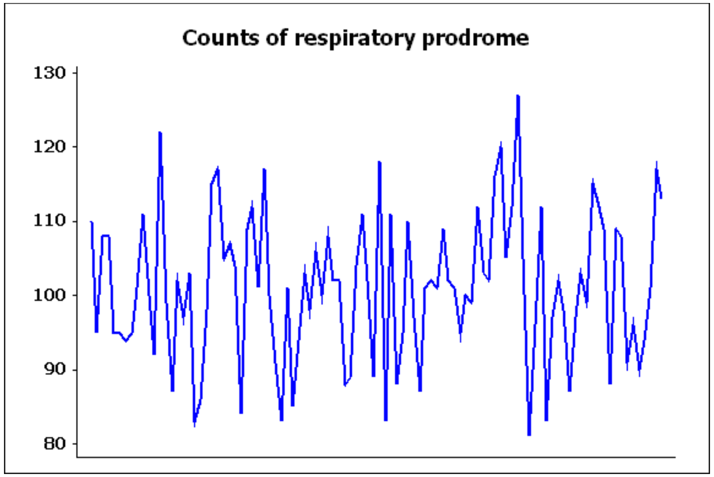


Image Source: [Syndromic Surveillance for Meaningful Use: Background and Resources](#) (ISDS, 2012)

Syndromic Surveillance Can Dilute the Signal of Novel Outbreaks

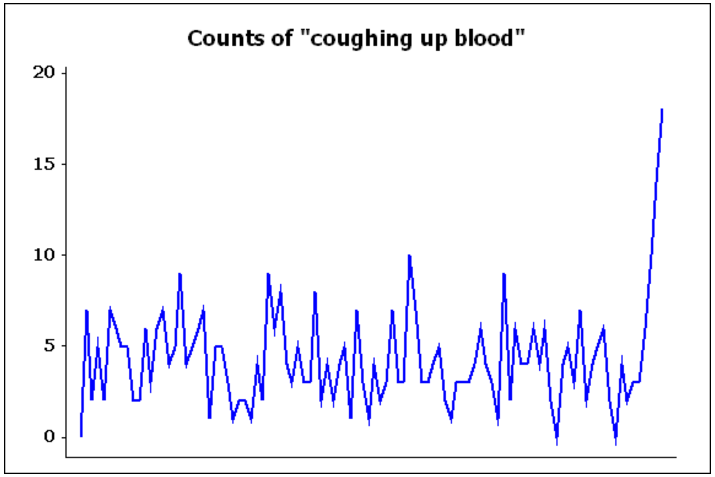
Syndromic Surveillance Approach

Chief Complaint	Syndromic Classification
coughing up blood	Respiratory
fatigue, coughing up blood	Flu
coughing up blood 2 days	Respiratory



Ideal Approach

Chief Complaint	Syndromic Classification
coughing up blood	Coughing up blood
fatigue, coughing up blood	Coughing up blood
coughing up blood 2 days	Coughing up blood



Public Health Needs Define Goals for a New Method

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 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.

Three main goals of Multidimensional Semantic Scan (MUSES):

1. **Learn syndromes** to describe emerging patterns of keywords.
2. **Detect emerging outbreaks** of novel, rare and more common diseases.
3. **Characterize detected events** by identifying the affected time duration, locations and subpopulations.

Learn New Syndromes from Textual ED Notes Using Topic Modeling

- Topic models are algorithms for discovering the main themes that are present in a large and unstructured collection of documents

- Topics learned by these models can act as syndromes since they group symptoms which often co-occur
  - Topics give the relative frequency of symptoms within an illness

Symptoms in Fever Topic	P(symptom)
fever	.68
sweating	.14
chills	.12

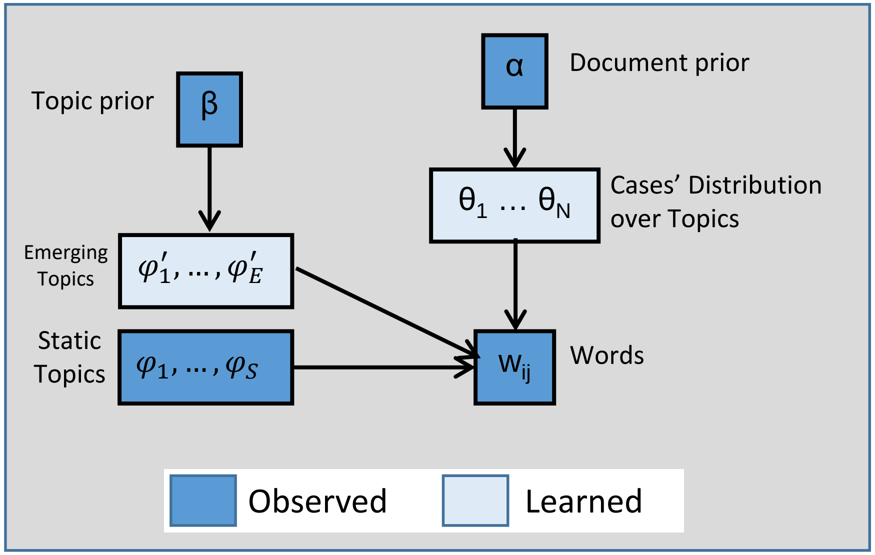
Symptoms in Respiratory Topic	P(symptom)
cough	.42
breathing	.21
shortness	.16

Symptoms in Flu Topic	P(symptom)
fever	.55
aches	.27
fatigue	.09

Multidimensional Semantic Scan Learns Two Sets of Topics

- Static Topics
  - Designed to capture common illnesses, like the flu.
  - Learned over a large set of data.
  - Learned using a standard topic model.

- Emerging Topics
  - Designed to capture rare or novel diseases that aren't well explained by static topics.
  - Learned over the most recent set of data.
  - Learned using a new topic model.



Learned Illness Categories Create Time Series to Monitor for Anomalies

Date/time	Hosp.	Age	Complaint
Jan 1 08	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

Static Topics

φ_1: vomiting, nausea, diarrhea,

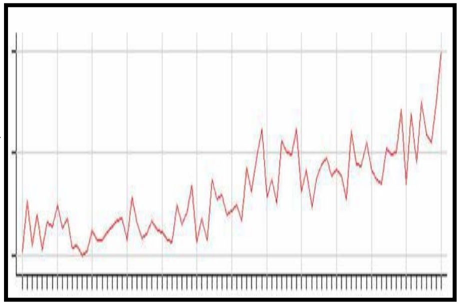
…

φ_2: dizzy, lightheaded, weak, …

Emerging Topics

φ'_1:green, nose, hands…

Classify cases to topics



Hourly counts for each learned syndrome and for each subpopulation

Scan Statistics Identify Anomalous Outbreaks

We consider subsets S that are a combination of a topic, time duration, set of hospitals, and age range.

For each hour of data and each subset S, we compute:

- Count
  - C(S) = number of cases in that time interval matching on hospital, age range, topic
- Baseline
  - B(S) = expected count (28-day moving average)

- Log Likelihood Ratio Score
  - $F(S) = \log \frac{\Pr(data \mid H_1(S))}{\Pr(data \mid H_0)} = \begin{cases} \mathrm{C(S)} * \log \frac{C(S)}{B(S)} + B(S) - C(S) & if\ C(S) > B(S) \\ 0 & otherwise \end{cases}$
  - Null H_0: No outbreak occurring in S, counts have Poisson distribution where mean is baseline
  - Alternative H_1: Outbreak in S, counts have Poisson distribution where mean is multiplicative increase over baseline

We return cases corresponding to the top-scoring subsets S.

Applying Method To Data From New York City

- New York City's Department of Health and Mental Hygiene provided us with ~28 million chief complaint cases from 53 hospitals in NYC from 2010-2016.

- For each case, we have data on the patient's chief complaint, date and time of arrival, age group, gender, and discharge ICD-10 code.

- The chief complaint data required substantial pre-processing.
  - 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

VOIMITING	VOMITINIG	VOMITINGN
VOIMITTING	VOMITINNG	VOMITINGQ
VOIMTING	VOMITIONG	VOMITINGS
VOMIITING	VOMITITING	VOMITINGT
VOMIITNG	VOMITITNG	VOMITINGX
VOMINITING	VOMITN	VOMITINGX1
VOMINTING	VOMITNG	VOMITINGX2
VOMIOTING	VOMITNIG	VOMITINGX3
VOMITE	VOMITNING	VOMITINGX4
VOMITED	VOMITO	VOMMITTING
VOMITG	VOMITOS	VOMNITING
VOMITHING	VOMITS	VOMOITING
VOMITI	VOMITT	VOMTIING
VOMITIG	VOMITTE	VOMTIN
VOMITIGN	VOMITTI	VOMTITING
VOMITIING	VOMITTING	VONMITING
VOMITIN	VOMITTTING	VOOMITING
VOMITING3	VOMITUS	VOPMITING
VOMITINGA	VOMMIT	VVOMITING
VOMITINGG	VOMMITING	VOMITINGM

Variations of the words "vomit" and "vomiting" that appear > 15 times in data

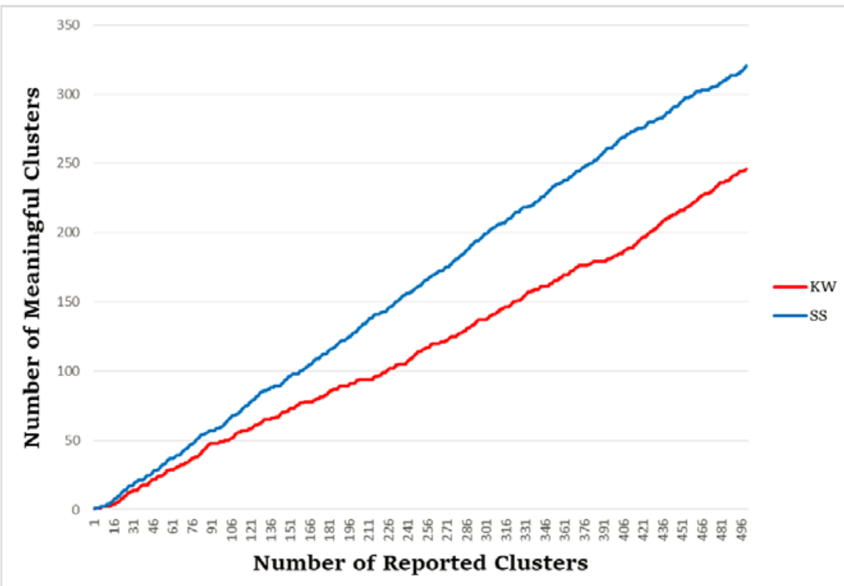
New York City Public Health Practitioners Performed Blinded Evaluation of Results

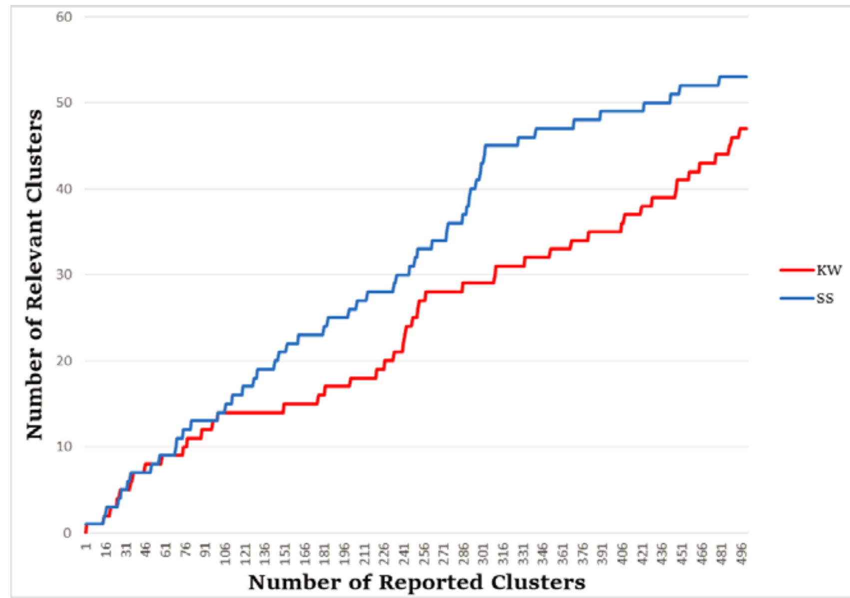
We used Multidimensional Semantic Scan and a keyword-based method (representing the current state of the art) to identify potential outbreaks in the NYC data.

- For each method's 500 highest scoring clusters, NYC DOHMH public health officials indicated if the cluster is a relevant or meaningful cluster

	Relevant Clusters of Interest	Meaningful Clusters of Potential Interest	Clusters Not of Interest
	Examples: bacterial meningitis, synthetic drugs use	Examples: flu, rashes, motor vehicle accidents	Examples: misspellings, non-specific words (i.e. "left")
Multidimensional Semantic Scan	53	267	180
Keyword Based Method	47	199	254

Multidimensional Semantic Scan has Higher Precision than Competing Method





Example of Detected Cluster

Arrival Date	Arrival Time	Chief Complaint	ICD-10	Patient Sex	Patient Age
11/28/2014	6:04	I DRANK VODKA AND NEED DETOX.		F	35-39
11/28/2014	7:52	EVAUATION, DRANK COFFEE WITH CRUSHED GLASS THIS MORNING		M	45-49
11/28/2014	7:53	DRANK TAINTED COFFEE		M	65-69
11/28/2014	7:57	DRANK TAINTED COFFEE		F	20-24
11/28/2014	7:59	INGESTED TAINTED COFFEE		M	35-39
11/28/2014	8:01	DRANK TAINTED COFFEE		M	45-49
11/28/2014	8:03	DRANK TAINTED COFFEE		M	40-44
11/28/2014	8:04	DRANK TAINTED COFFEE		M	30-34
11/28/2014	8:06	DRANK TAINTED COFFEE		M	35-39
11/28/2014	8:09	INGESTED TAINTED COFFEE		M	25-29

After Hurricane Sandy, Detected Clusters Consistent with Retrospective Analysis

Summary of Chief Complaints in Cluster	Date Range	
Shortness of Breath, Asthma	October 29 - 30	Acute Cases
Falls	October 30	
Lower Leg Injury	October 30	
Trouble Sleeping	October 30	Mental Health Disturbances
Depression	October 30 - November 6	
Agitation or Anxiety	November 4 - 5	
Transfer Cases	November 5	Burden on Medical Infrastructure
Methadone Maintenance	November 6	
Needs Dialysis	November 7	

Results consistent with Lall et al. (OJPHI, 2014):

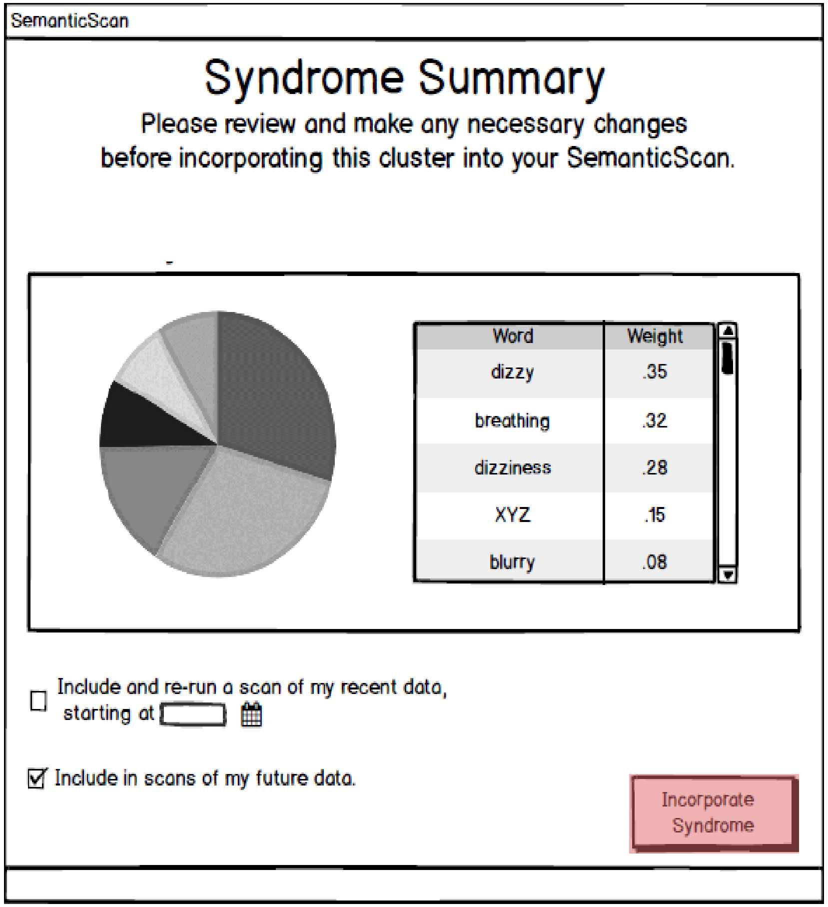
- Manual inspection of ED data immediately following Hurricane Sandy uncovered increase in words “methadone”, “dialysis”, “oxygen”.

Multidimensional Semantic Scan Identified Many Other Clusters in NYC Data

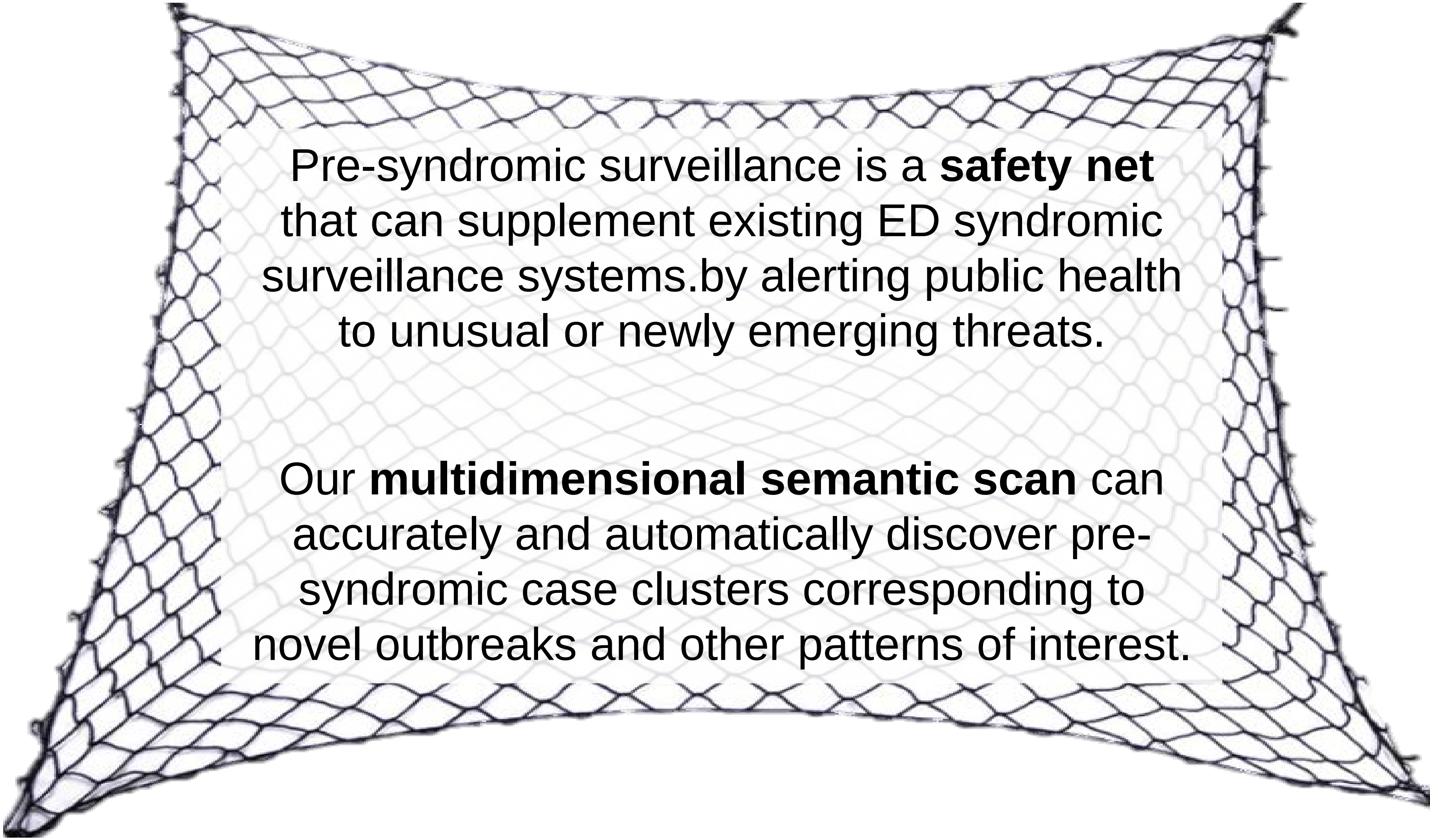
Contagious Diseases	Accidents	Other
Meningitis	Motor Vehicle	Drug overdoses
Scabies	Ferry	Smoke inhalation
Ringworm	School bus	Carbon monoxide poisoning
Lice	Elevator	Exposure to bats
Pink Eye	Pedestrians Struck by Cars	Animal Bites/Rabies Shots
Hepatitis		Crime related, e.g., pepper spray attacks
Sexually Transmitted Diseases		Concern over Ebola
		Food Poisoning

Ongoing Work to Integrate User Feedback

- Improve performance by including a human in the loop and incorporating feedback
  - Practitioners can indicate detected topics that they would like to monitor in the future.
  - If public health officials indicate that a detected outbreak is not of interest, model will not learn this type of outbreak in the future.
  - Public health officials can also indicate terms to exclude in future.



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