Identifying Emerging Novel Disease Outbreaks In Textual Emergency Department Data

Daniel B. Neill
Event and Pattern Detection Laboratory
Carnegie Mellon University
E-mail: neill@cs.cmu.edu

Joint work with:

Mallory Nobles (CMU- EPD Lab)
Lana Deyneka (NC Dept. of Health)
Amy Ising (UNC- NC DETECT)

Scaling up surveillance

The landscape of event surveillance is changing rapidly, due to increased availability of huge amounts of data at the societal scale.

Increasing use of detailed **electronic medical records** for patient data.

Informal, Web-based data sources such as Internet search queries and Twitter feeds.

New data sources have enormous **potential** for enabling more timely and accurate event detection, but also pose many **challenges**.

Massive amounts of data…

Integrating many data sources…

Data mostly exists as **unstructured free text**!

The NC DETECT use case

Created by Amy Ising, Lana Deyneka, Jenna Waggoner, and Anna Waller.

Use case development facilitated by the ISDS Technical Conventions Committee.

UNC Carolina Center for Health Informatics and NC Department of Health and Human Services

See panel discussions today and tomorrow.

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.

Monitor **aggregate counts** of cases in space and time (e.g., by spatial scanning).

Monitor hospital ED visits for **time of arrival clusters**.

Identify differentially affected **subpopulations** (e.g., by age and gender)

Track novel and rare **keyword** counts.

Our approach: detect emerging **topics** (patterns of keywords).

The NC DETECT use case

Created by Amy Ising, Lana Deyneka, Jenna Waggoner, and Anna Waller.

Use case development facilitated by the ISDS Technical Conventions Committee.

UNC Carolina Center for Health Informatics and NC Department of Health and Human Services

See panel discussions today and tomorrow.

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.

Dataset: ~200K de-identified ED visits over one year at 3 NC hospitals.

Attributes: arrival date/time (altered), hospital (A/B/C), age group, CC.

Goal: to detect any clusters of interest. (symptoms, events, place names, arrival time, hospital location, …)

*** ~40 examples of such clusters were injected into the data. ***

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?

Possible approach: map ED cases to broad syndrome categories ("prodromes") and do a **multidimensional scan**.

a fib

n v d

tired weak

food
poisoning

diarrhea

fever

Multidimensional scanning

(for known prodromes)

For each hour of data (~8K):

For each combination S of:

- Hospital (A/B/C)
- Time duration (1-3 hours)
- Age range (9 groups → 73 ranges)
- Prodrome

Count: C(S) = # of cases in that time interval matching on hospital, age range, prodrome.

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.

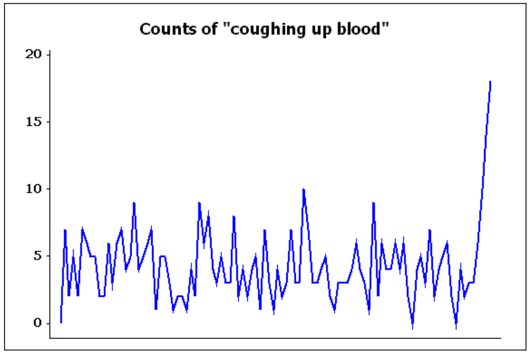
Where do existing methods fail?

The typical, prodrome-based scan statistic approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

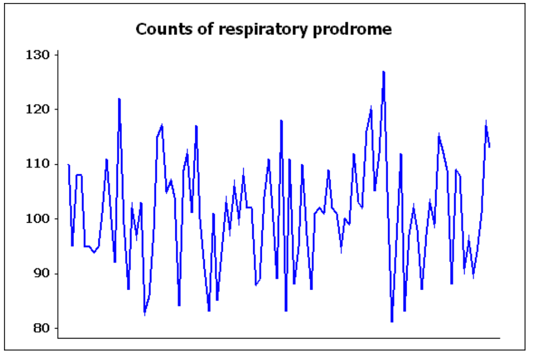
What happens when something new and scary comes along?

- **More specific symptoms** ("coughing up blood")
- **Previously unseen symptoms** ("nose falls off")

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



Mapping specific chief complaints to a broader symptom category can dilute the outbreak signal, delaying or preventing detection.



Where do existing methods fail?

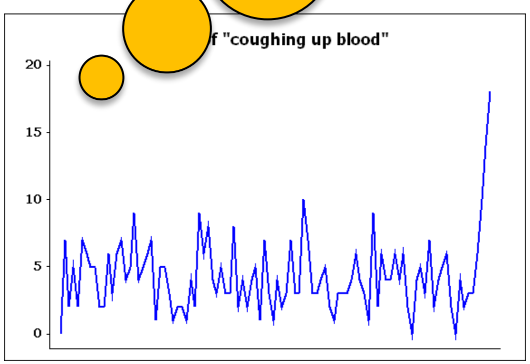
The typical, prodr…
scan statisti…
effectively …
outbrea…
seen…
sy…

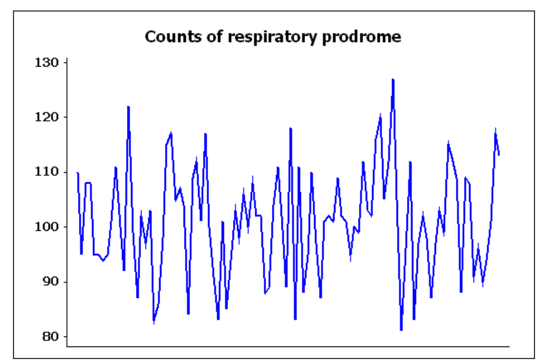
…en something
… along?
…toms
…d”)
…n
… off”)

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

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

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





The semantic scan statistic

Date/time	Hosp.	Age	Complaint
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp

1 year of free-text ED chief complaint data from 3 hospitals in North Carolina.

The semantic scan statistic

Date/time	Hosp.	Age	Complaint
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp

Bayesian inference using LDA model

Case prior: α

Distribution over topics per case: $\theta_1 \ldots \theta_N$

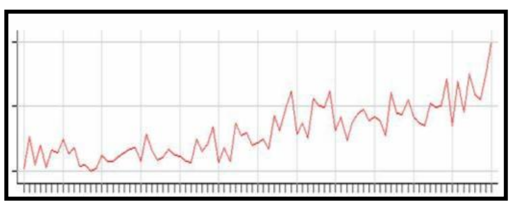
Topic prior: β

Topics: $\Phi_1 \ldots \Phi_K$

Observed words: w_{ij}

φ_1: vomiting, nausea, diarrhea, …
φ_2: dizzy, lightheaded, weak, …
φ_3: cough, throat, sore, …

Classify cases to topics



Time series of hourly counts for each combination of hospital and age group, for each topic φ_j.

Now we can do a multidimensional scan, using the learned topics instead of pre-specified prodromes!

Multidimensional scanning

(for learned topics)

For each hour of data (~8K):

For each combination S of:

- Hospital (A/B/C)
- Time duration (1-3 hours)
- Age range (9 groups → 73 ranges)
- **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.

Multidimensional scanning

(for emerging topics)

For each hour of data (~8K):

For each combination S of:

- Hospital (A/B/C)
- Time duration (1-3 hours)
- Age range
- **Emerging topic**

We can do even better by:

1) Learning a set of "static" topics from historical data.
2) Identifying "emerging topics" that are maximally different from the static topics.

Count: C(S) = # of cases in that time interval matching on hospital, age range, emerging 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.

Multidimensional scanning

(for keywords)

For each hour of data (~8K):

For each combination S of:

- Hospital (A/B/C)
- Time duration (1-3 hours)
- Age range
- **Keyword**

Just using **keyword matching** does not do as well:

1) Huge # of subsets S to score
2) Picks up noise (e.g., typos) and more typical symptoms (e.g., cold/flu).

Count: C(S) = # of cases in that time interval matching on hospital, age range, keyword.

Baseline: B(S) = expected count (28-day moving average).

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

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

Semantic scan use case results

We applied the multidimensional semantic scan (with emerging topics) on data provided by the North Carolina Department of Health, with simulated novel outbreaks of interest injected by the NC DETECT group.

We identified clusters of cases referring to specific locations, unusual sets of symptoms, or affected subpopulations.   Here are some highlights:

<u>Location and symptoms</u>:
"sudden onset of rashes
at the beach"

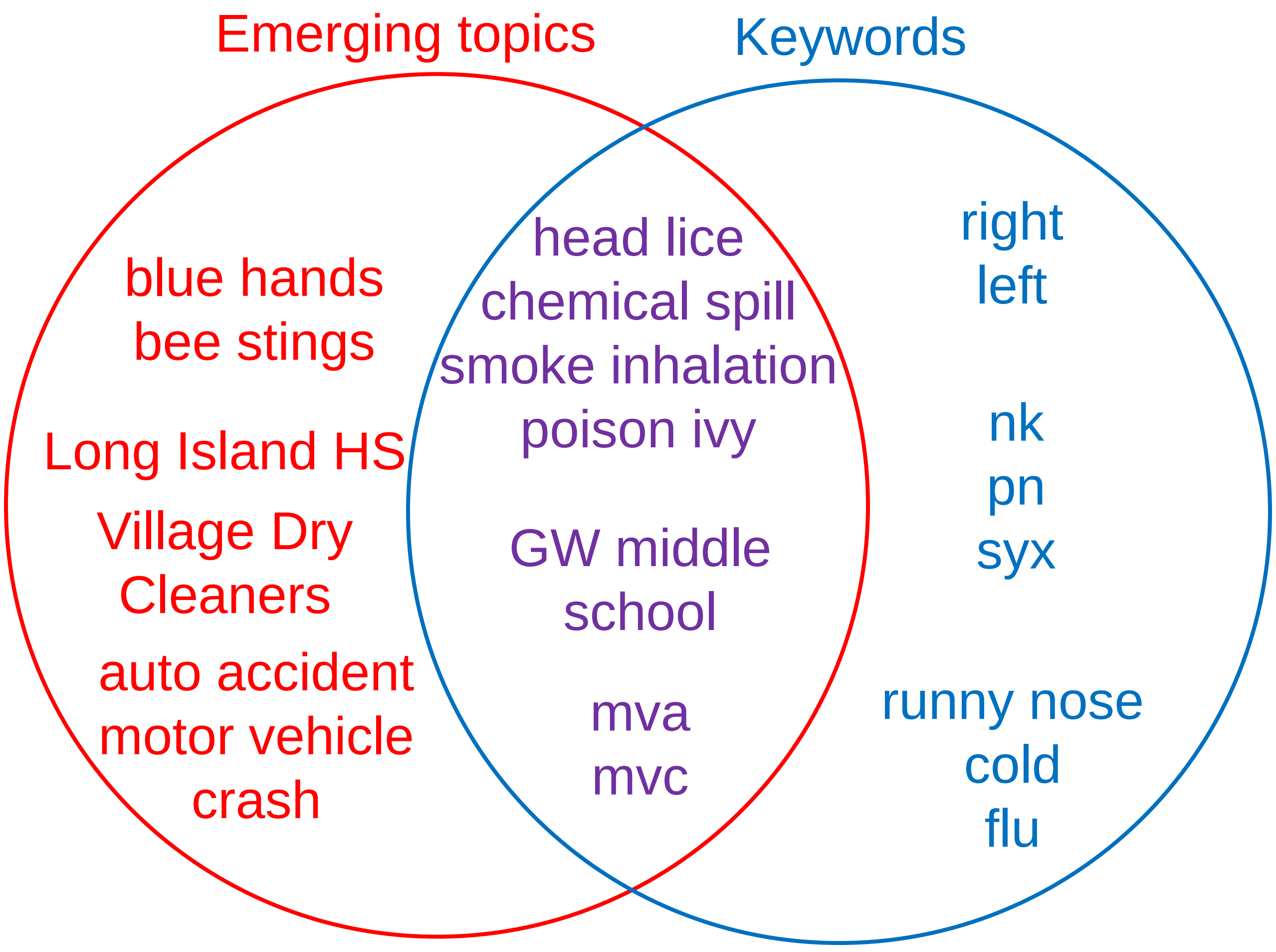
Ten cases that mentioned
a local middle school
within a four-hour span

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

<u>Specific subpopulations</u>:
Seven young adults
suffering from smoke
inhalation

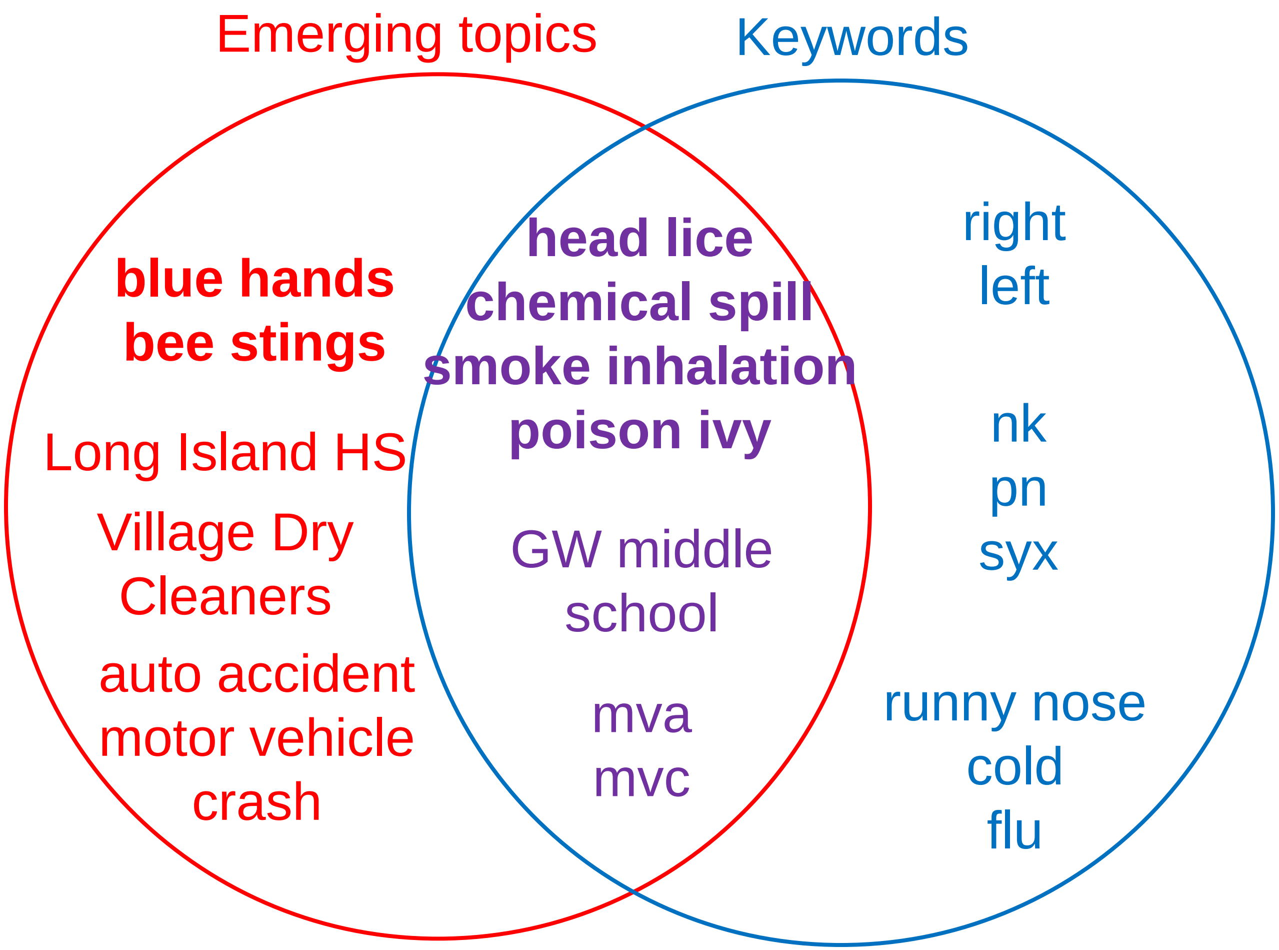
Preliminary comparison results

We compared the top-20 clusters detected by the emerging topic semantic scan and keyword-based scan for each hospital.



Preliminary comparison results

We compared the top-20 clusters detected by the emerging topic semantic scan and keyword-based scan for each hospital.

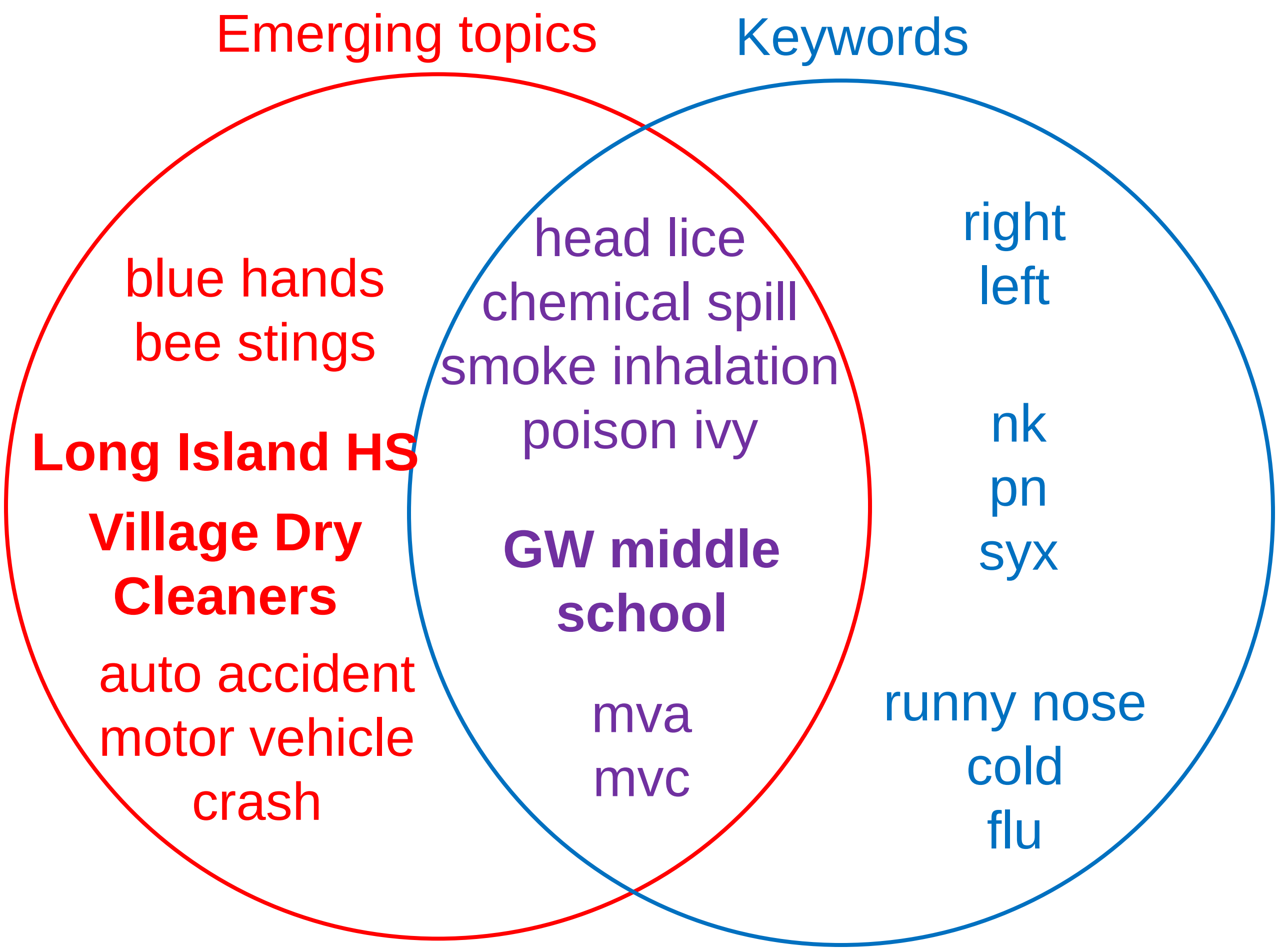


Both methods detected unusual symptom patterns including at least one rare word.

Semantic scan was also able to detect unusual combinations of more common words, e.g., blue hands.

Preliminary comparison results

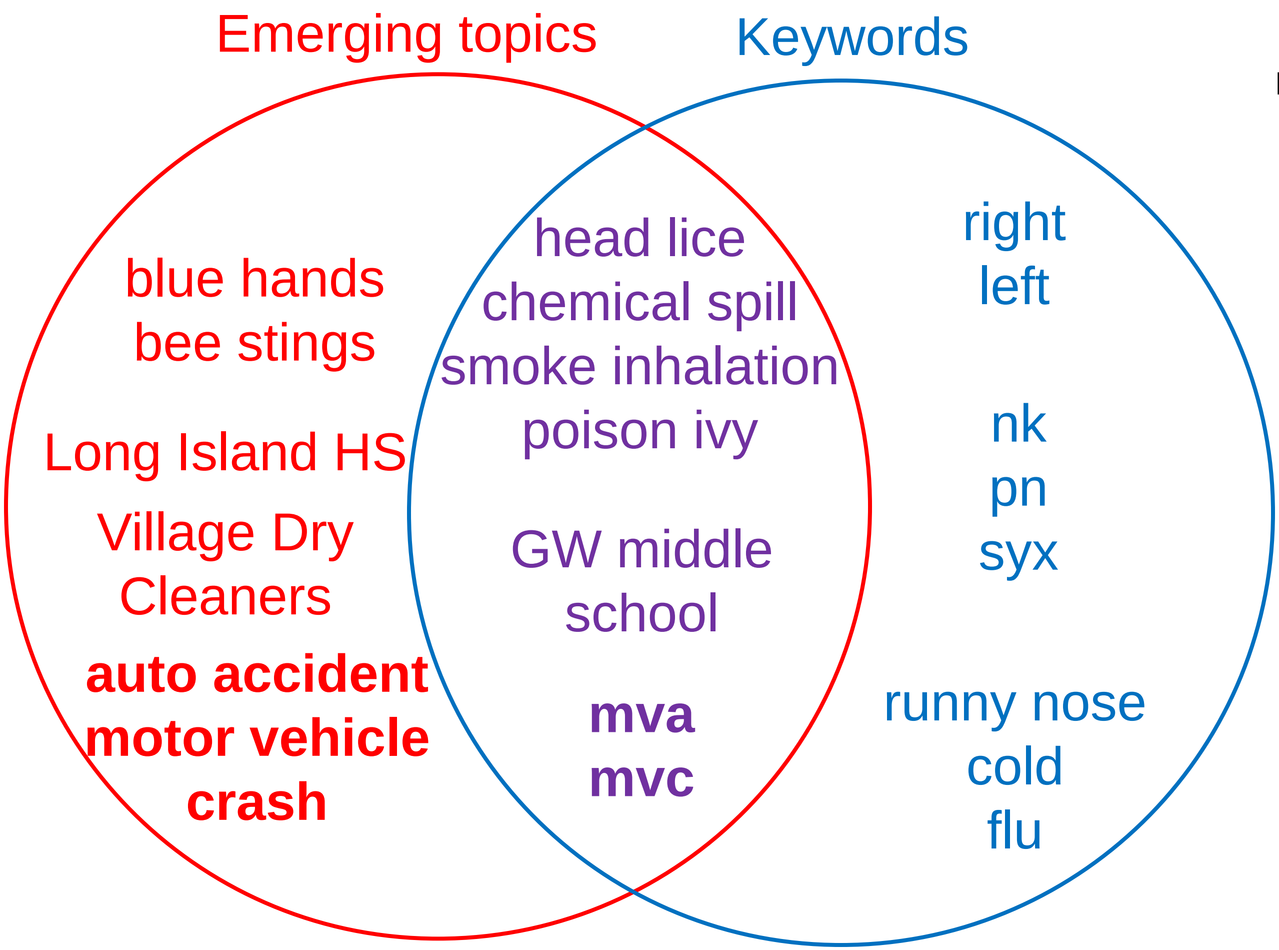
We compared the top-20 clusters detected by the emerging topic semantic scan and keyword-based scan for each hospital.



Similarly, place names for common-source exposures may be missed by the keyword approach if consisting of only common words.

Preliminary comparison results

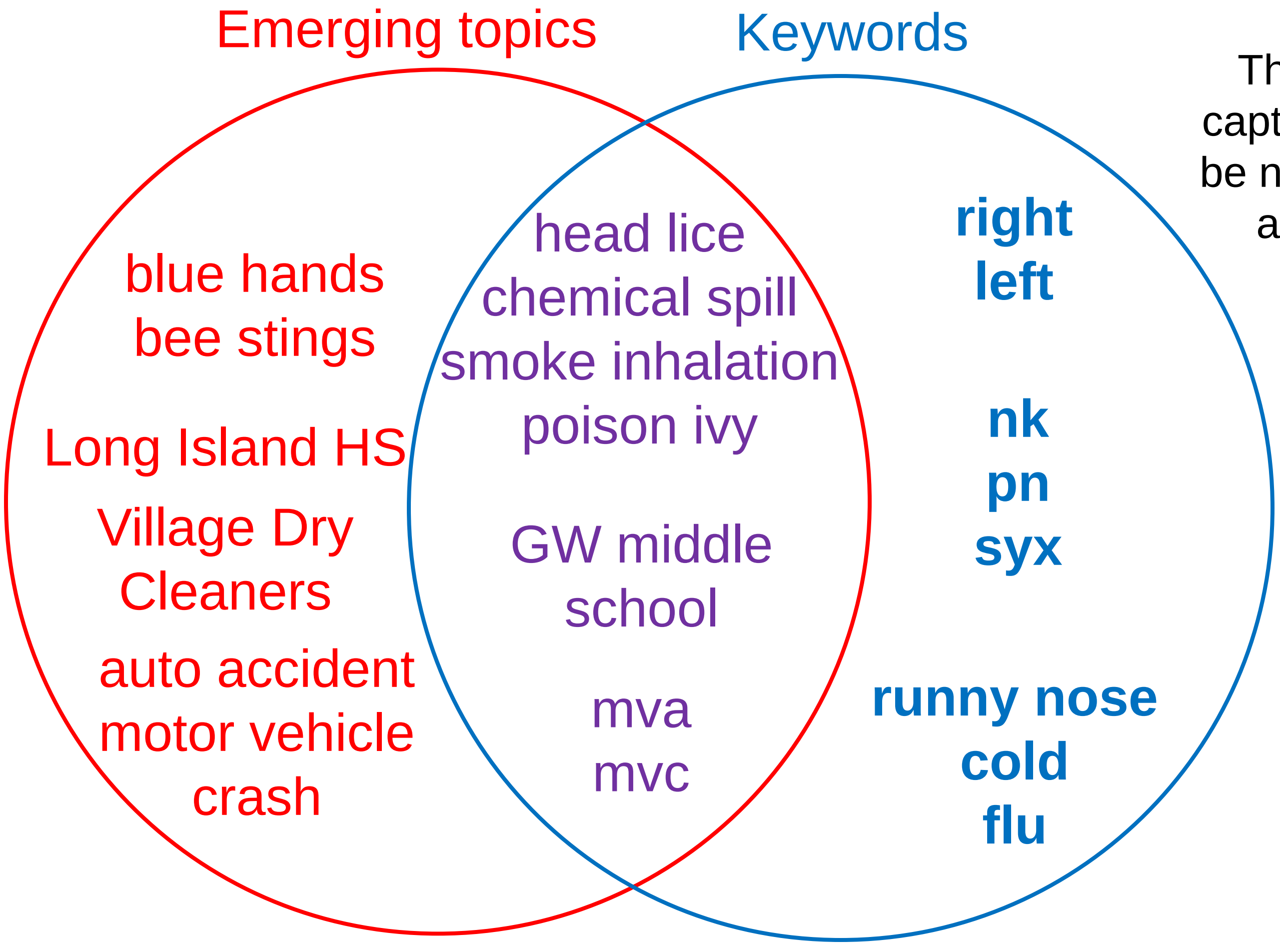
We compared the top-20 clusters detected by the emerging topic semantic scan and keyword-based scan for each hospital.



Both methods picked up multiple clusters of "mva" and "mvc" cases from auto accidents, but semantic scan also detected clusters using multiple different words.

Preliminary comparison results

We compared the top-20 clusters detected by the emerging topic semantic scan and keyword-based scan for each hospital.



The keyword approach also captured some clusters likely to be noise, unusual abbreviations and typos, and clusters of common symptoms.

Conclusions

Semantic scan with emerging topics is a promising approach to detection of novel emerging clusters of disease in free-text ED visit data.

A full evaluation and comparison of methods using gold standard data (injected clusters, true clusters of interest to NC DPH) is in progress.

Preliminary results suggest that our approach outperforms both simpler keyword-based methods, and methods that do not use the free text data.

The work has potential for incorporation into deployed surveillance systems such as NC DETECT, and should ideally be used to supplement (not replace) prodrome-based outbreak detection methods.

Acknowledgements

- We gratefully acknowledge **funding support** from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330.
- **Data** was provided by the NC DHHS/DPH NC DETECT system. The NC DETECT Data Oversight Committee does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.
- **Kenton Murray**, **Yandong Liu**, and **Chris Dyer** contributed to development of the semantic scan approach.
- A special thanks to **Howard Burkom**, for feedback and his role in leading the ISDS Technical Conventions Committee.

Thanks for listening!

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

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