

# Analytical Methods for Large Scale Surveillance of Unstructured Data

Daniel B. Neill, Ph.D.  
Event and Pattern Detection Laboratory  
Carnegie Mellon University  
neill@cs.cmu.edu

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330.

# Scaling up Disease Surveillance

The landscape of 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.

# Scaling up Disease Surveillance

The landscape of 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 outbreak detection, but also pose many **challenges**.

Massive amounts of data...

Integrating many data sources...

# Scaling up Disease Surveillance

The landscape of 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 outbreak detection, but also pose many **challenges**.

Massive amounts of data...

Integrating many data sources...

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

# Scaling up Disease Surveillance

The landscape of surveillance is changing, due to increased availability of huge amounts of data from many sources.

Key message: New, cool data sources are not enough!

New methods are needed to deal with the **scale** and **complexity** of the new data.

New data sources are appearing more timely and in greater volume, so pose many **challenges**.

Massive amount of data...  
Integrating many data sources...

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

# 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. influenza-like illness).

# 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. influenza-like illness).

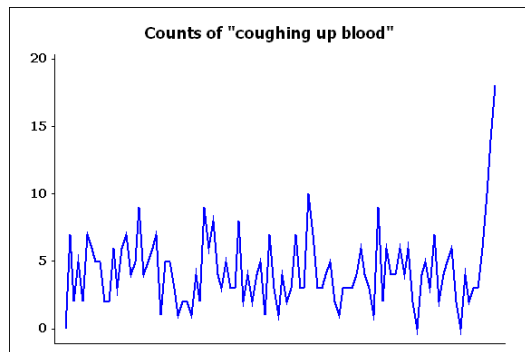
What happens when something new and scary comes along?

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

# 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. influenza-like illness).

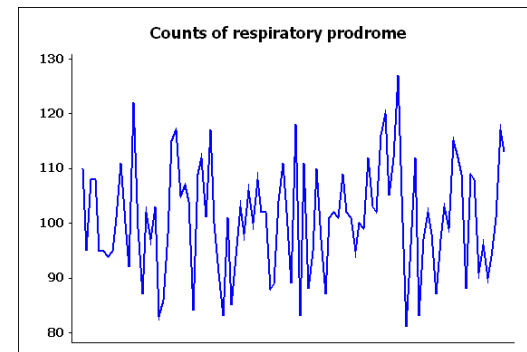
If we were monitoring these particular symptoms, it would only take a few such cases to realize that an outbreak is occurring!



What happens when something new and scary comes along?

- **More specific symptoms** ("coughing up blood")
- **Previously unseen symptoms** ("nose turns green and 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?

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

The

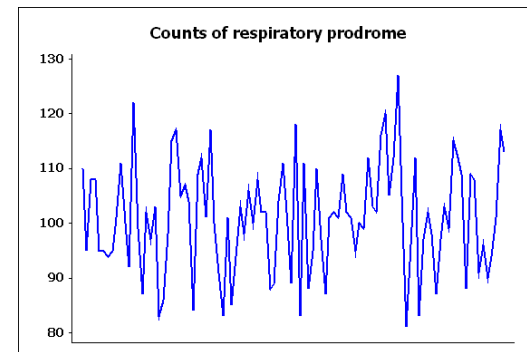
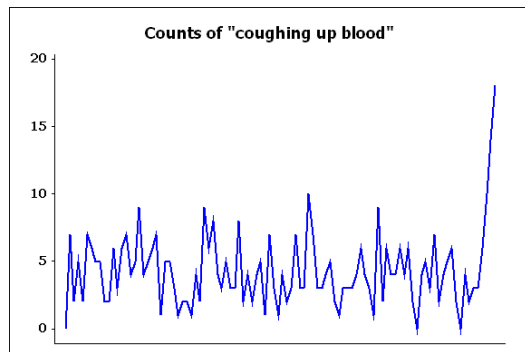
(e.g. n.

ptoms

turns green and 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.



# Semantic Scan Statistic

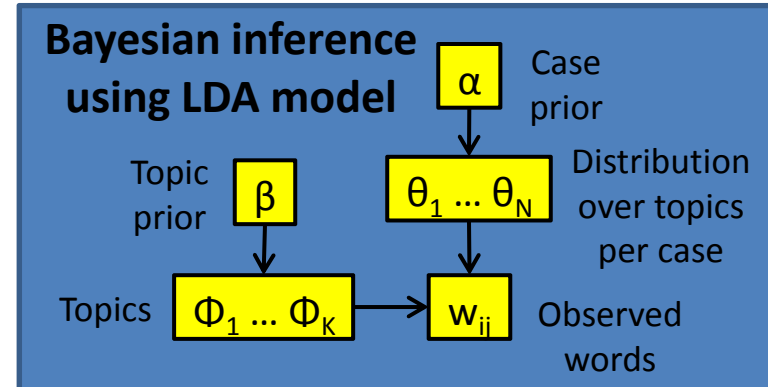
<u>Date</u>	<u>Location</u>	<u>Complaint</u>
1/1/11	15213	runny nose
1/1/11	15217	fever and chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp

2 years of free-text ED chief complaint data from 10 hospitals in Allegheny County, PA



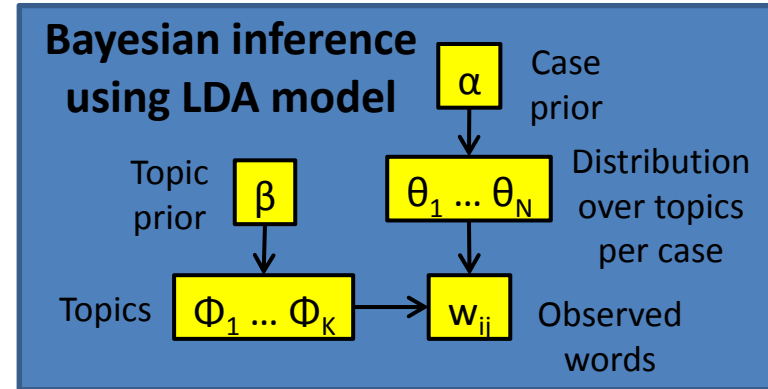
# Semantic Scan Statistic

<u>Date</u>	<u>Location</u>	<u>Complaint</u>
1/1/11	15213	runny nose
1/1/11	15217	fever and chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp



# Semantic Scan Statistic

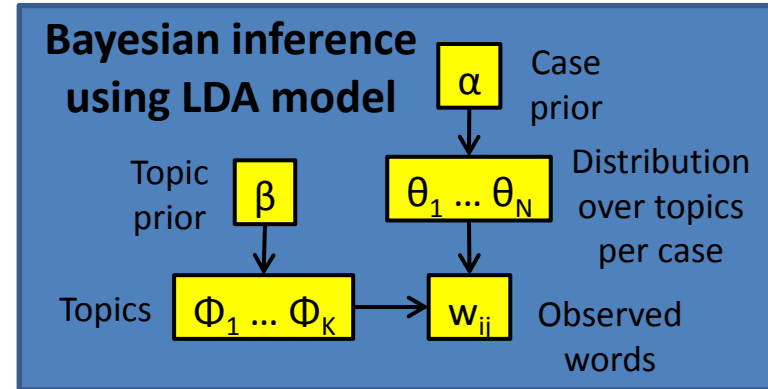
<u>Date</u>	<u>Location</u>	<u>Complaint</u>
1/1/11	15213	runny nose
1/1/11	15217	fever and chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp



$\phi_1$ : vomiting, nausea, diarrhea, ...
$\phi_2$ : dizzy, lightheaded, weak, ...
$\phi_3$ : cough, throat, sore, ...

# Semantic Scan Statistic

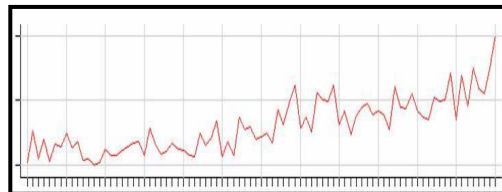
Date	Location	Complaint
1/1/11	15213	runny nose
1/1/11	15217	fever and chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp



$\phi_1$ : vomiting, nausea, diarrhea, ...  
 $\phi_2$ : dizzy, lightheaded, weak, ...  
 $\phi_3$ : cough, throat, sore, ...



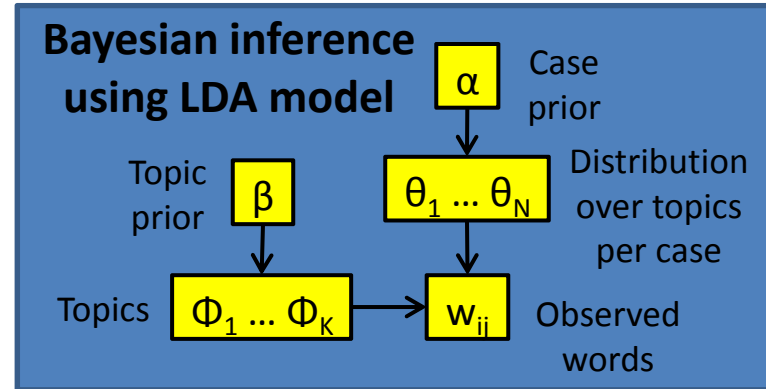
Classify cases to topics



Time series of counts for each location, for each topic T

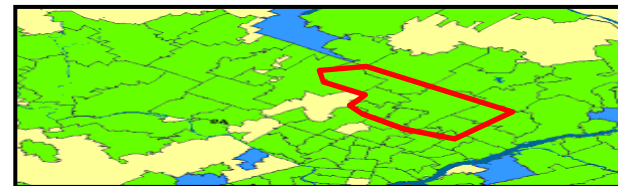
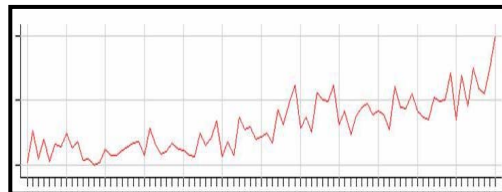
# Semantic Scan Statistic

Date	Location	Complaint
1/1/11	15213	runny nose
1/1/11	15217	fever and chills
1/1/11	15218	broken arm
1/2/11	15101	vomited 3x
1/2/11	15217	high temp



Classify cases to topics

$\phi_1$ : vomiting, nausea, diarrhea, ...  
 $\phi_2$ : dizzy, lightheaded, weak, ...  
 $\phi_3$ : cough, throat, sore, ...

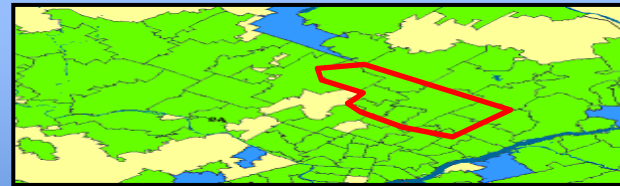


Time series of counts for each location, for each topic T

Find topic T and region S maximizing the likelihood ratio statistic,  $F(S, T)$

# Fast Subset Scanning

We want to perform a constrained search over **subsets** of locations and data streams, but this is computationally infeasible to do exhaustively.

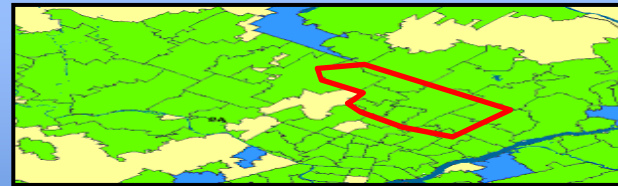


Find topic  $T$  and region  $S$  maximizing the likelihood ratio statistic,  $F(S, T)$

# Fast Subset Scanning

We show\* that it is possible to scan over the exponentially many subsets of the data in linear time, reducing run time from years to milliseconds in practice.

\*D.B. Neill, "Fast subset scan for spatial pattern detection," *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 2012, e-pub ahead of print.



Find topic  $T$  and region  $S$  maximizing the likelihood ratio statistic,  $F(S, T)$



# Results

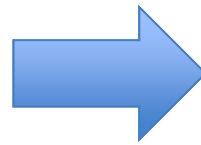
Semantic scan detected outbreaks **more than twice as fast** as the standard prodrome-based method (5.3 days vs. 10.9 days to detect)

# Results

Semantic scan detected outbreaks **more than twice as fast** as the standard prodrome-based method (5.3 days vs. 10.9 days to detect)




Simulated novel outbreak: "green nose"



green  
nose  
possible  
color  
greenish  
nasal  
...

Top words from detected topic

A satellite view of the Earth, showing the Americas and surrounding oceans. The image is centered on the Western Hemisphere, with North America and South America visible. The oceans are a deep blue, and the continents are green and brown. A semi-transparent blue box with a grid pattern is overlaid on the center of the image, containing white text.

For digital disease detection,  
novel methods are essential to  
harness the richness of free text  
data and to address problems at  
the Web- and societal-scale.



THANKS!!!

MORE INFO:  
[NEILL@CS.CMU.EDU](mailto:NEILL@CS.CMU.EDU)

(or see my article in the Jan/Feb 2012  
issue of *IEEE Intelligent Systems*)