Event & Pattern Detection

Carnegie Mellon
HeinzCollege
iLab: Interdisciplinary Research

Research Challenges for Biosurveillance: The Next Ten Years

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

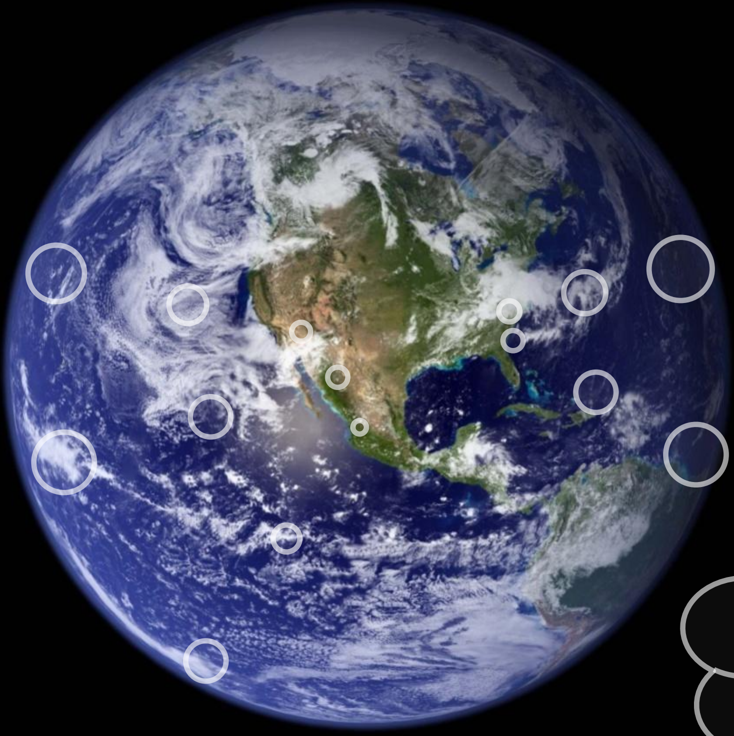
We gratefully acknowledge funding support from the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330.

Goals of this talk

To start the conversation, and get us all thinking about the big picture of disease surveillance research.

What is the current state of the field?

Where are we going in the next ten years?



How is the landscape of biosurveillance changing?

How will these changes affect research and practice?

What will be the new challenges and opportunities?

Three facets of biosurveillance research

Translational: moving discoveries from research to public health practice.

Evolutionary: doing the things we already do well, even better.

Revolutionary: what will be the game-changers in the next ten years?

All three are important – but we probably spend too much time on "evolutionary" research, at the expense of the other two.

Evolutionary research

The past ten years have been marked by many major advances in analytical methods for outbreak detection.

Analysis of **spatial/temporal** data

1) Spatial and space-time scans

2) Irregularly shaped regions

3) Expectation-based approaches

Fusion of multiple streams

1) Multivariate temporal analysis

2) Multivariate spatial scans

3) Network-based approaches

Bayesian modeling and detection

1) Differentiating multiple event types

2) Learning models from data

3) Incorporating user feedback

And many more…

1) Improved NLP & case classification

2) Realistic outbreak simulations

3) Better evaluation metrics

Further refinement of these methods, and their translation into practical tools, will continue to be important focus areas in the years to come.

Translational research

Moving biosurveillance methods from research discoveries to practical and useful tools requires **cooperation** and **communication** between researchers and practitioners.

From a researcher’s perspective, one of the main challenges is still availability of health data.

Available data lacks sufficient **resolution** and **richness**.

Lack of a **gold standard** makes evaluation difficult.

Risk of **overfitting** the available datasets.

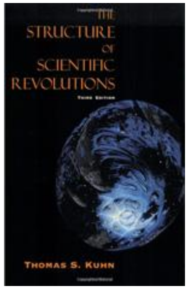
To bridge the gap between research and practice, we need to **systematize** sharing of data and tools within the community.

Need development framework to plug in new methods, test, and provide feedback.

Design more flexible and responsive methods: **learn** what's relevant, limit false positives.

Prioritize software development, documentation, outreach.

Revolutionary research



Kuhn: scientific revolutions result from changes in the underlying landscape/context.

New circumstances =
New possibilities!

Data-Driven World

Increased digitization (EMR, lab tests, pathology slides)

Data from emerging technologies (cellular telephones, Internet search queries, user-generated web content)

Increased individual participation (e-health, HealthMap)

In the next decade, potentially available data sources will continue to increase in **number**, **size**, and **complexity**.

Even applying existing methods to novel data sources can advance biosurveillance practice…

But new methods will be needed to fully harness their incredible potential.

This will require development of increasingly advanced analytics:

Statistical methods to distinguish relevant from irrelevant patterns

Computational algorithms to process the massive quantities of data

Machine learning approaches to improve performance from user feedback

Harnessing complexity

Increasing **richness** and **complexity** of available data, thanks to the rapid growth of new and transformative technologies.

We must develop methods to exploit the new, richer data sources of the future:

Text data from individual patient records

Location data tracking movements

Network data describing social relationships and interactions.

Online data- Google, Facebook, Twitter…

These will not just improve the timeliness and quality of detection, but will fundamentally change what disease surveillance can accomplish!

Rich text data from individual patient records will enable us to detect **newly emerging infections**, with patterns of symptoms that do not correspond to typical groupings, from a very small number of cases.

“fever and green skin”

Possible solution: identify emerging topics (probability distributions over words) from data, and search over multiple topics as well as over time and space.

Harnessing complexity

Increasing **richness** and **complexity** of available data, thanks to the rapid growth of new and transformative technologies.

We must develop methods to exploit the new, richer data sources of the future:

Text data from individual patient records

Location data tracking movements

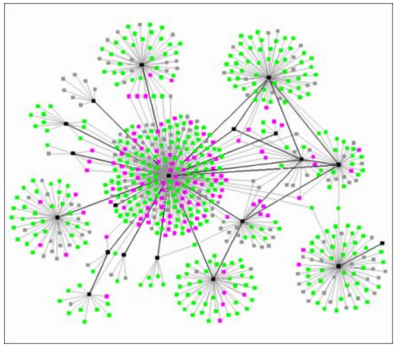
Network data describing social relationships and interactions.

Online data- Google, Facebook, Twitter…

Detailed graph and network data describing individuals, their movements, social relationships, and day-to-day interactions will enable us to automate the challenging task of **epidemiological contact tracing**.

Reverse 911 (Krishnan et al.): Use proximity data from cellular phones to identify who an infectious individual has been in possible contact with.

Call and SMS data could be used to estimate the probability of infection.



Harnessing complexity

Increasing **richness** and **complexity** of available data, thanks to the rapid growth of new and transformative technologies.

We must develop methods to exploit the new, richer data sources of the future:

Text data from individual patient records

Location data tracking movements

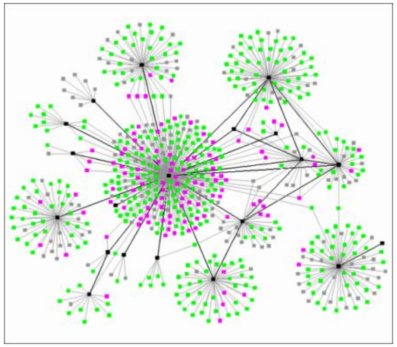
Network data describing social relationships and interactions.

Online data- Google, Facebook, Twitter…

Detailed graph and network data describing individuals, their movements, social relationships, and day-to-day interactions will enable us to automate the challenging task of **epidemiological contact tracing**.

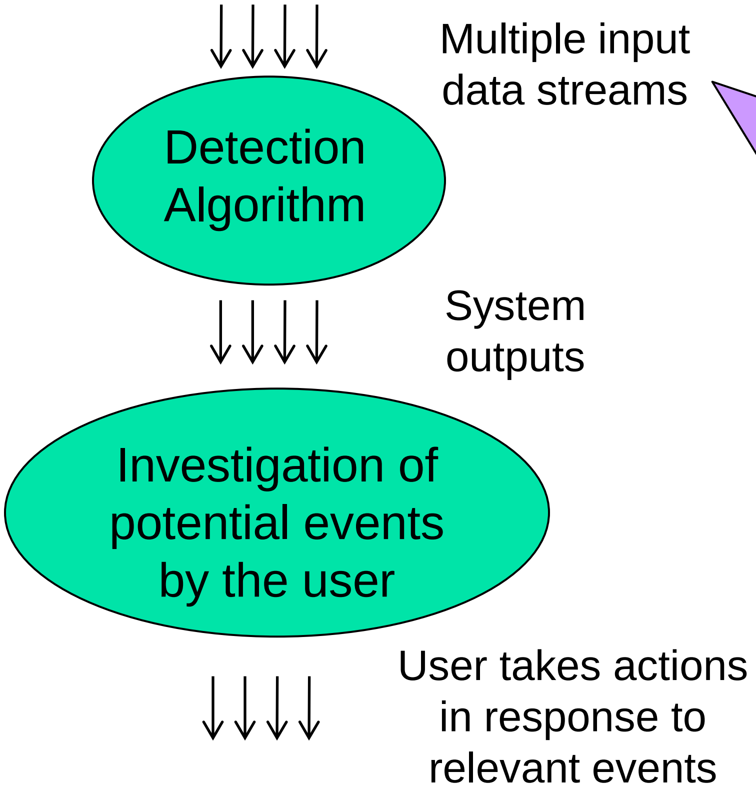
An even larger challenge is to combine multiple, noisy sources of data (e.g. contacts, movement patterns, and health data) from <u>all</u> individuals in the population.

<u>Goals</u>: automatically determine whether an outbreak is emerging, and which individuals are potentially infected.



End-to-end surveillance

Enabling a timely and appropriate response to emerging threats requires more than clever detection algorithms… we must optimize the entire end-to-end process of outbreak detection, investigation and response.



Multiple input data streams

Detection Algorithm

System outputs

Investigation of potential events by the user

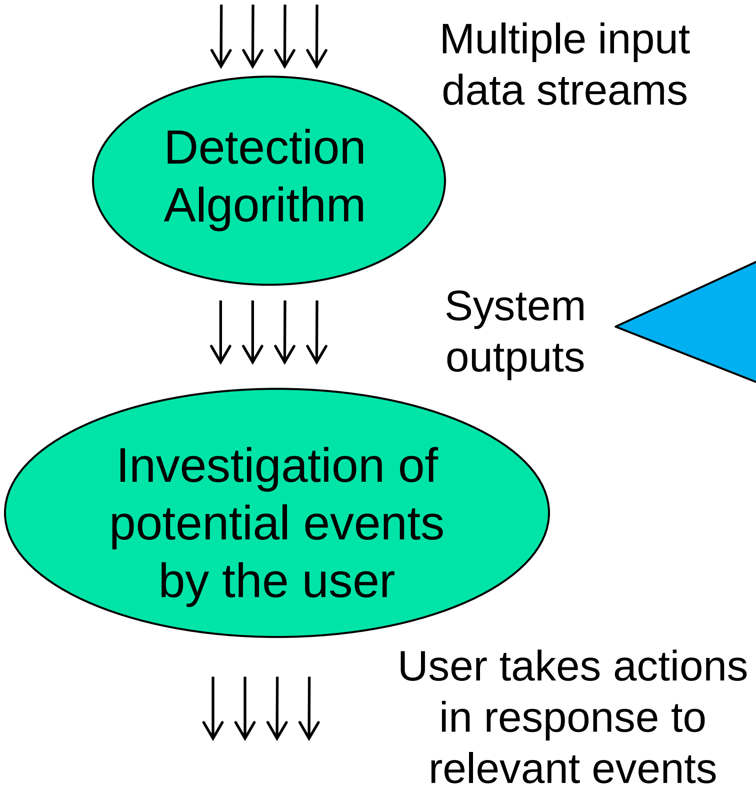
User takes actions in response to relevant events

Which data to collect? While the number of potentially available data sources will be huge, data acquisition will still be costly, particularly in developing countries with limited resources and little existing health infrastructure.

Challenge: develop new methods to **prioritize data sources** for acquisition, via simulation and probabilistic inference.

End-to-end surveillance

Enabling a timely and appropriate response to emerging threats requires more than clever detection algorithms… we must optimize the entire end-to-end process of outbreak detection, investigation and response.

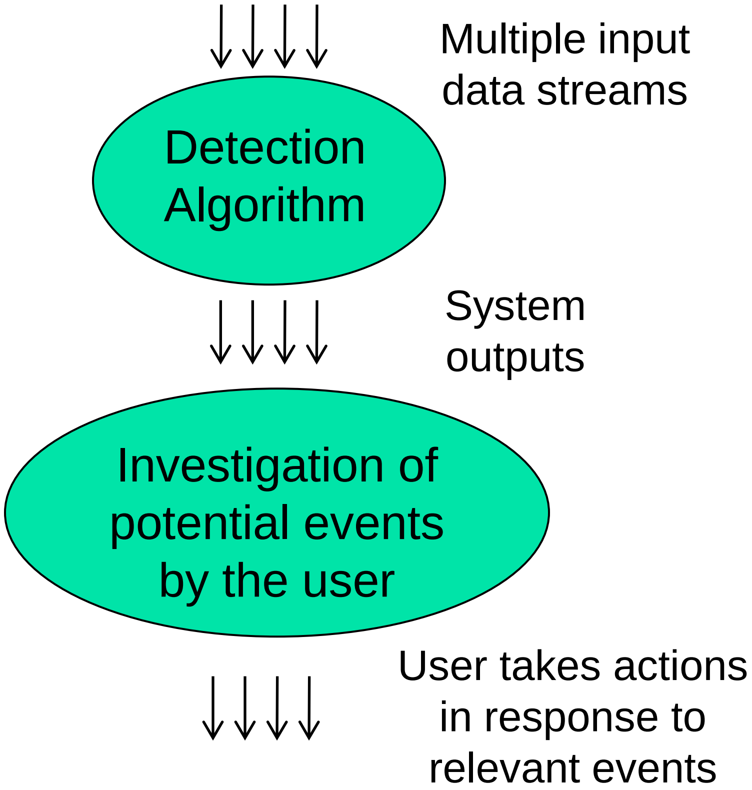

Multiple input data streams
Detection Algorithm
System outputs
Investigation of potential events by the user
User takes actions in response to relevant events

Future systems will not only **detect** events, but **characterize** these events, **explain** them to the user, and provide useful **visualizations** and **exploratory analysis tools**.

In addition to contact tracing, graph-based methods could be used for **back-tracing** (to identify the source of a food-borne or water-borne illness), and **tracking** events over time.

End-to-end surveillance

Enabling a timely and appropriate response to emerging threats requires more than clever detection algorithms… we must optimize the entire end-to-end process of outbreak detection, investigation and response.



Multiple input data streams

Detection Algorithm

System outputs

Investigation of potential events by the user

User takes actions in response to relevant events

Challenge: can we exploit the collective brainpower and spare cycles of the millions of non-expert users via the Internet?

“Harnessing the Wisdom of Crowds”

As the amount of available data increases, leveraging the expertise of multiple users by providing new tools for **collaboration** and **communication** will become increasingly important.

Scaling up surveillance

New, fast detection algorithms will be needed to deal with the increasingly huge amounts of available data.

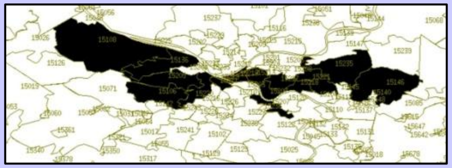
Huge number of data records
Monitoring every individual in a population, or monitoring content from the entire Web.

Many data sources, fine-grained streams, very high spatial and temporal resolution.

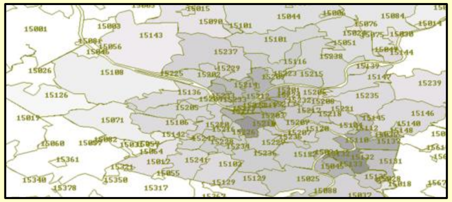
Key idea: subset scan
Detection can be treated as a **search** problem: optimize a function over all subsets of the data.

We can perform computations over all subsets without evaluating each one individually!

Fast subset scan: find the **best** subset of locations and data streams, subject to spatial proximity or graph connectivity constraints.



Fast subset sums: compute and visualize the posterior probability distribution over multiple locations and multiple event types.



Scaling up surveillance

New, fast detection algorithms will be needed to deal with the increasingly huge amounts of available data.

Huge number of data records

Monitoring every individual in a population, or monitoring content from the entire Web.

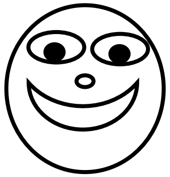
Many data sources, fine-grained streams, very high spatial and temporal resolution.

Key idea: subset scan

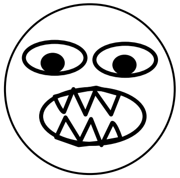
Detection can be treated as a **search** problem: optimize a function over all subsets of the data.

We can perform computations over all subsets without evaluating each one individually!

These methods reduce computation time for some detection problems from millions of years to milliseconds…



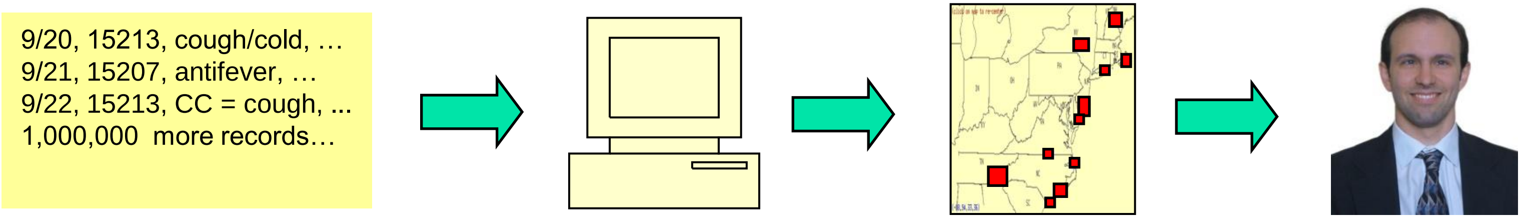
… but for the **society-scale data** of the future, where we cannot even look at all the data in a reasonable time, these algorithms will be insufficient.



Other algorithmic techniques (e.g. sub-linear time methods based on sampling, multi-resolution and multi-scale methods, and parallel/distributed processing) must be developed and incorporated into practical surveillance systems.

Detecting relevant and useful patterns

We must make disease surveillance systems more **adaptive** and more **interactive**, helping users to make sense of the ever-increasing mass of data.



9/20, 15213, cough/cold, …
9/21, 15207, antifever, …
9/22, 15213, CC = cough, ...
1,000,000 more records…

Huge mass of data

Detection algorithm

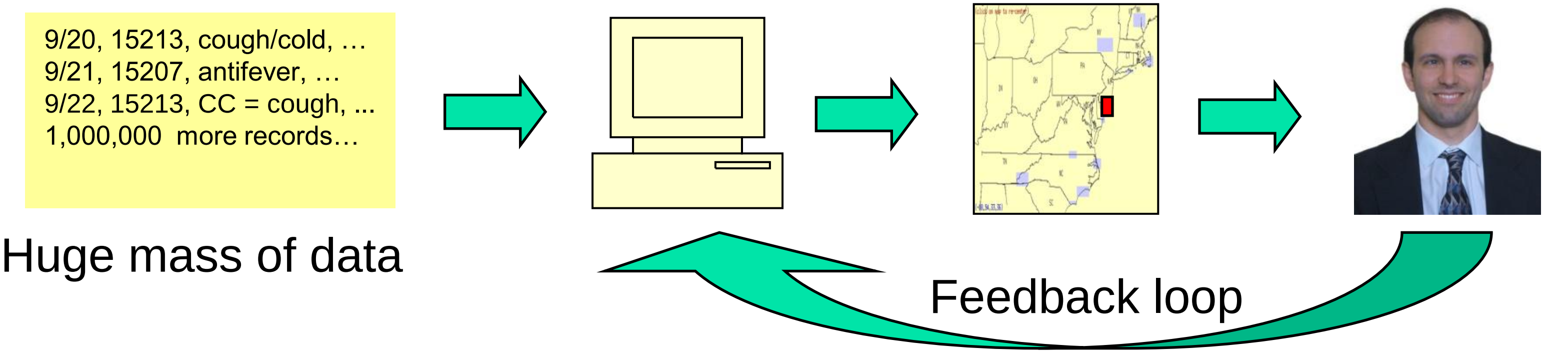
Too many alerts

“What am I supposed to do with this?”

Most existing methods are based on **anomaly detection**.
But any real-world dataset contains a huge number of anomalous observations, very few of which are actually relevant to the user.

Detecting relevant and useful patterns

We must make disease surveillance systems more **adaptive** and more **interactive**, helping users to make sense of the ever-increasing mass of data.



9/20, 15213, cough/cold, …
9/21, 15207, antifever, …
9/22, 15213, CC = cough, ...
1,000,000 more records…

Huge mass of data

Feedback loop

Challenge: incorporate user feedback to **learn** which of the many anomalous patterns in the data are actually relevant, reducing the number of false positives.

Possible solution: Learn **models** of multiple event types and incorporate into event detection.

Report patterns corresponding to any known and relevant event type, such as an anthrax bio-attack, as well as highly anomalous unknown patterns, such as a newly emerging infection.

Distinguish these from events which are known and irrelevant, such as promotional sales.

Incorporate feedback to continually expand and improve the set of models over time.

Conclusions

The increasingly **data-driven** landscape in which future disease surveillance systems will operate presents both great opportunities and great challenges for the next decade of biosurveillance research.

We must prepare to take advantage of these opportunities, developing practical new methods which not only improve detection, but enable fundamental changes in what biosurveillance systems can accomplish.

To do so, we must deal not only with existing challenges such as data availability and gaps between theory/practice, but anticipate the new statistical and computational challenges resulting from massive data.

Most of all, we must never lose sight of our most fundamental goals as disease surveillance researchers: to look into the future of biosurveillance, and bring that future into the present.