Machine Learning and Event Detection for Urban Public Health

Daniel B. Neill, Ph.D.

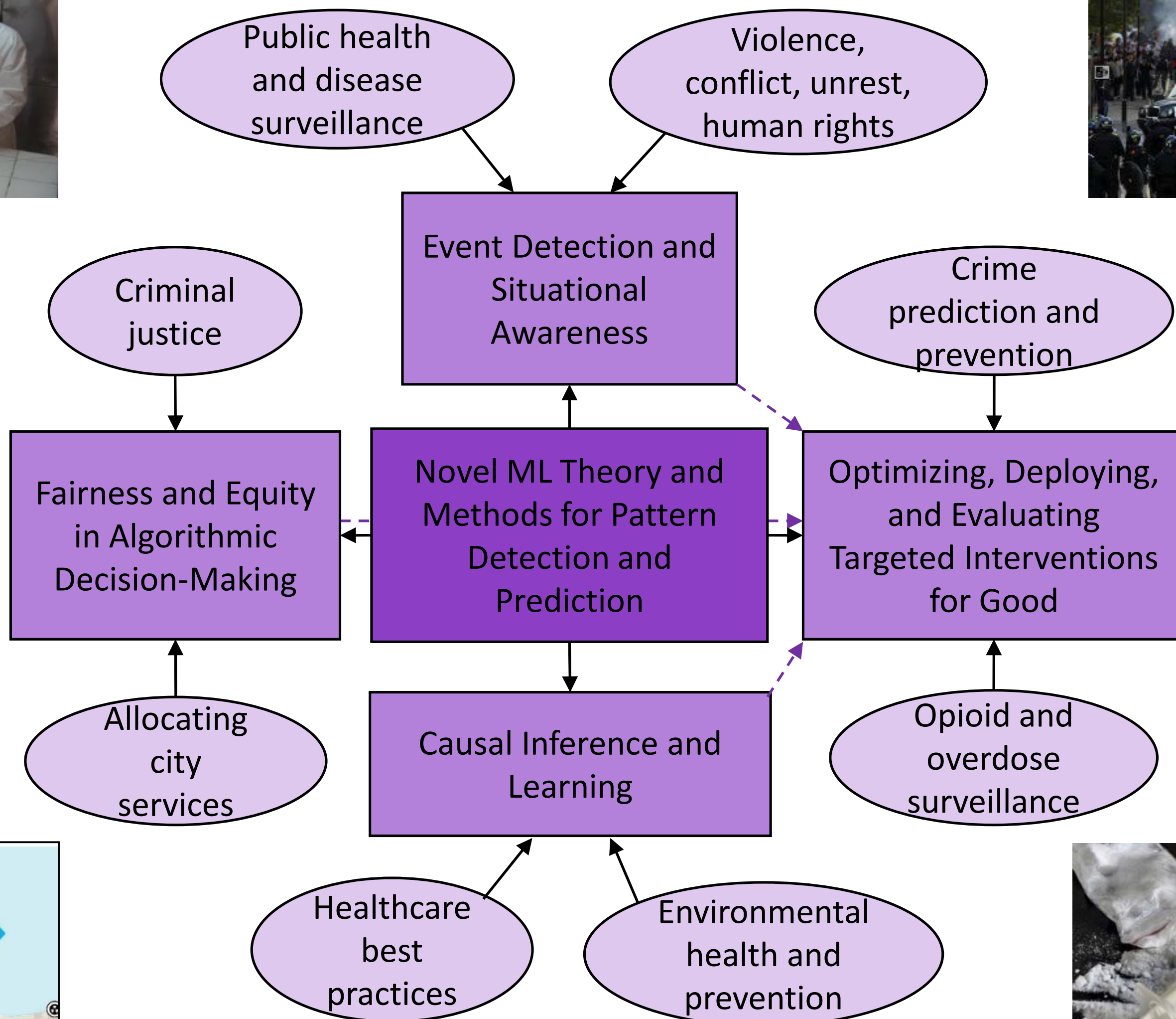
Associate Professor of Computer Science and Public Service
Associate Professor of Urban Analytics, NYU CUSP
Director, Machine Learning for Good (ML4G) Laboratory

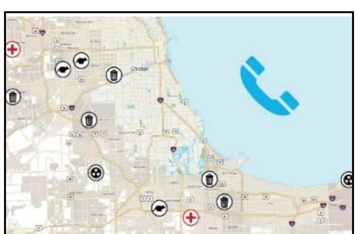
New York University

E-mail: daniel.neill@nyu.edu
Web: http://www.cs.nyu.edu/~neill
http://wp.nyu.edu/ml4good

NYU | Center for Urban Science + Progress

The Machine Learning for Good Lab @ NYU





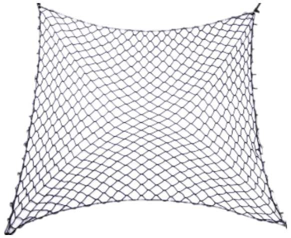
How can machine learning improve urban public health?

Event & pattern detection

Better situational awareness

Earlier & more precisely targeted interventions

Interventions to combat the opioid overdose crisis



Providing a safety net for novel disease outbreaks and emerging public health threats

Drug overdoses

- Drug overdoses are an increasingly serious problem in the United States and worldwide.
  - In 2020, more than 93,000 drug overdose deaths occurred in the U.S., more than any year in recorded history.
  - Nearly three-quarters of these overdose deaths involved opioids.
  - Economic costs of the crisis have been estimated at between $78.5 billion and >$1 trillion annually.
- These statistics motivate public health to identify and predict emerging trends in overdoses (geographic, demographic, and behavioral) to better target interventions.
  - **Prevention** of high-risk prescribing and opioid use behaviors
  - **Treatment** of opioid addiction, e.g., medication-assisted therapy
  - **Rescue**, e.g., access to life-saving naloxone
  - **Recovery**, e.g., peer recovery coaches

Drug overdoses

- Drug overdoses are an increasingly serious problem in the United States and worldwide.
  - In 2020, more than 93,000 drug overdose deaths occurred in the U.S., more than any year in recorded history.
  - Nearly three-quarters of these overdose deaths involved opioids.
  - Economic costs of the crisis have been estimated at between $78.5 billion and >$1 trillion annually.
- These statistics motivate public health to identify and predict emerging trends in overdoses (geographic, demographic, and behavioral) to better target interventions.
- Machine learning has potential to **save lives** by detecting subtle, emerging patterns of overdoses in their early stages and targeting an effective public health response at the geographic, subpopulation, individual, and network levels.

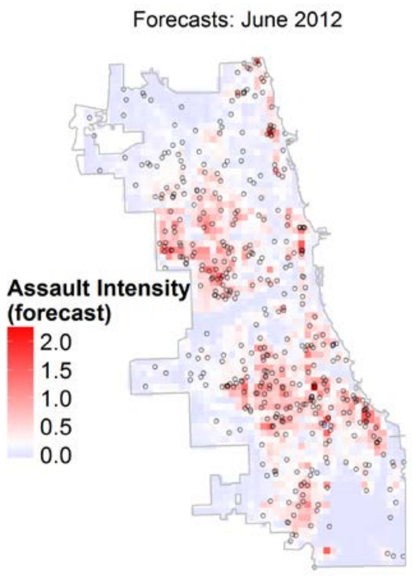
Geographic surveillance

- Answers the question, **where** should I intervene?
- <u>Main goals</u>: estimate predicted overdose trends in space and time; identify anomalous spikes in overdose deaths.

Our work on **scalable Gaussian processes*** achieves state-of-the-art accuracy for long-term, small-area forecasting.

Useful predictors include neighborhood characteristics and recent spatio-temporal trends in overdoses and leading indicators.

We are currently integrating multiple data sources (ME, EMS, PDMP, census) to **predict overdose risk** and **target interventions** in RI.



*SR Flaxman, AG Wilson, DB Neill, H Nickisch, AJ Smola. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. Proc. 32nd Intl. Conf. on Machine Learning, *PMLR* 37: 607-616, 2015.

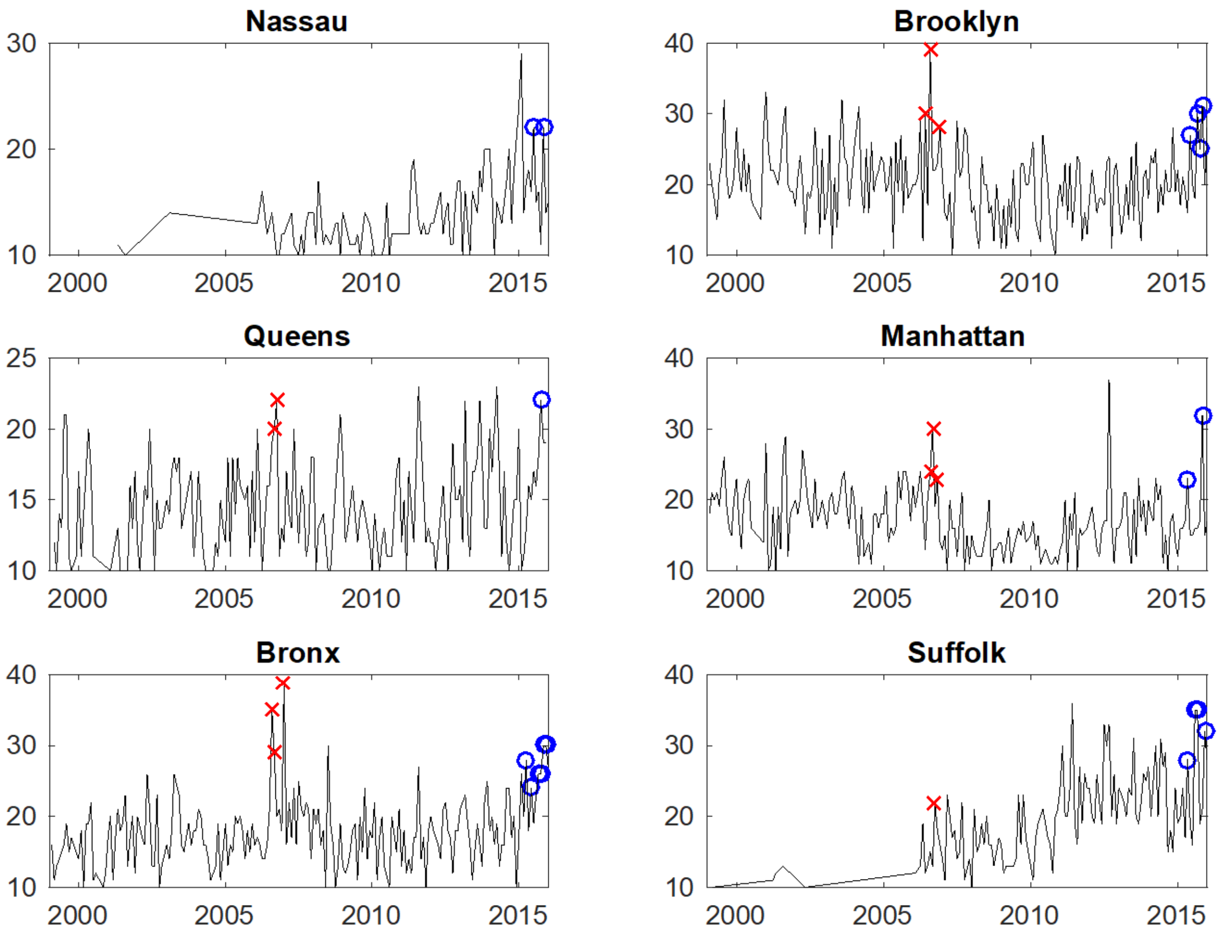
Case study: Geographic surveillance

- We analyzed **aggregate monthly counts** of fatal opioid overdoses for six New York counties from 1999-2015.
- We developed a new detection approach* which combines **Gaussian processes** (to model correlations) and **subset scan** (to identify the most anomalous space-time regions).
- We compared our new method to typical anomaly detection approaches on real and synthetic datasets.
  - GPSS > GP alone: nearby points matter for subtle anomalies
  - GPSS > SS alone: covariance structure matters for correlated data

*W Herlands, E McFowland III, AG Wilson, DB Neill. Gaussian process subset scanning for anomalous pattern detection in non-iid data. *Proc. 21st Intl. Conf. on Artificial Intelligence and Statistics, PMLR 84*: 425-434, 2018.

Case study: Geographic surveillance

Two statistically significant spikes in overdose cases:

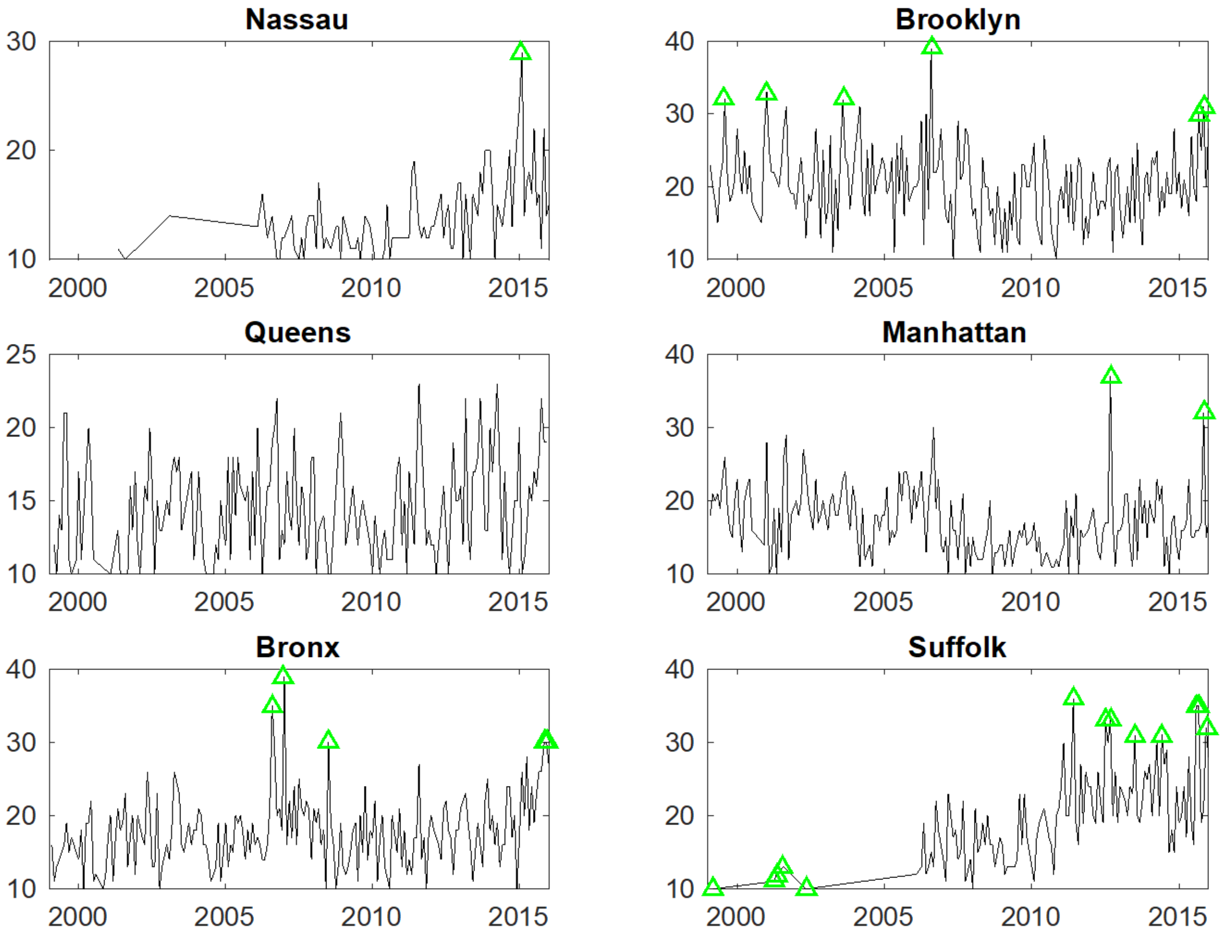


Mid 2006. Just before naloxone programs.

End of 2015. Recent surge due to fentanyl.

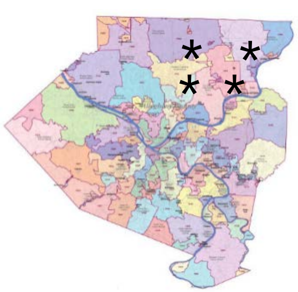
Case study: Geographic surveillance

Simpler anomaly detection methods fail to capture the relevant trends.



Subpopulation-level monitoring

- Answers the question, **for whom** should I intervene?
- <u>Main goal</u>: provide early warning for newly emerging subpopulation-level spikes/clusters of overdose deaths.
- We developed a novel detection method, **multi-dimensional tensor scan**, to detect emerging geographic, demographic, and behavioral patterns.
  - **Earlier detection** of emerging overdose clusters through daily surveillance runs.
  - Better characterization of **where** and **who** is affected.



X

white
males
aged
20-49

X

Subpopulation-level monitoring

- Answers the question, **for whom** should I intervene?
- Main goal: provide early warning for newly emerging subpopulation-level spikes/clusters of overdose deaths.
- We developed a novel detection method, **multi-dimensional tensor scan**, to detect emerging geographic, demographic, and behavioral patterns.
  - **Earlier detection** of emerging overdose clusters through daily surveillance runs.
  - Better characterization of **where** and **who** is affected.
- Analyzed eight years of data from Allegheny County, PA.

Changing demographics of risk

Cluster of 11 fentanyl-related deaths in 2015, elderly Black males in downtown Pittsburgh.

Impacts of policy

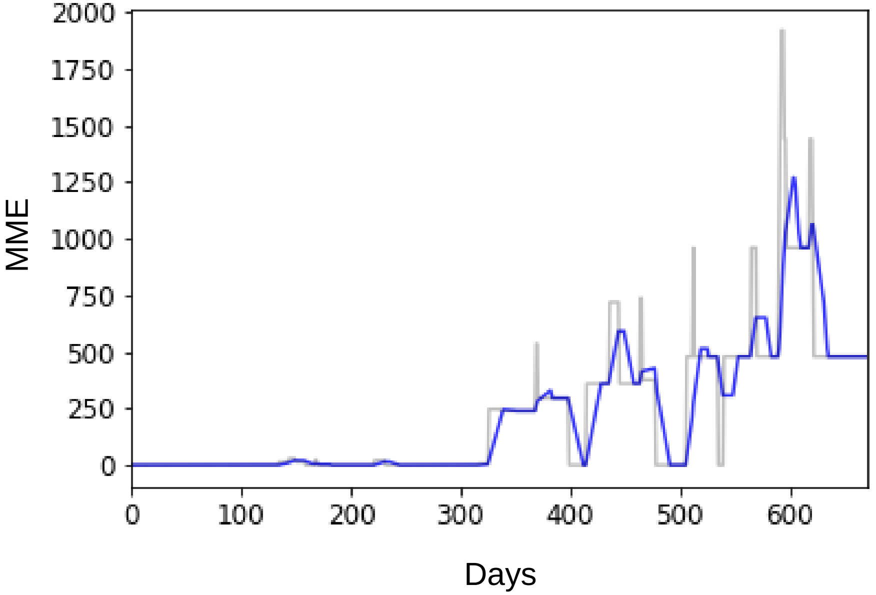
"Methadone + Xanax" overdose clusters were reduced by PA's passage of the Methadone Death & Incident Review Act.

Individual-level opioid use monitoring

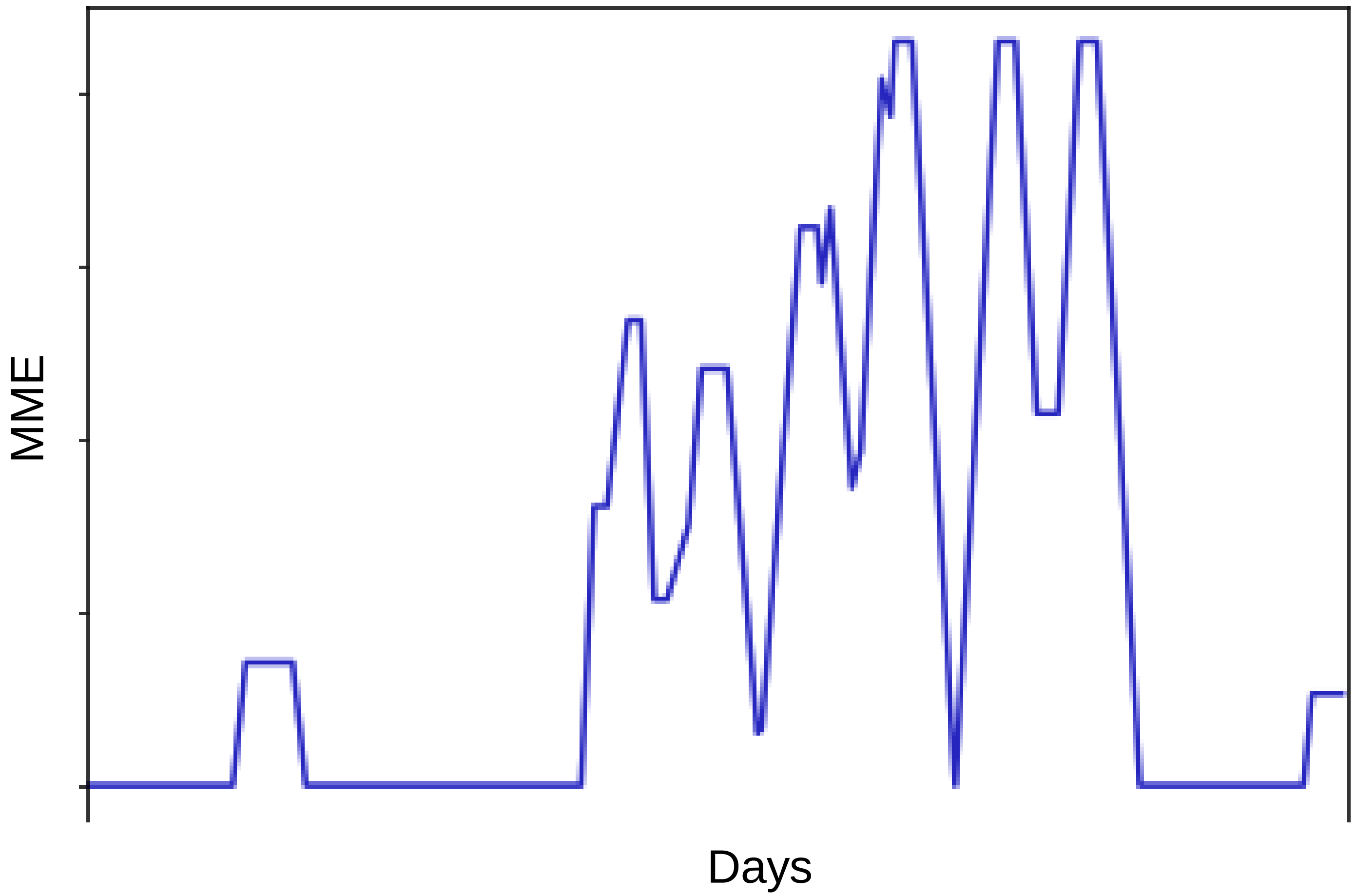
(Joint work with Dylan Fitzpatrick)

- Seven years of de-identified data from over 1M individuals provided by Kansas prescription drug monitoring program (PDMP).
- Duration and quantity of prescribed opioids are used to create timelines of morphine milligram equivalents (MME) for individual patients.
- We were able to identify **early indicators** in patient MME timelines which were highly predictive of later opioid misuse.

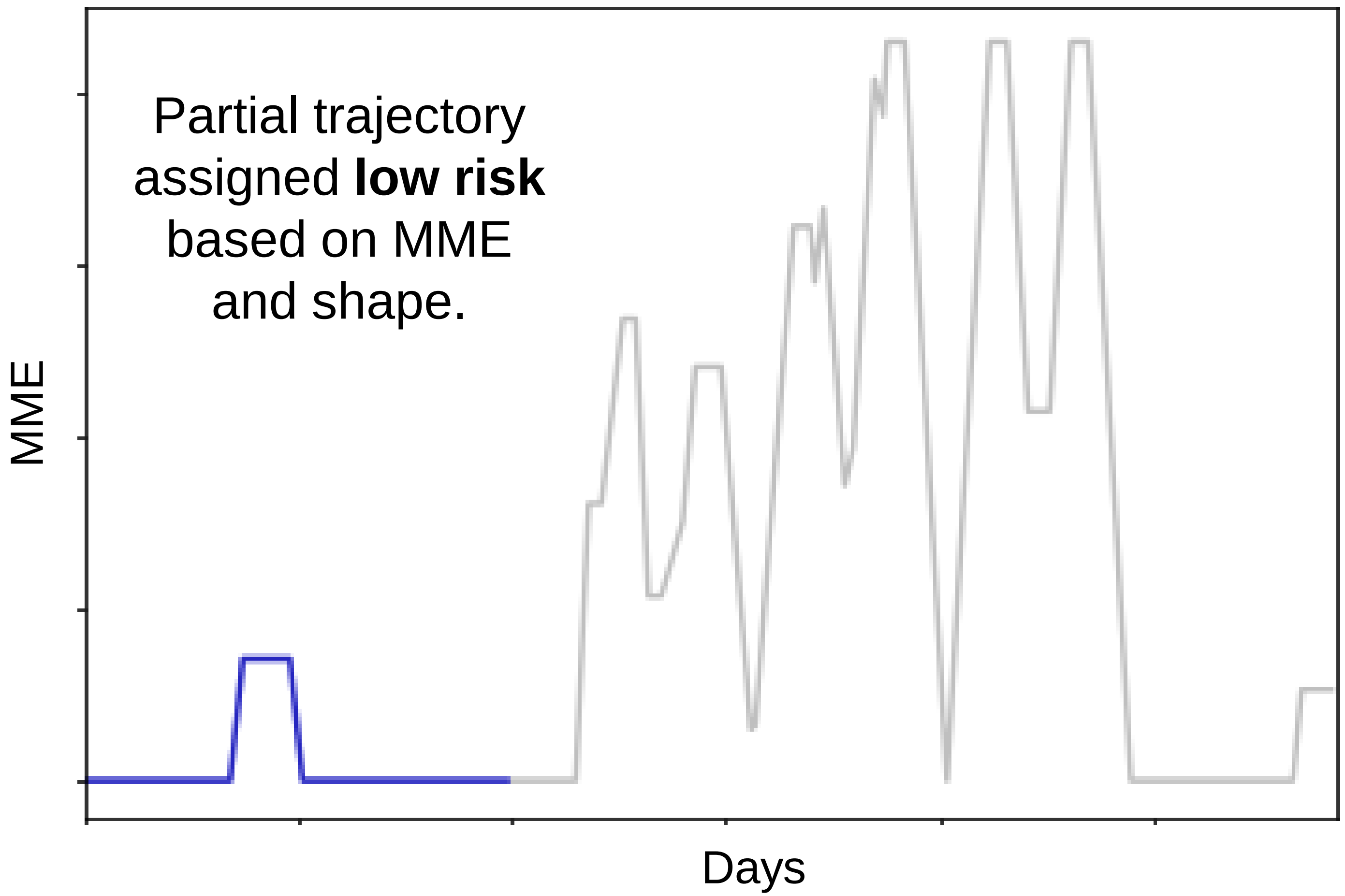
Smoothed MME Timeline for a Single Patient



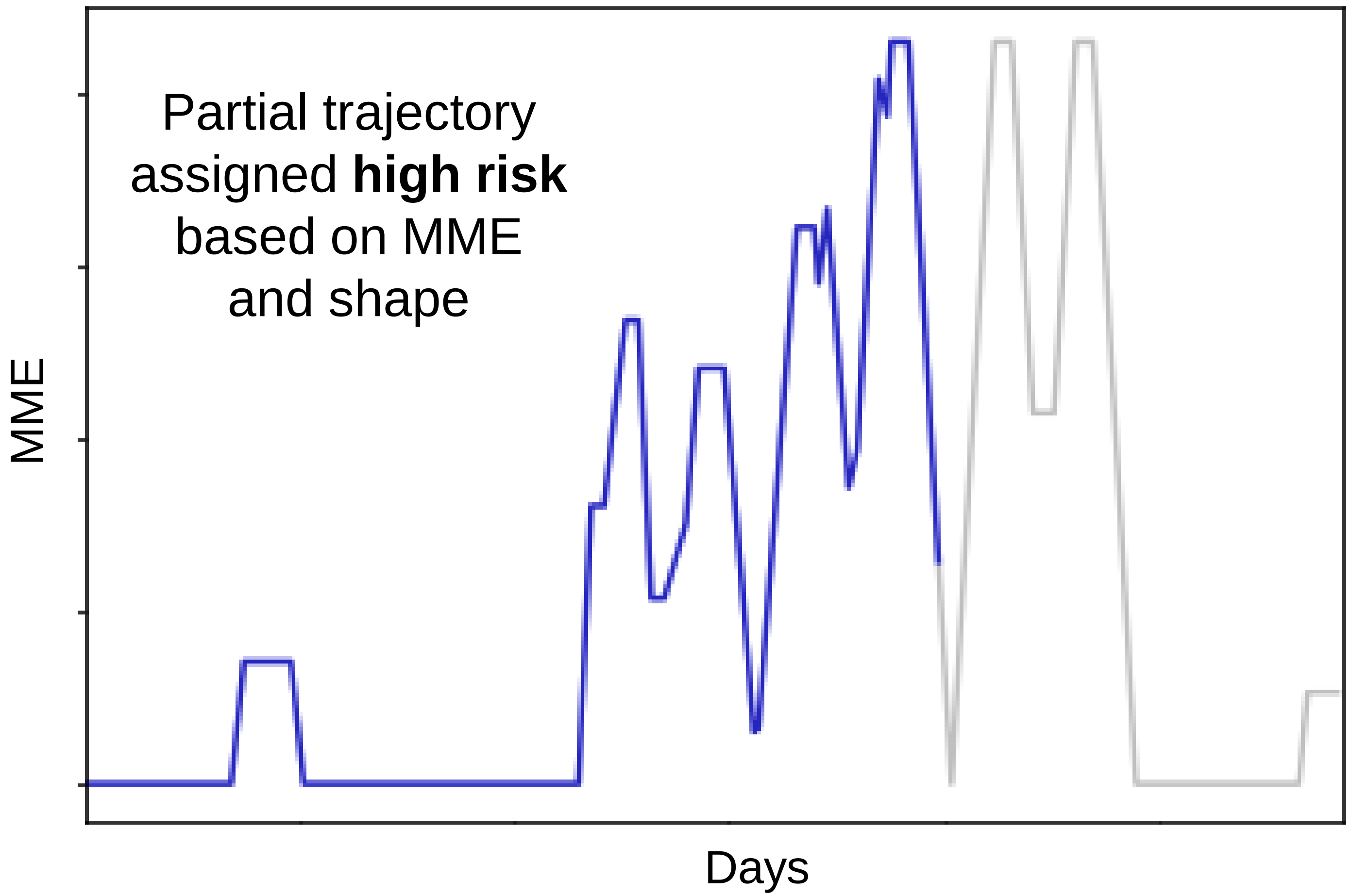
Early individual-level risk assessment by classifying partial trajectories



Early individual-level risk assessment by classifying partial trajectories



Early individual-level risk assessment by classifying partial trajectories



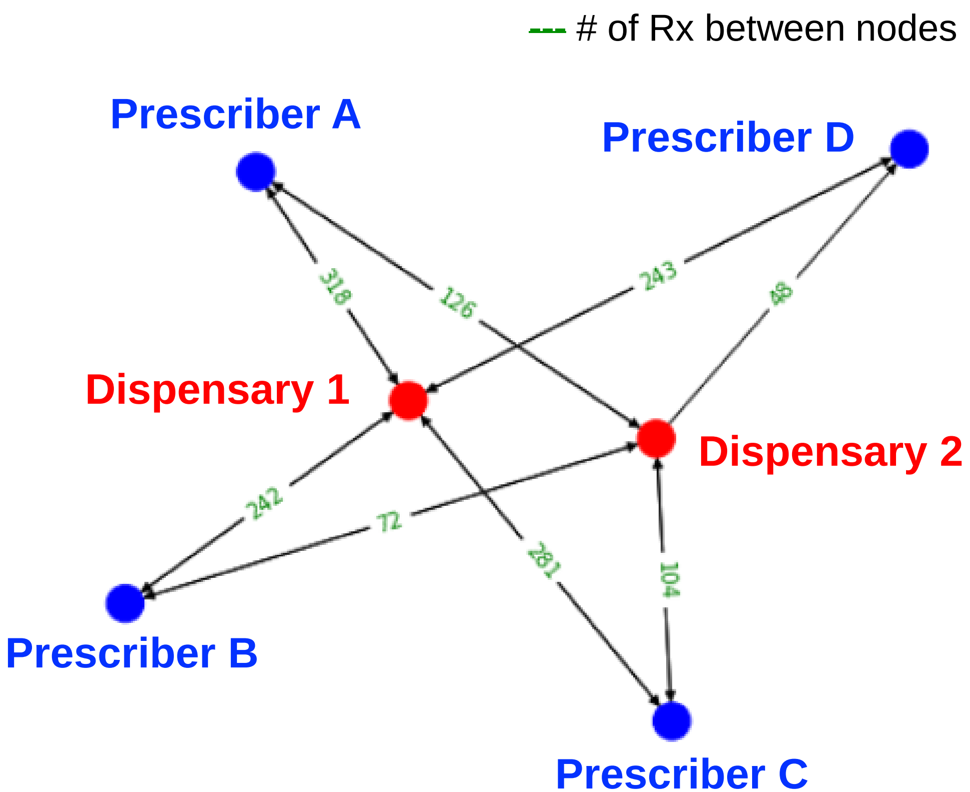
Monitoring networks of prescribers

(Joint work with Katie Rosman)

We are also using the PDMP data for **network analysis**: we identify connected networks of prescribers and dispensaries who are engaging in high-risk and possibly illicit prescribing behaviors.

Step 1: compute the **anomalousness** of each prescriber and dispensary based on Rx and patient-level attributes.

Step 2: Identify the **most anomalous clusters** by maximizing a nonparametric scan statistic over connected subgraphs.



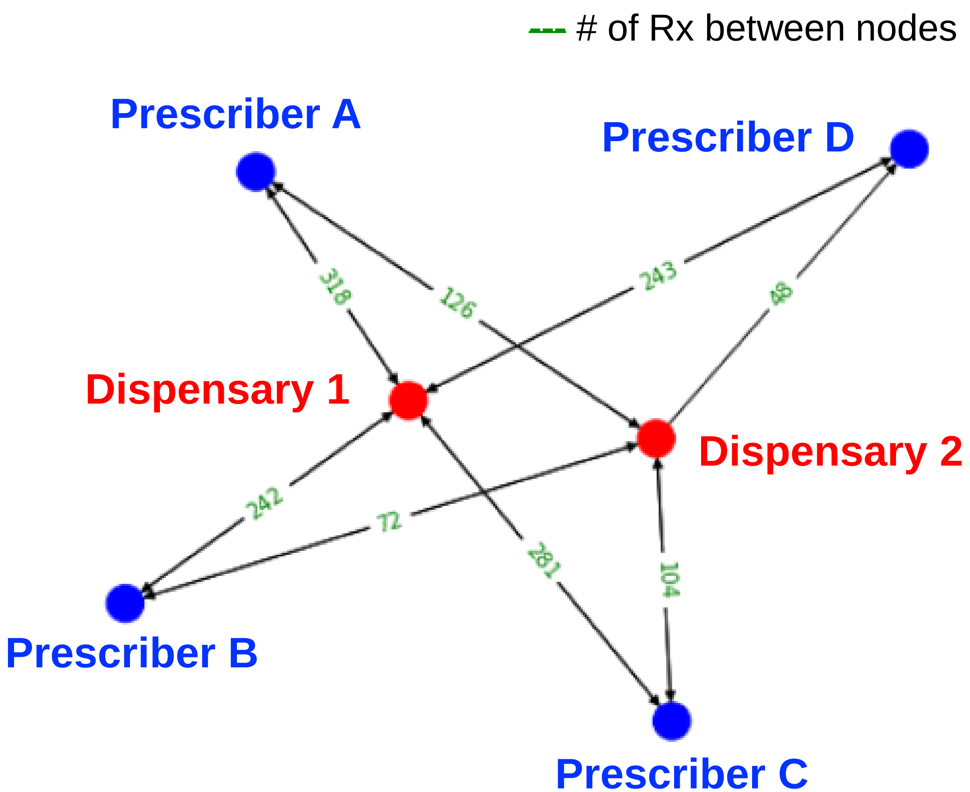
Monitoring networks of prescribers

(Joint work with Katie Rosman)

We are also using the PDMP data for **network analysis**: we identify connected networks of prescribers and dispensaries who are engaging in high-risk and possibly illicit prescribing behaviors.

This detected subgraph of four prescribers and two dispensaries had ~8K prescriptions and ~1,800 patients associated with it.

- 77% of prescriptions were opioids (1.5x expected)
- Average daily dose of opioids per patient was 135 MME (6x expected).
- 30% of prescriptions paid for in cash (3x expected).



Discussion

Here we described several new methods that can be used for **early warning** and **advance forecasting** of overdoses at geographic, subpopulation, individual, & network levels.

Our retrospective analyses of overdose and opioid use data from Pennsylvania, New York, and Kansas suggest high potential utility for **prospective** drug overdose surveillance systems, to facilitate targeted and effective interventions.

We are currently collaborating with an interdisciplinary team of investigators and public health practitioners, with the goals of deploying targeted interventions to prevent overdoses and evaluating their effectiveness through randomized trials.

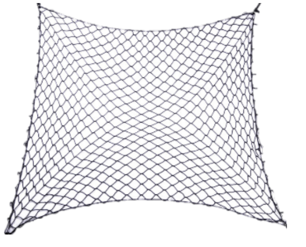
How can machine learning improve urban public health?

Event & pattern detection

Better situational awareness

Earlier & more precisely targeted interventions

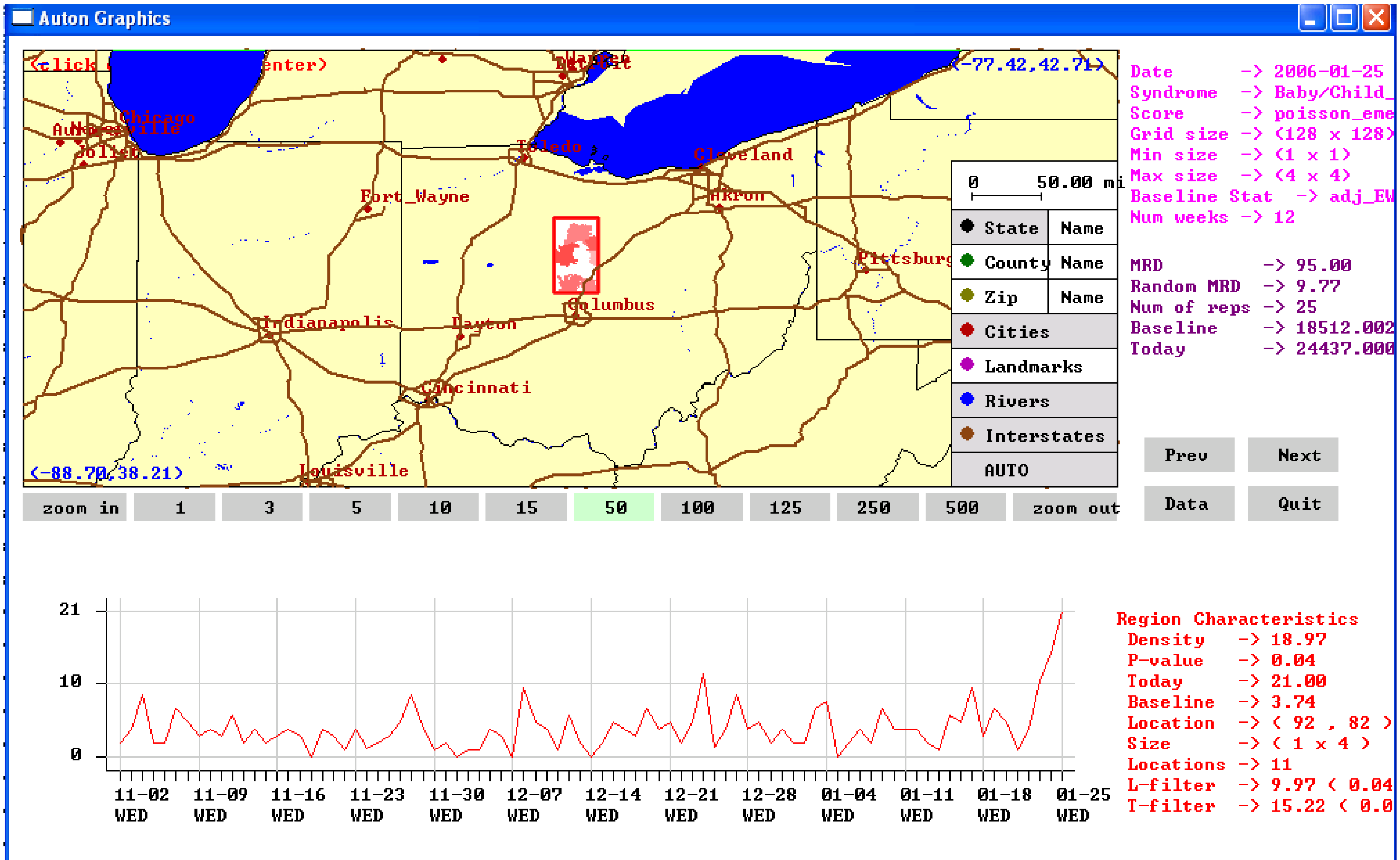
Interventions to combat the opioid overdose crisis



Providing a safety net for novel disease outbreaks and emerging public health threats

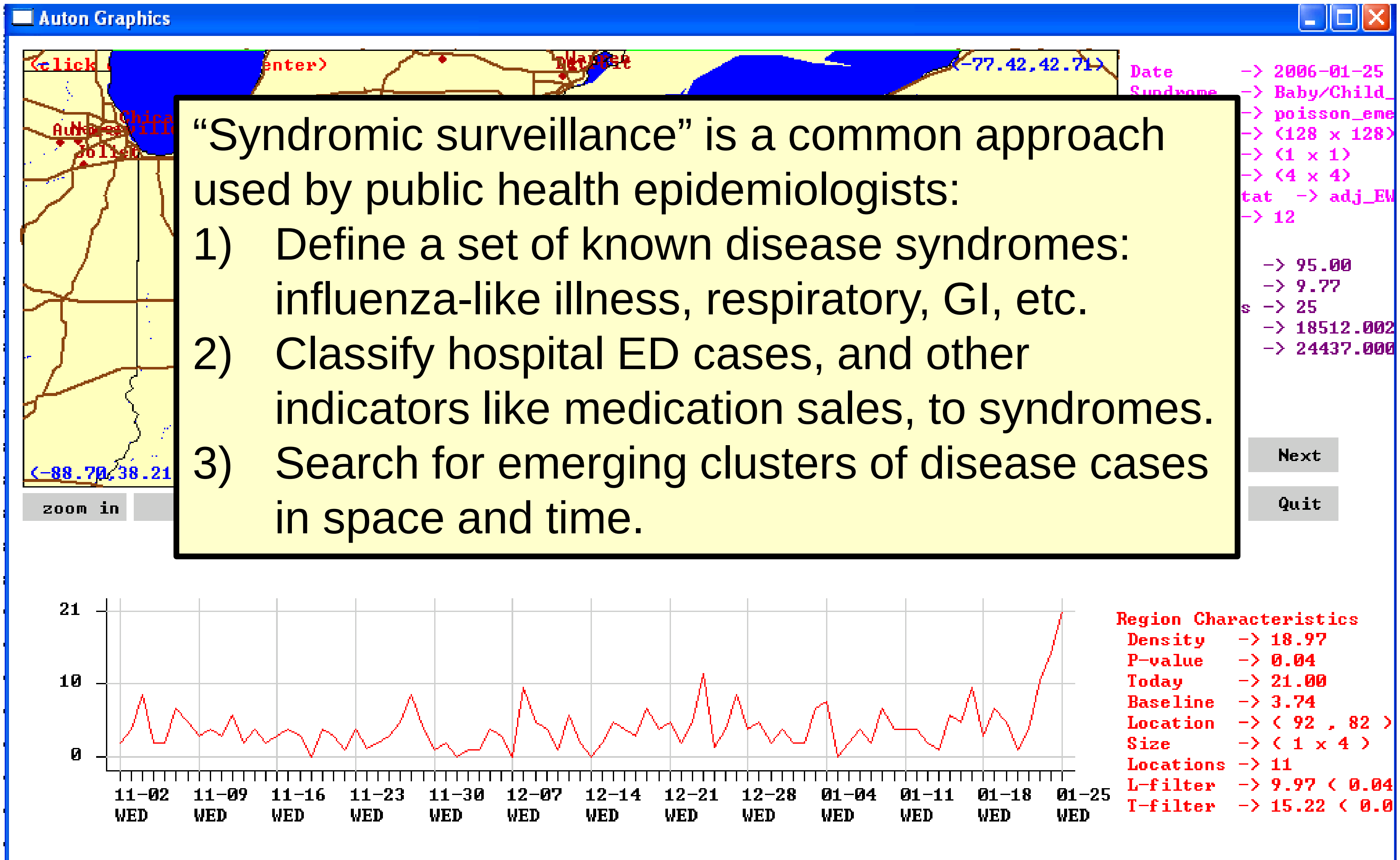
Disease surveillance example

Spike in gastrointestinal (GI) illness near Columbus, Ohio



Disease surveillance example

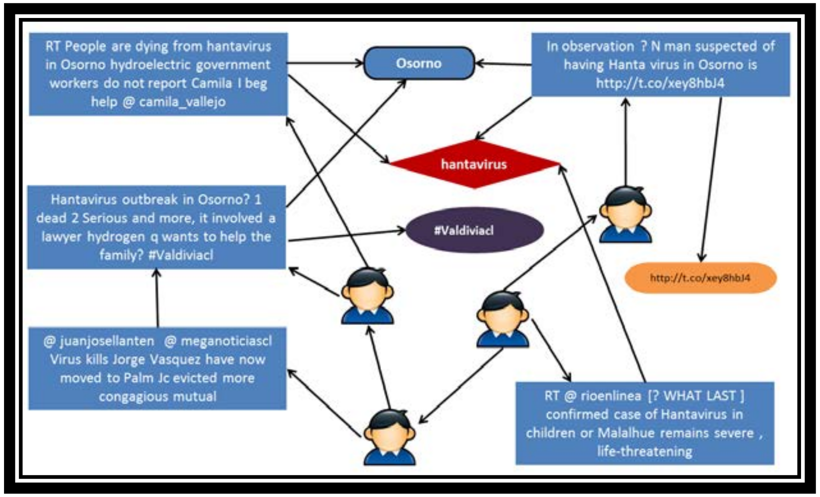
Spike in gastrointestinal (GI) illness near Columbus, Ohio



"Syndromic surveillance" is a common approach used by public health epidemiologists:

1) Define a set of known disease syndromes: influenza-like illness, respiratory, GI, etc.
2) Classify hospital ED cases, and other indicators like medication sales, to syndromes.
3) Search for emerging clusters of disease cases in space and time.

Detecting rare disease outbreaks with Twitter

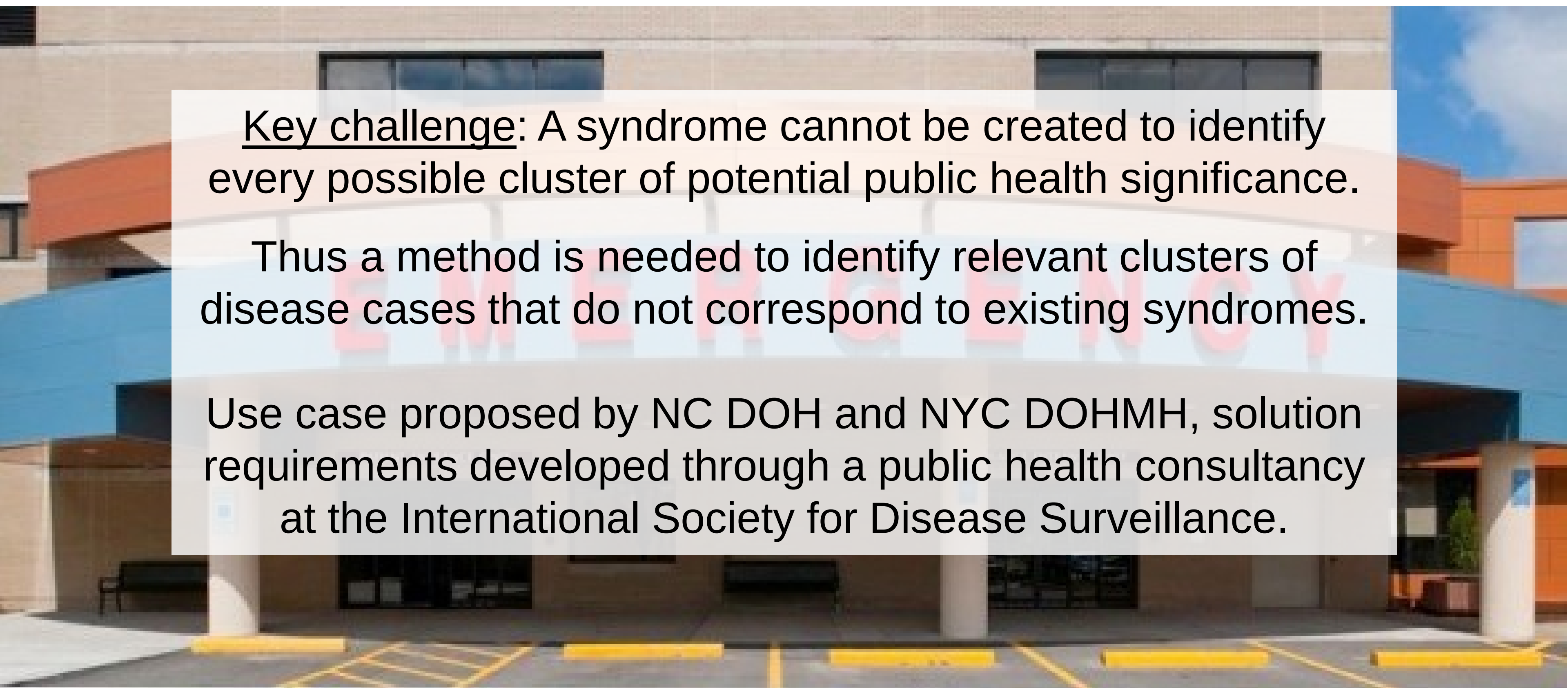


Locations
Users
Keywords
Hashtags
Links
Videos

F. Chen and D.B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. Proc. 20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 1166-1175, 2014.

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



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

Thus 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.

Where do existing methods fail?

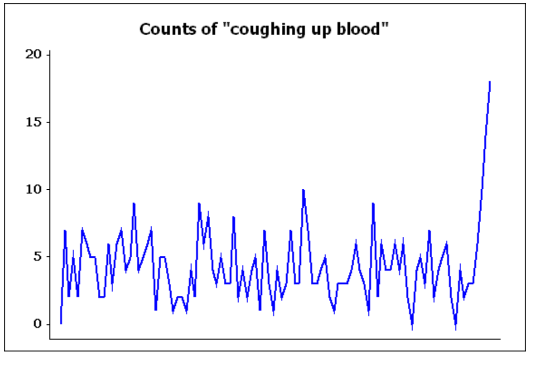
The typical syndromic surveillance approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

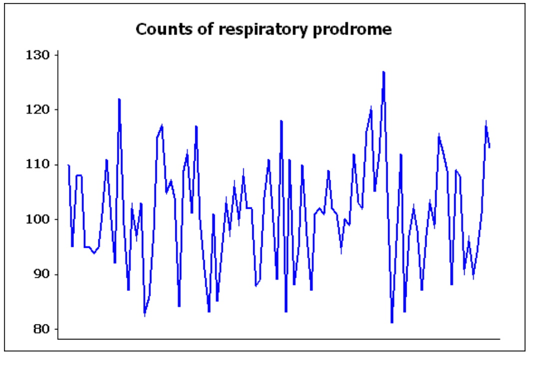
What happens when something new and scary comes along?

- **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.





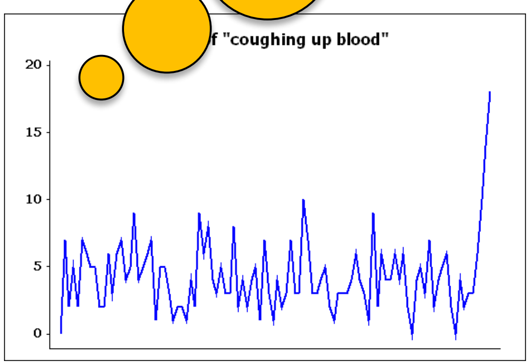
Where do existing methods fail?

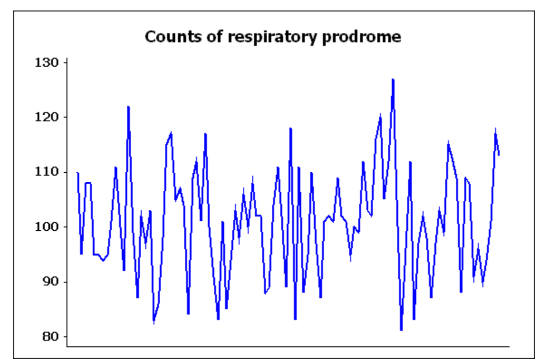
The typical sy... ic
surveillance
effectively
outbrea
seen
sy

...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**.

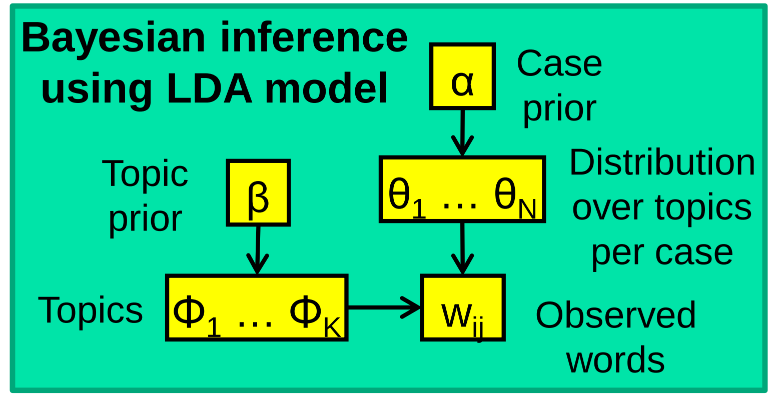
If we w... ief complaints
particular sy... ptom category
take a few such ... ilute the outbreak signal,
that an outb... s occurring! ... ying or preventing detection.





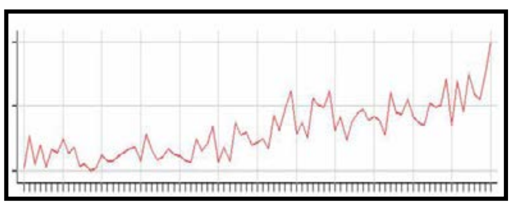
The semantic scan statistic

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



Classify cases to topics

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

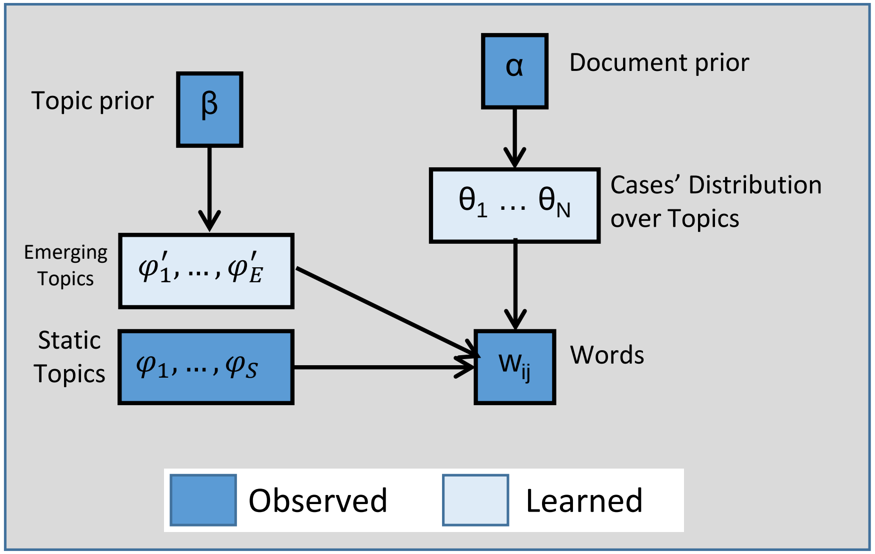


Time series of hourly counts for each combination of hospital and age group, for each topic φ_j.

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

Multidimensional Semantic Scan Learns Two Sets of Topics

- Static Topics
  - Designed to capture common illnesses like flu.
  - Learned over a large set of historical data using a standard LDA topic model.
- Emerging Topics
  - Designed to capture rare or novel diseases that are not well explained by the static topics.
  - Learned over the most recent set of data using a new variant of LDA.



NYC DOHMH dataset

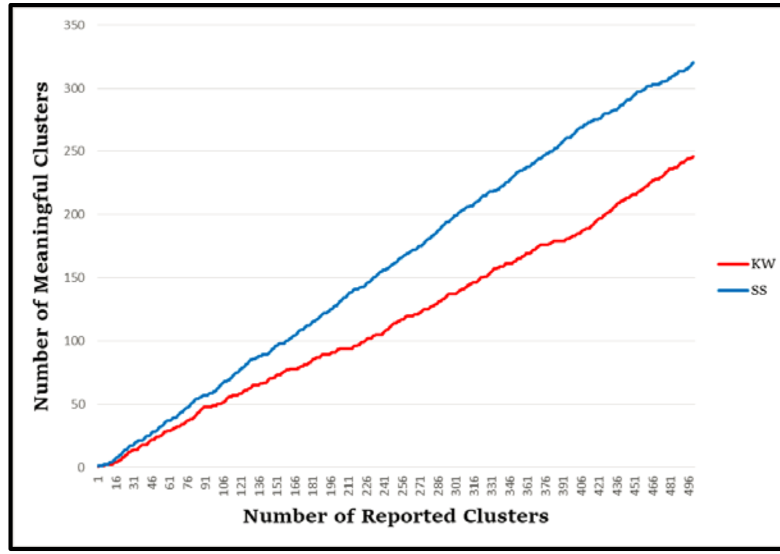
- New York City's Department of Health and Mental Hygiene, Bureau of Communicable Disease, provided us with 6 years of data (2010-2016) consisting of ~28M chief complaint cases from 53 hospitals in NYC.
- For each case, we have data on the patient's chief complaint (free text), date and time of arrival, age group, gender, and discharge ICD-9 code.
- Substantial pre-processing of the chief complaint field was necessary because of the size and messiness of the data (typos, abbreviations, etc.).

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

Evaluation on NYC DOHMH data

- Blinded evaluation by NYC DOHMH public health practitioners, comparing our multidimensional semantic scan approach to a state-of-the-art keyword-based scan approach.
- For each method's 500 highest scoring clusters, users indicated if the cluster is relevant, meaningful, or not of interest.



	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

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:

Acute cases: falls, SOB, leg Injuries → Mental health disturbances: depression, anxiety → Burden on medical infrastructure: methadone, dialysis

Many other events of public health interest were identified:

Accidents	Contagious Diseases	Other
Motor vehicle	Meningitis	Drug overdoses
Ferry	Scabies	Smoke inhalation
School bus	Ringworm	Carbon monoxide poisoning
Elevator	Hepatitis	Crime related, e.g., pepper spray attacks

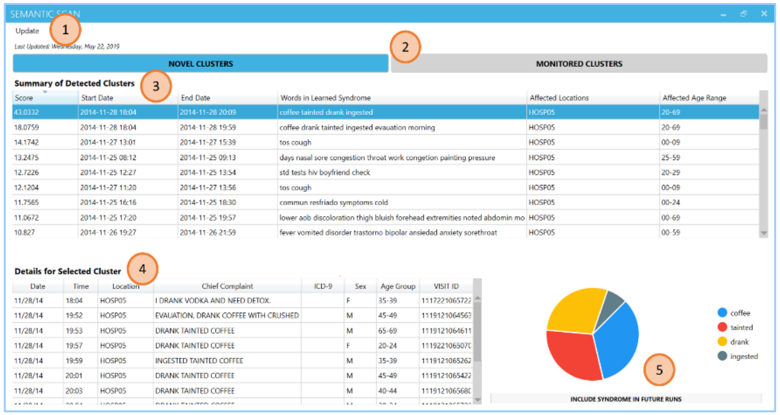
Example of a detected cluster

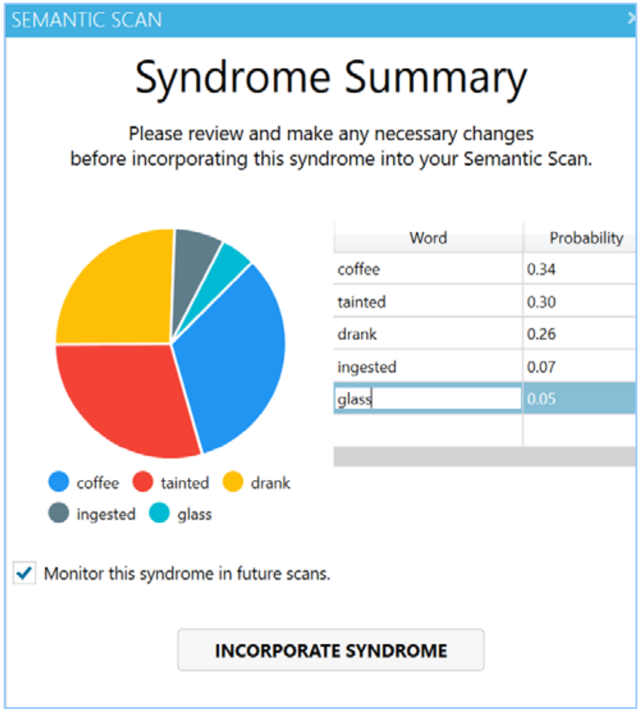
Arrival Date	Arrival Time	Hospital ID	Chief Complaint	Patient Sex	Patient Age
11/28/2014	7:52:00	HOSP5	EVAUATION, DRANK COFFEE WITH CRUS	M	45-49
11/28/2014	7:53:00	HOSP5	DRANK TAINTED COFFEE	M	65-69
11/28/2014	7:57:00	HOSP5	DRANK TAINTED COFFEE	F	20-24
11/28/2014	7:59:00	HOSP5	INGESTED TAINTED COFFEE	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.

Incorporating user feedback

- Our system enables continual improvement of performance by including public health practitioners in the loop and incorporating their feedback.
- Users can add new syndromes and specify if they would like the system to monitor or ignore them in the future.
- Blinded user studies show that this Practitioner in the Loop approach enables the system to report more **relevant** clusters and to avoid overwhelming the user with irrelevant findings.





Discussion

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.

How can machine learning improve urban public health?

Event & pattern detection

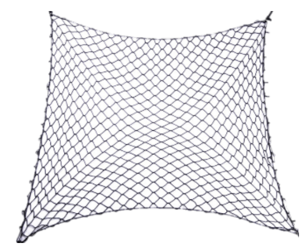
→ Better situational awareness

→ Earlier & more precisely targeted interventions

Interventions to combat the opioid overdose crisis

Many other ways ML can help, e.g.:

- **Causal inference** methods to assess impacts of environmental exposures, such as poor-quality housing, on health.
- **Algorithmic fairness** to allocate resources and reduce health disparities.



Providing a safety net for novel disease outbreaks and emerging public health threats

Identifying causal effects of environmental exposures

We are using Medicaid data linked to detailed building characteristics in order to identify impacts of poor-quality housing on chronic health.

"Which housing conditions impact which health conditions, for which subpopulations, to what extent?"

Key idea: treatment effects may be **heterogeneous**; use multidimensional scan to identify most affected subpopulations.

Must adjust for known confounders, selection into treatment (exposure).

Must account for multiple hypothesis testing to bound false positive rate.

Step 1: Predictive model at building level

X = 65 diagnoses x {adult, child}

Y = building on landlord watch list?

Adult asthma and COPD

Mental health (ADHD, adjust. disorder)

Injuries (children and adults)

Step 2: Heterogeneous treatment effect scan

"**Crowded housing** is associated with increased respiratory conditions & injuries among Asians living in Manhattan."

We have also developed an alternative scan-based approach to causal inference, based on automated discovery of natural experiments.

Thanks for listening!

More details on our web site:

http://wp.nyu.edu/ml4good

Or e-mail me at:

daniel.neill@nyu.edu