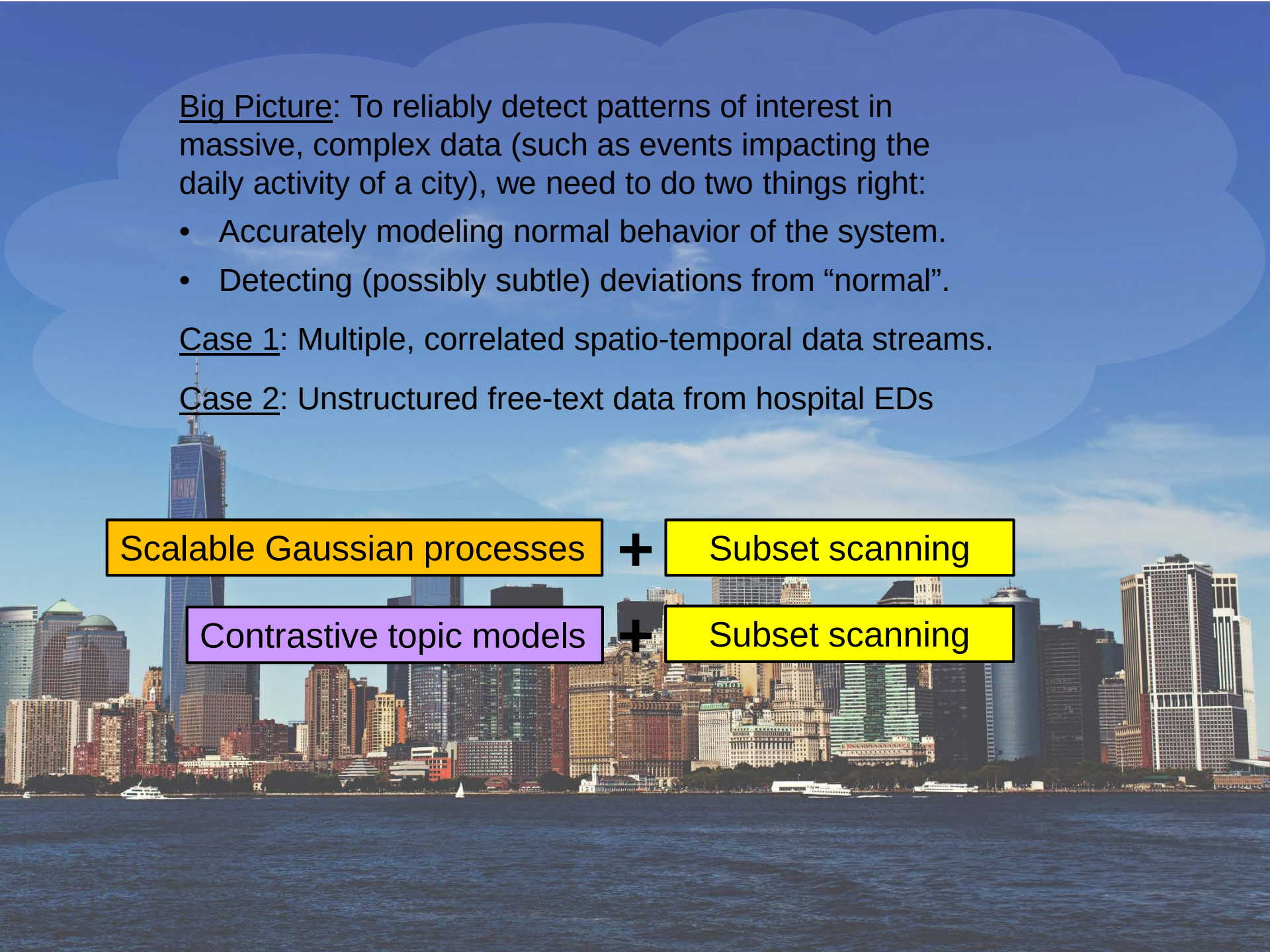
Modeling and Detecting Patterns in Complex Urban Data

Daniel B. Neill, Ph.D.
Center for Urban Science + Progress
New York University
E-mail: daniel.neill@nyu.edu

This talk is based on joint work with William Herlands and Mallory Nobles (Event and Pattern Detection Laboratory, Carnegie Mellon University).

We gratefully acknowledge funding support from NSF, RK Mellon Foundation, and NCSU Laboratory for Analytical Sciences.



Big Picture: To reliably detect patterns of interest in massive, complex data (such as events impacting the daily activity of a city), we need to do two things right:

- Accurately modeling normal behavior of the system.
- Detecting (possibly subtle) deviations from “normal”.

Case 1: Multiple, correlated spatio-temporal data streams.

Case 2: Unstructured free-text data from hospital EDs

Scalable Gaussian processes + Subset scanning

Contrastive topic models + Subset scanning

Today's talk:

- **Subset scanning** for event and pattern detection.
- **Gaussian processes** for modeling correlated data.
- Gaussian Process Subset Scan (GPSS)
- **Contrastive topic modeling** for identifying emerging patterns in free-text.
- Detecting emerging novel disease outbreaks.

Subset scanning

A machine learning* approach for **detecting subtle patterns** in complex data (spatial, graphs, multidimensional, …)

* Unsupervised learning = no labeled data, incl. clustering and anomaly detection as well as pattern detection. This is distinct from supervised learning, which focuses on classification and regression.

Key idea: A group of data records may be **collectively** interesting or anomalous even if each record is "normal" when considered individually.

Typical anomaly detection methods fail in this scenario. So do clustering methods, if the pattern of interest is small compared to the entire dataset.

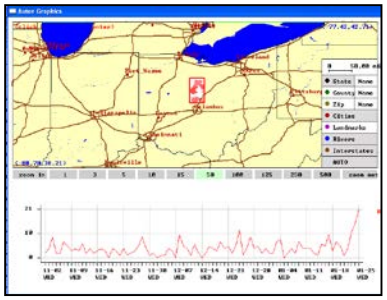
Very general framework:

1) Define a score function F(S) to evaluate interestingness of a subset.
2) Maximize F(S) over all subsets S which satisfy constraints C.
3) Evaluate statistical significance of detected subsets by randomization.

Challenge 1: Choose F(S) and C for high detection power and accuracy.

Challenge 2: Computational efficiency given exponentially large search space.

Subset scanning

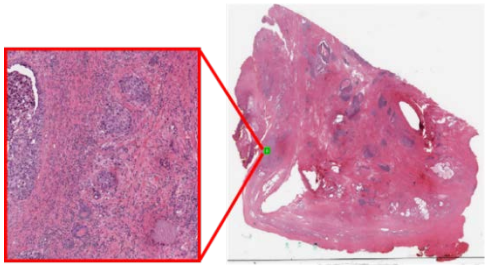


Early outbreak detection

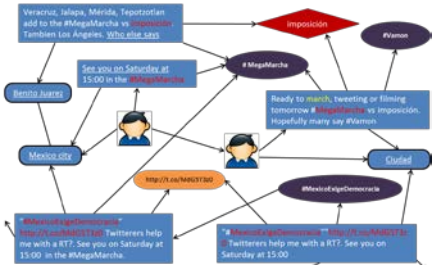
Drug overdose surveillance

Discovery of novel outbreaks

Rodent prevention



Prostate cancer detection



Predicting civil unrest

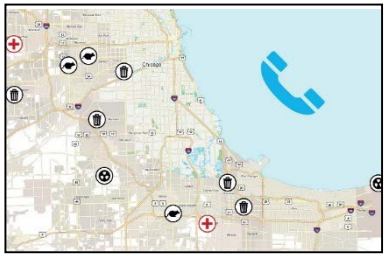
<u>Ongoing work (examples):</u>

Scaling to massive graphs

Detecting natural experiments

Discovering heterogeneous treatment effects

“Auditing” black box classifiers



Improved city services

Crime prevention- randomized field trial with Pittsburgh police

Gaussian processes

Gaussian processes are a machine learning* approach for **modeling** and **prediction** with non-iid data (e.g., spatial/temporal).

* Supervised learning = prediction from labeled data, including both classification and regression.

Advantages: Flexibility to accurately model complex non-linear functions given sufficient training data.

Can obtain **prediction intervals** rather than just point estimates, and can **learn** dependence structure from data.

Disadvantages: computationally expensive, $O(N^3)$.

A Gaussian process defines a prior over functions (infinite-dimensional vectors), parameterized by the mean function m(x) and kernel (covariance) function k(x,x').

Any draw of observed data from a GP, i.e., a function f(x) evaluated at a finite set of x values, has a **multivariate Gaussian** prior distribution given by the mean vector μ and covariance matrix K, where $\mu_i = m(x_i)$ and $K_{ij} = k(x_i, x_j)$.

Can then combine this prior distribution with observed data $\{(x_i, y_i)\}$, using Bayes Theorem to obtain the posterior mean & covariance.

Why GPs for modeling urban data?

Typical machine learning assumption: data points are drawn i.i.d. (independently, and identically distributed) from some distribution.

How reasonable is this assumption for urban systems?

Congestion may propagate from an initial traffic event, leading to spatial and temporal dependence.

Environmental monitoring: we expect similar sensor readings if the measurements are close together in space and time.

While dependent data can arise in many contexts (such as structured prediction or correlated data streams), the most common sources are dependence over time (serial autocorrelation) and spatial correlation.

Why GPs for modeling urban data?

Typical machine learning assumption: data points are drawn i.i.d. (independently, and identically distributed) from some distribution.

How reasonable is this assumption for urban systems?

Congestion may propagate from an initial traffic event, leading to spatial and temporal dependence.

Environmental monitoring: we expect similar sensor readings if the measurements are close together in space and time.

First law of geography: “Everything is related to everything else, but nearby things are more related than more distant things.”

Why GPs for modeling urban data?

Typical machine learning assumption: data points are drawn i.i.d. (independently, and identically distributed) from some distribution.

How reasonable is this assumption for urban systems?

Congestion may propagate from an initial traffic event, leading to spatial and temporal dependence.

Environmental monitoring: we expect similar sensor readings if the measurements are close together in space and time.

To accurately detect emerging events or discover anomalous patterns, we need a very accurate model of what is “normal”.

Otherwise, rather than detecting events of interest, we simply pick up where the model has failed to adequately capture normal behavior of the system.

Gaussian Process Regression

Given a set of M training examples, (X,Y) = {(x_i, y_i)} for i=1..M,
and a set of M* test examples, (X*,Y*) = {(x_i^*,y_i^*)} for i=1..M*.

Assume $y_i = f(x_i) + \varepsilon_i$, where ε_i are independent noise variables drawn from N(0, σ^2), and $f(\cdot) \sim \mathcal{GP}\big(0, k(\cdot,\cdot)\big)$, where $k(\cdot,\cdot)$ is the kernel (e.g., RBF).

We can now compute the conditional distribution of the unobserved outputs Y* given the observed outputs Y and all inputs (X and X*).

This can be computed in closed form by specifying the joint distribution: $[Y\ Y^*] \mid X, X^* \sim N(0, K + \sigma^2 I)$, where K can be decomposed as:

$$K = \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}.$$

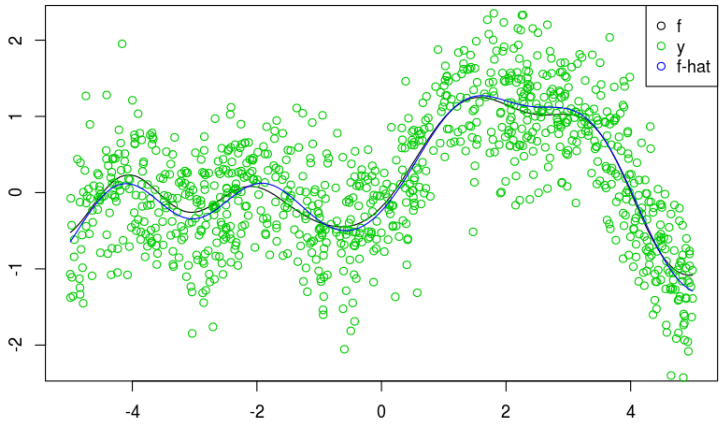
Using properties of multivariate Gaussians, we can compute the conditional distribution $Y^* \mid Y, X, X^* \sim N(\bar{\mu}, \bar{K})$, where:

$$\bar{\mu} = K(X^*,X)(K(X,X) + \sigma^2 I)^{-1} Y$$

$$\bar{K} = K(X^*,X^*) - K(X^*,X)(K(X,X) + \sigma^2 I)^{-1} K(X,X^*)$$

Acknowledgement: this slide was adapted from lecture notes by Chuong B. Do and Honglak Lee, and from Flaxman et al., Gaussian processes for independence tests with non-iid data in causal inference, *ACM TIST*, 2015.

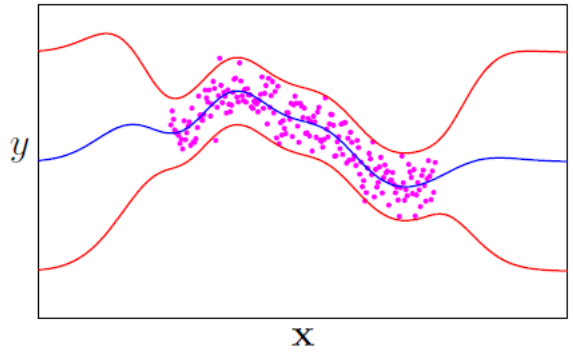
Gaussian Process Regression



(figure by Seth Flaxman)

Here's an example of fitting the true function f(x) with the mean function $\bar{\mu}$ estimated from the GP.

With lots of data, GPs can very accurately estimate a complex, non-linear function of x.



(figure by Zoubin Ghahramani)

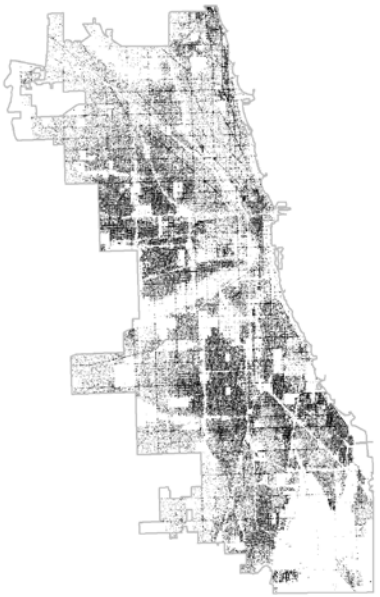
We can also sample from the posterior predictive distribution $N(\bar{\mu}, \bar{K})$, and use the samples to compute a 95% confidence interval for prediction.

Computation is exact but expensive: $O(N^3)$ training, $O(N^2)$ test.
But various approximations can be used to speed things up.

Scaling up Gaussian processes

(Flaxman et al., ICML 2015)

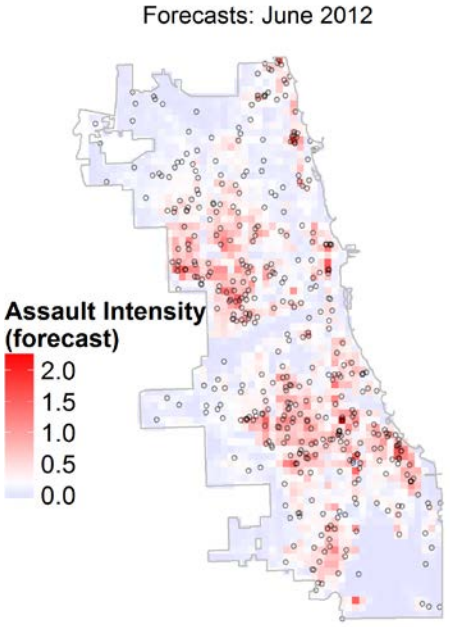
Application: long-term, local-area **crime forecasting** in Chicago. We applied a new, scalable GP model to 10 years of geocoded, date-stamped crime reports.



New methods for very fast inference and learning with non-Gaussian likelihoods (count data) and interpretable (spectral mixture) kernels.

⇒

Approximations plus exploiting grid structure for fast matrix operations.



n = 233,088 reported incidents of assault.

After discretization (½ mile x ½ mile x week): 1.6 million observations in total, much too large for standard GP formulations.

Very accurate, small-area crime forecasts up to 12 months in advance: scalability and accuracy higher than previous state of the art.

Scaling up Gaussian processes

(Flaxman et al., ICML 2015)

Application: long-term, local-area **crime forecasting** in Chicago. We applied a new, scalable GP model to 10 years of geocoded, date-stamped crime reports.

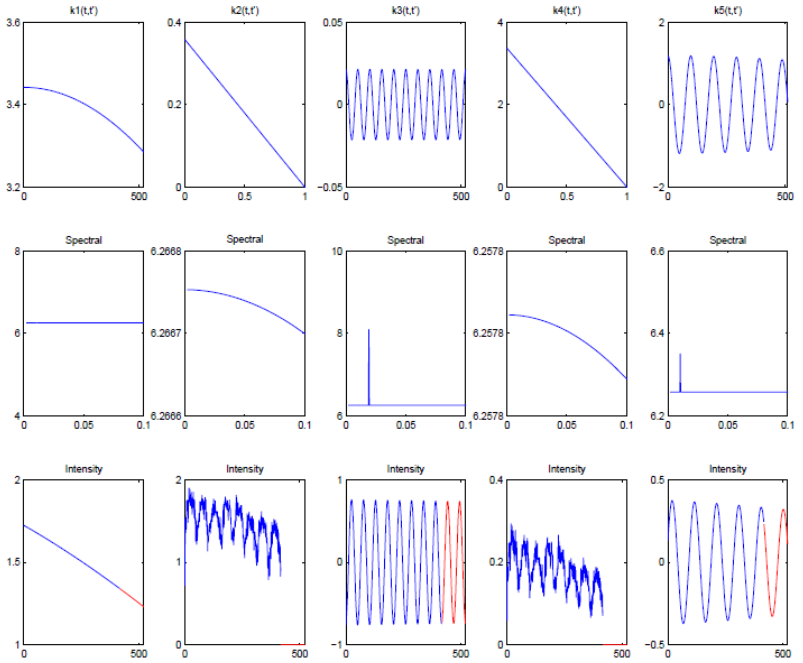


Figure 4. The five spectral mixture components with highest weights learned by our model are shown as a covariance (top) and spectral density (middle). In the bottom row, time series predictions were made on the dataset (ignoring space) using only that component. Red indicates out-of-sample forecasts.

Learning an **interpretable kernel** tells us a lot about the correlation structure in space and time.

Component 1 picks up long-term trend of decreasing crime.

Component 3 picks up yearly periodic (seasonal) trend.

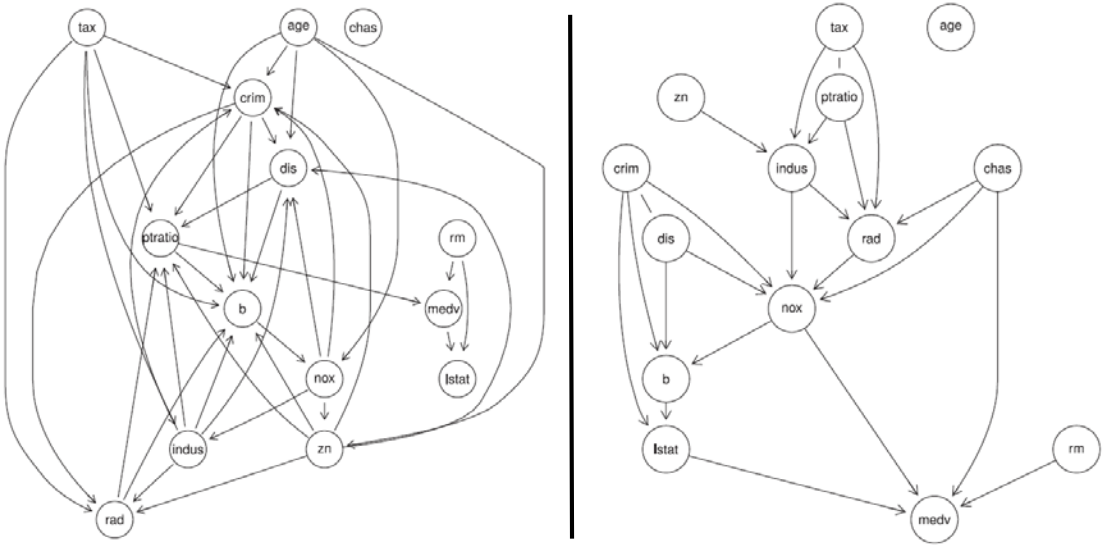
Components 2 and 4 represent correlation on short time-scales (no effect on long-term forecasts).

Component 5 picks up another, more subtle periodic trend.

Application to causal inference

(Flaxman et al., *ACM TIST*, 2015)

Causal inference in non-iid data, using pre-whitening with GPs to remove non-causal relationships resulting from spatial and temporal dependencies.



Causal inference on Boston housing data.

Left: result of PC algorithm.

Right: PC after using our GP inference method.

Note many fewer edges than the previous graph.

Causes of median house value (medv):
Percent of lower SES in the population (lstat), number of rooms (rm), whether located on the Charles River (chas), and pollution as measured by nitric oxide concentration (nox).

Also note the edge from industrial activity (indus) to pollution (nox), while the previous graph had this edge reversed.

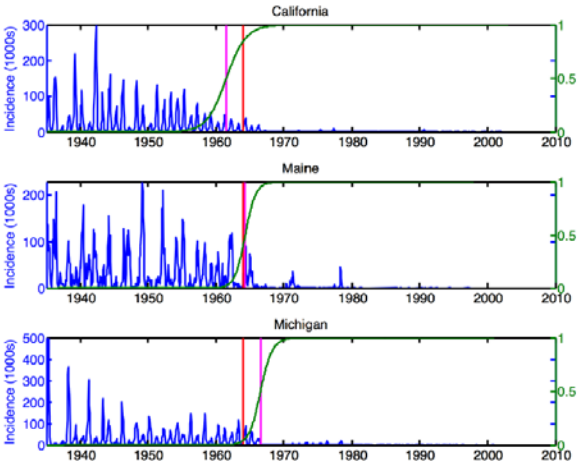
GPs for change point detection

(Herlands et al., AISTATS 2016)

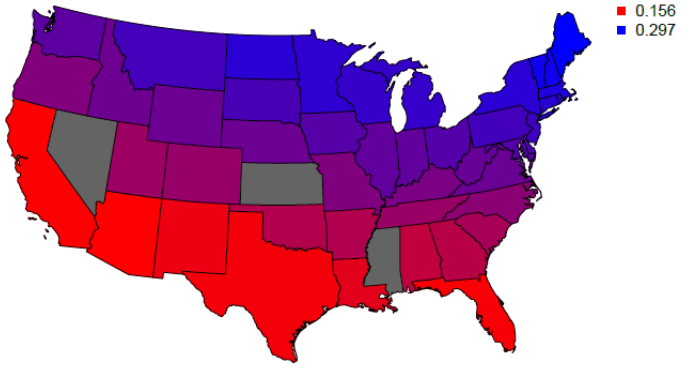
We developed new, scalable GP methods to detect **multidimensional**, **gradual**, and **heterogeneous** changes in the data distribution ("change surface detection").

We used our approach to model state-level, monthly measles incidence data from 1935 to 2003 (~33K data points in 3 dimensions: long, lat, time).

As expected, we identified a significant change from the introduction of the measles vaccine in 1963. But the impacts of the vaccine were shown to be gradual and heterogeneous: different states had different change points and rates of change from pre-vaccine to post-vaccine.



Change points: mid-1961 to early 1967.



Rate of change by state varied by 2x (Maine fastest, Arizona slowest).

Gaussian Process Subset Scan

(Herlands et al., AISTATS 2018)

Goals:
a) Model “normal” behavior of urban systems (correlated → learn GP)
b) Identify localized deviations from normal, enabling targeted response.

General approach:
1) Learn GP parameters from the entire dataset D.
2) Scan over neighborhoods:
   - For each neighborhood S, perform GP inference given the data in $D \setminus S$ to estimate the mean vector μ and covariance matrix Σ for locations $s_i \in S$.
3) Maximize a likelihood ratio statistic over neighborhoods (GP Neighborhood Scan) or over subsets of locations within a neighborhood (GP Subset Scan) to find most anomalous areas.

LLR comparison

Covariates $x_i \in R^D$
e.g. space & time

Outputs
$y_i \in R^1$

Subset
$w_i \in \{0,1\}$

$$LLR = \frac{H_1(x, y, w)}{H_0(x, y)}$$

Alternative model: anomaly for locations with $\mathbf{w_i = 1}$ in (**x**, **y**)

Null model: no anomalies for all locations in (**x**, **y**)

LLR comparison

$$f(x) \sim GP$$

Parameters learned from all data

Null model

$$y = f(x) + \varepsilon \longrightarrow y \sim N(\mu, \Sigma)$$

Alternative model

$$w_i = 0 : \; y = f(x) + \varepsilon$$

$$w_i = 1 : \; y = f(x) + \beta + \varepsilon$$

$$\longrightarrow \; y \sim N(\mu + \beta w, \Sigma)$$

LLR comparison

- Log-likelihood ratio of alternative to null model

- Maximization of $LLR(w)$ provides most anomalous subset of the data

$$\arg\max_w LLR(w) = \arg\max_w -\frac{\beta^2}{2} w^T E w + \beta w^T E(y - \mu)$$

- Integer Quadratic Program: optimizing $LLR(w)$ over subsets requires $O(2^n)$ comparisons.

Gaussian Process Neighborhood Scan (GPNS)

- Consider the k-neighborhood consisting of a point and its (k – 1)-nearest neighbors, for each center point and each $k = \{1, 2, \ldots, k_{max}\}$.
- Compute LLR for each neighborhood.
  - Highest LLR = most anomalous neighborhood

Pros	Cons
- Substantially reduced computation, $O(nk_{max})$. - Considers collective anomalies consisting of multiple nearby points.	- Constraining assumptions about anomaly shape! - Reduces precision and detection power.

Gaussian Process Subset Scan (GPSS)

- Fix neighborhood size, k.
- Conduct an unconstrained search within each neighborhood of size k.
  - Naively requires $O(n2^k)$ evaluations.
  - Reduce to O(nk) evaluations, using one of several new (approximate) optimization methods compared in the paper.

Technical details can be found in Herlands et al. (AISTATS 2018).

Randomization testing

1) Create multiple simulated datasets $y^{(r)} \sim GP$ assuming that the null hypothesis is true.

2) Scan over $(x, y^{(r)})$

   Find max LLR for each simulated dataset.

3) After R trials, determine α-level significance using the null max-LLR values.

Multiple Streams

- Searching for patterns over multiple data sets or data streams
  - E.g. reports of downed trees and sewer flooding

General approach

1) Learn GP parameters θ_s for each stream **independently**
2) Scan over neighborhoods **jointly**
   - Compute posteriors (μ_s, Σ_s) independently
   - Combine them with block diagonal matrix

$$\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_S \end{bmatrix}, \begin{bmatrix} \Sigma_1 & 0 & 0 & 0 \\ 0 & \Sigma_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \Sigma_S \end{bmatrix} \right)$$

Synthetic Experiments

- Compared methods w.r.t. accuracy and detection power, on synthetic data with known anomalies.
- Main conclusions:

1) GP+SS > individual anomaly detection (SVM, GP)

→ Nearby points matter for subtle anomalies

2) GP+SS > SS alone

→ Covariance structure matters for non-iid data

3) GPSS > GPNS

→ Flexible detection matters for complex patterns

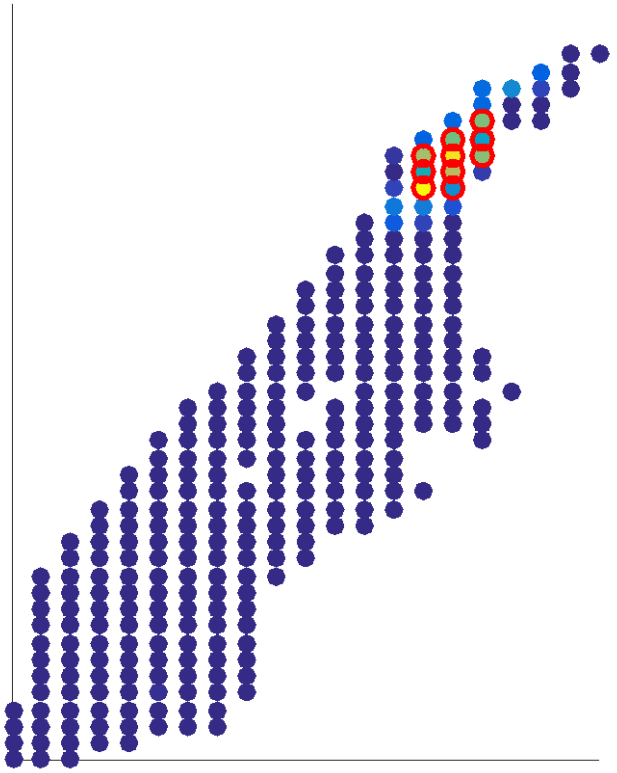
Real-world experiments

- Using NYC data: opioid overdoses and 311 calls.
- Challenging to get ground truth
  - Even with known anomaly – precisely where and when is unknown.
  - Manual validation of anomalies.

Evaluation: 311 calls for service

- 1/22/2016: Washington Heights residents concerned due to brown tap water.
- Data: 311 calls
  - Aggregate all 311 types

  (call type for evaluation only).
  - Daily for January 2016
  - Spatial grid of ~ 0.08 mi^2



Evaluation: 311 calls for service

- No ground truth for evaluation!
- Consider the “signal-to-noise ratio” within the detected region.
  - Signal = 311 water-related complaints
  - Noise = all other 311 complaints

Model	Signal-to-Noise
GRQ	7.22
Stepwise	7.22
β_{MAX}	7.22
Independent SS	7.06
Baseline GP	0.44
One-class SVM	0.23
RobustCov	0.12

GP and SS methods

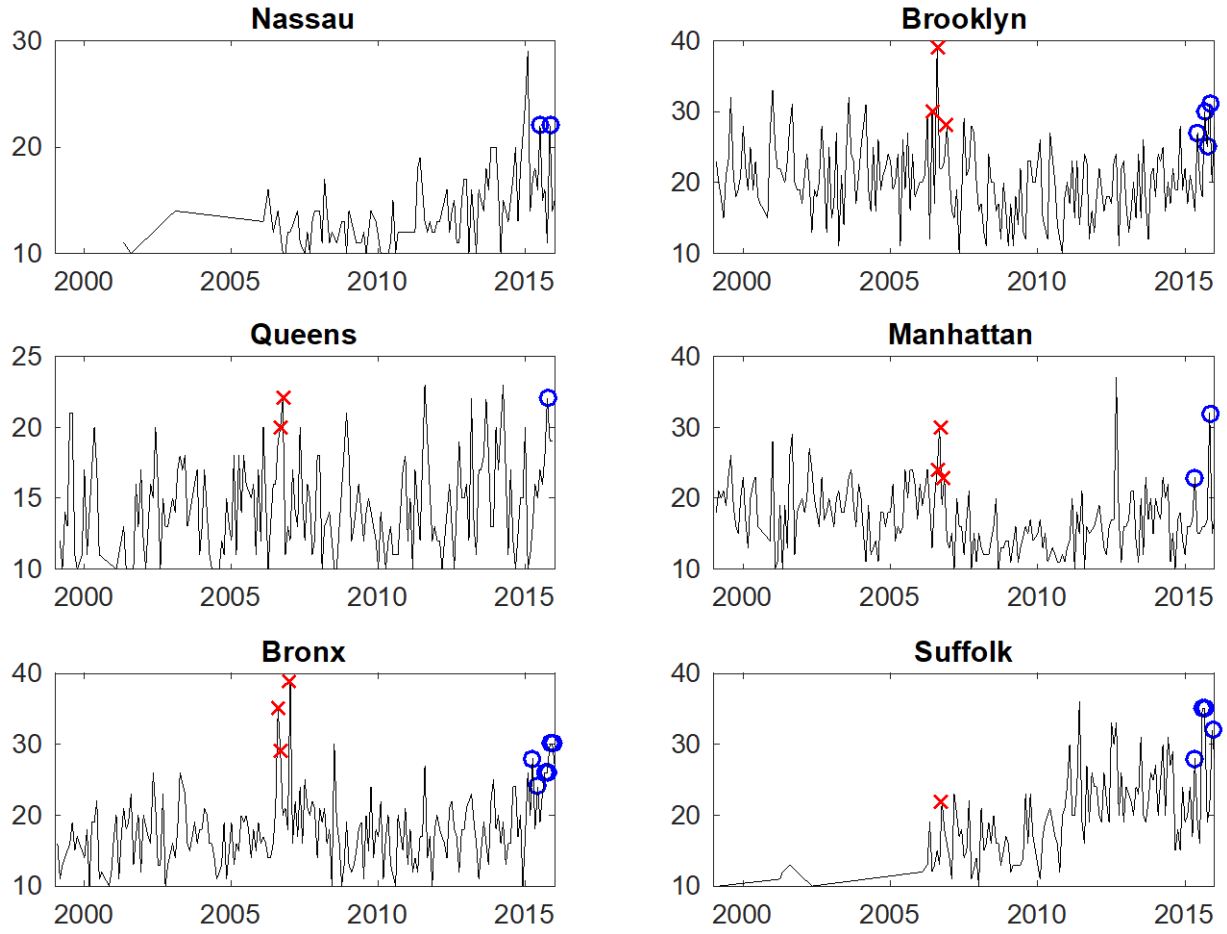
Anomaly detection methods

Evaluation: opioid overdoses

- US opioid epidemic has affected NYC area
- Data: opioid overdose deaths
  - Monthly counts, 1999-2015
  - Counties: Man, Qns, Bky, Bnx, Suffolk, Nassau

Evaluation: opioid overdoses

- Two statistically significant anomalies.



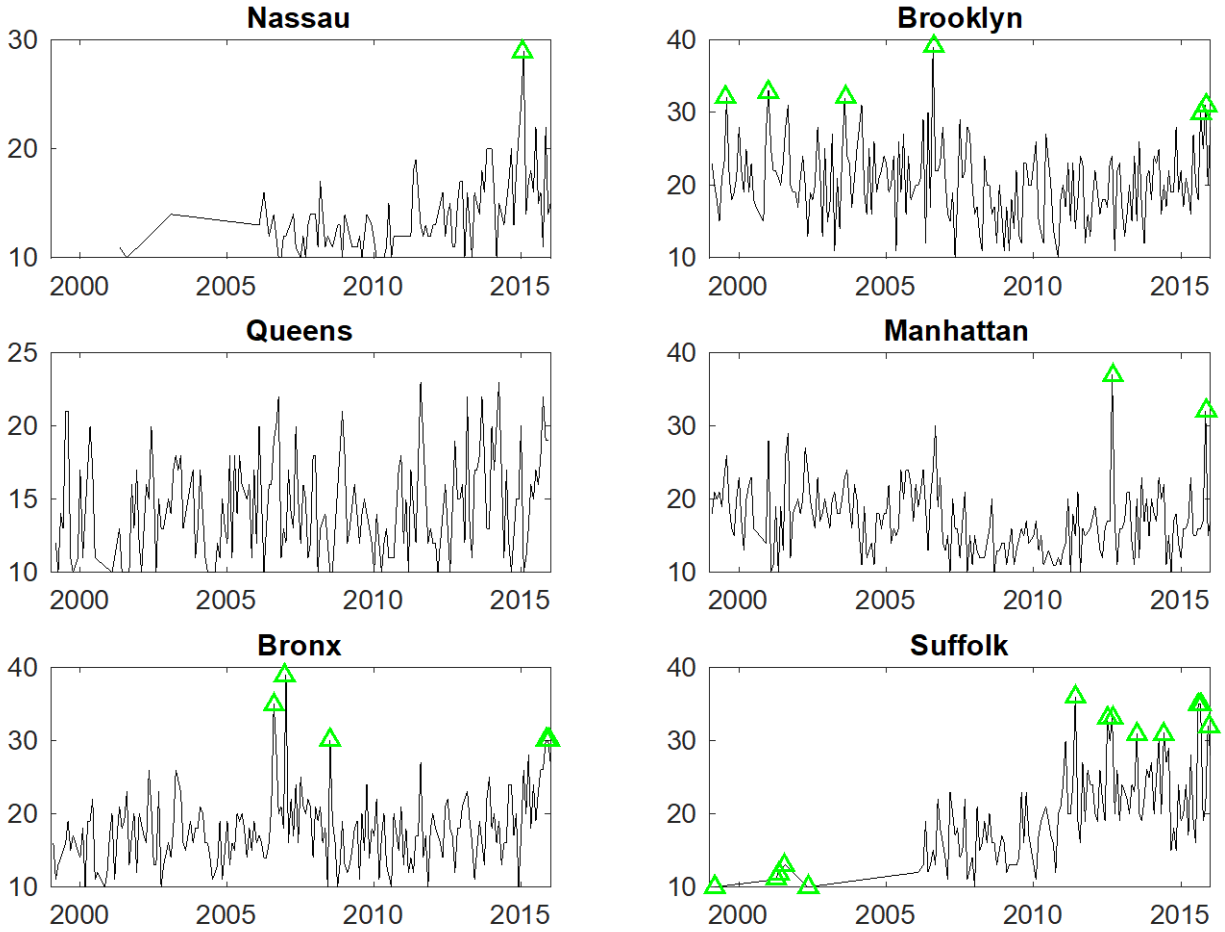
✖ Mid 2006. Just before naloxone programs.

⭕ End of 2015. Recent surge due to fentanyl.

Recall we jointly search all space-time.

Opioid overdose deaths

- SVM provides no coherent anomalies

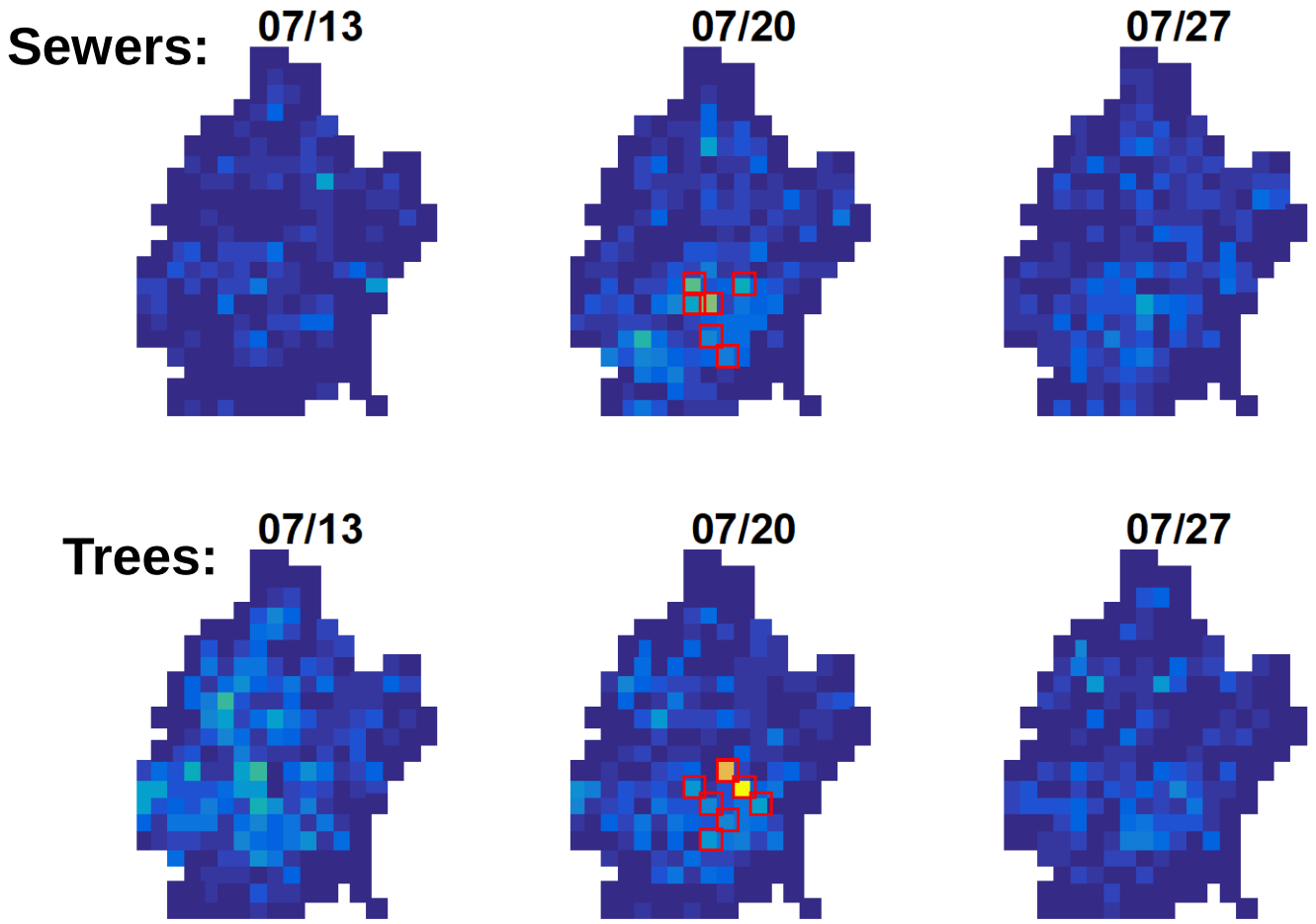


Multi-stream: trees and sewers

- Consider 311 reports of damaged tree and sewer issues.
  - Fit each stream independently
  - Scanned for anomalies jointly

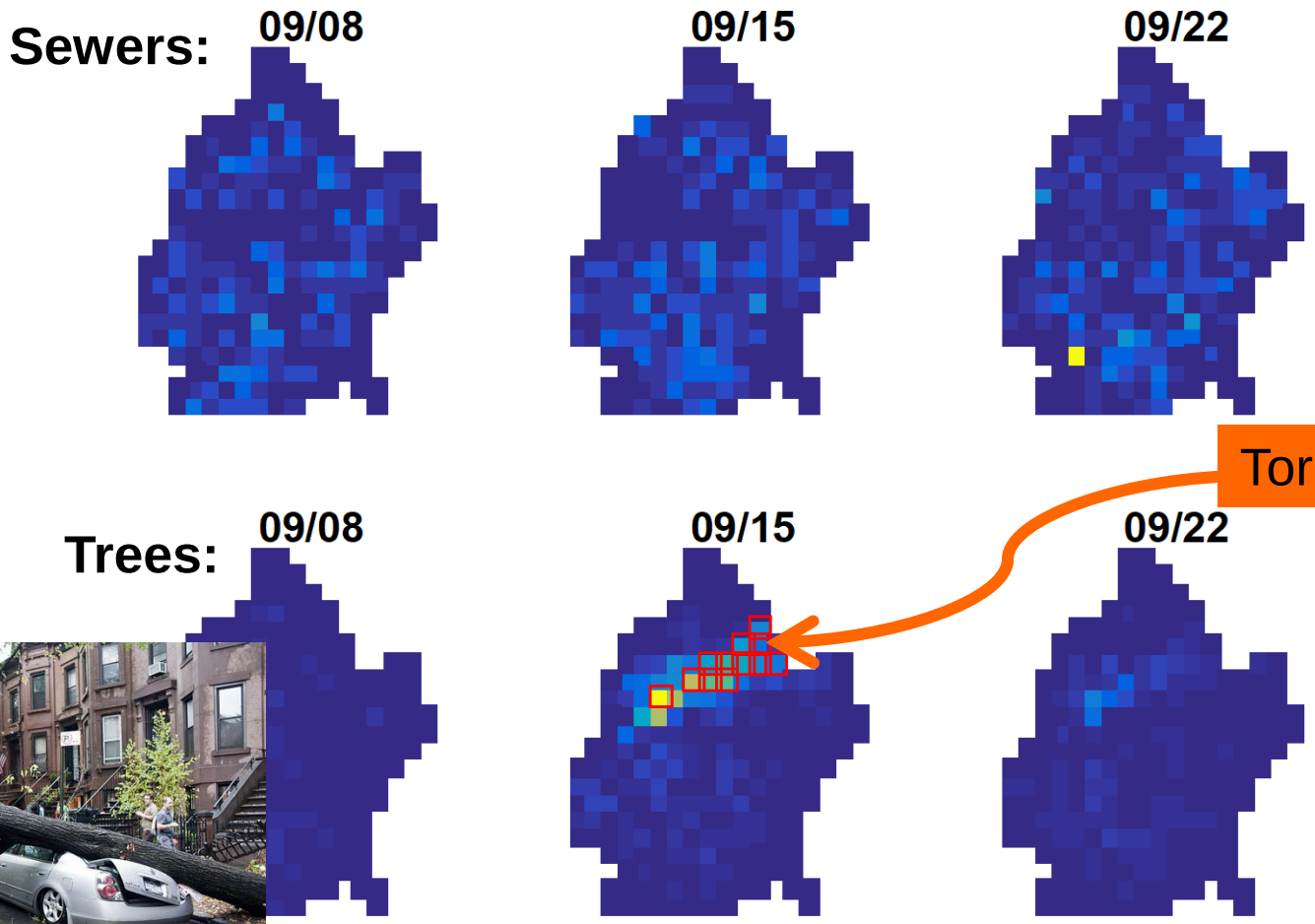
- Aggregated at weekly level for Brooklyn
  - Analysis for 2016 and 2010

Multi-stream: trees and sewers 2016



- Anomalies during July summer storm

Multi-stream: trees and sewers 2010



- Urban tornado in Brooklyn!

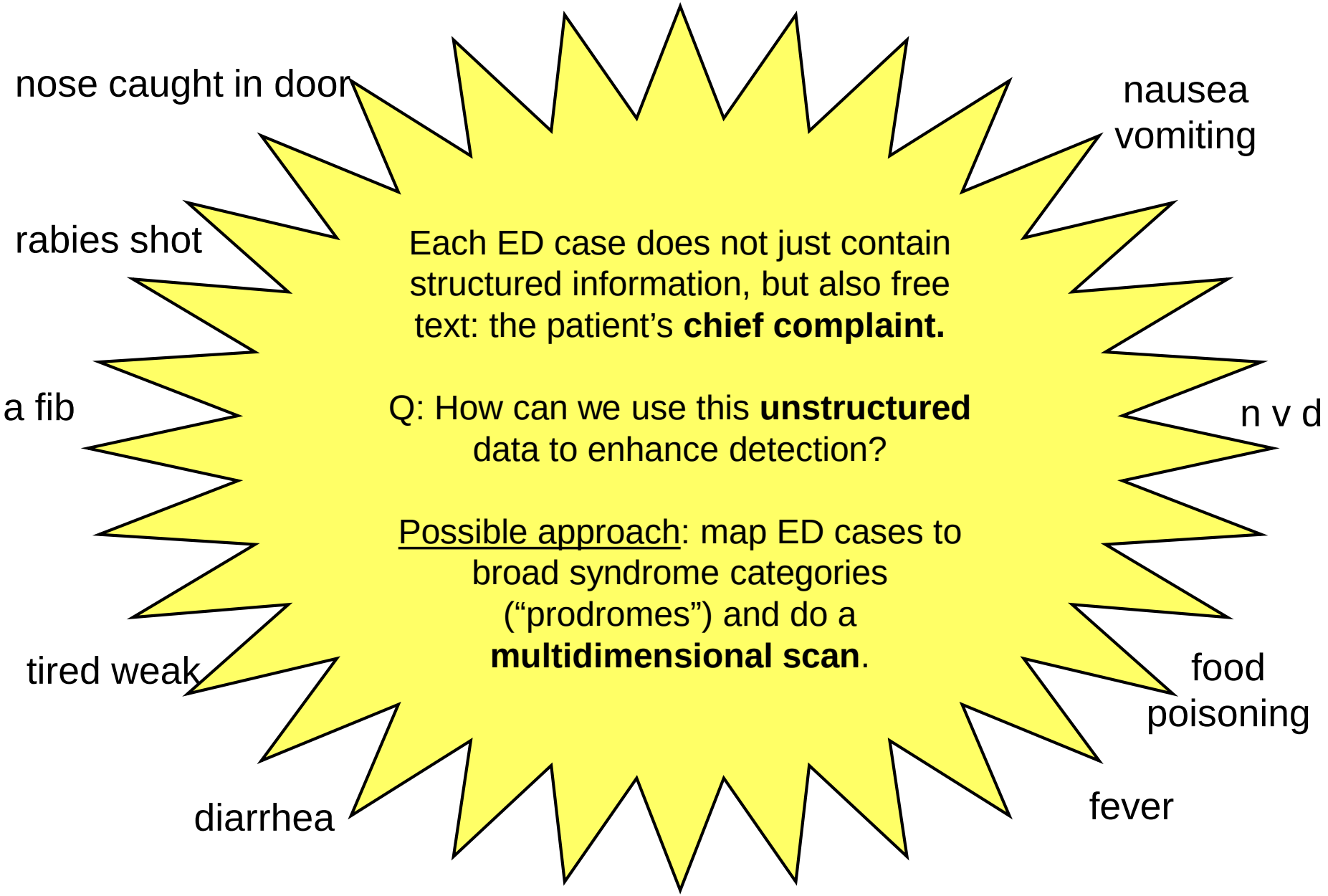
Incorporating unstructured data

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

Free-text ED chief complaint data from hospitals in NYC and North Carolina.

Key challenge: public health agencies must be able to identify relevant clusters of disease cases that may not correspond to known syndromes (e.g., rare or novel outbreaks)

From structured to unstructured…



Where do existing methods fail?

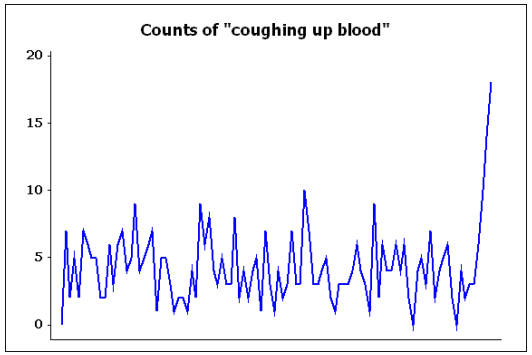
The typical syndromic surveillance approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

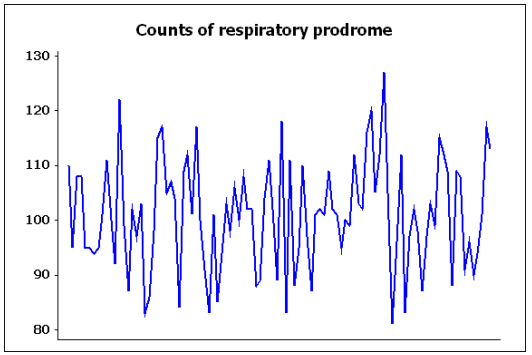
What happens when something new and scary comes along?

- **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.





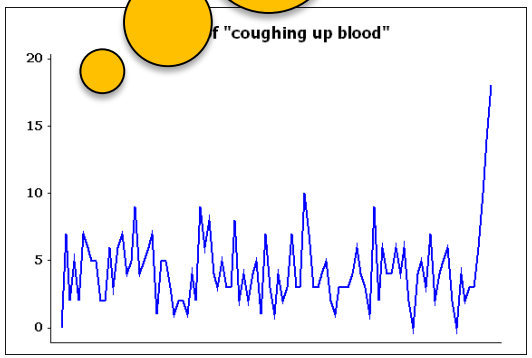
Where do existing methods fail?

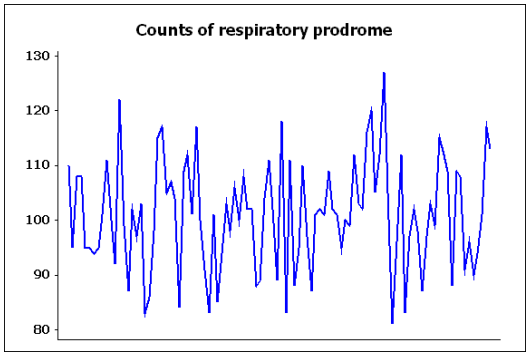
The typical sy… surveillance… effectively … outbrea… seen… sy…

…en something …s 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**.

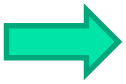
If we w… …ief complaints particular sy… …ptom category take a few such … …nute the outbreak signal, that an outb… …s occurring! …ying or preventing detection.

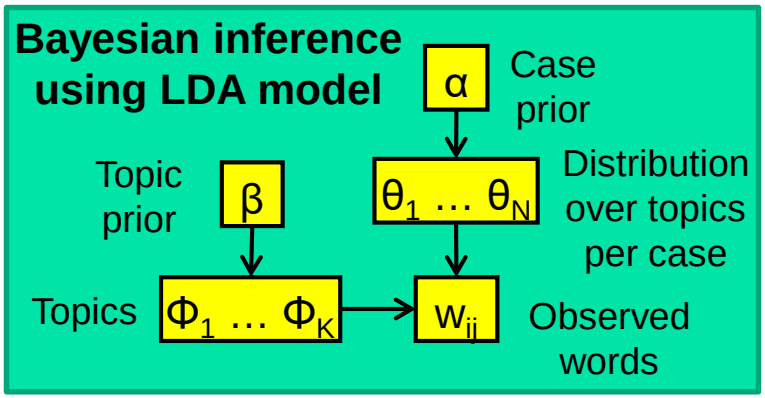




The semantic scan statistic

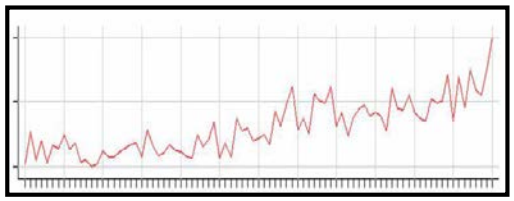
Date/time	Hosp.	Age	Complaint
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp





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 scanning

For each hour of data:

For each combination S of:

- Hospital
- Time duration
- Age range
- Topic

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

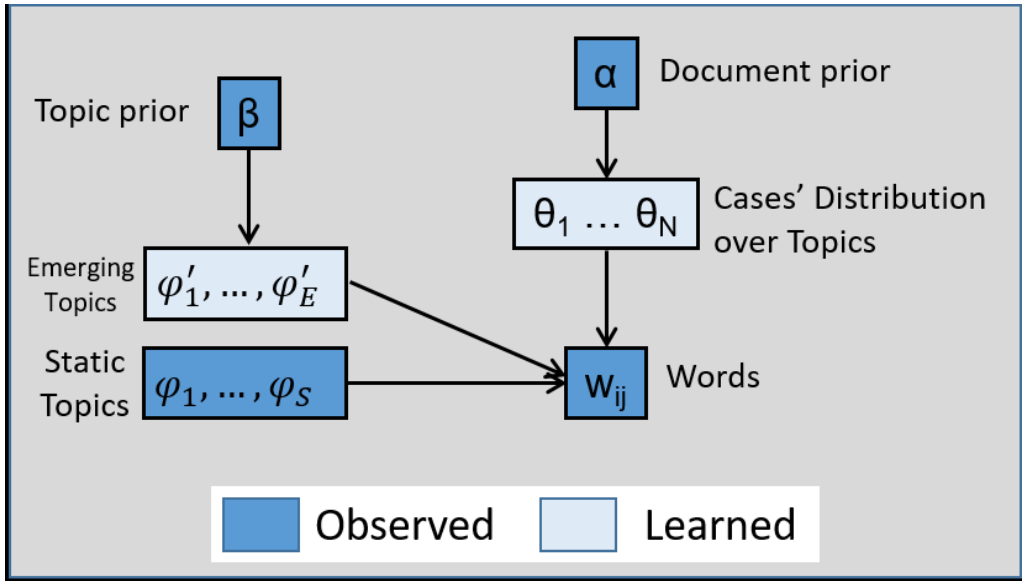
Baseline: B(S) = expected count (28-day moving average).

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.

Detecting emerging topics

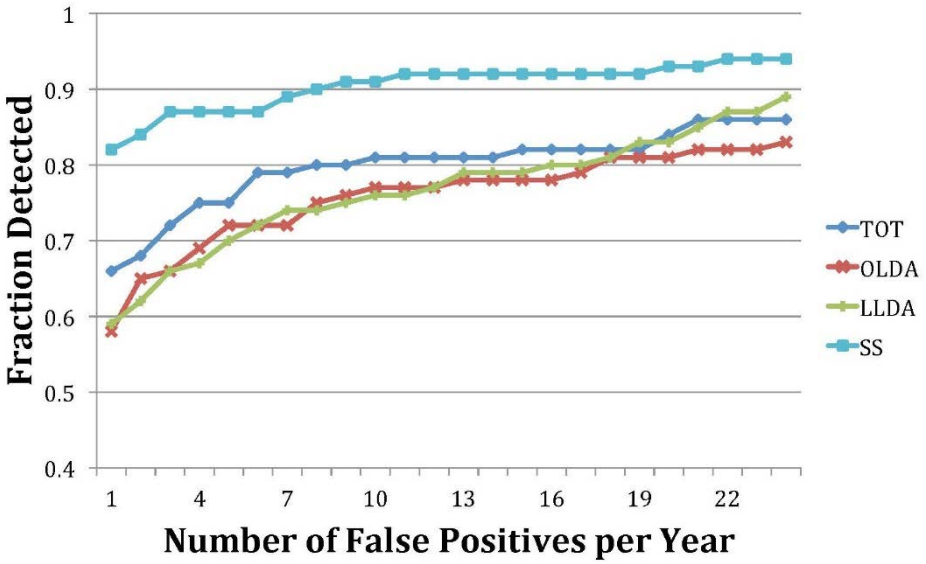
Our **contrastive topic model** is a novel extension of the LDA approach, designed to identify **newly emerging topics**.



1) Learning a set of "background" topics from historical data.
2) Learning a set of "foreground" topics from recent data.
3) Combined LDA inference, holding the background topics constant, leads to discovery of foreground topics that are maximally different.

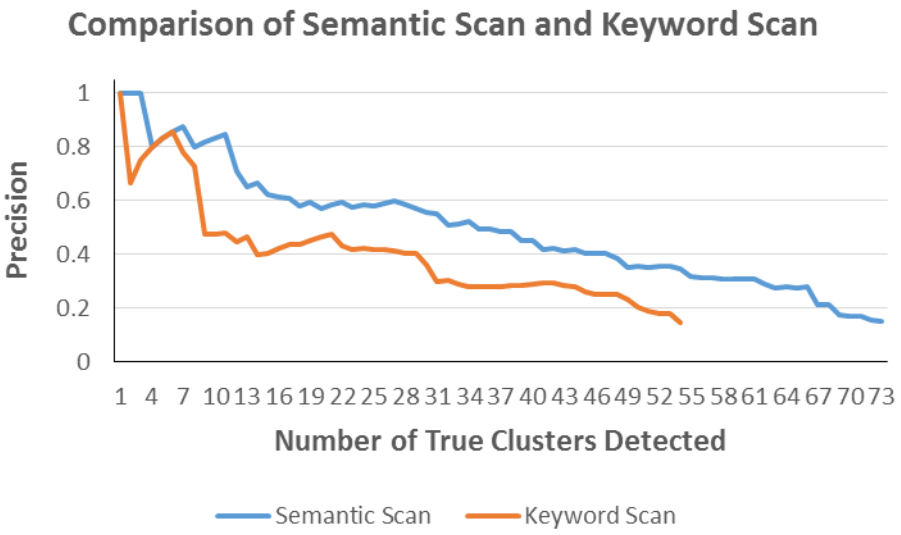
Detecting emerging topics

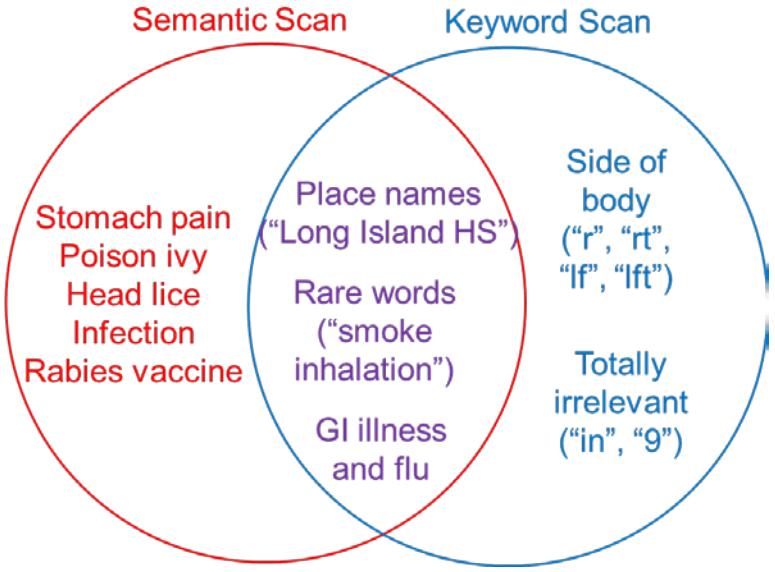
Using a “leave one out” approach in which we hold out one International Classification of Diseases (ICD) code and inject cases as if from a novel outbreak, we observe huge improvements in detection power and accuracy vs. competing methods, including Online LDA, Topics Over Time, and Labeled LDA.



NC DOH evaluation results

We compared the top 500 clusters found by semantic scan and a keyword-based scan on data provided by the NC DOH in a blinded evaluation, with DOH labeling each cluster as “relevant” or “not relevant”.





Semantic scan: for 10 true clusters, had to report 12;
for 30 true clusters, had to report 54.

Keyword scan: for 10 true clusters, had to report 21;
for 30 true clusters, had to report 83.

NYC DOHMH dataset

- New York City's Department of Health and Mental Hygiene provided us with 5 years of data (2010-2014) consisting of ~20M chief complaint cases from 50 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 size and messiness of data (typos, abbreviations, etc.).
  - 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.

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.

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
Motor vehicle
Ferry
School bus
Elevator

Contagious Diseases
Meningitis
Scabies
Ringworm

Other
Drug overdoses
Smoke inhalation
Carbon monoxide poisoning
Crime related, e.g., pepper spray attacks

Conclusions

Detecting emerging events in complex data requires us to accurately model “normal” behavior and to detect subtle deviations from “normal”.

Challenges include non-iid data, subtle signals, and unstructured text.

Three building blocks for modeling and detection:

Scalable Gaussian processes can model complex correlations across space, time, and multiple data streams.

(Multidimensional) subset scanning can accurately detect subsets of data elements that deviate subtly from the expected distribution.

Contrastive topic modeling can detect newly emerging patterns of keywords; **semantic scan** identifies affected areas & subpopulations.

Together, these tools allow us to detect emerging patterns of interest, integrating a wide variety of structured and unstructured urban data sources to detect emerging outbreaks and other relevant events.

References

- D.B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.
- S.R. Flaxman, A.G. Wilson, D.B. Neill, H. Nickisch, and A.J. Smola. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. *Proc. ICML,* 2015.
- S.R. Flaxman, D.B. Neill, and A.J. Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM TIST*, 2015.
- W. Herlands, A.G. Wilson, H. Nickisch, S. Flaxman, D.B. Neill, W. van Panhuis, and E.P. Xing. Scalable Gaussian processes for characterizing multidimensional change surfaces. *Proc. AISTATS*, 2016.
- W. Herlands, A.G. Wilson, E. McFowland III, and D.B. Neill. Gaussian process subset scanning for anomalous pattern detection in non-iid data. *Proc. AISTATS*, 2018.
- A. Maurya, K. Murray, C. Dyer, Y. Liu, and D.B. Neill. A semantic scan statistic for novel disease outbreak detection. Submitted for publication. Winner of the Yelp Dataset Challenge.
- D.B. Neill and M. Nobles. Pre-syndromic surveillance. Working paper, 2018. Runner-up in Department of Homeland Security's Hidden Signals Challenge.

Thanks for listening!

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

Or e-mail me at:
daniel.neill@nyu.edu