

# Subset Scanning for Event and Pattern Detection

Daniel B. Neill, Ph.D.

Machine Learning for Good Laboratory  
New York University

E-mail: [daniel.neill@nyu.edu](mailto:daniel.neill@nyu.edu)

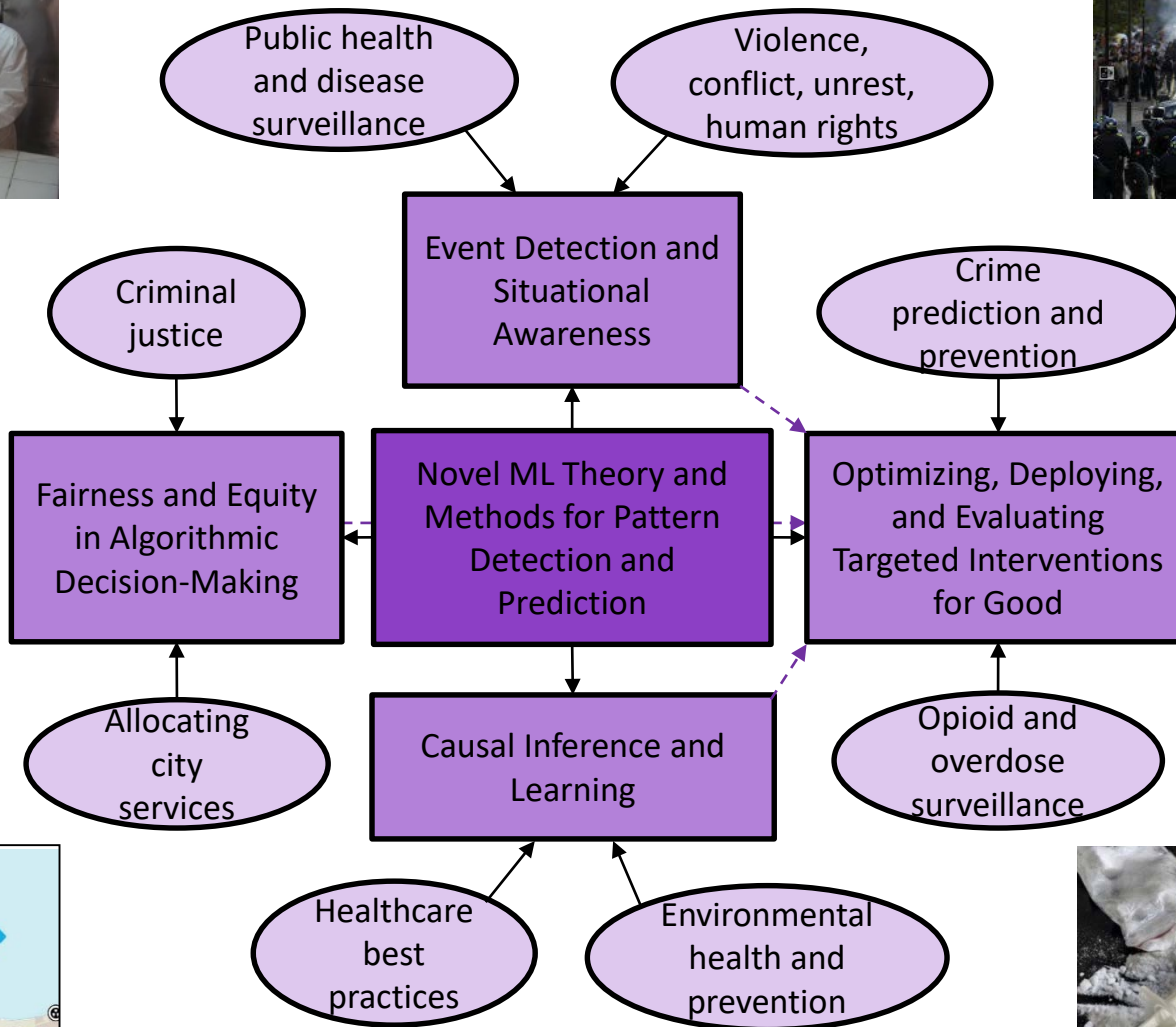
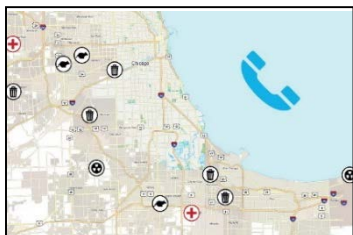
Web: <http://www.cs.nyu.edu/~neill>  
<http://wp.nyu.edu/ml4good>



**NYU**

Center for Urban  
Science + Progress

# The Machine Learning for Good Lab @ NYU



# Pattern detection by subset scan

One key insight that underlies much of my work is that pattern detection can be viewed as a **search** over subsets of the data.

## Statistical challenges:

Which subsets to search?  
Is a given subset anomalous?  
Which anomalies are relevant?

## Computational challenge:

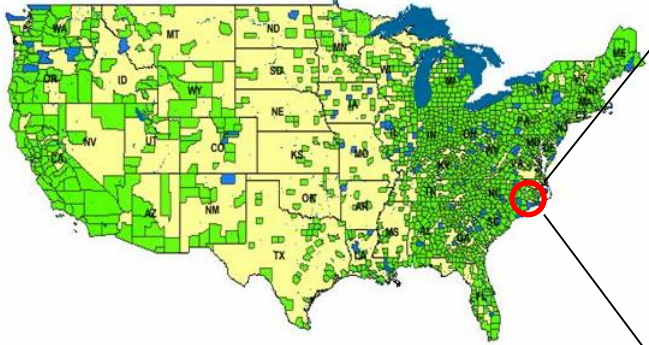
How to make this search over subsets efficient for massive, complex, high-dimensional data?

New statistical methods enable more timely and more accurate detection by integrating **multiple data sources**, incorporating **spatial** and **temporal** information, and using **prior knowledge** of a domain.

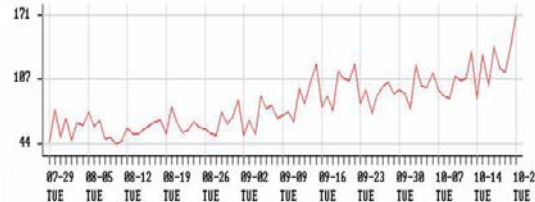
New algorithms and data structures make previously impossible detection tasks computationally feasible and fast.

New machine learning methods enable our systems to learn from user feedback, modeling and distinguishing between relevant and irrelevant types of anomaly.

# Multivariate event detection



Spatial time series data from spatial locations  $s_i$  (e.g. zip codes)



Time series of counts  $c_{i,m}^t$  for each zip code  $s_i$  for each data stream  $d_m$ .

## Outbreak detection

- $d_1$  = respiratory ED
- $d_2$  = constitutional ED
- $d_3$  = OTC cough/cold
- $d_4$  = OTC anti-fever  
(etc.)

## Main goals:

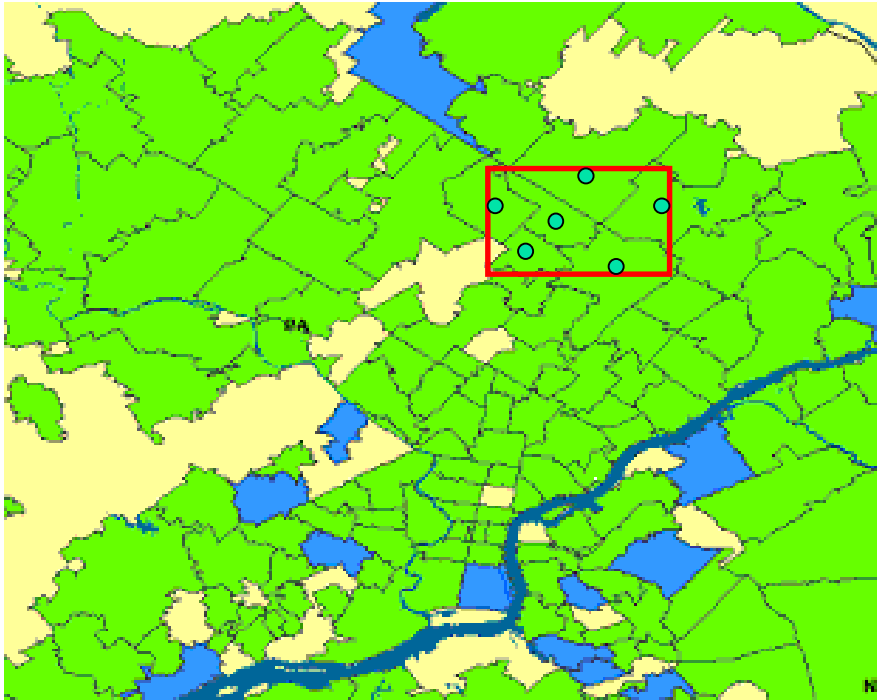
- Detect** any emerging events.
- Pinpoint** the affected subset of locations and time duration.
- Characterize** the event, e.g., by identifying the affected streams.

## Compare hypotheses:

- $H_1(D, S, W)$
- $D$  = subset of streams
- $S$  = subset of locations
- $W$  = time duration
- vs.  $H_0$ : no events occurring

# Expectation-based scan statistics

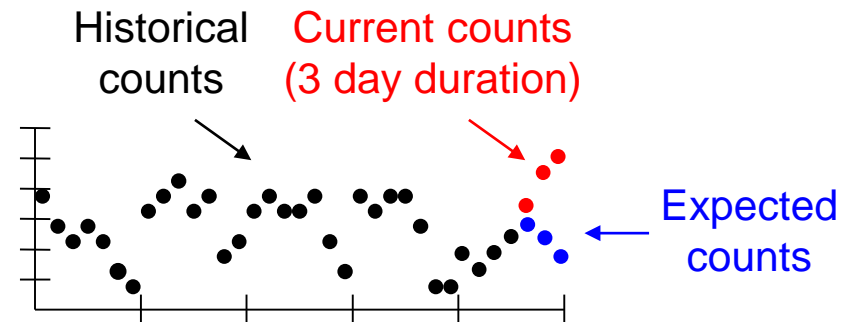
(Kulldorff, 1997; Neill and Moore, 2005)



We search for spatial regions (subsets of locations) where the recently observed counts for some subset of streams are significantly higher than expected.

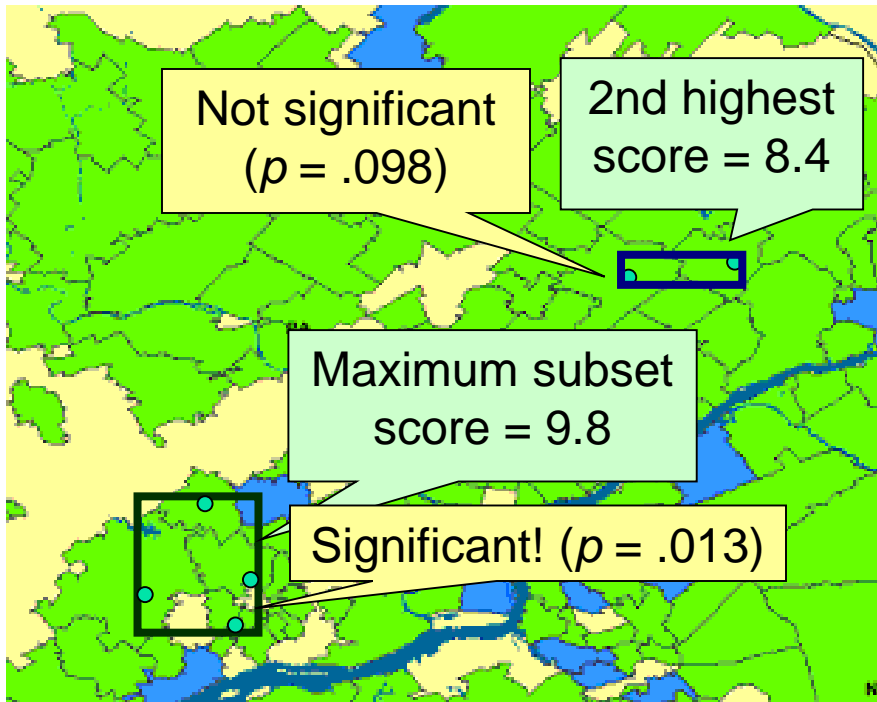
We perform **time series analysis** to compute expected counts (“baselines”) for each location and stream for each recent day.

We then compare the actual and expected counts for each subset (D, S, W) under consideration.



# Expectation-based scan statistics

(Kulldorff, 1997; Neill and Moore, 2005)

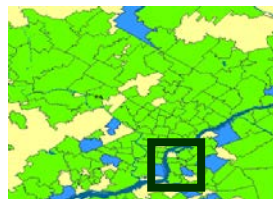


We find the subsets with highest values of a **likelihood ratio statistic**, and compute the  $p$ -value of each subset by randomization testing.

$$F(D, S, W) = \frac{\Pr(\text{Data} \mid H_1(D, S, W))}{\Pr(\text{Data} \mid H_0)}$$

To compute p-value  
Compare subset score to maximum subset scores of simulated datasets under  $H_0$ .

$F_1^* = 2.4$



$F_2^* = 9.1$



...

$F_{999}^* = 7.0$



# Likelihood ratio statistics

For our expectation-based scan statistics, the null hypothesis  $H_0$  assumes “business as usual”: each count  $c_{i,m}^t$  is drawn from some parametric distribution with mean  $b_{i,m}^t$ .  $H_1(S)$  assumes a multiplicative increase for the affected subset  $S$ .

## Expectation-based Poisson

$$H_0: c_{i,m}^t \sim \text{Poisson}(b_{i,m}^t)$$

$$H_1(S): c_{i,m}^t \sim \text{Poisson}(qb_{i,m}^t)$$

$$\text{Let } C = \sum_S c_{i,m}^t \text{ and } B = \sum_S b_{i,m}^t.$$

$$\text{Maximum likelihood: } q = C / B.$$

$$F(S) = C \log (C/B) + B - C$$

## Expectation-based Gaussian

$$H_0: c_{i,m}^t \sim \text{Gaussian}(b_{i,m}^t, \sigma_{i,m}^t)$$

$$H_1(S): c_{i,m}^t \sim \text{Gaussian}(qb_{i,m}^t, \sigma_{i,m}^t)$$

$$\text{Let } C' = \sum_S c_{i,m}^t b_{i,m}^t / (\sigma_{i,m}^t)^2 \\ \text{and } B' = \sum_S (b_{i,m}^t)^2 / (\sigma_{i,m}^t)^2.$$

$$\text{Maximum likelihood: } q = C' / B'.$$

$$F(S) = (C')^2 / 2B' + B'/2 - C'$$

Many possibilities: exponential family, nonparametric, Bayesian...

# Which regions to search?

Typical approach: “spatial scan” (Kulldorff, 1997)

Each search region  $S$  is a **sub-region** of space.

- Choose some region shape (e.g. circles, rectangles) and consider all regions of that shape and varying size.
- Low power for true events that do not correspond well to the chosen set of search regions (e.g. irregular shapes).

Our approach: “subset scan” (Neill, 2012)

Each search region  $S$  is a **subset** of locations.

- Find the highest scoring subset, subject to some constraints (e.g. spatial proximity, connectivity).
- For multivariate, also optimize over subsets of streams.
- Exponentially many possible subsets,  $O(2^N \times 2^M)$ : computationally infeasible for naïve search.



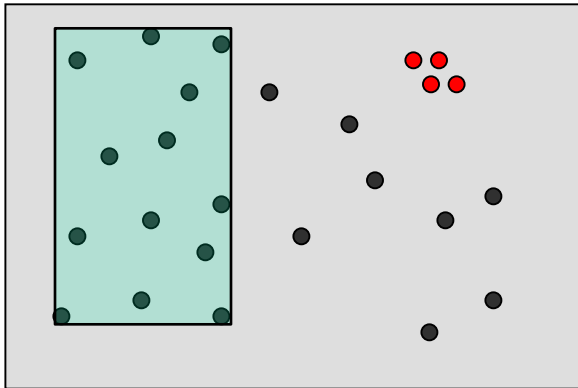
# Question: Why search over subsets?

## Answer: Simpler approaches can fail.

### Top-down detection approaches

Are there any globally interesting patterns? If so, recursively search the most interesting sub-partition.

Two examples: bump hunting;  
“cluster then detect”.

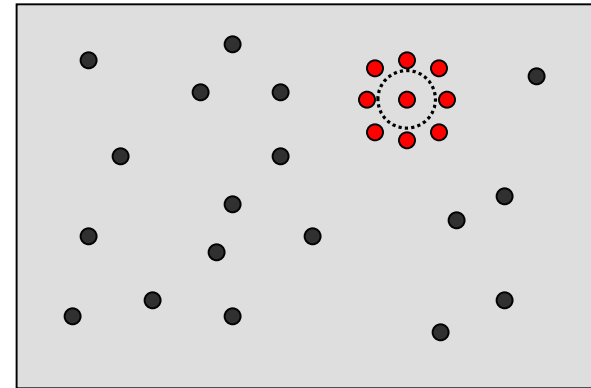


Top-down fails for **small-scale patterns** that are not evident from the global aggregates.

### Bottom-up detection approaches

Find individually (or locally) anomalous data points, and optionally, aggregate into clusters.

Two examples: anomaly/outlier detection;  
density-based clustering.



Bottom-up fails for **subtle patterns** that are only evident when a group of data records are considered collectively.

# Question: Why search over subsets? Answer: Simpler approaches can fail.

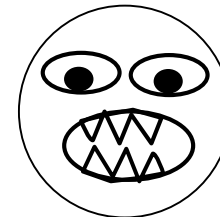
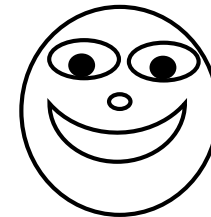
Top-down detection approaches

Are there any patterns?  
the most interesting

So here's where we are so far:

Treating pattern detection as a subset scan problem is statistically desirable for maximizing detection power...

but computationally infeasible (for exhaustive search at least).



Top-down fails to find **subtle patterns** that are not evident when a group of data words are considered collectively.

# Fast subset scan (Neill, 2012)

- In certain cases, we can optimize  $F(S)$  over the exponentially many subsets of the data, while evaluating only  $O(N)$  rather than  $O(2^N)$  subsets.
- Many commonly used scan statistics have the property of linear-time subset scanning:
  - Just sort the data records (or spatial locations, etc.) from highest to lowest priority according to some function...
  - ... then search over groups consisting of the top-k highest priority records, for  $k = 1..N$ .

The highest scoring subset is **guaranteed** to be one of these!

Sample result: we can find the **most anomalous** subset of Allegheny County zip codes in **0.03 sec** vs.  **$10^{24}$  years**.

# Linear-time subset scanning

- Example: Expectation-Based Poisson statistic
  - Sort data locations  $s_i$  by the ratio of observed to expected count,  $c_i / b_i$ .
  - Given the ordering  $s_{(1)} \dots s_{(N)}$ , we can **prove** that the top-scoring subset  $F(S)$  consists of the locations  $s_{(1)} \dots s_{(k)}$  for some  $k$ ,  $1 \leq k \leq N$ .
  - Key step: if there exists some location  $s_{\text{out}} \notin S$  with higher priority than some location  $s_{\text{in}} \in S$ , then we can show that  $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$ .
- Theorem: LTSS holds for expectation-based scan statistics in any exponential family. (Speakman et al., 2016)

$$F(S) = \max_{q>1} \log \frac{P(\text{Data} \mid H_1(S))}{P(\text{Data} \mid H_0)} \quad \begin{array}{l} H_0 : x_i \sim \text{Dist}(\mu_i) \\ H_1 : x_i \sim \text{Dist}(q\mu_i) \end{array}$$

# Linear-time subset scanning

- Example: Expectation-Based Poisson statistic
  - Sort data locations  $s_i$  by the ratio of observed to expected count,  $c_i / b_i$ .
  - Given the ordering  $s_{(1)} \dots s_{(N)}$ , we can **prove** that the top-scoring subset  $F(S)$  consists of the locations  $s_{(1)} \dots s_{(k)}$  for some  $k$ ,  $1 \leq k \leq N$ .
  - Key step: if there exists some location  $s_{\text{out}} \notin S$  with higher priority than some location  $s_{\text{in}} \in S$ , then we can show that  $F(S) \leq \max(F(S \cup \{s_{\text{out}}\}), F(S \setminus \{s_{\text{in}}\}))$ .
- Even better theorem: We can also maximize the **penalized** scan statistic  $F(S) + \sum_{s_i \in S} \Delta_i$  in  $O(N \log N)$  time, evaluating only  $2N$  of the  $2^N$  subsets.

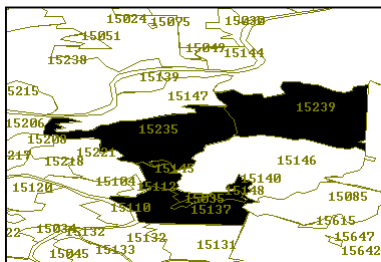
(Speakman et al., 2016)

# Constrained fast subset scanning

LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

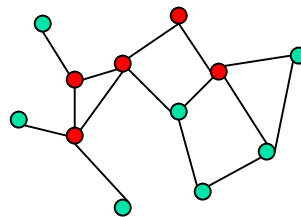
Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + Proximity constraints → Fast spatial scan (irregular regions)
- + Multiple data streams → Fast multivariate scan
- + Connectivity constraints → Fast graph scan
- + Group self-similarity → Fast generalized subset scan

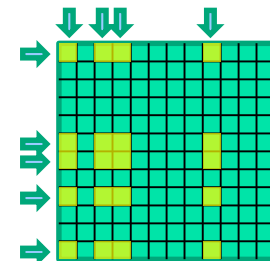


(Neill, *JRSS-B*, 2012)

(Neill et al., *Stat. Med.*, 2013)



(Speakman et al., *JCGS*, 2015)



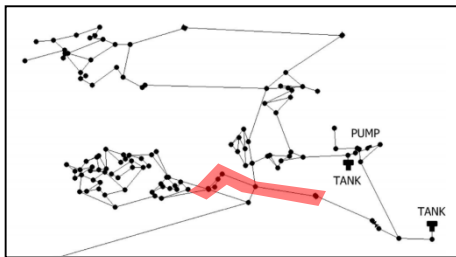
(McFowland et al., *JMLR*, 2013)

# Constrained fast subset scanning

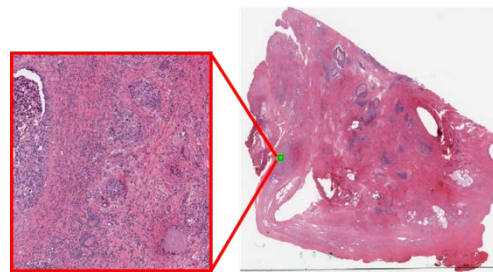
LTSS is a new and powerful tool for **exact** combinatorial optimization (as opposed to approximate techniques such as submodular function optimization). But it only solves the “best unconstrained subset” problem, and cannot be used directly for constrained optimization.

Many of our recent papers have focused on how LTSS can be extended to the many real-world problems with (hard or soft) constraints on our search.

- + Temporal dynamics → Spreading contamination in water supply
- + Hierarchical scanning → Prostate cancer in digital pathology slides
- + Scalable GP regression → Predicting and preventing rat infestations



(Speakman et al., ICDM 2013)



(Somanchi et al.,  
*Stat. Med.*, 2018)



(Flaxman et al., 2015;  
Neill et al., in preparation)

# Fast subset scan with spatial proximity constraints

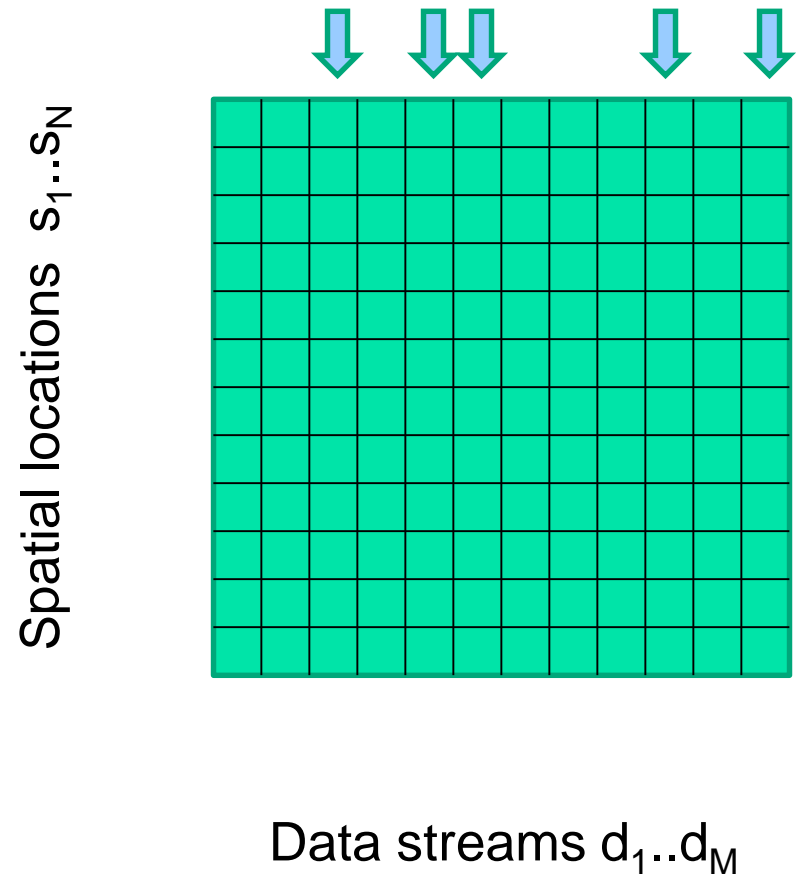
- Maximize a likelihood ratio statistic over all subsets of the “local neighborhoods” consisting of a center location  $s_i$  and its  $k-1$  nearest neighbors, for a fixed neighborhood size  $k$ .
- Naïve search requires  $O(N \cdot 2^k)$  time and is computationally infeasible for  $k > 25$ .
- For each center, we can search over all subsets of its local neighborhood in  $O(k)$  time using LTSS, thus requiring a total time complexity of  $O(Nk) + O(N \log N)$  for sorting the locations.
- In Neill (2012), we show that this approach dramatically improves the timeliness and accuracy of outbreak detection for irregularly-shaped disease clusters.



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

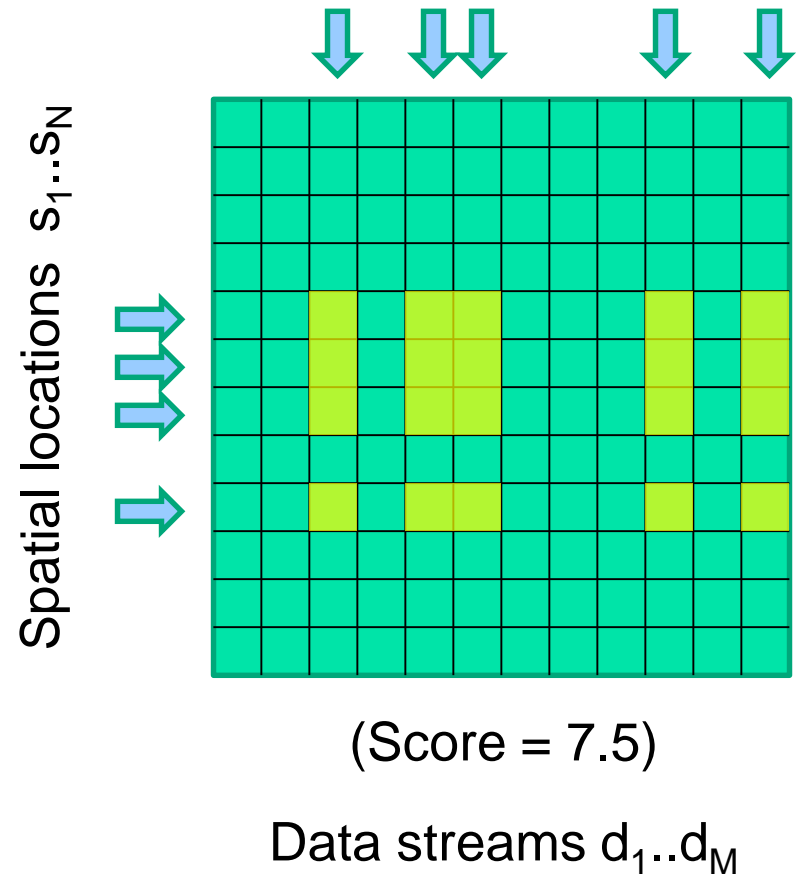
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

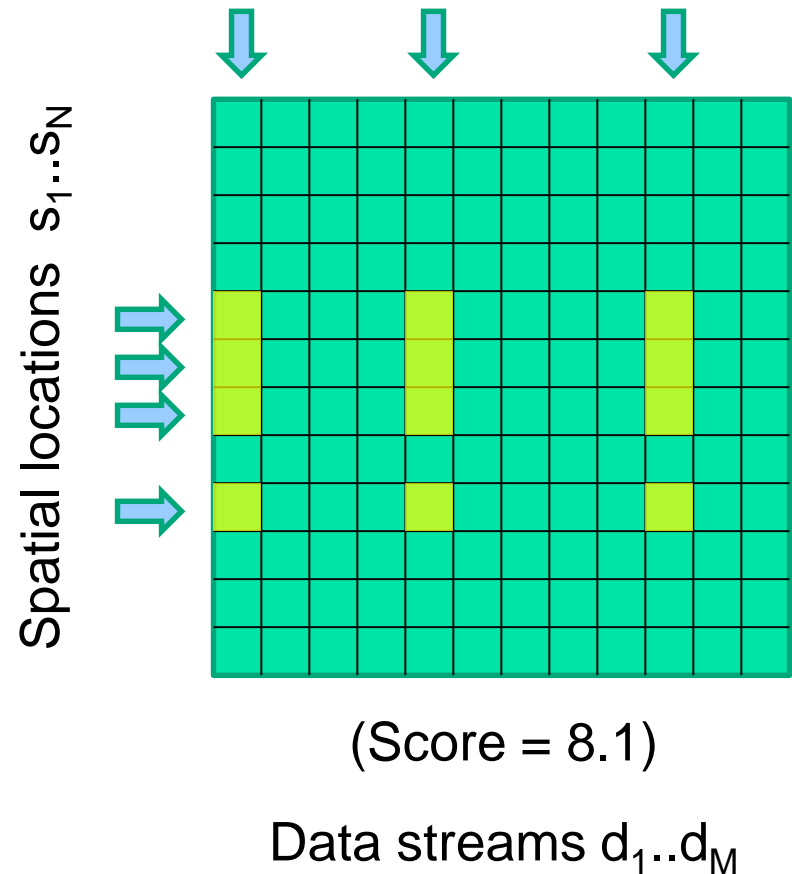
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

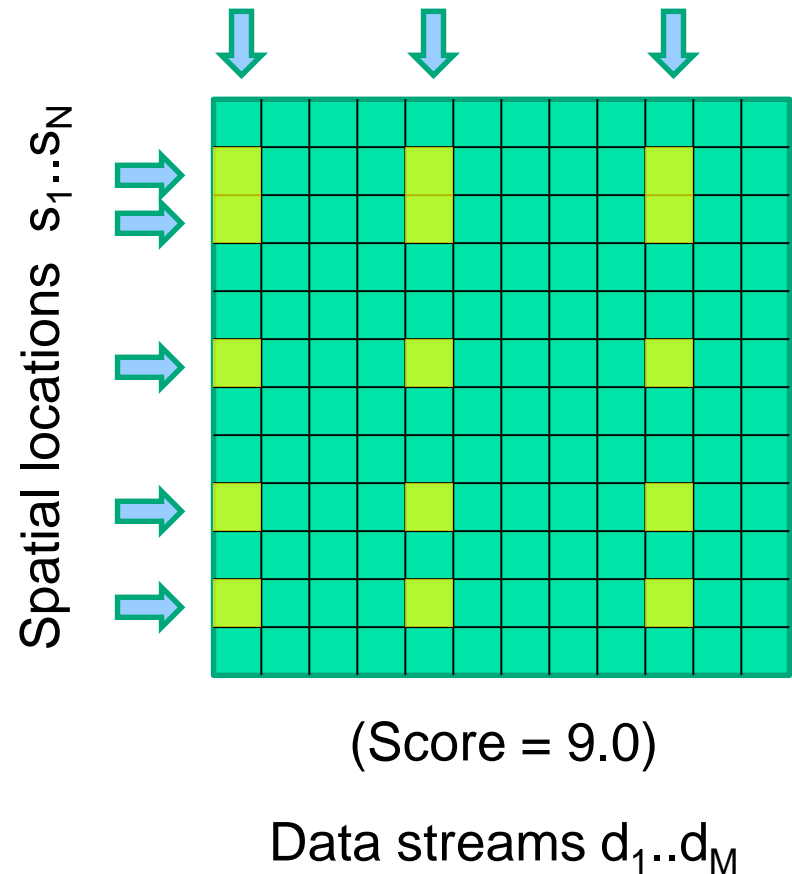
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

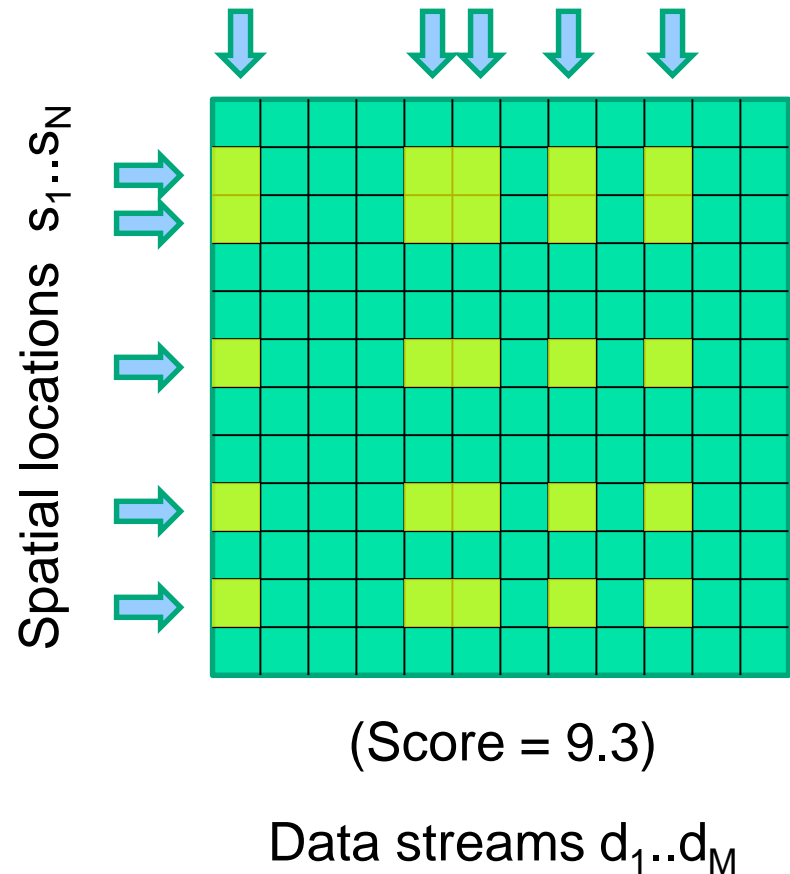
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!



# Multivariate fast subset scan

(Neill, McFowland, and Zheng, 2013)

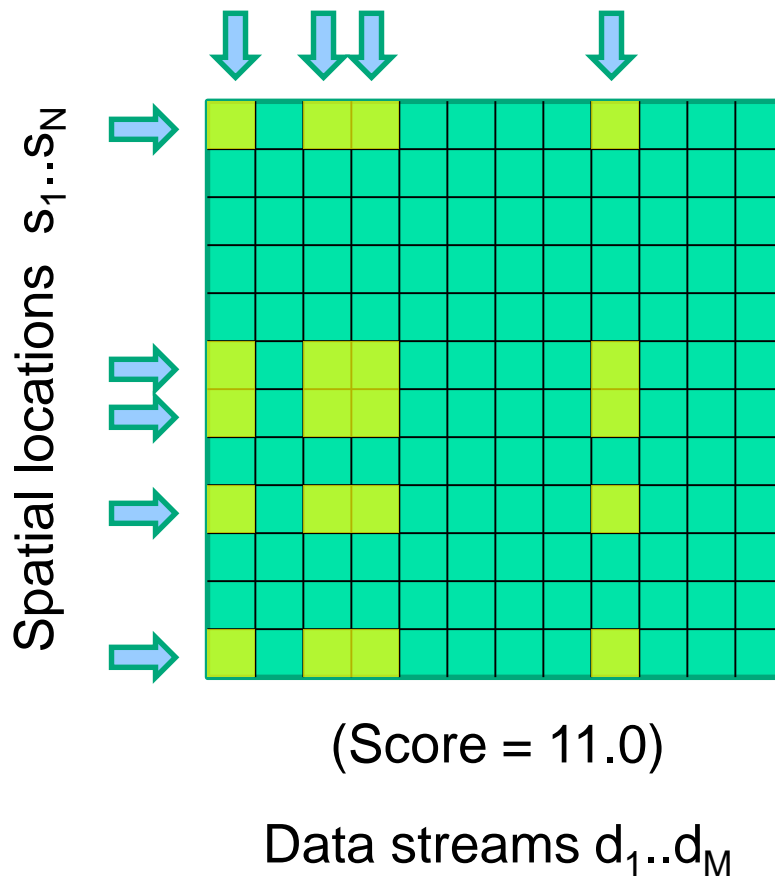
- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!



# Multivariate fast subset scan

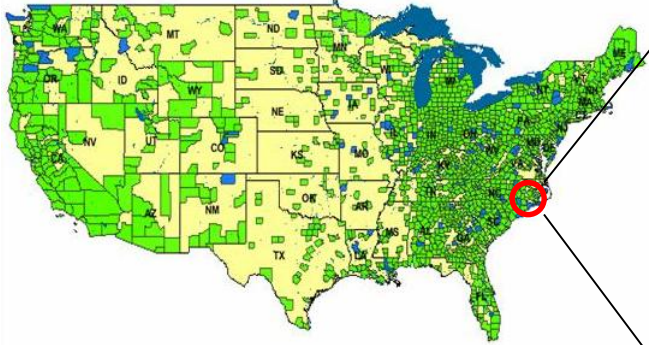
(Neill, McFowland, and Zheng, 2013)

- The LTSS property allows us to efficiently optimize over subsets of spatial locations for a given subset of streams.
- But it also allows us to efficiently optimize over subsets of **streams** for a given subset of **locations**...
- So we can jointly optimize over subsets of streams **and** locations by iterating between these steps!
- Converges to local maximum: we do multiple random restarts to approach the global maximum.
- For general datasets, a similar approach\* can be used to jointly optimize over subsets of data records and attributes.

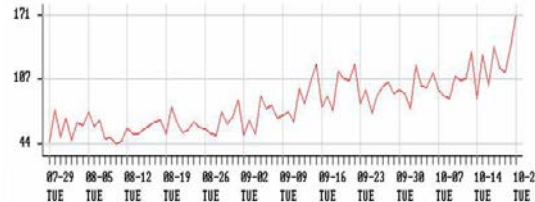


\*McFowland, Speakman, and Neill, *JMLR*, 2013

# Multivariate event detection



Spatial time series data from spatial locations  $s_i$  (e.g. zip codes)



Time series of counts  $c_{i,m}^t$  for each zip code  $s_i$  for each data stream  $d_m$ .

## Outbreak detection

- $d_1$  = respiratory ED
- $d_2$  = constitutional ED
- $d_3$  = OTC cough/cold
- $d_4$  = OTC anti-fever  
(etc.)

## Main goals:

- Detect** any emerging events.
- Pinpoint** the affected subset of locations and time duration.
- Characterize** the event, e.g., by identifying the affected streams.

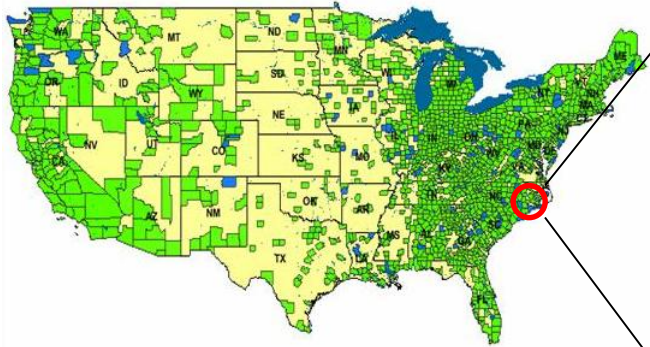
## Compare hypotheses:

$$H_1(D, S, W)$$

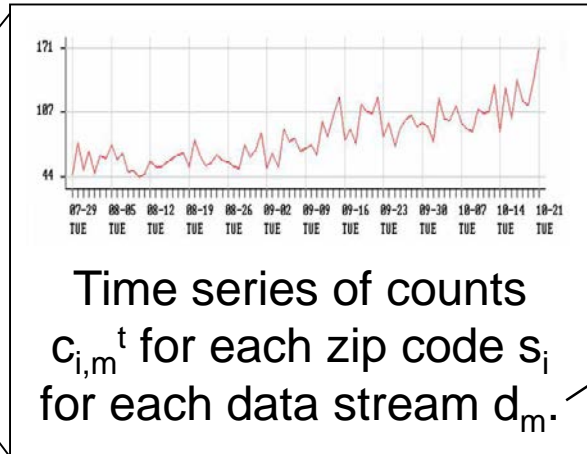
- $D$  = subset of streams
- $S$  = subset of locations
- $W$  = time duration

vs.  $H_0$ : no events occurring

# Multidimensional event detection



Spatial time series data from spatial locations  $s_i$  (e.g. zip codes)



## Outbreak detection

- $d_1$  = respiratory ED
- $d_2$  = constitutional ED
- $d_3$  = OTC cough/cold
- $d_4$  = OTC anti-fever  
(etc.)

Additional goal: identify any differentially affected **subpopulations**  $P$  of the monitored population.

- Gender (male, female, both)
- Age groups (children, adults, elderly)
- Ethnic or socio-economic groups
- Risk behaviors: e.g. intravenous drug use, multiple sexual partners

More generally, assume that we have a set of additional discrete-valued attributes  $A_1..A_J$  observed for each individual case. We identify not only the affected streams, locations, and time window, but also a **subset** of values for each attribute.



# Multidimensional subset scan

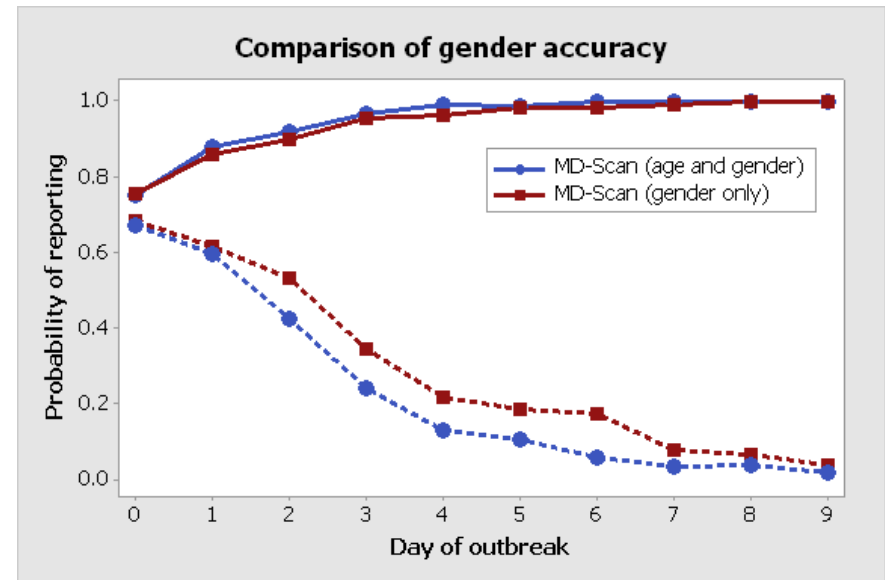
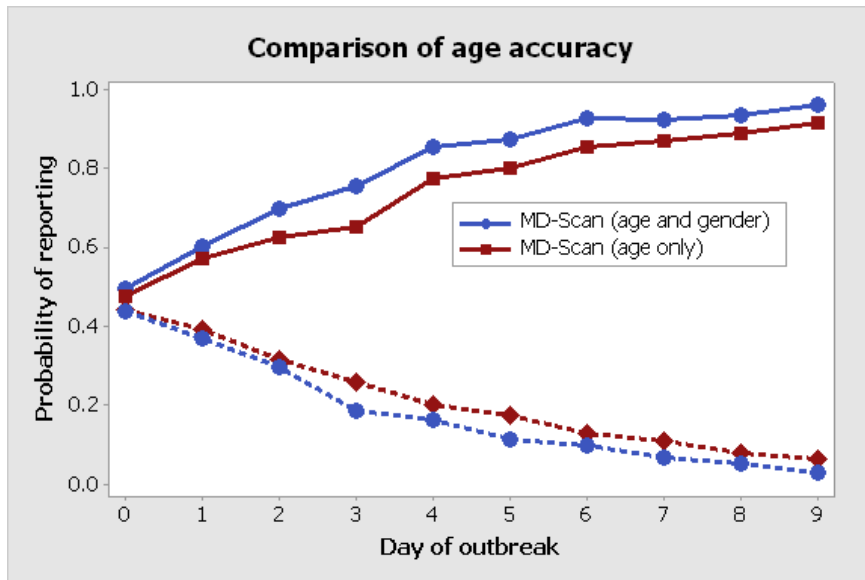
- Our **MD-Scan** framework (Neill & Kumar, 2013) extends LTSS to the multidimensional case:
  - For each time window and spatial neighborhood (center + k-nearest neighbors), we do the following:
    1. Start with randomly chosen subsets of **locations**  $S$ , **streams**  $D$ , and **values**  $V_j$  for each attribute  $A_j$  ( $j=1..J$ ).
    2. Choose an attribute  $A$  (randomly or sequentially) and use LTSS to find the highest scoring subset of values, locations, or streams, conditioned on all other attributes.  
**\*\*\* Linear rather than exponential in arity of A \*\*\***
    3. Iterate step 2 until convergence to a local maximum of the score function  $F(D, S, W, \{V_j\})$ , and use multiple restarts to approach the global maximum.

# Evaluation of MD-Scan

- We first evaluated the detection performance of MD-Scan for detecting simulated disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- For outbreaks with differential effects by age and gender, MD-Scan demonstrated **more timely** and **more accurate** detection, and accurately **characterized** the affected subpopulations.

# 1) Identifying affected subpopulations

By the midpoint of the outbreak, MD-Scan is able to correctly identify the affected gender and age deciles with high probability, without reporting unaffected subpopulations.



**Proportions of correct and incorrect groups reported vs. time since start of outbreak.**

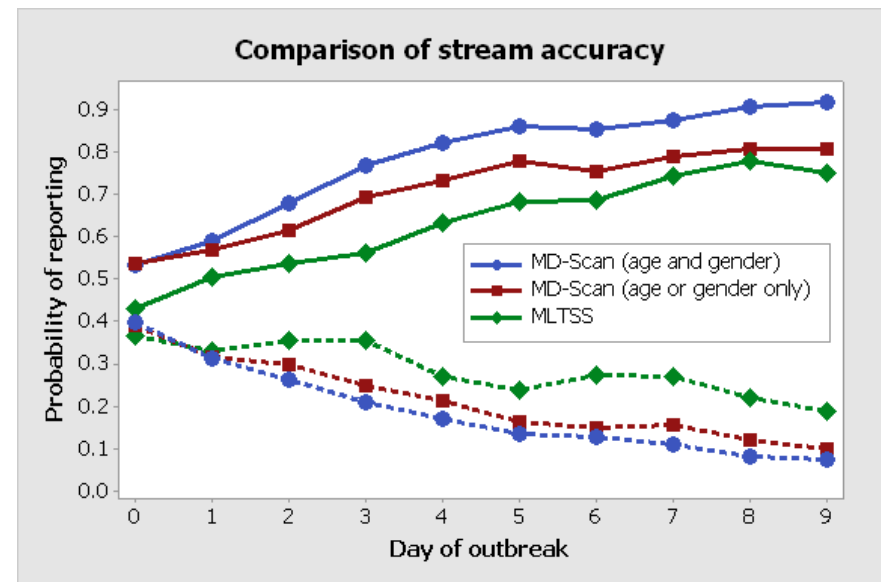
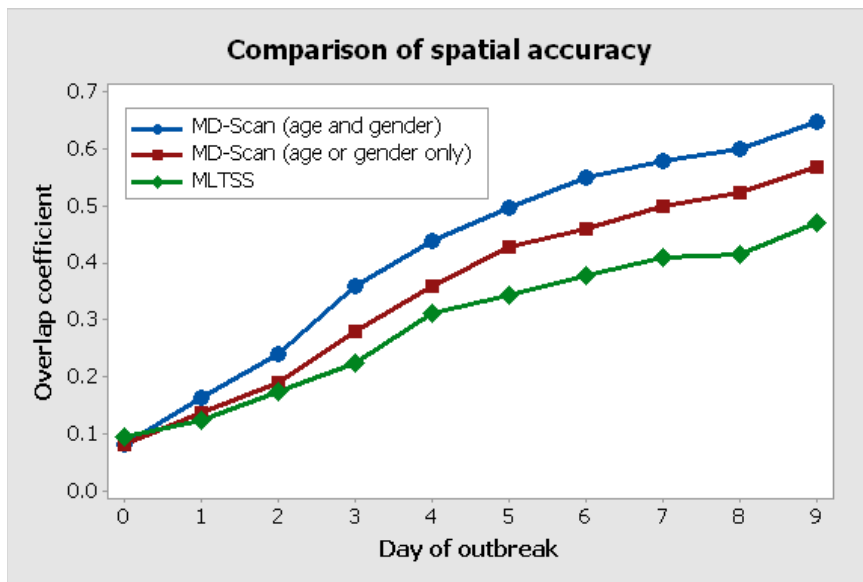
Solid lines: affected gender and/or age deciles. Dashed lines: unaffected.

Blue lines: outbreaks with differential effects by both age and gender (easier).

Red lines: outbreaks with differential effects by age or gender only (harder).

## 2) Characterizing affected streams

As compared to the previous state of the art (multivariate linear-time subset scanning), MD-Scan is better able to characterize the affected spatial locations and subset of the monitored streams.



**Left: overlap coefficient between true and detected subsets of spatial locations.**  
**Right: Proportions of correct and incorrect streams reported vs. day of outbreak.**

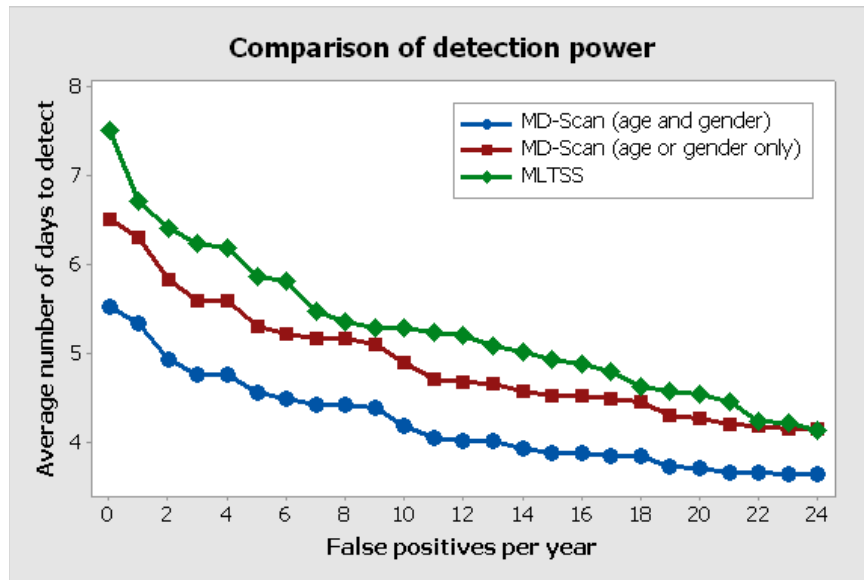
Blue lines: outbreaks with differential effects by both age and gender (easier).

Red lines: outbreaks with differential effects by age or gender only (harder).

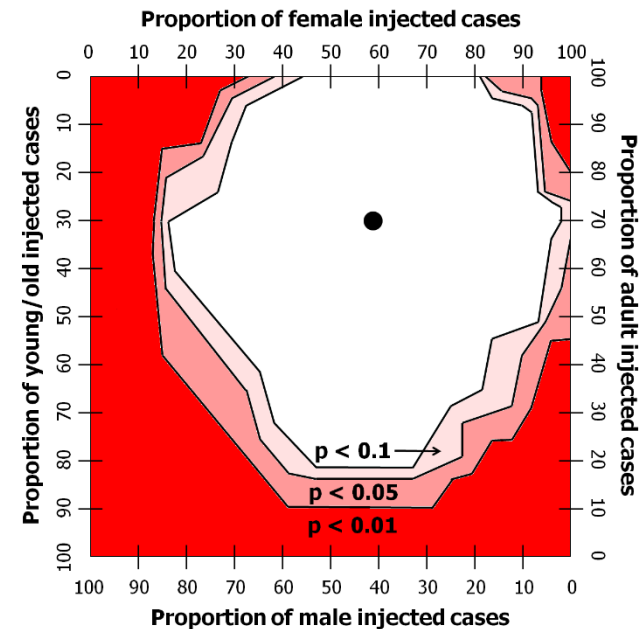
Green lines: MLTSS, ignoring age and gender information

# 3) Timeliness of outbreak detection

MD-Scan achieved significantly more timely detection for outbreaks that were sufficiently biased by age and/or gender.



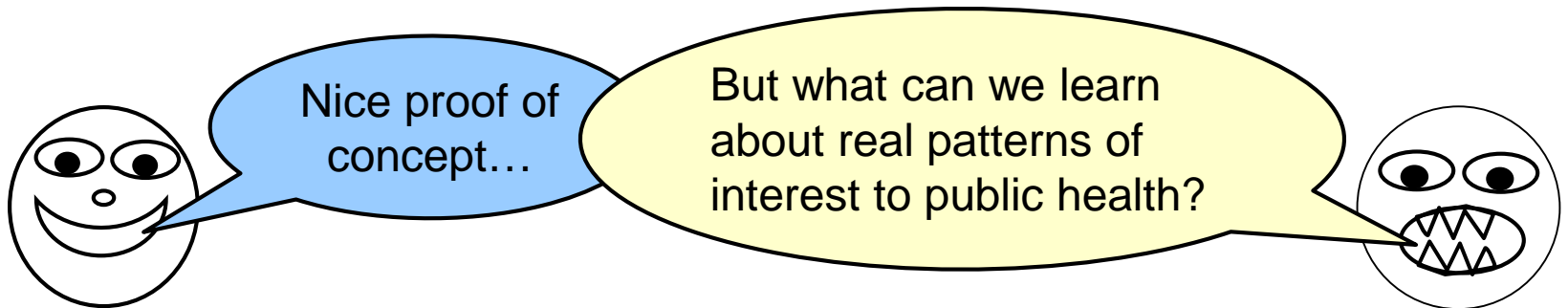
For outbreaks with strong age and gender biases, time to detection improved from 5.2 to 4.0 days at a fixed false positive rate of 1/month.



Smaller biases in age or gender were sufficient for significant improvements; even when no age/gender signal is present, MD-Scan performs comparably to MLTSS.

# Evaluation of MD-Scan

- We first evaluated the detection performance of MD-Scan for detecting simulated disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- For outbreaks with differential effects by age and gender, MD-Scan demonstrated **more timely** and **more accurate** detection, and accurately **characterized** the affected subpopulations.



# Allegheny County Overdose Data

- We analyzed county medical examiner data for fatal accidental drug overdoses, 2008-2015.
- ~2000 cases: for each overdose victim, we have date, location (zip), age, gender, race, and the set of drugs present in their system.
- Reduced to 30 dimensions (age decile, gender, race, presence/absence of 27 common drugs) plus space and time.
- Clusters discovered by MD-Scan were shared with Allegheny County Dept. of Human Services.

# MD-Scan Overdose Results (1)



**Fentanyl** is a dangerous drug which has been a huge problem in western PA.

It is often mixed with white powder heroin, or sold disguised as heroin.

January 16-25, 2014:

14 deaths county-wide from fentanyl-laced heroin.

March 27 to April 21, 2015:

26 deaths county-wide from fentanyl, heroin only present in 11.

January 10 to February 7, 2015:

Cluster of 11 fentanyl-related deaths, mainly black males over 58 years of age, centered in Pittsburgh's downtown Hill District.

Very unusual demographic: common dealer / shooting gallery?

Started in the SE suburbs of Pittsburgh, including a cluster of 5 cases around McKeesport between March 27 and April 8.

Cluster score became significant March 29<sup>th</sup> (4 nearby cases, white males ages 20-49) and continued to increase through April 20<sup>th</sup>.

Fentanyl, heroin, and combined deaths remained high through end of June (>100).



# MD-Scan Overdose Results (2)

Another set of discovered overdose clusters each involved a combination of Methadone and Xanax.



The combination produces a strong high but can be deadly (~30% of methadone fatal ODs).

From 2008-2012: multiple M&X OD clusters, 3-7 cases each, localized in space and time.

Methadone: an opioid used for chronic pain relief and to treat heroin addiction, but also addictive and risk of OD.

From 2013-2015: no M&X overdose clusters; 33% and 47% drops in yearly methadone and M&X deaths respectively.



Xanax (alprazolam): a benzodiazepine prescribed for panic and anxiety disorders.

Why did these deaths cluster, when methadone and methadone + other benzo deaths did not?

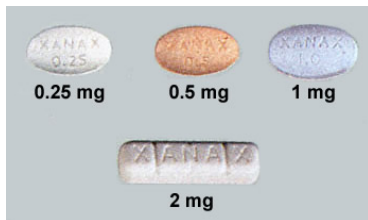
What factors could explain the dramatic reduction in M&X overdose clusters?

# MD-Scan Overdose Results (2)

Another set of discovered overdose clusters each involved a combination of Methadone and Xanax.



Methadone: an opioid used for chronic pain relief and to treat heroin addiction, but also addictive and risk of OD.



Xanax (alprazolam): a benzodiazepine prescribed for panic and anxiety disorders.

Increased state oversight of methadone clinics and prescribing physicians after passage of the Methadone Death and Incident Review Act (Oct 2012).

Approval of generic suboxone (buprenorphine + naloxone) in early 2013 lowered cost of suboxone treatment as an alternative to methadone clinics.

Why did these deaths cluster, when methadone and methadone + other benzo deaths did not?

What factors could explain the dramatic reduction in M&X overdose clusters?

# Identifying causal effects of treatments and exposures

## 1. Healthcare treatments

We are using health insurance claims data from ~125K individuals to identify anomalous patterns of patient care that impact health outcomes.

- \* Correct suboptimal care \*
- \* Identify new best practices \*

Key idea: treatment effects may be **heterogeneous**; look for most positively and negatively affected subpopulations.

“**Glucocorticoids** significantly increase hospitalizations following treatment in the subpopulation of hypertensive, overweight males with endocrine disorders.”



## 2. Environmental health

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?”

Must adjust for known confounders, selection into treatment/exposure.

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



Another application of MD-Scan: auditing algorithms for fairness.



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Source:  
Julia Angwin,  
Jeff Larson,  
Surya Mattu and  
Lauren Kirchner, *ProPublica*

# Machine Bias

There's software used across the country to predict future criminals.  
And it's biased against blacks.

# Motivating questions

- Is the COMPAS algorithm for predicting re-offending risk **fair**, or is it **biased** against some subpopulation defined by observed characteristics?
  - **Black box** algorithm. All we observe is predictions vs. gold standard (re-offending) for a sample of individuals (ProPublica data from Broward County, FL).
  - Many possible biases: race, gender, age, past offenses...
  - Combinations of factors, e.g., “elderly white females”
- This led us to develop a **general approach** to auditing black box algorithms for fairness or bias.

# Broward County data

- Source: ProPublica's data on criminal defendants in Broward County, FL, in 2013-2014
- Outcome: re-arrests (!) assessed through April 2016.
- Score: **COMPAS** score from 1 (low risk) to 10 (high risk)

<b>Background</b>	Black ( $n = 3696$ )		White ( $n = 2454$ )
Age	32.7 (10.9)	<	37.7 (12.8)
Male (%)	82.4	>	76.9
Number of Priors	4.44 (5.58)	>	2.59 (3.8)
Any priors? (%)	76.4	>	65.9
Felony (%)	68.9	>	60.3
COMPAS Score	5.37 (2.83)	>	3.74 (2.6)

# What does it mean to be “fair”?

There are at least three possibilities (and probably more):

**1) Group Fairness:** The same proportion of each group should be classified as “high risk”.

- Doesn’t seem reasonable for COMPAS: observed reoffending rates are not constant across groups. For Broward County, 51% of black defendants and 39% of white defendants reoffended.

**2) Disparate Impacts:** Comparing false positive and false negative rates across groups.

- Impacts depend on how predictions are used (particularly if the prediction is a probability). Can we separate **fairness of prediction** from **fair decisions** using these predictions?

# What does it mean to be “fair”?

There are at least three ways to think about fairness (see):

1) G  
sh

3) We focus on **unbiasedness** of probability estimates.

Individual risk probabilities should be predicted accurately, **without systematic biases** based on any observed attributes or combinations of attributes.

→ Are there any **statistically significant** biases?

→ Can we automatically **correct** these systematic biases, in order to improve fairness of prediction?

prediction is fair only if the  
prediction is fair and  
**prediction from fair** of these predictions?



# Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the multidimensional subset scan to identify subgroups where classifier predictions are significantly biased.

Assume a dataset with inputs  $x_i$ , binary labels  $y_i \in \{0, 1\}$ , and the classifier's risk predictions  $\hat{p}_i = \Pr(y_i = 1)$ .

Search space: subspaces defined by a subset of values for each attribute (e.g., “white and Asian males under 25”)

Score function: a log-likelihood ratio statistic.  $H_0$ :  $\hat{p}_i$  correctly calibrated;  $H_1(S)$ : constant multiplicative increase or decrease in odds of  $y_i = 1$  for subspace  $S$ .

$$F(S) = \max_q \log \prod_{s_i \in S} \frac{\Pr\left(y_i \sim \text{Bernoulli}\left(\frac{q\hat{p}_i}{1 - \hat{p}_i + q\hat{p}_i}\right)\right)}{\Pr(y_i \sim \text{Bernoulli}(\hat{p}_i))}$$

# Bias scan (Zhang and Neill, 2016)

Our goal is to **detect** and **correct** any **systematic biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

We developed a new variant of the multidimensional subset scan to identify subgroups where classifier predictions are significantly biased.

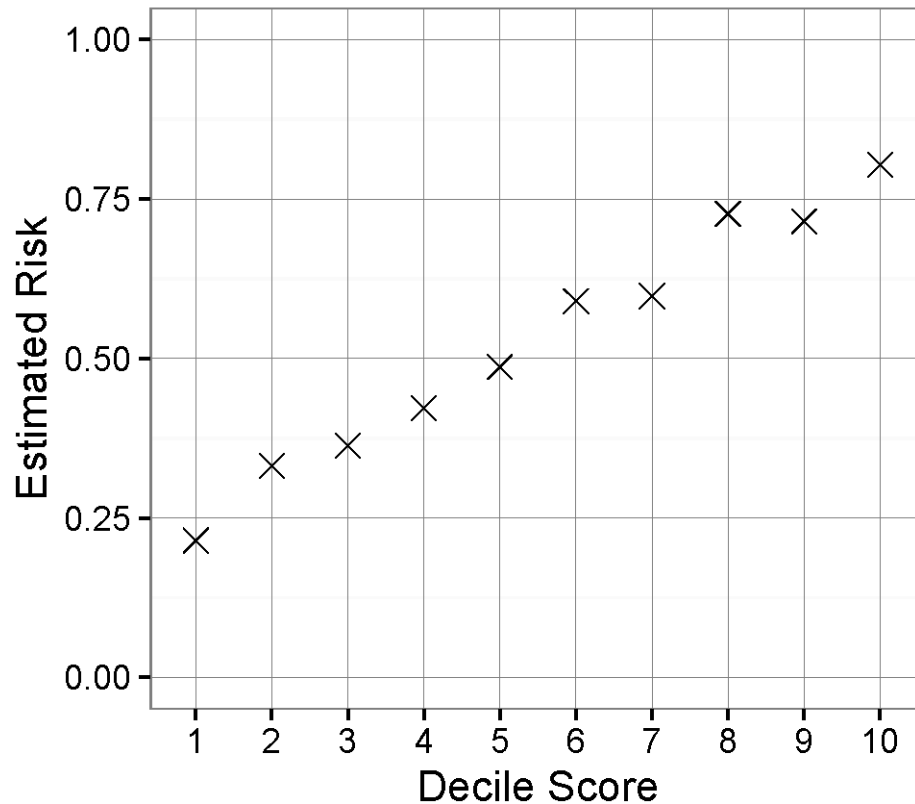
Assume a dataset with inputs  $x_i$ , binary labels  $y_i \in \{0, 1\}$ , and the classifier's risk predictions  $\hat{p}_i = \Pr(y_i = 1)$ .

Search space: subspaces defined by a subset of values for each attribute (e.g., “white and Asian males under 25”)

Score function: a log-likelihood ratio statistic.  $H_0$ :  $\hat{p}_i$  correctly calibrated;  $H_1(S)$ : constant multiplicative increase or decrease in odds of  $y_i = 1$  for subspace  $S$ .

For interpretability, we maximize the penalized score  $F(S) - \log \prod |S_j|$ , where attributes with no excluded values are ignored. For each conditional optimization, we can use the simple penalty,  $\log(|S_j|) 1\{|S_j| < \text{arity}(A_j)\}$ .

# Results of bias scan on COMPAS

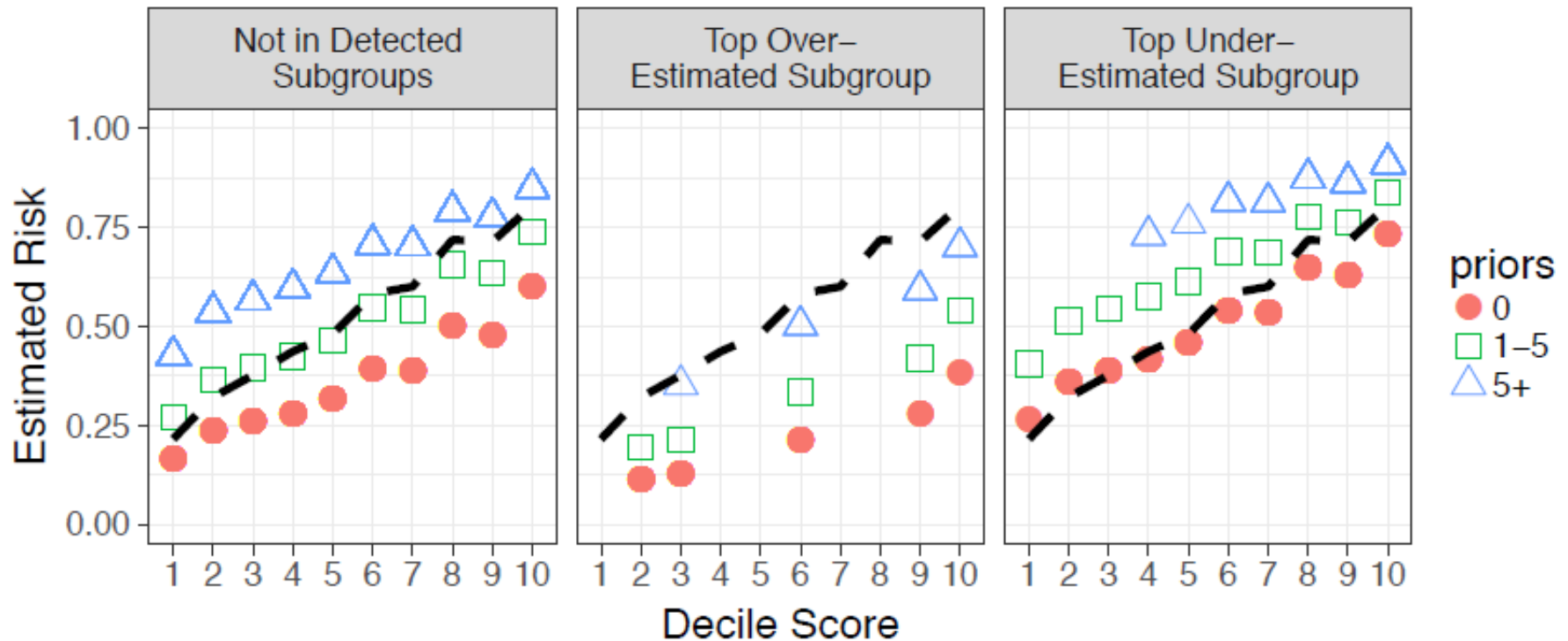


Start with maximum likelihood risk estimates for each COMPAS decile score.

Detection result 1: COMPAS underestimates the importance of prior offenses, overestimating risk for 0 priors, and underestimating risk for 5 or more priors.

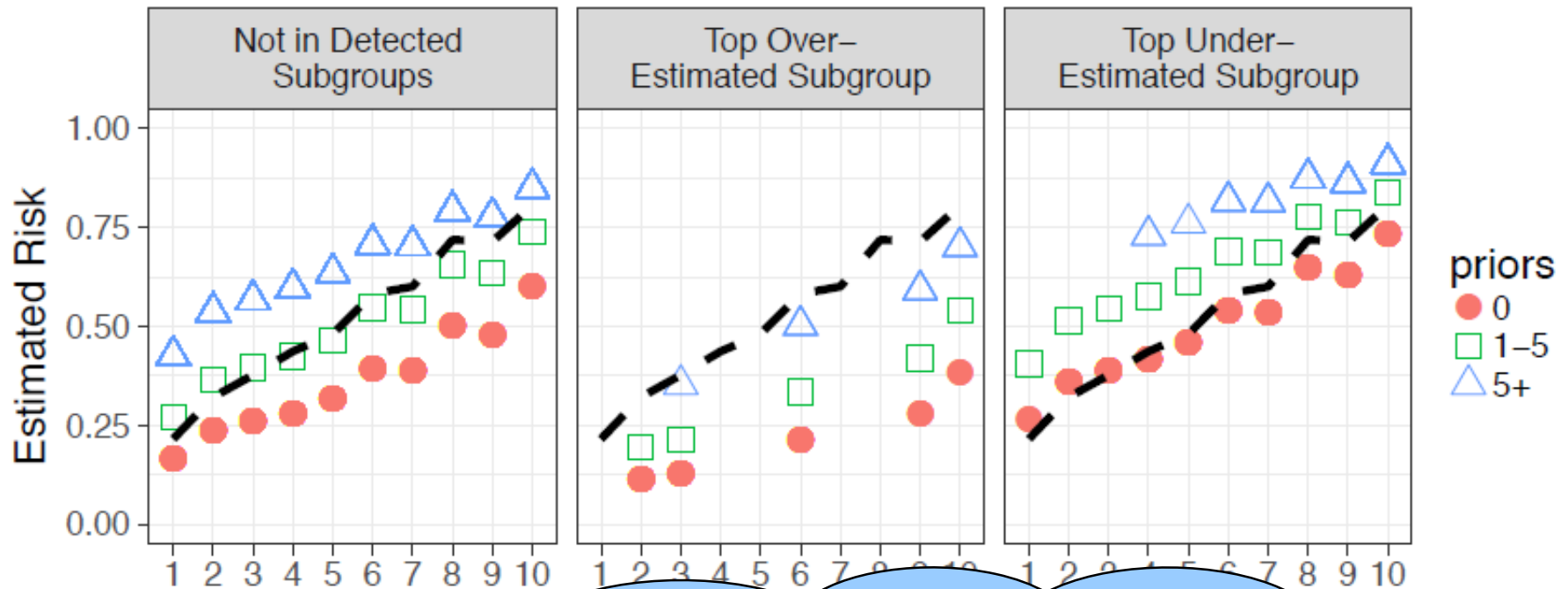
Detection result 2: Even controlling for prior offenses, COMPAS still underestimates risk for males under 25, and overestimates risk for females who committed misdemeanors.

# Results of bias scan on COMPAS



After controlling for number of prior offenses and for membership in the two detected subgroups, there are no significant systematic biases in prediction.

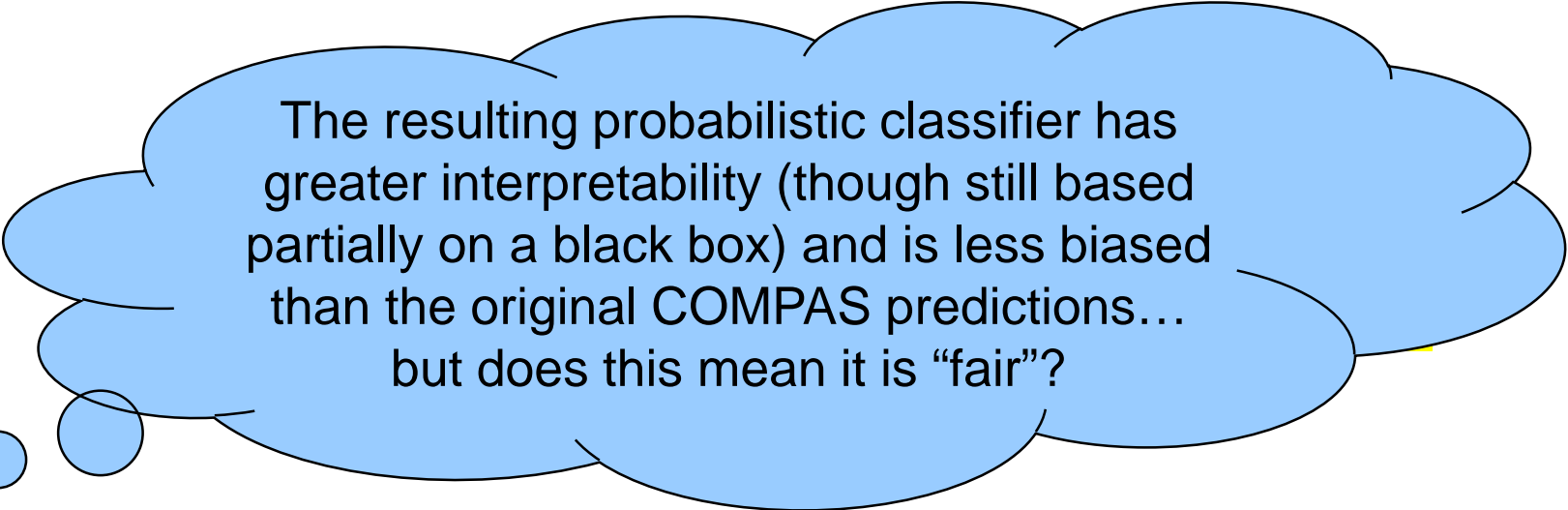
# Results of bias scan on COMPAS



The resulting probabilistic classifier has greater interpretability (though still based partially on a black box) and is less biased than the original COMPAS predictions... but does this mean it is “fair”?

# Discussion: predictive fairness in context

- The method does not account for **target variable bias**: we predict re-offending risk but the gold standard is based on re-arrests not re-offenses.
  - Big problem with drug possession, weapon possession charges. Leads to feedback loops.
- How to avoid **disparate impacts** when making decisions based on even unbiased predictions?



The resulting probabilistic classifier has greater interpretability (though still based partially on a black box) and is less biased than the original COMPAS predictions... but does this mean it is “fair”?

# Conclusions

Real-world problems at the societal scale require new computational methods to deal with both the **size** and the **complexity** of data.



**Fast subset scanning** (with constraints) can serve as a fundamental building block for efficient, scalable pattern detection in massive data.

Practical solutions to societal challenges also require an understanding of complex data (text, networks, images, streams, ...), leading to **new statistical and algorithmic tools** for extracting relevant patterns.

# 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. Speakman, S. Somanchi, E. McFowland III, D.B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics* 25: 382-404, 2016.
- D.B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. *Statistics in Medicine* 32: 2185-2208, 2013.
- E. McFowland III, S. Speakman, and D.B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14: 1533-1561, 2013.
- F. Chen and D.B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. *Proc. 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.
- F. Chen and D.B. Neill. Human rights event detection from heterogeneous social media graphs. *Big Data* 3(1): 34-40, 2015.
- D.B. Neill and W. Herlands. Machine learning for drug overdose surveillance. *Journal of Technology in Human Services* 36(1): 8-14, 2018.
- Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. <https://arxiv.org/abs/1611.08292>





**Thanks for listening!**

More details on my web site:  
<http://www.cs.nyu.edu/~neill>

Or e-mail me at:  
[daniel.neill@nyu.edu](mailto:daniel.neill@nyu.edu)