

Tutorial on Event Detection

KDD 2009

Daniel B. Neill
H. J. Heinz III College
Carnegie Mellon University
neill@cs.cmu.edu

Weng-Keen Wong
School of EECS
Oregon State University
wong@eecs.oregonstate.edu

Introduction

- Many real-world tasks in surveillance, scientific discovery and data cleaning involve monitoring routinely collected data
- Want to detect “**events of interest**” which are usually anomalous events that rarely occur
- These events typically affect a *subgroup* of the data rather than an *individual* data point
- Examples to follow...

Introduction

Early detection of disease outbreaks

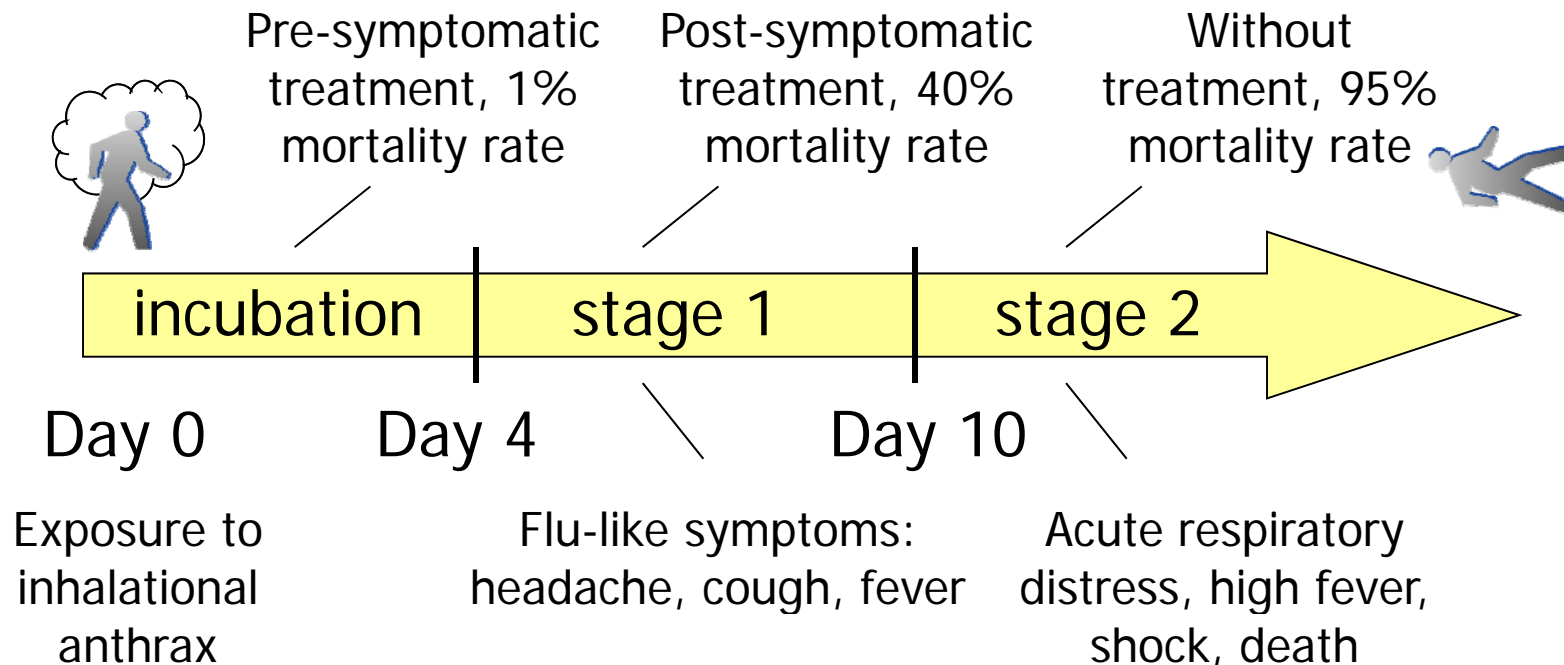
- Bioterrorist attacks are a very real, and scary, possibility
 - 100 kg anthrax, released over D.C., could kill 1-3 million and hospitalize millions more.
- Emerging infectious diseases
 - “Conservative estimate” of 2-7 million deaths from pandemic avian influenza.
- Better response to common outbreaks (seasonal flu, GI)



Introduction

Benefits of early detection:

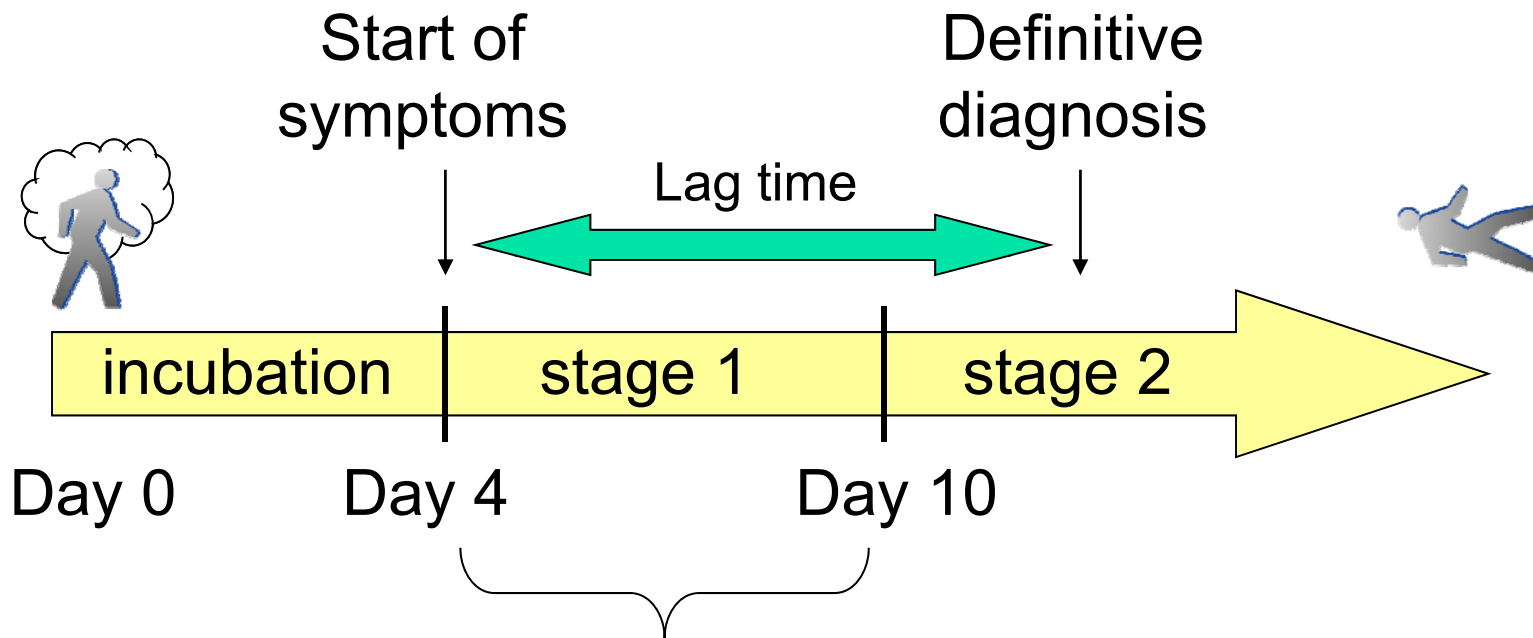
Reduces **cost to society**, both in lives and in dollars!



DARPA estimate: a two-day gain in detection time could reduce fatalities by a factor of six.

Introduction

Early detection is hard

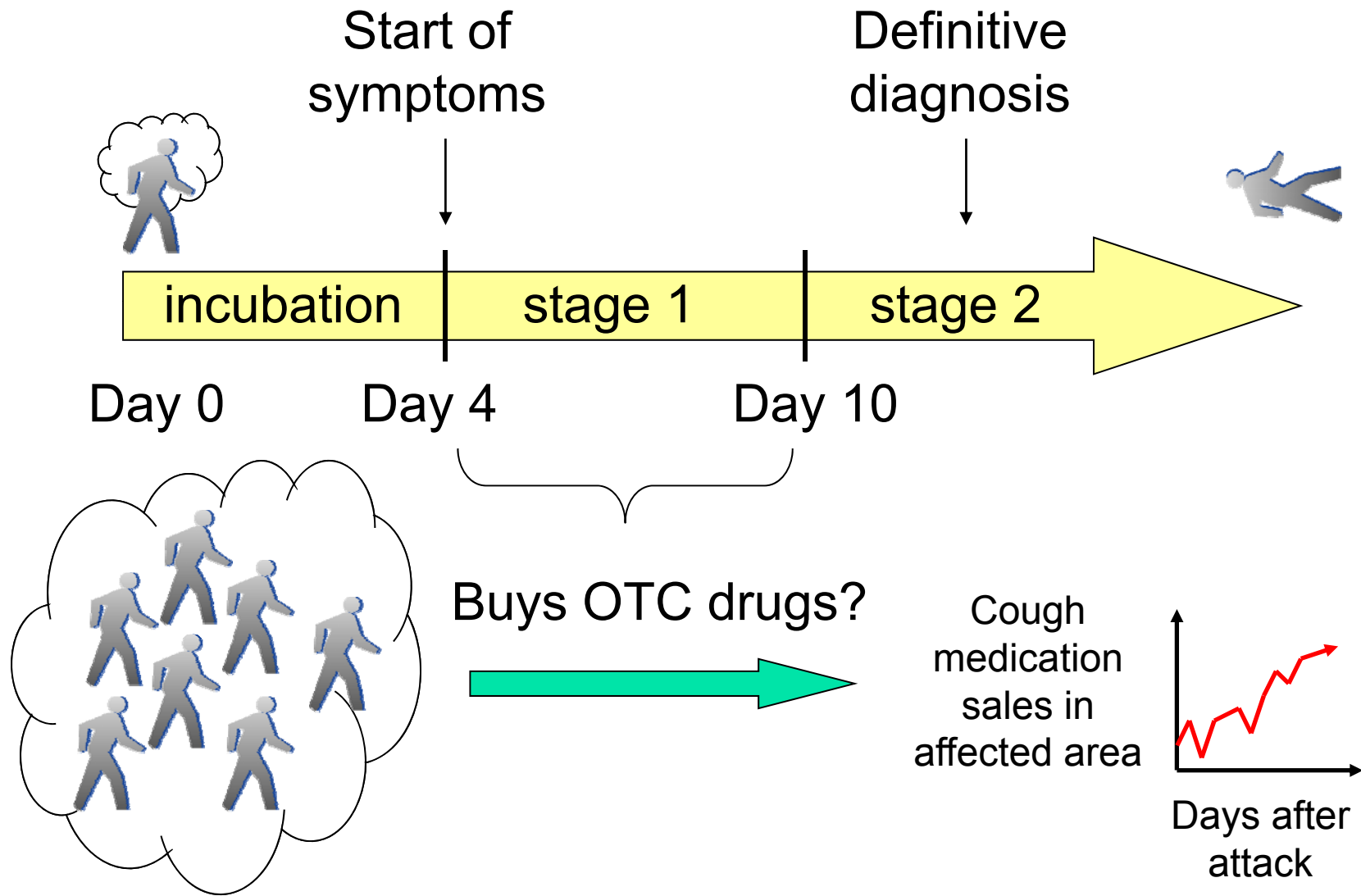


Buys OTC drugs?

Skips work/school?

Visits doctor/hospital/ED?

Introduction



Introduction

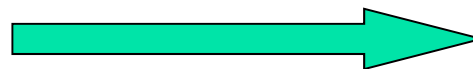
Start of
symptoms

Definitive
diagnosis

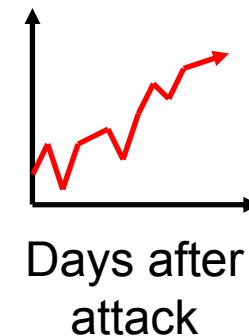
We can achieve very early detection of outbreaks by gathering syndromic data, and identifying emerging spatial clusters of symptoms.



Buys OTC drugs?

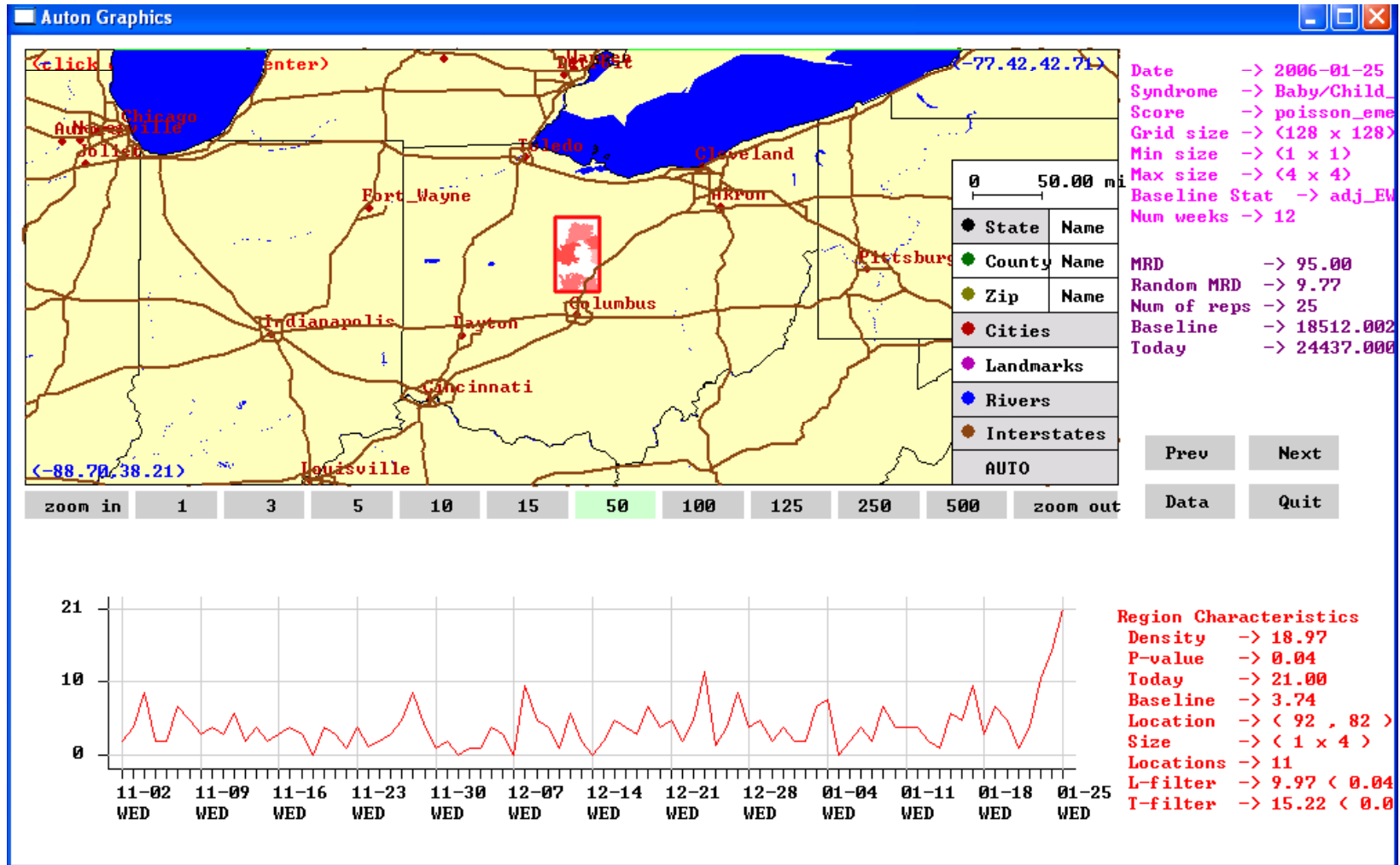


Cough
medication
sales in
affected area



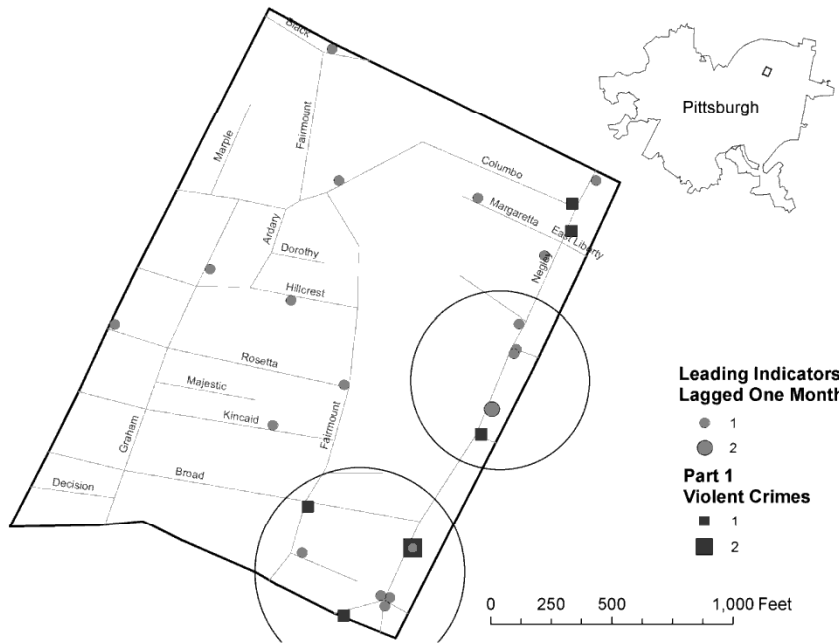
Introduction

Spike in sales of pediatric electrolytes near Columbus, Ohio



Introduction

Crime hot-spot detection



Application to law enforcement:
detecting crime hot-spots.

Hot-spot = neighborhood or other spatial area with an unexpected rise in crime.

Goal: early detection to enable targeted enforcement.

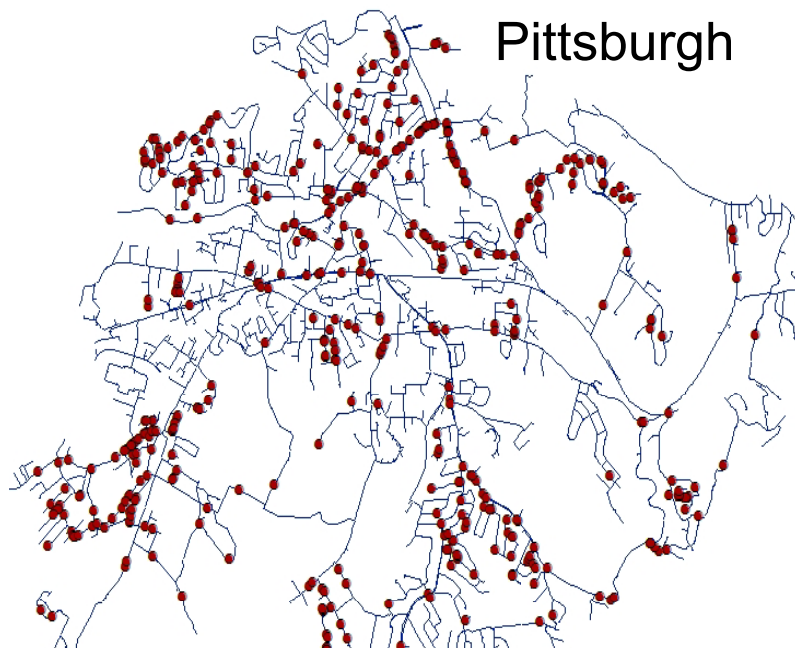
Even better goal: predict where hot spots of crime are going to occur, and prevent them.

We demonstrated¹ that hot-spots of violent crime can be predicted 1-3 weeks in advance, by detecting clusters of “leading indicator” crimes such as disorderly conduct, trespass, and simple assault.

¹D.B. Neill and W.L. Gorr, Proc. ISDS Annual Conf. 2007.

Introduction

Detecting clusters of pipe breaks



Application to civil engineering:
Monitoring a city's water distribution system to detect anomalous clusters of pipe breakage.¹

Different distance metric: flow distance along pipes, not Euclidean distance.

Must account for pipe age, dimensions, and material when computing expected number of breaks.

¹D. Olivera, et al., in preparation. Thanks to Daniel Olivera for providing this picture.

Introduction

Detecting illicit container shipments



Goal is to detect patterns of suspicious shipments corresponding to illegal activity (terrorism, smuggling, etc.)

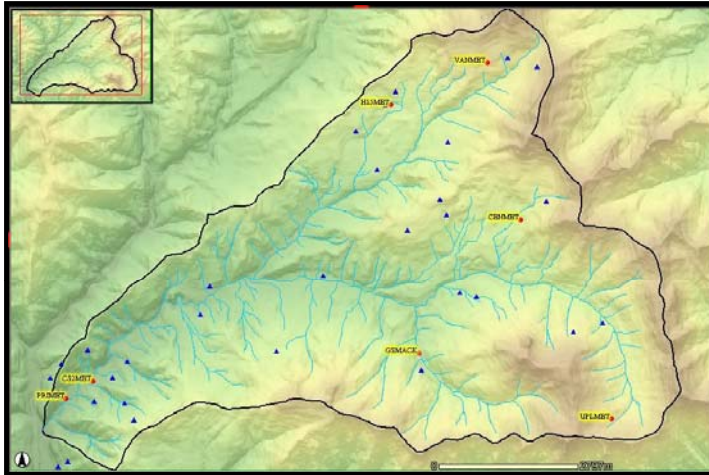
We can achieve this goal by detecting anomalous, self-similar groups of records.

No “spatial” dimension in the standard sense, but we can define a dissimilarity metric between shipments and detect anomalous patterns in metric space.

FPORT	USPORT	COUNTRY	SLINE	VESSEL	SHIPPER NAME	F NAME	COMMODITY	SIZE	MTONS	VALUE
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	AMERICAN_TRI_NET_EXPRES	TRI_NET	EMPTY_RACK	0	5.6	27579
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRE	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRE	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	AMERICAN_TRI_NET_EXPRES	TRI_NET	CRUDE_IODINE_PURITY	1	17.68	251151
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRES	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	CHINA_OCEAN_SHPG	CHINA_OC	EMPTY_CONTAINERS	0	0	0
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	CHINA_OCEAN_SHPG	CHINA_OC	EMPTY_CONTAINERS	0	0	0

Introduction

Environmental Monitoring



Remote Sensors are becoming the new standard for collecting field data

Nearly continuous observation of a given domain, generating large volumes of data

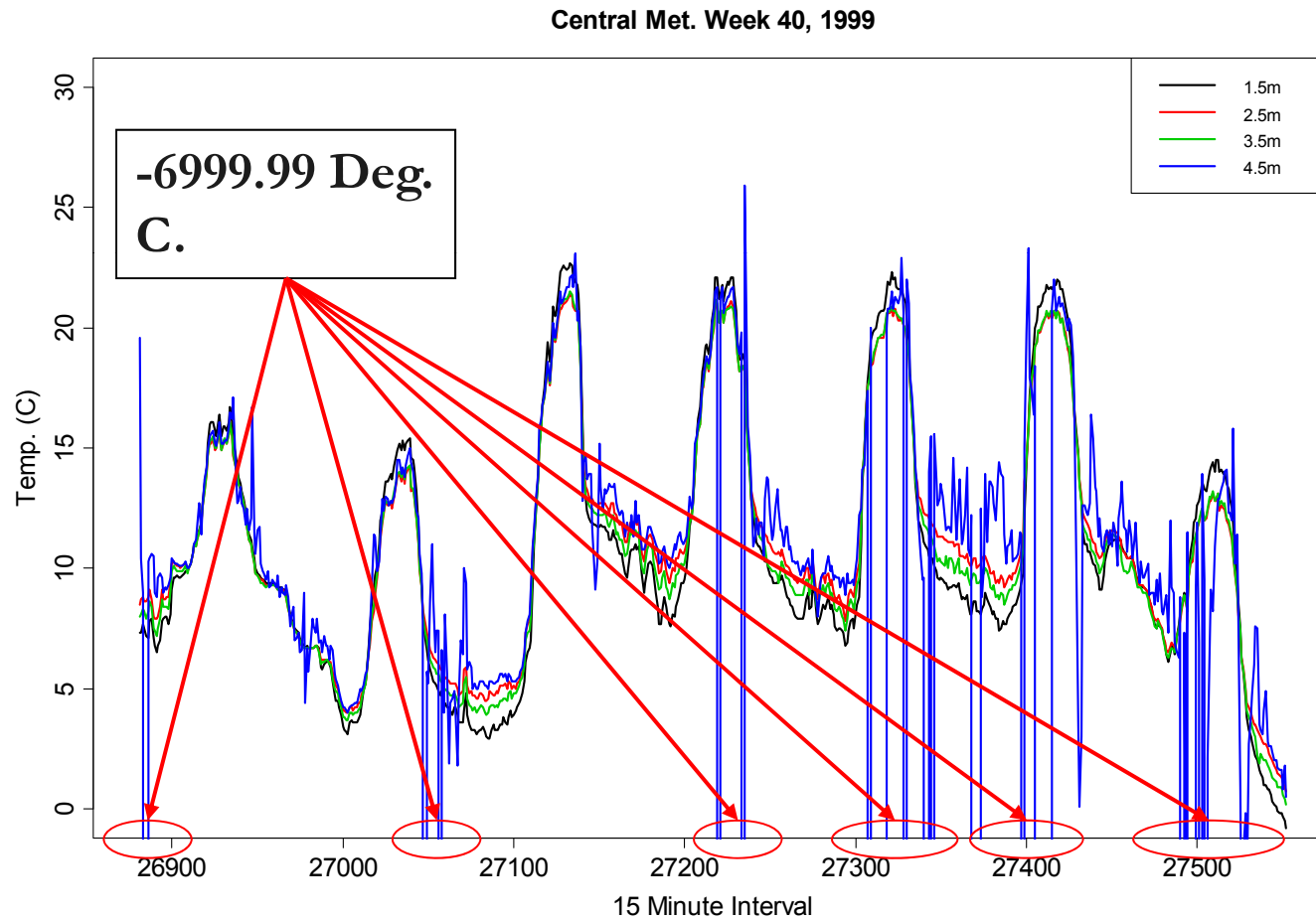
Data must be cleaned before being given to outside researchers.
Requires removal of anomalous data points.

Anomalies can be simple or very challenging!

From: Dereszynski, E., Dietterich, T. (2007). **Probabilistic Models for Anomaly Detection in Remote Sensor Data Streams**. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007)*. 75-82. Thanks to Ethan Dereszynski for the slide materials.

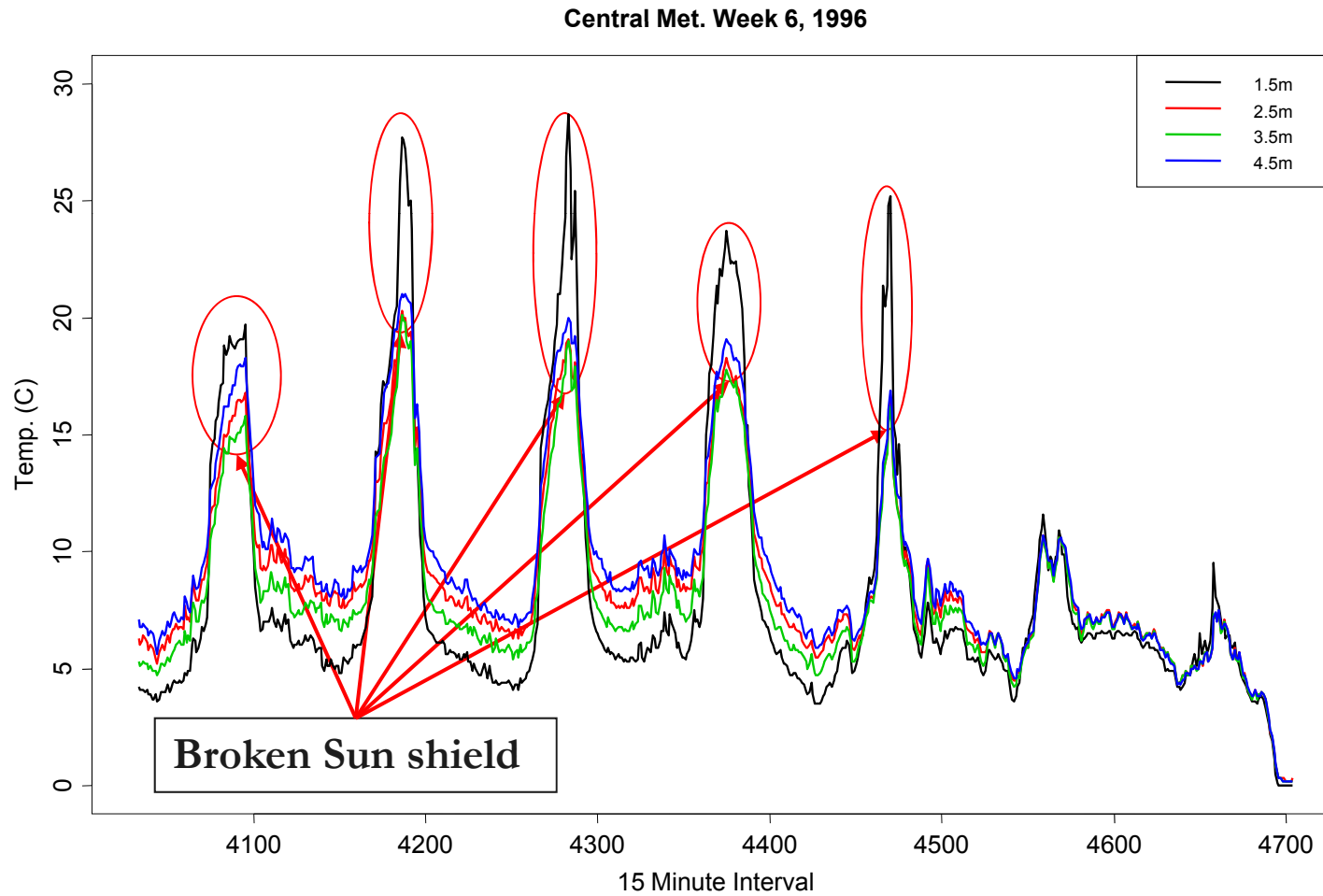
Introduction

Simple Anomaly Types



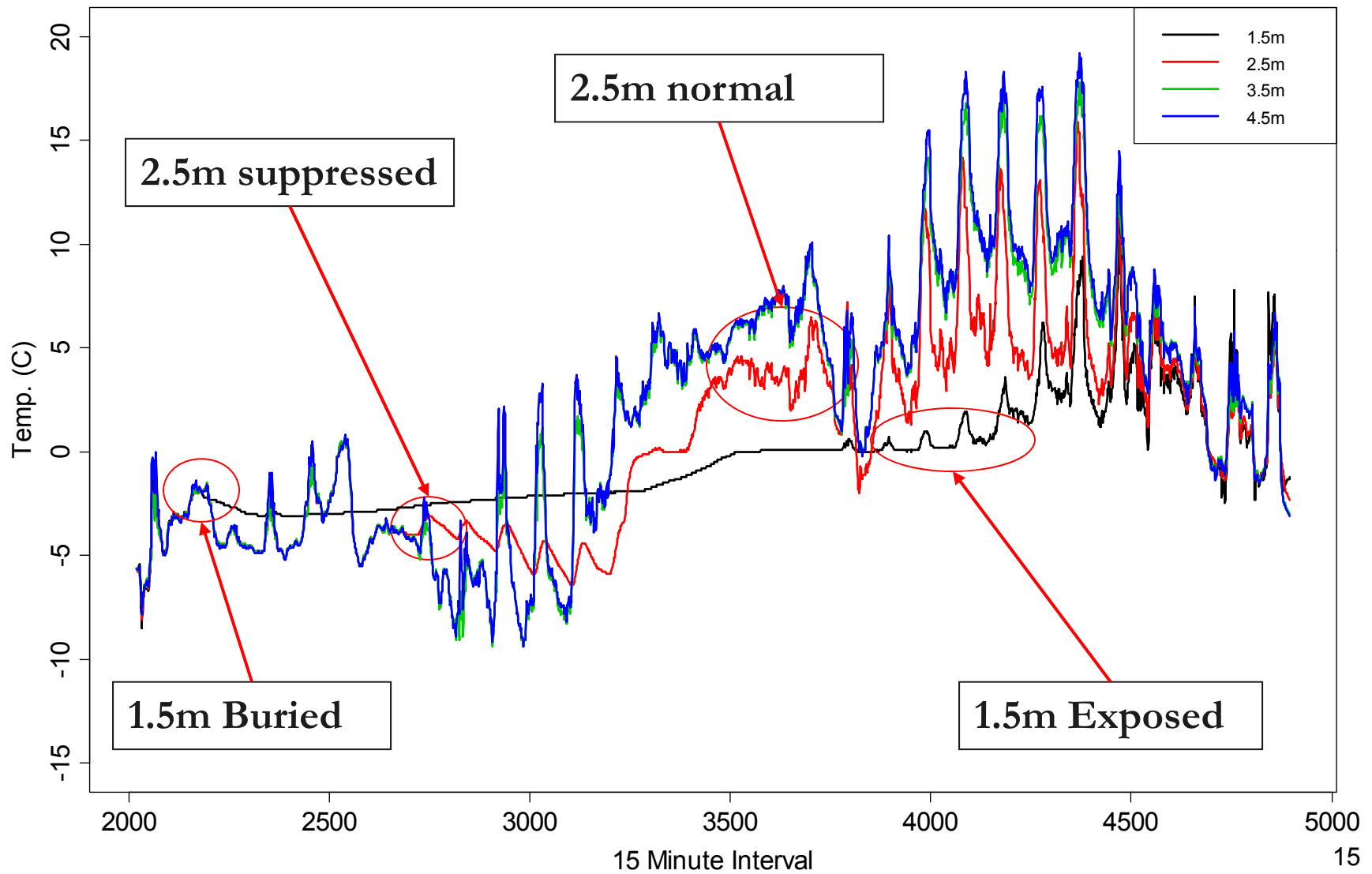
Introduction

More difficult anomaly types



Introduction

Upper Lookout Met. Weeks 3-7, 1996



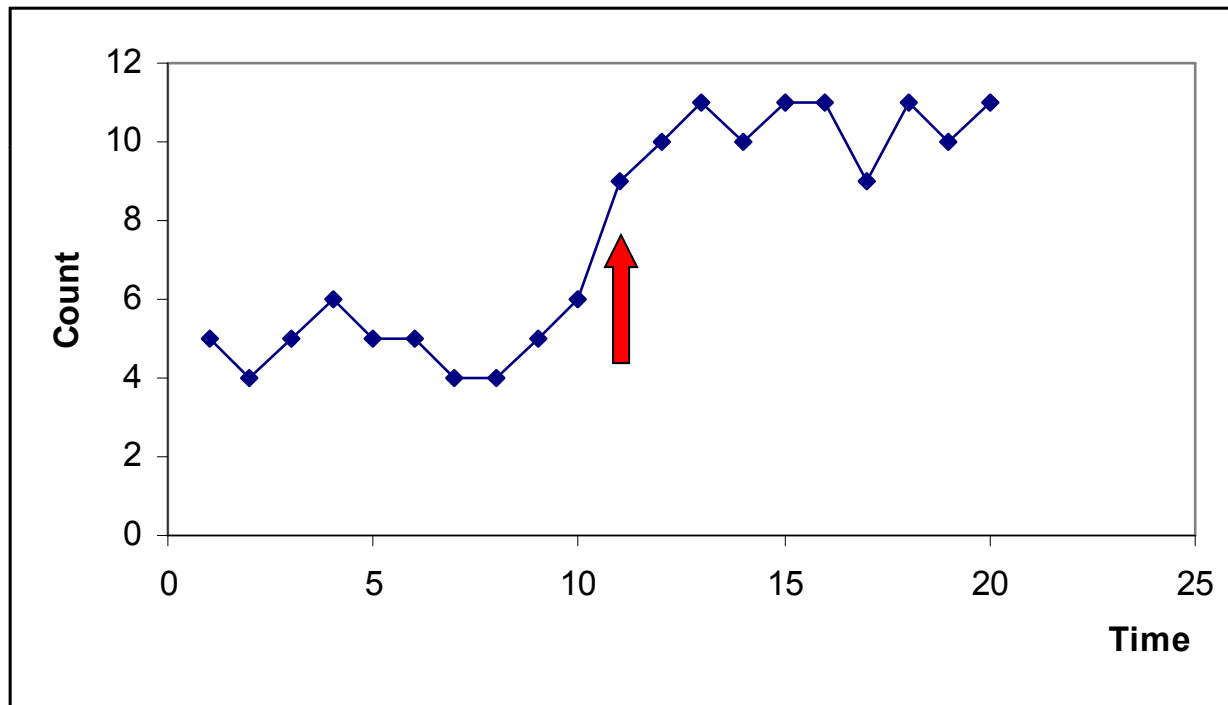
Introduction

Suppose you have data $\mathbf{D} = \{x_1, \dots, x_n\}$ where x_i arrives over time

- Can we detect a time t when an event of interest occurs?
- The question we are asking is: at what point in time is the data is “different”?

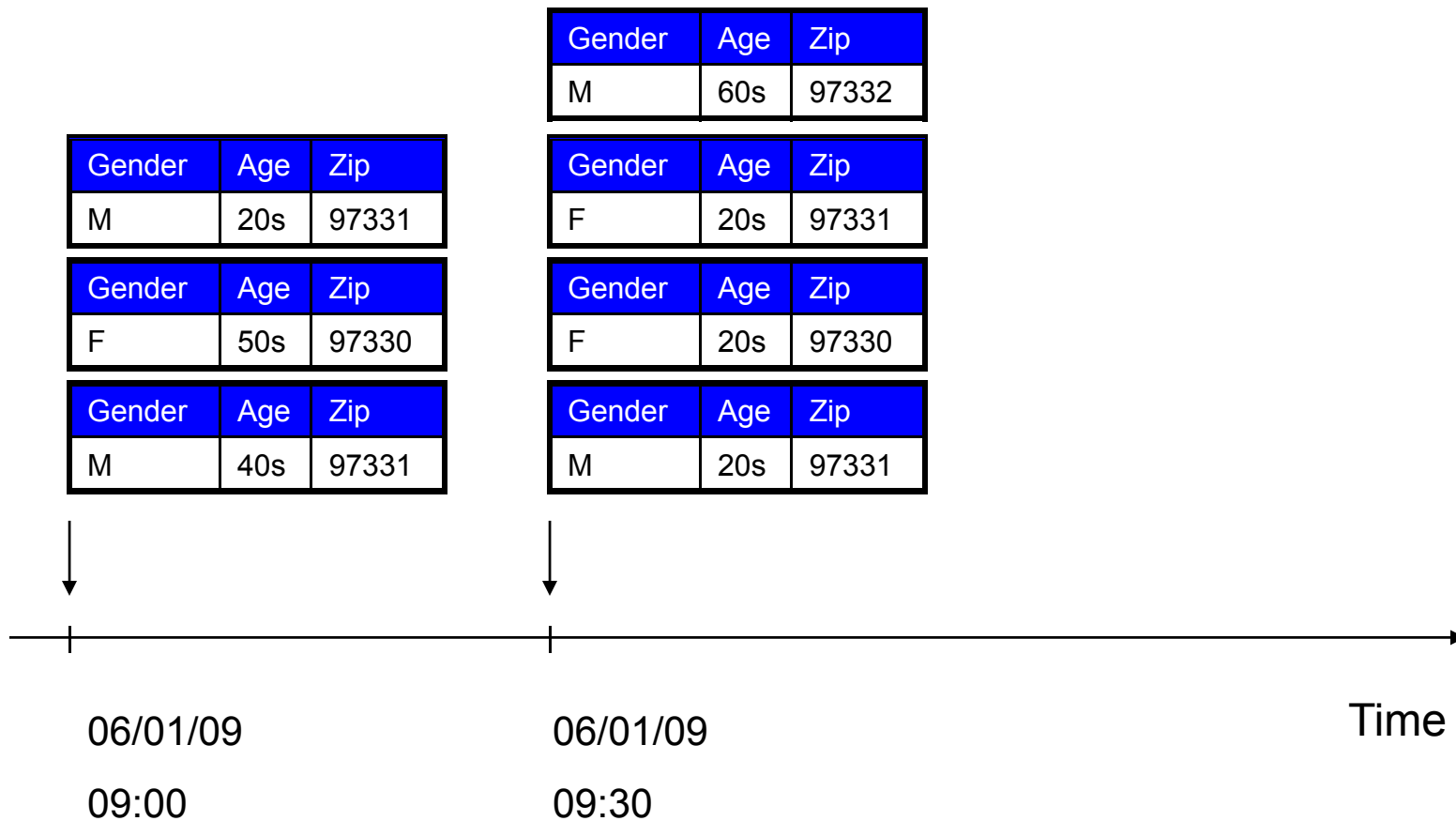
Introduction

Simple example: x_i is a scalar eg. a count



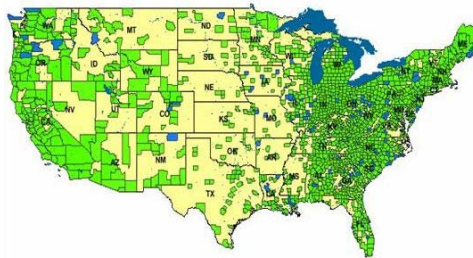
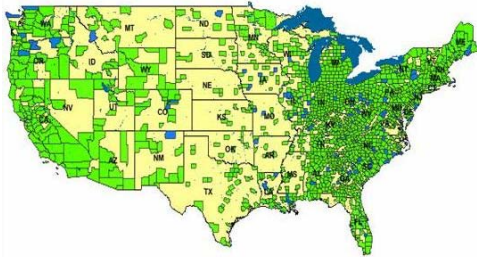
Introduction

Harder example: x_i is a vector of categorical values

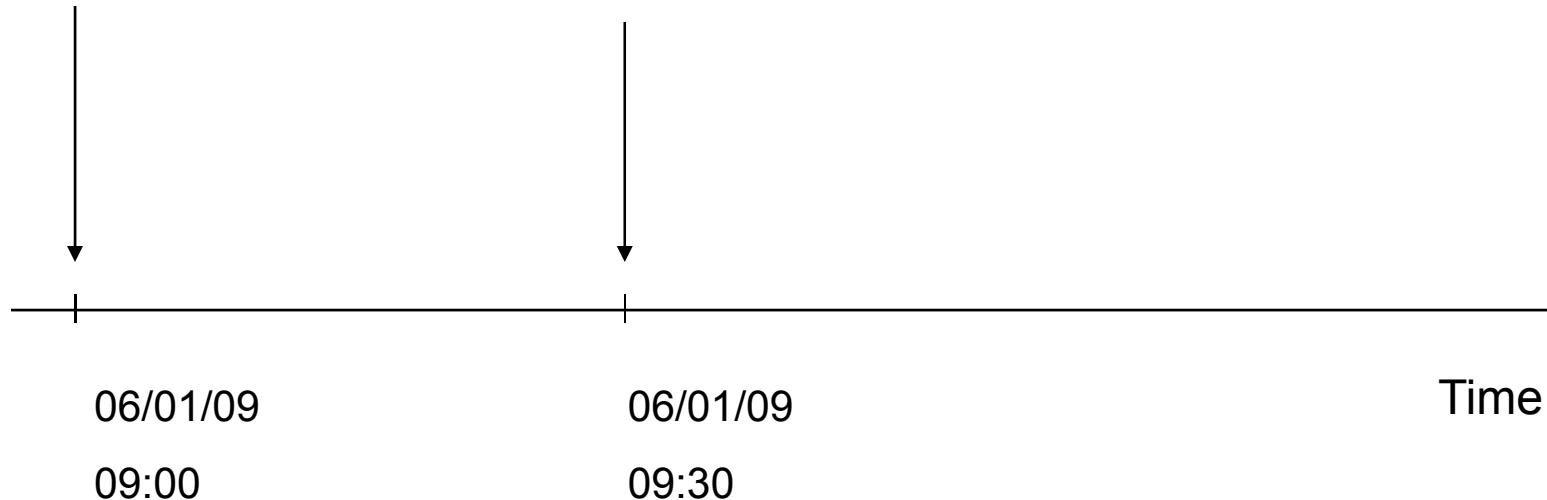


Introduction

Even harder example: x_i is spatial data



1. How do we determine if the data is “different”?
2. How can we figure out what makes the data different?



Introduction

Goals of event detection:

- Identify if an event of interest has occurred
- Characterize the event
 - Pinpoint the affected subgroup of the data ie. what features describe the event (eg. spatial area, time duration)?
 - What is the severity/magnitude of the event?
- Detect as accurately as possible
- Detect as early as possible

Introduction

How is event detection different from:

1. **Supervised Learning:**

- Abnormal events are extremely rare, normal events are plentiful

2. **Clustering:**

- Clustering = partitioning data into groups
- Not the same as finding statistically anomalous groups

3. **Outlier Detection:**

- Events of interest are usually not individual outliers
- The event typically affects a subgroup of the data rather than a single data point

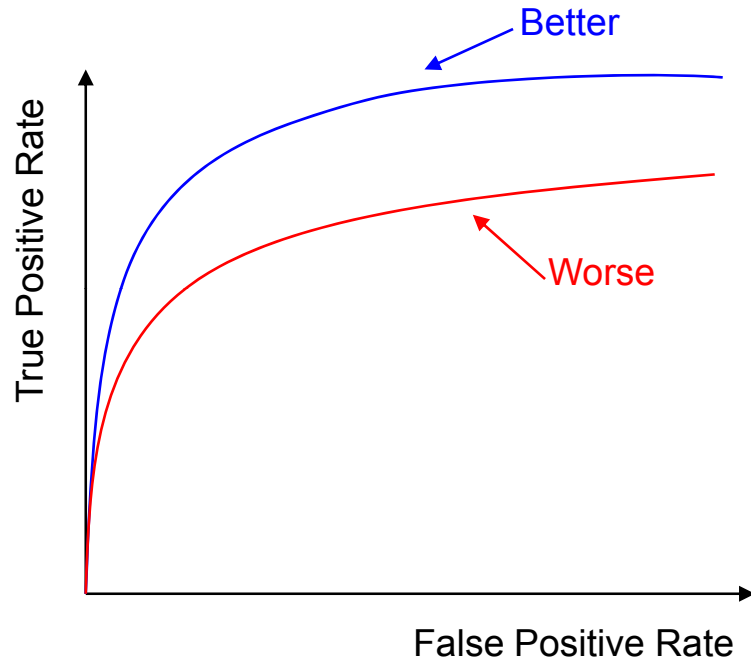
Introduction

How do we evaluate event detection algorithms?

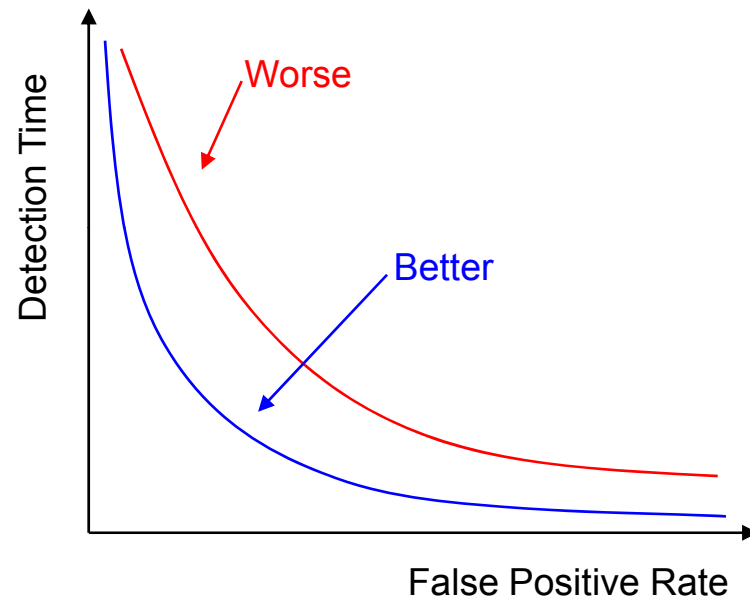
- Can't use prediction accuracy for "event" vs "non-event"
 - Class imbalance: many more "non-events" than "events"
 - Guessing "non-event" all the time results in very good accuracy
- Most event detection algorithms have a tunable threshold for when an alarm is raised
 - Trades off accuracy and false alarm rate
 - Need performance over multiple thresholds

Introduction

How do we evaluate event detection algorithms?



To evaluate accuracy, use a Receiver Operating Characteristic (ROC) curve



To evaluate timeliness of detection, use an Activity Monitoring Operating Characteristic (AMOC) curve (Fawcett and Provost 1999)

Introduction

Challenges:

- Incorporating spatial and/or temporal information
- Integrating information from multiple features or data streams
- Distinguishing between multiple event types
- Computational complexity

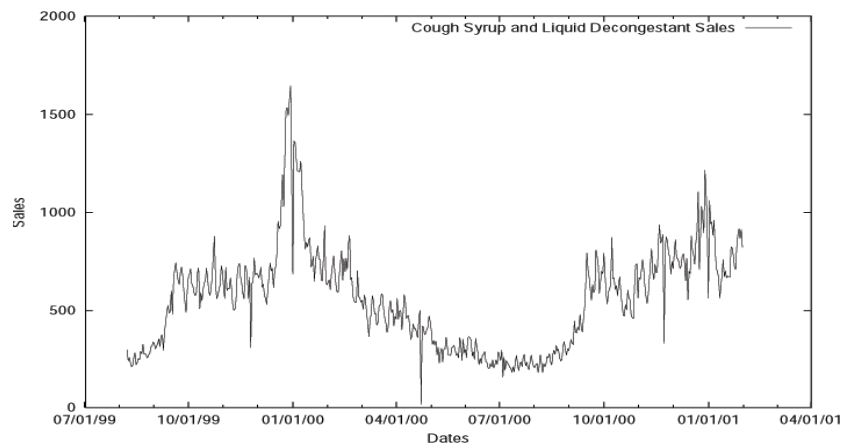
Outline

1. Introduction
- ➔ 2. Temporal Event Detection
3. Spatio-Temporal Event Detection
4. Future Work

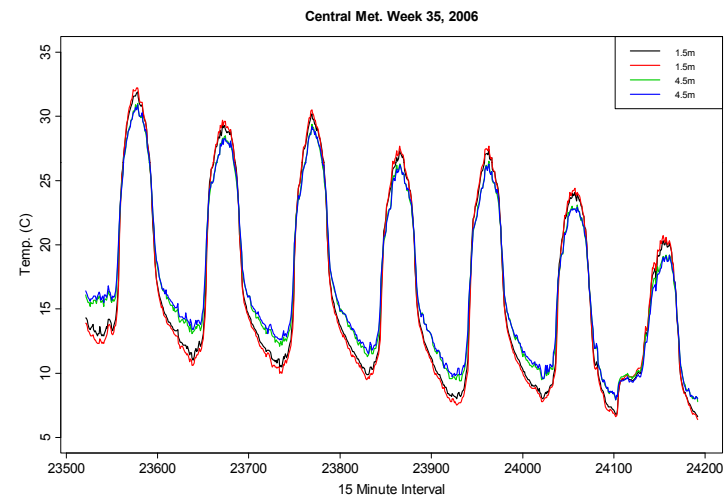
Univariate Temporal Methods

Univariate Methods

Examples of univariate time series



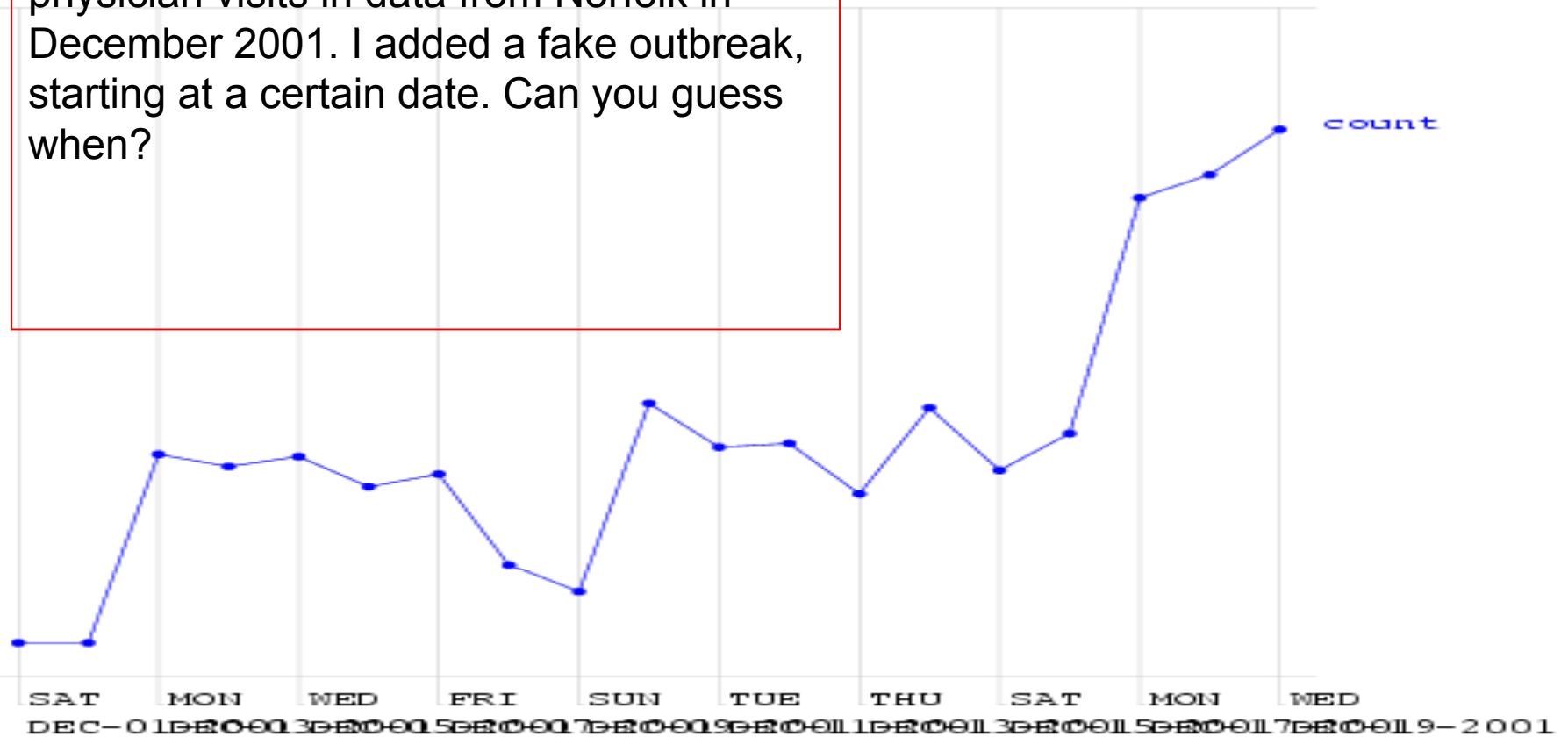
From: Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002). **Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales.** *Proceedings of the National Academy of Sciences* (pp. 5237-5249)



From: Dereszynski, E., Dieterich, T. (2007). **Probabilistic Models for Anomaly Detection in Remote Sensor Data Streams.** *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007)*. 75-82.

Univariate Methods

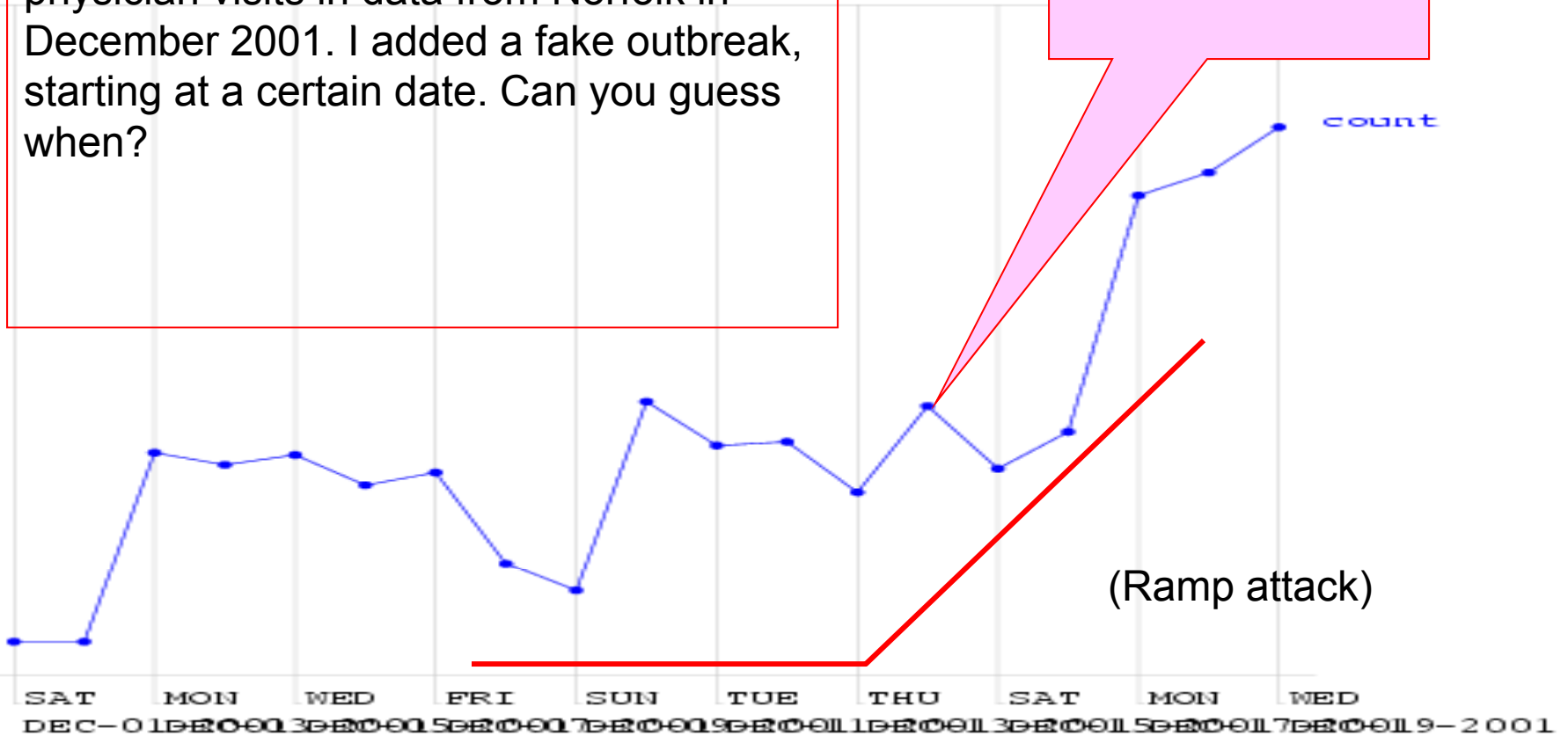
This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?



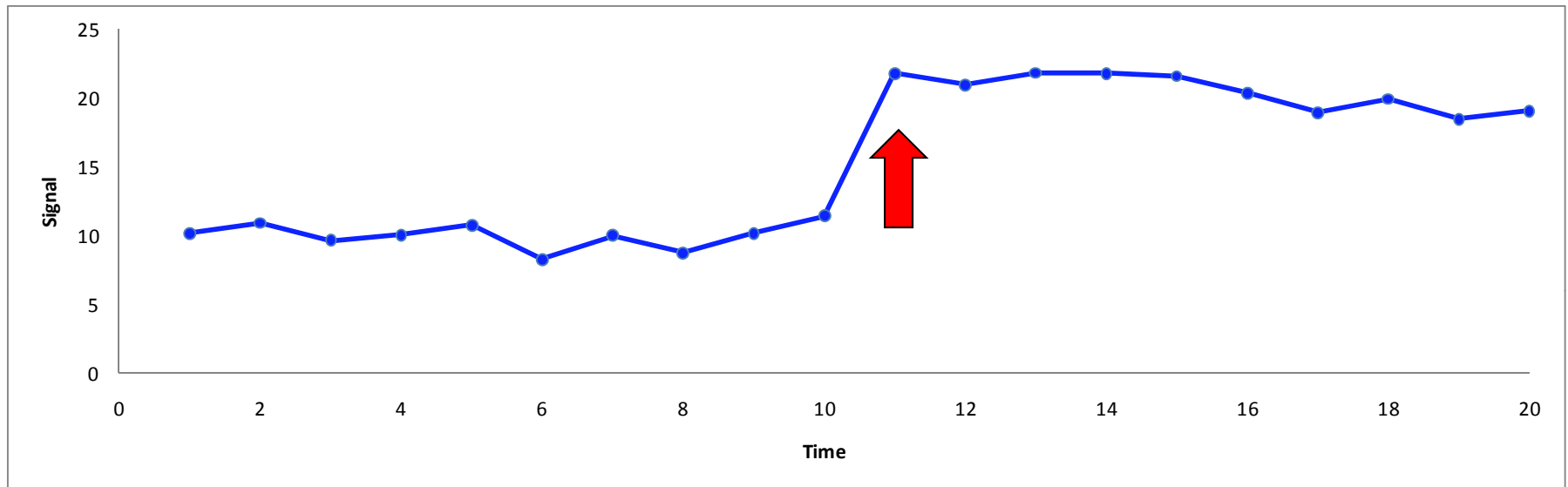
Univariate Methods

This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?

Here (much too high for a Friday)

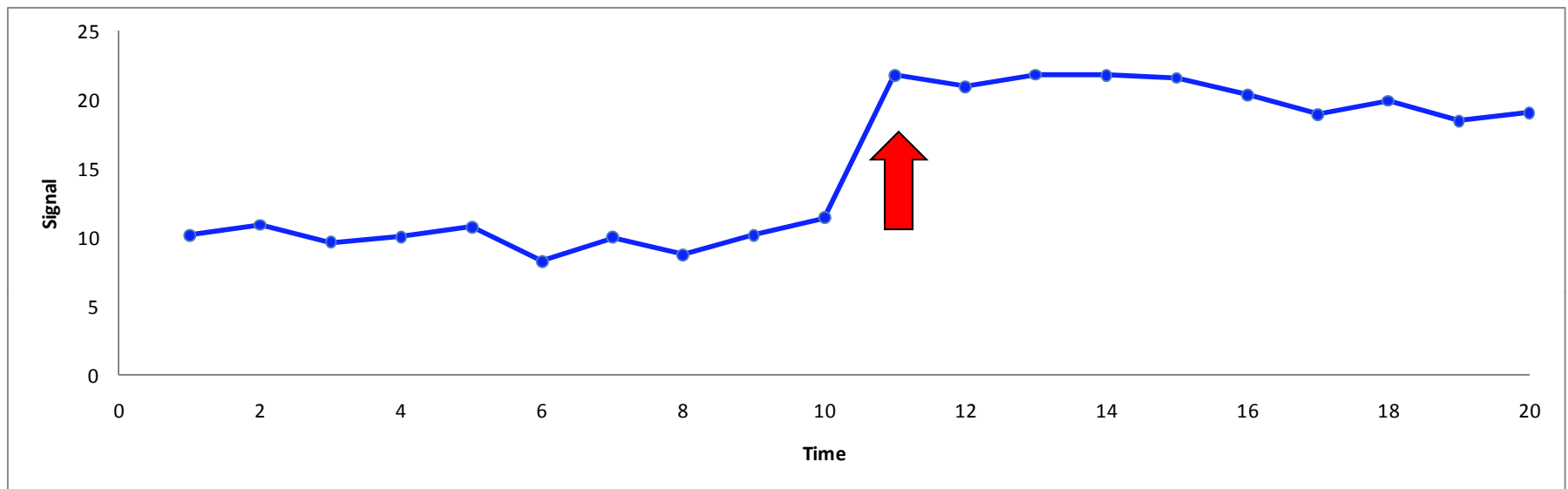


Univariate Methods



- Easy case: when does an “event” happen?
- How can we detect this with an algorithm?

Univariate Methods



General framework:

1. Learn model to predict *expected* signal value
2. Measure difference between *actual* and *expected*
3. Compute alarm value

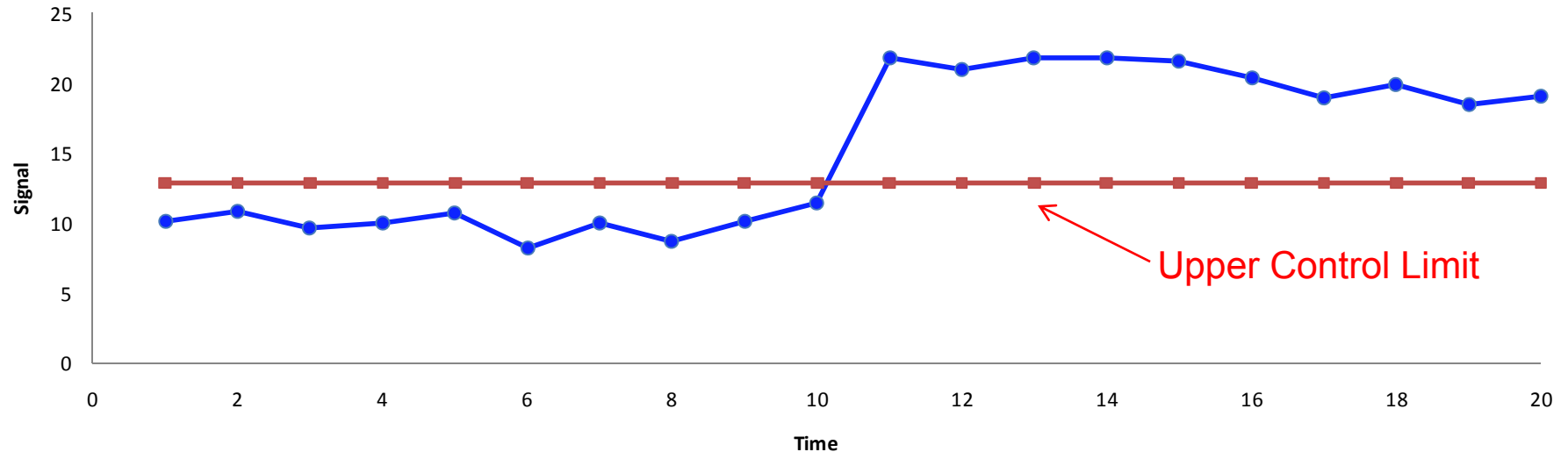
Univariate Methods

Methods we will discuss

- Control Chart (Shewhart 1931)
- Moving Average
- Exponentially Weighted Moving Average (Roberts 1959)
- CUSUM (Page 1954)
- Regression

For a reference on Statistical Quality Control techniques such as control charts, EWMA and CUSUM, see (Montgomery 2001)

Univariate Methods (Control Chart)



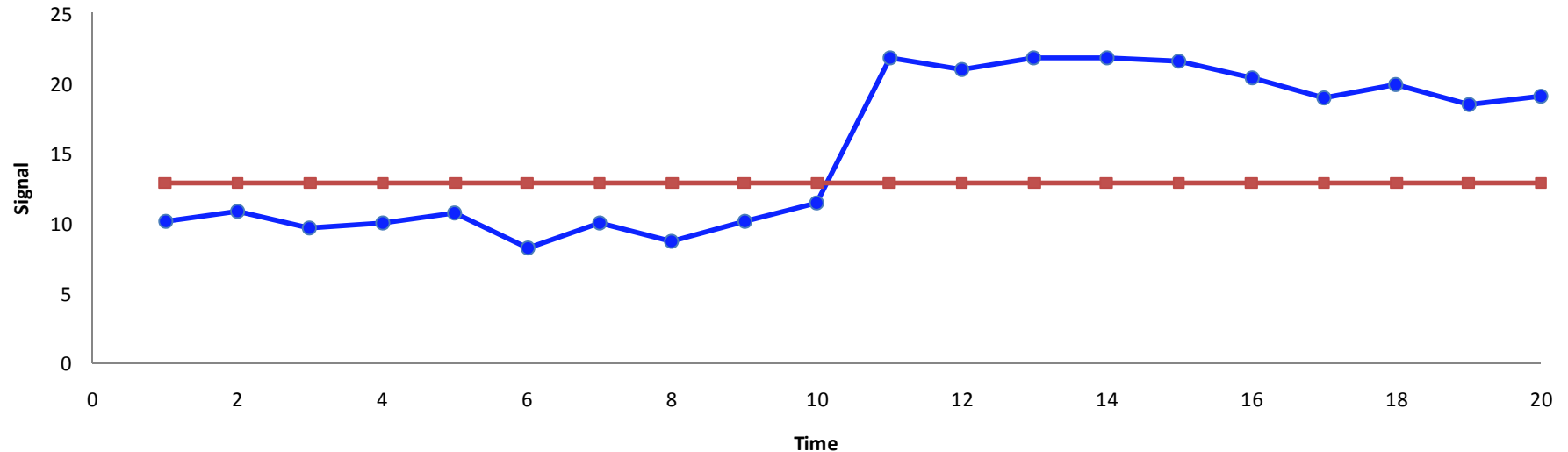
Control chart (from Statistical Quality Control)

- Estimate $\hat{\mu}$ and $\hat{\sigma}$ from data up to current time

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i \quad \hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \hat{\mu})^2}$$

- **Upper control limit** = $\hat{\mu} + 3\hat{\sigma}$
- Raise alarm if upper control limit exceeded

Univariate Methods (Control Chart)



Control chart (from Statistical Quality Control)

- Alternately, use

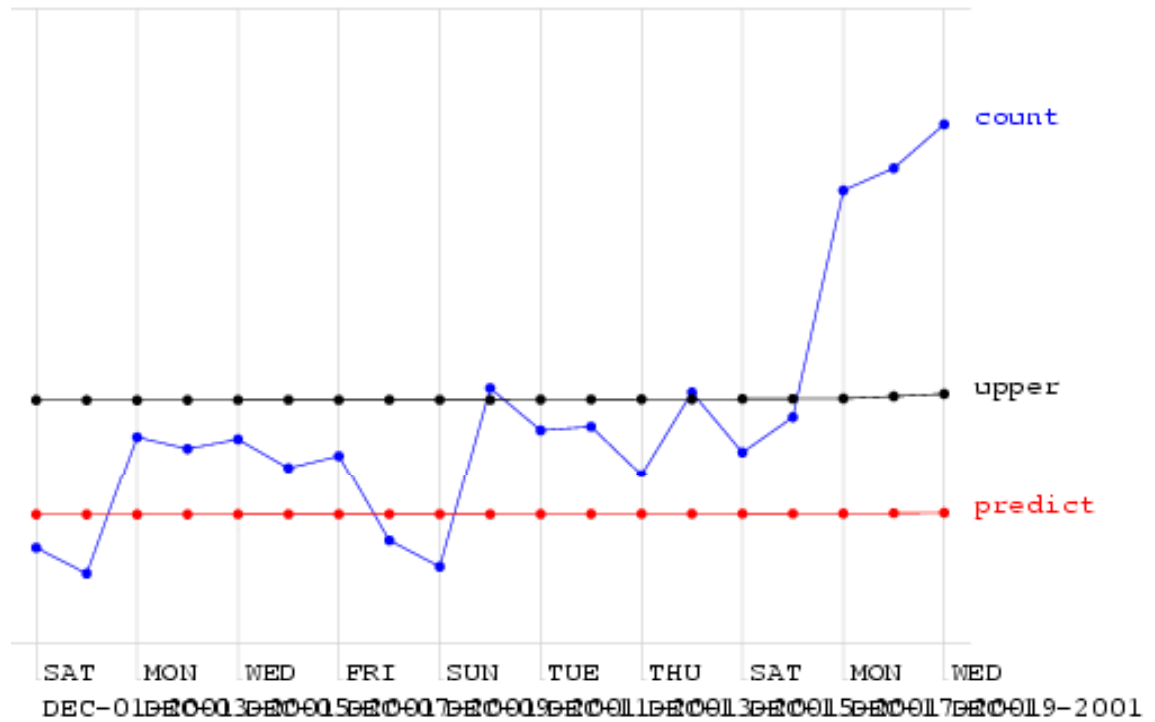
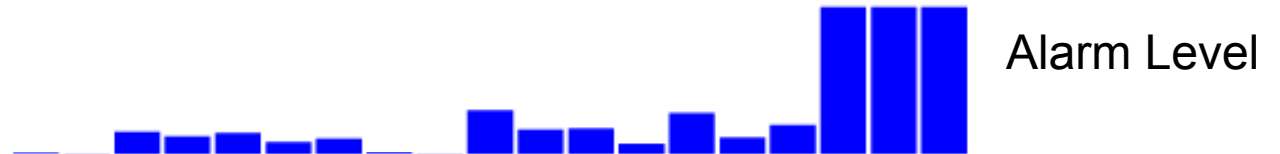
$$\text{Alarm level} = \Phi\left(\frac{\max(0, X_i - \hat{\mu})}{\hat{\sigma}}\right) \quad \text{where } \Phi = \text{CDF for } N(0,1)$$

- And signal alarm when alarm level > threshold

Univariate Methods (Control Chart)

Control chart
applied to Norfolk
data

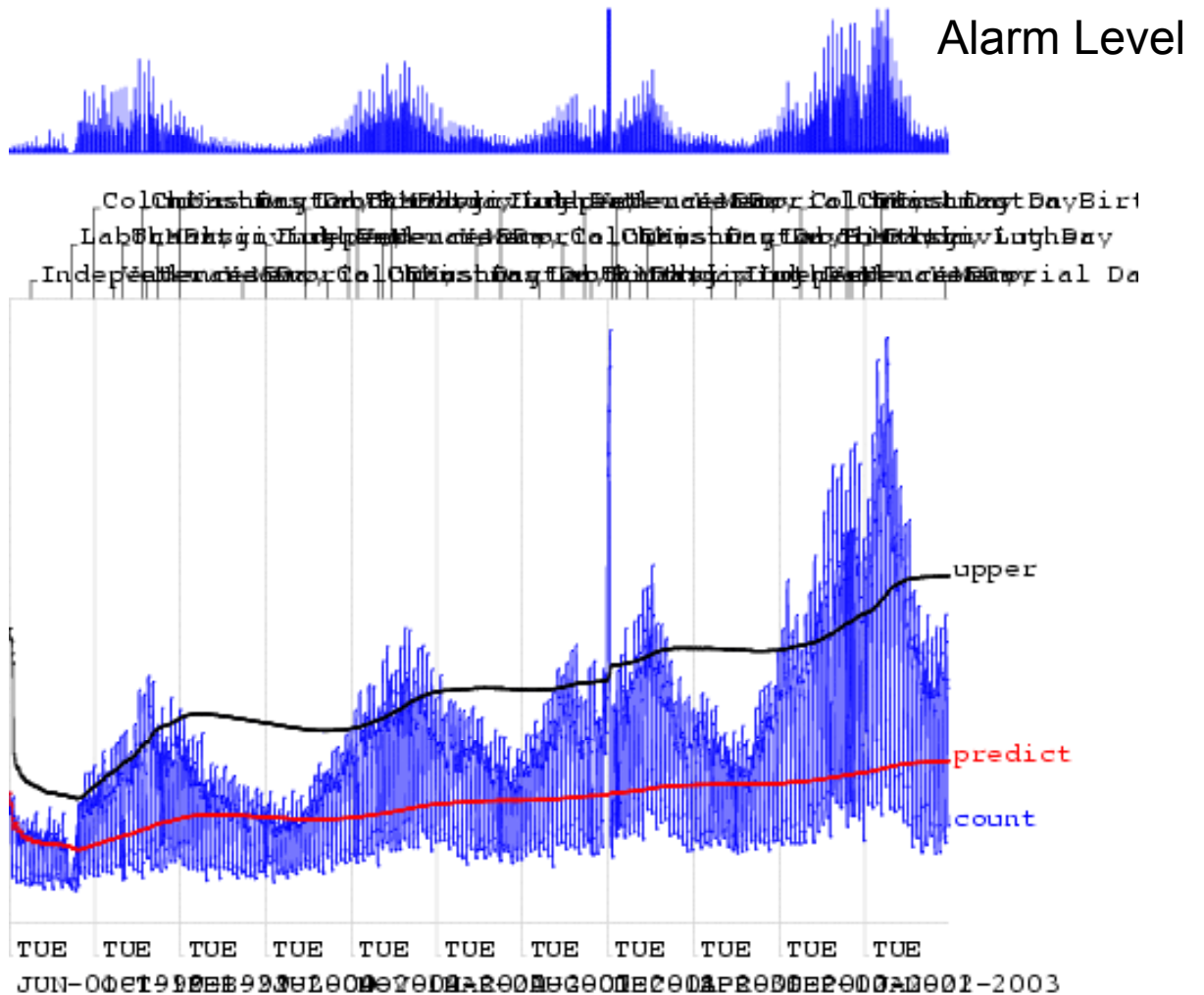
Bus stop demands: nr:=10



Univariate Methods (Control Chart)

Control chart applied to Norfolk data (long term)

Bus stop demands: nr:=10



Univariate Methods (Moving Average)

- Let W be the window size
- A moving average window predicts the following:

$$X_{t+1} = \frac{1}{W} (X_t + X_{t-1} + \dots + X_{t-W+1})$$

Setting the alarm value:

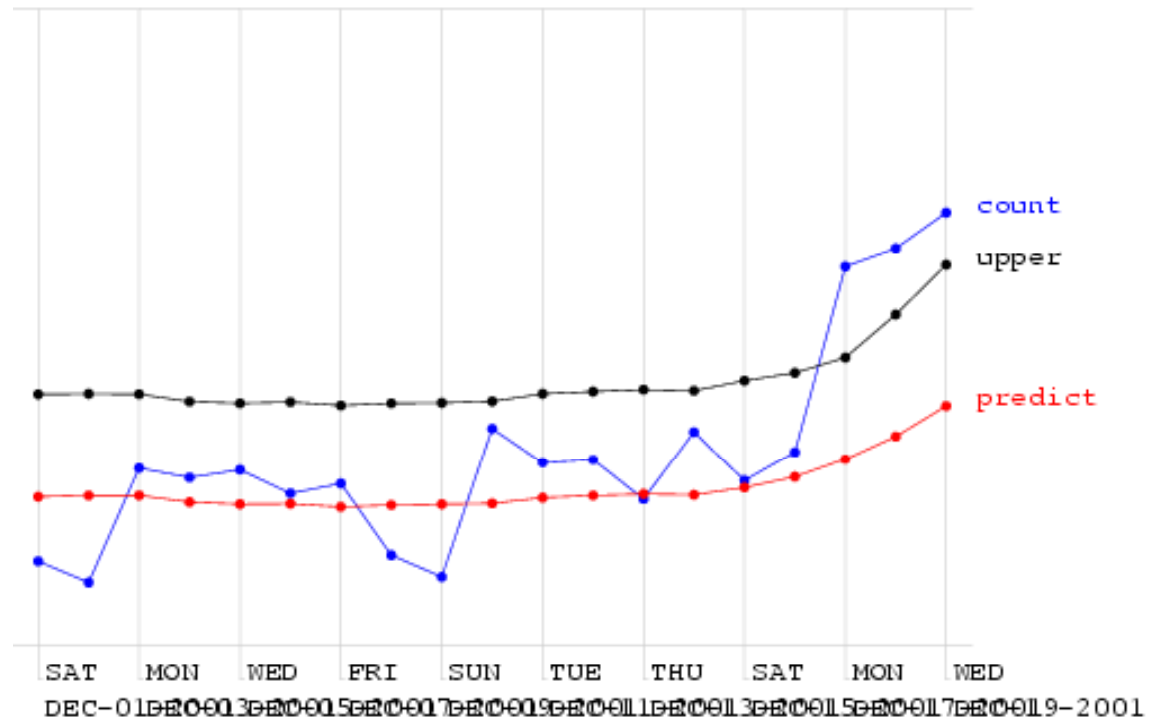
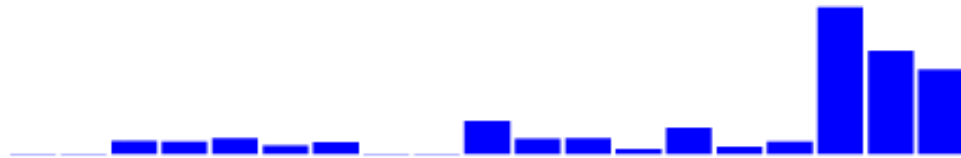
- Fit a Gaussian to the W observations within the window ie. estimate $\hat{\mu}$ and $\hat{\sigma}$
- Calculate the alarm level as before

$$\text{Alarm level} = \Phi\left(\frac{\max(0, X_i - \hat{\mu})}{\hat{\sigma}}\right)$$

Univariate Methods (Moving Average)

Moving Average applied to Norfolk data

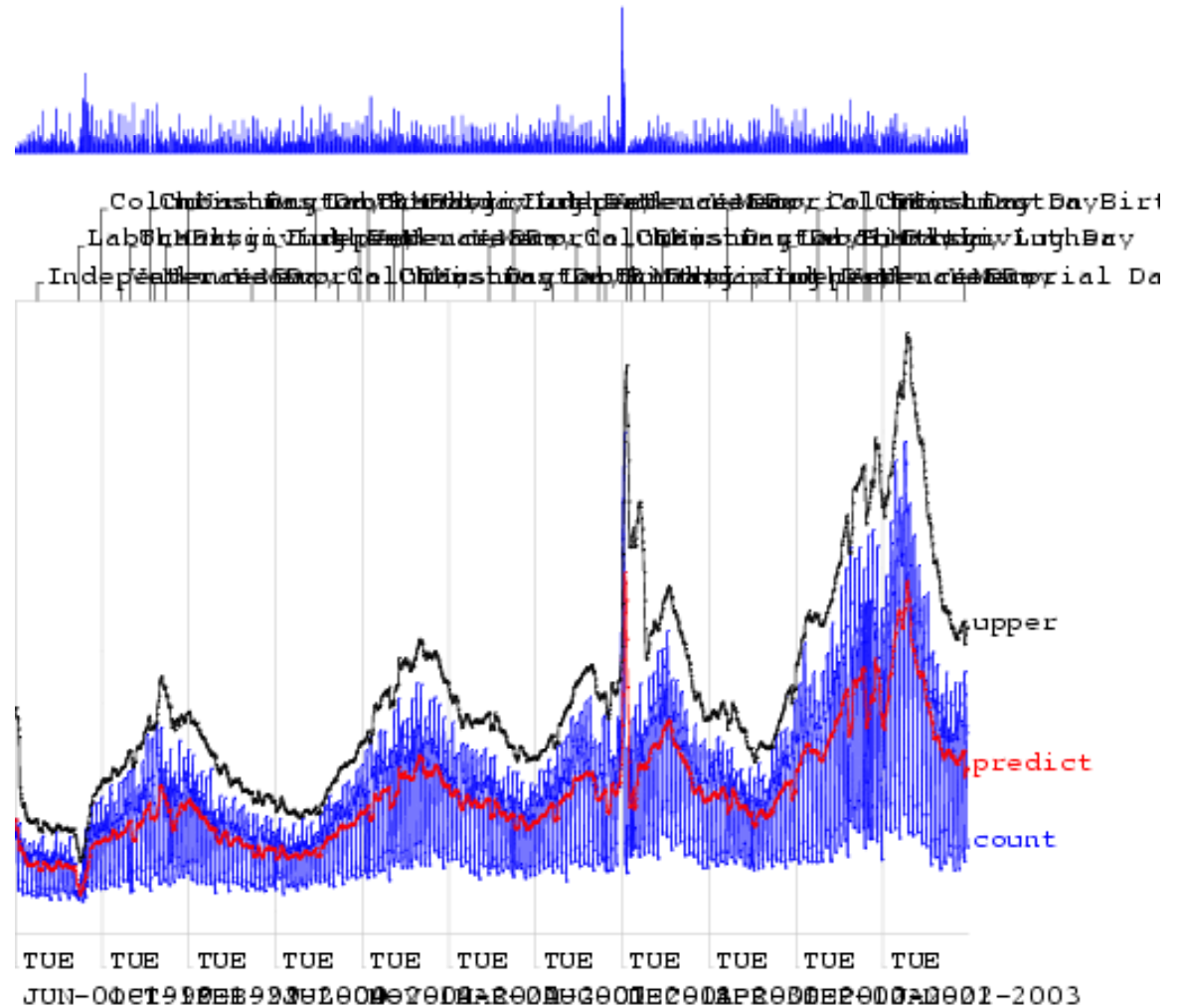
Bus stop demands: nr:=7,3807



Univariate Methods (Moving Average)

Moving Average applied to Norfolk data (long term)

Bus stop downloads: nr:=73807



Univariate Methods (EWMA)

- Exponentially Weighted Moving Average (EWMA) - a variation on the moving average
- Let Z_i be the EWMA statistic (which is monitored):

$$Z_i = \lambda X_i + (1 - \lambda) Z_{i-1} \quad \text{where } 0 < \lambda \leq 1$$



Observations in the past receive a decreasing amount of weight

Univariate Methods (CUSUM)

- Cumulative SUM Statistics
- Good at detecting shifts from the mean more quickly than control chart
- Keep a running sum of “surprises”: a sum of excesses each day over the mean
- When this sum exceeds threshold H , signal alarm and reset sum

Univariate Methods (CUSUM)

- r = reference value eg. mean
- X_i = i th observation
- S_i = i th cumulative sum

$$S_1 = X_1 - r$$

$$S_2 = (X_2 - r) + (X_1 - r) = (X_2 - r) + S_1$$

⋮

$$S_k = \sum_{i=1}^k (X_i - r) + S_{k-1}$$

When a shift from the mean occurs, S_i will start to increase

Univariate Methods (CUSUM)

- If we are only tracking increases, we can do the following:

$$S_k = \max(0, (X_k - r) + S_{k-1})$$

↑
Ensures we don't go below 0

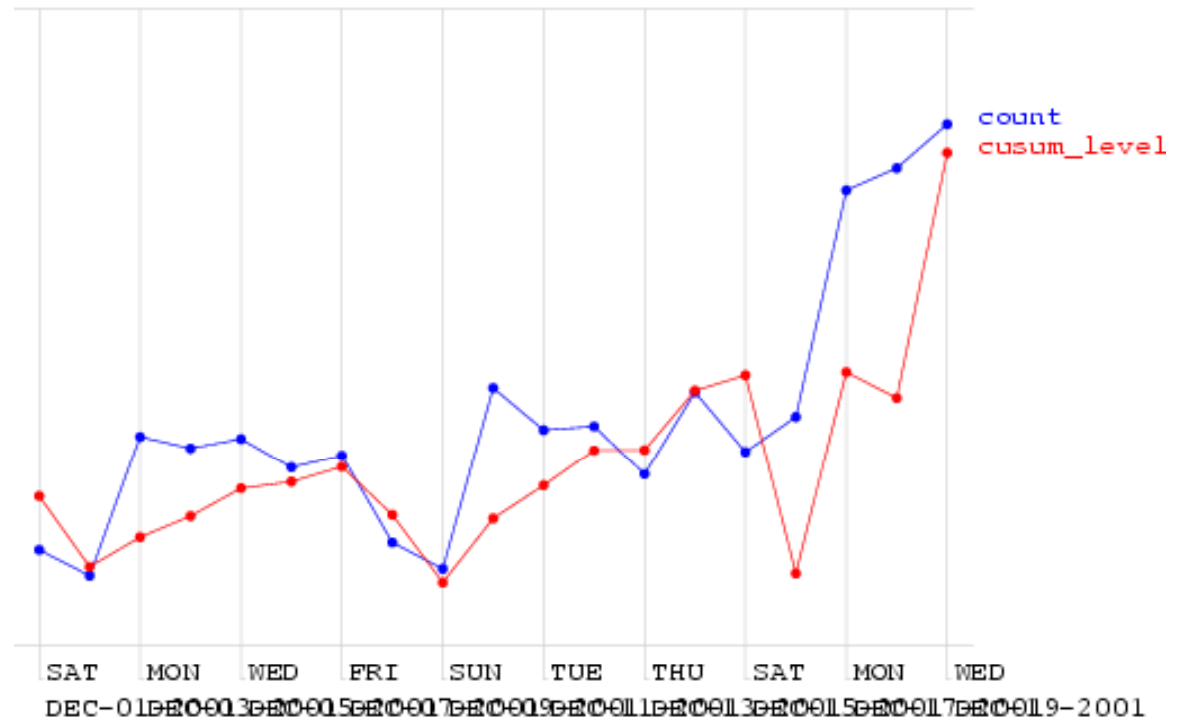
- We can also add a tolerance or a slack K

$$S_k = \max(0, X_k - (r + K) + S_{k-1})$$

Univariate Methods (CUSUM)

Bus stop demands: nr=1

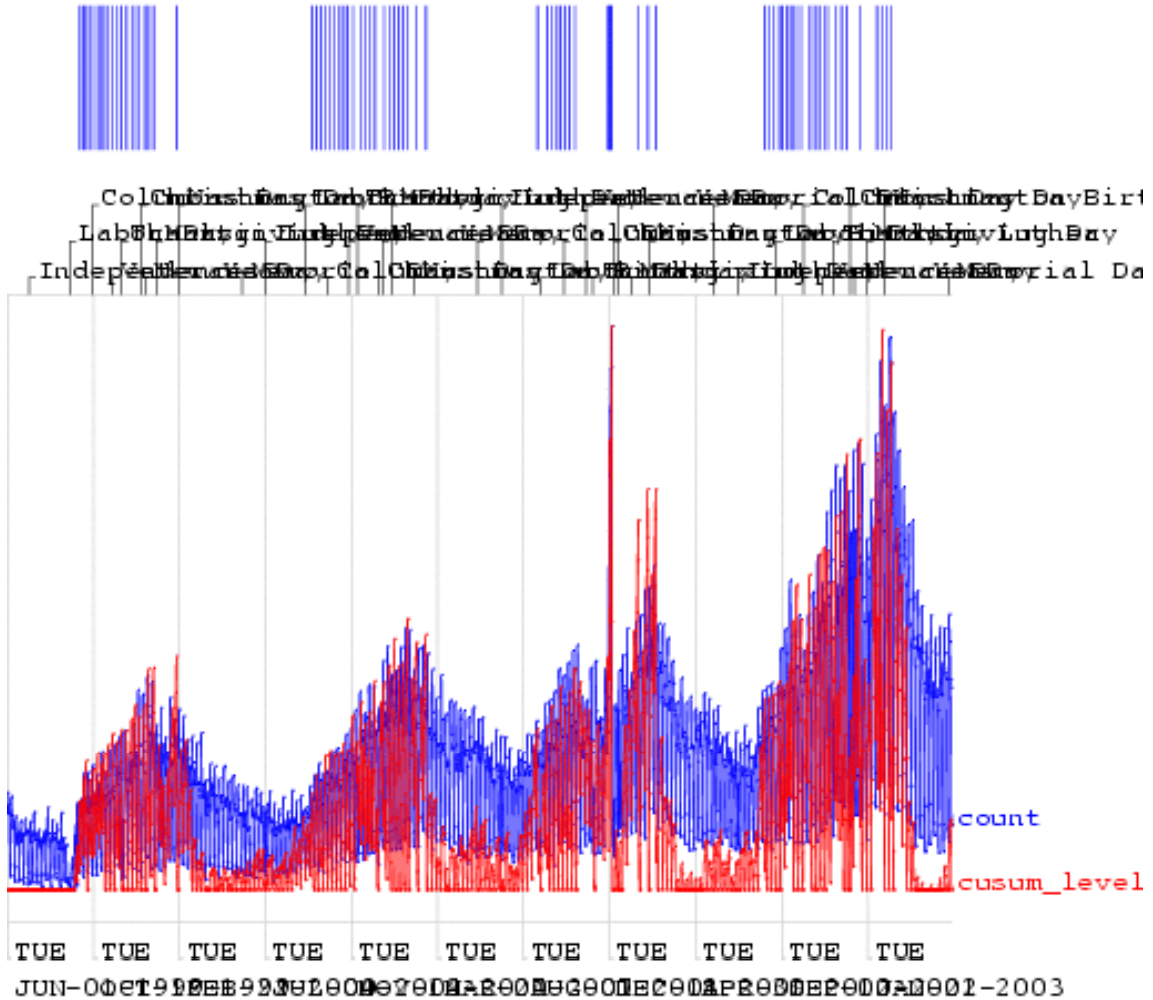
CUSUM applied
to Norfolk data



Univariate Methods (CUSUM)

CUSUM applied
to Norfolk data
(long term)

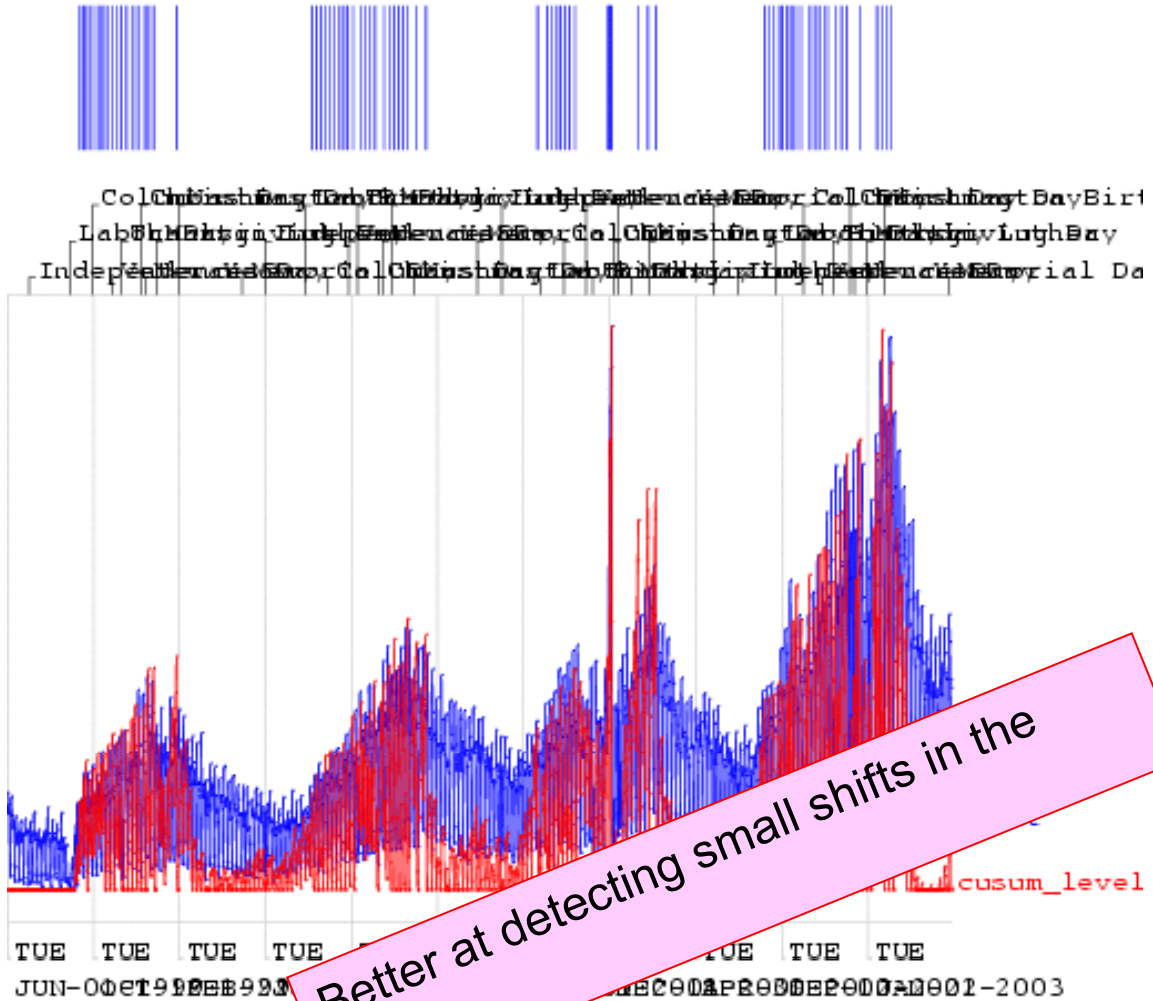
Bus to docks: nr=1



Univariate Methods (CUSUM)

CUSUM applied
to Norfolk data
(long term)

Bus to docks: nr=1



Univariate Methods

- Data often consists of trends eg.
 - Seasonal effect
 - Day-of-week effects
 - Holiday effect
- None of the methods discussed so far explicitly model these trends
- Regression (eg. linear regression) can be used with extra terms for the trends

Univariate Methods (Regression)

Regression example to model seasonal effects and Monday effects:

$$Y_i = \beta_0 + \beta_1(\text{HoursOfDaylight}_i) + \beta_2(\text{IsMonday}_i) + \varepsilon_i$$

Could be defined as:

$$\sin\left(\frac{2\pi(\text{num days since July 31})}{365.25} - \frac{\pi}{2}\right)$$

Boolean feature – adds a “bump” to the value of Y if it is a Monday

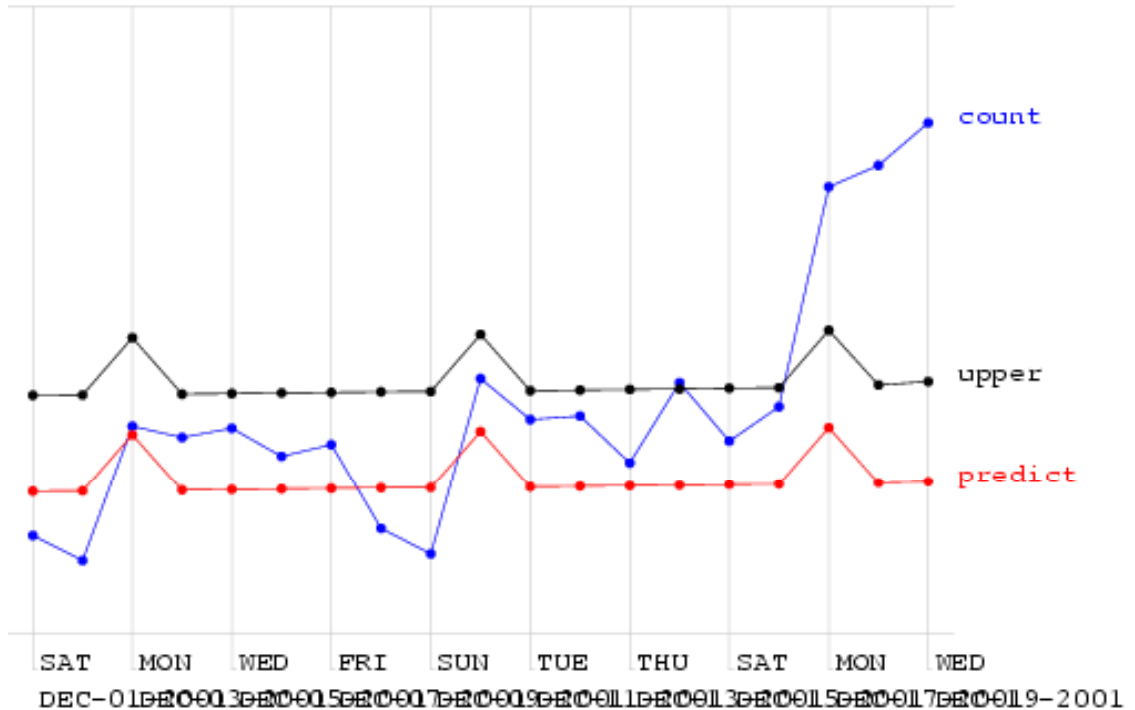
Normally distributed noise with mean 0, known variance σ^2

Regression learns the β parameters from data to minimize the residual sum of squares

Univariate Methods (Regression)

Bus stop downloads: nr=10

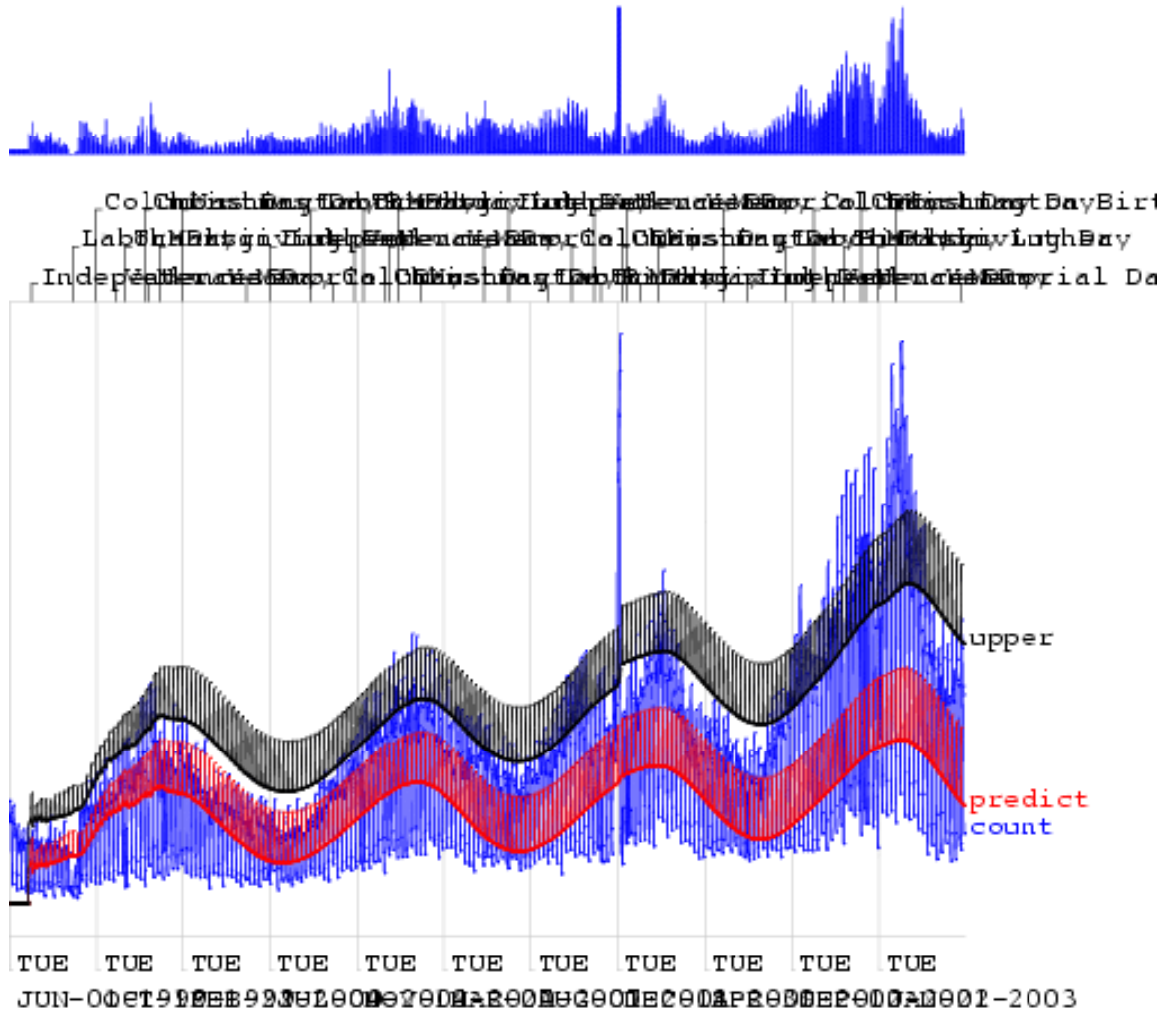
Regression applied to Norfolk data using *HoursOfDaylight* and *IsMonday* terms



Univariate Methods (Regression)

Regression applied to Norfolk data using *HoursOfDaylight* and *IsMonday* terms (long term)

Bus stop downloads: nr:=10



Univariate Methods (Other)

Other state-of-the-art methods not discussed in this tutorial

- Box-Jenkins models eg. ARMA, ARIMA
- Wavelets
- Change-point detection
- Kalman filters
- Hidden Markov Models

Multivariate Temporal Methods

Multivariate Methods

Each data point (recorded at some time point) is now a multivariate vector
eg. patient records from an Emergency Department

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/1/09	9:12	M	20s	Fever	NE	NE	...
6/1/09	10:45	F	40s	Diarrhea	NE	NE	...
6/1/09	11:03	F	60s	Respiratory	NE	N	...
6/1/09	11:07	M	60s	Diarrhea	E	W	...
:	:	:	:	:	:	:	:

Multivariate Methods

How are patient records from 6/1/09 different

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/1/09	9:12	M	20s	Fever	NE	NE	...
6/1/09	10:45	F	40s	Diarrhea	NE	NE	...
6/1/09	11:03	F	60s	Respiratory	NE	N	...
:	:	:	:	:	:	:	:

from patient records from 6/2/09?

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/2/09	9:15	M	60s	Respiratory	E	NE	...
6/2/09	10:01	F	50s	Respiratory	N	NW	...
6/2/09	13:05	F	40s	Respiratory	SW	SW	...
:	:	:	:	:	:	:	:

Multivariate Methods

Note: need to split data into two groups according to time:

1. **Training**: used to learn model

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/1/09	9:12	M	20s	Fever	NE	NE	...
6/1/09	10:45	F	40s	Diarrhea	NE	NE	...
6/1/09	11:03	F	60s	Respiratory	NE	N	...
:	:	:	:	:	:	:	:

2. **Testing**: used to identify events with respect to the learned model

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/2/09	9:15	M	60s	Respiratory	E	NE	...
6/2/09	10:01	F	50s	Respiratory	N	NW	...
6/2/09	13:05	F	40s	Respiratory	SW	SW	...
:	:	:	:	:	:	:	:

Multivariate Methods

We make the following distinction

- **Multivariate Changepoint Detection:**
 - Detects that a change has happened
 - Does not identify the subgroup of data that has changed the most
- **Multivariate Event Detection**
 - Detects that a change has happened
 - Identifies the subgroup that has changed the most

Multivariate Methods

Outline

- Multivariate Changepoint Detection
 - Multivariate Statistical Quality Control
 - Others
- Multivariate Event Detection
 - Emerging Patterns
 - STUCCO
 - WSARE 2.0
 - WSARE 3.0

Multivariate Changepoint Detection (Hotelling's T^2)

Multivariate version of control chart is Hotelling's T^2 statistic (Hotelling 1931)

$$T^2 = c(\mathbf{X}_i - \hat{\boldsymbol{\mu}})\hat{\mathbf{S}}^{-1}(\mathbf{X}_i - \hat{\boldsymbol{\mu}})$$

Sample size that covariance matrix was estimated from

Estimated covariance matrix

Estimated mean

Other multivariate statistical quality control methods:

- Multivariate CUSUM (Crosier 1988)
- Multivariate EWMA (Lowry et al. 1992)

All make strong assumptions about the underlying model

Multivariate Changepoint Detection

Other methods:

- Cross-match test (Rosenbaum 2005)
- kdq-tree (Dasu et al. 2006)
- Density Test (Song et al. 2007)

Multivariate Event Detection

General framework:

1. Learn model to predict *expected* signal value **for the given subgroup**
2. Measure difference between *actual* and *expected*
3. **Compute alarm value (now more involved)**

Multivariate Event Detection

Algorithm	Data	Model	Measuring Differences
Emerging Patterns (Dong and Li 1999)	Categorical	Counts	Increase in support ratio
STUCCO (Bay and Pazzani 1999)	Categorical	Counts	Chi-square, Bonferroni
WSARE 2.0 (Wong et al. 2005)	Categorical	Counts	Fisher's Exact test, Randomization test
WSARE 3.0 (Wong et al. 2005)	Categorical	Bayesian network	Fisher's Exact test, Randomization Test

Multivariate Event Detection (Categorical Data)

How can we find differences in multivariate categorical data?

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/2/09	9:15	M	60s	Respiratory	E	NE	...
6/2/09	10:01	F	50s	Respiratory	N	NW	...
6/2/09	13:05	F	40s	Respiratory	SW	SW	...
:	:	:	:	:	:	:	:

Idea from association rule mining:
Characterize differences by rules ie.
conjunctions of attribute-value pairs

Multivariate Event Detection (Categorical Data)

Training

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/1/09	9:12	M	20s	Fever	NE	NE	...
6/1/09	10:45	F	40s	Diarrhea	NE	NE	...
6/1/09	11:03	F	60s	Respiratory	NE	N	...
:	:	:	:	:	:	:	:

Testing

Date	Time	Gender	Age	Prodrome	Home Location	Work Location	Many more...
6/2/09	9:15	M	60s	Respiratory	E	NE	...
6/2/09	10:01	F	50s	Respiratory	N	NW	...
6/2/09	13:05	F	40s	Respiratory	SW	SW	...
:	:	:	:	:	:	:	:

Find which rules predict unusually high proportions in test data when compared to the training data eg.

92/180 records from Testing have *Gender = Male AND Age = 60s*

43/200 records from Training have *Gender = Male AND Age = 60s*

Multivariate Event Detection (Emerging Patterns)

- Let $\mathbf{D} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ be a data set with N data points
- Define the **support** of a rule R to be:

$$\text{supp}_D(R) = \frac{\text{count}_D(R)}{|D|}$$

where $\text{count}_D(R)$ = number of data points that match rule R

Multivariate Event Detection (Emerging Patterns)

Suppose we are given data sets $D1$ and $D2$.
Define the *GrowthRate*(R) from $D1$ to $D2$ as:

$$GrowthRate(R) = \begin{cases} 0, & \text{if } \text{supp}_{D1}(R) = 0 \text{ and } \text{supp}_{D2}(R) = 0 \\ \infty, & \text{if } \text{supp}_{D1}(R) = 0 \text{ and } \text{supp}_{D2}(R) \neq 0 \\ \frac{\text{supp}_{D2}(R)}{\text{supp}_{D1}(R)}, & \text{otherwise} \end{cases}$$

Multivariate Event Detection (Emerging Patterns)

- Given $\rho > 1$, a rule R is said to be a ρ -emerging pattern from $D1$ to $D2$ if

$$GrowthRate(R) \geq \rho$$

- Goal: For a given ρ , find all ρ -emerging patterns

See (Dong and Li 1999) for efficient algorithms to find emerging patterns using borders to describe large collections of itemsets

Search and Testing for Understandable Consistent Contrasts (STUCCO)

Multivariate Methods (STUCCO)

- Define a **contrast set** as a conjunction of attribute-value pairs (ie. what we defined as a *rule*)
- Search for contrast sets CS such that:
 1. $P(CS | Training) \neq P(CS | Testing)$

Multivariate Methods (STUCCO)

- Define a **contrast set** as a conjunction of attribute-value pairs
- Search for contrast sets CS such that:
 1. $P(CS | Training) \neq P(CS | Testing)$



This says that the distribution of the contrast set CS is different in the Training and Testing data. “Different” will be defined shortly.

Multivariate Methods (STUCCO)

- Define a **contrast set** as a conjunction of attribute-value pairs
- Search for contrast sets CS such that:
 1. $P(CS | Training) \neq P(CS | Testing)$
 2. $|Support(CS, Training) - Support(CS, Testing)| \geq \delta$

Multivariate Methods (STUCCO)

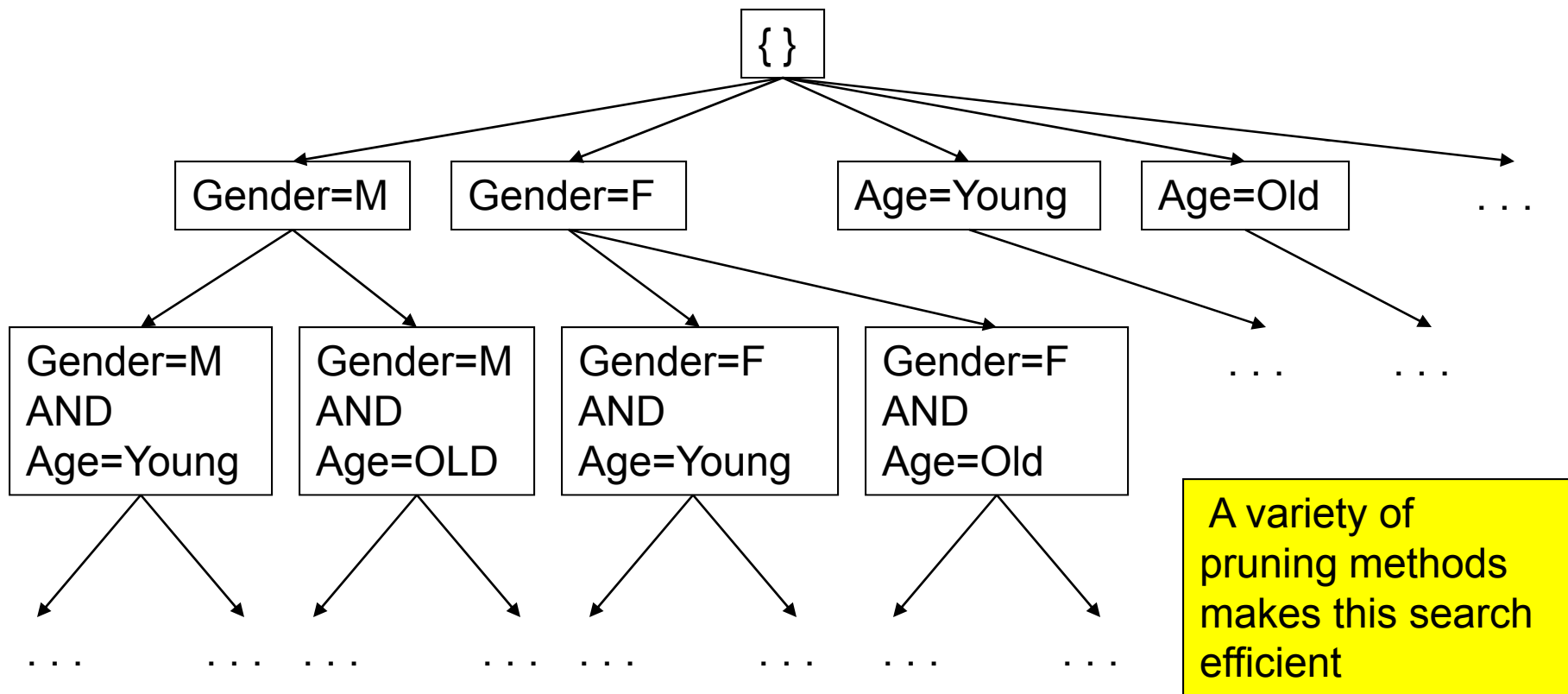
- Define a **contrast set** as a conjunction of attribute-value pairs
- Search for contrast sets CS such that:
 1. $P(CS | Training) \neq P(CS | Testing)$
 2. $|Support(CS, Training) - Support(CS, Testing)| \geq \delta$



The support of a contrast set is the percentage of data points (in the Training / Testing data) where the contrast set is true

Multivariate Methods (STUCCO)

Search for contrast sets involves efficient breadth-first search of a set enumeration tree (Rymon 1992)



Multivariate Methods (STUCCO)

- How do we determine that the distributions of contrast sets are different between training and testing?
- For each contrast set, construct a 2x2 contingency table

	Count (Training)	Count (Testing)
Age = Young	25	52
Age \neq Young	101	206

Multivariate Methods (STUCCO)

- Perform Chi-Square test of independence

- Compute χ^2 statistic:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Where O_{ij} = observed frequency count for the cell in row i and column j
- E_{ij} is the expected frequency count for the cell in row i and column j given independence of the row and column variables ie.

$$E_{ij} = \left(\sum_{i=1}^2 O_{ij} \sum_{j=1}^2 O_{ij} \right) / N$$

Where N = total number of observations in all cells

- To obtain a p-value, compare χ^2 statistic to a chi-square distribution

Multivariate Methods (STUCCO)

- If we perform a Chi-Square test on the 2x2 contingency table below, we get a p-value of 0.0074
- This is significant at the $\alpha = 0.05$ level
- Report contrast sets with p-value < 0.05

	Count (Training)	Count (Testing)
Age = Young	25	100
Age \neq Young	101	206

Multivariate Methods (STUCCO)

- But...we can't interpret this p-value at face value
- The search suffers from a multiple hypothesis testing problem
- Need to correct the p-values to compensate for multiple hypothesis testing

Multivariate Methods (STUCCO)

Multiple Hypothesis Testing

- Suppose we reject null hypothesis when p-value $< \alpha$, where $\alpha = 0.05$
- For a single hypothesis test, the probability of a false positive = α
- Suppose we do 1000 tests, one for each possible rule
- Probability(false positive) could be as bad as:
$$1 - (1 - 0.05)^{1000} \gg 0.05$$

Multivariate Methods (STUCCO)

Bonferroni Correction:

- If you are performing n hypothesis tests for hypotheses h_1, \dots, h_n ,
 - Adjust significance level for test i to be

$$\alpha_i = \frac{\alpha}{n}$$

- Reject hypothesis h_i if

$$\text{pvalue}_i \leq \alpha_i$$

Multivariate Methods (STUCCO)

Two problems:

1. We do not know n as we incrementally mine each level of the tree
2. Same cutoff for contrast sets with different numbers of attribute-value pairs (want more power on smaller conjunctions)

STUCCO's solution

$$\alpha_l = \min\left(\frac{\alpha}{2^l / |C_l|}, \alpha_{l-1}\right)$$

Significance threshold at level l of the tree

Number of candidates at level l

Multivariate Methods (STUCCO)

Summary

Traverse set-enumeration tree using Breadth-First Search

For each contrast set at depth l of tree:

1. Form 2x2 contingency table
2. Compute p-value using chi-square test
3. Account for multiple hypothesis testing by computing significance level α_1
4. If p-value $< \alpha_1$, report contrast set

Multivariate Methods (WSARE 2.0)

Summary

For each rule in rule set:

1. Form 2x2 contingency table
2. Compute rule score using Fisher's Exact test / Chi-square test
3. Account for multiple hypothesis testing by randomization test
4. If p-value from randomization test $<$ alarm value, report rule

Multivariate Methods (WSARE 2.0)

Summary

For each rule in rule set:

1. Form 2x2 contingency table
2. Compute rule score using Fisher's Exact test / Chi-square test
3. Account for multiple hypothesis testing by randomization test
4. If p-value from randomization test $<$ alarm value, report rule

Difference #1

Difference #2

Difference #3

Multivariate Methods (WSARE 2.0)

Difference #1 (from STUCCO)

- Rule set defined as all rules with a maximum of k conjunctions of attribute-value pairs
- No need to search set-enumeration tree

Multivariate Methods (WSARE 2.0)

Difference #2 (from STUCCO)

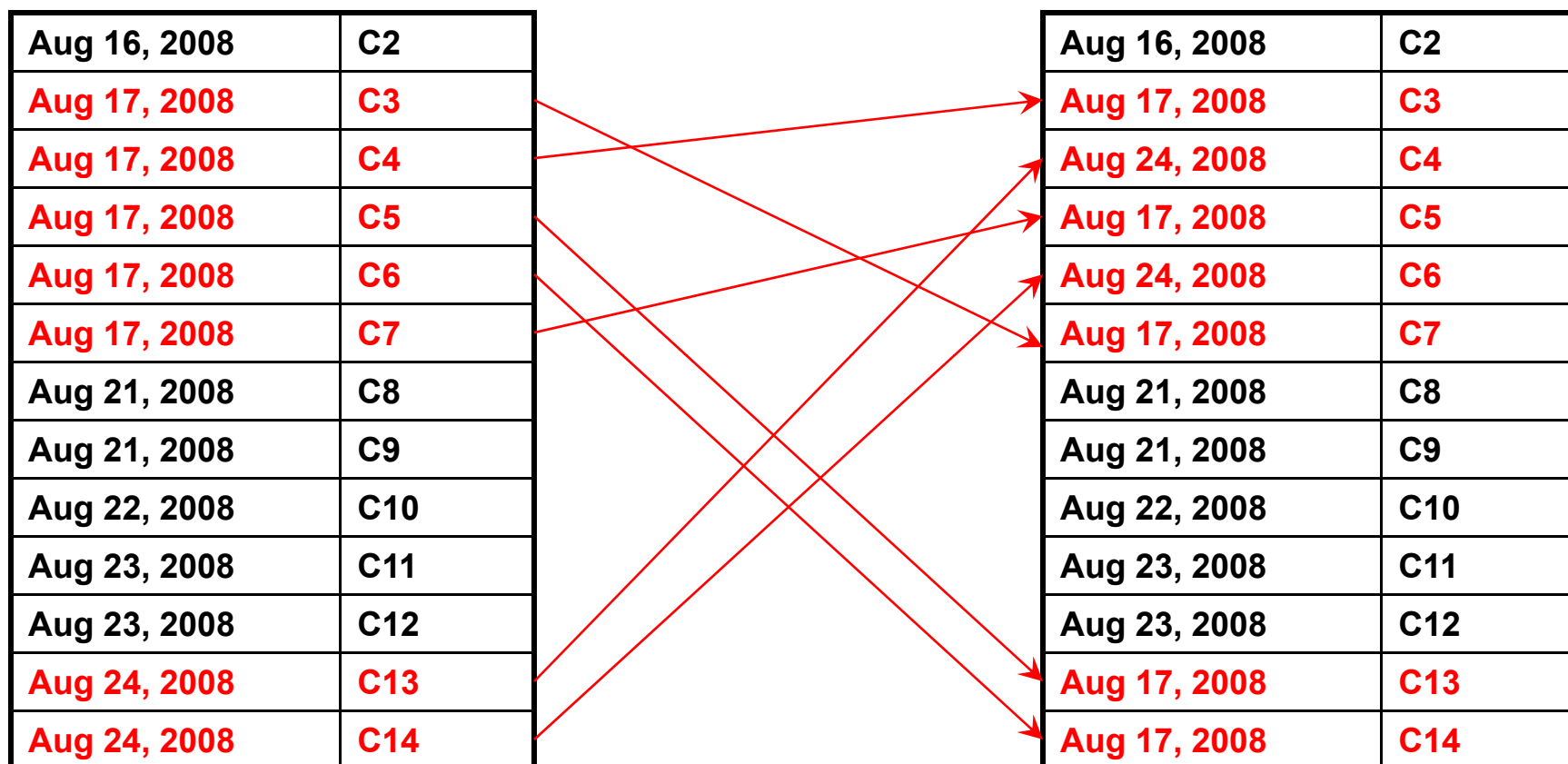
- Use of Fisher's Exact Test for rules with small counts that violate assumptions of Chi-square test

Multivariate Methods (WSARE 2.0)

Difference #3 (from STUCCO)

- Bonferroni correction is very conservative
- Increases risk of Type II errors (not rejecting null hypothesis when it is false)
- Very hesitant to declare something as an “event”
- Randomization test is a better alternative with more statistical power

Multivariate Methods (WSARE 2.0)

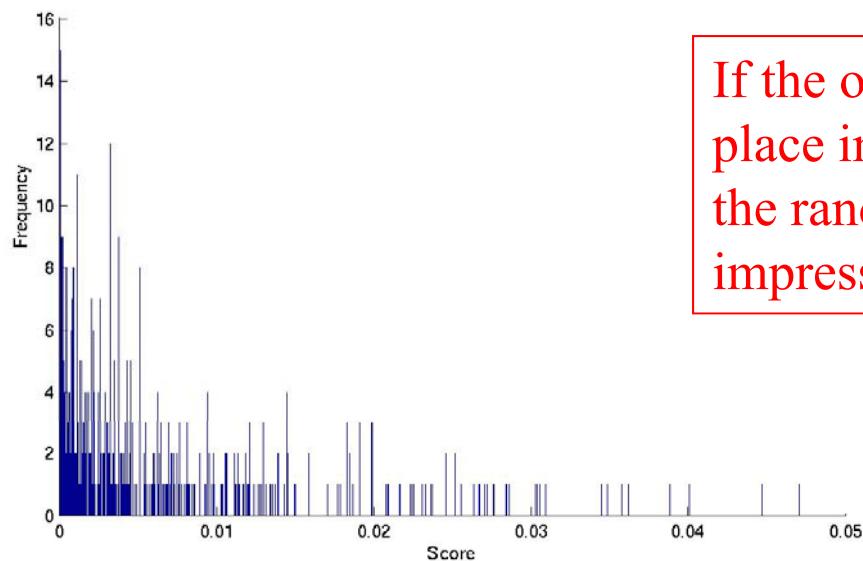
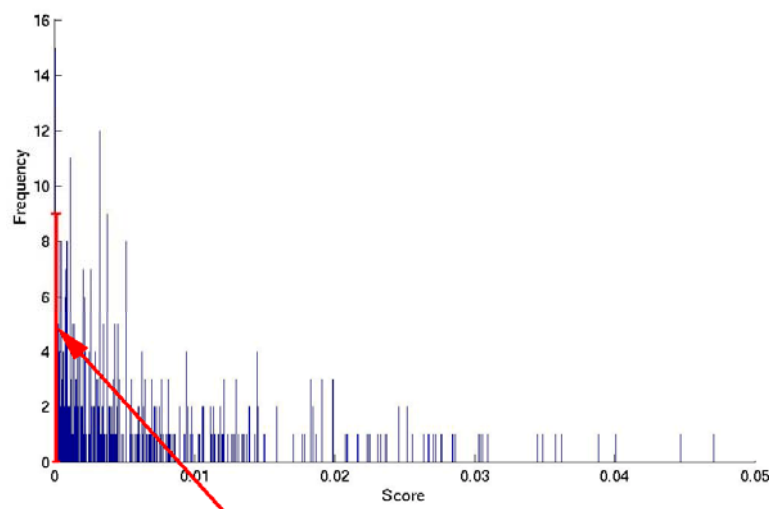


Randomization Test

- Take the training data points and the testing data points. Shuffle the date field to produce a randomized dataset called DB_{Rand}
- Find the rule with the best score on DB_{Rand} .

Multivariate Methods (WSARE 2.0)

Repeat the procedure on the previous slide for 1000 iterations. Determine how many scores from the 1000 iterations are better than the original score.



If the original score were here, it would place in the top 1% of the 1000 scores from the randomization test. We would be impressed and an alert should be raised.

Corrected p-value of the rule is:
 $\# \text{ better scores} / \# \text{ iterations}$

Multivariate Methods (WSARE 3.0)

Summary

For each rule in rule set:

1. Learn a Bayesian network from training data. Sample a baseline data set using Bayesian network.
2. Form 2x2 contingency table using counts from Baseline data and Testing.
3. Compute rule score using Fisher's Exact test / Chi-square test
4. Account for multiple hypothesis testing by randomization test
5. If p-value from randomization test $<$ alarm value, report rule

Multivariate Methods (WSARE 3.0)

Summary

Difference #1



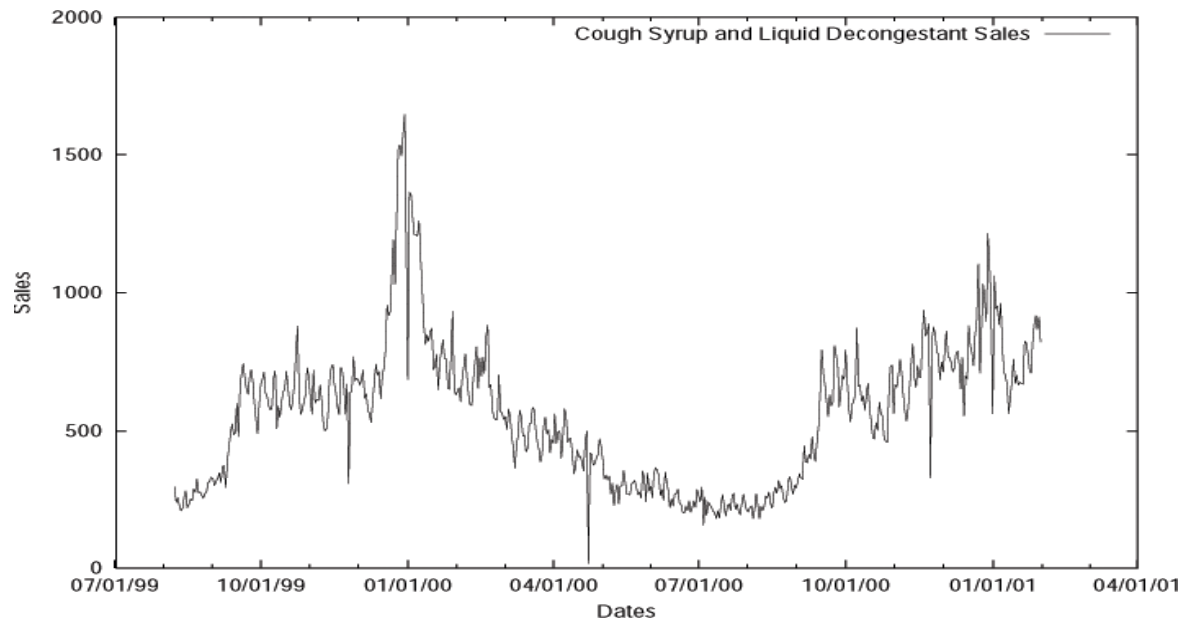
For each rule in rule set:

1. Learn a Bayesian network from training data. Sample a baseline data set using Bayesian network.
2. Form 2x2 contingency table using counts from Baseline data and Testing.
3. Compute rule score using Fisher's Exact test / Chi-square test
4. Account for multiple hypothesis testing by randomization test
5. If p-value from randomization test $<$ alarm value, report rule

Multivariate Methods (WSARE 3.0)

Difference #1 (from WSARE 2.0)

- Need to account for trends in the data



From: Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002). Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceedings of the National Academy of Sciences (pp. 5237-5249)

Multivariate Methods (WSARE 3.0)

- Temporal trends in health care data:
 - Seasonal effects in temperature and weather
 - Day of Week effects
 - Holidays
 - Etc.
- Not accounting for these trends can adversely affect the detection time and false positives rate

Multivariate Methods (WSARE 3.0)

Generating the Baseline:

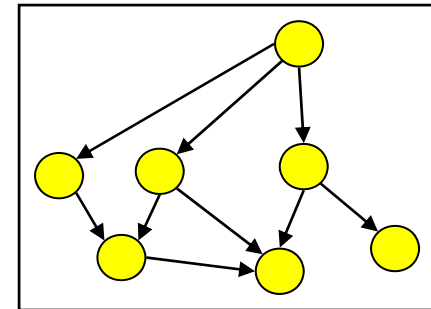
- “Taking into account that today is a public holiday...”
- “Taking into account that this is Spring...”
- “Taking into account recent heatwave...”
- “Taking into account recent flu levels...”
- “Taking into account that there’s a known natural Food-borne outbreak in progress...”

Multivariate Methods (WSARE 3.0)

Generating the Baseline:

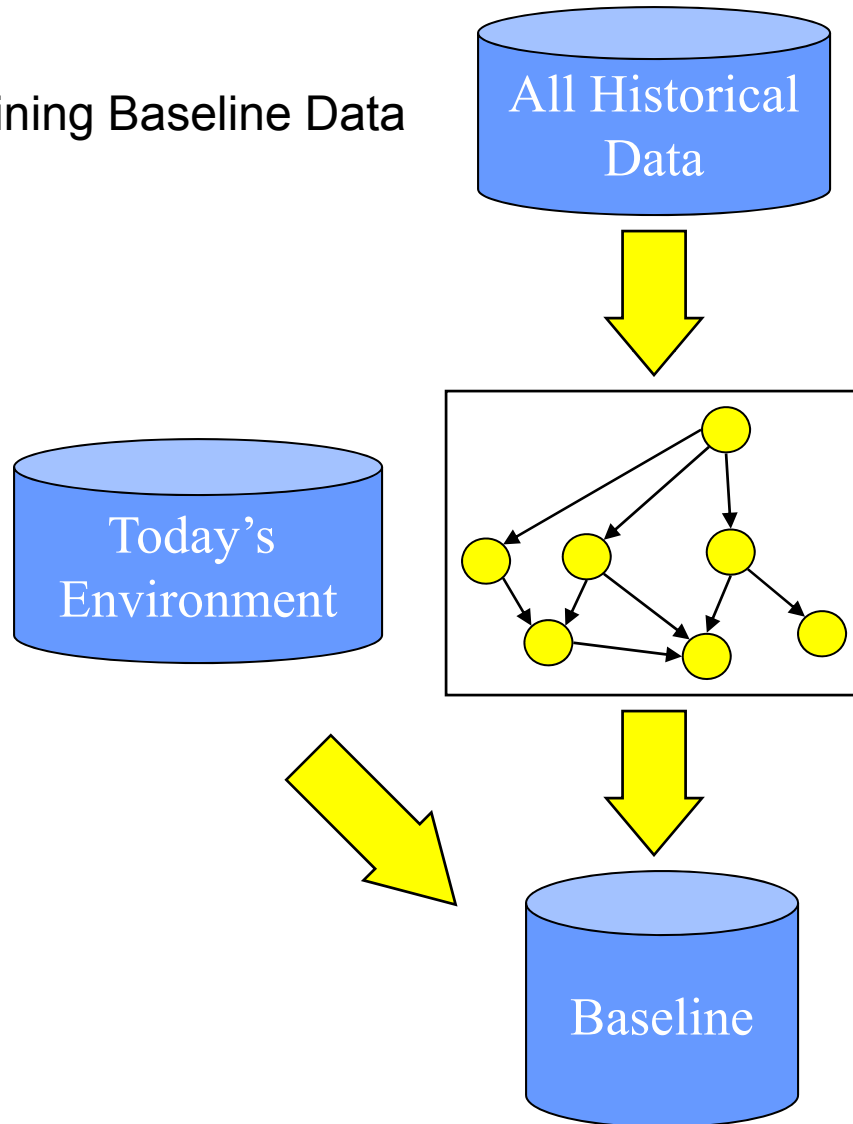
- “Taking into account that today is a public holiday...”
- “Taking into account that this is Spring...”
- “Taking into account recent heatwave...”
- “Taking into account recent flu levels...”
- “Taking into account that there’s a known natural Food-borne outbreak in progress...”

Use a Bayesian network (Pearl 1988) to model the joint probability distribution of the attributes.



Multivariate Methods (WSARE 3.0)

Obtaining Baseline Data

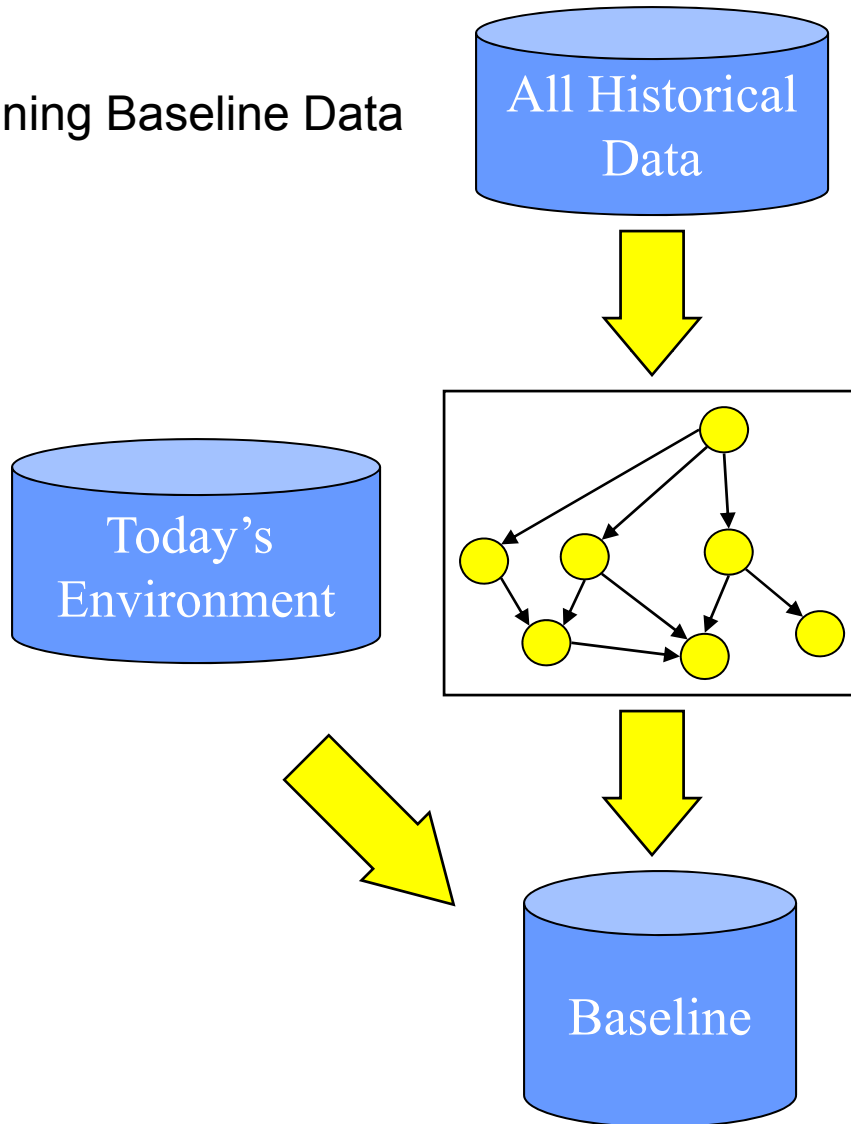


1. Learn Bayesian Network using Optimal Reinsertion [Moore and Wong 2003]

2. Generate baseline given today's environment

Multivariate Methods (WSARE 3.0)

Obtaining Baseline Data



1. Learn Bayesian Network using Optimal Reinsertion [Moore and Wong 2003]

2. Generate baseline given today's environment

Multivariate Methods (WSARE 3.0)

Divide the data into two types of attributes:

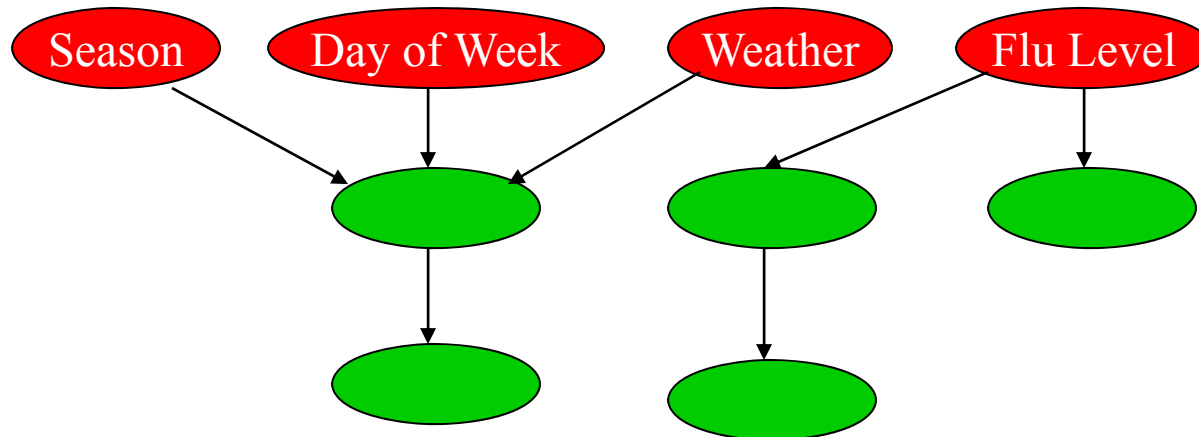
- **Environmental attributes:** attributes that cause trends in the data eg. day of week, season, weather, flu levels
- **Response attributes:** all other non-environmental attributes eg. age, gender

Multivariate Methods (WSARE 3.0)

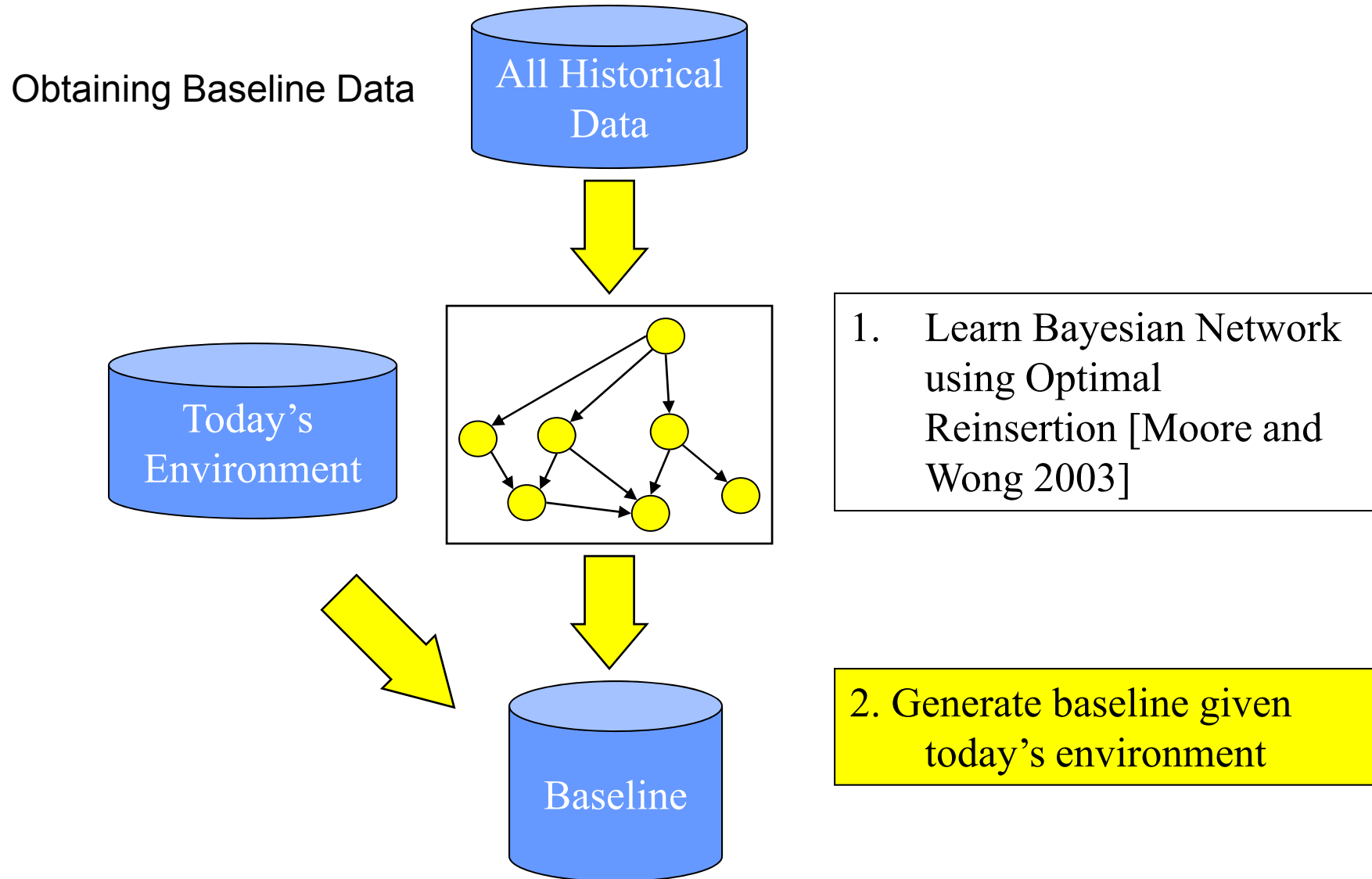
When learning the Bayesian network structure, do not allow environmental attributes to have parents.

Why?

- We are not interested in predicting their distributions
- Instead, we use them to predict the distributions of the response attributes



Multivariate Methods (WSARE 3.0)

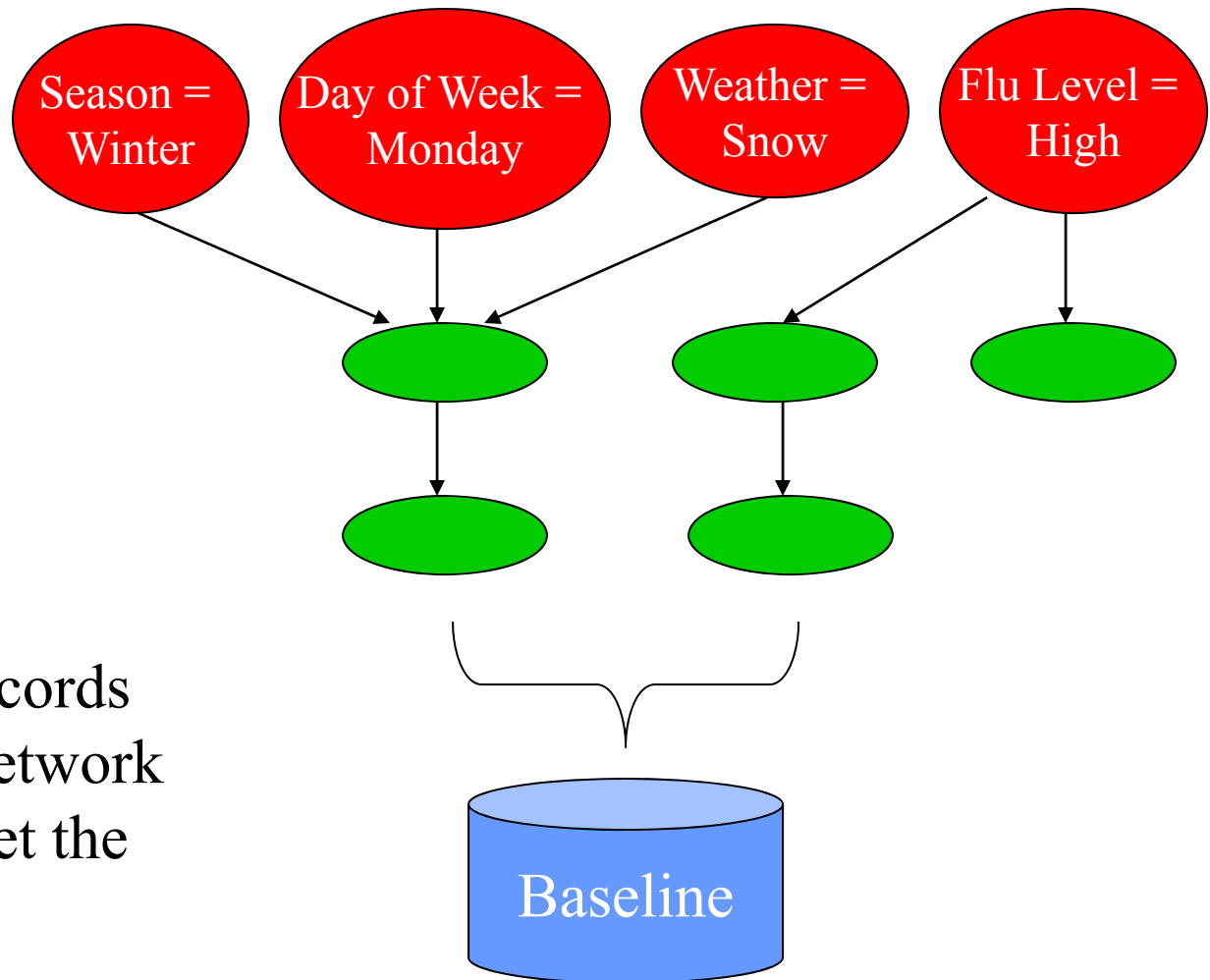


Multivariate Methods (WSARE 3.0)

Suppose we know the following for today:

	Season	Day of Week	Weather	Flu Level
Today	Winter	Monday	Snow	High

We fill in these values for the environmental attributes in the learned Bayesian network



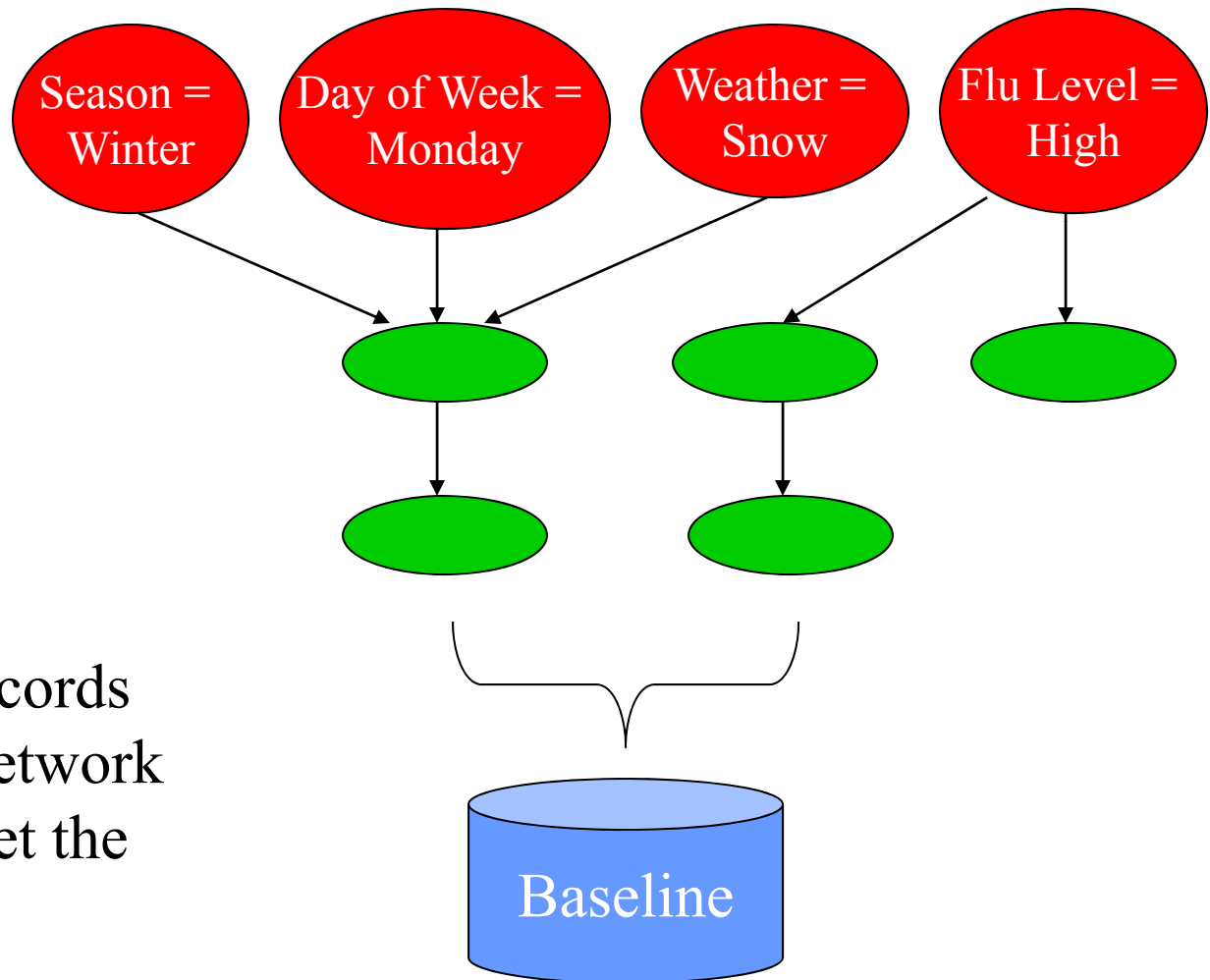
We sample 10000 records from the Bayesian network and make this data set the baseline

Multivariate Methods (WSARE 3.0)

Suppose we know the following for today:

	Season	Day of Week	Weather	Flu Level
Today	Winter	Monday	Snow	High

Sampling is easy because environmental attributes are at the top of the Bayes Net



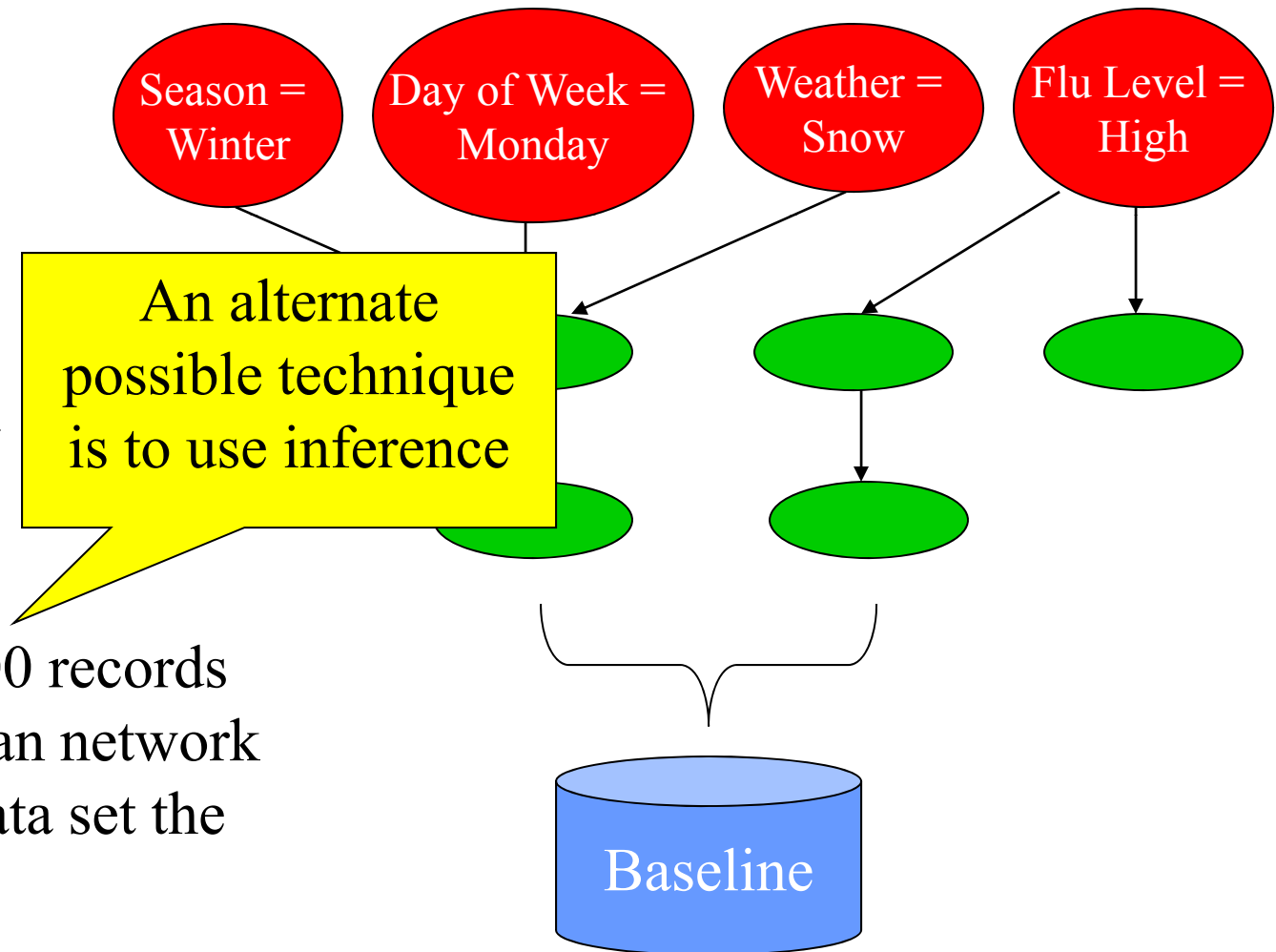
We sample 10000 records from the Bayesian network and make this data set the baseline

Multivariate Methods (WSARE 3.0)

Suppose we know the following for today:

	Season	Day of Week	Weather	Flu Level
Today	Winter	Monday	Snow	High

We fill in these values for the environmental attributes in the learned Bayesian network



We sample 10000 records from the Bayesian network and make this data set the baseline

Multivariate Methods (WSARE 3.0)

Summary

For each rule in rule set:

1. Learn a Bayesian network from training data. Sample a baseline data set using Bayesian network.
2. Form 2x2 contingency table using counts from Baseline data and Testing.
3. Compute rule score using Fisher's Exact test / Chi-square test
4. Account for multiple hypothesis testing by randomization test
5. If p-value from randomization test $<$ alarm value, report rule

Multivariate Methods (WSARE 3.0)

Side Note:

- Conditional Anomaly Detection (Song et al. 2007) is a similar approach
- Uses a Gaussian Mixture Model instead of a Bayesian Network
- Applicable to continuous multivariate data

But:

It discovers individual data points that are anomalies, not anomalous groups of data points

Multivariate Methods

Open Questions

- What about continuous features?
- What about mixed discrete and continuous features?
- Can we develop faster methods, especially those that can avoid randomization testing?
- Can we discover **interesting** (not just statistically significant) events?

Acknowledgements

- We would like to thank the following individuals/groups for their slide materials:
 - AUTON Lab (Carnegie Mellon University)
 - RODS Lab (University of Pittsburgh)
 - Ethan Dereszynski
- Univariate temporal methods section based in part on an earlier tutorial on detection algorithms for biosurveillance by Andrew Moore

References

- [Bay and Pazzani 1999] Bay, S. D., and Pazzani, M. J., Detecting change in categorical data: Mining contrast sets. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 302–306, New York, NY, 1999. ACM.
- [Crosier 1988] Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30, 291-303.
- [Dasu et al. 2006] Dasu, T., Krishnan, S., Venkatasubramanian, S., and Yi, K. (2006). An information-theoretic approach to detecting changes in multi-dimensional data streams. In Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications (Interface 06).
- [Dereszynski and Dietterich] Dereszynski, E., and Dietterich, T. (2007). Probabilistic Models for Anomaly Detection in Remote Sensor Data Streams. Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007). 75-82.
- [Dong and Li 1999] Dong, G. and Li, J. Efficient mining of emerging patterns: discovering trends and differences. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 43–52, New York, NY, 1999. ACM.
- [Fawcett and Provost 1999] Fawcett, T., and Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 53-62).
- [Goldenberg et al. 2002] Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002). Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceedings of the National Academy of Sciences (pp. 5237-5249)
- [Hotelling 1931] Hotelling, H. (1931). The generalization of Student's ratio, *Ann. Math. Statist.*, Vol. 2, pp 360–378.
- [Lowry et al. 1992] Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992). A Multivariate Exponentially Weighted Moving Average Chart, *Technometrics*, 34, 46-53.

References

- [Montgomery 2001] Montgomery, D. C. Introduction to Statistical Quality Control. John Wiley and Sons, Inc., 2001.
- [Moore and Wong 2003] Moore, A., and Wong, W.-K. (2003). Optimal Reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003) (pp. 552-559). Menlo Park, CA: AAAI Press.
- [Page 1954] Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41, 100-115.
- [Pearl 1988] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco, CA: Morgan Kaufmann Publishers, Inc.
- [Roberts 1959] Roberts, S. W. (1959). Control chart tests based on geometric moving averages. Technometrics, 1, 239-250.
- [Rosenbaum 2005] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of the Royal Statistical Society Series B, 67(4): 515-530.
- [Shewhart 1931] Shewhart, W. A. (1931). Economic control of quality of manufactured product. New York: D. Van Nostrand Company.
- [Song et al. 2007] Song, X., Wu, M., and Jermaine, C. (2007). Conditional Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering, 19(5), 631-645.
- [Song et al. 2007b] Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Statistical Change Detection for Multi-Dimensional Data. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 667-676, New York, NY: ACM Press.
- [Wong et al. 2005] Wong, W.-K., Moore, A., Cooper, G. and Wagner, M. (2005). What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Learning Research, 6, 1961-1998.

Tutorial on Event Detection

Part II: Spatial Event Detection

Daniel B. Neill
Carnegie Mellon University
H.J. Heinz III College
neill@cs.cmu.edu

This work was partially supported by NSF grant IIS-0325581 and CDC grant 8-R01-HK000020-02.

Outline of this part

A. Introduction to spatial event detection

Problem statement and overview of approaches.

B. Univariate scan statistic approaches

Spatial and space-time.

C. Multivariate scan statistic approaches

Parametric, non-parametric, and Bayesian.

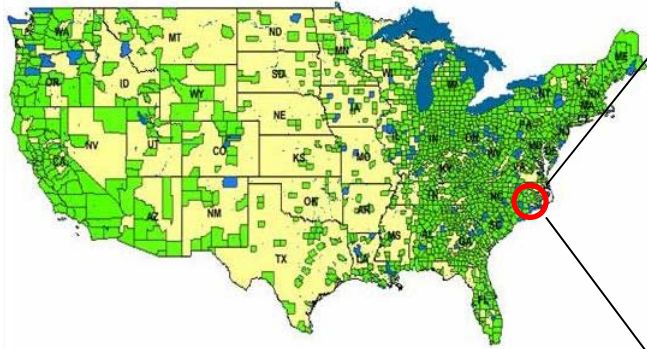
D. Current and future directions

Incorporating learning; fast algorithms.

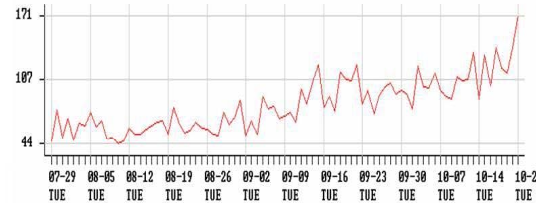
A. Introduction to Spatial Event Detection

- 1. The spatial event detection problem**
- 2. Approaches to spatial event detection**
 - a. Top-down and bottom-up approaches**
 - b. Parallel monitoring approaches**
 - c. Scan statistic approaches**
 - d. Other approaches from spatial statistics**

Spatial 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 location 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.)

Goals of detection task: **detect** any emerging events (e.g. disease outbreaks), **pinpoint** the affected spatial area, and **characterize** the type of event.

Informally, we want to know:

Is there anything happening?

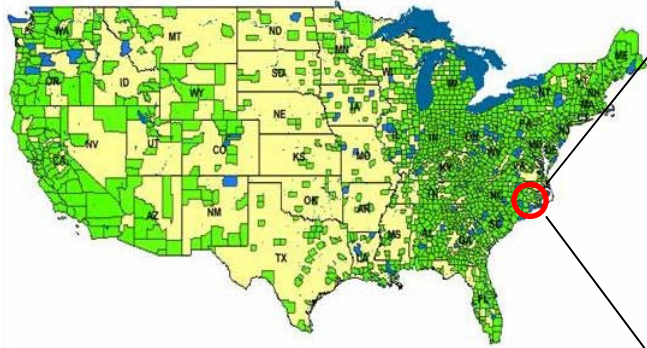
If so, **what** and **where**?

Formally, we will distinguish between:

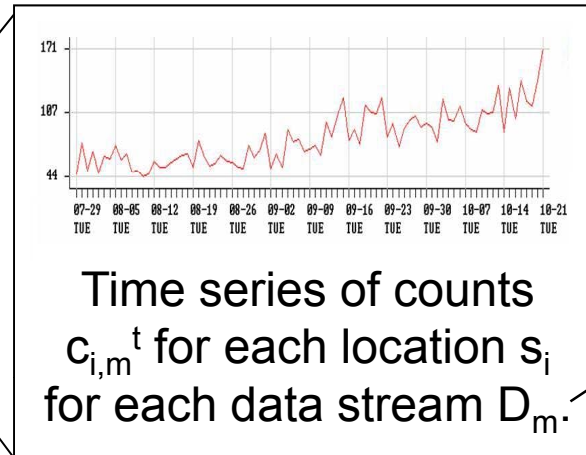
Null hypothesis H_0 (no events)

Set of alternative hypotheses $H_1(\mathbf{S}, \mathbf{E}_k)$
= event of type E_k in spatial region S .

Spatial 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

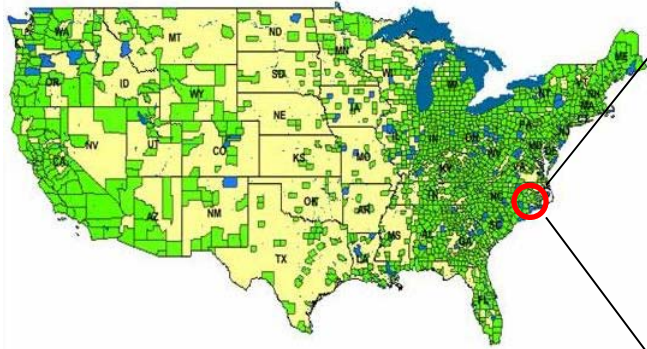
(etc.)

Goals of detection task: **detect** any emerging events (e.g. disease outbreaks), **pinpoint** the affected spatial area, and **characterize** the type of event.

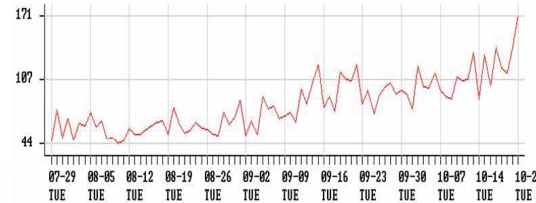
This formulation assumes **count** data aggregated to discrete time steps (e.g. days) and small areas (e.g. zips).

More generally, we can have a set of data records (observations) where each observation has a time-stamp, location information, and possibly other attributes. Each count represents the **number of observations** with given attributes in a given area and time interval.

Spatial 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 location 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

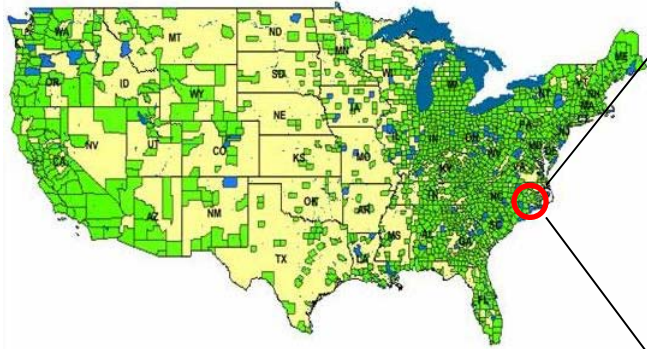
(etc.)

Goals of detection task: **detect** any emerging events (e.g. disease outbreaks), **pinpoint** the affected spatial area, and **characterize** the type of event.

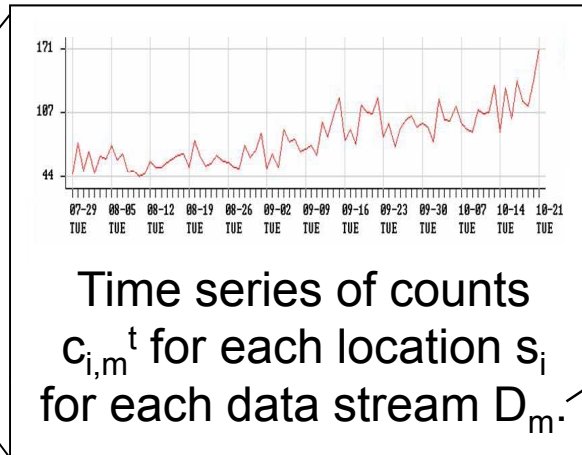
This formulation assumes **count** data aggregated to discrete time steps (e.g. days) and small areas (e.g. zips).

We assume that an event will result in anomalously high counts for some subset of data streams for the affected spatial region and time interval.

Spatial 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.)

Goals of detection task: **detect** any emerging events (e.g. disease outbreaks), **pinpoint** the affected spatial area, and **characterize** the type of event.

We will initially make three additional assumptions:

Purely spatial detection problem
(only a single time interval to consider)

Monitoring a single data stream D_m

Attempting to detect a single event type E_k

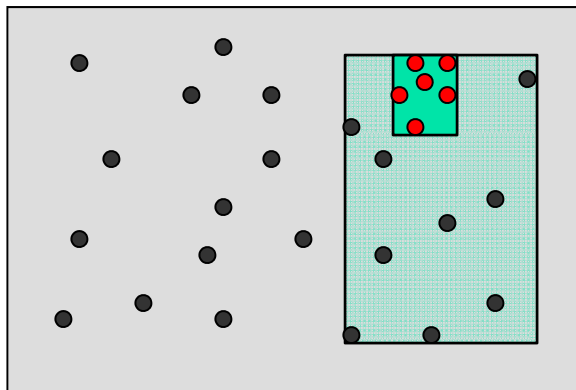
A. Introduction to Spatial Event Detection

1. The spatial event detection problem
2. Approaches to spatial event detection
 - a. Top-down and bottom-up approaches
 - b. Parallel monitoring approaches
 - c. Scan statistic approaches
 - d. Other approaches from spatial statistics

Top-down and bottom-up detection

Top-down detection approaches

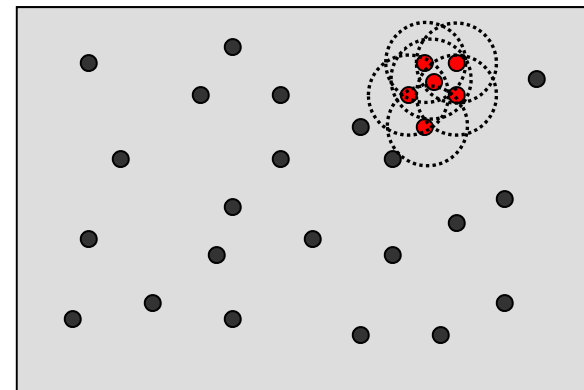
1. Are there any globally interesting patterns?
2. If so, find the most interesting sub-partition of the data and search it recursively.



Top-down: bump hunting¹

Bottom-up detection approaches

1. Find individual data points with “interesting” local neighborhoods
2. Aggregate interesting points into clusters.



Bottom-up: density-based clustering (e.g. DBSCAN²)

Thanks to Daniel Olivera for these examples.

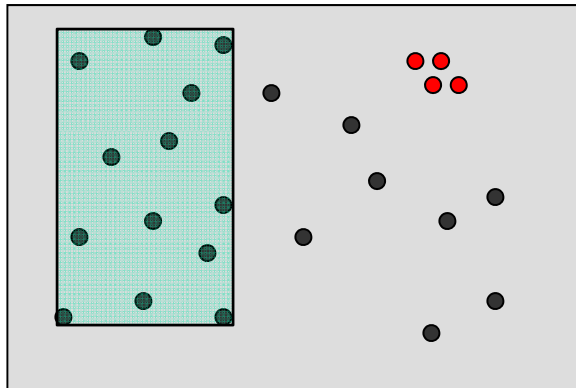
¹J. Friedman and N. Fisher, 1999.

²M. Ester et al., KDD 1996.

Greedy approaches can fail!

Top-down detection approaches

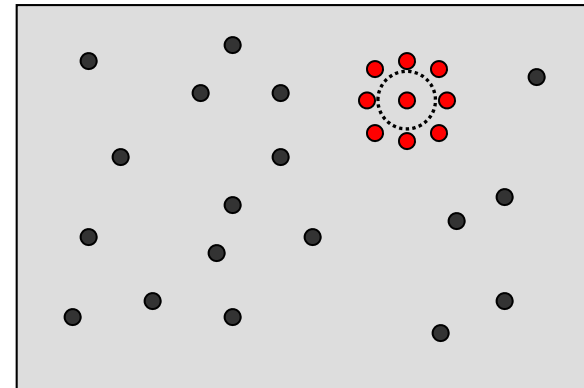
1. Are there any globally interesting patterns?
2. If so, find the most interesting sub-partition of the data and search it recursively.



Top-down fails when the affected region is too small to significantly affect the global aggregate statistics.

Bottom-up detection approaches

1. Find individual data points with “interesting” local neighborhoods
2. Aggregate interesting points into clusters.



Bottom-up fails when the affected region is not dense enough for the local neighborhoods to be interesting.

Greedy approaches can fail!

Top-down detection approaches

1. Are there any globally interesting patterns?
2. If so, find the most interesting sub-partition of the data and search it recursively.

How can we detect both small, dense clusters and larger, less dense clusters?

One answer: Parallel Monitoring

Partition the monitored area into subregions.

Then separately monitor each subregion using **purely temporal** detection methods (see Part 1!)

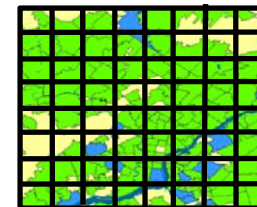
Bottom-up detection approaches

1. Find individual data points with “interesting” local neighborhoods
2. Aggregate interesting points into clusters.

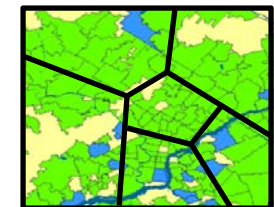
How can we move beyond cluster detection, to detect events that **emerge** in time?



Fixed partition:
zip code
boundaries



Fixed partition:
uniform grid



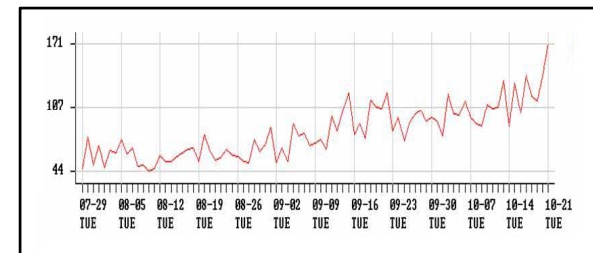
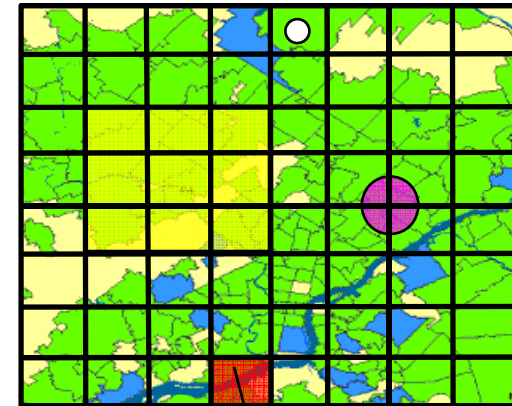
Ad-hoc partition:
data clustering

Challenges of parallel monitoring

One major challenge of parallel monitoring is choosing an appropriate partitioning of the monitored area.

A given partitioning has high power to detect events corresponding to a single partition (red), but is suboptimal for events which affect multiple partitions (yellow), part of a partition (white), or parts of multiple partitions (pink).

Coarse partitions lose power for small regions, fine partitions lose power for large regions, and both lose power for unaligned regions.



Challenges of parallel monitoring

One major challenge of parallel monitoring is choosing an appropriate partitioning of the monitored area.

A given partitioning has high power to detect events corresponding to a single partition (red), but is suboptimal for events which affect multiple partitions (yellow), part of a partition (white), or parts of multiple partitions (pink).

Coarse partitions lose power for small regions, fine partitions lose power for large regions, and both lose power for unaligned regions.

A second challenge of parallel monitoring is the problem of **multiple hypothesis testing**.

Monitoring thousands of spatial partitions, and performing a separate significance test for each, leads to huge numbers of false positive alerts.

The Bonferroni correction for multiple tests leads to greatly reduced detection power.

Solution to the first challenge: the spatial scan statistic.

1. Form a very fine partitioning of the monitored area into individual **locations** (e.g. zip codes or census tracts, depending on spatial resolution of the data).
2. Rather than monitoring each partition separately, examine a huge number of overlapping spatial **regions**, each consisting of a **group** of locations.

Challenges of parallel monitoring

One major challenge of parallel monitoring is choosing an appropriate partitioning of the monitored area.

A second challenge of parallel monitoring is the problem of **multiple hypothesis testing**.

A given partitioning has high power to detect events corresponding to a single partition (red).

Monitoring thousands of spatial partitions simultaneously is a

but is

multiple

(w

Coar

fine partiti

both lose power for unaligned regions,

Searching over so many regions makes the multiple hypothesis testing problem even worse...

leads to greatly reduced detection power.

Solution to

1. Form a very fine (e.g. zip codes)
2. Rather than more overlapping spat

But we can solve the multiple testing problem by:

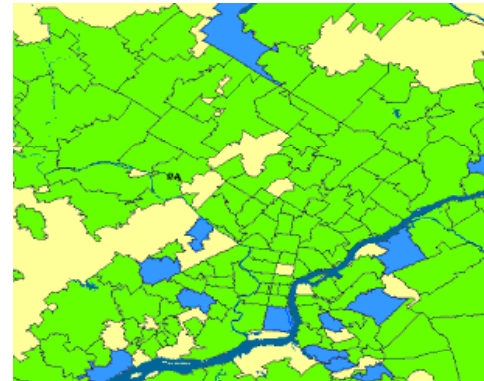
1. Finding the most significant regions.
2. Determining how likely we would be to see any regions that significant due to chance.

Other methods from spatial statistics

Tests for “general clustering”

(Whittemore, Tango, Knox, Mantel, etc.)

Determine whether there is sufficient evidence of spatial or space-time clustering in the data, but without detecting specific clusters.



Clustering?

Not clustered?

Tests for “focused clustering”

(Lawson, Stone, Waller, Diggle, etc.)

Determine whether the risk is significantly increased near a given point (e.g. possible environmental hazard).

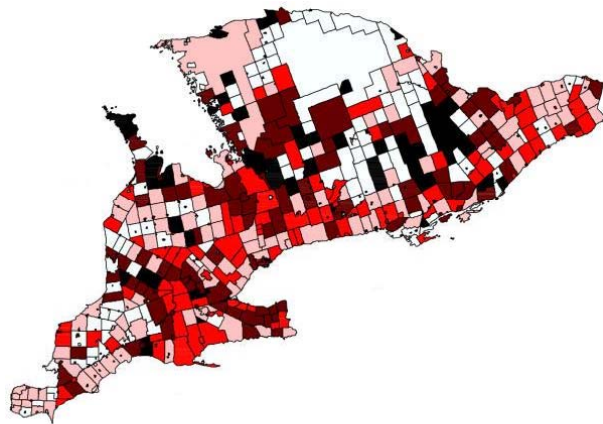


Focus?

Not a focus?

Neither method detects cluster locations!

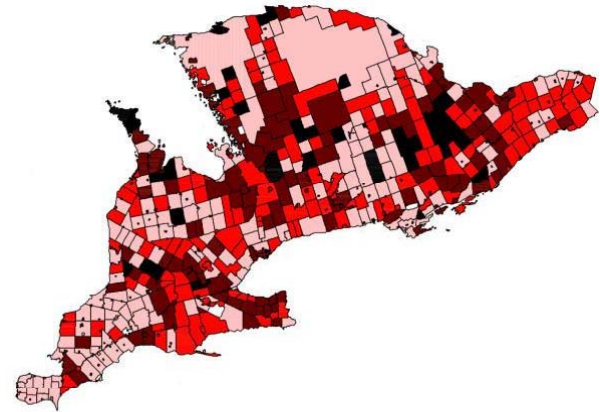
Spatial risk mapping approaches



Original risk map
(observed / expected)



Spatial
smoothing



Smoothed risk map

Often based on Bayesian modeling



EB (Gangnon & Clayton; Mollie)

FB (Lawson et al.)

Advantages: Explicit modeling of spatial correlation structure, useful for data visualization, can detect areas with high risk.

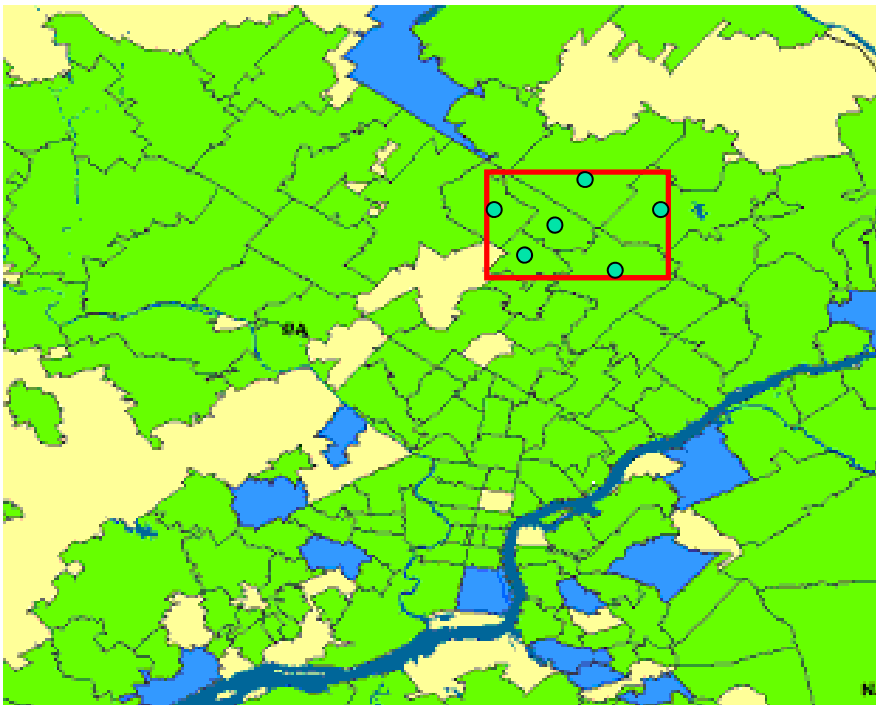
Disadvantages: Cannot automatically determine whether an event has occurred; cannot identify the spatial area and time duration.

B. Univariate Scan Statistic Approaches

1. **Kulldorff's spatial scan statistic**
2. **Variants of spatial scan:**
 - Which spatial regions to search?
 - How to evaluate the score of a region?
3. **Extensions to space-time scanning
(expectation-based scan statistic)**

The spatial scan statistic

(Kulldorff, 1997)

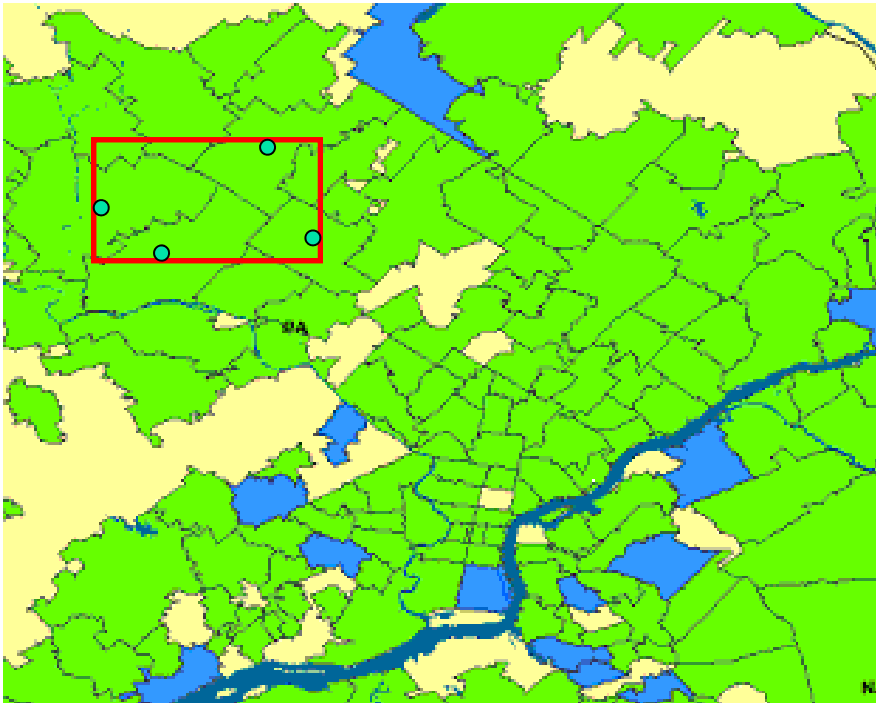


Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

The spatial scan statistic

(Kulldorff, 1997)

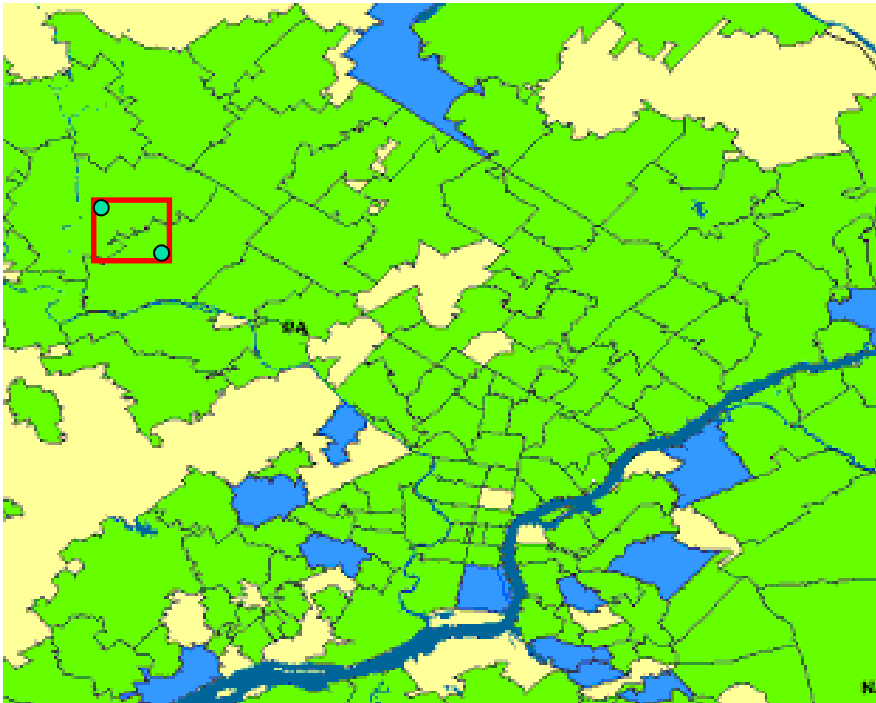


Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

The spatial scan statistic

(Kulldorff, 1997)

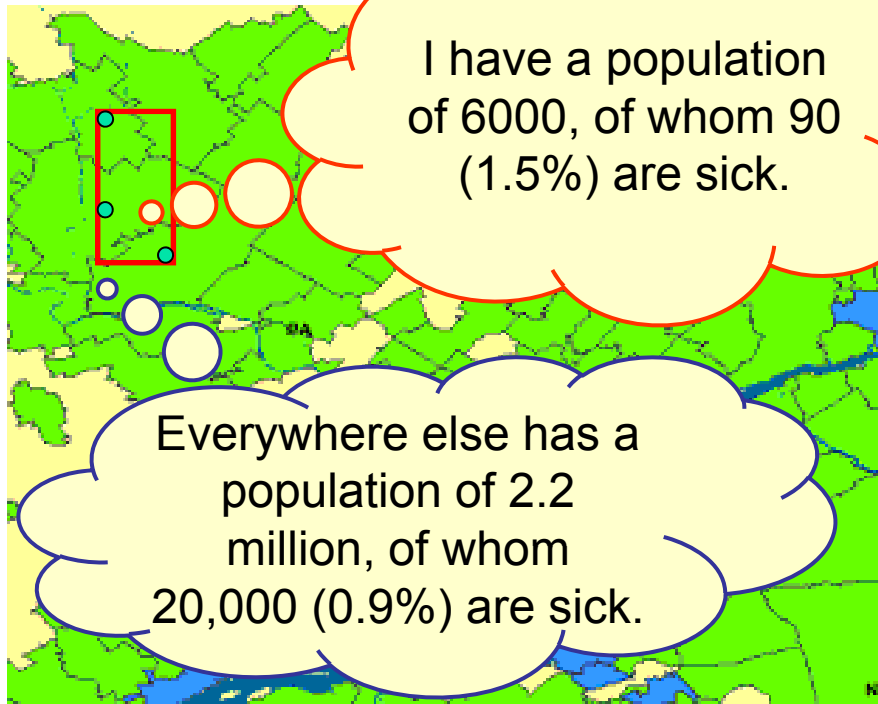


Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

The spatial scan statistic

(Kulldorff, 1997)



Rather than monitoring individual locations, we examine groups of locations.

Imagine moving a spatial window around the monitored area, allowing the size and shape of the window to vary.

Is there any position of the window such that the points inside form a significant cluster?

How to evaluate a region?
Which regions to search?
How to search them efficiently?

We compute a **score** for each spatial region, and then test whether the highest scoring regions are significant.

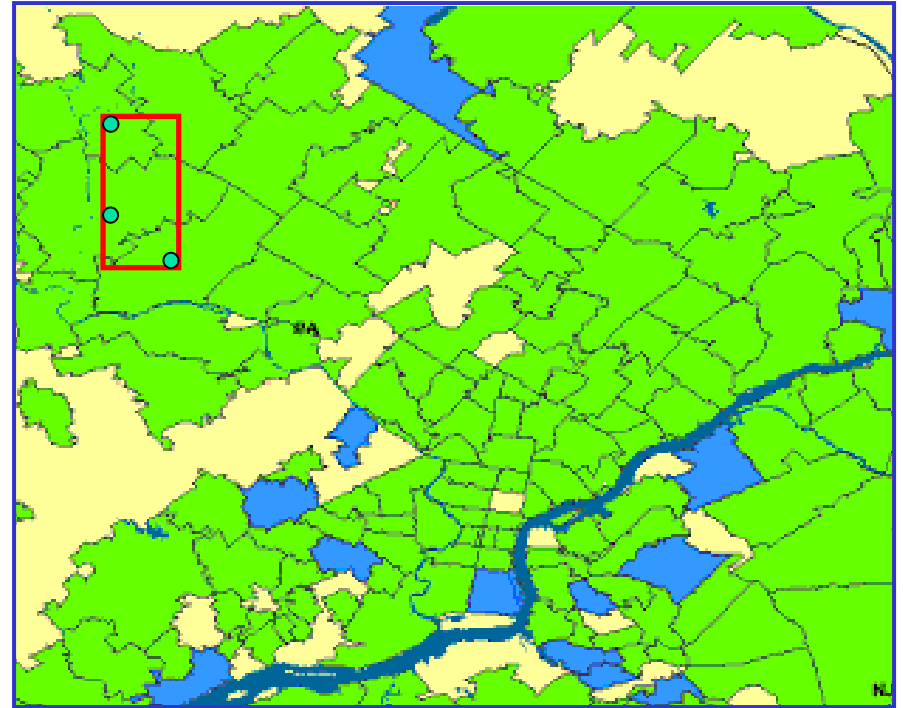
Finding the most significant regions

- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .

c_i = **count** for location s_i (e.g. number of disease cases)

b_i = **baseline** for location s_i (e.g. population at-risk, or expected count computed from historical data)

q = **risk** (expected ratio of count to baseline)



Kulldorff's model

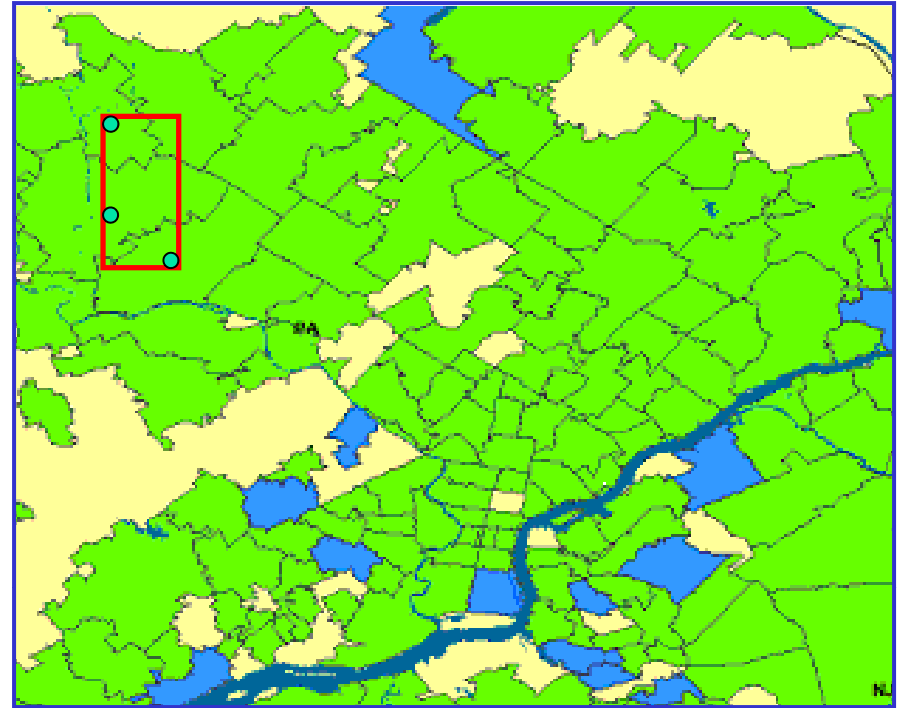
$$c_i \sim \text{Poisson}(qb_i)$$

H_0 : $q = q_{\text{all}}$ everywhere

$H_1(S)$: $q = q_{\text{in}}$ inside S ,
 $q = q_{\text{out}}$ outside,
 $q_{\text{in}} > q_{\text{out}}$.

Finding the most significant regions

- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .



Kulldorff's model

$$c_i \sim \text{Poisson}(qb_i)$$

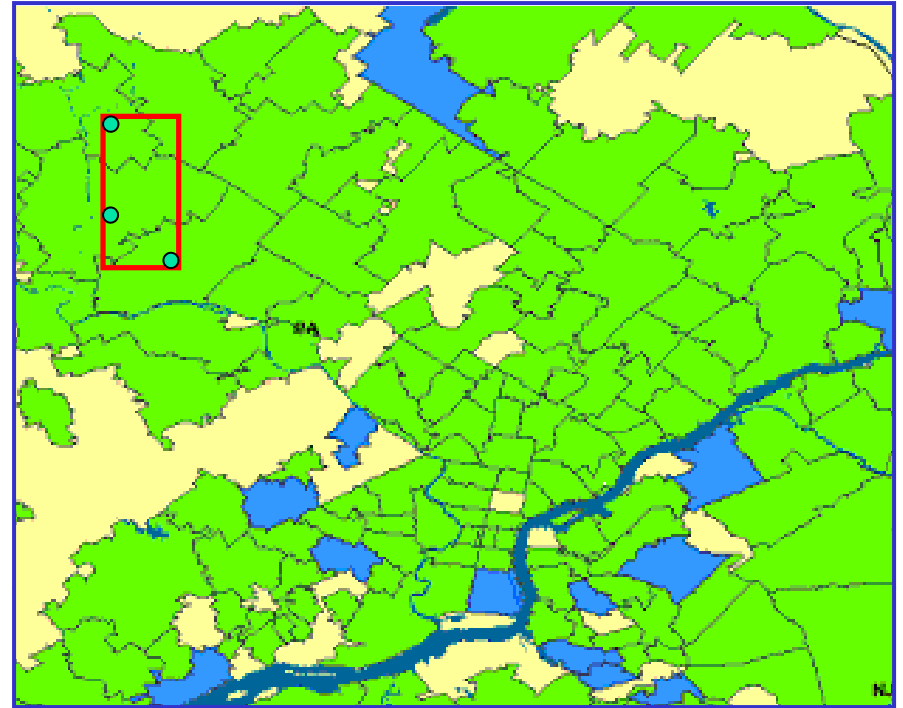
H_0 : $q = q_{\text{all}}$ everywhere

$H_1(S)$: $q = q_{\text{in}}$ inside S ,
 $q = q_{\text{out}}$ outside,
 $q_{\text{in}} > q_{\text{out}}$.

Finding the most significant regions

- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .
- Derive a score function:
 - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$



Kulldorff's model

$$c_i \sim \text{Poisson}(qb_i)$$

H_0 : $q = q_{\text{all}}$ everywhere

$H_1(S)$: $q = q_{\text{in}}$ inside S ,
 $q = q_{\text{out}}$ outside,
 $q_{\text{in}} > q_{\text{out}}$.

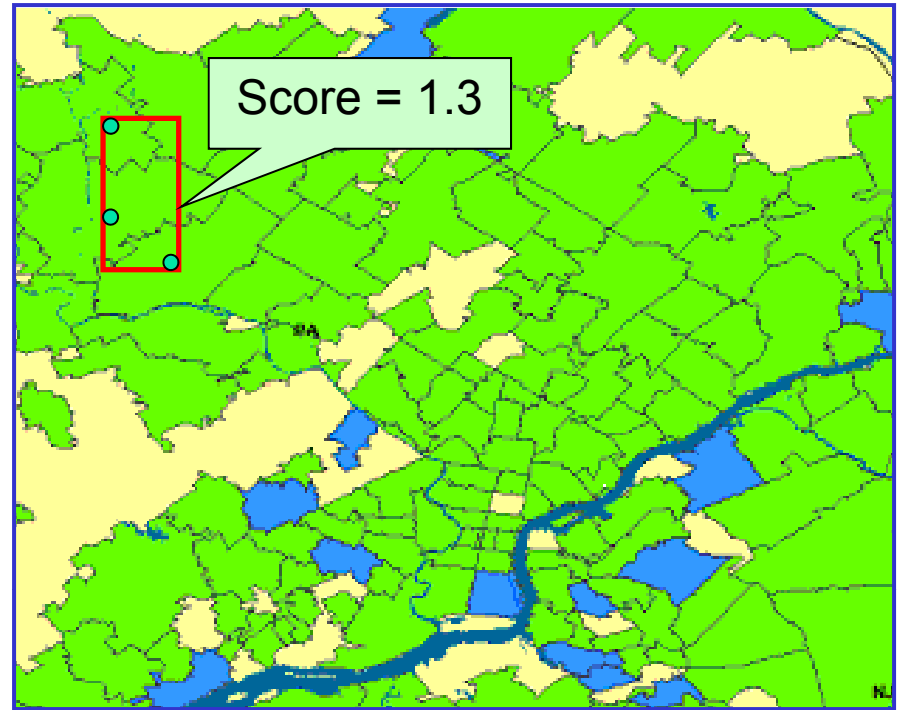
Finding the most significant regions

- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .

- Derive a score function:
 - Likelihood ratio:

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

$$F(S) = \left(\frac{C}{B}\right)^C \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$



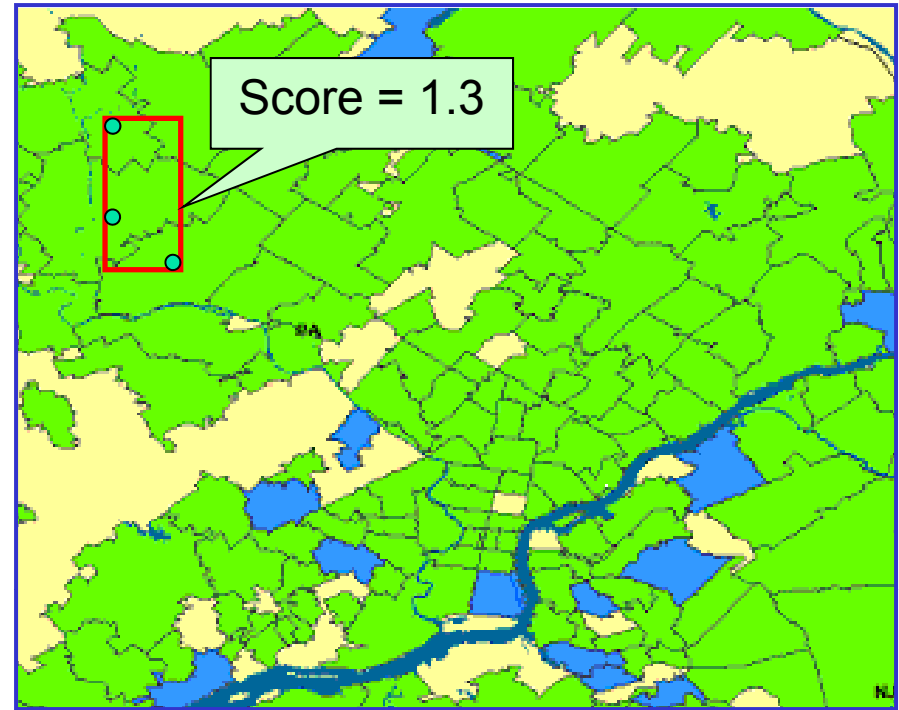
Kulldorff's model

$$c_i \sim \text{Poisson}(qb_i)$$

H_0 : $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S ,
 $q = q_{out}$ outside,
 $q_{in} > q_{out}$.

Finding the most significant regions



- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .

- Derive a score function:
 - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

Total count and baseline of region S

Total count and baseline of search area

$$F(S) = \left(\frac{C}{B}\right)^C \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

Kulldorff's model

$c_i \sim \text{Poisson}(qb_i)$

$H_0: q = q_{all}$ everywhere

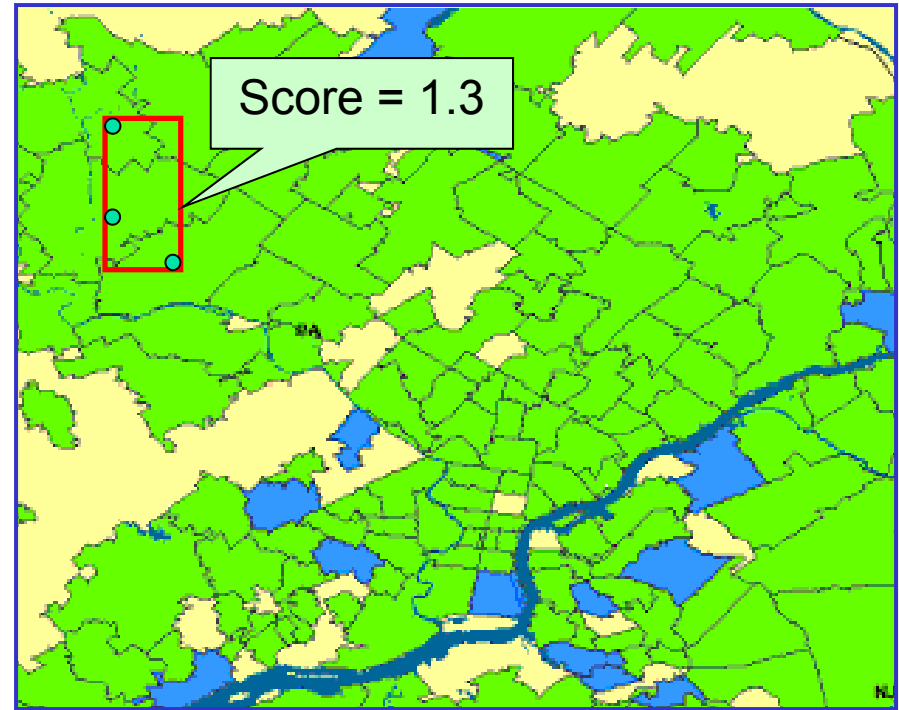
$H_1(S): q = q_{in}$ inside S ,
 $q = q_{out}$ outside,
 $q_{in} > q_{out}$.

Finding the most significant regions

- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .
- Derive a score function:
 - Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

$$F(S) = \left(\frac{C}{B}\right)^C \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$



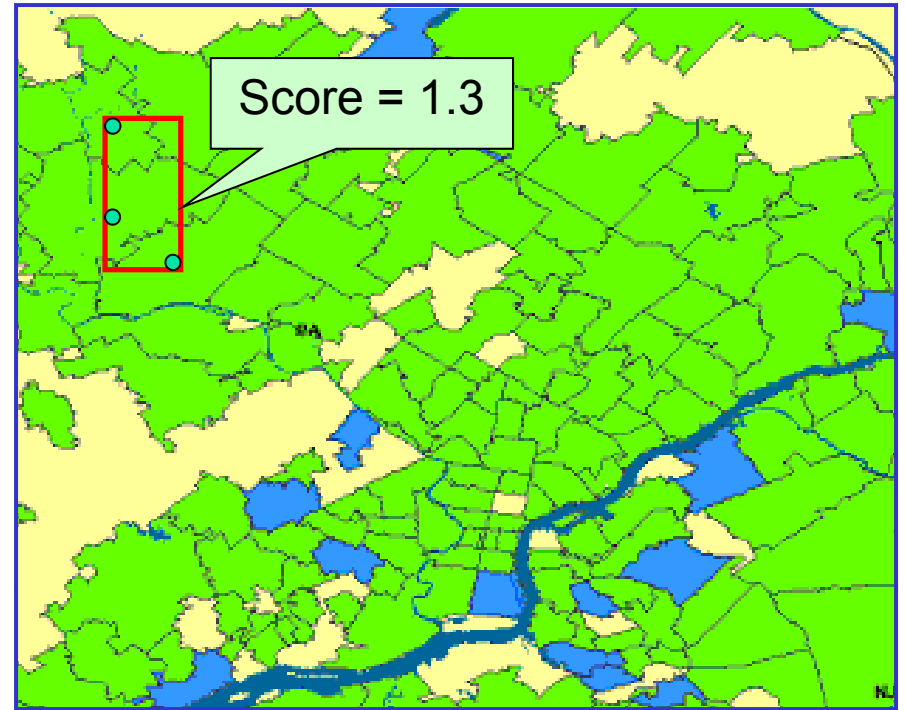
Kulldorff's model

$$c_i \sim \text{Poisson}(qb_i)$$

H_0 : $q = q_{\text{all}}$ everywhere

$H_1(S)$: $q = q_{\text{in}}$ inside S ,
 $q = q_{\text{out}}$ outside,
 $q_{\text{in}} > q_{\text{out}}$.

Finding the most significant regions



- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .

- Derive a score function:

- Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

- To find the most significant regions:

$$S^* = \arg \max_S F(S)$$

$$F(S) = \left(\frac{C}{B}\right)^C \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$

Kulldorff's model

$$c_i \sim \text{Poisson}(q b_i)$$

H_0 : $q = q_{all}$ everywhere

$H_1(S)$: $q = q_{in}$ inside S ,

$q = q_{out}$ outside,

$$q_{in} > q_{out}$$

Finding the most significant regions

- Define models:
 - of the null hypothesis H_0 : no events.
 - of the alternative hypotheses $H_1(S)$: event in region S .

- Derive a score function:

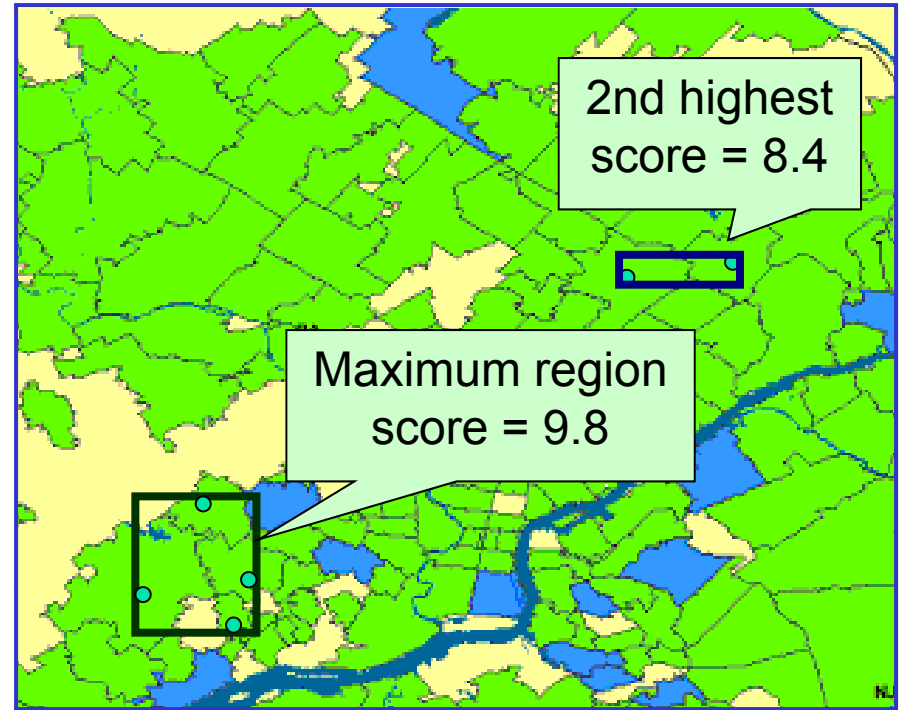
- Likelihood ratio:

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

- To find the most significant regions:

$$S^* = \arg \max_S F(S)$$

$$F(S) = \left(\frac{C}{B}\right)^C \left(\frac{C_{tot} - C}{B_{tot} - B}\right)^{C_{tot} - C} \left(\frac{C_{tot}}{B_{tot}}\right)^{-C_{tot}}$$



Kulldorff's model

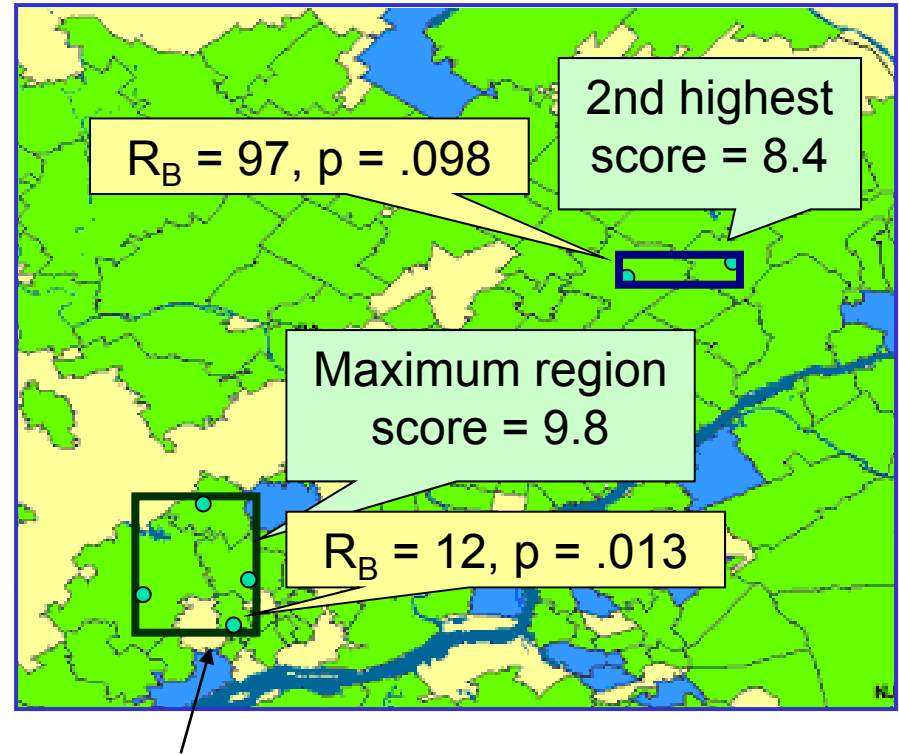
$$c_i \sim \text{Poisson}(qb_i)$$

H_0 : $q = q_{all}$ everywhere

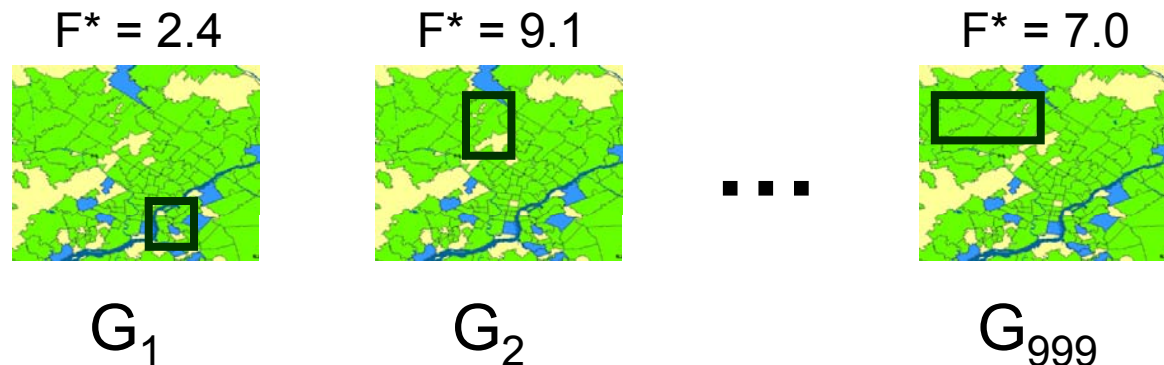
$H_1(S)$: $q = q_{in}$ inside S ,
 $q = q_{out}$ outside,
 $q_{in} > q_{out}$.

Which regions are significant?

- Randomly generate counts for $R = 999$ replica datasets under H_0 (i.e. assuming no events).
- Find maximum region score $F^* = \max_S F(S)$ of each replica.
- p-value of region $S = (R_B + 1) / (R + 1)$, where $R_B = \#$ of replicas with $F^* \geq F(S)$.
- All regions with p-values $< \alpha$ are significant at level α .



This region is significant at $\alpha = .05$; no other regions are significant.



B. Univariate Scan Statistic Approaches

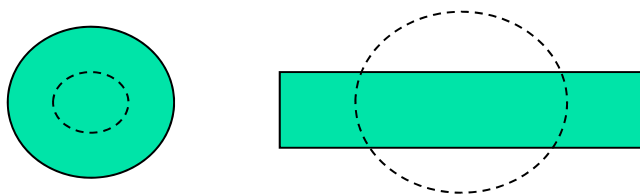
1. Kulldorff's spatial scan statistic
2. **Variants of spatial scan:**
 - Which spatial regions to search?
 - How to evaluate the score of a region?
3. Extensions to space-time scanning
(expectation-based scan statistic)

Choosing the set of search regions

- Some practical considerations:
 - Set of regions should cover entire search space.
 - Regions should overlap, not partition the space.
- Choose a set of regions that corresponds well with the size/shape of the clusters we want to detect.
 - Typical approaches consider some fixed shape (circles, rectangles) and vary the location and dimensions.

Don't search too few regions:

Reduced power to detect clusters outside the search space.



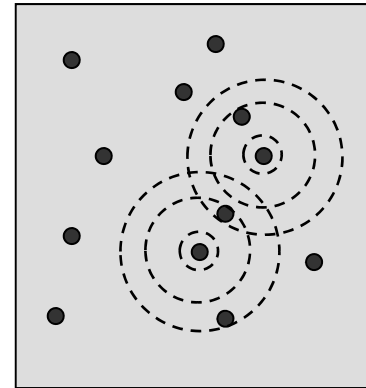
Don't search too many regions:

Overall power to detect any given subset of regions reduced because of multiple hypothesis testing.

Computational infeasibility!

Choosing the set of search regions

- Kulldorff's original spatial scan searches over circular regions of varying radius, centered at each spatial location s_i .
- Since the score function $F(S)$ depends only on which locations are included, we need to search $O(N^2)$ regions, each consisting of a center location and its k -NN.
- Advantages: computationally efficient, generalizable to arbitrary metric spaces, high detection power for compact clusters.
- Disadvantage: low power for elongated/irregular clusters.



April 1979: inadvertent release of anthrax from a Soviet biological weapons facility, 77 cases confirmed. Disease cluster elongated due to wind.

Choosing the set of search regions

- Kulldorff's original spatial scan searches over circular regions of varying radius, centered at each spatial location s_i .
- Since the score function $F(S)$ depends only on which locations are included, we need to search $O(N^2)$ regions, each consisting of a center location and its k -NN.
- Advantages: computationally efficient, generalizable to arbitrary metric spaces, high detection power for compact clusters.
- Disadvantage: low power for elongated/irregular clusters.

Many recent spatial scan variants search over elongated clusters, e.g. rectangles¹ or ellipses²

Other variants: heuristic search over all connected regions³, or exhaustive search over a subset of connected regions^{4,5}

**Main challenge:
efficient computation!**

¹Neill and Moore, KDD 2004

²Kulldorff et al., Stat. Med., 2007

³Duczmal and Assuncao, CSDA, 2004

⁴Tango and Takahashi, IJHG, 2005

⁵Patil and Taillie, EES, 2004

Computing the score function

Method 1 (Frequentist, hypothesis testing approach):

Use likelihood ratio $F(S) = \frac{\Pr(Data | H_1(S))}{\Pr(Data | H_0)}$

Prior probability of region S

Method 2 (Bayesian approach):

Use posterior probability $F(S) = \frac{\Pr(Data | H_1(S)) \Pr(H_1(S))}{\Pr(Data)}$

Computing the score function

Method 1 (Frequentist, hypothesis testing approach):

Use likelihood ratio $F(S) = \frac{\Pr(Data | H_1(S))}{\Pr(Data | H_0)}$

Prior probability of region S

Method 2 (Bayesian approach):

Use posterior probability $F(S) = \frac{\Pr(Data | H_1(S)) \Pr(H_1(S))}{\Pr(Data)}$

What to do when each hypothesis has a parameter space Θ ?

Method A (Maximum likelihood approach)

$$\Pr(Data | H) = \max_{\theta \in \Theta(H)} \Pr(Data | H, \theta)$$

Method B (Marginal likelihood approach)

$$\Pr(Data | H) = \int_{\theta \in \Theta(H)} \Pr(Data | H, \theta) \Pr(\theta)$$

Computing the score function

Method 1 (Frequentist, hypothesis testing approach):

Use likelihood ratio $F(S) = \frac{\Pr(Data | H_1(S))}{\Pr(Data | H_0)}$

Most common (frequentist) approach: use likelihood ratio statistic, with maximum likelihood estimates of any free parameters, and compute statistical significance by randomization^{1,2}

Method A (Maximum likelihood approach)

$$\Pr(Data | H) = \max_{\theta \in \Theta(H)} \Pr(Data | H, \theta)$$

¹Kulldorff, 1997

²Neill and Moore, ADKDD 2005.

Many possible variants, depending on how we model the likelihood of the data under each hypothesis $H_1(S)$ and H_0 (Poisson, Gaussian, exponential, negative binomial, etc.)

Computing the score function

Advantages: Randomization testing unnecessary (1000x speedup), can be extended to multiple data streams and multiple event types (more on this later).

Method 2 (Bayesian approach):

Use posterior probability $F(S) = \frac{\Pr(Data | H_1(S)) \Pr(H_1(S))}{\Pr(Data)}$

Bayesian spatial scan statistic^{1,2}:

A Bayesian marginal likelihood approach, efficiently computable using conjugate priors (Gamma-Poisson).

Method B (Marginal likelihood approach)

$$\Pr(Data | H) = \int_{\theta \in \Theta(H)} \Pr(Data | H, \theta) \Pr(\theta)$$

¹Neill et al., NIPS 2005

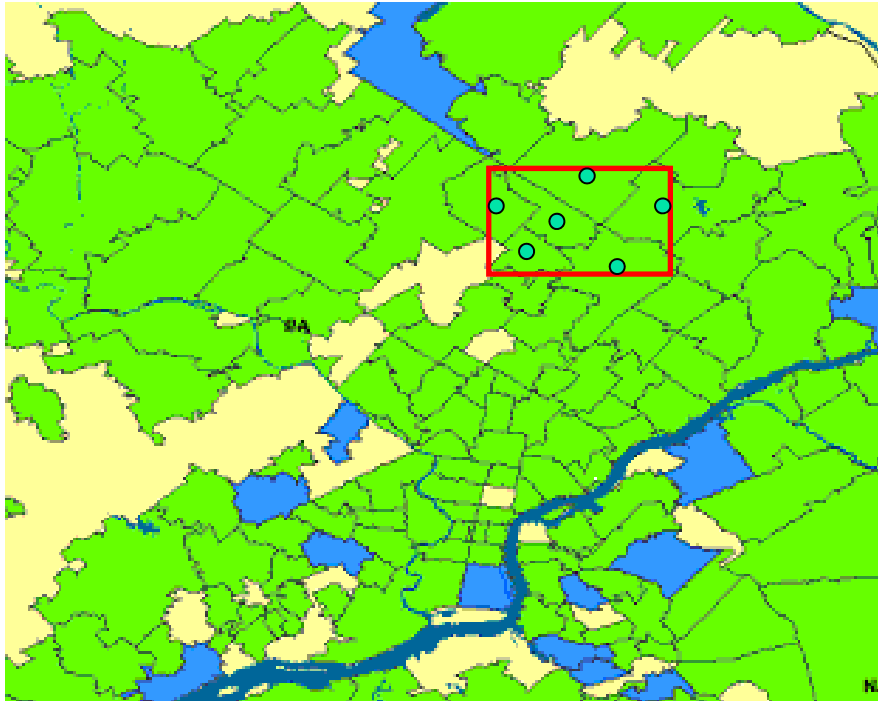
²Neill and Cooper, *Machine Learning*, 2009, in press.

B. Univariate Scan Statistic Approaches

1. Kulldorff's spatial scan statistic
2. Variants of spatial scan:
 - Which spatial regions to search?
 - How to evaluate the score of a region?
3. **Extensions to space-time scanning
(expectation-based scan statistic)**

Expectation-based scan statistics

(Kulldorff, 2001; Neill et al., KDD 2005)

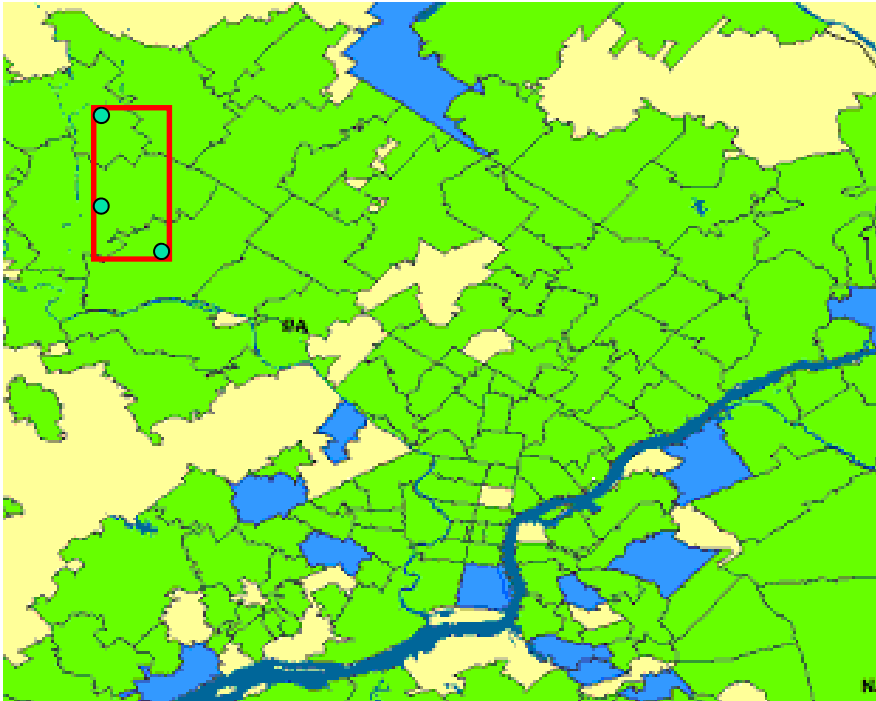


To detect emerging events, we can search for **space-time regions** where the recently observed counts are significantly higher than expected.

Imagine moving a **space-time** window around the scan area, allowing the window size, shape, and **duration** to vary.

Expectation-based scan statistics

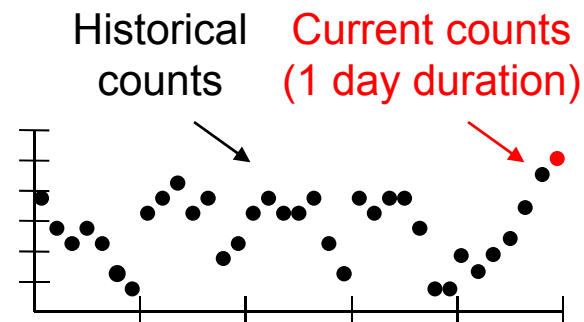
(Kulldorff, 2001; Neill et al., KDD 2005)



To detect emerging events, we can search for **space-time regions** where the recently observed counts are significantly higher than expected.

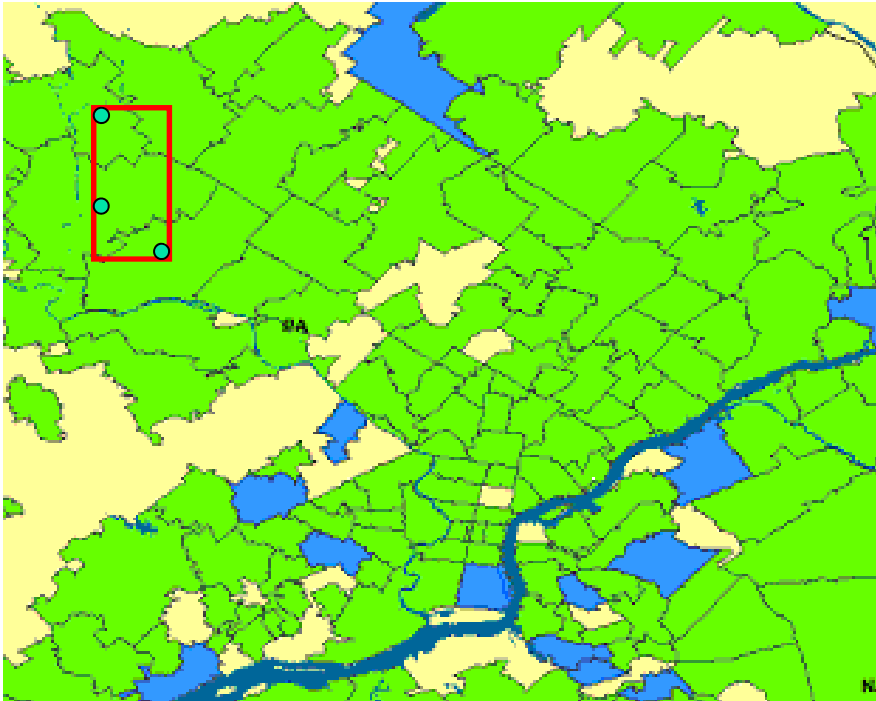
Imagine moving a **space-time** window around the scan area, allowing the window size, shape, and **duration** to vary.

(Consider most recent w days, $w = 1 \dots W_{\max}$)



Expectation-based scan statistics

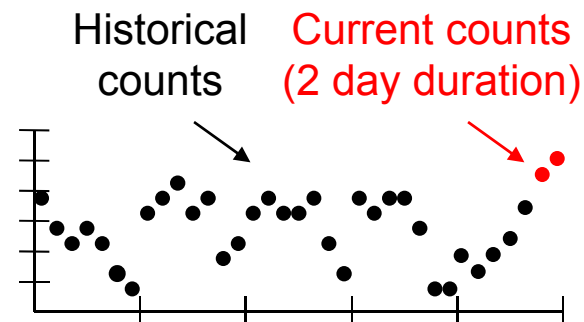
(Kulldorff, 2001; Neill et al., KDD 2005)



To detect emerging events, we can search for **space-time regions** where the recently observed counts are significantly higher than expected.

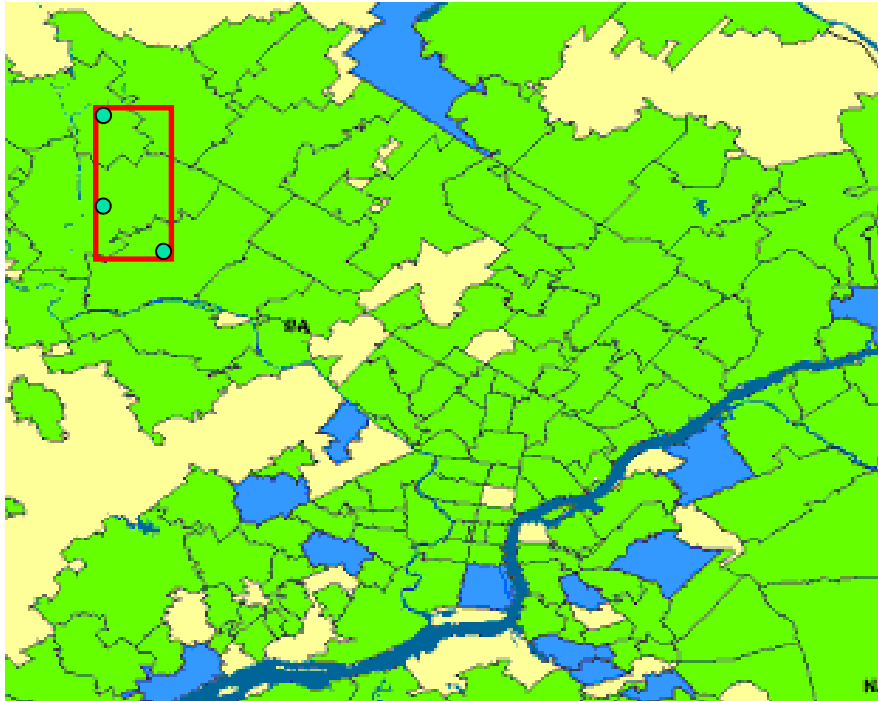
Imagine moving a **space-time** window around the scan area, allowing the window size, shape, and **duration** to vary.

(Consider most recent w days, $w = 1 \dots W_{\max}$)



Expectation-based scan statistics

(Kulldorff, 2001; Neill et al., KDD 2005)

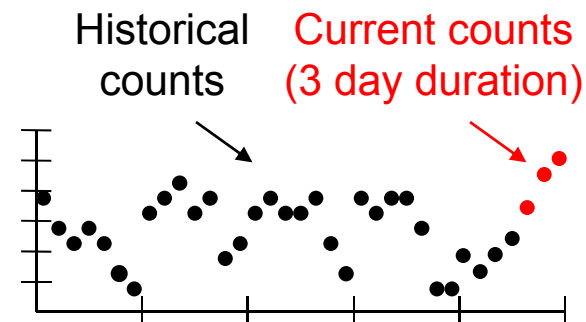


To detect emerging events, we can search for **space-time regions** where the recently observed counts are significantly higher than expected.

Imagine moving a **space-time** window around the scan area, allowing the window size, shape, and **duration** to vary.

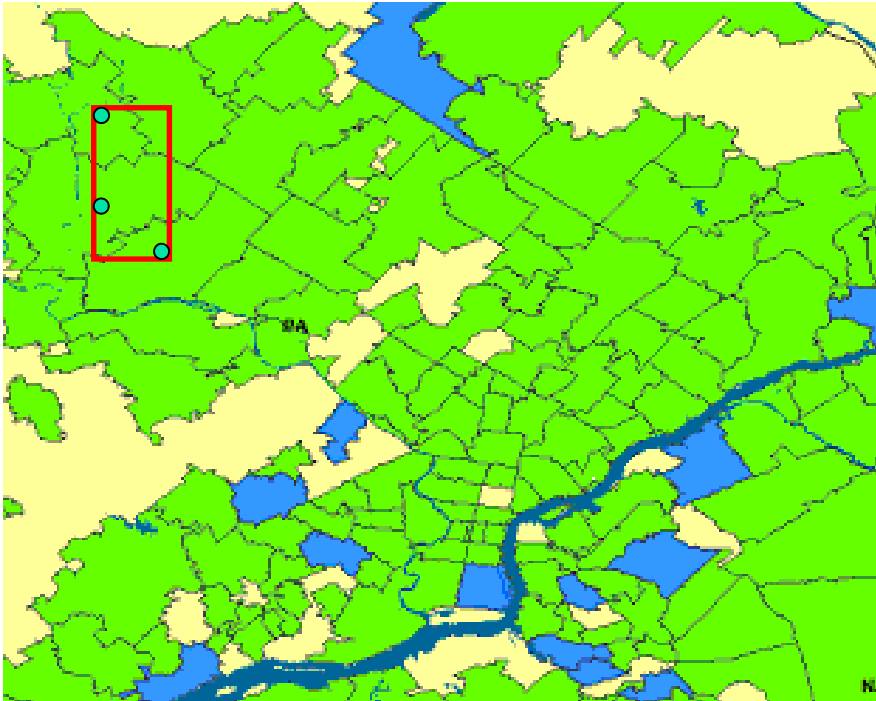
(Consider most recent w days, $w = 1 \dots W_{\max}$)

For each space-time region, we compare the current counts for each location to the time series of historical counts for that location.



Expectation-based scan statistics

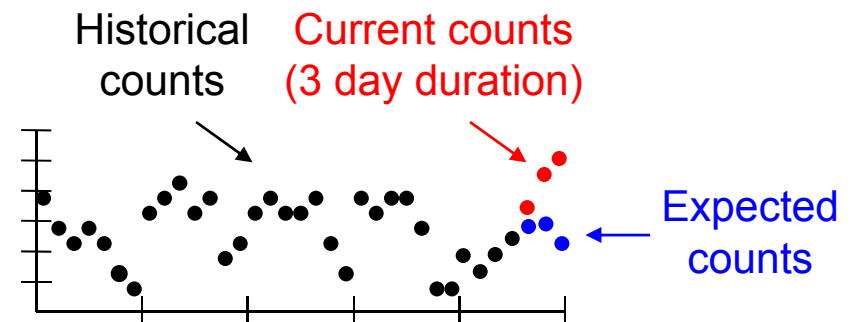
(Kulldorff, 2001; Neill et al., KDD 2005)



For the standard scan statistic approach, we assume that each count is drawn from a Poisson distribution with unknown mean.

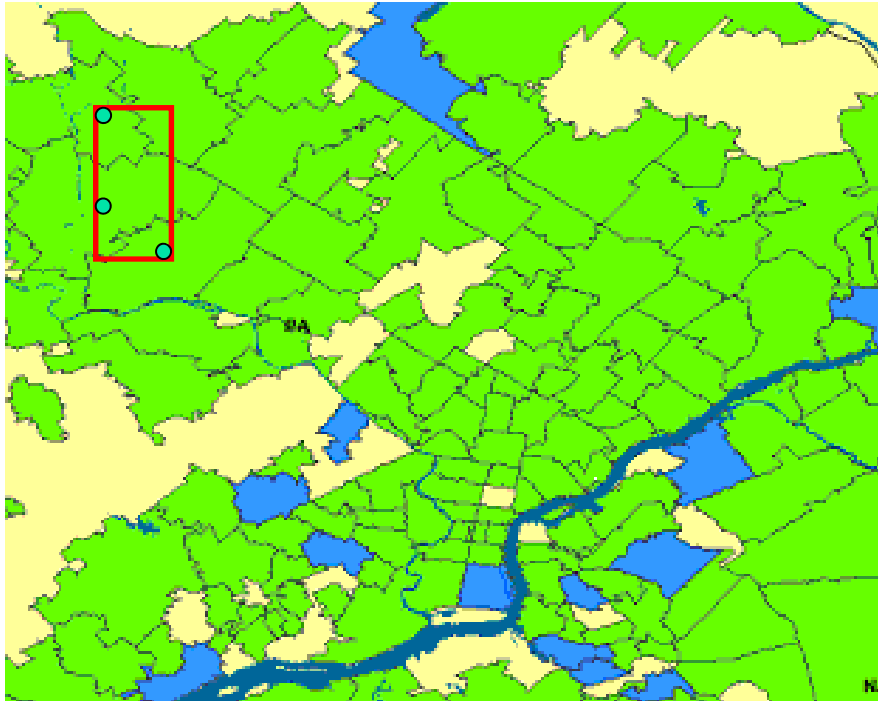
We perform time series analysis to find the expected counts for each recent day, then compare actual to expected counts.

For each space-time region, we compare the current counts for each location to the time series of historical counts for that location.



Expectation-based scan statistics

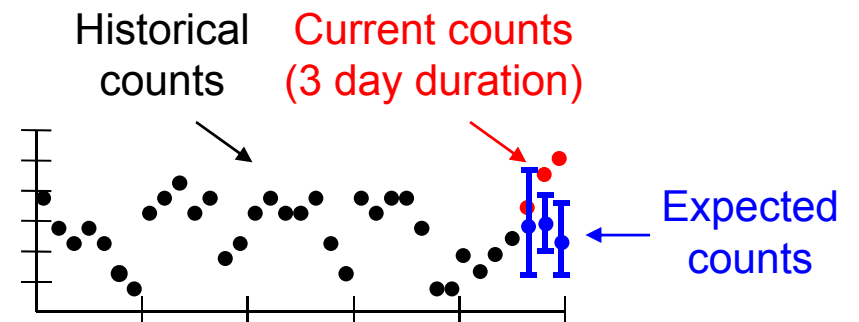
(Kulldorff, 2001; Neill et al., KDD 2005)



For the standard scan statistic approach, we assume that each count is drawn from a Poisson distribution with unknown mean.

Similarly, we can compute a Gaussian scan statistic by obtaining the expectations and variances from historical data.

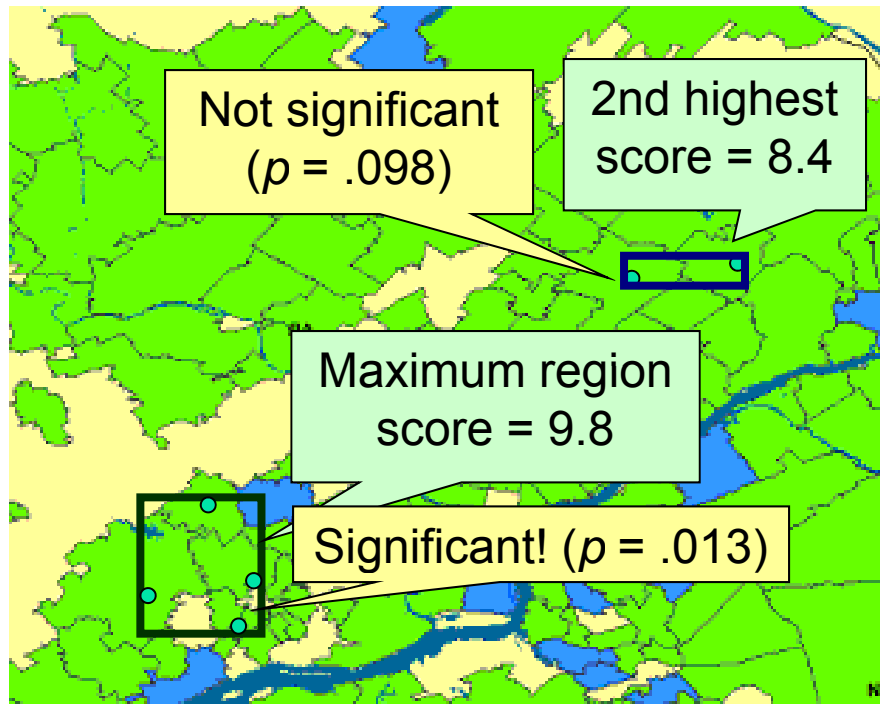
For each space-time region, we compare the current counts for each location to the time series of historical counts for that location.



Expectation-based scan statistics

(Kulldorff, 2001; Neill et al., KDD 2005)

As before, we find the regions with highest values of the likelihood ratio statistic, and compute the p -value of each region by randomization.



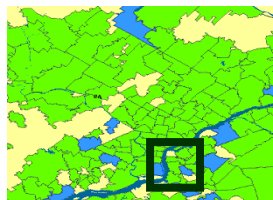
Alternative hypothesis:
event in region S

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

Null hypothesis:
no events

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

$$F_1^* = 2.4$$

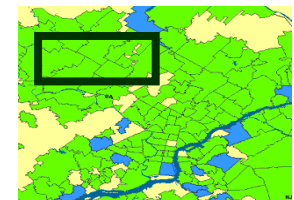


$$F_2^* = 9.1$$



...

$$F_{999}^* = 7.0$$



Poisson scan statistic models

Counts are Poisson distributed: $c_i^t \sim \text{Poisson}(q_i^t b_i^t)$ —

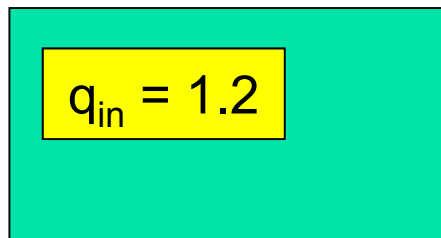
q_i^t is relative risk,
 b_i^t is expected
count under H_0

Expectation-based Poisson (EBP)

(Neill et al., KDD 2005)

H_0 : $q_i^t = 1$ everywhere
(counts = expected)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = 1$
outside, for some $q_{in} > 1$.
(counts > expected in S)

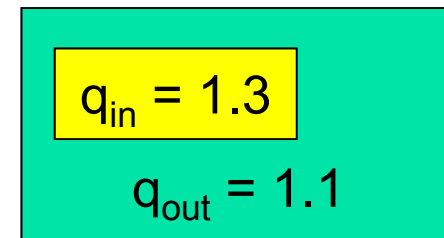


Population-based Poisson (PBP)

(Kulldorff, 1997, 2001)

H_0 : $q_i^t = q_{all}$ everywhere
(inside = outside)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = q_{out}$
outside, for some $q_{in} > q_{out}$.
(inside > outside)



Poisson scan statistic models

Counts are Poisson distributed: $c_i^t \sim \text{Poisson}(q_i^t b_i^t)$ —

q_i^t is relative risk,
 b_i^t is expected
count under H_0

Expectation-based Poisson (EBP)

(Neill et al., KDD 2005)

H_0 : $q_i^t = 1$ everywhere
(counts = expected)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = 1$
outside, for some $q_{in} > 1$.
(counts > expected in S)

$$F(S) = \left(\frac{C}{B}\right)^C e^{B-C}$$

(if $C > B$)

Population-based Poisson (PBP)

(Kulldorff, 1997, 2001)

H_0 : $q_i^t = q_{all}$ everywhere
(inside = outside)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = q_{out}$
outside, for some $q_{in} > q_{out}$.
(inside > outside)

$$F(S) = \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}} \left(\frac{C_{all}}{B_{all}}\right)^{-C_{all}}$$

(if $C_{in} / B_{in} > C_{out} / B_{out}$)

Gaussian scan statistic models

Counts are Gaussian distributed: $c_i^t \sim \text{Gaussian}(q_i^t b_i^t, \sigma_i^t)$

Let $C' = \sum c_i^t b_i^t / (\sigma_i^t)^2$ and $B' = \sum (b_i^t)^2 / (\sigma_i^t)^2$

Expectation-based Gaussian (EBG)

(Neill, Ph.D. thesis, 2006)

H_0 : $q_i^t = 1$ everywhere
(counts = expected)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = 1$
outside, for some $q_{in} > 1$.
(counts > expected in S)

$$F(S) = \exp\left(\frac{(C')^2}{2B'} + \frac{B'}{2} - C'\right)$$

(if $C' > B'$)

Population-based Gaussian (PBG)

(Neill, Ph.D. thesis, 2006)

H_0 : $q_i^t = q_{all}$ everywhere
(inside = outside)

$H_1(S)$: $q_i^t = q_{in}$ in S and $q_i^t = q_{out}$
outside, for some $q_{in} > q_{out}$.
(inside > outside)

$$F(S) = \exp\left(\frac{(C'_{in})^2}{2B'_{in}} + \frac{(C'_{out})^2}{2B'_{out}} - \frac{(C'_{all})^2}{2B'_{all}}\right)$$

(if $C'_{in} / B'_{in} > C'_{out} / B'_{out}$)

Comparison of models and methods

- Expectation-based space-time scan statistics typically outperform purely spatial and purely temporal scans¹ and parallel monitoring².
- EBP and EBG statistics have consistently high detection power whether the affected region is large or small in size.³
- Kulldorff's statistic (PBP) has very low detection power for large regions. For small regions, PBP beats EBP and EBG for large-count datasets, while EBP wins for small-count datasets.³
- Different time series methods are best for computing baselines for different datasets; it is important to adjust for seasonal and day-of-week trends if these are present.^{2,3}
- Randomization testing is often miscalibrated for public health datasets, resulting in lower detection power and high false positive rates. We suggest using the empirical distribution of maximum scores from historical data instead.²
- Bayesian^{4,5} and nonparametric⁶ approaches often outperform typical parametric scan statistics (more on these later).

¹Neill, Ph.D. thesis, 2006 ⁴Neill et al., NIPS 2005

²Neill, IJF, 2009

⁵Neill and Cooper, MLJ, 2009

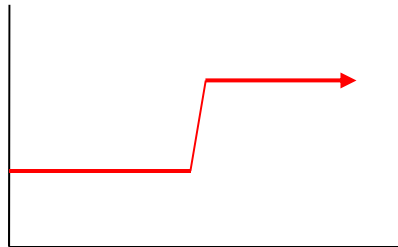
³Neill, IJHG, 2009

⁶Neill and Lingwall, ISDS 2007

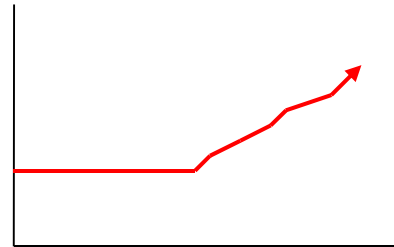
Persistent vs. emerging clusters

Most space-time scan approaches assume that the relative risks q_i^t are spatially uniform over the affected region, and constant over the duration of the event.

Good for detecting
persistent clusters
(e.g. shift in mean)



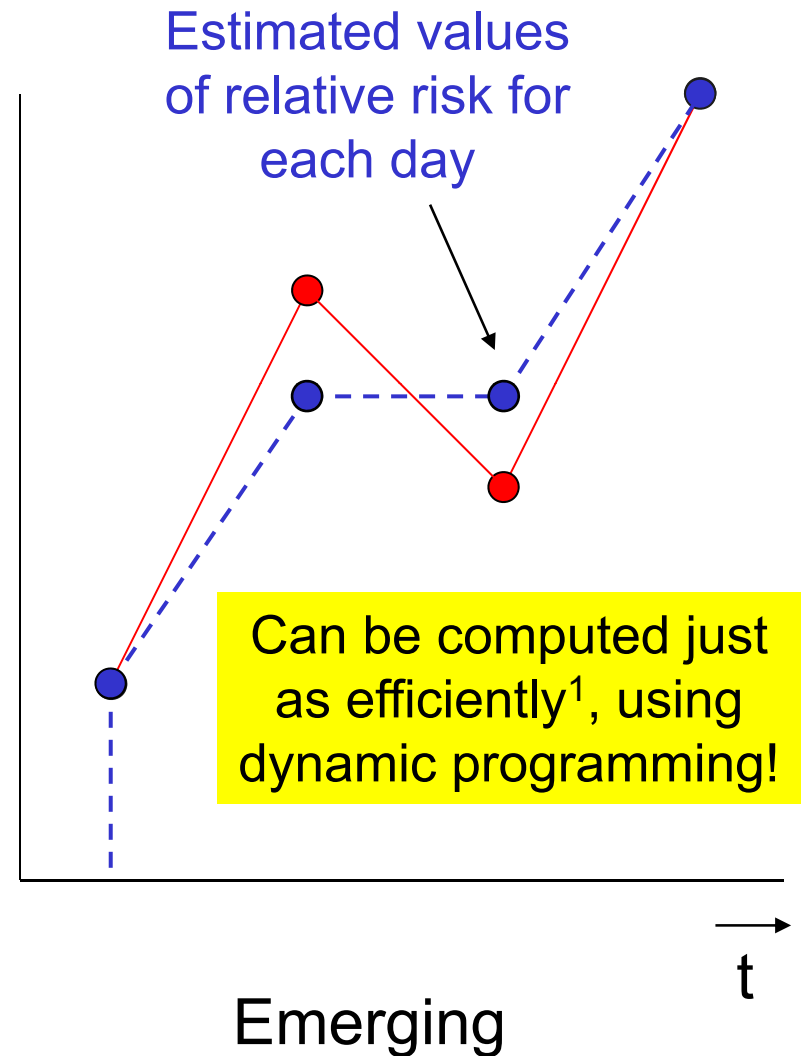
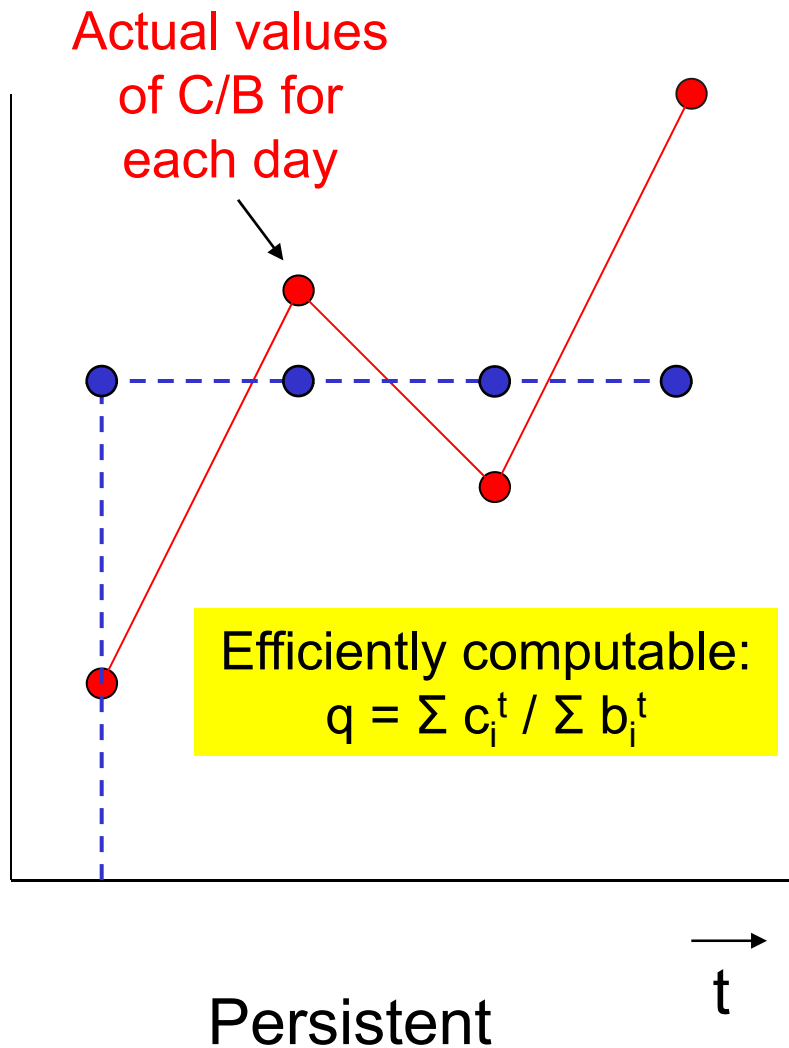
Not as good for detecting
clusters that **emerge**
gradually over time



Better idea: assume that relative risk is non-decreasing over the duration of the event.¹

¹Neill et al., KDD 2005.

Persistent vs. emerging example



¹Neill et al., KDD 2005.

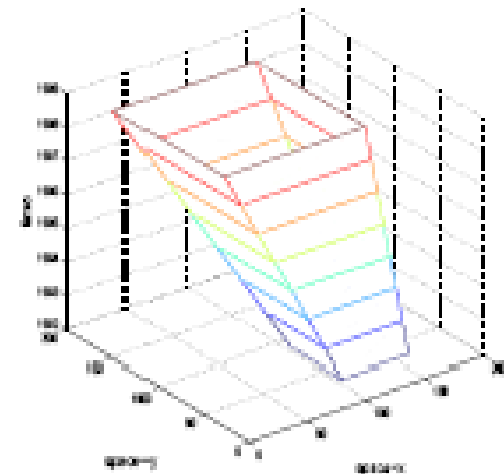
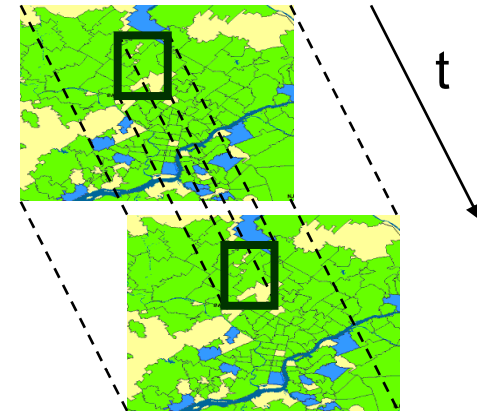
Static vs. dynamic clusters

Most space-time scan approaches assume that the affected set of spatial locations remains constant over time.

We can think of this as a search over regions shaped like right prisms (with both bases the same) in 3-D space-time.

Iyengar (KDD 2004) considers regions with truncated pyramidal shapes in space-time. This models regions which move, grow, or shrink linearly over time.

Exact search is computationally infeasible; heuristic search is used to obtain an approximate solution.

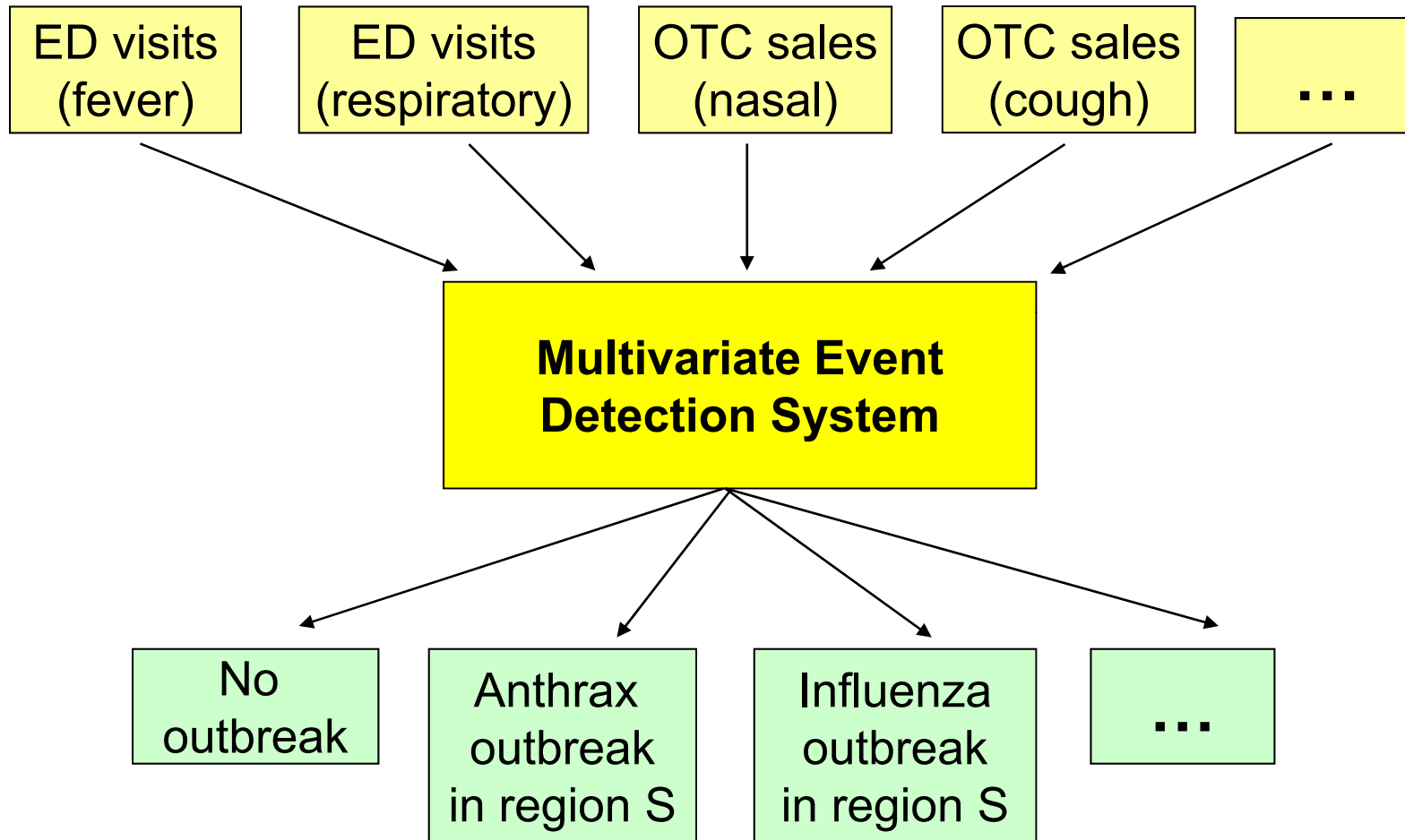


From Iyengar, KDD 2004

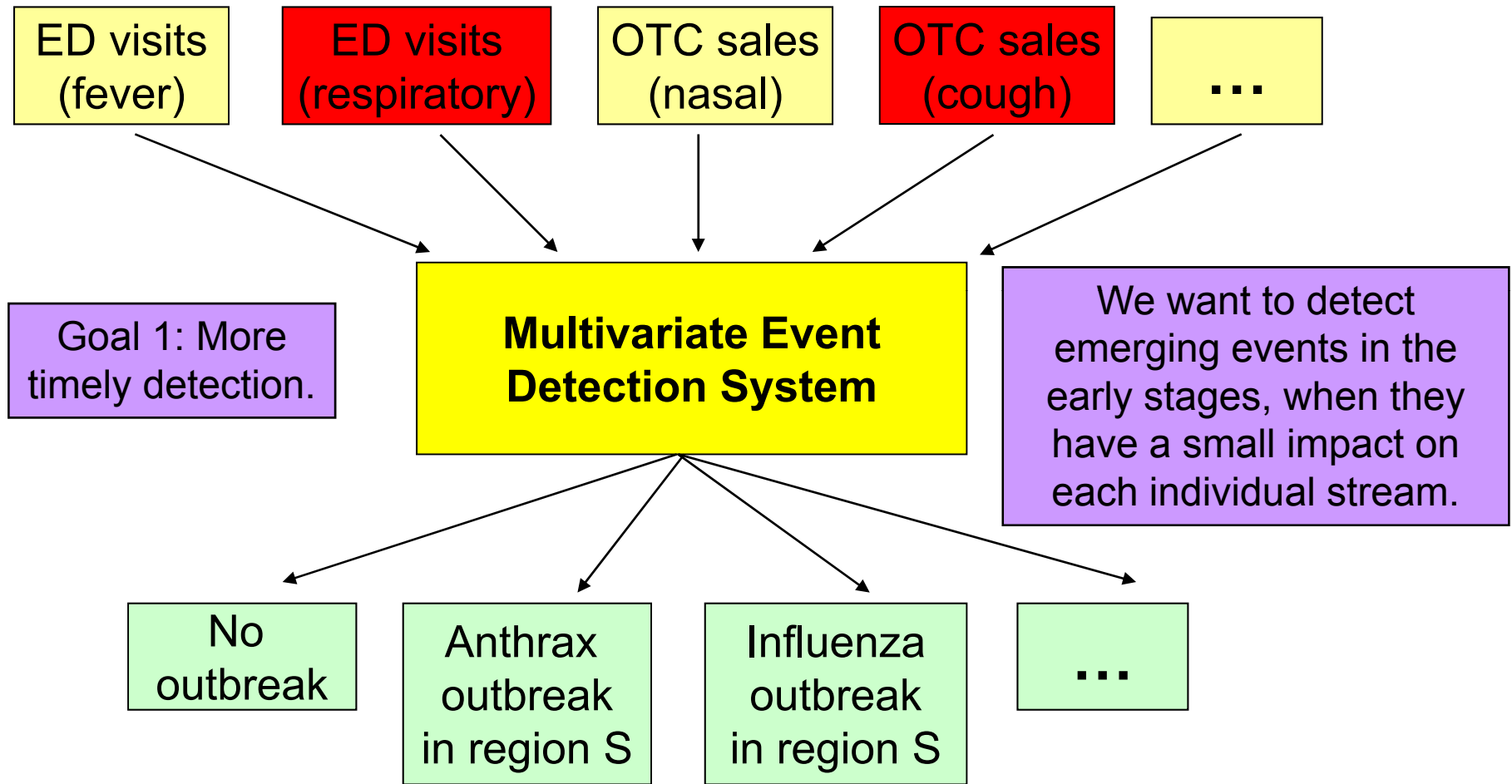
C. Multivariate Scan Statistic Approaches

1. Advantages of multivariate approaches
2. Parametric multivariate scan statistics
3. Non-parametric scan statistics (NPSS)
4. Multivariate Bayesian scan statistics (MBSS)

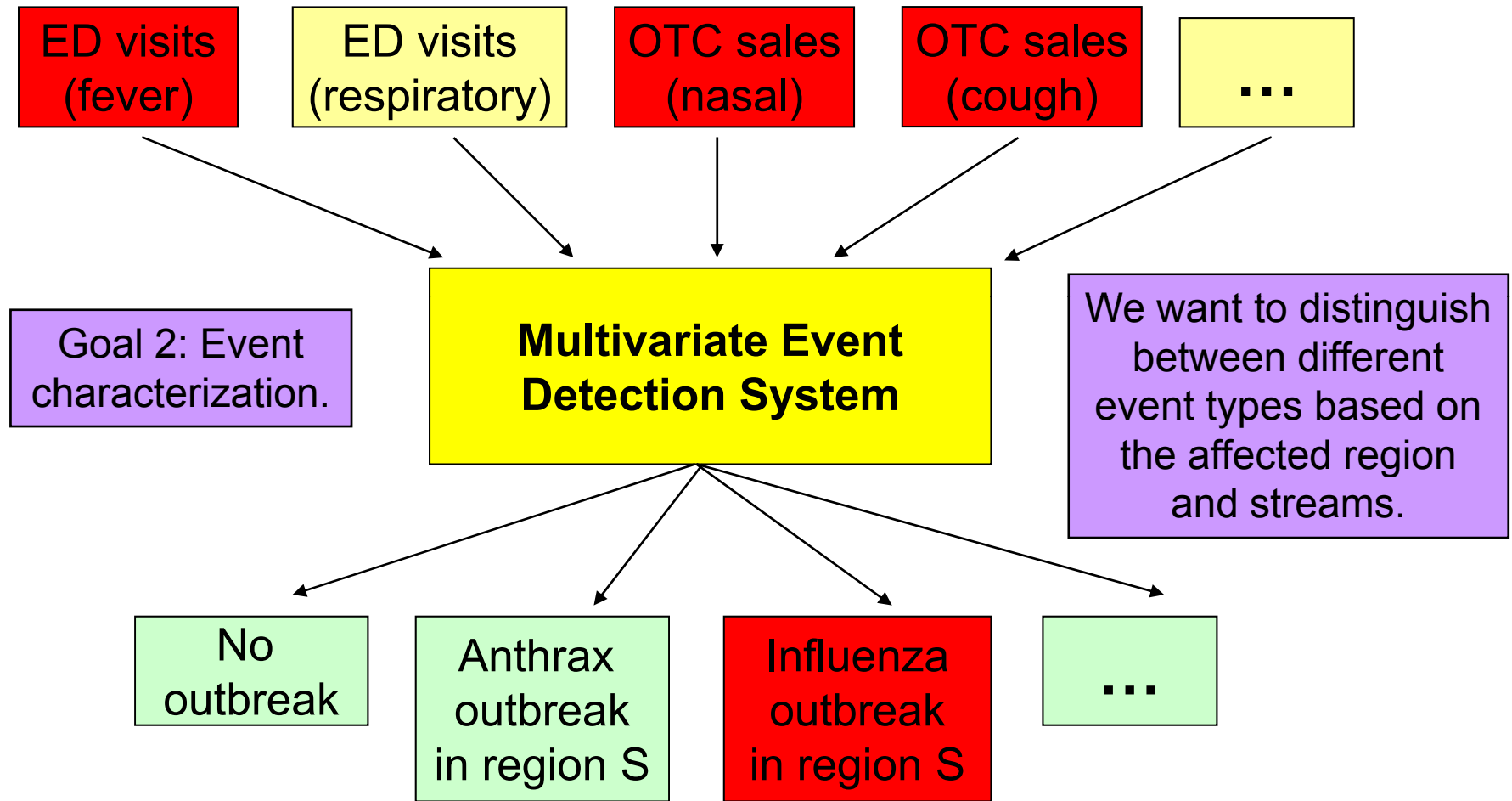
Why multivariate event detection?



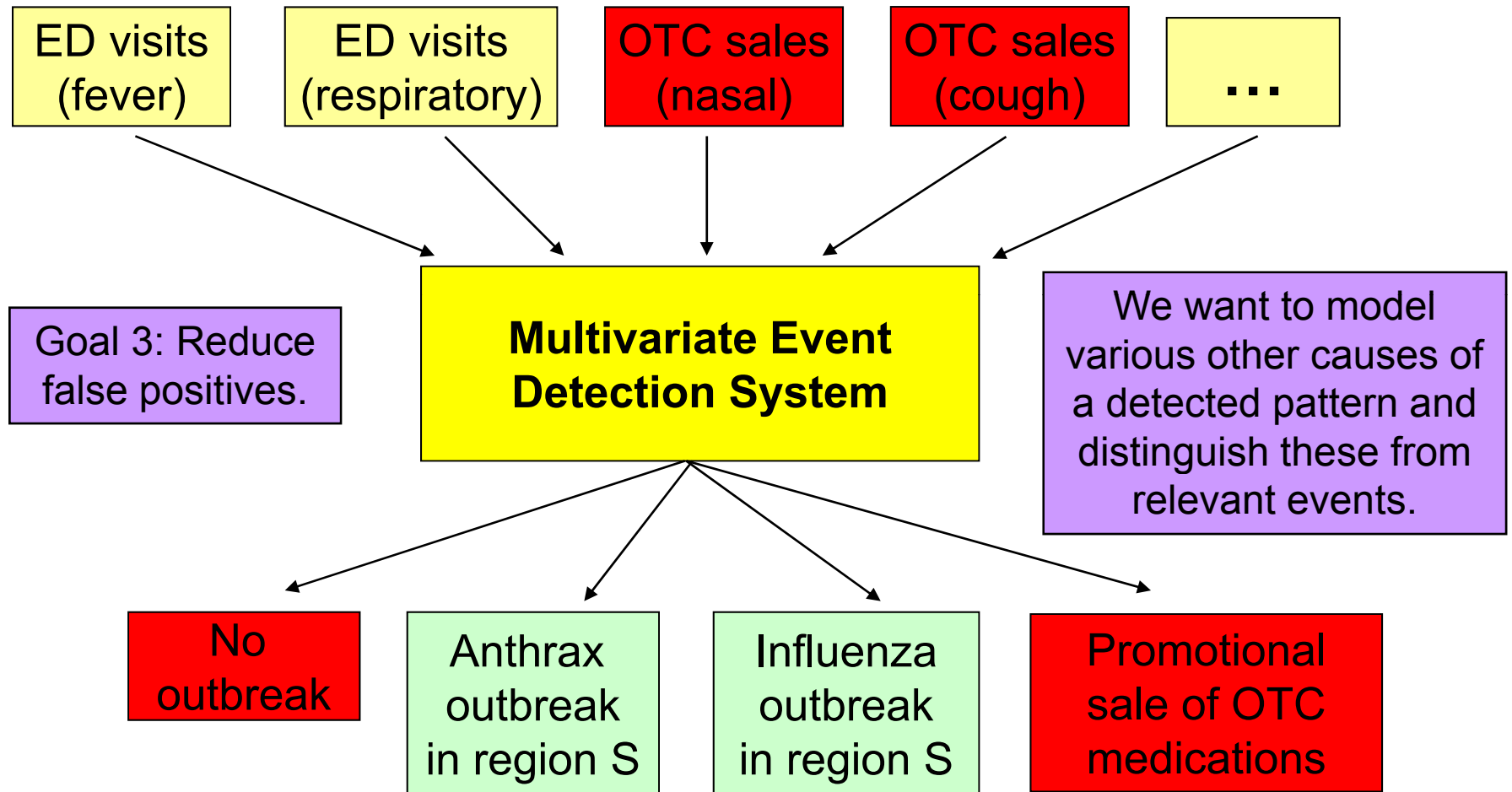
Why multivariate event detection?



Why multivariate event detection?



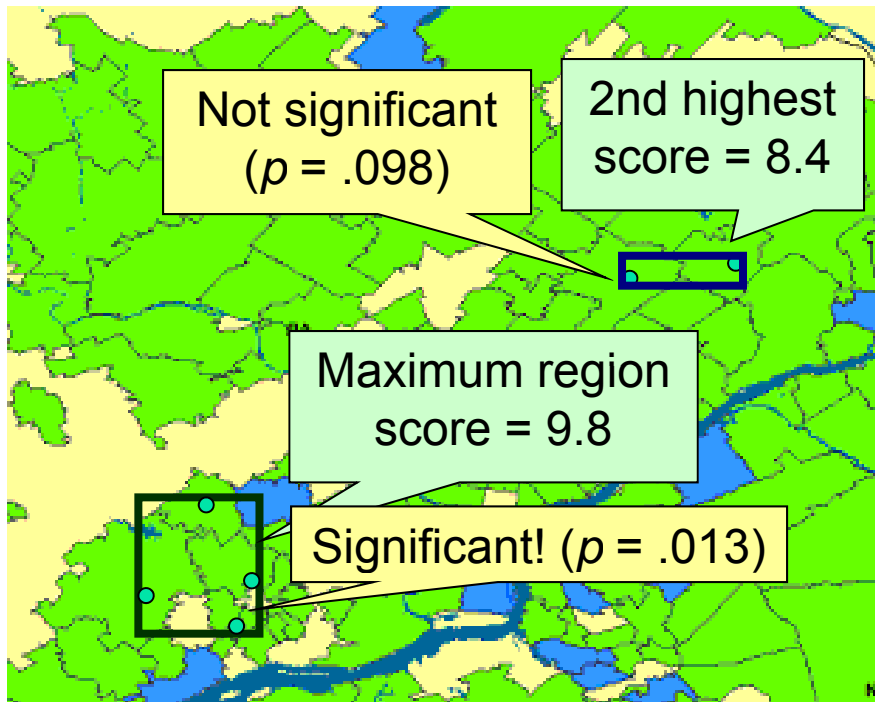
Why multivariate event detection?



C. Multivariate Scan Statistic Approaches

1. Advantages of multivariate approaches
2. **Parametric multivariate scan statistics**
3. Non-parametric scan statistics (NPSS)
4. Multivariate Bayesian scan statistics (MBSS)

Parametric scan statistics



Parametric scan statistics find the regions with highest values of a likelihood ratio statistic, and compute statistical significance of each region by randomization.

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

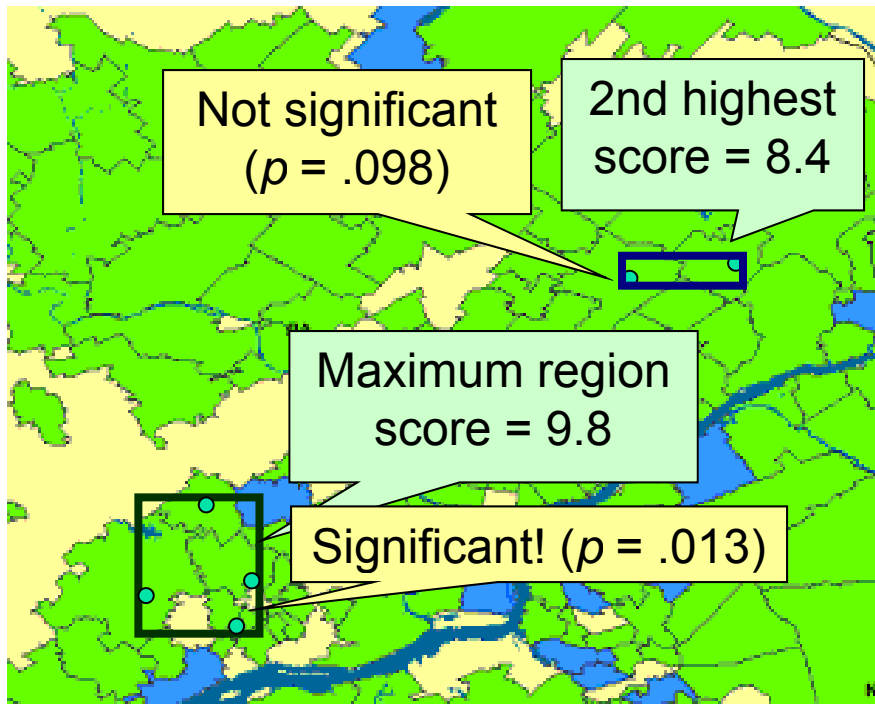
Alternative hypothesis:
outbreak in region S

Null hypothesis:
no outbreak

Typical multivariate approach¹: assume streams are **independent**, conditional on whether an event has occurred and the affected space-time region S.

¹Kulldorff et al., Stat. Med., 2007

Parametric scan statistics



Typical multivariate approach¹: assume streams are **independent**, conditional on whether an event has occurred and the affected space-time region S.

Parametric scan statistics find the regions with highest values of a likelihood ratio statistic, and compute statistical significance of each region by randomization.

$$F(S) = \frac{\Pr(\text{Data} | H_1(S))}{\Pr(\text{Data} | H_0)}$$

Alternative hypothesis:
outbreak in region S

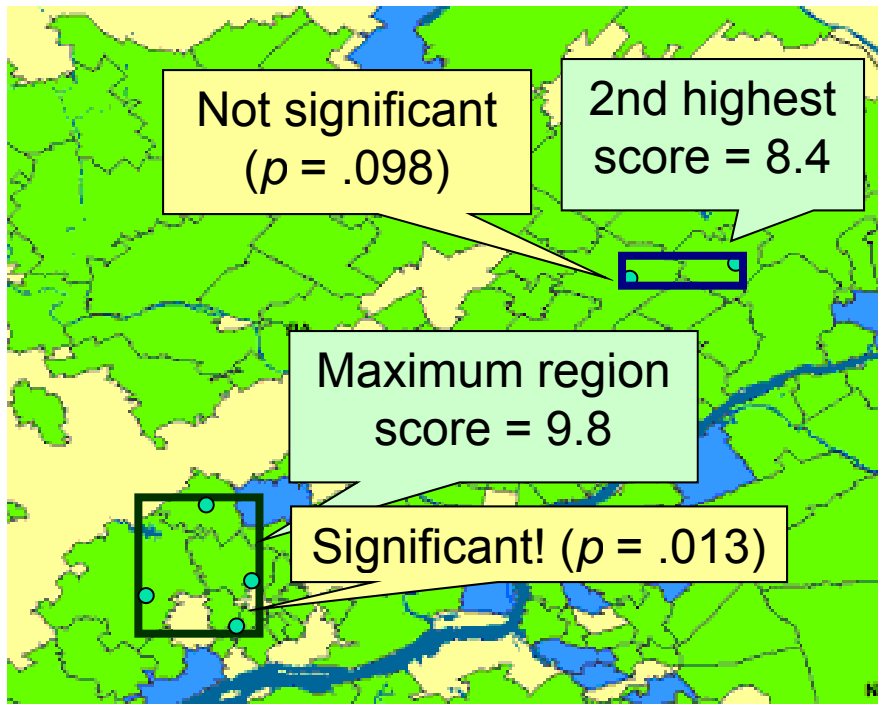
Null hypothesis:
no outbreak

Under this assumption, we can multiply the likelihood ratios for each stream:

$$F(S) = \prod_{m=1}^M \frac{\Pr(D_m | H_1(S))}{\Pr(D_m | H_0)}$$

¹Kulldorff et al., Stat. Med., 2007

Parametric scan statistics



Typical multivariate approach¹: assume streams are **independent**, conditional on whether an event has occurred and the affected space-time region S .

This multivariate approach has several disadvantages:

- 1) Does not account for correlations between streams.
- 2) Cannot determine which subset of streams have been affected.
- 3) Tends to focus detection on streams with highest counts.
- 4) Cannot distinguish between multiple **types** of event.

Under this assumption, we can multiply the likelihood ratios for each stream:

$$F(S) = \prod_{m=1}^M \frac{\Pr(D_m | H_1(S))}{\Pr(D_m | H_0)}$$

¹Kulldorff et al., Stat. Med., 2007

C. Multivariate Scan Statistic Approaches

1. Advantages of multivariate approaches
2. Parametric multivariate scan statistics
3. **Non-parametric scan statistics (NPSS)**
4. Multivariate Bayesian scan statistics (MBSS)

The nonparametric scan statistic

Neill and Lingwall, ISDS 2007

Rather than assuming a parametric distribution and learning the mean and variance parameters from past counts, NPSS compares the current counts to the entire empirical distribution of historical counts.

Simple assumption: under H_0 , all counts for a given location and data stream are drawn independently from the same distribution.

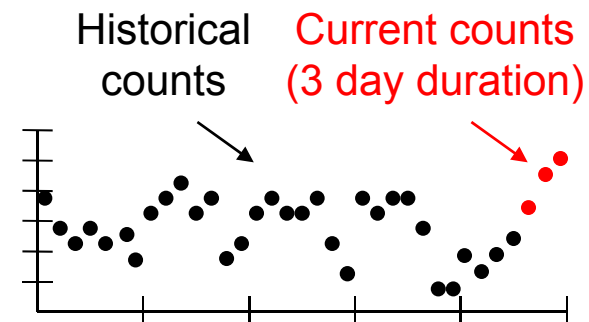
In this case, the **proportion** of historical counts that are greater than current count $c_{i,m}^t$ will be asymptotically uniformly distributed on $[0,1]$.

Compute the empirical p-value $p_{i,m}^t$ corresponding to each current count $c_{i,m}^t$:

$$p_{i,m}^t = (T_{\text{beat}} + 1) / (T + 1)$$

of historical counts $> c_{i,m}^t$

Total # of historical counts



The nonparametric scan statistic

Neill and Lingwall, ISDS 2007

Rather than assuming a parametric distribution and learning the mean and variance parameters from past counts, NPSS compares the current counts to the entire empirical distribution of historical counts.

Simple assumption: under H_0 , all counts for a given location and data stream are drawn independently from the same distribution.

In this case, the **proportion** of historical counts that are greater than current count $c_{i,m}^t$ will be asymptotically uniformly distributed on $[0,1]$.

Compute the empirical p-value $p_{i,m}^t$ corresponding to each current count $c_{i,m}^t$:

Under H_0 , $p_{i,m}^t \sim U[0,1]$

$$p_{i,m}^t = (T_{\text{beat}} + 1) / (T + 1)$$

of historical counts $> c_{i,m}^t$

Total # of historical counts

Under $H_1(S)$, the counts in region S will be higher than expected under H_0 , and thus the empirical p-values will be lower than expected.

The nonparametric scan statistic

We search for regions (D, S, W) with a surprisingly large number of low empirical p-values.

← D: subset of data streams
S: set of spatial locations
W: duration (number of days)

Total number of empirical p-values in region: $N = |D| \times |S| \times W$

How many low empirical p-values ($p_{i,m}^t < \alpha$) do we expect under H_0 ?

Let $N_\alpha = \# \{p_{i,m}^t < \alpha\}$. Then $N_\alpha \sim \text{Binomial}(N, \alpha)$, with mean $N\alpha$ and variance $N\alpha(1 - \alpha)$.

Following Donoho and Jin (2004), we define the higher criticism statistic $F(D, S, W) = \max_\alpha (N_\alpha - N\alpha) / \sqrt{N\alpha(1 - \alpha)}$.

We find the multivariate space-time regions (D, S, W) with highest scores $F(D, S, W)$, and compute statistical significance by randomization.

The nonparametric scan statistic

Advantages of the nonparametric scan statistic (NPSS)

No parametric model assumptions.

Can easily combine information from multiple streams and identify which subset of streams are affected.

Randomization testing is easy (draw each $p_{i,m}^t \sim U[0,1]$).

NPSS assumes that all of the counts for a given time series are drawn from the same (unknown) distribution, which will not be true if the time series is nonstationary.

Solution: use the standardized residuals $r_{i,m}^t = (c_{i,m}^t - b_{i,m}^t) / \sqrt{b_{i,m}^t}$, where the expected counts $b_{i,m}^t$ are inferred by time series analysis.

Other nonparametric score functions $F(D, S, W) = \max_{\alpha} F_{\alpha}(N_{\alpha}, N)$ can be defined, and in some cases these outperform higher criticism.

Preliminary comparison of methods

- We compared the parametric and nonparametric scan statistics on a variety of outbreak detection tasks using Emergency Department data from Allegheny County.
- Univariate detection performance was comparable; NPSS outperformed parametric scans for larger outbreaks, and for data that did not fit the parametric model assumptions.
- NPSS achieved significant gains in detection power on multivariate tasks, especially when only a subset of the monitored streams were affected.
- NPSS was able to accurately identify the affected streams.
- A naïve implementation of NPSS is more computationally expensive than parametric scan, $O(2^M)$ for M streams.
- However, we have developed an efficient implementation that scales linearly with M , using newly developed methods for **linear-time subset scanning (LTSS)**.¹

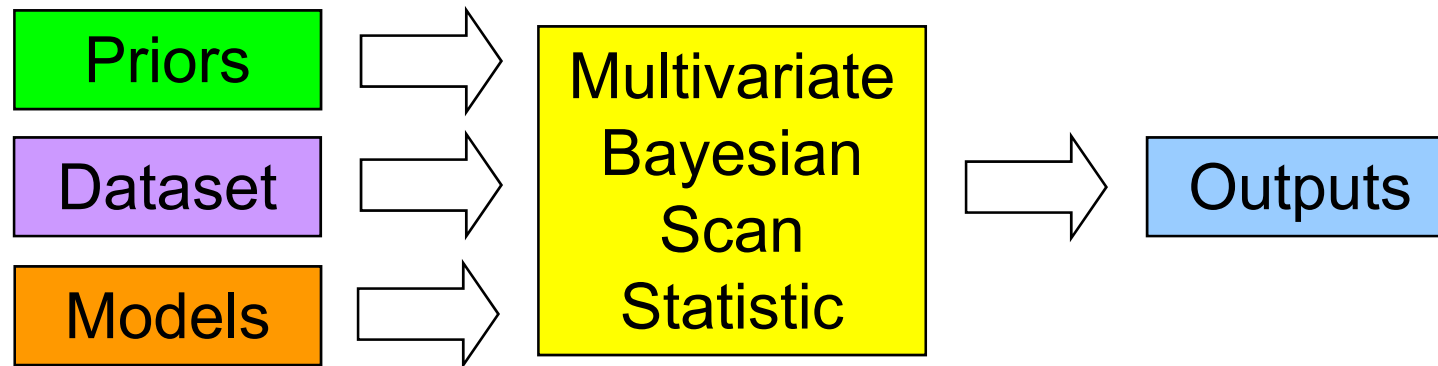
¹Neill, ISDS 2008

C. Multivariate Scan Statistic Approaches

1. Advantages of multivariate approaches
2. Parametric multivariate scan statistics
3. Non-parametric scan statistics (NPSS)
4. **Multivariate Bayesian scan statistics (MBSS)**

Overview of the MBSS method

Neill and Cooper, MLJ, 2009

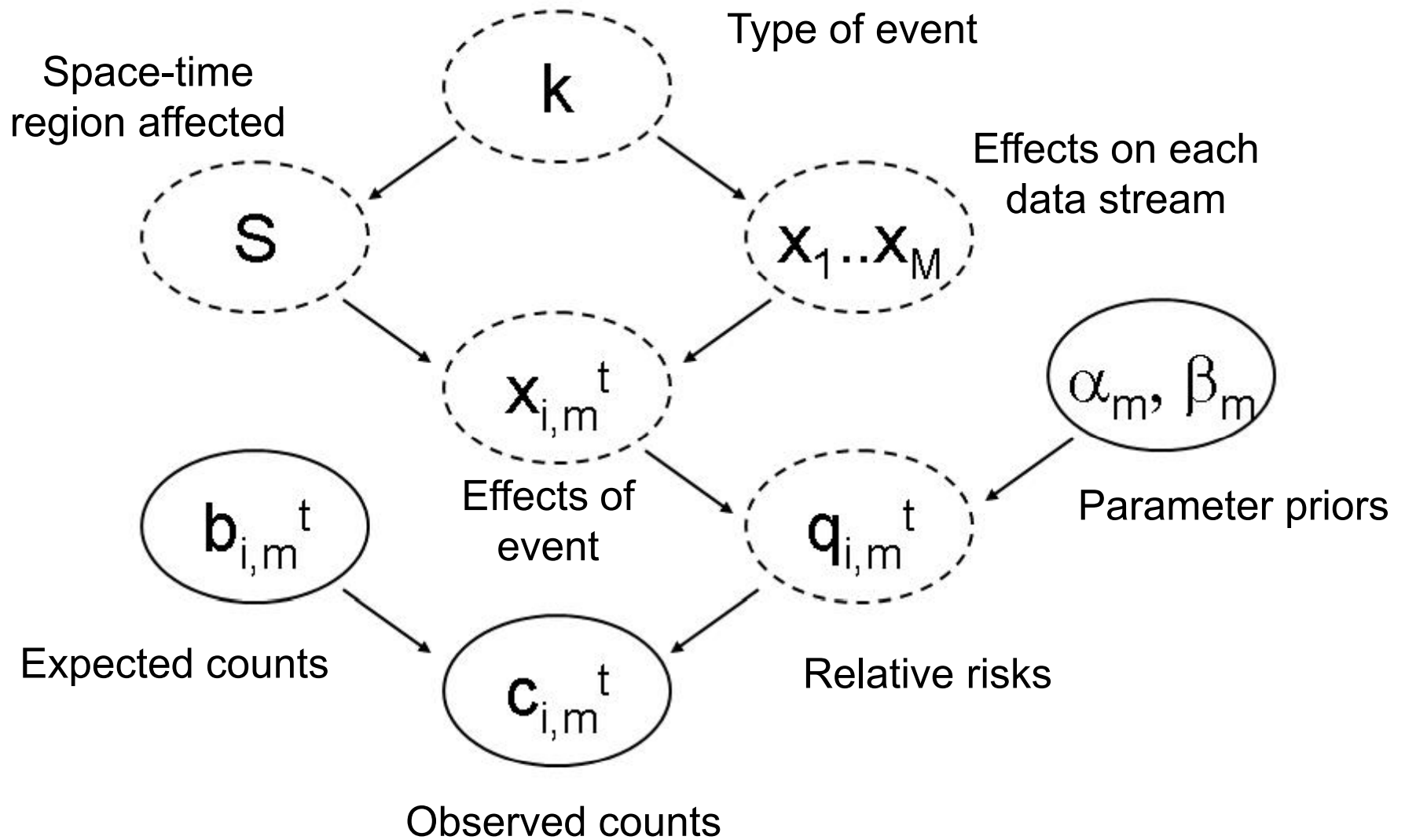


Given a set of event types E_k , a set of spatial regions S , and the multivariate dataset D , MBSS outputs the posterior probability $\Pr(H_1(S, E_k) | D)$ of each type of event in each region, as well as the probability of no event, $\Pr(H_0 | D)$.

We must provide the prior probability $\Pr(H_1(S, E_k))$ of each event type E_k in each region S , as well as the prior probability of no event, $\Pr(H_0)$.

MBSS uses Bayes' Theorem to combine the data likelihood given each hypothesis with the prior probability of that hypothesis: $\Pr(H | D) = \Pr(D | H) \Pr(H) / \Pr(D)$.

The Bayesian hierarchical model



The Bayesian hierarchical model

Count for data stream d_m in location s_i at time t

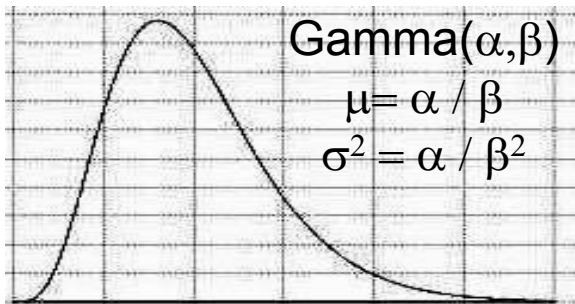
$$c_{i,m}^t \sim \text{Poisson}(q_{i,m}^t b_{i,m}^t)$$

$b_{i,m}^t$ is expected value of $c_{i,m}^t$ under the null hypothesis, predicted from historical data.
 $q_{i,m}^t$ is relative risk.

Null hypothesis H_0
 (no events)

$$q_{i,m}^t \sim \text{Gamma}(\alpha_m, \beta_m) \text{ everywhere}$$

α_m and β_m are learned from historical data for stream d_m .



Alternative hypothesis $H_1(S, E_k)$
 (event of type E_k in region S)

$$q_{i,m}^t \sim \text{Gamma}(x_m \alpha_m, \beta_m) \text{ inside region } S, \\ q_{i,m}^t \sim \text{Gamma}(\alpha_m, \beta_m) \text{ elsewhere}$$

Event type E_k multiplies expected counts in S by some constant x_m for each stream d_m .

$$\text{Simple event model: } x_m = 1 + \theta (x_{km, \text{avg}} - 1)$$

Event severity

Average effect of event type E_k on stream d_m .

Computing Bayesian likelihoods

- **Marginal likelihood** approach: integrate over possible values of the relative risks $q_{i,m}^t$, weighted by their prior probabilities.
- **Conjugate priors** allow a closed form solution.
 - Gamma priors, Poisson counts \rightarrow Negative binomial likelihoods.

Computing Bayesian likelihoods

- **Marginal likelihood** approach: integrate over possible values of the relative risks $q_{i,m}^t$, weighted by their prior probabilities.
- **Conjugate priors** allow a closed form solution.
 - Gamma priors, Poisson counts \rightarrow Negative binomial likelihoods.

$$\Pr(D | H_0) \propto \prod_{i,m,t} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, \alpha_m, \beta_m)$$

where $\text{NegBin}(c, b, \alpha, \beta) = \int \Pr(q \sim \text{Ga}(\alpha, \beta)) \Pr(c \sim \text{Po}(qb)) dq$

$$\propto \frac{\beta^\alpha \Gamma(\alpha + c)}{(\beta + b)^{\alpha+c} \Gamma(\alpha)}$$

Computing Bayesian likelihoods

- **Marginal likelihood** approach: integrate over possible values of the relative risks $q_{i,m}^t$, weighted by their prior probabilities.
- **Conjugate priors** allow a closed form solution.
 - Gamma priors, Poisson counts \rightarrow Negative binomial likelihoods.

$$\Pr(D | H_0) \propto \prod_{i,m,t} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, \alpha_m, \beta_m)$$

$$\begin{aligned} \Pr(D | H_1(S, E_k), \{x_m\}) &\propto \prod_{i,m,t \in S} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, x_m \alpha_m, \beta_m) \\ &\times \prod_{i,m,t \notin S} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, \alpha_m, \beta_m) \end{aligned}$$

where $\text{NegBin}(c, b, \alpha, \beta) = \int \Pr(q \sim \text{Ga}(\alpha, \beta)) \Pr(c \sim \text{Po}(qb)) dq$

$$\propto \frac{\beta^\alpha \Gamma(\alpha + c)}{(\beta + b)^{\alpha+c} \Gamma(\alpha)}$$

Computing Bayesian likelihoods

- **Marginal likelihood** approach: integrate over possible values of the relative risks $q_{i,m}^t$, weighted by their prior probabilities.
- **Conjugate priors** allow a closed form solution.
 - Gamma priors, Poisson counts \rightarrow Negative binomial likelihoods.

$$\Pr(D | H_0) \propto \prod_{i,m,t} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, \alpha_m, \beta_m)$$

$$\begin{aligned} \Pr(D | H_1(S, E_k), \{x_m\}) &\propto \prod_{i,m,t \in S} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, x_m \alpha_m, \beta_m) \\ &\times \prod_{i,m,t \notin S} \text{NegBin}(c_{i,m}^t, b_{i,m}^t, \alpha_m, \beta_m) \end{aligned}$$

To compute the data likelihood given the alternative hypothesis $H_1(S, E_k)$, we marginalize over the values of x_m .

Comparison to prior methods

We compared MBSS to the parametric multivariate scan statistic on outbreak detection using OTC medication sales.

Using uninformative priors, MBSS achieves similar detection performance to parametric scans, enabling its use as a general detector with high performance across many event types.

However, we can also incorporate prior information into event models, and thus use MBSS as a specific detector with much higher detection power for the given event type, achieving an average of 1.3 days faster detection than parametric scans.

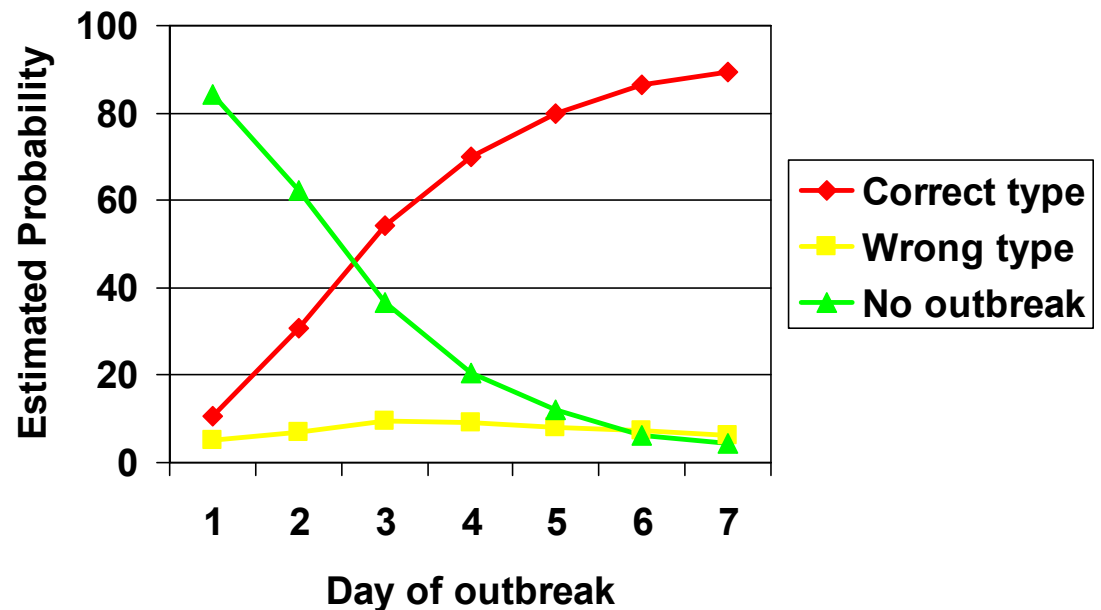
Additionally, MBSS can be used to characterize events by specifying models for multiple event types and computing the probability that each type of event has occurred.

Testing discrimination power

- We examined the ability of MBSS to differentiate between two types of influenza-like illness using two streams of OTC data (cough/cold, antifever).

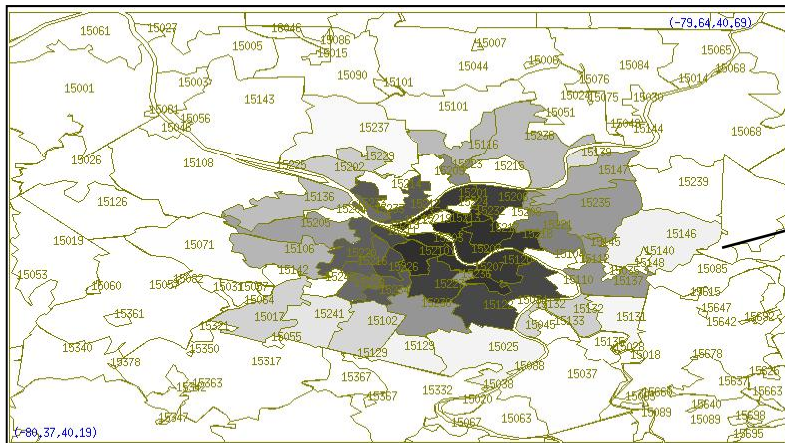
Outbreak E_1 affects cough/cold twice as much as fever.
Outbreak E_2 affects fever twice as much as cough/cold.

MBSS was able to accurately discriminate between the two event types by the second outbreak day.



Interpretation and visualization

- MBSS gives the total posterior probability of each event type E_k , and the distribution of this probability over space-time regions S .
- Probabilistic basis for decision-making, given costs of false positives and false negatives.
- Visualization: $\Pr(H_1(s_i, E_k)) = \sum \Pr(H_1(S, E_k))$ for all regions S containing location s_i .



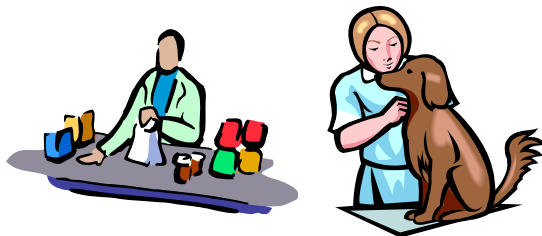
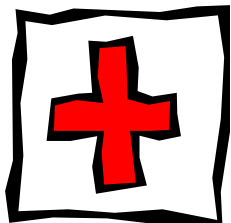
Total posterior probability of a respiratory outbreak in each Allegheny County zip code on 6/3/05. Darker shading = higher probability.

Advantages of MBSS

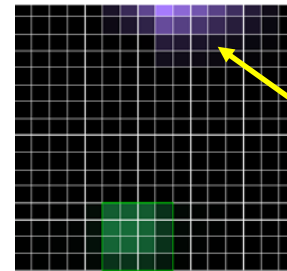
Can incorporate prior knowledge of event prevalence, size, shape, duration, spread, and impact.

Computation is fast in the Bayesian framework, and randomization testing is not necessary.

We can detect faster and more accurately by integrating multiple data streams.



Results are easy to interpret, visualize, and use for decision-making.



$P(\text{anthrax}) = 22\%$
 $P(\text{influenza}) = 13\%$
 $P(\text{other ILI}) = 33\%$

We can model and differentiate between multiple potential causes of an event.

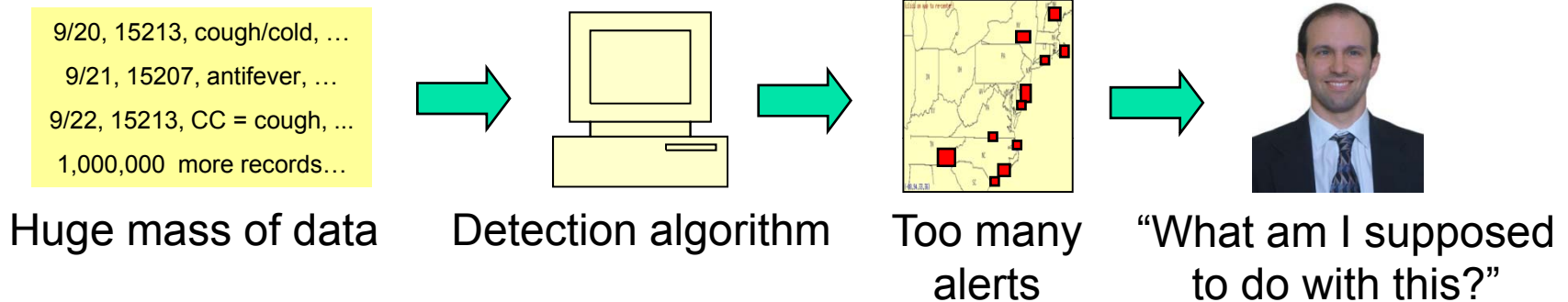


D. Current Directions in Spatial Event Detection

- ➔ 1. Incorporating learning into detection
- ➔ 2. Very fast detection algorithms
- 3. Generalization of spatial methods to non-spatial data
- 4. Non-aggregated spatial and temporal data
- 5. Incorporating rich information from observations
- 6. Detecting multiple, dynamic, irregular clusters
- 7. Integrating detection, tracking, and response
- 8. Combining sensor placement and sensor fusion

Incorporating learning into detection

We have made major advances in detecting anomalous patterns, but not in determining which of these anomalies are relevant.

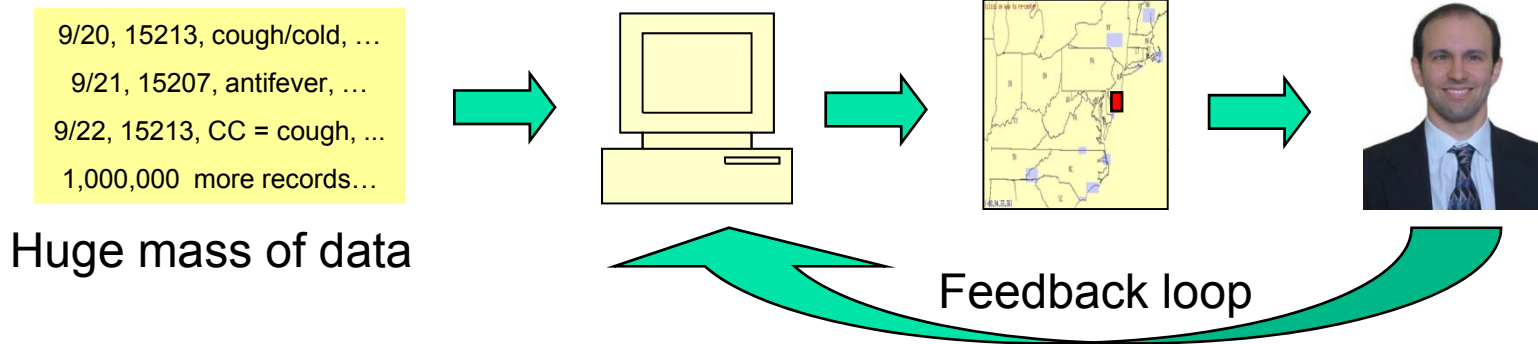


We must model and differentiate between multiple causes of a detected pattern, and provide only the relevant patterns to the user.

How can we classify patterns, and determine which ones are relevant to a given user at a given time?

Incorporating learning into detection

We have made major advances in detecting anomalous patterns, but not in determining which of these anomalies are relevant.



We must model and differentiate between multiple causes of a detected pattern, and provide only the relevant patterns to the user.

How can we classify patterns, and determine which ones are relevant to a given user at a given time?

Incorporate user feedback into the detection process, and use it to learn models!

Incorporating learning into MBSS

Many aspects of the MBSS framework can be learned from data.

The set of event types E_k , and the prevalence of each event type $\Pr(E_k)$.

The space-time pattern of each event type, $\Pr(H_1(S, E_k) | E_k)$.

The effects of each event type on the multiple data streams, $\Pr(D | H_1(S, E_k))$.

The relevance of each type of event to the user.

We first consider the passive learning scenario, in which the user provides a label (S, E_k) for each day.

This label can be then used by the system to update its models and improve its performance for future days.

Incorporating learning into MBSS

Many aspects of the MBSS framework can be learned from data.

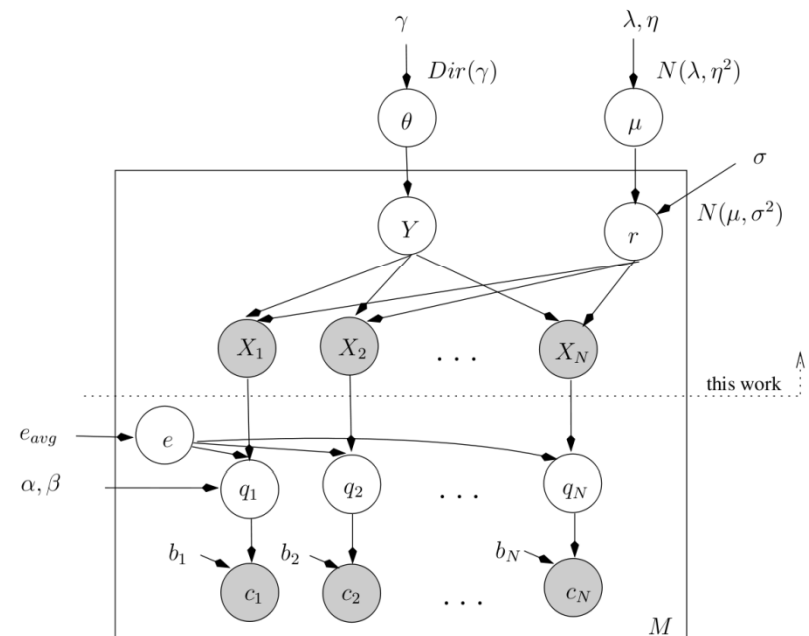
The set of event types E_k , and the prevalence of each event type $\Pr(E_k)$.

The space-time pattern of each event type, $\Pr(H_1(S, E_k) | E_k)$.

The effects of each event type on the multiple data streams, $\Pr(D | H_1(S, E_k))$.

The relevance of each type of event to the user.

We need to generalize over huge # of possible regions S .



Latent center model
(Makatchev and Neill, 2008)

Incorporating learning into MBSS

Many aspects of the MBSS framework can be learned from data.

The set of event types E_k , and the prevalence of each event type $\Pr(E_k)$.

The space-time pattern of each event type, $\Pr(H_1(S, E_k) | E_k)$.

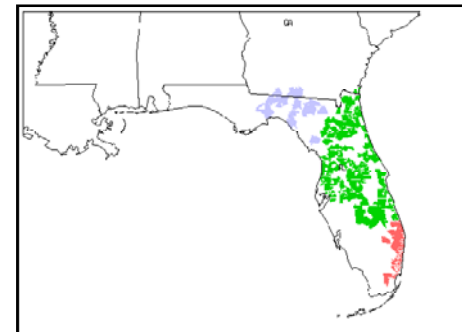
The effects of each event type on the multiple data streams, $\Pr(D | H_1(S, E_k))$.

The relevance of each type of event to the user.

We can learn different temporal patterns for different event types.



We can detect events where the affected region changes over time.



Passive vs. active learning



So far, we have considered the passive learning scenario, in which each day is assigned a label by the user (no event, or event type E_k and affected area S) independently of the system's output.

However, having a user in the loop allows for much interaction and learning than this simple framework. In the active learning scenario, the system presents its output (detected clusters) to the user, and receives feedback on these clusters, each day.

This presents an interesting challenge: the system must strike a balance between exploration, presenting “unknown” examples that will best inform its models, and exploitation, presenting “known” examples that have highest probability of relevance to the user.

Active learning of new event types

This scenario allows the user to define new classes “on the fly”, by assigning a new label type to an example. The system can then find other potential examples of the new class in historical data, ask the user to label these, and learn a model for the new class.



“Any clusters of interest today?”

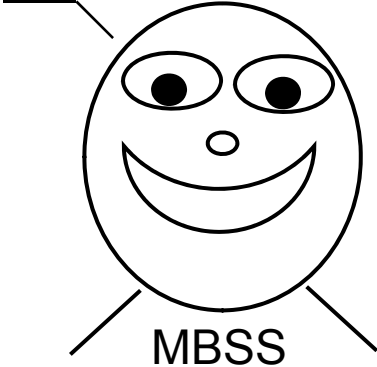
“Yes, this appears to be an anthrax attack, based on increased OTC cough and fever.”

“No, we don’t see a corresponding increase in ED visits. I think this cluster is just a promotional sale.”

“OK. Can you identify which of these historical clusters also correspond to promotional sales?”

“Yes, all of these. Any other clusters of interest?”

“Not really, just more seasonal flu and another promotional sale.”



MBSS

The end goal is to transform the process of pattern discovery into a “conversation” between the user and system, in which the system takes an active role in identifying and explaining potentially interesting patterns.

D. Current Directions in Spatial Event Detection

1. Incorporating learning into detection
2. **Very fast detection algorithms**
3. Generalization of spatial methods to non-spatial data
4. Non-aggregated spatial and temporal data
5. Incorporating rich information from observations
6. Detecting multiple, dynamic, irregular clusters
7. Integrating detection, tracking, and response
8. Combining sensor placement and sensor fusion

Which subsets to scan?

Since there are exponentially many subsets of the data, it is often computationally infeasible to search all of them.

The most common approach is to use domain knowledge to restrict our search space: for example, we assume that an event will affect a contiguous spatial region, and often further restrict the region size and shape.

e.g. “search over circular regions centered at a data point” → only N^2 regions instead of 2^N .

Another common approach is to perform a heuristic search. For example, we can greedily grow subsets starting from each data record, repeatedly adding the additional record that gives the highest scoring subset.¹

Tradeoff: much more efficient than naïve search, but not guaranteed to find highest scoring region.

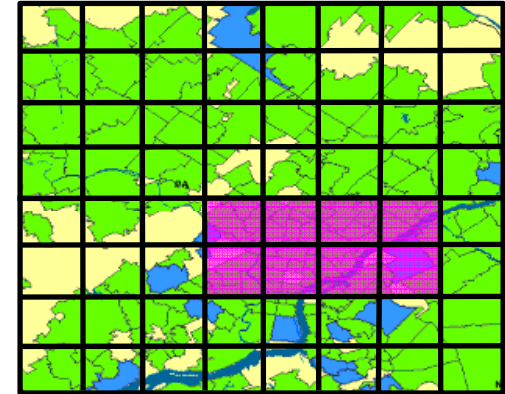
In some cases, we can find the highest-scoring subsets *without* computing the scores of all possible subsets!

¹Neill et al., in *Scan Statistics: Methods and Applications*, 2009.

Fast spatial scan over rectangles

Consider searching over all rectangular regions for data aggregated to a $N \times N$ grid.

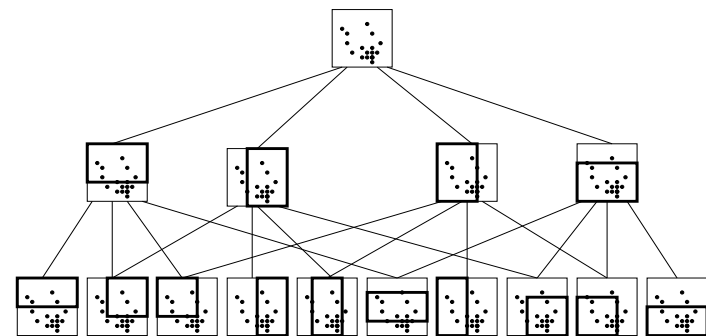
The number of search regions scales as $O(N^4)$, making an exhaustive search computationally infeasible for large N .



We can find the highest scoring clusters without an exhaustive search using **branch and bound**: we keep track of the highest region score found so far, and prune sets of regions with provably lower scores.^{1,2}

A new multi-resolution data structure, the **overlap-kd tree**, enables us to make this search efficient.

We can now monitor nationwide health data in 20 minutes (vs. 1 week).



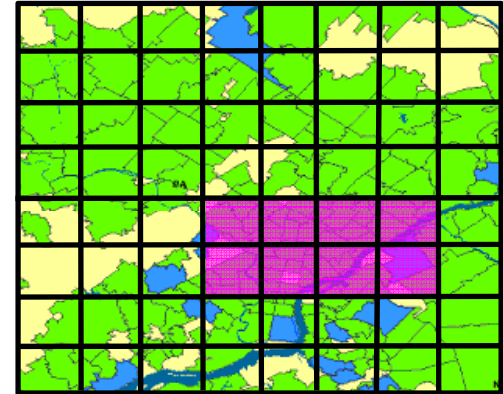
¹Neill and Moore, KDD 2004

²Neill et al., NIPS 2004

Fast spatial scan over rectangles

Consider searching over all rectangular regions for data aggregated to a $N \times N$ grid.

The number of search regions scales as $O(N^4)$, making an exhaustive search computationally infeasible for large N .



We can find the highest scoring clusters without an exhaustive search using **branch and bound**: we keep track of the highest region score found so far, and prune sets of regions with provably lower scores.^{1,2}

Other recent work on efficiently maximizing scan statistics over gridded rectangles

Agarwal et al., SODA 2006, KDD 2006

Fast approximate optimization of convex score functions. Solution within ε of optimal, runtime $O((1/\varepsilon) N^2 \log^2 N)$.

Wu et al., KDD 2009

Compute scores for subset of rectangles, bound scores of other rectangles by **tiling** with evaluated rectangles.

Can be used for non-convex score functions.

Linear-time subset scanning

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

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

¹Neill, ISDS 2008

The LTSS property

- Example: Poisson statistics (Kulldorff, EBP)
 - $F(S) = F(C, B)$, where $C = \sum c_i$ and $B = \sum b_i$ are the aggregate count and baseline of region S .
 - Sort 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 consists of the locations $s_{(1)} \dots s_{(k)}$ for some k , $1 \leq k \leq N$.
 - This follows from the facts that $F(S)$ is convex, increasing with C and decreasing with B .

How to use LTSS in practice?

- Simplest case: assume all subsets are equally likely (e.g. outbreak that does not cluster spatially)
 - LTSS gives highest-scoring subset by evaluating N subsets instead of 2^N for naïve search.
- But what if we want to use spatial information to constrain our search over subsets?
 - Soft constraints: some subsets of locations are more likely than others (non-uniform priors).
 - Hard constraints: some subsets of locations are not allowed (e.g. non-contiguous or highly irregular regions).
- In most cases, we cannot use LTSS directly to find the optimal subset subject to these constraints.

How to use LTSS in practice?

- We can use LTSS to speed up the constrained search problem in three ways:
 - 1) For some hard constraints, we can compute the optimal subset by **maximizing** over multiple LTSS searches (e.g. fast localized scan).
 - 2) We can use the unconstrained maximum score as an **upper bound** on the constrained maximum (e.g. fast scan over rectangles).
 - 3) For **heuristic search**, we can use the unconstrained maximum as a starting point, or use the LTSS ordering to guide our search.

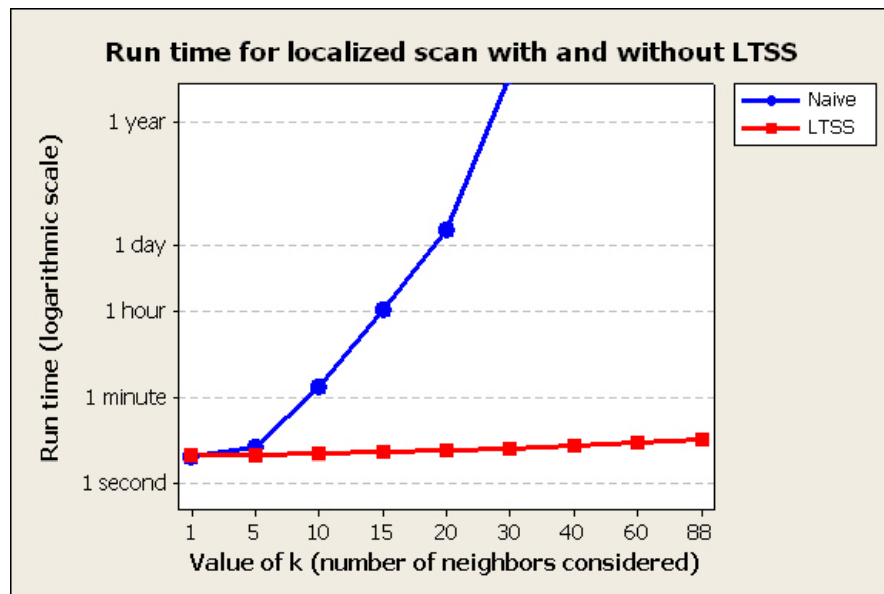
Fast localized scan

- Maximize the spatial scan statistic over regions consisting of a “center” location s_i and **any subset** of its k -nearest neighbors, for a fixed constant k .
- This is similar to FlexScan¹ but does not force the region to be contiguous.
- Naïve search requires $O(N \cdot 2^k)$ time and is computationally infeasible for $k > 20$.
- For each center, we can search over all subsets of its k -nearest neighbors in $O(k)$ using LTSS, thus requiring total time $O(Nk) + O(N \log N)$ for sorting by priority.

¹Tango and Takahashi, IJHG, 2005

Evaluation on ED data

We compared the time needed to perform localized scans with and without LTSS, as a function of the number of neighbors k , for 281 days of Emergency Department visit data from Allegheny County.



$k = 15$: **869x** speedup
(4.42 sec. vs. over 1 hour)

$k = 20$: **38,460x** speedup
(4.69 sec. vs. over 50 hours)

$k = 30$: 5.21 sec. vs. ~9 yrs.

$k = 88$: 8.12 sec. vs. $\sim 10^{26}$ yrs.

Linear-time subset scanning

Linear-time subset scanning is a powerful and useful tool that enables us to speed up a wide variety of spatial event detection methods.

The Poisson, Gaussian, and nonparametric spatial scan statistics all satisfy the LTSS property, as do many other possible statistics.

LTSS makes “all subsets” search computationally feasible, makes localized scans feasible even for large values of k , and speeds up searches over “all distinct rectangles” by 2-3 orders of magnitude.

Many other LTSS-enabled searches are possible, and these will enable huge speedups for an even wider range of problems.

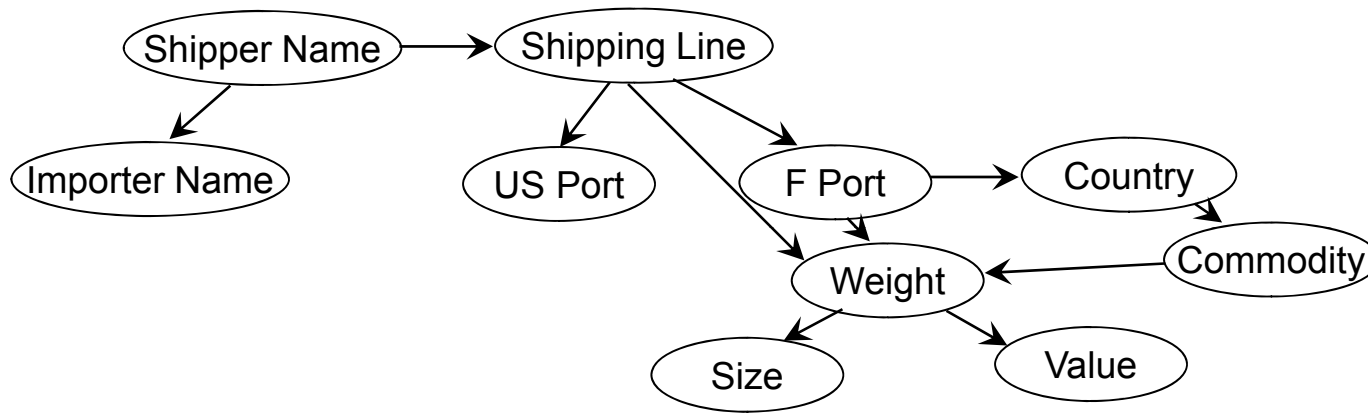
Current work includes extending LTSS to the multivariate and space-time scan statistics, developing fast graph scan algorithms, and incorporating LTSS into our Bayesian scan framework.

D. Current Directions in Spatial Event Detection

1. Incorporating learning into detection
2. Very fast detection algorithms
3. **Generalization of spatial methods to non-spatial data**
4. Non-aggregated spatial and temporal data
5. Incorporating rich information from observations
6. Detecting multiple, dynamic, irregular clusters
7. Integrating detection, tracking, and response
8. Combining sensor placement and sensor fusion

Anomalous Group Detection

1. Learn a Bayesian Network model for the null hypothesis H_0 (no events) from the training data.



2. To evaluate a group of records S :

1. Fit the alternate hypothesis Bayesian Network ($H_1(S)$) parameters using $Data_S$.

2. Compute the group score using the likelihood ratio: $F(S) = \frac{P(S|H_1(S))}{P(S|H_0)}$

3. Greedily grow a group from each record, and output the groups with highest score.

Neill et al., in *Scan Statistics: Methods and Applications*, 2009.

D. Current Directions in Spatial Event Detection

- 1. Incorporating learning into detection**
- 2. Very fast detection algorithms**
- 3. Generalization of spatial methods to non-spatial data**
- 4. Non-aggregated spatial and temporal data**
- 5. Incorporating rich information from observations**
- 6. Detecting multiple, dynamic, irregular clusters**
- 7. Integrating detection, tracking, and response**
- 8. Combining sensor placement and sensor fusion**

References

- D. Agarwal, A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu. Spatial scan statistics: approximations and performance study. *Proc. 12th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 24–33, 2006.
- D. Agarwal, J. M. Phillips, and S. Venkatasubramanian. The hunting of the bump: On maximizing statistical discrepancy. *Proc. Symposium on Discrete Algorithms*, 1137–1144, 2006.
- D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3): 962–994, 2004.
- L. Duczmal and R. Assuncao. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, 45:269–286, 2004.
- M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Intl. Conf. on Knowledge Discovery and Data Mining*, 1996.
- J. Friedman and N. Fisher. Bump hunting in high dimensional data. *Statistics and Computing*, 9(2):1–20, 1999.
- V. Iyengar. On detecting space-time clusters. *Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 587–592, 2004.
- M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6): 1481–1496, 1997.
- M. Kulldorff. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A*, 164: 61–72, 2001.
- M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal. An elliptic spatial scan statistic. *Statistics in Medicine*, 25:3929–3943, 2006.
- M. Kulldorff, F. Mostashari, L. Duczmal, W. K. Yih, K. Kleinman, and R. Platt. Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, 26: 1824–1833, 2007.
- M. Makatchev and D.B. Neill. Learning outbreak regions in Bayesian spatial scan statistics. *Proc. ICML/UAI/COLT Workshop on Machine Learning for Health Care Applications*, 2008.
- D.B. Neill and A.W. Moore. Rapid detection of significant spatial clusters. *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 256–265, 2004.
- D.B. Neill, A.W. Moore, F. Pereira, and T. Mitchell. Detecting significant multidimensional spatial clusters. In L.K. Saul, et al., eds., *Adv. Neural Information Processing Systems 17*, 969–976, 2005.

References

- D.B. Neill, A.W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. *Proc. 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 218-227, 2005.
- D.B. Neill and A.W. Moore. Anomalous spatial cluster detection. *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.
- D.B. Neill, A.W. Moore, and G.F. Cooper. A Bayesian spatial scan statistic. In Y. Weiss, et al., eds. *Advances in Neural Information Processing Systems 18*, 1003-1010, 2006.
- D.B. Neill. Detection of spatial and spatio-temporal clusters. Ph.D. thesis, Carnegie Mellon University, Department of Computer Science, 2006.
- D.B. Neill. Incorporating learning into disease surveillance systems. *Advances in Disease Surveillance* 4: 107, 2007.
- D.B. Neill and W.L. Gorr. Detecting and preventing emerging epidemics of crime. *Advances in Disease Surveillance* 4: 13, 2007.
- D.B. Neill and J. Lingwall. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* 4: 106, 2007.
- D.B. Neill. Fast and flexible outbreak detection by linear-time subset scanning. *Advances in Disease Surveillance* 5: 48, 2008.
- D.B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics* 8: 20, 2009.
- D.B. Neill, G.F. Cooper, K. Das, X. Jiang, and J. Schneider. Bayesian network scan statistics for multivariate pattern detection. In J. Glaz, et al., eds., *Scan Statistics: Methods and Applications*, 2009.
- D.B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 2009, in press.
- D.B. Neill and G.F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning*, 2009, in press.
- G. P. Patil and C. Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Envir. Ecol. Stat.*, 11: 183–197, 2004.
- T. Tango and K. Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *Intl. Journal of Health Geographics*, 4: 11, 2005.
- M. Wu, X. Song, C. Jermaine, S. Ranka and J. Gums. A LRT Framework for Fast Spatial Anomaly Detection. *Proc. 15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 2009.