Detection Algorithms for Biosurveillance: A tutorial

Andrew Moore	Professor	Computer Science, Carnegie Mellon	awm@cs.cmu.edu

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials . Comments and corrections gratefully received.

Tutorial compiled with much help from…

Greg Cooper	Professor	Computer Science and RODS lab, U. Pitt	gfc@cbmi.upmc.edu
Bill Hogan	Assistant Professor	RODS lab, U. Pitt	wrh@cbmi.pitt.edu
Rich Tsui	Research Professor and associate Director of RODS lab	RODS lab, U. Pitt	tsui@cbmi.pitt.edu
Mike Wagner	Professor and Director of RODS lab	RODS lab, U. Pitt	mmw@cbmi.pitt.edu

RODS: http://www.health.pitt.edu/rods

Auton Lab: http://www.autonlab.org

Many Methods!

Method	Has Pitt/ CMU tried it?	Tried but little used	Tried and used	Under development	Multivariate signal tracking?	Spatial?
Time-weighted averaging	Yes	Yes				
Serfling	Yes		Yes			
ARIMA	Yes	Yes				
SARIMA + External Factors	Yes		Yes			
Univariate HMM	Yes		Yes			
Kalman Filter	Yes	Yes				
Recursive Least Squares	Yes		Yes			
Support Vector Machine	Yes	Yes				
Neural Nets	Yes	Yes				
Randomization	Yes		Yes	Yes		
Spatial Scan Statistics	Yes			(w/ Howard Burkom)	Yes	Yes
Bayesian Networks	Yes			Yes	Yes	
Contingency Tables	Yes		Yes			
Scalar Outlier (SQC)	Yes	Yes				
Multivariate Anomalies	Yes		Yes		Yes	
Change-point statistics	Yes			Yes		
FDR Tests	Yes		Yes		Yes	
WSARE (Recent patterns)	Yes		Yes	Yes	Yes	Yes
PANDA (Causal Model)	Yes			Yes	Yes	Yes
FLUMOD (space/Time HMM)				Yes	Yes	Yes

Details of these methods and bibliography available from "Summary of Biosurveillance-relevant statistical and data mining technologies" by Moore, Cooper, Tsui and Wagner. Downloadable (PDF format) from www.cs.cmu.edu/~awm/biosurv-methods.pdf

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

Boring Names.

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

Boring Names.

Univariate Anomaly Detection

Multivariate Anomaly Detection

Spatial Scan Statistics

WSARE

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

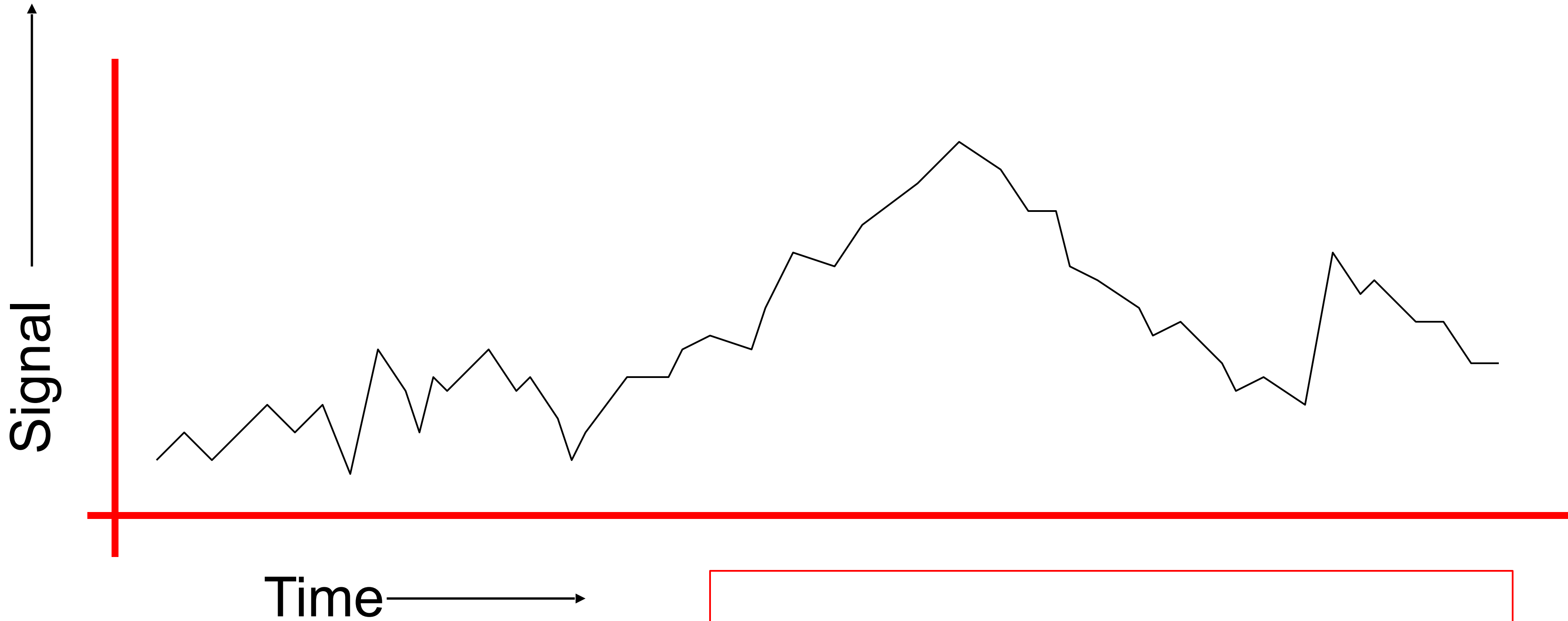
Univariate Anomaly Detection

Multivariate Anomaly Detection

Spatial Scan Statistics

WSARE

Univariate Time Series



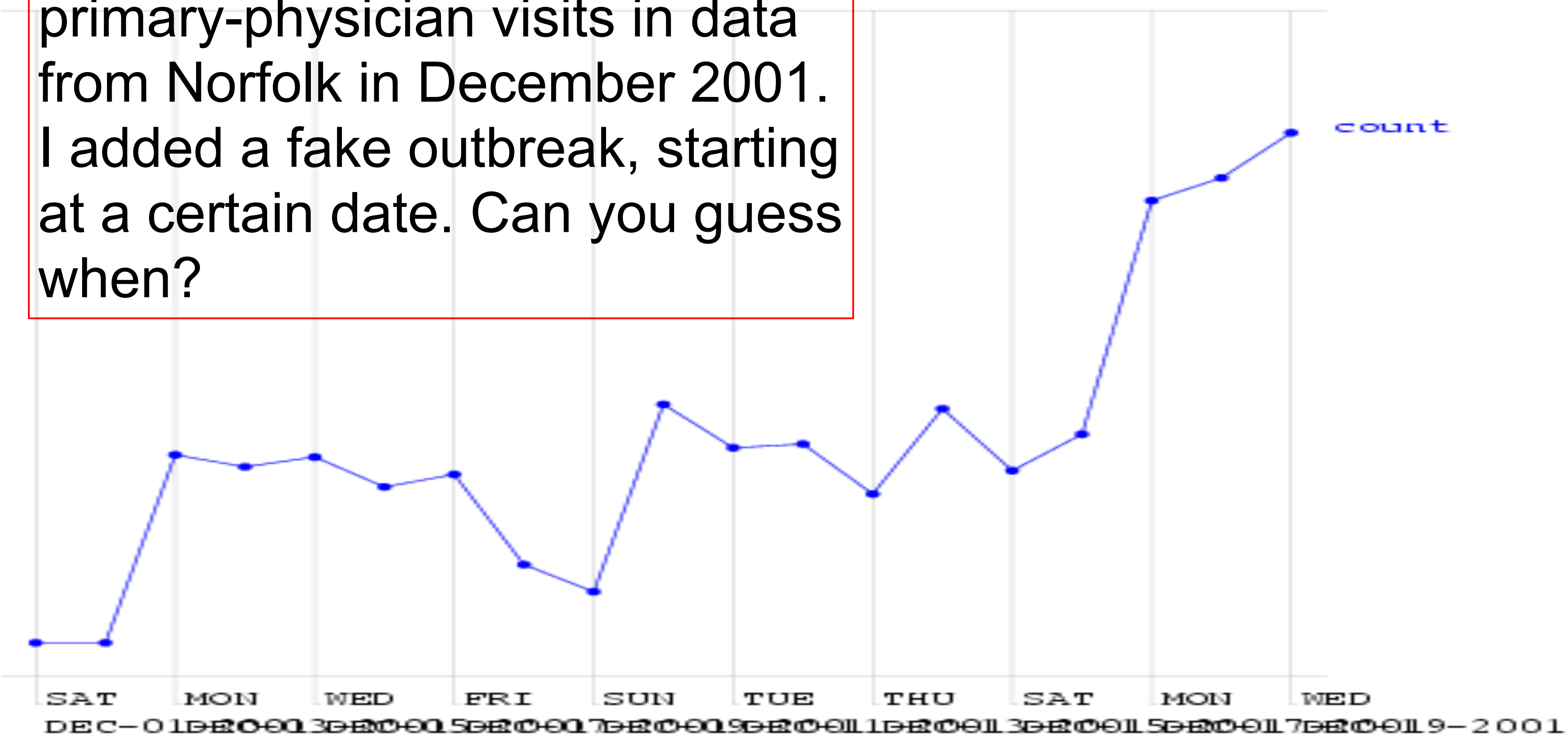
Example Signals:

- Number of ED visits today
- Number of ED visits this hour
- Number of Respiratory Cases Today
- School absenteeism today
- Nyquil Sales today

(When) is there an anomaly?

(When) is there an anomaly?

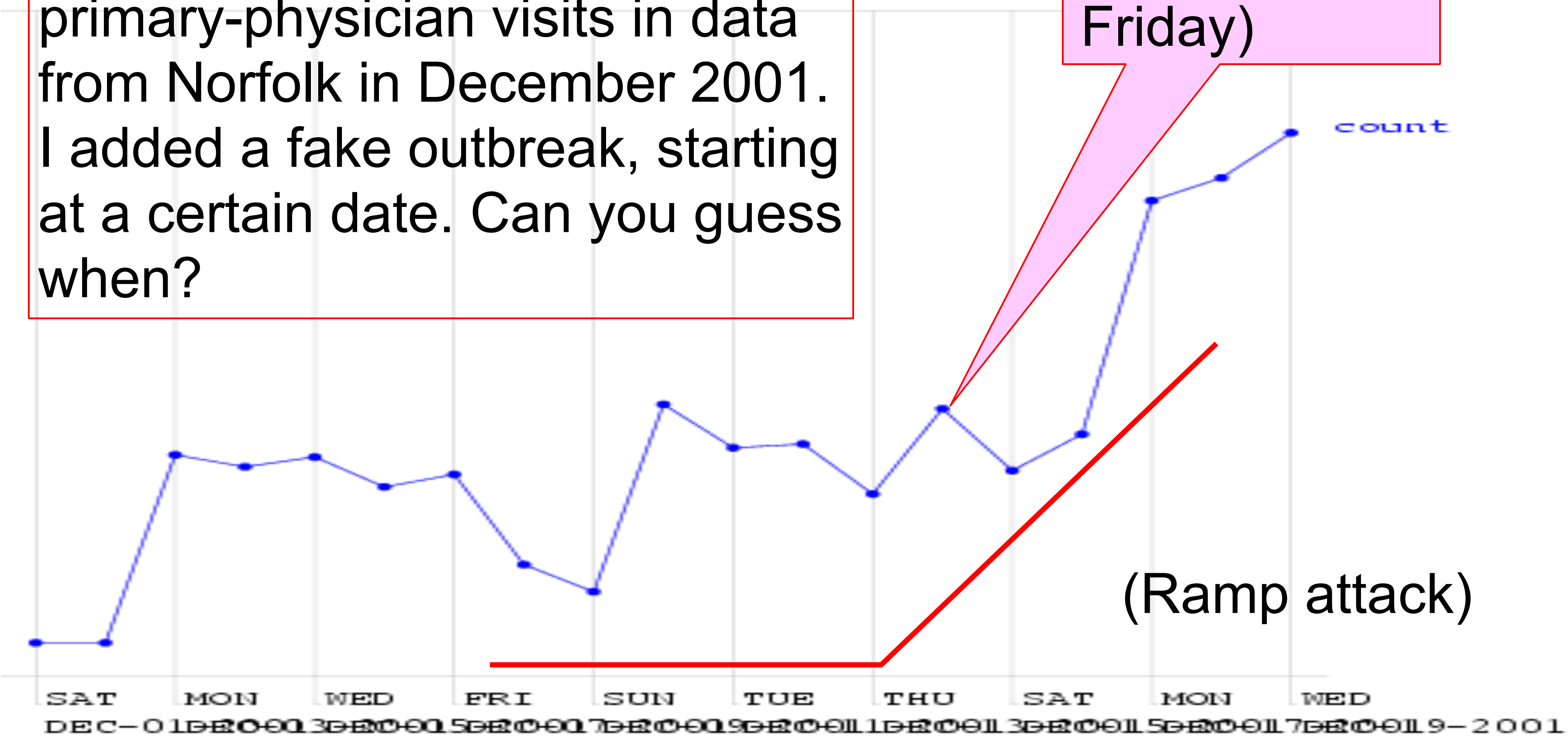
This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?



(When) is there an anomaly?

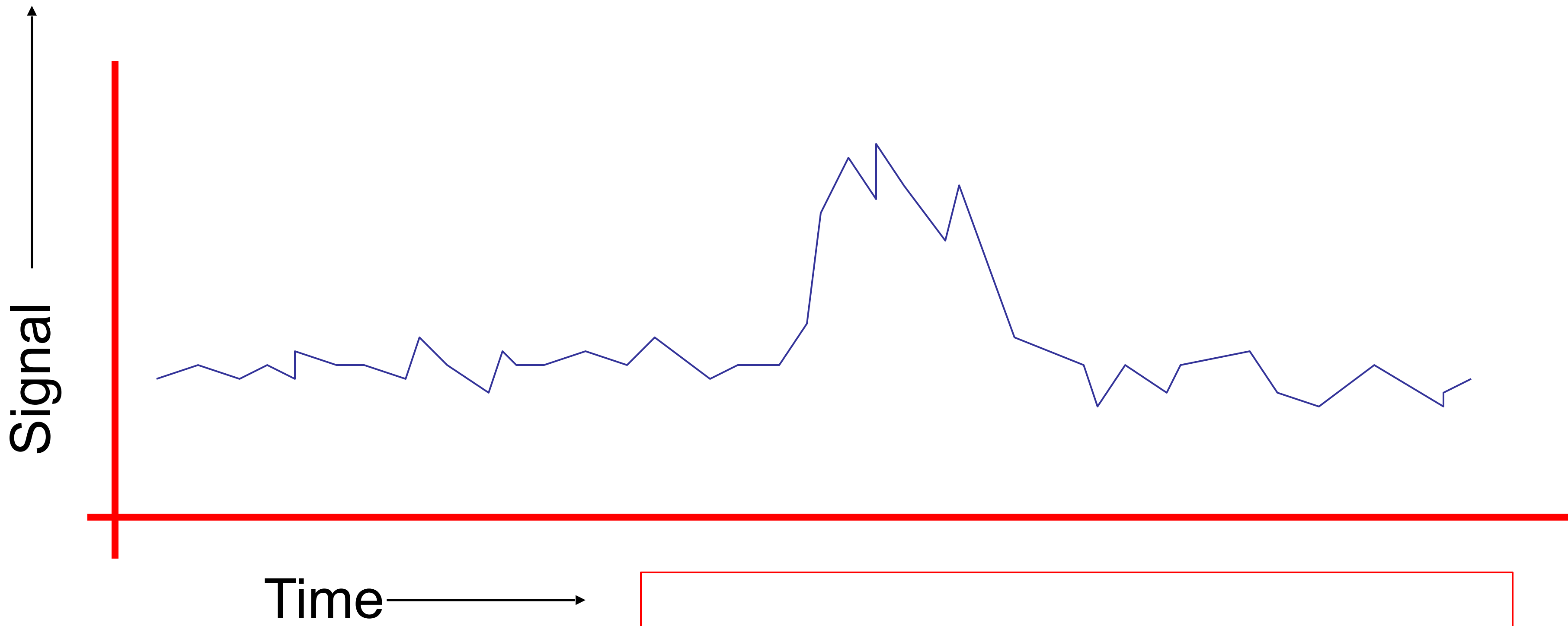
This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?

Here (much too high for a Friday)



(Ramp attack)

An easy case

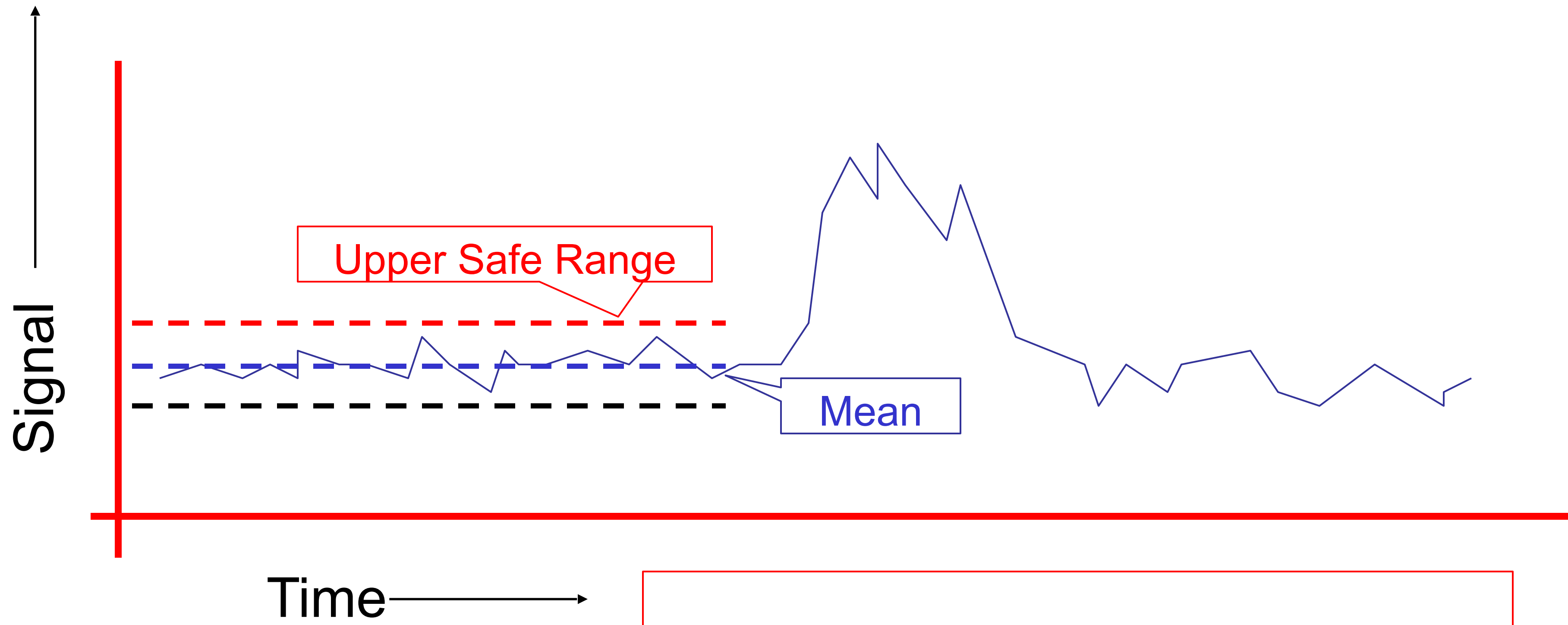


Dealt with by Statistical Quality Control

Record the mean and standard deviation up the the current time.

Signal an alarm if we go outside 3 sigmas

An easy case: Control Charts

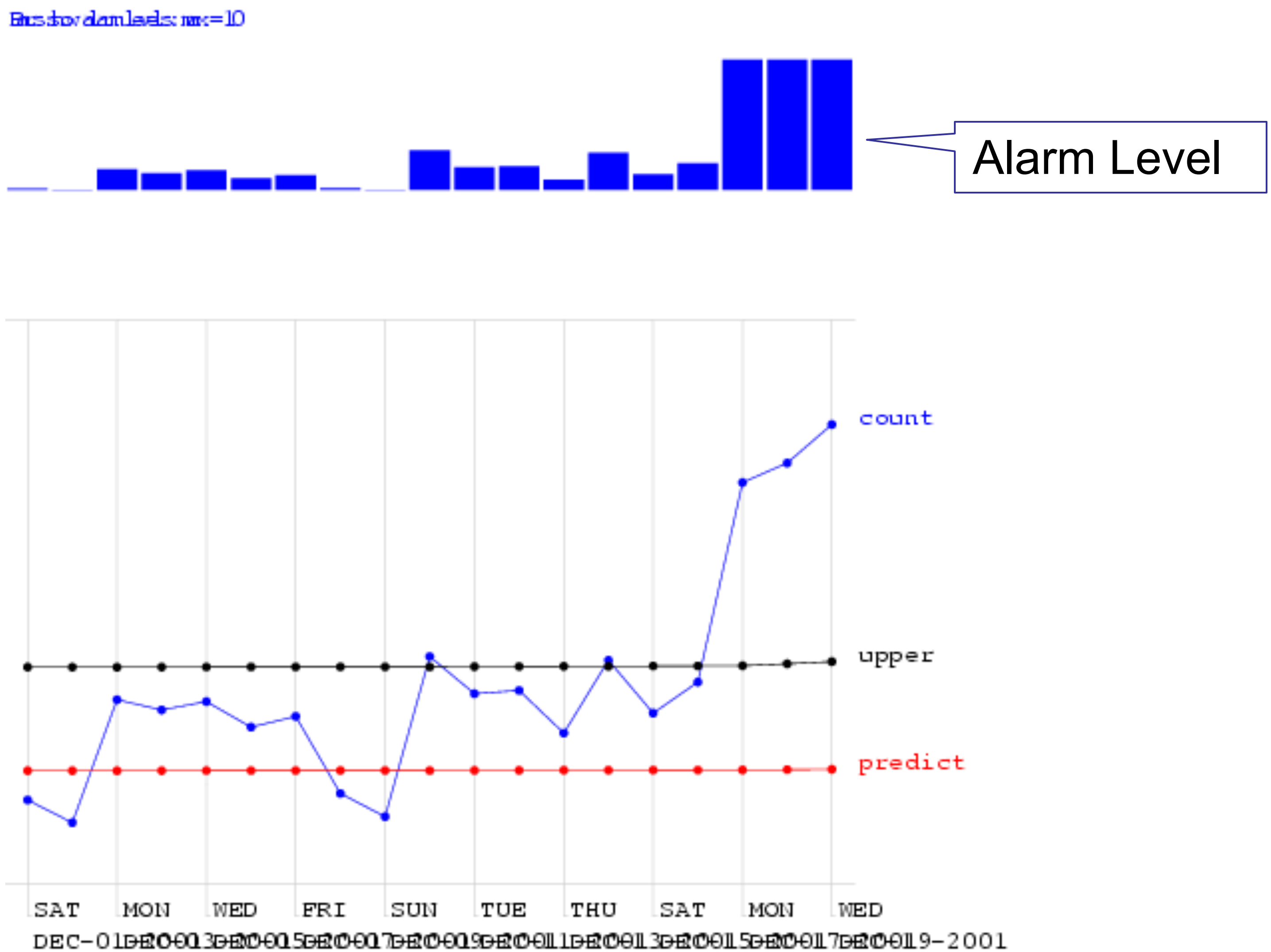


Dealt with by Statistical Quality Control

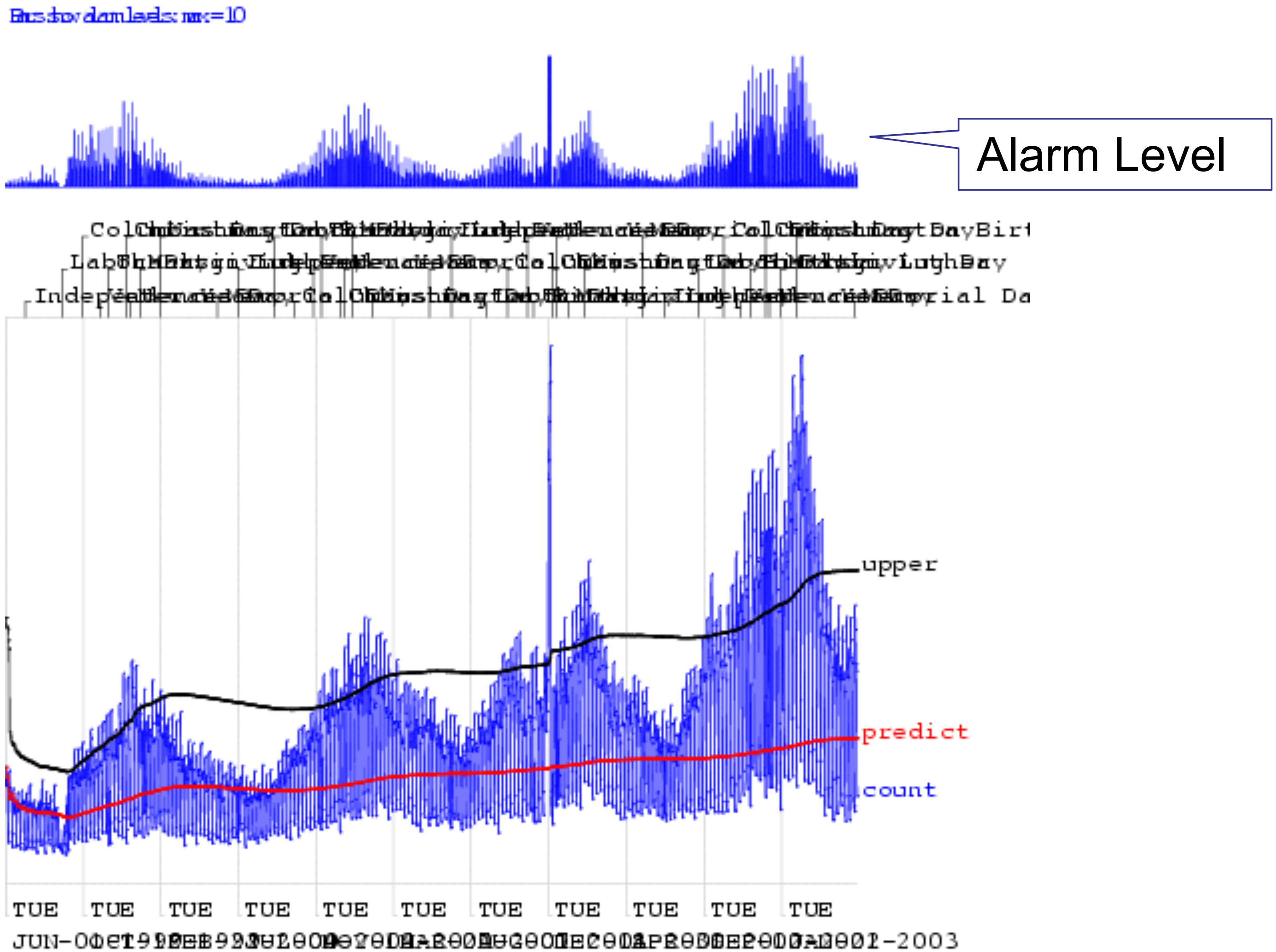
Record the mean and standard deviation up the the current time.

Signal an alarm if we go outside 3 sigmas

Control Charts on the Norfolk Data

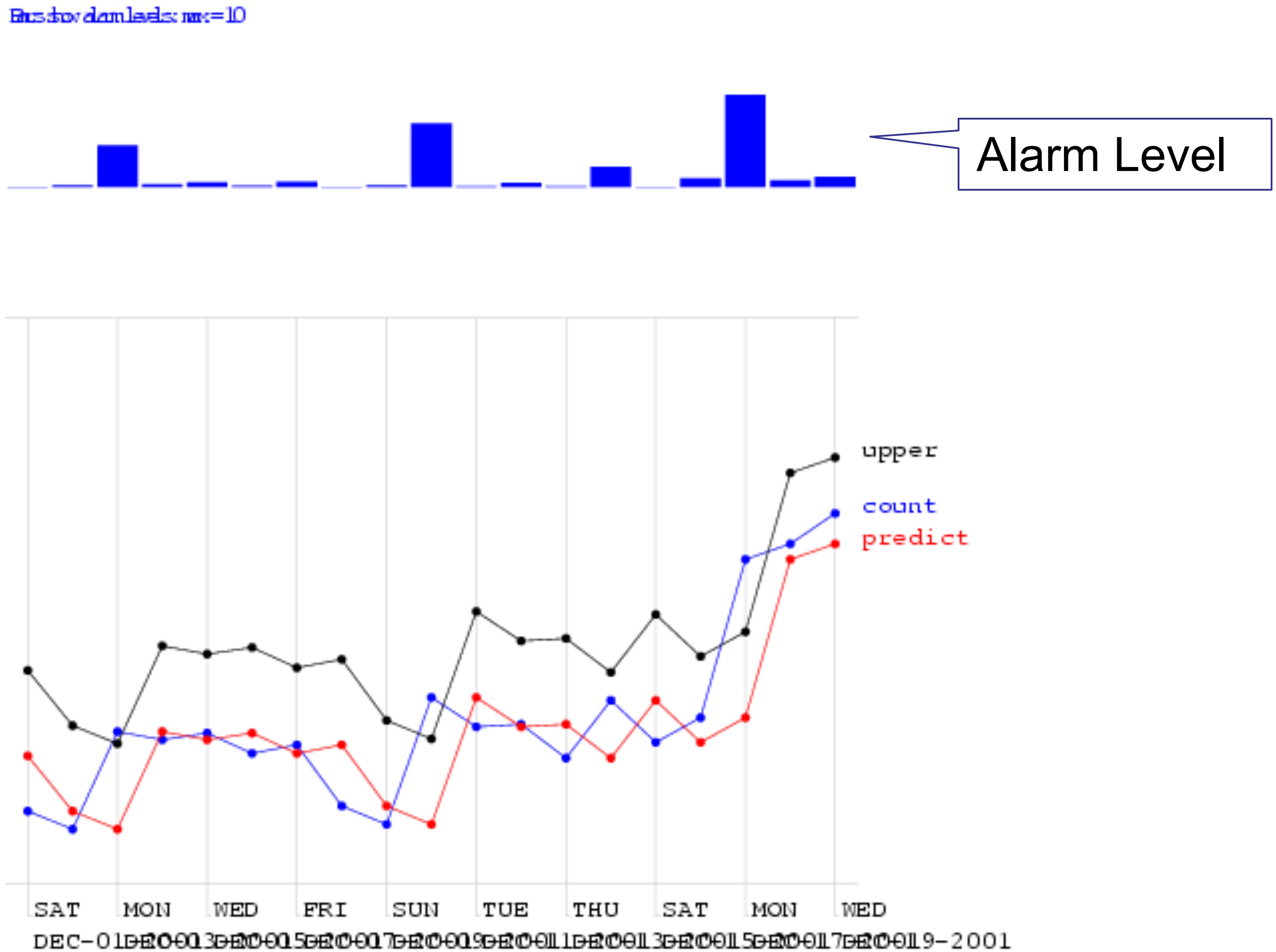


Control Charts on the Norfolk Data

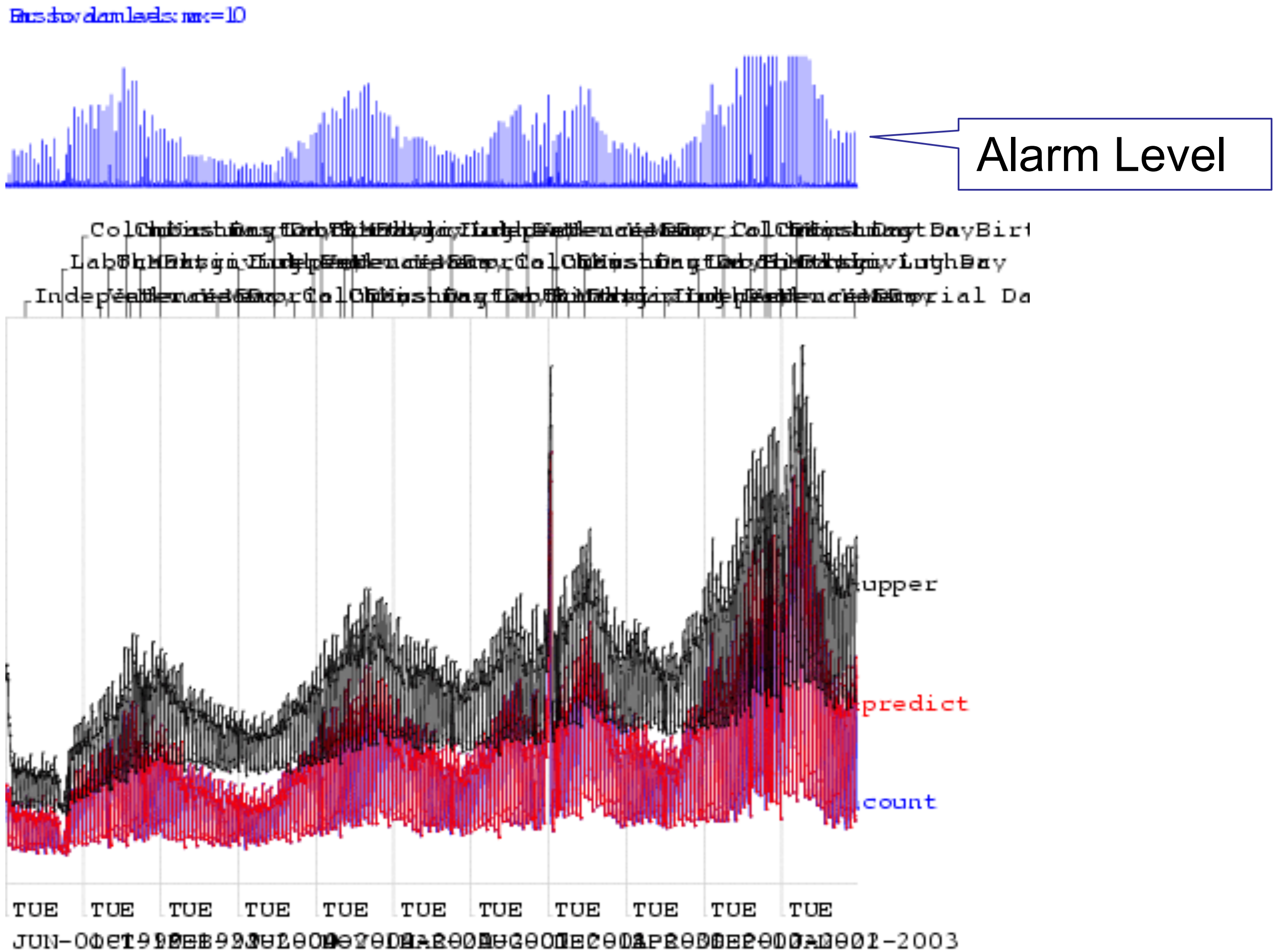


Looking at changes from yesterday

Looking at changes from yesterday

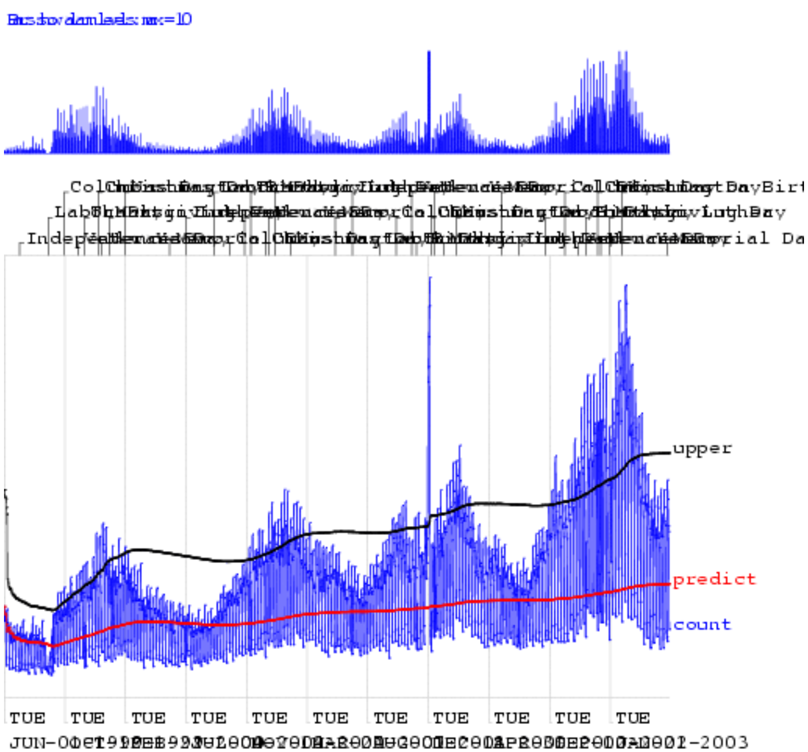


Looking at changes from yesterday

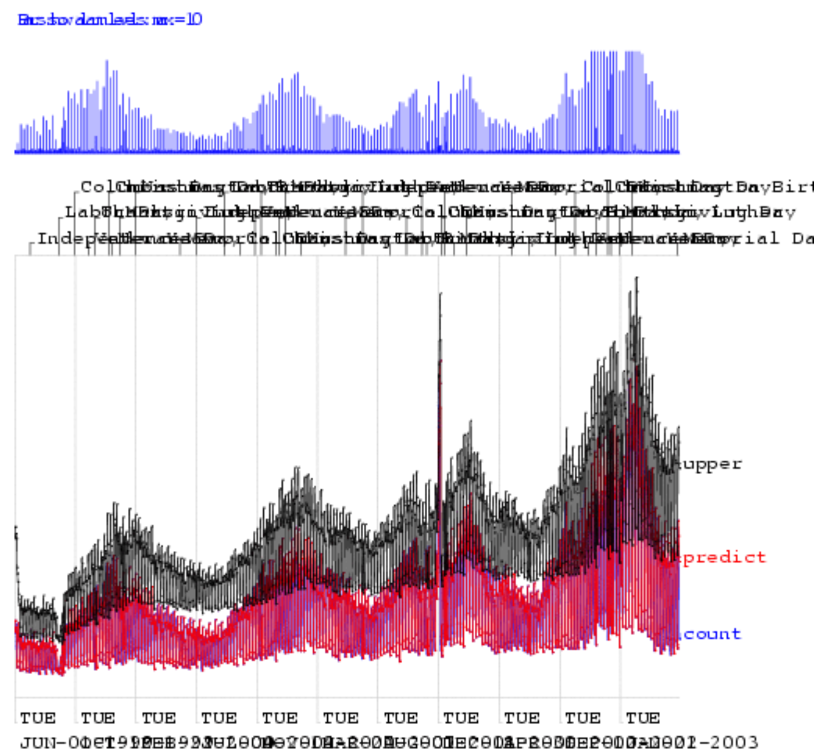


We need a happy medium:

Control Chart: Too insensitive to recent changes

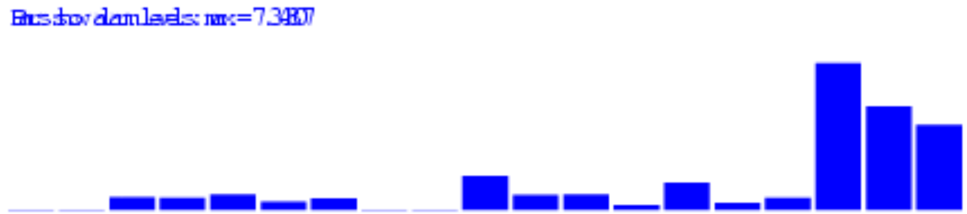


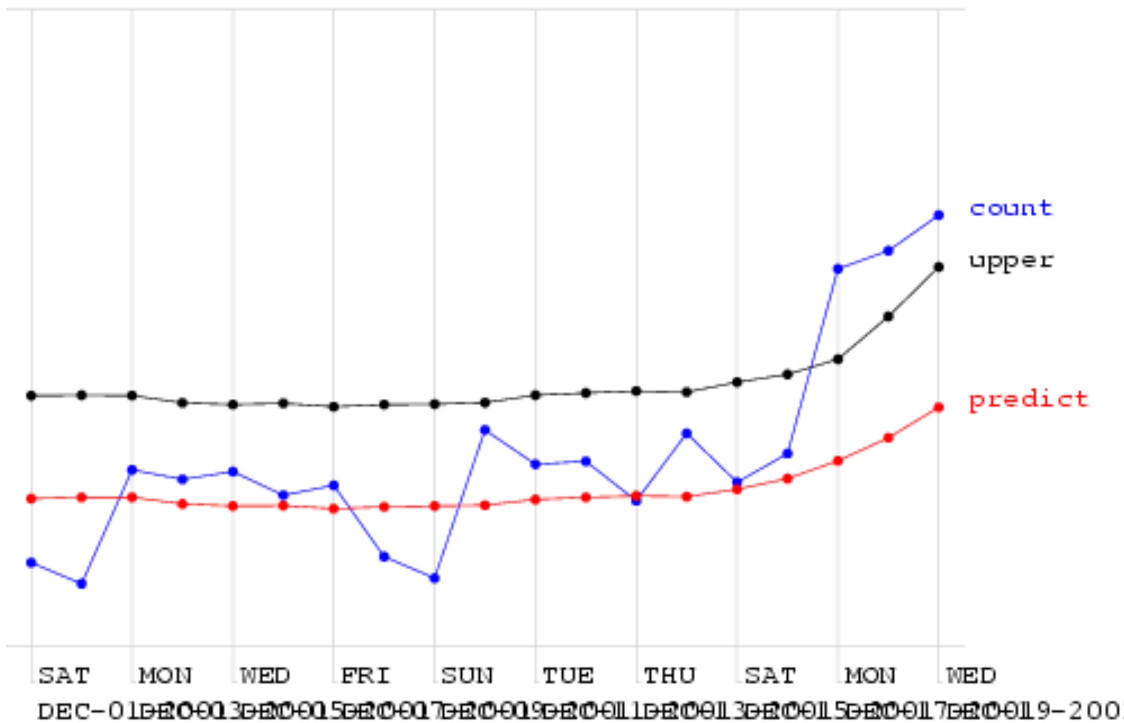
Change from yesterday: Too sensitive to recent changes



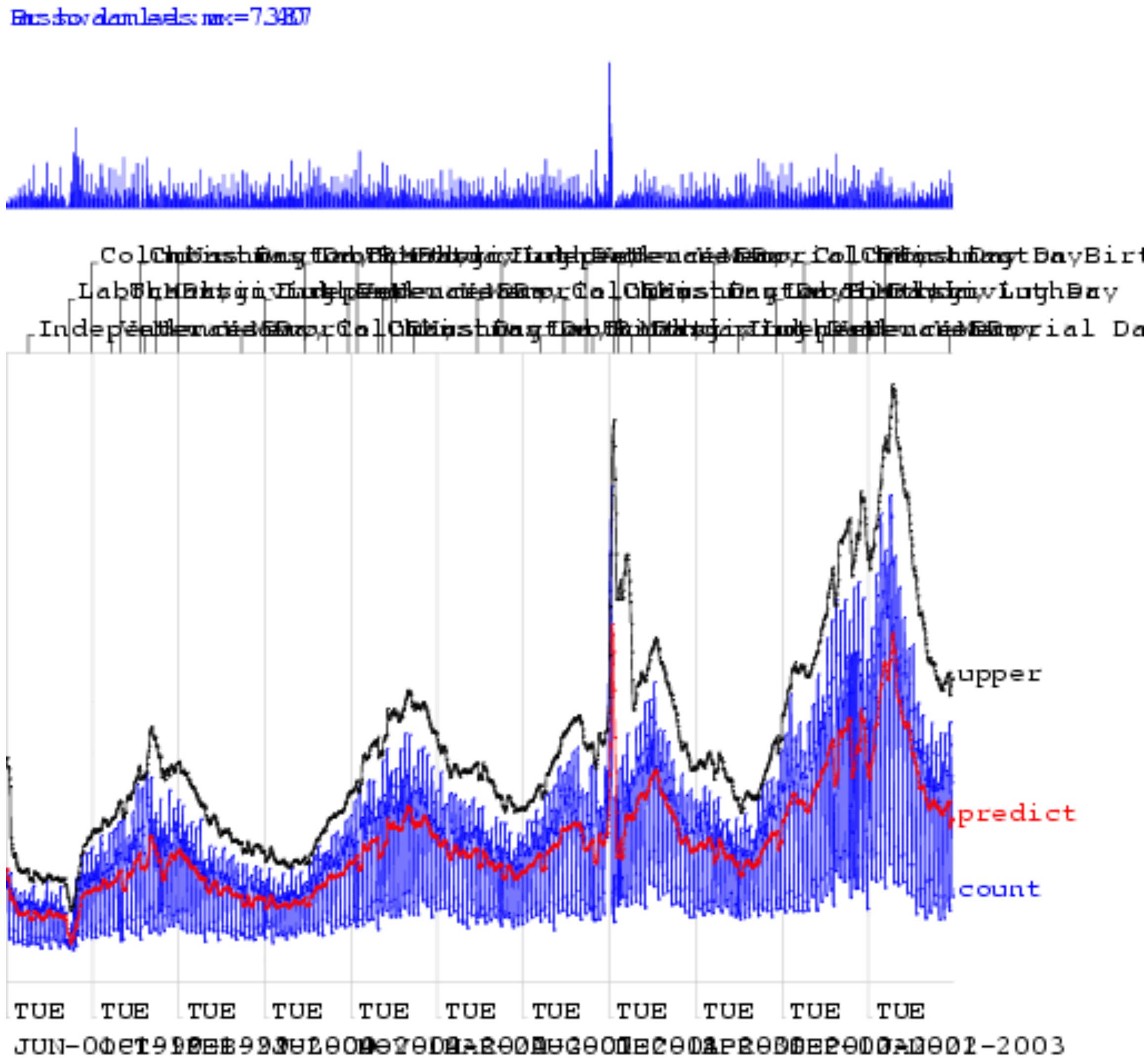
Moving Average

Moving Average

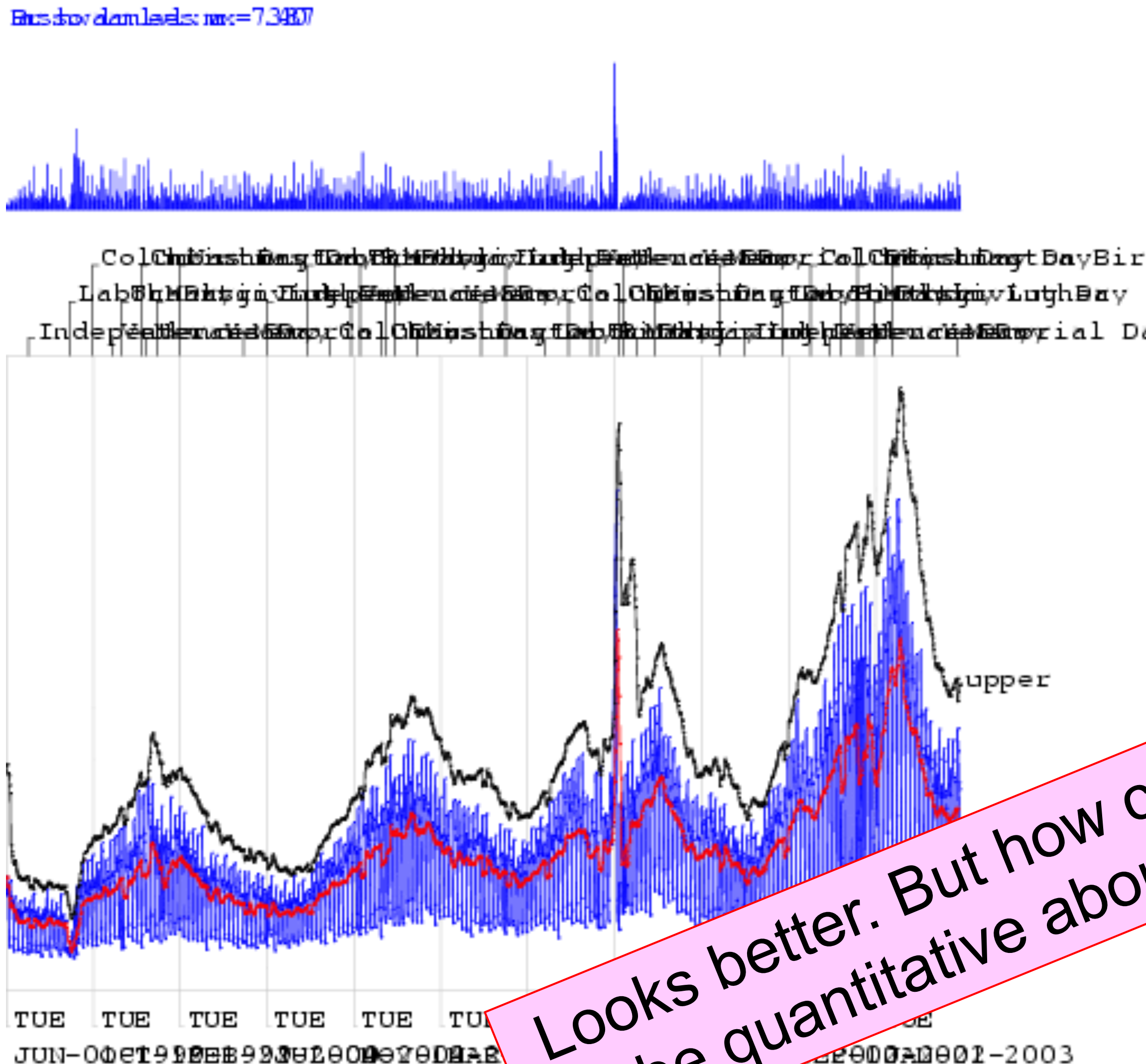




Moving Average



Moving Average



Looks better. But how can we be quantitative about this?

Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7

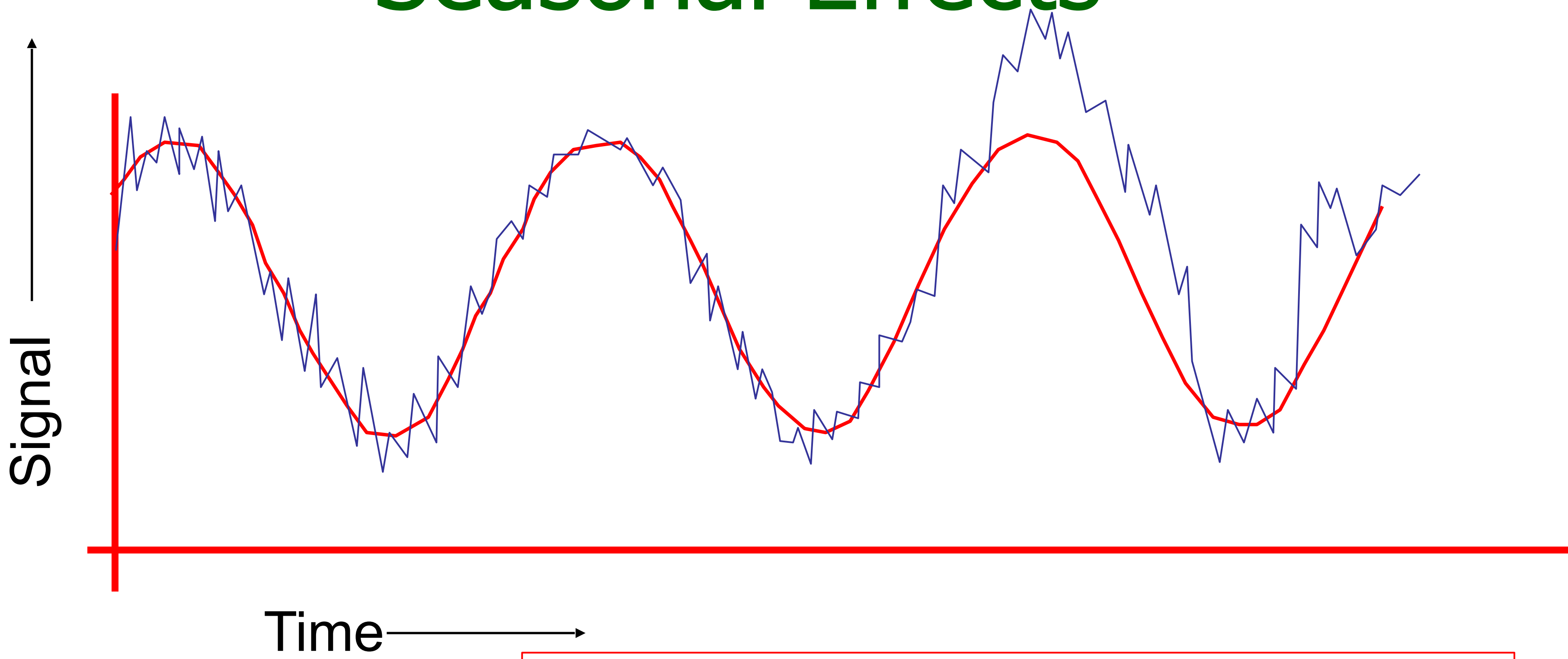
Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 7	0.58	2.79	0.51	3.31

Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54

Seasonal Effects



Fit a periodic function (e.g. sine wave) to previous data. Predict today's signal and 3-sigma confidence intervals. Signal an alarm if we're off.

Reduces False alarms from Natural outbreaks.

Different times of year deserve different thresholds.

Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9

Day-of-week effects

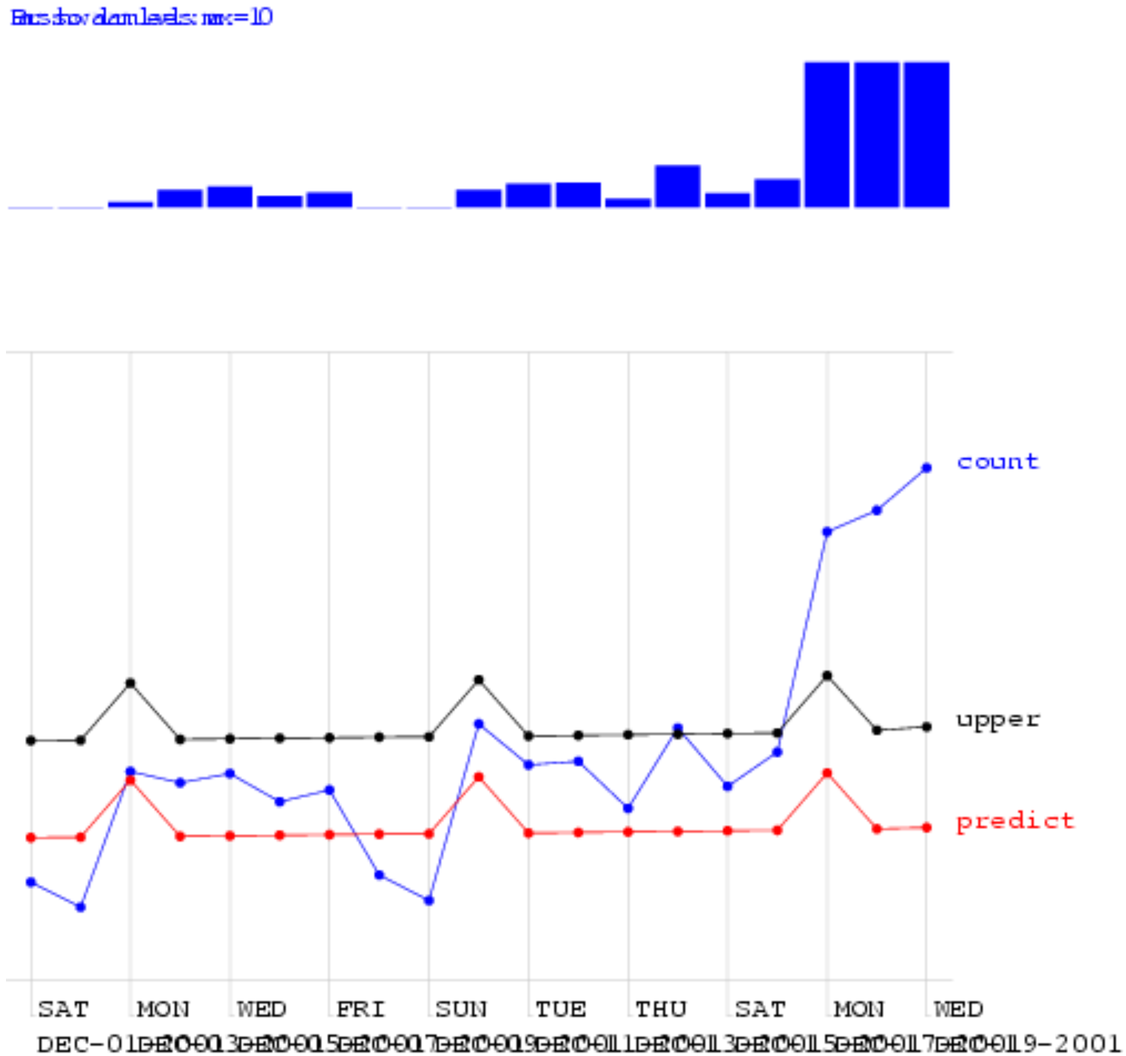
Fit a day-of-week component

E[Signal] = a + delta_{day}

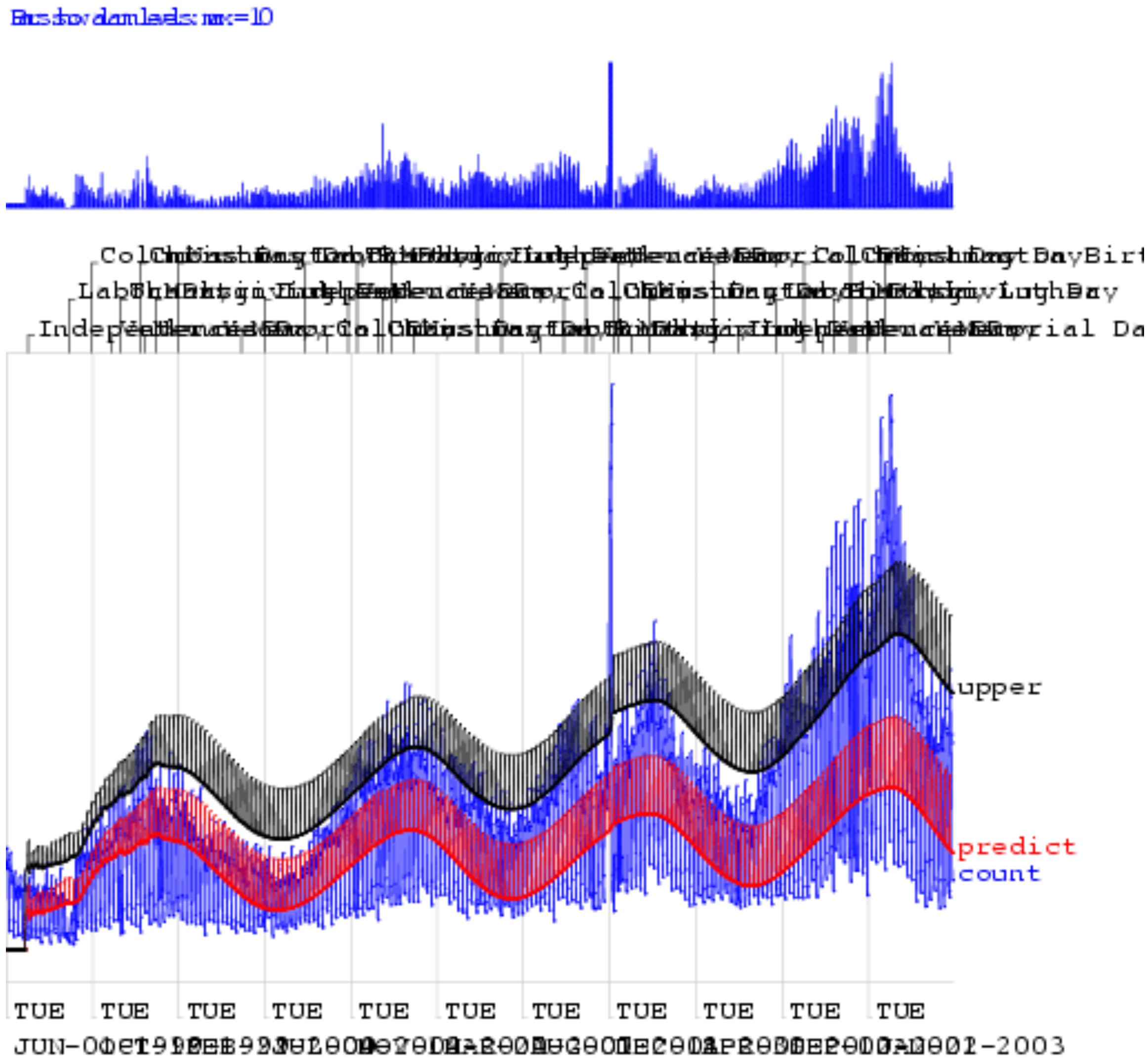
E.G: delta_{mon}= +5.42, delta_{tue}= +2.20, delta_{wed}= +3.33, delta_{thu}= +3.10, delta_{fri}= +4.02, delta_{sat}= -12.2, delta_{sun}= -23.42

A simple form of ANOVA

Regression using Hours-in-day & IsMonday



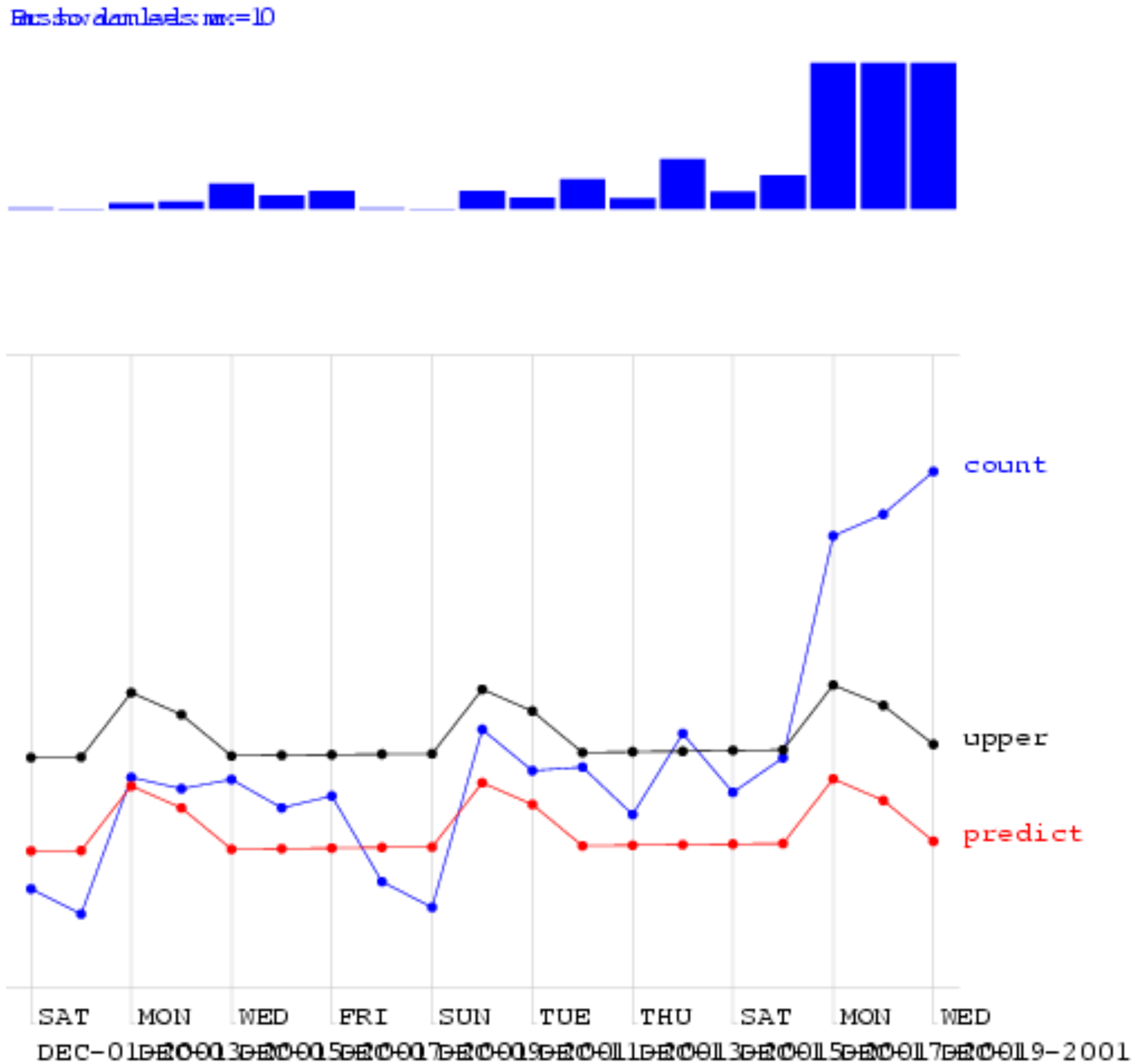
Regression using Hours-in-day & IsMonday



Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9
hours_of_daylight is_mon	0.7	2.25	0.57	3.12

Regression using Mon-Tue



Algorithm Performance

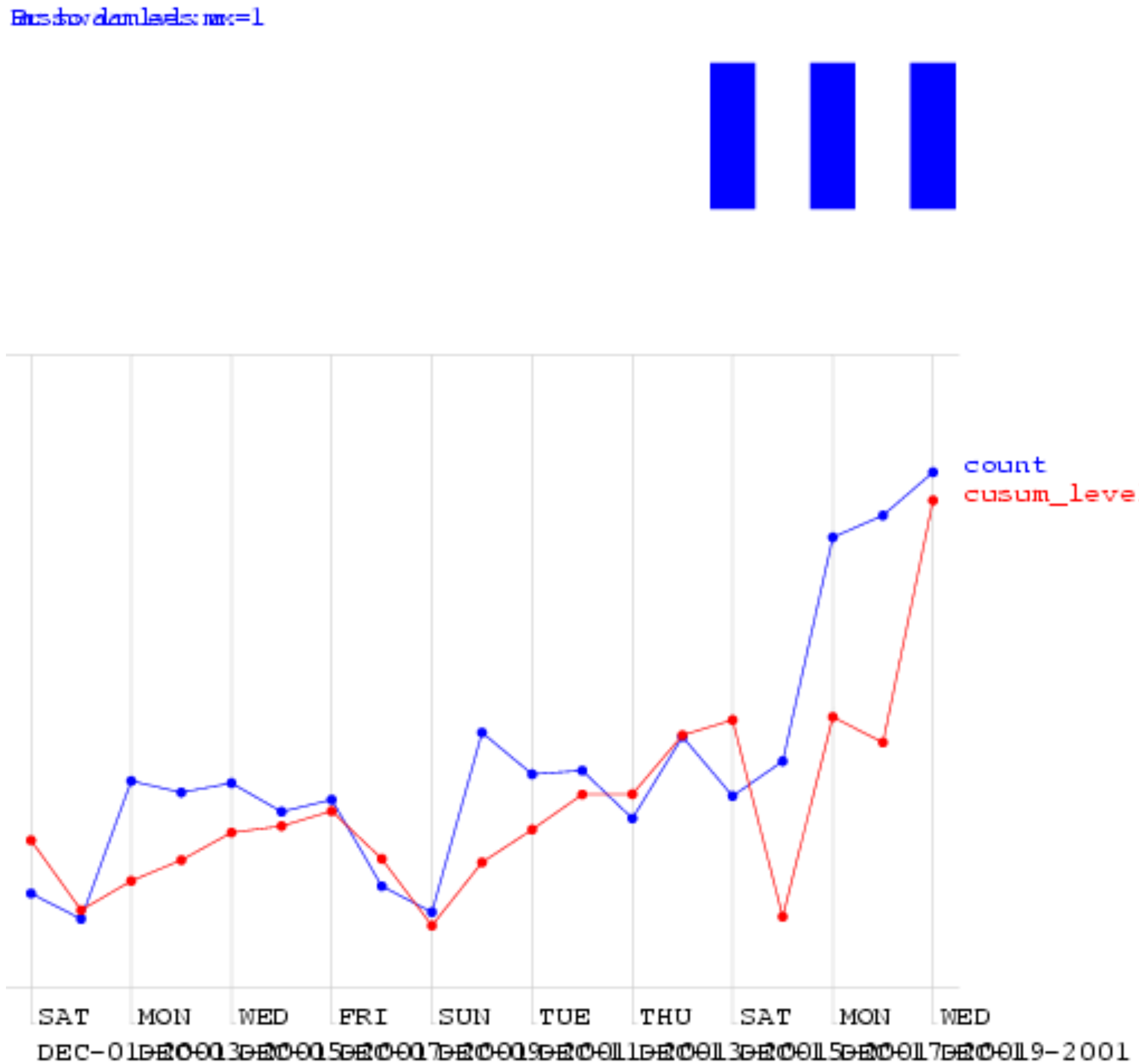
	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9
hours_of_daylight is_mon	0.7	2.25	0.57	3.12
hours_of_daylight is_mon … is_tue	0.72	1.83	0.57	3.16
hours_of_daylight is_mon … is_sat	0.77	2.11	0.59	3.26

CUSUM

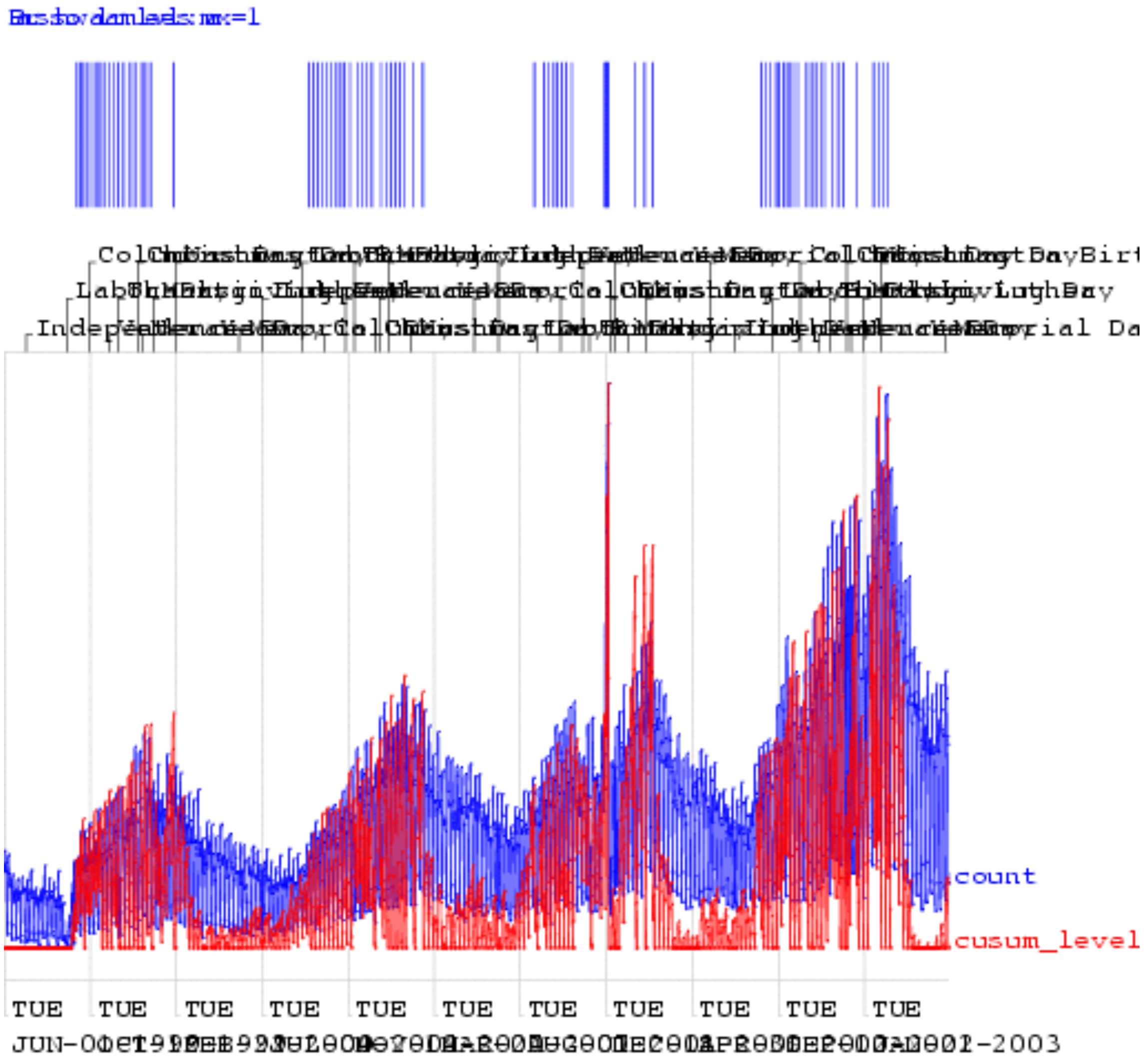
- <u>CU</u>mulative <u>SUM</u> Statistics

- Keep a running sum of “surprises”: a sum of excesses each day over the prediction
- When this sum exceeds threshold, signal alarm and reset sum

CUSUM



CUSUM

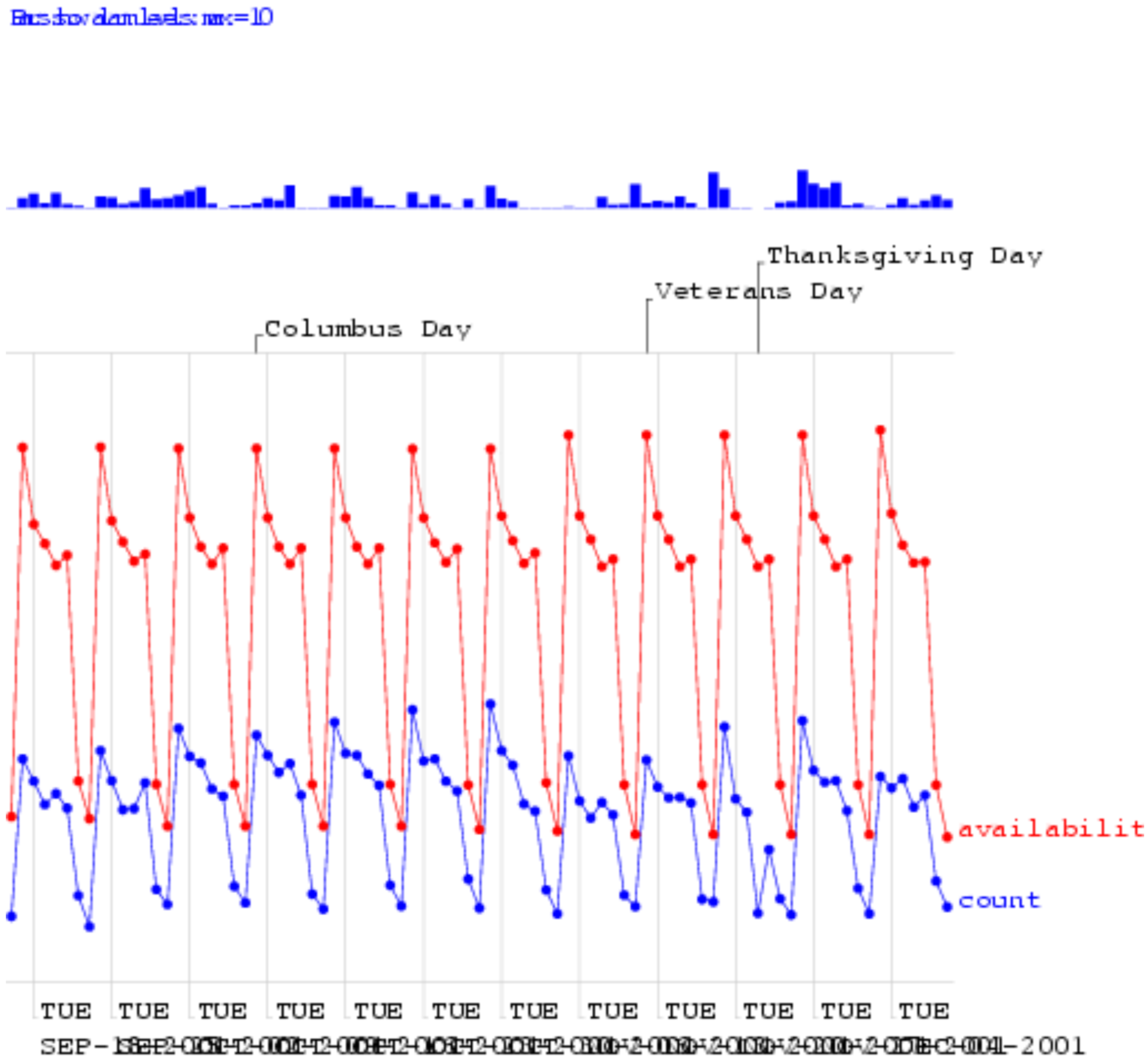


Algorithm Performance

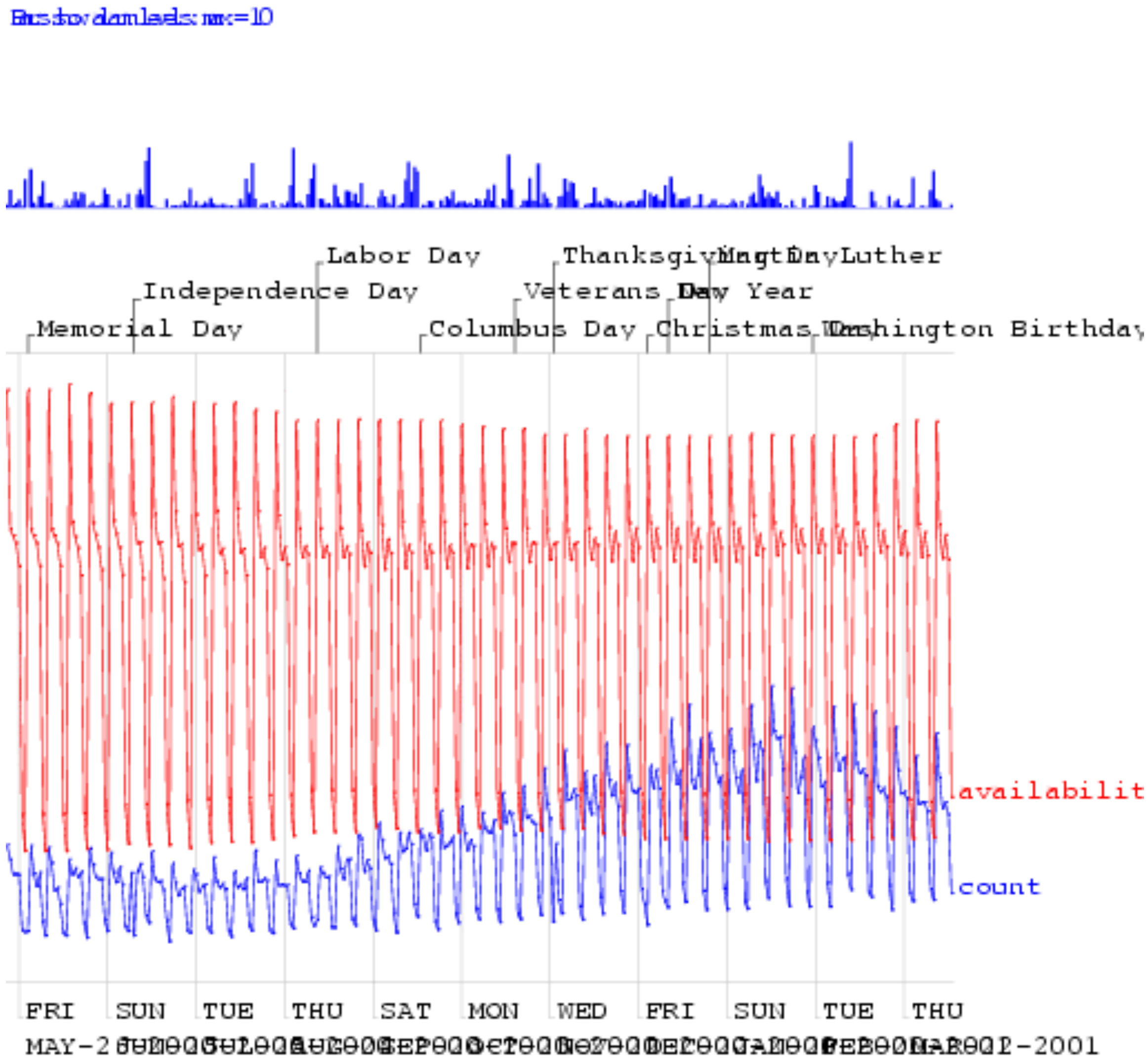
	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9
hours_of_daylight is_mon	0.7	2.25	0.57	3.12
hours_of_daylight is_mon … is_tue	0.72	1.83	0.57	3.16
hours_of_daylight is_mon … is_sat	0.77	2.11	0.59	3.26
CUSUM	0.45	2.03	0.15	3.55

The Sickness/Availability Model

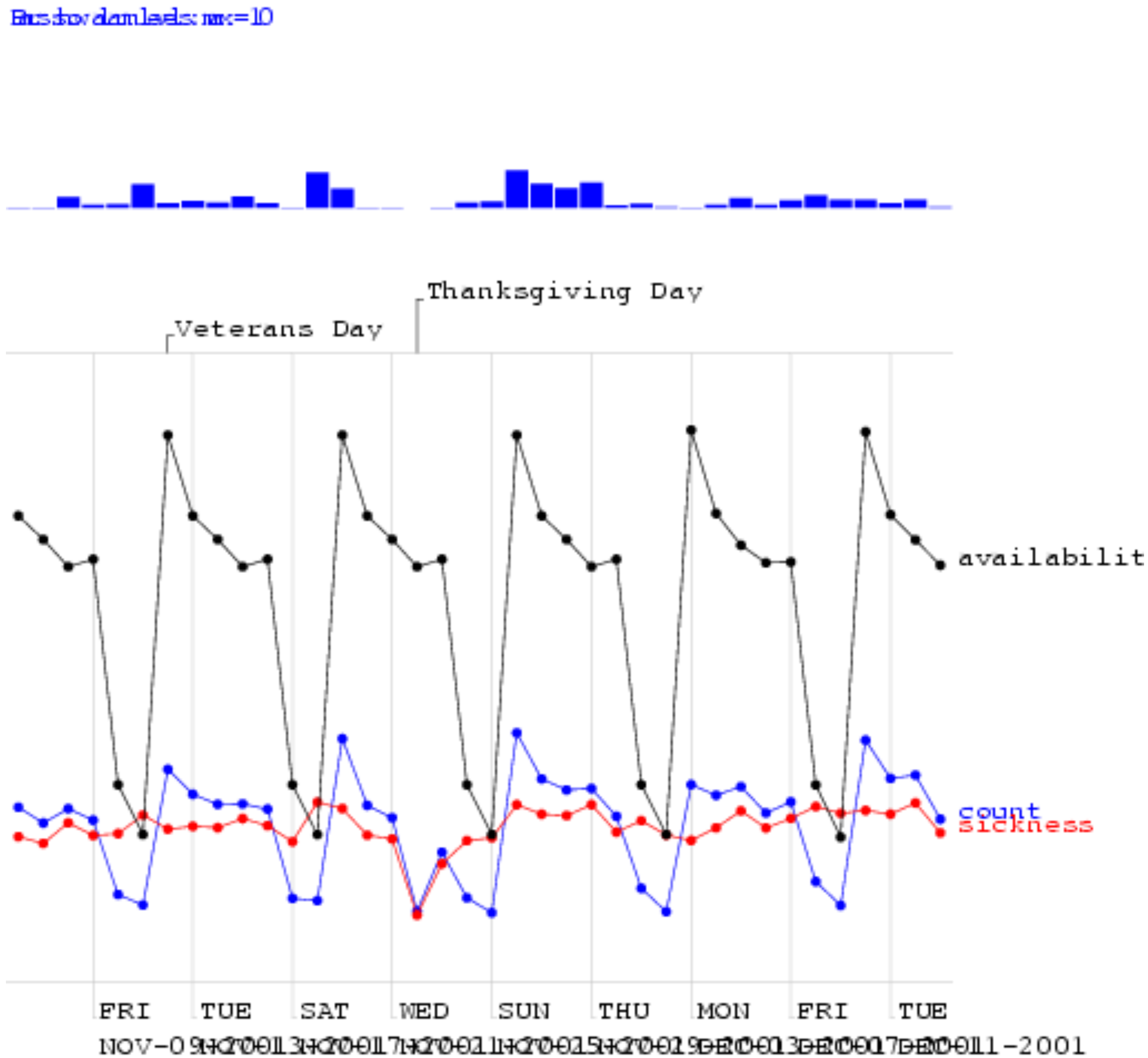
The Sickness/Availability Model



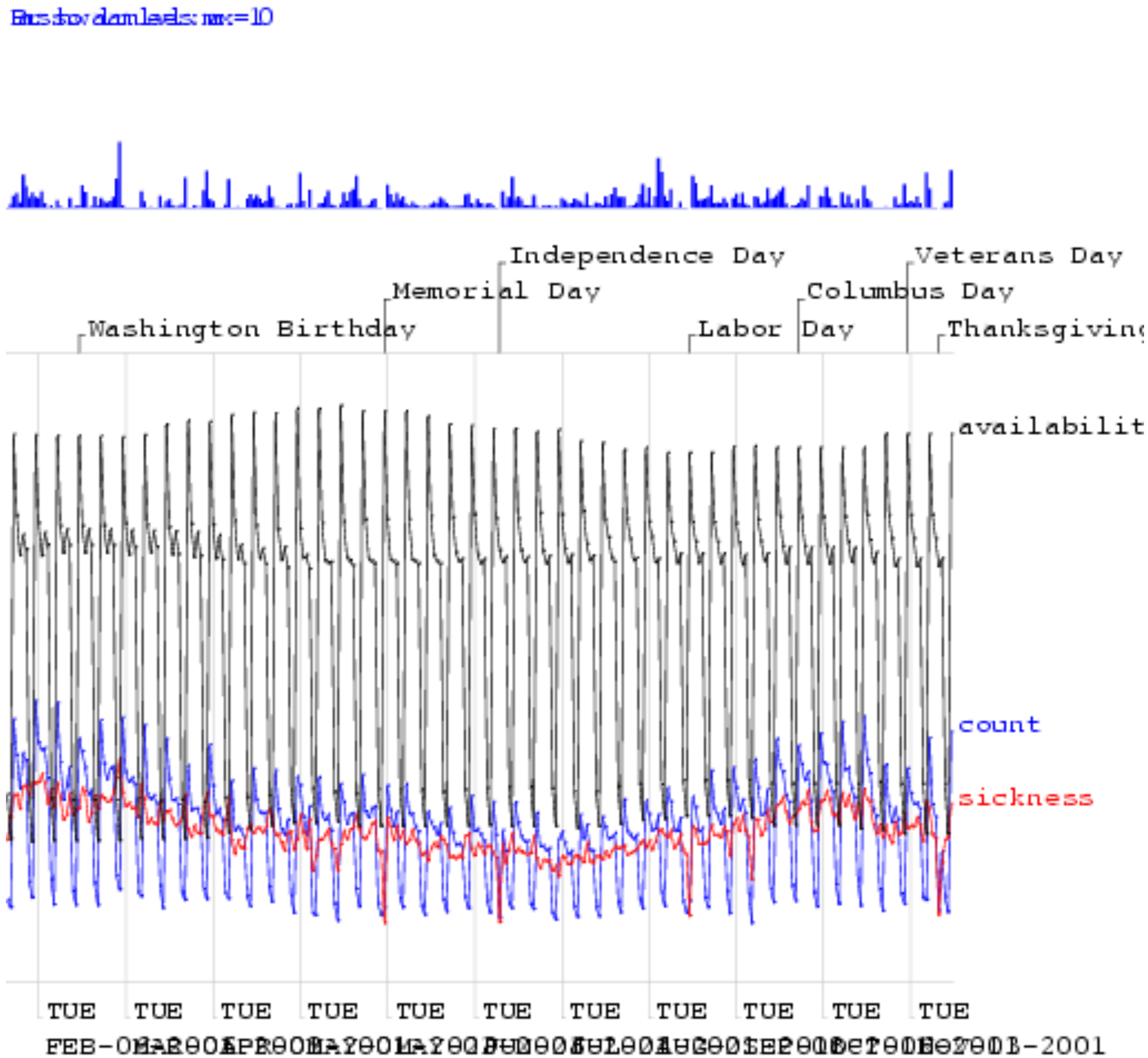
The Sickness/Availability Model



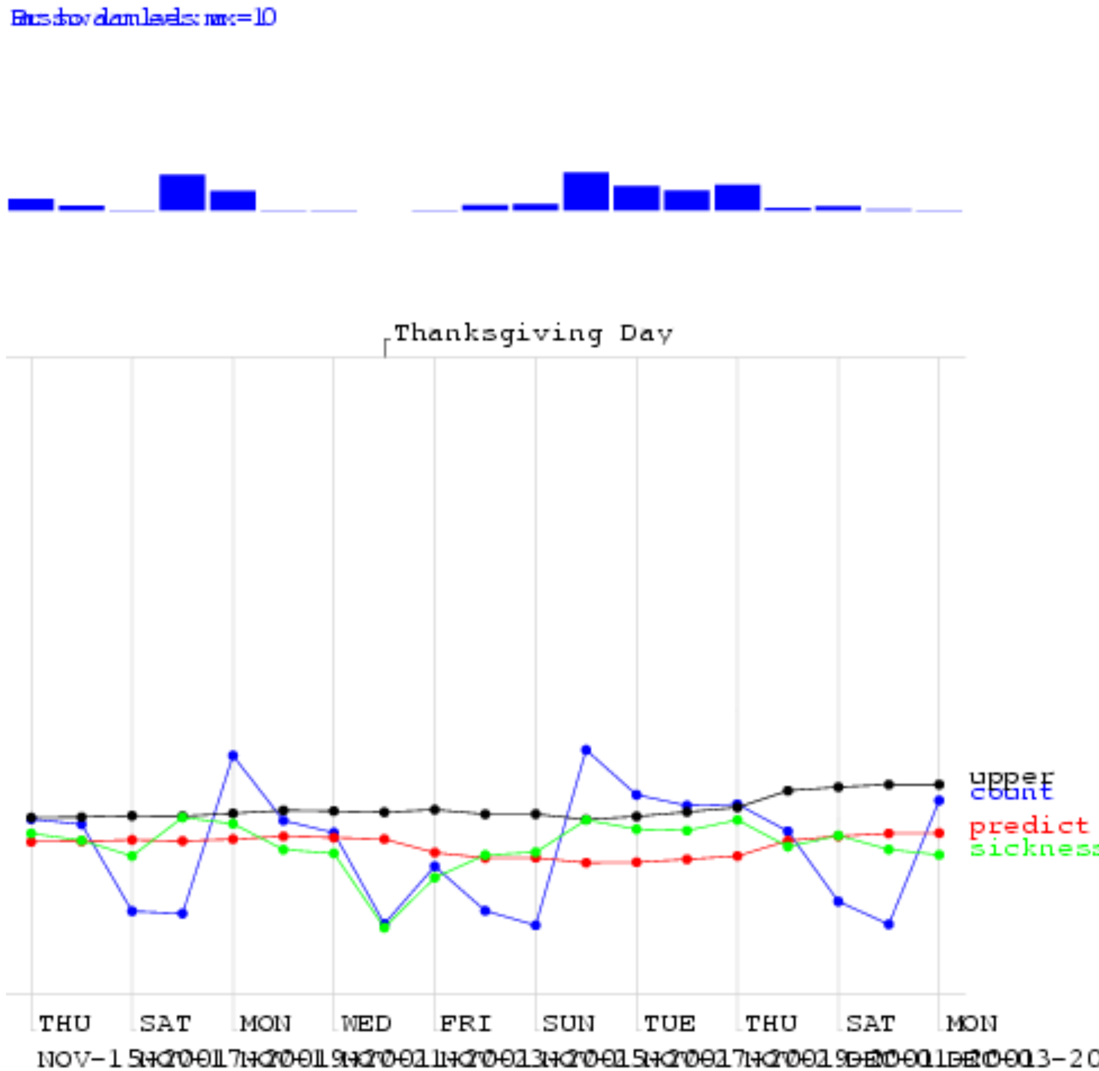
The Sickness/Availability Model



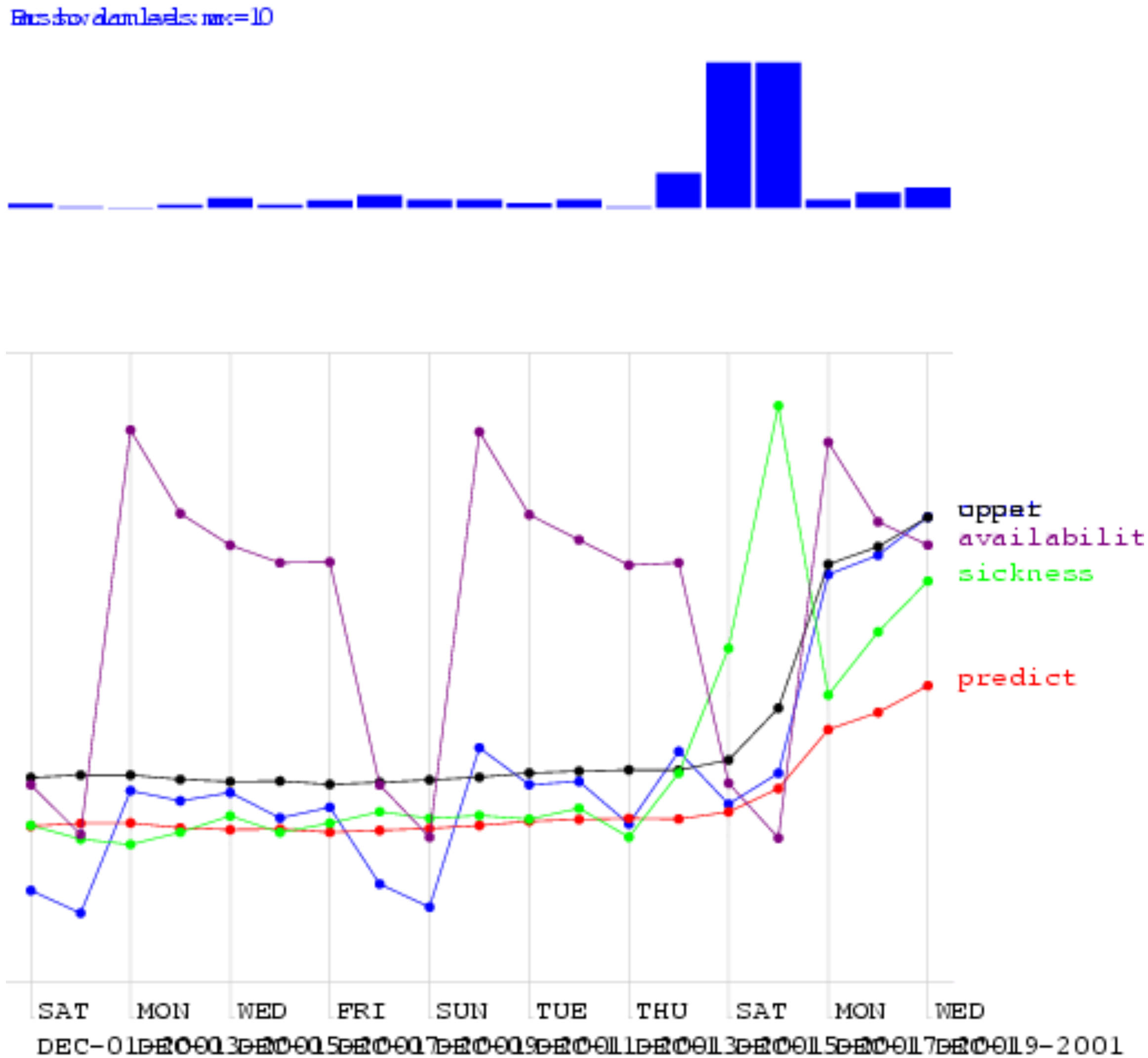
The Sickness/Availability Model



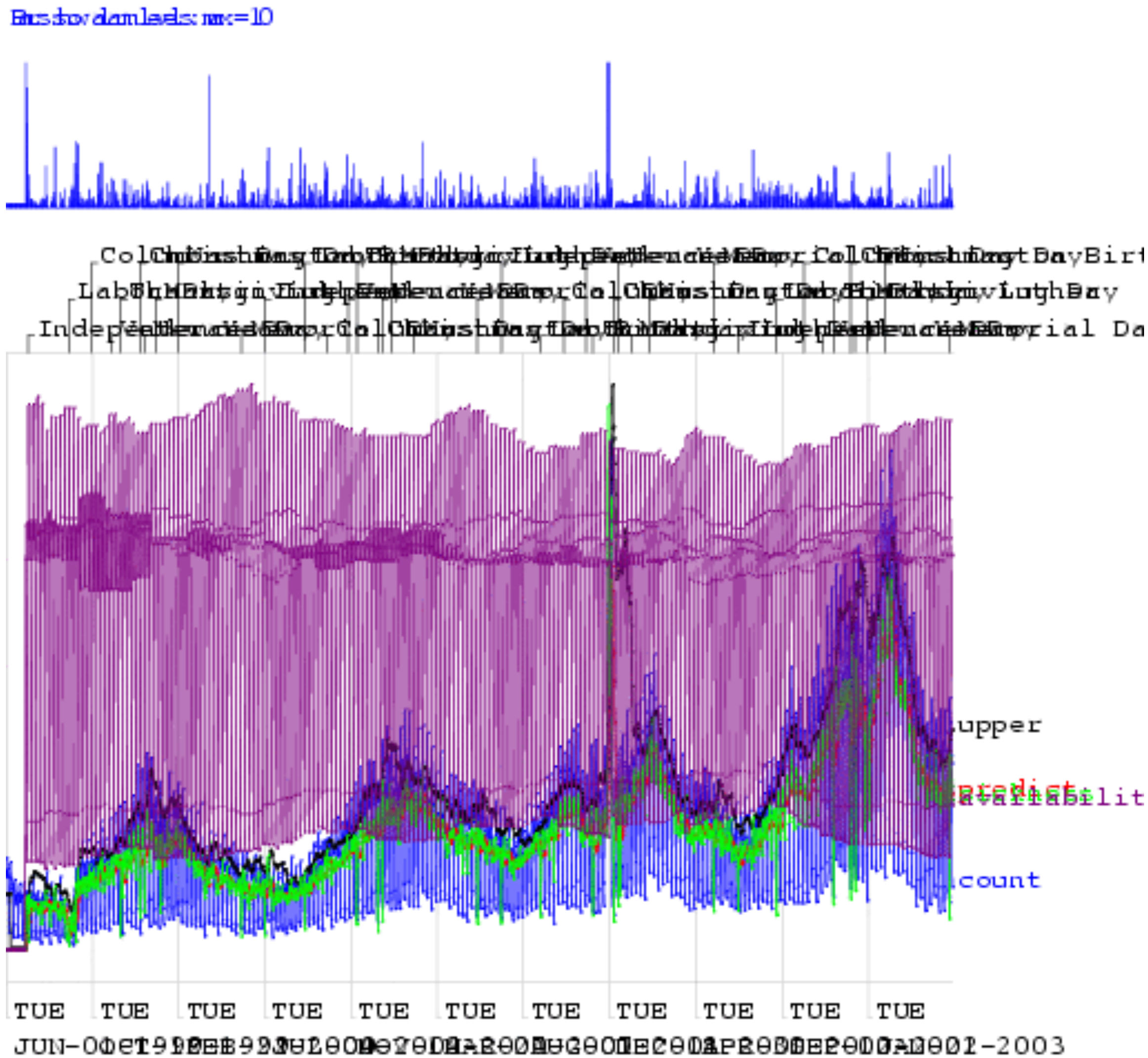
The Sickness/Availability Model



The Sickness/Availability Model



The Sickness/Availability Model



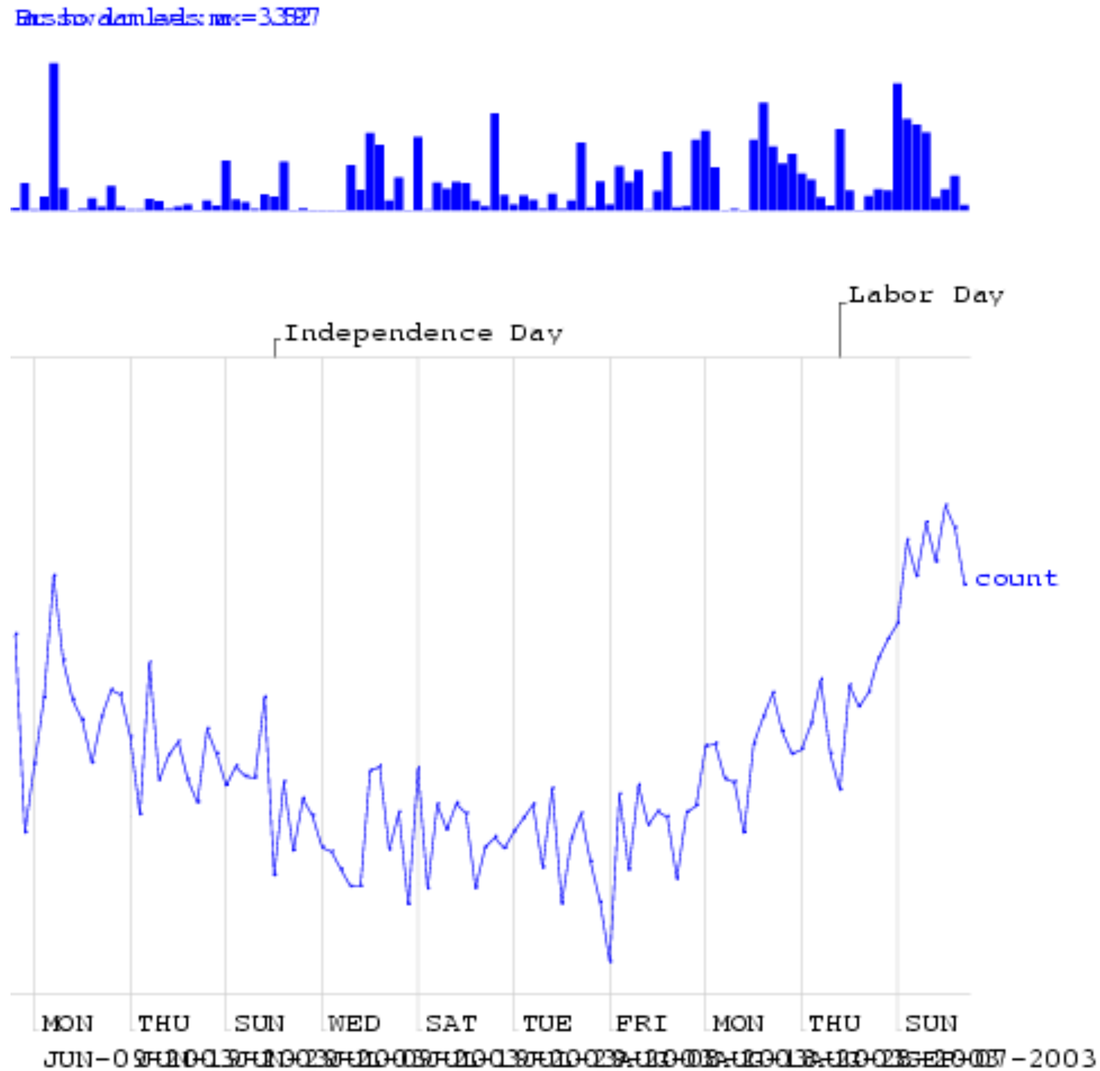
Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9
hours_of_daylight is_mon	0.7	2.25	0.57	3.12
hours_of_daylight is_mon … is_tue	0.72	1.83	0.57	3.16
hours_of_daylight is_mon … is_sat	0.77	2.11	0.59	3.26
CUSUM	0.45	2.03	0.15	3.55
sa-mav-1	0.86	1.88	0.74	2.73
sa-mav-7	0.87	1.28	0.83	1.87
sa-mav-14	0.86	1.27	0.82	1.62

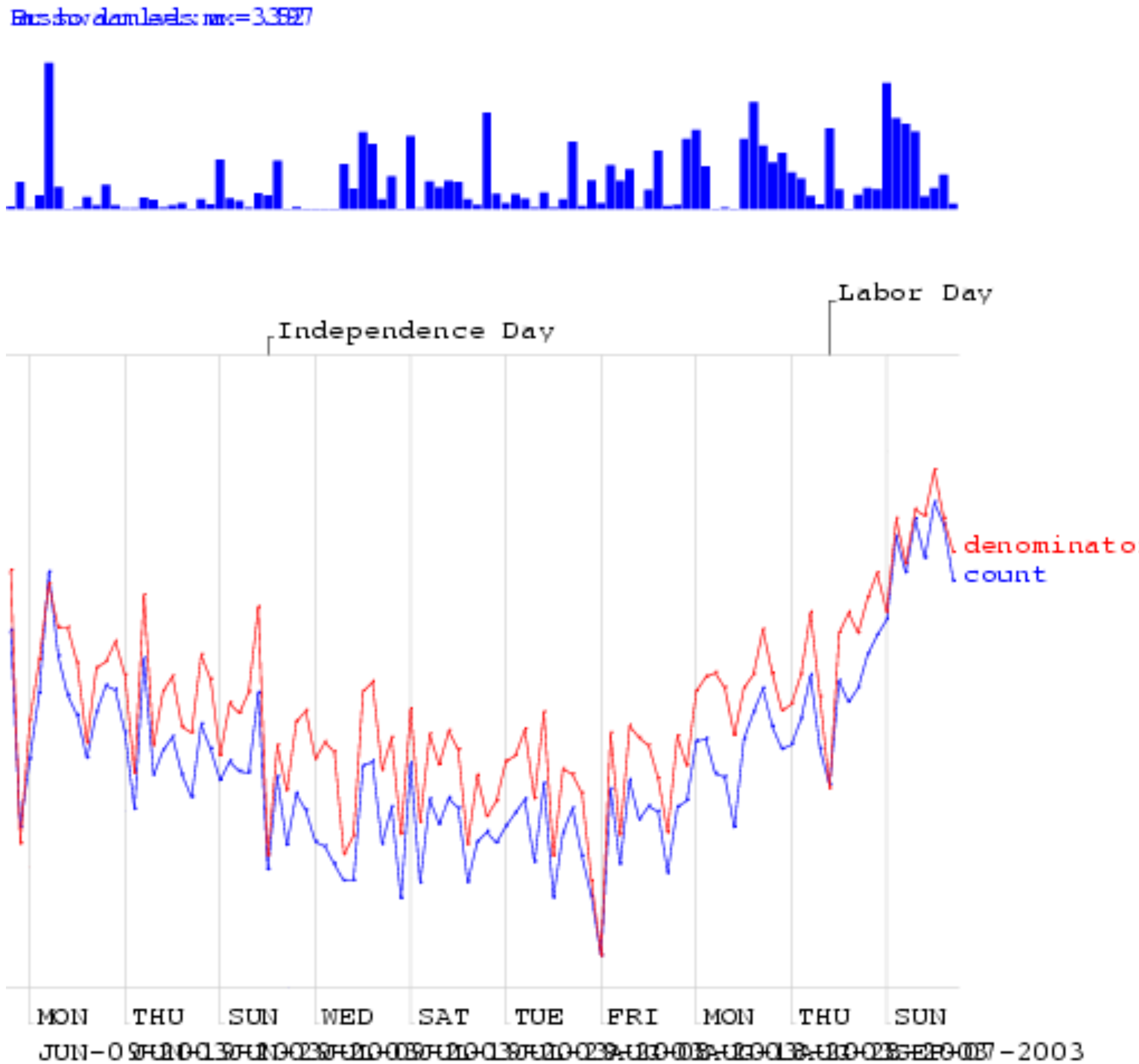
Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9
hours_of_daylight is_mon	0.7	2.25	0.57	3.12
hours_of_daylight is_mon … is_tue	0.72	1.83	0.57	3.16
hours_of_daylight is_mon … is_sat	0.77	2.11	0.59	3.26
CUSUM	0.45	2.03	0.15	3.55
sa-mav-1	0.86	1.88	0.74	2.73
sa-mav-7	0.87	1.28	0.83	1.87
sa-mav-14	0.86	1.27	0.82	1.62
sa-regress	0.73	1.76	0.67	2.21

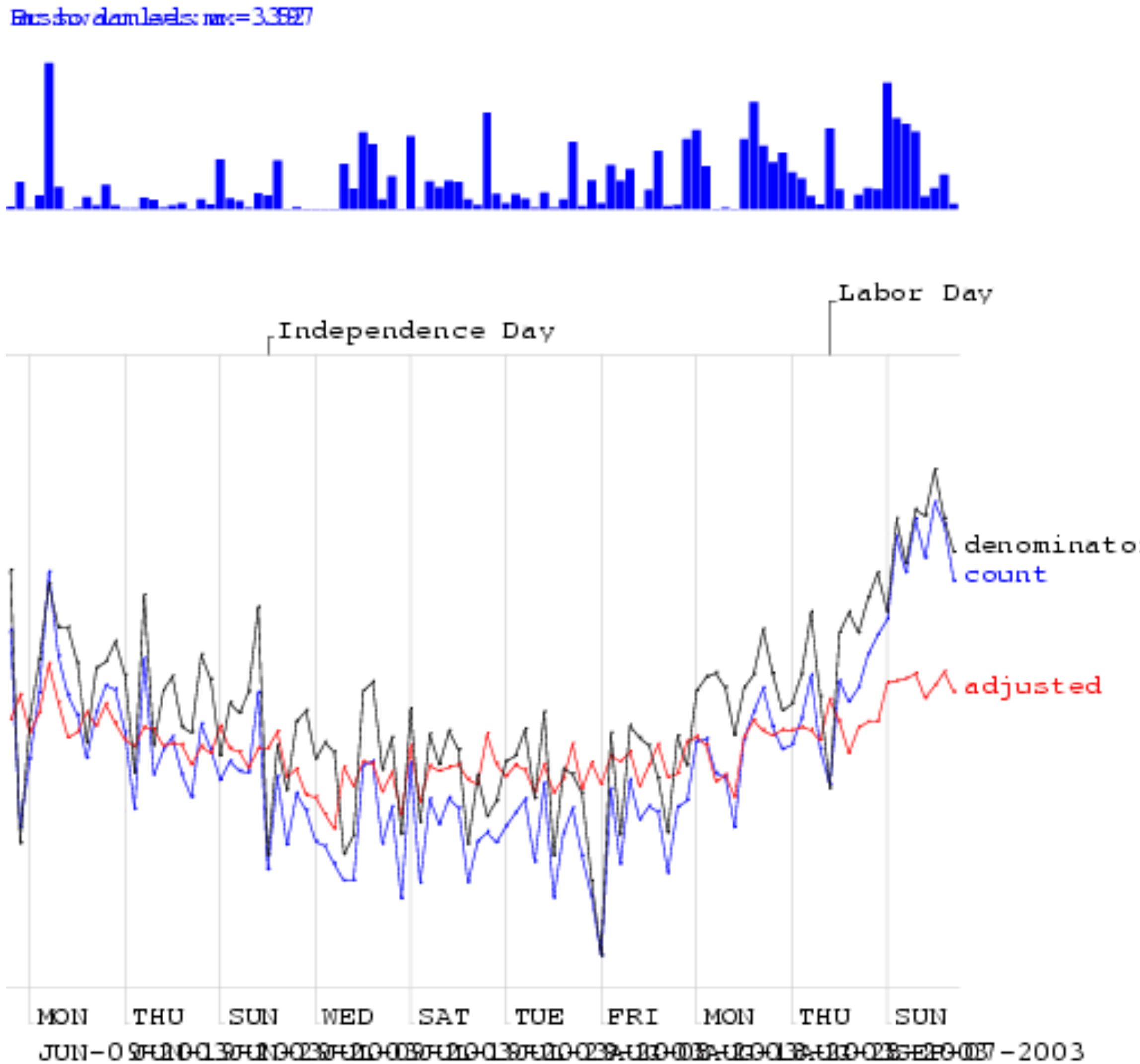
Exploiting Denominator Data



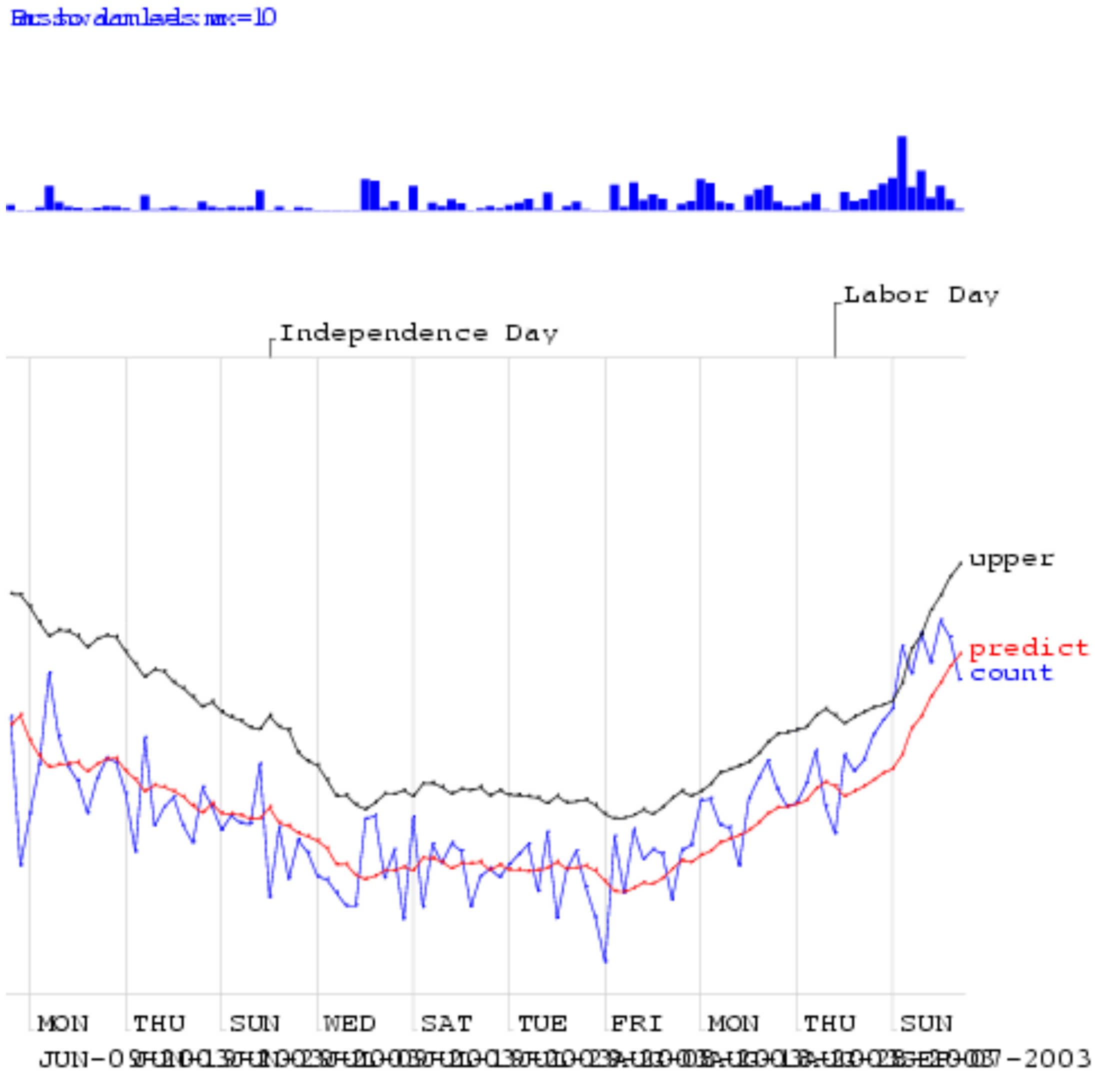
Exploiting Denominator Data



Exploiting Denominator Data



Exploiting Denominator Data



Algorithm Performance

	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	0.33	3.79
Moving Average 7	0.58	2.79	0.51	3.31
Moving Average 56	0.54	2.72	0.44	3.54
hours_of_daylight	0.58	2.73	0.43	3.9
hours_of_daylight is_mon	0.7	2.25	0.57	3.12
hours_of_daylight is_mon ... is_tue	0.72	1.83	0.57	3.16
hours_of_daylight is_mon ... is_sat	0.77	2.11	0.59	3.26
CUSUM	0.45	2.03	0.15	3.55
sa-mav-1	0.86	1.88	0.74	2.73
sa-mav-7	0.87	1.28	0.83	1.87
sa-mav-14	0.86	1.27	0.82	1.62
sa-regress	0.73	1.76	0.67	2.21
Cough with denominator	0.78	2.15	0.59	2.41
Cough with MA	0.65	2.78	0.57	3.24

Other state-of-the-art methods

- Wavelets
- Change-point detection
- Kalman filters
- Hidden Markov Models

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

Univariate Anomaly Detection

Multivariate Anomaly Detection

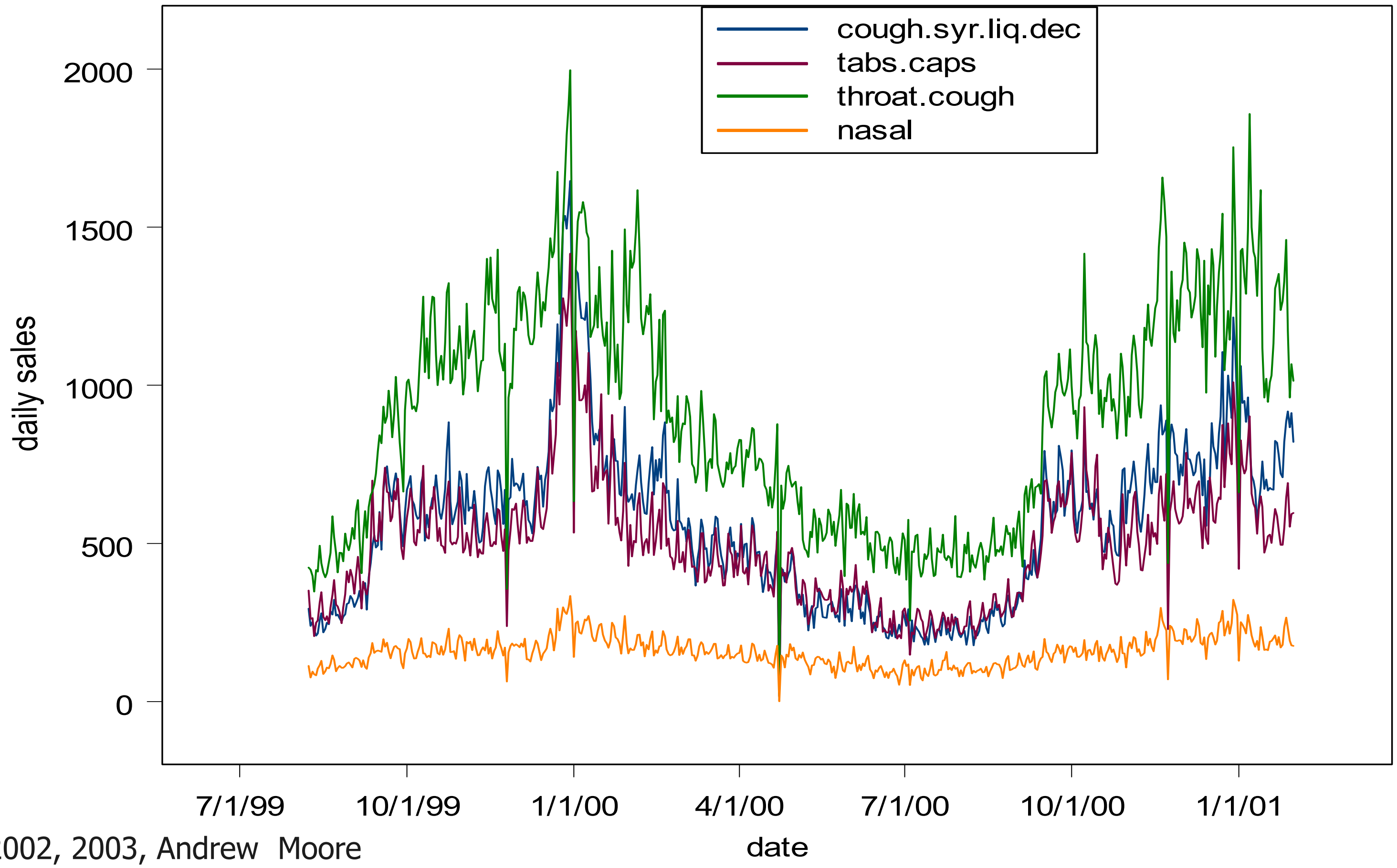
Spatial Scan Statistics

WSARE

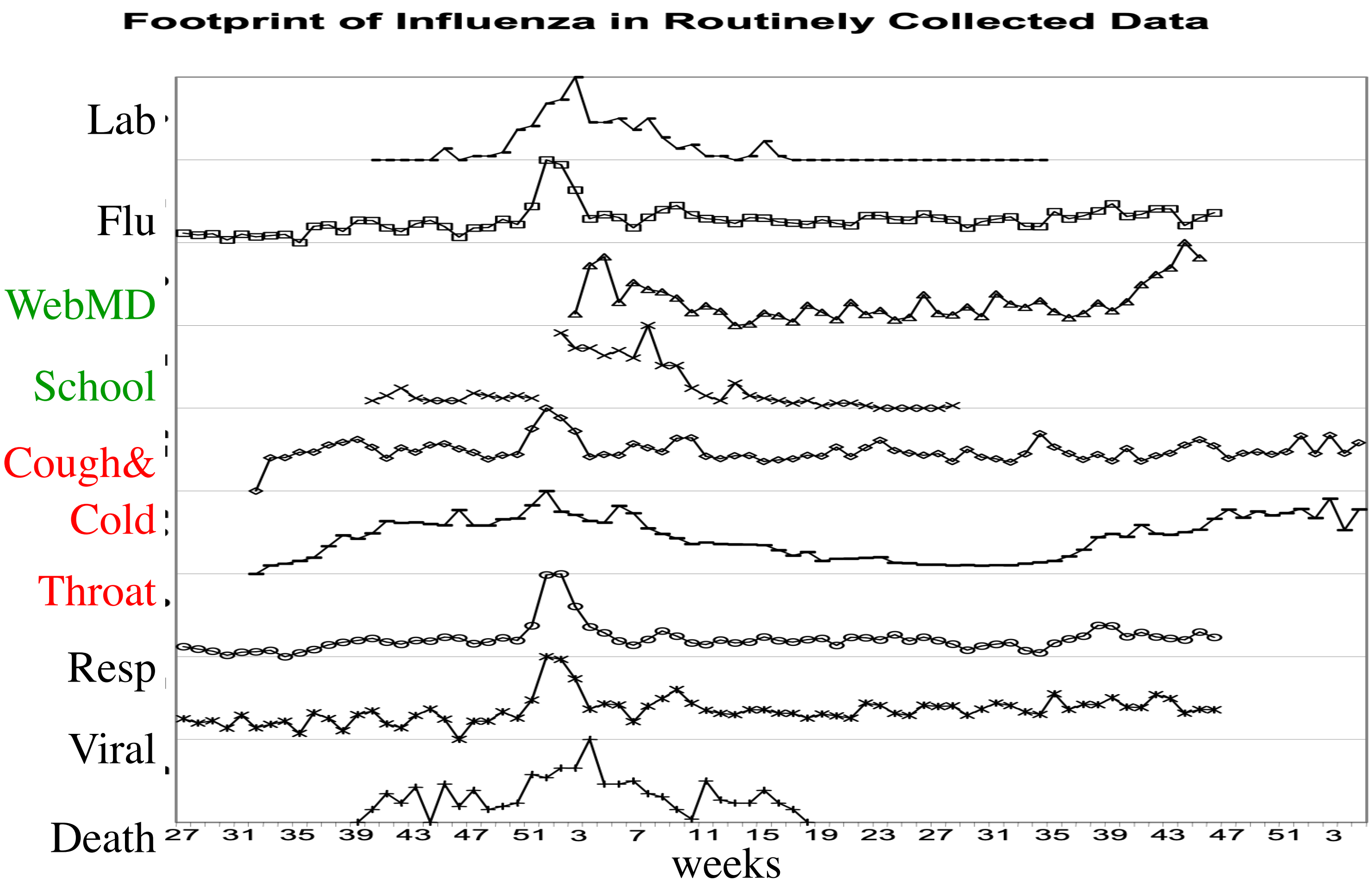
Multiple Signals

Multivariate Signals

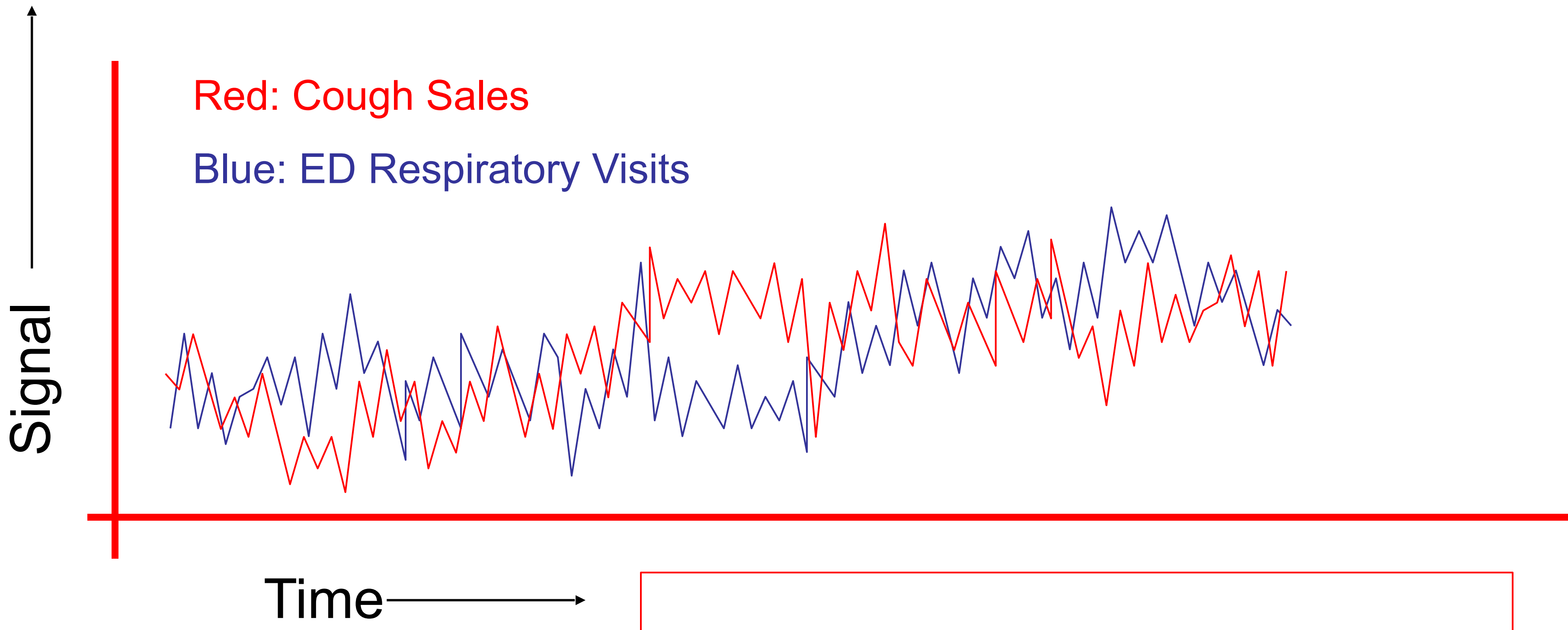
(relevant to inhalational diseases)



Multi Source Signals



What if you've got multiple signals?

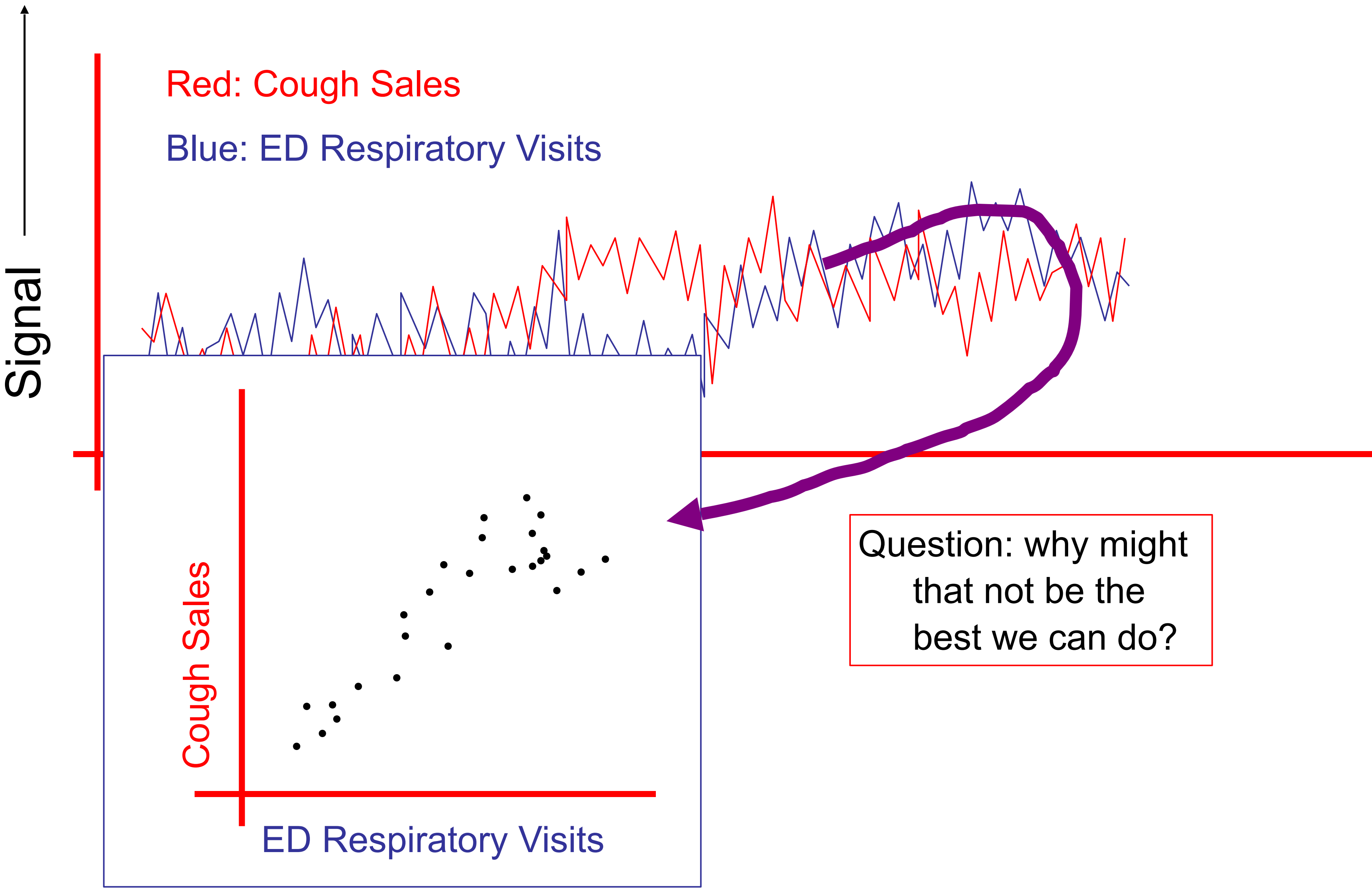


Idea One:

Simply treat it as two separate alarm-from-signal problems.

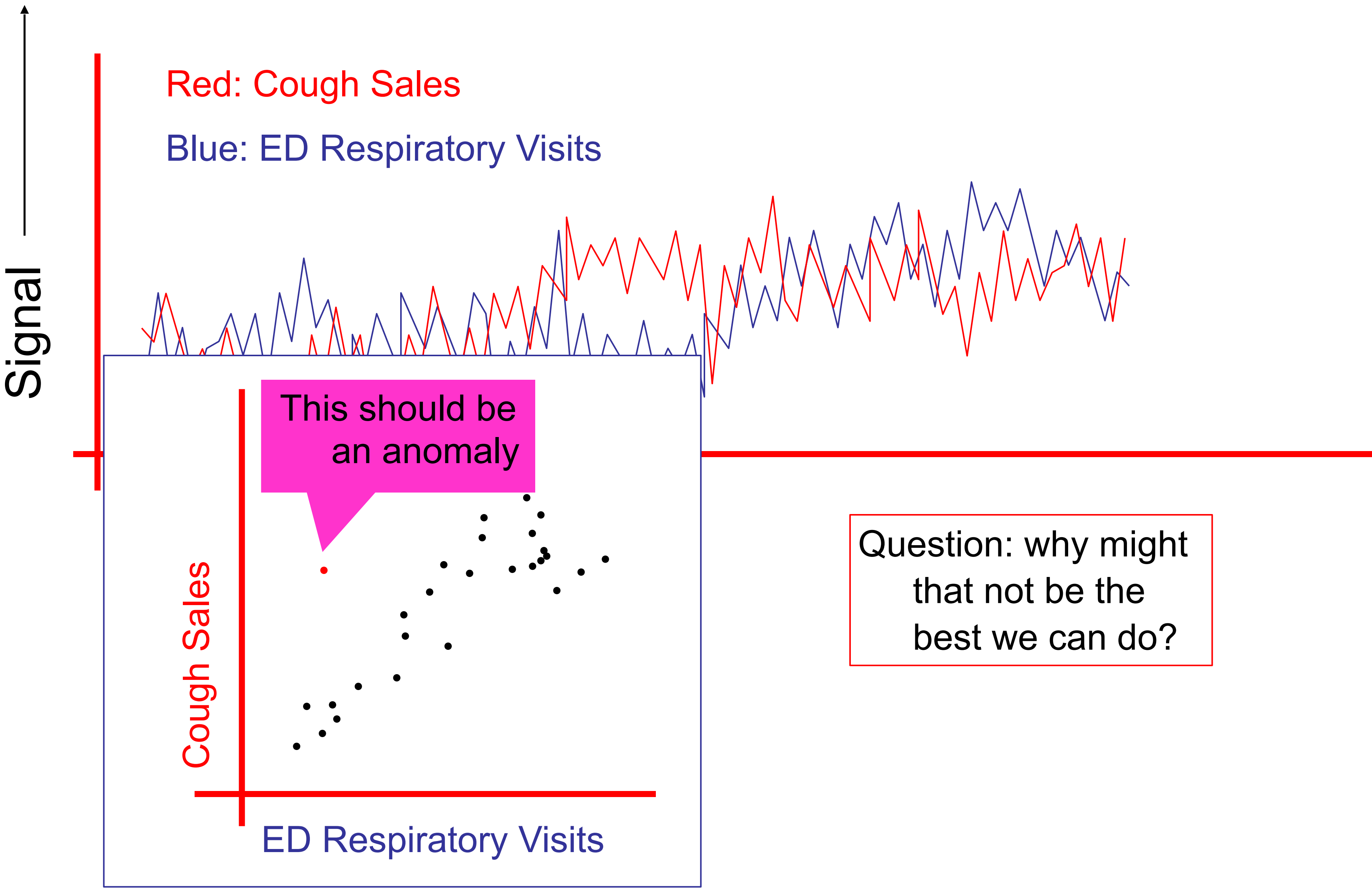
…Question: why might that not be the best we can do?

Another View



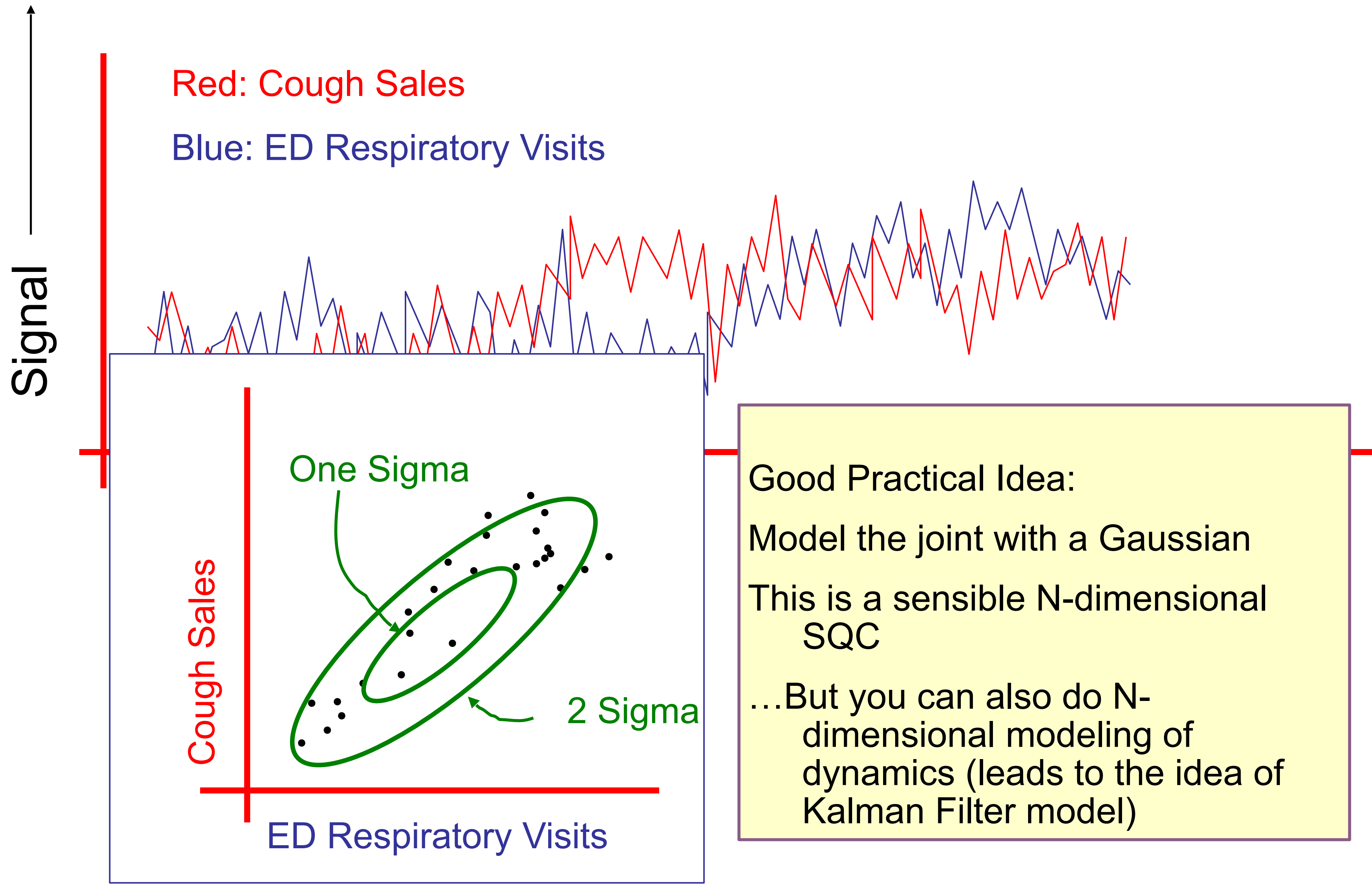
Question: why might that not be the best we can do?

Another View



Question: why might that not be the best we can do?

N-dimensional Gaussian



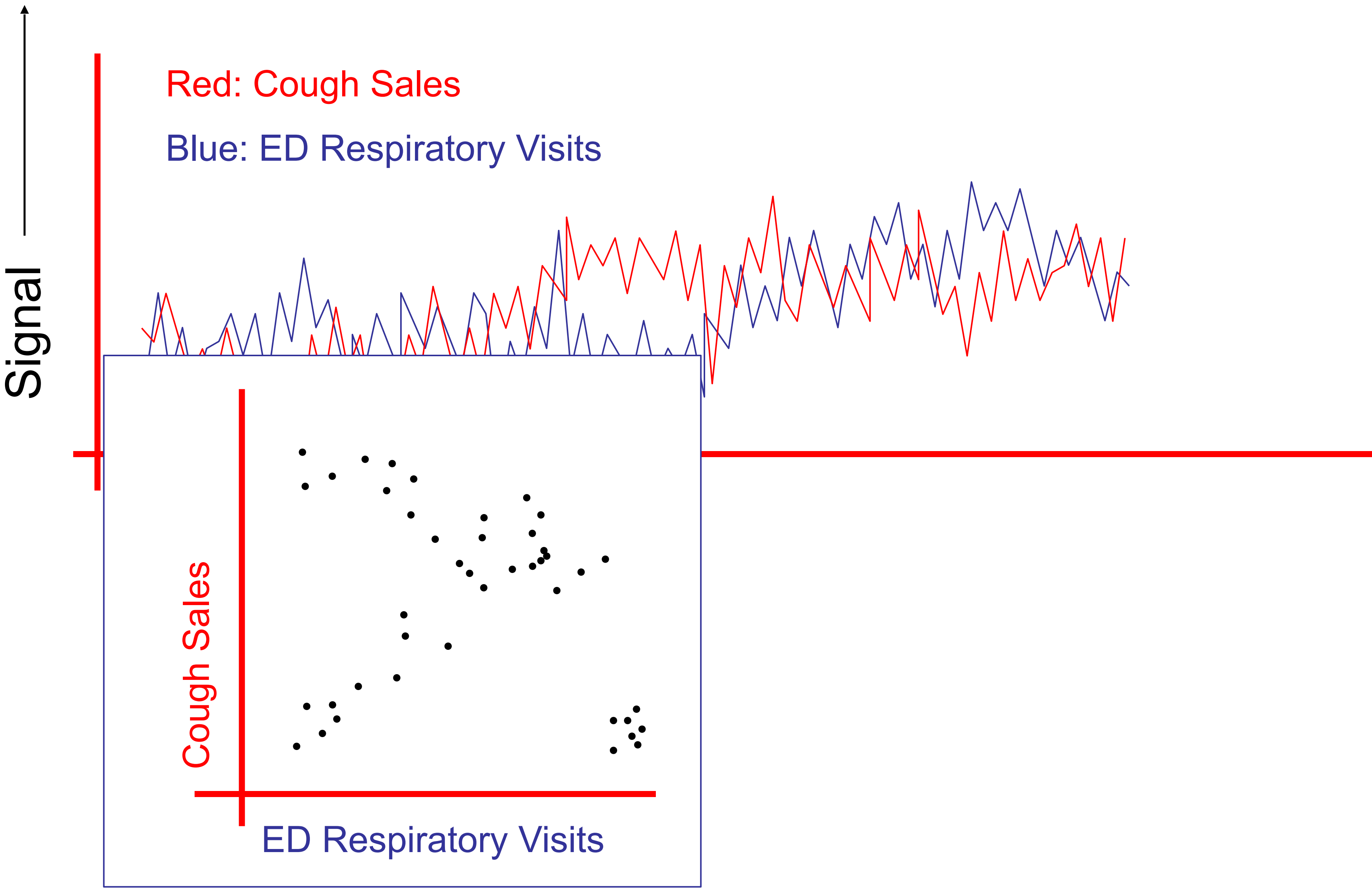
Good Practical Idea:

Model the joint with a Gaussian

This is a sensible N-dimensional SQC

…But you can also do N-dimensional modeling of dynamics (leads to the idea of Kalman Filter model)

But what if joint N-dimensional distribution is highly non-Gaussian?



What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

Univariate Anomaly Detection

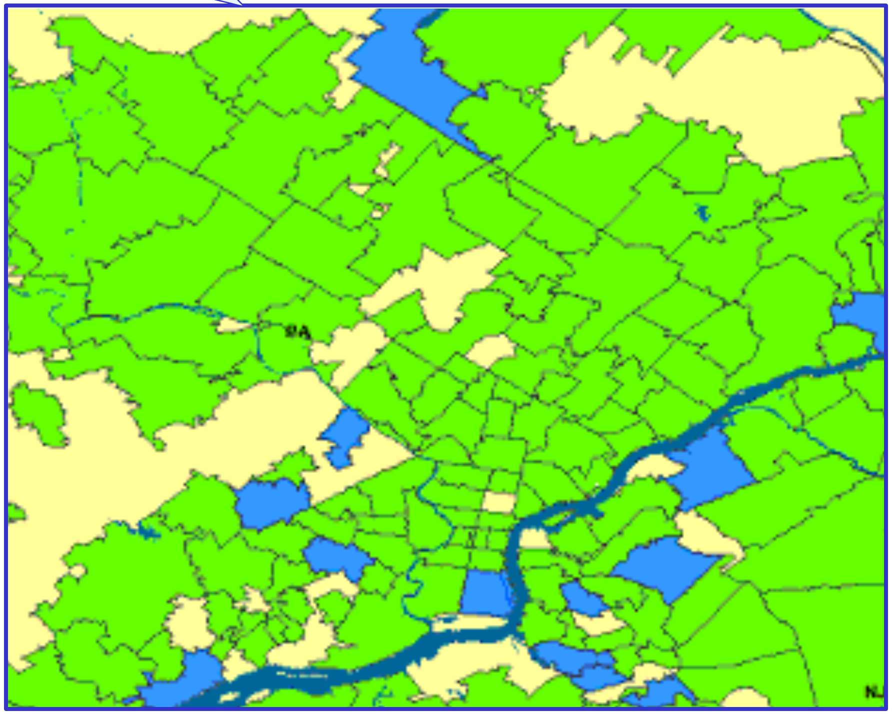
Multivariate Anomaly Detection

Spatial Scan Statistics

WSARE

One Step of Spatial Scan

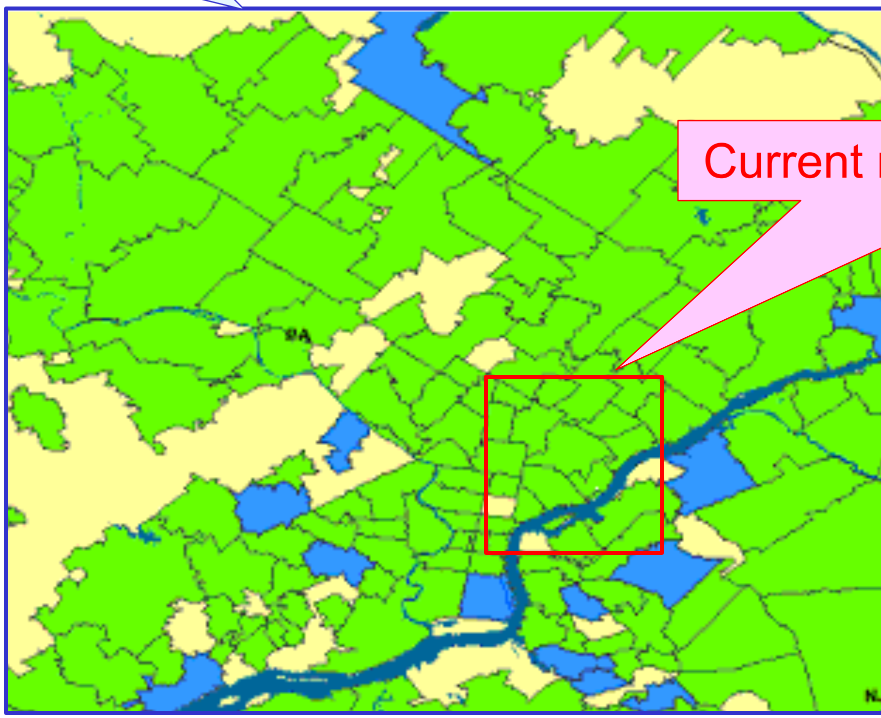
Entire area being scanned



One Step of Spatial Scan

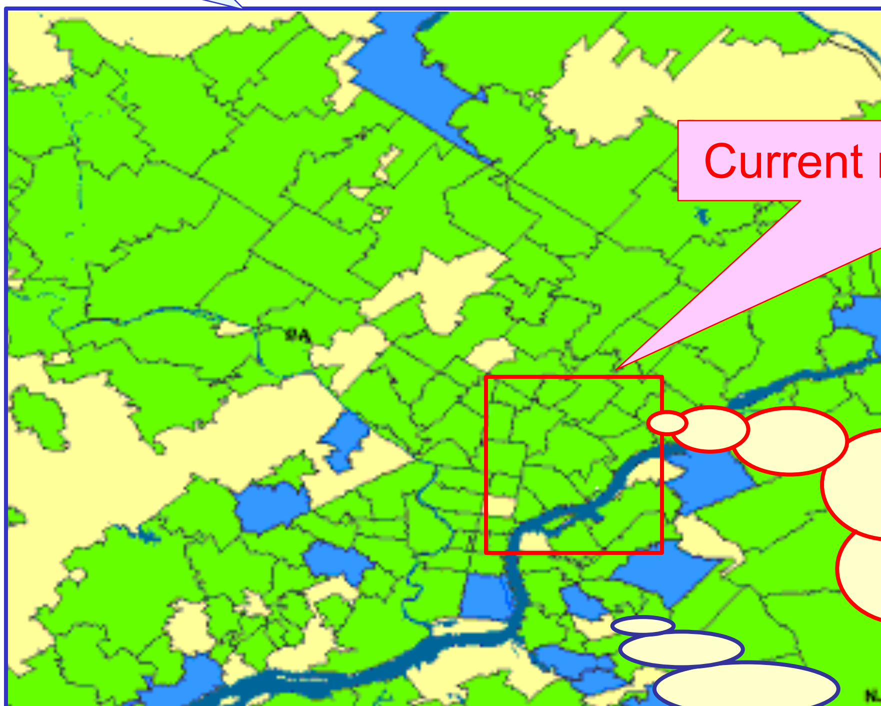
Entire area being scanned

Current region being considered



One Step of Spatial Scan

Entire area being scanned



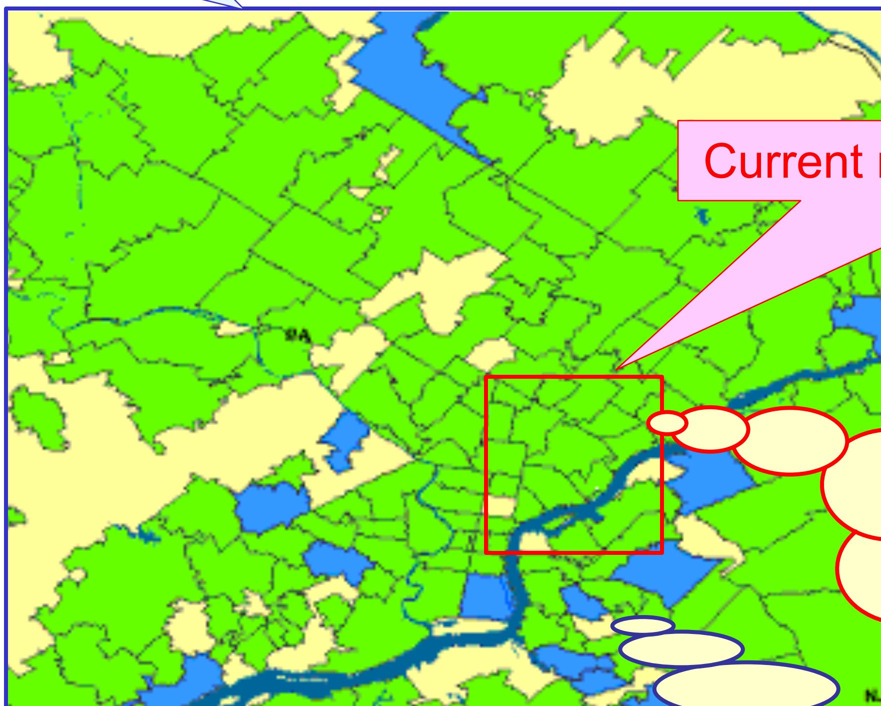
Current region being considered

I have a population of 5300 of whom 53 are sick (1%)

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

One Step of Spatial Scan

Entire area being scanned



Current region being considered

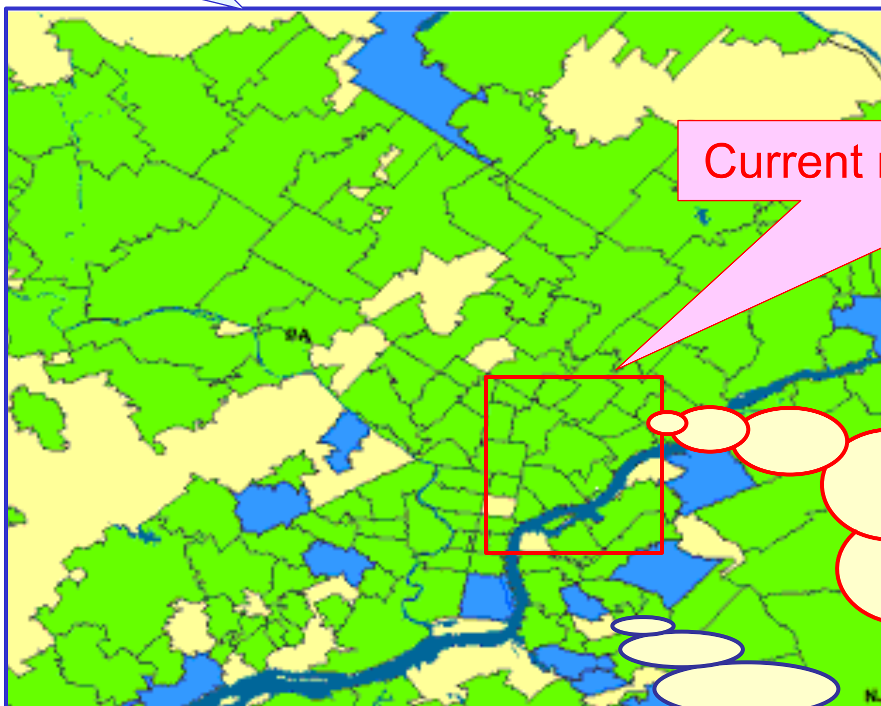
I have a population of 5300 of whom 53 are sick (1%)

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

So... *is that a big deal? Evaluated with Score function (e.g. Kulldorf's score)*

One Step of Spatial Scan

Entire area being scanned



Current region being considered

I have a population of 5300 of whom 53 are sick (1%)

[Score = 1.4]

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

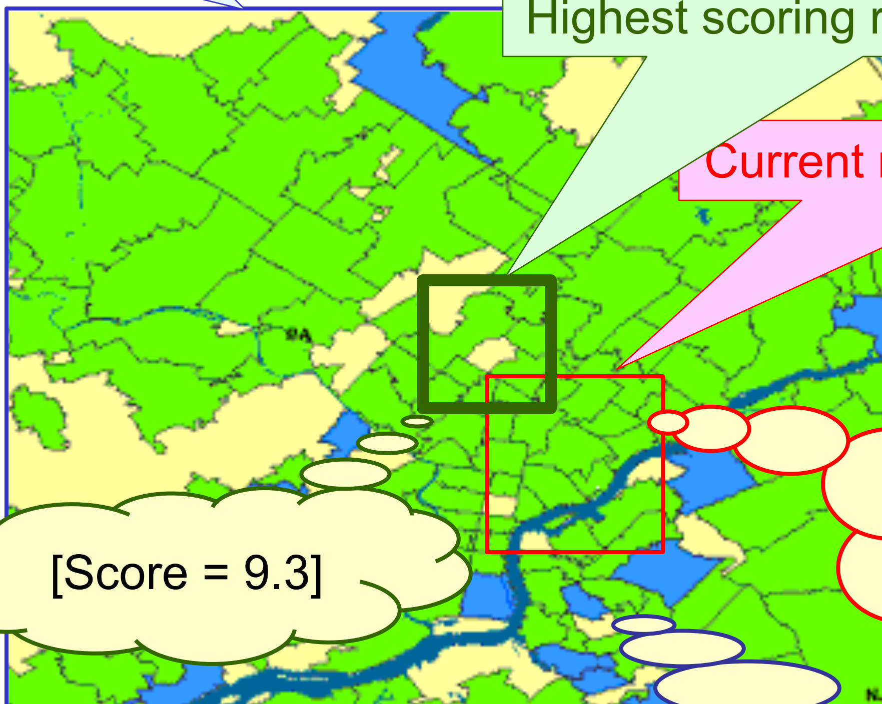
So… *is that a big deal? Evaluated with Score function (e.g. Kulldorf's score)*

Many Steps of Spatial Scan

Entire area being scanned

Highest scoring region in search so far

Current region being considered



[Score = 9.3]

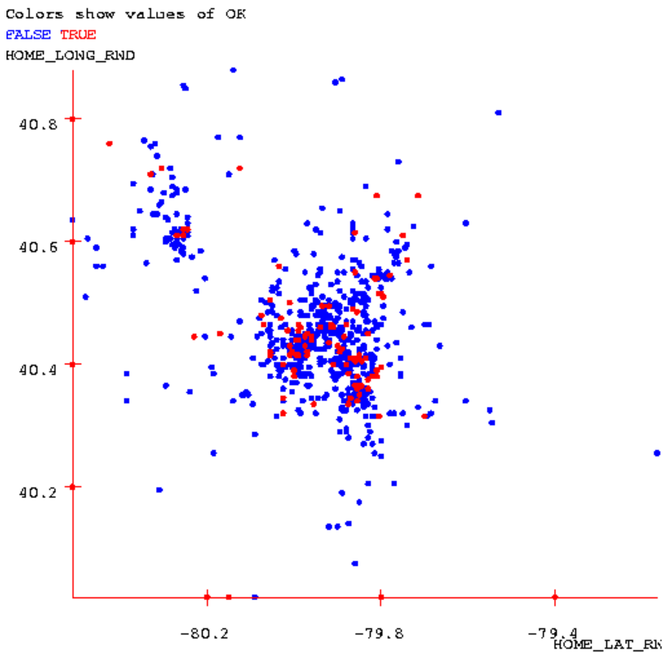
I have a population of 5300 of whom 53 are sick (1%)

[Score = 1.4]

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

So… *is that a big deal? Evaluated with Score function (e.g. Kulldorf's score)*

Scan Statistics



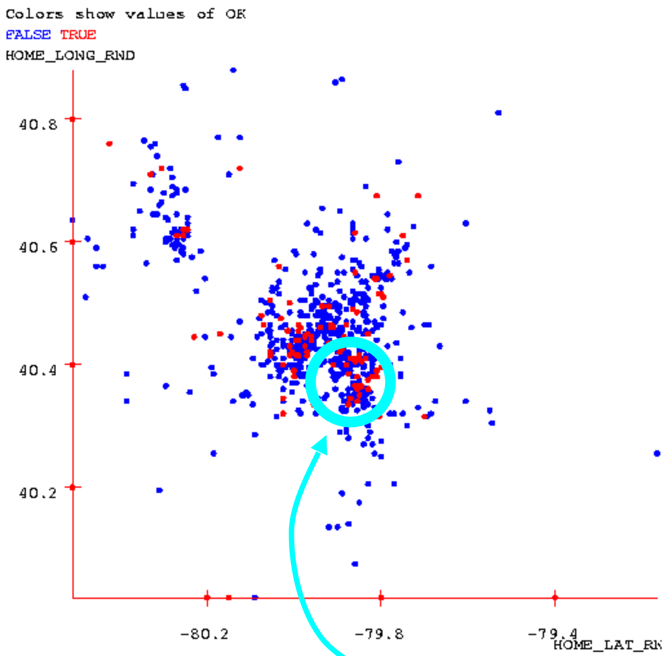
Standard scan statistic question:
Given the geographical locations of occurrences of a phenomenon, is there a region with an unusually high (low) rate of these occurrences?

Standard approach:

1. Compute the likelihood of the data given the hypothesis that the rate of occurrence is uniform everywhere, L_0
2. For some geographical region, W, compute the likelihood that the rate of occurrence is uniform at one level inside the region and uniform at another level outside the region, L(W).
3. Compute the likelihood ratio, $L(W)/L_0$
4. Repeat for all regions, and find the largest likelihood ratio. This is the scan statistic, S^*_W
5. Report the region, W, which yielded the max, S^*_W

See [Glaz and Balakrishnan, 99] for details

Significance testing

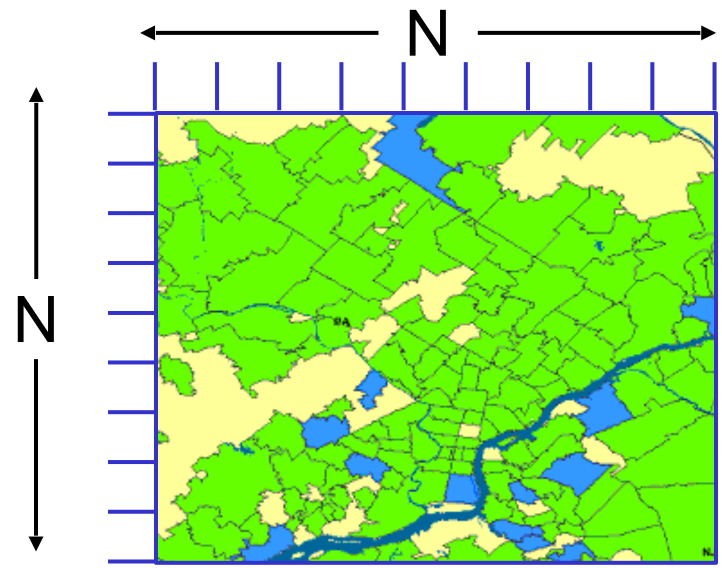


Given that region W is the most likely to be abnormal, is it significantly abnormal?

Standard approach:

1. Generate many randomized versions of the data set by shuffling the labels (positive instance of the phenomenon or not).
2. Compute S^*_W for each randomized data set. This forms a baseline distribution for S^*_W if the null hypothesis holds.
3. Compare the observed value of S^*_W against the baseline distribution to determine a p-value.

Fast squares speedup



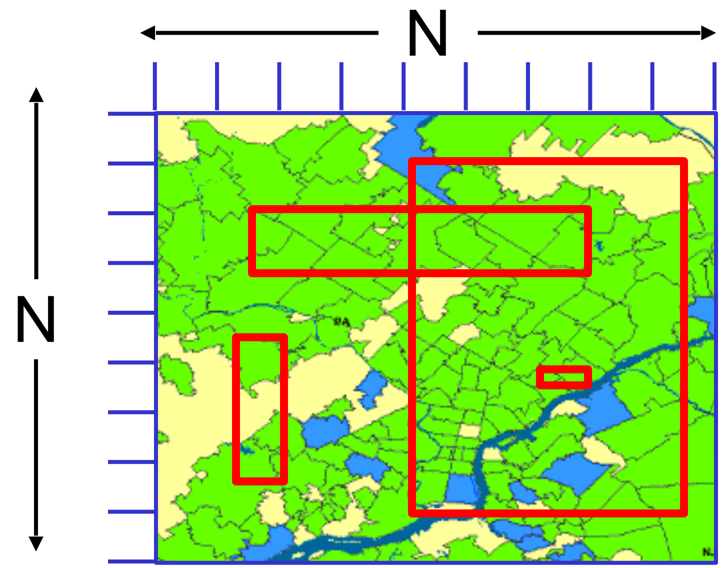
- Theoretical complexity of fast squares: $O(N^2)$ (as opposed to naïve N^3), if maximum density region sufficiently dense.

  If not, we can use several other speedup tricks.
- In practice: 10-200x speedups on real and artificially generated datasets.

  Emergency Dept. dataset (600K records): 20 minutes, versus 66 hours with naïve approach.

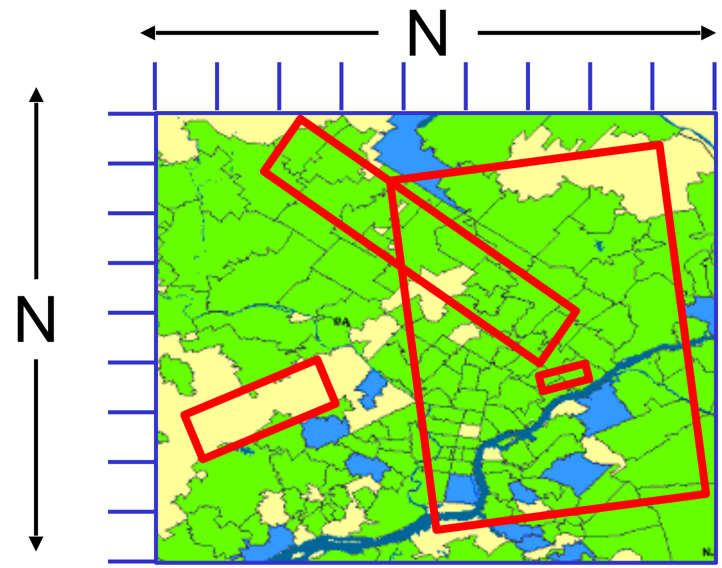
Fast rectangles speedup

Work in progress



- Theoretical complexity of fast rectangles: $O(N^2 \log N)$ (as opposed to naïve N^4)

Fast oriented rectangles speedup



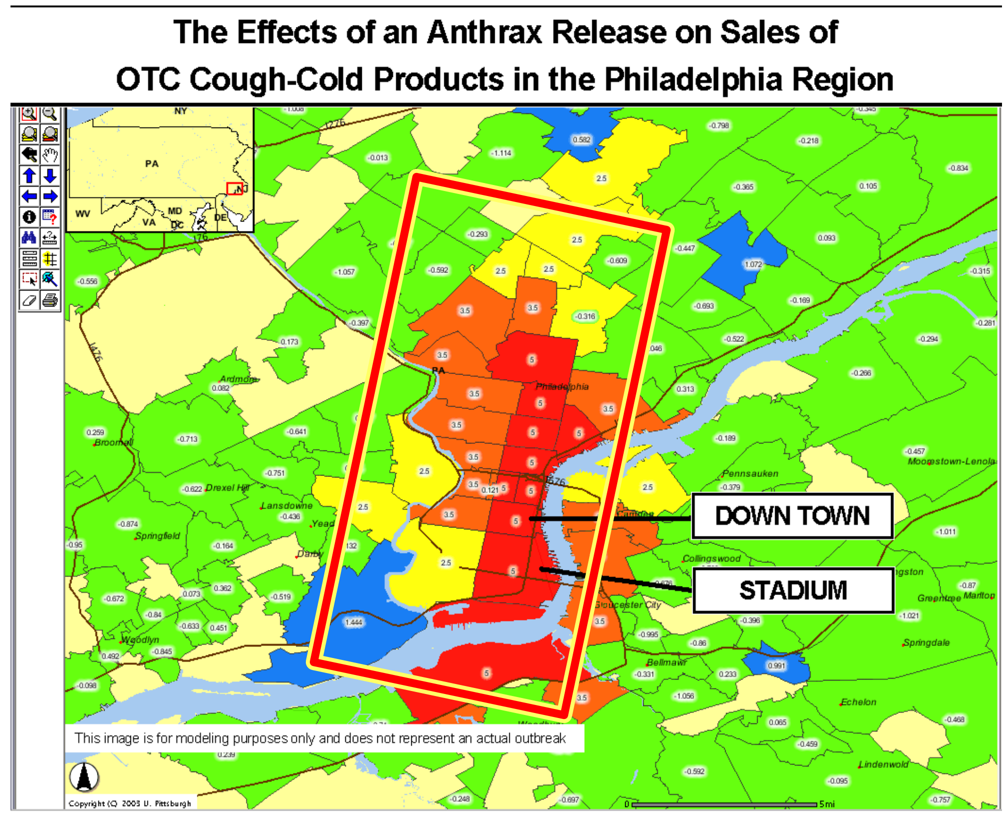
Work in progress

- Theoretical complexity of fast rectangles: $18N^2 \log N$ (as opposed to naïve $18N^4$)

(Angles discretized to 5 degree buckets)

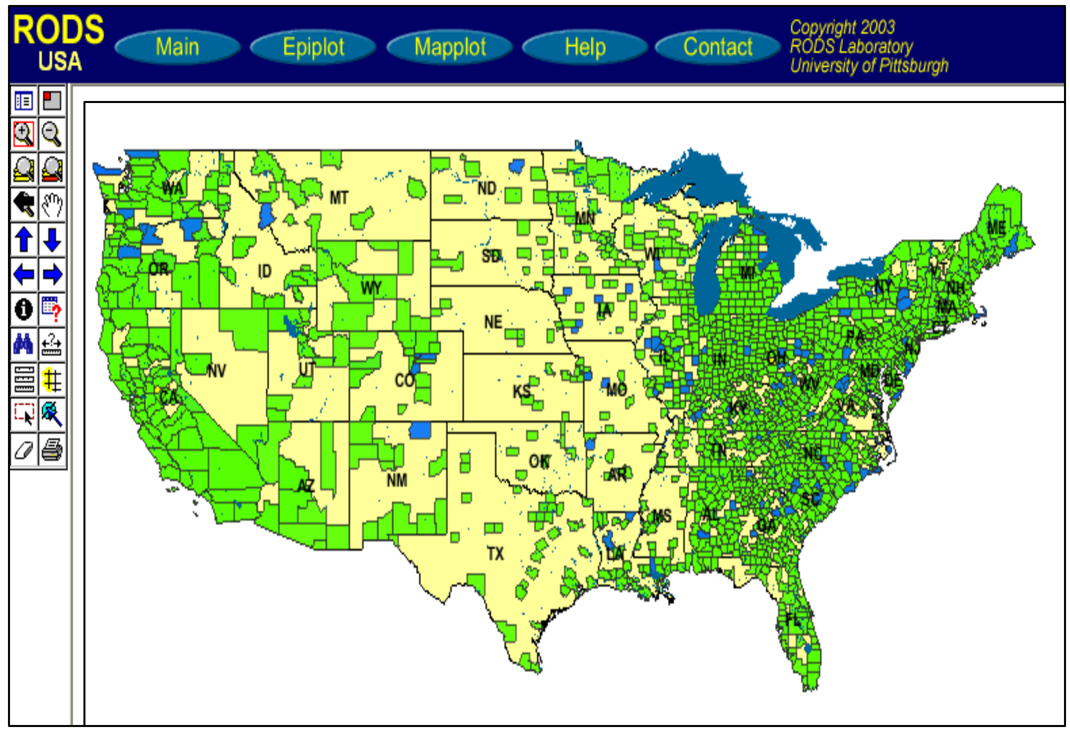
Why the Scan Statistic speed obsession?

- Traditional Scan Statistics very expensive, especially with Randomization tests
- New "Historical Model" Scan Statistics
- Proposed new WSARE/Scan Statistic hybrid



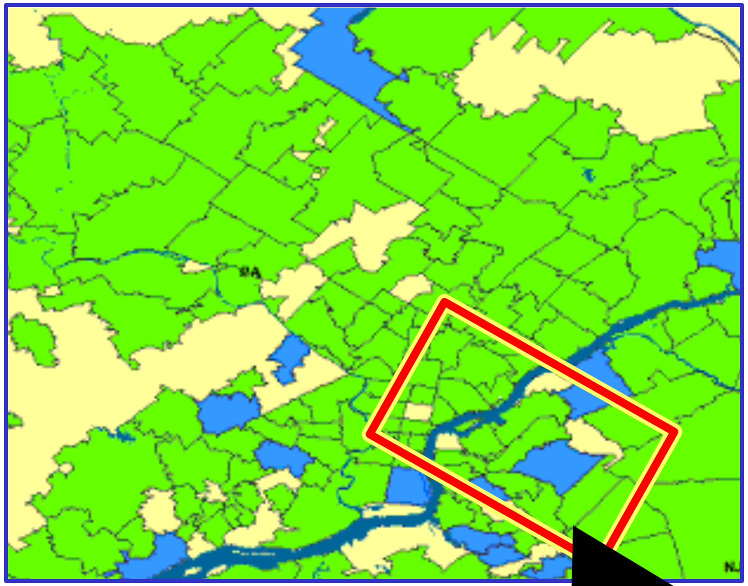
Why the Scan Statistic speed obsession?

- Traditional Scan Statistics very expensive, especially with Randomization tests
- New "Historical Model" Scan Statistics
- Proposed new WSARE/Scan Statistic hybrid



Why the Scan Statistic speed obsession?

- Traditional Scan Statistics very expensive, especially with Randomization tests
- New "Historical Model" Scan Statistics
- Proposed new WSARE/Scan Statistic hybrid



This is the strangest region because the age distribution of respiratory cases has changed dramatically for no reason that can be explained by known background changes

What you'll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

Univariate Anomaly Detection

Multivariate Anomaly Detection

Spatial Scan Statistics

WSARE

But there's potentially more data than aggregates

Suppose we know that today in the ED we had…

- 421 Cases
- 78 Respiratory Cases
- 190 Males
- 32 Children
- 21 from North Suburbs
- 2 Postal workers

(etc etc etc)

Have we made best use of all possible information?

There are so many things to look at

Human Analysts

Massive Computer Analysis

WSARE v2.0

- What's Strange About Recent Events?
- Designed to be easily applicable to any date/time-indexed biosurveillance-relevant data stream.

WSARE v2.0

- Inputs:

1. Date/time-indexed biosurveillance-relevant data stream

2. Time Window Length

3. Which attributes to use?

WSARE v2.0

- Inputs:

1. Date/time-indexed biosurveillance-relevant data stream

2. Time Window Length

3. Which attributes to use?

Example

"last 24 hours"

"ignore key and weather"

Primary Key	Date	Time	Hospital	ICD 9	Prodrome	Gender	Age	Home			Work			Recent Flu Levels	Recent Weather	(Many more…)
								Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale			
h6r32	6/2/2	14:12	Down-town	781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	…
t3q15	6/2/2	14:15	River-side	717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	…
t5hh5	6/2/2	14:15	Smith-field	622	Respiratory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	…
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

WSARE v2.0

- Inputs:

1. Date/time-indexed biosurveillance-relevant data stream

2. Time Window Length

3. Which attributes to use?

- Outputs:

1. Here are the records that most surprise me

2. Here's why

3. And here's how seriously you should take it

Primary Key	Date	Time	Hospital	ICD 9	Prodrome	Gender	Age	Home			Work			Recent Flu Levels	Recent Weather	(Many more…)
								Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale			
h6r32	6/2/2	14:12	Down-town	781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	…
t3q15	6/2/2	14:15	River-side	717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	…
t5hh5	6/2/2	14:15	Smith-field	622	Respiratory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	…
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Simple WSARE

- Given 500 day's worth of ER cases at 15 hospitals…

Date	Cases
Thu 5/22/2000	C1, C2, C3, C4 …
Fri 5/23/2000	C1, C2, C3, C4 …
:	:
:	:
Sat 12/9/2000	C1, C2, C3, C4 …
Sun 12/10/2000	C1, C2, C3, C4 …
:	:
Sat 12/16/2000	C1, C2, C3, C4 …
:	:
Sat 12/23/2000	C1, C2, C3, C4 …
:	:
:	:
Fri 9/14/2001	C1, C2, C3, C4 …

Simple WSARE

- Given 500 day's worth of ER cases at 15 hospitals…
- For each day…
  - Take today's cases

Date	Cases
Thu 5/22/2000	C1, C2, C3, C4 …
Fri 5/23/2000	C1, C2, C3, C4 …
:	:
:	:
Sat 12/9/2000	C1, C2, C3, C4 …
Sun 12/10/2000	C1, C2, C3, C4 …
:	:
Sat 12/16/2000	C1, C2, C3, C4 …
:	:
Sat 12/23/2000	C1, C2, C3, C4 …
:	:
:	:
Fri 9/14/2001	C1, C2, C3, C4 …

Simple WSARE

- Given 500 day's worth of ER cases at 15 hospitals…
- For each day…
  - Take today's cases
  - The cases one week ago
  - The cases two weeks ago

Date	Cases
Thu 5/22/2000	C1, C2, C3, C4 …
Fri 5/23/2000	C1, C2, C3, C4 …
:	:
:	:
Sat 12/9/2000	C1, C2, C3, C4 …
Sun 12/10/2000	C1, C2, C3, C4 …
:	:
Sat 12/16/2000	C1, C2, C3, C4 …
:	:
Sat 12/23/2000	C1, C2, C3, C4 …
:	:
:	:
Fri 9/14/2001	C1, C2, C3, C4 …

Simple WSARE

- Given 500 day's wort of ER cases at 15 hospitals…
- For each day…
  - Take today's cases
  - The cases one week ago
  - The cases two weeks ag
- Ask: "What's different about today?"

DATE_ADI	ICD9	PRODRON	GENDER	place2	…	…
					…	…
12/9/00	786.05	3	F	s-e	…	…
12/9/00	789	1	F	s-e	…	…
12/9/00	789	1	M	n-w	…	…
12/9/00	786.05	3	M	s-e	…	…
:	:	:	:	:	…	…
12/16/00	787.02	2	M	n-e	…	…
12/16/00	782.1	4	F	s-w	…	…
12/16/00	789	1	M	s-e	…	…
12/16/00	786.09	3	M	n-w	…	…
12/23/00	789.09	1	M	s-w	…	…
12/23/00	789.09	1	F	s-w	…	…
12/23/00	782.1	4	M	n-w	…	…
:	:	:	:	:	…	…
12/23/00	786.09	3	M	s-e	…	…
12/23/00	786.09	3	M	s-e	…	…
12/23/00	780.9	2	F	n-w	…	…
12/23/00	V40.9	7	M	s-w	…	…

Simple WSARE

- Given 500 day's wort
  of ER cases at 15
  hospitals…
- For each day…

DATE_ADI	ICD9	PRODRON	GENDER	place2	…	…
					…	…
12/9/00	786.05	3	F	s-e	…	…
12/9/00	789	1	F	s-e	…	…
12/9/00	789	1	M	n-w	…	…
12/9/00	786.05	3	M	s-e	…	…
:	:	:	:	:	…	…
12/16/00	787.02	2	M	n-e	…	…
12/16/00	782.1	4	F	s-w	…	…
12/16/00	789	1	M	s-e	…	…
12/16/00	786.09	3	M	n-w	…	…
12/23/00	[illegible]		[illegible]	s-w	…	…
		1	F	s-w	…	…
				n-w	…	…
				:	…	…
				s-e	…	…
				s-e	…	…
				n-w	…	…
				s-w	…	…

Fields we use:

Date, Time of Day, Prodrome, ICD9, ***Symptoms***, Age, Gender, Coarse Location, Fine Location, ***ICD9 Derived Features, Census Block Derived Features, Work Details, Colocation Details***

Example

```
Sat 12-23-2001 (daynum 36882, dayindex 239)

35.8% ( 48/134) of today's cases have 30 <= age < 40
17.0% ( 45/265) of   other cases have 30 <= age < 40
```

Example

```
Sat 12-23-2001 (daynum 36882, dayindex 239)
FISHER_PVALUE = 0.000051
35.8% ( 48/134) of today's cases have 30 <= age < 40
17.0% ( 45/265) of   other cases have 30 <= age < 40
```

Table 1: A sample 2x2 Contingency Table

	C_{today}	C_{other}
$Age\_Decile = 3$	48	45
$Age\_Decile \neq 3$	86	220

Searching for the best score…

- Try ICD9 = x for each value of x
- Try Gender=M, Gender=F
- Try CoarseRegion=NE, =NW, SE, SW..
- Try FineRegion=AA,AB,AC, … DD (4x4 Grid)
- Try Hospital=x, TimeofDay=x, Prodrome=X, …
- [In future… features of census blocks]

Overfitting Alert!

Example

```
Sat 12-23-2001 (daynum 36882, dayindex 239)
FISHER_PVALUE = 0.000051 RANDOMIZATION_PVALUE = 0.031
35.8% ( 48/134) of today's cases have 30 <= age < 40
17.0% ( 45/265) of   other cases have 30 <= age < 40
```

Table 1: A sample 2x2 Contingency Table

	C_{today}	C_{other}
$Age\_Decile = 3$	48	45
$Age\_Decile \neq 3$	86	220

Multiple component rules

- We would like to be able to find rules like:

There are a surprisingly large number of children with respiratory problems today

or

There are too many skin complaints among people from the affluent neighborhoods

- These are things that would be missed by casual screening
- BUT
  - The danger of overfitting could be much worse
  - It's very computationally demanding
  - How can we be sure the entire rule is meaningful?

Checking two component rules

Table 2: 2x2 Contingency Table 1 for a two component rule

Records from Today matching C_0 and C_1	Records from Other matching C_0 and C_1
Records from Today matching C_1 and differing on C_0	Records from Other matching C_1 and differing on C_0

Table 3: 2x2 Contingency Table 2 for a two component rule

Records from Today matching C_0 and C_1	Records from Other matching C_0 and C_1
Records from Today matching C_0 and differing on C_1	Records from Other matching C_0 and differing on C_1

- Must pass both tests to be allowed to live.

WSARE v2.0

- Inputs:

1. Date/time-indexed biosurveillance-relevant data stream

2. Time Window Length

3. Which attributes to use?

- Outputs:

1. Here are the records that most surprise me

2. Here's why

3. And here's how seriously you should take it

Primary Key	Date	Time	Hospital	ICD 9	Prodrome	Gender	Age	Home			Work			Recent Flu Levels	Recent Weather	(Many more…)
								Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale			
h6r32	6/2/2	14:12	Down-town	781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	…
t3q15	6/2/2	14:15	River-side	717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	…
t5hh5	6/2/2	14:15	Smith-field	622	Respiratory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	…
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

WSARE v2.0

- Inputs:
  - 1. Date/time-indexed biosurveillance-relevant data stream
  - 2. Time Window Length
  - 3. Which attributes to use?
- Outputs:
  - 1. Here are the records that most surprise me
  - 2. Here's why
  - 3. And here's how seriously you should take it

Primary Key	Date	Time	Hospit	ICD	P		Ag	Home			Work				Recent Weath	(Many more …)
								Large Scale	Medium Scale	Fine Scal	Large Scal	Medium Scale	Fine Scal			
h6r32								NE	15 7							
t3q15	6/2/2	14:15	River-side	717	Respiratory	M	60s	NE	15 2							
t5hh5	6/2/2	14:15	Smith-field	622	Respiratory	F	80s	SE	15 0							
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Normally, 8% of cases in the East are over-50s with respiratory problems.

But today it's been 15%

Don't be too impressed!

Taking into account all the patterns I've been searching over, there's a 20% chance I'd have found a rule this dramatic just by chance

WSARE on recent Utah Data

Saturday June 1st in Utah:

The most surprising thing about recent records is:

Normally:

- 0.8% of records (50/6205) have time before 2pm and prodrome = Hemorrhagic

But recently:

- 2.1% of records (19/907) have time before 2pm and prodrome = Hemorrhagic

Pvalue = 0.0484042

Which means that in a world where nothing changes we'd

expect to have a result this significant about once

every 20 times we ran the program

Results on Emergency Dept Data

Rule 1: Tue 05-16-2000 (daynum 36661, dayindex 18)
SCORE = -0.00000000 PVALUE = 0.00000000
32.84% (44/134) of today's cases have Time Of Day4 after 6:00 pm
90.00% (27/ 30) of other cases have Time Of Day4 after 6:00 pm

Rule 2: Fri 06-30-2000 (daynum 36706, dayindex 63)
SCORE = -0.00000000 PVALUE = 0.00000000
19.40% (26/134) of today's cases have Place2 = NE and Lat4 = d
5.71% (16/280) of other cases have Place2 = NE and Lat4 = d

Rule 3: Wed 09-06-2000 (daynum 36774, dayindex 131)
SCORE = -0.00000000 PVALUE = 0.00000000
17.16% (23/134) of today's cases have Prodrome = Respiratory and age2 less than 40
4.53% (12/265) of other cases have Prodrome = Respiratory and age2 less than 40

Rule 4: Fri 12-01-2000 (daynum 36860, dayindex 217)
SCORE = -0.00000000 PVALUE = 0.00000000
22.88% (27/118) of today's cases have Time Of Day4 after 6:00 pm and Lat2 = s
8.10% (20/247) of other cases have Time Of Day4 after 6:00 pm and Lat2 = s

Rule 5: Sat 12-23-2000 (daynum 36882, dayindex 239)
SCORE = -0.00000000 PVALUE = 0.00000000
18.25% (25/137) of today's cases have ICD9 = shortness of breath and Time Of Day2 before 3:00 pm
5.12% (15/293) of other cases have ICD9 = shortness of breath and Time Of Day2 before 3:00 pm

Rule 6: Fri 09-14-2001 (daynum 37147, dayindex 504)
SCORE = -0.00000000 PVALUE = 0.00000000
66.67% (30/ 45) of today's cases have Time Of Day4 before 10:00 am
18.42% (42/228) of other cases have Time Of Day4 before 10:00 am

WSARE 3.0

- "Taking into account recent flu levels…"
- "Taking into account that today is a public holday…"
- "Taking into account that this is Spring…"
- "Taking into account recent heatwave…"
- "Taking into account that there's a known natural Food-borne outbreak in progress…"

Bonus: More efficient use of historical data

Analysis of variance

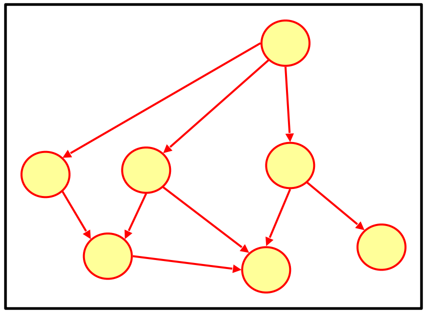
- Good news:

  If you're tracking a daily aggregate (e.g. number of flu cases in your ED, or Nyquil Sales)…then ANOVA can take care of many of these effects.

- But…

  What if you're tracking a whole joint distribution of transactional events?

Idea: Bayesian Networks



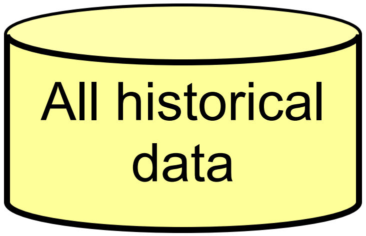
"Patients from West Park Hospital are less likely to be young"

"On Cold Tuesday Mornings the folks coming in from the North part of the city are more likely to have respiratory problems"

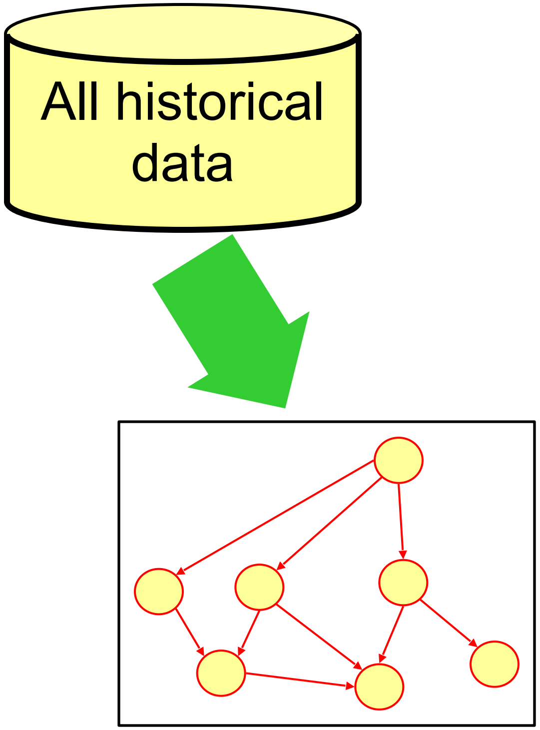
"The Viral prodrome is more likely to co-occur with a Rash prodrome than Botulinic"

"On the day after a major holiday, expect a boost in the morning followed by a lull in the afternoon"

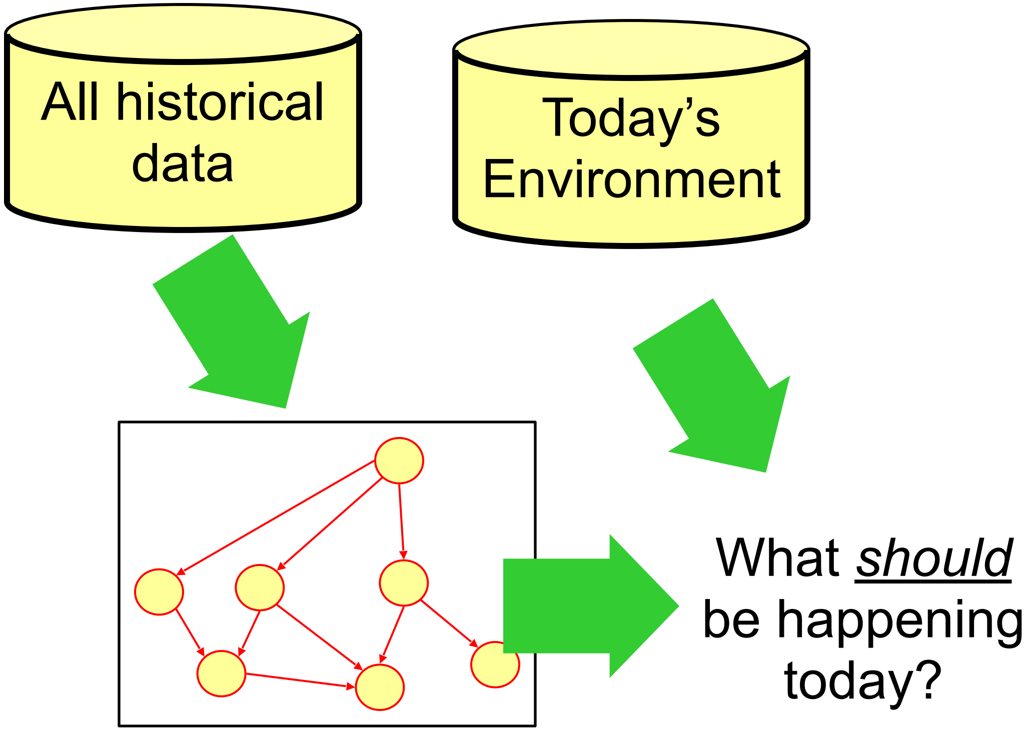
WSARE 3.0



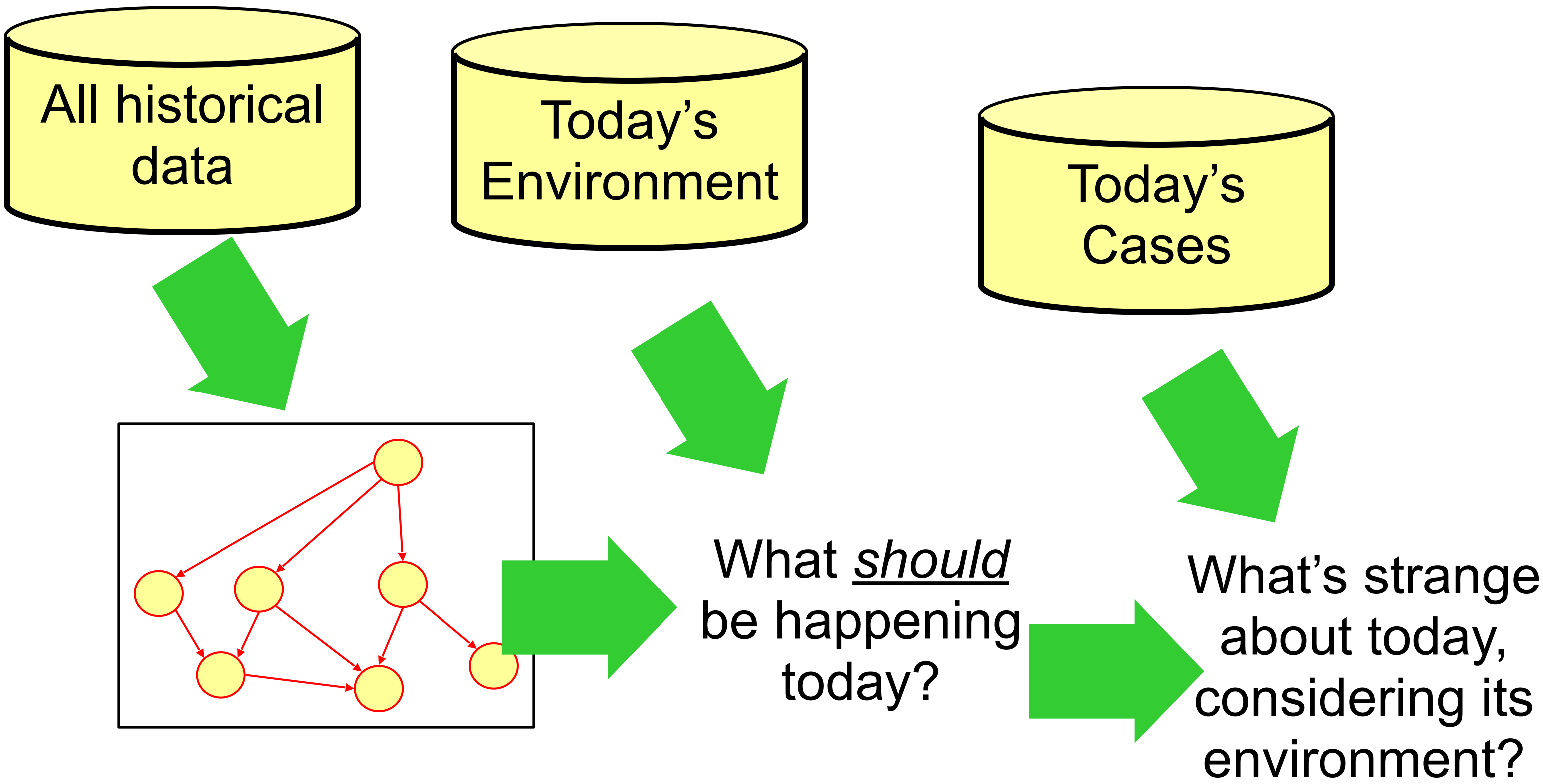
WSARE 3.0



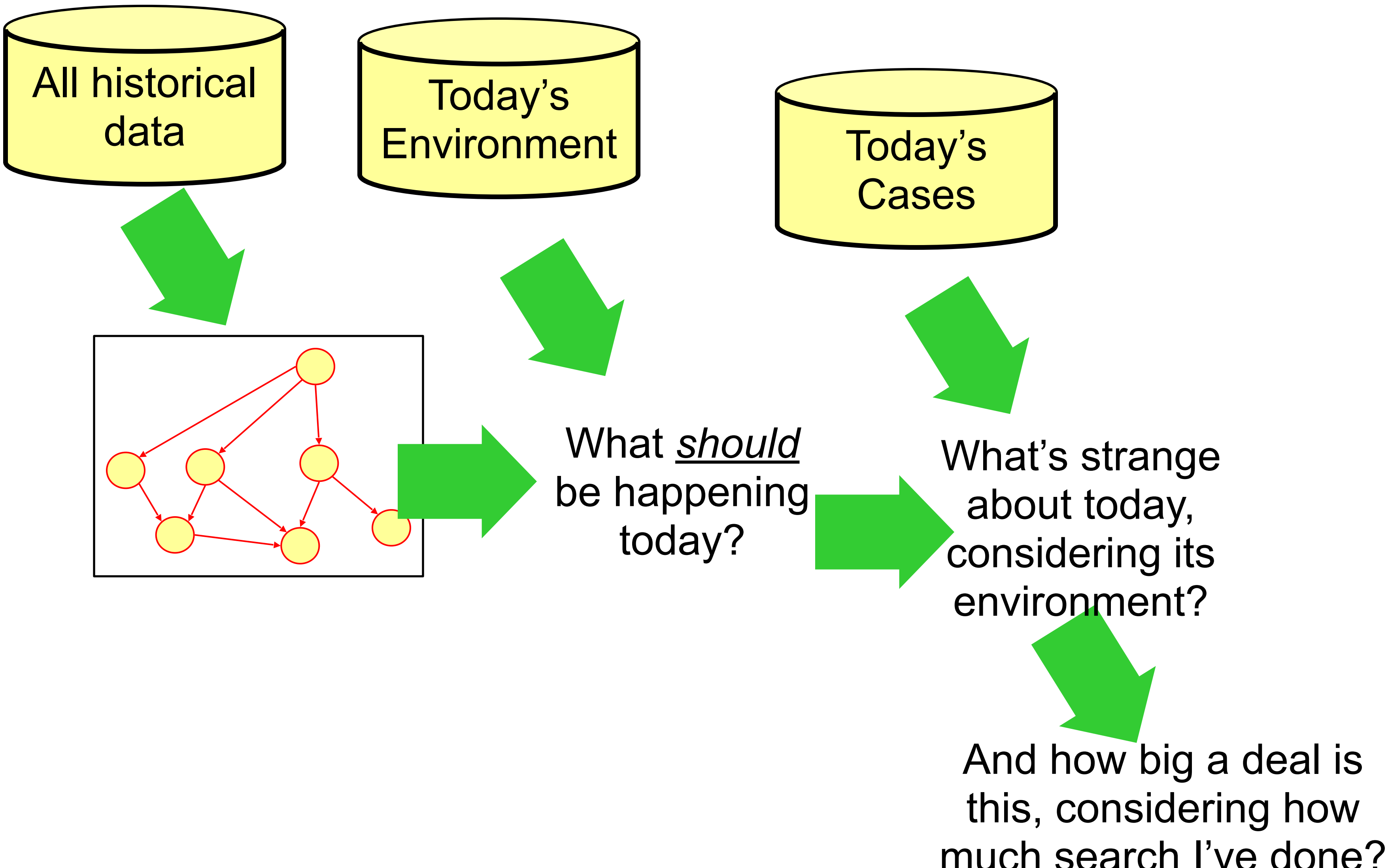
WSARE 3.0


All historical data
Today's Environment
What *should* be happening today?

WSARE 3.0

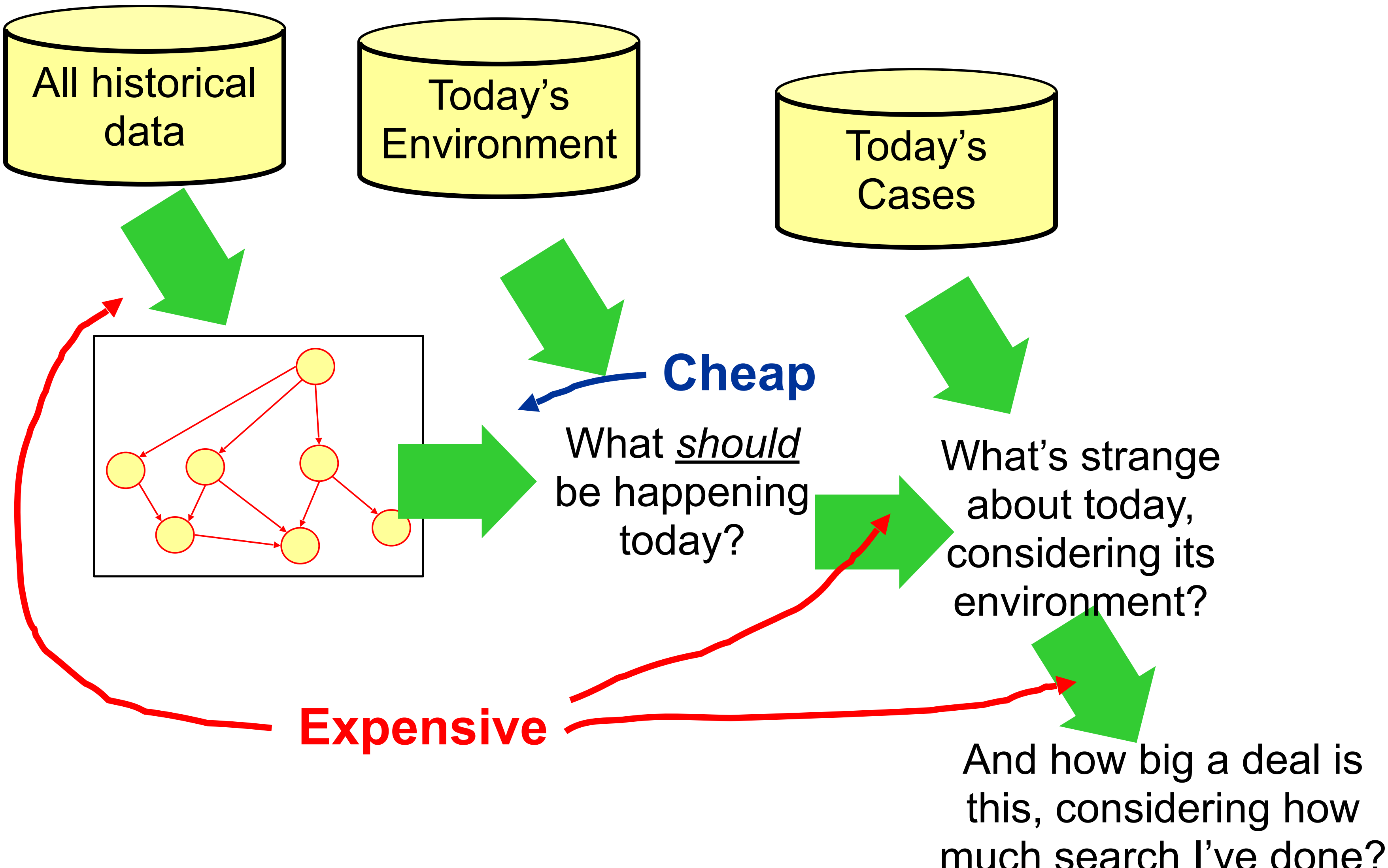

All historical data
Today's Environment
Today's Cases
What *should* be happening today?
What's strange about today, considering its environment?

WSARE 3.0

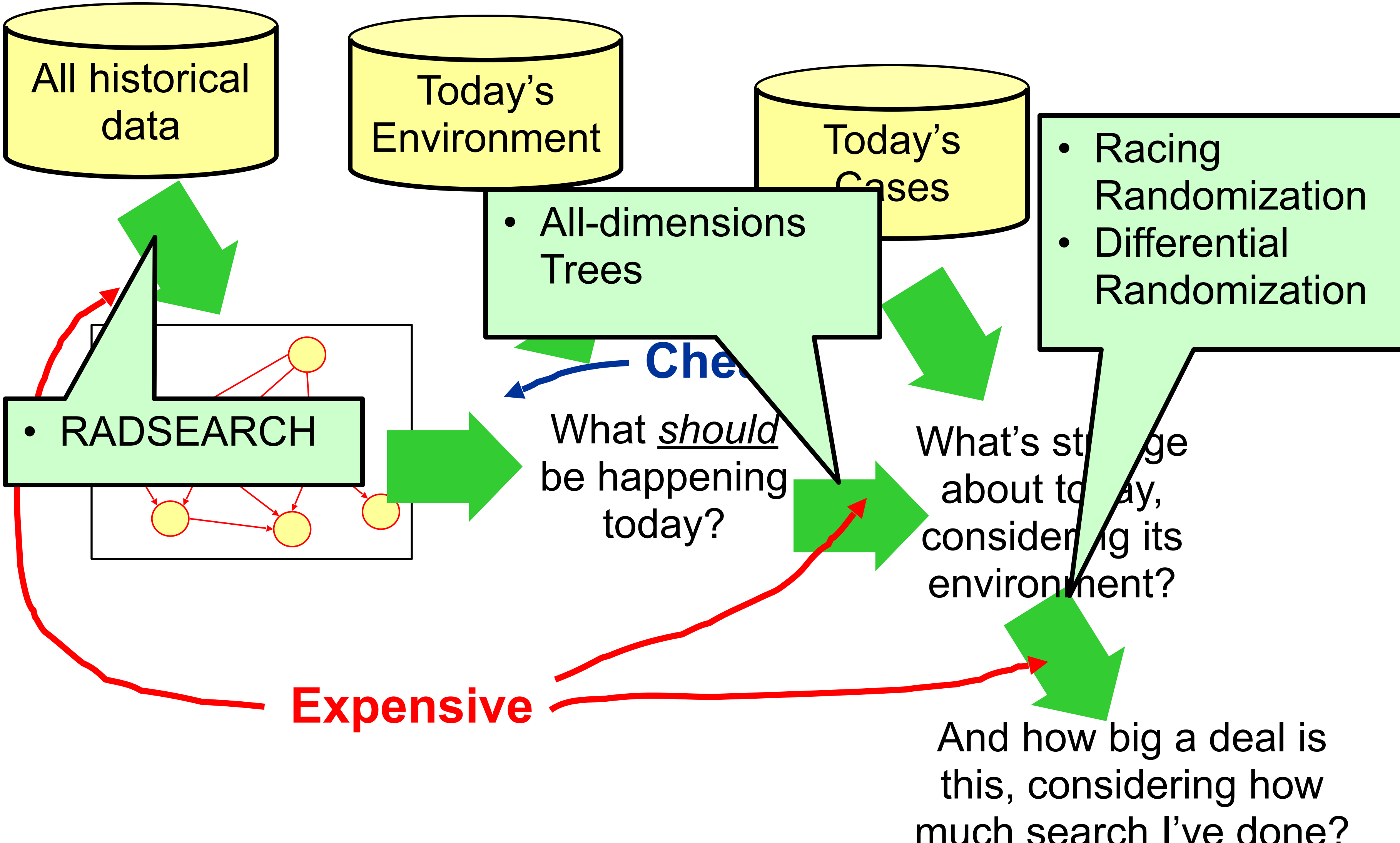


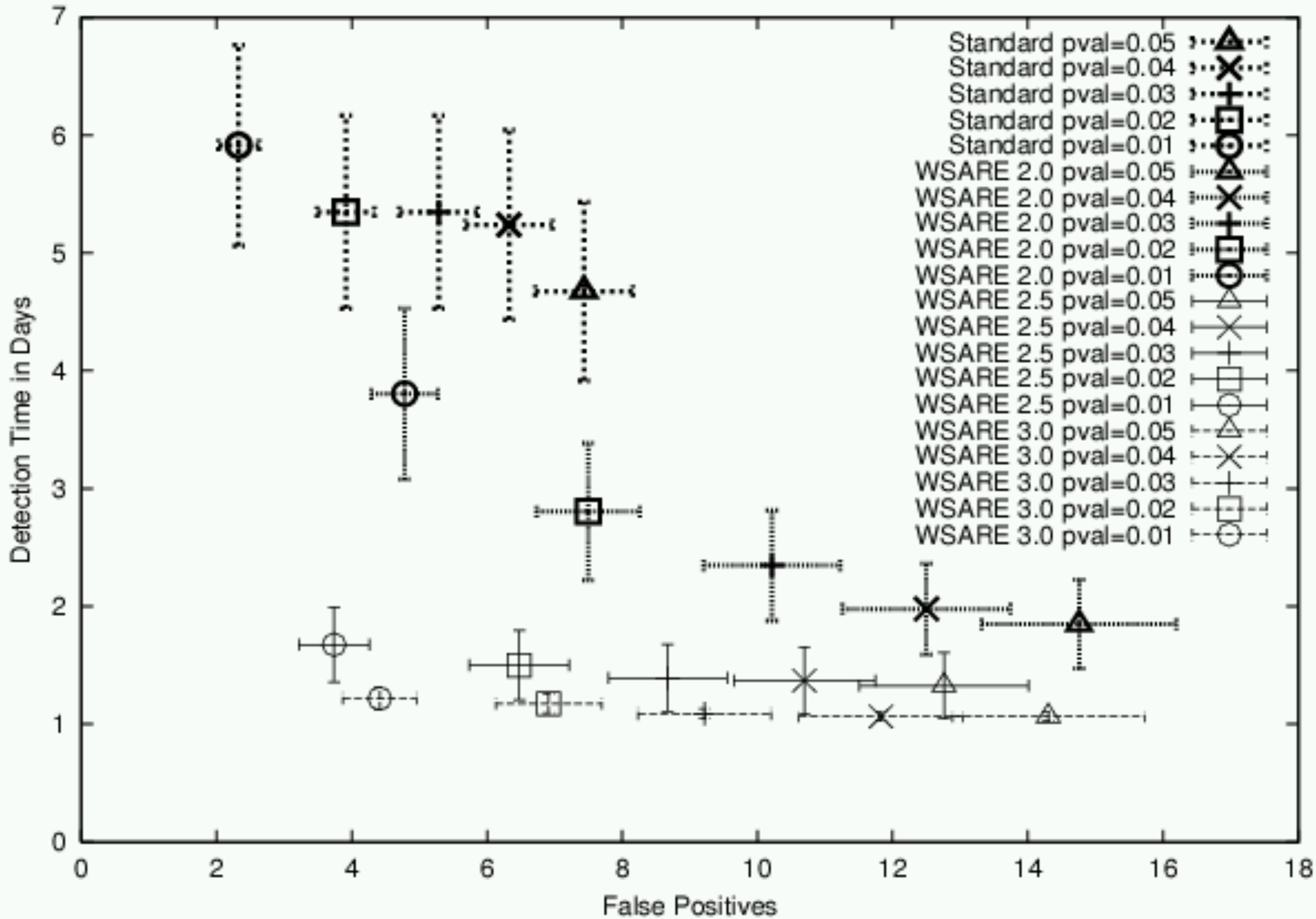
All historical data
Today's Environment
Today's Cases
What *should* be happening today?
What's strange about today, considering its environment?
And how big a deal is this, considering how much search I've done?

WSARE 3.0

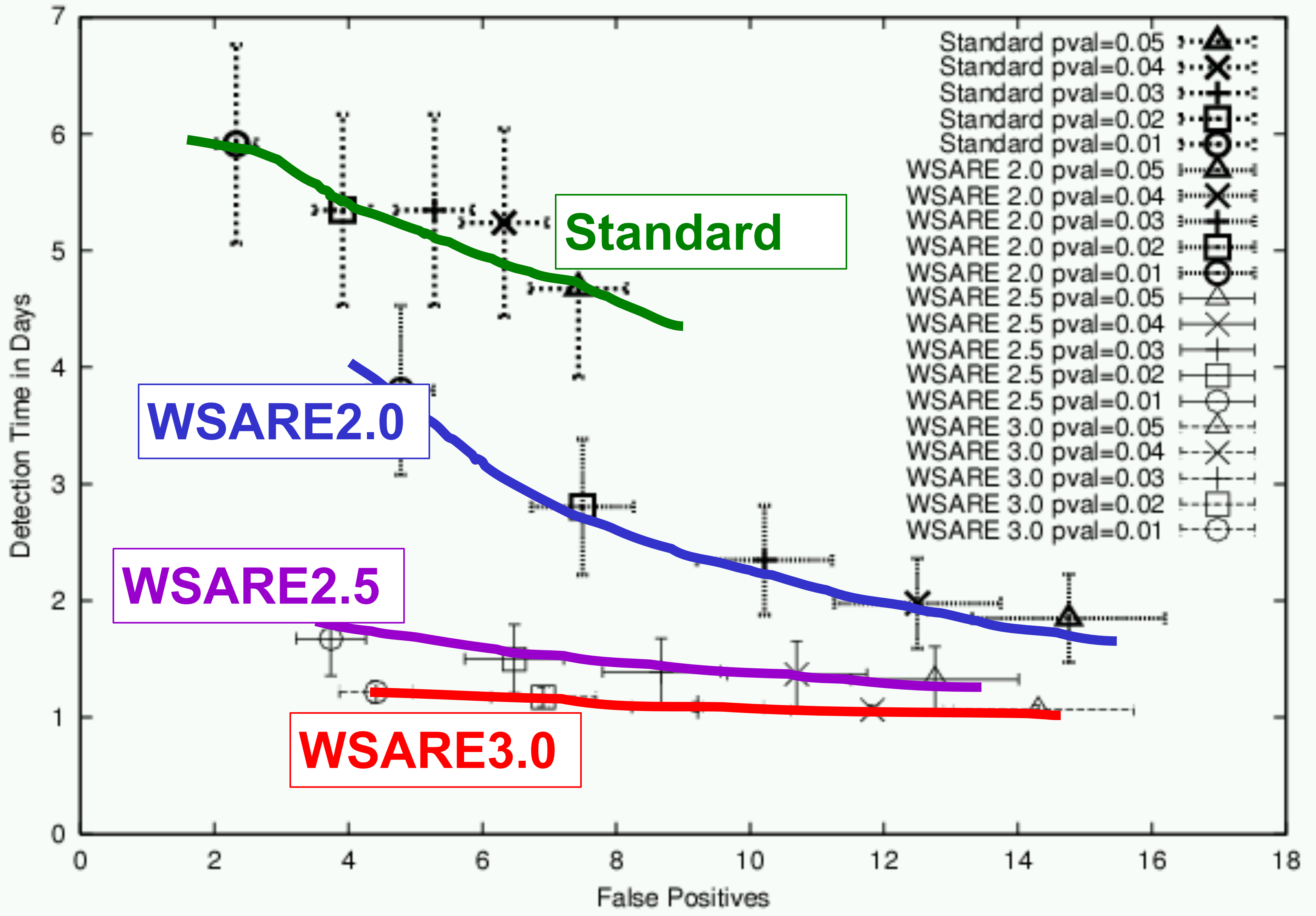

All historical data
Today's Environment
Today's Cases
Cheap
What *should* be happening today?
What's strange about today, considering its environment?
Expensive
And how big a deal is this, considering how much search I've done?

WSARE 3.0


All historical data
Today's Environment
Today's Cases
• All-dimensions Trees
• Racing Randomization
• Differential Randomization
• RADSEARCH
Che
What *should* be happening today?
What's st ge about to ay, consider g its environ ent?
And how big a deal is this, considering how much search I've done?
Expensive



Results on Simulation



Results on Simulation

Conclusion

- One approach to biosurveillance: one algorithm monitoring millions of signals derived from multivariate data
  instead of
  Hundreds of univariate detectors

- Modeling historical data with Bayesian Networks to allow conditioning on unique features of today
- Computationally intense unless we're tricksy!

- Searching over thousands of contingency tables on a large database...
- ...only we have to do it 10,000 times on the replicas during randomization
- ...we also need to learn Bayes Nets from databases with millions of records...
- ...and keep relearning them as data arrives online...
- ...in the end we typically search about a billion alternative Bayes net structures for modeling 800,000 records in 10 minutes

- allow conditi unique features of today
- Computationally intense unless we’re tricksy!

Conclusion

- One approach to biosurveillance: one algorithm monitoring millions of signals derived from multivariate data

  instead of

  Hundreds of univariate detectors

- Modeling historical data with Bayesian Networks to allow conditioning on unique features of today
- Computationally intense unless we're tricksy!
- WSARE 2.0 Deployed during the past year
- WSARE 3.0 about to go online
- WSARE now being extended to additionally exploit over the counter medicine sales

Other New Algorithmic Developments

Specific Detectors

PANDA2: Patient-based Bayesian Network
[Cooper, Levander et. al]

BARD: Airborne Attack Detection
[Hogan, Cooper]

Possible Future Connection

General Detectors

WSARE meets Scan Statistics

Fast Scan Statistic
[Neill, Moore]

Fast Scan for Oriented Regions
[Neill, Moore et al.]

Historical Model Scan Statistic
[Hogan, Moore, Neill, Tsui, Wagner]

Bayesian Network Spatial Scan
[Neill, Moore, Schneider, Cooper Wagner, Wong]

Question: How do we use all this information?
How can we "plug in" new streams?
How can we exploit multiattribute form?

Other New Algorithmic Developments

Specific Detectors

PANDA2: Patient-based Bayesian Network
[Cooper, Levander et. al]

BARD: Airborne Attack Detection
[Hogan, Cooper]

General Detectors

WSARE meets Scan Statistics

Fast Scan for Oriented Regions
[Neill, Moore et al.]

Historical Model Scan Statistic

[Neill, Moore, Schneider, Cooper Wagner, Wong]

Please contact Greg Cooper gfc@cbmi.upmc.edu for information

Please contact Bill Hogan wrh@cbmi.pitt.edu for information

ible Future

Question: How do we use all this information?
How can we "plug in" new streams?
How can we exploit multiattribute form?





For further info

- Papers on these and other anti-terror applications: www.cs.cmu.edu/~awm/antiterror
- Papers on scaling up many of these analysis methods: www.cs.cmu.edu/~awm/papers.html
- Software implementing the above: www.autonlab.org
- Copies of 18 lectures on 25 statistical data mining topics: www.cs.cmu.edu/~awm/781
- CD-ROM, powerpoint-synchronized video/audio recordings of the above lectures: awm@cs.cmu.edu

Information Gain, Decision Trees

Probabilistic Reasoning, Bayes Classifiers, Density Estimation

Probability Densities in Data Mining

Gaussians in Data Mining

Maximum Likelihood Estimation

Gaussian Bayes Classifiers

Regression, Neural Nets

Overfitting: detection and avoidance

The many approaches to cross-validation

Locally Weighted Learning

Bayes Net, Bayes Net Structure Learning, Anomaly Detection

Andrew's Top 8 Favorite Regression Algorithms (Regression Trees, Cascade Correlation, Group Method Data Handling (GMDH), Multivariate Adaptive Regression Splines (MARS), Multilinear Interpolation, Radial Basis Functions, Robust Regression, Cascade Correlation + Projection Pursuit

Clustering, Mixture Models, Model Selection

K-means clustering and hierarchical clustering

Vapnik-Chervonenkis (VC) Dimensionality and Structural Risk Minimization

PAC Learning

Support Vector Machines

Time Series Analysis with Hidden Markov Models

References

1. WSARE 3.0 : Bayesian Network based Anomaly Pattern Detection

   Wong, Moore, Cooper and Wagner [ICML/KDD 2003]

2. Fast Grid Based Computation of Spatial Scan Statistics

   Neill and Moore [NIPS 2003]

These and other Biosurveillance algorithms papers and free software available from

http://www.autonlab.org/

See also: http://www.health.pitt.edu/rods