Outlier and Anomaly Detection
Case Study: Biosurveillance

Excerpts/Adaptations from:
Detection Algorithms for Biosurveillance: A tutorial

Andrew Moore, CMU tutorial:
http://www.autonlab.org/tutorials/biosurv.html

Also Excerpts/Adaptations from:
Russell and Norvig, *Intro. to AI* slides

Biosurveillance

- Attempt to quickly detect significant disease outbreaks
  - Allows for quick action to try to contain disease

- Approach:

  Collect data from many sources ...

  (such as emergency room visits, purchases of particular types of over-the-counter medicine, school absences, pharmacy sales, etc.)

  ...and recognize statistical anomalies

Biosurveillance is a difficult topic

- April 6, 2006: "House Government Reform Committee scolds Centers for Disease Control Prevention" for "technologically backward" system of biosurveillance [Government Health IT]

  http://govhealthit.com/article93982-04-06-06-Web

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

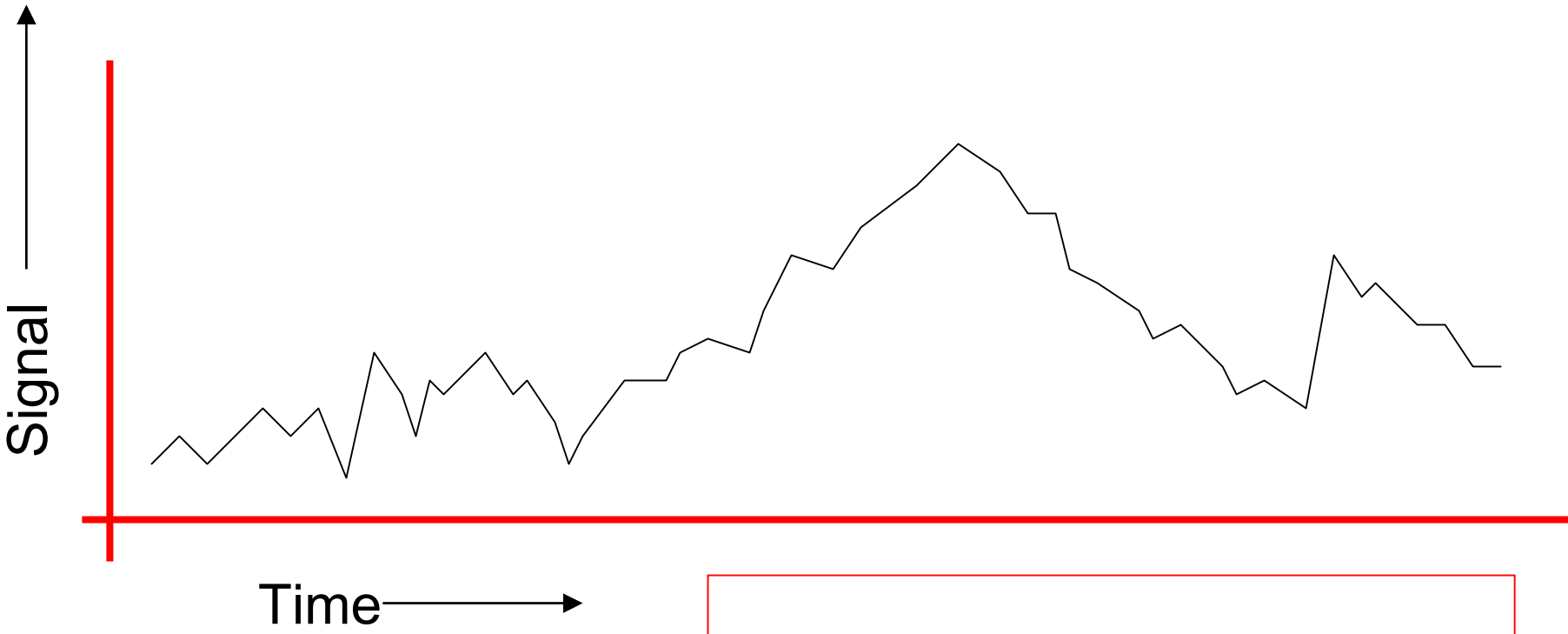
Univariate Anomaly Detection

Multivariate Anomaly Detection

Spatial Scan Statistics

WSARE

Univariate Time Series



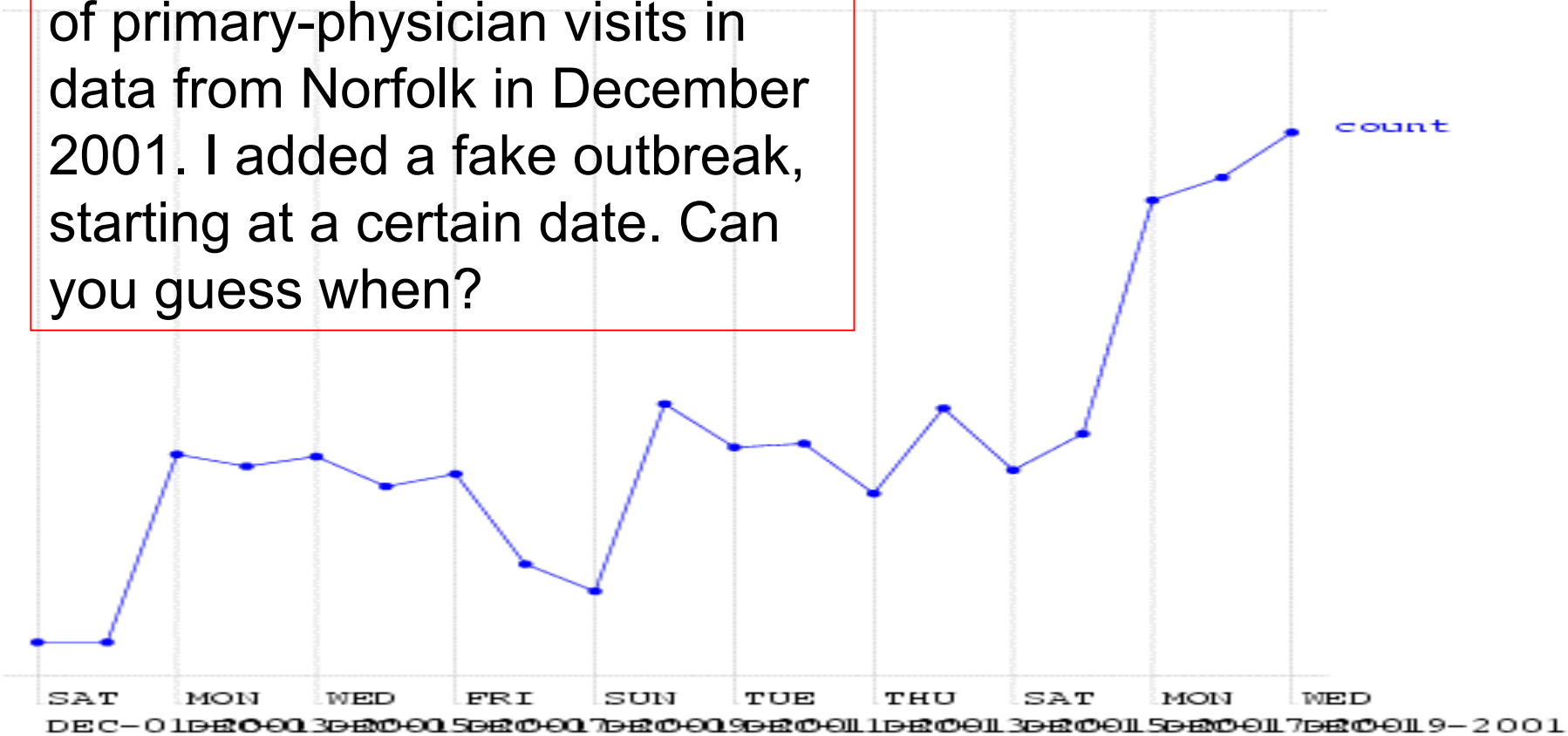
Example Signals:

- 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?

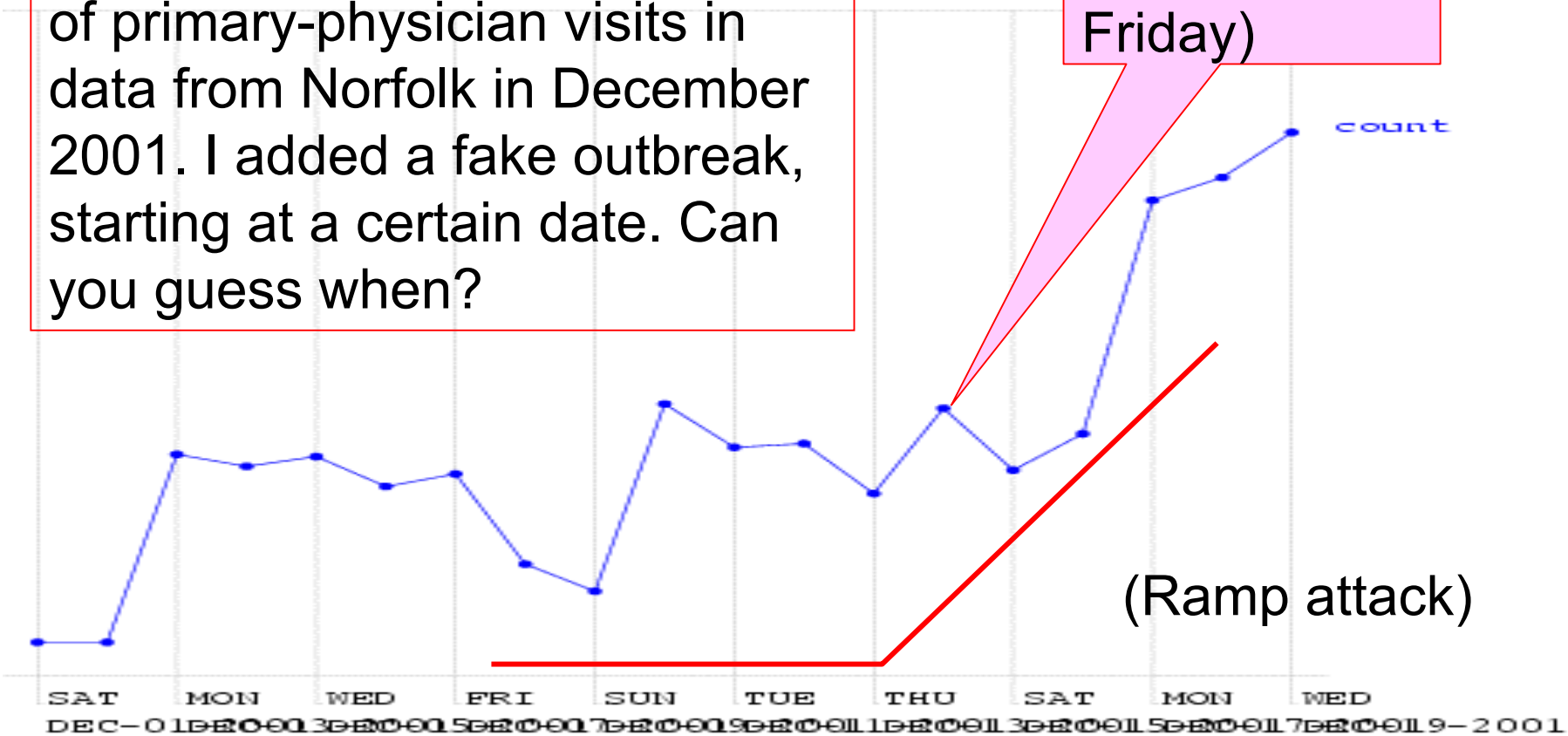
This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?



(When) is there an anomaly?

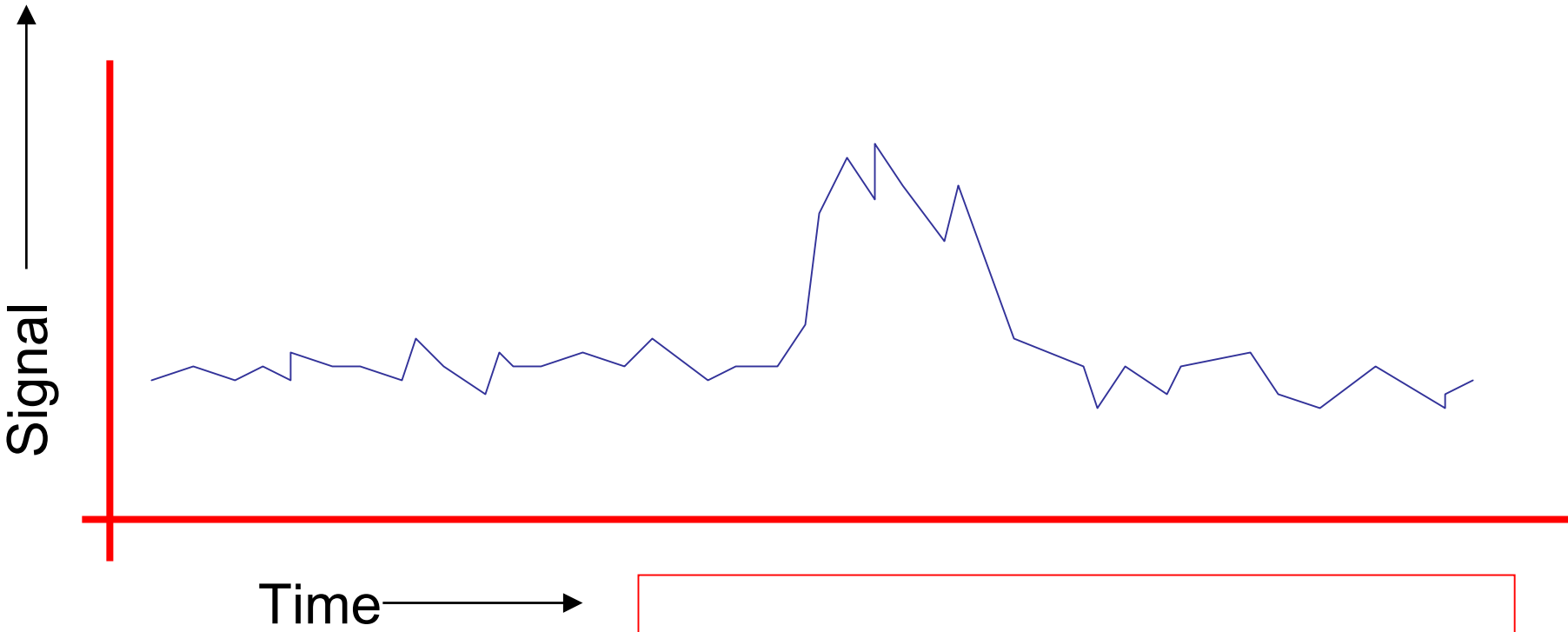
This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?

Here (much too high for a Friday)



(Ramp attack)

An easy case

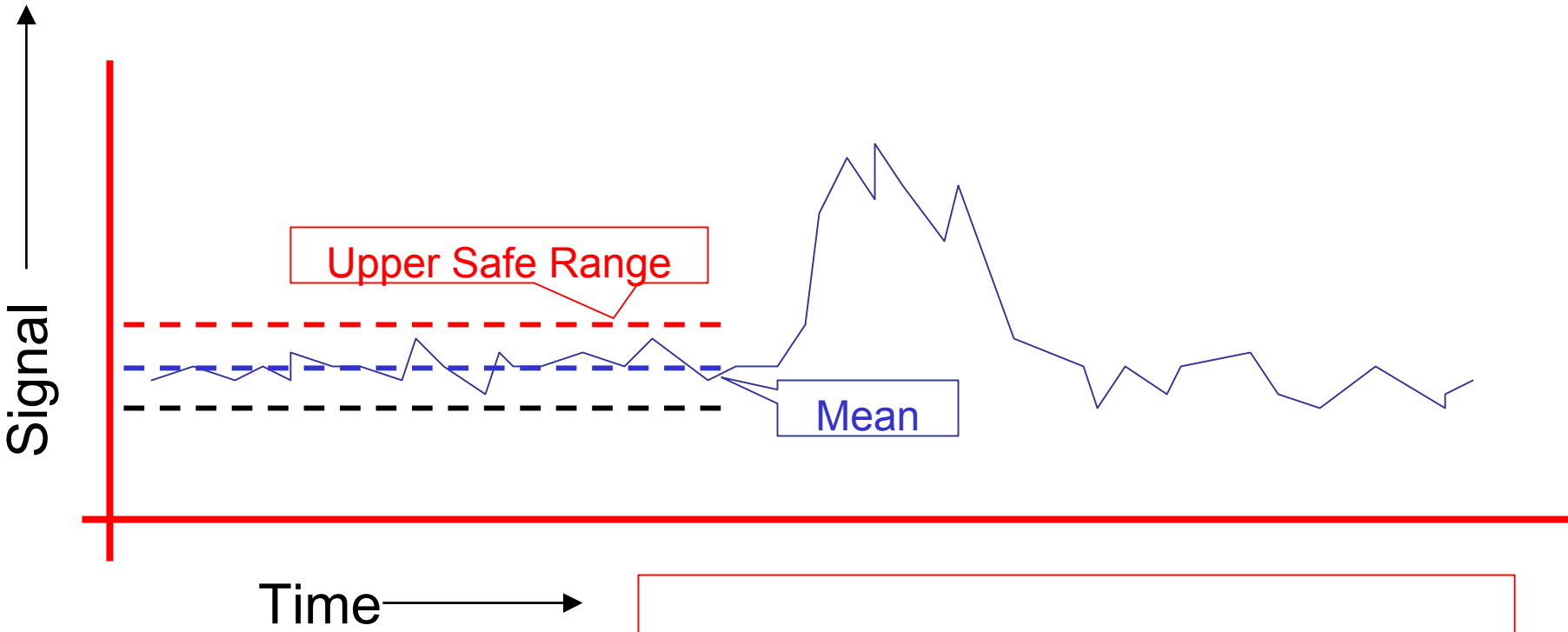


Dealt with by Statistical Quality Control

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

Signal an alarm if we go outside 3 sigmas

An easy case: Control Charts

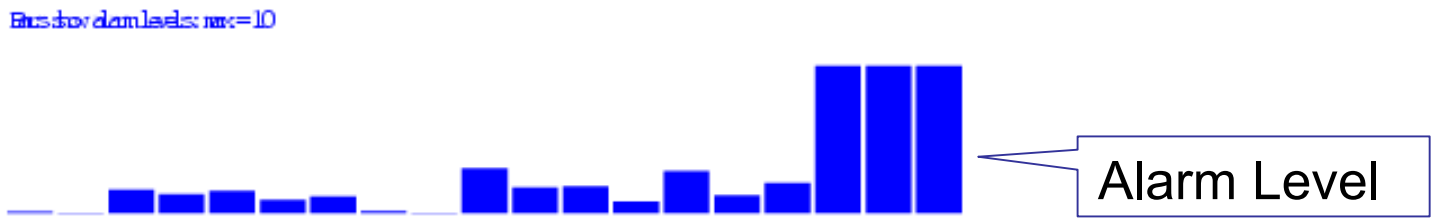


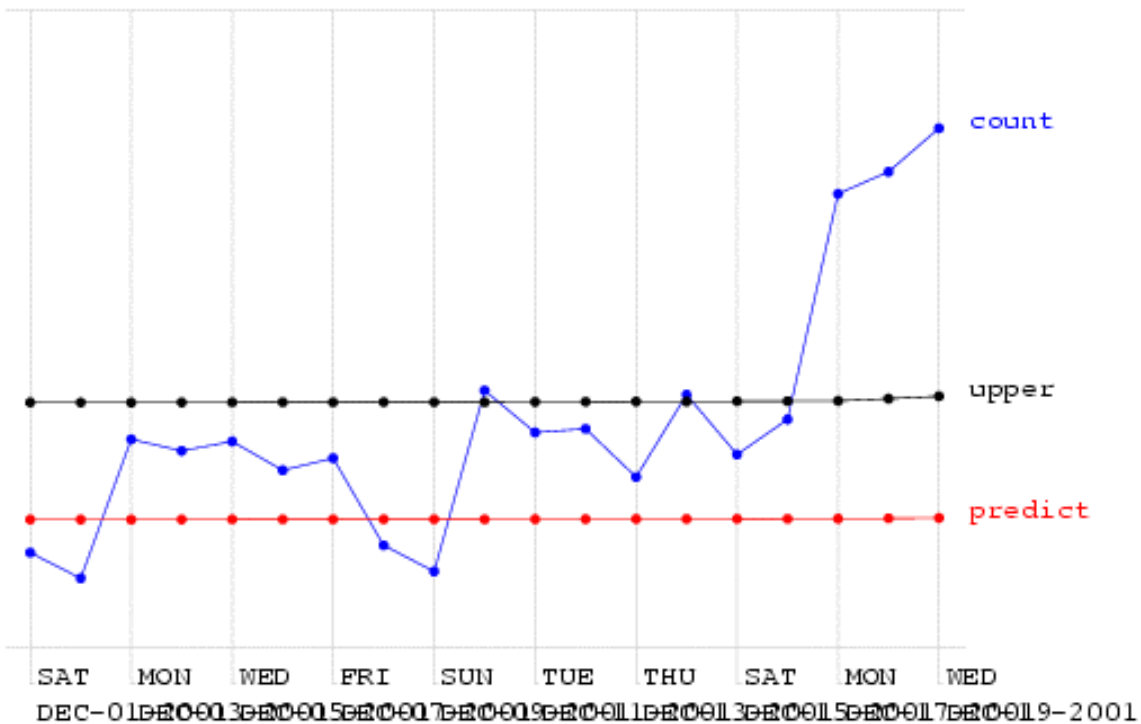
Dealt with by Statistical Quality Control

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

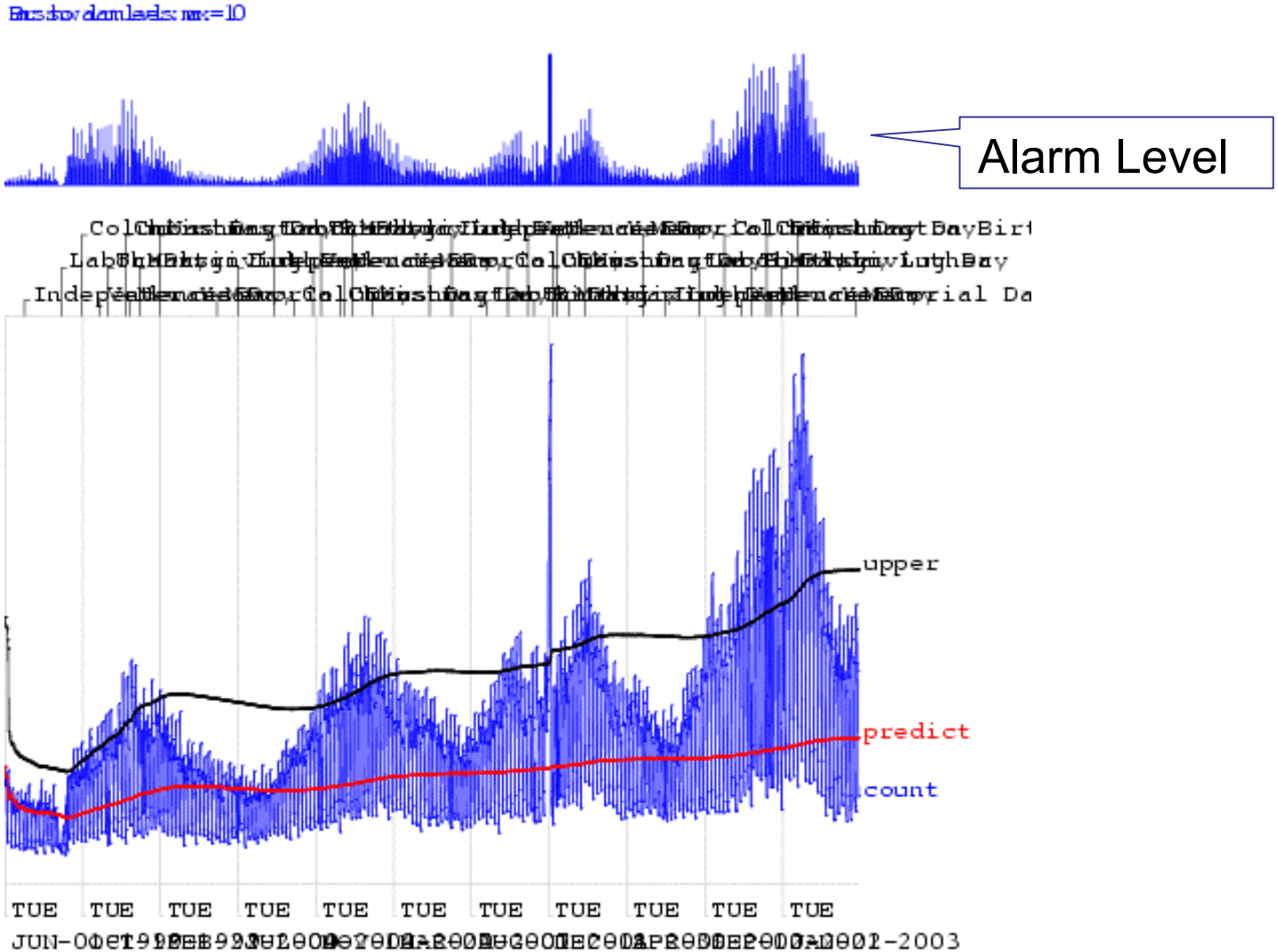
Signal an alarm if we go outside 3 sigmas

Control Charts on the Norfolk Data



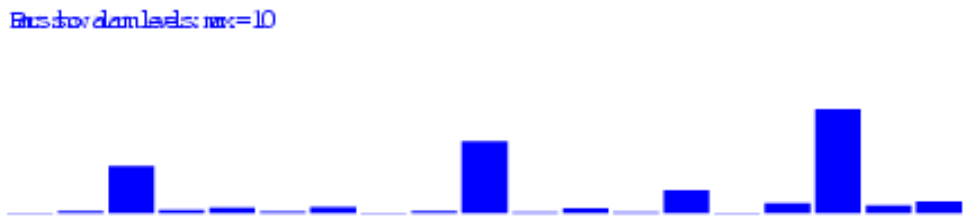


Control Charts on the Norfolk Data

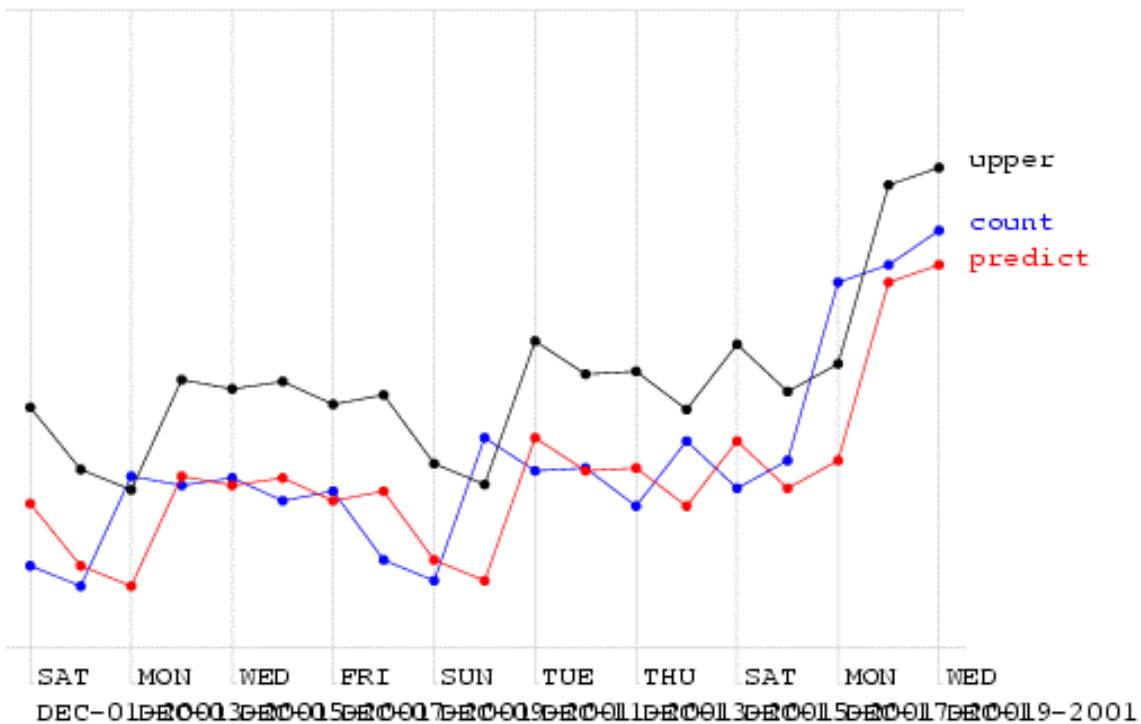


Looking at changes from yesterday

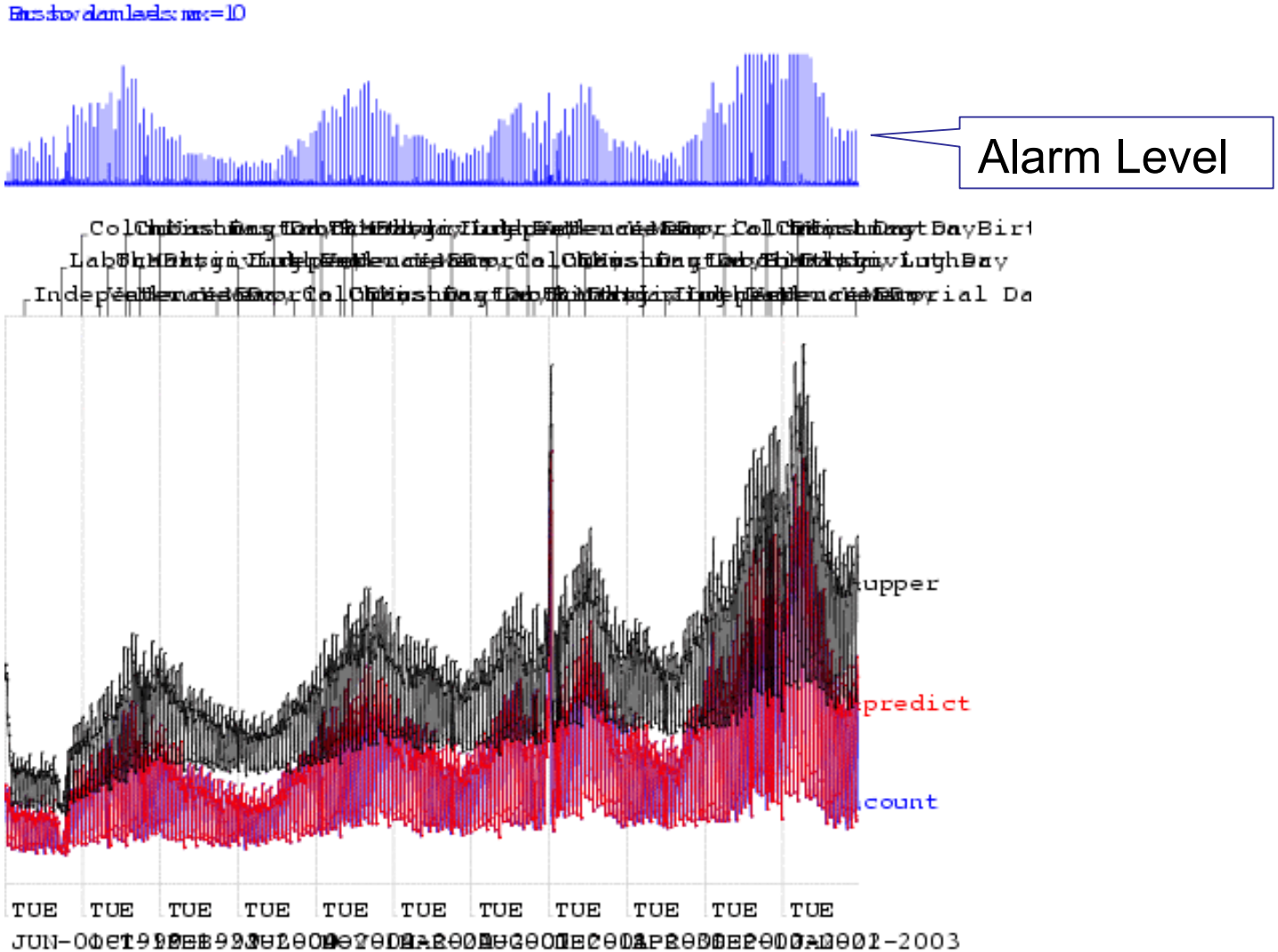
Looking at changes from yesterday



Alarm Level

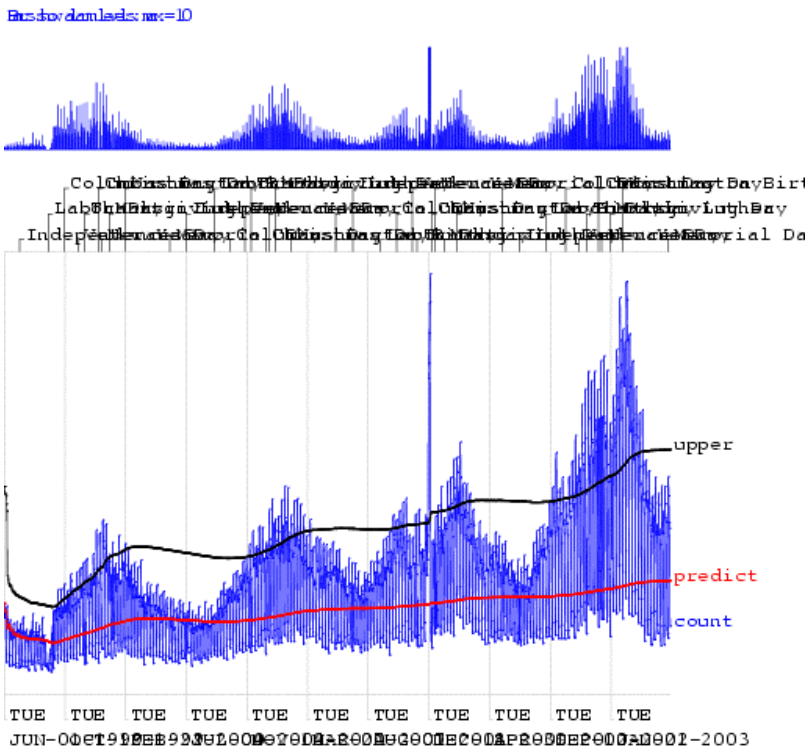


Looking at changes from yesterday

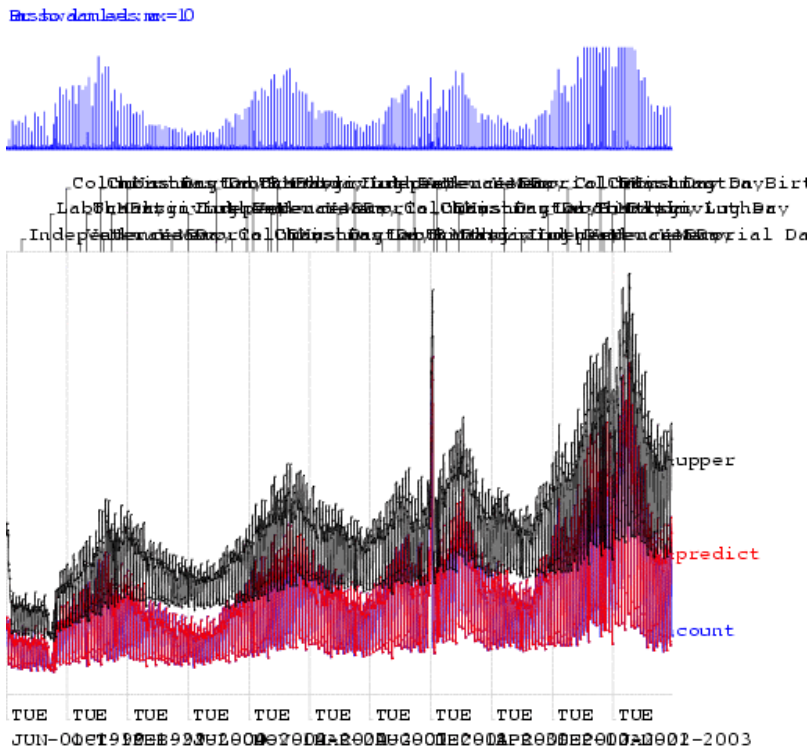


We need a happy medium:

Control Chart:
Too <u>in</u>sensitive to recent changes

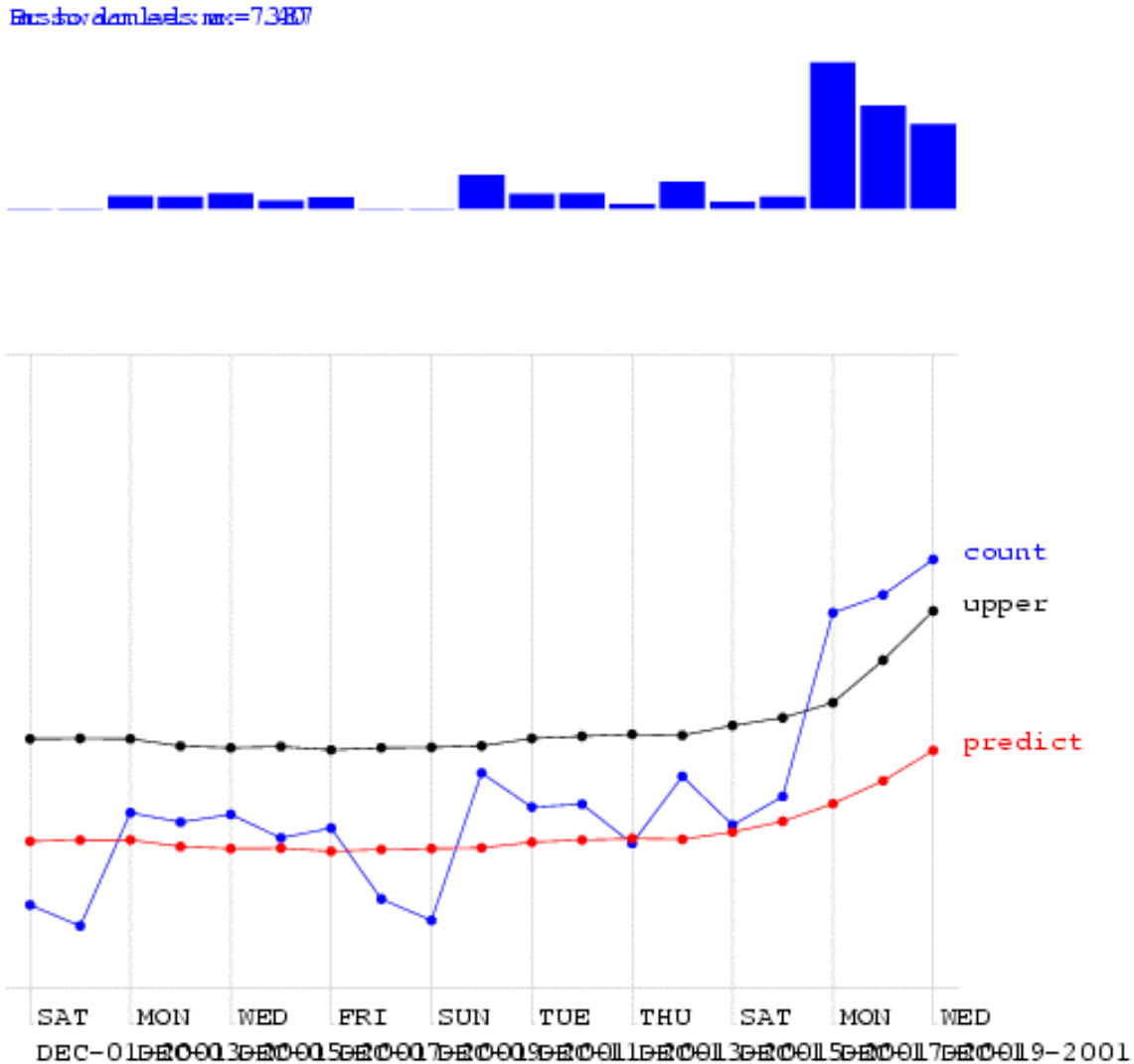


Change from yesterday:
Too sensitive to recent changes

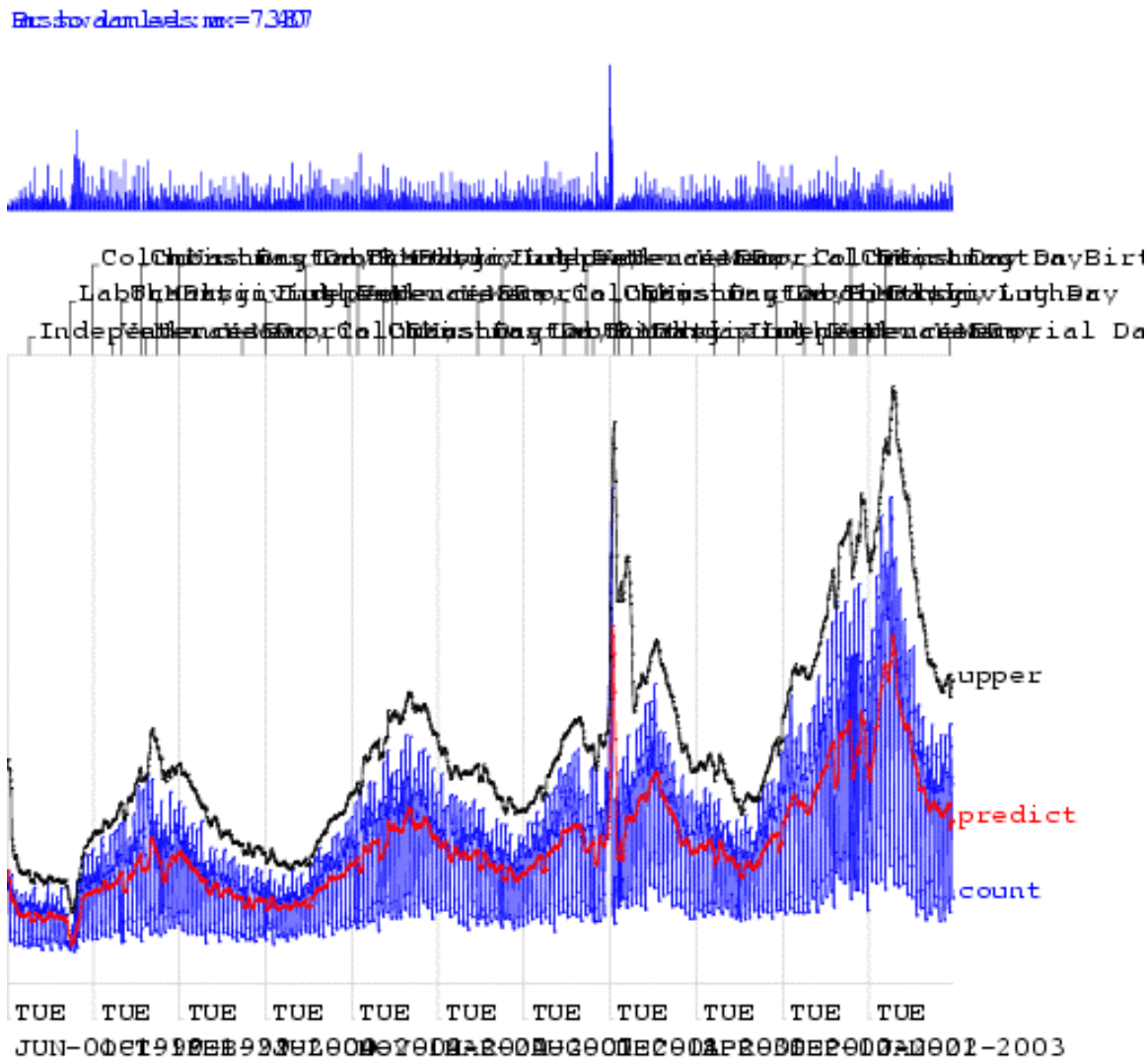


Moving Average

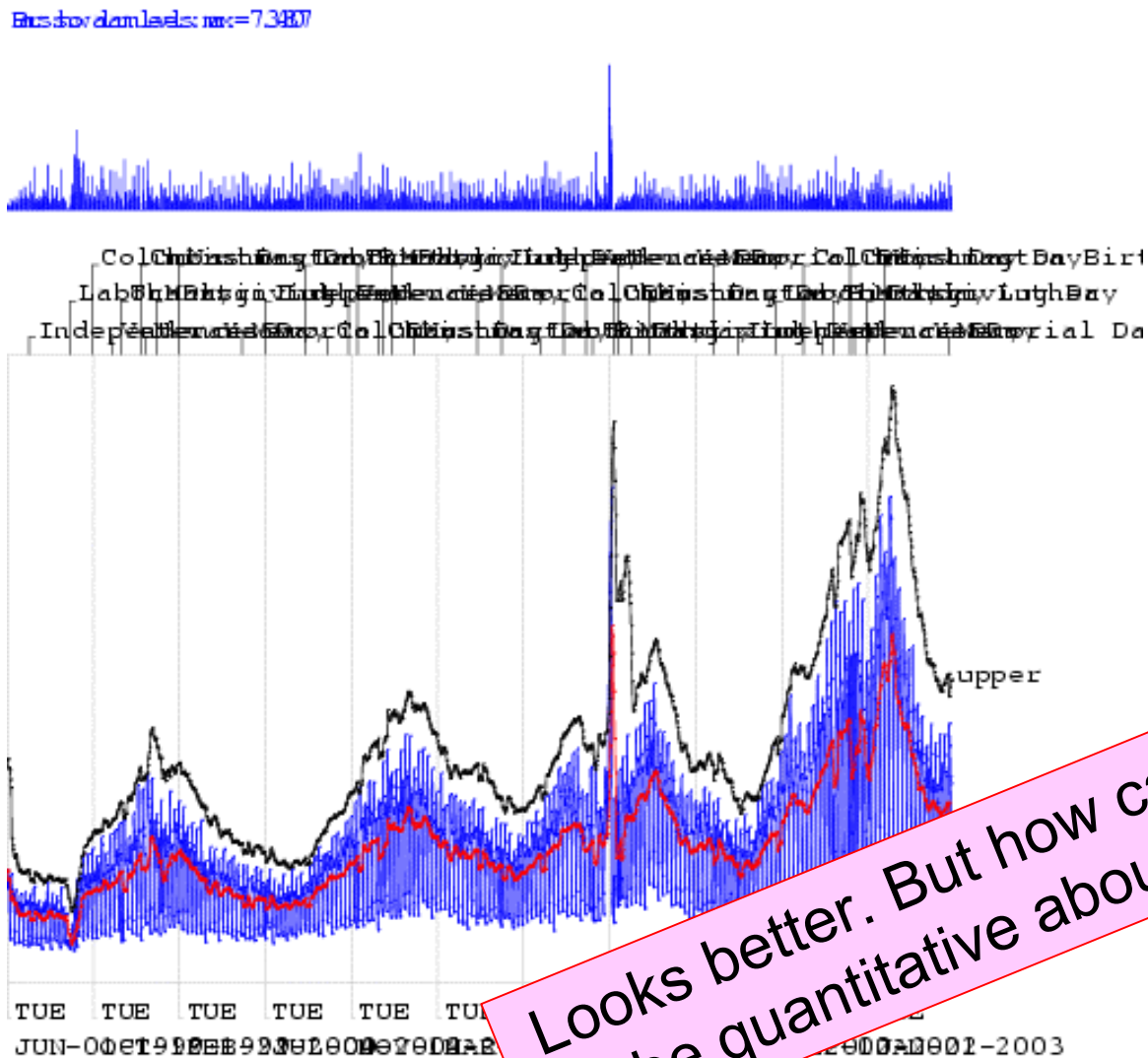
Moving Average



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

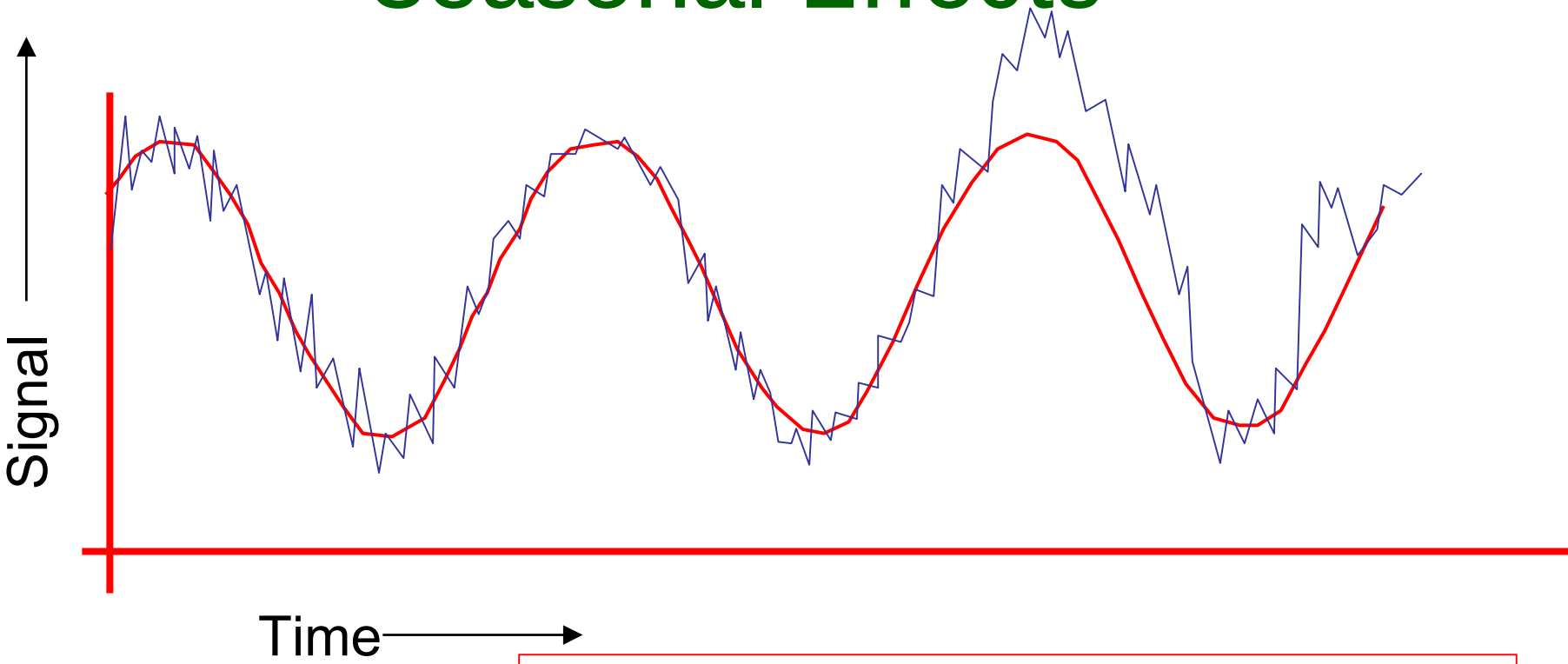
Algorithm Performance

	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 + $delta_{day}$

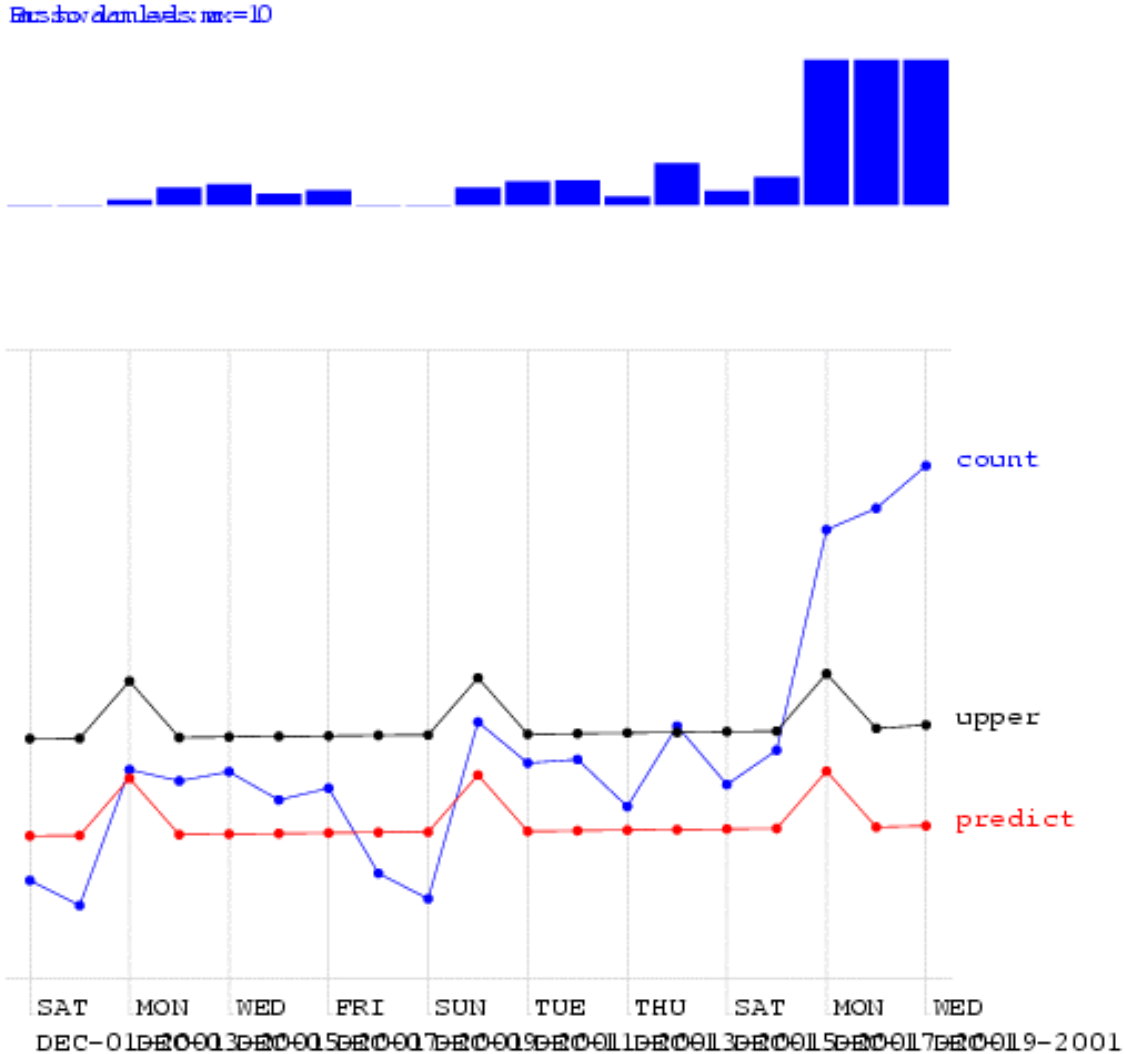
E.G:

$delta_{mon}$ = +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

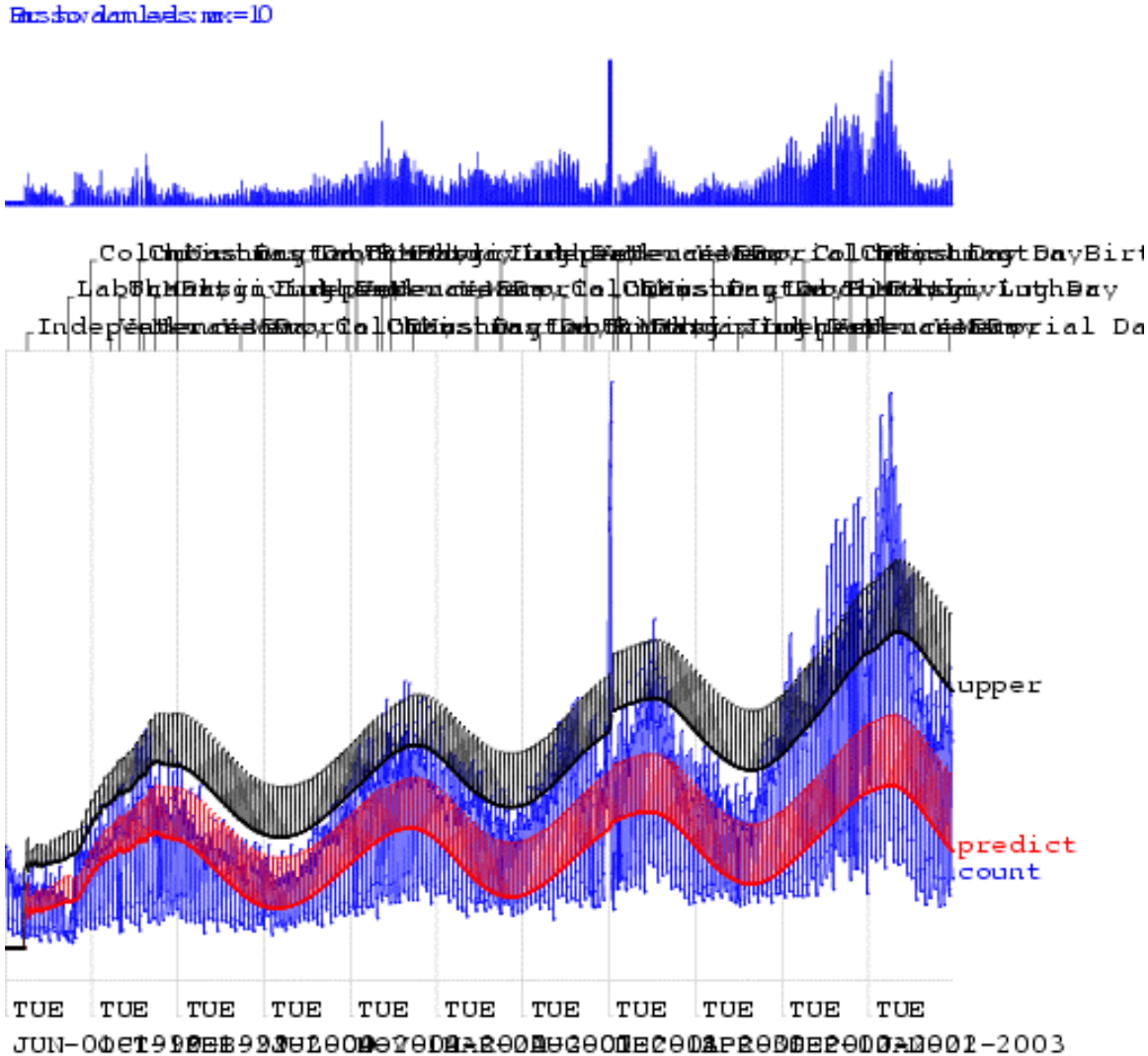
A simple form of ANOVA

(ANOVA: “Analysis of Variance” -- technique for testing differences among several means)

Regression using Hours-in-day & IsMonday



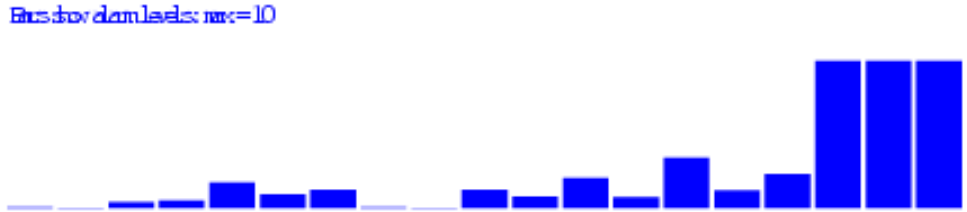
Regression using Hours-in-day & IsMonday

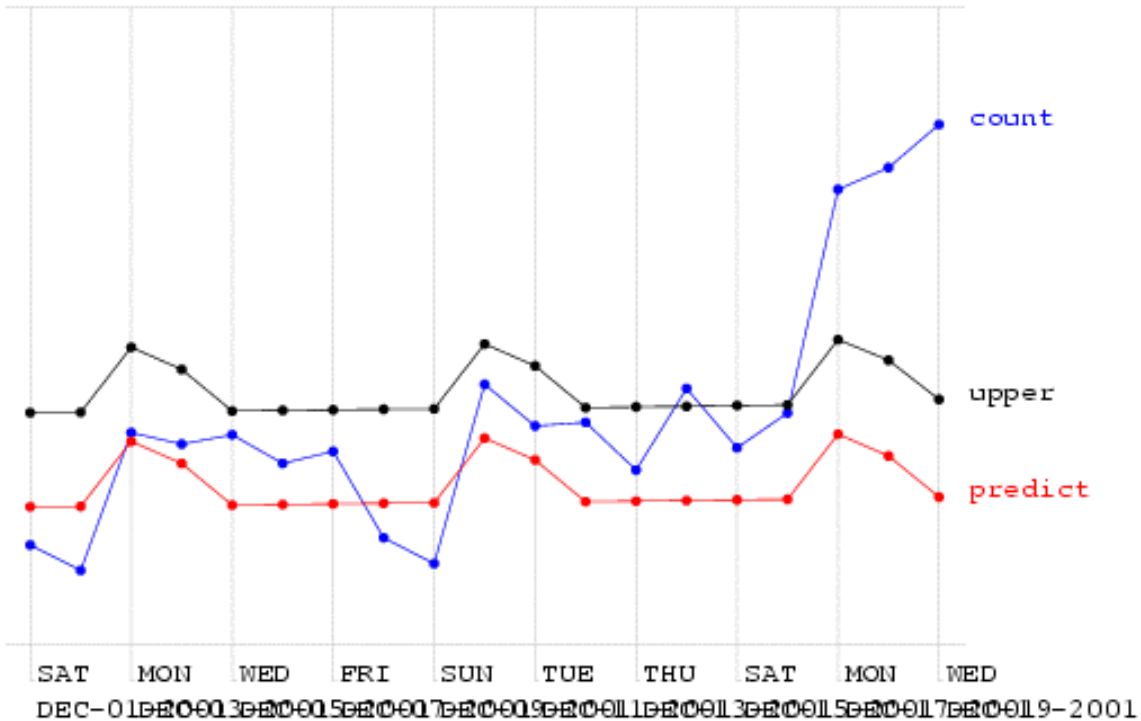


Algorithm Performance

	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

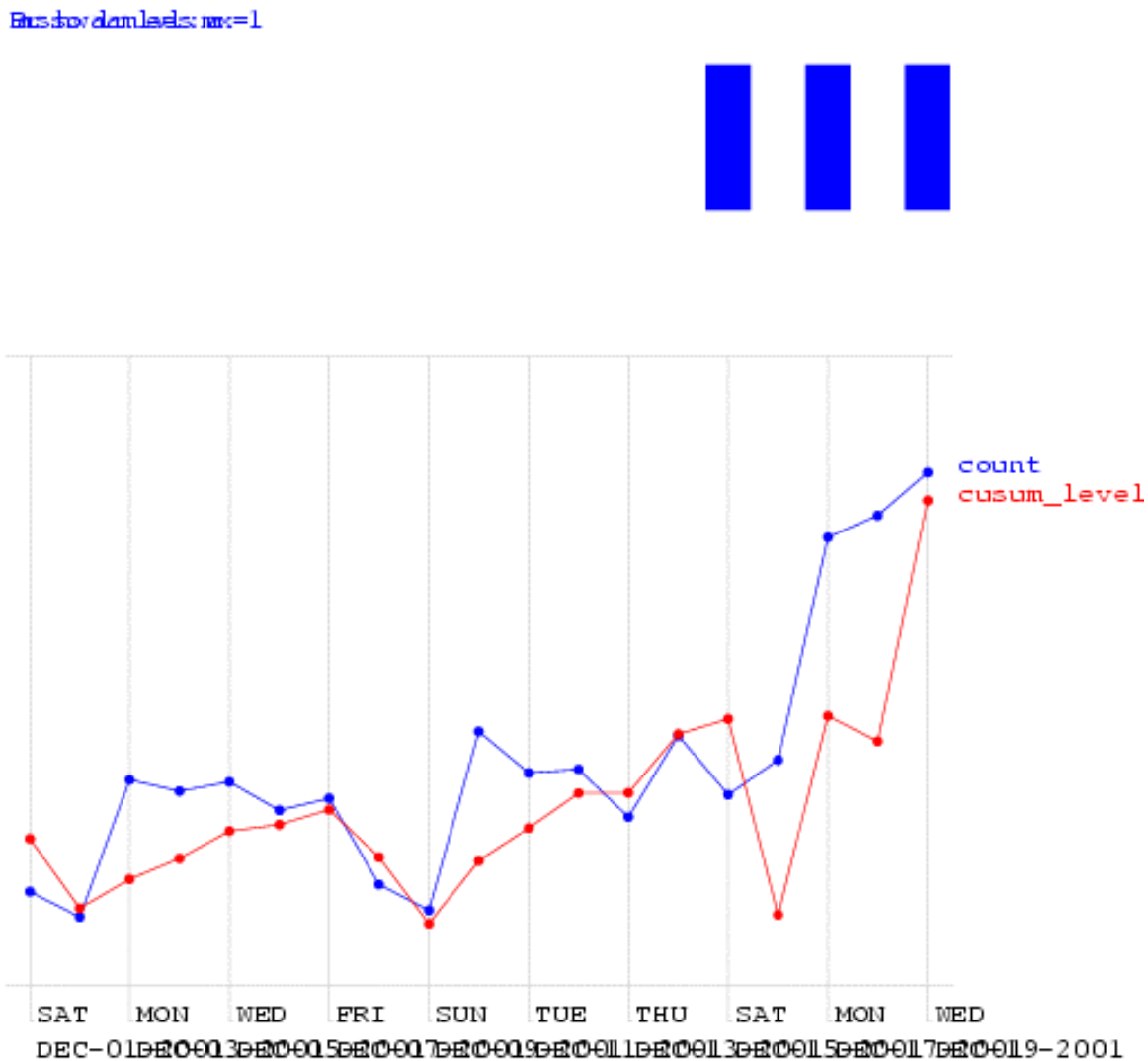
	Allowing one False Alarm per TWO weeks…		Allowing one False Alarm per SIX weeks…	
	Fraction of spikes detected	Days to detect a ramp attack	Fraction of spikes detected	Days to detect a ramp attack
standard control chart	0.39	3.47	0.22	4.13
using yesterday	0.14	3.83	0.1	4.7
Moving Average 3	0.36	3.45	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

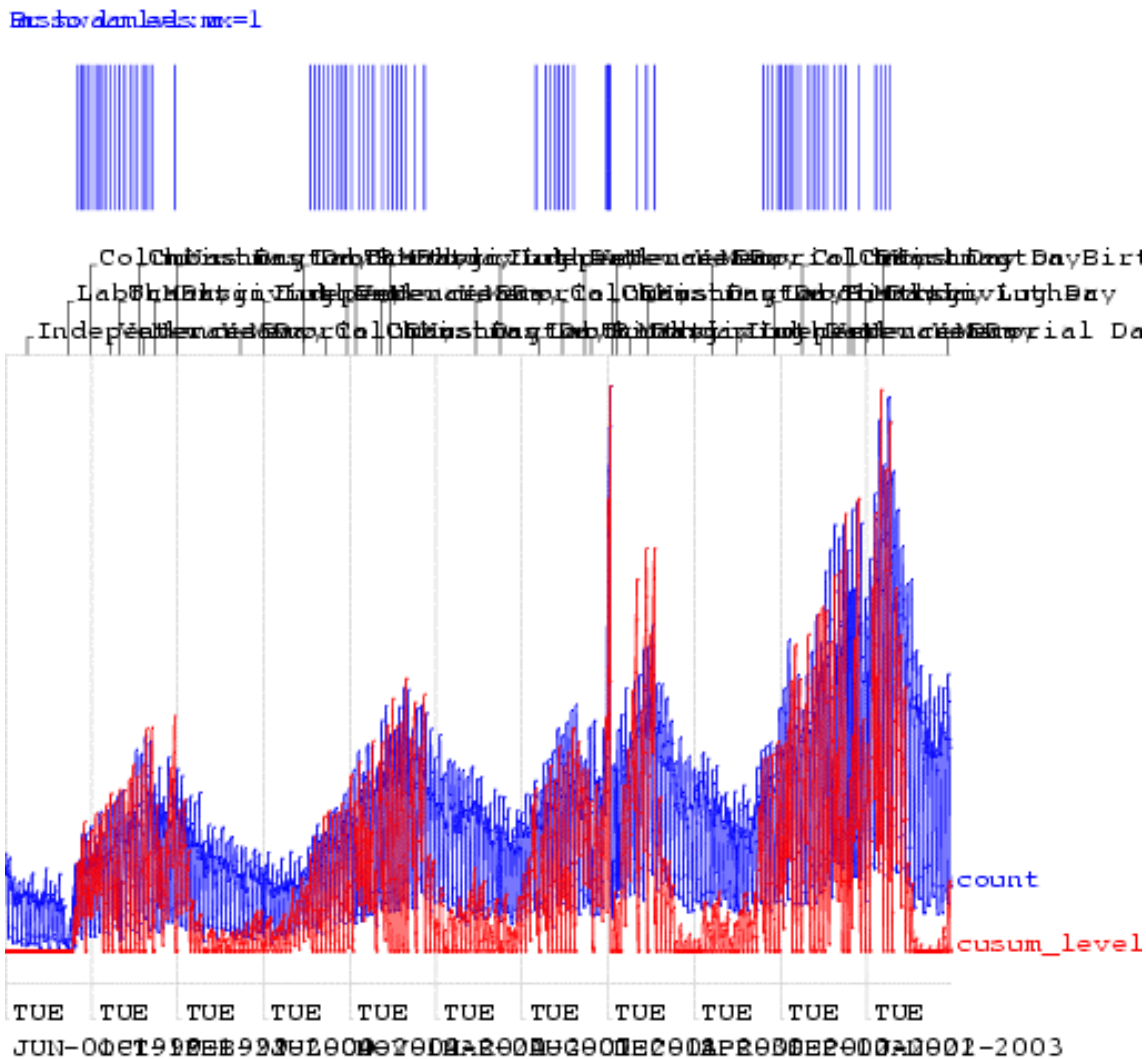
- CUmulative SUM Statistics

- 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

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

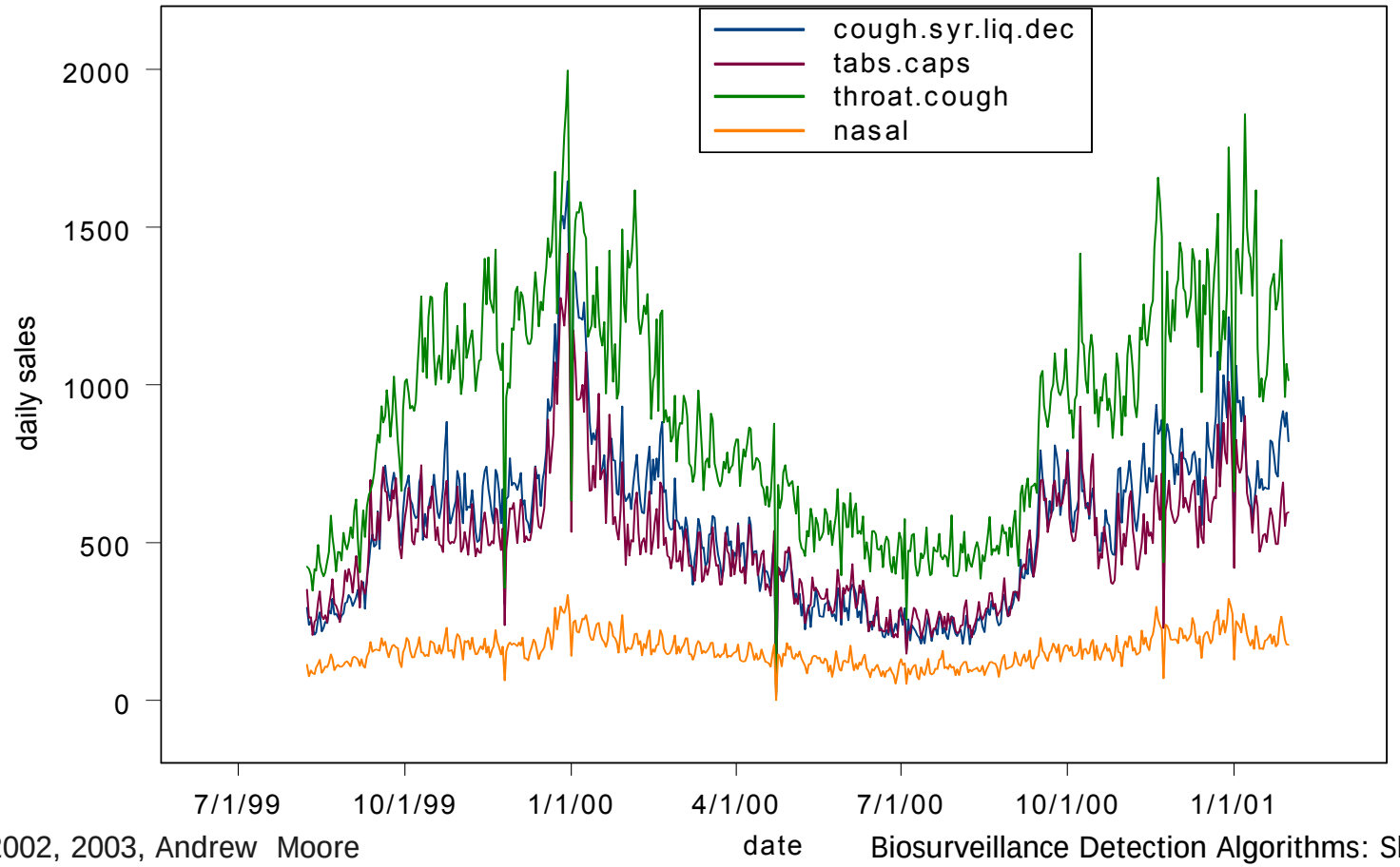
Spatial Scan Statistics

WSARE

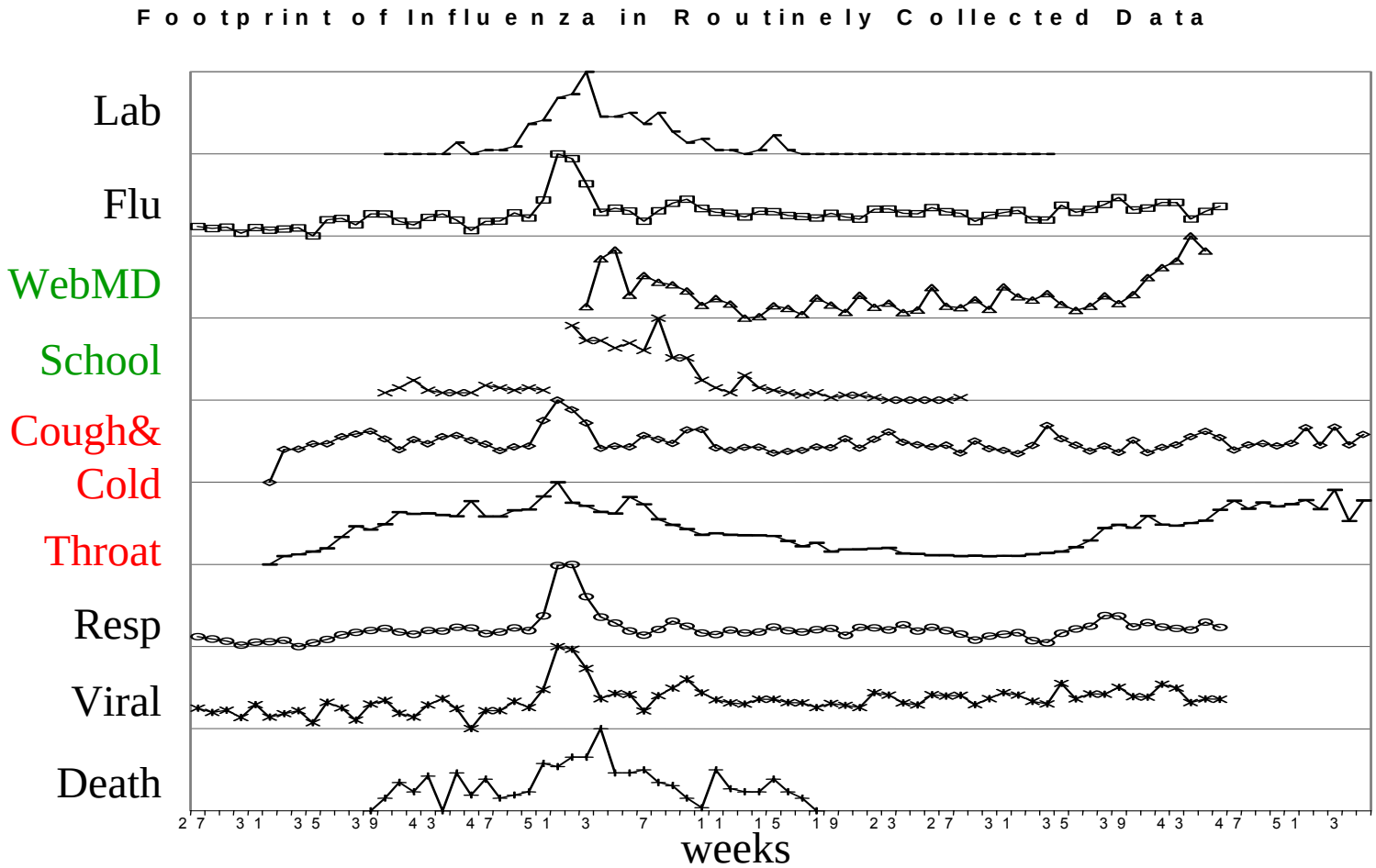
Multiple Signals

Multivariate Signals

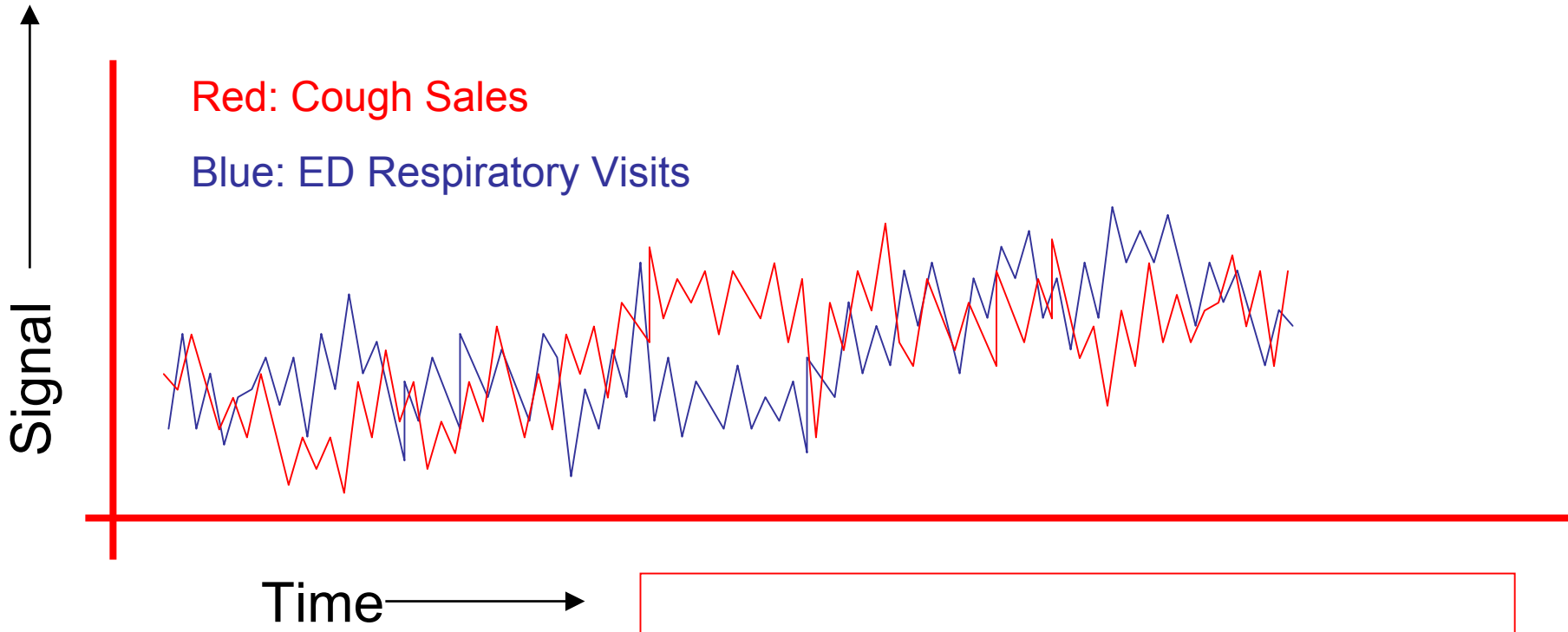
(relevant to inhalational diseases)



Multi Source Signals



What if you've got multiple signals?

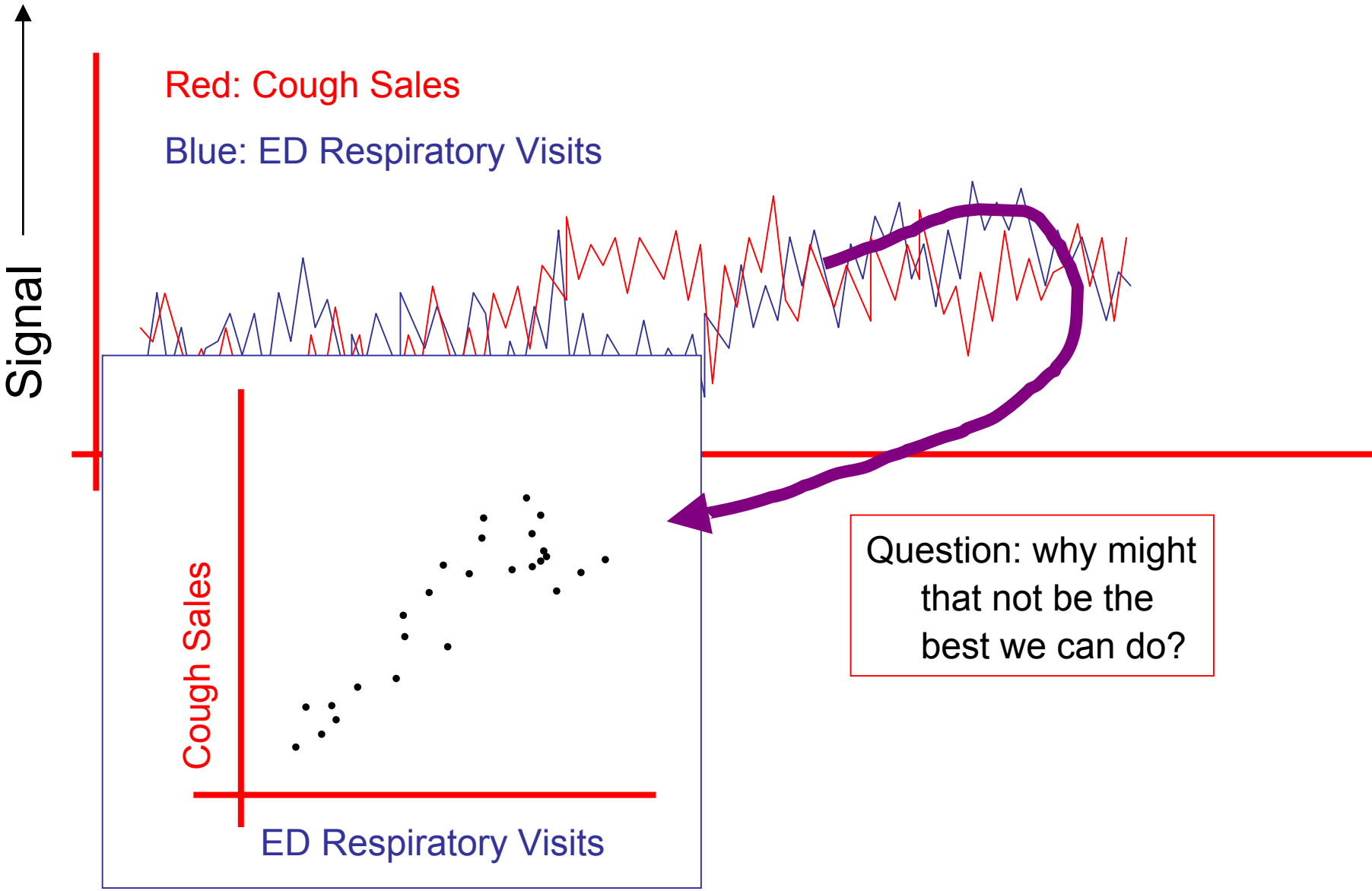


Idea One:

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

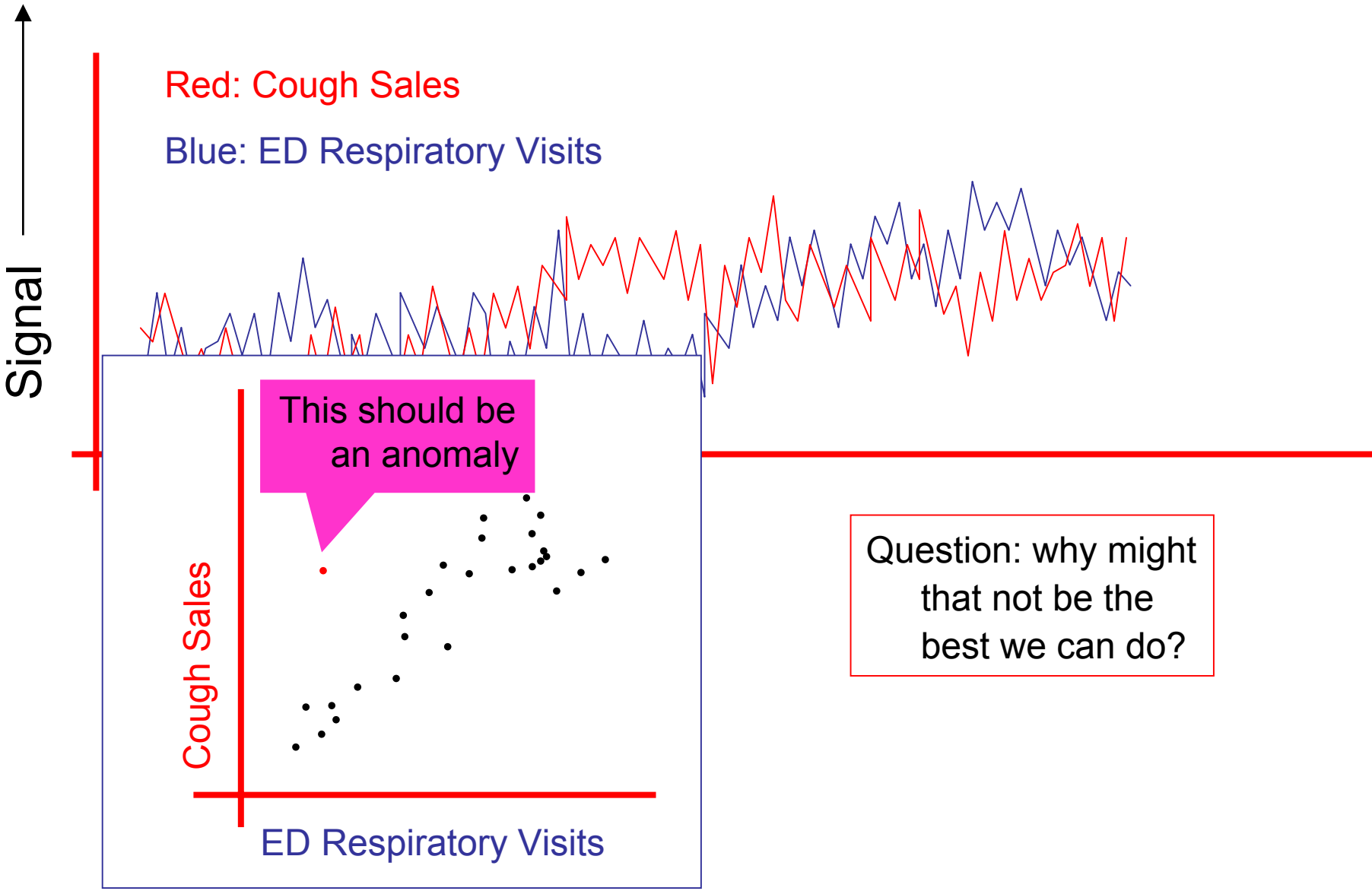
…Question: why might that not be the best we can do?

Another View



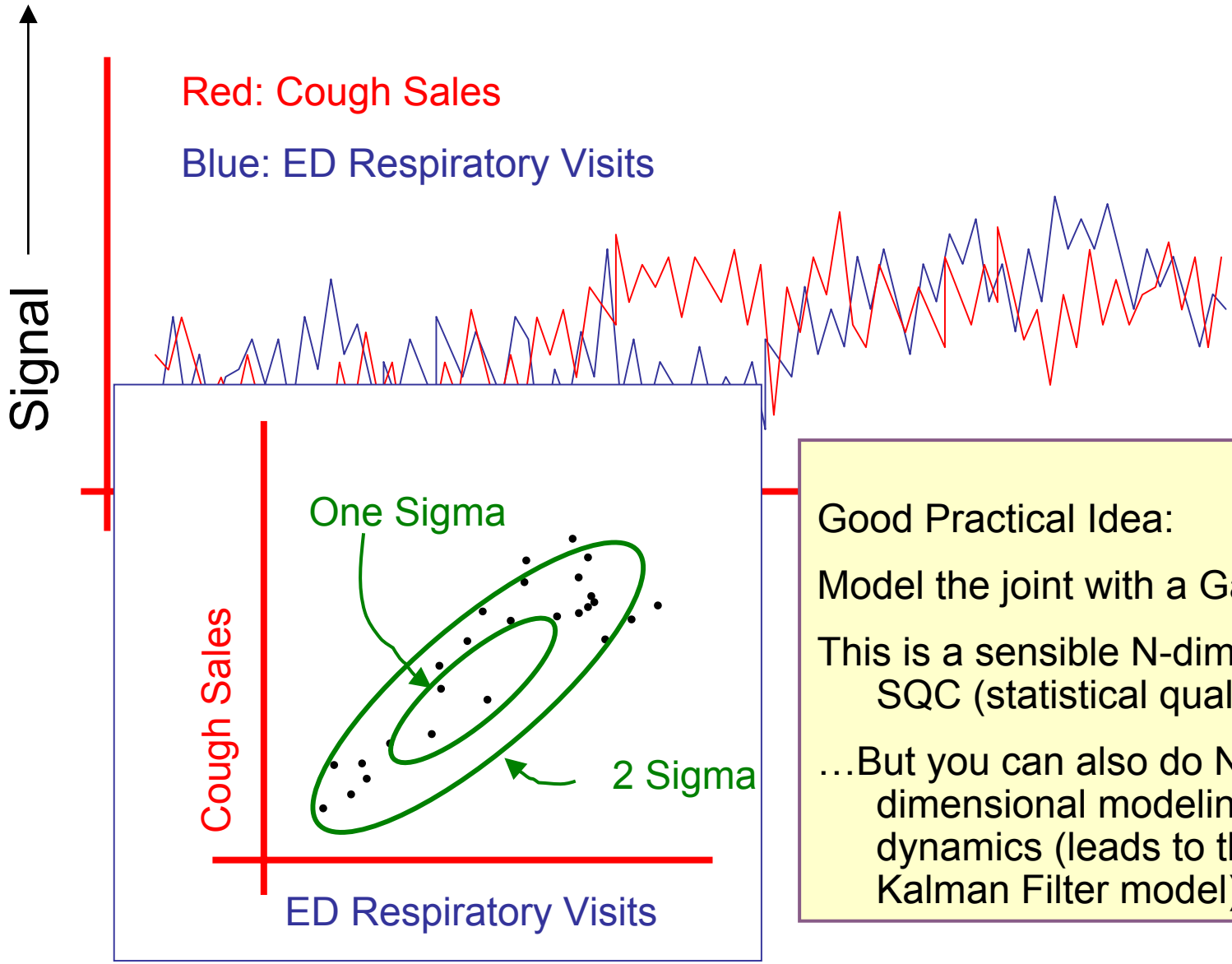
Question: why might that not be the best we can do?

Another View



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

N-dimensional Gaussian



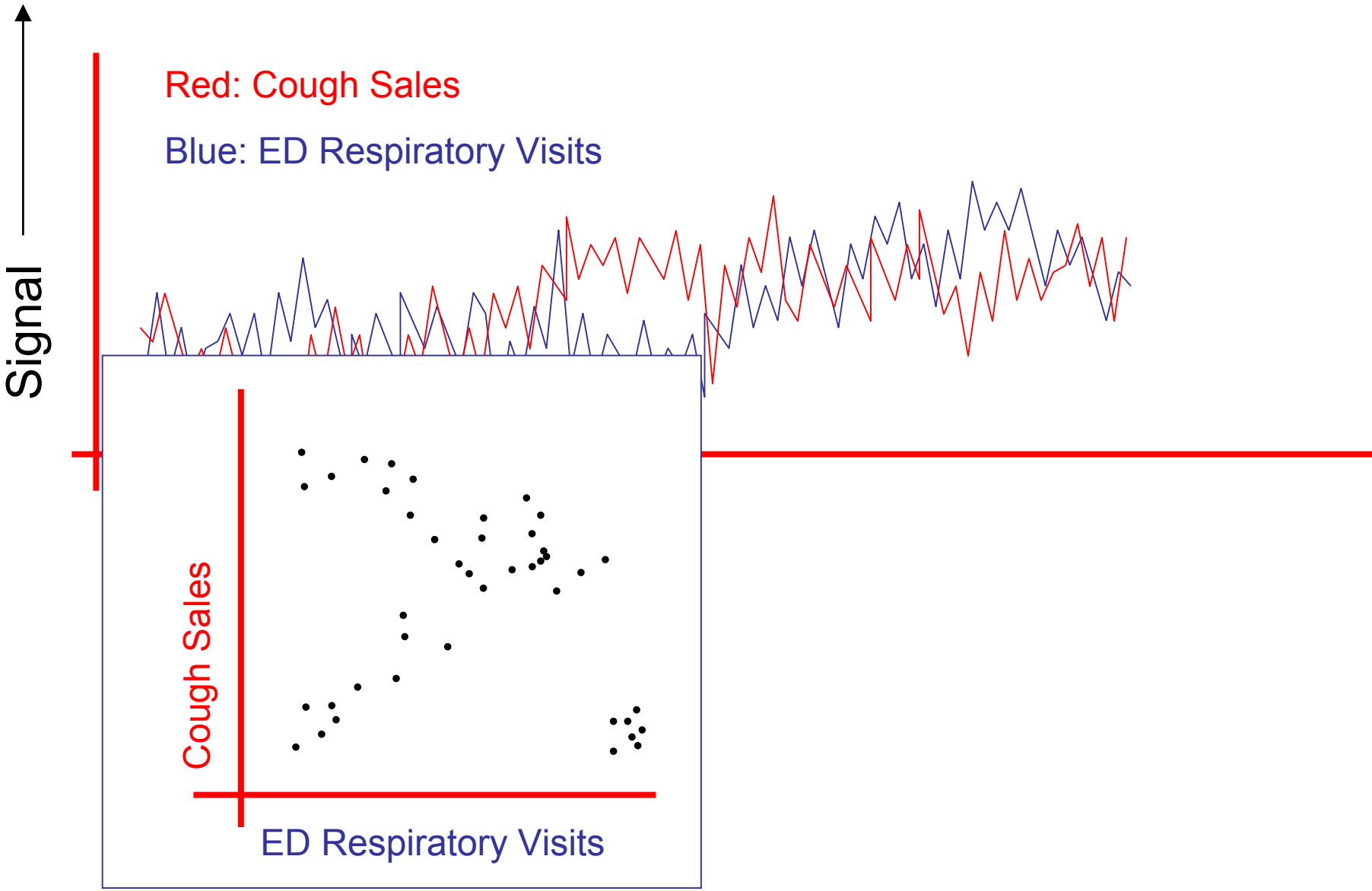
Good Practical Idea:

Model the joint with a Gaussian

This is a sensible N-dimensional SQC (statistical quality control)

…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

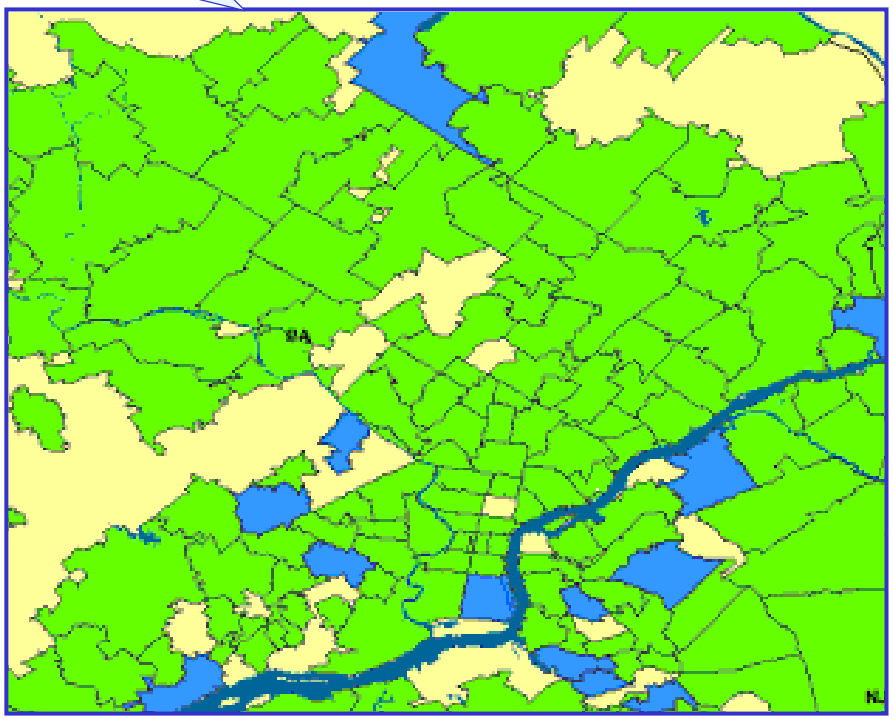
Multivariate Anomaly Detection

Spatial Scan Statistics

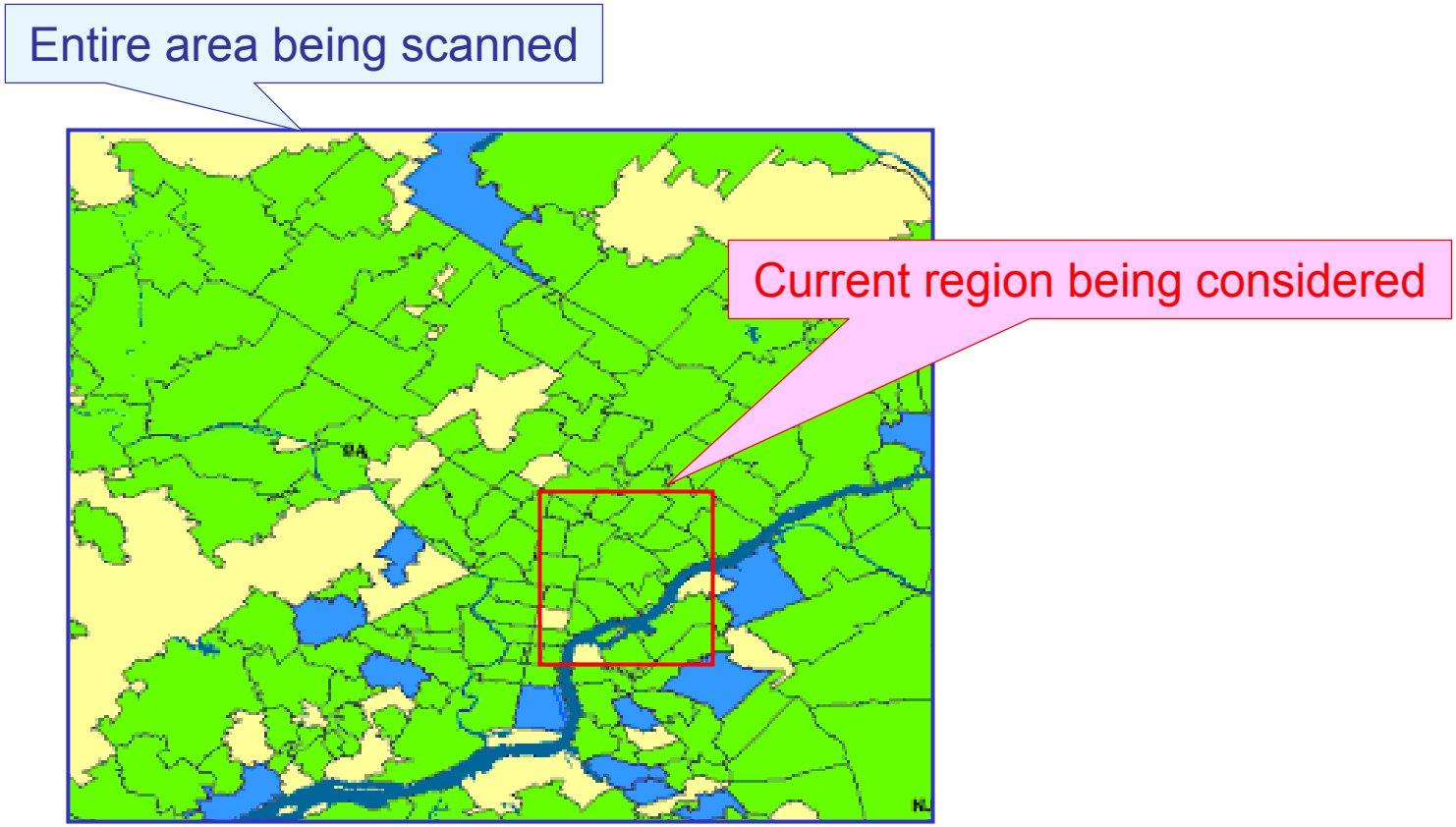
WSARE

One Step of Spatial Scan

Entire area being scanned

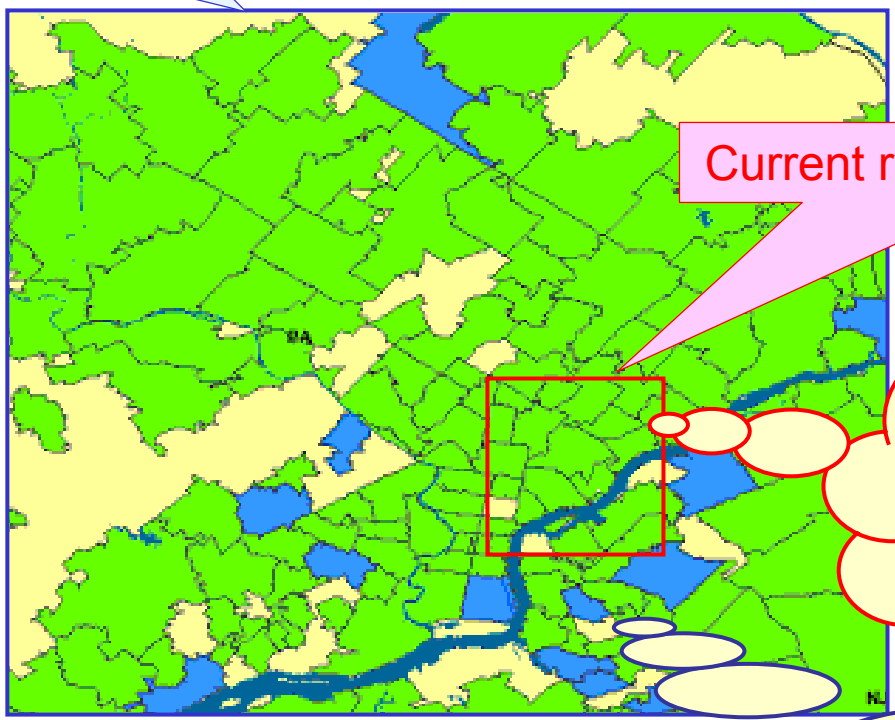


One Step of Spatial Scan



One Step of Spatial Scan

Entire area being scanned



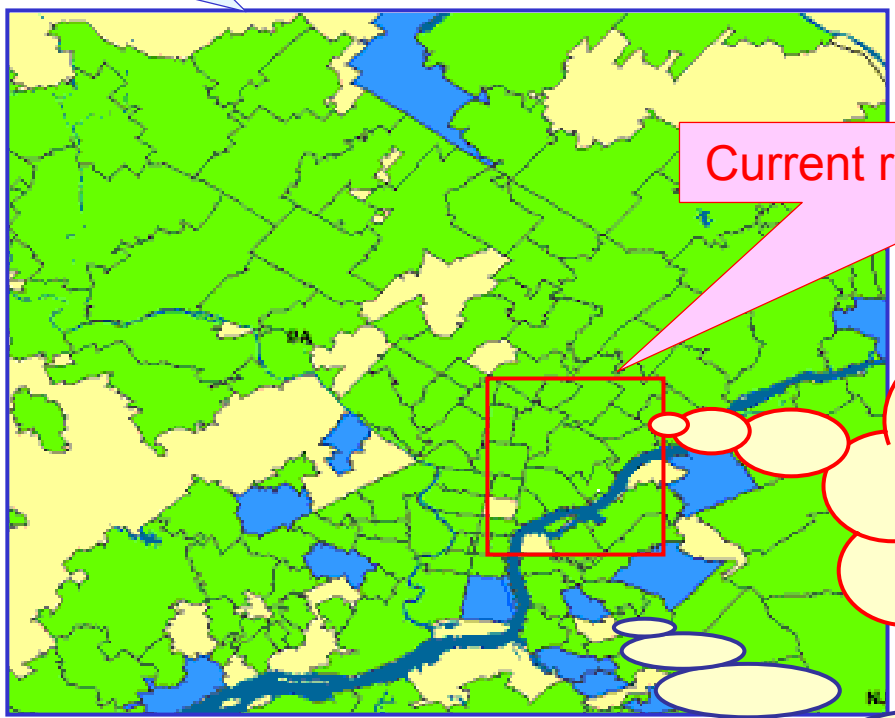
Current region being considered

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

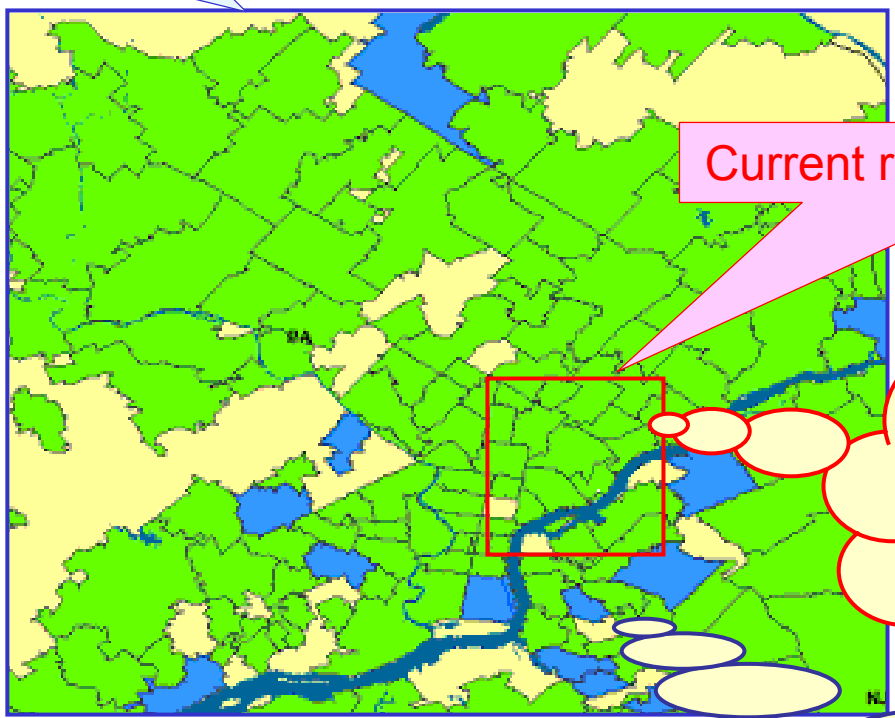
I have a population of 5300 of whom 53 are sick (1%)

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

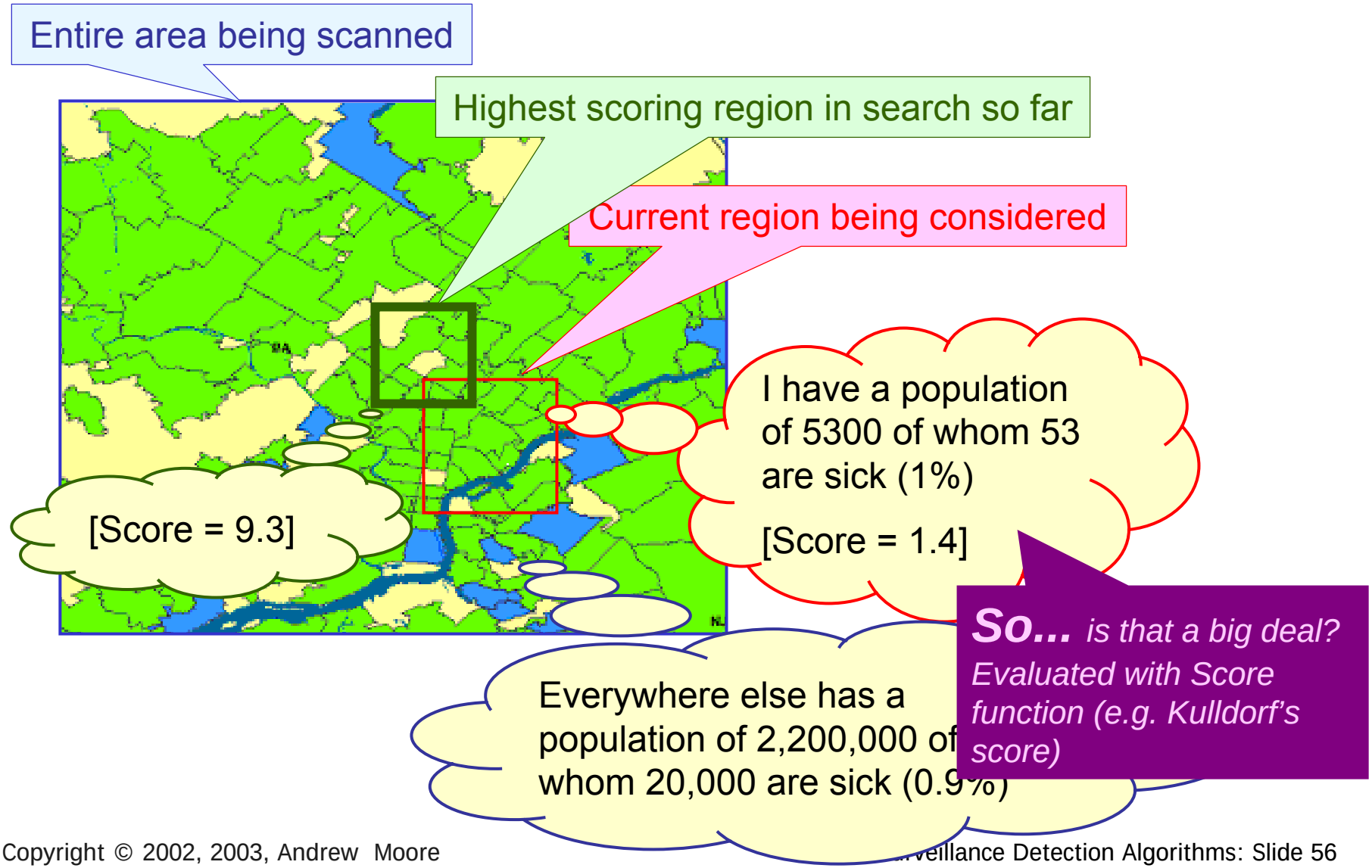
I have a population of 5300 of whom 53 are sick (1%)

[Score = 1.4]

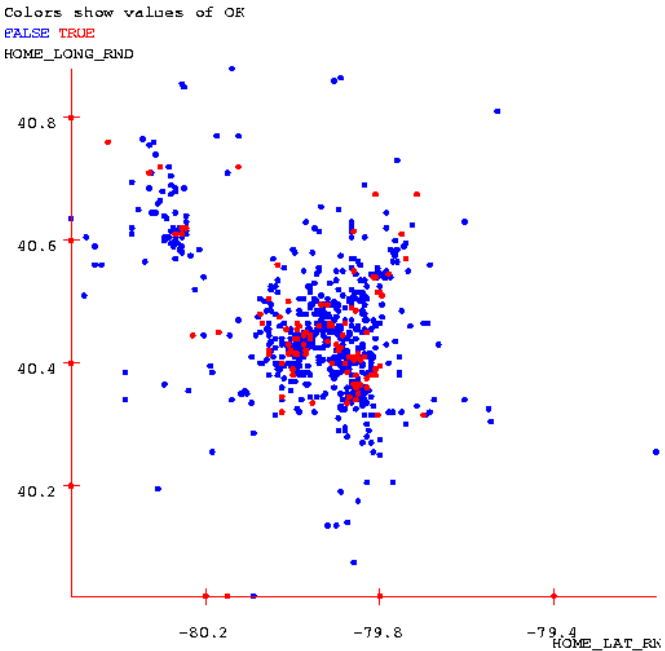
Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

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

Many Steps of Spatial Scan



Scan Statistics



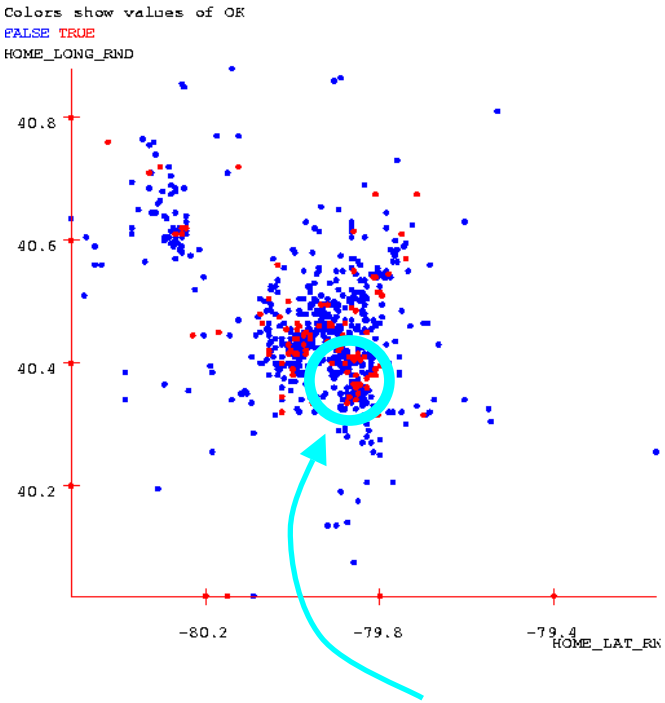
Standard scan statistic question:
Given the geographical locations of occurrences of a phenomenon, is there a region with an unusually high (low) rate of these occurrences?

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

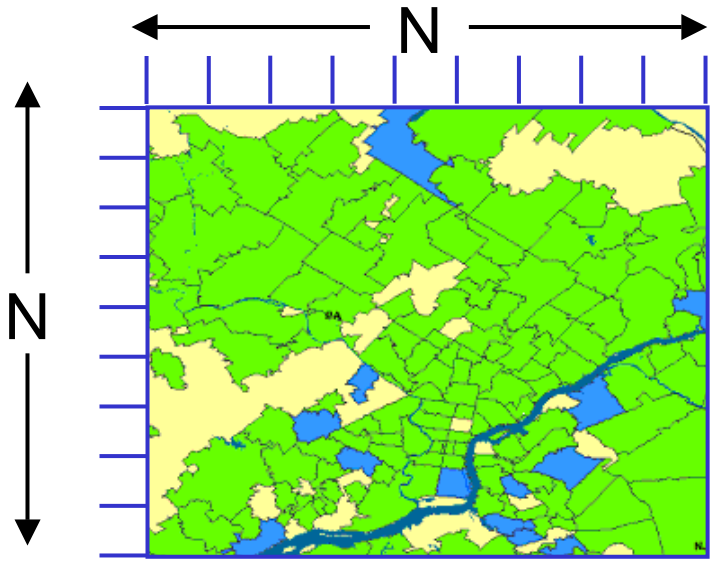


Given that region W is the most likely to be abnormal, is it significantly abnormal?

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.

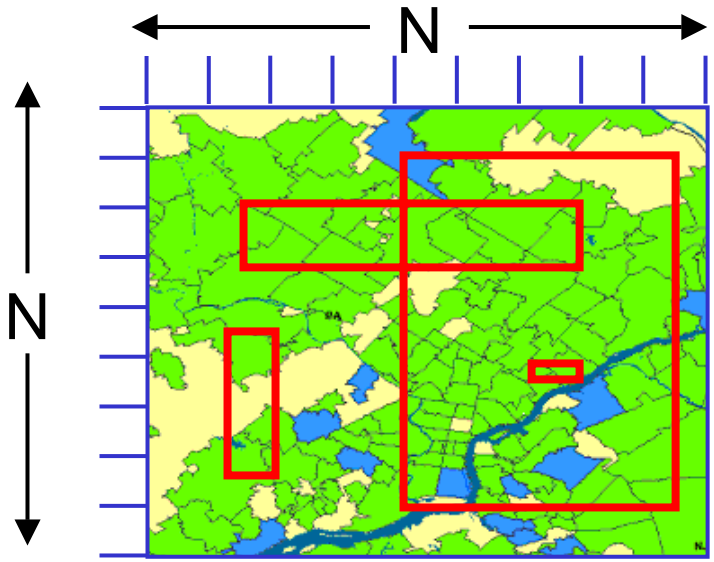
  If not, we can use several other speedup tricks.

- 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

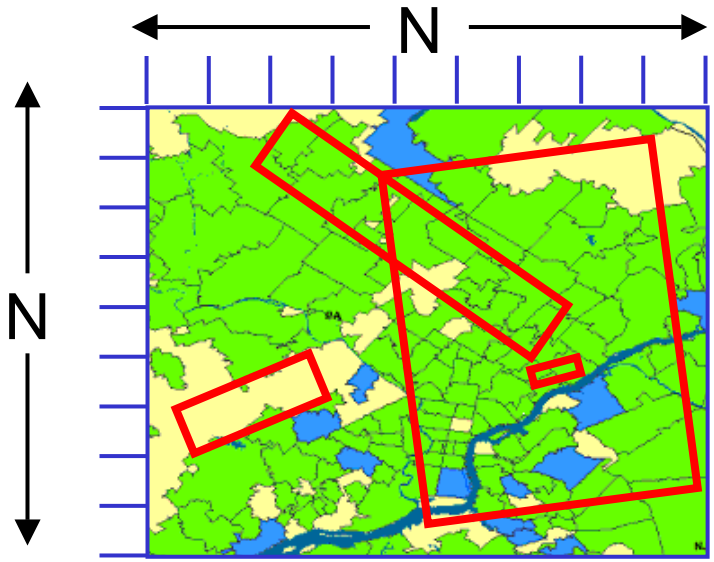
Work in progress



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

Fast oriented rectangles speedup

Work in progress

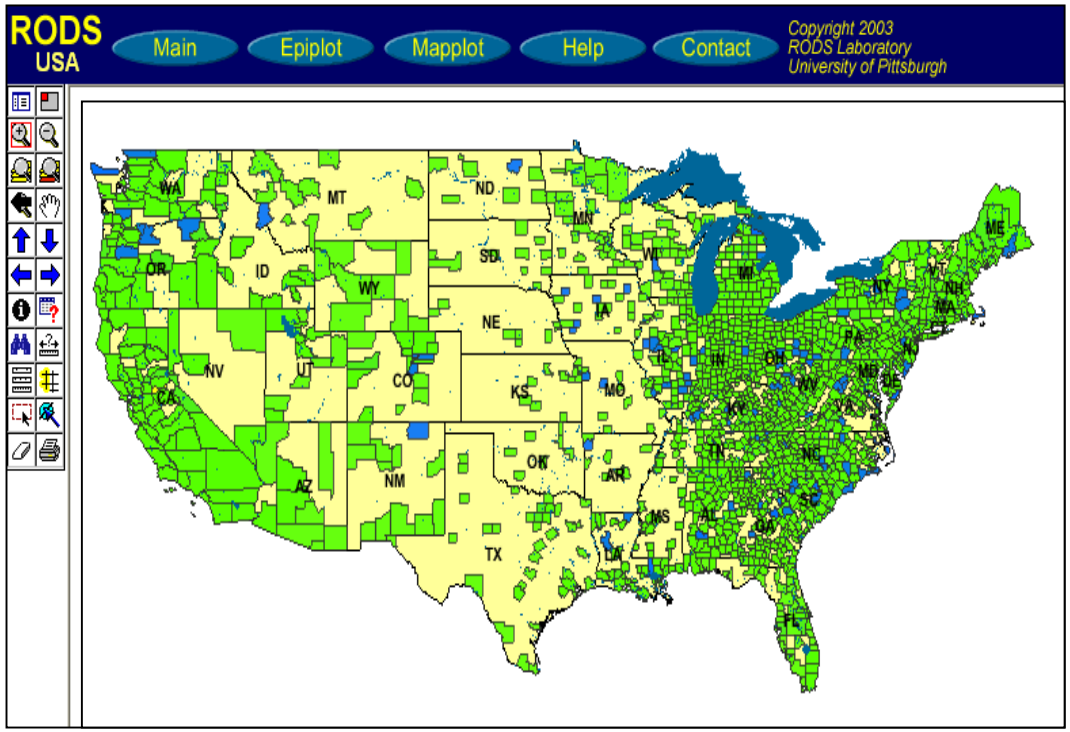


- Theoretical complexity of fast rectangles: $18N^2 \log N$ (as opposed to naïve $18N^4$)

(Angles discretized to 5 degree buckets)

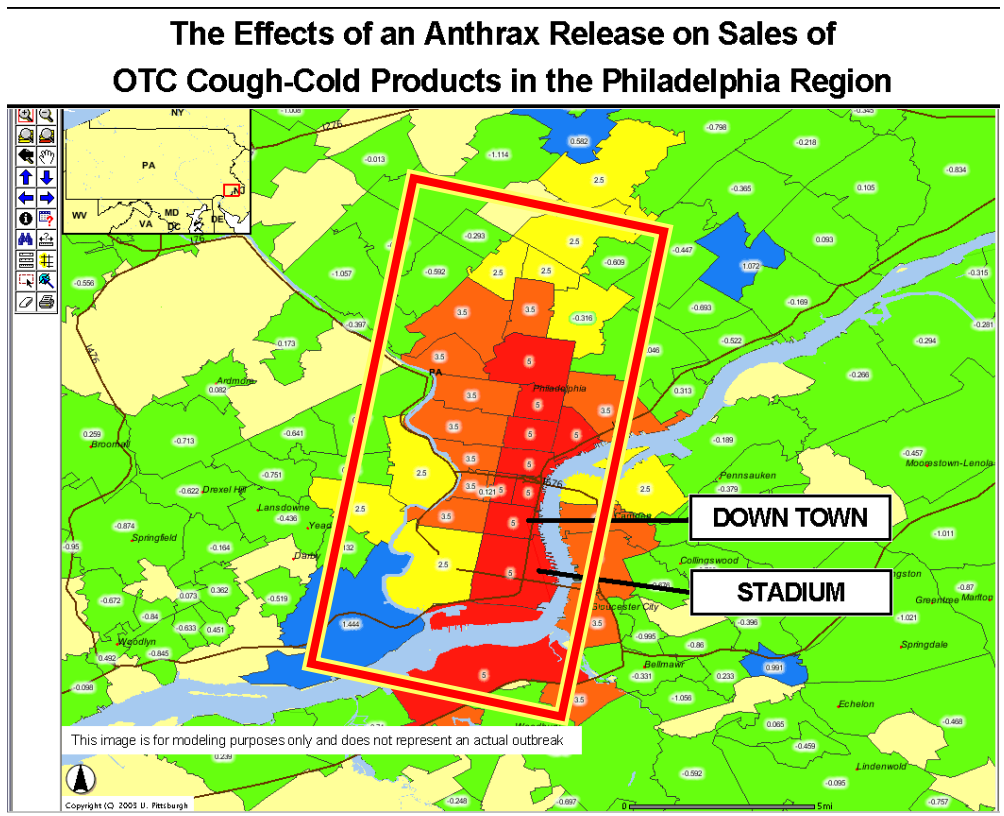
Why the Scan Statistic speed obsession?

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



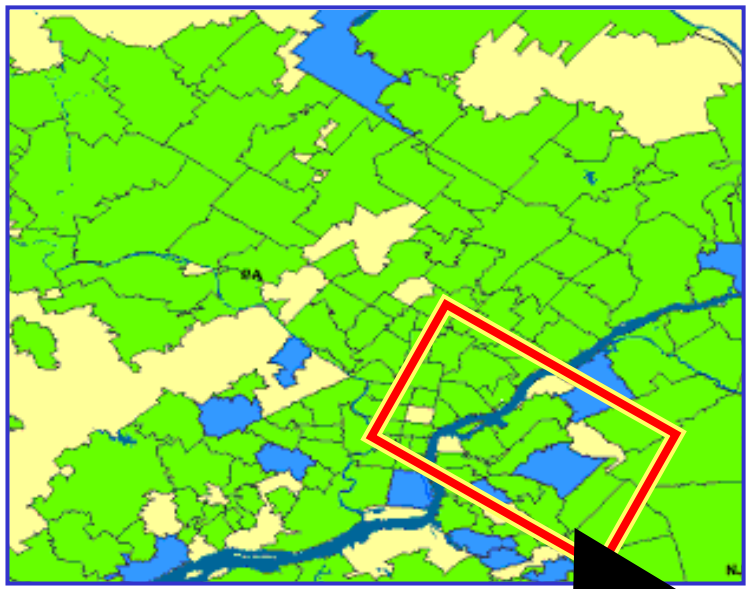
Why the Scan Statistic speed obsession?

- 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

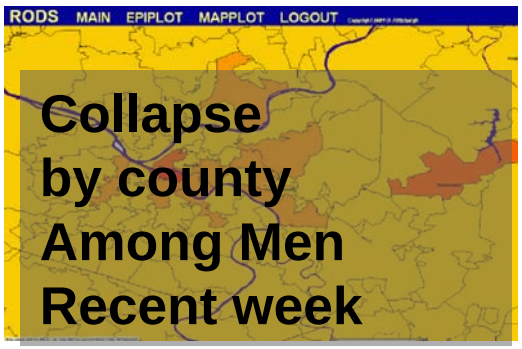
Suppose we know that today in the ED we had...

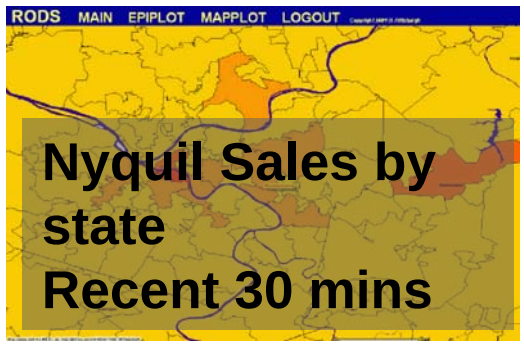
- 421 Cases
- 78 Respiratory Cases
- 190 Males
- 32 Children
- 21 from North Suburbs
- 2 Postal workers

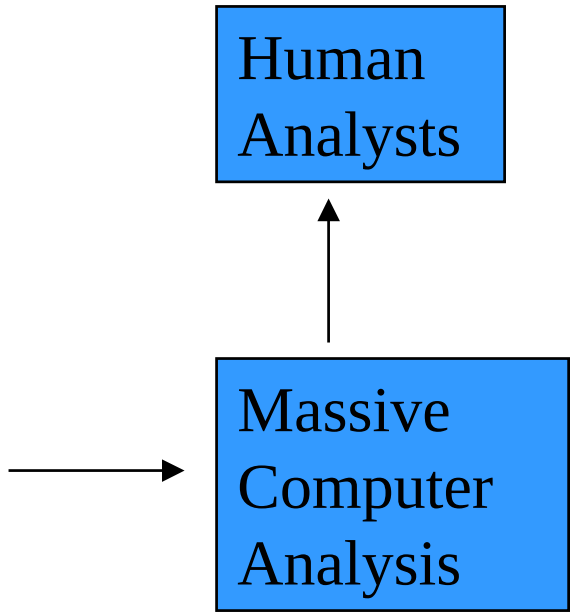
(etc etc etc)

Have we made best use of all possible information?

There are so many things to look at







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?

"last 24 hours"

"ignore key and weather"

Example

Primary Key	Date	Time	Hospital	ICD9	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	ICD9	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 wor
  of ER cases at 15
  hospitals...
- For each day...
  - Take today's cases
  - The cases one week ago
  - The cases two weeks ago
- Ask: "What's different
  about today?"

DATE_ADI	ICD9	PRODROI	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 wor
  of ER cases at 15
  hospitals...
- For each day...

DATE_ADI	ICD9	PRODROI	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]	[illegible]	s-w	…	…
[illegible]	[illegible]	1	F	s-w	…	…
[illegible]	[illegible]	[illegible]	[illegible]	n-w	…	…
[illegible]	[illegible]	[illegible]	[illegible]	:	…	…
[illegible]	[illegible]	[illegible]	[illegible]	s-e	…	…
[illegible]	[illegible]	[illegible]	[illegible]	s-e	…	…
[illegible]	[illegible]	[illegible]	[illegible]	n-w	…	…
[illegible]	[illegible]	[illegible]	[illegible]	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 cens

Overfitting Alert!

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?

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

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

Primary Key	Date	Time	Hospital	ICD	[illegible]	[illegible]	Age	Home			Work			[illegible]	Recent Weather	(Many more...)
								Large Scale	Medium Scale	Fine Scale	Large Scale	Medium Scale	Fine Scale			
h6r32								[illegible]E	19							
t3q15								[illegible]E	15							
t5hh5	6/2/2	14:15	Smith-field	622	Respiratory	F	80s	SE	15							
:	:	:	:	:	:	:	:	:	:							

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

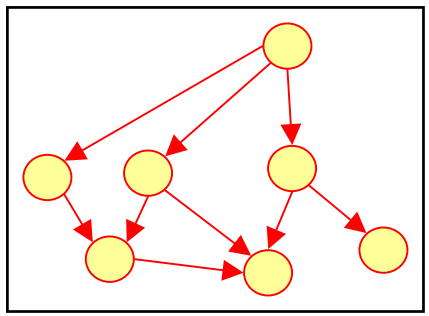
- Good news:

  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"

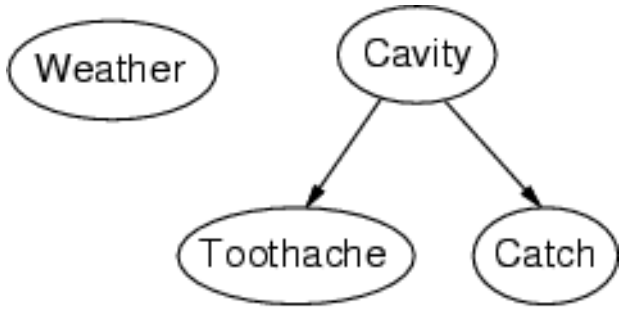
"On Cold Tuesday Mornings the folks coming in from the North part of the city are more likely to have respiratory problems"

"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"

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions



- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

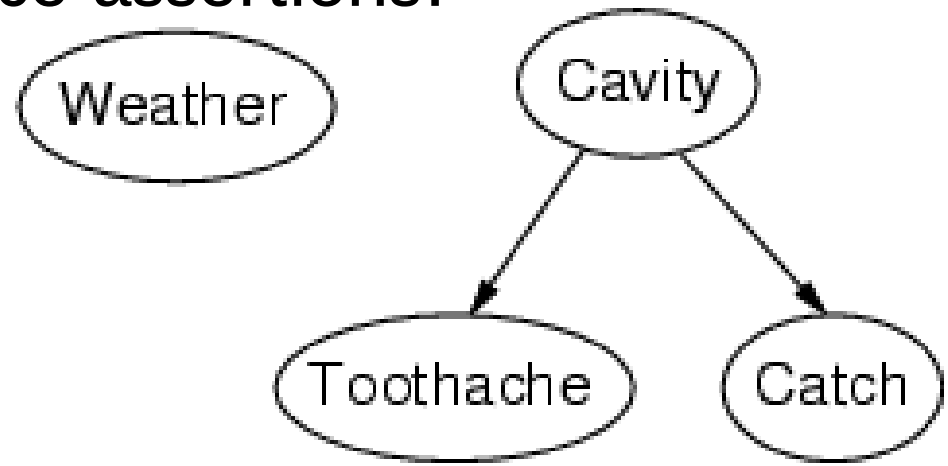
$$\mathbf{P}\,(X_i \mid \text{Parents}\,(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

(from Russell and Norvig)

Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

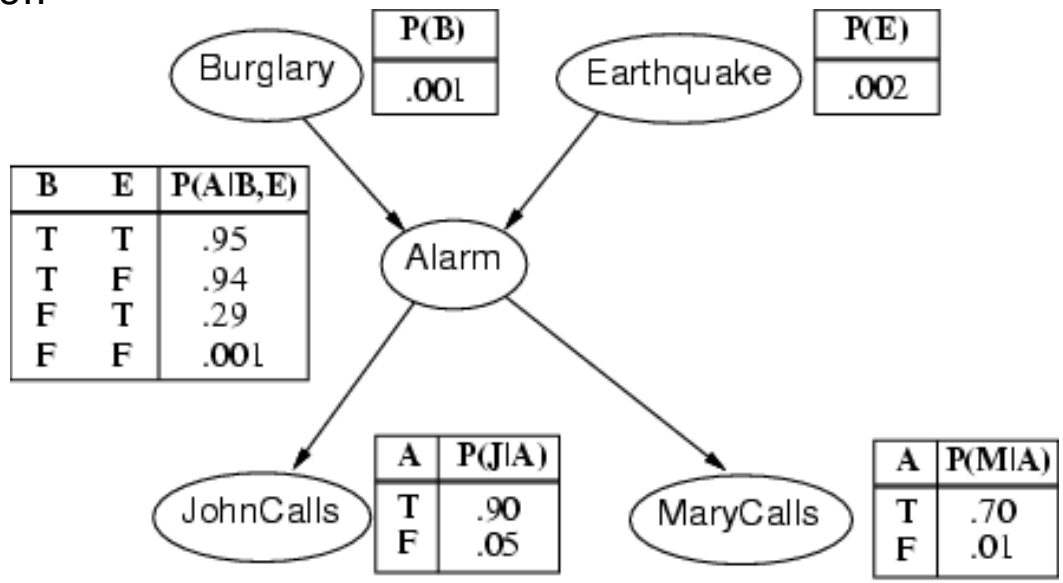
(from Russell and Norvig)

Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

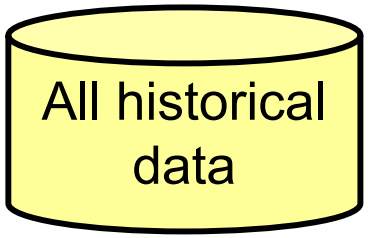
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call



(from Russell and Norvig)

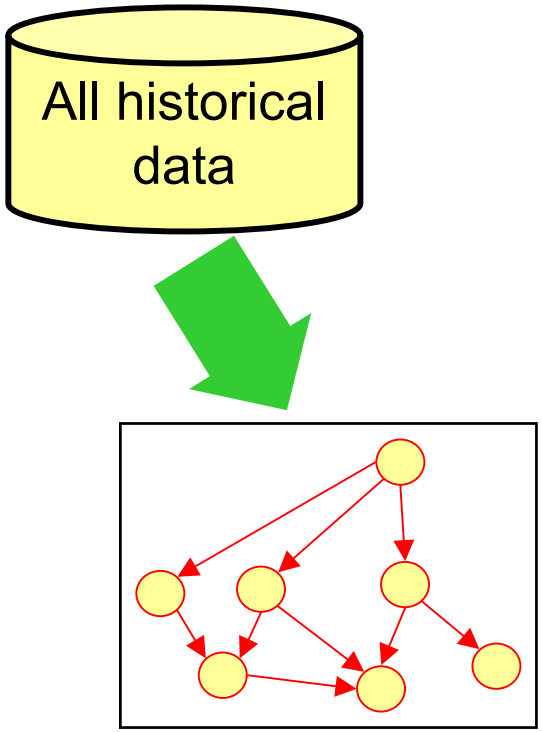
WSARE 3.0

(software developed by A. Moore, et al, at CMU)



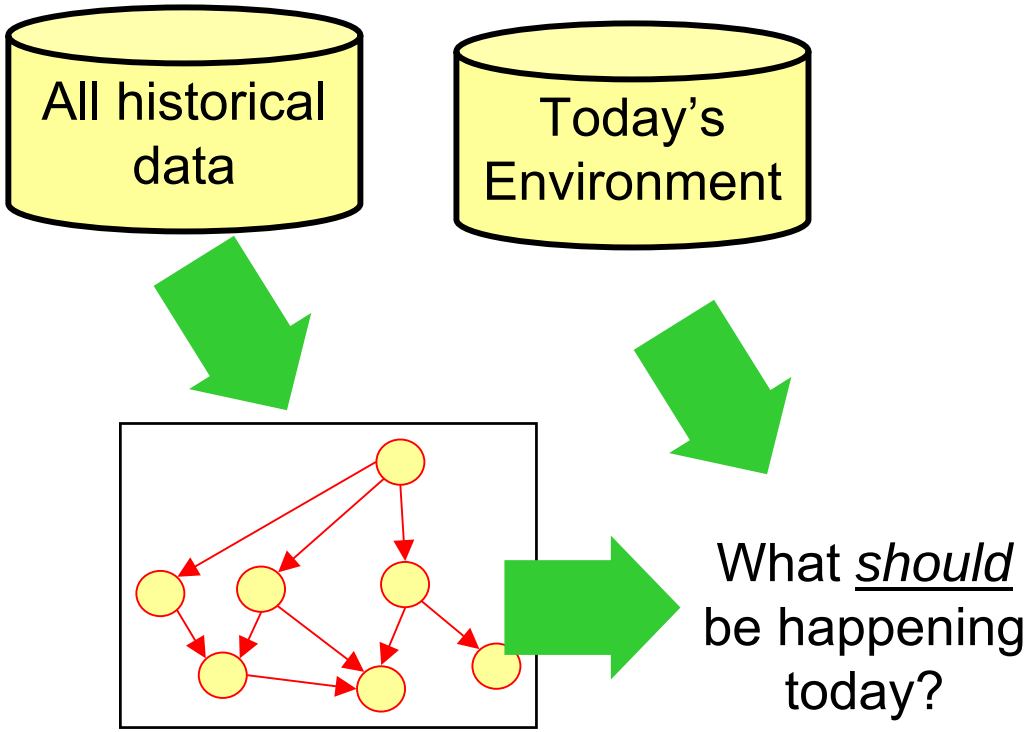
WSARE 3.0

(software developed by A. Moore, et al, at CMU)



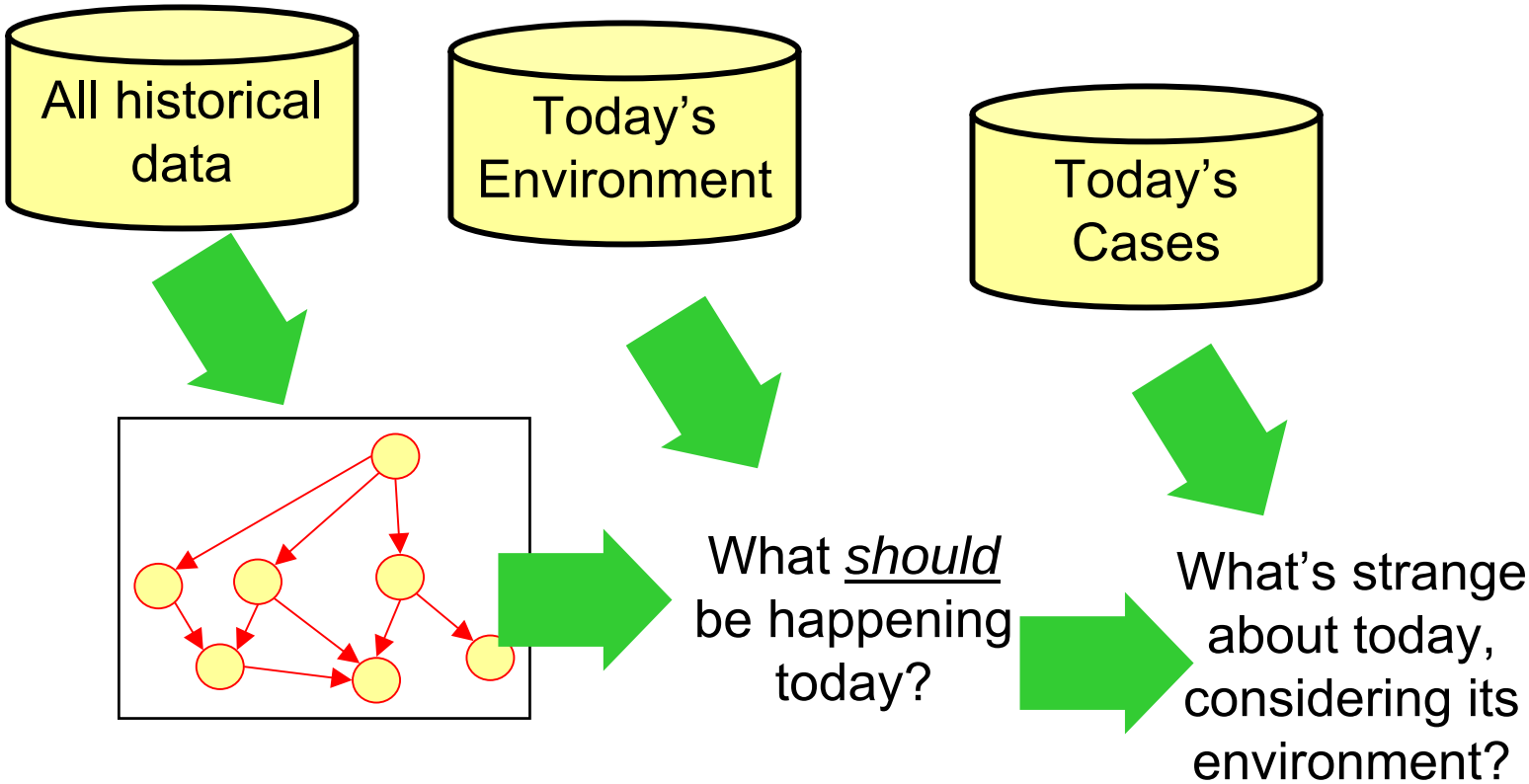
WSARE 3.0

(software developed by A. Moore, et al, at CMU)



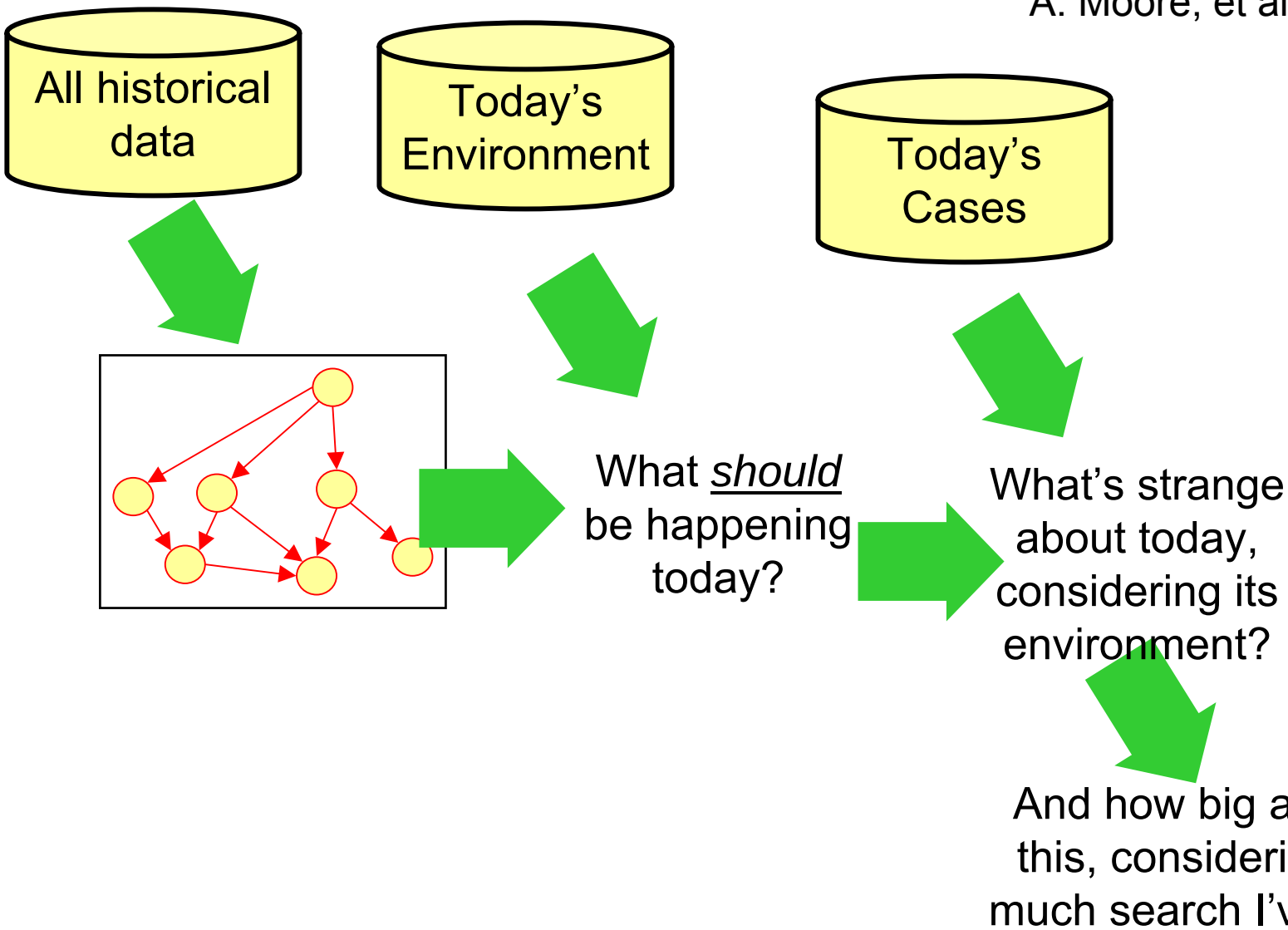
WSARE 3.0

(software developed by A. Moore, et al, at CMU)



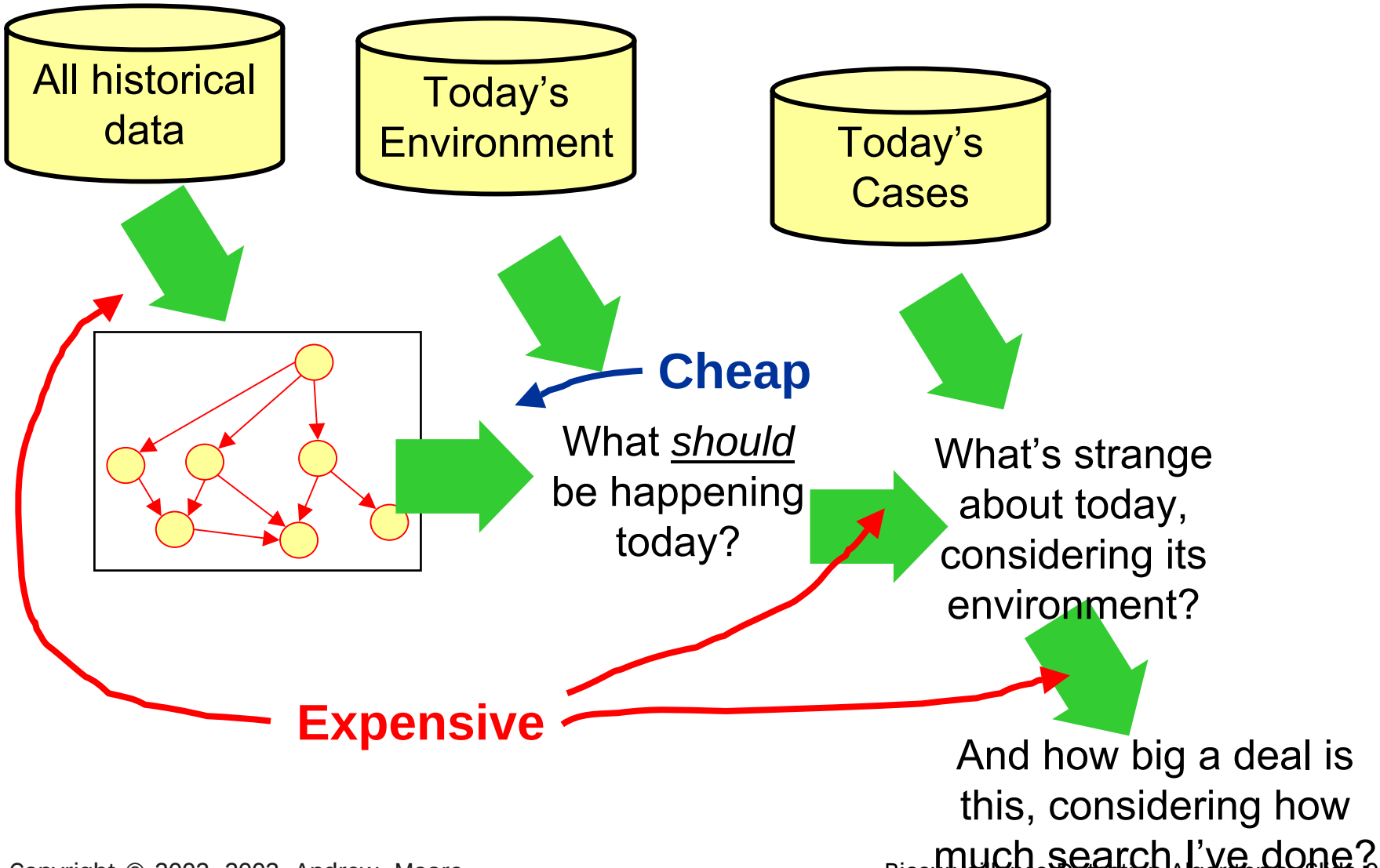
WSARE 3.0

(software developed by A. Moore, et al, at CMU)



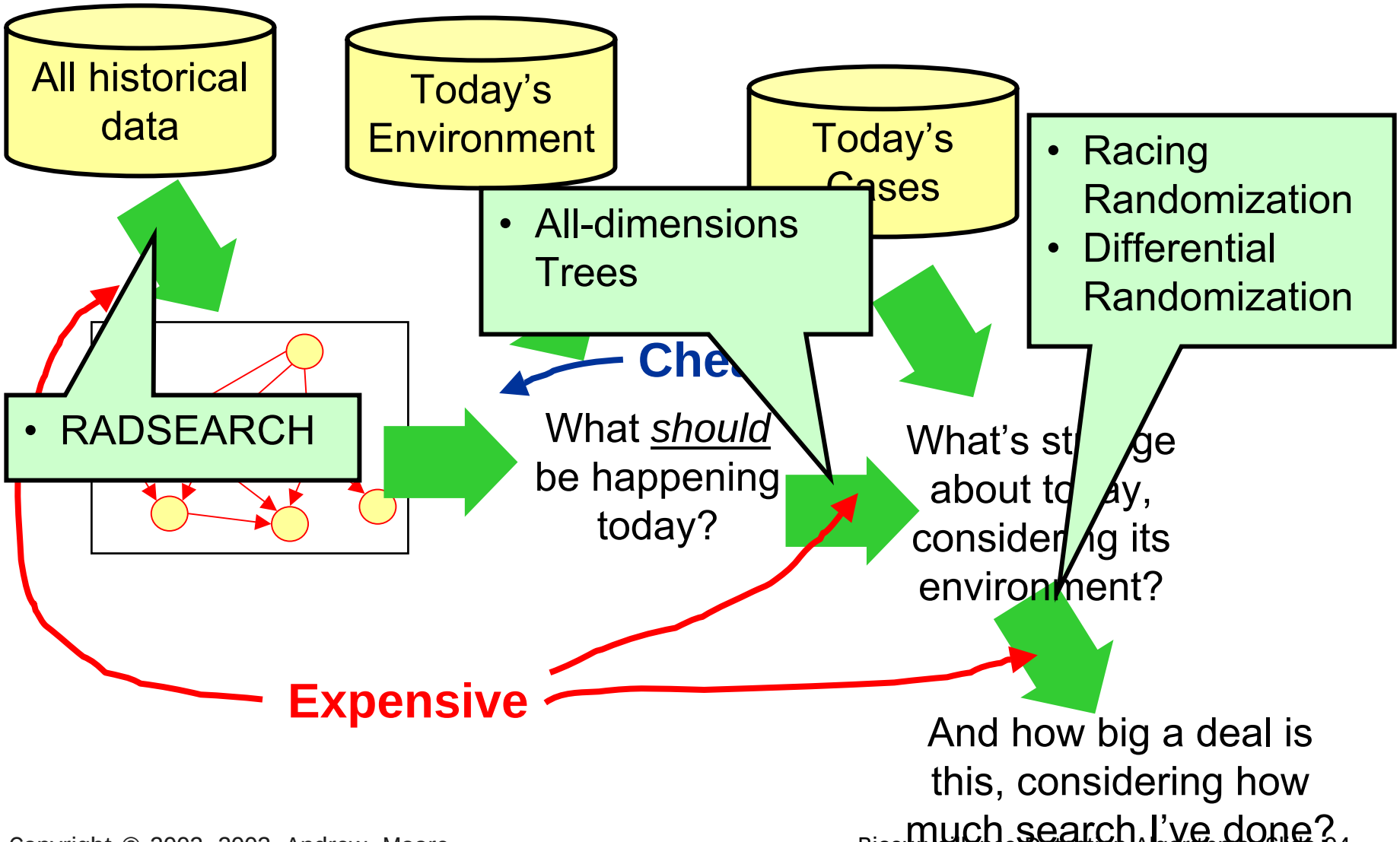
WSARE 3.0

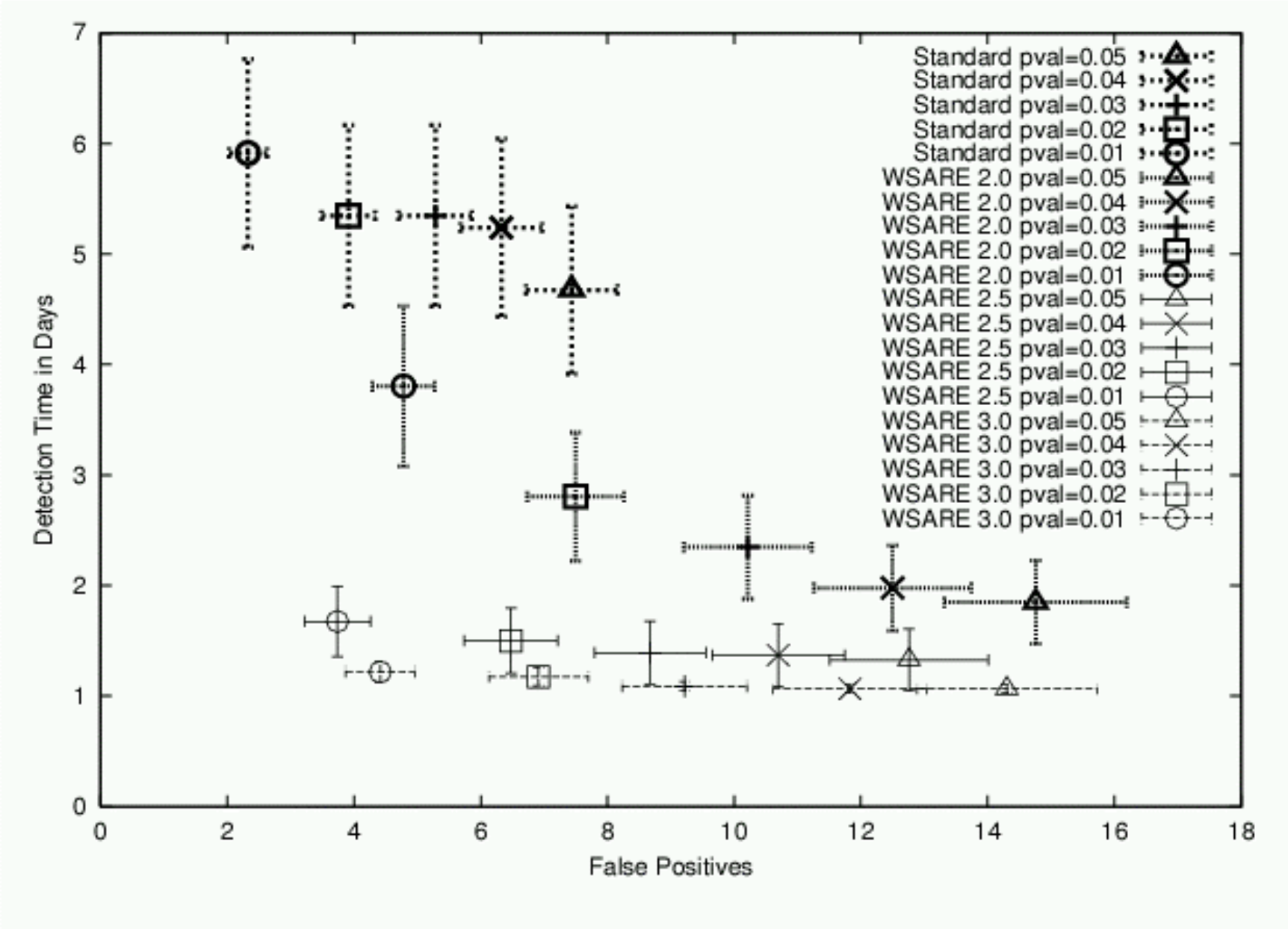
(software developed by A. Moore, et al, at CMU)



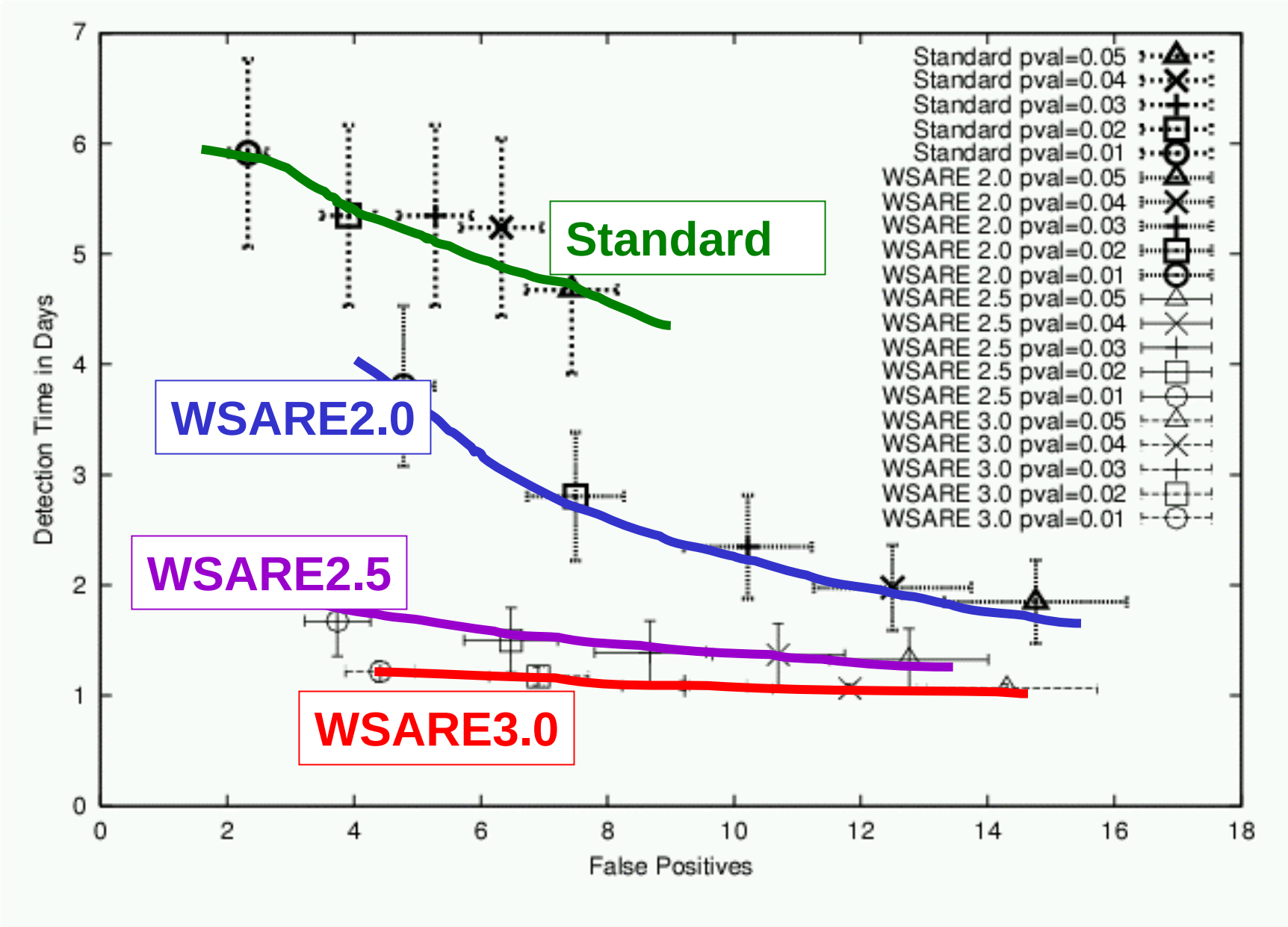
WSARE 3.0

(software developed by A. Moore, et al, at CMU)





Results on Simulation



Results on Simulation

Conclusion

- WSARE: 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 conditioning unique features of today

- Computationally intense unless we're tricksy!

The End... ☺