

# Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks

Your presenter today

**Andrew Moore**  
**Weng-Keen Wong**  
**Mike Wagner**  
**Greg Cooper**

In collaboration with many others, including Wendy Chapman, Bill Hogan, Bob Olszewski, Jeff Schneider, Rich Tsui, Weng-Keen Wong

University of Pittsburgh and Carnegie Mellon University  
[www.cs.cmu.edu/~awm](http://www.cs.cmu.edu/~awm)

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 1

## The Auton Lab

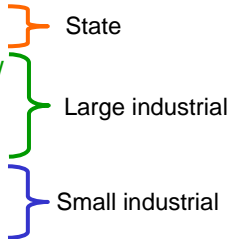
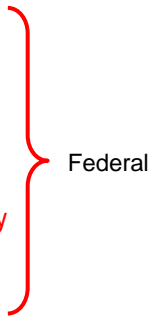
<b>Faculty:</b>	Andrew Moore, Jeff Schneider, Artur Dubrawski
<b>Postdoctoral Fellows:</b>	Brigham Anderson, Alexander Gray, Paul Komarek, Dan Pelleg
<b>Graduate Students:</b>	Brent Bryan, Kaustav Das, Khalid El-Arini, Anna Goldenberg, Jeremy Kubica, Ting Liu, Daniel Neill, Sajid Siddiqi, Purna Sarkar, Ajit Singh, Weng-Keen Wong
<b>Head of Software Development:</b>	Jeanie Komarek
<b>Programmers:</b>	Patrick Choi, Adam Goode, Pat Gunn, Joey Liang, John Ostlund, Robin Sabhnani, Rahul Sankathar
<b>Executive Assistant:</b>	Kristen Schrauder
<b>Head of Sys. Admin:</b>	Jacob Joseph
<b>Undergraduate and Masters Interns:</b>	Kenny Daniel, Sandy Hsu, Dongryeol Lee, Jennifer Lee, Avilay Parekh, Chris Rotella, Jonathan Terleski
<b>Recent Alumni:</b>	Drew Bagnell (RI faculty), Scott Davies (Google), David Cohn (Google), Geoff Gordon (CMU), Paul Hsiung (USC), Marina Meila (U. Washington), Remi Munos (Ecole Polytechnique), Malcolm Strens (Qinetiq)

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 2

# Current Sponsors

- National Science Foundation (NSF)
- NASA
- Defense Advanced Research Projects Agency (DARPA)
- Other Government Sponsor
- Department of Homeland Security (DHS)
- Homeland Security Advanced Research Projects Agency (HSARPA)
- United States Department of Agriculture (USDA)
- Health Canada
- State of Pennsylvania
- A Fortune-50 Pharmaceutical company
- Caterpillar Corporation
- A Fortune-50 Petrochemical company
- Psychogenics Corporation
- Transform Pharmaceuticals



Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 3

# Collaborators...

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 4

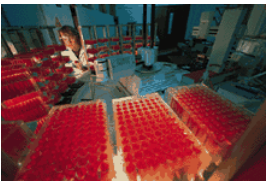
1996	1997	1998	1999	2000	2001	2002	2003
M&M Mars Line control Adrenaline (NOX minimization)	Kodak (Image stabilization) Digital Equipment (pregnancy monitoring)	M&M Mars (manufacturing) NASA/NSF (Astrophysics mining) 3M: Textile tension control	Caterpillar (Spare parts) US Army (biotoxin detection) M&M Mars: Scheduling with uncertainty 3M (Adhesive design)	DigitalMC (Music tastes) Caterpillar (emissions) SmartMoney (anomalies) Unilever (Brand Management) Phillips Petroleum (work-force optimization) Cellomics (screened anomaly detection)	Biometrics company (health monitor) Boeing (intrusion) Masterfoods (new product development) Cellomics (pro-teomics screen) ABB (Circuit-breaker supply chain) SwissAir (Flight delays) 3M (secret) Washington Public Hospital System (ER delays) Unilever (targeted marketing)	NASA (National Virtual Observatory) NSF (astrostatistics software) DARPA (national disease monitor) Masterfoods (biochemistry) Pfizer (High-throughput screen) Caterpillar Inc. (Self-optimizing Engines) Beverage Company (Ingredients/Manufacturing/Marketing/Sales Bayes Net) Transform Pharma (massive autonomous experiment design) Census Bureau (privacy protection) Psychogenics Inc. (Effects of psychotropic drugs on rats)	NSF (astrostatistics software) Masterfoods (biochemistry) State of PA (National Disease Monitor [with Mike Wagner of U. Pitt]) State of PA (Anti Cancer [collaboration with CMU Biology]) DARPA (detecting patterns in links) Other Government Departments (identifying dangerous people, potential collaborators, and aliases) Other Government Departments (detecting a class of clusters) Other Pharma Research Co. Life Science specific data mining United States Department of Agriculture: Early warning system for food terrorism NSF: Biosurveillance Algorithms

Auton/SPR Deployments


Copyright © 2002-2004, Andrew Moore Biosurveillance: Slide 5

### Our 5 biggest applications in 2004

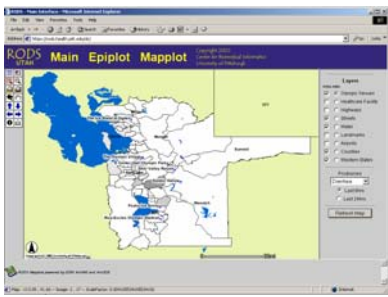
Drug Screening




Big Astrophysics Automated Science



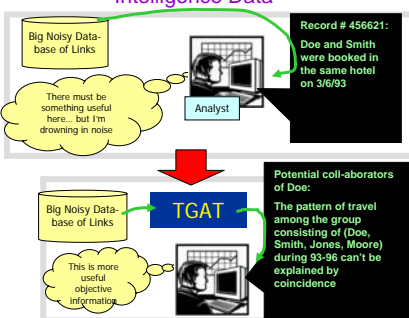
Biomedical Security (with Mike Wagner, University of Pittsburgh)



Autonomous self-tweaking engines



Intelligence Data

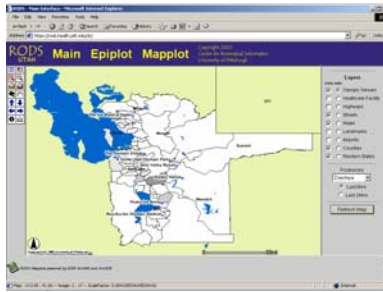
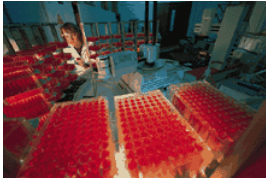


Copyright © 2002-2004, Andrew Moore Biosurveillance: Slide 6

**Our 5 biggest applications in 2002**

Biomedical Security (with Mike Wagner, University of Pittsburgh)

Drug Screening



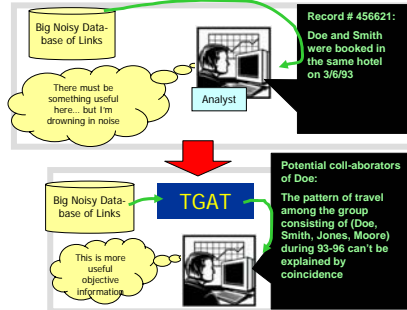
Autonomous self-tweaking engines



Big Astrophysics Automated Science



Security Surveillance Mining



Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 7

*..Early Thursday Morning. Russia. April 1979...*



Copyright © 2002-2004, Andrew Moore

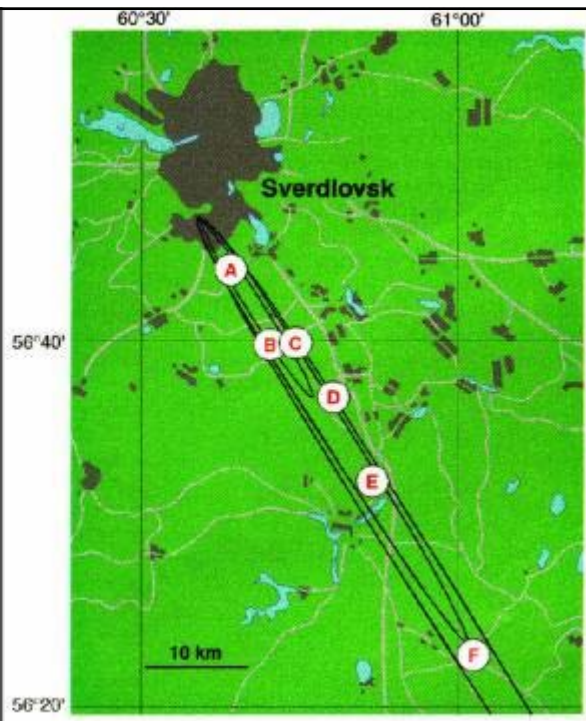
Biosurveillance: Slide 8

## Sverdlovsk: Aerial View

- During April and May 1979, there were 77 Confirmed cases of inhalational anthrax

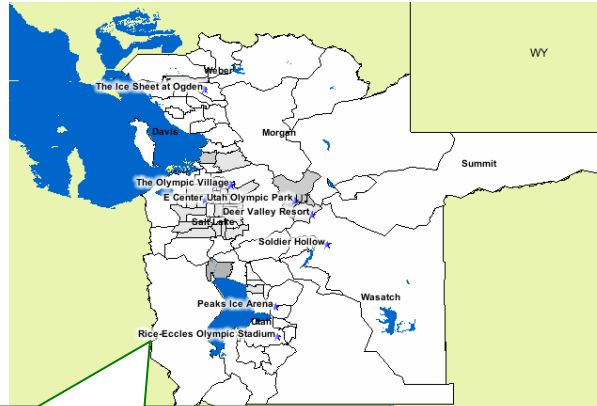


## Sverdlovsk Region: Epi-map



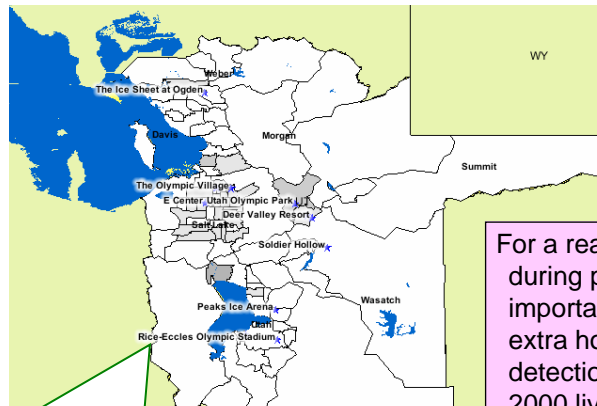


What if this happened in 2002?



Salt-lake city, Feb 2002

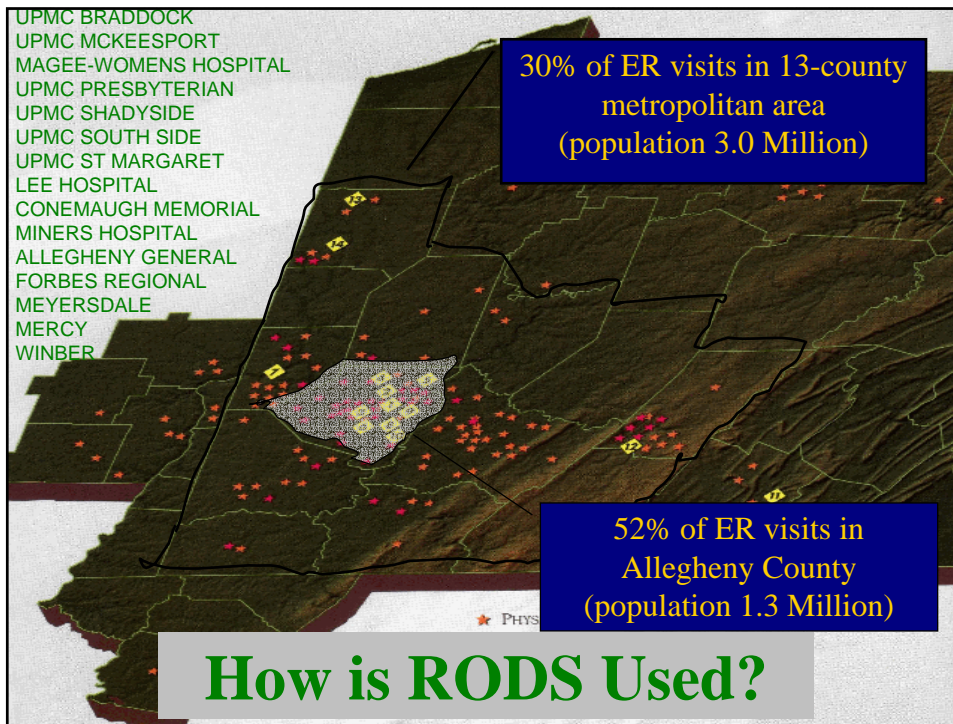
What if this happened in 2002?



Salt-lake city, Feb 2002

For a realistic attack, during peak period of importance, each extra hour of early detection could save 2000 lives.

[Tsui et al., 2000]



## When an alarm sounds...

RODS - Main Interface - Microsoft Internet Explorer

Address: <http://ultra.cbmi.upmc.edu/~sue/rods/main.htm>

Main EpiPlot MapPlot Advise Report Help About Logout

Electronic Medical Record Review

Search Results from Period=2 County=All Syndrome = Diarrhea

2001-05-03 - 35 year old F cc: DIARRHEA, BLO. PAIN - DEMO PATIENT
2002-02-02 - 25 year old F cc: NAUSEA AND VOMITING
2002-02-02 - 64 year old F cc: DIARRHEA
2002-02-02 - 68 year old F cc: NAUSEA/VOMITING
2002-02-02 - 74 year old F cc: NV
2002-02-02 - 31 year old F cc: NAUSEA/VOMITING
2002-02-02 - 18 year old M cc: VOMITING
2002-02-02 - 44 year old F cc: DIARRHEA X 2DAYS, NAUSEA, WEAKNESS
2002-02-02 - 91 year old F cc: NV
2002-02-02 - 87 year old M cc: DIARRHEA NOT EATING CONFUSED PER WIFE

Request Medical Record

EpiPlot View Controls

Period: last 2 days Start: Feb 2 2001 End: 2001 County: All Counties

Syndrome / Culture Data: Diarrhea Syndromes Plot:  open in new window  hospital mode  Get Cases

About RODS and its creators  Internet

# When an alarm sounds...

The screenshot shows the 'RODS Main Interface' in Microsoft Internet Explorer. The address bar shows 'http://ultra.cbmc.upmc.edu/~jue/rods/main.htm'. The main content area is titled 'Electronic Medical Record Review' and displays search results for 'Period=2 County=All Syndrome = Diarrhea'. A list of patient records is shown, with the first entry selected: '2002-02-02 - 68 year old F cc: NAUSEA/VOMITING'. An 'Enter Network Password' dialog box is overlaid on the interface, prompting the user to enter their username and password. The dialog box contains fields for 'Site' (mars.upmc.edu), 'Realn' (Patient Record Query), 'User Name' (myname), and 'Password' (masked with asterisks). There is also a checkbox for 'Save this password in your password list' and 'OK' and 'Cancel' buttons.

Copyright © 2002-2004, Andre... Biosurveillance: Slide 15

# How is RODS Used?

The screenshot shows the 'UPMC Version of the MARS System' in Microsoft Internet Explorer. The main content area displays a patient record for 'Report #: 02/02/02 Emergency Room'. The record includes fields for NAME, IDNO, SIBTYPE, DOCTOR, and PROCEDURE BY. The 'ATTENDING PHYSICIAN ADDENDUM' section contains a detailed clinical note: 'This patient that I examined here for complaints of diarrhea with the resident. Confirmed the history and physical with the resident and examined here by myself. She has got loose stools for a couple days. Her husband has had a similar illness and it is not improving. She feels she is able to keep up with p.o. liquids has got poor appetite. No nausea. No vomiting. REVIEW OF SYSTEMS: Otherwise negative for so. PHYSICAL EXAMINATION: I find the woman's temp at 36.6, initial triage pulse is 104, when I examine her is 88. Neck supple. Bilateral breath sounds clear. Abdomen is a bit obese. Nontender. No focal tenderness, guarding, or rebound. Back nontender. Rectal is per the resident and did not suggest blood. She is neurologically intact. Skin is warm and dry. IMPRESSION: Multiple loose stools consistent with a diarrheal illness. No vomiting. She is taking p.o. She does not appear clinically dehydrated. Recommended that we check a potassium given her use of Lasix and this was ordered. Also, gave her some fluids while we were waiting. She was given some symptomatic treatment which will continue as an outpatient. Unfortunately due to a lab error, the potassium was hemolyzed and the patient did not wish to wait for a redraw potassium. She was satisfied with her care at that point and wished to follow up with her own PCP. She is, therefore, discharged in good condition with diagnosis of diarrhea.'

The 'EMR Navigation Windows' on the right side of the screen shows a 'Category Selection' list with the following items: HP History/Physical, DS Discharge Summary, ER Emergency Room, RAD Radiology, SP Surgical Pathology, PCN Progress Note, LEFT Letter, LAB Laboratory, LABS Spread Sheets, and MICRO Microbiology. Below this is an 'ER Reports' list showing a series of reports with dates ranging from 02/02/02 to 11/28/94.

Copyright © 2002-2004, Andre... ce: Slide 16




## The task

How strange is everything that's happened in the last  $n$  hours?

## The task

How strange is everything that's happened in the last  $n$  hours?



Should we launch active data collection?

## The task

How strange is everything that's happened in the last  $n$  hours?

How strange is it, under the hypothesis that we have the following population of diseases with the following characteristics?

Should we launch active data collection?

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 19

## The task

How strange is everything that's happened in the last  $n$  hours?

How strange is it, under the hypothesis that we have the following population of diseases with the following characteristics?

Should we launch active data collection?

What is the chance that disease X is currently in the population?

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 20

## The task

How strange is everything that's happened in the last  $n$  hours?

Can we find a pattern in the strange stuff?

Should we launch active data collection?

How strange is it, under the hypothesis that we have the following population of diseases with the following characteristics?

What is the chance that disease X is currently in the population?

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 21

## WSARE v2.0

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

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 22

# 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	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	Downtown	781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	...
t3q15	6/2/2	14:15	Riverside	717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	...
t5hh5	6/2/2	14:15	Smithfield	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	Downtown	781	Fever	M	20s	NE	15217	A5	NW	15213	B8	2%	70R	...
t3q15	6/2/2	14:15	Riverside	717	Respiratory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	...
t5hh5	6/2/2	14:15	Smithfield	622	Respiratory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	...
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 25

## 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 ...

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 26



## 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 ...

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 27

## 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 ...

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 28



# Example

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

35.8% ( 48/134) of today's cases have  $30 \leq \text{age} < 40$

17.0% ( 45/265) of other cases have  $30 \leq \text{age} < 40$

# Example

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

FISHER\_PVALUE = 0.000051

35.8% ( 48/134) of today's cases have  $30 \leq \text{age} < 40$

17.0% ( 45/265) of other cases have  $30 \leq \text{age} < 40$

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

**Overfitting Alert!**

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 33

## 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

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 34

## 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?

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 35

## 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.

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 36



# 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	...
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 37

# 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	...
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 38

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

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 39

## Which of the 500 days have Irregularities?

- WARNING: Yet another overfitting opportunity.
- If we took 500 days of randomly generated data, then about 5 days would have a p-value below 0.01
- This can be solved with...
  - A Bonferroni correction
  - The FDR (False Discovery Rate) method [Benjamini & Hochberg, 1995, J. R. Stat Soc, 57 289]

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 40

## 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
```

Copyright © 2002-2004, Andrew Moore

## Sanity check

- What happens if we generate a fake database in which we know that there can be no relation between date and case features?
- This can be achieved by shuffling all the dates in the database.
- The days detected by FDR are then...

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 42

## Survived Randomized

- No days reported as abnormal

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 43

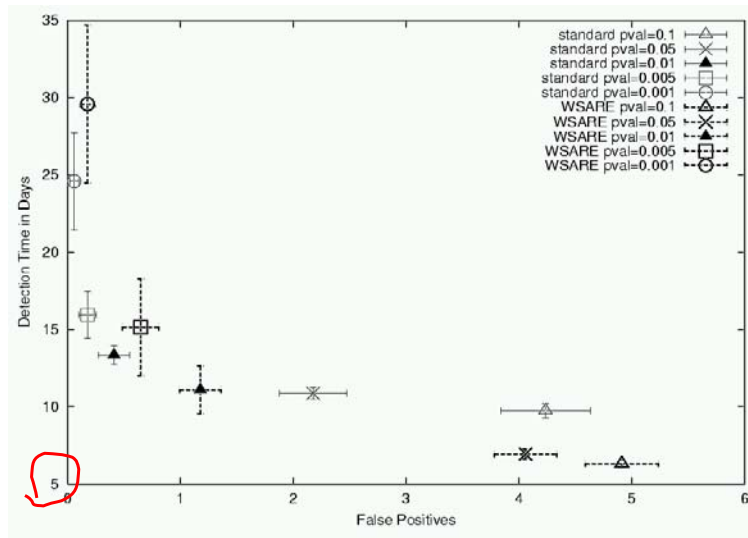
## Simulation-based validation

- We need to understand sensitivity and specificity.
- Have developed a very fast simulator in which people go to work, go out for food, live in families.
- Background of mild diseases: some food-borne, some occupational, some contagious etc.
- A nasty disease is introduced at some point.
  - Can WSARE detect it?
  - How does WSARE compare to a standard sensible epidemiological alternative?

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 44

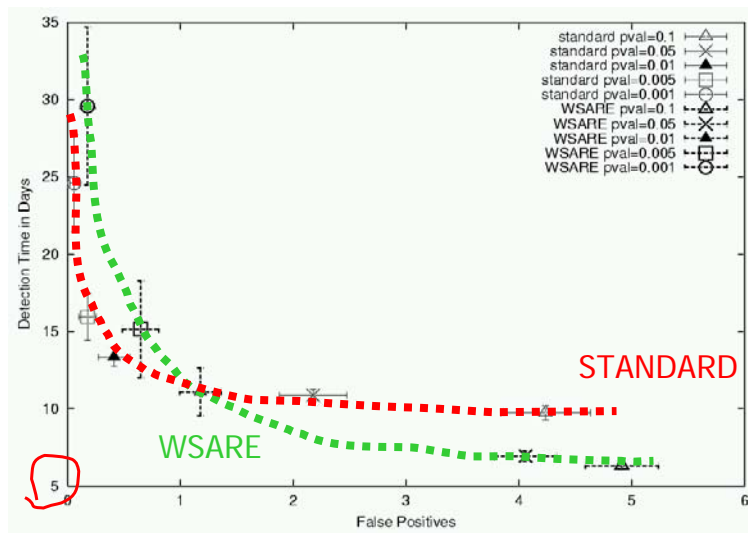
## WSARE Simulation Sensitivity vs Specificity on GRIDSIM



Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 45

## WSARE Simulation Sensitivity vs Specificity on GRIDSIM



Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 46



## WSARE 3.0

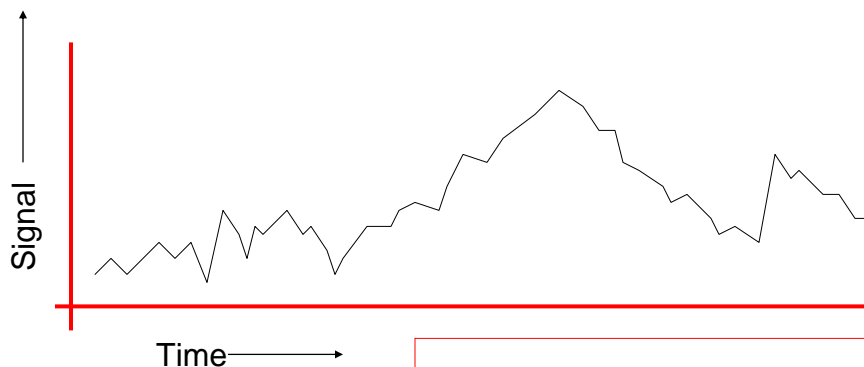
- "Taking into account recent flu levels..."
- "Taking into account that today is a public holiday..."
- "Taking into account that this is Spring..."
- "Taking into account recent heatwave..."
- "Taking into account that there's a known natural Food-borne outbreak in progress..."

Bonus: More efficient use of historical data

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 47

## Conditioning on observed environment: Well understood for Univariate Time Series



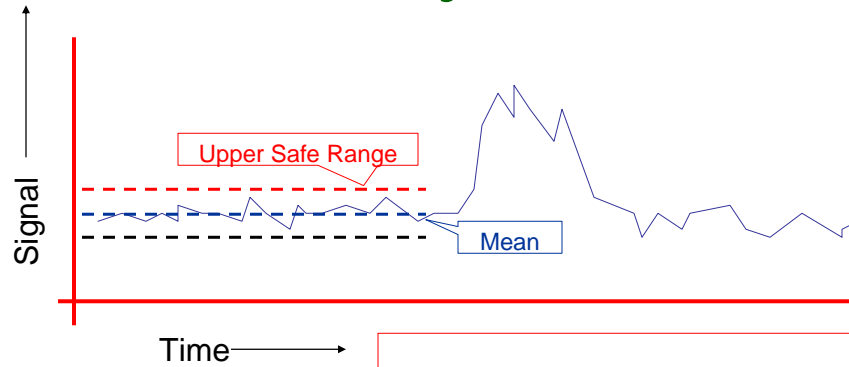
### Example Signals:

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

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 48

## An easy case

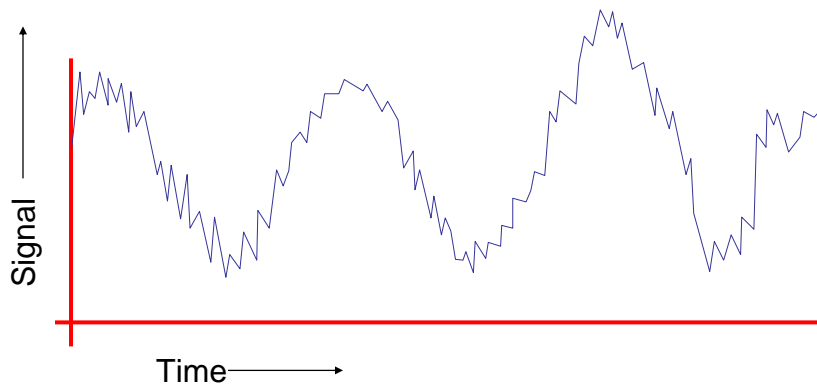


Dealt with by Statistical Quality Control  
Record the mean and standard deviation up to the current time.  
Signal an alarm if we go outside 3 sigmas

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 49

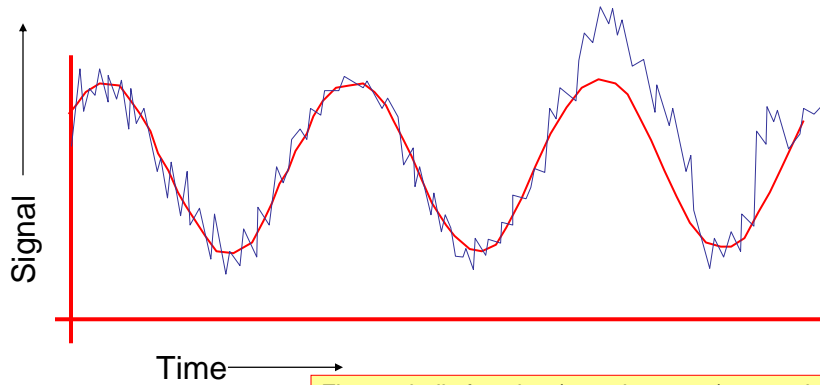
## Conditioning on Seasonal Effects



Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 50

## Seasonal Effects

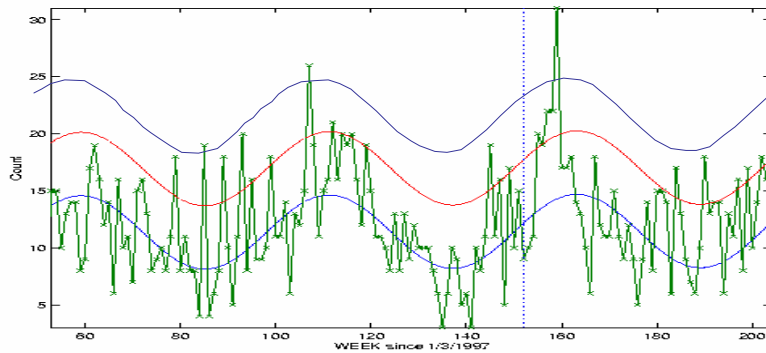


Fit a periodic function (e.g. sine wave) to previous data. Predict today's signal and 3-sigma confidence intervals. Signal an alarm if we're off.  
Reduces False alarms from Natural outbreaks.  
Different times of year deserve different thresholds.

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 51

## Example [Tsui et. Al]



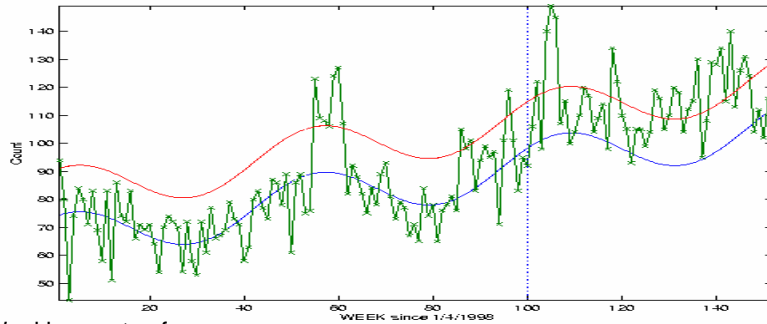
Weekly counts of P&I from week 1/98 to 48/00

From: "Value of ICD-9-Coded Chief Complaints for Detection of Epidemics", Fu-Chiang Tsui, Michael M. Wagner, Virginia Dato, Chung-Chou Ho Chang, AMIA 2000

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 52

## Seasonal Effects with Long-Term Trend



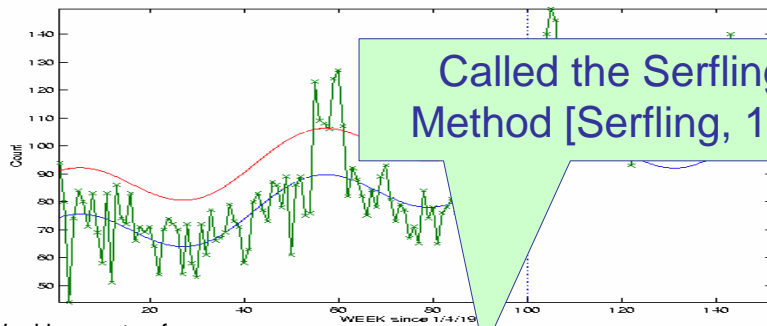
Weekly counts of IS from week 1/98 to 48/00.

From: "Value of ICD-9-Coded Chief Complaints for Detection of Epidemics", Fu-Chiang Tsui, Michael M. Wagner, Virginia Dato, Chung-Chou Ho Chang, AMIA 2000

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 53

## Seasonal Effects with Long-Term Trend



Weekly counts of IS from week 1/98 to 48/00.

From: "Value of ICD-9-Coded Chief Complaints for Detection of Epidemics", Fu-Chiang Tsui, Michael M. Wagner, Virginia Dato, Chung-Chou Ho Chang, AMIA 2000

Copyright © 2002-2004, Andrew Moore

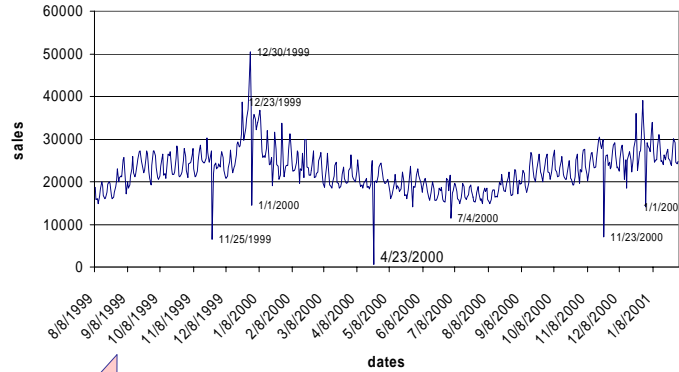
Fit a periodic function (e.g. sine wave) plus a linear trend:

$$E[\text{Signal}] = a + bt + c \sin(d + t/365)$$

Good if there's a long term trend in the disease or the population.

Biosurveillance: Slide 54

## Day-of-week effects

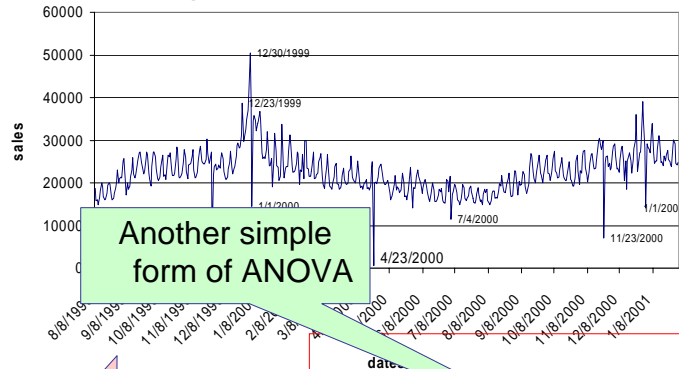


From: "Using Grocery Sales Data for the Detection of Bio-Terrorist Attacks", Goldenberg, Shmueli and Caruana, 2002

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 55

## Day-of-week effects



Another simple form of ANOVA

From: "Using Grocery Sales Data for the Detection of Bio-Terrorist Attacks", Goldenberg, Shmueli and Caruana, 2002

Copyright © 2002-2004, Andrew Moore

Fit a day-of-week component

$$E[\text{Signal}] = a + \delta_{\text{day}}$$

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

Biosurveillance: Slide 56

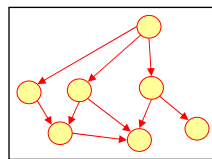
## Analysis of variance

- **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?

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 57

## 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"

Copyright © 2002-2004, Andrew Moore

Biosurveillance: Slide 58

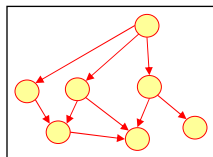
# WSARE 3.0



Copyright © 2002-2004, Andrew Moore

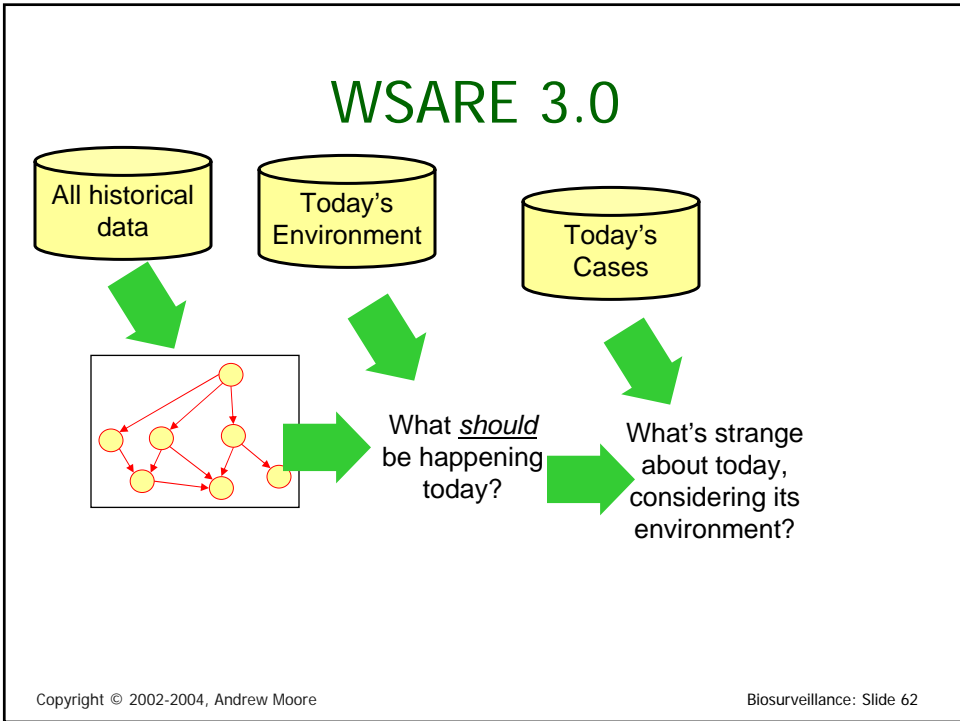
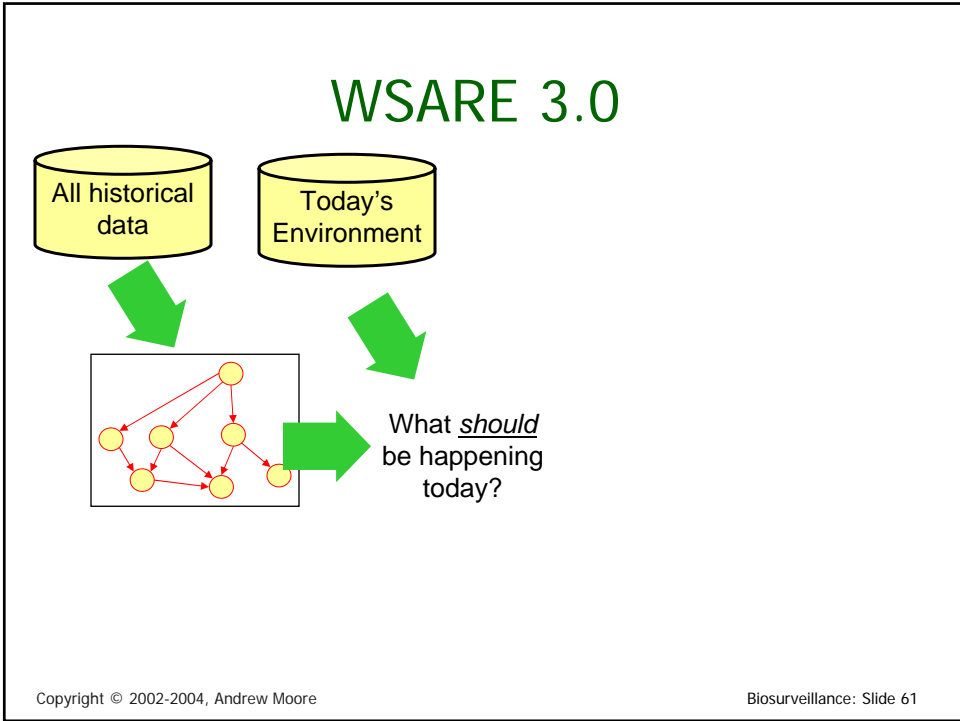
Biosurveillance: Slide 59

# WSARE 3.0

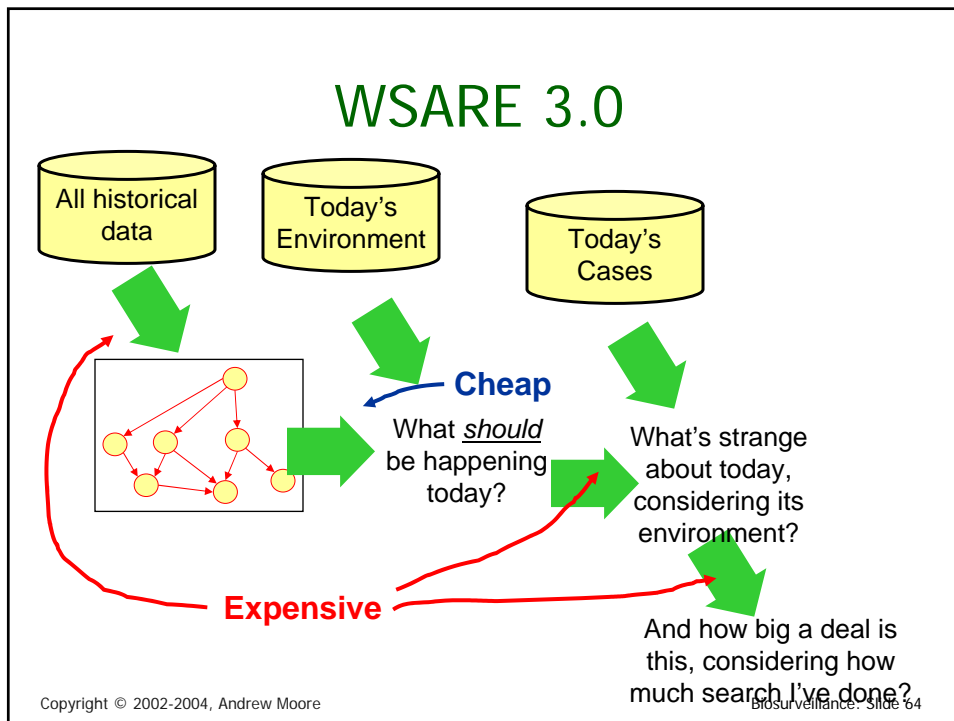
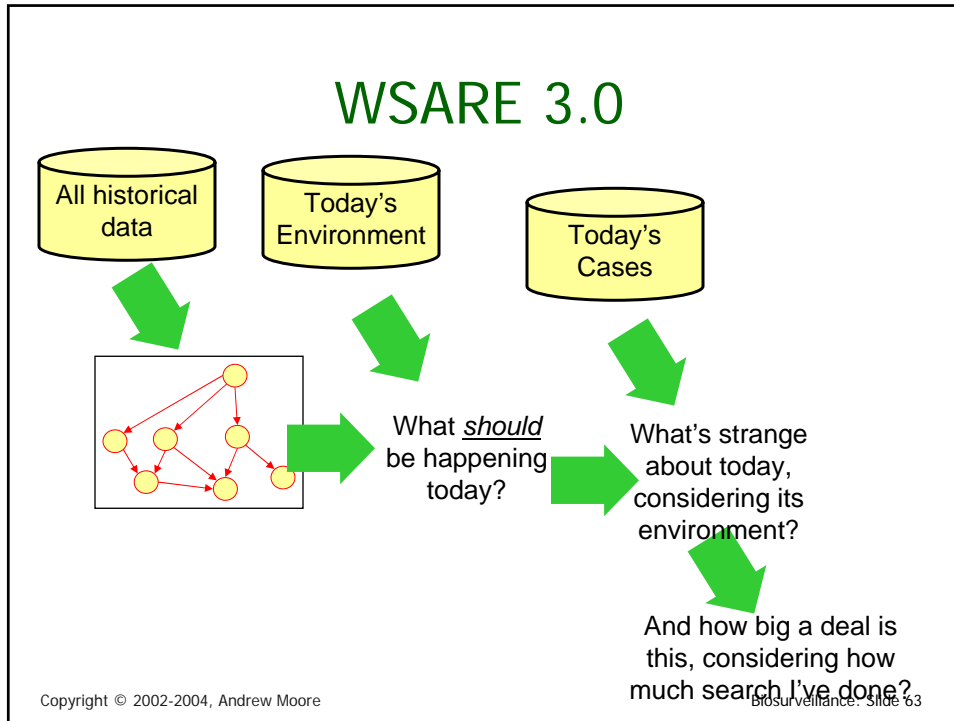


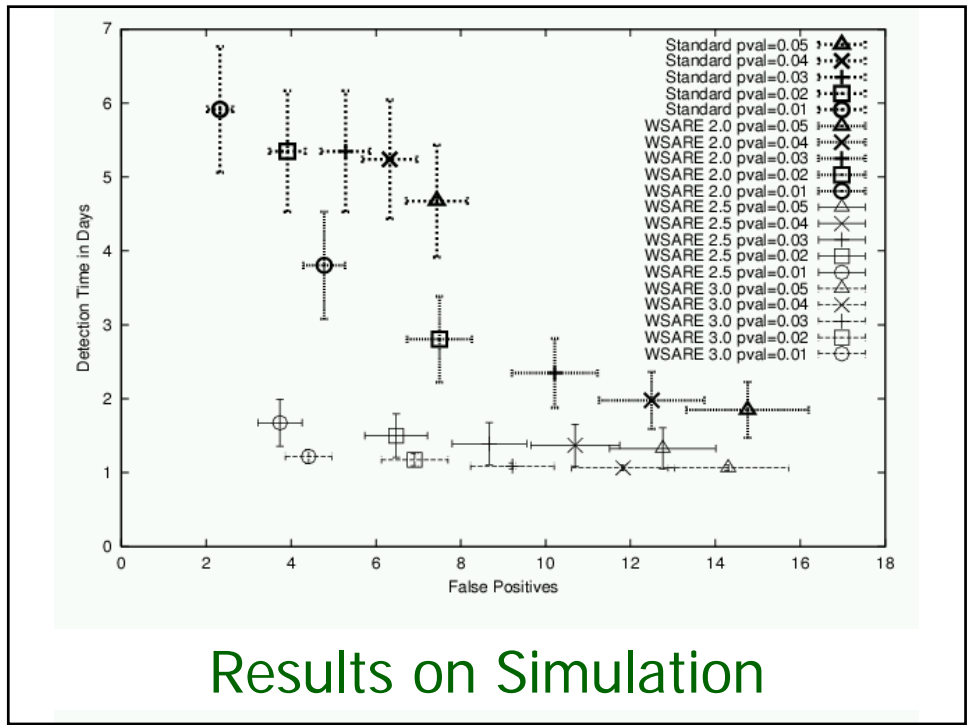
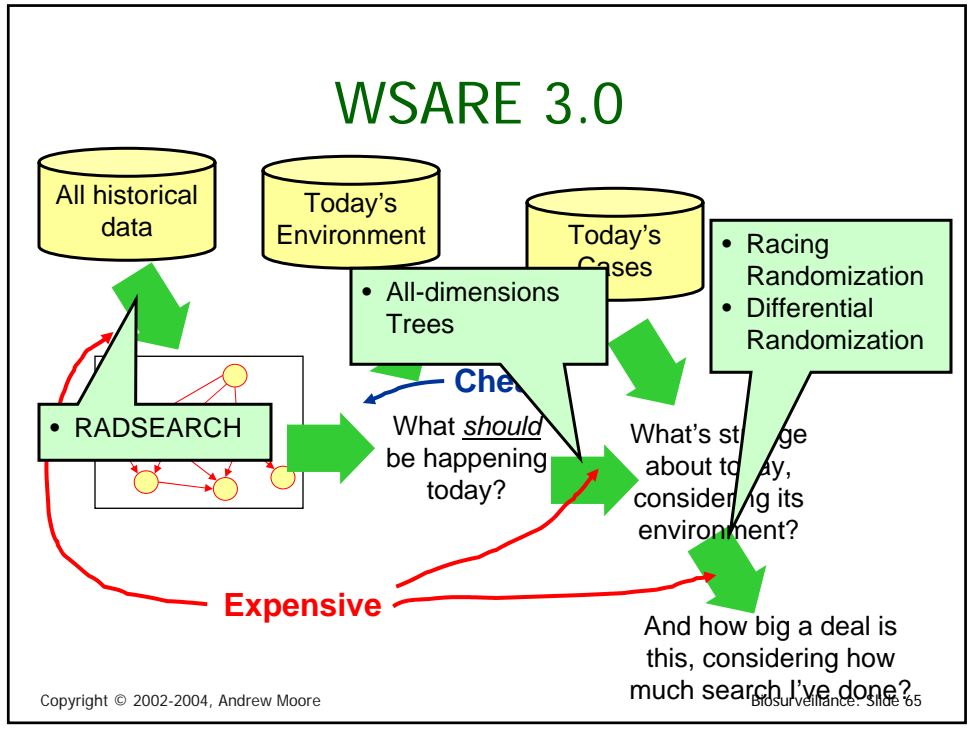
Copyright © 2002-2004, Andrew Moore

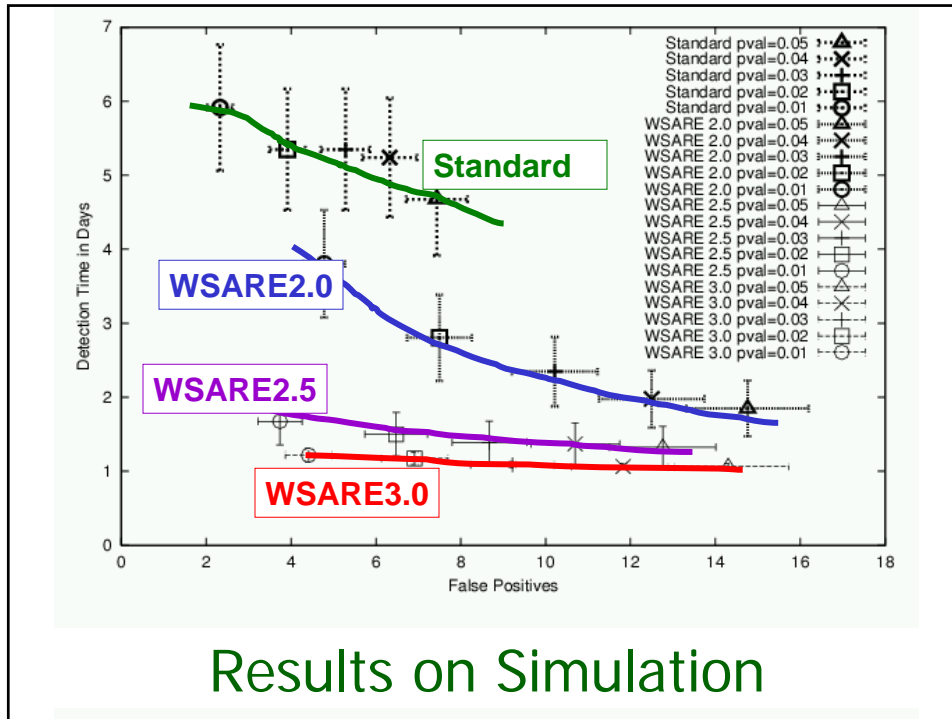
Biosurveillance: Slide 60











## 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 tricky!

- 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

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

## 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 tricky!
- 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