

What's Strange About Recent Events (WSARE) v3.0: Adjusting for a Changing Baseline

Weng-Keen Wong (Carnegie Mellon University)

Andrew Moore (Carnegie Mellon University)

Gregory Cooper (University of Pittsburgh)

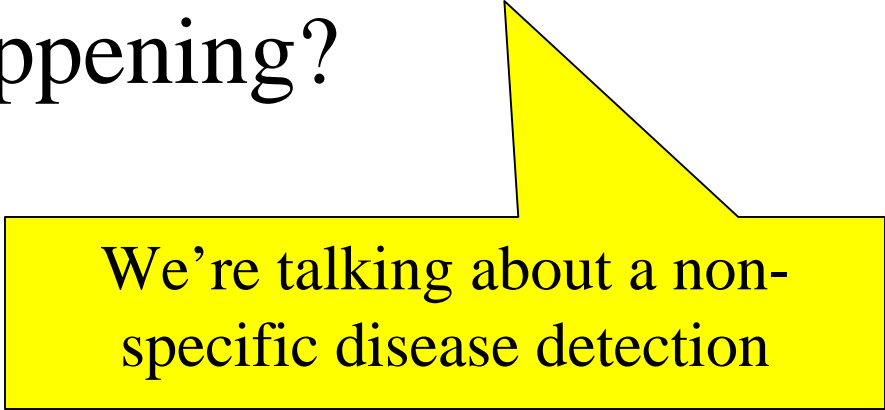
Michael Wagner (University of Pittsburgh)

The Problem

From this data, can we detect if a disease outbreak is happening?

The Problem

From this data, can we detect if a disease outbreak is happening?



We're talking about a non-specific disease detection

The Problem

From this data, can we detect if a disease outbreak is happening? How early can we detect it?

The Problem

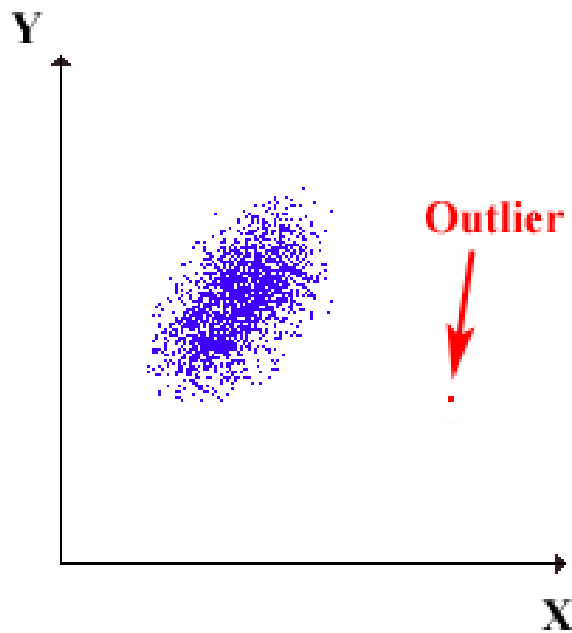
From this data, can we detect if a disease outbreak is happening? How early can we detect it?

The question we're really asking:
In the last n hours, has anything strange happened?

Traditional Approaches

What about using traditional anomaly detection?

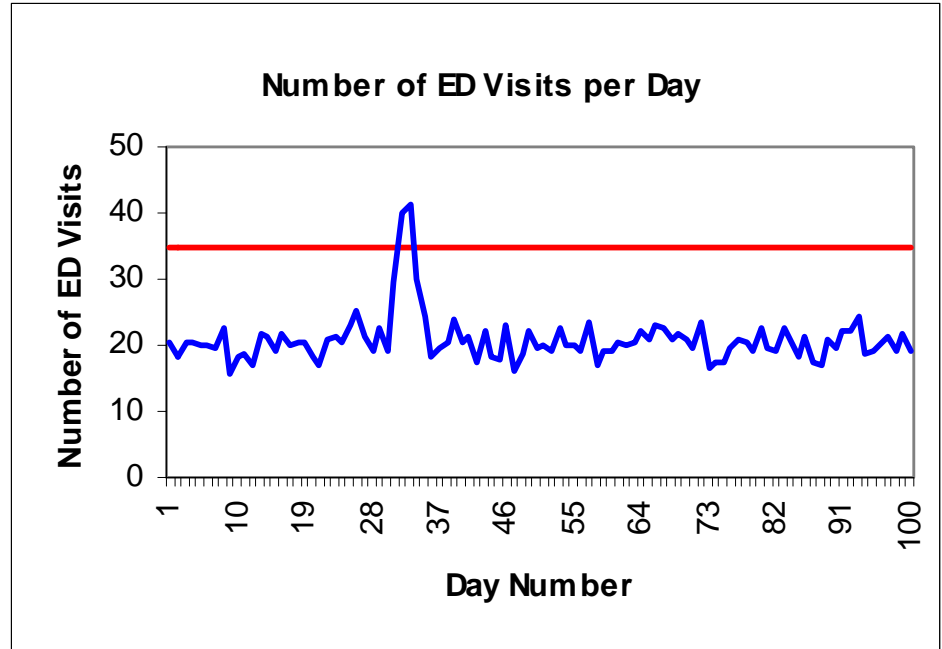
- Typically assume data is generated by a model
- Finds individual data points that have low probability with respect to this model
- These outliers have rare attributes or combinations of attributes
- Need to identify anomalous *patterns* not isolated data points



Traditional Approaches

What about monitoring aggregate daily counts of certain attributes?

- We've now turned multivariate data into univariate data
- Lots of algorithms have been developed for monitoring univariate data:
 - Time series algorithms
 - Regression techniques
 - Statistical Quality Control methods
- Need to know *apriori* which attributes to form daily aggregates for!



Traditional Approaches

What if we don't know what attributes to monitor?

Traditional Approaches

What if we don't know what attributes to monitor?

What if we want to exploit the spatial, temporal and/or demographic characteristics of the epidemic to detect the outbreak as early as possible?

Traditional Approaches

We need to build a univariate detector to monitor each interesting combination of attributes:

Diarrhea cases among children	Number of cases involving people working in southern part of the city
Respiratory syndrome cases among females	Number of cases involving teenage girls living in the western part of the city
Viral syndrome cases involving senior citizens from eastern part of city	Botulinic syndrome cases
Number of children from downtown hospital	And so on...

Traditional Approaches

We need to build a univariate detector to monitor each interesting combination of attributes:

Diarrhea cases among children	Number of cases involving people working in southern part of the city
Respiratory syndrome involving senior citizens from eastern part of city	Botulinic syndrome cases
Number of children from downtown hospital	And so on...

You'll need hundreds of univariate detectors!
We would like to identify the groups with the strangest behavior in recent events.

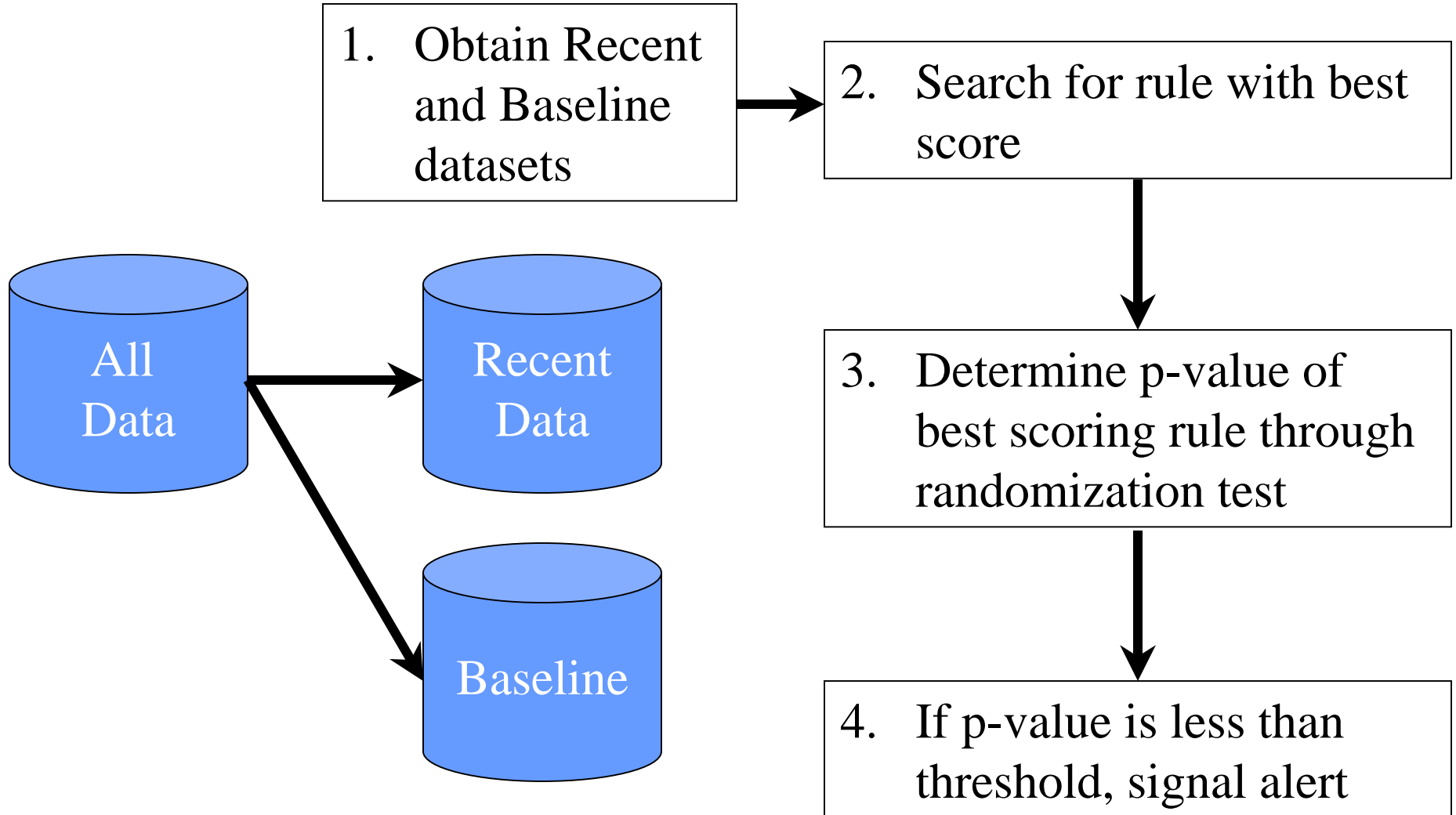
Our Approach

- We use Rule-Based Anomaly Pattern Detection
- Association rules used to characterize anomalous patterns. For example, a two-component rule would be:

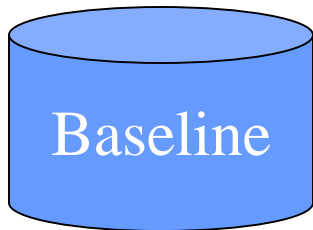
Gender = Male AND $40 \leq \text{Age} < 50$

- Related work:
 - Market basket analysis [Agrawal et. al, Brin et. al.]
 - Contrast sets [Bay and Pazzani]
 - Spatial Scan Statistic [Kulldorff]
 - Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance [Brossette et. al.]

WSARE v2.0 Overview



Obtaining the Baseline (WSARE v2.0)



Assumed to capture non-epidemic behavior. We use raw historical data.

June 2002

S	M	Tu	W	Th	F	S
			5	6	7	8
2	3	4	12	13	14	15
9	10	11	19	20	21	22
16	17	18	26	27	28	29
23	24	25				
30						

Compare against "usual" behaviour

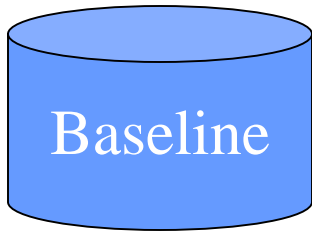
July 2002

S	M	Tu	W	Th	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

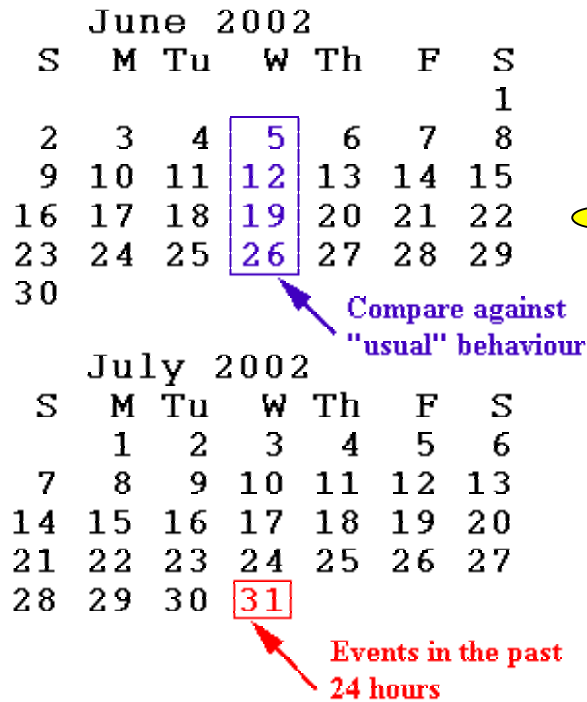
Events in the past 24 hours

Here we use data from 35,42, 49 and 56 days ago

Obtaining the Baseline (WSARE v2.0)

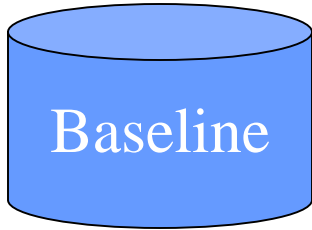


Assumed to capture non-epidemic behavior. We use raw historical data.



What if data from 7, 14, 21 and 28 days ago is better?

Obtaining the Baseline (WSARE v2.0)



Assumed to capture non-epidemic behavior. We use raw historical data.

June 2002

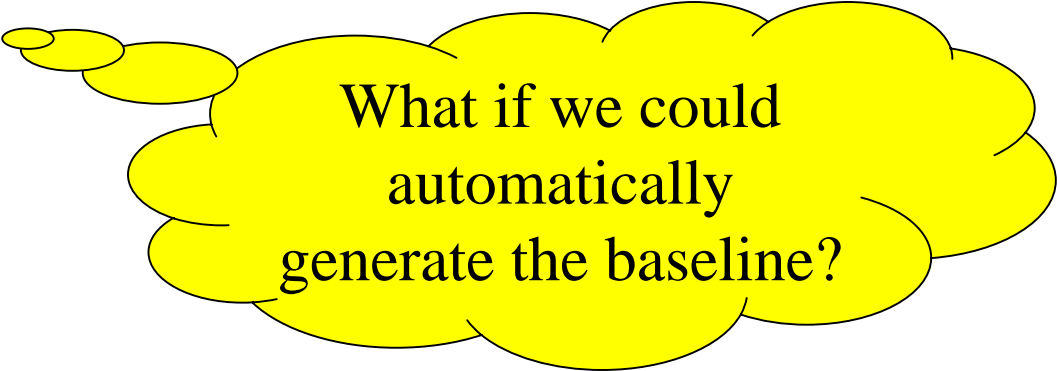
S	M	Tu	W	Th	F	S
			5	6	7	8
2	3	4	12	13	14	15
9	10	11	19	20	21	22
16	17	18	26	27	28	29
23	24	25				
30						

Compare against "usual" behaviour

July 2002

S	M	Tu	W	Th	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

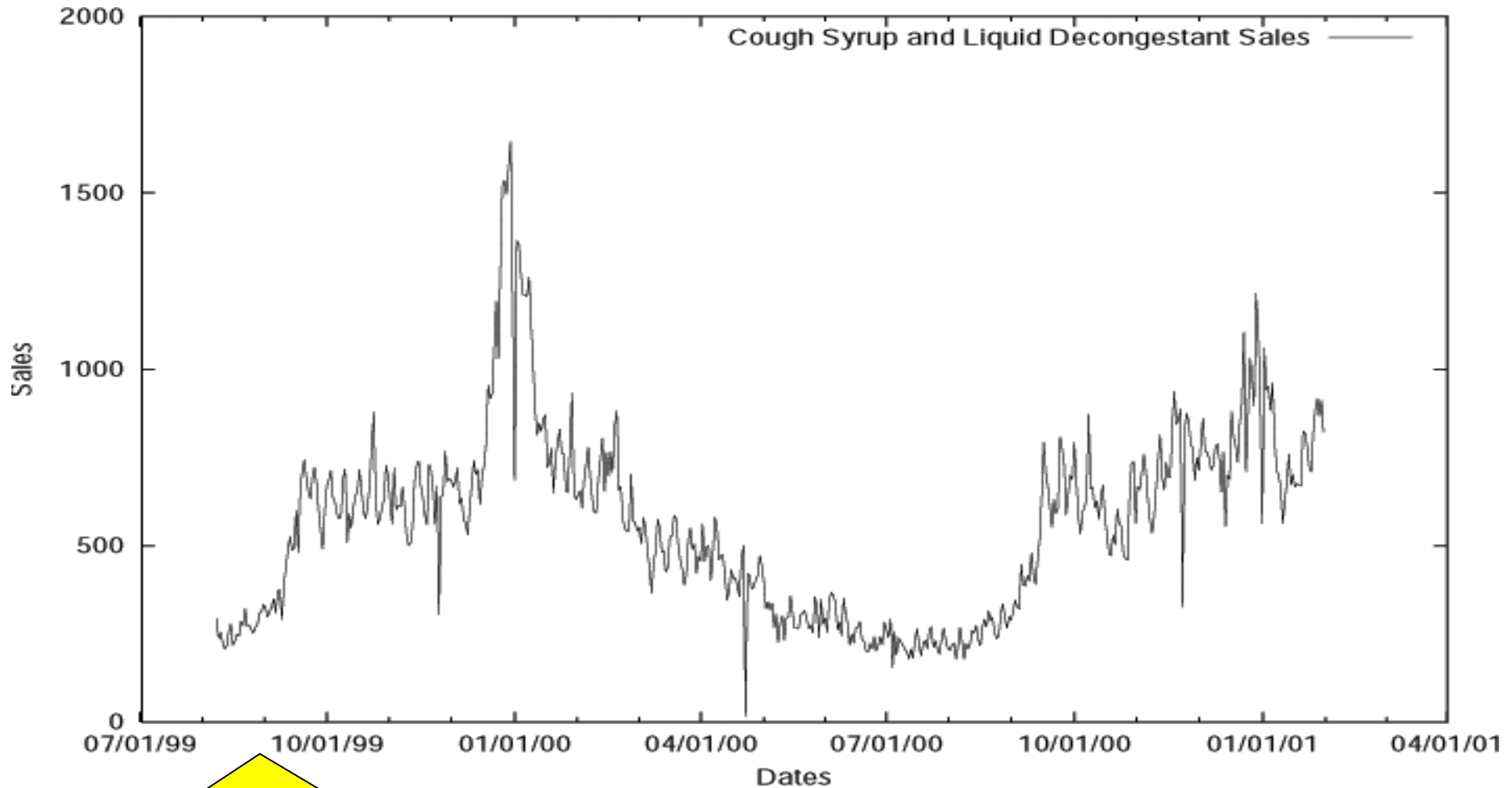
Events in the past 24 hours



Temporal Trends

- But health care data has many different trends due to
 - Seasonal effects in temperature and weather
 - Day of Week effects
 - Holidays
 - Etc.
- Allowing the baseline to be affected by these trends may dramatically alter the detection time and false positives of the detection algorithm

Temporal Trends




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

WSARE v3.0

Generate the baseline...

- “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...”

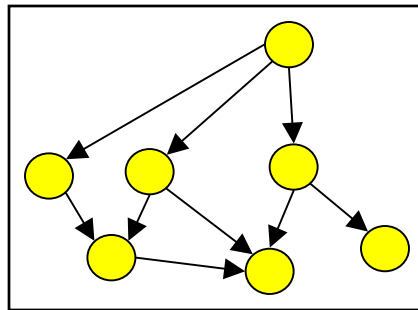


Bonus: More
efficient use of
historical data

Idea: Bayesian Networks

“Patients from West Park Hospital are less likely to be young”

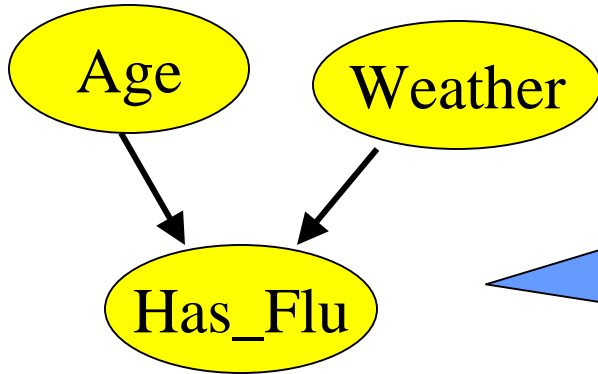
“On Cold Tuesday Mornings the folks coming in from the North part of the city are more likely to have respiratory problems”



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

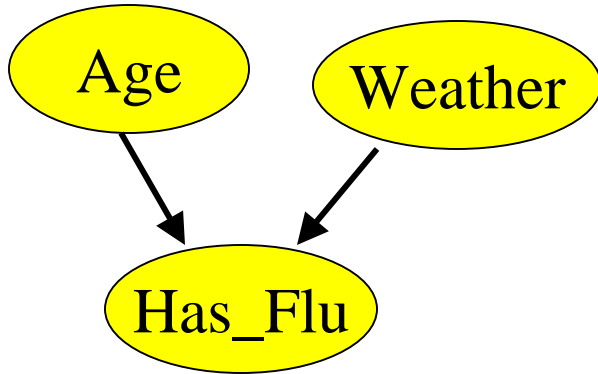
“On the day after a major holiday, expect a boost in the morning followed by a lull in the afternoon”

What is a Bayesian Network?



The arrows say something about the conditional independence structure of the attributes. They do not necessarily say anything about causality.

What is a Bayesian Network?



Bayesian Network: A graphical model representing the joint probability distribution of a set of random variables

From the Bayesian Network above, we can get:

$P(\text{Age} = \text{Senior}, \text{Weather} = \text{Cold}, \text{Has_Flu} = \text{True})$

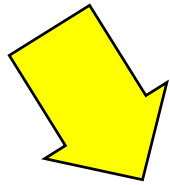
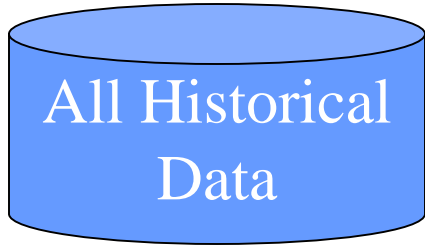
More importantly, we can get:

$P(\text{Has_Flu} = \text{True} \mid \text{Age} = \text{Senior}, \text{Weather} = \text{Cold})$

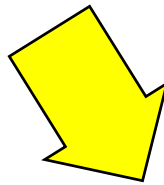
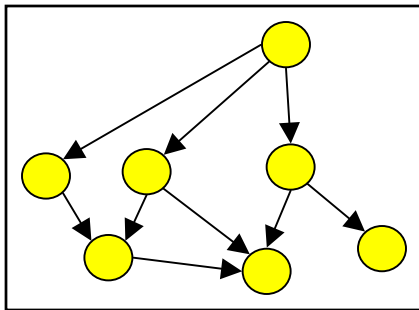
How do we come up with the Bayesian Network Structure?

1. By hand
2. By learning it from historical data
 - Lots of different algorithms for doing this
 - We use Optimal Reinsertion [Moore and Wong 2003]

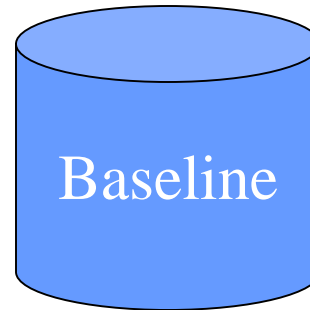
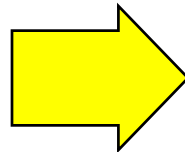
Obtaining Baseline Data



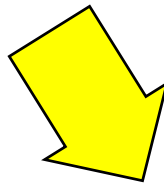
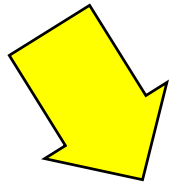
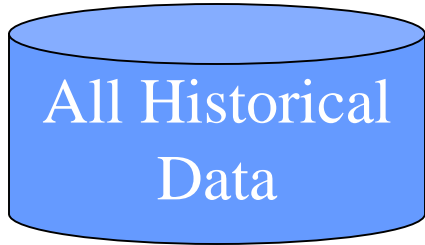
1. Learn Bayesian Network



2. Generate baseline given today's environment

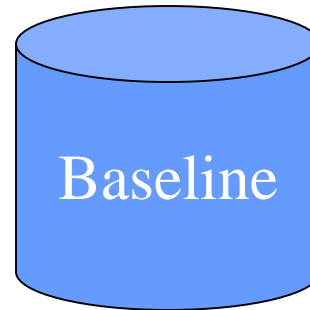
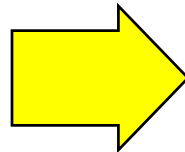
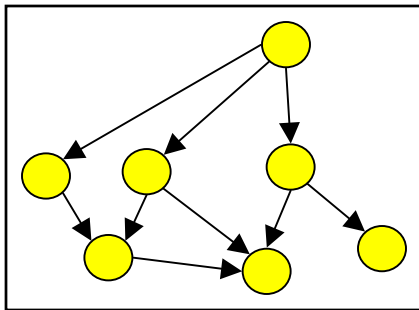


Obtaining Baseline Data



1. Learn Bayesian Network

2. Generate baseline given today's environment



What should be happening today given today's environment

Environmental Attributes

Divide the data into two types of attributes:

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

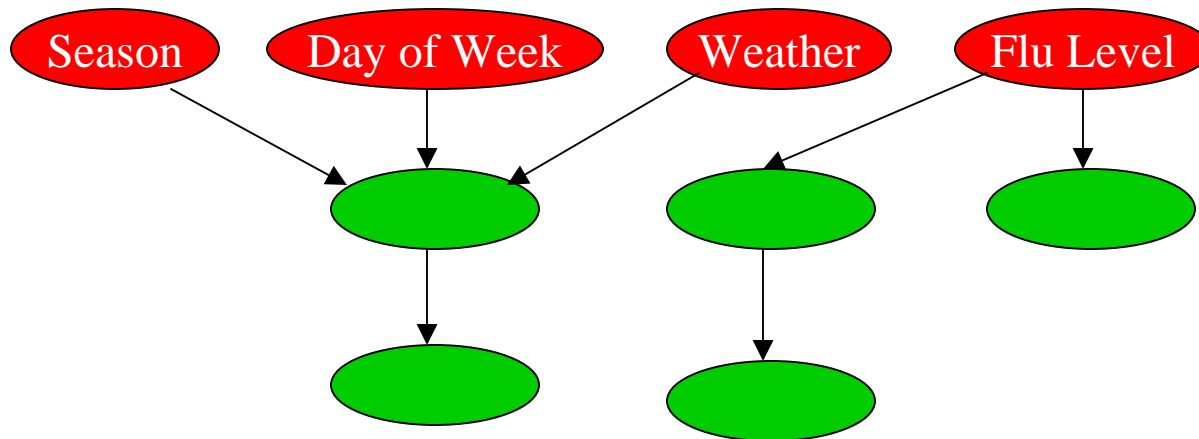
Environmental Attributes

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

Why?

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

Side Benefit: We can speed up the structure search by avoiding DAGs that assign parents to the environmental attributes

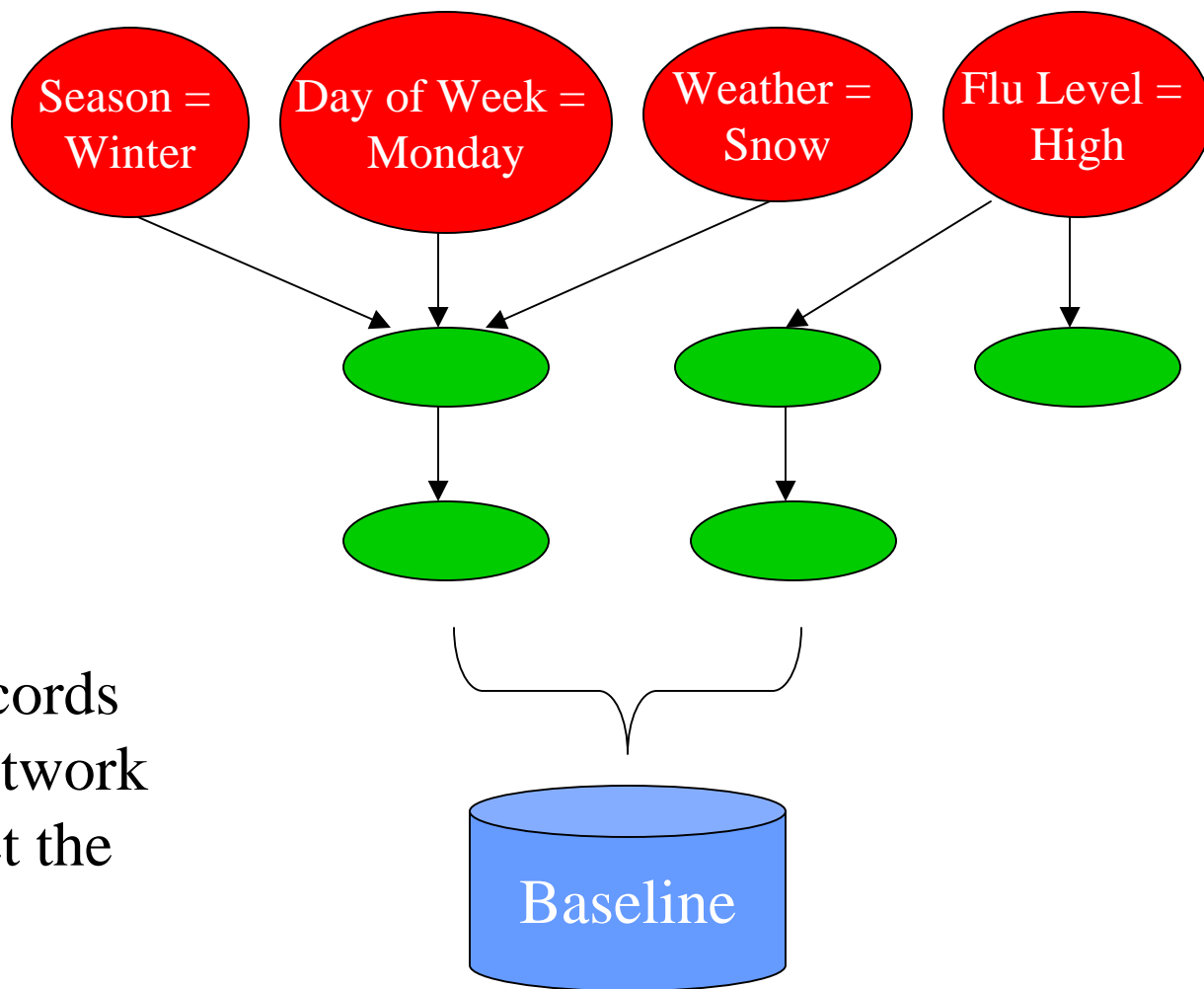


Step 2: Generate Baseline Given Today's Environment

Suppose we know the following for today:

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

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



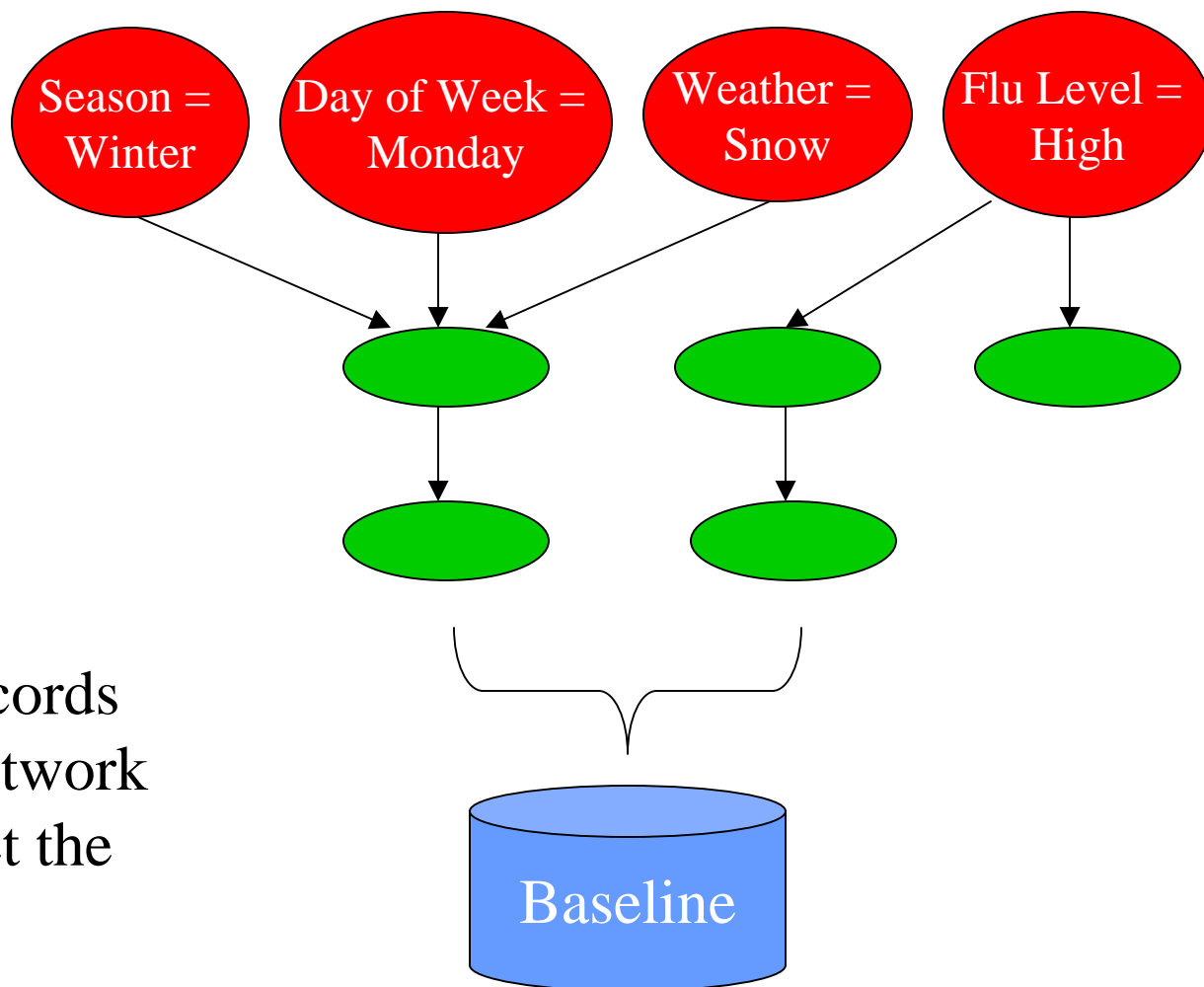
We sample 10000 records from the Bayesian network and make this data set the baseline

Step 2: Generate Baseline Given Today's Environment

Suppose we know the following for today:

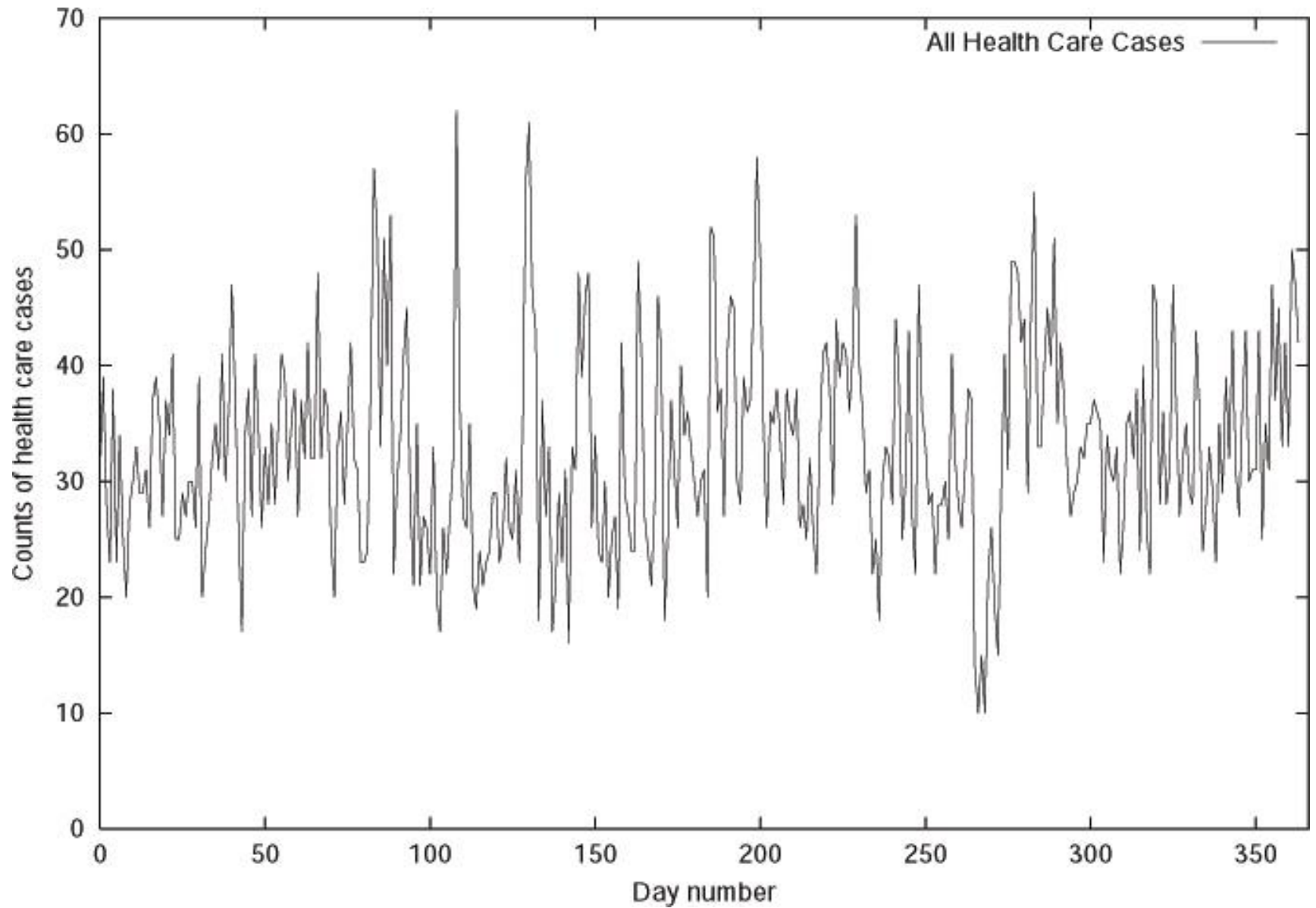
	Season	Day of Week	Weather	Flu Level
Today	Winter	Monday	Snow	High

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



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

Simulator

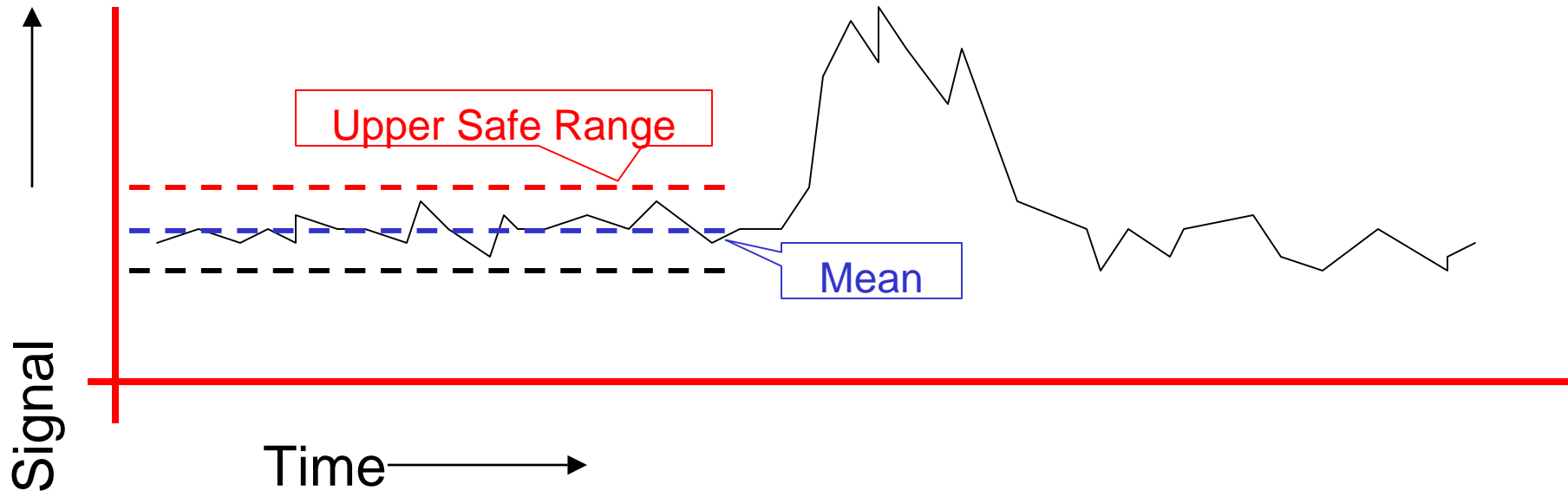


Simulation

- 100 different data sets (available on web)
- Each data set consisted of a two year period
- Anthrax release occurred at a random point during the second year
- Algorithms allowed to train on data from the current day back to the first day in the simulation
- Any alerts before actual anthrax release are considered a false positive
- Detection time calculated as first alert after anthrax release. If no alerts raised, cap detection time at 14 days

Other Algorithms used in Simulation

1. Control Chart



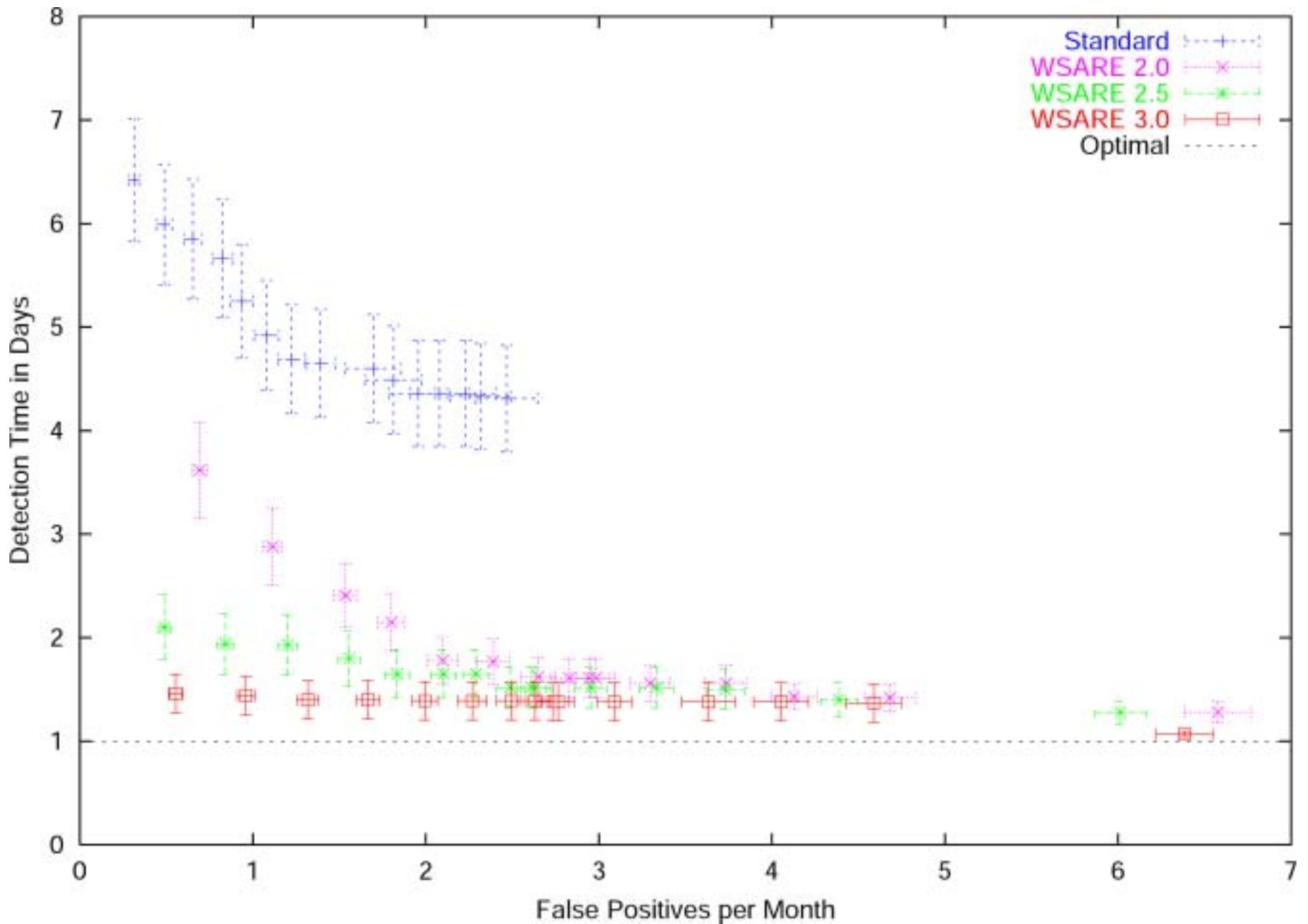
2. WSARE 2.0

Create baseline using historical data from 7, 14, 21 and 28 days ago

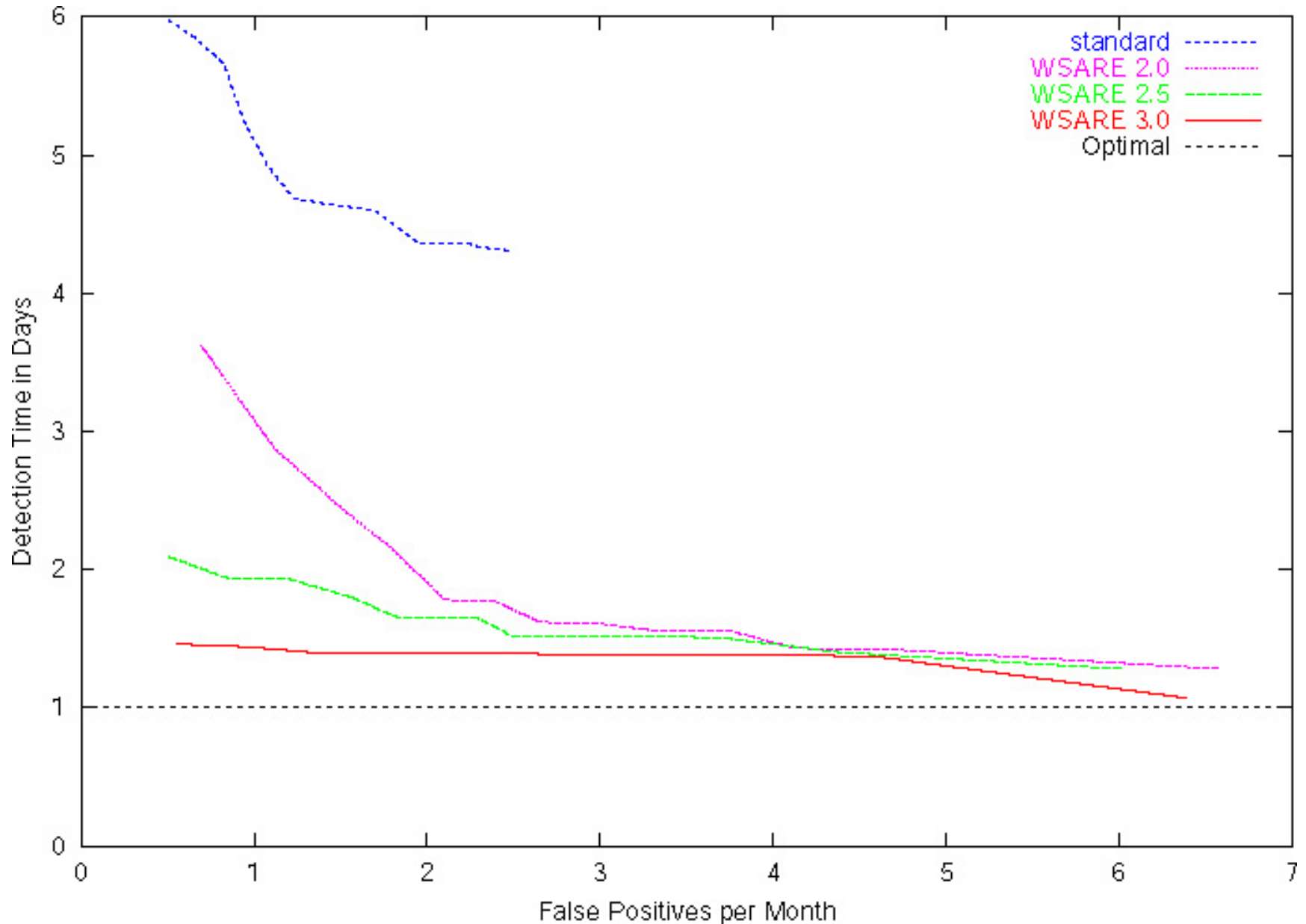
3. WSARE 2.5

Use all past data but condition on environmental attributes

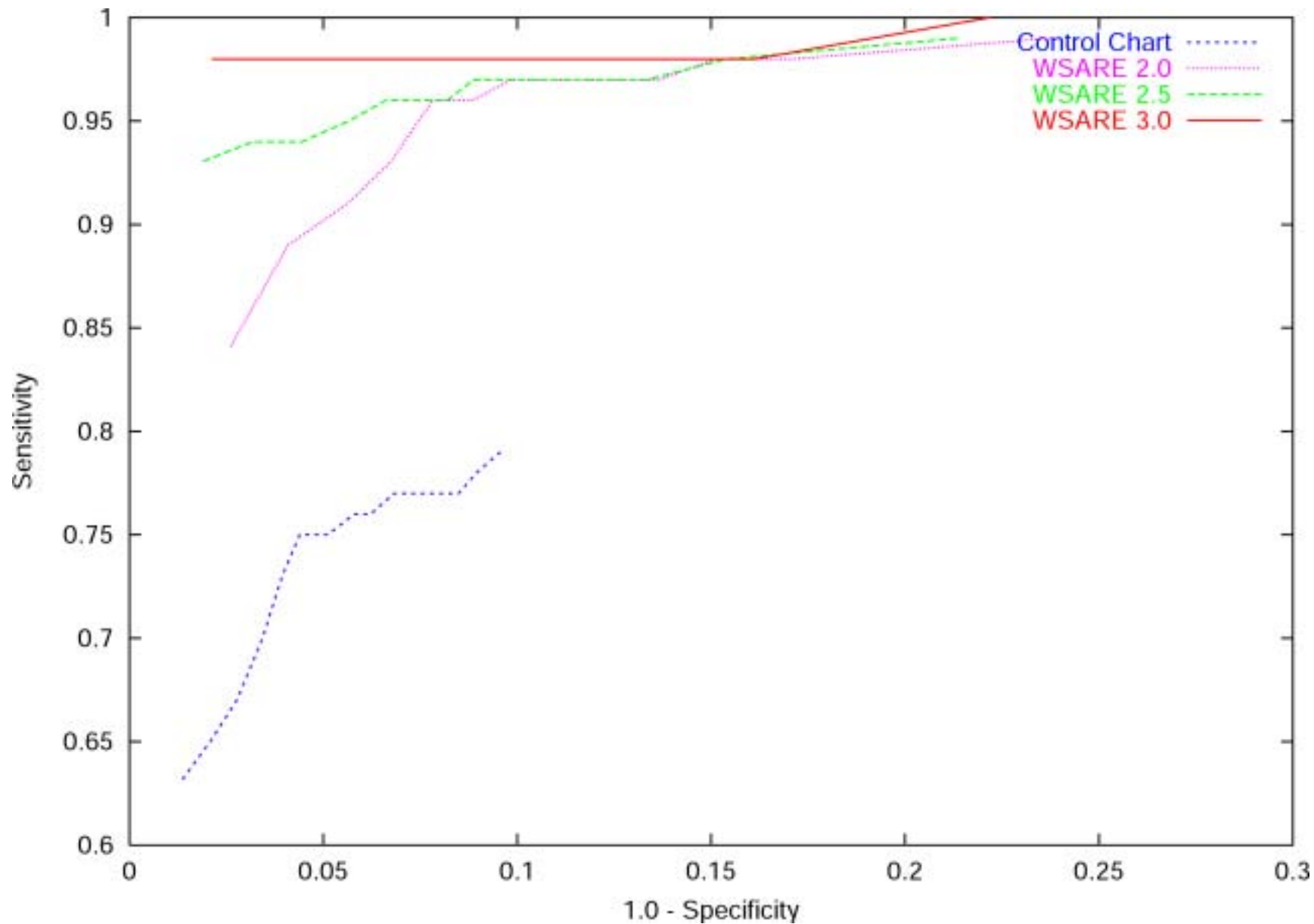
Results on Simulation



Results on Simulation



Results on Simulation



Results on Actual ED Data from 2001

1. Sat 2001-02-13: SCORE = -0.00000004 PVALUE = 0.00000000
14.80% (74/500) of today's cases have Viral Syndrome = True and Encephalitic Prodome = False
7.42% (742/10000) of baseline have Viral Syndrome = True and Encephalitic Syndrome = False
2. Sat 2001-03-13: SCORE = -0.00000464 PVALUE = 0.00000000
12.42% (58/467) of today's cases have Respiratory Syndrome = True
6.53% (653/10000) of baseline have Respiratory Syndrome = True
3. Wed 2001-06-30: SCORE = -0.00000013 PVALUE = 0.00000000
1.44% (9/625) of today's cases have $100 \leq \text{Age} < 110$
0.08% (8/10000) of baseline have $100 \leq \text{Age} < 110$
4. Sun 2001-08-08: SCORE = -0.00000007 PVALUE = 0.00000000
83.80% (481/574) of today's cases have Unknown Syndrome = False
74.29% (7430/10001) of baseline have Unknown Syndrome = False
5. Thu 2001-12-02: SCORE = -0.00000087 PVALUE = 0.00000000
14.71% (70/476) of today's cases have Viral Syndrome = True and Encephalitic Syndrome = False
7.89% (789/9999) of baseline have Viral Syndrome = True and Encephalitic Syndrome = False
6. Thu 2001-12-09: SCORE = -0.00000000 PVALUE = 0.00000000
8.58% (38/443) of today's cases have Hospital ID = 1 and Viral Syndrome = True
2.40% (240/10000) of baseline have Hospital ID = 1 and Viral Syndrome = True

Conclusion

- One approach to biosurveillance: one algorithm monitoring millions of signals derived from multivariate data
instead of
Hundreds of univariate detectors
- WSARE is best used as a general purpose safety net in combination with other detectors
- Modeling historical data with Bayesian Networks to allow conditioning on unique features of today
- Computationally intense unless we use clever algorithms

Conclusion

- WSARE 2.0 deployed during the past year
- WSARE 3.0 to be deployed online
- WSARE now being extended to additionally exploit over the counter medicine sales

For more information

References:

- Wong, W. K., Moore, A. W., Cooper, G., and Wagner, M. (2002). Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks. Proceedings of AAAI-02 (pp. 217-223). MIT Press.
- Wong, W. K., Moore, A. W., Cooper, G., and Wagner, M. (2003). Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. Proceedings of ICML 2003.
- Moore, A., and Wong, W. K. (2003). Optimal Reinsertion: A New Search Operator for Accelerated and More Accurate Bayesian Network Structure Learning. Proceedings of ICML 2003.

AUTON lab website: <http://www.autonlab.org/wsare>

Email: wkw@cs.cmu.edu