RODS and Multiple Data Streams

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
Andrew Moore	Professor	Computer Science, Carnegie Mellon	awm@cs.cmu.edu
Daniel Neill	Graduate Student	Computer Science, Carnegie Mellon	neill@cs.cmu.edu
Jeff Schneider	Research Professor	Computer Science, Carnegie Mellon	schneide@cs.cmu.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
Weng-Keen Wong	Graduate Student	Computer Science, Carnegie Mellon	wkw@cs.cmu.edu

RODS: http://www.health.pitt.edu/rods
Auton Lab: http://www.autonlab.org

New Biosurveillance Algorithms

An interesting feature of large Biosurveillance Programs:

Multiple, rich, new streams of data

New Biosurveillance Algorithms

A unique feature of the BioAlirt Program:

Multiple, rich, new streams of data

Multiattribute

New Biosurveillance Algorithms

A unique feature of the BioAlirt Program:

Multiple, rich, new streams of data

Multiattribute

Question: How do we use all this information?

How can we "plug in" new streams?

How can we exploit multiattribute form?

New Biosurveillance Algorithms

Specific Detectors

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

BARD: Airborne Attack Detection
[Hogan, Cooper]

General Detectors

What's Strange about Recent Events

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]

Possible Future Connection

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

What's Strange about Recent Events

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]

Possible Future Connection

Question: How do we use all this information?

How can we "plug in" new streams?

How can we exploit multiattribute form?

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	Respirat ory	M	60s	NE	15222	J3	NE	15222	J3	2%	70R	...
t5hh5	6/2/2	14:15	Smith-field	622	Respirat ory	F	80s	SE	15210	K9	SE	15210	K9	2%	70R	...
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

WSARE v2.0

- Inputs:
- Outputs:

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

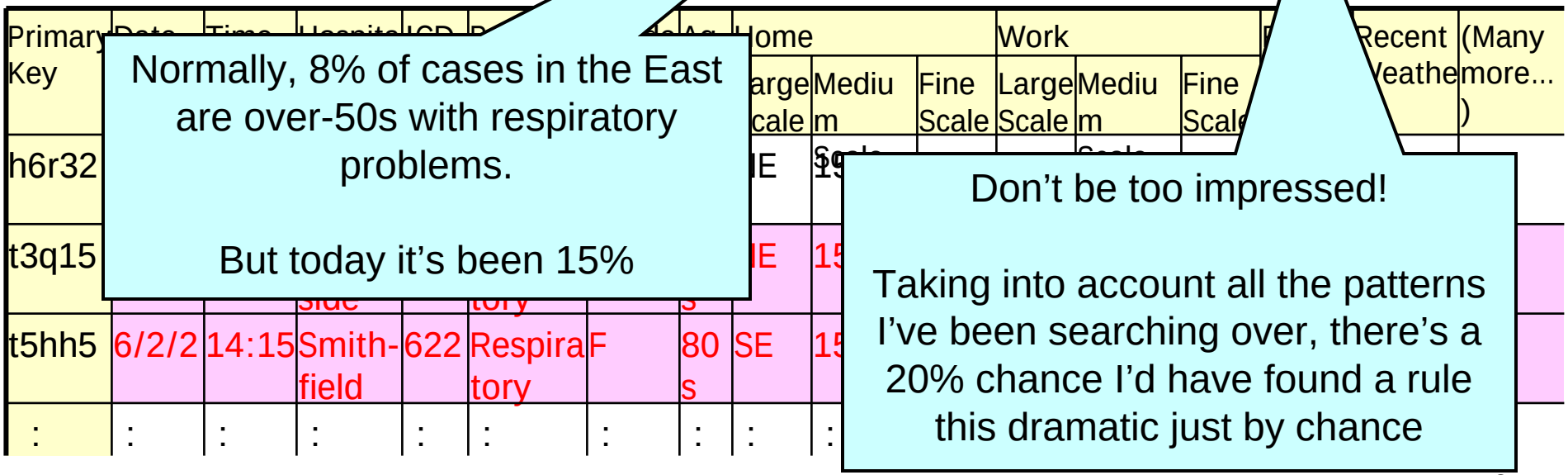
2. Time Window Length

3. Which attributes to use?

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!

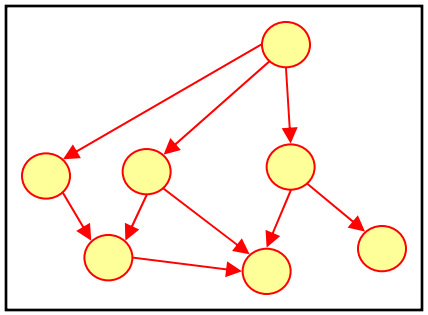
Taking into account all the patterns I've been searching over, there's a 20% chance I'd have found a rule this dramatic just by chance

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

Idea: Bayesian Networks



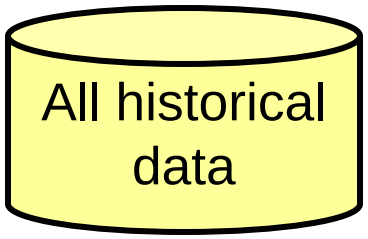
"Patients from West Park Hospital are less likely to be young"

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

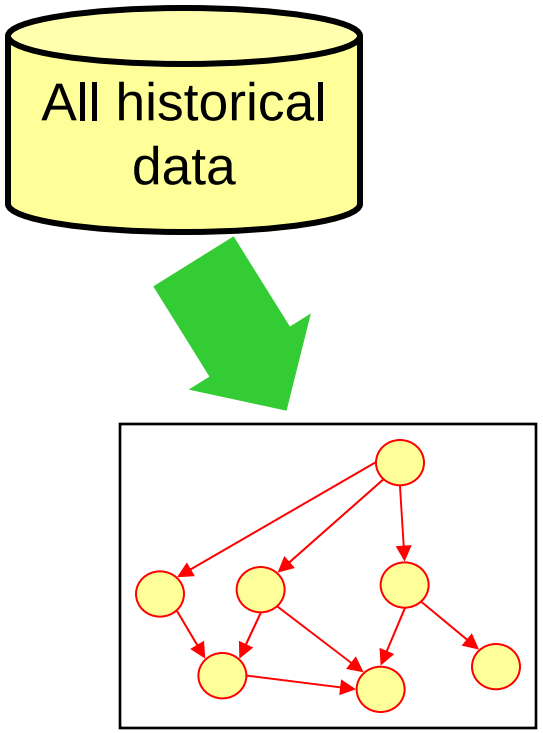
"The Viral prodrome is more likely to co-occur with a Rash prodrome than Botulinic"

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

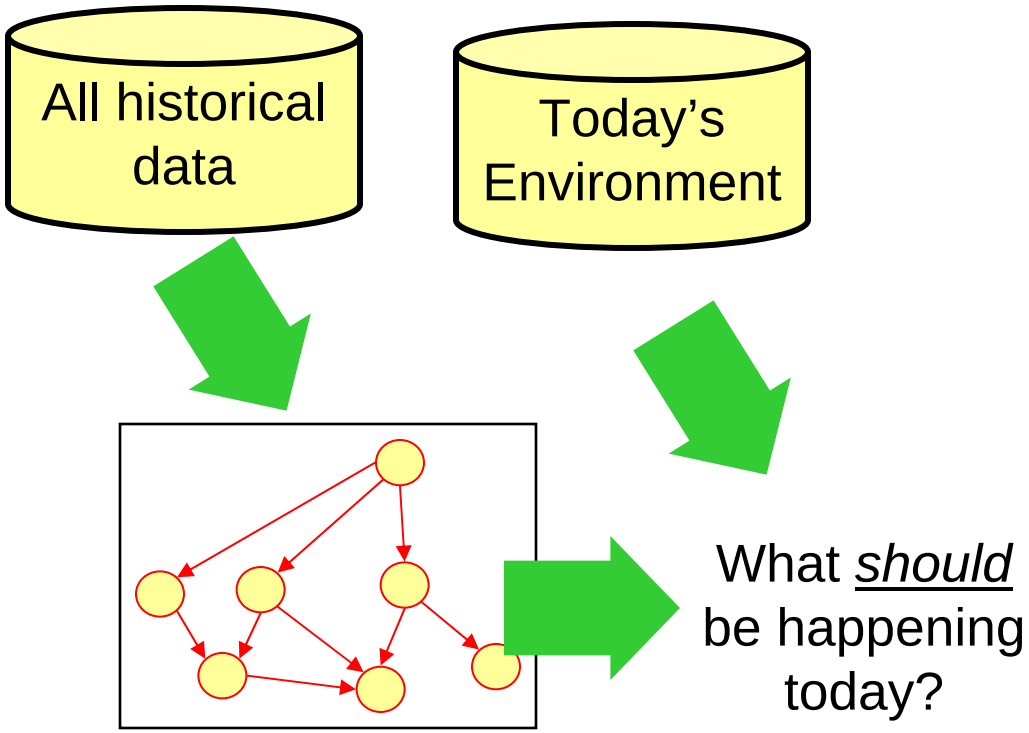
WSARE 3.0



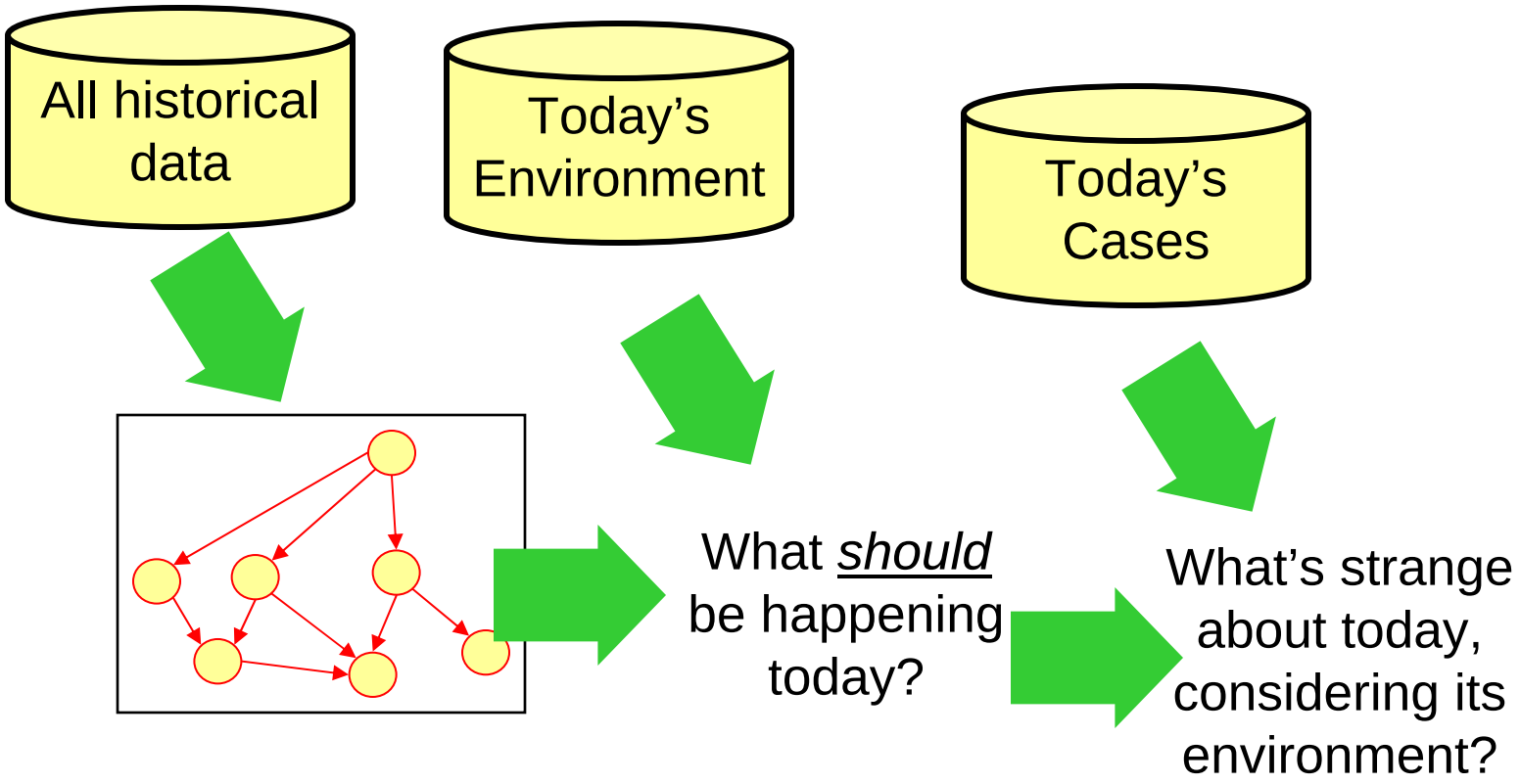
WSARE 3.0



WSARE 3.0


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

WSARE 3.0

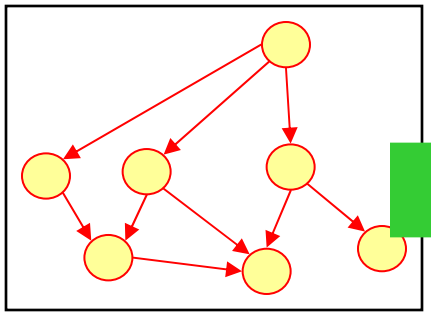

All historical data
Today's Environment
Today's Cases
What *should* be happening today?
What's strange about today, considering its environment?

WSARE 3.0

All historical data

Today's Environment

Today's Cases

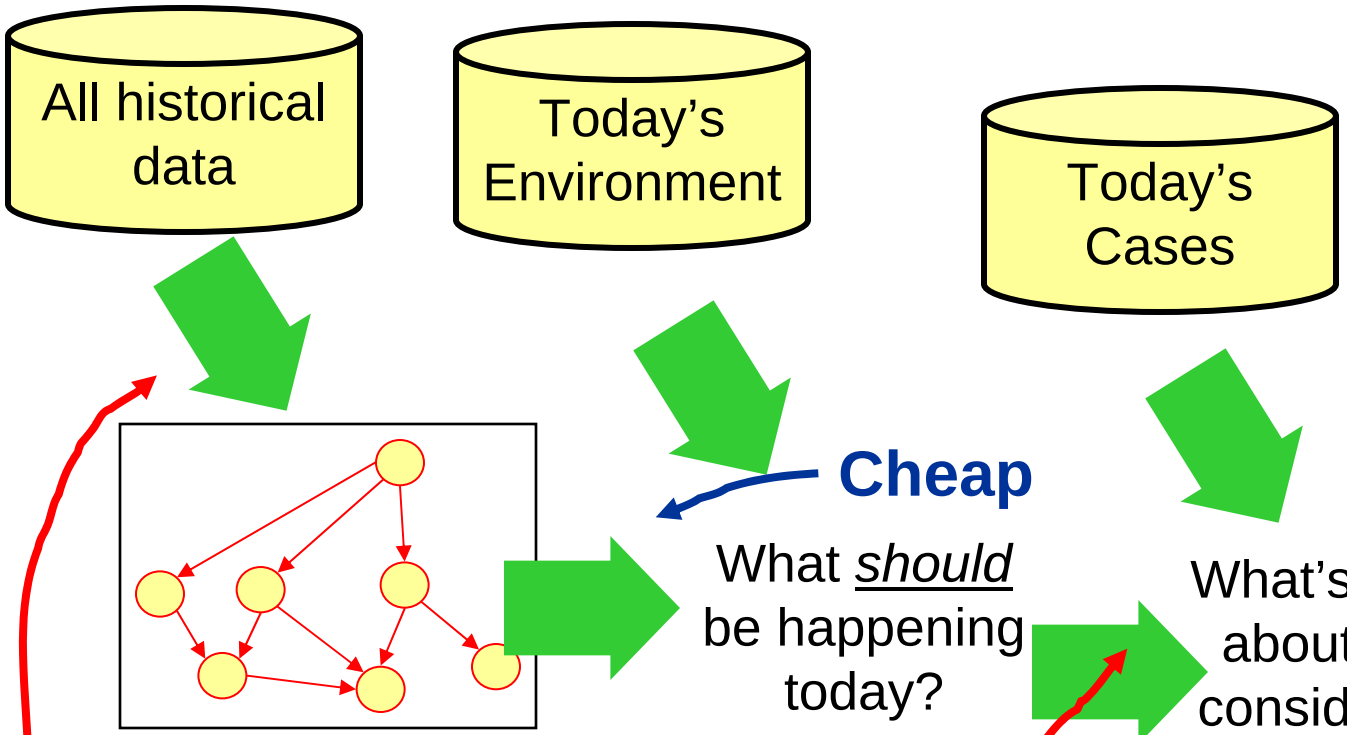


What *should* be happening today?

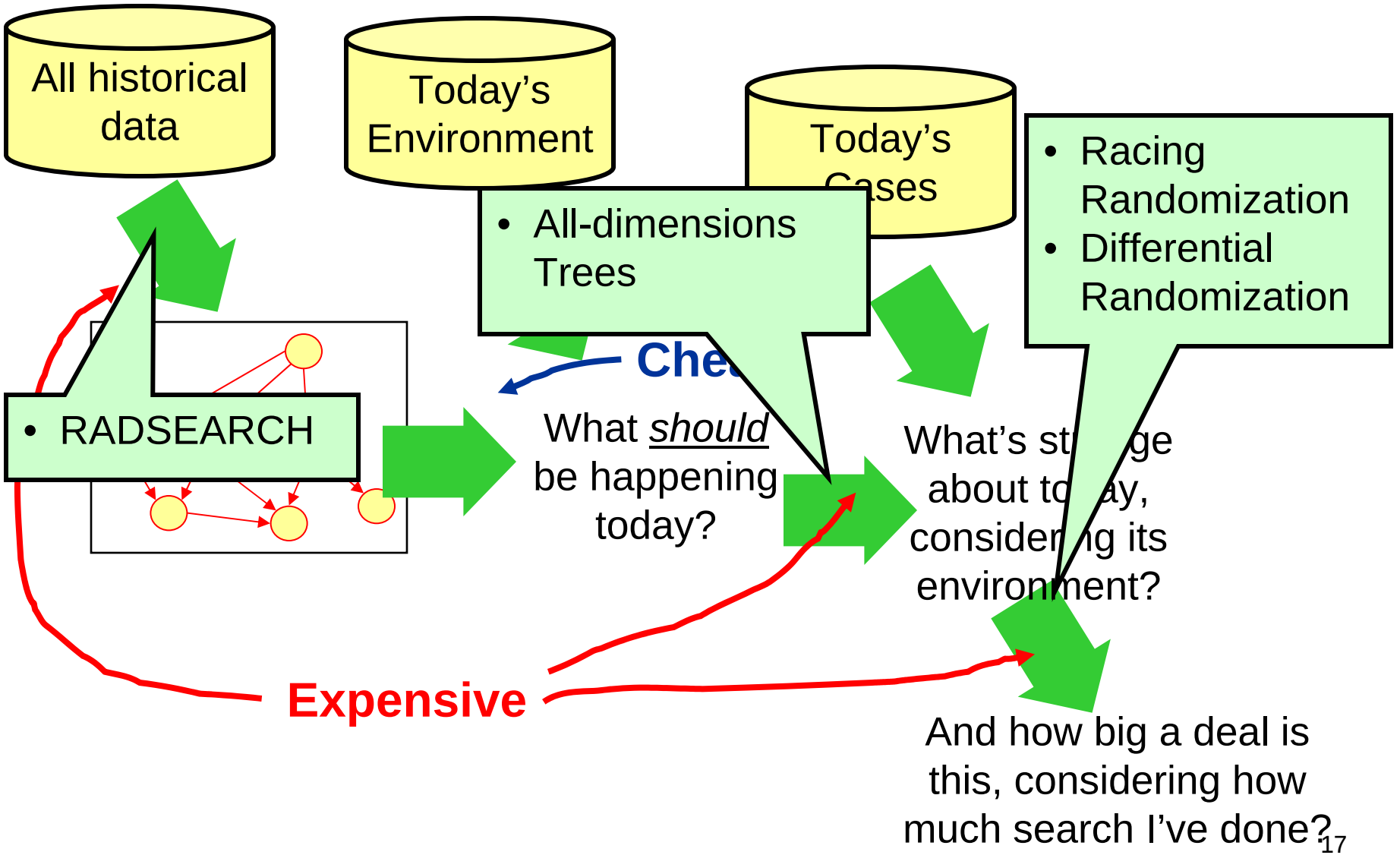
What's strange about today, considering its environment?

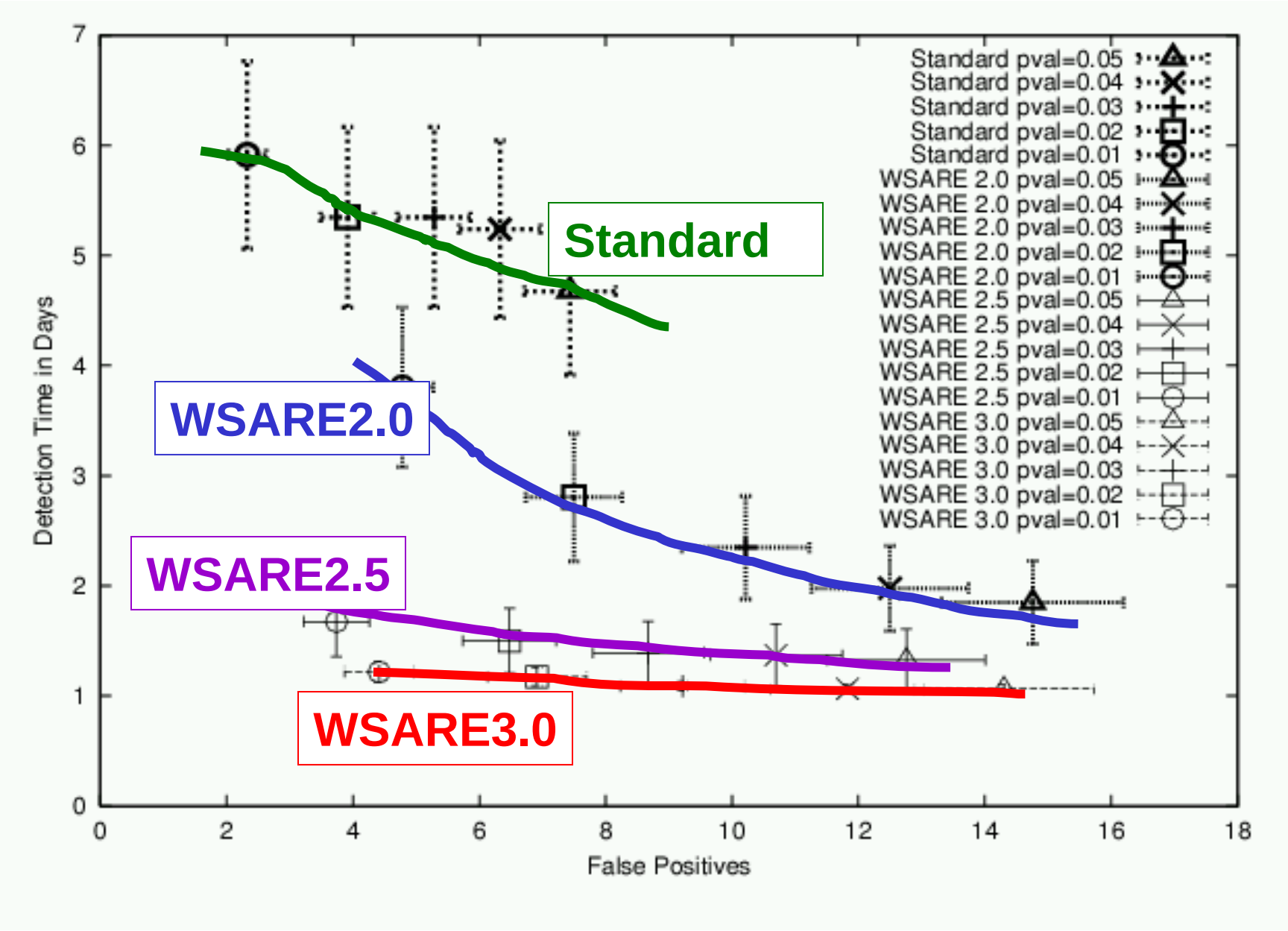
And how big a deal is this, considering how much search I've done?

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 environment?
Expensive
And how big a deal is this, considering how much search I've done?



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!

Other New Algorithmic Developments

Specific Detectors

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

BARD: Airborne Attack Detection
[Hogan, Cooper]

General Detectors

What's Strange about Recent Events

Fast Scan Statistic
[Neill, Moore]

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

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

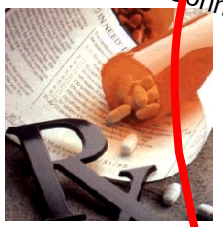
Bayesian Network Spatial Scan
[Neill, Moore, Schneider, Cooper Wagner, Wong]

Possible Future Connection

Question: How do we use all this information?

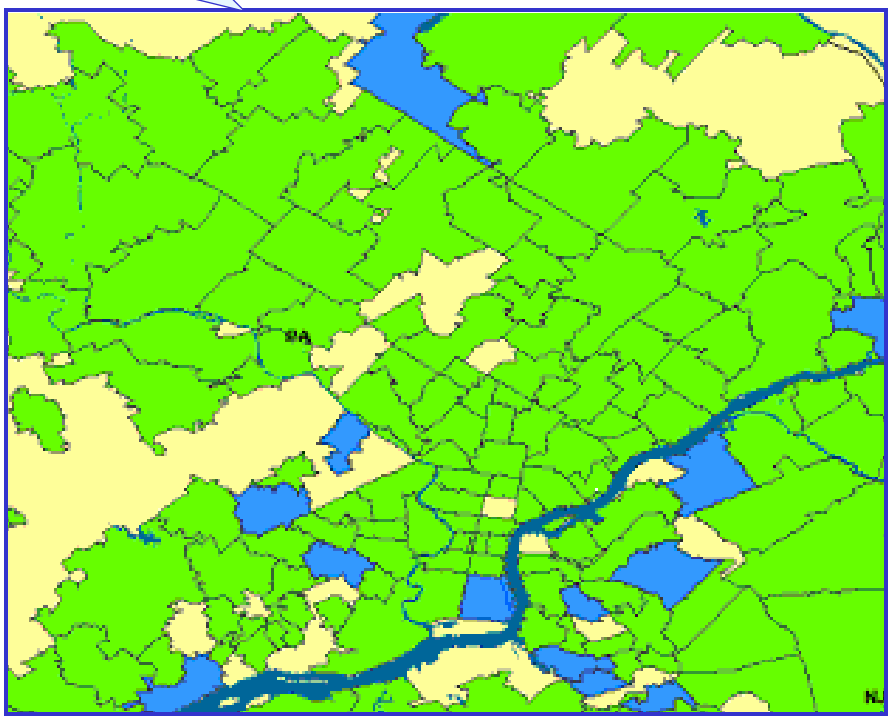
How can we "plug in" new streams?

How can we exploit multiattribute form?

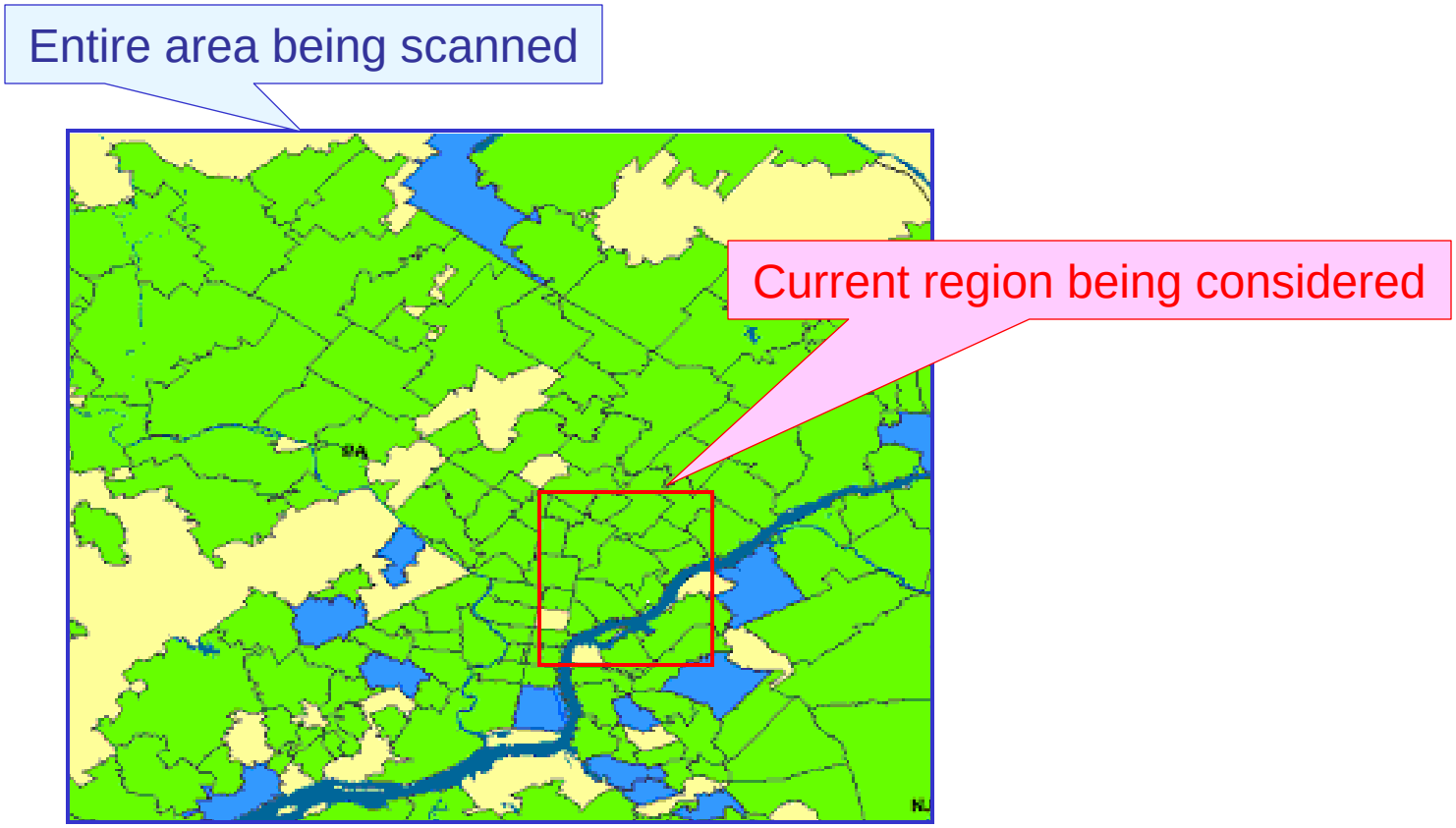


One Step of Spatial Scan

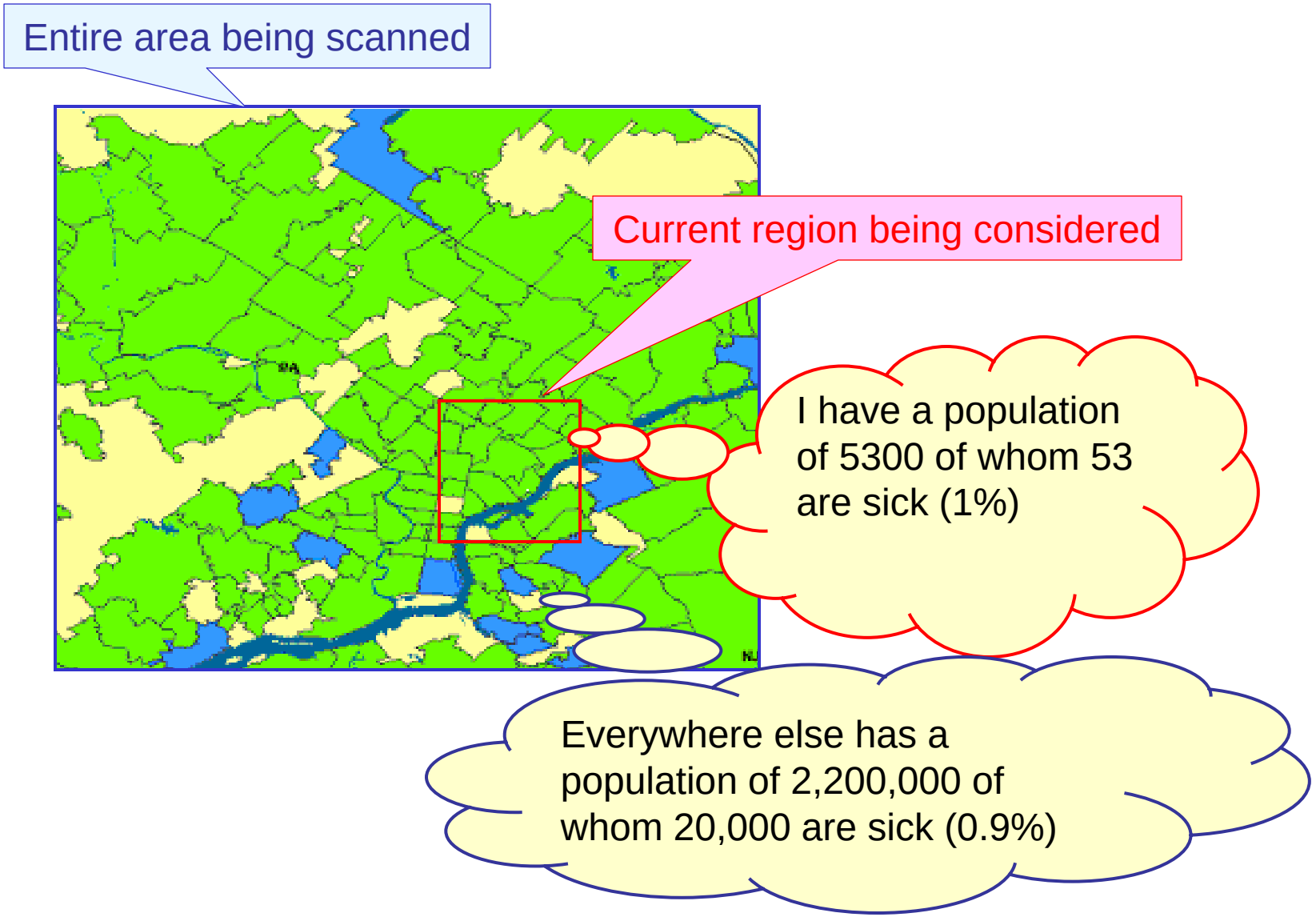
Entire area being scanned



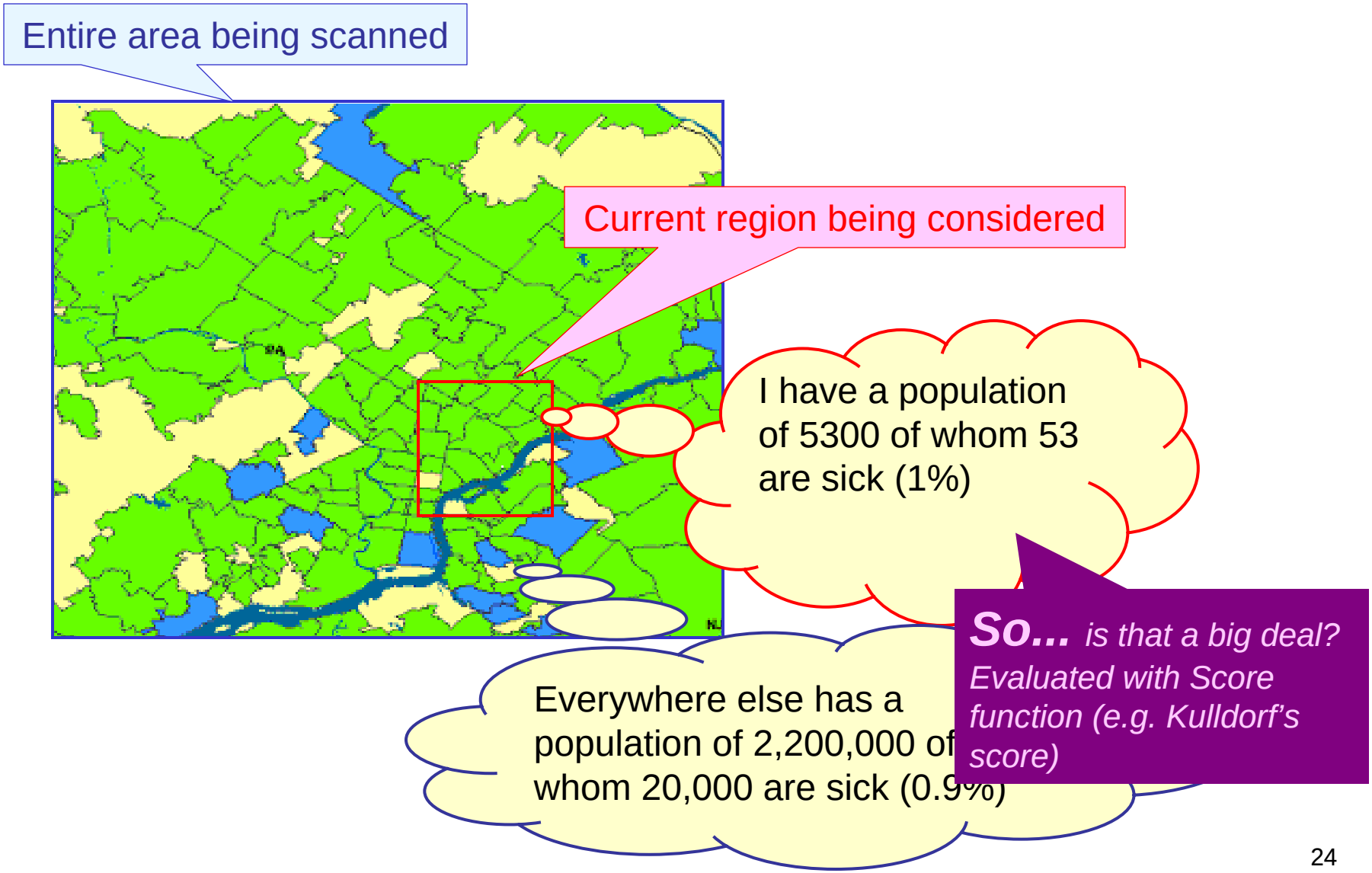
One Step of Spatial Scan



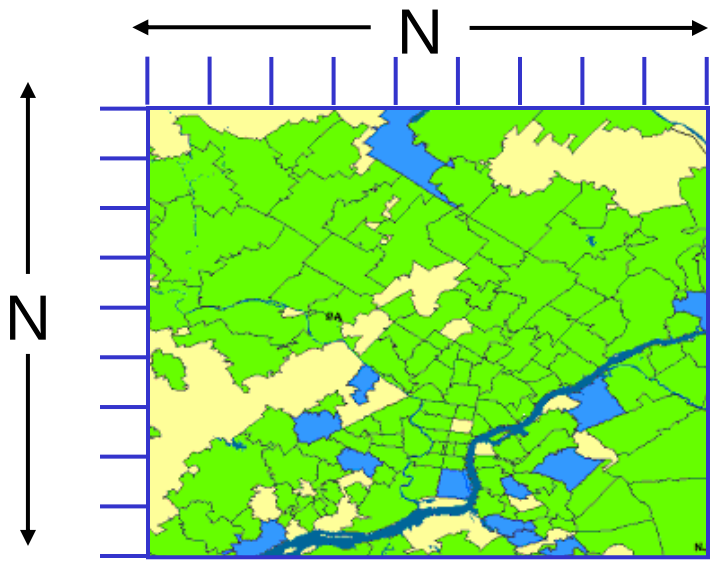
One Step of Spatial Scan



One Step of Spatial Scan



Fast squares speedup



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

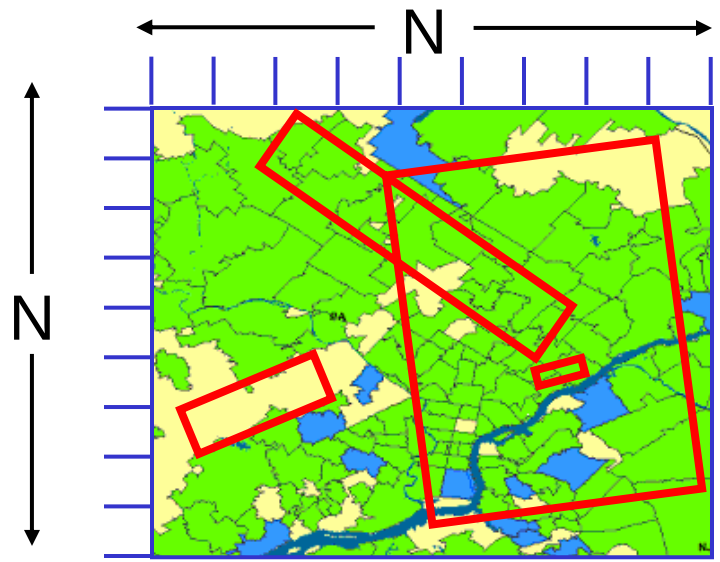
  If not, we can use several other speedup tricks.

- 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 oriented rectangles speedup

Work in progress

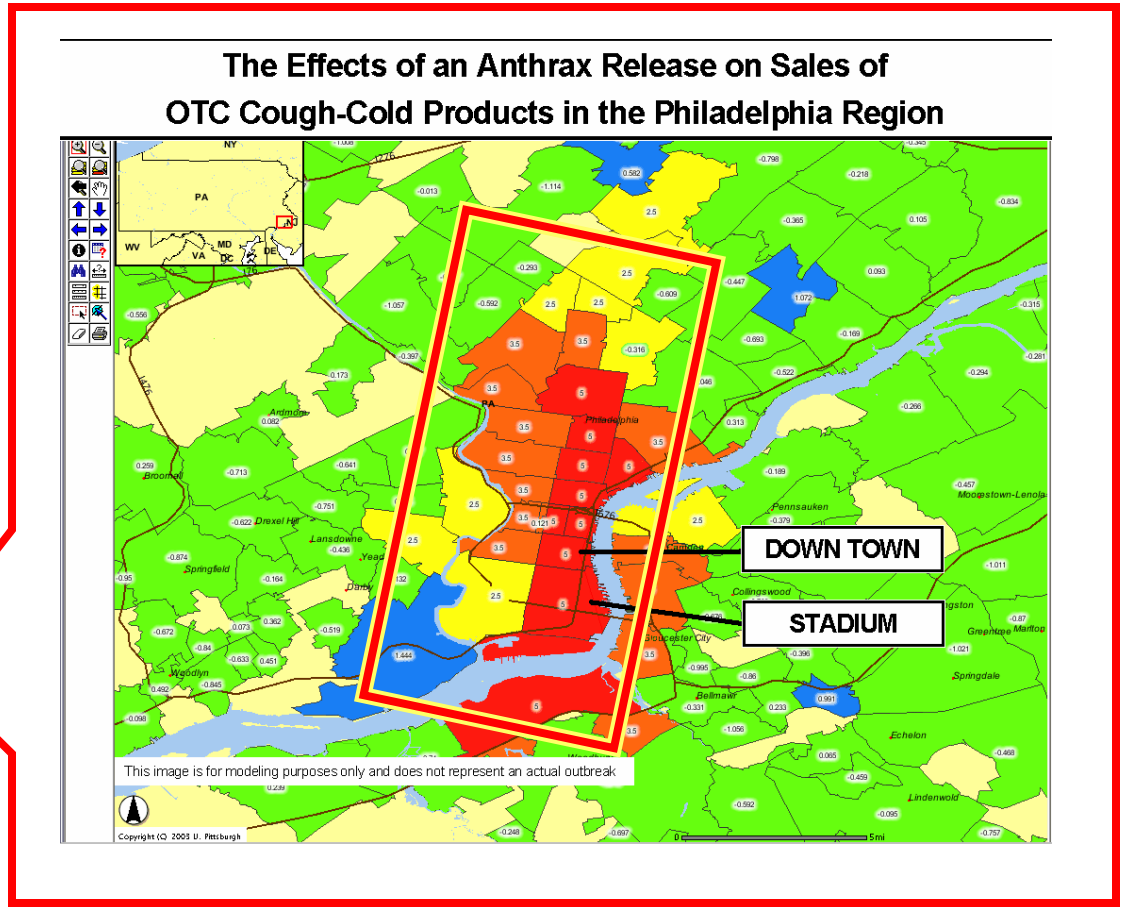


- Theoretical complexity of fast rectangles: $18N^2 \log N$ (as opposed to naïve $18N^4$)

(Angles discretized to 5 degree buckets)

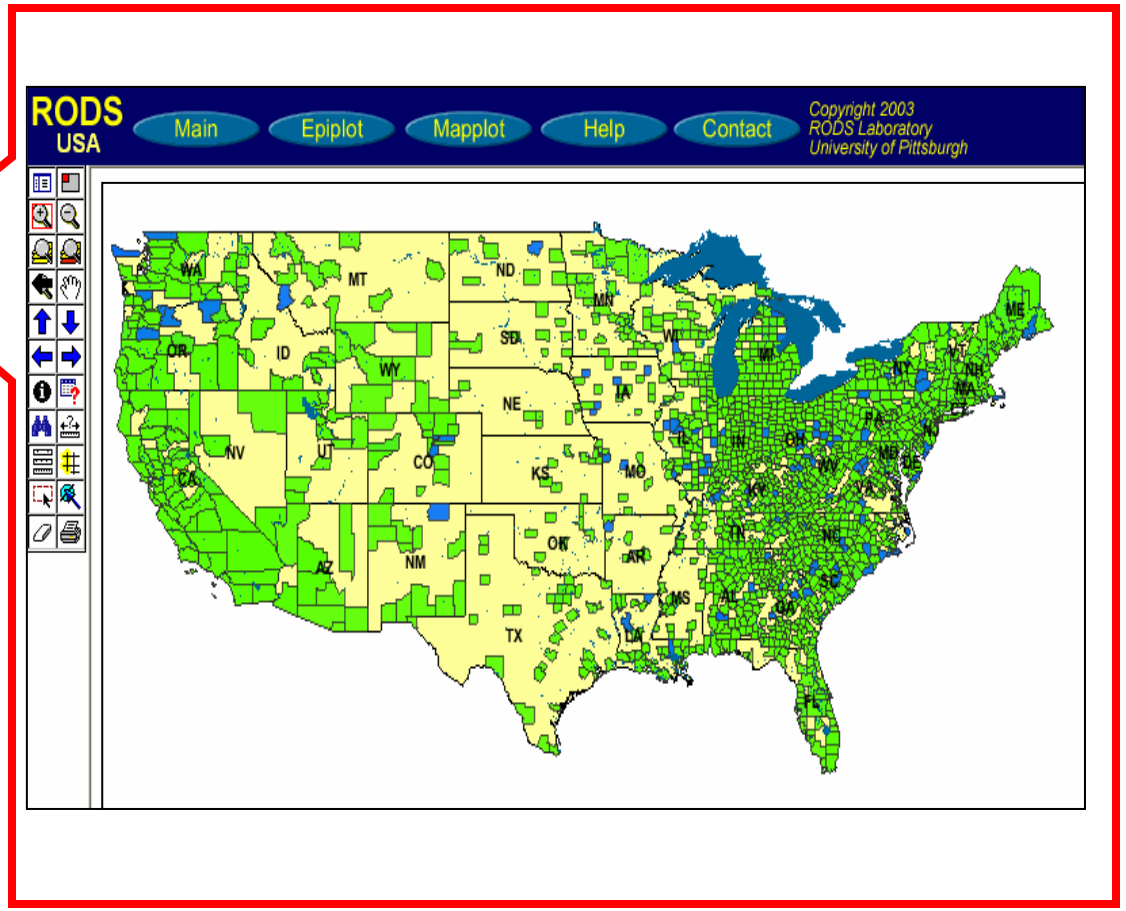
Why the Scan Statistic speed obsession?

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



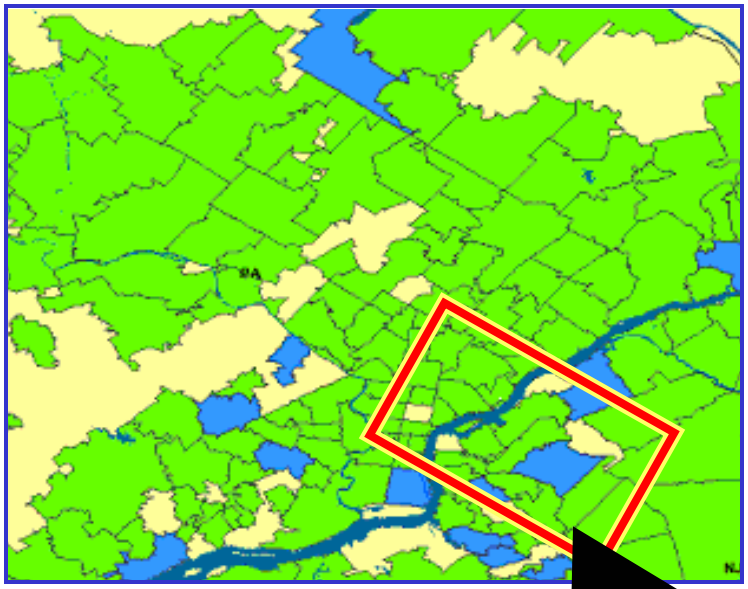
Why the Scan Statistic speed obsession?

- Traditional Scan Statistics very expensive, especially with Randomization tests
- "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
- "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

PANDA: A Few Details about Its Current Status

- *Data* consists of census information about a population, plus emergency department (ED) information about patients
- The *population* currently being modeled consists of all ~1.4M people in Allegheny County
- The outbreak being modeled is roughly based on an airborne anthrax release – it requires (and will receive) significant refinement and extension.

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

What's Strange about Recent Events

Fast Scan Statistic
[Neill, Moore]

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

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

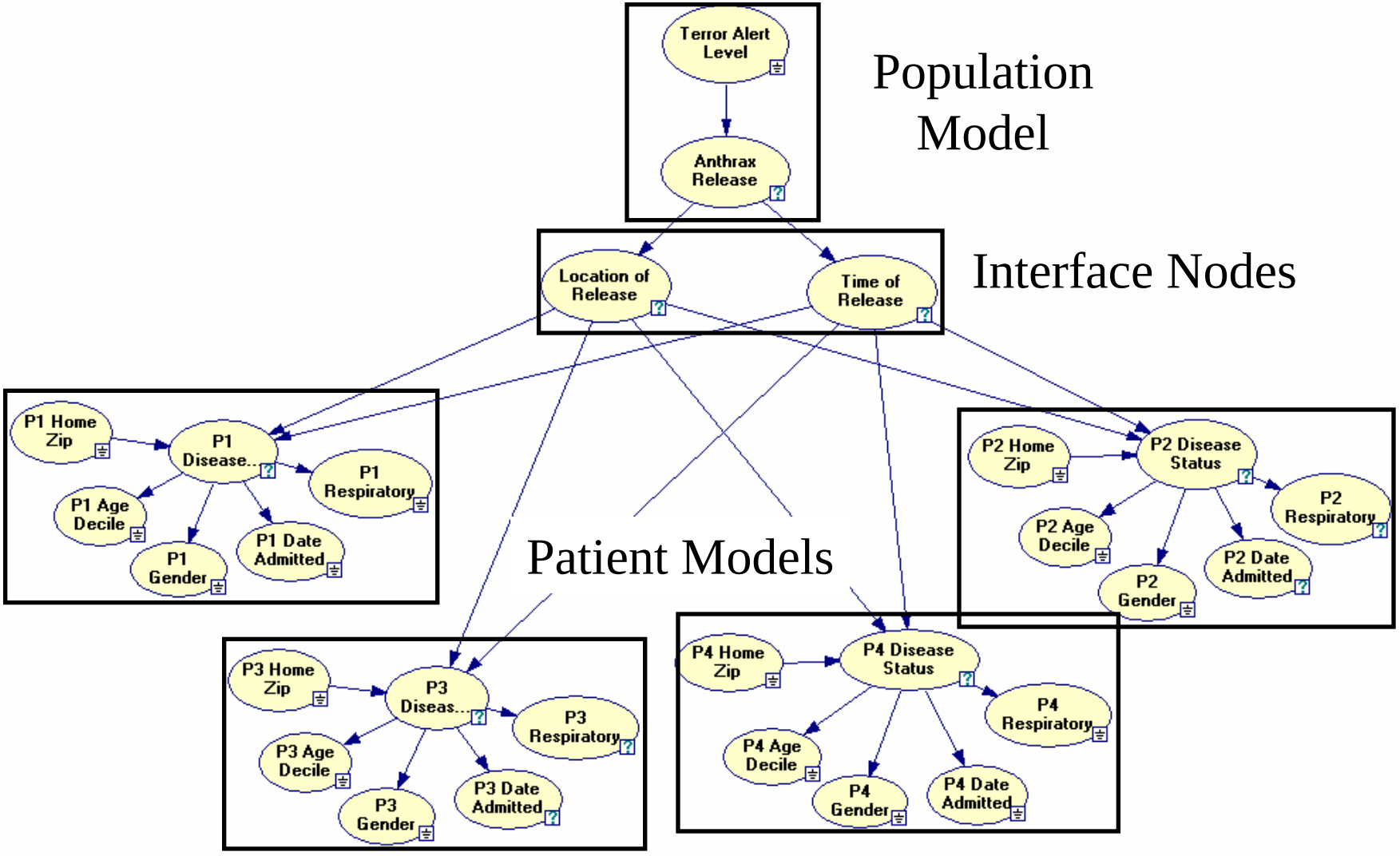
Bayesian Network Spatial Scan
[Neill, Moore, Schneider, Cooper Wagner, Wong]

Question: How do we use all this information?

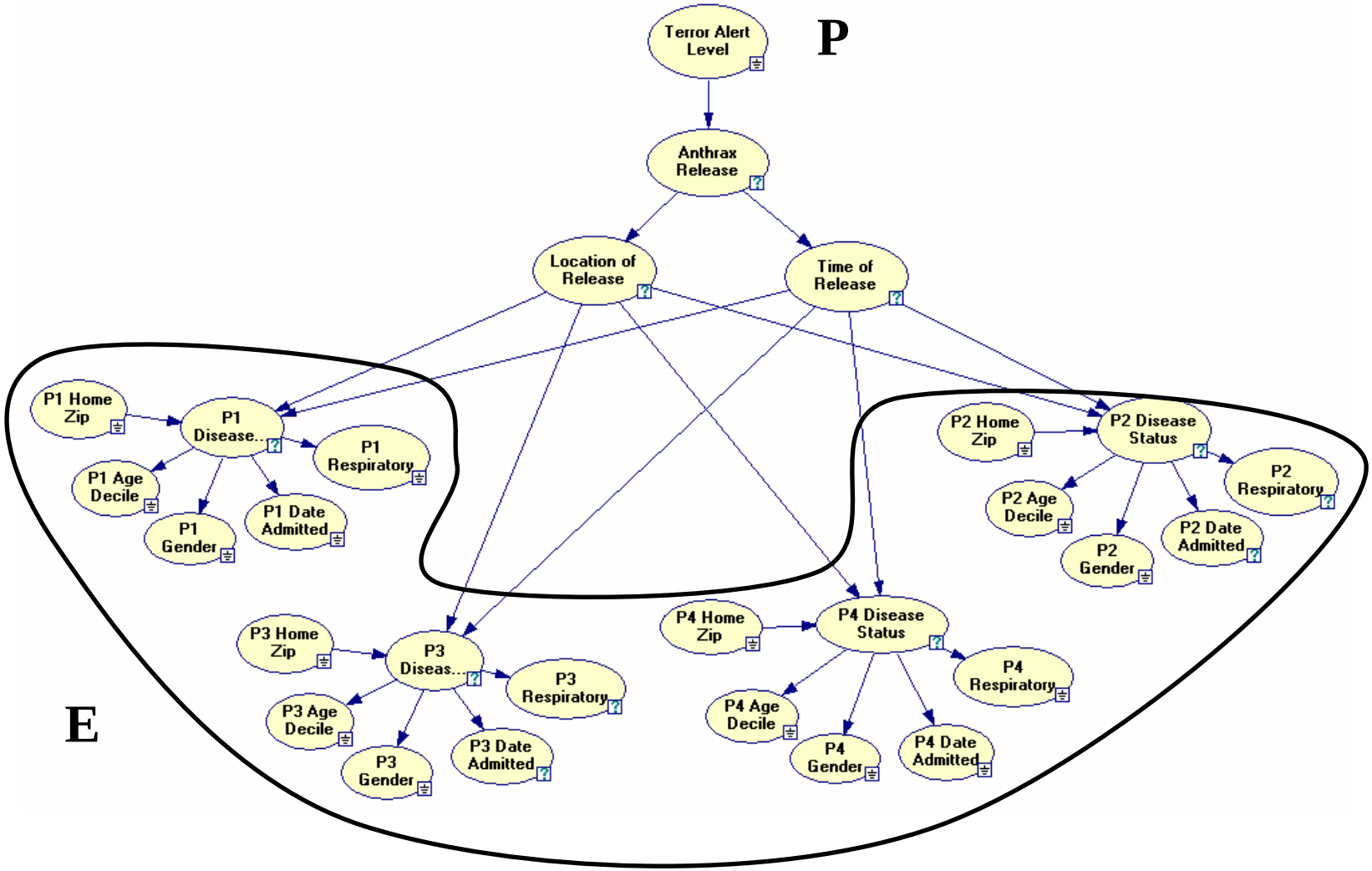
How can we "plug in" new streams?

How can we exploit multiattribute form?

Example Model for “Anthrax-like” Airborne Release

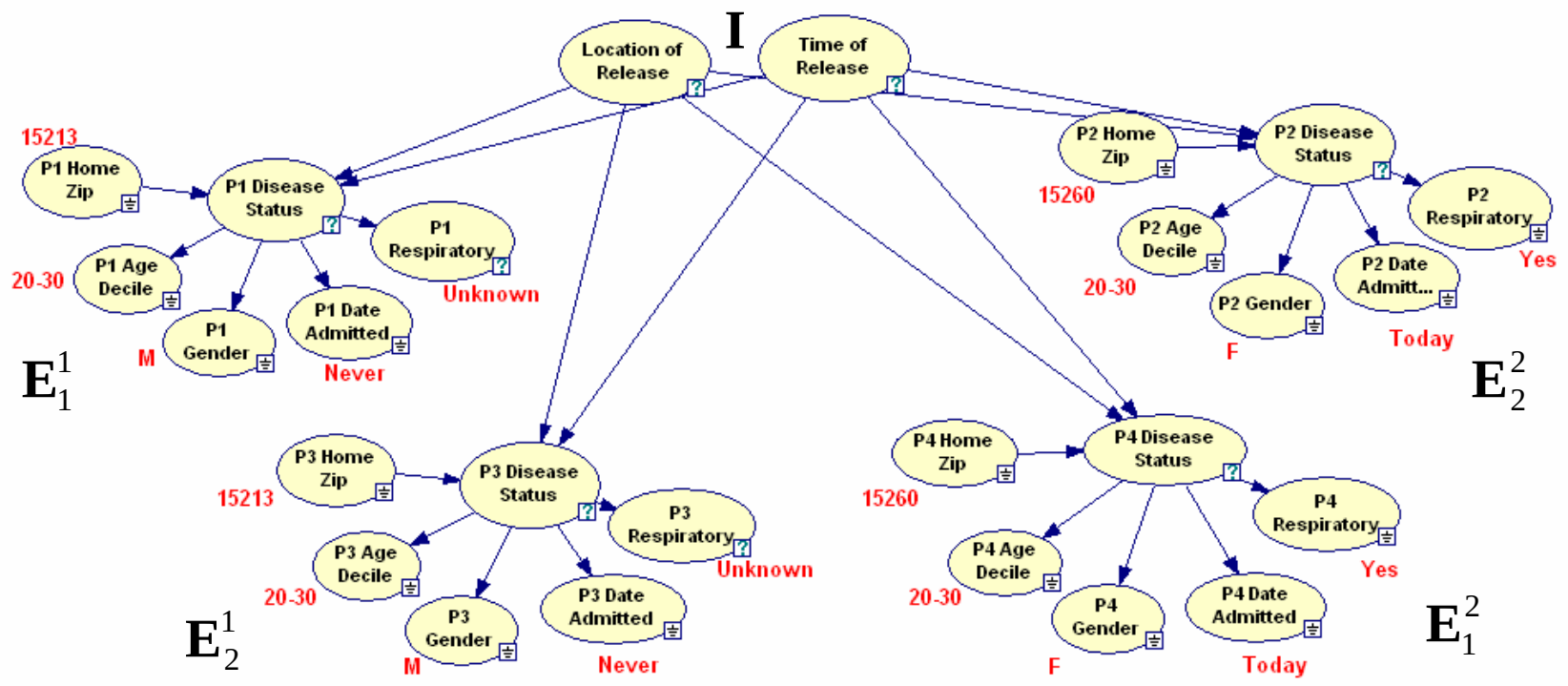


Calculating Probability of a Release



$$P(\text{Release} \mid \mathbf{E}; \mathbf{P}) \propto P(\mathbf{E} \mid \text{Release}; \mathbf{P}) P(\text{Release} \mid \mathbf{P})$$

Calculating $P(\mathbf{E} \mid \mathbf{I})$

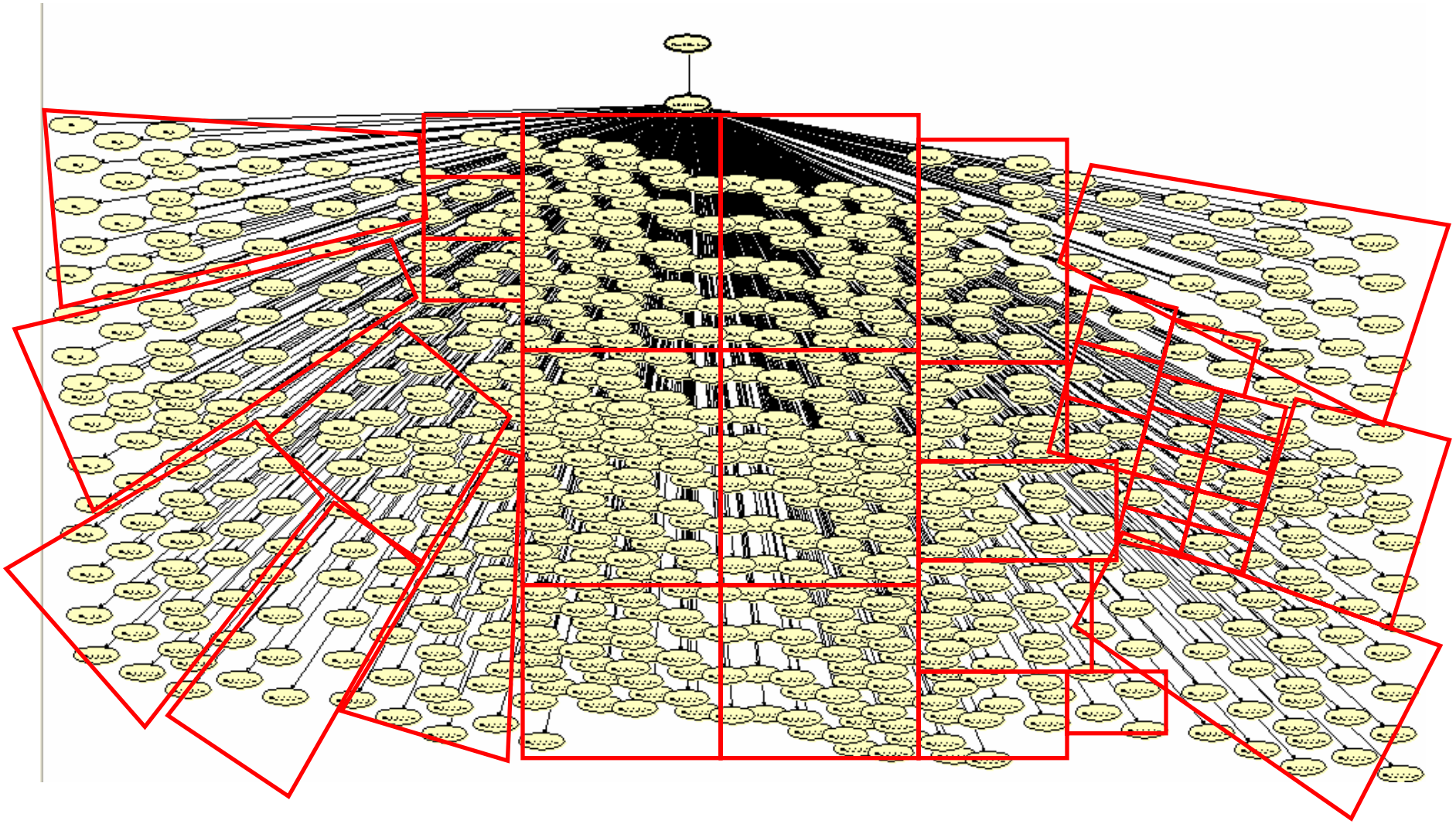


$$P(\mathbf{E} \mid \mathbf{I}) = P(\mathbf{E}_1^1, \mathbf{E}_2^1, \mathbf{E}_1^2, \mathbf{E}_2^2 \mid \mathbf{I})$$

$$= P(\mathbf{E}_1^1 \mid \mathbf{I}) \cdot P(\mathbf{E}_2^1 \mid \mathbf{I}) \cdot P(\mathbf{E}_1^2 \mid \mathbf{I}) \cdot P(\mathbf{E}_2^2 \mid \mathbf{I}) \qquad \text{(Assumption 3)}$$

$$= P(\mathbf{E}^1 \mid \mathbf{I})^2 \cdot P(\mathbf{E}^2 \mid \mathbf{I})^2$$

Equivalence Classes



Millions of people in a population can be partitioned into 48,000 or fewer equivalence classes

Conclusions

- The easy way to combine data streams is to insert them into one relational table.
- Can do spatial scans that evaluate multiple sources per region.
- Can use a huge probabilistic model to rationally combine multiple data streams.

E.G. WSARE

E.G. WSARE-SCAN

E.G. Panda

RODS: http://www.health.pitt.edu/rods
Auton Lab: http://www.autonlab.org

Conclusions

- The easy way to combine data streams is to insert them into one relational table.
- Can do spatial scans that evaluate multiple sources per region.
- Can use a huge probabilistic model to rationally combine multiple data streams.

E.G. WSARE

E.G. WSARE-SCAN

E.G. Panda

Challenge: Managing complexity

Challenge: Computational tractability

RODS: http://www.health.pitt.edu/rods
Auton Lab: http://www.autonlab.org