## RODS and Multiple Data Streams

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Institution</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greg Cooper</td>
<td>Professor</td>
<td>Computer Science and RODS lab, U. Pitt</td>
<td><a href="mailto:gfc@cbmi.upmc.edu">gfc@cbmi.upmc.edu</a></td>
</tr>
<tr>
<td>Bill Hogan</td>
<td>Assistant Professor</td>
<td>RODS lab, U. Pitt</td>
<td><a href="mailto:wrh@cbmi.pitt.edu">wrh@cbmi.pitt.edu</a></td>
</tr>
<tr>
<td>Andrew Moore</td>
<td>Professor</td>
<td>Computer Science, Carnegie Mellon</td>
<td><a href="mailto:awm@cs.cmu.edu">awm@cs.cmu.edu</a></td>
</tr>
<tr>
<td>Daniel Neill</td>
<td>Graduate Student</td>
<td>Computer Science, Carnegie Mellon</td>
<td><a href="mailto:neill@cs.cmu.edu">neill@cs.cmu.edu</a></td>
</tr>
<tr>
<td>Jeff Schneider</td>
<td>Research Professor</td>
<td>Computer Science, Carnegie Mellon</td>
<td><a href="mailto:schneide@cs.cmu.edu">schneide@cs.cmu.edu</a></td>
</tr>
<tr>
<td>Rich Tsui</td>
<td>Research Professor and associate Director of RODS lab</td>
<td>RODS lab, U. Pitt</td>
<td><a href="mailto:tsui@cbmi.pitt.edu">tsui@cbmi.pitt.edu</a></td>
</tr>
<tr>
<td>Mike Wagner</td>
<td>Professor and Director of RODS lab</td>
<td>RODS lab, U. Pitt</td>
<td><a href="mailto:mmw@cbmi.pitt.edu">mmw@cbmi.pitt.edu</a></td>
</tr>
<tr>
<td>Weng-Keen Wong</td>
<td>Graduate Student</td>
<td>Computer Science, Carnegie Mellon</td>
<td><a href="mailto:wkw@cs.cmu.edu">wkw@cs.cmu.edu</a></td>
</tr>
</tbody>
</table>

**RODS**: [http://www.health.pitt.edu/rods](http://www.health.pitt.edu/rods)
**Auton Lab**: [http://www.autonlab.org](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

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]

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

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]
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]

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

<table>
<thead>
<tr>
<th>Primary Key</th>
<th>Date</th>
<th>Time</th>
<th>Hospital</th>
<th>ICD9</th>
<th>Prodrome</th>
<th>Gender</th>
<th>Age</th>
<th>Home</th>
<th>Large Scale</th>
<th>Medium Scale</th>
<th>Fine Scale</th>
<th>Work</th>
<th>Large Scale</th>
<th>Medium Scale</th>
<th>Fine Scale</th>
<th>Recent Flu Levels</th>
<th>Recent Weather</th>
<th>(Many more…)</th>
</tr>
</thead>
<tbody>
<tr>
<td>h6r32</td>
<td>6/2</td>
<td>14:12</td>
<td>Downtown</td>
<td>781</td>
<td>Fever</td>
<td>M</td>
<td>20s</td>
<td>NE</td>
<td>15217</td>
<td>A5</td>
<td>NW</td>
<td>15217</td>
<td>B8</td>
<td>2%</td>
<td>70R</td>
<td>…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t3q15</td>
<td>6/2</td>
<td>14:15</td>
<td>Riverside</td>
<td>717</td>
<td>Respiratory</td>
<td>M</td>
<td>60s</td>
<td>NE</td>
<td>15222</td>
<td>J3</td>
<td>NE</td>
<td>15222</td>
<td>J3</td>
<td>2%</td>
<td>70R</td>
<td>…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t5hh5</td>
<td>6/2</td>
<td>14:15</td>
<td>Smithfield</td>
<td>622</td>
<td>Respiratory</td>
<td>F</td>
<td>80s</td>
<td>SE</td>
<td>15210</td>
<td>K9</td>
<td>SE</td>
<td>15210</td>
<td>K9</td>
<td>2%</td>
<td>70R</td>
<td>…</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
WSARE v2.0

• Input:
  1. Date/time-indexed biosurveillance-relevant data stream
  2. Time Window Length
  3. Which attributes to use?

• Output:
  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 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

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

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

“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

All historical data
WSARE 3.0

All historical data
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

What should be happening today?

Cheap

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

- Racing Randomization
- Differential Randomization

- RADSEARCH

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?

Expensive
Results on Simulation

- **Standard**
- **WSARE 2.0**
- **WSARE 2.5**
- **WSARE 3.0**

The graph shows the detection time in days against the number of false positives for different versions of WSARE.
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]

**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

Entire area being scanned

Current region being considered
One Step of Spatial Scan

Entire area being scanned

Current region being considered

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

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

Entire area being scanned

Current region being considered

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

So... is that a big deal?

Evaluated with Score function (e.g. Kulldorf’s score)

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

- 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.
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]

Specific Detectors

PANDA2: Patient-based Bayesian Network
  - [Cooper, Levander et al.]
BARD: Airborne Attack Detection
  - [Hogan, Cooper]

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

Population Model

Interface Nodes

Patient Models
Calculating Probability of a Release

$$P(\text{Release} | E ; P) \propto P(E | \text{Release} ; P)P(\text{Release} | P)$$
Calculating \( P(E \mid I) \)

\[
P(E \mid I) = P(E_1^1, E_2^1, E_1^2, E_2^2 \mid I) \\
= P(E_1^1 \mid I) \cdot P(E_2^1 \mid I) \cdot P(E_1^2 \mid I) \cdot P(E_2^2 \mid I) \quad \text{(Assumption 3)} \\
= P(E_1^1 \mid I)^2 \cdot P(E_2^2 \mid I)^2
\]
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.

Challenge: Managing complexity

Challenge: Computational tractability

E.G. WSARE
E.G. WSARE-SCAN
E.G. Panda

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