Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks

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

The Auton Lab

Faculty: Andrew Moore, Jeff Schneider, Artur Dubrawski

Postdoctoral Fellows: Brigham Anderson, Alexander Gray, Paul Komarek, Dan Pelleg

Graduate Students: Brent Bryan, Kaustav Das, Khalid El-Arini, Anna Goldenberg, Jeremy Kubica, Ting Liu, Daniel Neill, Sajid Siddiqi, Purna Sarkar, Ajit Singh, Weng-Keen Wong

Head of Software Development: Jeanie Komarek

Programmers: Patrick Choi, Adam Goode, Pat Gunn, Joey Liang, John Ostlund, Robin Sabhnani, Rahul Sankathar

Executive Assistant: Kristen Schrauder

Head of Sys. Admin: Jacob Joseph

Undergraduate and Masters Interns: Kenny Daniel, Sandy Hsu, Dongyeol Lee, Jennifer Lee, Avilay Parekh, Chris Rotella, Jonathan Terleski

Recent Alumni: 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 Seres (Qnetia)
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

Collaborators...
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

Potential coll-aborators of Doe:
The pattern of travel among the group consisting of (Doe, Smith, Jones, Moore) during 93-96 can’t be explained by coincidence

Our 5 biggest applications in 2004

Big Noisy Data-base of Links

Potential coll-aborators of Doe:
The pattern of travel among the group consisting of Doe, Smith, Jones, Moore) during 93-96 can’t be explained by coincidence

Our 5 biggest applications in 2004

Big Noisy Data-base of Links

Potential coll-aborators of Doe:
The pattern of travel among the group consisting of Doe, Smith, Jones, Moore) during 93-96 can’t be explained by coincidence
Our 5 biggest applications in 2002

1. Drug Screening
2. Big Astrophysics Automated Science
3. Biomedical Security (with Mike Wagner, University of Pittsburgh)
4. Security Surveillance Mining
5. Autonomous self-tweaking engines

Potential coll-aborators of Doe:

The pattern of travel among the group consisting of Doe, Smith, Jones, Moore during 1993-1996 can’t be explained by coincidence.

Drug Screening

Security Surveillance Mining

Autonomous self-tweaking engines

Big Noisy Database of Links

There must be something useful here. I’m drowning in noise.

Analyst Record # 456621: Doe and Smith were booked in the same hotel on 3/6/93.

Potential coll-aborators of Doe:

The pattern of travel among the group consisting of Doe, Smith, Jones, Moore during 1993-1996 can’t be explained by coincidence.

Our 5 biggest applications in 2002

1. Big Astrophysics Automated Science
2. Biomedical Security (with Mike Wagner, University of Pittsburgh)
3. Security Surveillance Mining
4. Autonomous self-tweaking engines
5. Drug Screening

"..Early Thursday Morning. Russia. April 1979..."

Sverdlovsk
During April and May 1979, there were 77 confirmed cases of inhalational anthrax in Sverdlovsk.
What if this happened in 2002?

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

[Tsui et al., 2000]
30% of ER visits in 13-county metropolitan area (population 3.0 Million)

52% of ER visits in Allegheny County (population 1.3 Million)

How is RODS Used?

When an alarm sounds...
When an alarm sounds...

How is RODS Used?
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?

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

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

Can we find a pattern in the strange stuff?

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?

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.
### WSARE v2.0

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

---

### Input Examples:

- **Example**
  - "last 24 hours"
  - "ignore key and weather"

#### Table:

<table>
<thead>
<tr>
<th>Primary Key</th>
<th>Date</th>
<th>Time</th>
<th>Hospital</th>
<th>ICD9</th>
<th>Prodmote</th>
<th>Gender</th>
<th>Age</th>
<th>Home Large Scale</th>
<th>Home Medium Scale</th>
<th>Home Fine Scale</th>
<th>Work Large Scale</th>
<th>Work Medium Scale</th>
<th>Work Fine Scale</th>
<th>Recent Flu Levels</th>
<th>Recent Weather</th>
<th>Many more...</th>
</tr>
</thead>
<tbody>
<tr>
<td>H6r32</td>
<td>6/2/2</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>k5</td>
<td>NW</td>
<td>15213</td>
<td>B8</td>
<td>2%</td>
<td>70R</td>
<td>...</td>
</tr>
<tr>
<td>Eqq15</td>
<td>6/2/2</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>3</td>
<td>NE</td>
<td>15222</td>
<td>3</td>
<td>2%</td>
<td>70R</td>
<td>...</td>
</tr>
<tr>
<td>Sshh5</td>
<td>6/2/2</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>
</tr>
</tbody>
</table>
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>Work</th>
<th>Large Scale</th>
<th>Medium Scale</th>
<th>Fine Scale</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>5/12</td>
<td>14:12</td>
<td>Downtown</td>
<td>781</td>
<td>Fever</td>
<td>M</td>
<td>20s</td>
<td>NE</td>
<td>NE</td>
<td>15217</td>
<td>A5</td>
<td>NW</td>
<td>15213</td>
<td>B8</td>
<td>2%</td>
<td>70R</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>B3q15</td>
<td>5/12</td>
<td>14:15</td>
<td>Riverside</td>
<td>717</td>
<td>Respiratory</td>
<td>M</td>
<td>60s</td>
<td>NE</td>
<td>NE</td>
<td>15222</td>
<td>3</td>
<td>NE</td>
<td>15222</td>
<td>3</td>
<td>2%</td>
<td>70R</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>B3hh5</td>
<td>5/12</td>
<td>14:15</td>
<td>Smithfield</td>
<td>622</td>
<td>Respiratory</td>
<td>F</td>
<td>80s</td>
<td>SE</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>
</tr>
</tbody>
</table>

Simple WSARE

• Given 500 day’s worth of ER cases at 15 hospitals...

<table>
<thead>
<tr>
<th>Date</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu 5/22/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Fri 5/23/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Sat 12/9/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sun 12/10/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Sat 12/16/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Sat 12/23/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Fri 9/14/2001</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
</tbody>
</table>
Simple WSARE

- Given 500 day’s worth of ER cases at 15 hospitals...
- For each day...
  - Take today’s cases

<table>
<thead>
<tr>
<th>Date</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu 5/22/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Fri 5/23/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sat 12/9/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sun 12/10/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sat 12/16/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sat 12/23/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Fri 9/14/2001</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu 5/22/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Fri 5/23/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sat 12/9/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sun 12/10/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sat 12/16/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Sat 12/23/2000</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
<tr>
<td>Fri 9/14/2001</td>
<td>C1, C2, C3, C4 ...</td>
</tr>
</tbody>
</table>
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
• Ask: “What’s different about today?”

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

<table>
<thead>
<tr>
<th>Age_Decile</th>
<th>(C_{\text{today}})</th>
<th>(C_{\text{other}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age_Decile = 3</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>Age_Decile (\neq) 3</td>
<td>86</td>
<td>220</td>
</tr>
</tbody>
</table>
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!

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

<table>
<thead>
<tr>
<th>Age_Decile = 3</th>
<th>( C_{\text{today}} )</th>
<th>( C_{\text{other}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Age_{\text{Decile}} \neq 3 )</td>
<td>86</td>
<td>220</td>
</tr>
</tbody>
</table>
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
  - 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?

Checking two component rules

**Table 2: 2x2 Contingency Table 1 for a two component rule**

<table>
<thead>
<tr>
<th>Records from Today matching $C_0$ and $C_1$</th>
<th>Records from Other matching $C_0$ and $C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records from Today matching $C_1$ and differing on $C_0$</td>
<td>Records from Other matching $C_1$ and differing on $C_0$</td>
</tr>
</tbody>
</table>

**Table 3: 2x2 Contingency Table 2 for a two component rule**

<table>
<thead>
<tr>
<th>Records from Today matching $C_0$ and $C_1$</th>
<th>Records from Other matching $C_0$ and $C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records from Today matching $C_0$ and differing on $C_1$</td>
<td>Records from Other matching $C_0$ and differing on $C_1$</td>
</tr>
</tbody>
</table>

- Must pass both tests to be allowed to live.
### 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>Work</th>
<th>Recent Flu Levels</th>
<th>Recent Weather</th>
<th>Many more...</th>
</tr>
</thead>
<tbody>
<tr>
<td>H6r32</td>
<td>6/2/2</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>88</td>
</tr>
<tr>
<td>E3q15</td>
<td>6/2/2</td>
<td>14:15</td>
<td>Riverside</td>
<td>717</td>
<td>Respiratory</td>
<td>NE</td>
<td>60s</td>
<td>NE</td>
<td>15222</td>
<td>3</td>
<td>NE</td>
<td>2%</td>
</tr>
<tr>
<td>5Shh5</td>
<td>6/2/2</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>2%</td>
</tr>
</tbody>
</table>

*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

P-value = 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

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]
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...
Survived Randomized

• No days reported as abnormal

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?
WSARE Simulation Sensitivity vs Specificity on GRIDSIM

WSARE Simulation Sensitivity vs Specificity on GRIDSIM

Copyright © 2002-2004, Andrew Moore
Biosurveillance: Slide 45

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..."

Extra: More efficient use of historical data

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
An easy case

- **Signal** vs **Time**
- **Upper Safe Range**
- **Mean**

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

Conditioning on Seasonal Effects

- **Signal** vs **Time**
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.

Example [Tsui et. Al]

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

Seasonal Effects with Long-Term Trend


Called the Serfling Method [Serfling, 1963]

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.
Day-of-week effects

Fit a day-of-week component
E[Signal] = a + deltaday

E.G: deltamon = +5.42, deltatur = +2.20, deltawe = +3.33, deltathu = +3.10, deltafri = +4.02, deltasa = -12.2, deltasure = -23.42


Another simple form of ANOVA
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”
WSARE 3.0

All historical data

Copyright © 2002-2004, Andrew Moore
Biosurveillance: Slide 59

WSARE 3.0

All historical data

Copyright © 2002-2004, Andrew Moore
Biosurveillance: Slide 60
WSARE 3.0

1. All historical data
2. Today's Environment
3. 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 → What should be happening today? → Today’s Cases → 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 → What should be happening today? → Today’s Cases → What’s strange about today, considering its environment? → And how big a deal is this, considering how much search I’ve done?

Copyright © 2002-2004, Andrew Moore
WSARE 3.0

All historical data

Today's Environment

Today's Cases

- All-dimensions Trees
- Racing Randomization
- Differential Randomization

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
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!
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!

• 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