Outlier and Anomaly Detection

Case Study: Biosurveillance

Excerpts/Adaptations from:
Detection Algorithms for Biosurveillance: A tutorial

Andrew Moore, CMU tutorial:
http://www.autonlab.org/tutorials/biosurv.html

Also Excerpts/Adaptations from:
Russell and Norvig, *Intro. to AI* slides
Biosurveillance

• Attempt to quickly detect significant disease outbreaks
  • Allows for quick action to try to contain disease

• Approach:
  
  Collect data from many sources …
  (such as emergency room visits, purchases of particular types of over-the-counter medicine, school absences, pharmacy sales, etc.)

  …and recognize statistical anomalies
Biosurveillance is a difficult topic

• April 6, 2006: “House Government Reform Committee scolds Centers for Disease Control Prevention” for “technologically backward” system of biosurveillance [Government Health IT]

http://govhealthit.com/article93982-04-06-06-Web
What you’ll learn about

• Noticing events in bio-event time series
• Tracking many series at once
• Detecting geographic hotspots
• Finding emerging new patterns
What you’ll learn about

• Noticing events in bio-event time series
• Tracking many series at once
• Detecting geographic hotspots
• Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…
What you’ll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

*Boring Names.*
What you’ll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

These are all powerful statistical methods, which means they all have to have one thing in common…

**Boring Names.**

- Univariate Anomaly Detection
- Multivariate Anomaly Detection
- Spatial Scan Statistics
- WSHARE
What you’ll learn about

• Noticing events in bio-event time series
• Tracking many series at once
• Detecting geographic hotspots
• Finding emerging new patterns
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
(When) is there an anomaly?
(When) is there an anomaly?

This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?
(When) is there an anomaly?

This is a time series of counts of primary-physician visits in data from Norfolk in December 2001. I added a fake outbreak, starting at a certain date. Can you guess when?

Here (much too high for a Friday)

(Ramp attack)
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
An easy case: Control Charts

Signal

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
Control Charts on the Norfolk Data
Control Charts on the Norfolk Data
Looking at changes from yesterday
Looking at changes from yesterday
Looking at changes from yesterday
We need a happy medium:

Control Chart:
Too insensitive to recent changes

Change from yesterday:
Too sensitive to recent changes
Moving Average
Moving Average
Moving Average
Moving Average

Looks better. But how can we be quantitative about this?
## Algorithm Performance

<table>
<thead>
<tr>
<th></th>
<th>Fraction of Spikes Detected</th>
<th>Days to Detect Spikes Detected</th>
<th>Allowing one False Alarm per TWO weeks</th>
<th>Allowing one False Alarm per SIX weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard control chart</td>
<td>0.39</td>
<td>3.47</td>
<td>0.22</td>
<td>4.13</td>
</tr>
<tr>
<td>using yesterday</td>
<td>0.14</td>
<td>3.83</td>
<td>0.1</td>
<td>4.7</td>
</tr>
</tbody>
</table>
## Algorithm Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fraction of spikes detected</th>
<th>Days to detect a ramp attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard control chart using yesterday</td>
<td>0.39</td>
<td>3.47</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>3.83</td>
</tr>
<tr>
<td>Moving Average 7</td>
<td>0.58</td>
<td>2.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>0.51</td>
<td>3.31</td>
</tr>
</tbody>
</table>
# Algorithm Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fraction of Spikes Detected</th>
<th>Days to Detect a Ramp Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Control Chart</td>
<td>0.39</td>
<td>3.47</td>
</tr>
<tr>
<td>Using Yesterday</td>
<td>0.14</td>
<td>3.83</td>
</tr>
<tr>
<td>Moving Average 3</td>
<td>0.36</td>
<td>3.45</td>
</tr>
<tr>
<td>Moving Average 7</td>
<td>0.58</td>
<td>2.79</td>
</tr>
<tr>
<td>Moving Average 56</td>
<td>0.54</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Allowing one False Alarm per TWO weeks...

Allowing one False Alarm per SIX weeks...
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.
## Algorithm Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Days to detect a ramp attack</th>
<th>Fraction of spikes detected</th>
<th>Days to detect a ramp attack</th>
<th>Fraction of spikes detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard control chart</td>
<td>3.47</td>
<td>0.39</td>
<td>4.13</td>
<td>0.22</td>
</tr>
<tr>
<td>using yesterday</td>
<td>3.83</td>
<td>0.14</td>
<td>4.7</td>
<td>0.1</td>
</tr>
<tr>
<td>Moving Average 3</td>
<td>3.45</td>
<td>0.36</td>
<td>3.79</td>
<td>0.33</td>
</tr>
<tr>
<td>Moving Average 7</td>
<td>2.79</td>
<td>0.58</td>
<td>3.31</td>
<td>0.51</td>
</tr>
<tr>
<td>Moving Average 56</td>
<td>2.72</td>
<td>0.54</td>
<td>3.54</td>
<td>0.44</td>
</tr>
<tr>
<td>hours_of_daylight</td>
<td>2.73</td>
<td>0.58</td>
<td>3.9</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Day-of-week effects

Fit a day-of-week component

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

E.G:

\[
\begin{align*}
\delta_{\text{mon}} &= +5.42 \\
\delta_{\text{tue}} &= +2.20 \\
\delta_{\text{wed}} &= +3.33 \\
\delta_{\text{thu}} &= +3.10 \\
\delta_{\text{fri}} &= +4.02 \\
\delta_{\text{sat}} &= -12.2 \\
\delta_{\text{sun}} &= -23.42
\end{align*}
\]

A simple form of ANOVA

(ANOVA: “Analysis of Variance” -- technique for testing differences among several means)
Regression using Hours-in-day & IsMonday
Regression using Hours-in-day & IsMonday
# Algorithm Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Days to detect a ramp attack</th>
<th>Fraction of spikes detected</th>
<th>Days to detect a ramp attack</th>
<th>Fraction of spikes detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard control chart</td>
<td>3.47</td>
<td>0.39</td>
<td>4.13</td>
<td>0.22</td>
</tr>
<tr>
<td>using yesterday</td>
<td>3.83</td>
<td>0.14</td>
<td>4.7</td>
<td>0.1</td>
</tr>
<tr>
<td>Moving Average 3</td>
<td>3.45</td>
<td>0.36</td>
<td>3.79</td>
<td>0.33</td>
</tr>
<tr>
<td>Moving Average 7</td>
<td>2.79</td>
<td>0.58</td>
<td>3.31</td>
<td>0.51</td>
</tr>
<tr>
<td>Moving Average 56</td>
<td>2.72</td>
<td>0.54</td>
<td>3.54</td>
<td>0.44</td>
</tr>
<tr>
<td>hours_of_daylight</td>
<td>2.73</td>
<td>0.58</td>
<td>3.9</td>
<td>0.43</td>
</tr>
<tr>
<td>hours_of_daylight is_mon</td>
<td>2.25</td>
<td>0.7</td>
<td>3.12</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Regression using Mon-Tue
# Algorithm Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fraction of Spikes Detected</th>
<th>Days todetect a ramp attack</th>
<th>Days to detect a ramp attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard control chart</td>
<td>0.39</td>
<td>3.47</td>
<td>0.22</td>
</tr>
<tr>
<td>using yesterday</td>
<td>0.14</td>
<td>3.83</td>
<td>0.1</td>
</tr>
<tr>
<td>Moving Average 3</td>
<td>0.36</td>
<td>3.45</td>
<td>0.33</td>
</tr>
<tr>
<td>Moving Average 7</td>
<td>0.58</td>
<td>2.79</td>
<td>0.51</td>
</tr>
<tr>
<td>Moving Average 56</td>
<td>0.54</td>
<td>2.72</td>
<td>0.44</td>
</tr>
<tr>
<td>hours_of_daylight</td>
<td>0.58</td>
<td>2.73</td>
<td>0.43</td>
</tr>
<tr>
<td>hours_of_daylight is_mon</td>
<td>0.7</td>
<td>2.25</td>
<td>0.57</td>
</tr>
<tr>
<td>hours_of_daylight is_mon ... is_tue</td>
<td>0.72</td>
<td>1.83</td>
<td>0.57</td>
</tr>
<tr>
<td>hours_of_daylight is_mon ... is_sat</td>
<td>0.77</td>
<td>2.11</td>
<td>0.59</td>
</tr>
</tbody>
</table>
CUSUM

• **CUSUM** Statistics

• Keep a running sum of “surprises”: a sum of excesses each day over the prediction

• When this sum exceeds threshold, signal alarm and reset sum
CUSUM
### Algorithm Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Spike Detection</th>
<th>Days to detect a ramp attack when allowing one false alarm per two weeks</th>
<th>Days to detect a ramp attack when allowing one false alarm per six weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard control chart</td>
<td>0.39</td>
<td>3.47</td>
<td>0.22</td>
</tr>
<tr>
<td>using yesterday</td>
<td>0.14</td>
<td>3.83</td>
<td>0.1</td>
</tr>
<tr>
<td>Moving Average 3</td>
<td>0.36</td>
<td>3.45</td>
<td>0.33</td>
</tr>
<tr>
<td>Moving Average 7</td>
<td>0.58</td>
<td>2.79</td>
<td>0.51</td>
</tr>
<tr>
<td>Moving Average 56</td>
<td>0.54</td>
<td>2.72</td>
<td>0.44</td>
</tr>
<tr>
<td>hours_of_daylight</td>
<td>0.58</td>
<td>2.73</td>
<td>0.43</td>
</tr>
<tr>
<td>hours_of_daylight is_mon</td>
<td>0.7</td>
<td>2.25</td>
<td>0.57</td>
</tr>
<tr>
<td>hours_of_daylight is_mon ... is_tue</td>
<td>0.72</td>
<td>1.83</td>
<td>0.57</td>
</tr>
<tr>
<td>hours_of_daylight is_mon ... is_sat</td>
<td>0.77</td>
<td>2.11</td>
<td>0.59</td>
</tr>
<tr>
<td>CUSUM</td>
<td>0.45</td>
<td>2.03</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Other state-of-the-art methods

- Wavelets
- Change-point detection
- Kalman filters
- Hidden Markov Models
What you’ll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns
Multiple Signals
Multivariate Signals

(relevant to inhalational diseases)
Multi Source Signals

Footprint of Influenza in Routinely Collected Data

- Lab
- Flu
- WebMD
- School
- Cough & Cold
- Throat
- Resp
- Viral
- Death

weeks
What if you’ve got multiple signals?

Red: Cough Sales
Blue: ED Respiratory Visits

Idea One:
Simply treat it as two separate alarm-from-signal problems.

…Question: why might that not be the best we can do?
Another View

Red: Cough Sales
Blue: ED Respiratory Visits

Question: why might that not be the best we can do?

Signal

Cough Sales
ED Respiratory Visits
Another View

Red: Cough Sales
Blue: ED Respiratory Visits

This should be an anomaly

Question: why might that not be the best we can do?
Good Practical Idea:
Model the joint with a Gaussian

This is a sensible N-dimensional SQC (statistical quality control)

...But you can also do N-dimensional modeling of dynamics (leads to the idea of Kalman Filter model)
But what if joint N-dimensional distribution is highly non-Gaussian?

Red: Cough Sales
Blue: ED Respiratory Visits

Signal

Cough Sales
ED Respiratory Visits
What you’ll learn about

- Noticing events in bio-event time series
- Tracking many series at once
- Detecting geographic hotspots
- Finding emerging new patterns

Univariate Anomaly Detection
Multivariate Anomaly Detection
Spatial Scan Statistics
WSARE
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

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

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

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%)

So... is that a big deal?
Evaluated with Score function (e.g. Kulldorf’s score)
One Step of Spatial Scan

I have a population of 5300 of whom 53 are sick (1%)
[Score = 1.4]

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

So... is that a big deal?
Evaluated with Score function (e.g. Kulldorf’s score)
Many Steps of Spatial Scan

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

[Score = 1.4]

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

[Score = 9.3]

So... is that a big deal?

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

Standard approach:
1. Compute the likelihood of the data given the hypothesis that the rate of occurrence is uniform everywhere, $L_0$
2. For some geographical region, $W$, compute the likelihood that the rate of occurrence is uniform at one level inside the region and uniform at another level outside the region, $L(W)$.
3. Compute the likelihood ratio, $L(W)/L_0$
4. Repeat for all regions, and find the largest likelihood ratio. This is the scan statistic, $S^*_W$
5. Report the region, $W$, which yielded the max, $S^*_W$

See [Glaz and Balakrishnan, 99] for details
Significance testing

Standard approach:

1. Generate many randomized versions of the data set by shuffling the labels (positive instance of the phenomenon or not).

2. Compute $S^*_W$ for each randomized data set. This forms a baseline distribution for $S^*_W$ if the null hypothesis holds.

3. Compare the observed value of $S^*_W$ against the baseline distribution to determine a p-value.
Fast squares speedup

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

  \textit{If not, we can use several other speedup tricks.}

- In practice: 10-200x speedups on real and artificially generated datasets.

  \textit{Emergency Dept. dataset (600K records): 20 minutes, versus 66 hours with naïve approach.}
Fast rectangles speedup

Work in progress

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

• New “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
- New “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.
What you’ll learn about

• Noticing events in bio-event time series
• Tracking many series at once
• Detecting geographic hotspots
• Finding emerging new patterns

WSARE

Univariate Anomaly Detection
Multivariate Anomaly Detection
Spatial Scan Statistics
But there’s potentially more data than aggregates

Suppose we know that today in the ED we had...

- 421 Cases
- 78 Respiratory Cases
- 190 Males
- 32 Children
- 21 from North Suburbs
- 2 Postal workers
  (etc etc etc)

Have we made best use of all possible information?
There are so many things to look at.

- Diarrhea by Street Among Children Recent 3 hours
- Diarrhea by Neighborhood Among Elderly Recent 24 hrs
- Collapse by county Among Men Recent week
- Nyquil Sales by state Recent 30 mins
- Absenteeism by zipcode Farm Workers Recent month

Human Analysts

Massive Computer Analysis
WSARE v2.0

- What’s Strange About Recent Events?
- Designed to be easily applicable to any 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?
WSARE v2.0

- **Input:**
  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"

<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>A5</td>
<td>15217</td>
<td>70R</td>
<td>...</td>
</tr>
<tr>
<td>t3q15</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>J3</td>
<td>15222</td>
<td>70R</td>
<td>...</td>
</tr>
<tr>
<td>t5hh5</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>K9</td>
<td>15210</td>
<td>70R</td>
<td>...</td>
</tr>
</tbody>
</table>

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 70
### WSARE v2.0

#### Inputs:
1. Date/time-indexed biosurveillance-relevant data stream
2. Time Window Length
3. Which attributes to use?
3. And here’s how seriously you should take it

#### 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>2%</td>
<td>70R</td>
</tr>
<tr>
<td>t3q15</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>J3</td>
<td>2%</td>
<td>70R</td>
</tr>
<tr>
<td>t5hh5</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>2%</td>
<td>70R</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>
</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></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>

Copyright © 2002, 2003, Andrew Moore

Biosurveillance Detection Algorithms: Slide 73
**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?”

<table>
<thead>
<tr>
<th>DATE_AD</th>
<th>ICD9</th>
<th>GENDER</th>
<th>place2</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/9/00</td>
<td>786.05</td>
<td>3 F</td>
<td>s-e</td>
</tr>
<tr>
<td>12/9/00</td>
<td>789</td>
<td>1 F</td>
<td>s-e</td>
</tr>
<tr>
<td>12/9/00</td>
<td>789</td>
<td>1 M</td>
<td>n-w</td>
</tr>
<tr>
<td>12/9/00</td>
<td>786.05</td>
<td>3 M</td>
<td>s-e</td>
</tr>
<tr>
<td>12/16/00</td>
<td>787.02</td>
<td>2 M</td>
<td>n-e</td>
</tr>
<tr>
<td>12/16/00</td>
<td>782.1</td>
<td>4 F</td>
<td>s-w</td>
</tr>
<tr>
<td>12/16/00</td>
<td>789</td>
<td>1 M</td>
<td>s-e</td>
</tr>
<tr>
<td>12/16/00</td>
<td>786.09</td>
<td>3 M</td>
<td>n-w</td>
</tr>
<tr>
<td>12/23/00</td>
<td>789.09</td>
<td>1 M</td>
<td>s-w</td>
</tr>
<tr>
<td>12/23/00</td>
<td>789.09</td>
<td>1 F</td>
<td>s-w</td>
</tr>
<tr>
<td>12/23/00</td>
<td>782.1</td>
<td>4 M</td>
<td>n-w</td>
</tr>
<tr>
<td>12/23/00</td>
<td>786.09</td>
<td>3 M</td>
<td>s-e</td>
</tr>
<tr>
<td>12/23/00</td>
<td>786.09</td>
<td>3 M</td>
<td>s-e</td>
</tr>
<tr>
<td>12/23/00</td>
<td>780.9</td>
<td>2 F</td>
<td>n-w</td>
</tr>
<tr>
<td>12/23/00</td>
<td>V40.9</td>
<td>7 M</td>
<td>s-w</td>
</tr>
</tbody>
</table>
Simple WSARE

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

<table>
<thead>
<tr>
<th>DATE_AD</th>
<th>ICD9</th>
<th>PRODROME</th>
<th>GENDER</th>
<th>place2</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/9/00</td>
<td>786.05</td>
<td>3 F</td>
<td>s-e</td>
<td>...</td>
</tr>
<tr>
<td>12/9/00</td>
<td>789</td>
<td>1 F</td>
<td>s-e</td>
<td>...</td>
</tr>
<tr>
<td>12/9/00</td>
<td>789</td>
<td>1 M</td>
<td>n-w</td>
<td>...</td>
</tr>
<tr>
<td>12/9/00</td>
<td>786.05</td>
<td>3 M</td>
<td>s-e</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>12/16/00</td>
<td>787.02</td>
<td>2 M</td>
<td>n-e</td>
<td>...</td>
</tr>
<tr>
<td>12/16/00</td>
<td>782.1</td>
<td>4 F</td>
<td>s-w</td>
<td>...</td>
</tr>
<tr>
<td>12/16/00</td>
<td>789</td>
<td>1 M</td>
<td>s-e</td>
<td>...</td>
</tr>
<tr>
<td>12/16/00</td>
<td>786.09</td>
<td>3 M</td>
<td>n-w</td>
<td>...</td>
</tr>
<tr>
<td>12/23/00</td>
<td>789.09</td>
<td>1 F</td>
<td>s-w</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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

Table 1: A sample 2x2 Contingency Table

<table>
<thead>
<tr>
<th>Age_Decile = 3</th>
<th>$C_{today}$</th>
<th>$C_{other}$</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!
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?
**WSARE v2.0**

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

- **Output s:**
  1. Here are the records that most surprise me
  2. Here’s why
  3. And here’s how seriously you should take it

---

**Table:**

<table>
<thead>
<tr>
<th>Primary Key</th>
<th>Data</th>
<th>Time</th>
<th>Hospital</th>
<th>CD-ICD</th>
<th>Prodrome</th>
<th>Gender</th>
<th>Age</th>
<th>Large Scale</th>
<th>Mediu Scale</th>
<th>Fine Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>h6r32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t3q15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t5hh5</td>
<td>6/2/2 14:15</td>
<td>Smithfield</td>
<td>622 Respiratory</td>
<td>80 SE</td>
<td>15217</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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 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
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”
Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:
    \[ P(X_i | \text{Parents}(X_i)) \]

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over \( X_i \) for each combination of parent values

(from Russell and Norvig)
Example

- Topology of network encodes conditional independence assertions:

  - Weather is independent of the other variables
  - Toothache and Catch are conditionally independent given Cavity

(from Russell and Norvig)
Example

• I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

• Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

• Network topology reflects "causal" knowledge:
  • A burglar can set the alarm off
  • An earthquake can set the alarm off
  • The alarm can cause Mary to call
  • The alarm can cause John to call

(from Russell and Norvig)
WSARE 3.0 (software developed by A. Moore, et al, at CMU)

All historical data
WSARE 3.0

(software developed by A. Moore, et al, at CMU)
WSARE 3.0 (software developed by A. Moore, et al, at CMU)

All historical data → Today’s Environment → What *should* be happening today?
WSARE 3.0 (software developed by A. Moore, et al, at CMU)

- 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?

Software developed by A. Moore, et al, at CMU
All historical data → Today’s Environment → Today’s Cases

Expensive

Cheap

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 (software developed by A. Moore, et al, at CMU)

All historical data

Today's Environment

Today's Cases

• Racing Randomization
• Differential Randomization

• All-dimensions Trees

• RADSEARCH

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?
Results on Simulation
Results on Simulation
Conclusion

- **WSARE:** 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!
• 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 historical data with Bayesia Networks to allow conditioning on unique features of today.
• Computationally intense unless we’re tricksy!
The End... 😊