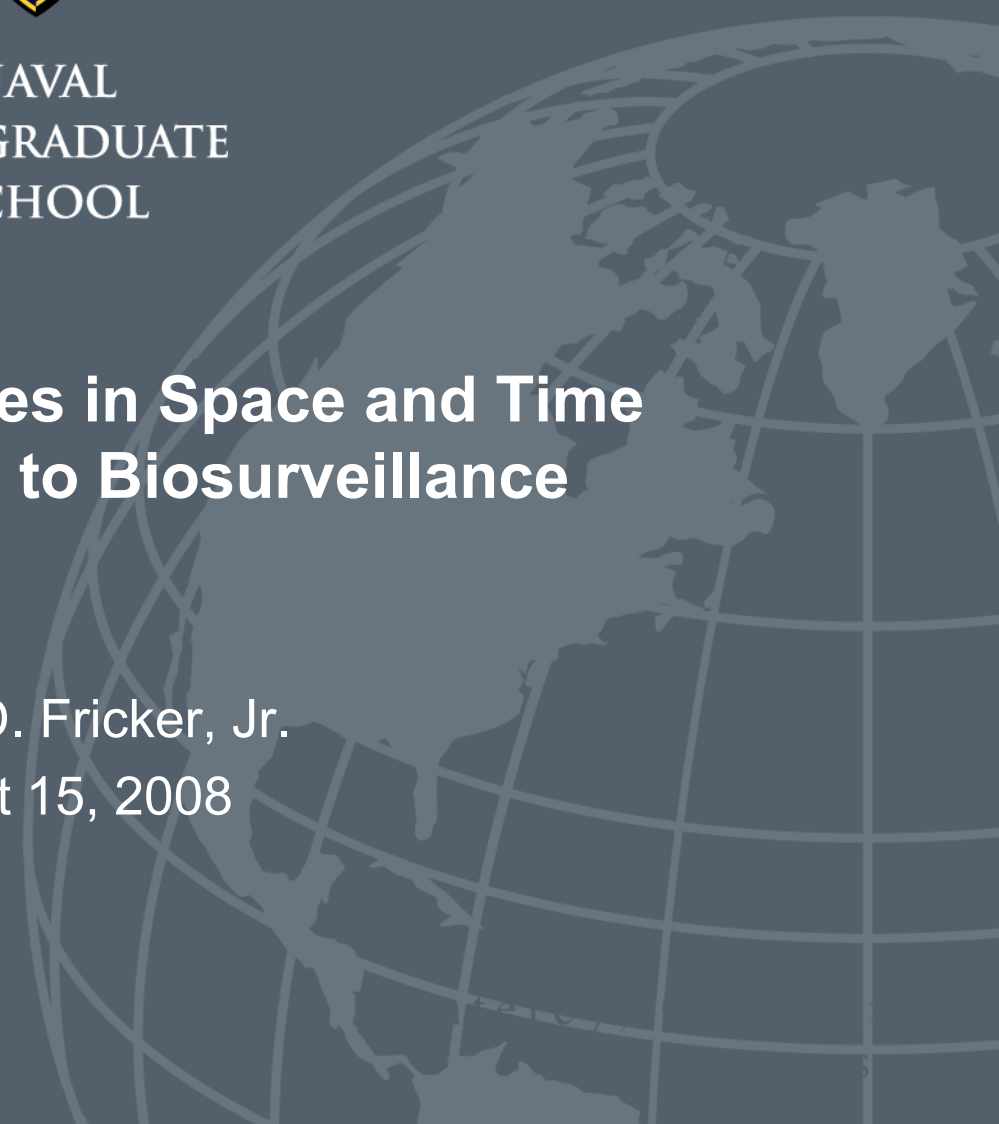




NAVAL  
POSTGRADUATE  
SCHOOL

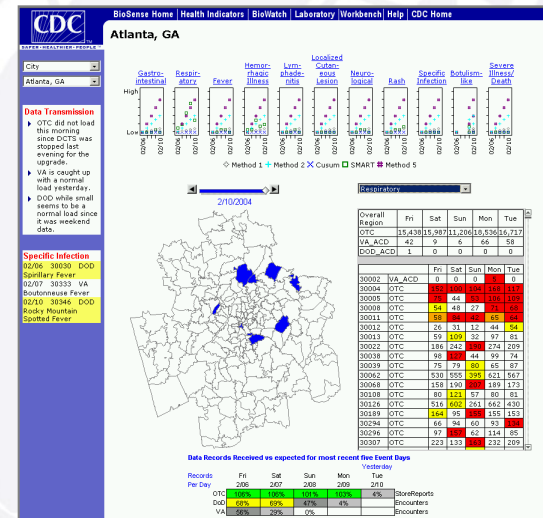
# Detecting Anomalies in Space and Time with Application to Biosurveillance

Ronald D. Fricker, Jr.  
August 15, 2008



# Motivating Problem: Biosurveillance

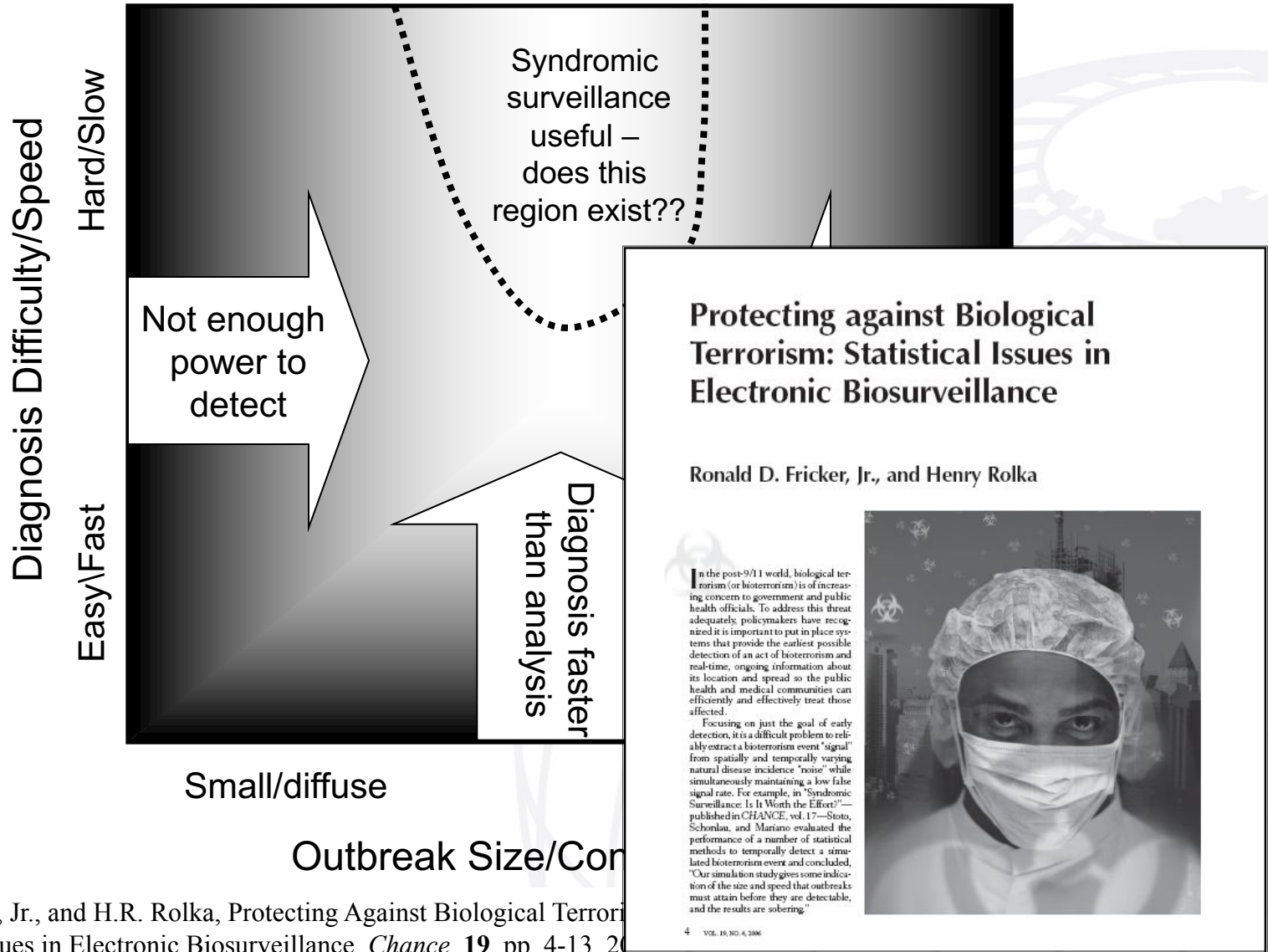
- “...surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response.” [1]
- Biosurveillance uses now encompass both “early event detection” and “situational awareness”



[1] CDC ([www.cdc.gov/epo/dphsi/syndromic.htm](http://www.cdc.gov/epo/dphsi/syndromic.htm), accessed 5/29/07)

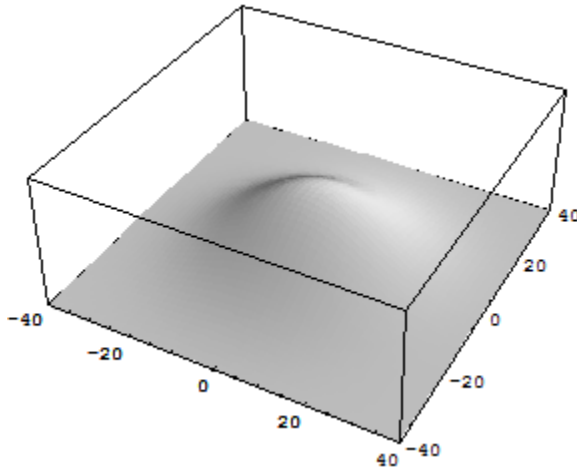


- *Early event detection*: gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak
- *Situational awareness*: the real-time analysis and display of health data to monitor the location, magnitude, and spread of an outbreak

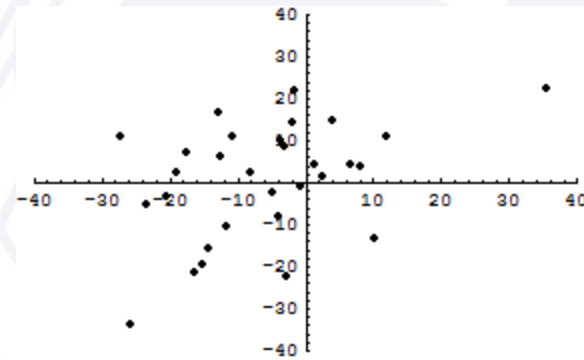


Fricker, R.D., Jr., and H.R. Rolka, Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, 19, pp. 4-13, 2006

- ER patients come from surrounding area
  - On average, 30 per day
    - More likely from closer distances
  - Outbreak occurs at (20,20)
    - Number of patients increase linearly by day after outbreak



(Unobservable) distribution of ER patients' home addresses



Observed distribution of ER patients' locations



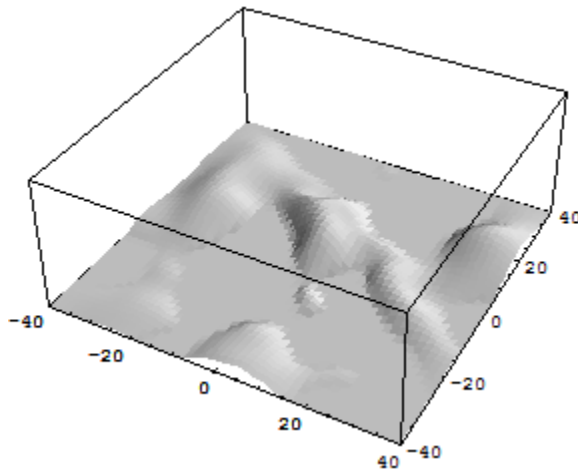
# A Couple of Major Assumptions

- Can geographically locate individuals *in a medically meaningful way*
  - Non-trivial problem
  - Data not currently available
- Data is reported in a timely and consistent manner
  - Public health community working this problem, but not solved yet
- Assuming the above problems away...



# Idea: Look at Differences in Kernel Density Estimates

- Construct kernel density estimate (KDE) of “normal” disease incidence using  $N$  historical observations
- Compare to KDE of most recent  $w+1$  obs



***But how to know when to signal?***



# Solution: Repeated Two-Sample Rank (RTR) Procedure

- Sequential hypothesis test of estimated density heights
- Compare estimated density heights of recent data against heights of set of historical data
  - Single density estimated via KDE on *combined* data
- If no change, heights uniformly distributed
  - Use nonparametric test to assess





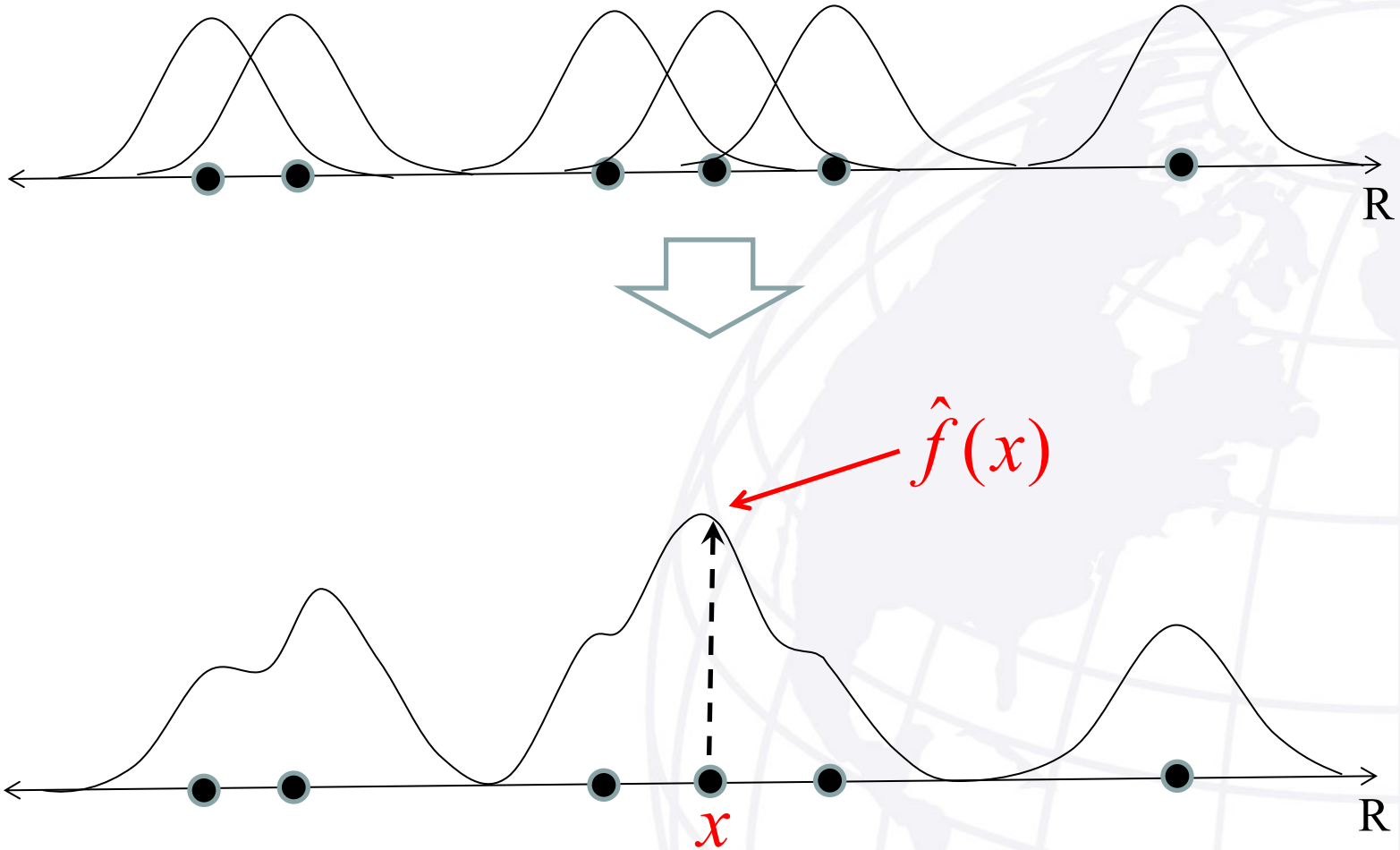
- Let  $\mathbf{X}_i = \{X_{1i}, X_{2i}\}$  be a sequence of bivariate observations
  - E.g., latitude and longitude of a case
- Assume a historical sequence  $\mathbf{X}_{1-N}, \dots, \mathbf{X}_0$  is available
  - Distributed iid according to  $f_0$
- Followed by  $\mathbf{X}_1, \mathbf{X}_2, \dots$  which may change from  $f_0$  to  $f_1$  at any time
- Densities  $f_0$  and  $f_1$  unknown

- Consider the  $w+1$  most recent data points
- At each time period estimate the density

$$\hat{f}_n(\mathbf{x}) = \begin{cases} \frac{1}{N+n} \sum_{i=1-N}^n k_h(\mathbf{x}, \mathbf{X}_i), & n < w+1 \\ \frac{1}{N+w+1} \sum_{i=n-w-N-1}^n k_h(\mathbf{x}, \mathbf{X}_i), & n \geq w+1 \end{cases}$$

where  $k$  is a kernel function on  $\mathbb{R}^2$  with bandwidth set to  $h_i = \sigma_i (1/(N+w+1))^{1/6}$

# Illustrating Kernel Density Estimation (in one dimension)



- The density estimate is evaluated at each historical and new point

- For  $n < w+1$

$$\underbrace{\hat{f}_n(\mathbf{X}_{1-N}), \dots, \hat{f}_n(\mathbf{X}_0)}_{\text{historical observations}}, \underbrace{\hat{f}_n(\mathbf{X}_1), \dots, \hat{f}_n(\mathbf{X}_n)}_{\text{new observations}}$$

- For  $n \geq w+1$

$$\underbrace{\hat{f}_n(\mathbf{X}_{n-w-N-1}), \dots, \hat{f}_n(\mathbf{X}_{n-w-1})}_{\text{historical observations}}, \underbrace{\hat{f}_n(\mathbf{X}_{n-w}), \dots, \hat{f}_n(\mathbf{X}_n)}_{\text{new observations}}$$



# Under the Null, Estimated Density Heights are Exchangeable

- *Theorem:* The RTR procedure is asymptotically distribution free
  - I.e., the estimated density heights are exchangeable, so all rankings are equally likely
  - Proof: See Fricker and Chang (2008)
- Means can do a hypothesis test on the ranks each time an observation arrives
  - Signal change in distribution first time test rejects

# Comparing Distributions of Heights

- Compute empirical distributions of the two sets of estimated heights:

$$\hat{J}_n(z) = \frac{1}{w+1} \sum_{i=n-w}^n I \left\{ \hat{f}_n(\mathbf{X}_i) \leq z \right\},$$

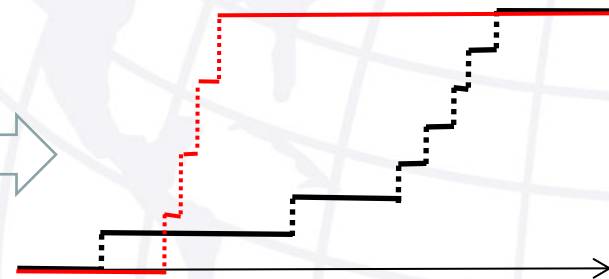
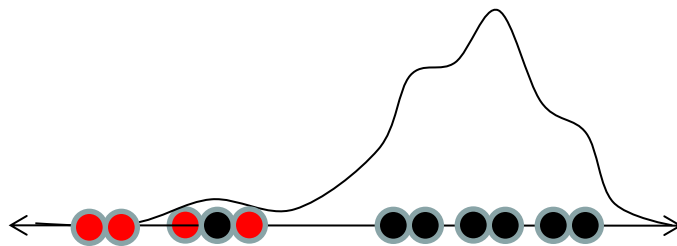
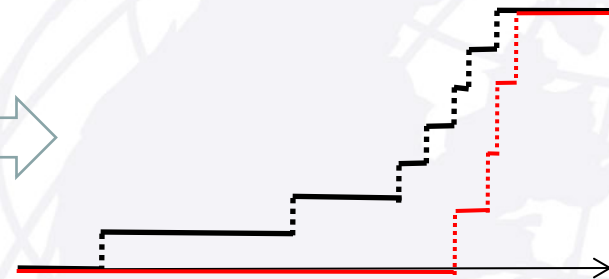
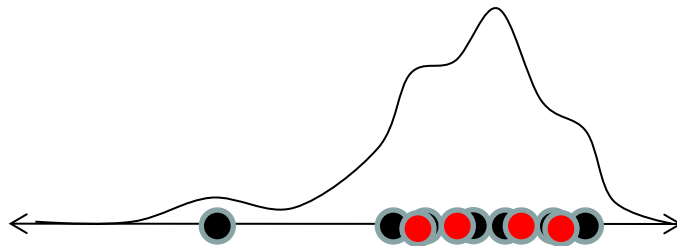
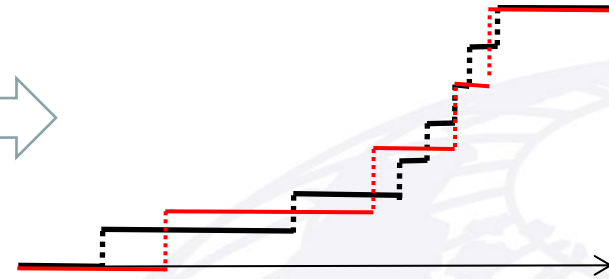
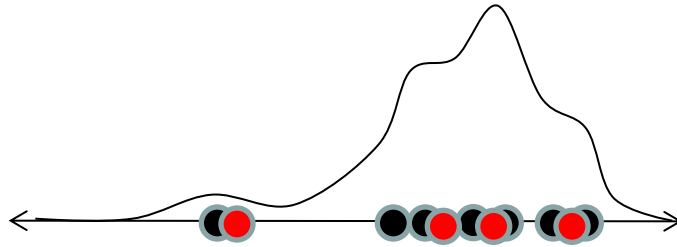
$$\hat{H}_n(z) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} I \left\{ \hat{f}_n(\mathbf{X}_i) \leq z \right\}$$

- Use Kolmogorov-Smirnov test to assess:

$$S_n = \max_z \left| \hat{J}_n(z) - \hat{H}_n(z) \right|$$

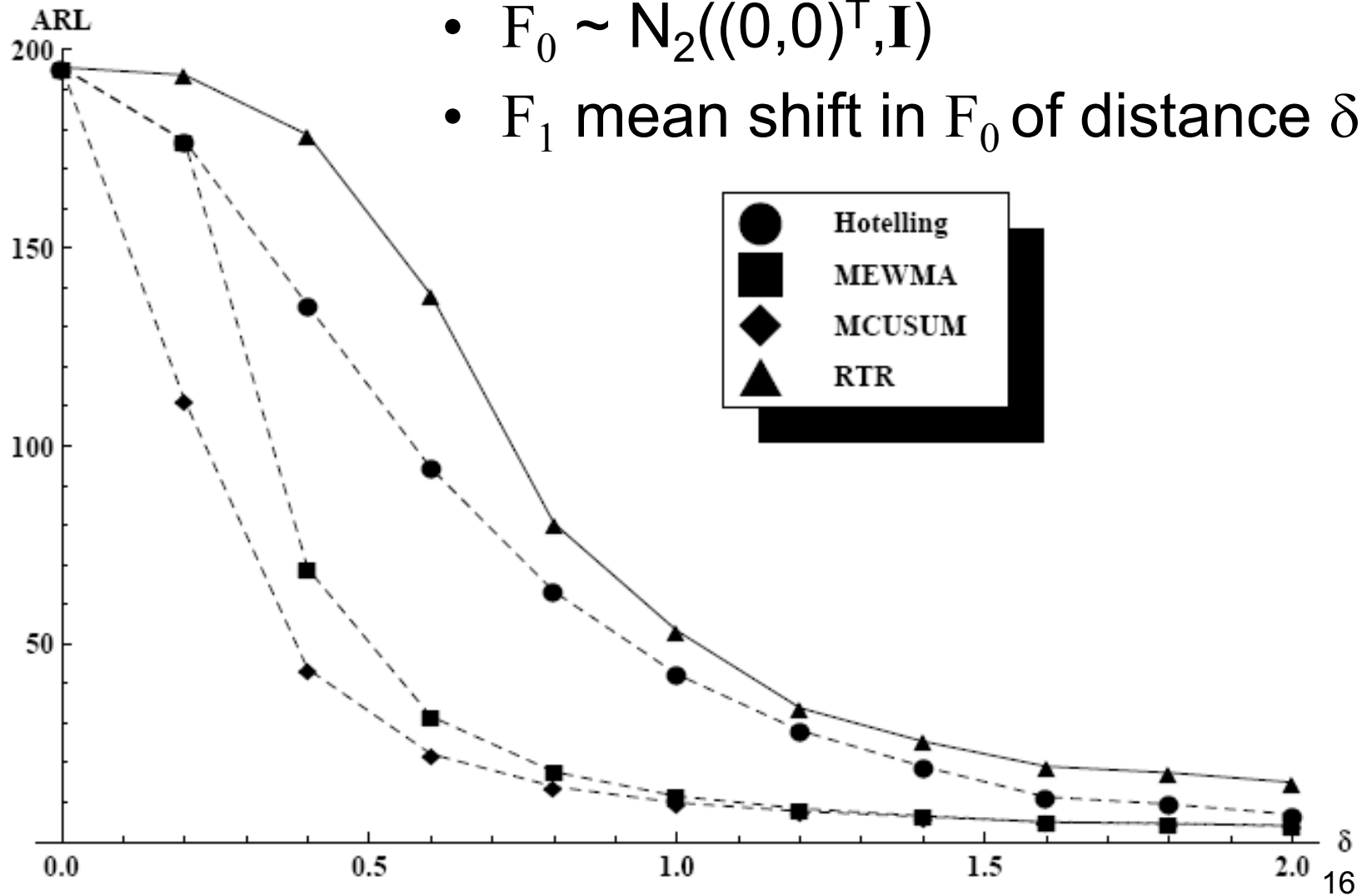
- Signal at time  $t = \min \{ n : S_n > c \}$

# Illustrating Changes in Distributions (again, in one dimension)



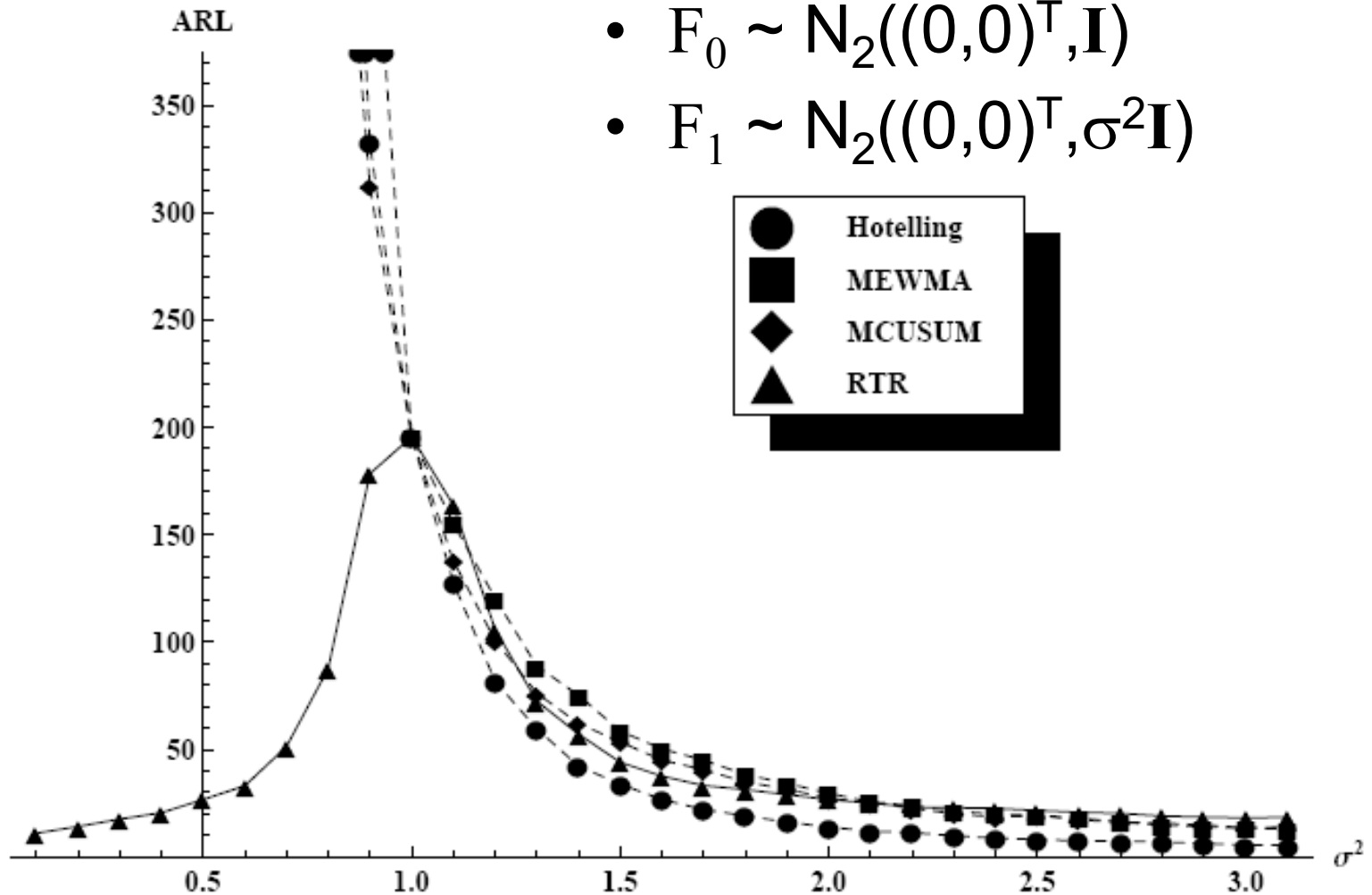
# Performance Comparison #1

- $F_0 \sim N_2((0,0)^T, \mathbf{I})$
- $F_1$  mean shift in  $F_0$  of distance  $\delta$





# Performance Comparison #2



- $F_0 \sim N_2((0,0)^T, \mathbf{I})$
- $F_1 \sim N_2((0,0)^T, \sigma^2 \mathbf{I})$



- How to find  $c$ ?
  - Use ARL approximation based on Poisson clumping heuristic:

$$A \approx \left[ \left( \frac{6.16c [c + 0.5/(w + 1)]}{1 + (w + 1)/N} \right) \exp \left\{ -2 \left( c + \frac{1}{2(w + 1)} \right)^2 \left( \frac{1}{w + 1} + \frac{1}{N} \right)^{-1} \right\} \right]^{-1}$$

- Example:  $c=0.07754$  with  $N=1,350$  and  $w+1=250$  gives  $A=900$ 
  - If 30 observations per day, gives average time between (false) signals of 30 days

- At signal, calculate optimal kernel density estimates and plot pointwise differences

$$\Delta_n(\mathbf{x}) = \max\left(\delta, \hat{h}_n(\mathbf{x}) - \hat{g}_n(\mathbf{x})\right)$$

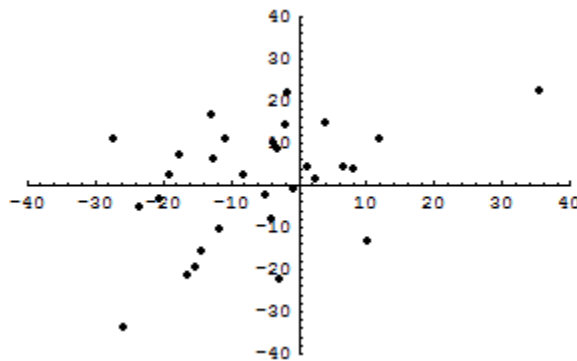
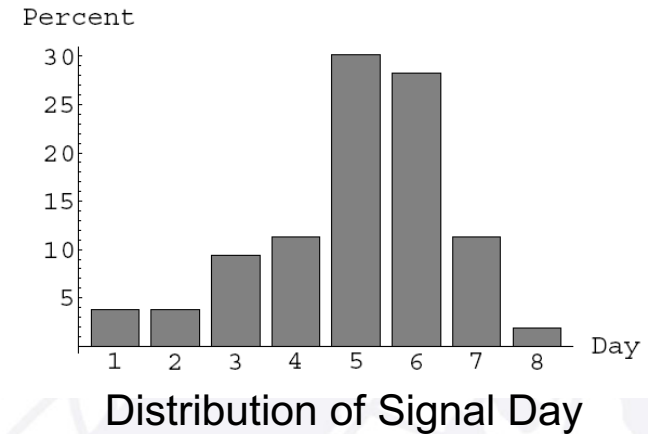
where

$$\hat{h}_n(\mathbf{x}) = \frac{1}{w+1} \sum_{i=n-w}^n k_h(\mathbf{x}, \mathbf{X}_i)$$

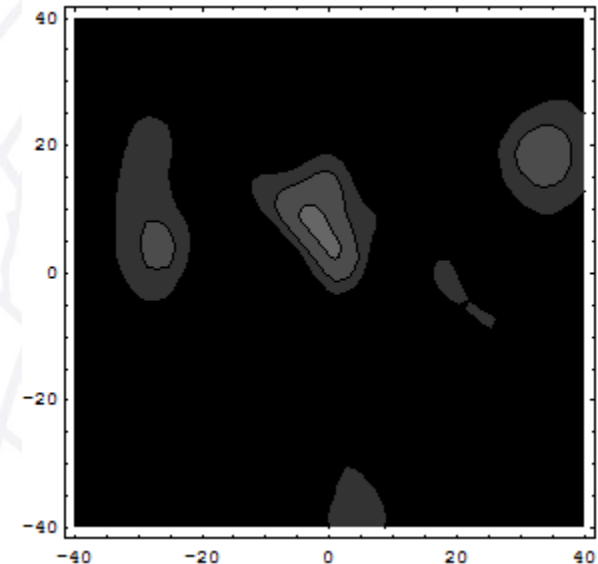
$$\hat{g}_n(\mathbf{x}) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} k_h(\mathbf{x}, \mathbf{X}_i)$$

and  $h_i = \sigma_i \left(\frac{1}{w+1}\right)^{1/6}$  or  $h_i = \sigma_i \left(\frac{1}{N}\right)^{1/6}$

- Assess performance by simulating outbreak multiple times, record when RTR signals
  - Signaled middle of day 5 on average
  - By end of 5<sup>th</sup> day, 15 outbreak and 150 non-outbreak observations
  - From previous example:



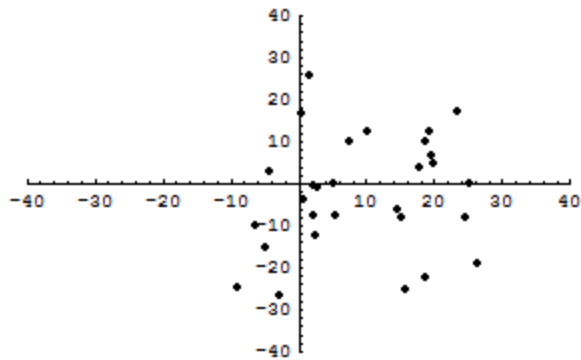
Daily Data



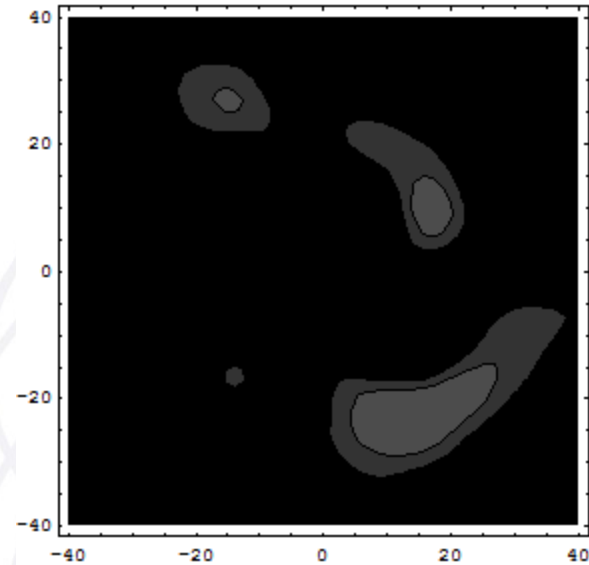
Outbreak Signaled on  
Day 7 (obs'n # 238)



# Same Scenario, Another Sample

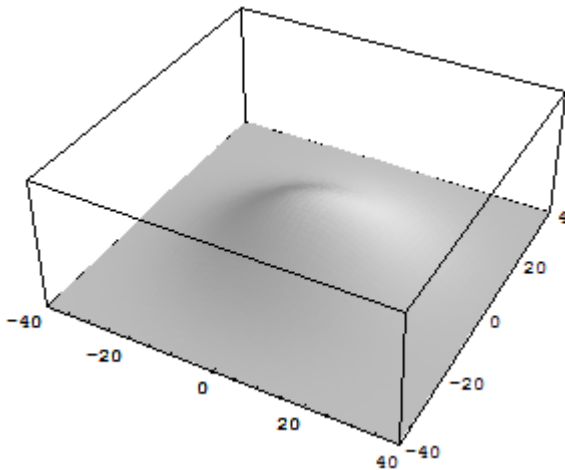


Daily Data

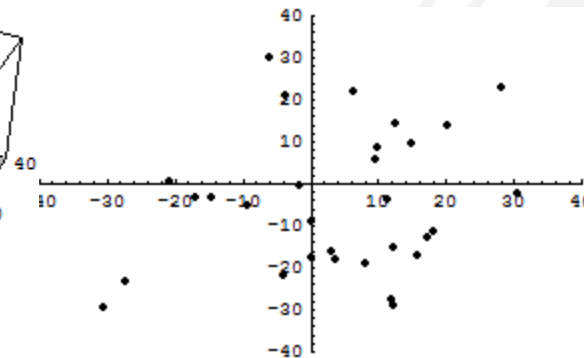


Outbreak Signaled on  
Day 5 (obs'n # 165)

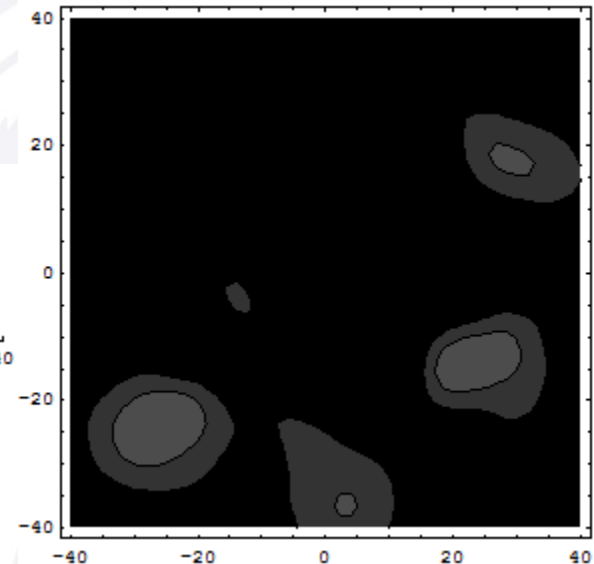
- Normal disease incidence  $\sim N(\{0,0\}^t, \sigma^2 \mathbf{I})$  with  $\sigma=15$ 
  - Expected count of 30 per day
- Outbreak incidence  $\sim N(\{20,20\}^t, 2.2d^2 \mathbf{I})$ 
  - $d$  is the day of outbreak
  - Expected count is  $30+d^2$  per day



Unobserved outbreak distribution



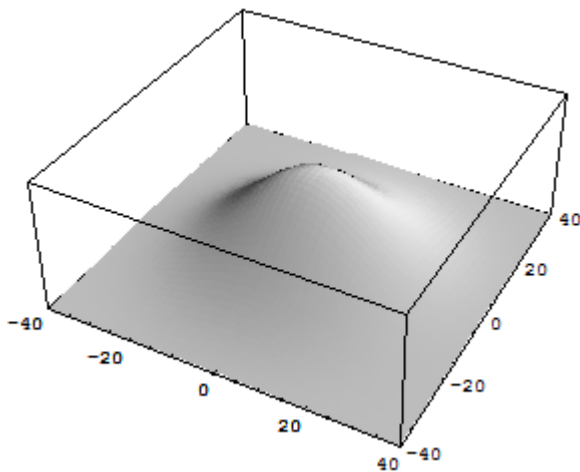
Daily data



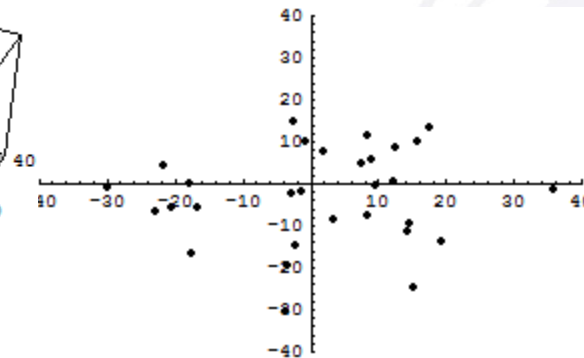
Outbreak signaled on day 1 (obs'n # 2)

(On average, signaled on day 3-1/2)

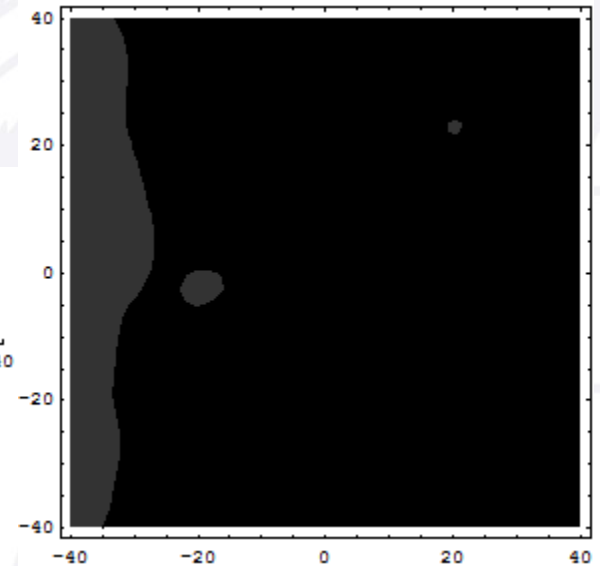
- Normal disease incidence  $\sim N(\{0,0\}^t, \sigma^2 \mathbf{I})$  with  $\sigma=15$ 
  - Expected count of 30 per day
- Outbreak sweeps across region from left to right
  - Expected count is  $30+64$  per day



Unobserved outbreak  
distribution



Daily data



Outbreak signaled on  
day 1 (obs'n # 11)

(On average, signaled 1/3  
of way into day 1)



# Advantages and Disadvantages

- Advantages
  - Methodology supports both biosurveillance goals: early event detection *and* situational awareness
  - Incorporates observations sequentially (singly)
    - Most other methods use aggregated data
  - Can be used for more than two dimensions
- Disadvantage?
  - Can't distinguish increase distributed according to  $f_0$ 
    - Unlikely for bioterrorism attack?
    - Won't detect an general increase in background disease incidence rate
      - E.g., Perhaps caused by an increase in population
      - In this case, advantage not to detect





## Detection Algorithm Development and Assessment:

- Fricker, R.D., Jr., and J.T. Chang, The Repeated Two-Sample Rank Procedure: A Multivariate Nonparametric Individuals Control Chart (in draft).
- Fricker, R.D., Jr., and J.T. Chang, A Spatio-temporal Method for Real-time Biosurveillance, *Quality Engineering* (to appear, November 2008).
- Fricker, R.D., Jr., Knitt, M.C., and C.X. Hu, Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance, *Quality Engineering* (to appear, November 2008).
- Jones, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr., A One-Sided MEWMA Chart for Health Surveillance, *Quality and Reliability Engineering International*, **24**, pp. 503-519, 2008.
- Fricker, R.D., Jr., Hegler, B.L., and D.A. Dunfee, Assessing the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms, *Statistics in Medicine*, **27**, pp. 3407-3429, 2008.
- Fricker, R.D., Jr., Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance, *Advances in Disease Surveillance*, **3:1**, 2007.

## Biosurveillance System Optimization:

- Fricker, R.D., Jr., and D. Banschbach, Optimizing a System of Threshold Detection Sensors, in submission.

## Background Information:

- Fricker, R.D., Jr., and H. Rolka, Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, pp. 4-13, 2006
- Fricker, R.D., Jr., Syndromic Surveillance, in *Encyclopedia of Quantitative Risk Assessment*, Melnick, E., and Everitt, B (eds.), John Wiley & Sons Ltd, pp. 1743-1752, 2008.