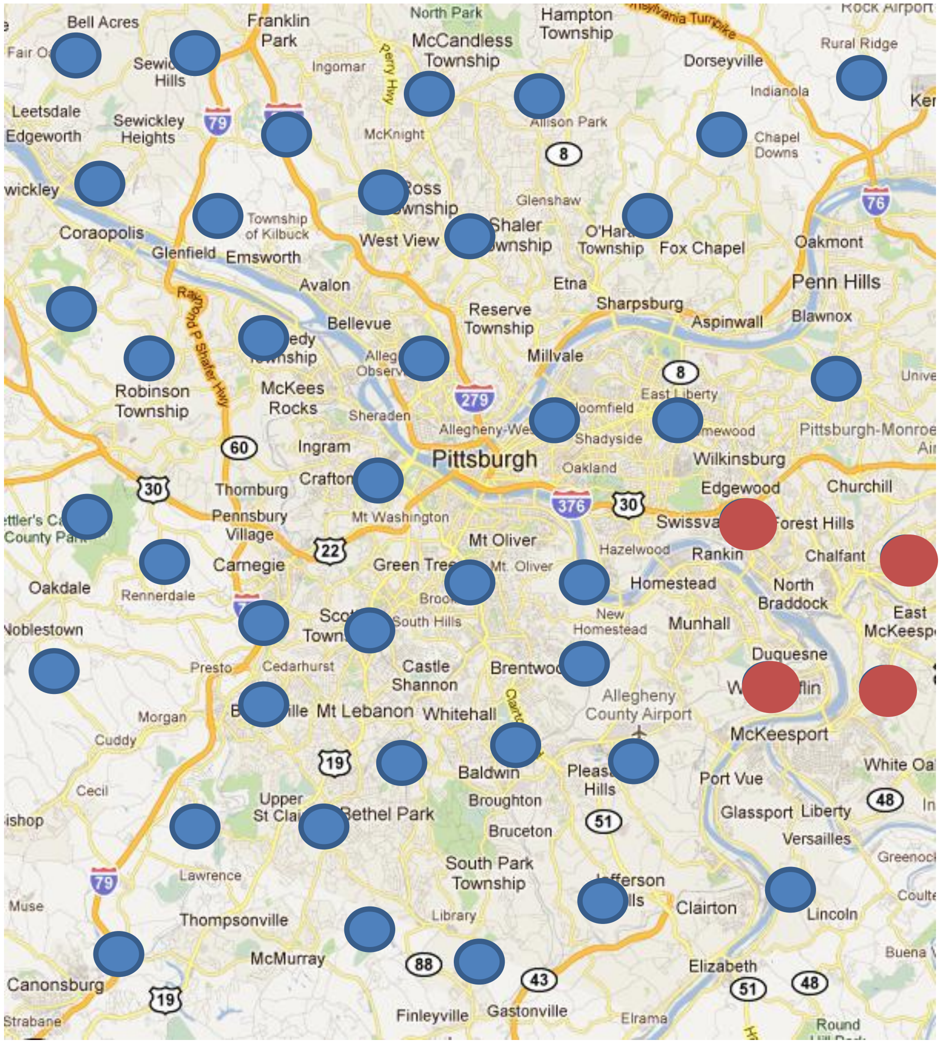
# Detecting Irregularly-Shaped Clusters with Star Scan Statistic

**Sriram Somanchi, Daniel B. Neill**
Event and Pattern Detection Laboratory
Carnegie Mellon University

This work was partially supported by NSF grants:
IIS-0916345, IIS-0911032, and IIS-0953330

# Detecting Disease Clusters



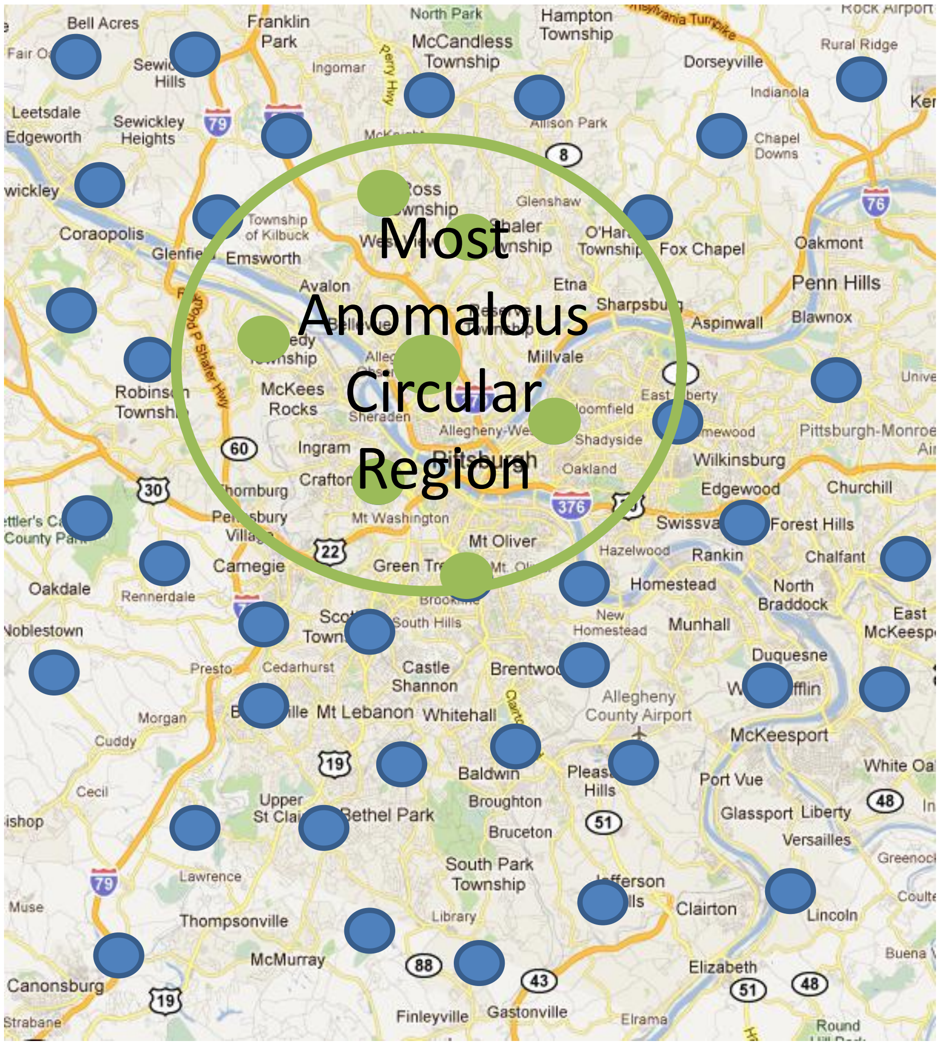

Location of an informative data stream

- \# of ER visits per Zip Code
- \# of OTC Drug sales per retailer
- Other novel data sources ...

**In the presence of an outbreak, we expect counts of the affected locations to increase.**

Effective methods should have high *detection power.*

# Detecting Disease Clusters




(Kulldorff, 1997)

Spatial Scan Statistic
(Circles)

Clusters locations by regions constrained by shape

High power to detect disease clusters of the corresponding shape

But what about irregular shaped clusters?

# Detecting Irregular Disease Clusters


(Neill, 2012)

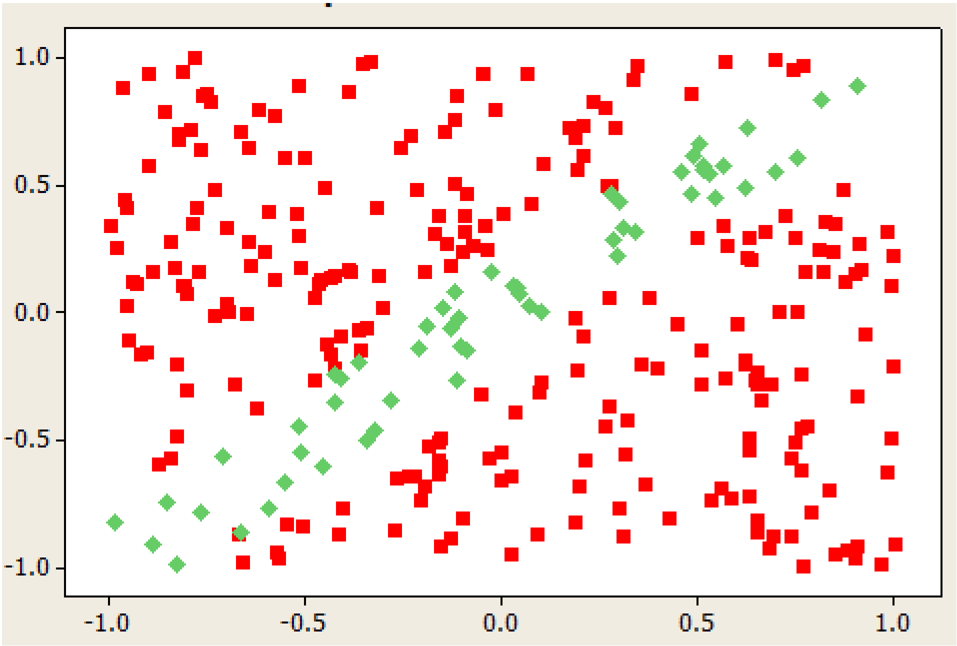
Fast Subset Scan

Instead of clustering ***ALL locations*** within the region together, only the ***most anomalous subset of locations*** within the region is used

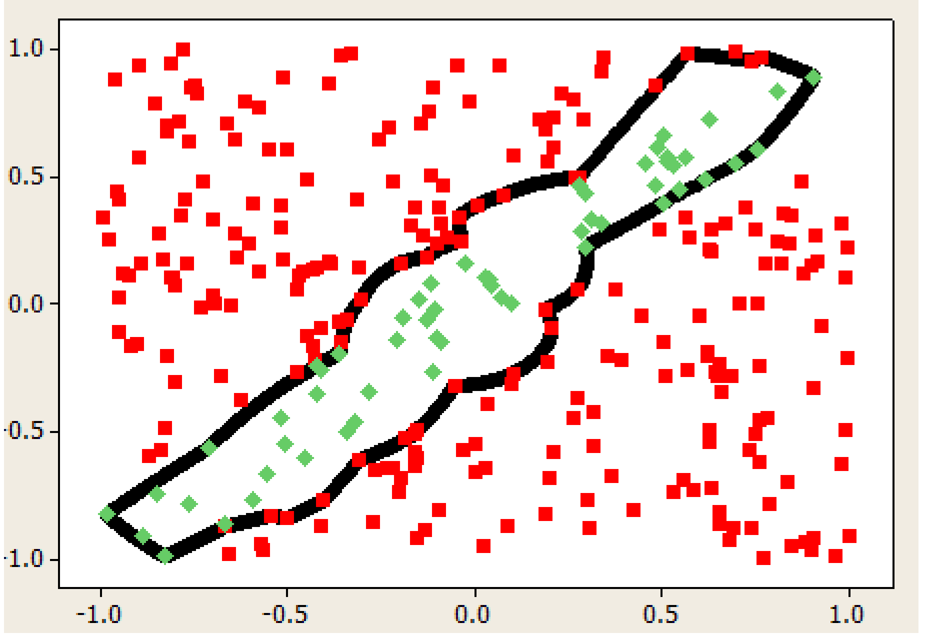
Increases power to detect irregularly shaped disease clusters

...but returns ***unconstrained subsets*** that may not reflect a pattern of interest

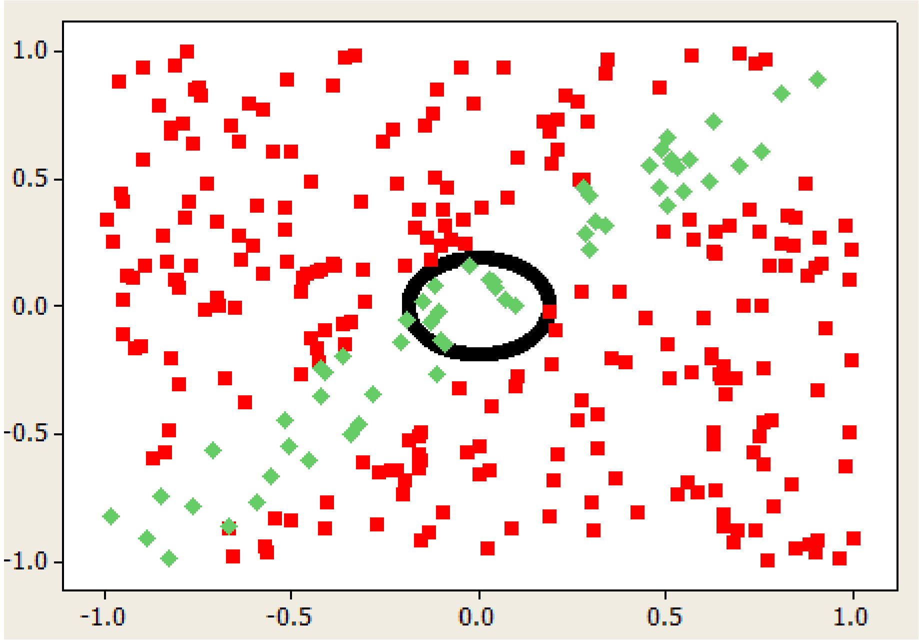
# Sample Data



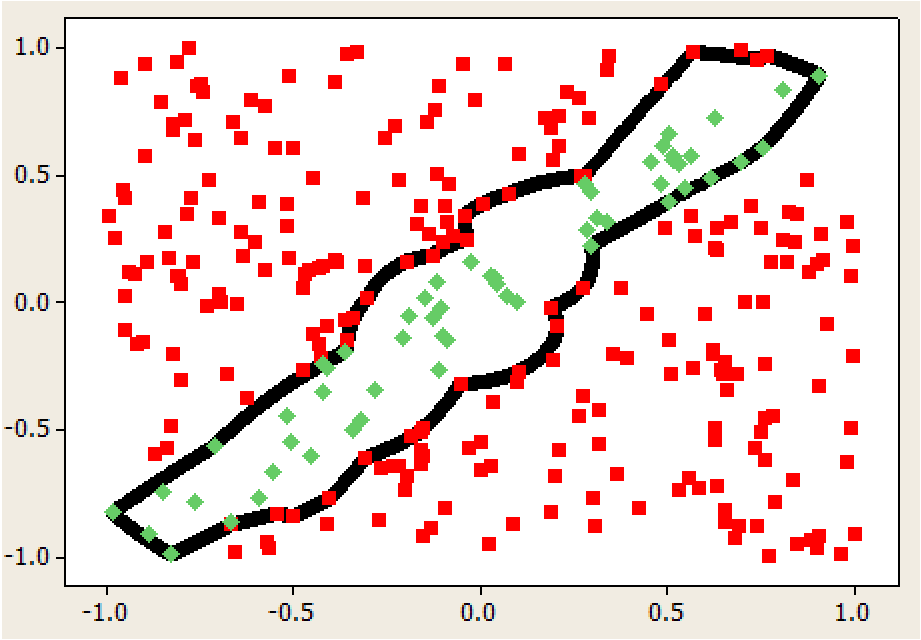
# Sample Data: Fast Subset Scanning



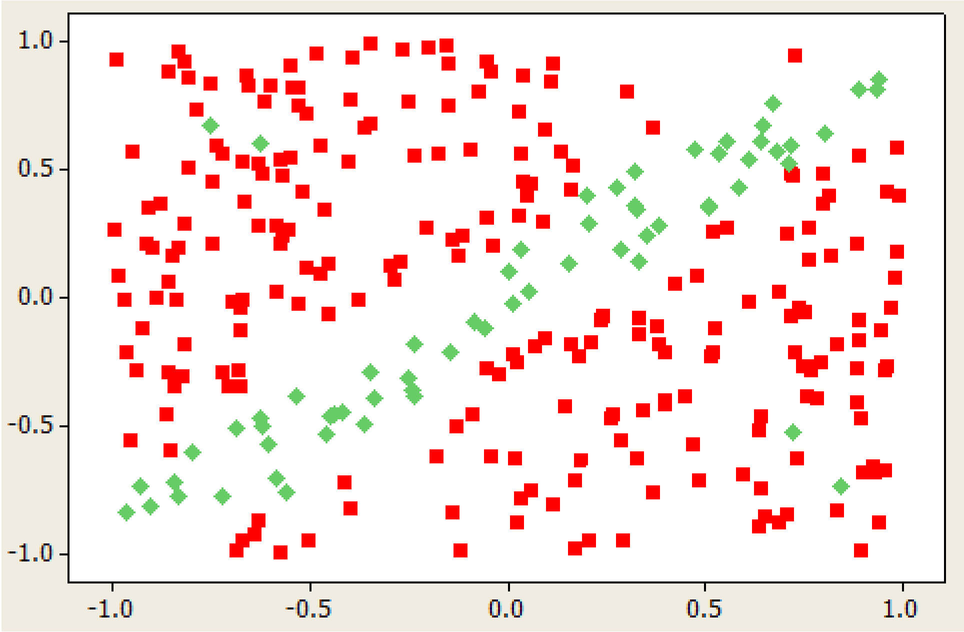
# Sample Data: Circles



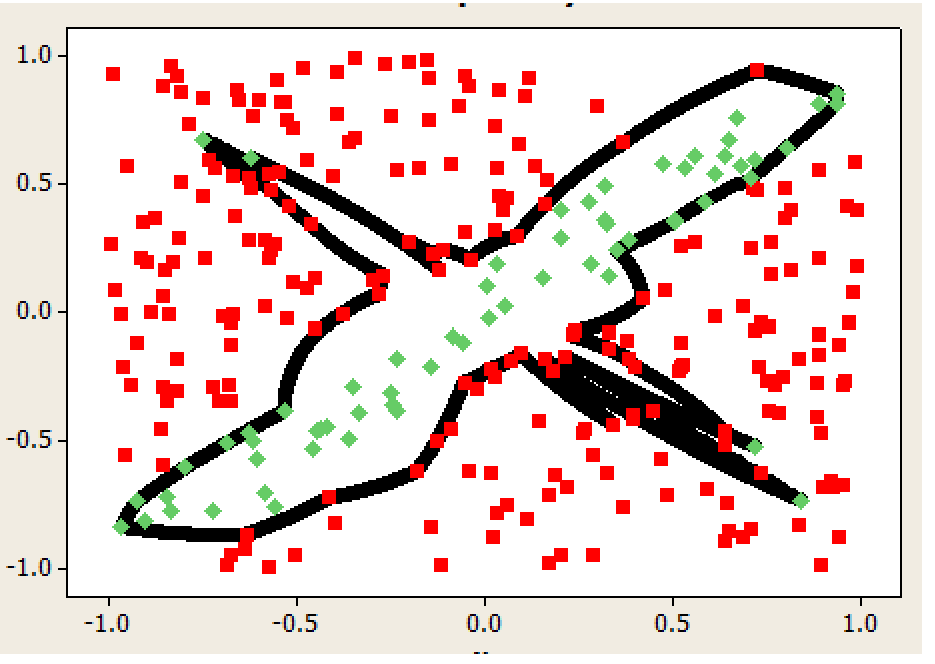
# Sample Data: Star Scan



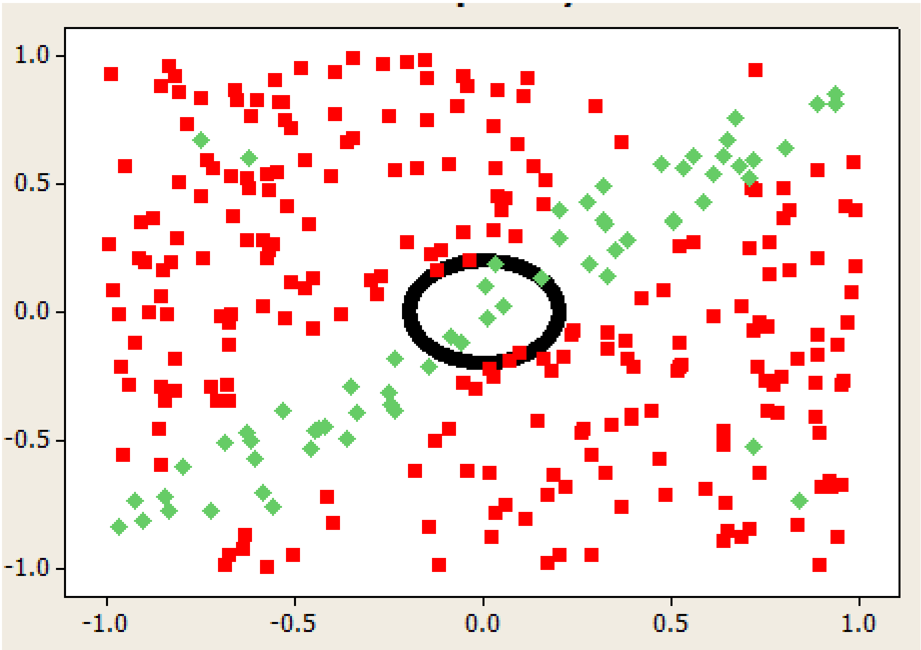
# Sample Data with Noise



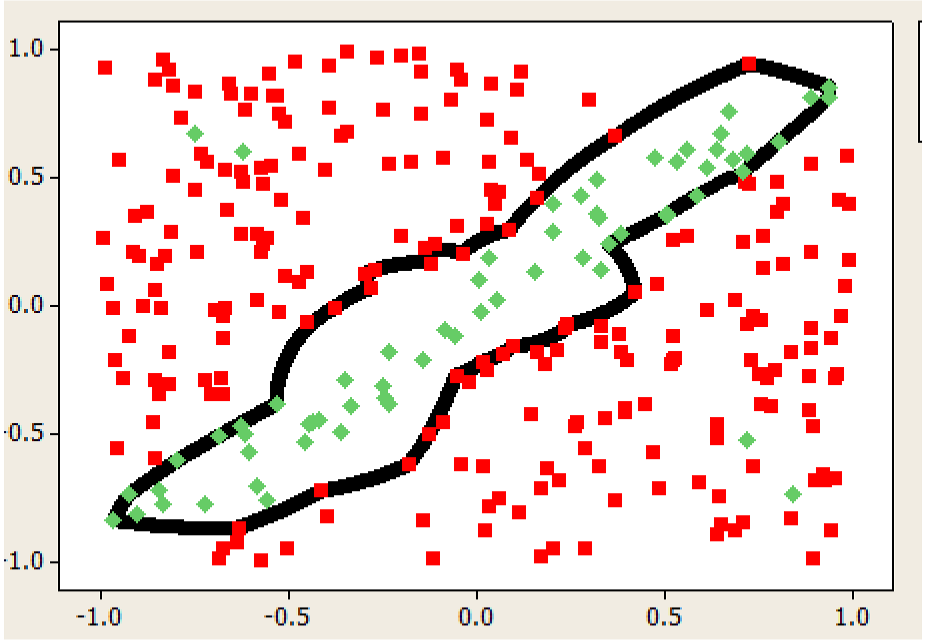
# Sample Data with Noise: Fast Subset Scanning



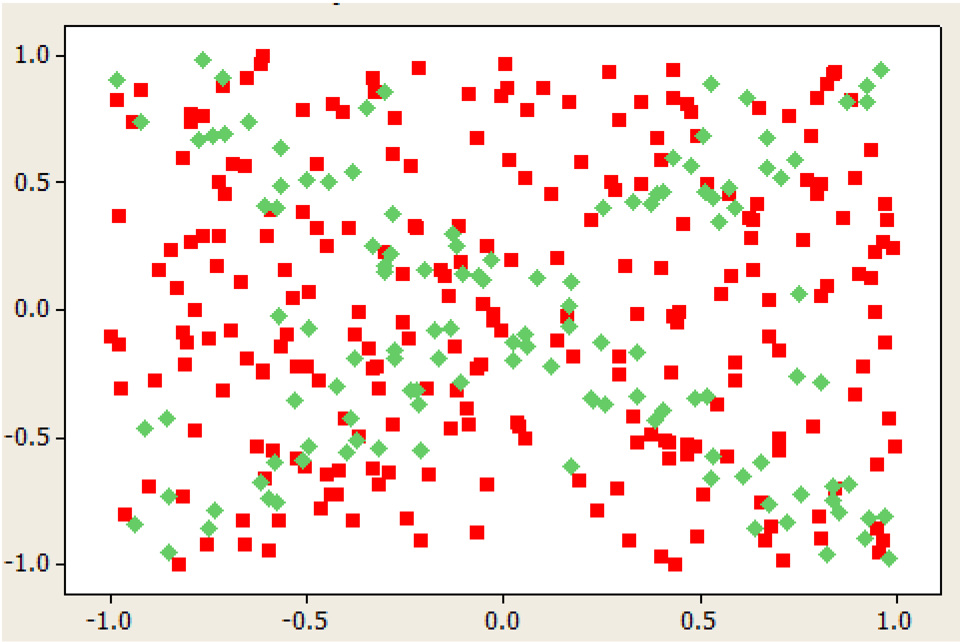
# Sample Data with Noise: Circles



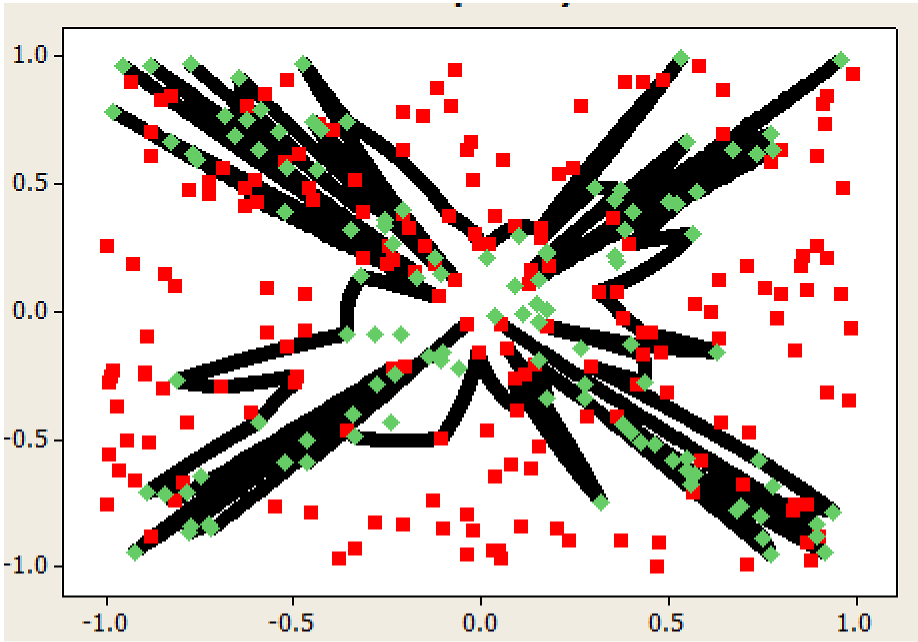
# Sample Data with Noise: Star Scan



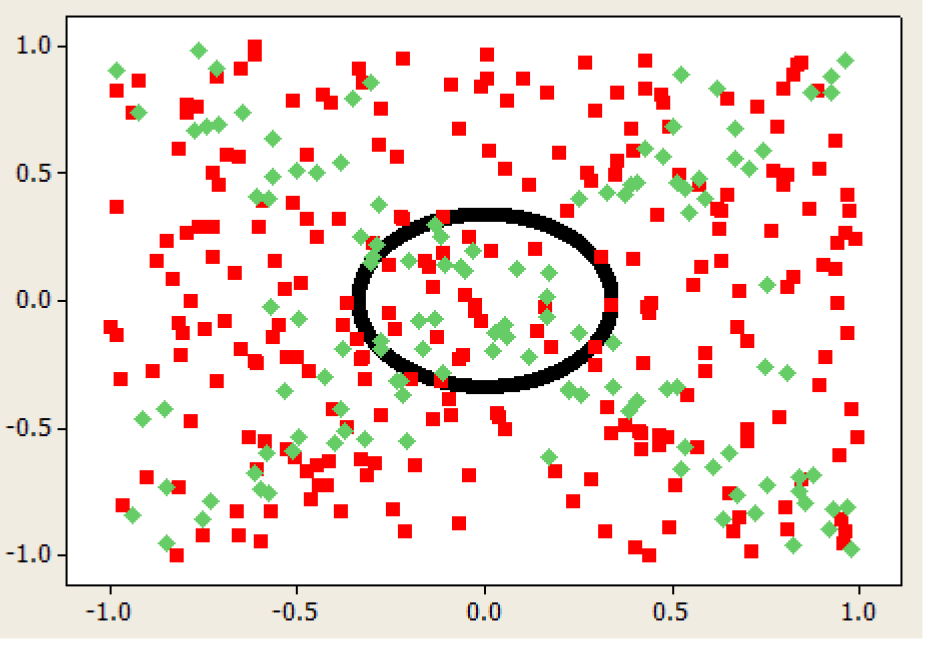
# Cross Pattern



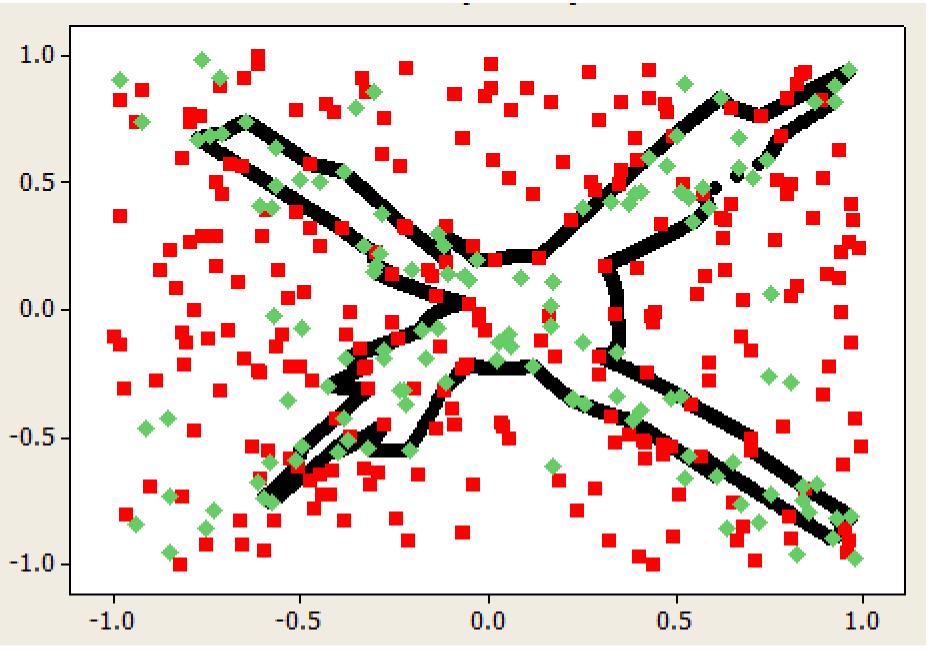
# Cross Pattern: Fast Subset Scanning



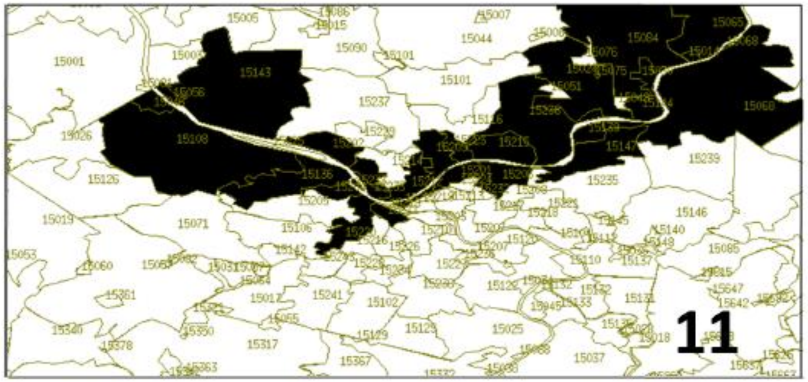
# Cross Pattern: Circles

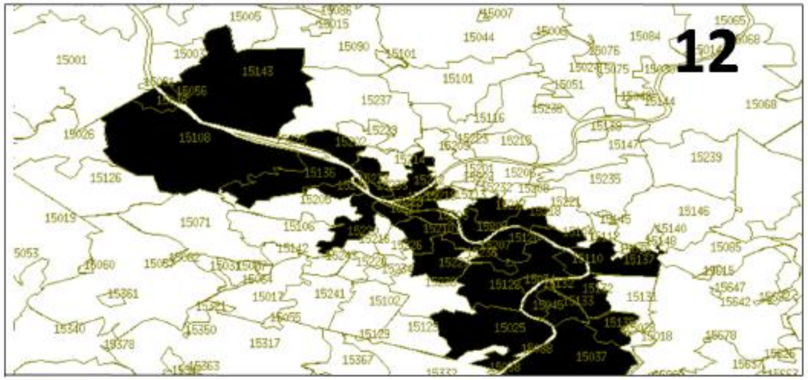


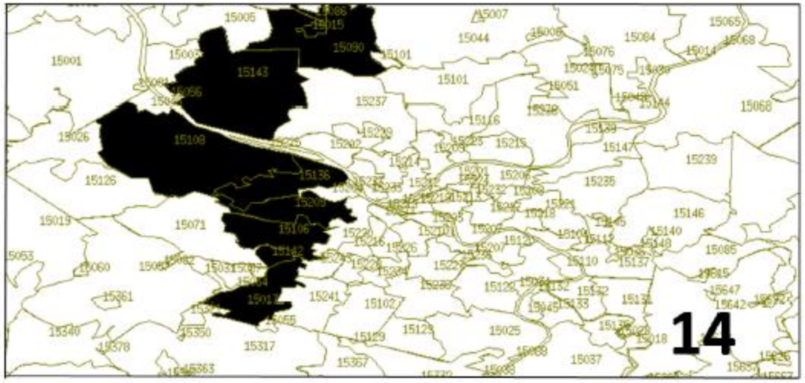
# Cross Pattern: Star Scan

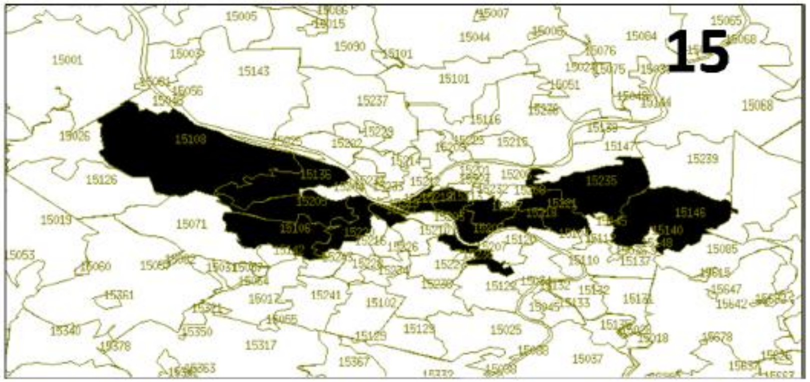


# Real-world examples



# Star Scan

- We propose a new technique to detect irregularly shaped clusters
- Star Scan maximizes the log-likelihood ratio of a cluster while penalizing the change in radius to form the cluster
- We propose a dynamic programming based solution to find optimal clusters, with penalty terms introduced to control smoothness in the circumference of cluster

# Expectation-Based Scan Statistics

For location $s_i$, $i = 1 \ldots N$

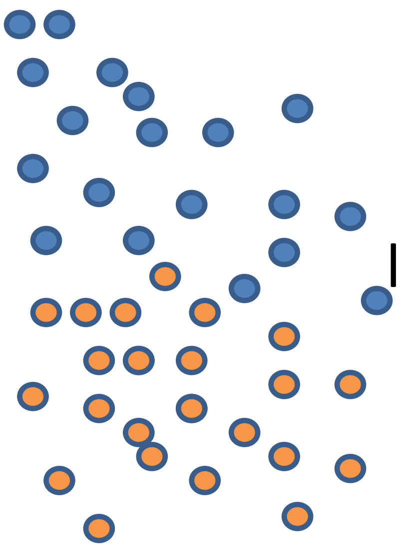
Observed : $x_i$

Expected : $\mu_i$

$H_0 : x_i \sim \mu_i$

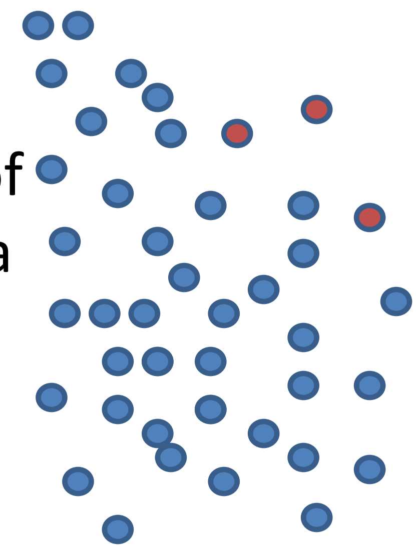
$H_1 : x_i \sim q\mu_i \quad q > 1$

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$



Large number locations with a moderate risk

Small number of locations with a high risk



# Additive Linear Time Subset Scanning

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$

$H_0 : x_i \sim \mu_i$

$H_1 : x_i \sim q\mu_i \quad q > 1$

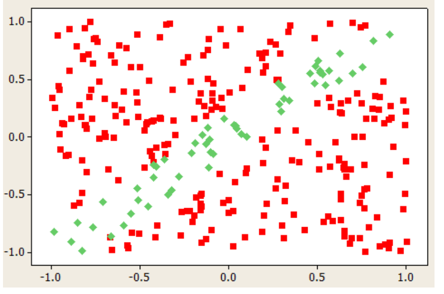
Conditioning ALTSS functions on the relative risk, *q*, allows the function to be written as an ***additive*** set function over the data elements $s_i$ contained in *S*.

**Poisson example:**

$$F(S) = \max_{q>1} \sum_{s_i \in S} x_i(\log q) + \mu_i(1 - q)$$

# Conditioning on relative risk

- By conditioning on relative risk (q) each element is either "positive" or "negative"



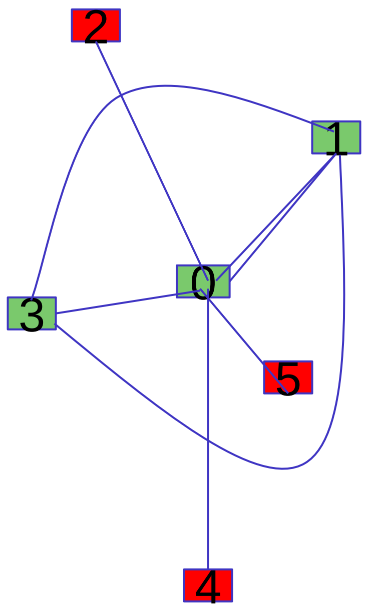

- This simplifies the maximization over subsets
  - Include only the points whose contribution to LLR are positive

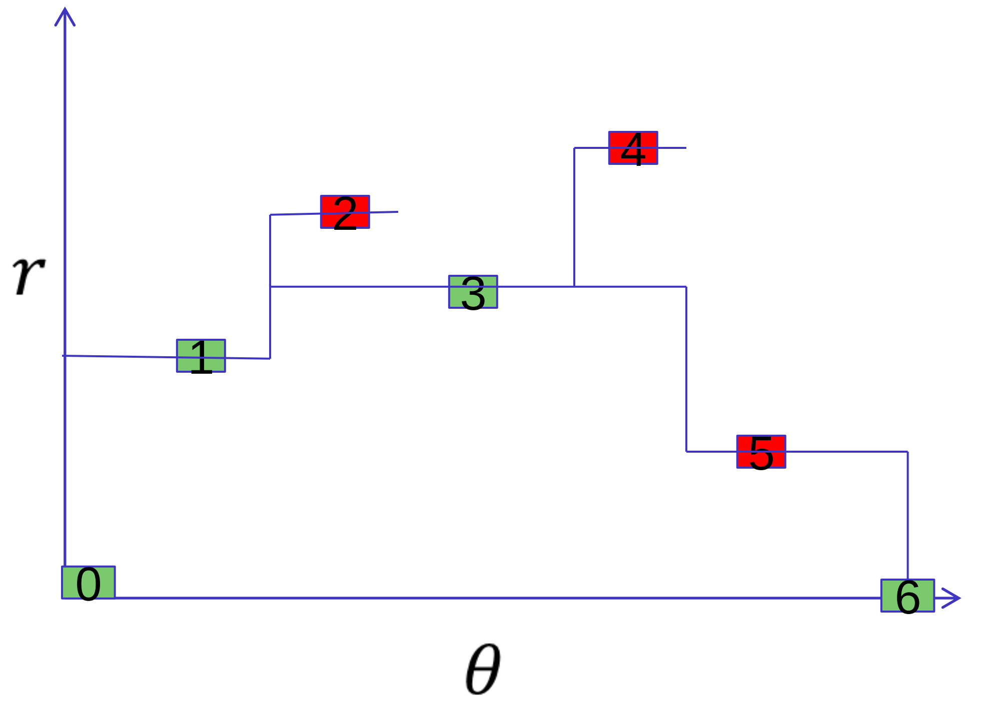
$$F(S) = \max_{q>1} \sum_{s_i \in S} [\, x_i(\log q) + \mu_i(1-q)\ ]$$

# Star Scan: Fundamentals

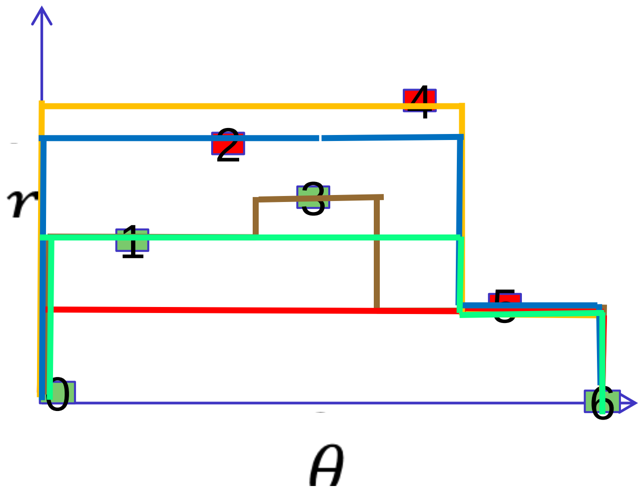
- The score of subset (S) is dependent on the following four characteristics
  - Cumulative sum of observed: $\sum x_i = X(S)$
  - Cumulative sum of expected: $\sum \mu_i = \mu(S)$
  - Total change in radius to form a subset : $R(S)$
- We propose a dynamic programming based solution to find optimal subset that maximizes the score of subset (S)
- $F_{starscan}(S \mid q) = F_{exp}(S|q) + \lambda * R(S)$

# Dynamic Programming for Star Scan



# Dynamic Programming for Star Scan




- Best of
  - Constant Radius
  - Best path via 1, 2, 3, 4, 5

**Steps Ahead**

| **Start Location** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| 0 | | | | | | $\boldsymbol{X(S), \mu(S)}$ $\boldsymbol{\Delta(S), \mathrm{R}(S)}$ |
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |

# Star Scan generalizes FSS and Circles

- The penalty parameter can be used to generalize Star Scan

$$F_{starscan}(S\,|q) = F_{exp}(S\,|q) + \lambda * R(S)$$

- $\lambda$ is the penalization parameter
  - High value of $\lambda$: Circles (Kulldorff, 1997)
  - Low value of $\lambda$: Fast Subset Scan (Neill, 2012)

# Star Scan: Challenges

- Dynamic programming is easy for a given relative risk (q) as each element is either “positive” or “negative”, that is,

$$F_{starscan}(S\,|q) = F_{exp}(S\,|q) + \lambda * R(S)$$

$$F^*(q) = \max_{S} F_{starscan}(S\,|q)$$

- However the optimal score $F^*$ is given by

$$F^* = \max_{q>1} \max_{S} F_{starscan}(S\,|q)$$

# DP for Star Scan: Solutions

- We can either grid search for the values of $q$ in the range of possible values
- Or use branch and bound technique in order to find the optimal value of $q$

# Bayesian Aerosol Release Detector (BARD)

Hogan et al; 2007

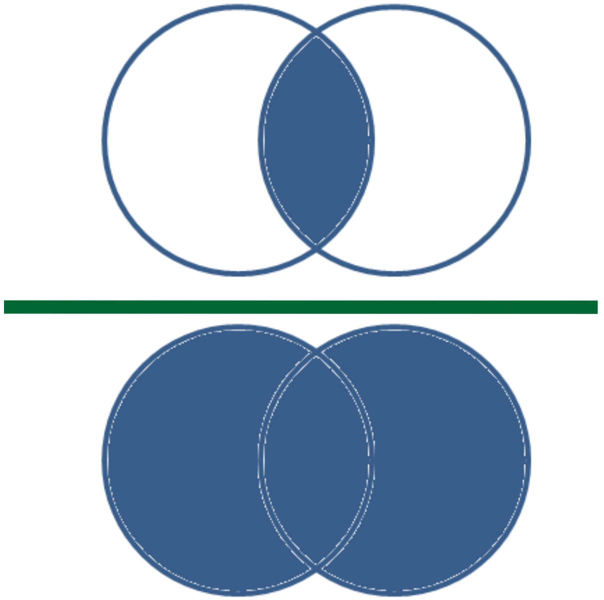
Simulates anthrax spores released over a city

Two models drive the simulator:

| Dispersion | Infection |
| --- | --- |
| Which areas will be affected? | How many infected people in an area? |
| Weather data | Demographic data |
| Gaussian plumes | Increased ER visits with respiratory complaints |

# Results: Spatial Overlap

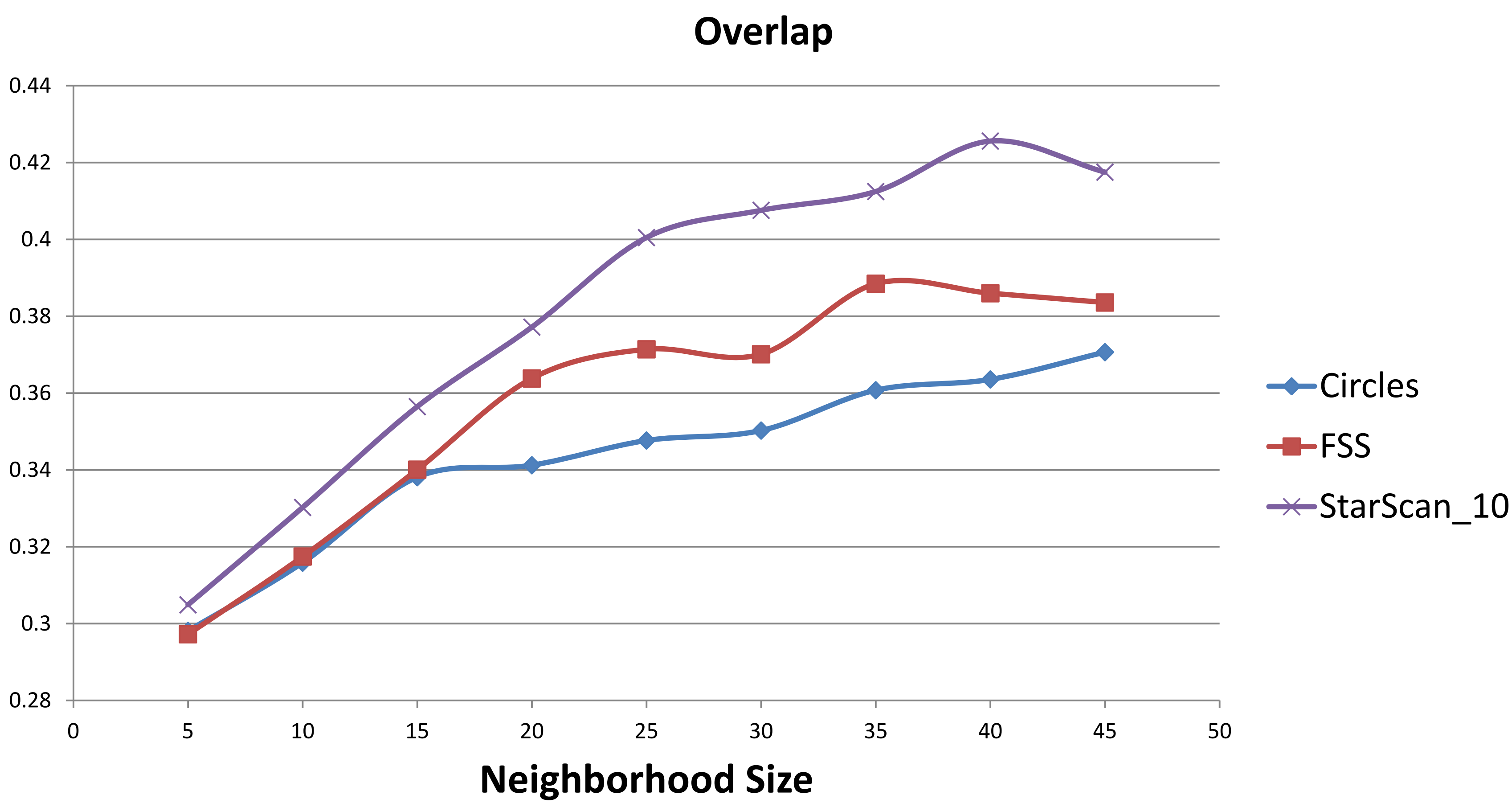
$$Overlap = \frac{A \cap B}{A \cup B} =$$



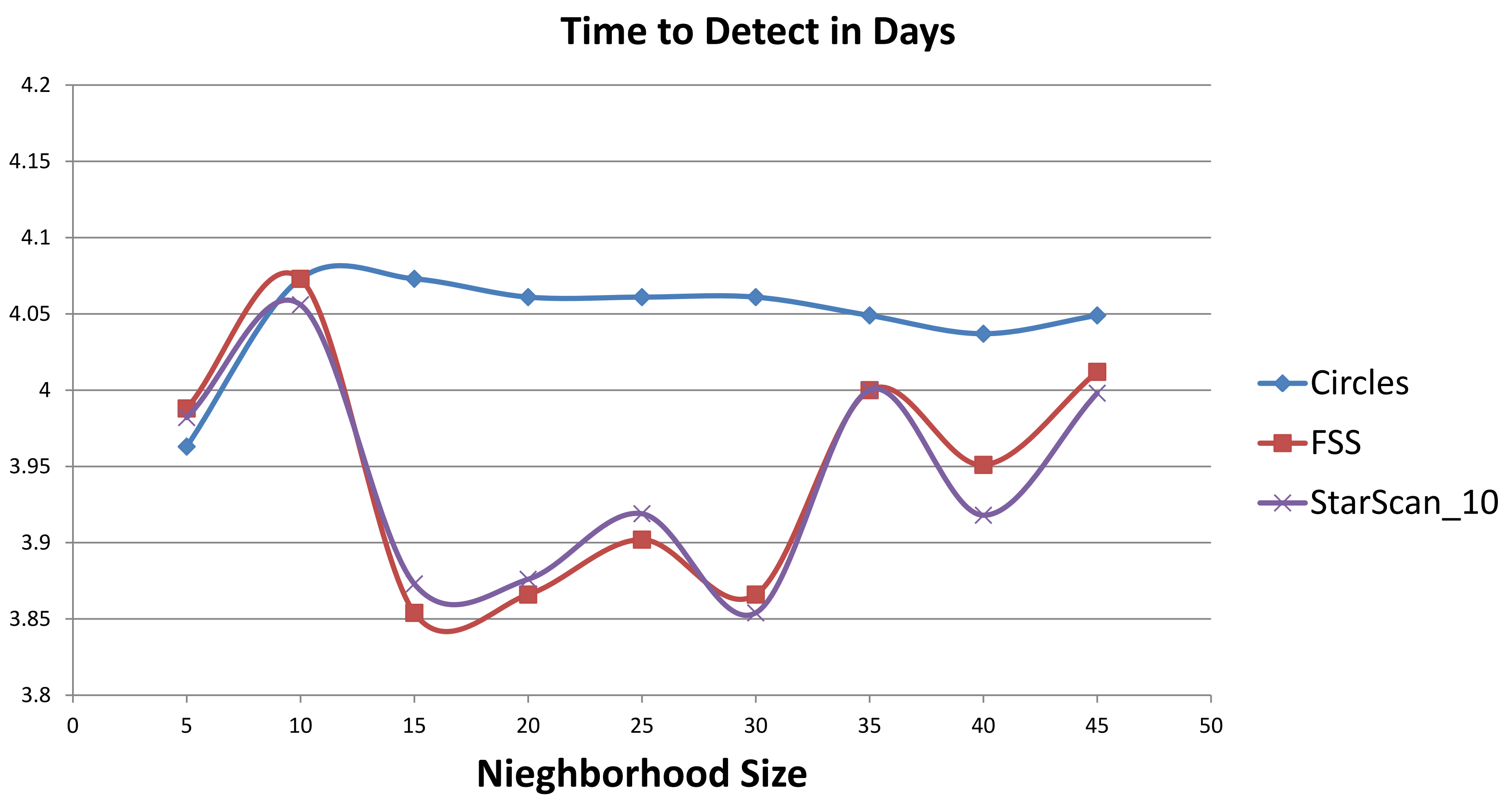
$$Overlap = 1$$ Perfect Match

$$Overlap = 0$$ Completely Disjoint

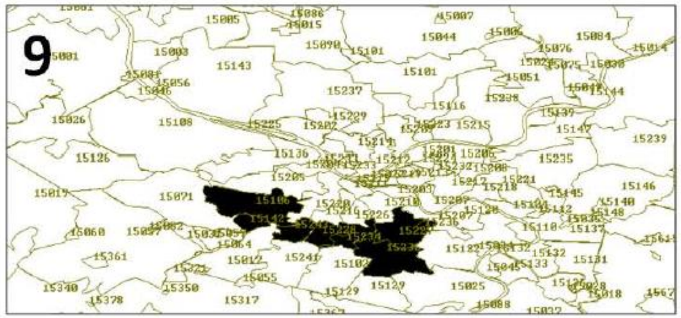
# BARD Results: Spatial Overlap

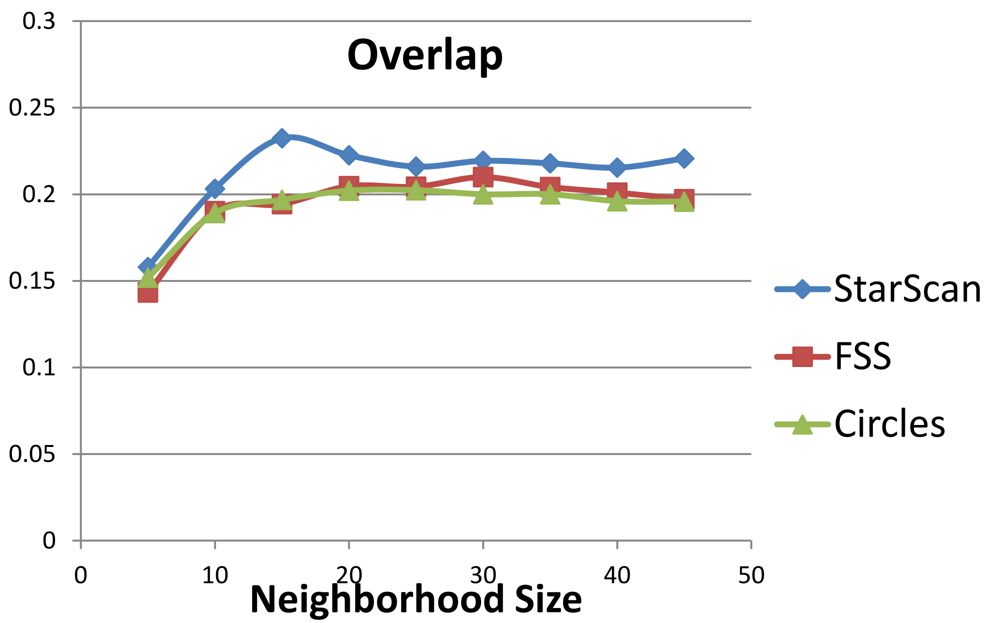


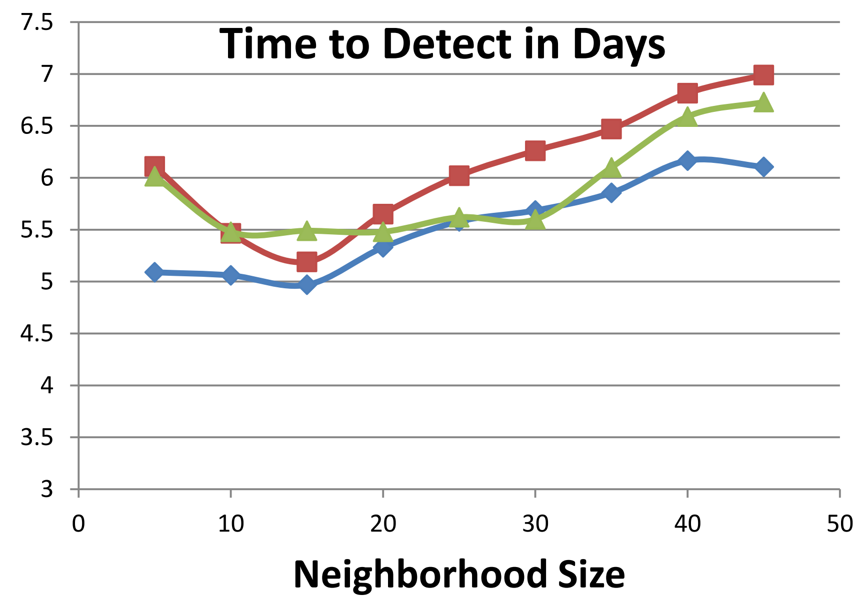
# BARD Results: Time to Detect at a fixed fpr



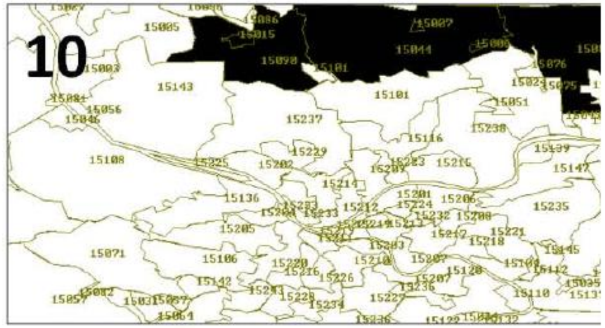
# Simulated Injects in real-world Emergency Department data.

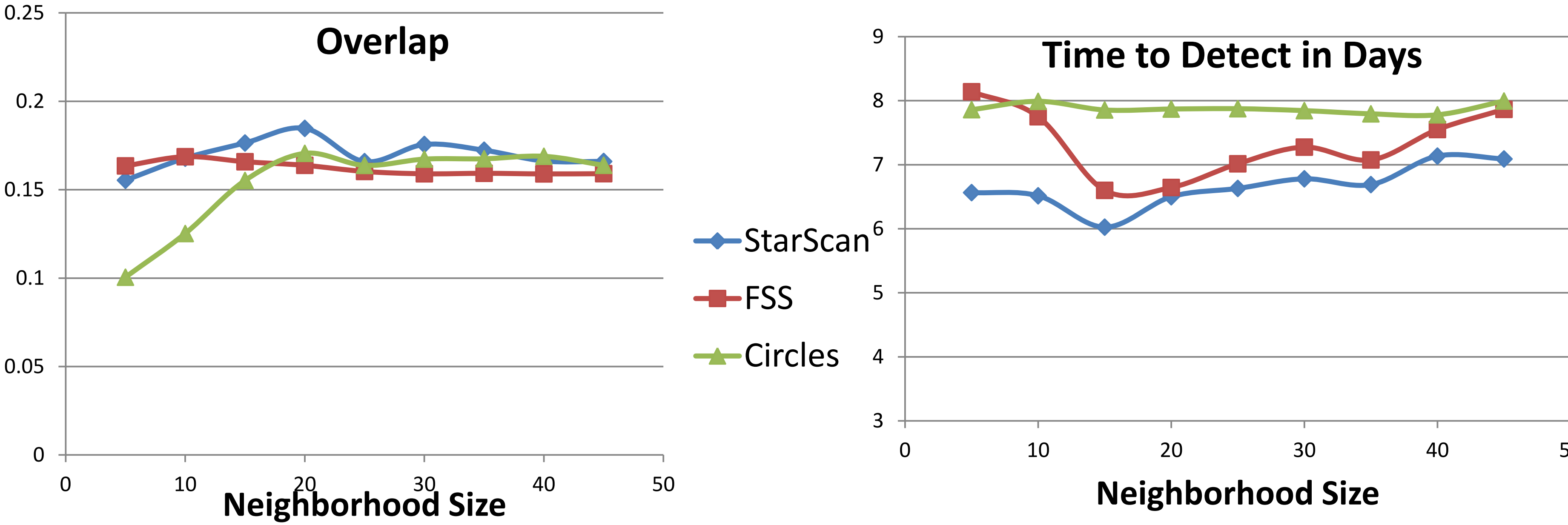


# Simulated Injects (continued)



# Conclusion

- We propose StarScan to find irregularly-shaped clusters more accurately than either the circular scan or unconstrained fast subset scan
- StarScan was compared with circular scan and fast localized subset scan on simulated respiratory outbreaks and bioterrorist anthrax attacks injected into real-world Emergency Department data
- Given a small amount of labeled training data, StarScan learns appropriate penalties for both compact and elongated clusters, resulting in improved detection performance