StarScan: A Novel Scan Statistic for Irregularly-Shaped Spatial Clusters

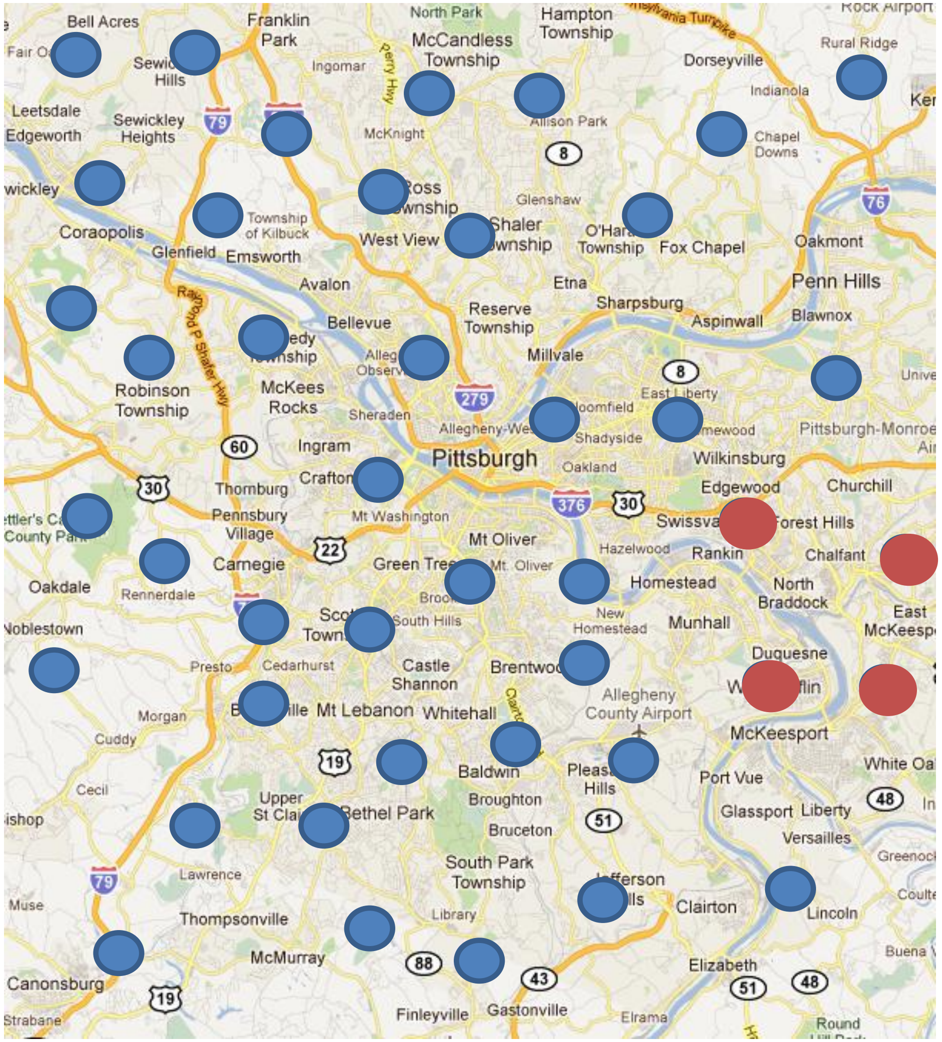
Sriram Somanchi, David Choi, Daniel B. Neill

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

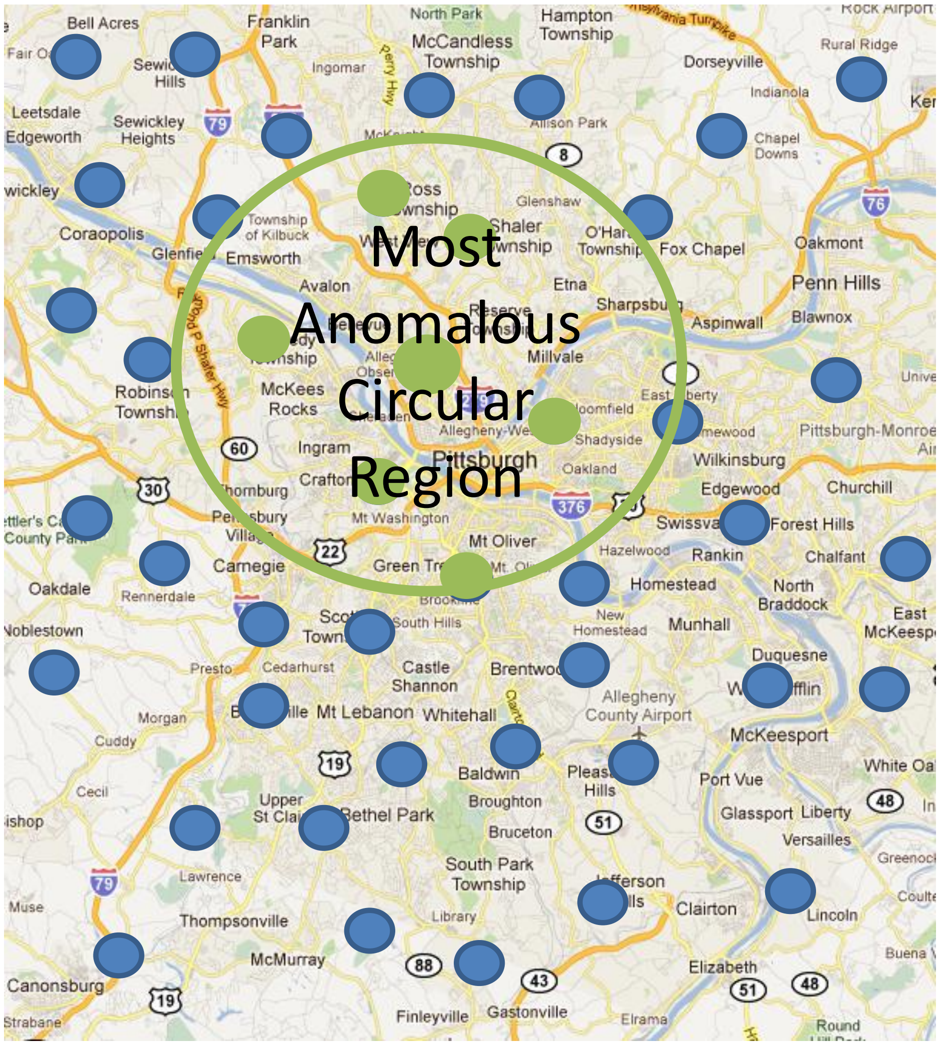


- Location of an informative data stream
  - # of ER visits per Zip Code
  - # of OTC Drug sales per retailer
  - Other novel data sources ...

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

Effective methods should have *high detection power & high spatial accuracy.*

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)

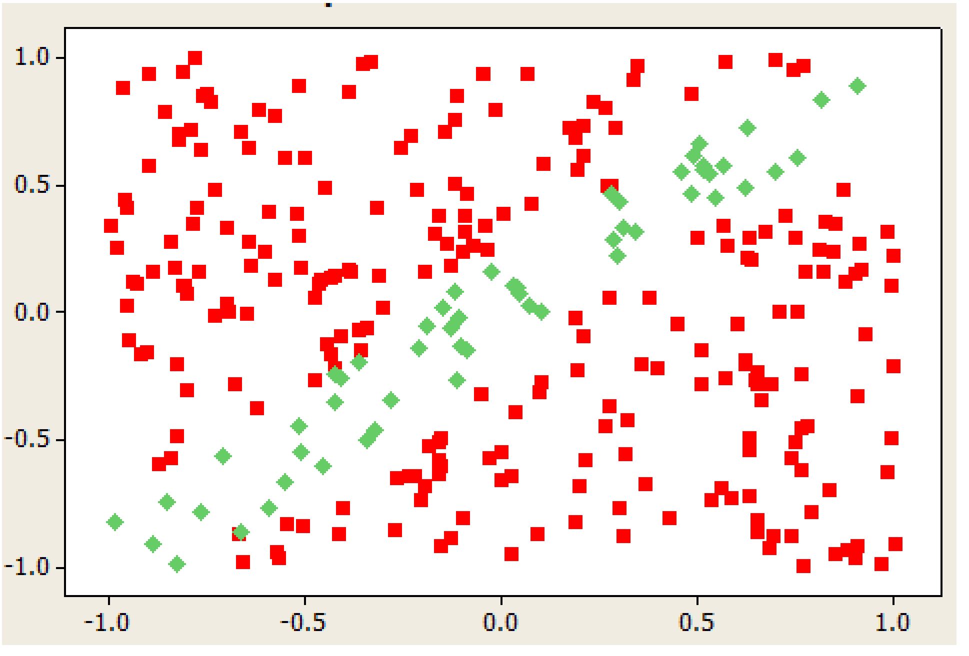
Fast Subset Scan

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

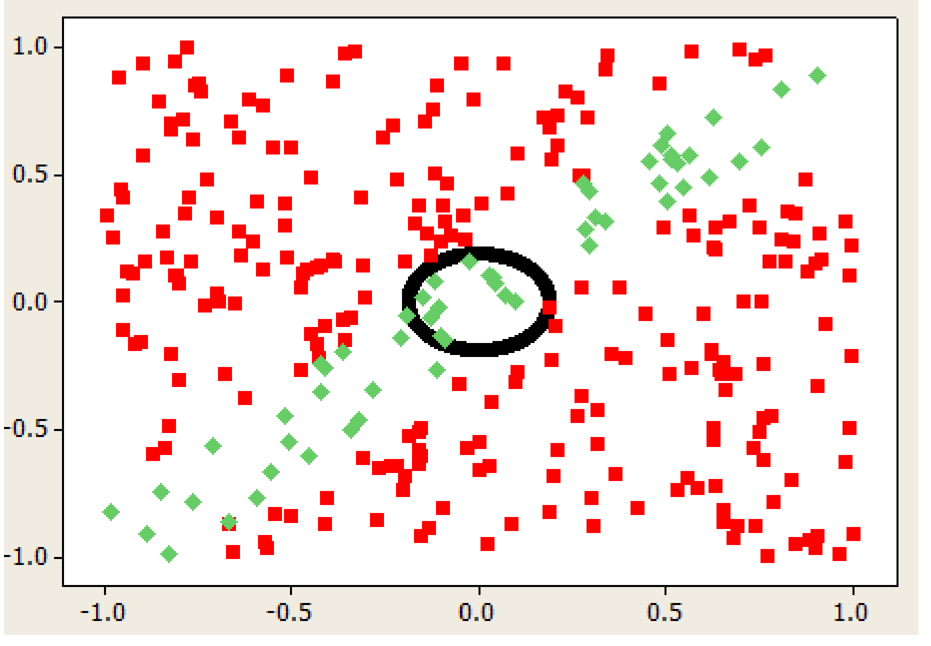
Increases power to detect irregularly shaped disease clusters

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

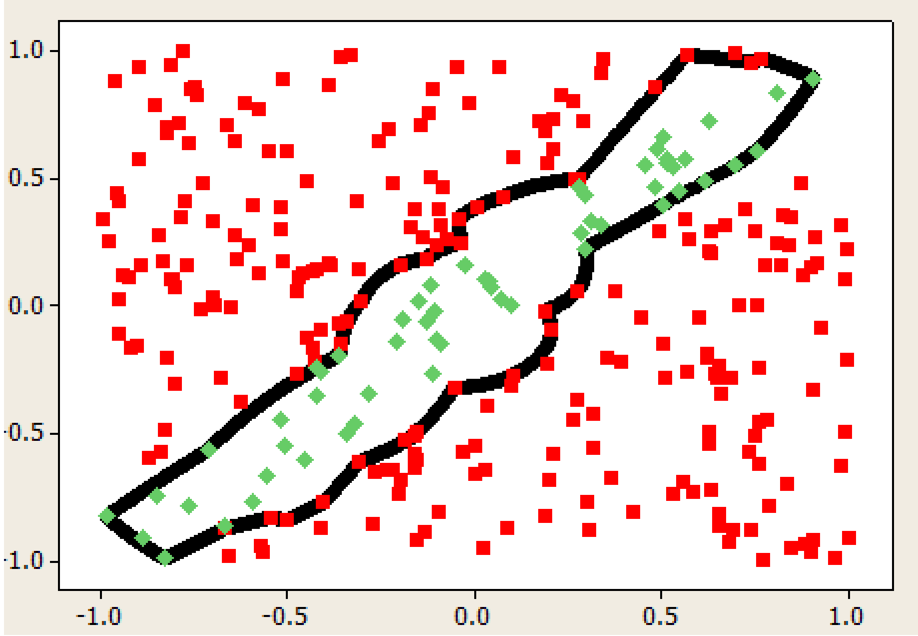
Sample Data



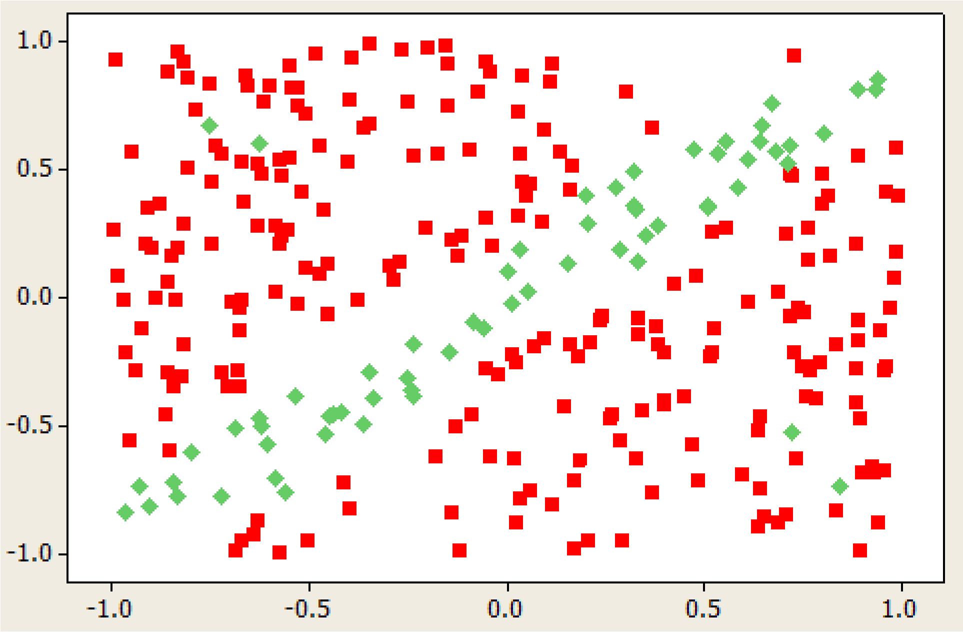
Sample Data: Circles



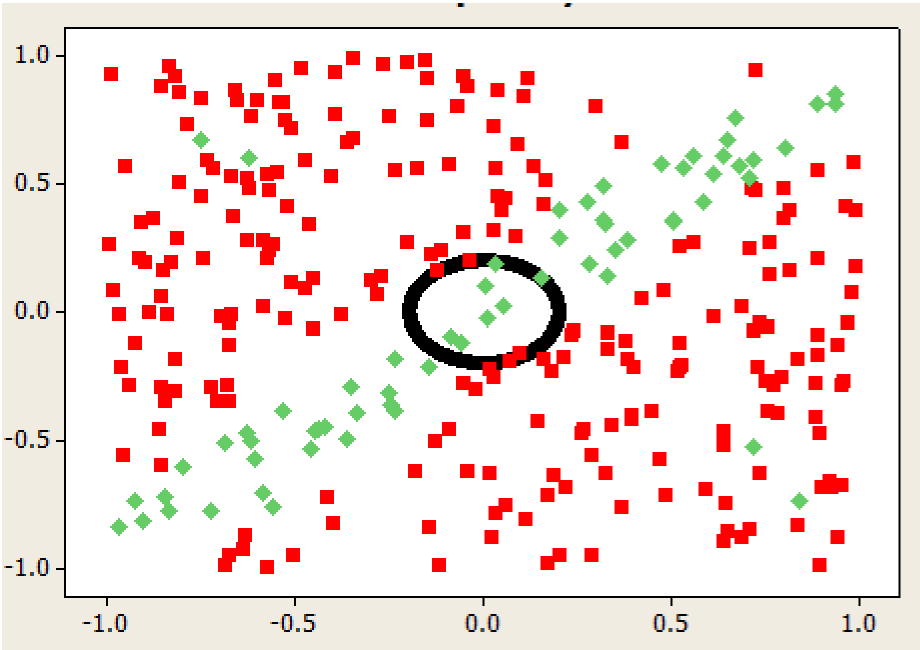
Sample Data: Fast Subset Scanning



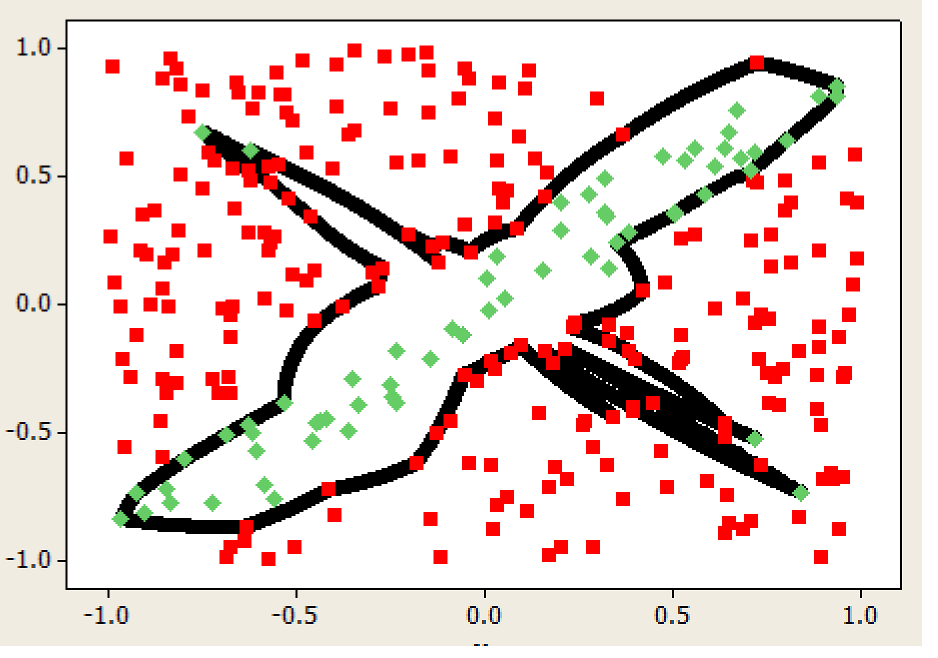
Sample Data with Noise



Sample Data with Noise: Circles

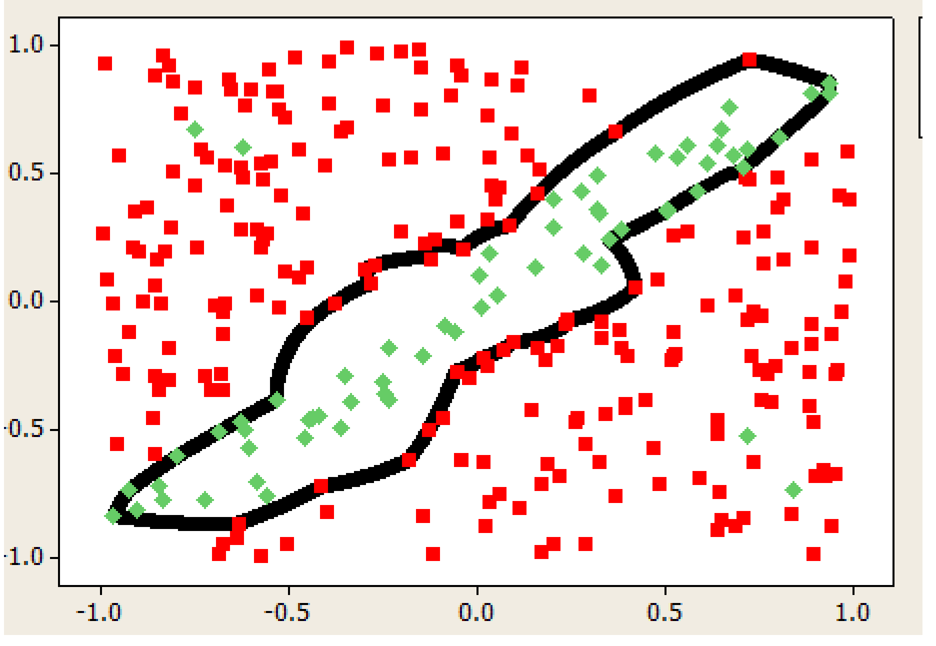


Sample Data with Noise: Fast Subset Scanning

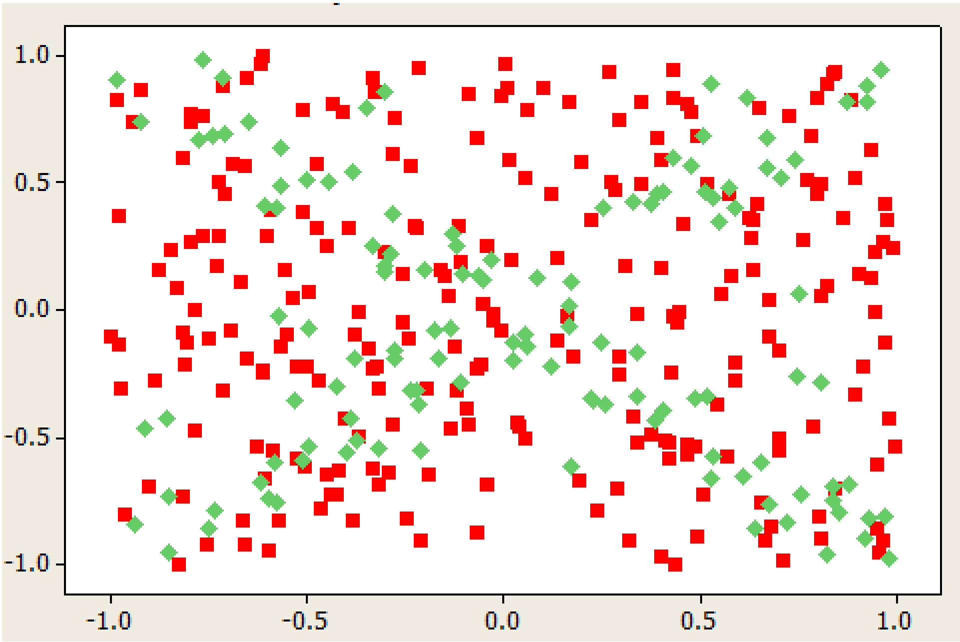


Sample Data with Noise: Star Scan

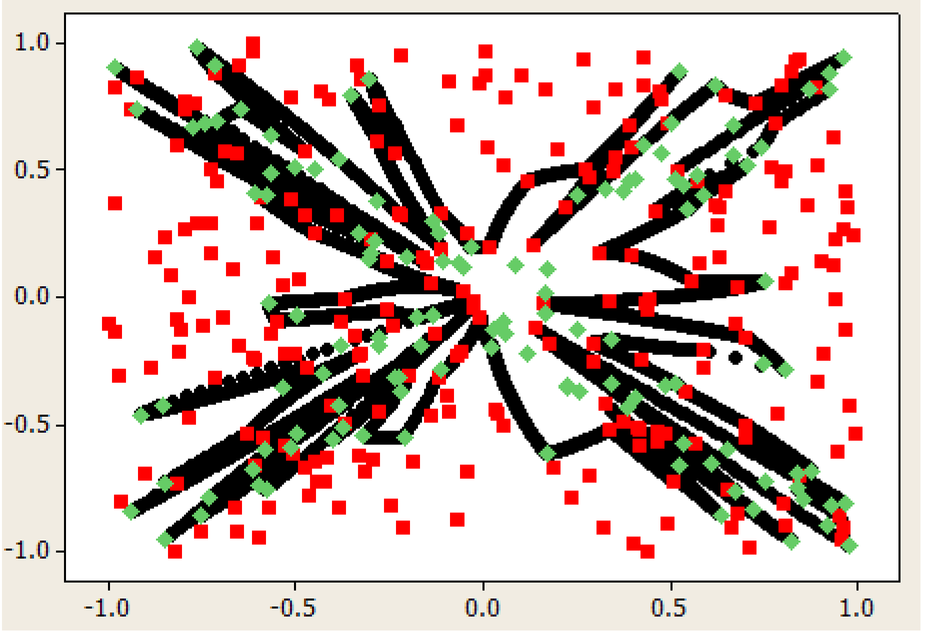
We propose a new star-shaped scan statistic ("StarScan") that can more accurately detect smooth irregularly-shaped clusters.



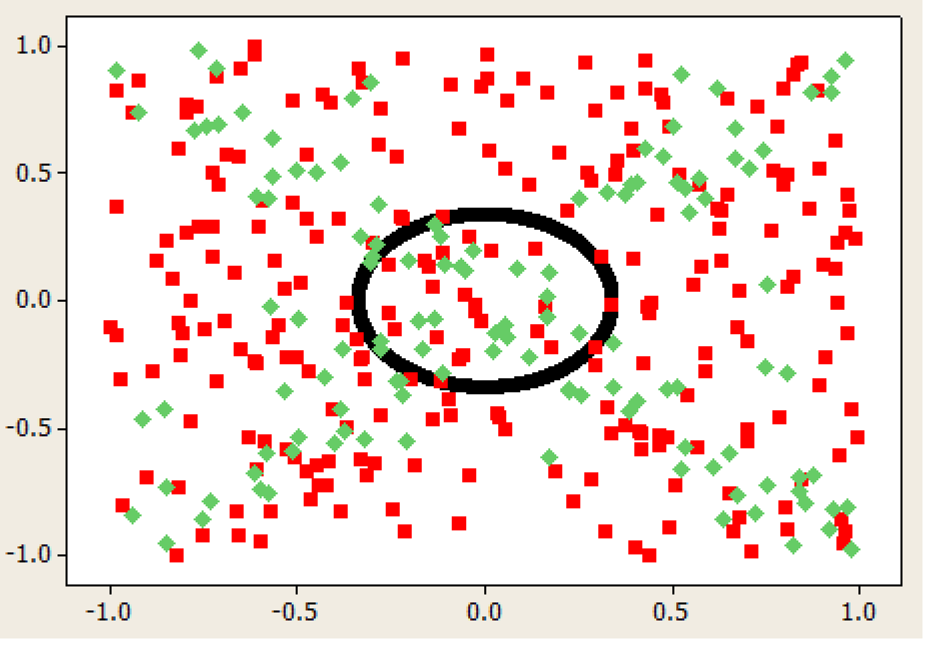
Cross Pattern



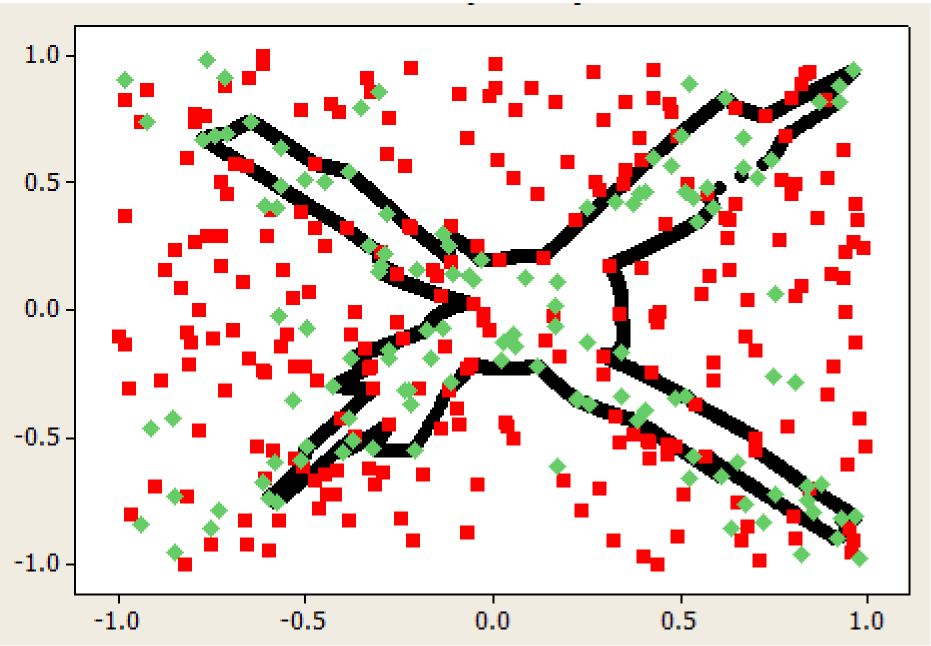
Cross Pattern: Fast Subset Scanning



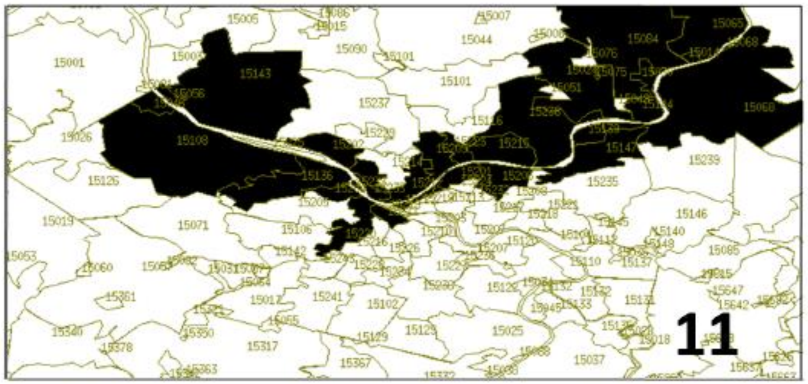
Cross Pattern: Circles

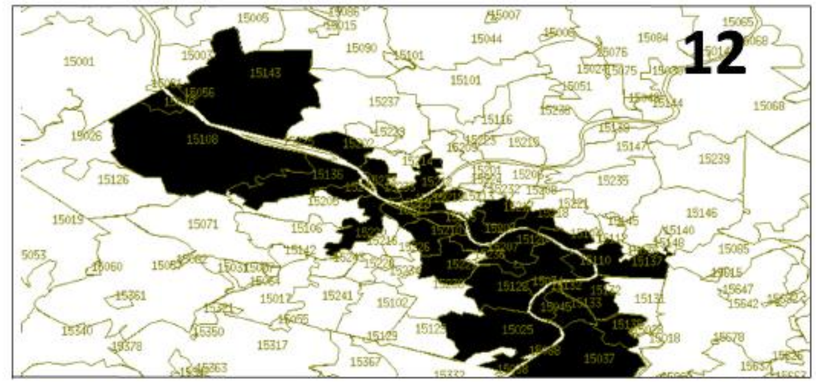


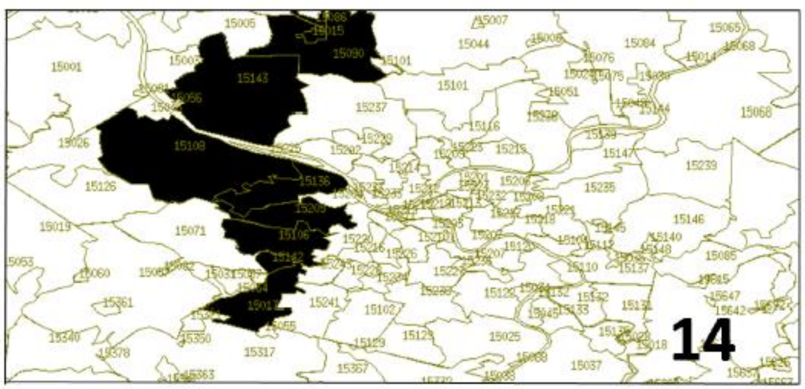
Cross Pattern: StarScan

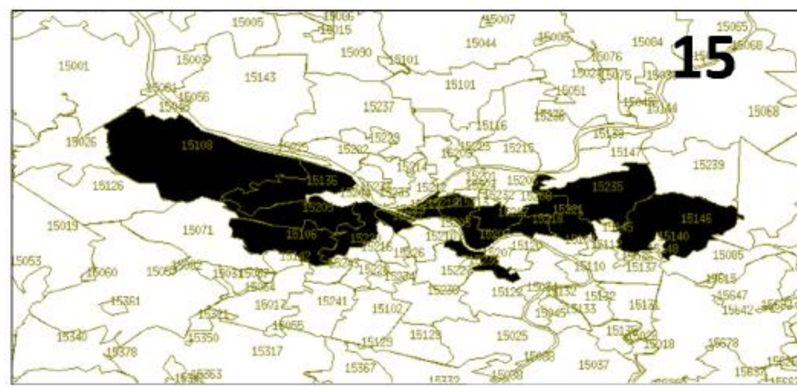


Real-world examples









Star Scan

- We propose a new technique to detect smooth irregularly shaped clusters.
- Instead of requiring a constant radius (as for circles), StarScan allows the radius to vary, but applies a penalty proportional to the total change in radius.
- We propose a new, dynamic programming-based solution to find the clusters that maximize the penalized log-likelihood ratio statistic.

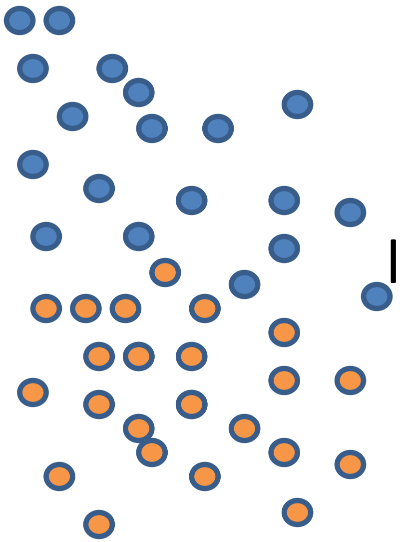
Expectation-Based Scan Statistics

For location $s_i \; {}_{i=1...N}$

Observed : x_i $\quad H_0 : x_i \sim Dist(\mu_i)$

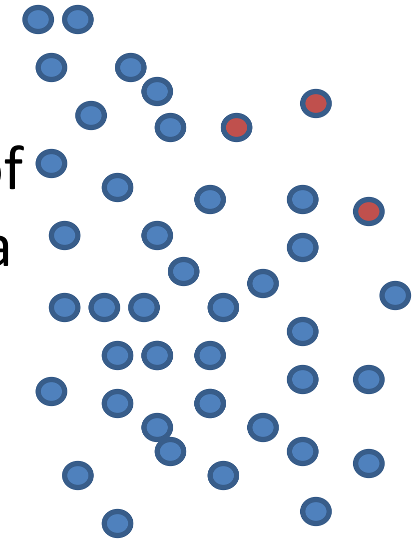
Expected : μ_i $\quad H_1(S) : x_i \sim Dist(q\mu_i) \;\; q > 1$

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



Large number locations with a moderate risk

Small number of locations with a high risk



Detour: Linear Time Subset Scanning

- Our goal in subset scanning is to find optimal subset S^* such that $F^* = F(S^*)$ where

$$F^* = \max_S F(S) = \max_S \max_{q>1} \log \frac{P(Data|H_1(S))}{P(Data|H_0)}$$

- In order to find best subset we need to evaluate exponential number of subsets.

For most (exponential family distributions) of the scoring functions, we need to evaluate only linear number of subsets

Additive Linear Time Subset Scanning

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

$$H_0 : x_i \sim Dist(\mu_i)$$

$$H_1(S) : x_i \sim Dist(q\mu_i) \quad q > 1$$

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*.

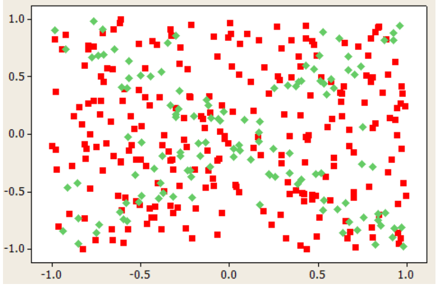
Poisson example:

$$F(S) = \max_{q>1} F(S|q)$$

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

Conditioning on relative risk

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



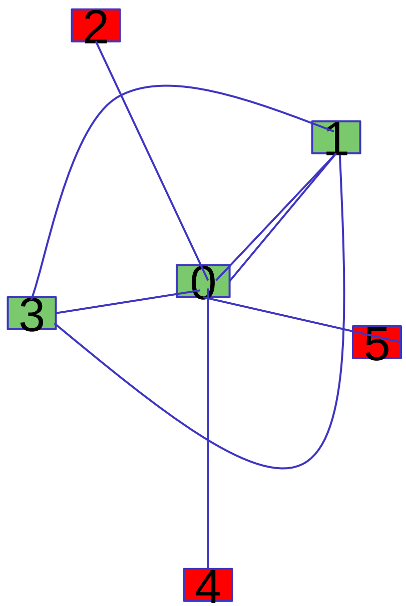
- This simplifies the maximization over subsets
  - Include only the points whose contribution to LLR are positive while minimizing change in radius

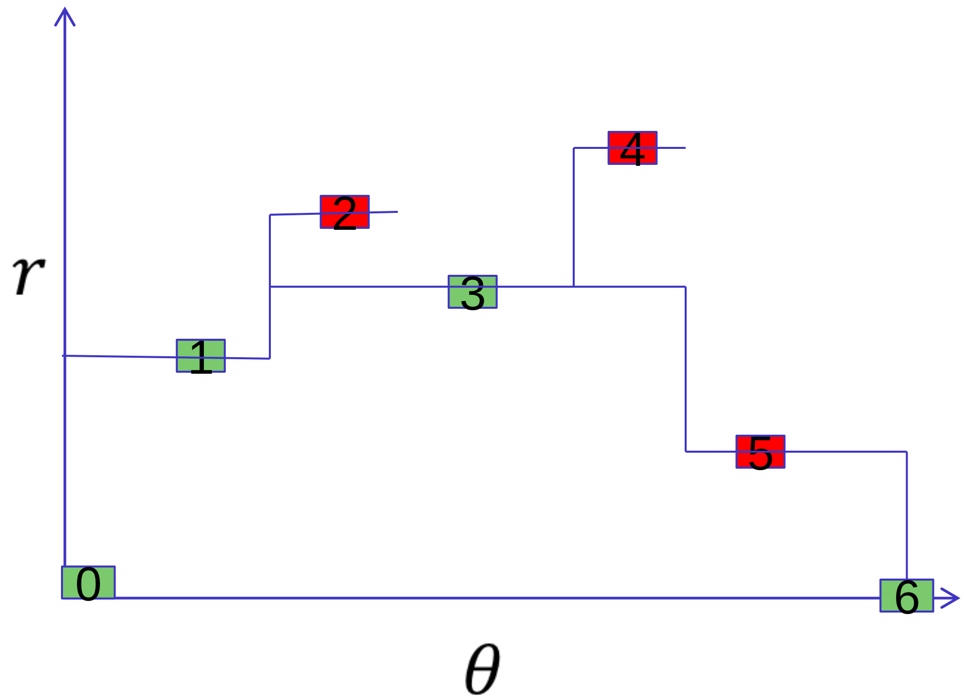
$$F(S) = \max_{q>1} \sum_{s_i \in S} [\, x_i(\log q) + \mu_i(1-q) \,]$$

Star Scan: Fundamentals

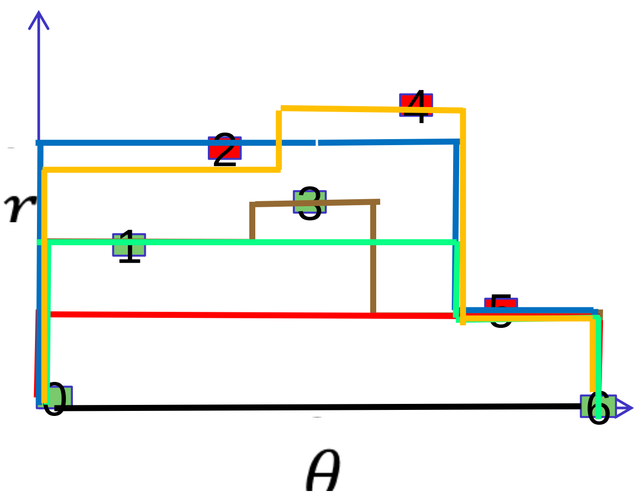
- The score of subset (S) is dependent on the following four characteristics
  - Cumulative sum of observed: $\sum x_i = X(S)$
  - Cumulative sum of expected: $\sum \mu_i = \mu(S)$
  - Total change in radius to form a subset : $R(S)$
- We propose a dynamic programming based solution to find optimal subset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

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						$X(S), \mu(S)$ $\Delta(S), \mathrm{R}(S)$
1						
2						
3						
4						
5						

Star Scan generalizes FSS and Circles

- The penalty parameter can be used to generalize Star Scan

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

- λ is the penalization parameter
  - High value of λ: Circles (Kulldorff, 1997)
  - Low value of λ: 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,

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

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

- 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

Which areas will be affected?

Weather data

Gaussian plumes

Infection

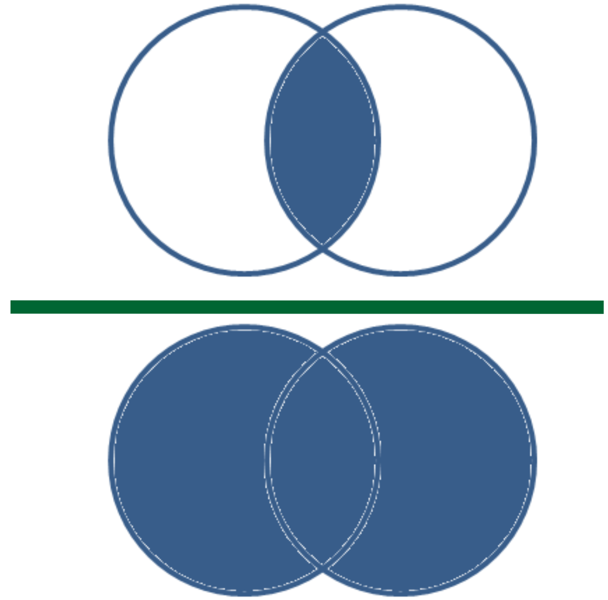
How many infected people in an area?

Demographic data

Increased ER visits with respiratory complaints

Results: Spatial Overlap

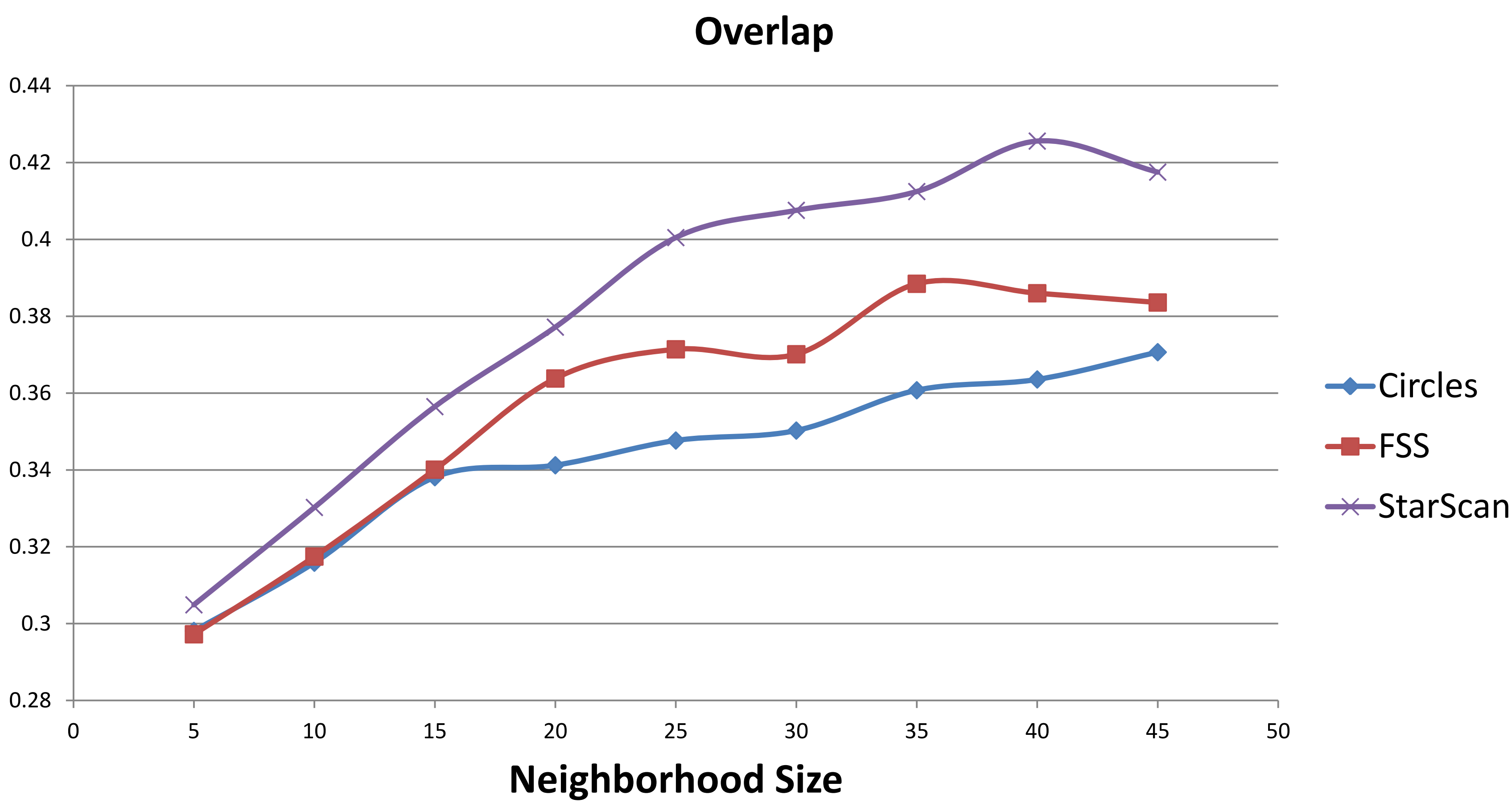
$$Overlap = \frac{A \cap B}{A \cup B} =$$



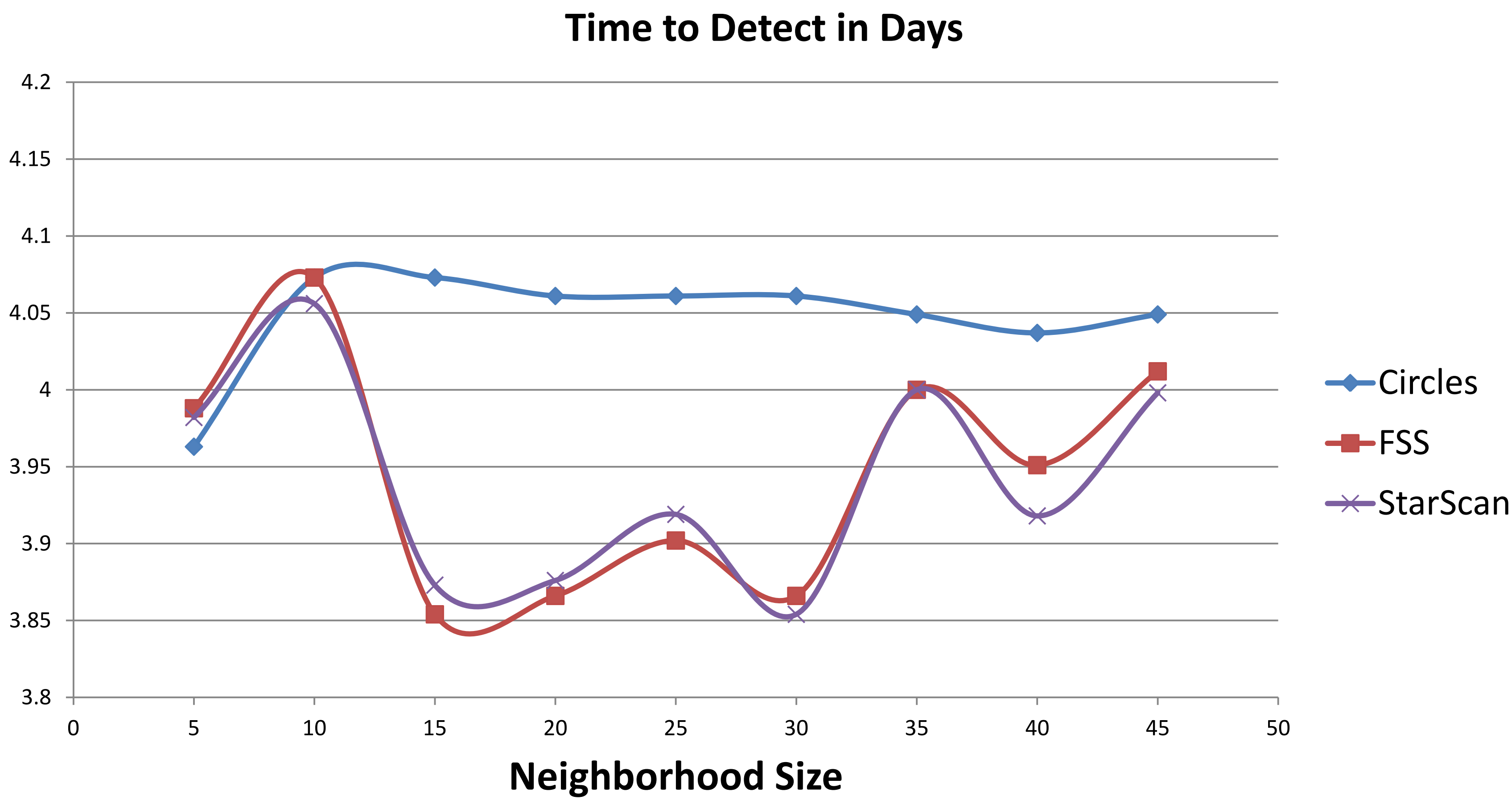
$$Overlap = 1$$ Perfect Match

$$Overlap = 0$$ Completely Disjoint

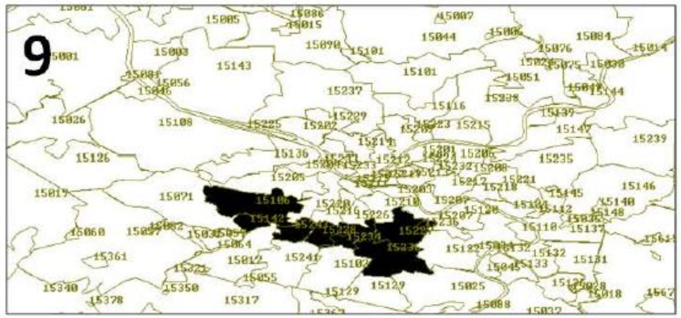
BARD Results: Spatial Overlap

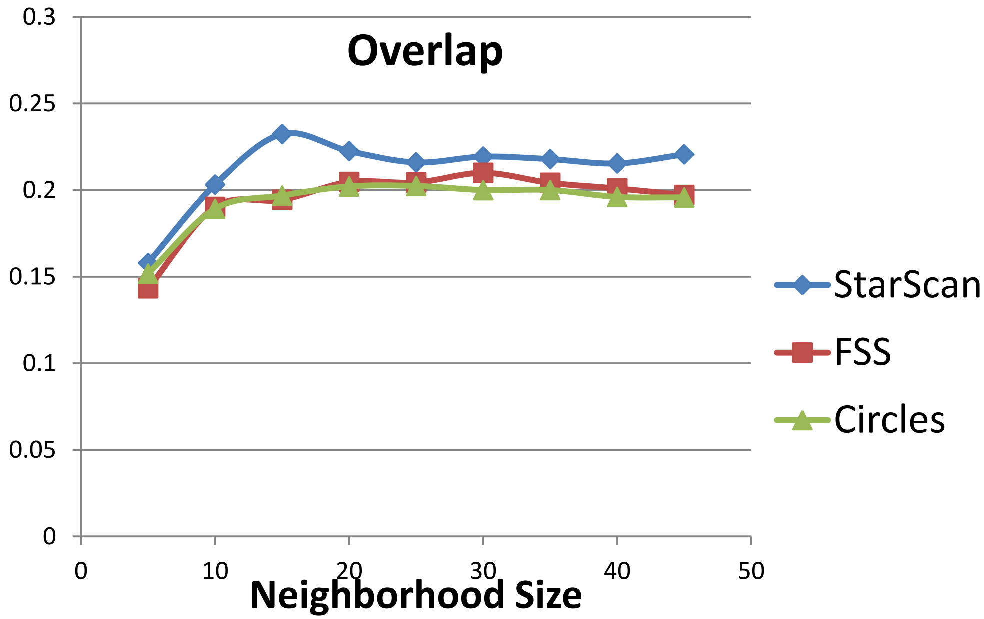


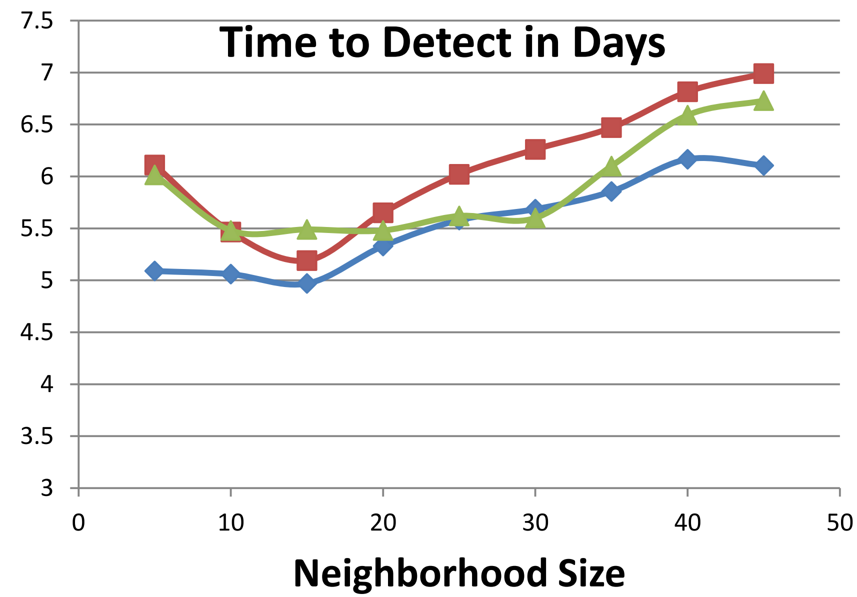
BARD Results: Time to Detect at a fixed false positive rate



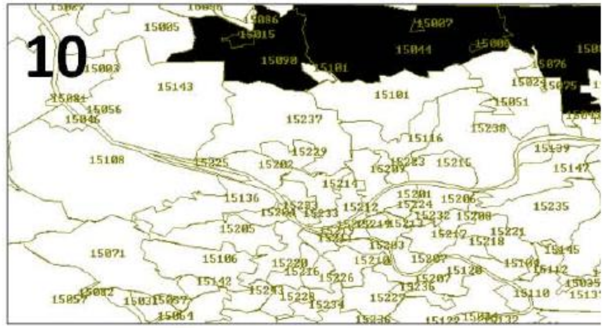
Simulated Injects in real-world Emergency Department data.

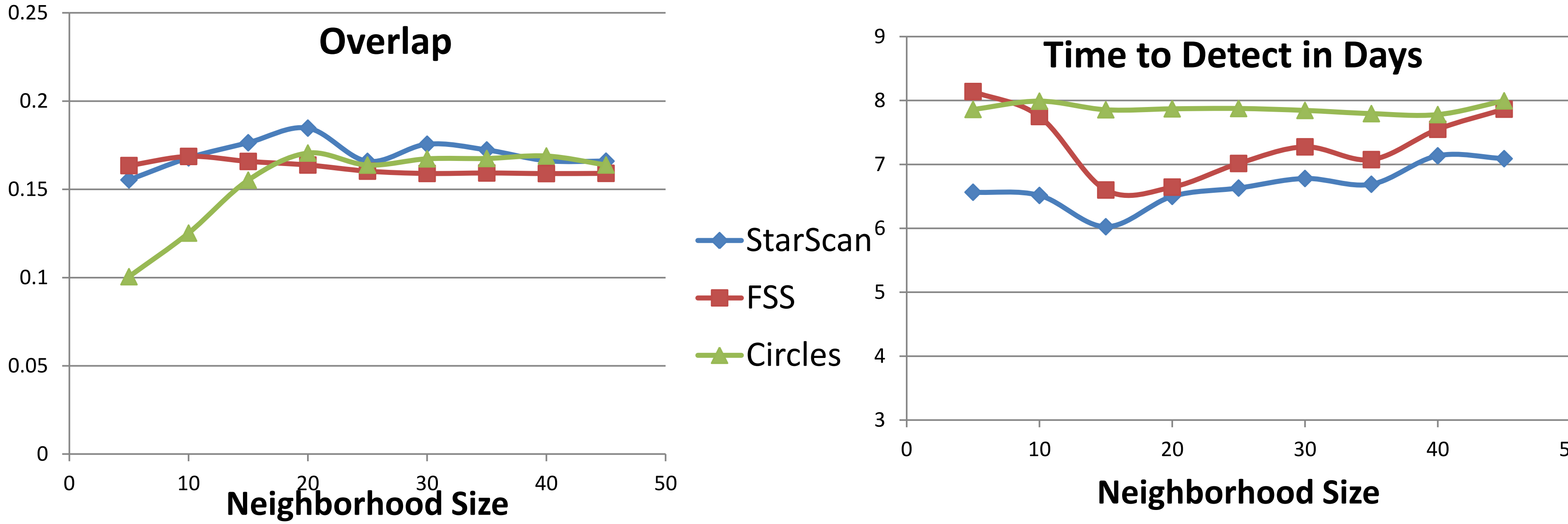






Simulated Injects (continued)





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

Thank Q?