Support Vector Subset Scan for Spatial Pattern Detection

Dylan Fitzpatrick, Yun Ni, and Daniel B. Neill
Event and Pattern Detection Laboratory
Carnegie Mellon University

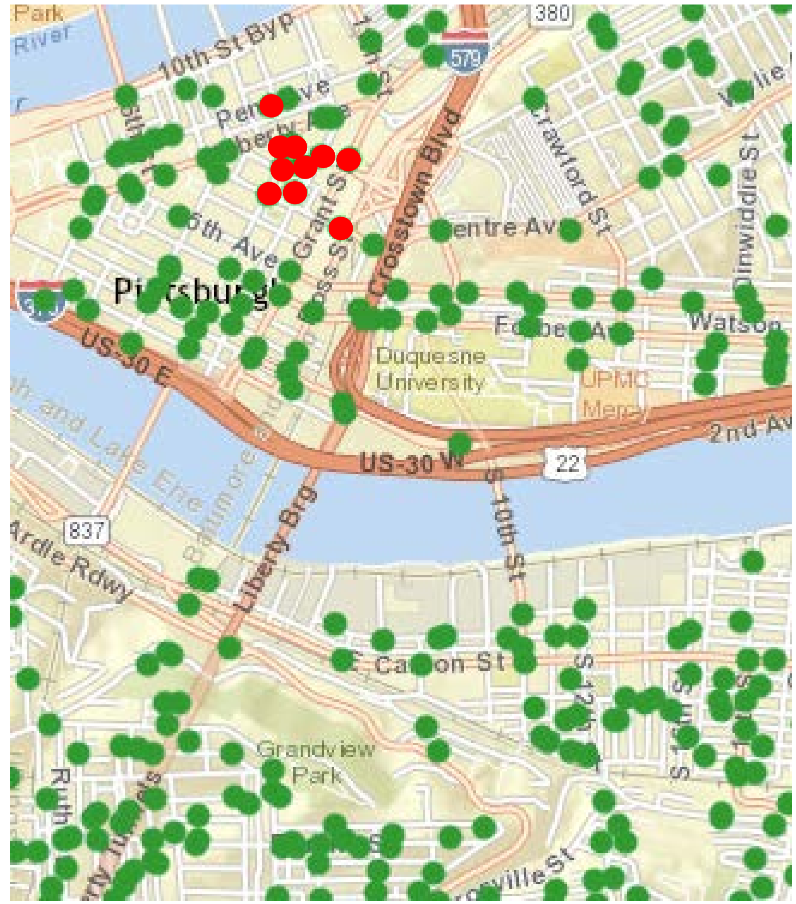
This work was partially supported by
NSF grant IIS-0953330

Detecting Spatial Clusters

Policy decisions often benefit from detecting and characterizing patterns in spatial or spatiotemporal data

E.g.,

- Detecting outbreaks of mosquito-borne disease through insect testing
- Identifying crime hot-spots at a city-block level from police reports



Detecting Spatial Clusters

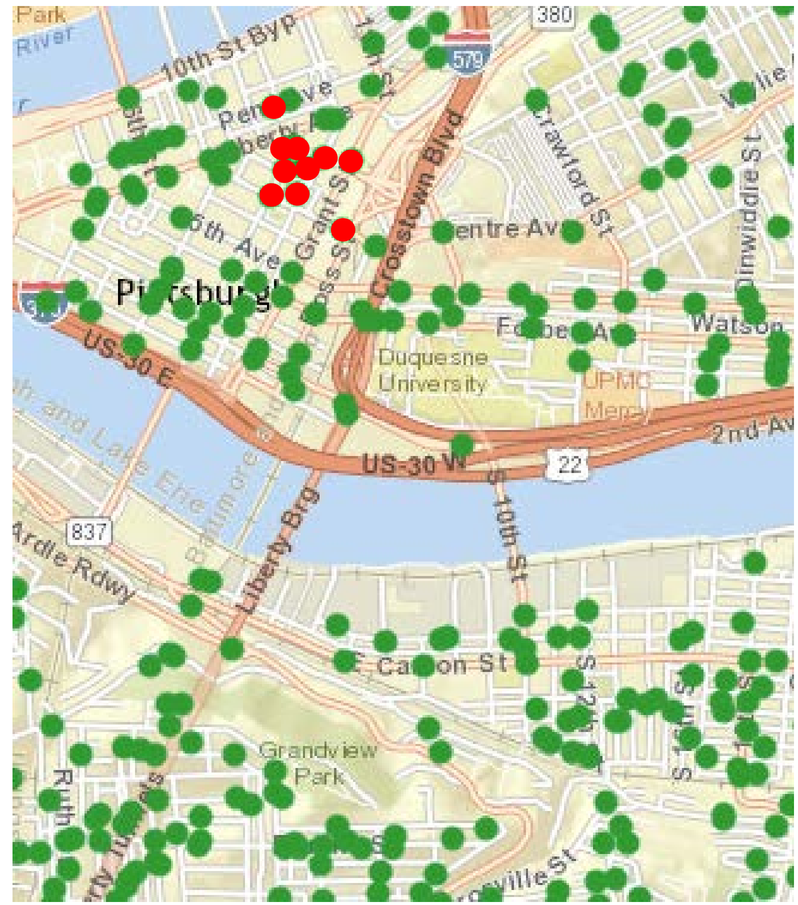
Given a data snapshot for spatial locations, can we find regions with observed values significantly higher than expected?

Goal:

- Method with high detection power that is computationally efficient

Challenges:

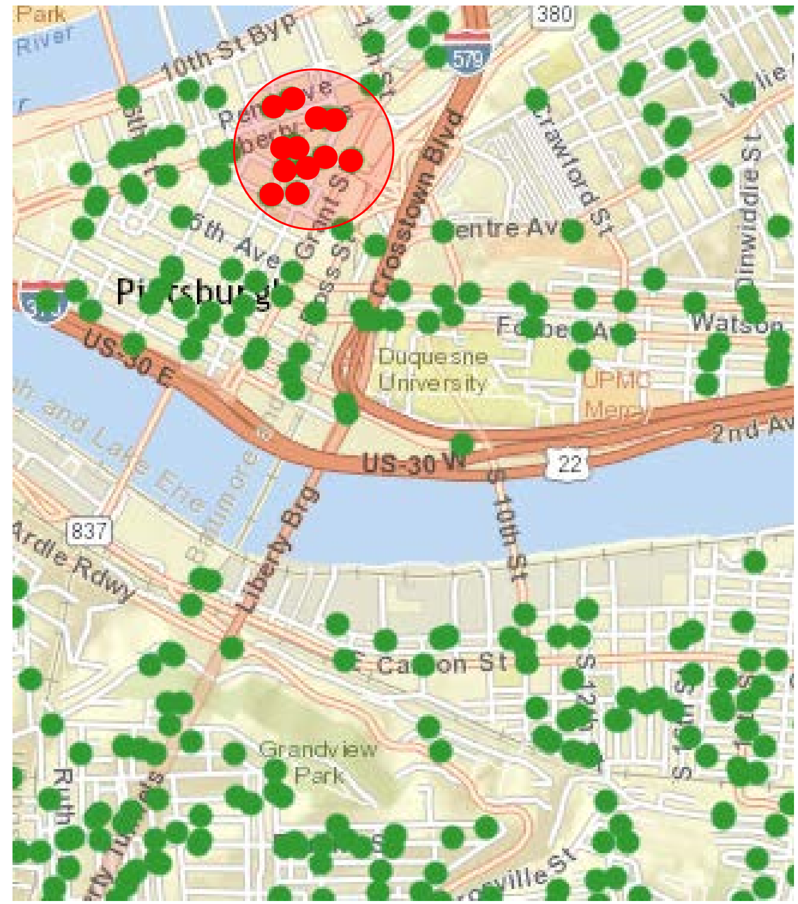
- 2^N different subsets for N locations.
- Regions may be highly irregular in shape.



Detecting Spatial Clusters

Spatial Scan Statistic (Kulldorff, 1997):

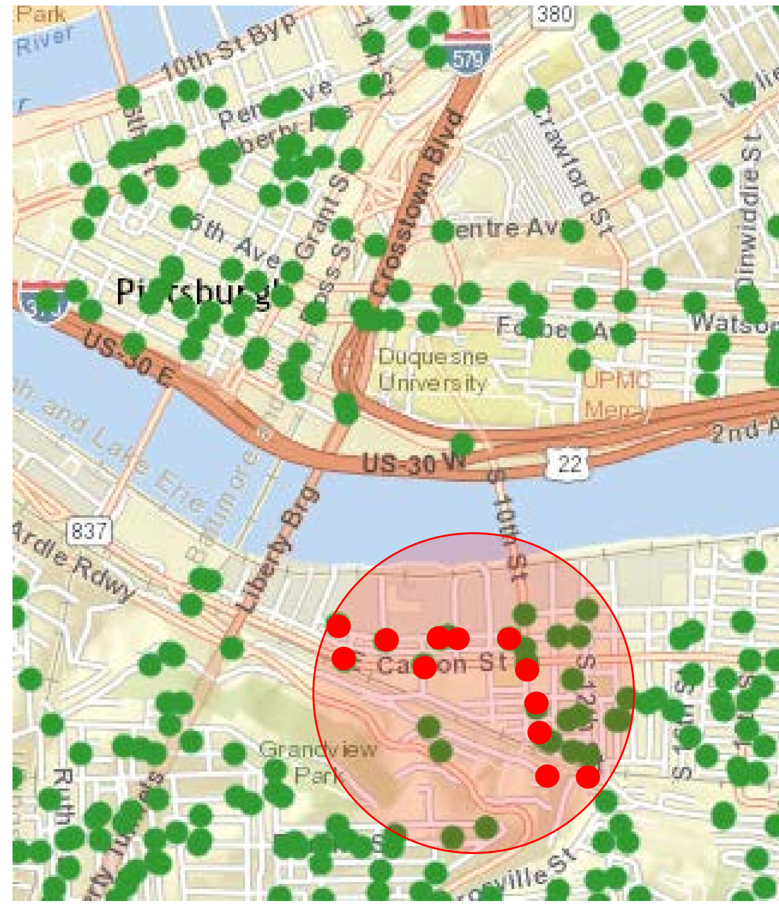
- Searches over circular regions
- **High detection power** for affected regions of corresponding shape
- **Low detection power** for irregular clusters



Detecting Irregular Spatial Clusters

Fast Subset Scan (Neill, 2012):

- Finds most anomalous subset over entire region (or constrained subregions) efficiently and exactly
- May result in sparse or spatially dispersed patterns
- Can we encourage spatial coherence without losing ability to detect subtle and irregular patterns?



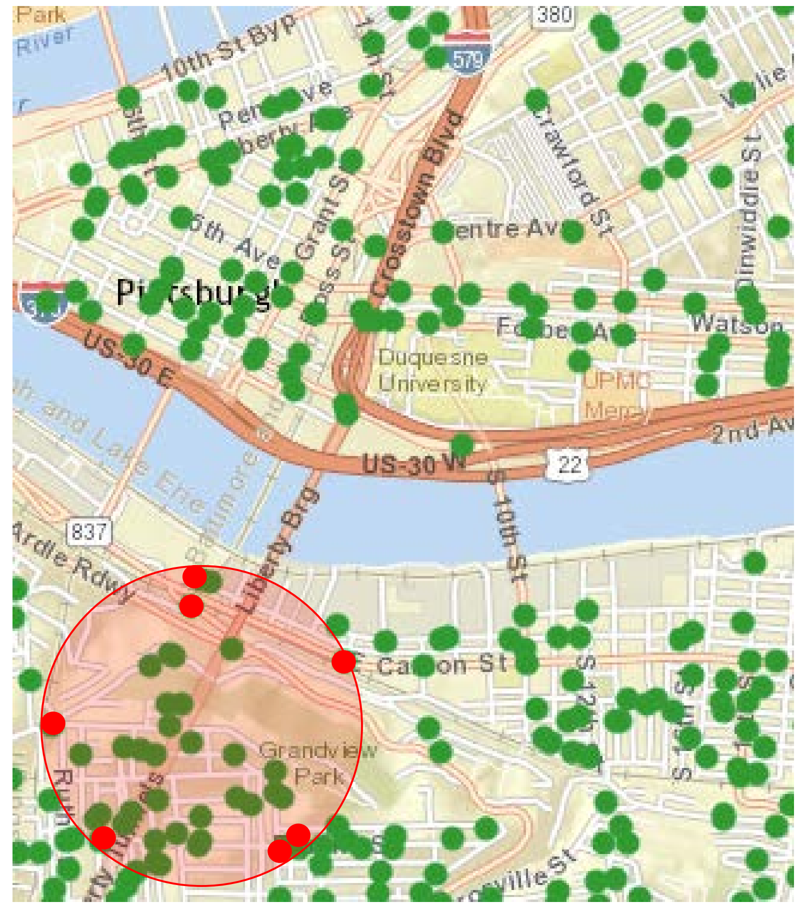
Detecting Irregular Spatial Clusters

Fast Subset Scan (Neill, 2012):

- Finds most anomalous subset over entire region (or constrained subregions) efficiently and exactly

- May result in sparse or spatially dispersed patterns

- Can we encourage spatial coherence without losing ability to detect subtle and irregular patterns?



Expectation-Based Scan Statistics

α_i – Binary variable indicating inclusion of location i in the subset being scored

c_i – Observed counts at location i

b_i – Expected counts at location i

q – Multiplicative increase for locations in subset (relative risk)

Poisson Example:

$$H_0 : c_i \sim Poisson(b_i)$$

$$H_1 : c_i \sim Poisson(qb_i), q > 1$$

$$F(\boldsymbol{\alpha}) = \max_{q>1} \log \frac{Pr(Data|H_1(\boldsymbol{\alpha}))}{Pr(Data|H_0)}$$

Adding Location-Specific Penalties

Penalized Fast Subset Scan (Speakman, McFowland, Somanchi, and Neill, 2016):

Location-specific terms can be added to score function:

$$F_{penalized}(\boldsymbol{\alpha}) = \max_{q>1} \sum_{i=1}^{N} \alpha_i(\lambda_i + \Delta_i)$$

Easy to interpret: Δ_i terms are the prior log-odds of location i being in the true affected subset.

Easy to maximize*:* For fixed relative risk q, only include points with positive overall contribution. Optimal subset can be found by considering *O*(*N*) values of q.

Support Vector Subset Scan (SVSS)

Intuition: Find anomalous subset with large margin between affected and unaffected points

Algorithm:

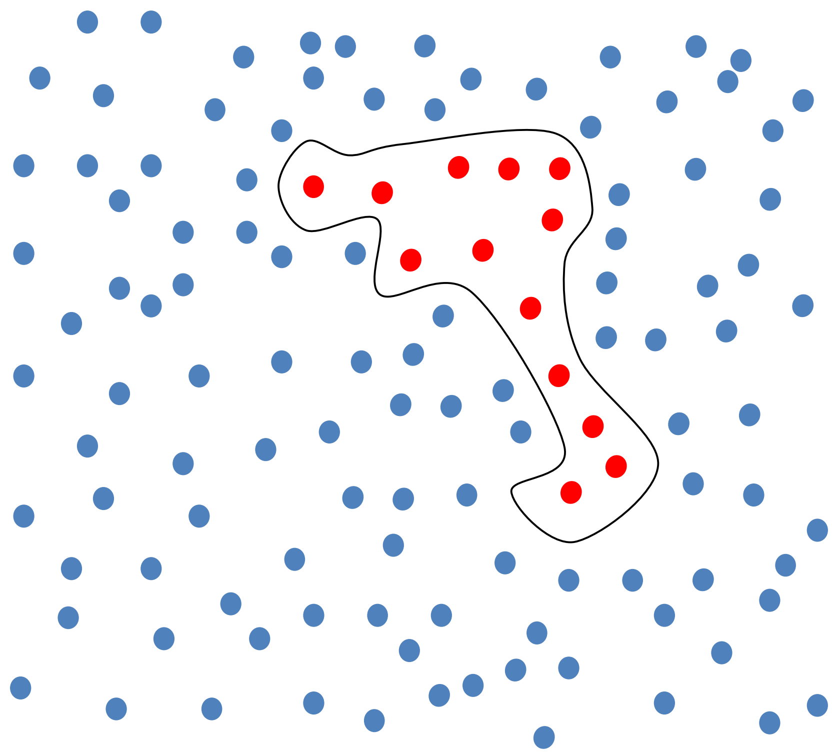
Alternately

(1) Run Penalized Fast Subset Scan (PFSS) to obtain an optimal subset, then

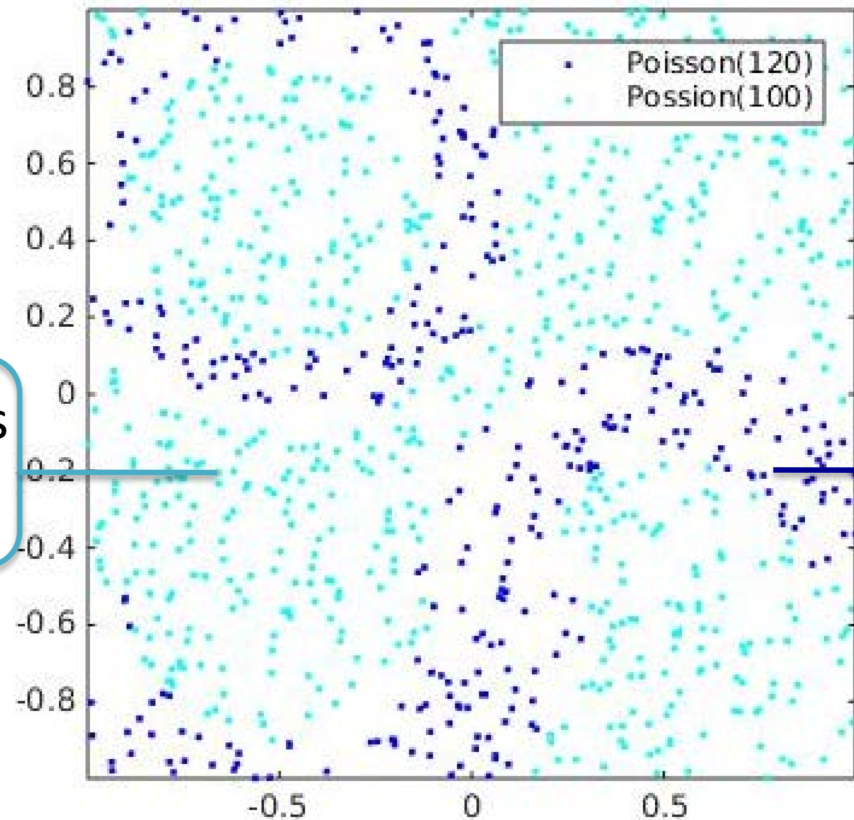
(2) Train a Support Vector Machine (SVM) classifier to maximize the margin between points within and outside of the subset.

On each iteration of PFSS, penalties are assigned based on distance to the SVM decision boundary.

Result: Irregular but spatially coherent regions



Improvement Over Iterations

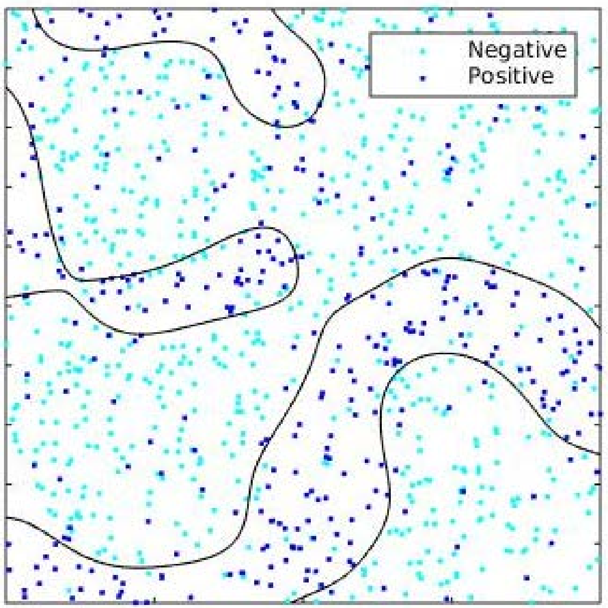


Unaffected points
~ Poisson(100)

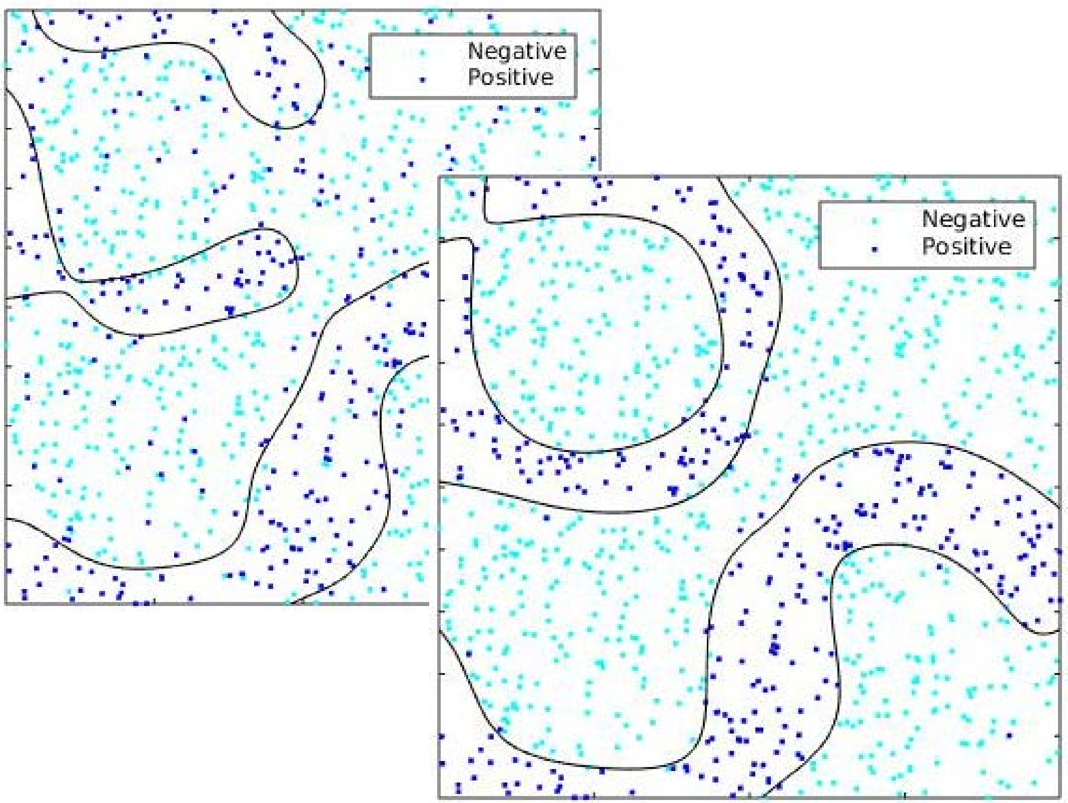
Affected points
~ Poisson(120)

Expectation = 100 for all sites

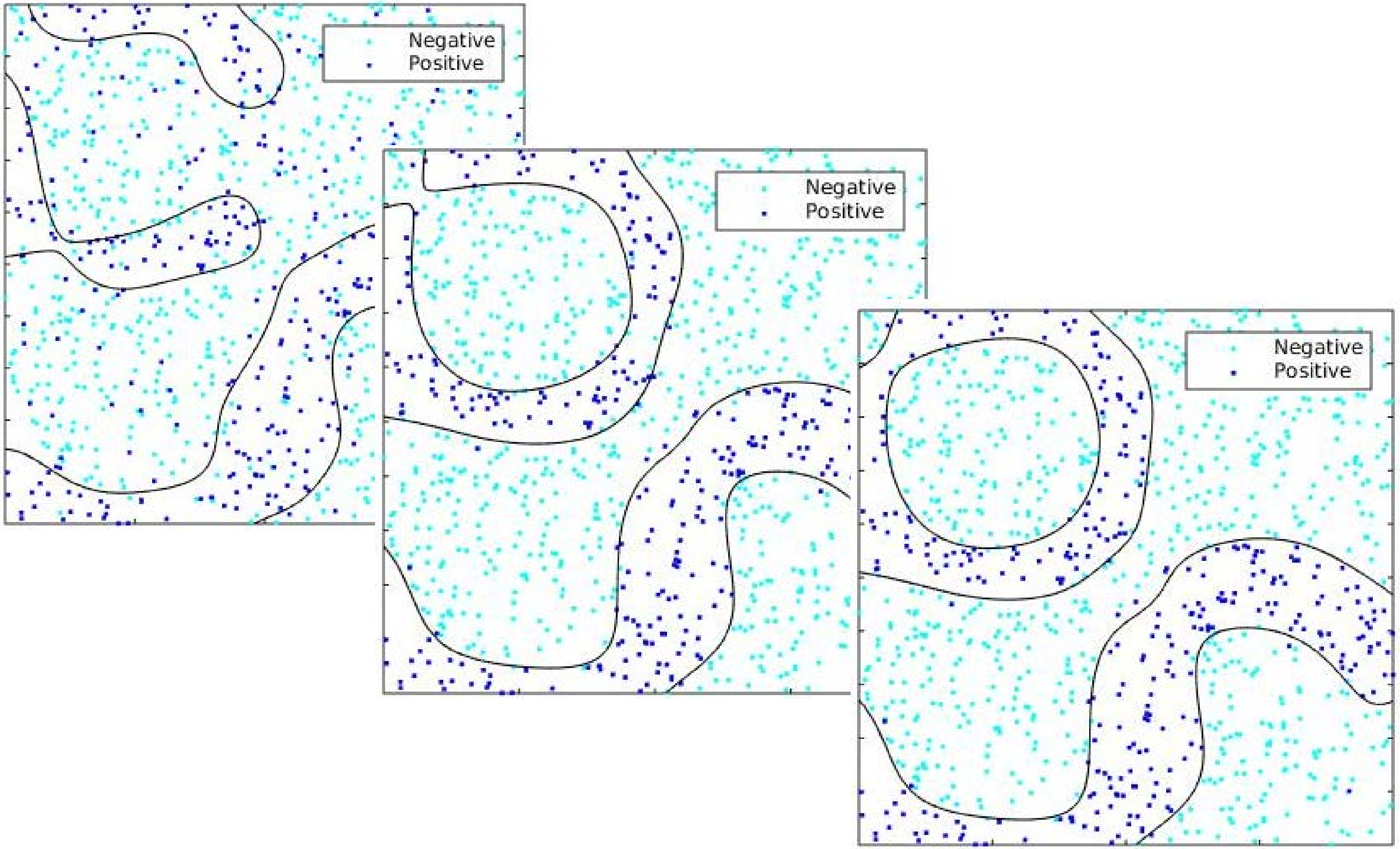
Improvement Over Iterations

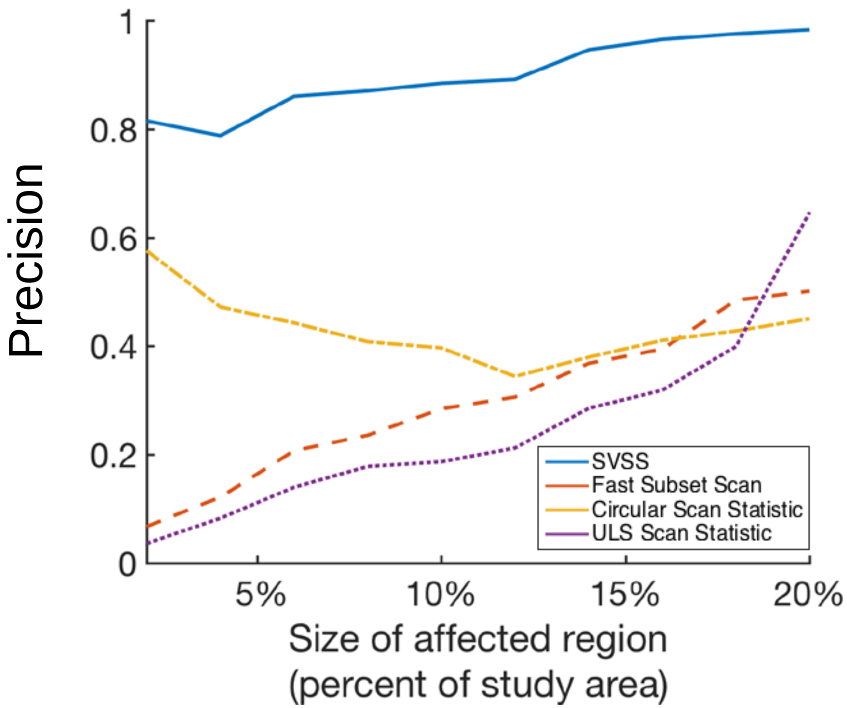


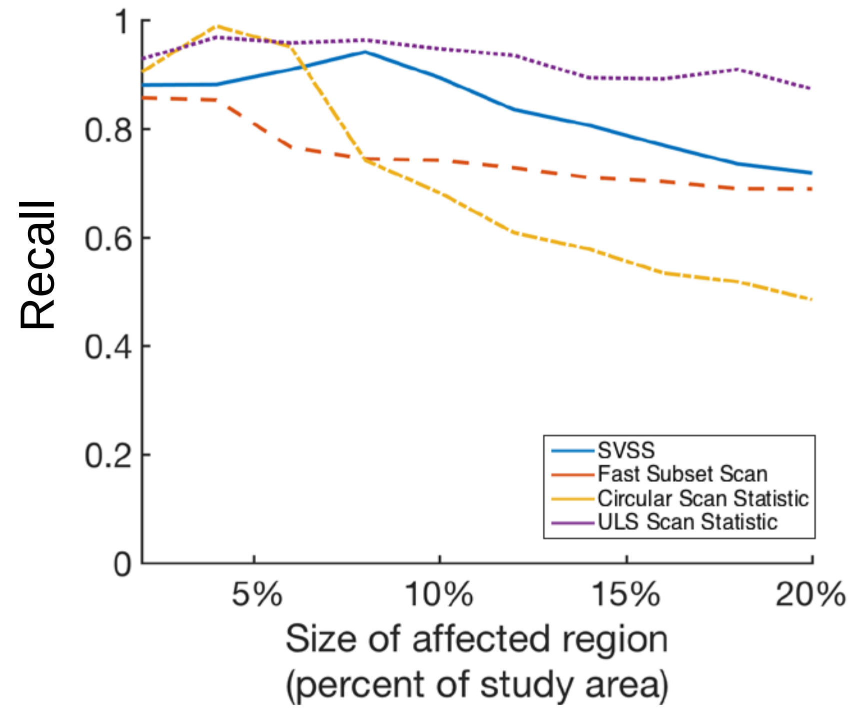
Improvement Over Iterations



Improvement Over Iterations







$$S_{true} = \text{true affected locations}$$

$$S^* = \text{detected locations}$$

$$Precision = \frac{|S_{true} \bigcap S^*|}{|S^*|}$$

$$Recall = \frac{|S_{true} \bigcap S^*|}{|S_{true}|}$$

Detecting Outbreaks of West Nile Virus (WNV)

Data:

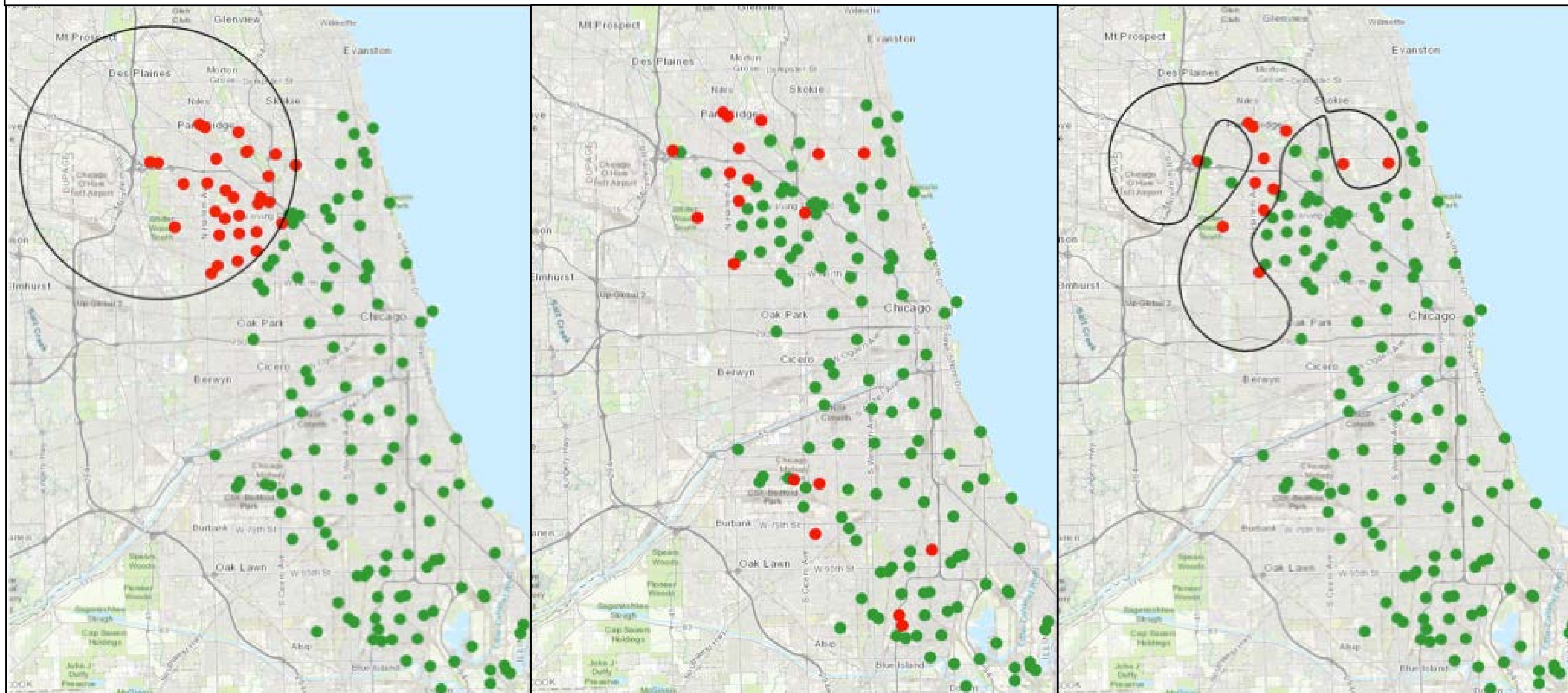
- Test results from mosquito pools tested for WNV in Chicago, IL

Timeframe:

- City-wide expected count estimated from entire 2007-2016 timeframe
- Observed counts for each pool generated from same period

Can we identify and characterize disease clusters across the city?

Clusters of West Nile Virus Detected by Three Scanning Algorithms



Circular scan | Fast subset scan | Support vector subset scan

Summary Statistics for Top WNV Patterns

- n_s – Number of points in pattern
- LLR – Unpenalized anomalousness score (log-likelihood ratio)
- q_{MLE} – Maximum likelihood estimate of relative risk
- K – Measure of geometric compactness

	n_S	LLR	q_{MLE}	K
Circular scan	30	79.9	1.75	0.86
ULS	24	108.2	1.81	0.09
FSS	19	125.9	2.02	0.08
SVSS	12	108.0	2.02	0.19

Conclusion

- **Support Vector Subset Scan** (SVSS) is a new method for detecting localized and irregularly shaped patterns which are spatially separated from non-anomalous data.
- In simulated experiments, SVSS outperforms competing methods on the task of detecting irregularly shaped patterns
- We demonstrate the utility of SVSS for disease surveillance by detecting clusters of West Nile Virus in Chicago

Thank you

djfitzpa@cmu.edu

SVSS Algorithm

Algorithm 1 Support Vector Subset Scan

procedure SVSS($\mathbf{c}, \mathbf{b}, \mathbf{x}, T_{max}, C_0, C_1$) ▷ Values $\mathbf{c}$, expectations $\mathbf{b}$, and coordinates $\mathbf{x}$

$min\_score \leftarrow \infty$

for $t := 1$ **to** T_{max} **do** ▷ T_{max} random restarts

$\xi_i(\alpha_i) \leftarrow \text{Uniform}(-C_0, C_0), \forall i = 1, ..., N$

while $\boldsymbol{\alpha}$ is changing **do**

PFSS

$\boldsymbol{\alpha} \leftarrow \underset{\boldsymbol{\alpha}}{\text{argmax}}\ F(\boldsymbol{\alpha}) - C_0/C_1 \sum_{i=1}^{N} \xi_i(\alpha_i)$ ▷ Fix $\mathbf{w}, b$ and optimize over $\boldsymbol{\alpha}$

$\boldsymbol{\xi}, \mathbf{w}, b \leftarrow \underset{\boldsymbol{\xi}, \mathbf{w}, b}{\text{argmin}} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i)$ ▷ Fix $\boldsymbol{\alpha}$, and optimize over $\mathbf{w}, b$

SVM

end while

$score \leftarrow \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\boldsymbol{\alpha})$

if $score < min\_score$ **then**

$min\_score \leftarrow score$

$\boldsymbol{\alpha}_{min} \leftarrow \boldsymbol{\alpha}$

end if

end for

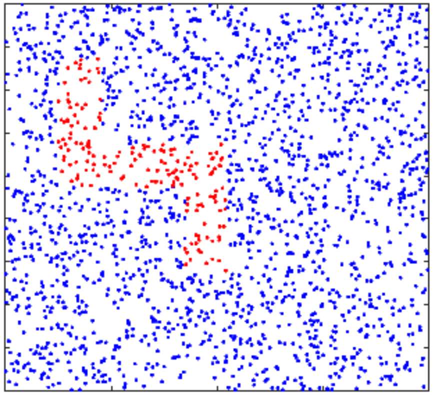
return $\boldsymbol{\alpha}_{min}$

end procedure

Evaluation Framework

- 2000 observations generated from Poisson distribution

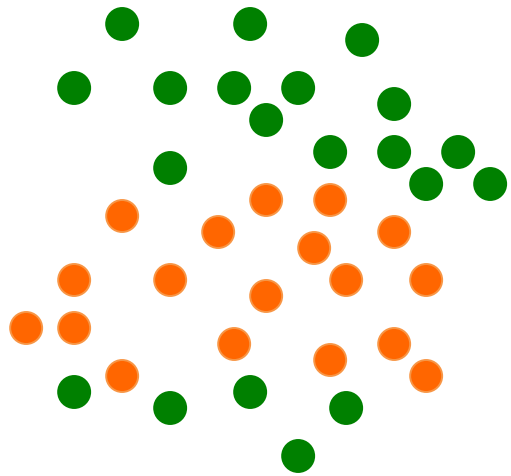
- Generated random, irregular-shaped regions of varying size with elevated counts
  - Unaffected points: $c_i \sim Poisson(100)$
  - Affected points: $c_i \sim Poisson(115)$
  - $b_i = 100$ for all points



- Compared precision and recall of top pattern at each size against:
  - Fast subset Scan (Neill, 2011)
  - Circular scan statistic (Kulldorff, 1997)
  - Upper level set scan statistic (Patil and Taillie, 2007)

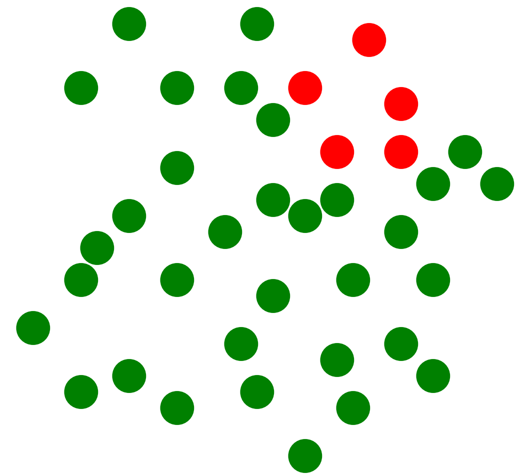
- Report averages over 100 simulations for each size

Expectation-Based Scan Statistics



Large subset, moderate risk

VS.



Small pattern, high risk

Poisson Example:

$$H_0 : c_i \sim Poisson(b_i)$$

$$H_1 : c_i \sim Poisson(qb_i), q > 1$$

$$F(\boldsymbol{\alpha}) = \max_{q>1} \log \frac{Pr(Data|H_1(\boldsymbol{\alpha}))}{Pr(Data|H_0)}$$

Adding Location-Specific Penalties

Penalized Fast Subset Scan (Speakman, McFowland, Somanchi, and Neill, 2016):

For expectation-based scan statistics in exponential family, score functions can be expressed as an additive function over points included in subset*:*

$$F(\boldsymbol{\alpha}) = \max_{q>1} F(\boldsymbol{\alpha}|q) \text{ where } F(\boldsymbol{\alpha}|q) = \sum_{i=1}^{N} \alpha_i \lambda_i$$

and λ_i depends only on observed count c_i, expected count b_i , and fixed relative risk q

SVSS Optimization Problem

Let $\boldsymbol{x}_i$ be the spatial coordinates of location i:

$$\min_{\alpha,\xi,\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\boldsymbol{\alpha})$$

$$\alpha_i \in \{0, 1\}, \forall i = 1, ..., N$$

$$\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - b))$$

SVSS Optimization Problem

Let $\boldsymbol{x}_i$ be the spatial coordinates of location i:

$$\min_{\alpha,\xi,\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\boldsymbol{\alpha})$$

$$\alpha_i \in \{0,1\}, \forall i = 1, ..., N$$

$$\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - b))$$

Problem: Objective is not convex. We optimize with alternate minimization and multiple random restarts.

SVSS Optimization Problem

Let x_i be the spatial coordinates of location i:

$$\min_{\alpha,\xi,\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\boldsymbol{\alpha})$$

$$\alpha_i \in \{0,1\}, \forall i = 1, ..., N$$

$$\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - b))$$

PFSS Problem

Location-specific penalties = Distance to SVM hyperplane

SVSS Optimization Problem

Let $\boldsymbol{x}_i$ be the spatial coordinates of location i:

$$\min_{\alpha,\xi,\mathbf{w},b} \boxed{\frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i)} - C_1 F(\boldsymbol{\alpha})$$

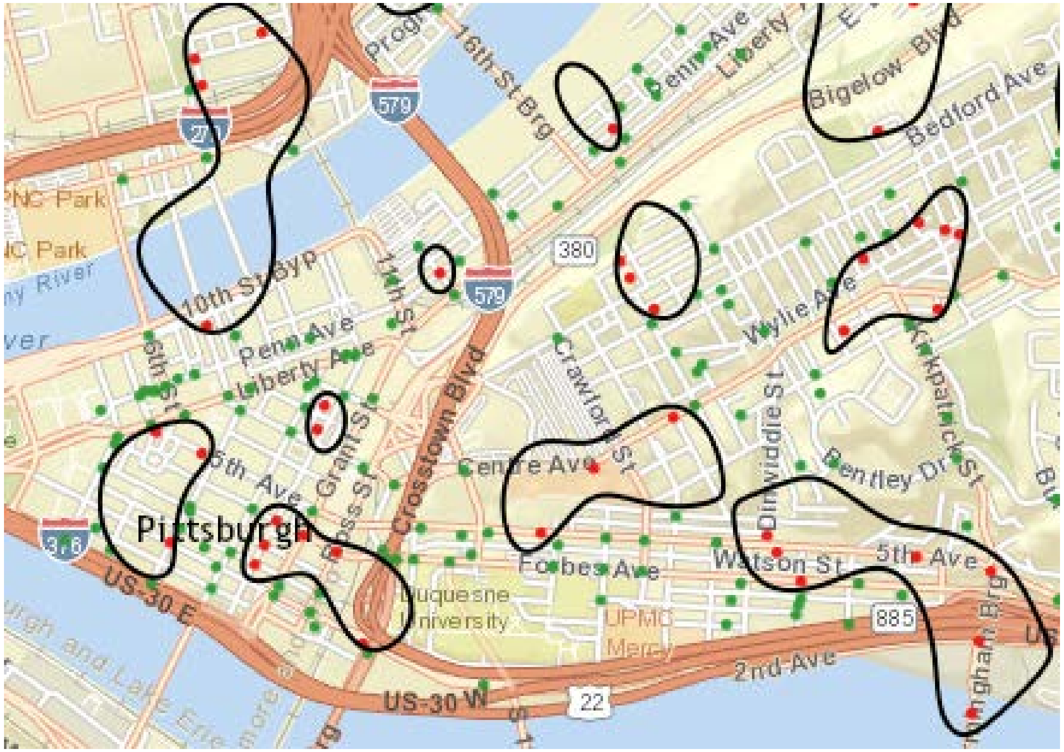
$$\alpha_i \in \{0,1\}, \forall i = 1, ..., N$$

$$\boxed{\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - b))}$$

SVM Problem

Binary data labels = Included/Not included in subset

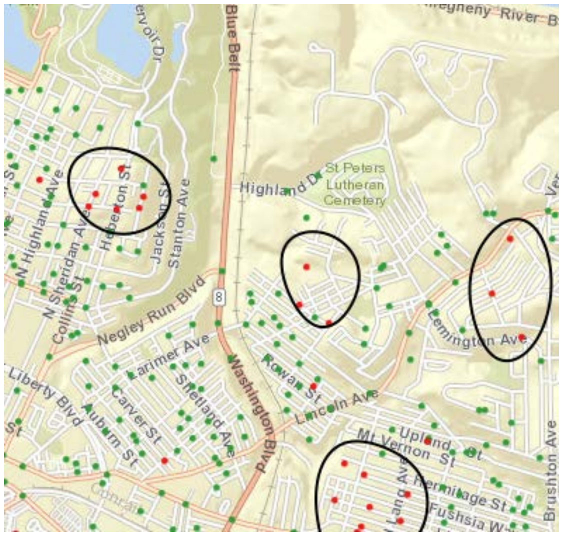
Ranking Disconnected Regions



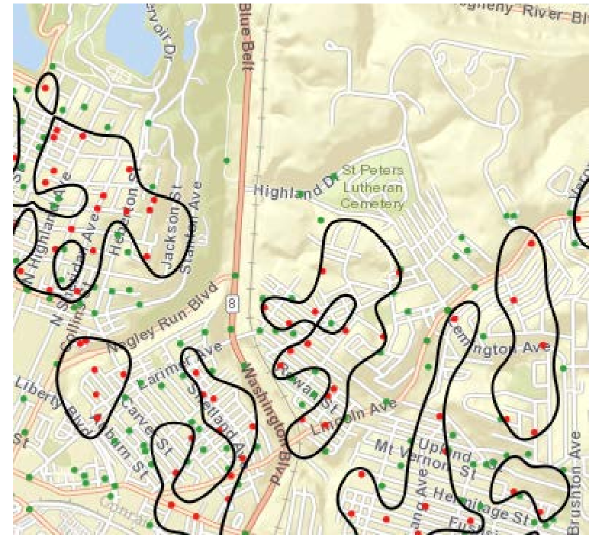
How can we rank the connected regions of the best subset?

Solution: Maximize penalized log-likelihood ratio over connected components of SVM decision boundary

Tuning model parameters

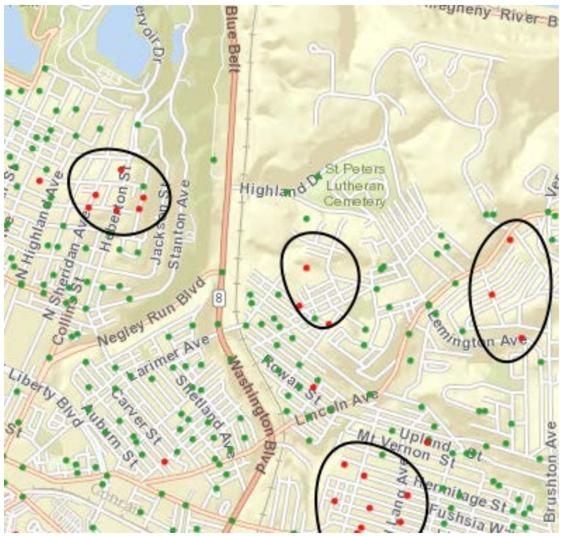


VS.

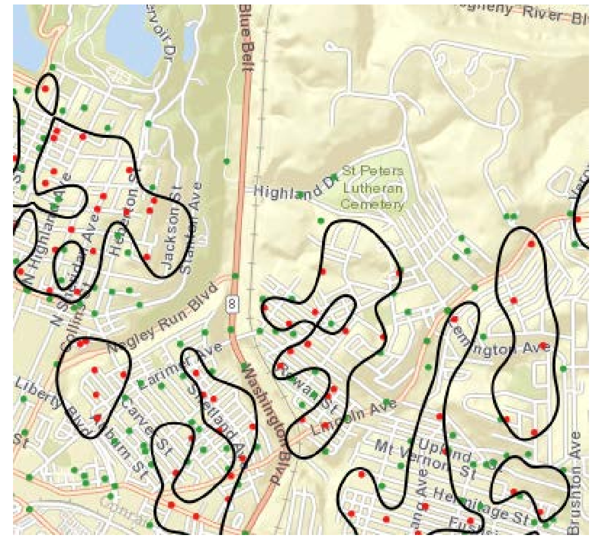


Goal: Find parameter combination that generates best subset with high log-likelihood ratio (LLR) and some minimum level of geometric compactness

Tuning model parameters



VS.



Tuning procedure:

1. Define measure of geometric compactness K (Duzcmal et al., 2006):

$$K(z) = \frac{4\pi \mathrm{A}(z)}{H(z)^2}$$ where $$A(z) = \text{Area of } z, \quad H(z) = \text{Perimeter of convex hull of } z$$

2. Maximize LLR of best subset over parameter settings with top SVM component meeting minimum compactness threshold

Support Vector Machine

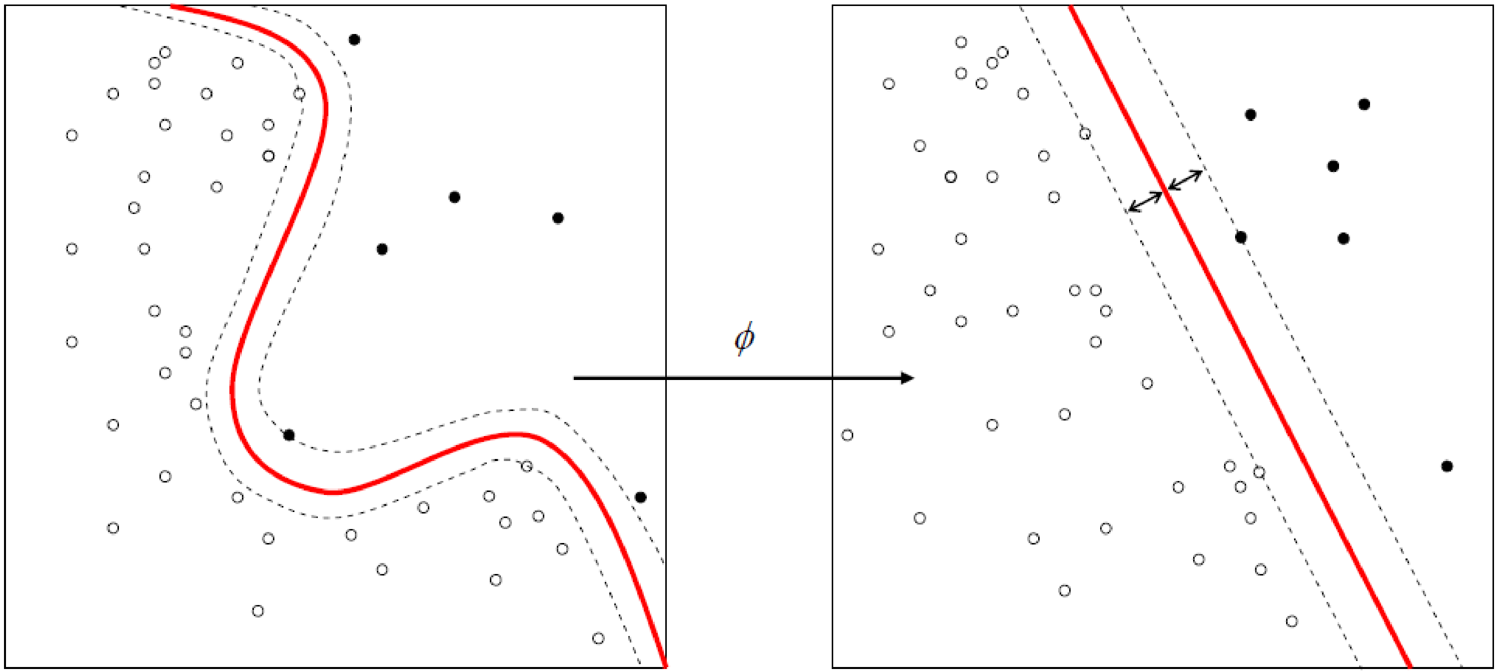


Image Source: Wikipedia

Classification algorithm that finds the separating hyperplane which maximizes the margin between positive and negative data points

Support Vector Machine

$$\min_{\xi,\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$$

$$\xi_i \geq 0, \forall i = 1, ..., N$$

$$y_i(\mathbf{w}\cdot\phi(\mathbf{x}_i) - b) \geq 1 - \xi_i, \forall i = 1, ..., N$$

where:

- weight vector **w** and bias term *b* define a hyperplane
- ξ_i terms allow for approximation in case data are not linearly separable
- ϕ is a transformation to high-dimensional feature space allowing for non-linear decision boundaries
- $\mathbf{w}\cdot\phi(\mathbf{x}_i) - b$ is a measure of distance from point x_i to the hyperplane

Adding Location-Specific Penalties

Penalized Fast Subset Scan (Speakman, McFowland, Somanchi, and Neill, 2015):

Distribution	$\lambda_i(q)$
Poisson	$c_i \log q + b_i(1-q)$
Gaussian	$c_i b_i \frac{(q-1)}{\sigma_i^2} + b_i^2(\frac{1-q^2}{2\sigma_i^2})$
exponential	$\frac{c_i}{b_i}(1-\frac{1}{q}) - \log q$
binomial	$c_i \log q + (n_i - c_i)\log(\frac{n_i - qb_i}{n_i - b_i})$
negative binomial	$c_i \log q + (r_i + c_i)\log(\frac{r_i + b_i}{r_i + qb_i})$

Computing Penalties

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\ F(\boldsymbol{\alpha}) - \frac{C_0}{C_1}\sum_{i=1}^{N} \xi_i(\alpha_i)$$

$$\xi_i(\alpha_i) = \begin{cases} \max(0, 1 - \mathbf{w}\cdot\phi(\mathbf{x}_i) + b), & y_i = 2\alpha_i - 1 = +1) \\ \max(0, 1 + \mathbf{w}\cdot\phi(\mathbf{x}_i) - b), & y_i = 2\alpha_i - 1 = -1) \end{cases}$$

How to fit into PFSS framework?

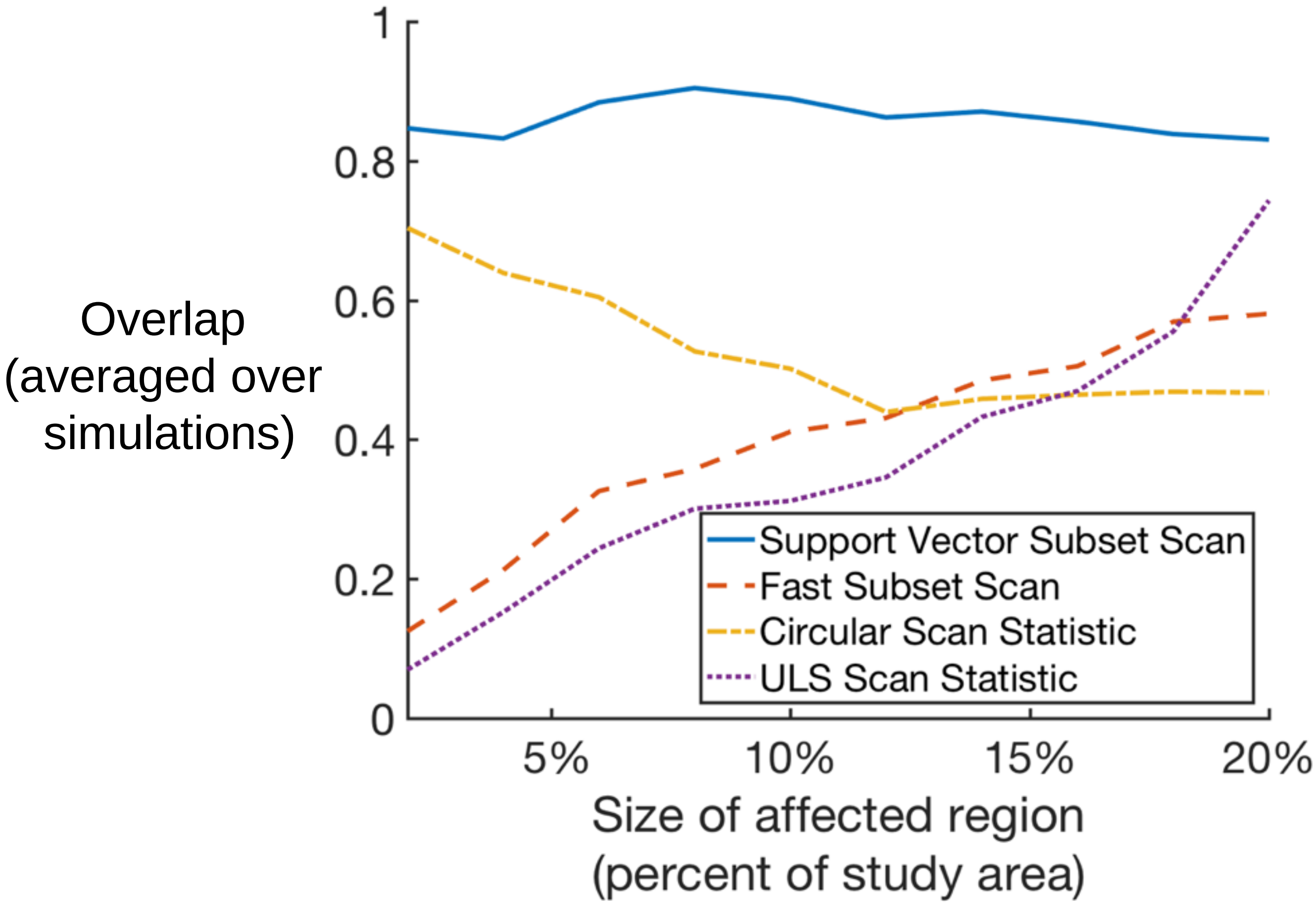
Needed: Element-specific penalties for included sites

Computing Penalties

EQUIVALENT:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\ F(\boldsymbol{\alpha}) - \frac{C_0}{C_1}\sum_{i=1}^{N}\alpha_i\Delta_i$$

$$\begin{aligned}
\Delta_i &= \max(0, 1 - \mathbf{w}\cdot\phi(\mathbf{x}_i) + b) - \max(0, 1 + \mathbf{w}\cdot\phi(\mathbf{x}_i) - b) \\
&= \begin{cases} \mathbf{w}\cdot\phi(\mathbf{x}_i) - b + 1, & \mathbf{w}\cdot\phi(\mathbf{x}_i) - b \geq 1 \\ 2(\mathbf{w}\cdot\phi(\mathbf{x}_i) - b), & \mathbf{w}\cdot\phi(\mathbf{x}_i) - b \in (-1, 1) \\ \mathbf{w}\cdot\phi(\mathbf{x}_i) - b - 1, & \mathbf{w}\cdot\phi(\mathbf{x}_i) - b \leq -1 \end{cases} \\
&= [\mathbf{w}\cdot\phi(\mathbf{x}_i) - b > -1](\mathbf{w}\cdot\phi(\mathbf{x}_i) - b + 1) + \\
&\quad [\mathbf{w}\cdot\phi(\mathbf{x}_i) - b < 1](\mathbf{w}\cdot\phi(\mathbf{x}_i) - b - 1)
\end{aligned}$$

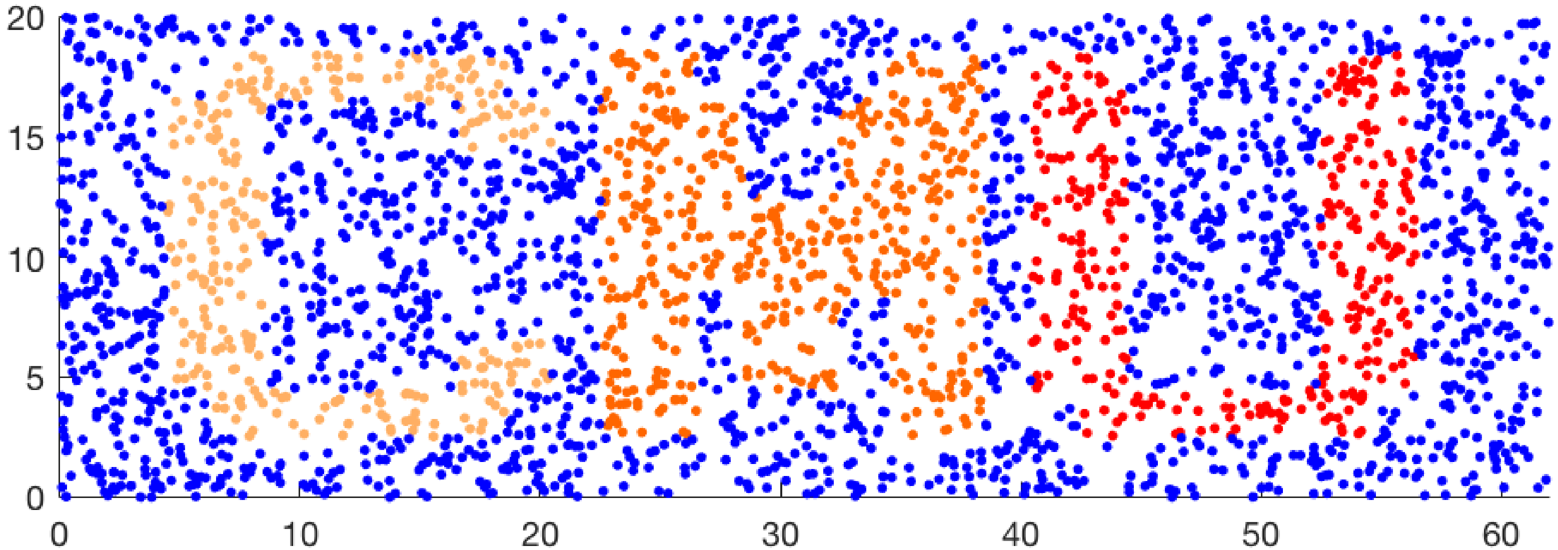


S_{true} = true affected locations

S^* = detected locations

$$Overlap = \frac{|S_{true} \bigcap S^*|}{|S_{true} \bigcup S^*|}$$

Detecting Letter-Shaped Regions

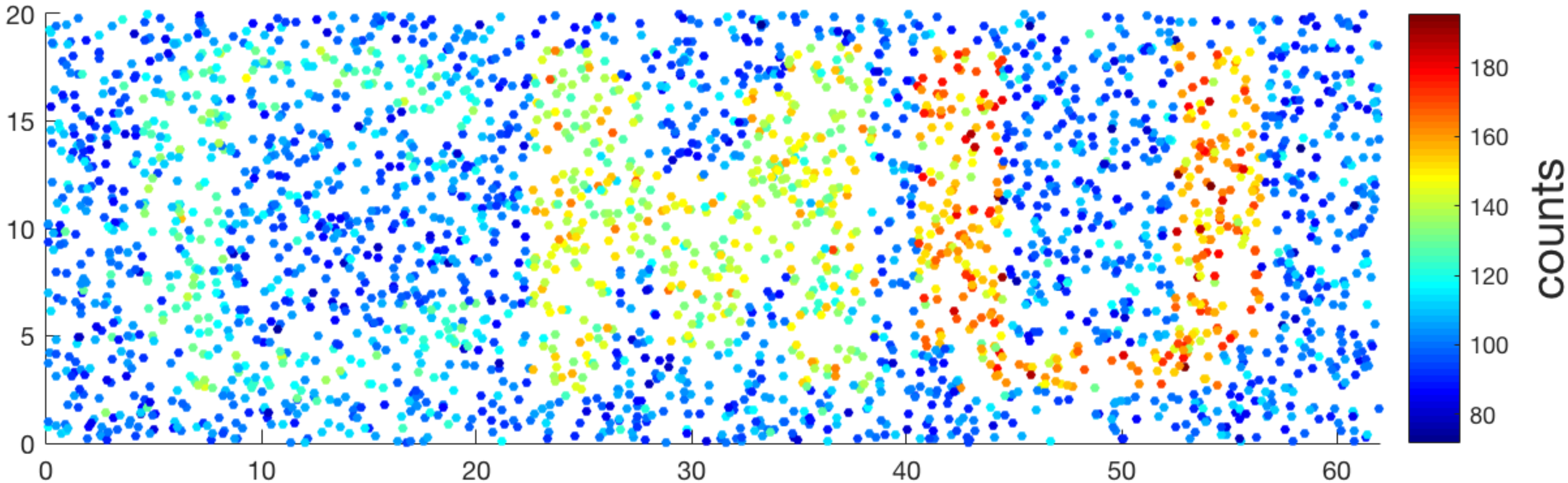


● $c_i \sim Poisson(100)$ ● $c_i \sim Poisson(140)$

● $c_i \sim Poisson(120)$ ● $c_i \sim Poisson(160)$

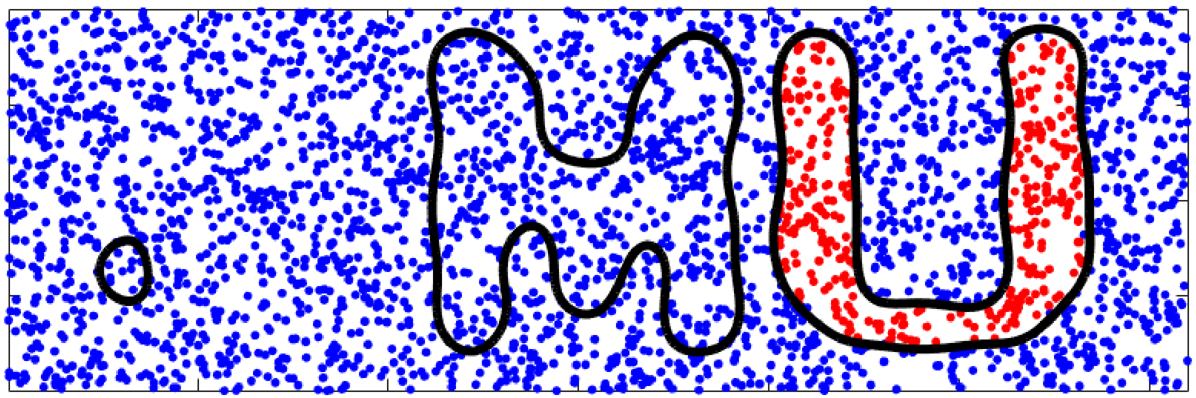
All points: b_i = 100

Detecting Letter-Shaped Regions

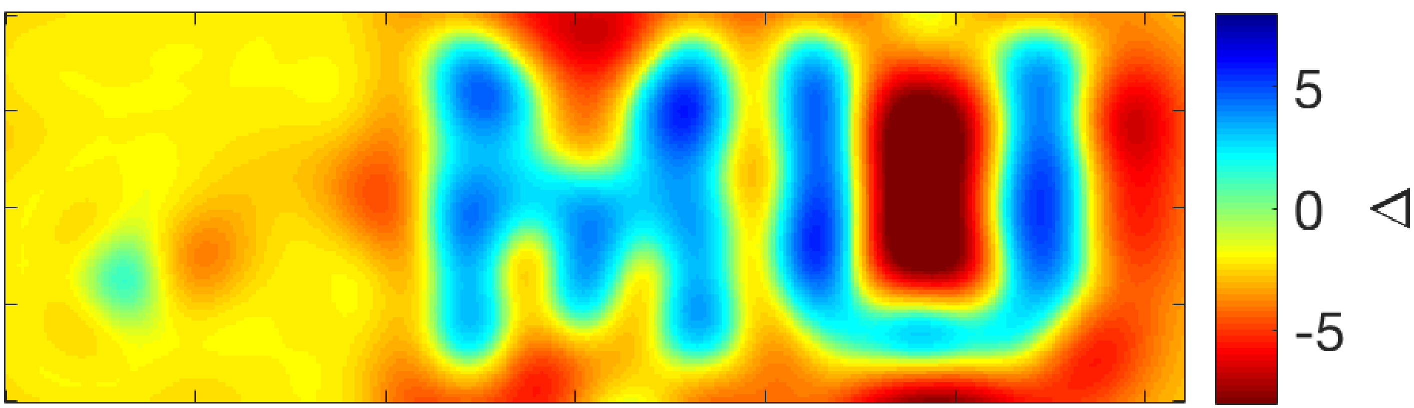


All points: $b_i = 100$

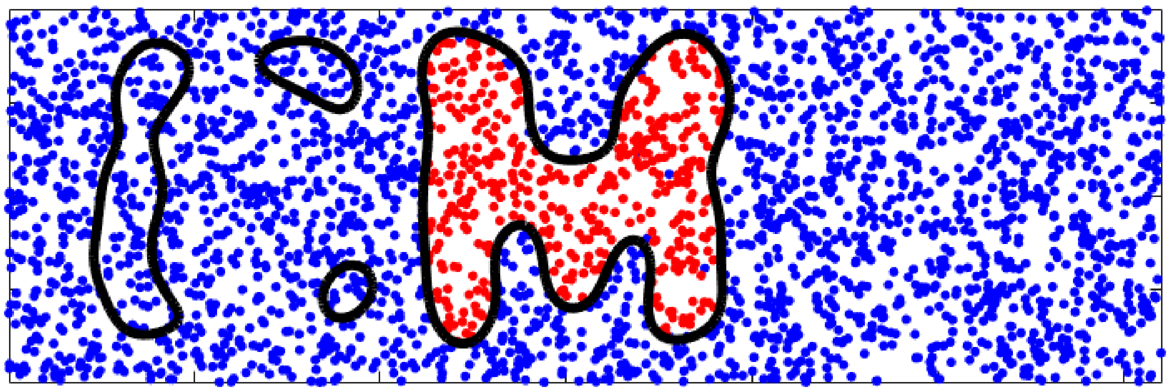
Detecting Letter-Shaped Regions



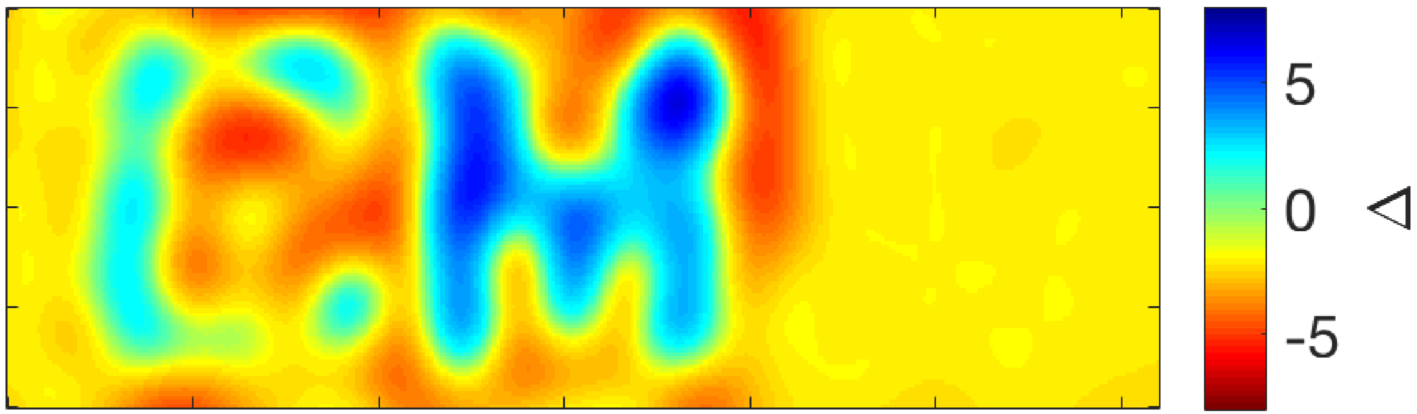
Best connected SVM region



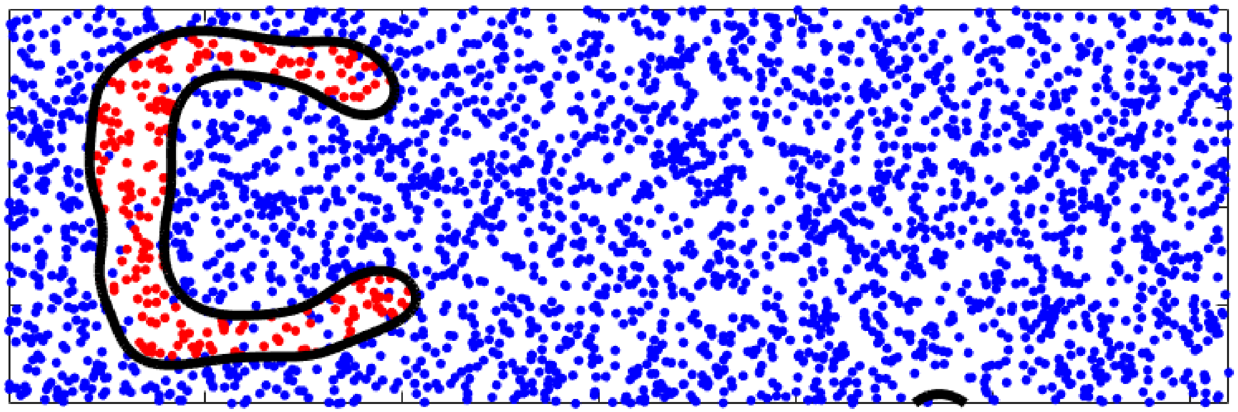
Detecting Letter-Shaped Regions



2nd Best connected SVM region



Detecting Letter-Shaped Regions



3rd Best connected SVM region