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

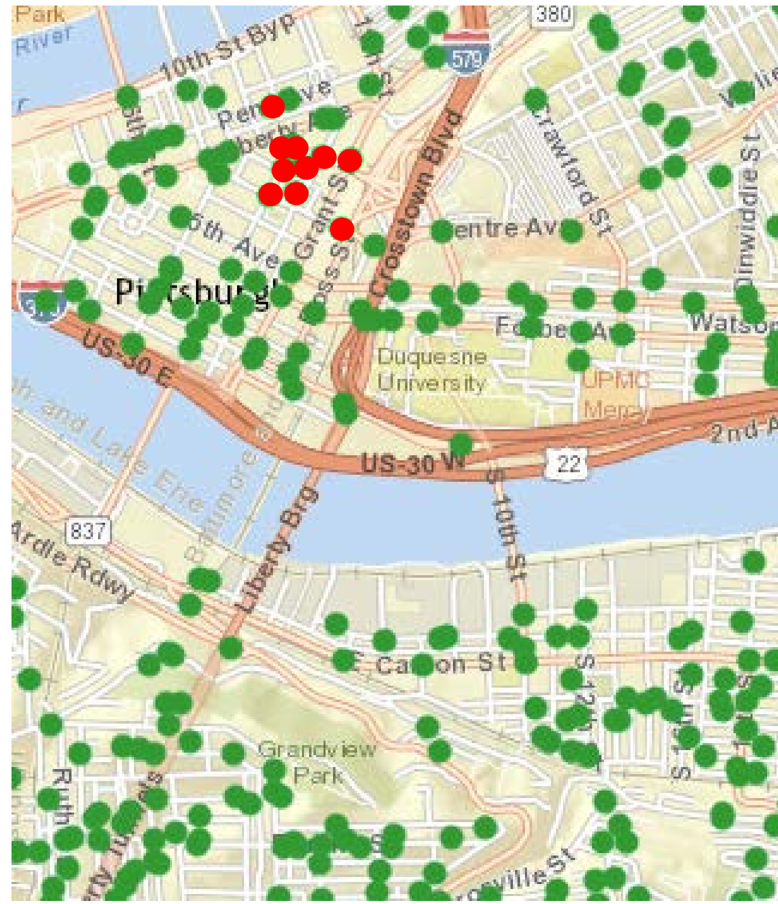
Carnegie Mellon University
EPD Lab
EVENT AND PATTERN DETECTION LABORATORY

Detecting Spatial Clusters

Given a set of data streams, can we find regions with counts significantly higher than expected?

Goal: Method with high detection power that is computationally efficient

Problem: Regions may be highly irregular in shape. 2^N different subsets.



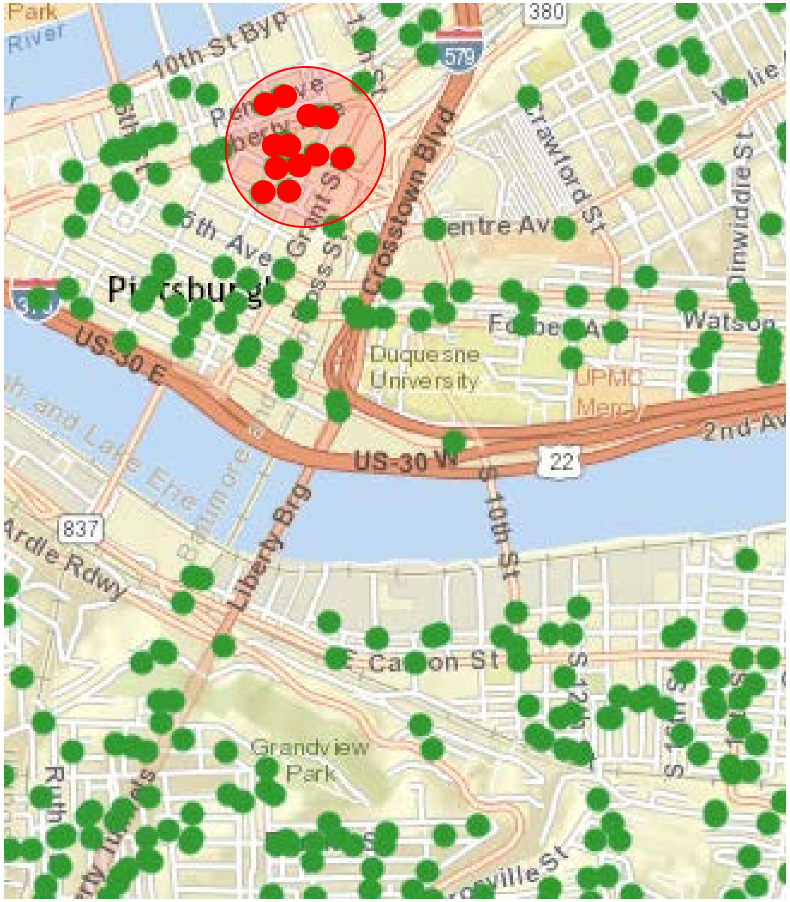
Detecting Spatial Clusters

Spatial Scan Statistic (Kulldorff, 1997):

Searches over circular regions

High detection power for
affected regions of corresponding shape

Low detection power for
irregular clusters

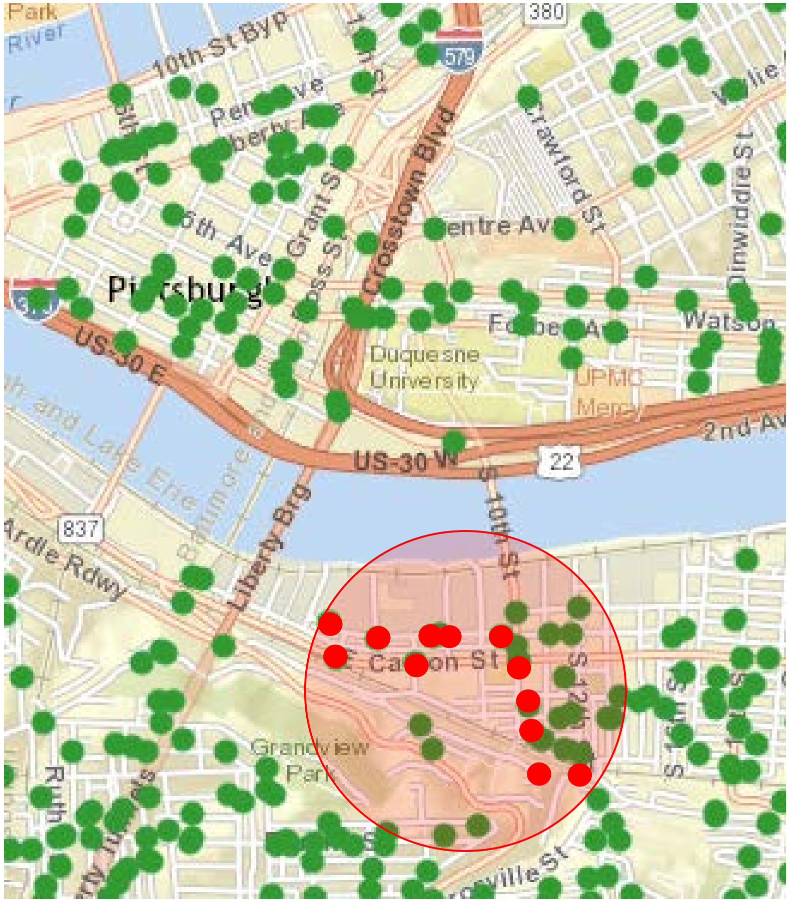


Detecting Irregular Spatial Clusters

Fast Subset Scan (Neill, 2011):

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

Can we impose spatial constraints without losing detection power for subtle and irregular patterns?

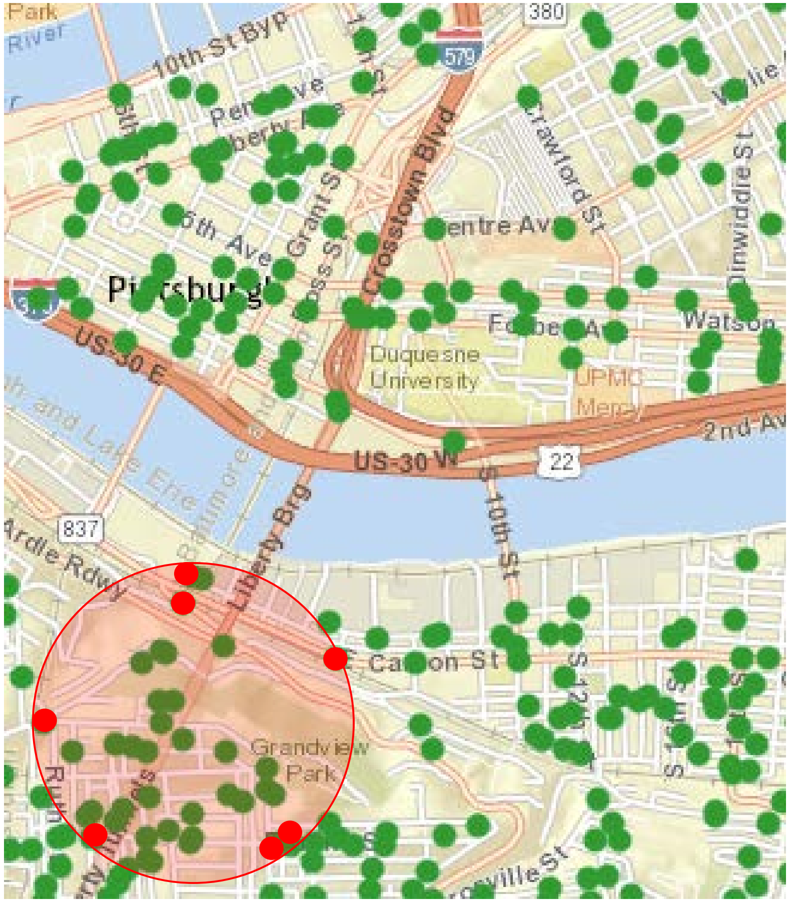


Detecting Irregular Spatial Clusters

Fast Subset Scan (Neill, 2011):

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

Can we impose spatial constraints without losing detection power for subtle and irregular patterns?

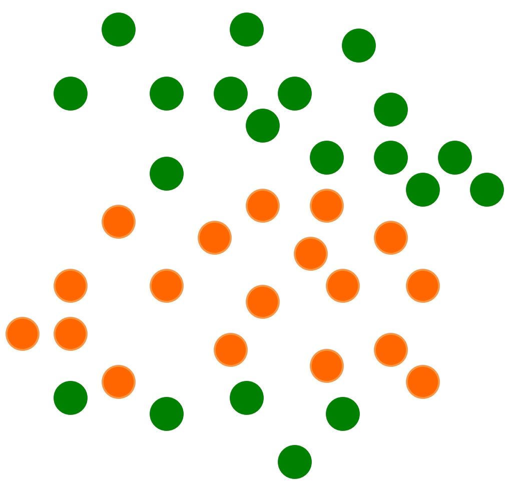


Expectation-Based Scan Statistics

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

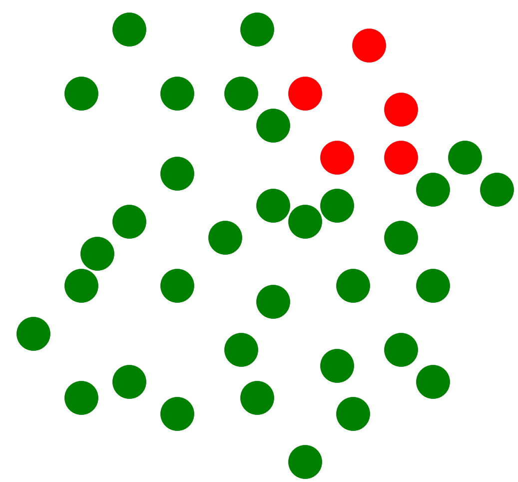
$H_1 : c_i \sim Poisson(qb_i), q > 1$

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



Large subset, moderate risk

VS.



Small pattern, high risk

Adding Element-Specific Penalties

Penalized Fast Subset Scan (Speakman et al., 2015):

For a data set *D,* score function *F(S)* satisfies the Additive Linear Subset Scanning (ALTSS) property if for all $S \subseteq D$,

$$F(S) = \max_{q>1} F(S|q) \text{ where } F(S|q) = \sum_{s_i \in S} \lambda_i$$

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

Adding Element-Specific Penalties

Penalized Fast Subset Scan (Speakman et al., 2015):

Distribution	$\lambda_i(q)$
Poisson	$x_i(\log q) + \mu_i(1-q)$
Gaussian	$x_i \frac{\mu_i}{\sigma_i^2}(q-1) + \mu_i \frac{\mu_i}{\sigma_i^2}(\frac{1-q^2}{2})$
exponential	$x_i \frac{1}{\mu_i}(1-\frac{1}{q}) + \mu_i \frac{1}{\mu_i}(-\log q)$
binomial(p_0)	$x_i \log(q \frac{1-p_0}{1-qp_0}) + \log(\frac{1-qp_0}{1-p_0})$

Adding Element-Specific Penalties

Penalized Fast Subset Scan (Speakman et al., 2015):

Element-specific terms can be added to score function while maintaining additive property

$$F_{penalized}(S) = \max_{q>1} \sum_{s_i \in S} (\lambda_i + \Delta_i)$$

Easy to interpret: Δ_i terms are the prior log-odds of data point s_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 Machine

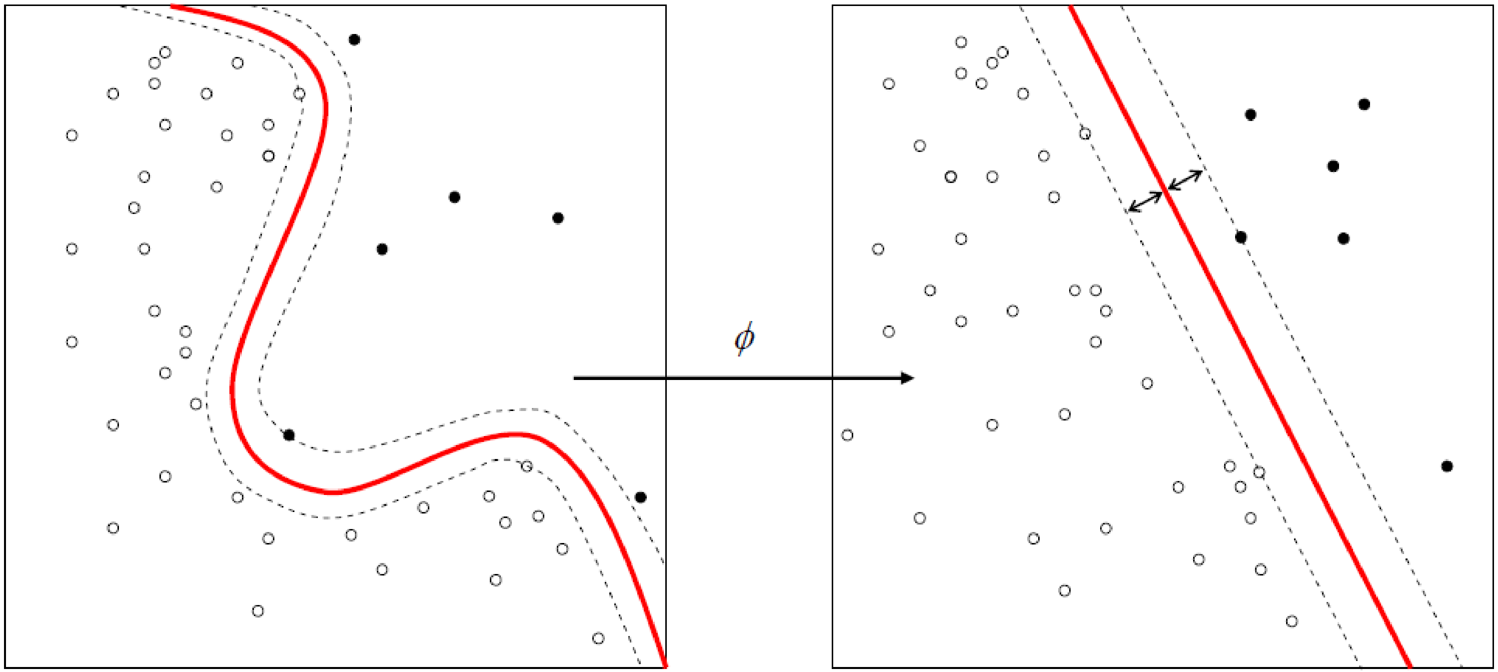


Image Source: Wikipedia

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

Support Vector Machine

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

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

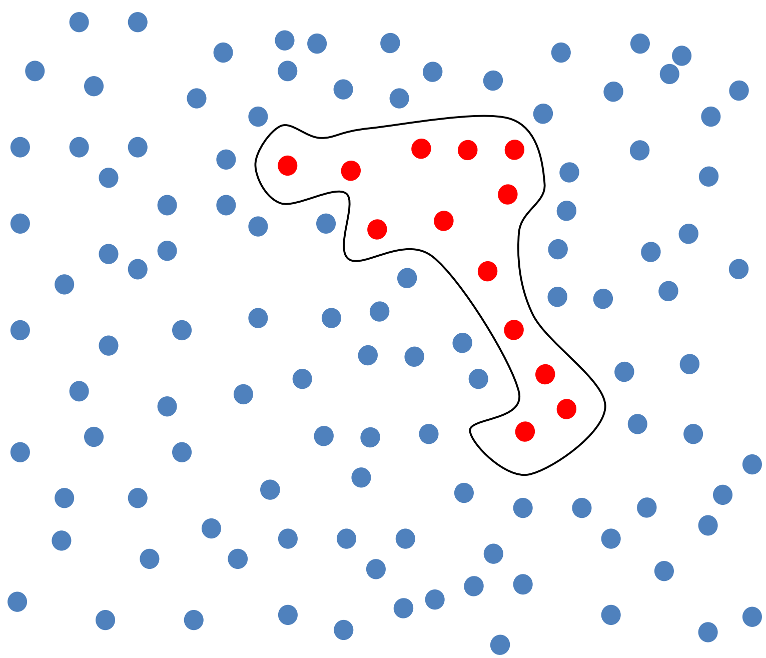
$$y_i(\mathbf{w}\cdot\phi(\mathbf{x}_i) - b) \geq 1 - \xi_i, \forall i = 1, ..., N$$

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

Support Vector Subset Scan (SVSS)

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



Result: Irregular but spatially coherent regions

SVSS Objective Function

Let $\mathbf{x}_i$ be the spatial coordinates of point s_i, let $\alpha_i \in \{0,1\}$ indicate presence/absence of point *i* in *S*, and let $y_i = 2\alpha_i - 1$

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

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

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

$$(2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - b) \geq 1 - \xi_i, \forall i = 1, ..., N$$

SVSS Objective Function

Equivalently,

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

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

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

SVSS Objective Function

Equivalently,

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

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

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

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

SVSS Objective Function

Equivalently,

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

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

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

PFSS Problem

Element-specific penalties = Distance to SVM hyperplane

SVSS Objective Function

Equivalently,

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

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

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

SVM Problem

Binary data labels = Included/Not included in subset

SVSS Algorithm

Algorithm 1 Support Vector Subset Scan

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

$\xi_i(\alpha_i) \leftarrow 0, \forall i = 1, ..., N$

while The optimal subset is changing **do**

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

$\min_{\xi, \mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i)$ ▷ Fix α, and optimize over $\mathbf{w}, b$

end while

return $\boldsymbol{\alpha}$

end procedure

SVSS Algorithm

Algorithm 1 Support Vector Subset Scan

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

  $\xi_i(\alpha_i) \leftarrow 0, \forall i = 1, ..., N$

  while The optimal subset is changing **do**

PFSS

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

    $\min_{\xi, \mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i)$ ▷ Fix α, and optimize over $\mathbf{w}, b$

SVM

  end while

  return α

end procedure

SVSS Algorithm

Algorithm 2 Support Vector Subset Scan (random restarts)

procedure SVSS($\mathbf{c}, \mathbf{b}, \mathbf{x}, T_{max}, C_0, C_1$) ▷ Counts $\mathbf{c}$, expectations $\mathbf{b}$, and coordinates $\mathbf{x}$
    $min\_score \leftarrow \infty$
    for $t := 1$ **to** T_{max} **do** ▷ T_{max} random restarts
        $\xi_i(\alpha_i) \leftarrow \text{Uniform}(-C_0, C_0), \forall i = 1, \ldots, N$
        while The optimal subset is changing **do**

            $\max_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}) - C_0/C_1 \sum_{i=1}^{N} \xi_i(\alpha_i)$ ▷ Fix $\mathbf{w}, b$ and optimize over $\boldsymbol{\alpha}$
            $\min_{\boldsymbol{\xi}, \mathbf{w}, b} \frac{1}{2}\|\mathbf{w}\|^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i)$ ▷ Fix α, and optimize over $\mathbf{w}, b$

        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

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

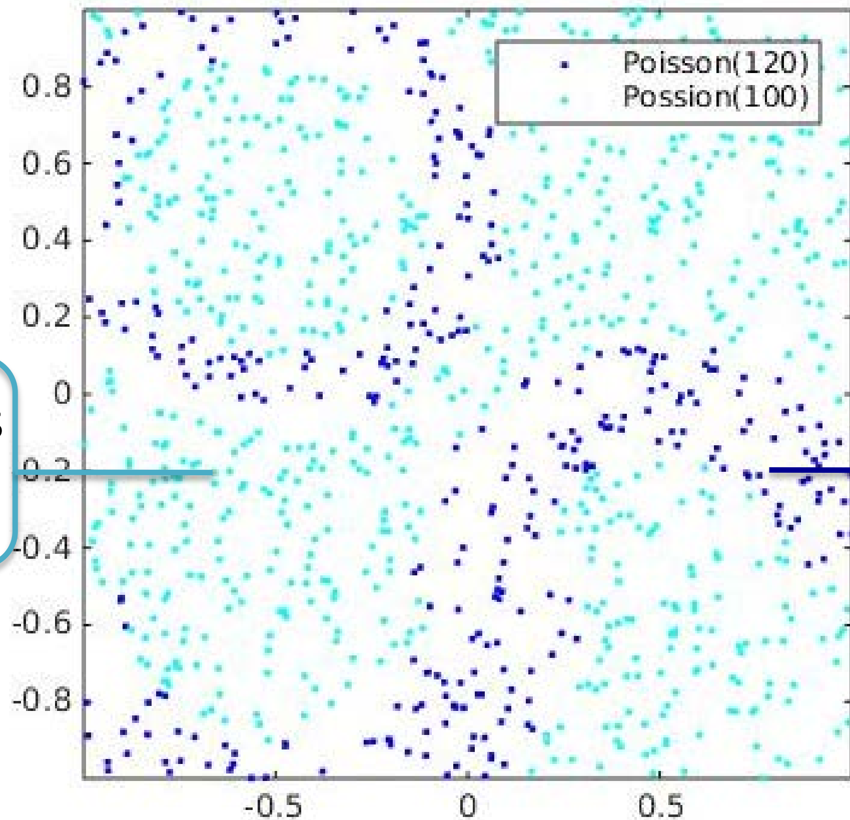
EQUIVALENT:

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

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

Improvement Over Iterations

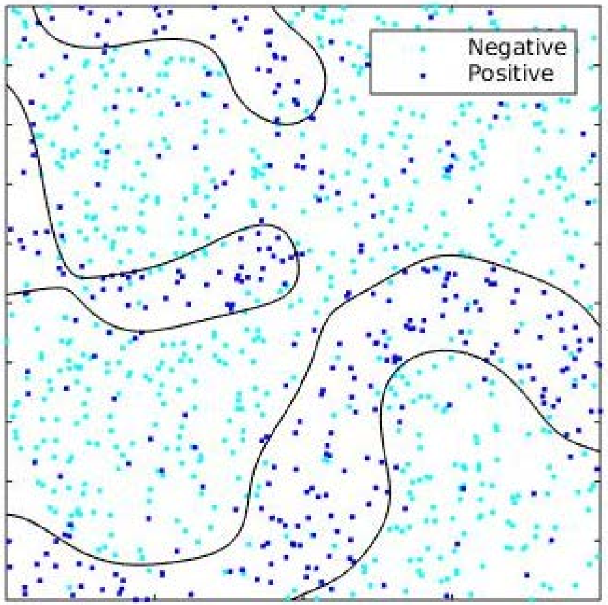
Unaffected points
~ Poisson(100)



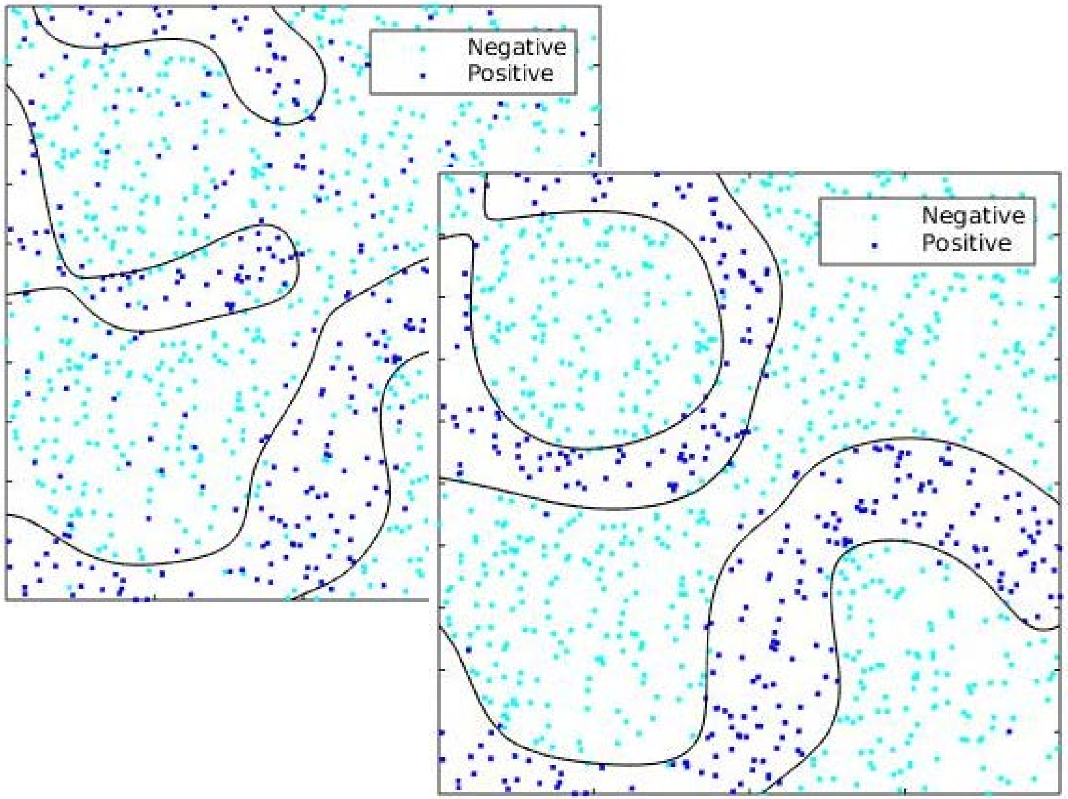
Affected points
~ Poisson(120)

Expectation = 100 for all sites

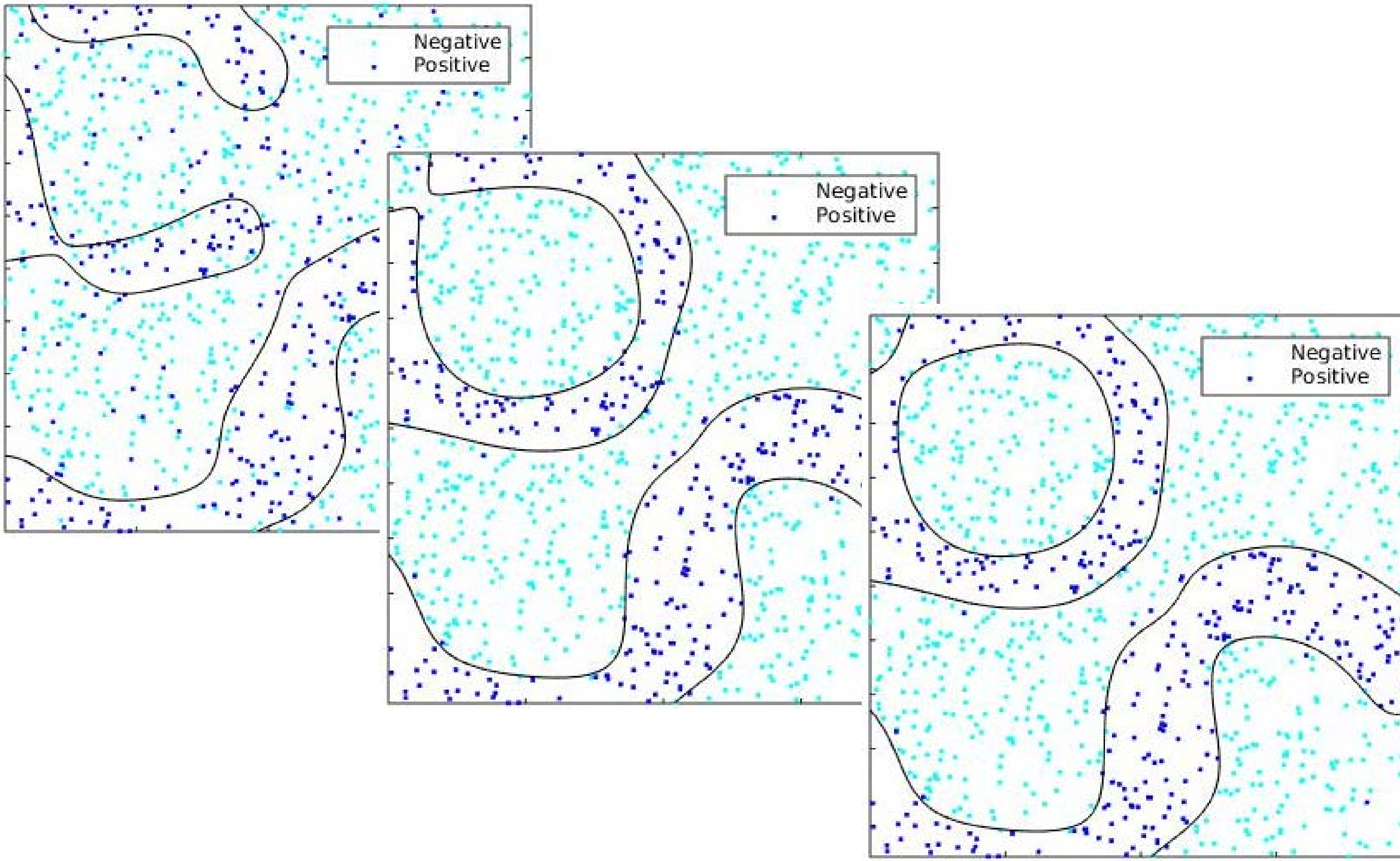
Improvement Over Iterations



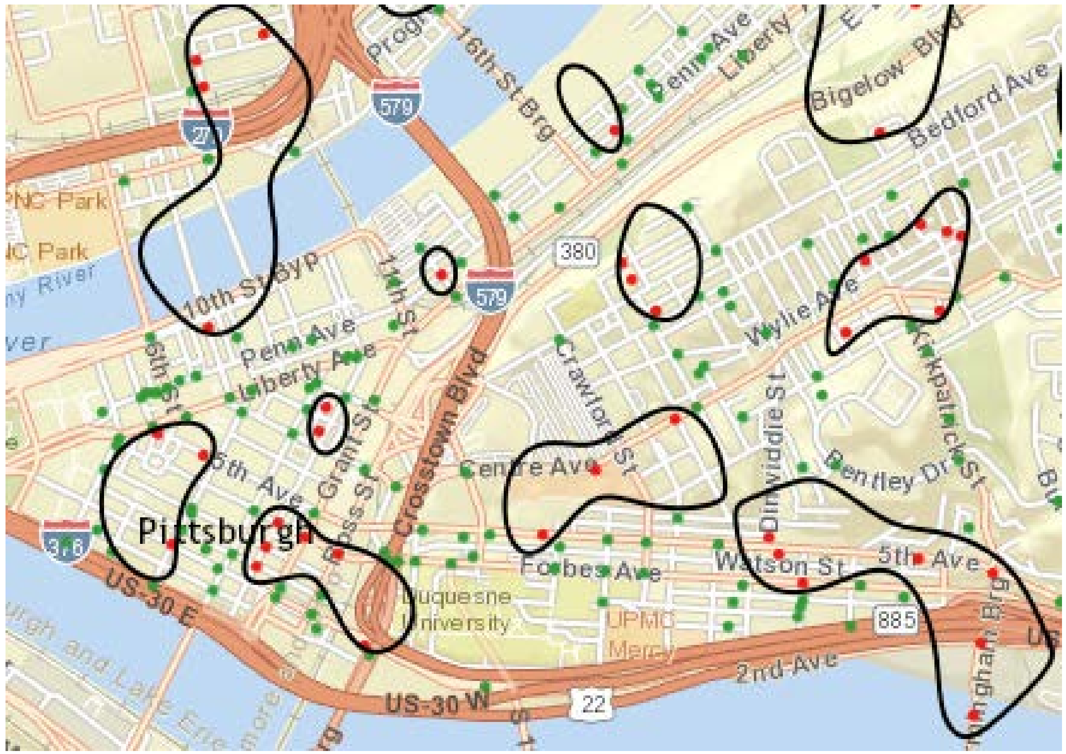
Improvement Over Iterations



Improvement Over Iterations



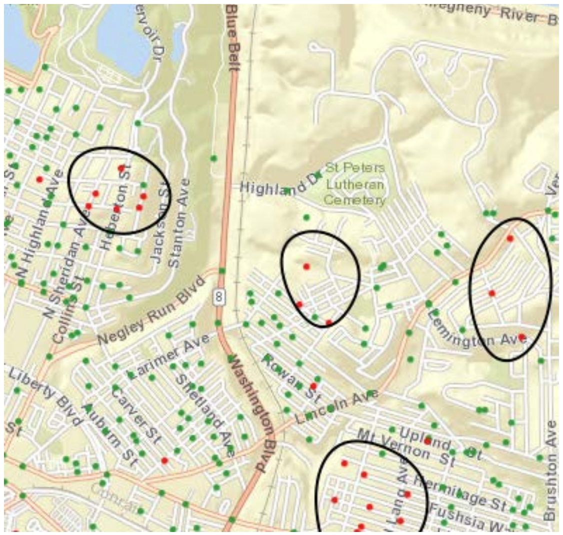
Ranking Disconnected Regions



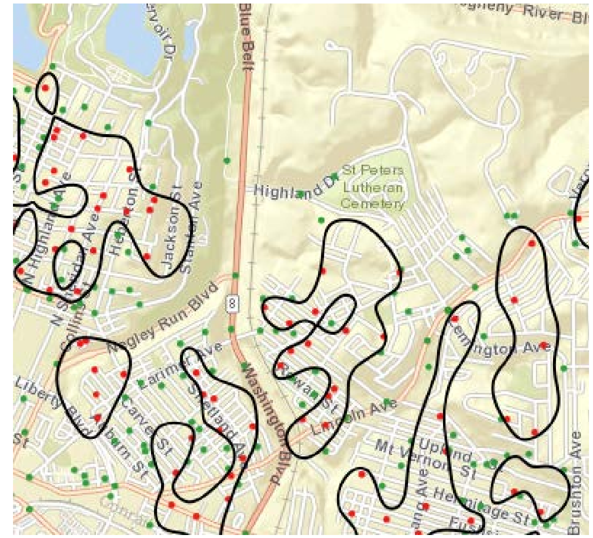
How can we rank the connected regions of the best subset?

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

Tuning model parameters

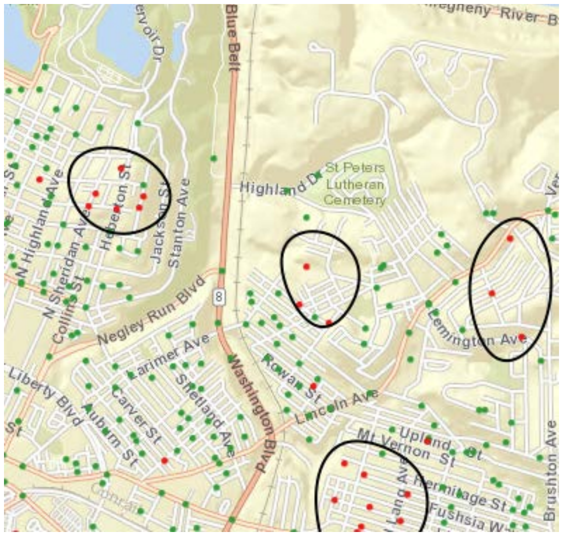


VS.

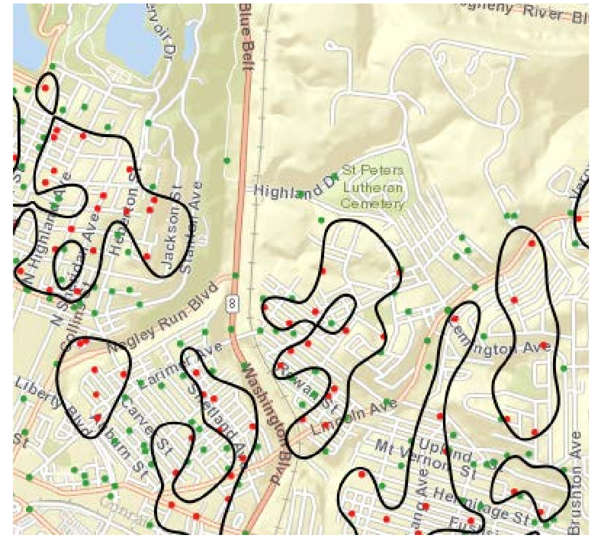


Goal: Find parameter combination that generates best subset with high log-likelihood ratio (LLR) and some minimum level of geometric compactness

Tuning model parameters



VS.



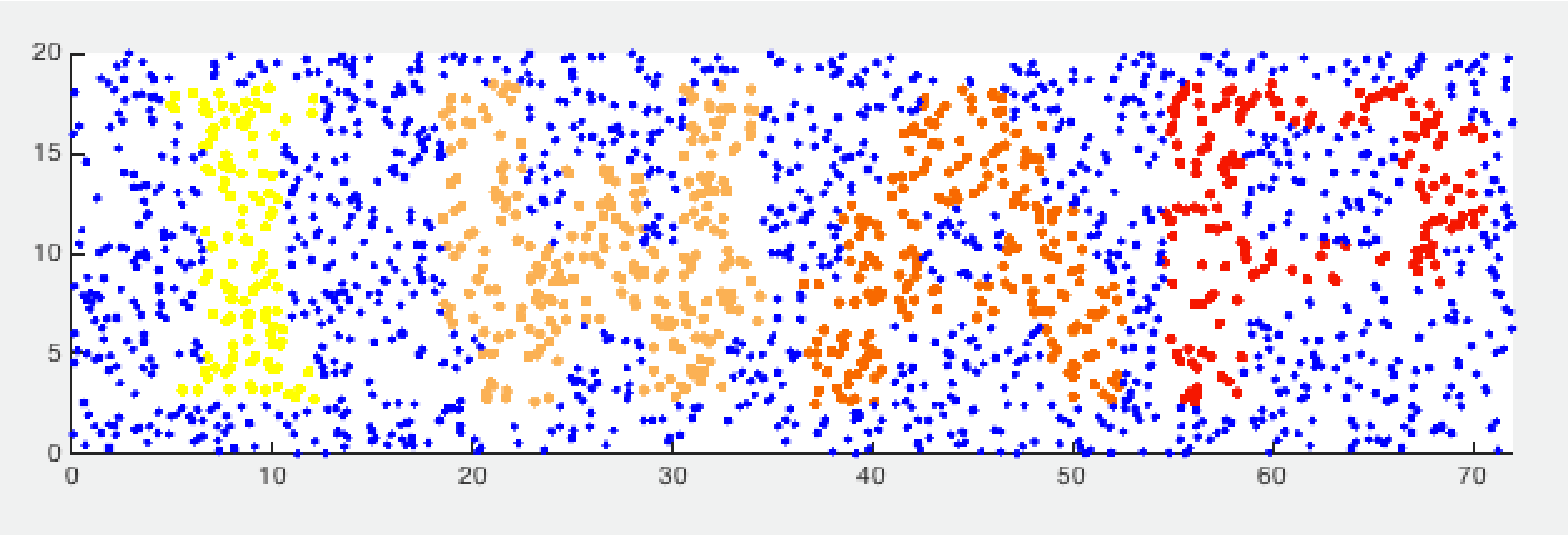
Tuning procedure:

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

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

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

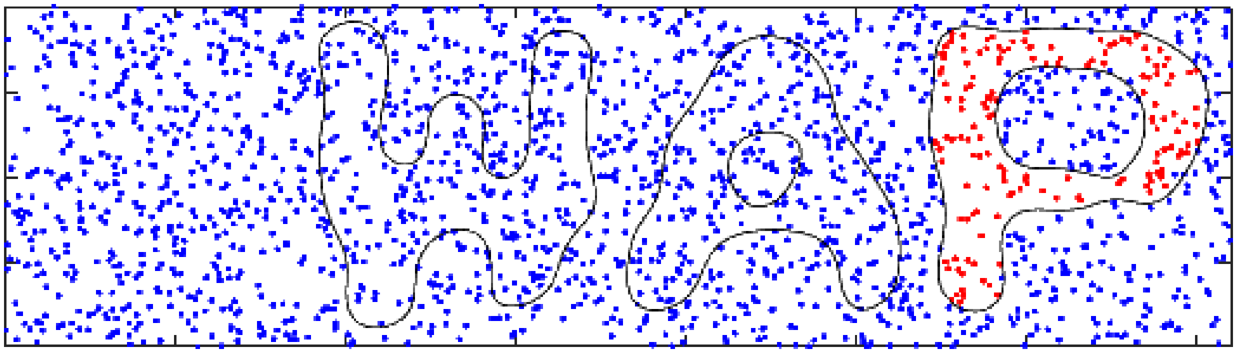
Detecting Letter-Shaped Regions



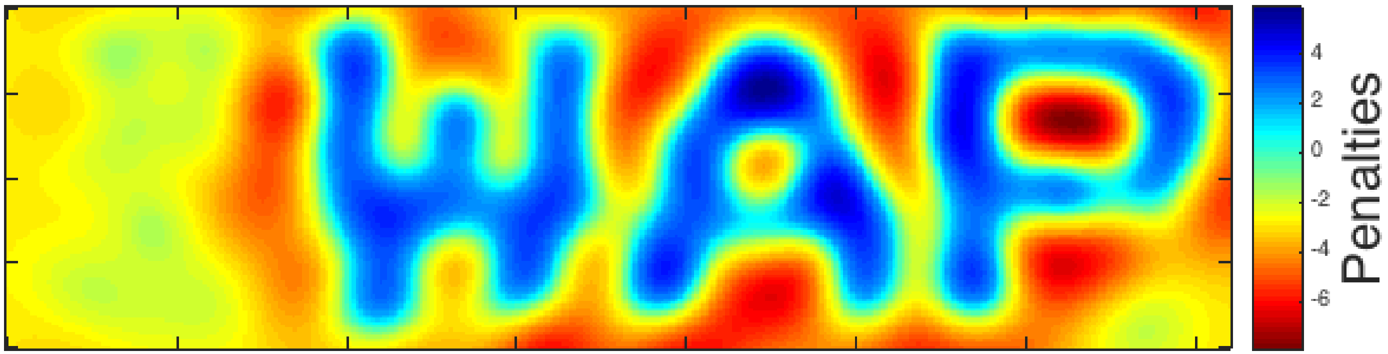
- $c_i \sim Poisson(100)$ (blue)
- $c_i \sim Poisson(120)$ (yellow)
- $c_i \sim Poisson(140)$ (light orange)
- $c_i \sim Poisson(160)$ (orange)
- $c_i \sim Poisson(180)$ (red)

All points: $b_i = 100$

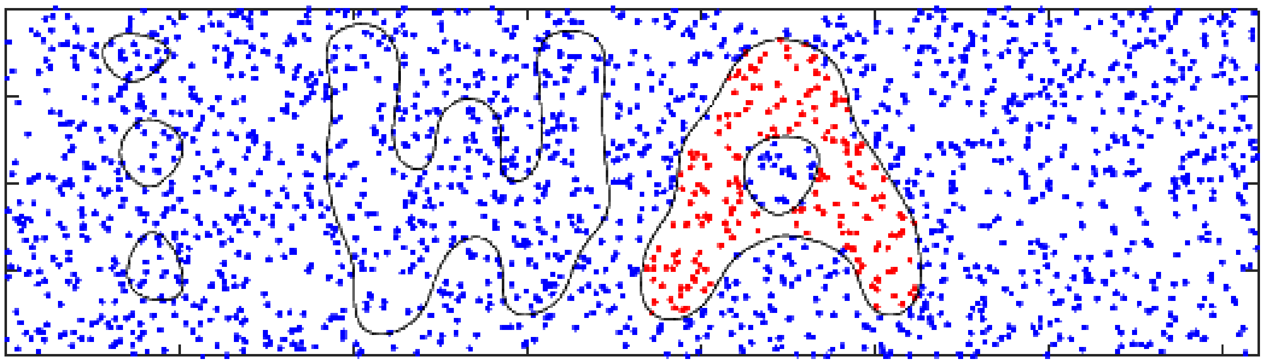
Detecting Letter-Shaped Regions



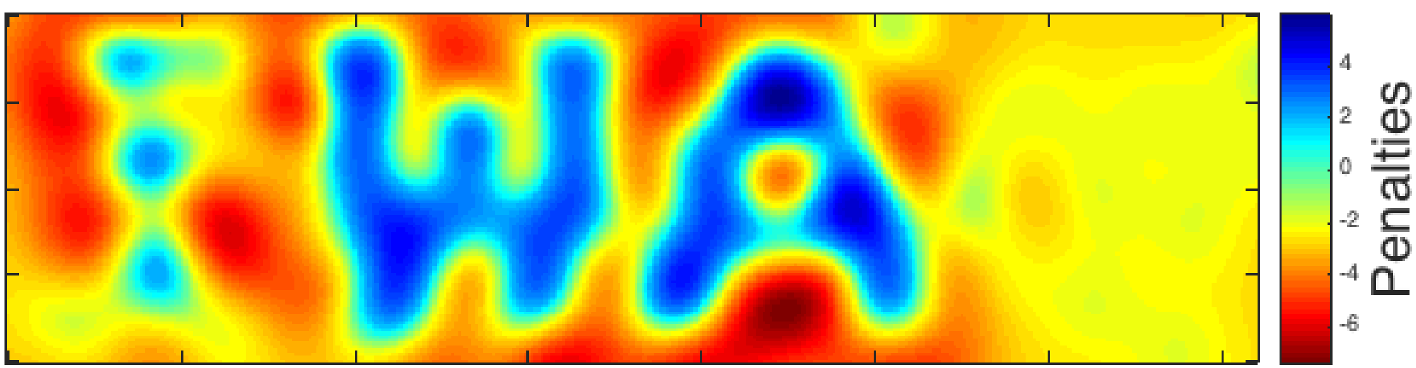
Best connected SVM region



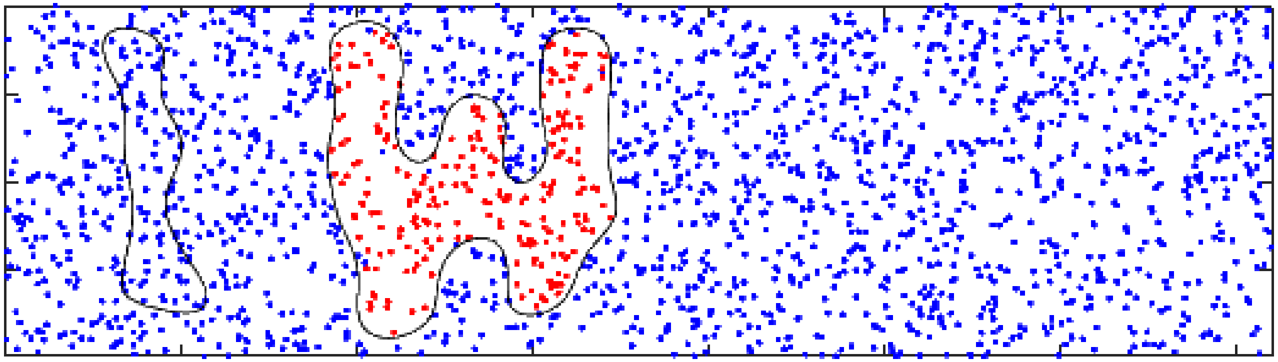
Detecting Letter-Shaped Regions



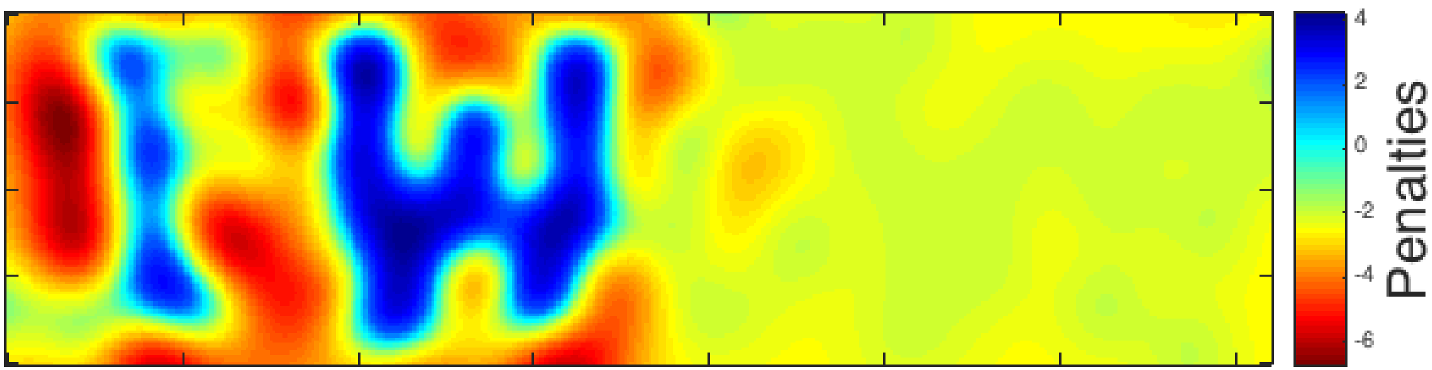
2nd Best connected SVM region



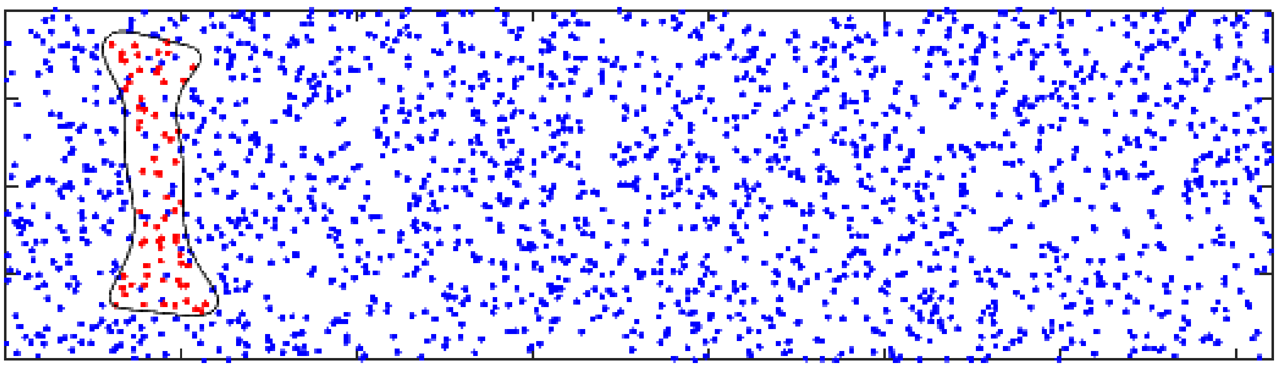
Detecting Letter-Shaped Regions



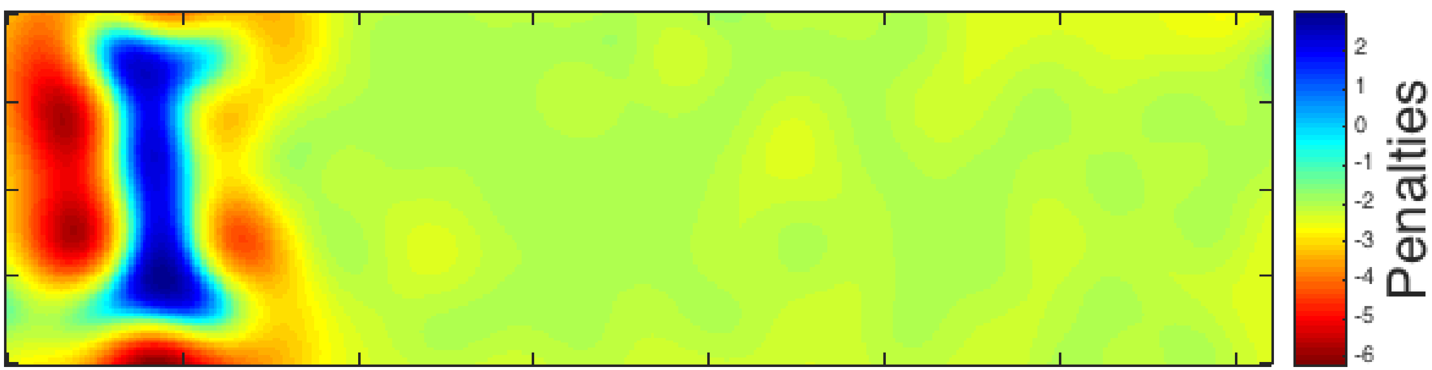
3rd Best connected SVM region



Detecting Letter-Shaped Regions

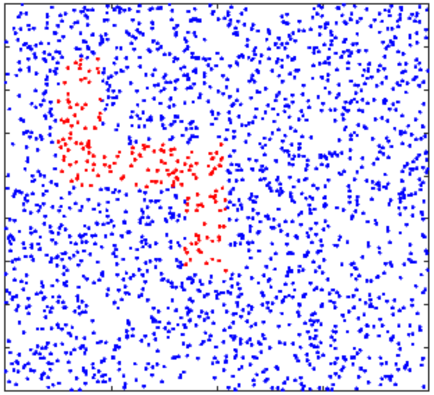


4th Best connected SVM region

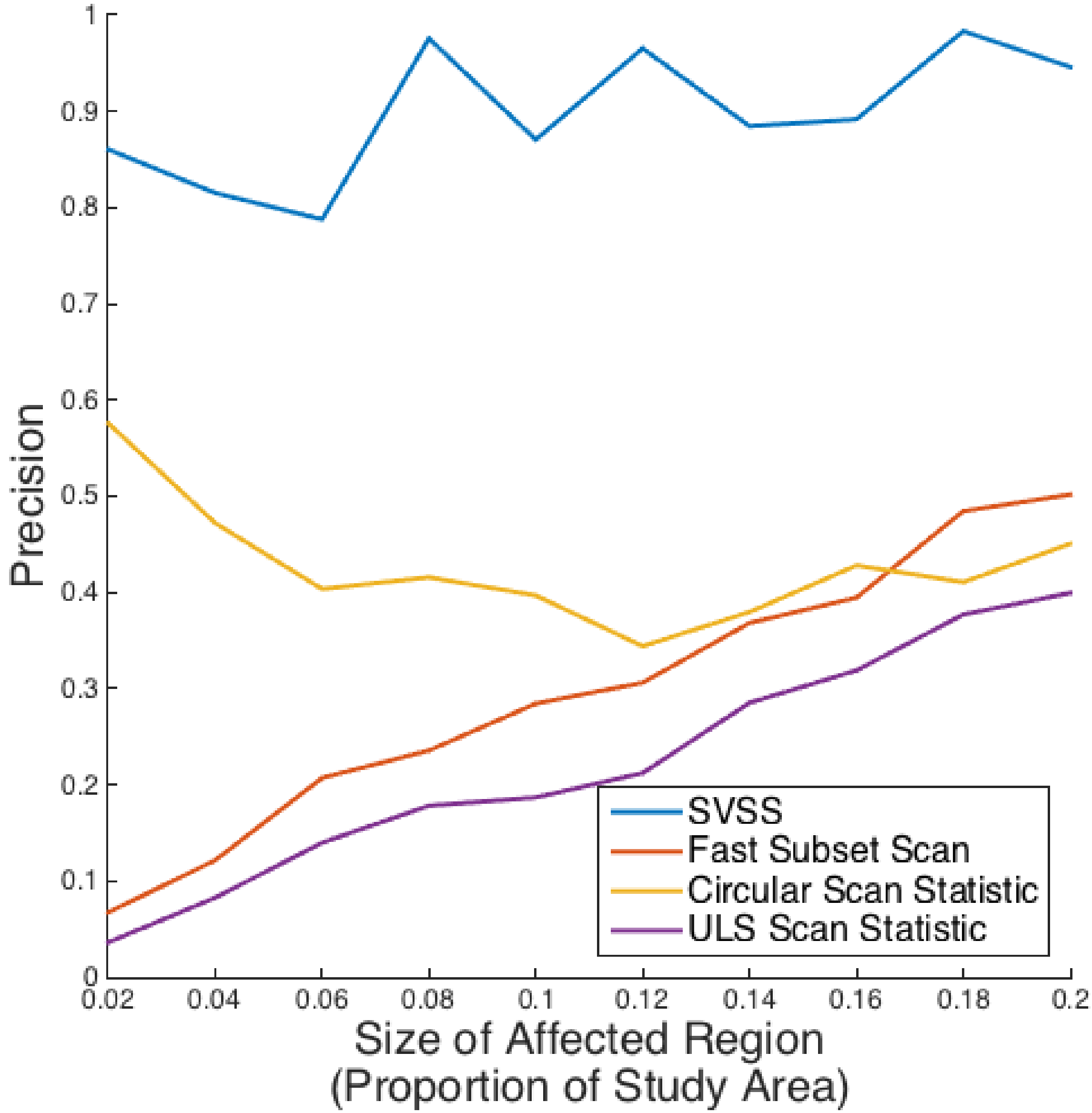


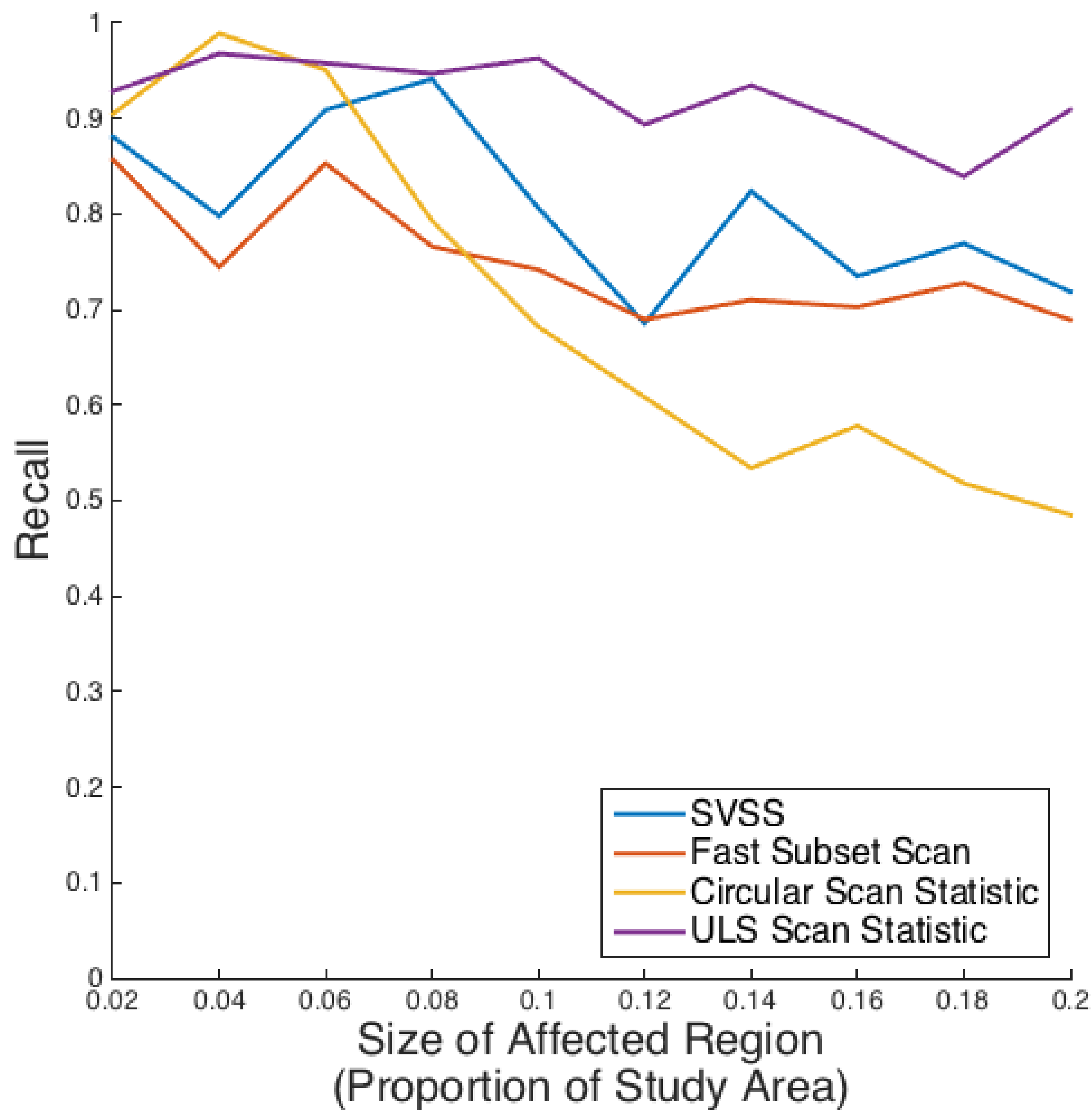
Evaluation Framework

- 2000 observations generated from Poisson distribution
- Generated random, irregular-shaped regions of varying length with elevated counts
  - Unaffected points: $c_i \sim \mathit{Poisson}(100)$
  - Affected points: $c_i \sim \mathit{Poisson}(115)$
  - $b_i = 100$ for all points



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


Precision
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
Size of Affected Region
(Proportion of Study Area)
SVSS
Fast Subset Scan
Circular Scan Statistic
ULS Scan Statistic



Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
Size of Affected Region
(Proportion of Study Area)
SVSS
Fast Subset Scan
Circular Scan Statistic
ULS Scan Statistic

Detecting Pothole Hotspots

Data:

- Pothole reports at city block level from City of Pittsburgh 311 system

Timeframe:

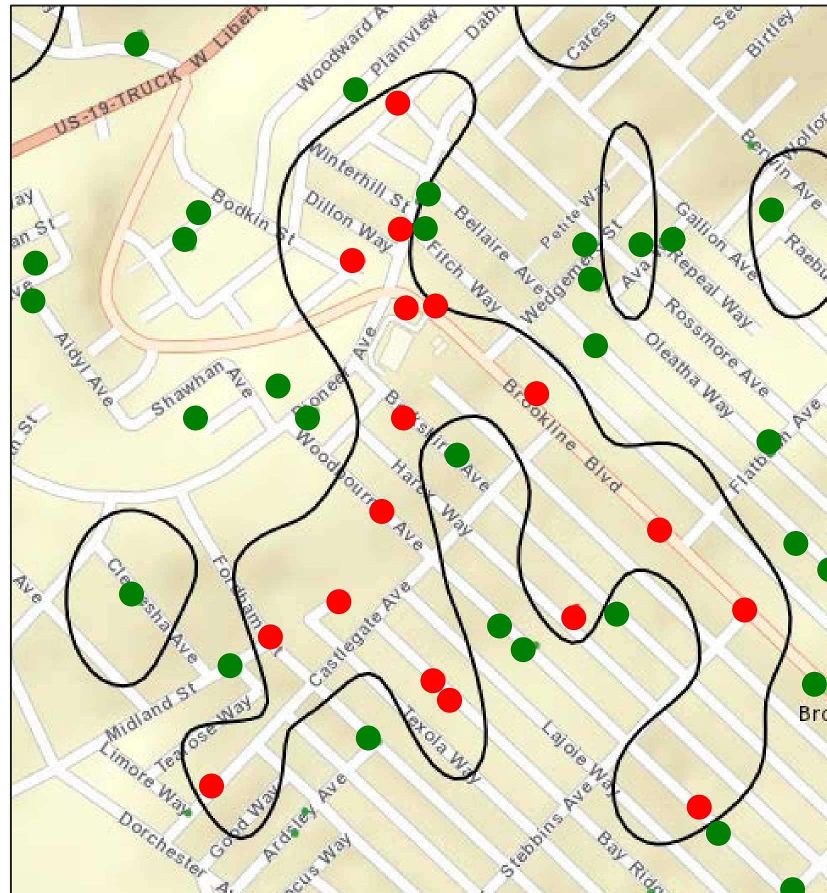
- Expected counts estimated from 2008-2011 control period
- Actual counts generated from 2012-2013

Can we identify roads or neighborhoods in need of maintenance?

Top 5 Pothole Hotspots

Rank	# of Points	Relative Risk (MLE)
1*	17	3.2
2	15	3.0
3	17	2.8
4	12	3.9
5	15	2.3

*Pattern shown to right



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 showed high precision and recall on the task of detecting irregularly shaped patterns relative to competing methods.

We demonstrated the real-world utility of SVSS by applying it to pothole hotspot detection in Pittsburgh roadways.

Thank you

djfitzpa@cmu.edu
neill@cs.cmu.edu