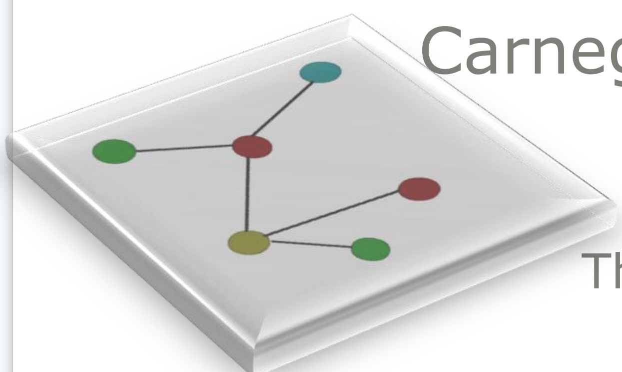
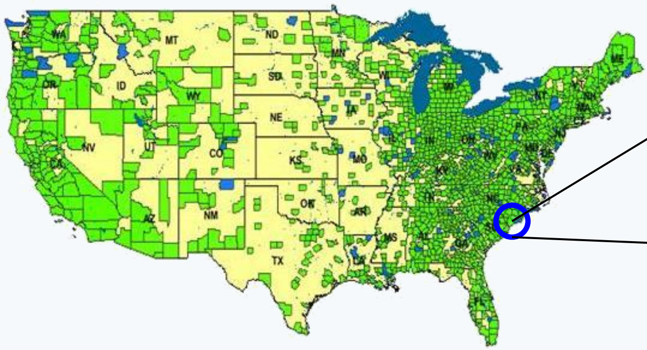
Fast Graph Scan for Scalable Detection of Arbitrary Connected Clusters

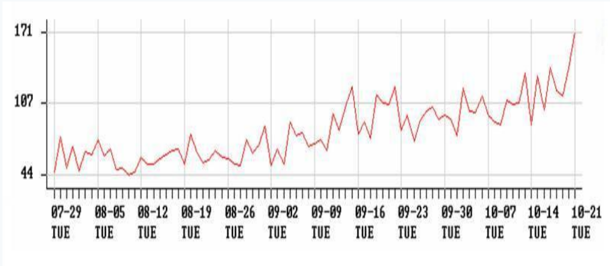
Skyler Speakman & Daniel B. Neill
Carnegie Mellon University, Heinz College
ISDS Annual Conference



This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0325581



Daily health data from thousands of hospitals and pharmacies nationwide



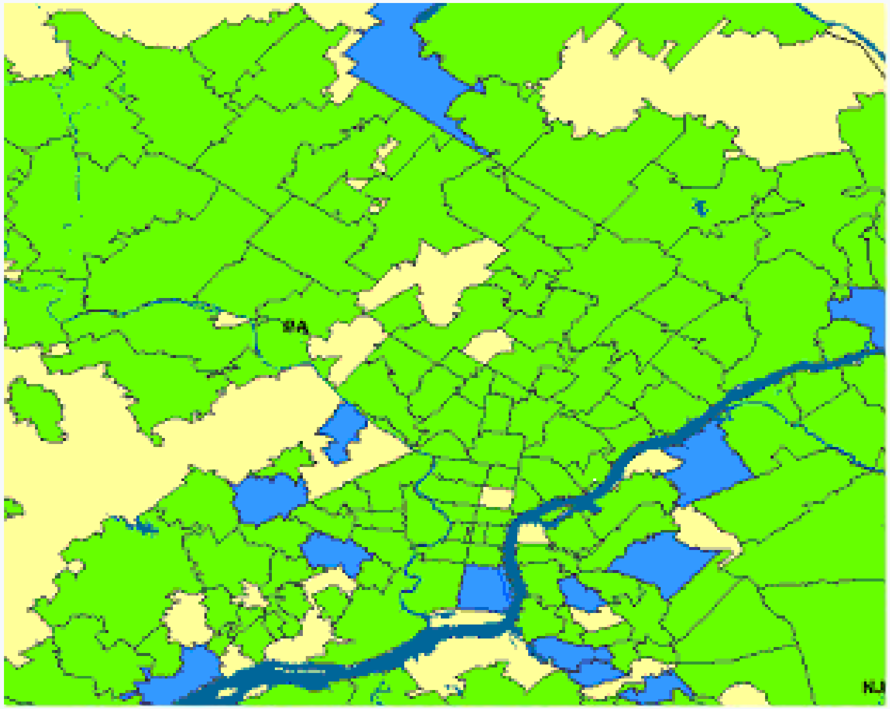
Time series of counts c_i^t for each zip code s_i

Use this data to detect **anomalous patterns**

Detect any emerging events (i.e. outbreaks of disease)
Pinpoint the affected areas

Biosurveillance

(Kulldorff, 1997; Neill and Moore, 2005)



Scan over multiple regions to detect where counts are higher than expected.

Aggregate the individual counts from each location within a region.

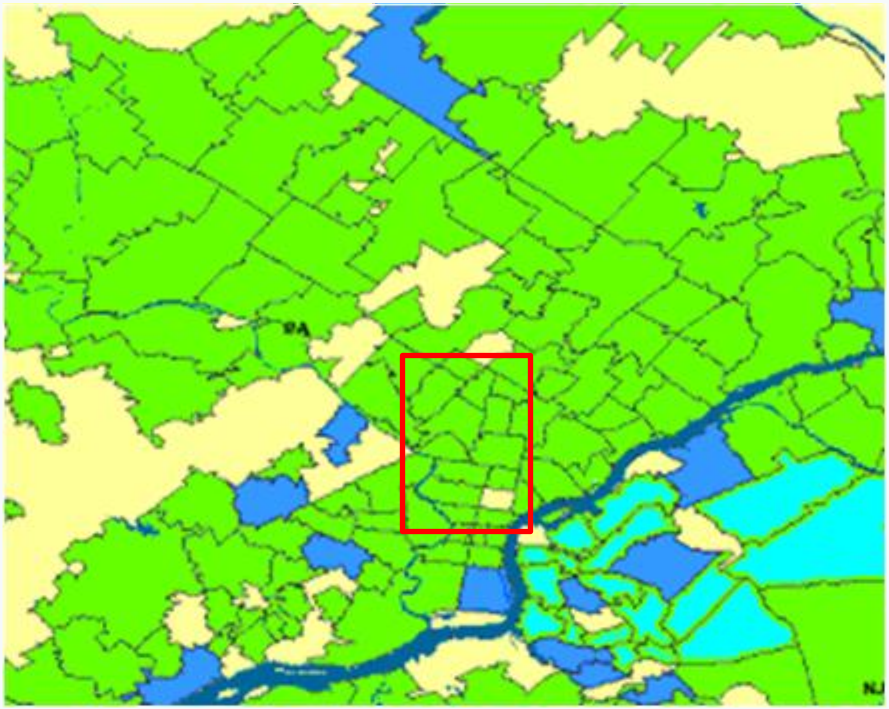
Circles

Choose a center location s_c and its k nearest neighbors.

Find the circle that maximizes a given score function of the aggregated counts and baselines.

Expectation-based Scan Statistics

(Kulldorff, 1997; Neill and Moore, 2005)



Scan over multiple regions to detect where counts are higher than expected.

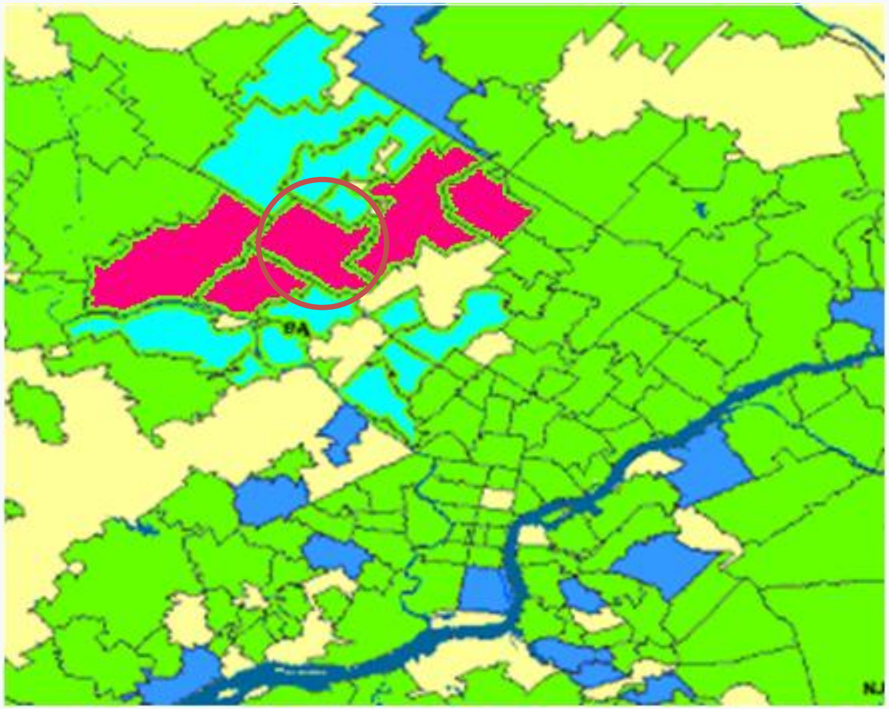
Aggregate the individual counts from each location within a region.

Rectangles

Find the rectangle that maximizes a given score function of the aggregated counts and baselines.

Expectation-based Scan Statistics

(Kulldorff, 1997; Neill and Moore, 2005)



Power to Detect

Circles are useful for detecting tightly clustered outbreaks

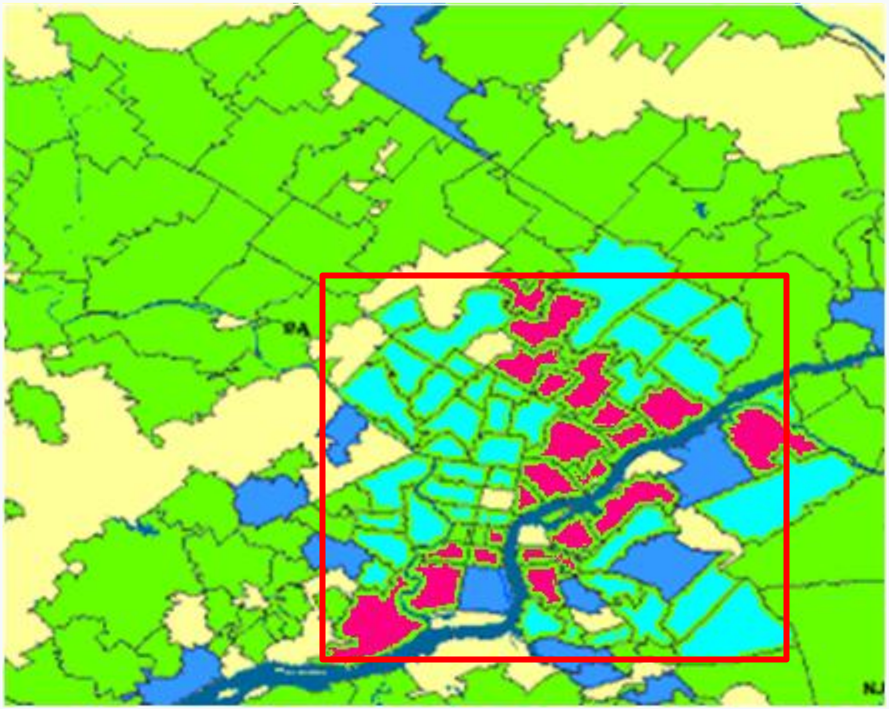
However, they lose power to detect abnormally shaped clusters

Affected locations

Un-affected locations contributing to region score

Expectation-based Scan Statistics

(Kulldorff, 1997; Neill and Moore, 2005)



Power to Detect

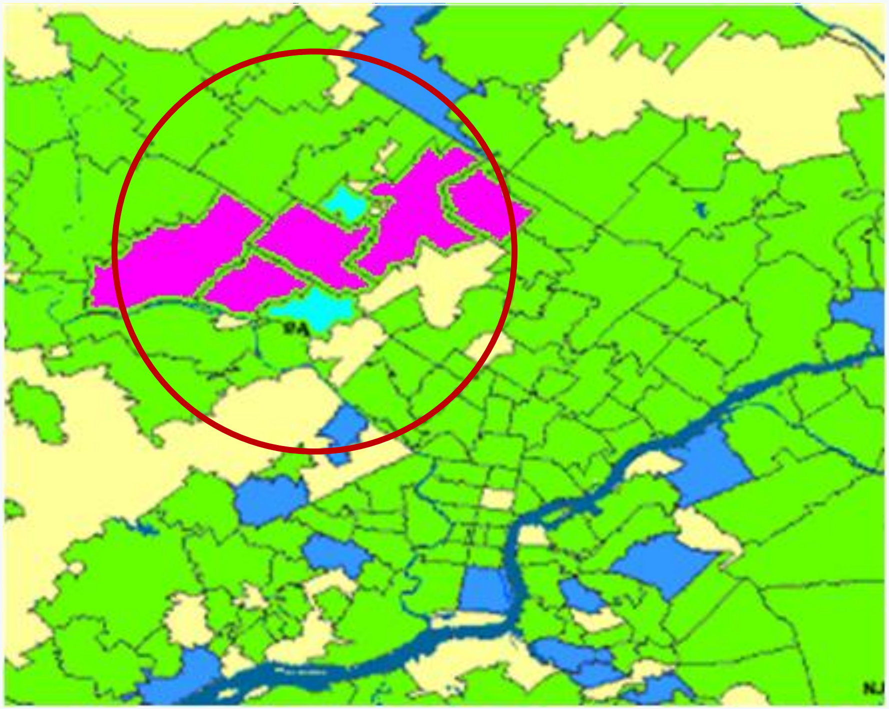
There are similar issues with rectangles for some outbreaks

Affected locations

Un-affected locations contributing to region score

Expectation-based Scan Statistics

(Neill, 2008)



An alternative to scanning over shapes of regions is to find the ***subset of locations*** for a given region that has the highest score

Affected locations

Un-affected locations contributing to region score

Pattern Detection through Subset Scanning

(Neill, 2008)

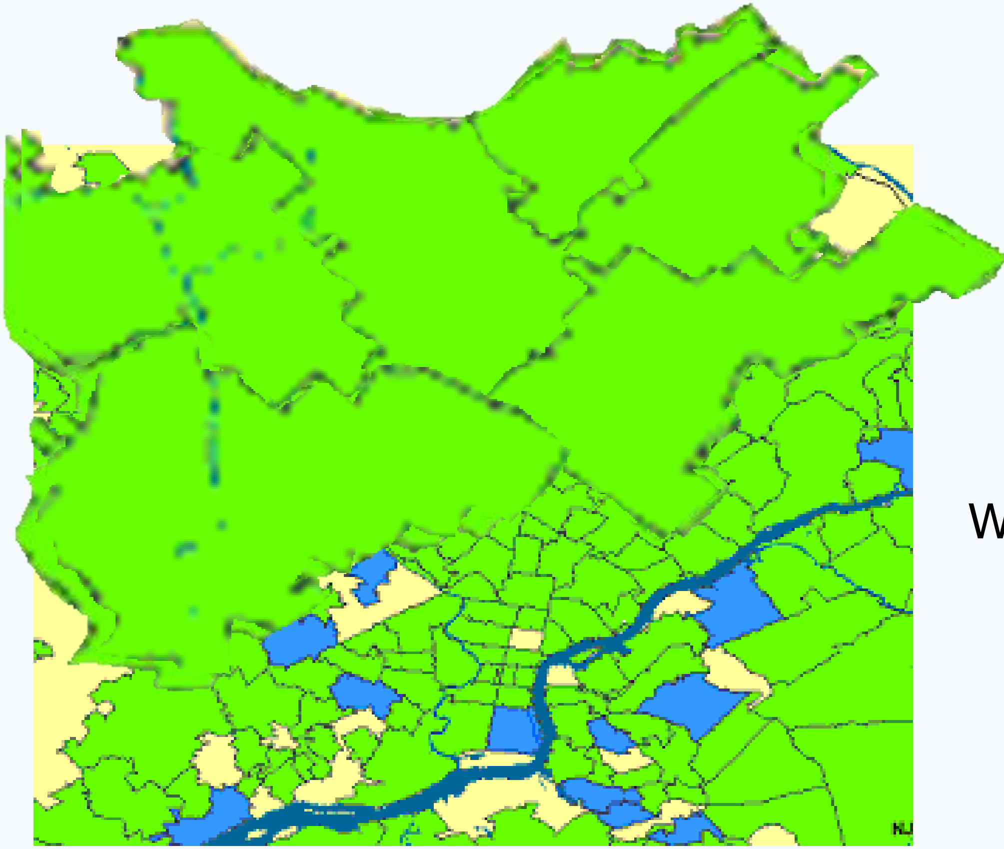
PROBLEM:	The number of subsets grows exponentially with the size of the region (2^n)

This makes it computationally infeasible for regions with more than ~30 locations

SOLUTION:	Exploit a property of scoring functions to rule out subsets that cannot obtain the highest score

This reduction in the search space allows for exact and efficient calculation of the highest scoring subset

Subset Scanning



(Neill, 2008)

We wish to maximize a scoring function

$$F(S) = F\left(\sum_{s_i \in S} c_i, \sum_{s_i \in S} b_i\right)$$

over all possible subsets, S

We sort the locations according to a relevance criteria

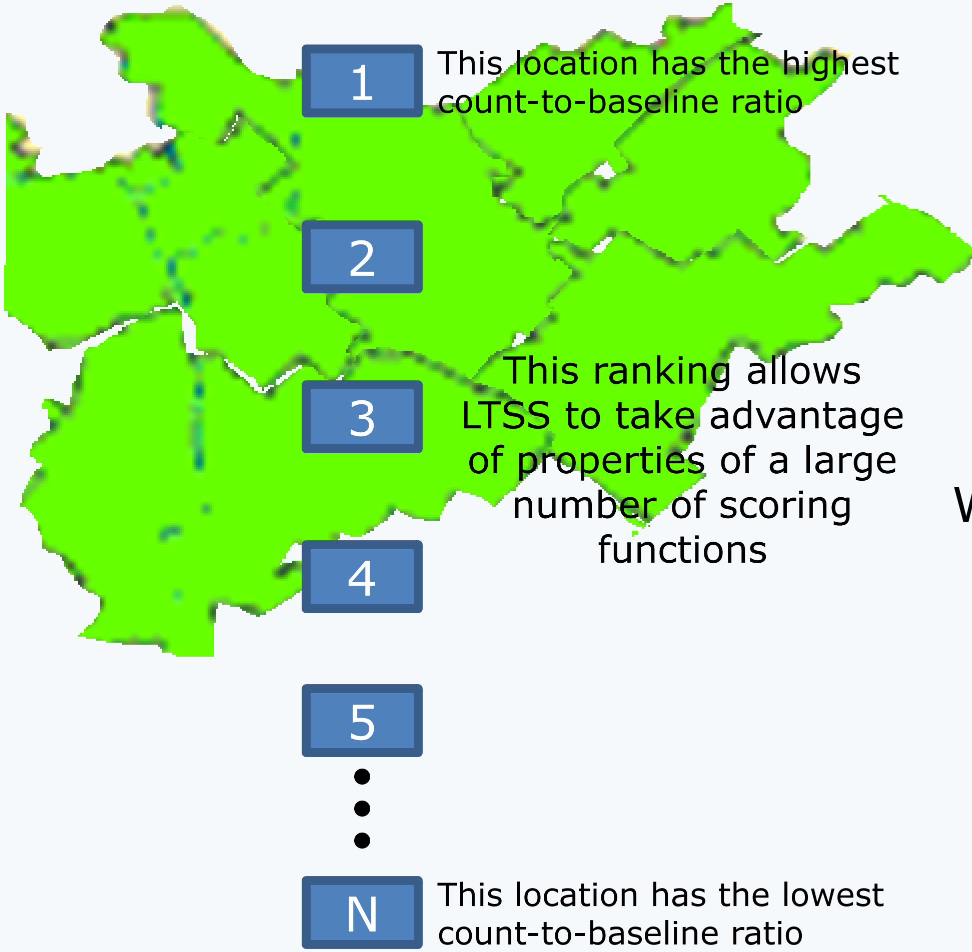
For example,

$$G(s_i) = \frac{c_i}{b_i}$$

works for Kulldorff's Statistic and Expectation-based Poisson

Linear Time Subset Scanning

Linear Time Subset Scanning



(Neill, 2008)

We wish to maximize a scoring function

$$F(S) = F\left(\sum_{s_i \in S} c_i, \sum_{s_i \in S} b_i\right)$$

over all possible subsets, S

We sort the locations according to a relevance criteria

For example,

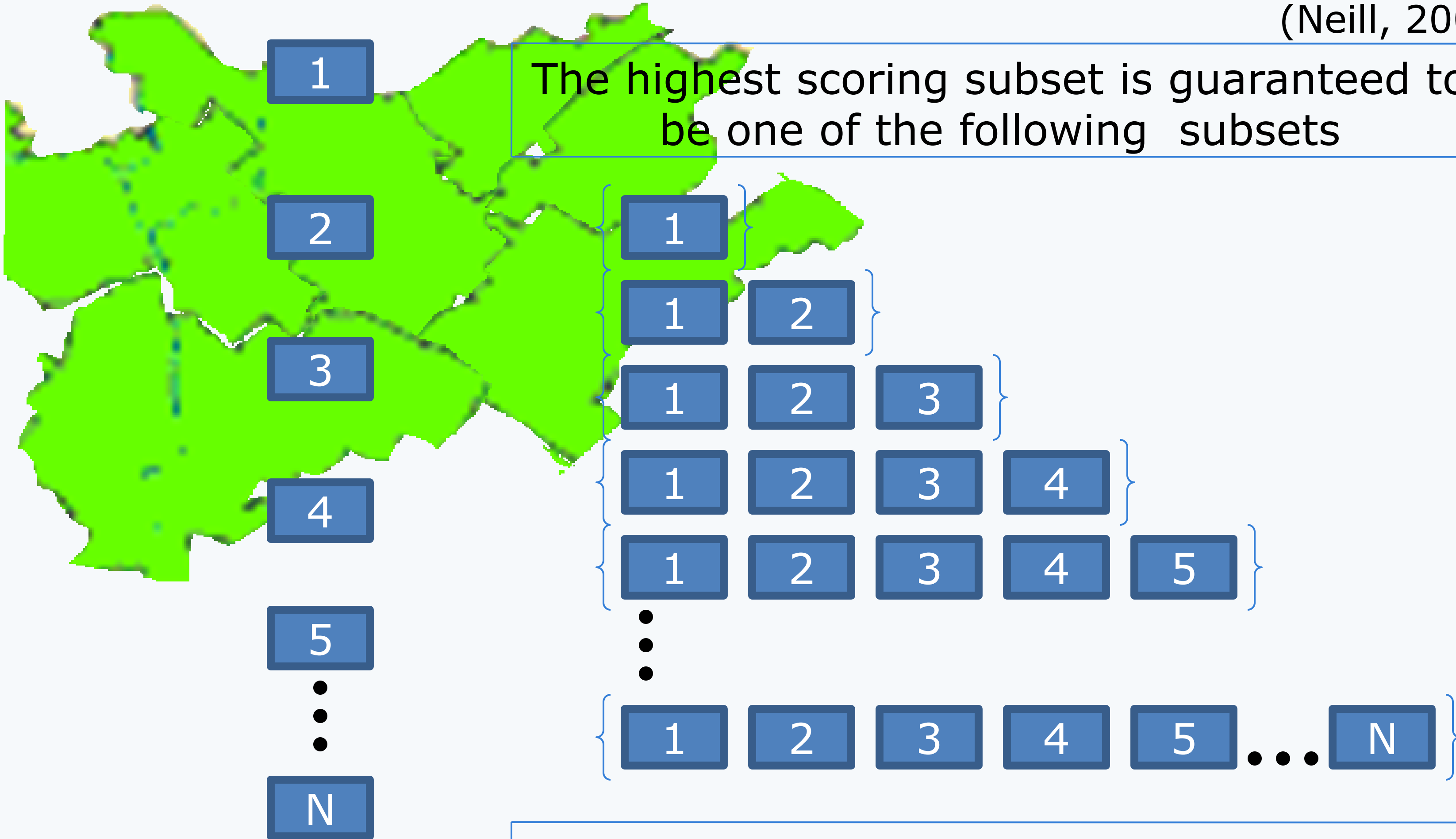
$$G(s_i) = \frac{c_i}{b_i}$$

works for Kulldorff's Statistic and Expectation-based Poisson

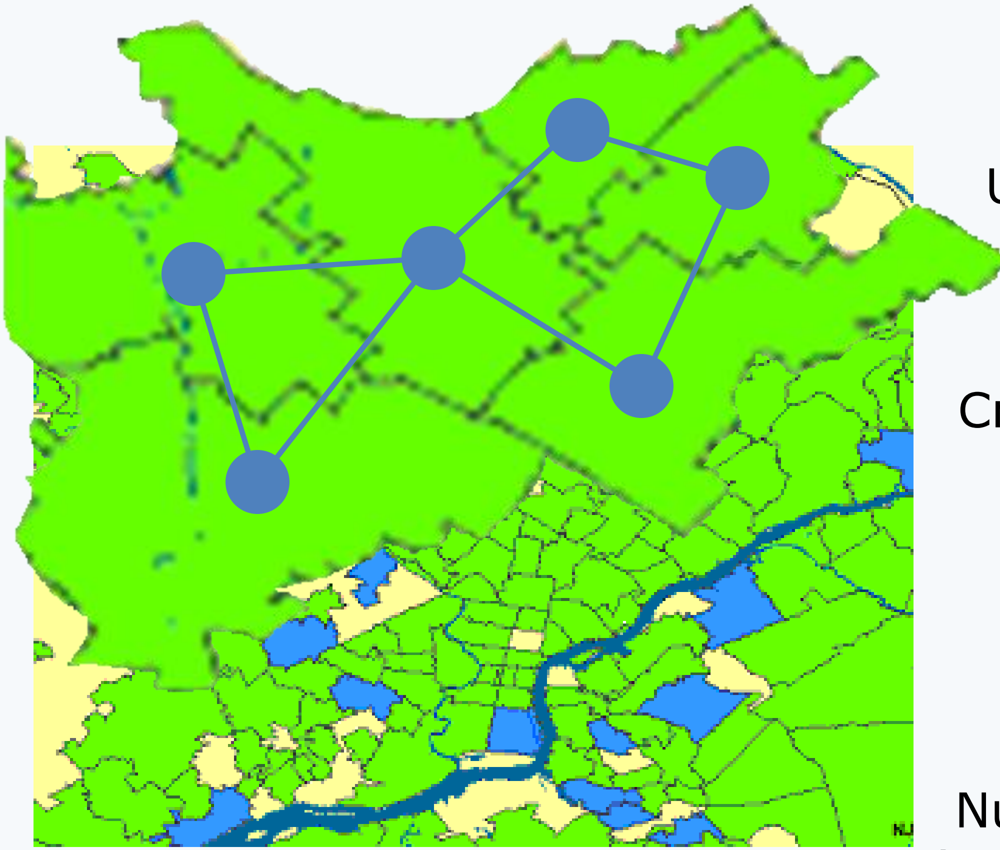
Linear Time Subset Scanning

(Neill, 2008)

The highest scoring subset is guaranteed to be one of the following subsets



Decreases the search space from 2^N to N



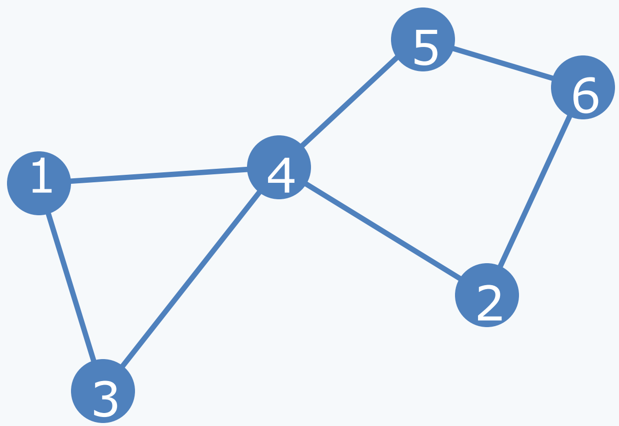
Use adjacency of locations to form a *flexible* scan statistic (Tango & Takahashi, 2005)

Create an adjacency graph of the locations and score ***every connected subset***

Increase power to detect non-circular clusters

Number of connected subsets is exponential in size of region. Infeasible for regions of >30 locations

Connectivity Constraints



Use property of LTSS to reduce the search space and rule out a large number of connected subsets

Rank the locations according to relevance critera

Only scan connected subsets that have *potential* for highest score

Graphscan:
If location s_i is contained in the optimal subset S*
and if removing s_i does not disconnect the subgraph,
any neighbor of s_i with higher relevance will also be contained in S*.

LTSS with Connectivity Constraints

The Graphscan algorithm would end up evaluating the sets:

[1]　　[2]　　[1] [2] [3]

Why not the sets [3] or [1] [3] or [2] [3] ?

Because these sets could include a higher ranked neighbor that would increase the set’s score

Brief Example

The GraphScan method was evaluated using Emergency Department data from 91 Allegheny County zip codes

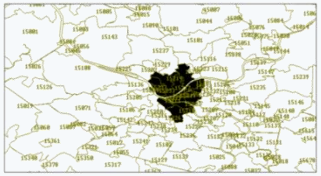
Runtimes

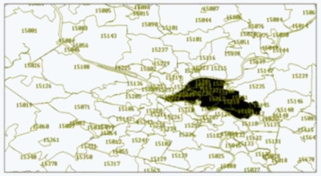
Original Graphscan		Branch & Bound GraphScan
For k=25 **0.24 seconds**	We can use LTSS to quickly determine the unconstrained bound of a given subset	For k=25 **0.08 seconds**
For k=50 **41.0 seconds**		For k=50 **1.1 seconds**
Single Region **87.9 seconds**	If the subset's bound is less than the current high score, we do not have to include it	Single Region **1.0 second**

…for a single day of data

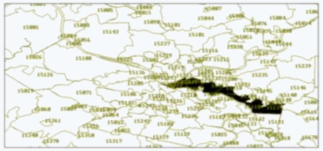
Results

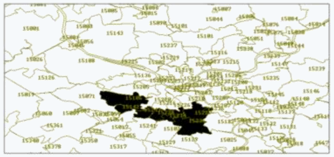
We compared the detection power and accuracy of GraphScan to the original Kulldorff scan statistic (circular regions) on multiple semi-synthetic outbreaks injected into the data

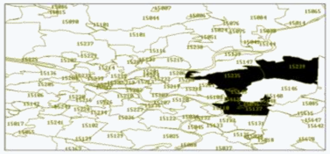




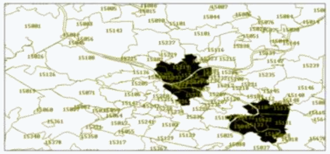
Average over all types of injects	% of Injects Detected	Days to detect
Circles	83.6%	8.6
GraphScan K=25	88.2%	8.2
GraphScan K=50	89.4%	8.1
GraphScan Single Region	88.6%	8.1



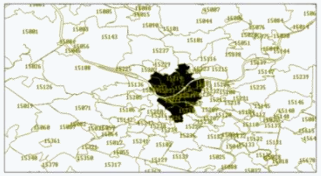


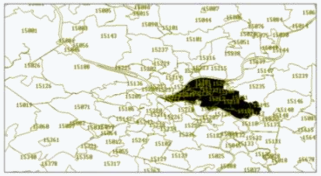


Results: Detection Power

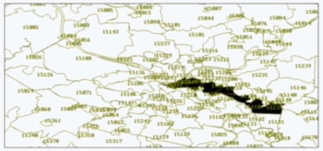


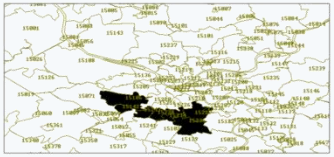
We compared the detection power and accuracy of GraphScan to the original Kulldorff scan statistic (circular regions) on multiple semi-synthetic outbreaks injected into the data

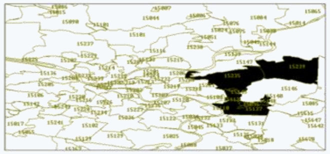




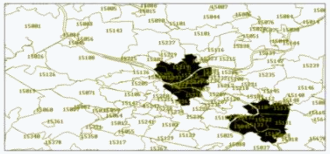
Compact Cluster	% Detected	Days to Detect
Circles	68%	10.4
Graphscan K=25	84%	9.3
Graphscan K=50	88%	8.3
Graphscan Single Region	88%	8.6







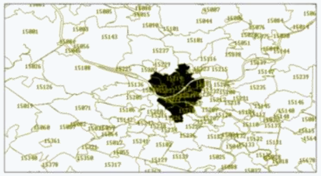
Results: Detection Power

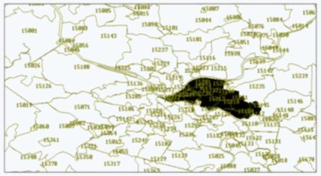


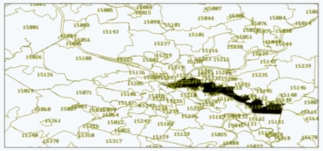
We compared the detection power and accuracy of GraphScan to the original Kulldorff scan statistic (circular regions) on multiple semi-synthetic outbreaks injected into the data

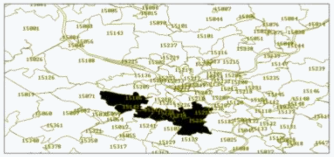
Elongated Cluster	% Detected	Days to Detect
Circles	66%	10.4
Graphscan K=25	87%	8.5
Graphscan K=50	92%	8.0
Graphscan Single Region	92%	8.2

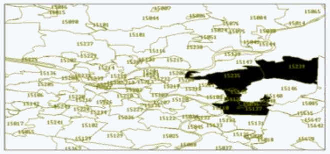
Results: Detection Power

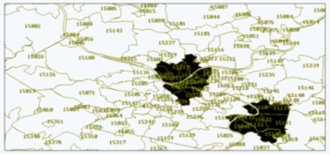








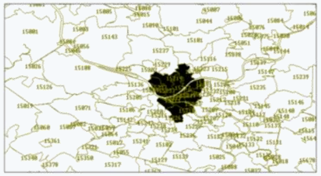
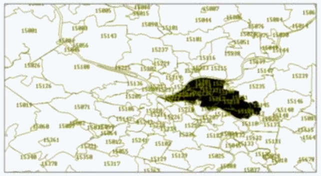
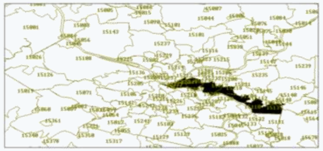
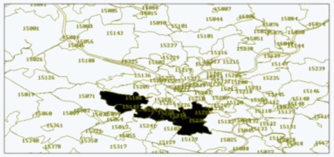
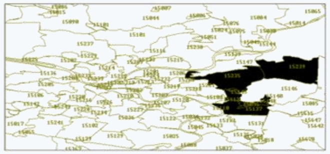
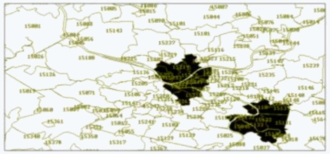




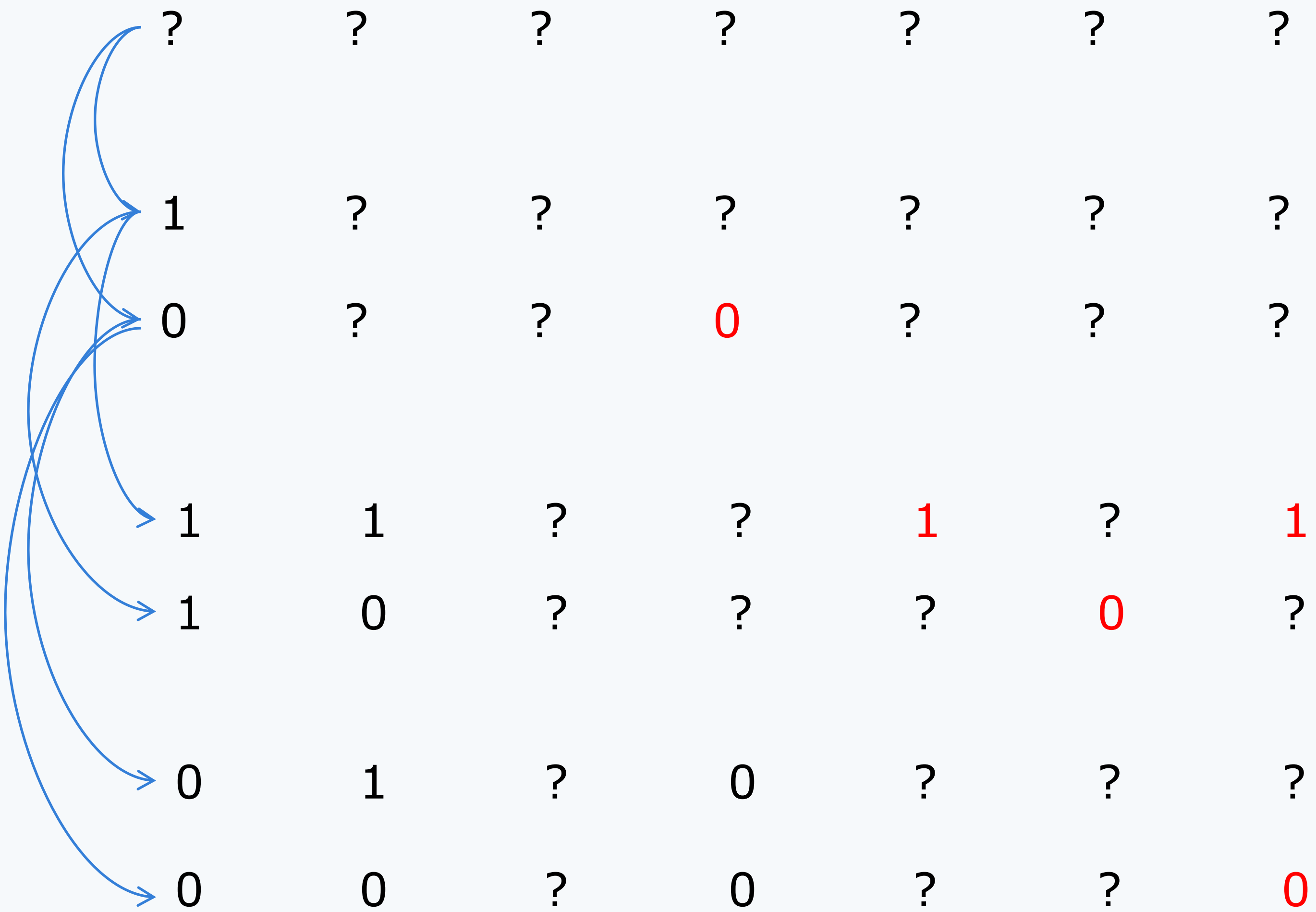
Results: Detection Power

We compared the detection power and accuracy of GraphScan to the original Kulldorff scan statistic (circular regions) on multiple semi-synthetic outbreaks injected into the data

Irregular Cluster	% Detected	Days to Detect
Circles	90%	8.7
Graphscan K=25	97%	7.6
Graphscan K=50	98%	7.5
Graphscan Single Region	96%	7.4

Thanks!


? ? ? ? ? ? ?
1 ? ? ? ? ? ?
0 ? ? 0 ? ? ?
1 1 ? ? 1 ? 1
1 0 ? ? ? 0 ?
0 1 ? 0 ? ? ?
0 0 ? 0 ? ? 0