Spatial biosurveillance

Authors of Slides

Andrew Moore
Carnegie Mellon
awm@cs.cmu.edu

Daniel Neill
Carnegie Mellon
d.neill@cs.cmu.edu

Slides and Software and Papers at:

http://www.autonlab.org

awm@cs.cmu.edu

412-268-7599

Thanks to:

Howard Burkom, Greg Cooper, Kenny Daniel, Bill Hogan, Martin Kulldorf, Robin Sabhnani, Jeff Schneider, Rich Tsui, Mike Wagner

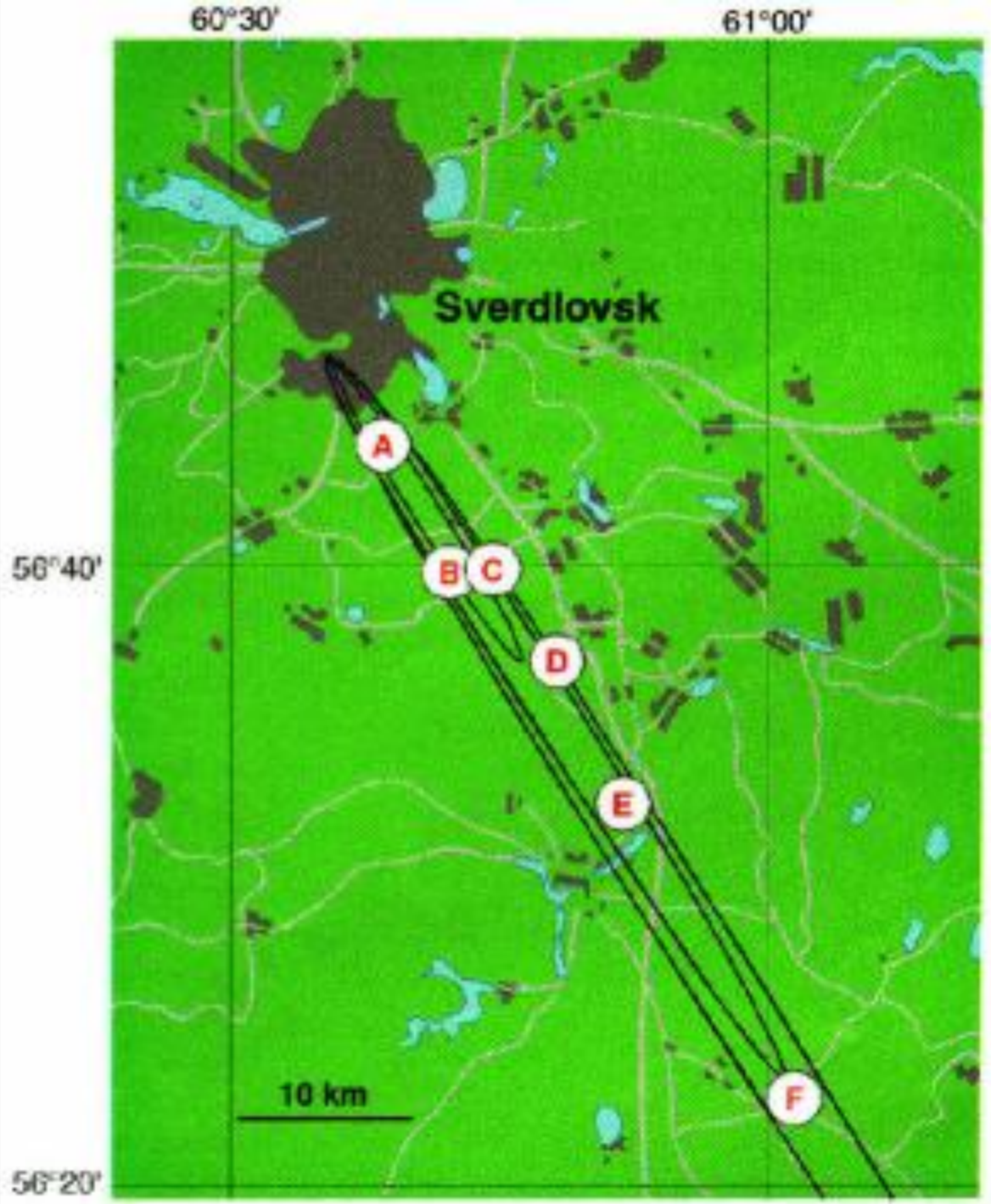
..Early Thursday Morning. Russia. April 1979...

Sverdlovsk: Aerial View

Sverdlovsk: Aerial View

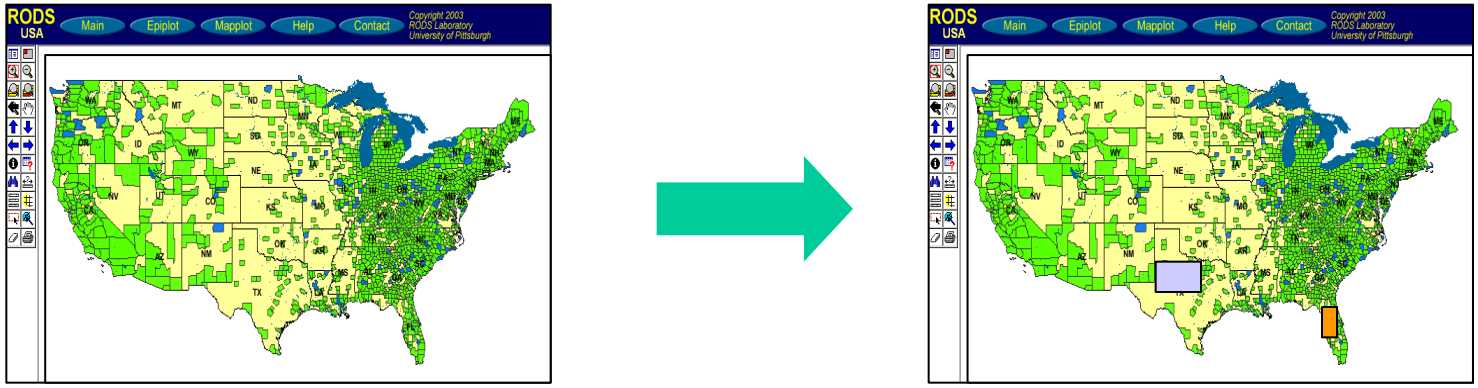
- During April and May 1979, there were 77 Confirmed cases of inhalational anthrax

Sverdlovsk Region: Epi-map



Goals of spatial cluster detection

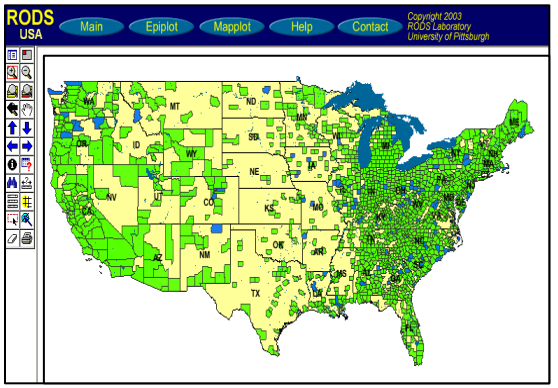
- To identify the locations, shapes, and sizes of potentially anomalous spatial regions.
- To determine whether each of these potential clusters is more likely to be a “true” cluster or a chance occurrence.
- In other words, is anything unexpected going on, and if so, where?



Disease surveillance

Given: count for each zip code

(e.g. number of Emergency Dept. visits, or over-the-counter drug sales, of a specific type)



Do any regions have sufficiently high counts to be indicative of an emerging disease epidemic in that area?

How many cases do we expect to see in each area?

Are there any regions with significantly more cases than expected?

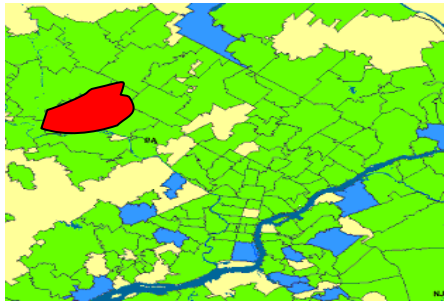
A simple approach

- For each zip code:
  - Infer how many cases we expect to see, either from given denominator data (e.g. census population) or from historical data (e.g. time series of previous counts).
  - Perform a separate statistical significance test on that zip code, obtaining its *p*-value.
- Report all zip codes that are significant at some level α.

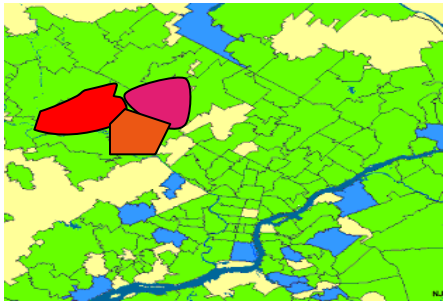
What are the potential problems?

A simple approach

- Fo
- Report all zip co are significant at some level α



One abnormal zip code might not be interesting…

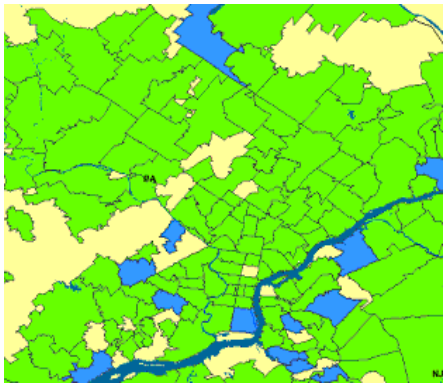


But a cluster of abnormal zip codes would be!

What are the potential problems?

A simple approach

- Fo
- Report all zip co are significant at some level α



Multiple hypothesis testing

Thousands of locations to test…

5% chance of false positive for each…

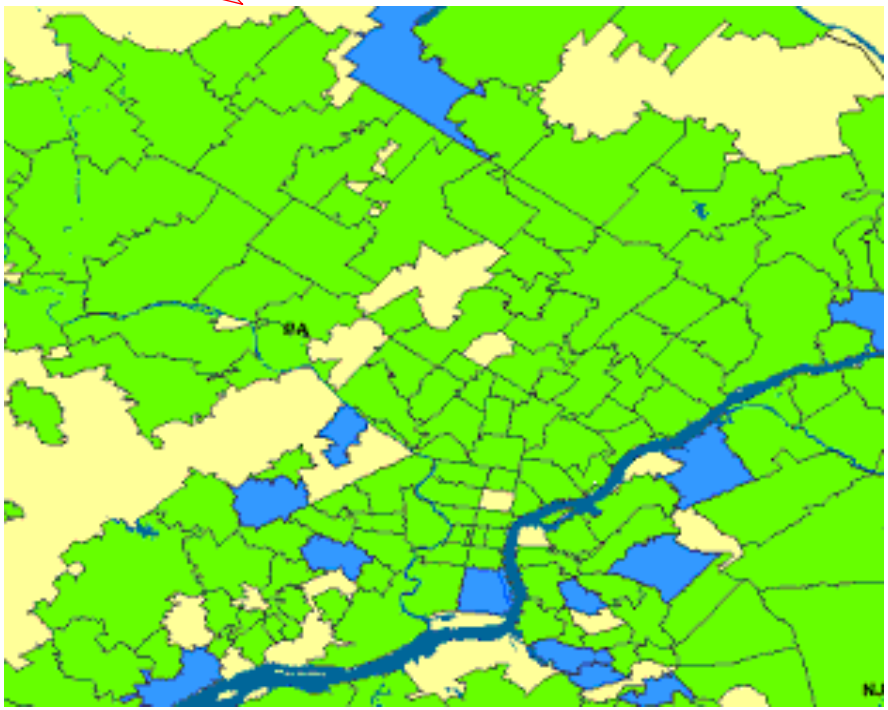
Almost certain to get large numbers of false alarms!

How do we bound the overall probability of getting any false alarms?

What are the potential problems?

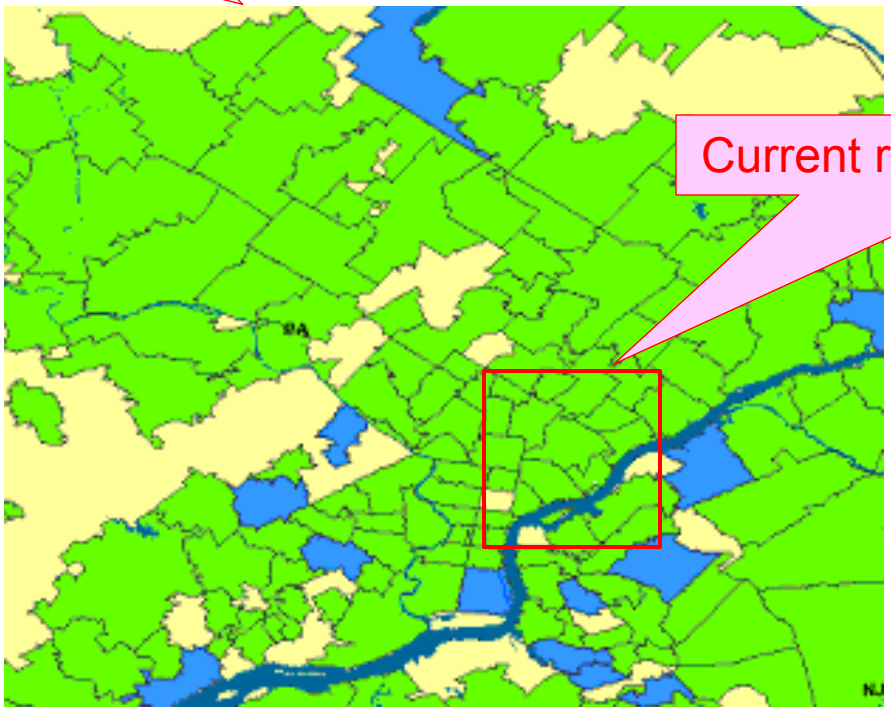
One Step of Spatial Scan

Entire area being scanned



One Step of Spatial Scan

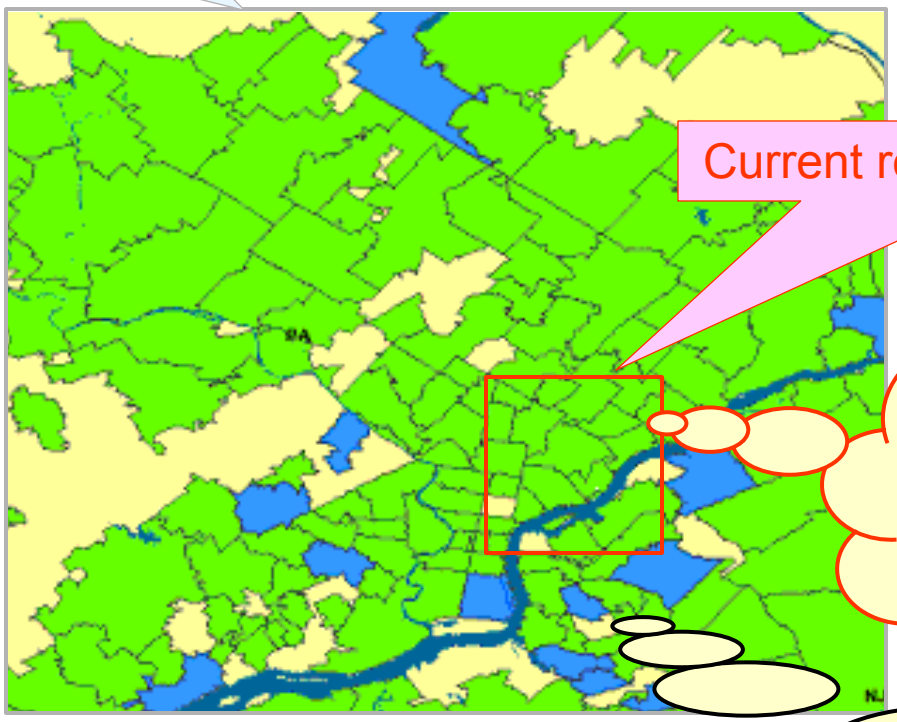
Entire area being scanned



Current region being considered

One Step of Spatial Scan

Entire area being scanned



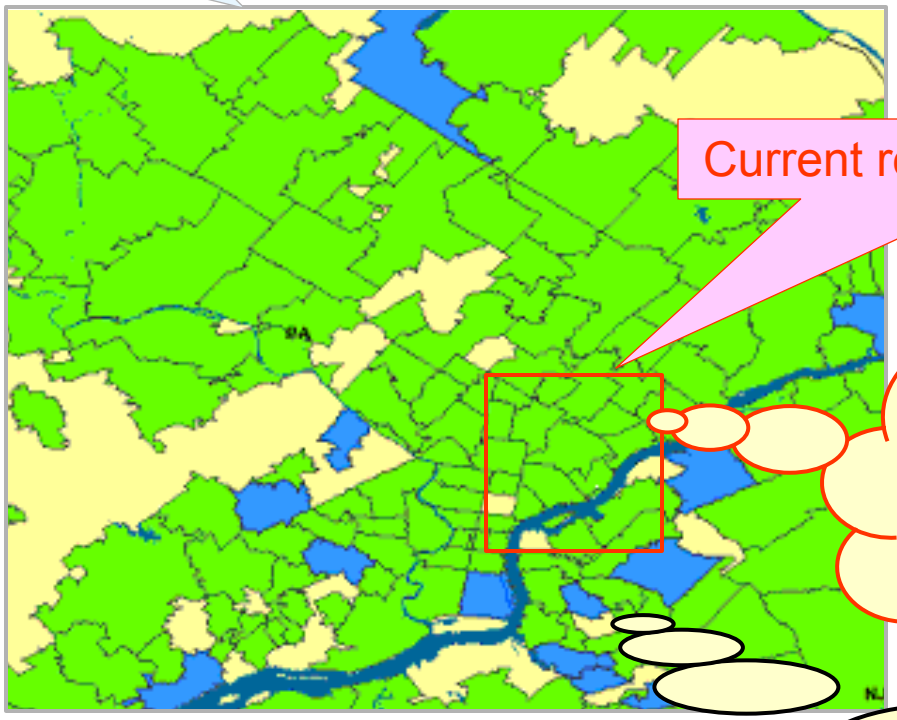
Current region being considered

I have a population of 5300 of whom 53 are sick (1%)

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

One Step of Spatial Scan

Entire area being scanned



Current region being considered

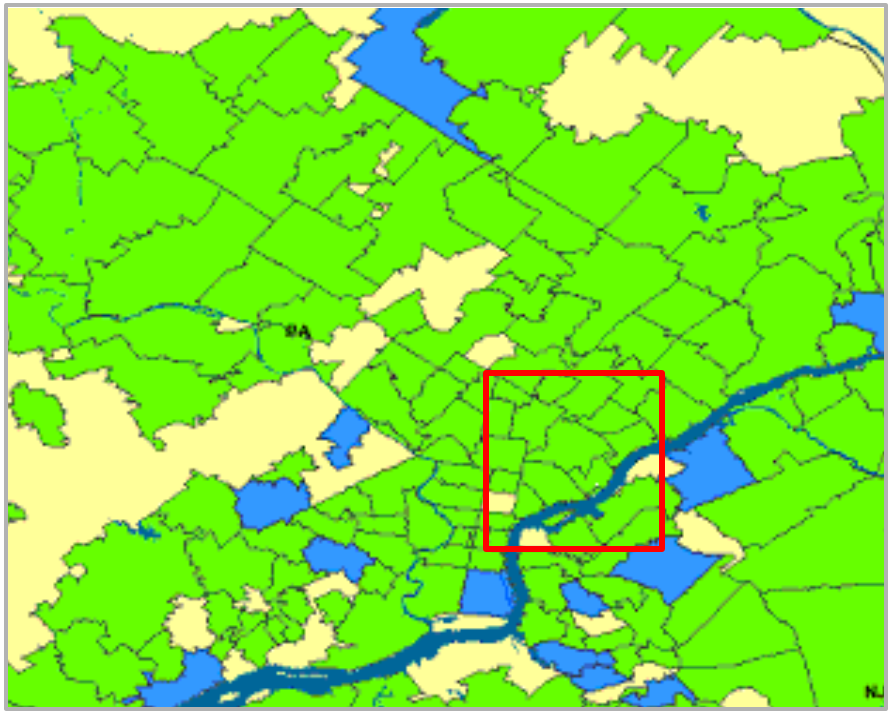
I have a population of 5300 of whom 53 are sick (1%)

Everywhere else has a population of 2,200,000 o whom 20,000 are sick (0.9%)

So... *is that a big deal? Evaluated with Score function.*

Scoring functions

- Define <u>models</u>:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.

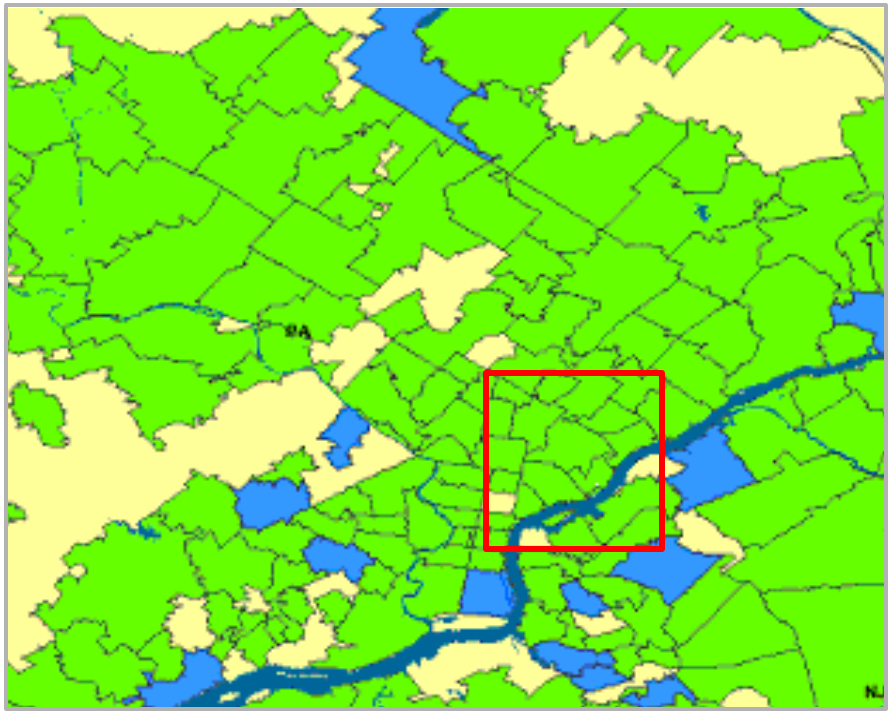


Scoring functions

- Define <u>models</u>:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a <u>score function</u> *Score(S) = Score(C, B).*
  - Likelihood ratio:
  - To find the most significant region:

$$\text{Score}(S) = \frac{L(\text{Data} \mid H_1(S))}{L(\text{Data} \mid H_0)}$$

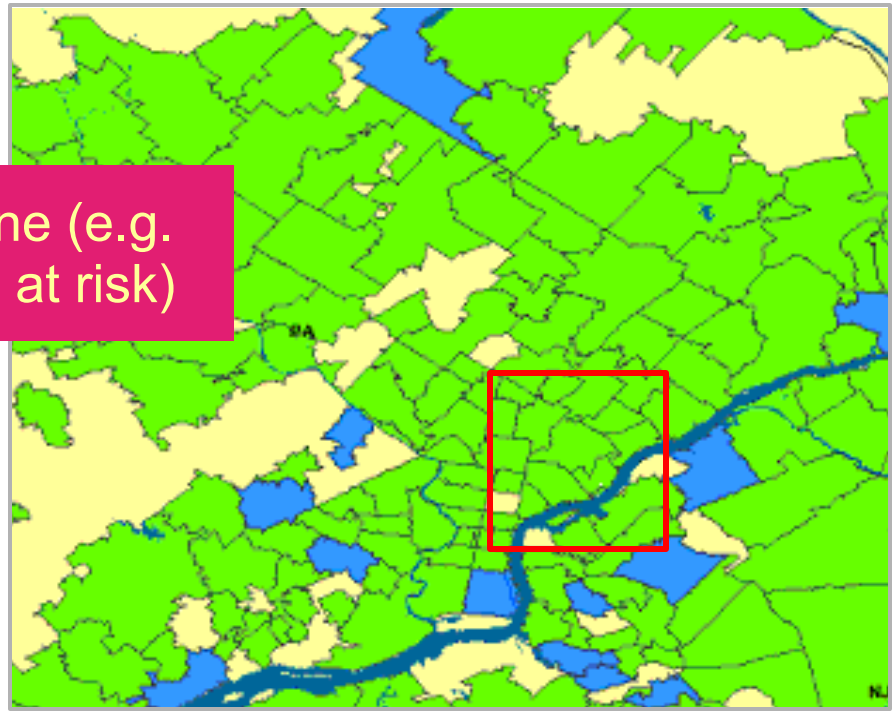
$$S^* = \arg\max_S \text{Score}(S)$$



Scoring function

C = Count in region

B = Baseline (e.g. Population at risk)



- Define <u>models</u>:
  - of the null hypothesis H_0: no attacks
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a <u>score function</u> *Score(S) = Score(C, B).*
  - Likelihood ratio:

$$\text{Score}(S) = \frac{L(\text{Data} \mid H_1(S))}{L(\text{Data} \mid H_0)}$$

  - To find the most significant region:

$$S^* = \arg\max_S \text{Score}(S)$$

Scoring functions

- Define <u>models</u>:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a <u>score function</u> *Score(S) = Score(C, B).*
  - Likelihood ratio:
  - To find the most significant region:

$$\text{Score}(S) = \frac{L(\text{Data} \mid H_1(S))}{L(\text{Data} \mid H_0)}$$

$$S^* = \arg\max_S \text{Score}(S)$$

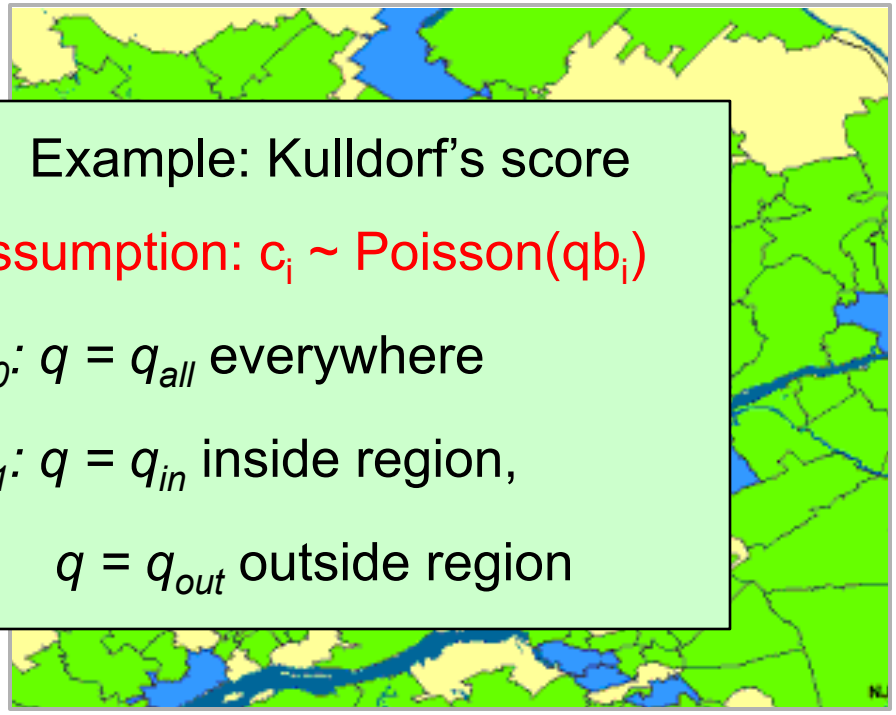
Example: Kulldorf's score

Assumption: $c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

H_1: $q = q_{in}$ inside region,

$q = q_{out}$ outside region

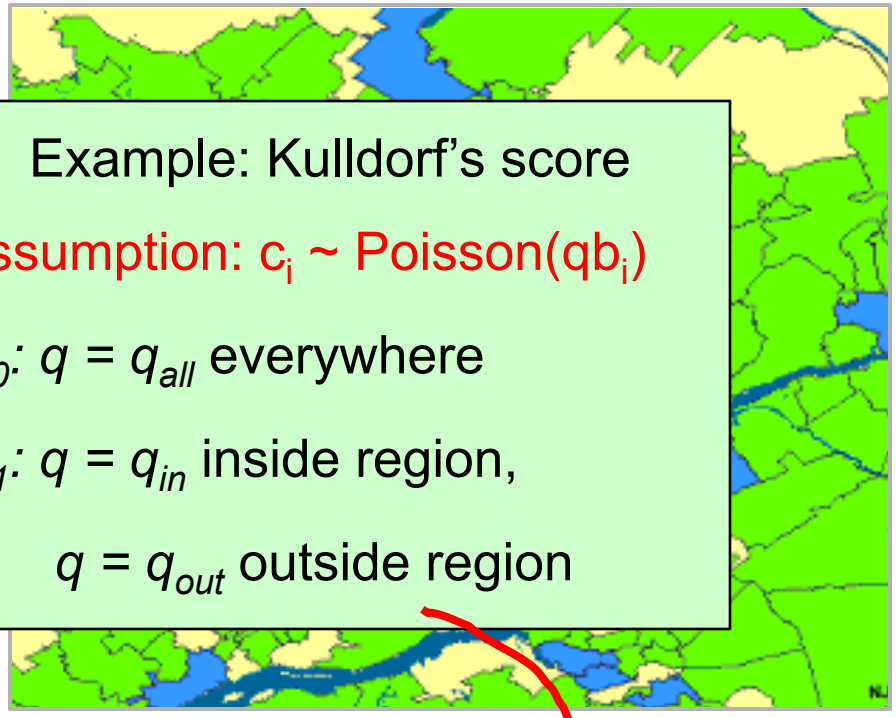


Scoring functions

- Define <u>models</u>:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a <u>score function</u> *Score(S) = Score(C, B).*
  - Likelihood ratio:
  - To find the most significant region:

$$\text{Score}(S) = \frac{L(\text{Data} \mid H_1(S))}{L(\text{Data} \mid H_0)}$$

$$S^* = \arg\max_S \text{Score}(S)$$



Example: Kulldorf's score

Assumption: $c_i \sim \text{Poisson}(qb_i)$

H_0: $q = q_{all}$ everywhere

H_1: $q = q_{in}$ inside region,

$q = q_{out}$ outside region

$$D(S) = C \log \frac{C}{B} + (C_{tot} - C) \log \frac{C_{tot} - C}{B_{tot} - B} - C_{tot} \log \frac{C_{tot}}{B_{tot}}$$

(Individually Most Powerful statistic for detecting significant increases) *(but still…just an example)*

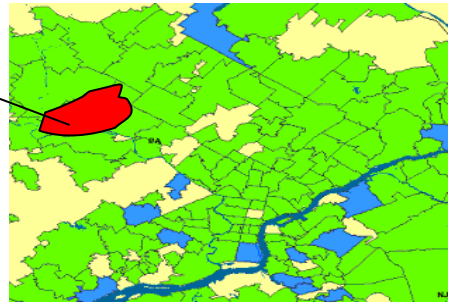
The generalized spatial scan

1. Obtain data for a set of spatial locations s_i.
2. Choose a set of spatial regions S to search.
3. Choose models of the data under null hypothesis H_0 (no clusters) and alternative hypotheses $H_1(S)$ (cluster in region S).
4. Derive a score function F(S) based on $H_1(S)$ and H_0.
5. Find the most anomalous regions (i.e. those regions S with highest F(S)).
6. Determine whether each of these potential clusters is actually an anomalous cluster.

1. Obtain data for a set of spatial locations s_i.

- For each spatial location s_i, we are given a count c_i and a baseline b_i.
- For example: c_i = # of respiratory disease cases, b_i = at-risk population.
- Goal: to find regions where the counts are higher than expected, given the baselines.

$c_i = 20$,
$b_i = 5000$



Population-based method:

Baselines represent population, whether given (e.g. census) or inferred (e.g. from sales); can be adjusted for age, risk factors, seasonality, etc.

Under null hypothesis, we expect counts to be proportional to baselines.

Compare disease rate (count / pop) inside and outside region.

Expectation-based method:

Baselines represent expected counts, inferred from the time series of previous counts, accounting for day-of-week and seasonality effects.

Under null hypothesis, we expect counts to be equal to baselines.

Compare region's actual count to its expected count.

1. Obtain data for a set of spatial locations s_i.

- For each spatial location s_i, we are given a count c_i and a baseline b_i.
- For example: c_i = # of [illegible] disease cases, b_i = a[illegible]
- Goal: to find regions [illegible] are higher than expected, g[illegible] baselines.

c_i = 20,
b_i = 5000

Discussion question: When is it preferable to use each method?

Population-based method:

Baselines represent population, whether given (e.g. census) or inferred (e.g. from sales); can be adjusted for age, risk factors, seasonality, etc.

Under null hypothesis, we expect counts to be proportional to baselines.

Compare disease rate (count / pop) inside and outside region.

Expectation-based method:

Baselines represent expected counts, inferred from the time series of previous counts, accounting for day-of-week and seasonality effects.

Under null hypothesis, we expect counts to be equal to baselines.

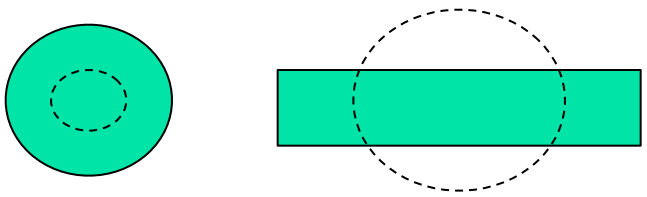
Compare region's actual count to its expected count.

2. Choose a set of spatial regions S to search.

- Some practical considerations:
  - Set of regions should cover entire search space.
  - Adjacent regions should partially overlap.
- Choose a set of regions that corresponds well with the size/shape of the clusters we want to detect.
  - Typically, we consider some fixed shape (e.g. circle, rectangle) and allow its location and dimensions to vary.

Don't search too few regions:

Reduced power to detect clusters outside the search space.



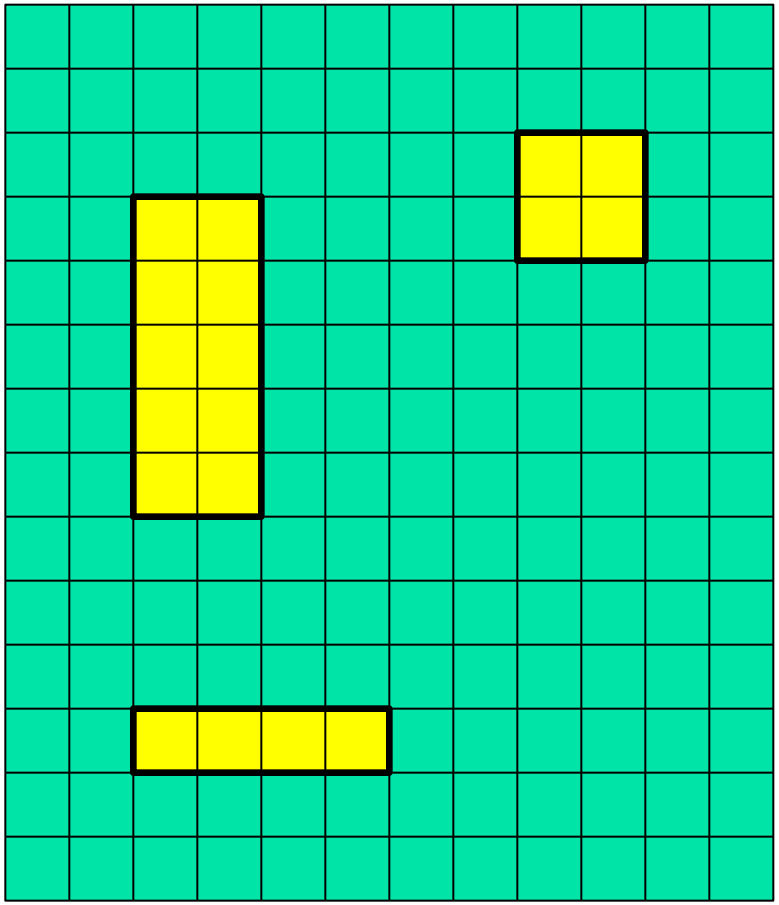
Don't search too many regions:

Overall power to detect any given subset of regions reduced because of multiple hypothesis testing.

Computational infeasibility!

2. Choose a set of spatial regions S to search.

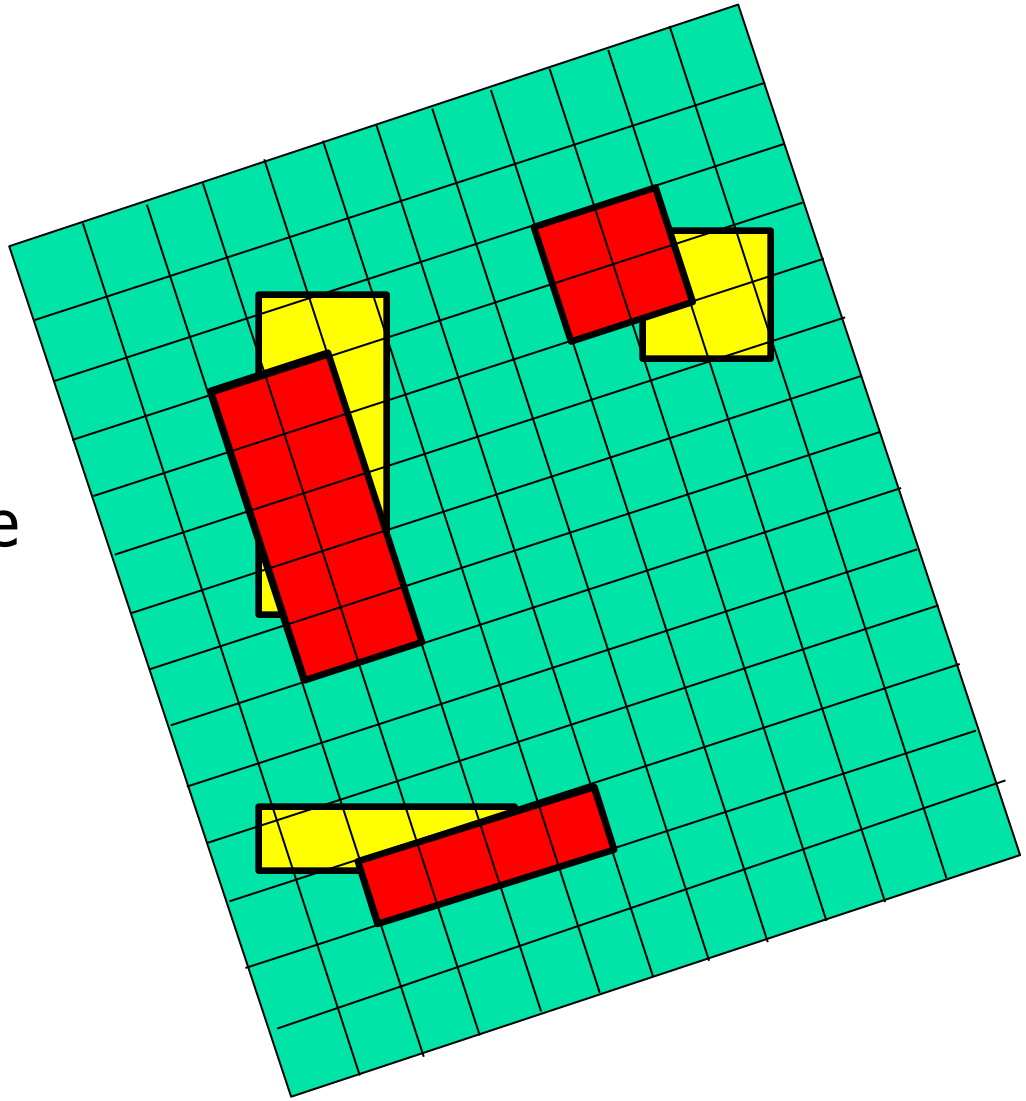
- Our typical approach for disease surveillance:
  - map spatial locations to grid
  - search over the set of all gridded rectangular regions.
- Allows us to detect both compact and elongated clusters (important because of wind- or water-borne pathogens).
- Computationally efficient
  - can evaluate any rectangular region in constant time
  - can use fast spatial scan algorithm



2. Choose a set of spatial regions S to search.

- Our typical approach for disease surveillance:
  - rid
  - ns.
  - use
  - can evaluate any rectangular region in constant time
  - can use fast spatial scan algorithm

Can also search over non-axis-aligned rectangles by examining multiple rotations of the data



3-4. Choose models of the data under H_0 and $H_1(S)$, and derive a score function F(S).

- Most difficult steps: must choose models which are efficiently computable and relevant.

3-4. Choose models of the data under H_0 and $H_1(S)$, and derive a score function F(S).

- Most difficult steps: must choose models which are efficiently computable and relevant.

F(S) must be computable as function of some additive sufficient statistics of region S, e.g. total count C(S) and total baseline B(S).

3-4. Choose models of the data under H_0 and $H_1(S)$, and derive a score function F(S).

- Most difficult steps: must choose models which are efficiently computable and relevant.

tradeoff!

F(S) must be computable as function of some additive sufficient statistics of region S, e.g. total count C(S) and total baseline B(S).

Any simplifying assumptions should not greatly affect our ability to distinguish between clusters and non-clusters.

3-4. Choose models of the data under H_0 and $H_1(S)$, and derive a score function F(S).

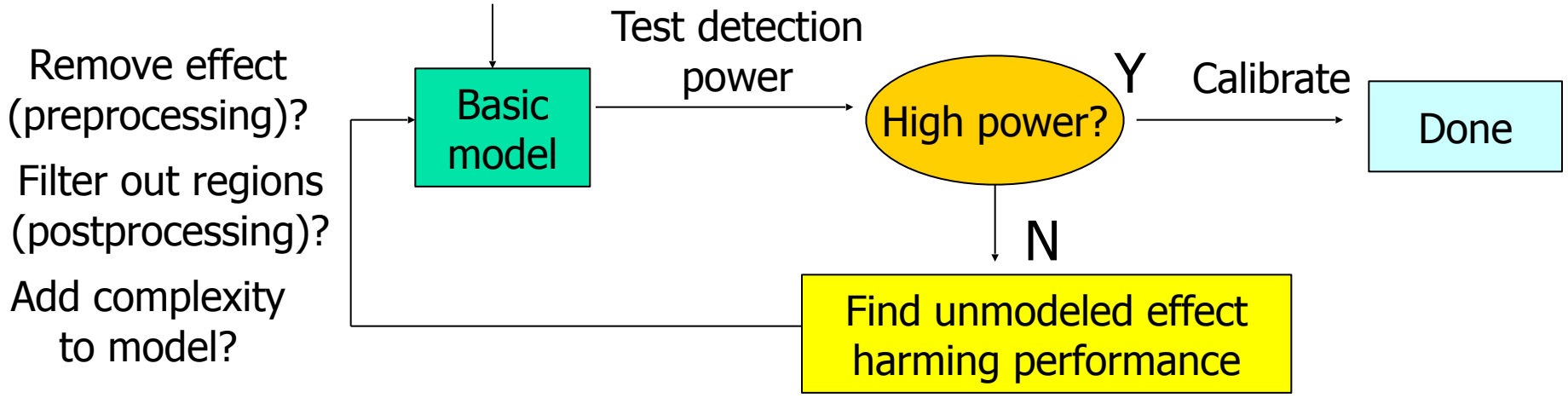
- Most difficult steps: must choose models which are efficiently computable and relevant.

tradeoff!

F(S) must be computable as function of some additive sufficient statistics of region S, e.g. total count C(S) and total baseline B(S).

Any simplifying assumptions should not greatly affect our ability to distinguish between clusters and non-clusters.

Iterative design process



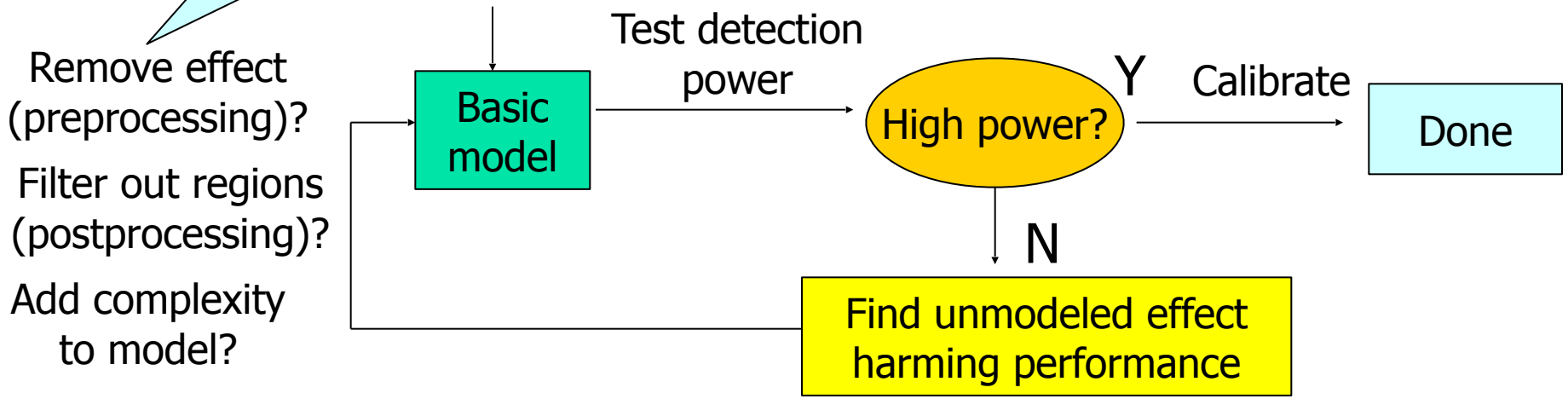
3-4. Choose models of the data under H_0 and $H_1(S)$, and derive a score function F(S).

- Most difficult steps: must choose models which are efficiently computable and relevant.

Discussion question: What effects should be treated with each technique?

F(S) m… …mptions should not some addit… …ability to distinguish e.g. total count C… …baseline B(S). between clusters and non-clusters.

Iterative design process



Computing the score function

Method 1 (Frequentist, hypothesis testing approach):

Use likelihood ratio $F(S) = \frac{\Pr(Data \mid H_1(S))}{\Pr(Data \mid H_0)}$

Method 2 (Bayesian approach):

Prior probability of region S

Use posterior probability $F(S) = \frac{\Pr(Data \mid H_1(S)) \Pr(H_1(S))}{\Pr(Data)}$

What to do when each hypothesis has a parameter space Θ?

Method A (Maximum likelihood approach)

$$\Pr(Data \mid H) = \max_{\theta \in \Theta(H)} \Pr(Data \mid H, \theta)$$

Method B (Marginal likelihood approach)

$$\Pr(Data \mid H) = \int_{\theta \in \Theta(H)} \Pr(Data \mid H, \theta) \Pr(\theta)$$

Computing the score function

Method 1 (Frequentist, hypothesis testing approach):

Use likelihood ratio $F(S) = \frac{\Pr(Data \mid H_1(S))}{\Pr(Data \mid H_0)}$

Most common (frequentist) approach: use likelihood ratio statistic, with maximum likelihood estimates of any free parameters, and compute statistical significance by randomization.

Method A (Maximum likelihood approach)

$$\Pr(Data \mid H) = \max_{\theta \in \Theta(H)} \Pr(Data \mid H, \theta)$$

5. Find the most anomalous regions, i.e. those regions S with highest F(S).

- Naïve approach: compute F(S) for each spatial region S.
- Better approach: apply fancy algorithms (e.g. Kulldorf's SatScan or the fast spatial scan algorithm (Neill and Moore, KDD 2004).

Problem: millions of regions to search!

5. Find the most anomalous regions, i.e. those regions S with highest F(S).

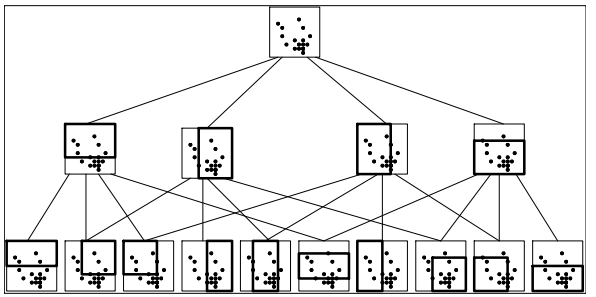
- Naïve approach: compute F(S) for each spatial region S.

Problem: millions of regions to search!

- Better approach: apply fancy algorithms (e.g. Kulldorf's SatScan or the fast spatial scan algorithm (Neill and Moore, KDD 2004).

Start by examining large rectangular regions S. If we can show that none of the smaller rectangles contained in S can have high scores, we do not need to individually search each of these subregions.

Using a multiresolution data structure (overlap-kd tree) enables us to efficiently move between searching at coarse and fine resolutions.



5. Find the most anomalous regions, i.e. those regions S with S).

S) for each spatial

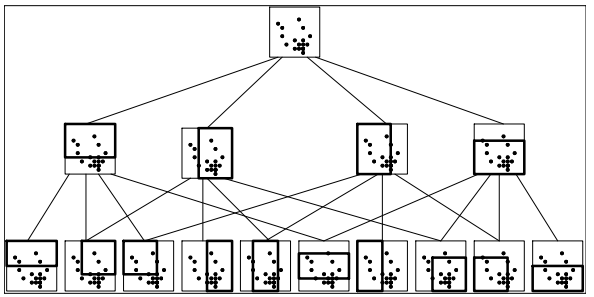
m: millions of regions to search!

- Better approach: fancy algorithms (e.g. Kulldorf's SatScan e fast spatial scan algorithm (Neill and Moore, KDD 2004).

Result: 20-2000x speedups vs. naïve approach, without any loss of accuracy

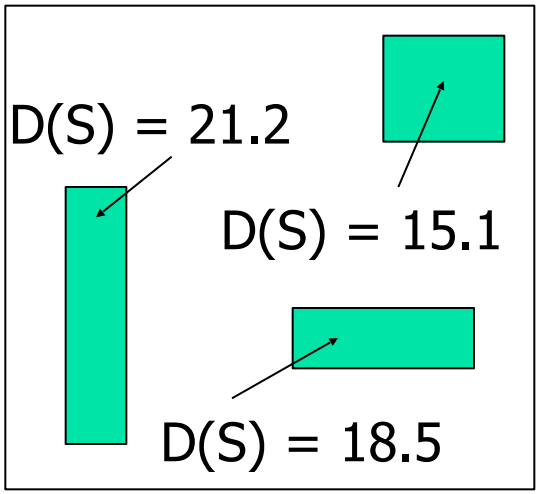
Start by examining large rectangular regions S. If we can show that none of the smaller rectangles contained in S can have high scores, we do not need to individually search each of these subregions.

Using a multiresolution data structure (overlap-kd tree) enables us to efficiently move between searching at coarse and fine resolutions.



6. Determine whether each of these potential clusters is actually an anomalous cluster.

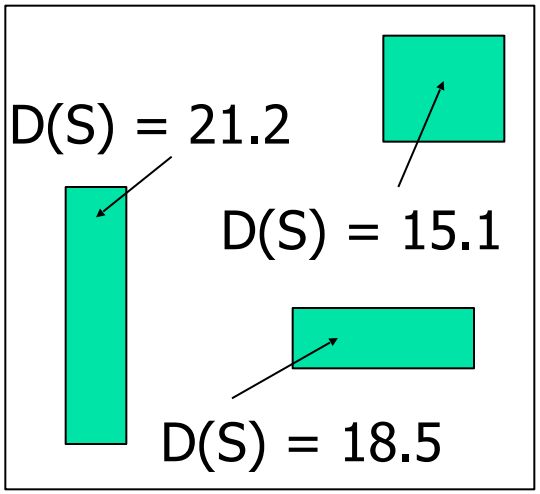
- Frequentist approach: calculate statistical significance of each region by randomization testing.



Original grid G

6. Determine whether each of these potential clusters is actually an anomalous cluster.

- Frequentist approach: calculate statistical significance of each region by randomization testing.

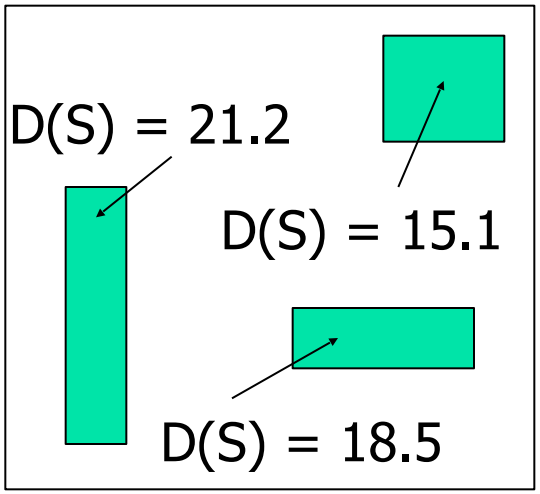


Original grid G

1. Create R = 999 replica grids by sampling under H_0, using max-likelihood estimates of any free params.

6. Determine whether each of these potential clusters is actually an anomalous cluster.

- Frequentist approach: calculate statistical significance of each region by randomization testing.

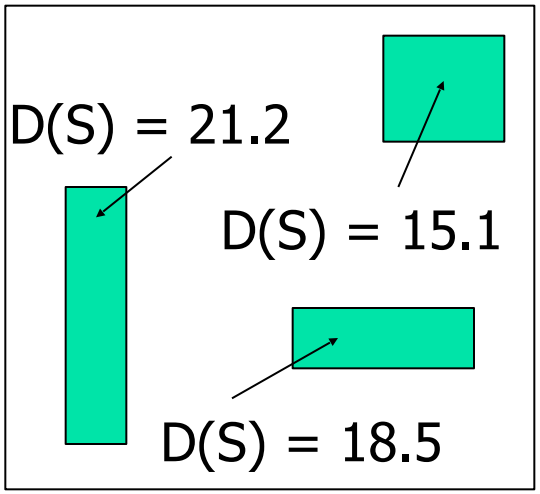


Original grid G

1. Create R = 999 replica grids by sampling under H_0, using max-likelihood estimates of any free params.
2. Find maximum region score D* for each replica.

6. Determine whether each of these potential clusters is actually an anomalous cluster.

- Frequentist approach: calculate statistical significance of each region by randomization testing.

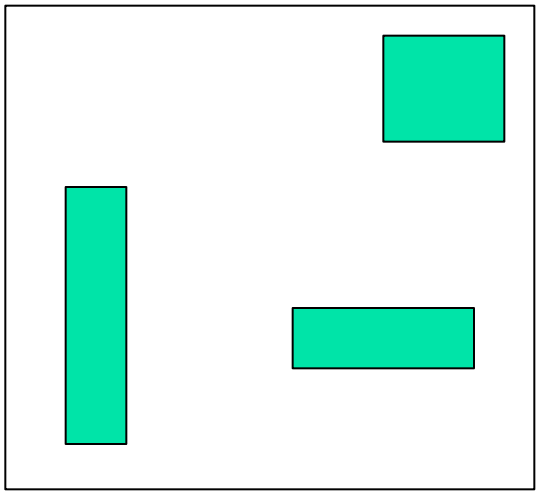


Original grid G

1. Create R = 999 replica grids by sampling under H_0, using max-likelihood estimates of any free params.
2. Find maximum region score D* for each replica.
3. For each potential cluster S, count R_{beat} = number of replica grids G' with D*(G') higher than D(S).
4. p-value of region S = $(R_{beat}+1)/(R+1)$.
5. All regions with p-value < α are significant at level α.

6. Determine whether each of these potential clusters is actually an anomalous cluster.

- Bayesian approach: calculate posterior probability of each potential cluster.



Original grid G

1. Score of region S = $Pr(Data \mid H_1(S))\ Pr(H_1(S))$

2. Total probability of the data: $Pr(Data) = Pr(Data \mid H_0)\ Pr(H_0) + \sum_S Pr(Data \mid H_1(S))\ Pr(H_1(S))$

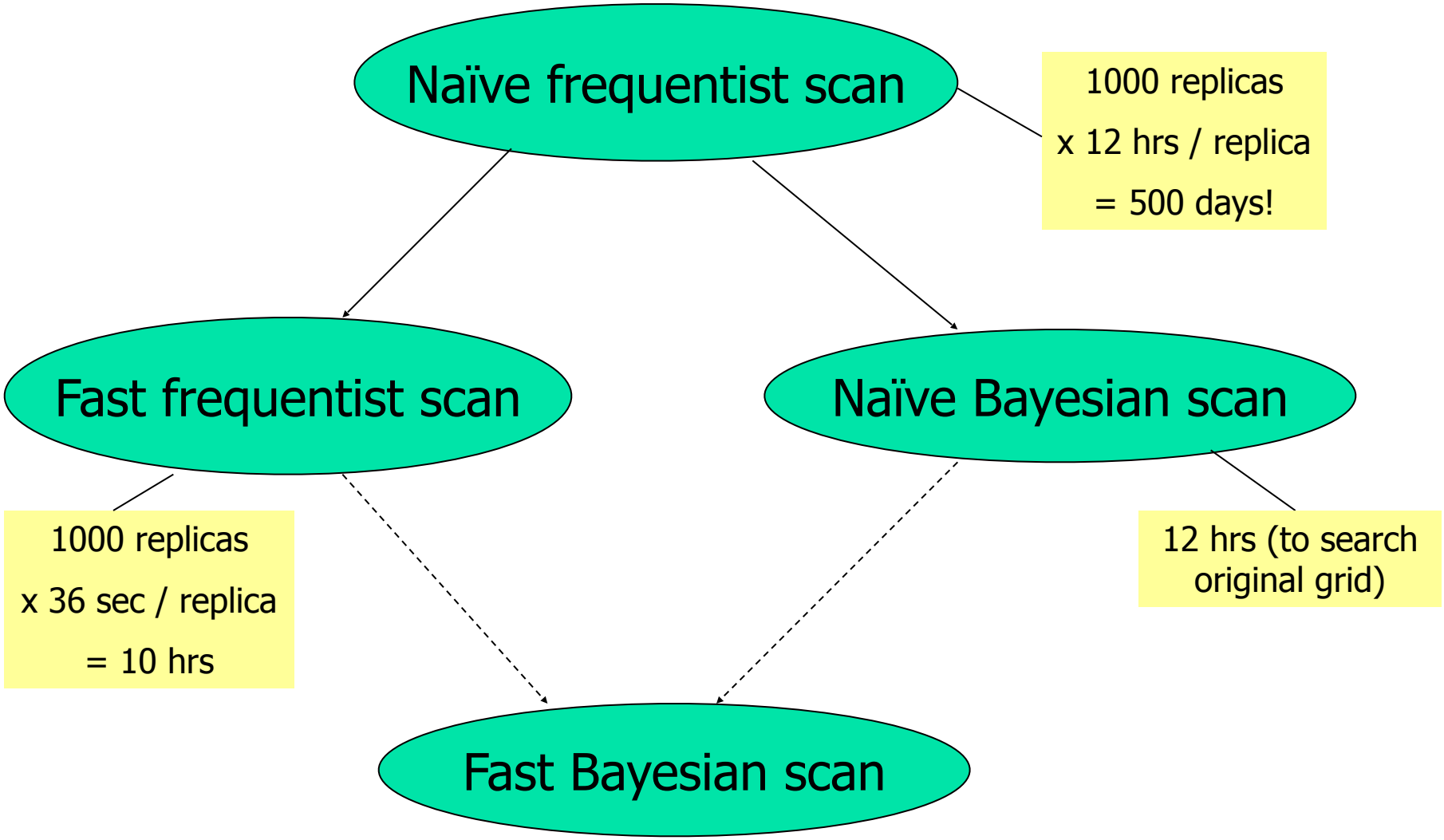
3. Posterior probability of region S: $Pr(H_1(S) \mid Data) = Pr(Data \mid H_1(S))\ Pr(H_1(S)) / Pr(Data)$.

4. Report all clusters with posterior probability > some threshold, or "sound the alarm" if total posterior probability of all clusters sufficiently high.

No randomization testing necessary… about 1000x faster than naïve frequentist approach!

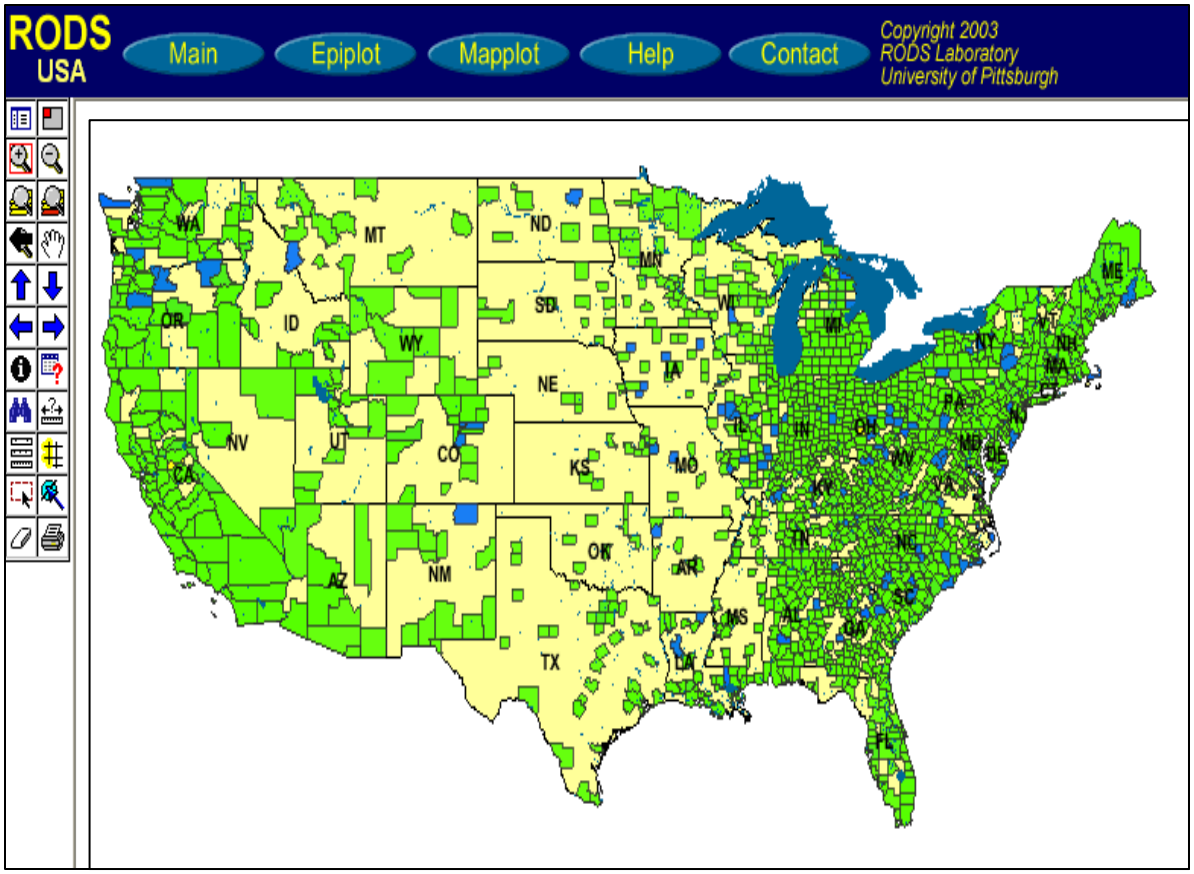
Making the spatial scan fast

256 x 256 grid = 1 billion regions!



Why the Scan Statistic speed obsession?

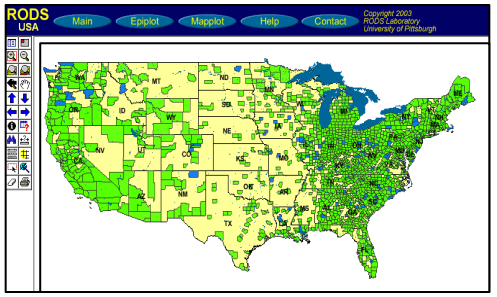
- Traditional Scan Statistics very expensive, especially with Randomization tests
- Going national
- A few hours could actually matter!



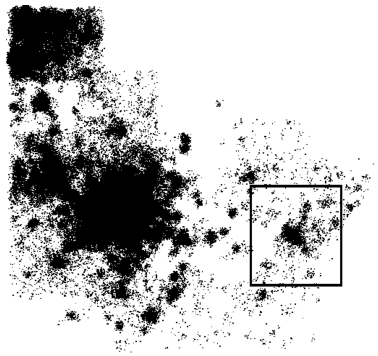
Results

Summary of results

- The fast spatial scan results in huge speedups (as compared to exhaustive search), making fast real-time detection of clusters feasible.
- No loss of accuracy: fast spatial scan finds the exact same regions and p-values as exhaustive search.



OTC data from National Retail Data Monitor



ED data

Performance comparison

Algorithm name	Search space	Number of regions	Search time (total)	Time / region	Likelihood ratio
SaTScan	Circles centered at datapts	150 billion	16 hours	400 ns	413.56
exhaustive	Axis-aligned rectangles	1.1 trillion	45 days	3600 ns	429.85
fast spatial scan	Axis-aligned rectangles	1.1 trillion	81 minutes	4.4 ns	429.85

- On ED dataset (600,000 records), 1000 replicas
- For SaTScan: M=17,000 distinct spatial locations
- For Exhaustive/fast: 256 x 256 grid

Performance comparison

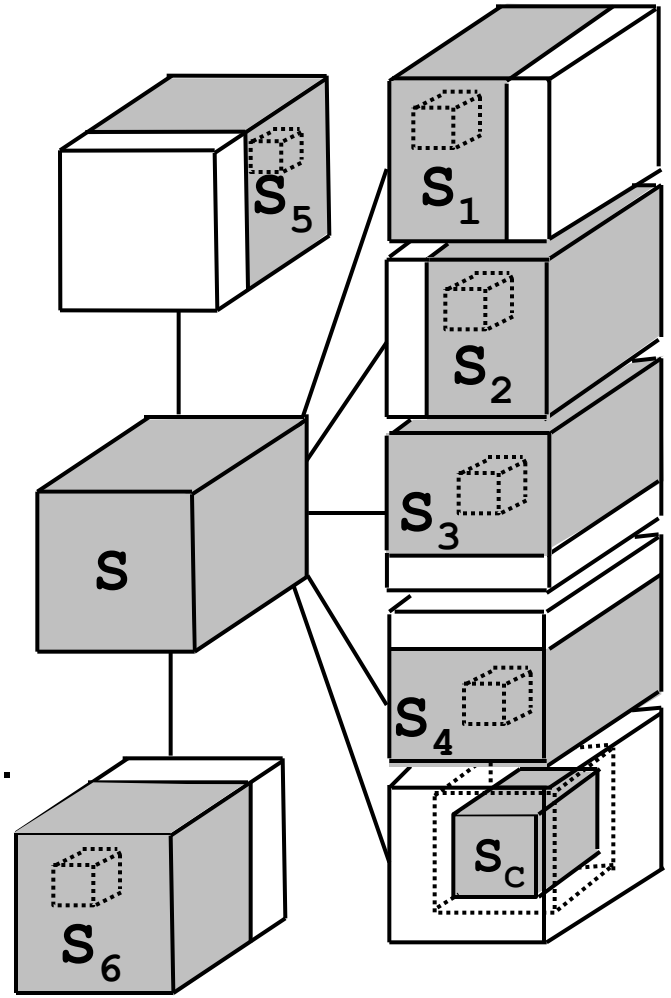
Algorithm name	Search space	Number of regions	Search time (total)	Time / region	Likelihood ratio
SaTScan	Circles centered at datapts	150 billion	16 hours	400 ns	413.56
exhaustive	Axis-aligned rectangles	1.1 trillion	45 days	3600 ns	429.85
fast spatial scan	Axis-aligned rectangles	1.1 trillion	81 minutes	4.4 ns	429.85

- On ED dataset (600,000 records), 1000 replicas
- For SaTScan: M=17,000 distinct spatial locations
- For Exhaustive/fast: 256 x 256 grid

- Algorithms: Neill and Moore, NIPS 2003, KDD 2004
- Deployment: *Neill, Moore, Tsui and Wagner, Morbidity and Mortality Weekly Report, Nov. '04*

d-dimensional partitioning

- Parent region S is divided into 2d overlapping children: an “upper child” and a “lower child” in each dimension.
- Then for any rectangular subregion S’ of S, exactly one of the following is true:
  - S’ is contained entirely in (at least) one of the children $S_1 \ldots S_{2d}$.
  - S’ contains the center region S_C, which is common to all the children.
- Starting with the entire grid G and repeating this partitioning recursively, we obtain the overlap-kd tree structure.



- Algorithm: Neill, Moore and Mitchell NIPS 2004

Limitations of the algorithm

- Data must be aggregated to a grid.
- Not appropriate for very high-dimensional data.
- Assumes that we are interested in finding (rotated) rectangular regions.
- Less useful for special cases (e.g. square regions, small regions only).
- Slower for finding multiple regions.

Related work

- non-specific clustering: evaluates general tendency of data to cluster
- focused clustering: evaluates risk w.r.t. a given spatial location (e.g. potential hazard)
- disease mapping: models spatial variation in risk by applying spatial smoothing.
- spatial scan statistics (and related techniques).