Some useful snippets!

The Auton Lab

Carnegie Mellon University

www.autonlab.org

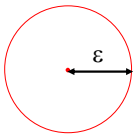
Priebe/BIRCH/Theisson

- Suppose you want to do clustering but can only afford one pass through the data.
- Step 1: Choose an epsilon: the radius of your cluster.
- Step 2: Slog through the data.

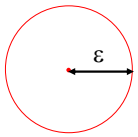
Priebe/BIRCH/Theisson

Priebe/BIRCH/Theisson

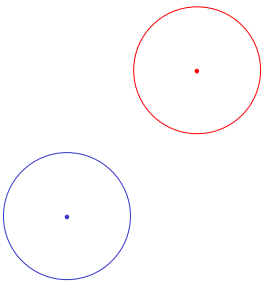
Priebe/BIRCH/Theisson



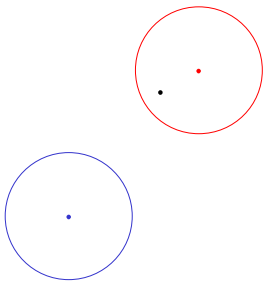
Priebe/BIRCH/Theisson



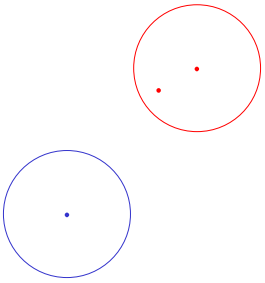
Priebe/BIRCH/Theisson



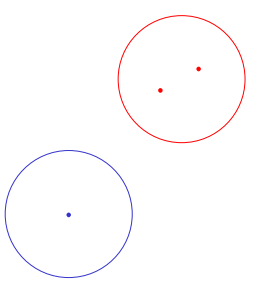
Priebe/BIRCH/Theisson



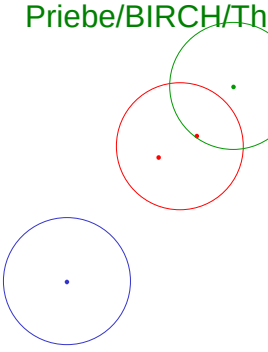
Priebe/BIRCH/Theisson



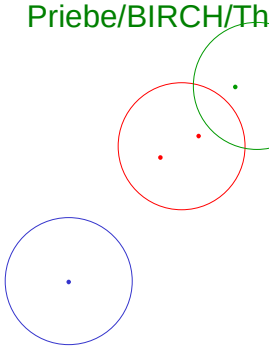
Priebe/BIRCH/Theisson



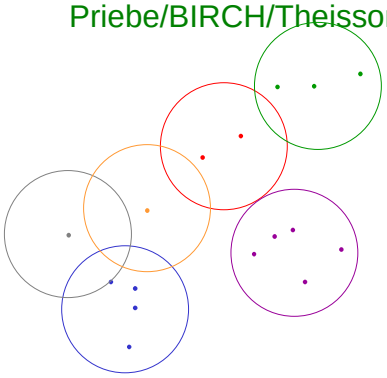
Priebe/BIRCH/Theisson



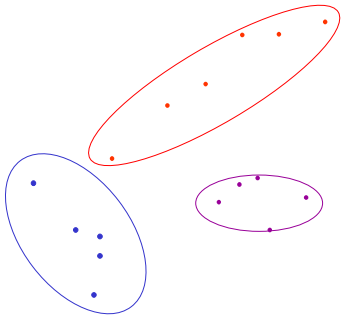
Priebe/BIRCH/Theisson



Priebe/BIRCH/Theisson



Priebe/BIRCH/Theisson



Sampling and Racing

M = 1 Million Records

A	B	C
1	0	1
0	1	0
0	1	1
0	0	1
1	1	1
0	1	1
0	1	1
0	0	1
1	1	1
:	:	:
1	1	1

Sampling and Racing

Who is most closely correlated with A. Is it B or C?

M = 1 Million Records

A	B	C
1	0	1
0	1	0
0	1	1
0	0	1
1	1	1
0	1	1
0	1	1
0	0	1
1	1	1
:	:	:
1	1	1

Sampling and Racing

Who is most closely correlated with A. Is it B or C?

- Idea One: Pick a random sample of 1000 records. Choose B if it's most closely correlated on the sample.
- Idea Two: Do same thing, but also perform significance analysis. If not 99.999% sure, repeat with a larger sample.

M = 1 Million Records

A	B	C
1	0	1
0	1	0
0	1	1
0	0	1
1	1	1
0	1	1
0	1	1
0	0	1
1	1	1
:	:	:
1	1	1

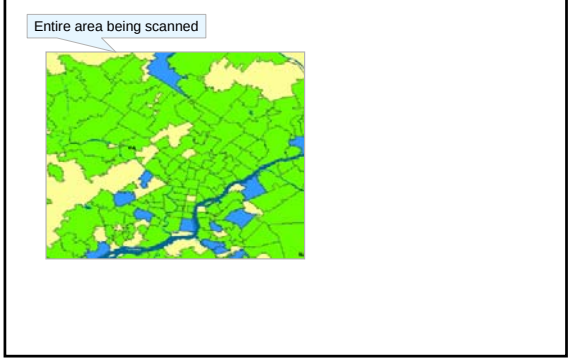
Sampling and Racing

Who is most closely correlated with A. Is it B or C?

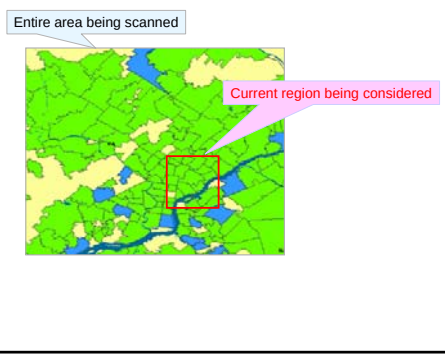
- Kaelbling used this for feature selection.
- Maron and Moore used this for cross-validation-based model selection
- Hulten and Domingos used this for k-means, mixtures, Bayes nets…
- Pelleg used this for dependency trees

- Idea One: Pick a sample of 10 Choose B if it's most correlated on the sam
- Idea Two: Do same thing, but also perform significance analysis. If not 99.999% sure, repeat with a larger sample.

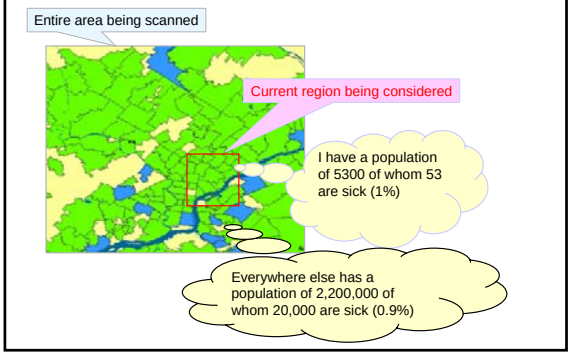
One Step of Spatial Scan



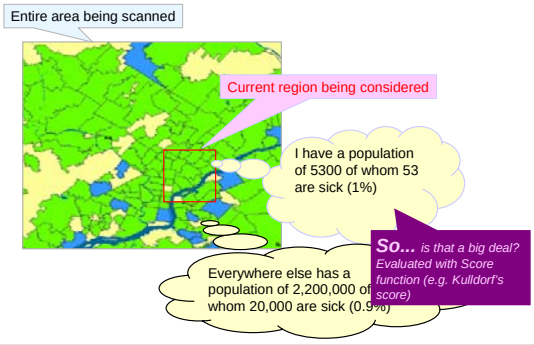
One Step of Spatial Scan



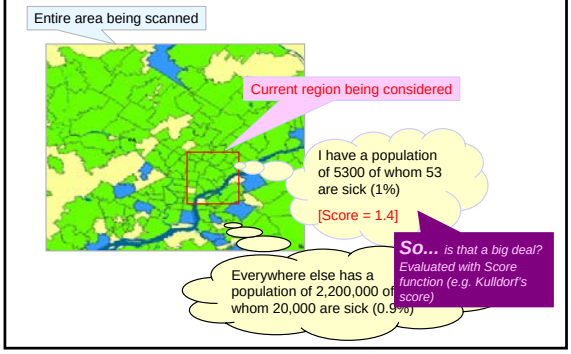
One Step of Spatial Scan



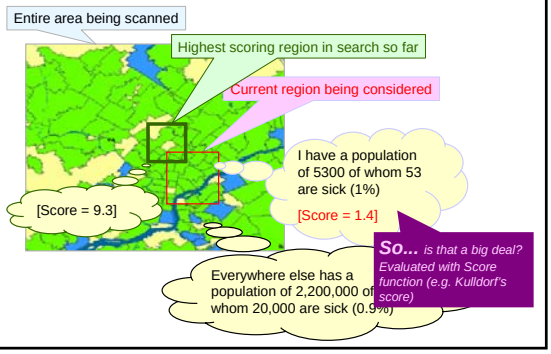
One Step of Spatial Scan



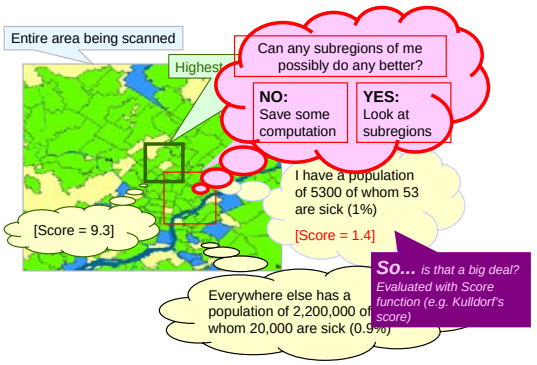
One Step of Spatial Scan



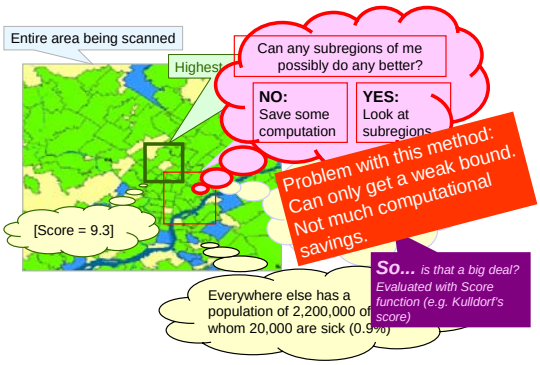
Many Steps of Spatial Scan



Many Steps of Spatial Scan



Many Steps of Spatial Scan



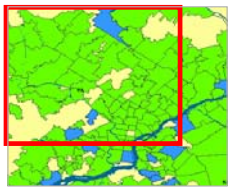
New Approach: Gridded then Exhaustive

Step 1: Gridded

Check a specific recursive overlapping set of regions called "Gridded Regions"

New Approach: Gridded then Exhaustive

Step 1: Gridded



Check a specific recursive overlapping set of regions called "Gridded Regions"

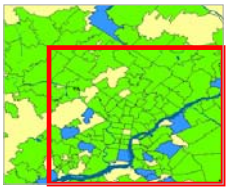
New Approach: Gridded then Exhaustive

Step 1: Gridded

Check a specific recursive overlapping set of regions called "Gridded Regions"

New Approach: Gridded then Exhaustive

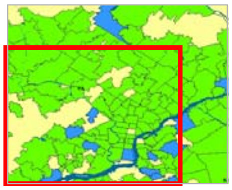
Step 1: Gridded



Check a specific recursive overlapping set of regions called "Gridded Regions"

New Approach: Gridded then Exhaustive

Step 1: Gridded



Check a specific recursive overlapping set of regions called "Gridded Regions"

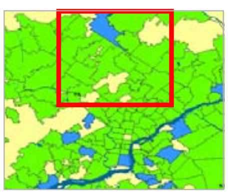
New Approach: Gridded then Exhaustive

Step 1: Gridded

Check a specific recursive overlapping set of regions called "Gridded Regions"

New Approach: Gridded then Exhaustive

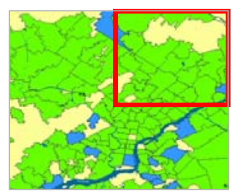
Step 1: Gridded



Check a specific recursive overlapping set of regions called "Gridded Regions"

New Approach: Gridded then Exhaustive

Step 1: Gridded



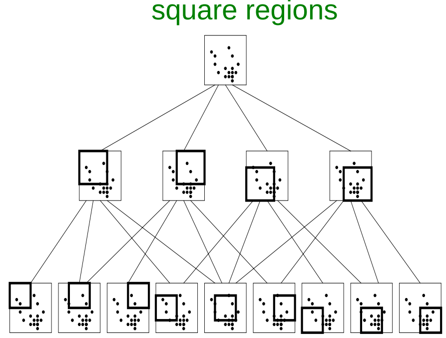
Check a specific recursive overlapping set of regions called "Gridded Regions"

New Approach: Gridded then Exhaustive

Step 1: Gridded

Check a specific recursive overlapping set of regions called "Gridded Regions"

The multi-resolution tree for square regions



New Approach: Gridded then Exhaustive

Step 1: Gridded

Check a specific recursive overlapping set of regions called "Gridded Regions"

Step 2: Exhaustive

Consider the set of subregions of a Gridded Region.

New Approach: Gridded then Exhaustive

A subregion of me could be one of five types...


...entirely inside my top left gridded child


...entirely inside my top right gridded child


...entirely inside my bottom left gridded child


...entirely in my bottom right gridded child


...not entirely inside any of my 4 gridded children

Step 2: Exhaustive

Consider the set of subregions of a Gridded Region.

New Approach: Gridded then Exhaustive

A subregion of me could be one of five types...


...entirely inside my top left gridded child


...entirely inside my top right gridded child


...entirely inside my bottom left gridded child


...entirely in my bottom right gridded child


...not entirely inside any of my 4 gridded children

Step 2: Exhaustive

Consider the set of subregions of a Gridded Region.

FACT: Any subregion of this type must be big...

...and we can put fairly tight bounds on how well any region of this type can score

New Approach: Gridded then Exhaustive

A subregion of me could be one of five types...


...entirely inside my top left gridded child


...entirely inside my top right gridded chil


...entirely inside my bottom left gridded c


...entirely in my botto right gridded child


...not entirely inside any of my 4 gridded children

Procedure: Exhaust(Gridded Region)

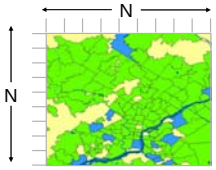
1. Exhaust(Region.TopLeft)
2. Exhaust(Region.TopRight)
3. Exhaust(Region.BottomLeft)
4. Exhaust(Region.BottomRight)
5. Is it possible that any "Type 5" subregion of "Gridded Region" could score better than best known score to date?

NO: Quit Procedure!

YES: Check all "Type 5" Subregions

region of this type can score

Fast squares speedup



- Theoretical complexity of fast squares: $O(N^2)$ (as opposed to naïve N^3), if maximum density region sufficiently dense.
  If not, we can use several other speedup tricks.
- In practice: 10-200x speedups on real and artificially generated datasets.
  Emergency Dept. dataset (600K records): 20 minutes, versus 66 hours with naïve approach.