New Cached-Sufficient Statistics Algorithms for quickly answering statistical questions

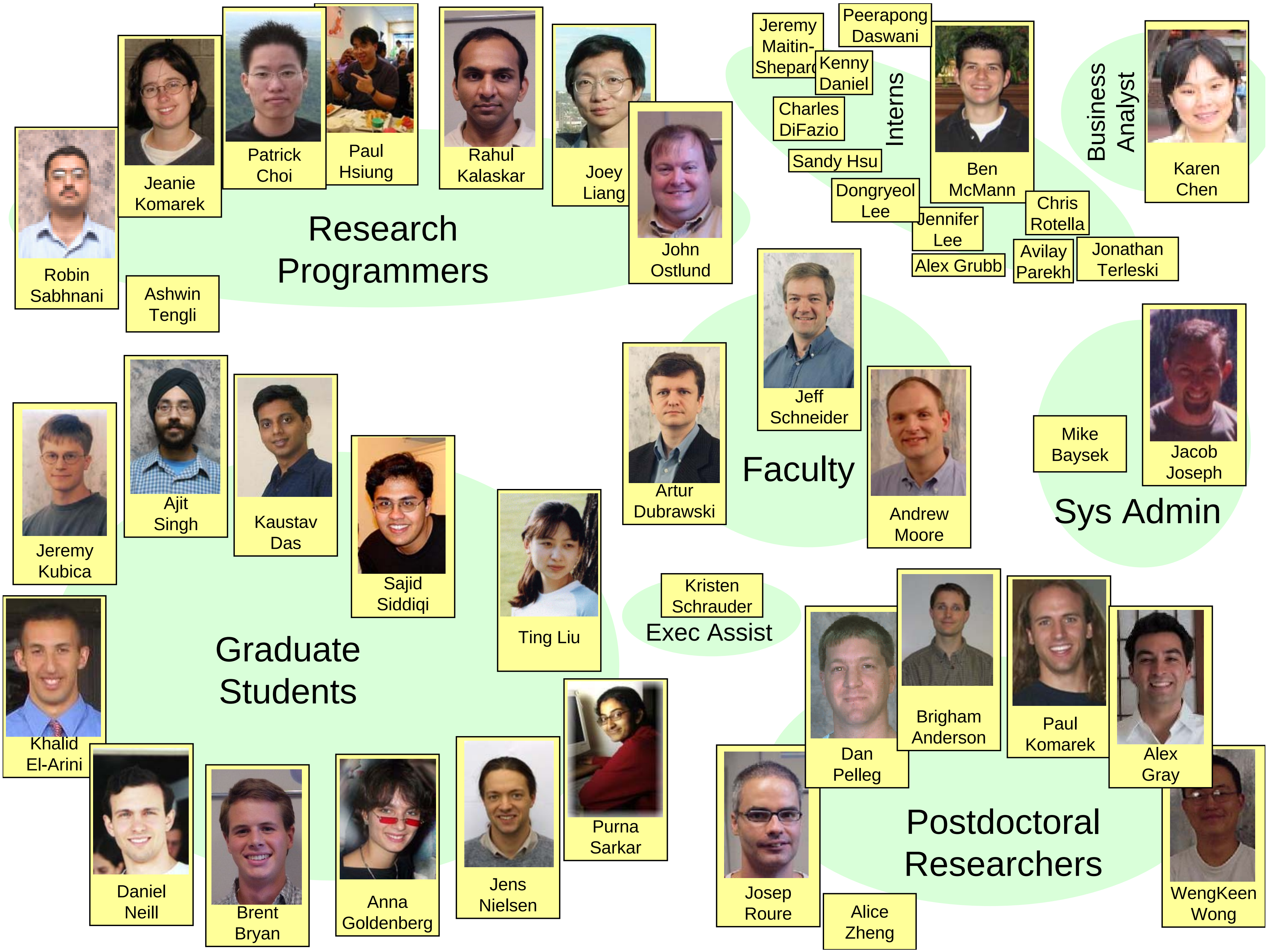
Jeremy Kubica
jkubica@google.com
Google Pittsburgh

Andrew Moore
awm@google.com
Google Pittsburgh

Daniel Neill
neill+@cs.cmu.edu
Auton Lab, CMU

Papers, Software, Example Datasets, Tutorials: www.autonlab.org

This is a condensed version of the invited talk at KDD 2006 in Philadelphia



Research Programmers
Robin Sabhnani
Jeanie Komarek
Patrick Choi
Paul Hsiung
Rahul Kalaskar
Joey Liang
John Ostlund
Ashwin Tengli
Interns
Jeremy Maitin-Shepard
Peerapong Daswani
Kenny Daniel
Charles DiFazio
Sandy Hsu
Dongryeol Lee
Jennifer Lee
Alex Grubb
Ben McMann
Chris Rotella
Avilay Parekh
Jonathan Terleski
Business Analyst
Karen Chen
Faculty
Jeff Schneider
Artur Dubrawski
Andrew Moore
Sys Admin
Mike Baysek
Jacob Joseph
Graduate Students
Jeremy Kubica
Ajit Singh
Kaustav Das
Sajid Siddiqi
Ting Liu
Khalid El-Arini
Daniel Neill
Brent Bryan
Anna Goldenberg
Jens Nielsen
Purna Sarkar
Exec Assist
Kristen Schrauder
Postdoctoral Researchers
Dan Pelleg
Brigham Anderson
Paul Komarek
Alex Gray
Josep Roure
Alice Zheng
WengKeen Wong

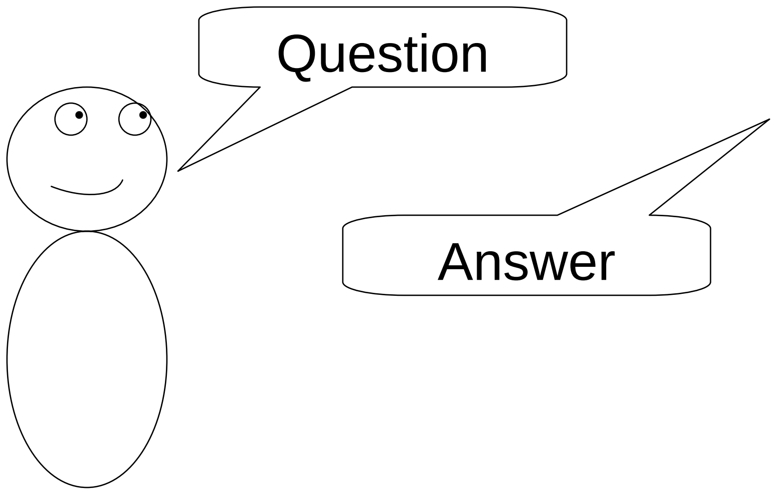
Outline

- Cached Sufficient Statistics
- New searches over cached statistics

- Biosurveillance and Epidemiology
- Scan Statistics
- Cached Scan Statistics
- Branch-and-Bound Scan Statistics
- Retail data monitoring
- Brain monitoring
- Entering Google

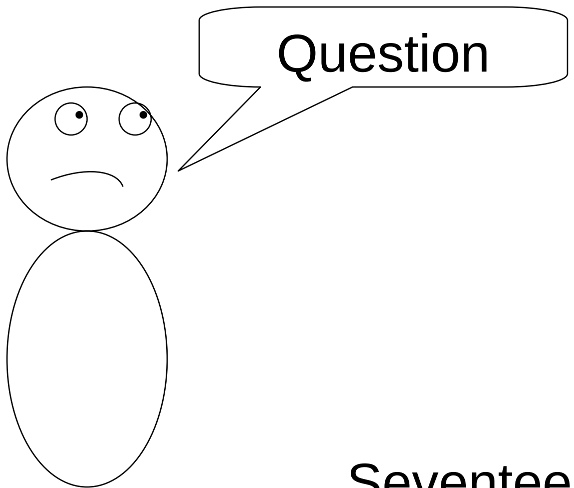
- Asteroids
- Multi (and I mean multi) object target tracking
- Multiple-tree search
- Entering Google

Data Analysis: The old days

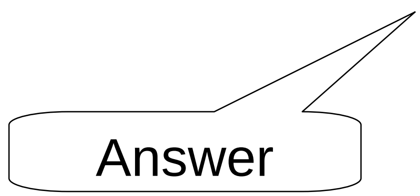


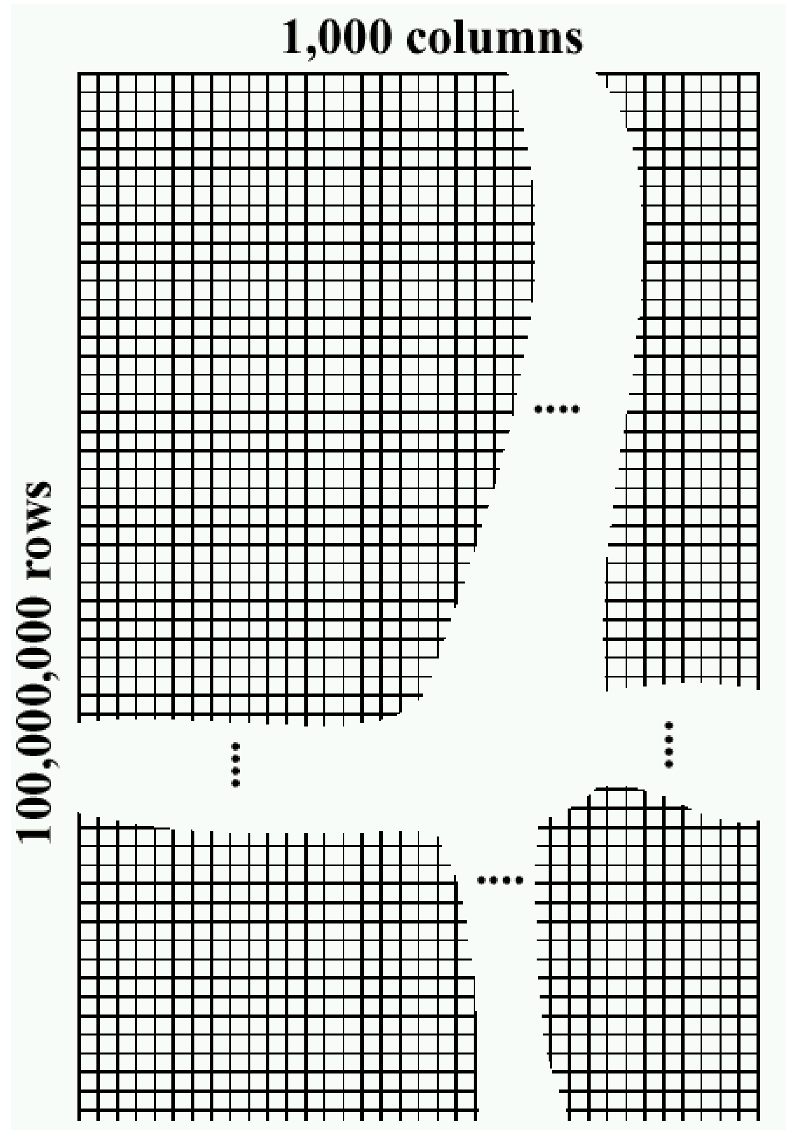
Size	Ellipticity	Color
23	0.96	Red
33	0.55	Red
36		Green
40		
20		
48		

Data Analysis: The new days

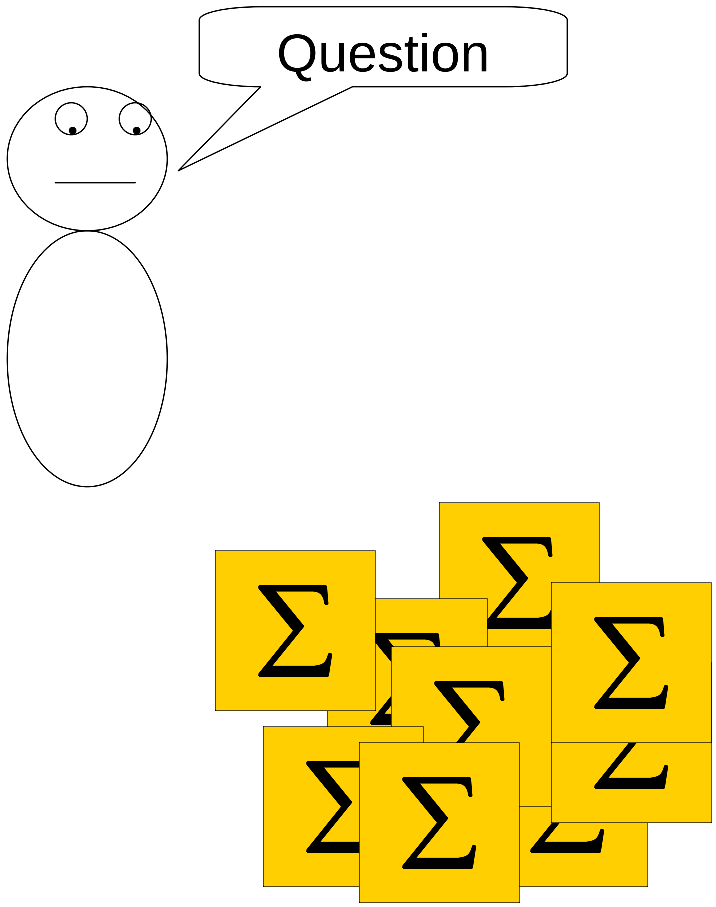


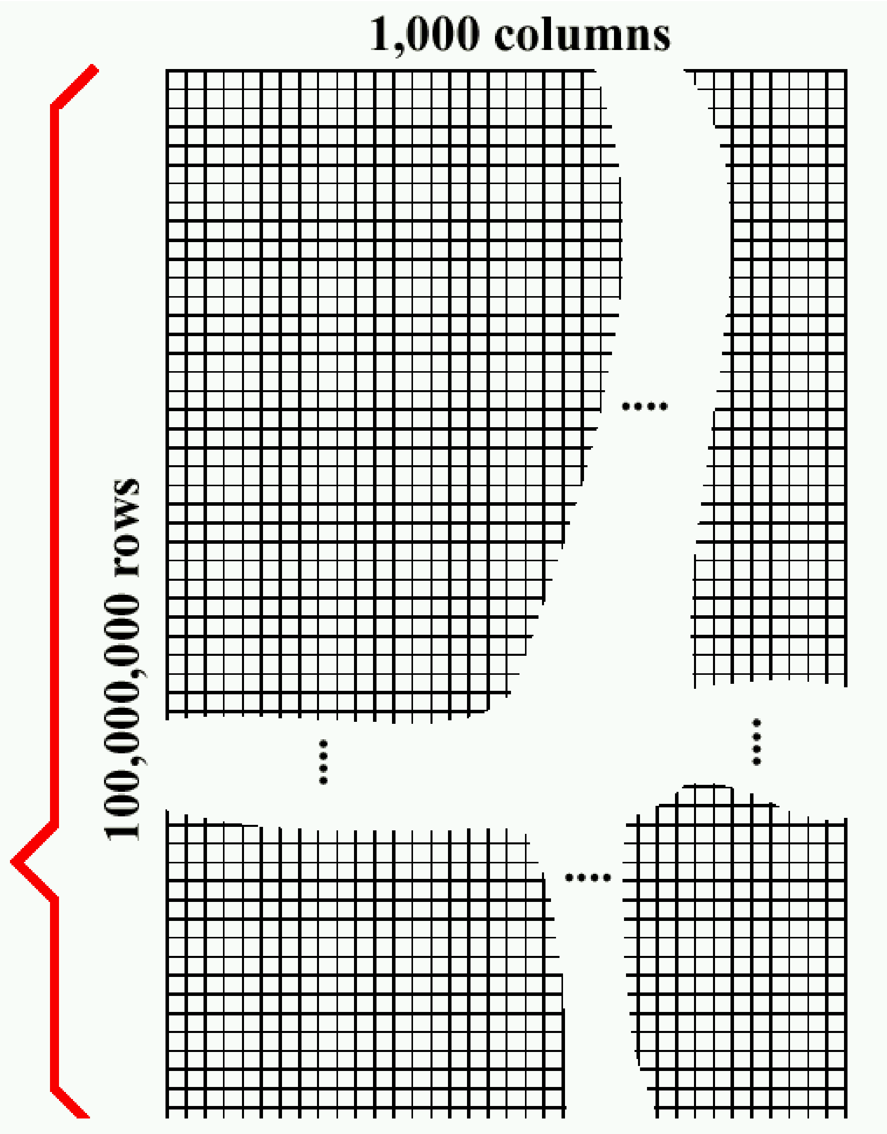
Seventeen
months later…



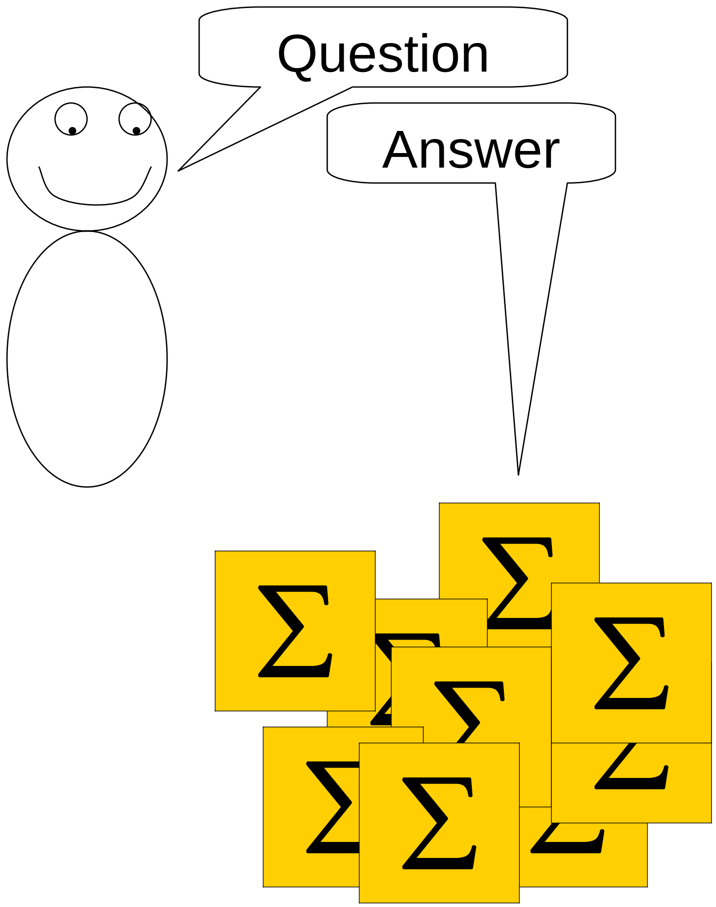


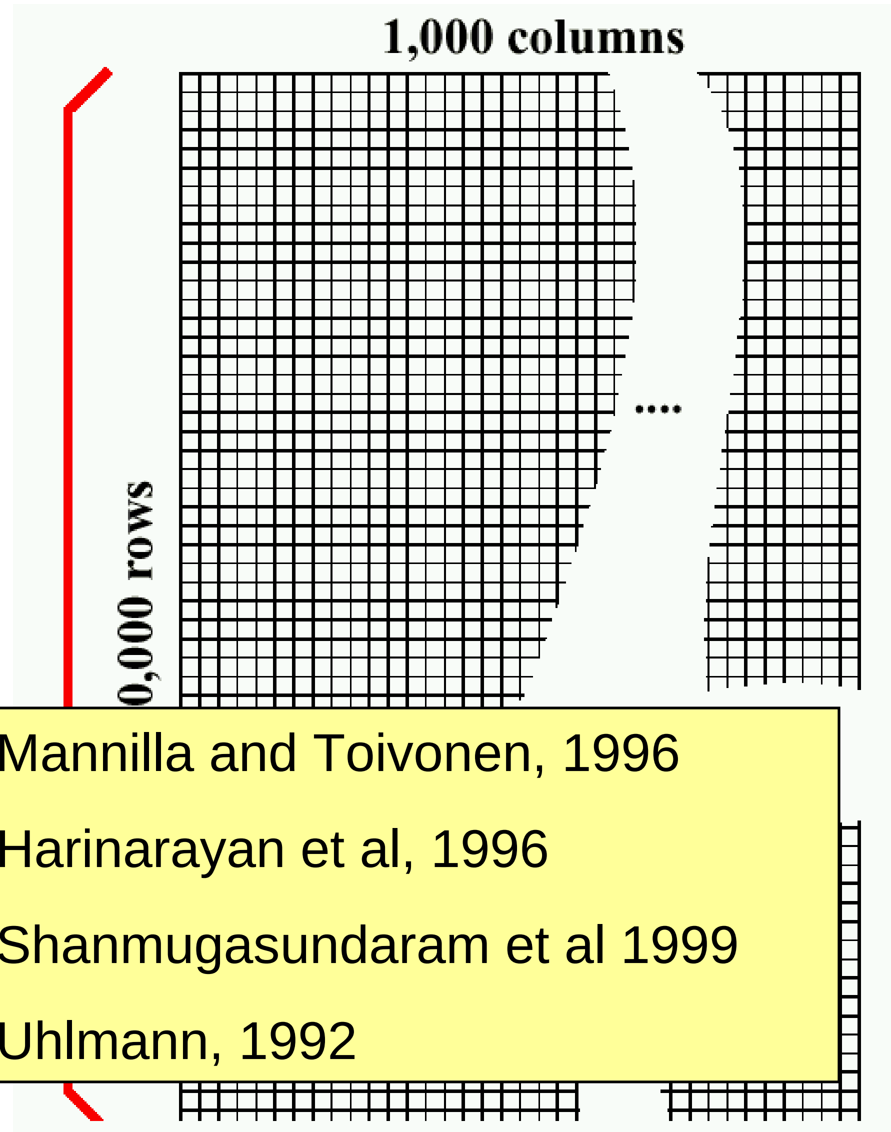
Cached Sufficient Statistics





Cached Sufficient Statistics





Mannilla and Toivonen, 1996

Harinarayan et al, 1996

Shanmugasundaram et al 1999

Uhlmann, 1992

Frequent Sets (Agrawal et al)

KD-trees (Friedman, Bentley, Finkel)

Multi-resolution KD-trees (Deng, Moore)

All-Dimensions Trees (Moore, Lee)

Multi-resolution metric trees (Liu, Moore)

Well-Separated Pairwise Decomposition (Callahan 1995)

TimeCube (Sabhnani, Moore)

Outline

Cached Sufficient Statistics

New searches over cached statistics

Biosurveillance and Epidemiology

Scan Statistics

Cached Scan Statistics

Branch-and-Bound Scan Statistics

Retail data monitoring

Brain monitoring

Entering Google

Asteroids

Multi (and I mean multi) object target tracking

Multiple-tree search

Entering Google

Outline

- Cached Sufficient Statistics
- ▶ New searches over cached statistics

Roberto Bayardo
Geoff Webb
Martin Kulldorf
Pregibon and DuMouchel

- Biosurveillance and ... demiology
- Scan Statistics
- Cached Scan S...
- Branch-and-Bou... Statistics
- Retail data monitoring
- Brain monitoring
- Entering Google

- Asteroids
- Multi (and I mean multi) object target tracking
- Multiple-tree search
- Entering Google

Outline

Cached Sufficient Statistics

New searches over cached statistics

Biosurveillance and Epidemiology

Scan Statistics

Cached Scan Statistics

Branch-and-Bound Scan Statistics

Retail data monitoring

Brain monitoring

Entering Google

Asteroids

Multi (and I mean multi) object target tracking

Multiple-tree search

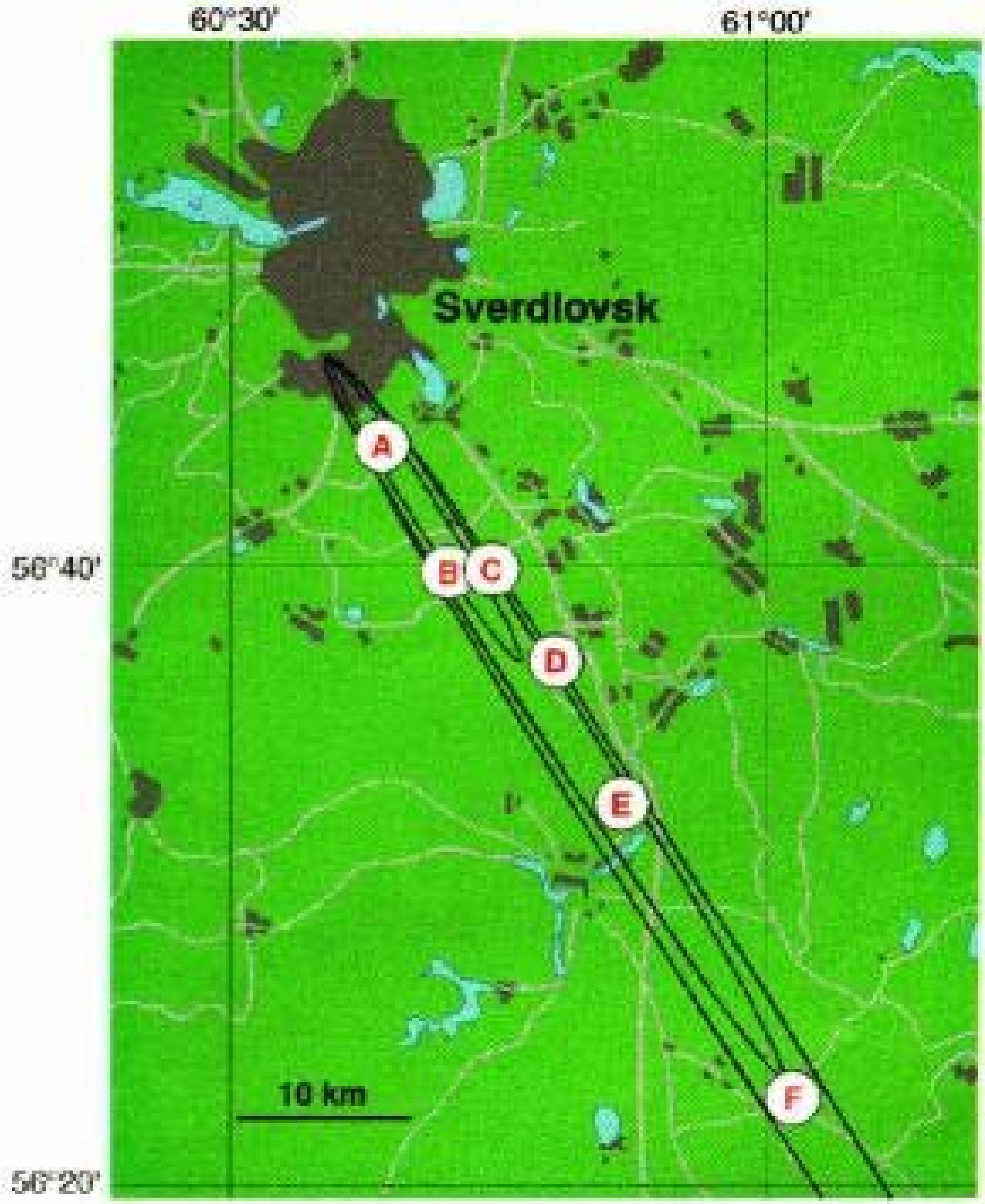
Entering Google

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

Sverdlovsk

collaboration with Daniel Neill <neill@cs.cmu.edu>

Sverdlovsk Region: Epi-map



Biosurveillance Algorithms

Biosurveillance Algorithms

Specific Detectors

CityDiagnosis (DBN-based surveillance): [Anderson, Moore]

EPFC: Emerging Patterns from food complaints: [Dubrawski, Sabhnani, Moore]

PANDA2: Patient-based Bayesian Network [Cooper, Levander et. al]

BARD: Airborne Attack Detection [Hogan, Cooper et al.]

General Detectors

What's Strange about Recent Events [Wong, Moore, Wagner and Cooper]

Fast Scan Statistic [Neill, Moore]

Fast Scan for Oriented Regions [Neill, Moore et al.]

Historical Model Scan Statistic [Hogan, Moore, Neill, Tsui, Wagner]

Bayesian Network Spatial Scan [Neill, Moore, Schneider, Cooper Wagner, Wong]

Biosurveillance Algorithms

Specific Detectors

PANDA2: Patient-based Bayesian Network
[Cooper, Levander et. al]

BARD: Airborne Attack Detection
[Hogan, Cooper]

General Detectors

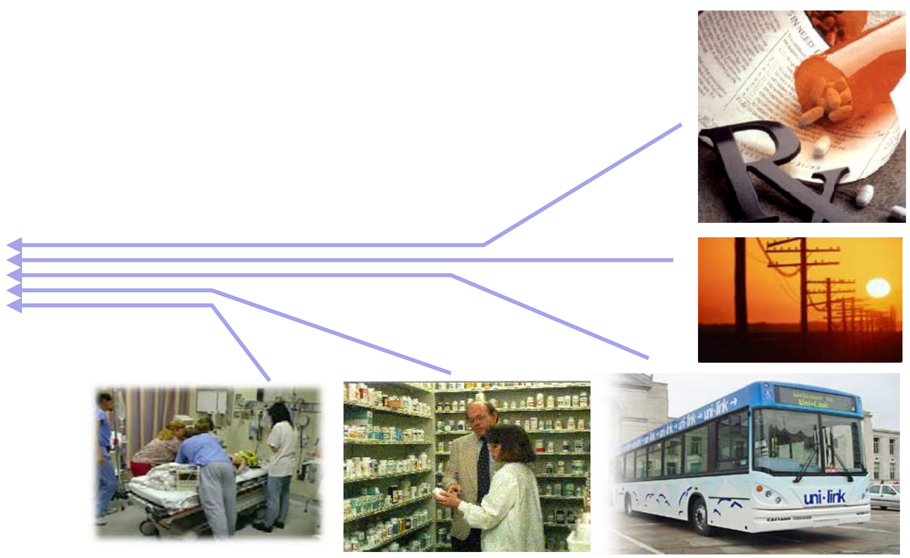
What's Strange about Recent Events

Fast Scan Statistic
[Neill, Moore]

Fast Scan for Oriented Regions
[Neill, Moore et al.]

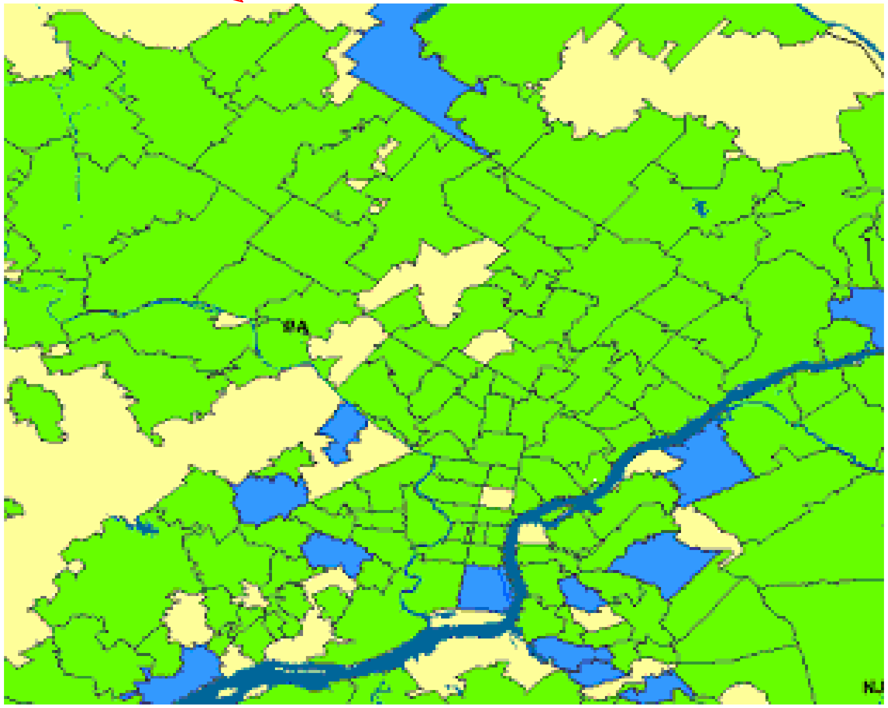
Historical Model Scan Statistic
[Hogan, Moore, Neill, Tsui, Wagner]

Bayesian Network Spatial Scan
[Neill, Moore, Schneider, Cooper Wagner, Wong]



One Step of Spatial Scan

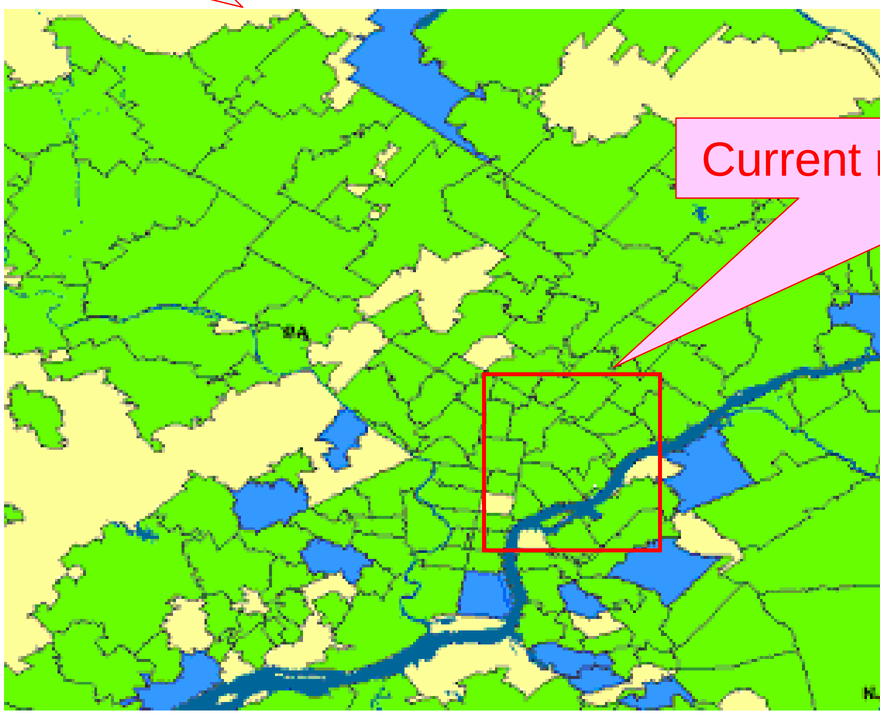
Entire area being scanned



collaboration with Daniel Neill <neill@cs.cmu.edu>

One Step of Spatial Scan

Entire area being scanned

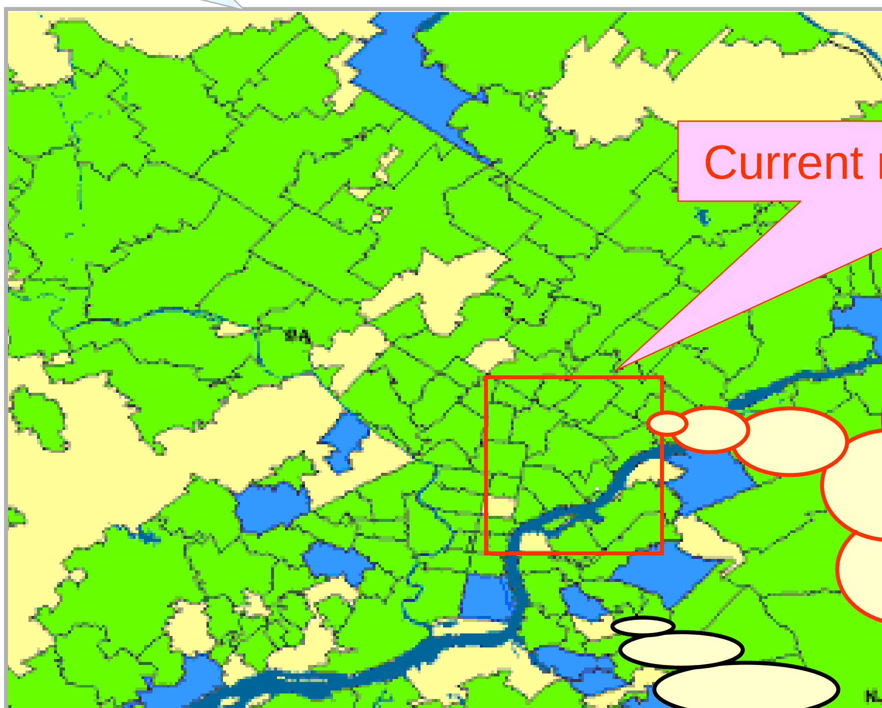


Current region being considered

collaboration with Daniel Neill <neill@cs.cmu.edu>

One Step of Spatial Scan

Entire area being scanned



Current region being considered

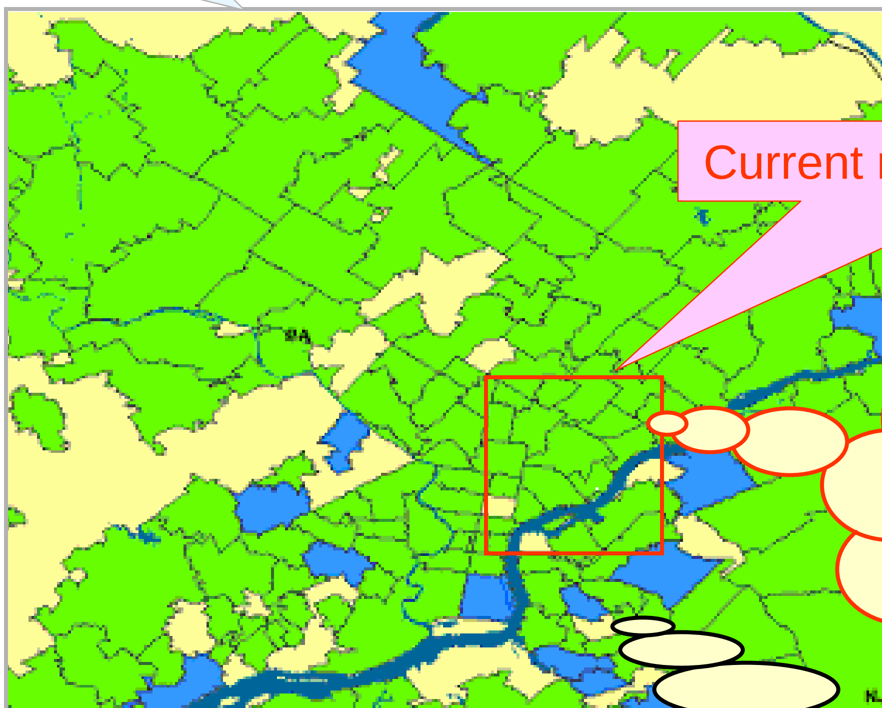
I have a population of 5300 of whom 53 are sick (1%)

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

collaboration with Daniel Neill <neill@cs.cmu.edu>

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%)

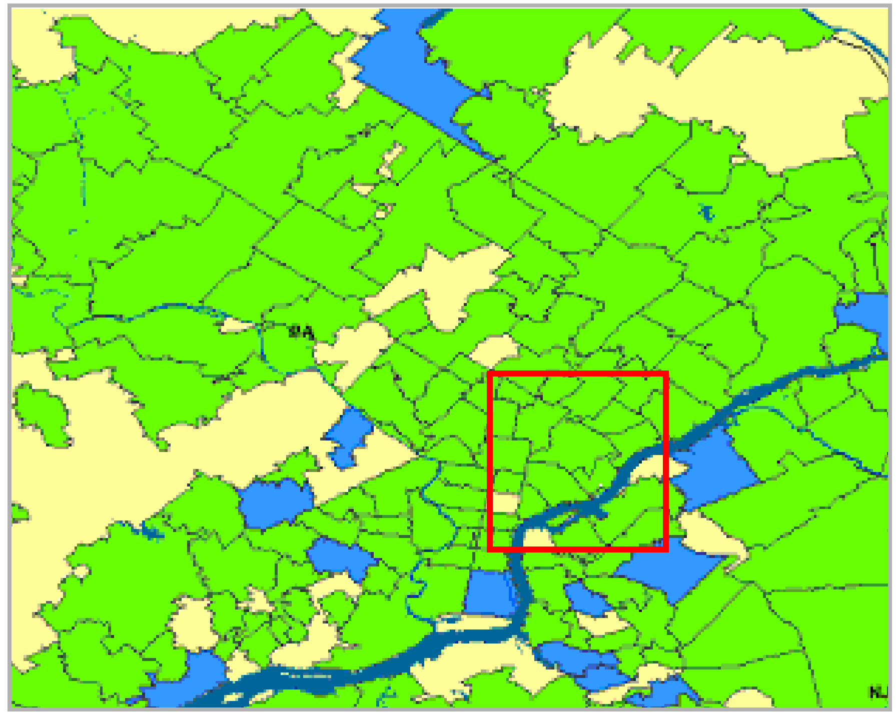
Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

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

collaboration with Daniel Neill <neill@cs.cmu.edu>

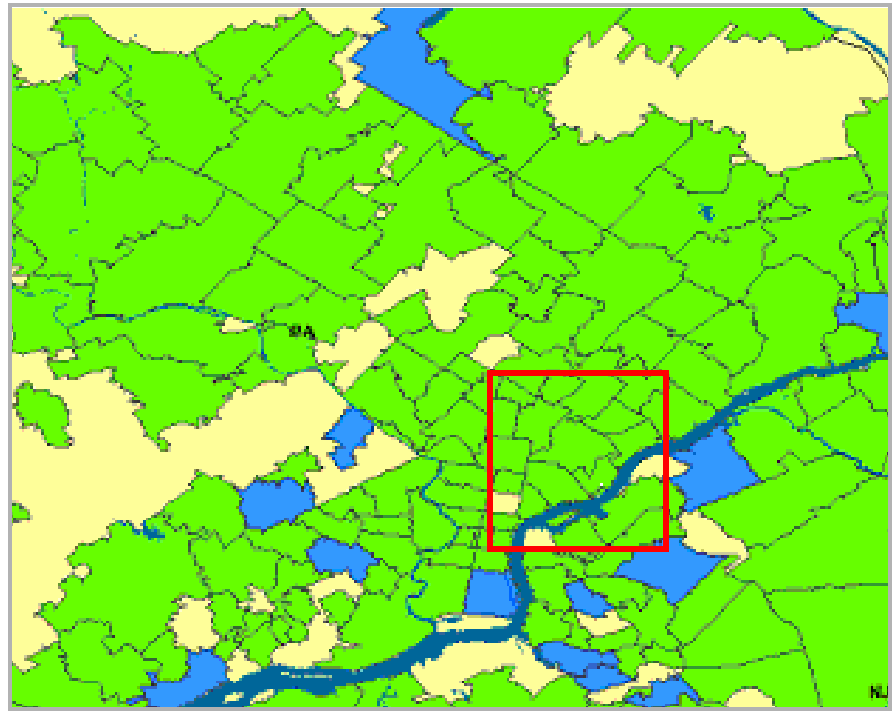
Scoring functions

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



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

Scoring functions

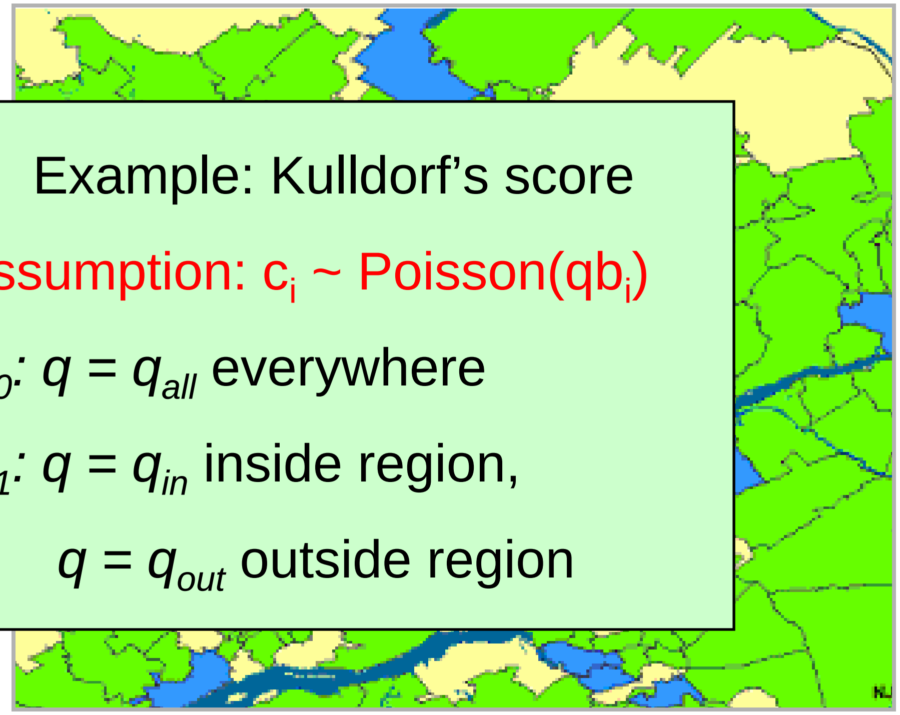


- Define <u>models</u>:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a <u>score function</u> *Score(S) = Score(C, B).*
  - Likelihood ratio: $$\text{Score}(S) = \frac{L(\text{Data} \mid H_1(S))}{L(\text{Data} \mid H_0)}$$
  - To find the most significant region: $$S^* = \arg\max_{S} \text{Score}(S)$$

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

Scoring functions

- Define models:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a score function *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: $$S^* = \arg\max_S \text{Score}(S)$$



Example: Kulldorf's score

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

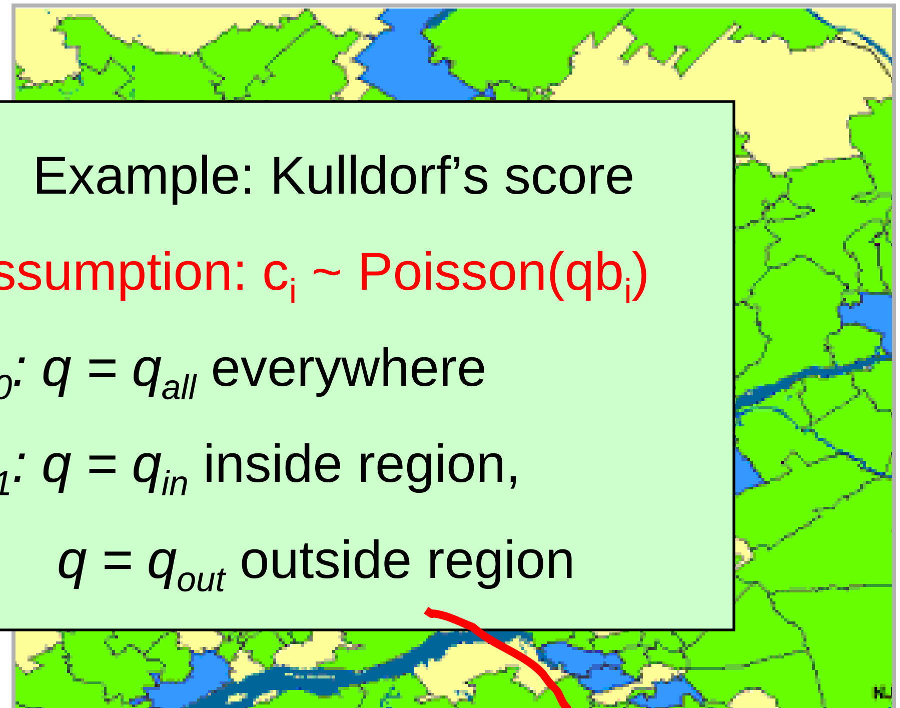
H_0: $q = q_{all}$ everywhere

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

$q = q_{out}$ outside region

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

Scoring functions



- Define models:
  - of the null hypothesis H_0: no attacks.
  - of the alternative hypotheses $H_1(S)$: attack in region S.
- Derive a score function *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: $$S^* = \arg\max_S \text{Score}(S)$$

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,

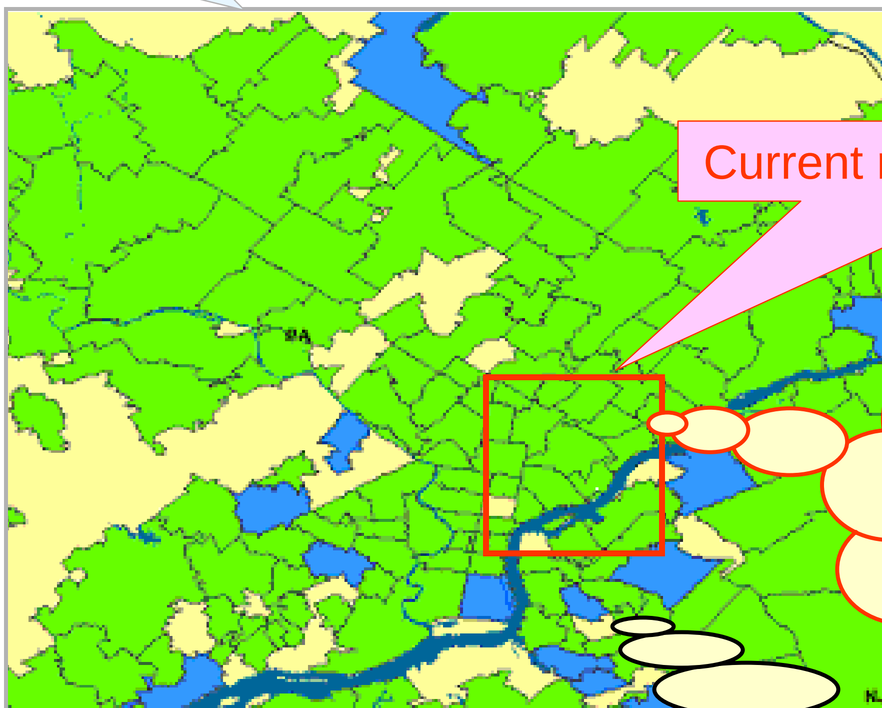
$q = q_{out}$ outside region

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

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

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%)

[Score = 1.4]

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

So... *is that a big deal? Evaluated with Score function (e.g. Kulldorf's score)*

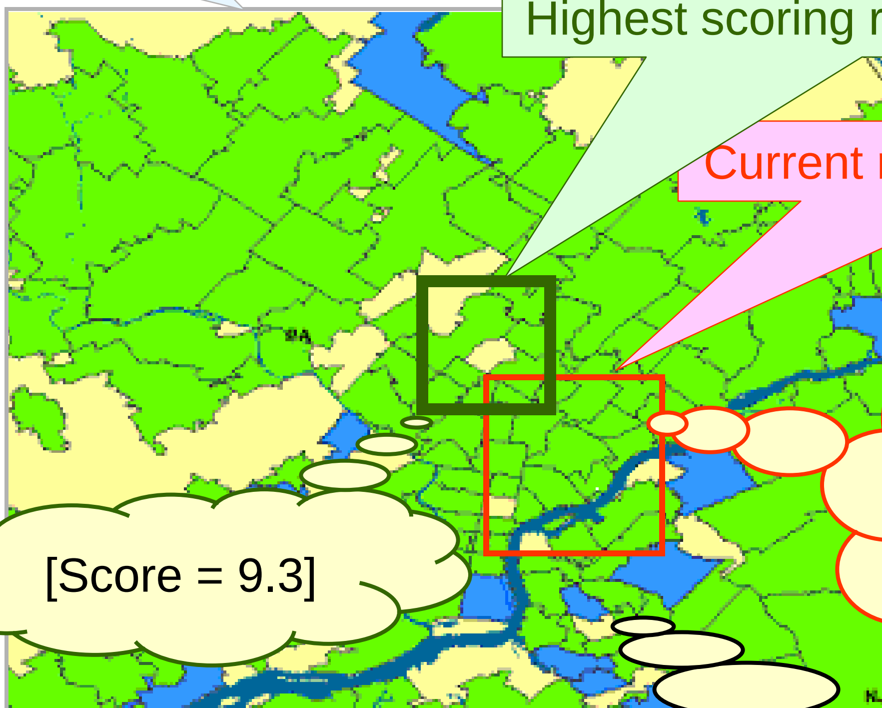
collaboration with Daniel Neill <neill@cs.cmu.edu>

Many Steps of Spatial Scan

Entire area being scanned

Highest scoring region in search so far

Current region being considered



[Score = 9.3]

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

[Score = 1.4]

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

So… *is that a big deal?*

Evaluated with Score function (e.g. Kulldorf's score)

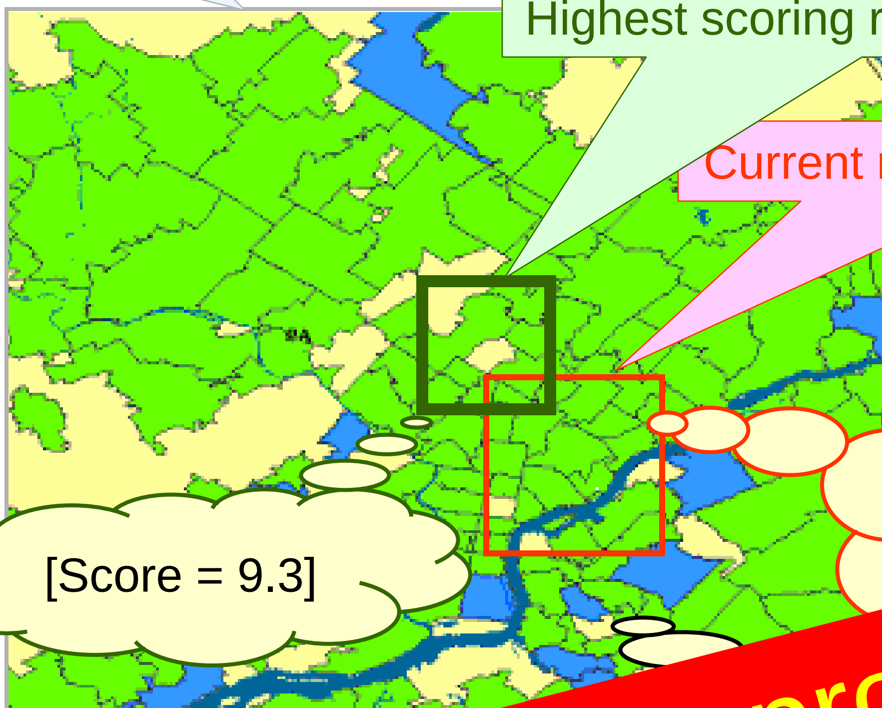
collaboration with Daniel Neill <neill@cs.cmu.edu>

Many Steps of Spatial Scan

Entire area being scanned

Highest scoring region in search so far

Current region being considered



[Score = 9.3]

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

[S

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

So… *is that a big deal? Evaluated with Score function (e.g. Kulldorf's score)*

So… a kd-tree problem, right?

collaboration with Daniel Neill <neill@cs.cmu.edu>

Computational framework

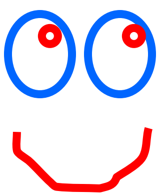
Data is aggregated to a grid.

B=25 C=27	B=18 C=14	B=22 C=22	B=14 C=15	B=5 C=5
B=25 C=26	B=20 C=17	B=6 C=9	B=20 C=12	B=5 C=4
B=25 C=19	B=25 C=26	B=20 C=43	B=15 C=37	B=20 C=20
B=24 C=18	B=24 C=20	B=19 C=40	B=15 C=32	B=19 C=16
B=23 C=20	B=15 C=17	B=14 C=8	B=10 C=10	B=2 C=3

collaboration with Daniel Neill <neill@cs.cmu.edu>

Computational framework

Data is aggregated to a grid.



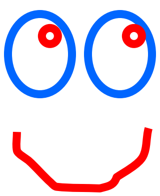
Cost of obtaining sufficient statistics for an arbitrary rectangle: O(1)

B=25 C=27	B=18 C=14	B=22 C=22	B=14 C=15	B=5 C=5
B=25 C=26	B=20 C=17	B=6 C=9	B=20 C=12	B=5 C=4
B=25 C=19	B=25 C=26	B=20 C=43	B=15 C=37	B=20 C=20
B=24 C=18	B=24 C=20	B=19 C=40	B=15 C=32	B=19 C=16
B=23 C=20	B=15 C=17	B=14 C=8	B=10 C=10	B=2 C=3

collaboration with Daniel Neill <neill@cs.cmu.edu>

Computational framework

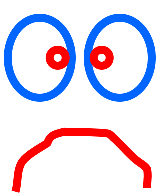
Data is aggregated to a grid.



Cost of obtaining sufficient statistics for an arbitrary rectangle: O(1)

n x n grid has

$$\left[\binom{n+1}{2}\right]^2 = O(n^4)$$



rectangles to search

B=25 C=27	B=18 C=14	B=22 C=22	B=14 C=15	B=5 C=5
B=25 C=26	B=20 C=17	B=6 C=9	B=20 C=12	B=5 C=4
B=25 C=19	B=25 C=26	B=20 C=43	B=15 C=37	B=20 C=20
B=24 C=18	B=24 C=20	B=19 C=40	B=15 C=32	B=19 C=16
B=23 C=20	B=15 C=17	B=14 C=8	B=10 C=10	B=2 C=3

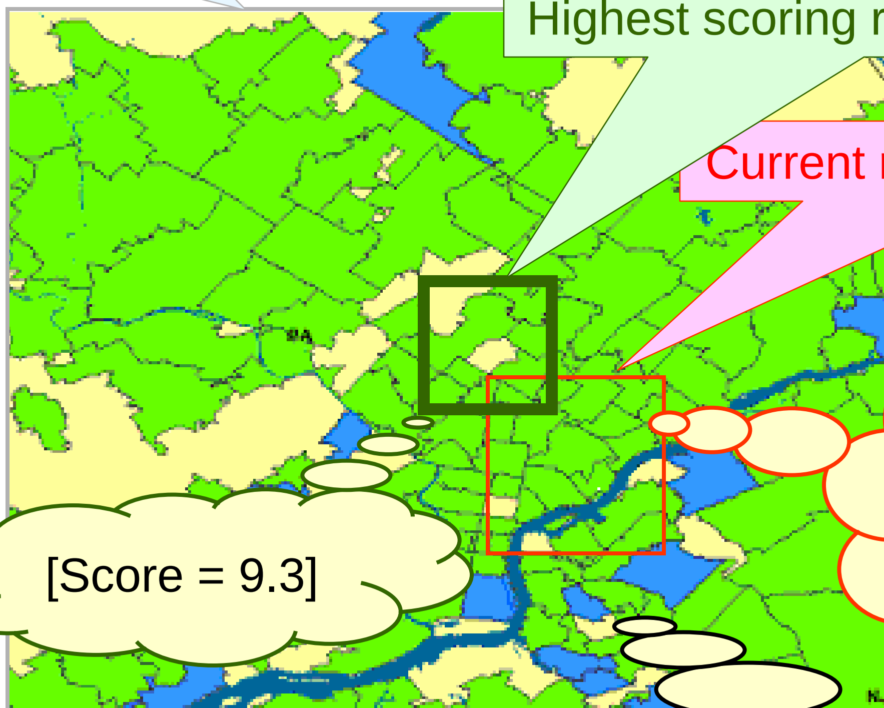
collaboration with Daniel Neill <neill@cs.cmu.edu>

Many Steps of Spatial Scan

Entire area being scanned

Highest scoring region in search so far

Current region being considered



[Score = 9.3]

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

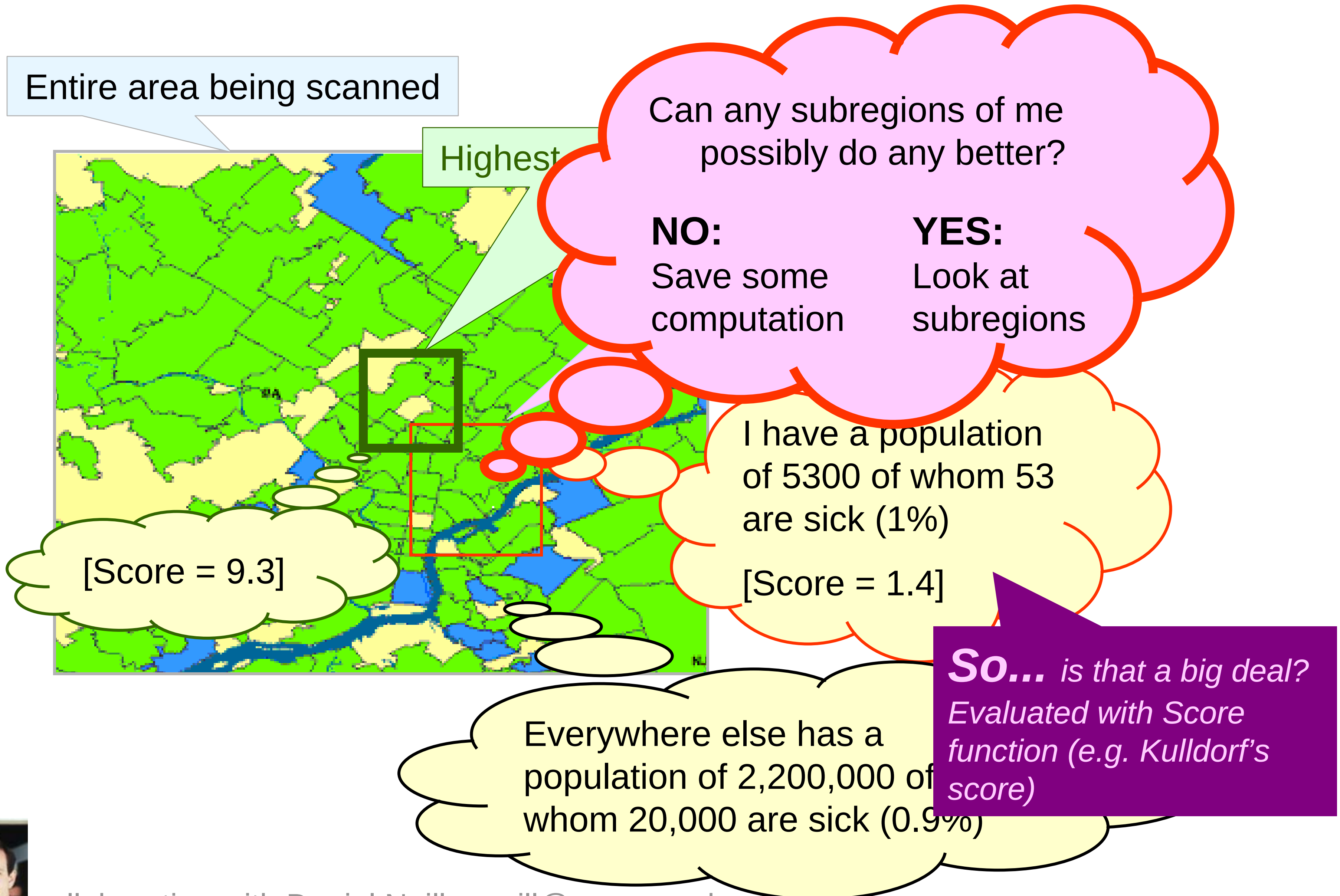
[Score = 1.4]

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

So… *is that a big deal?*
Evaluated with Score function (e.g. Kulldorf's score)

collaboration with Daniel Neill <neill@cs.cmu.edu>

Many Steps of Spatial Scan

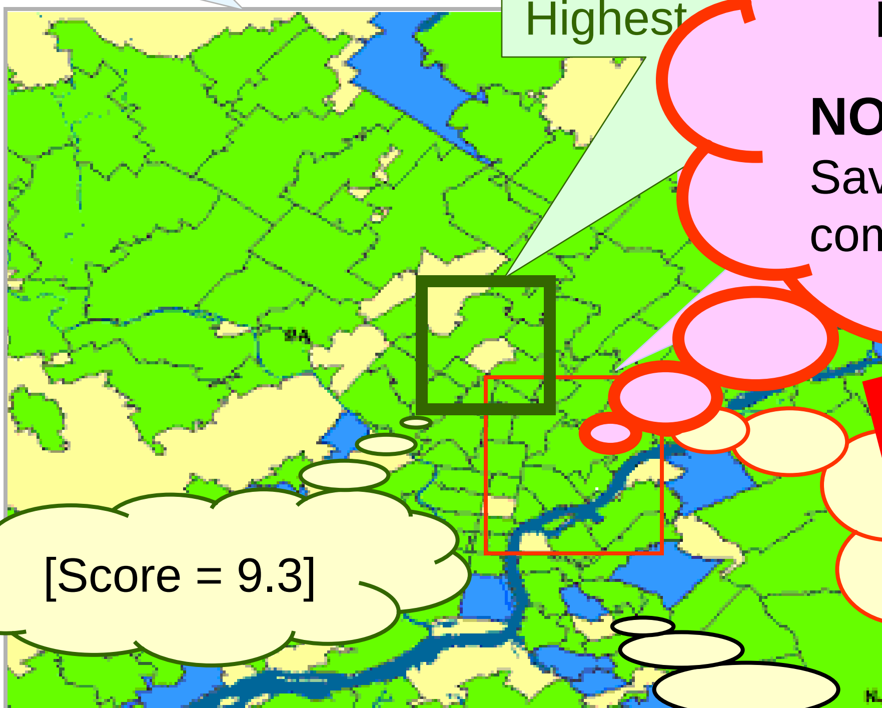


collaboration with Daniel Neill <neill@cs.cmu.edu>

Many Steps of Spatial Scan

Entire area being scanned

Highest



[Score = 9.3]

Can any subregions of me possibly do any better?

NO:
Save some computation

YES:
Look at subregions

Problem with this method:
Can only get a weak bound.
Not much computational benefit.

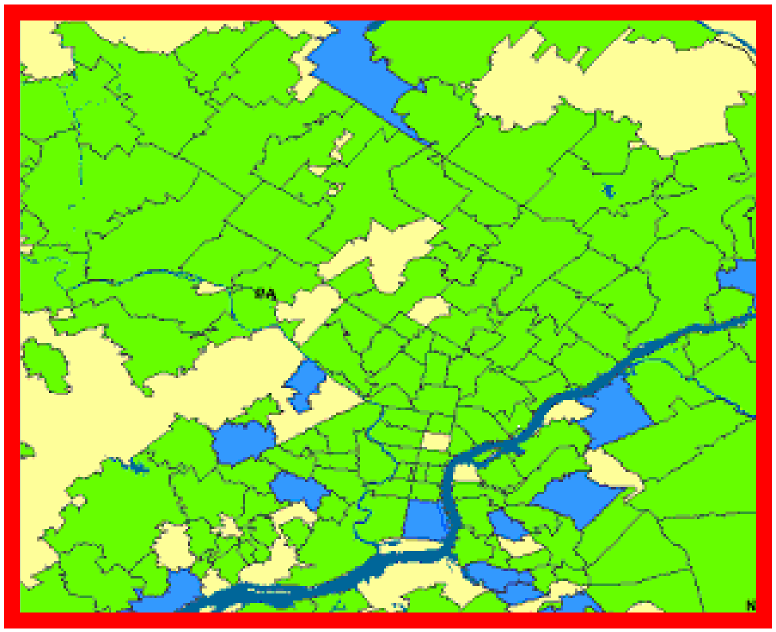
Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

So… *is that a big deal?*
Evaluated with Score function (e.g. Kulldorf's score)

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

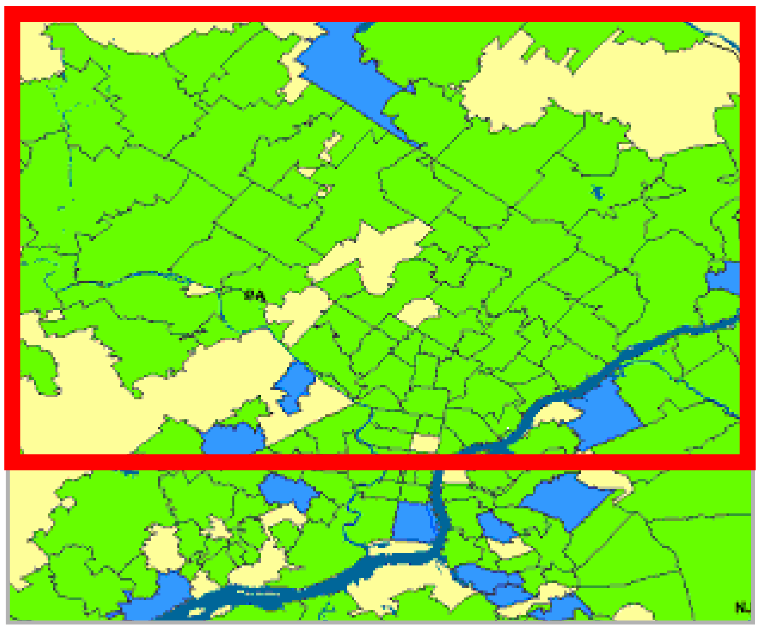


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

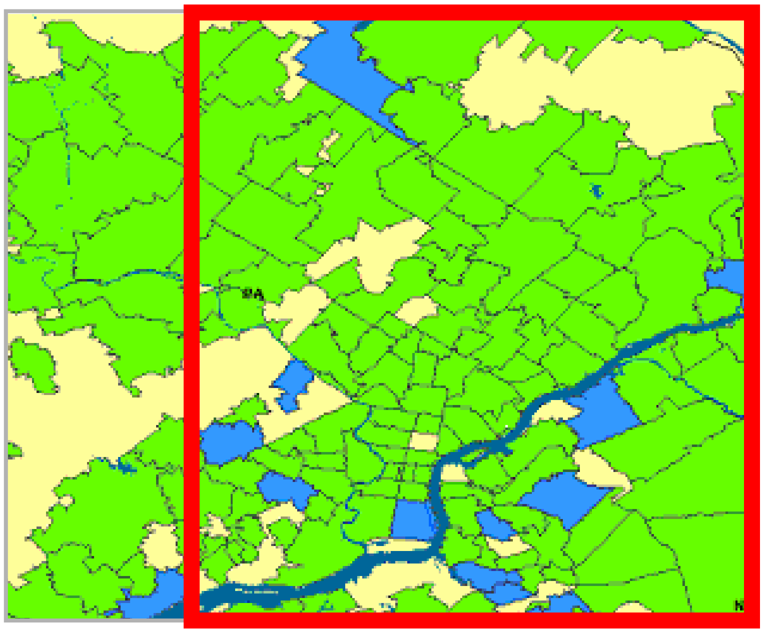


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

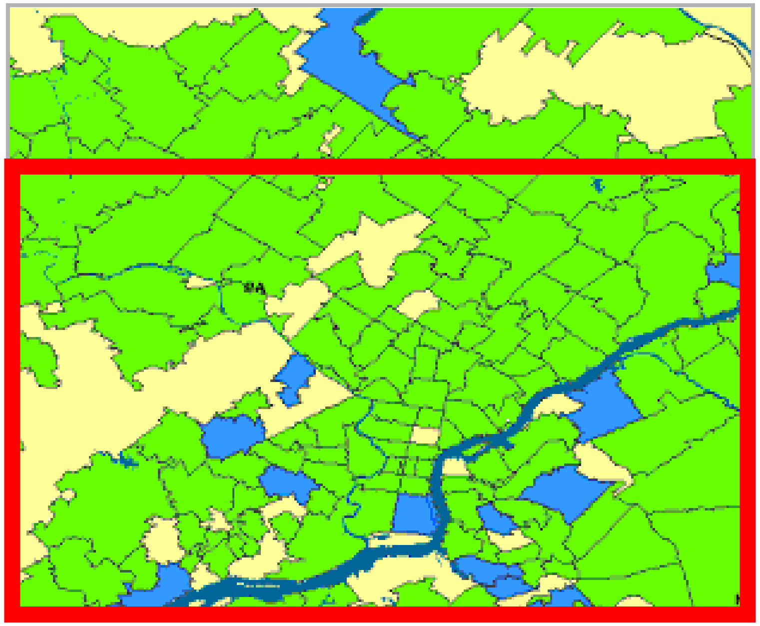


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

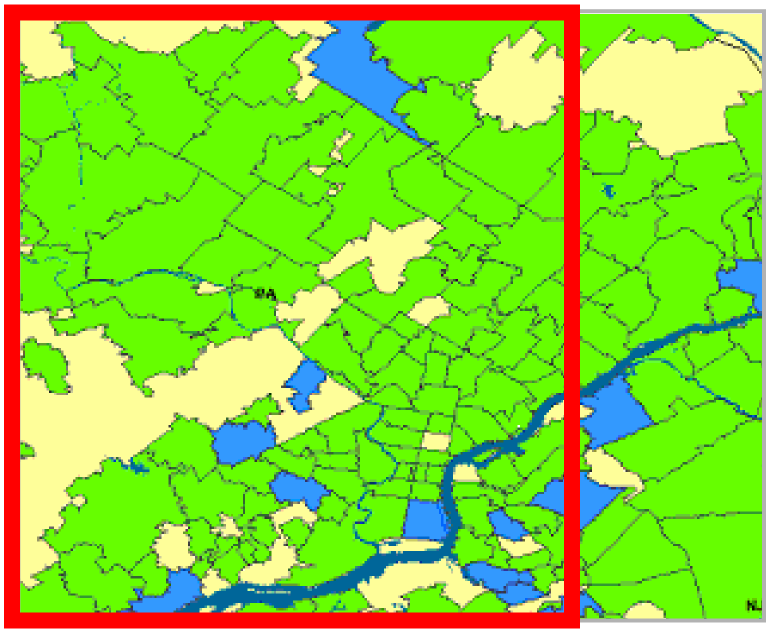
Step 1: Gridded



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

Gridded then Exhaustive

Step 1: Gridded

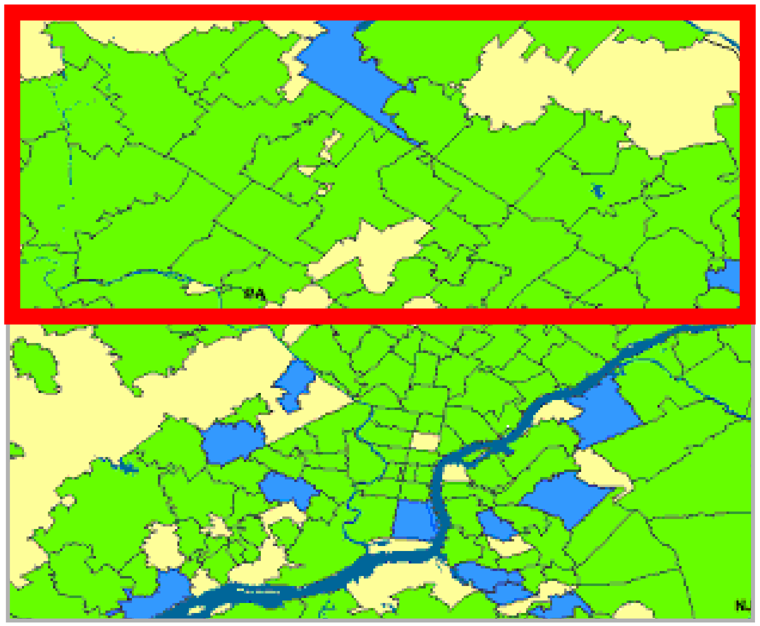


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

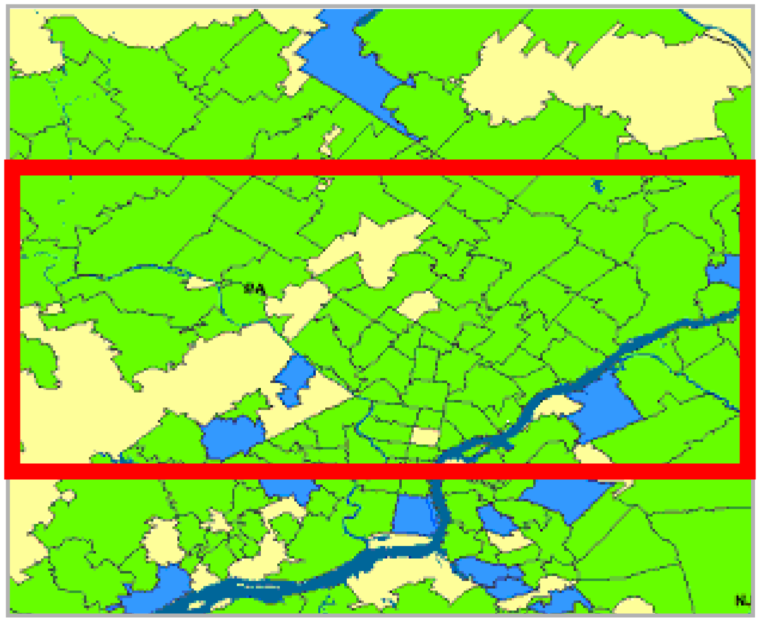


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

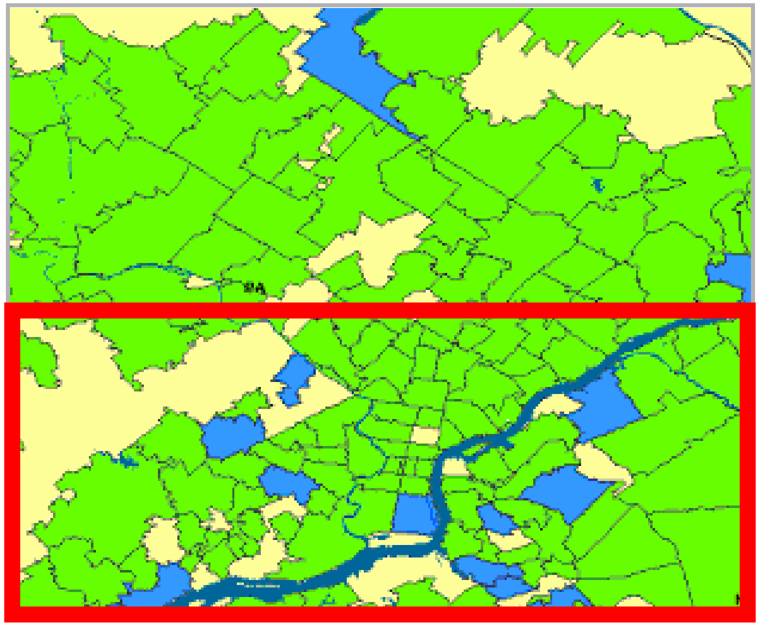


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

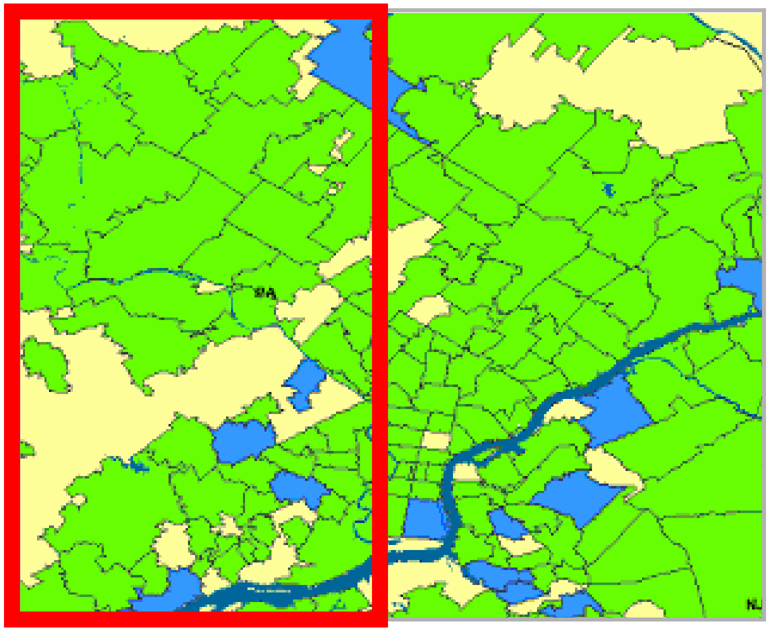
Step 1: Gridded



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

Gridded then Exhaustive

Step 1: Gridded

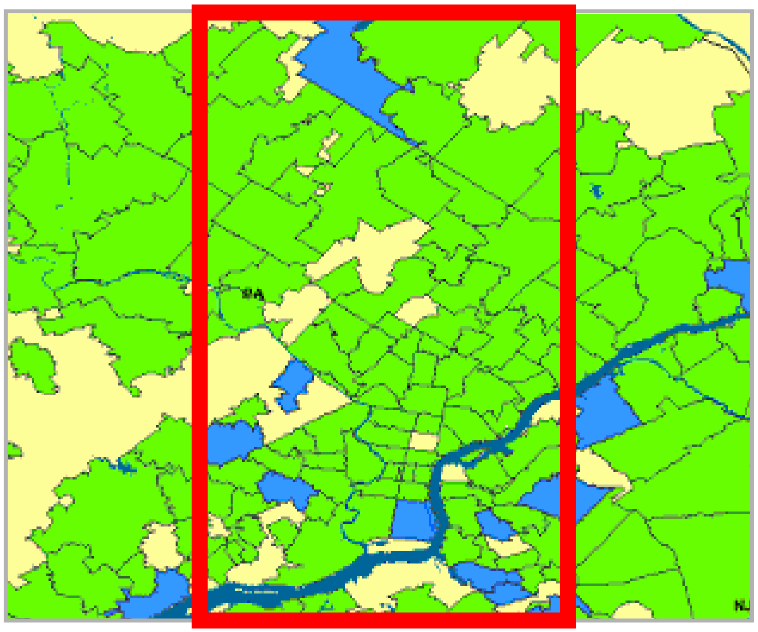


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

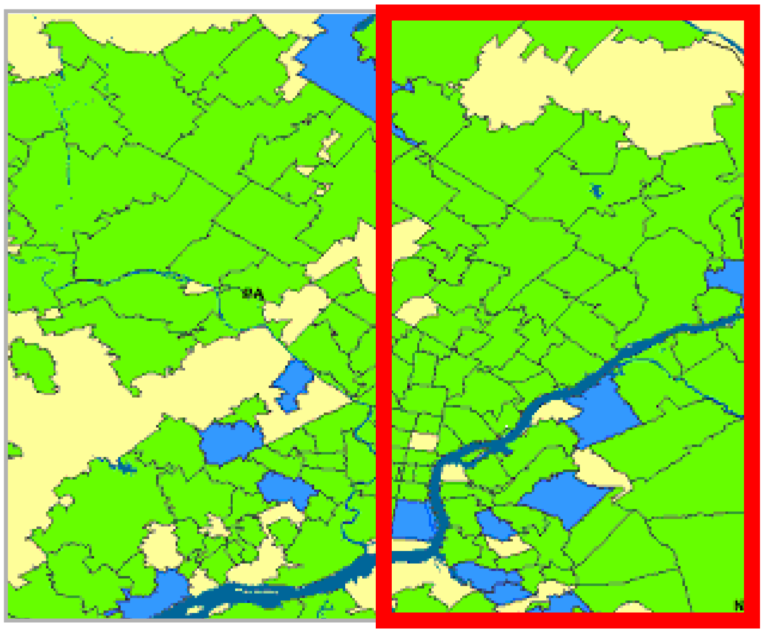


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

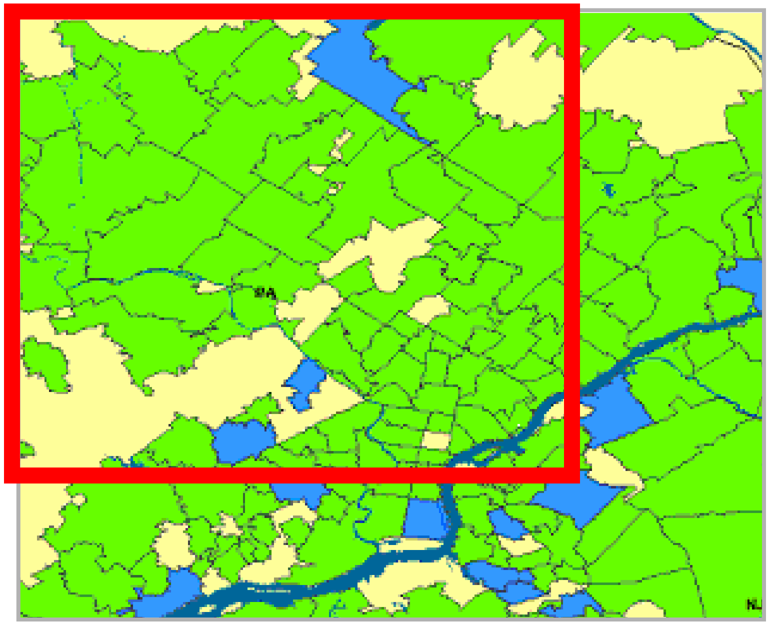


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

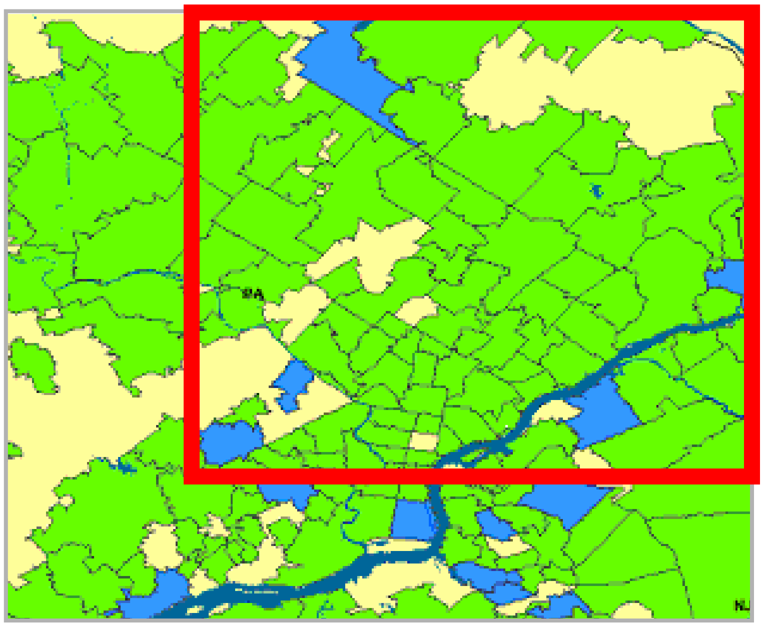
Step 1: Gridded



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

Gridded then Exhaustive

Step 1: Gridded

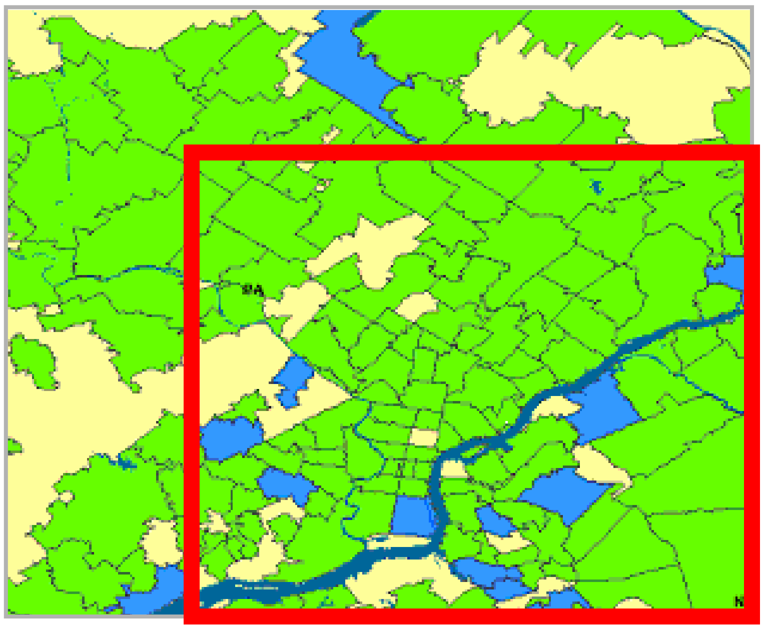


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

Step 1: Gridded

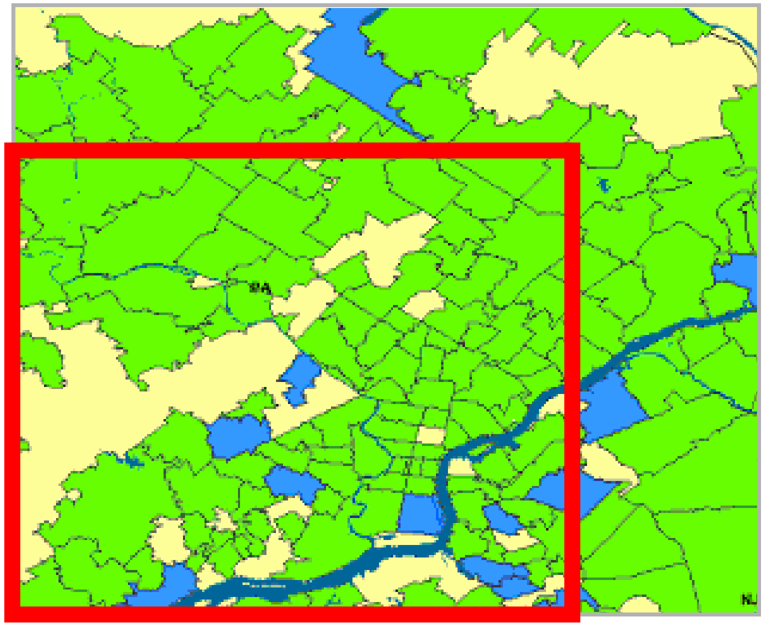


Check a specific recursive overlapping set of regions called "Gridded Regions"

collaboration with Daniel Neill <neill@cs.cmu.edu>

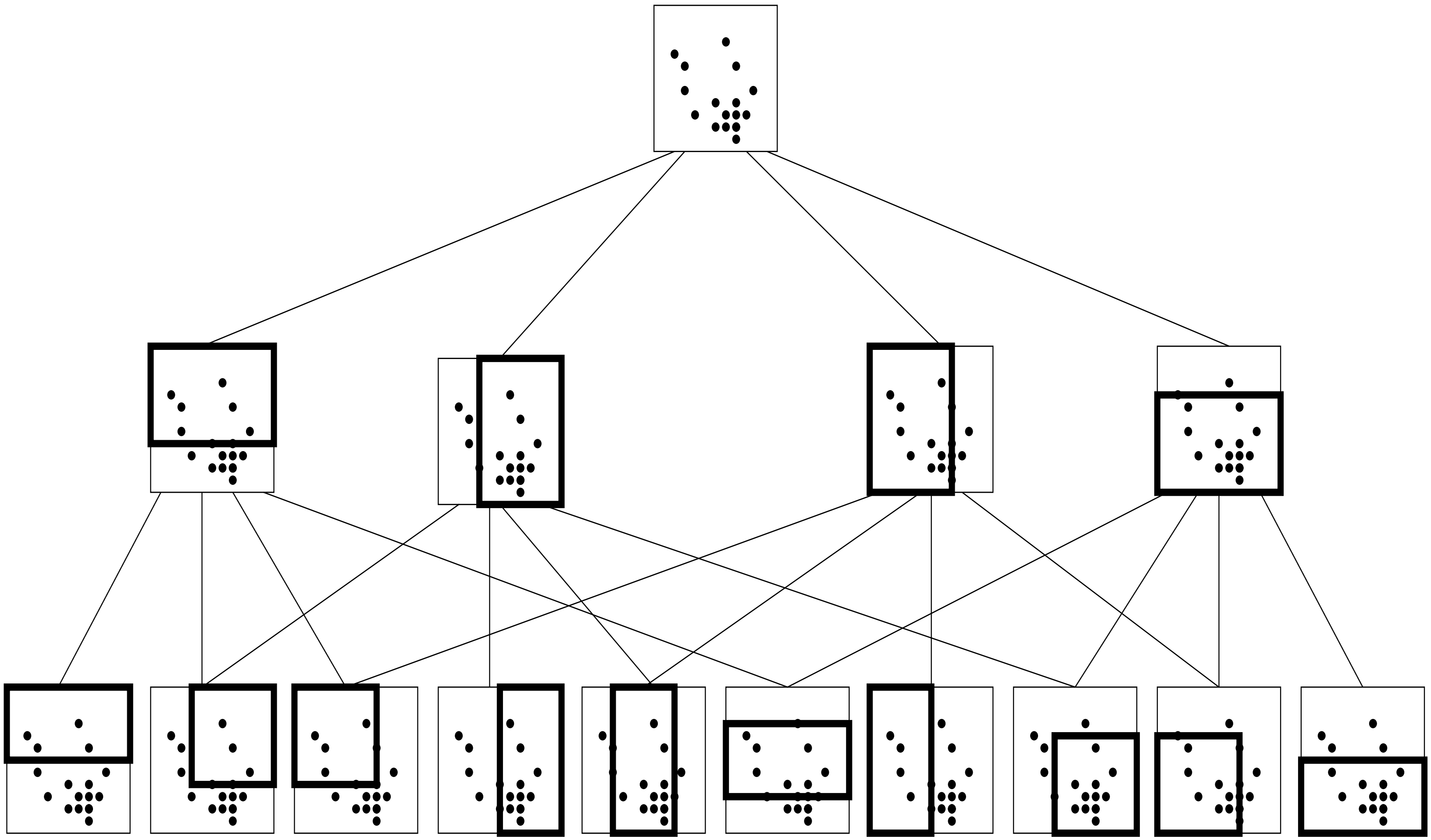
Gridded then Exhaustive

Step 1: Gridded



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

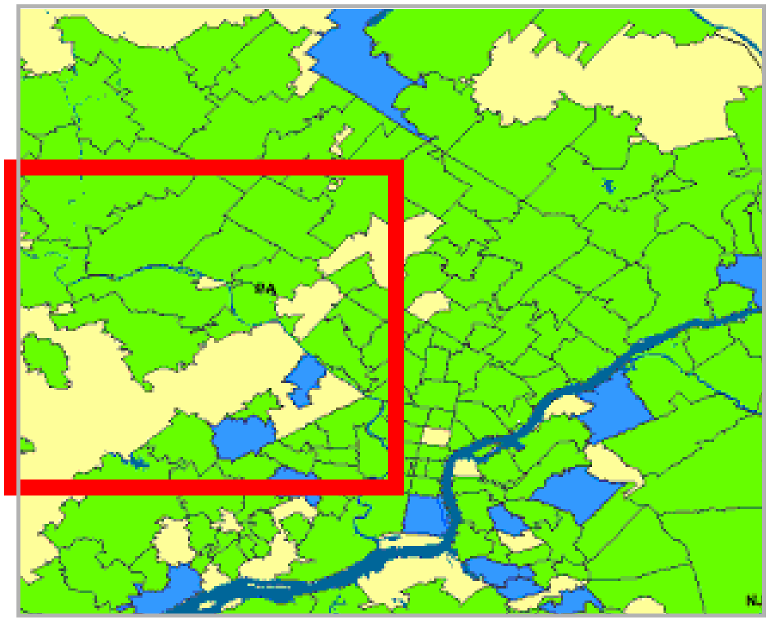
The multi-resolution tree for rectangular regions



collaboration with Daniel Neill <neill@cs.cmu.edu>

Gridded then Exhaustive

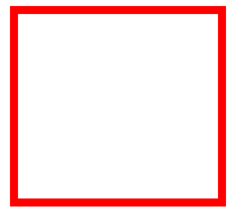
Step 1: Gridded



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

Step 2: Exhaustive

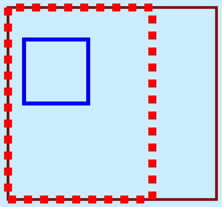
Consider the set of subregions of a Gridded Region.



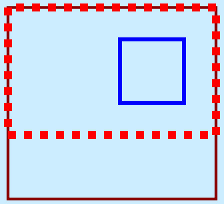
collaboration with Daniel Neill <neill@cs.cmu.edu>

l then Exhaustive

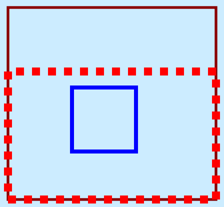
A subregion of me could be one of five types...



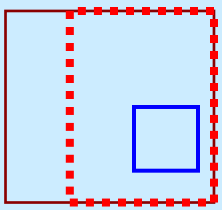
...entirely inside my left gridded child



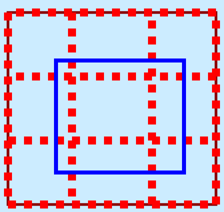
...entirely inside my top gridded child



...entirely inside my bottom gridded child



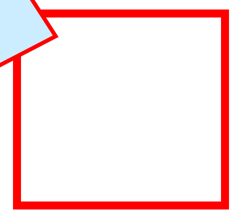
...entirely in my right gridded child



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

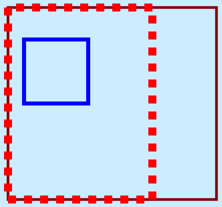
Step 2: Exhaustive

Consider the set of subregions of a Gridded Region.

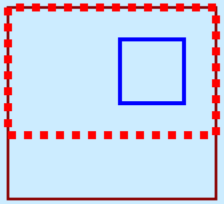


l then Exhaustive

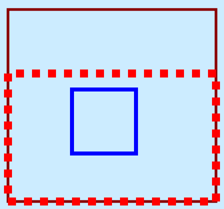
A subregion of me could be one of five types...



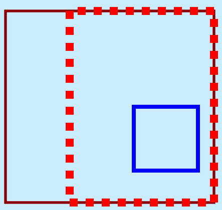
...entirely inside my left gridded child



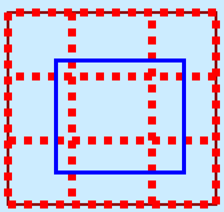
...entirely inside my top gridded child



...entirely inside my bottom gridded child



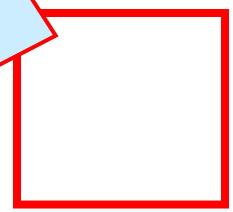
...entirely in my right gridded child



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

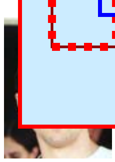
Step 2: Exhaustive

Consider the set of subregions of a Gridded Region.



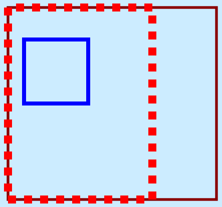
FACT: Any subregion of this type must include the middle...

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

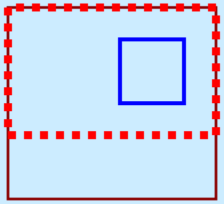


l then Exhaustive

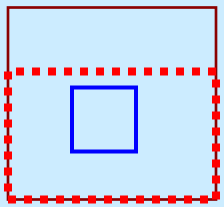
A subregion of me could be one of five types...



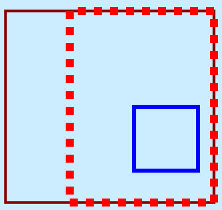
...entirely inside my left gridded child



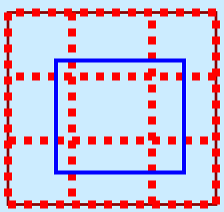
...entirely inside my top gridded child



...entirely inside my bottom gridded child



...entirely in my right gridded child



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

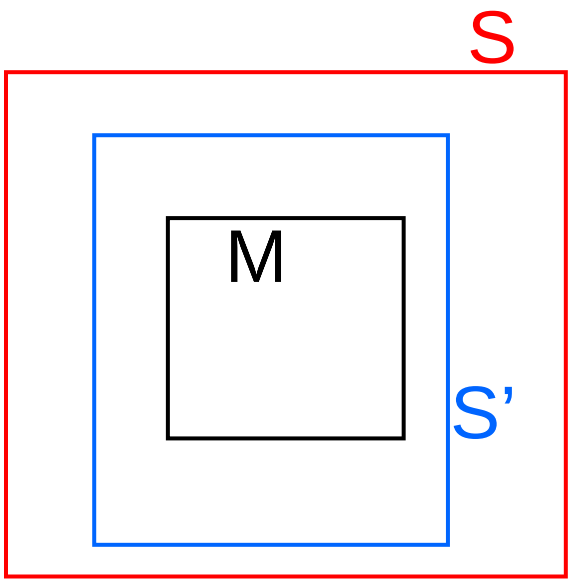
Procedure: Exhaust(Gridded Region)

1. Exhaust(Region.Left)
2. Exhaust(Region.Top)
3. Exhaust(Region.Bottom)
4. Exhaust(Region.Right)
5. Is it possible that any “Type 5” subregion of “Gridded Region” could score better than best known score to date?

    NO: Quit Procedure!

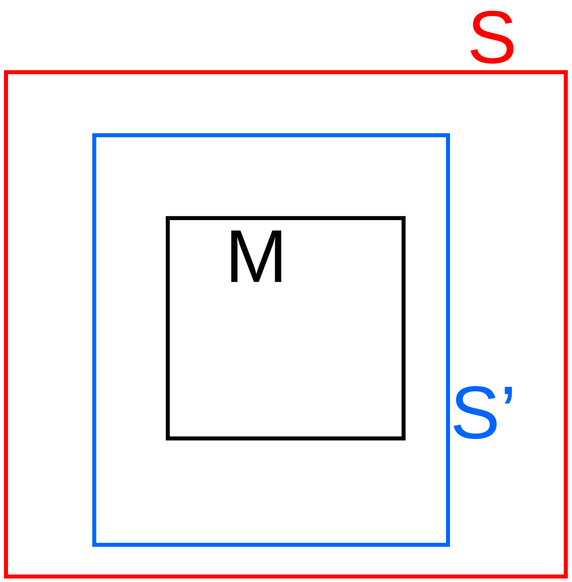
    YES: Check all “Type 5” Subregions

If S’ is a middle-containing subregion of S…



5. Is it possible that any “Type 5” subregion of “Gridded Region” could score better than best known score to date?

If S’ is a middle-containing subregion of S…

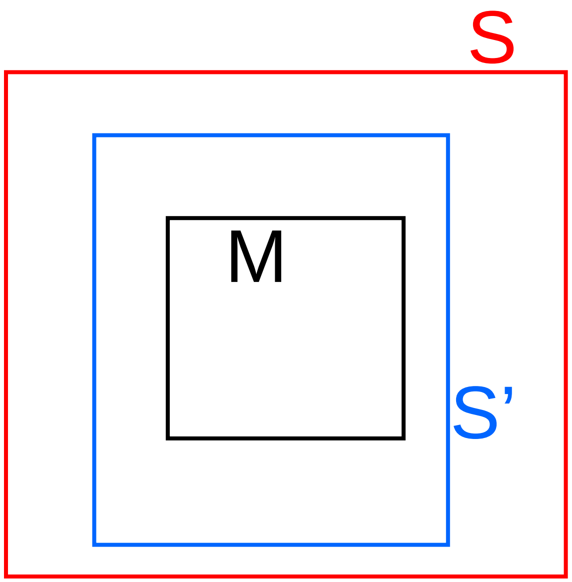


Score(S’) = Score(count(S’) , baseline(S’))

5. Is it possible that any “Type 5” subregion of “Gridded Region” could score better than best known score to date?

collaboration with Daniel Neill <neill@cs.cmu.edu>

If S’ is a middle-containing subregion of S…



An upper bound of c/b for any subregion of S-M

$$d_{inc} \geq \frac{c(S') - c(M)}{b(S') - b(M)}$$

$$b(M) \leq b(S') \leq b(S)$$

$$c(M) \leq c(S') \leq c(S)$$

An upper bound of c/b for any subregion of S that contains M

$$d_{\max} \geq \frac{c(S')}{b(S')}$$

A lower bound on c/b for any subregion of S that excludes M

$$d_{\min} \leq \frac{c(S) - c(S')}{b(S) - b(S')}$$

Score(S’) = Score(count(S’) , baseline(S’))

5. Is it possible that any “Type 5” subregion of “Gridded Region” could score better than best known score to date?

collaboration with Daniel Neill <neill@cs.cmu.edu>

If S’ is a middle-containing subregion of S…

Assume:

$$\frac{\partial}{\partial c}\text{Score}(c,b) \geq 0$$

$$\frac{\partial}{\partial b}\text{Score}(c,b) \leq 0$$

$$\frac{\partial}{\partial b}\text{Score}(c,b) + \frac{c}{b}\frac{\partial}{\partial c}\text{Score}(c,b) \geq 0$$

A lower bound on c/b for any subregion of S that excludes C

$$d_{inc} \geq \frac{c(S') - c(M)}{b(S') - b(M)}$$

$$b(M) \leq b(S') \leq b(S)$$

$$c(M) \leq c(S') \leq c(S)$$

$$d_{\max} \geq \frac{c(S')}{b(S')}$$

$$d_{\min} \leq \frac{c(S) - c(S')}{b(S) - b(S')}$$

Score(S’) = Score(count(S’) , baseline(S’))

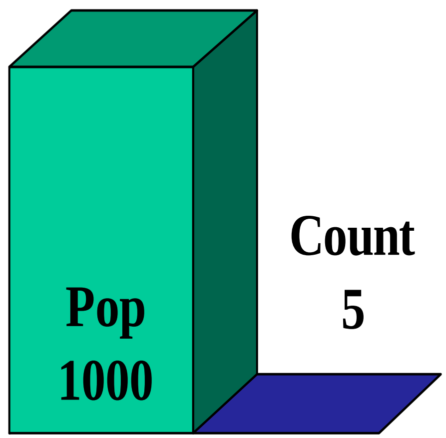
5. Is it possible that any “Type 5” subregion of “Gridded Region” could score better than best known score to date?

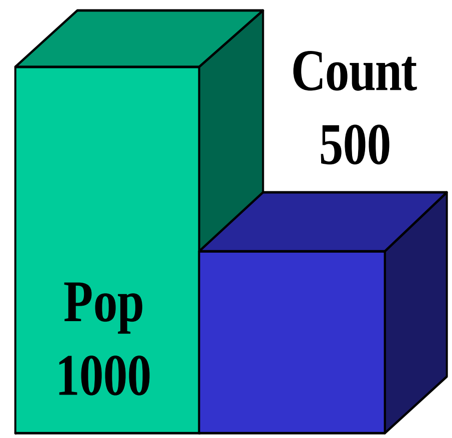
collaboration with Daniel Neill <neill@cs.cmu.edu>

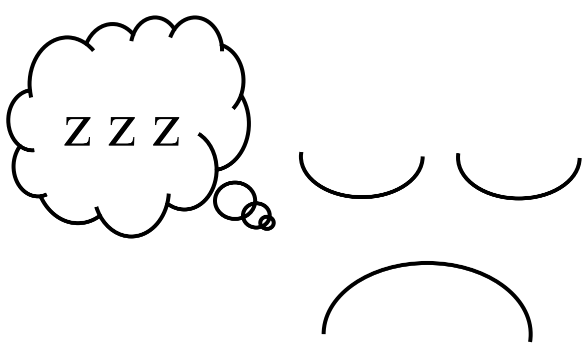
$$\frac{\partial}{\partial c}\mathrm{Score}(c,b) \geq 0$$

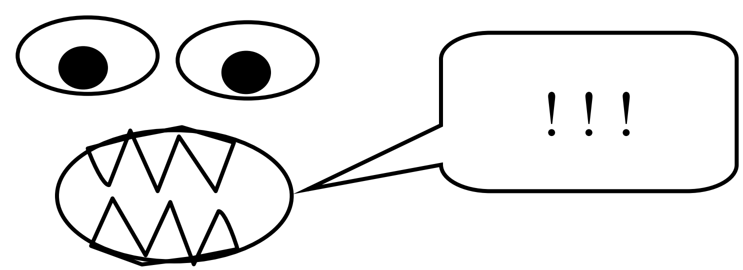
Properties of D(S)

Score(S) increases with the total count of S, $C(S) = \sum_S c_i$.







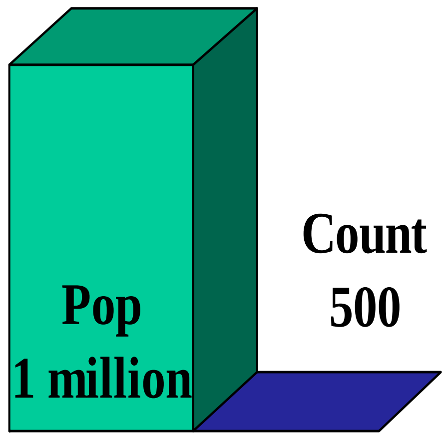


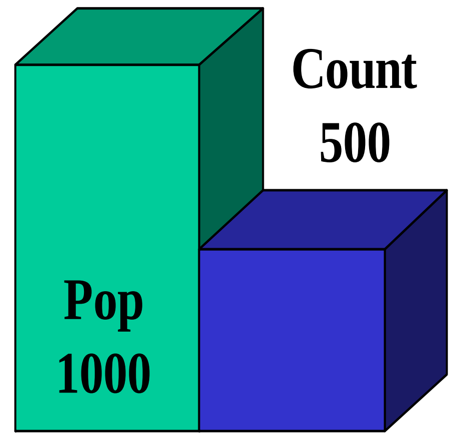
collaboration with Daniel Neill <neill@cs.cmu.edu>

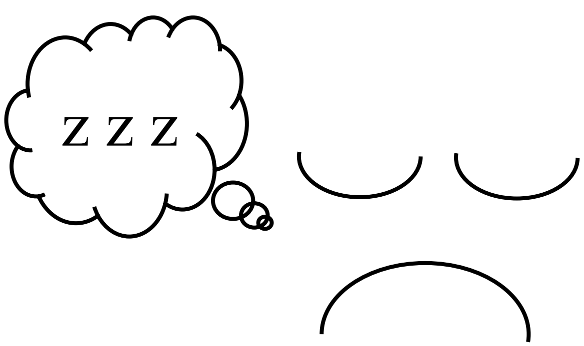
$$\frac{\partial}{\partial b}\mathrm{Score}(c,b) \le 0$$

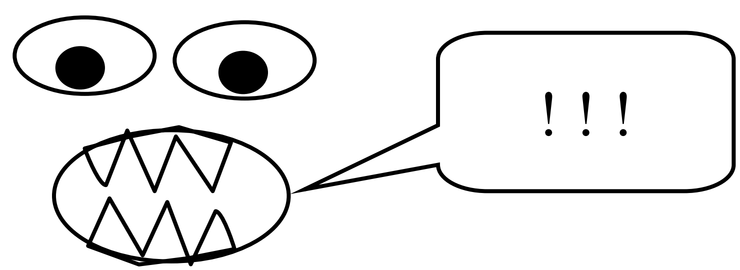
Properties of D(S)

Score(S) decreases with total baseline of S, B(S) = $\sum_S b_i$.







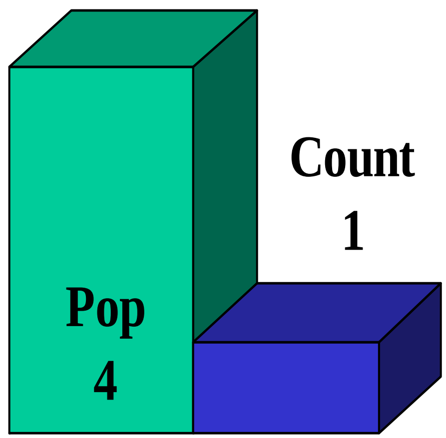


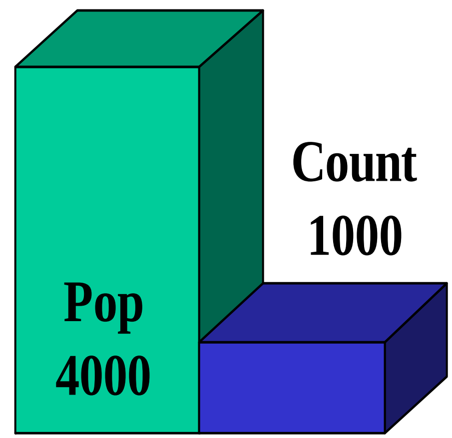
collaboration with Daniel Neill <neill@cs.cmu.edu>

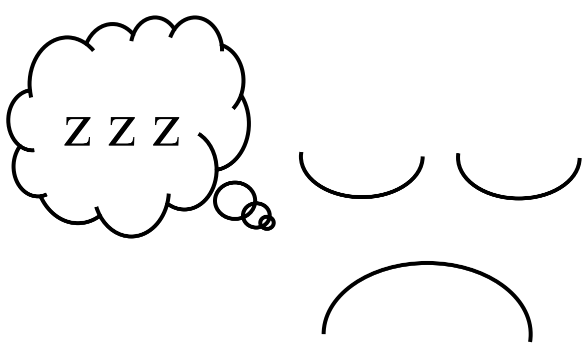
Properties of D(S)

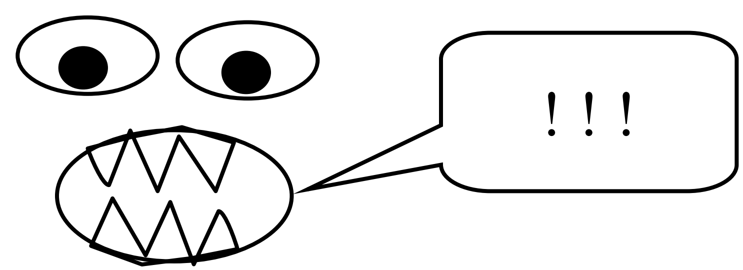
$$\frac{\partial}{\partial b}\text{Score}(c,b)+\frac{c}{b}\frac{\partial}{\partial c}\text{Score}(c,b)\geq 0$$

For a constant ratio C / B, *Score(S)* increases with C and B.









collaboration with Daniel Neill <neill@cs.cmu.edu>

If S’ is a middle-containing subregion of S…

Assume:

$$\frac{\partial}{\partial c}\mathrm{Score}(c,b) \geq 0$$

$$\frac{\partial}{\partial b}\mathrm{Score}(c,b) \leq 0$$

$$\frac{\partial}{\partial b}\mathrm{Score}(c,b) + \frac{c}{b}\frac{\partial}{\partial c}\mathrm{Score}(c,b) \geq 0$$

$$d_{inc} \geq \frac{c(S') - c(M)}{b(S') - b(M)}$$

$$b(M) \leq b(S') \leq b(S)$$

$$c(M) \leq c(S') \leq c(S)$$

$$d_{\max} \geq \frac{c(S')}{b(S')}$$

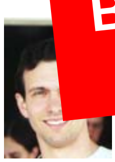
A lower bound on c/b for any subregion of S that excludes C

$$d_{\min} \leq \frac{c(S) - c(S')}{b(S) - b(S')}$$

Score(S’) = Score(count(S’) , baseline(S’))

Bottom Line: all the above lets us put a good upper bound on Score(S’)

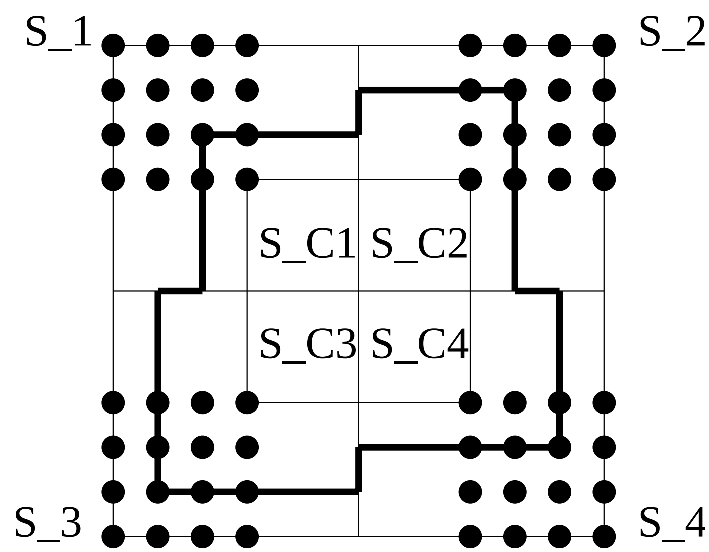
ssible that any “Type 5” subregion
dded Region” could score better
than best known score to date?



collaboration with Daniel Neill <neill@cs.cmu.edu>

Tighter score bounds by quartering

- We precompute global bounds on populations p_{ij} and ratios c_{ij} / p_{ij}, and use these for our initial pruning.
- If we cannot prune the outer regions of S using the global bounds, we do a second pass which is more expensive but allows much more pruning.
- We can use quartering to give much tighter bounds on populations and ratios, and compute a better score bound using these.
  - Requires time quadratic in region size; in effect, we are computing bounds for all irregular but rectangle-like outer regions.



Where are we?

- So we can find the most significant region by searching over the desired set of regions S, and finding the highest D(S).

- Now how can we find whether this region actually is a significant cluster?


collaboration with Daniel Neill <neill@cs.cmu.edu>

Where are we?

- So we can find the most significant region by searching over the desired set of regions S, and finding the highest D(S).

- Now how can we find whether this region actually is a significant cluster?

- Randomization testing

Can sometimes cost us 1000 times more computation!

Though there are further tricks…

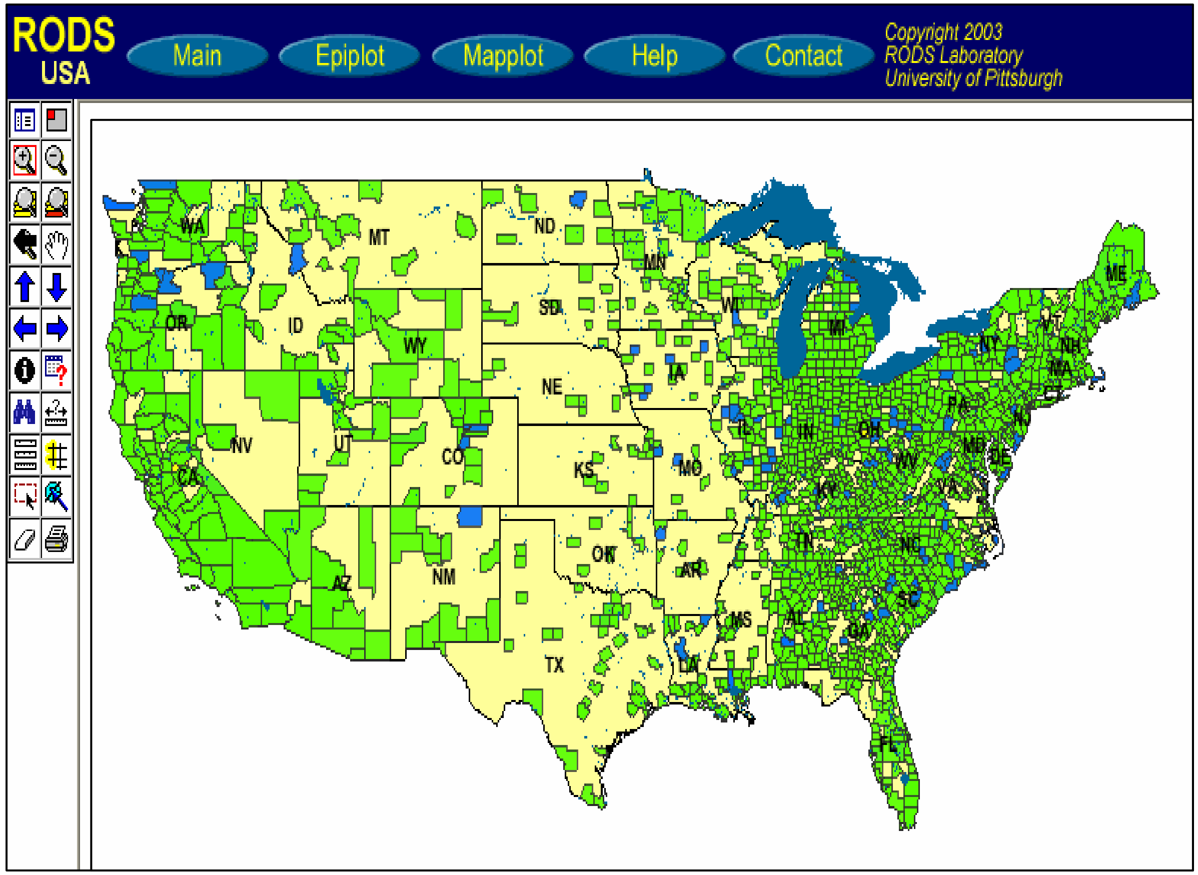
collaboration with Daniel Neill <neill@cs.cmu.edu>

Why the Scan Statistic speed obsession?

collaboration with Daniel Neill <neill@cs.cmu.edu>

Why the Scan Statistic speed obsession?

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

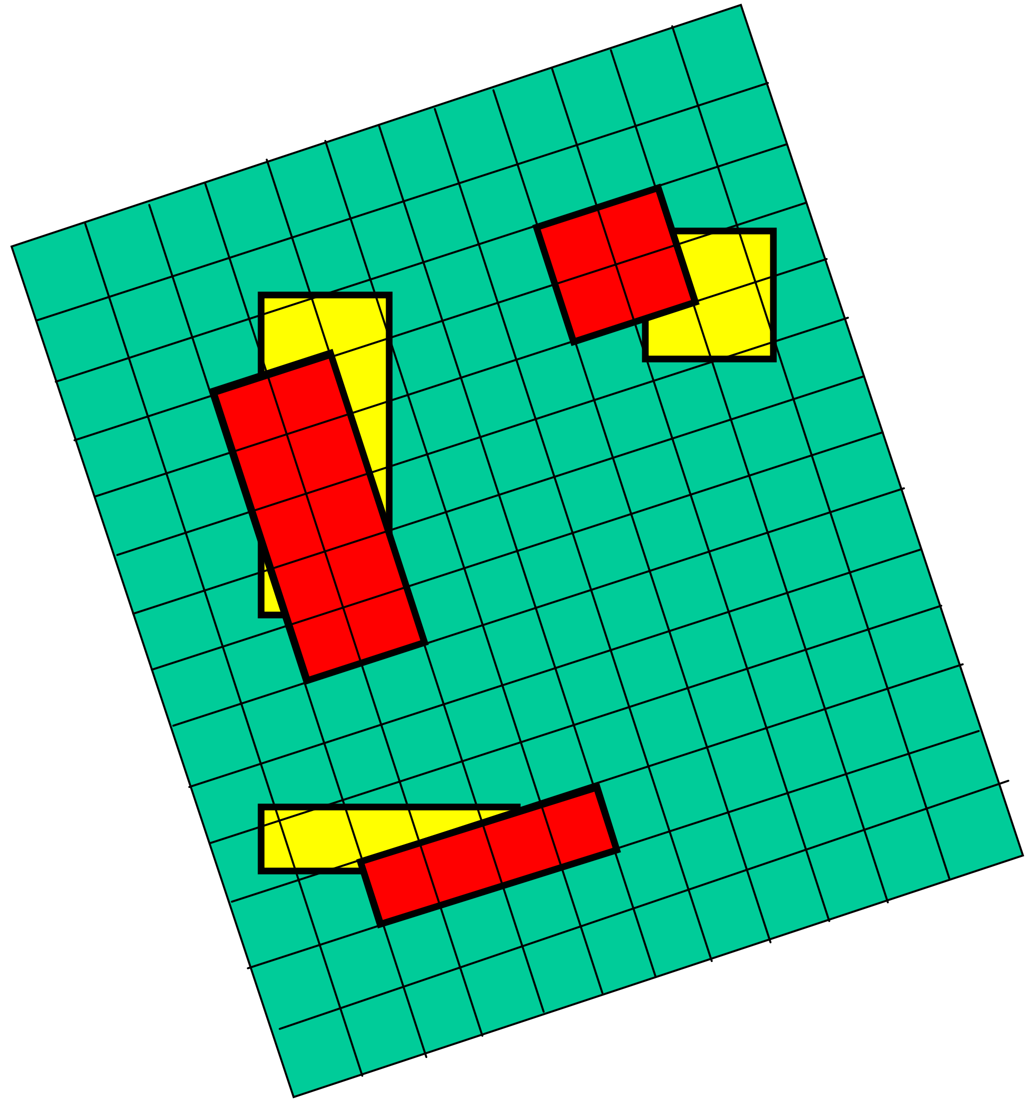



collaboration with Daniel Neill <neill@cs.cmu.edu>

Which regions to search?

- We choose to search over the space of all rectangular regi…
  …ers
  …es
  …ology
  …ers are
  …ated (e.g. from
  wind… pathogens).
  – Important in brain imaging because of the brain's "folded sheet" structure.

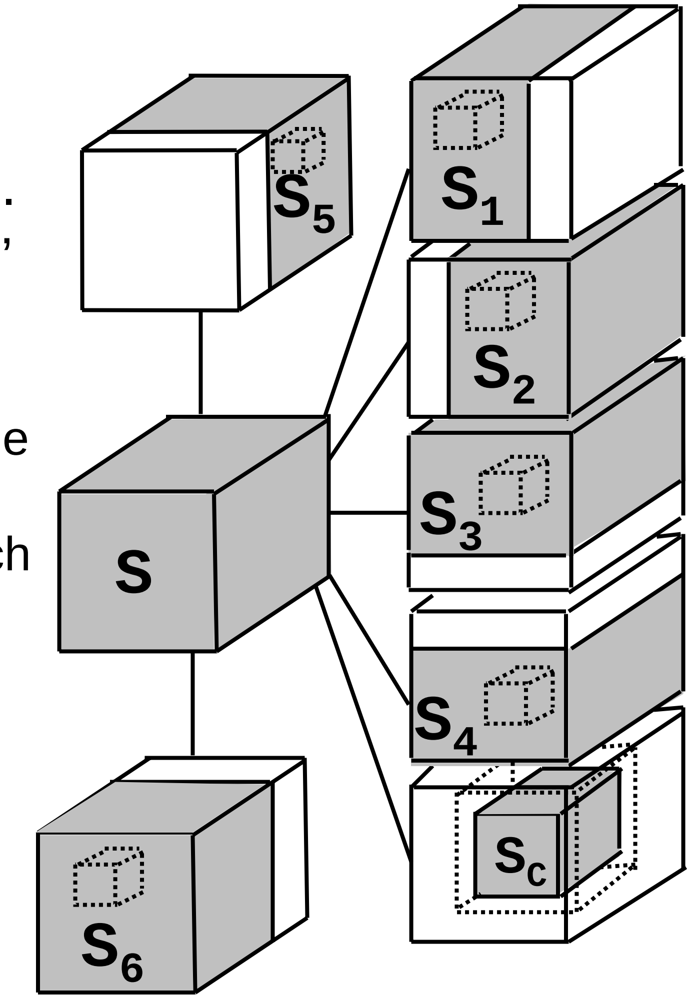
We can find non-axis-aligned rectangles by examining multiple rotations of the data.




collaboration with Daniel Neill <neill@cs.cmu.edu>

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.

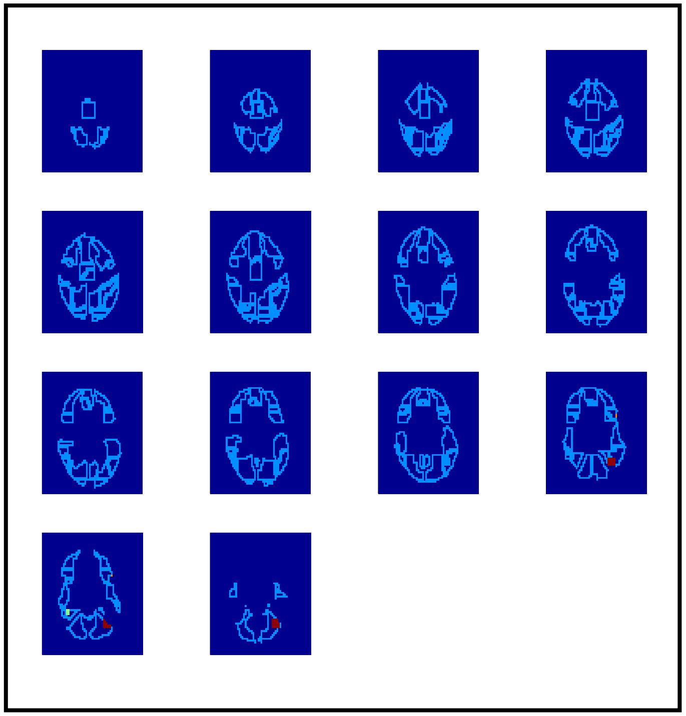


- Algorithm: Neill, Moore and Mitchell NIPS 2005

collaboration with Daniel Neill <neill@cs.cmu.edu>

Results: OTC, fMRI

- fMRI data (64 x 64 x 14 grid):
  - 7-148x speedups as compared to exhaustive search approach.



fMRI data from noun/verb word recognition task

collaboration with Daniel Neill <neill@cs.cmu.edu>

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.

collaboration with Daniel Neill <neill@cs.cmu.edu>

Density-based cluster detection

- Kernel density based detection
- Spatial statistics
- Connected component approaches
- Density optima
- Linear scan approximations


collaboration with Daniel Neill <neill@cs.cmu.edu>

Density-based cluster detection

- Kernel density based detection
- Spatial statistics
- Connected component approaches
- Density optima
- Linear scan approximations

- DBSCAN (Ester, Kriegel, Sander and Xu)
- CFF Clustering (Cuevas, Febrero and Fraiman)
- CLIQUE (Agrawal, Gehrke, Gunopulus, and Raghavan)
- Priebe's method (Priebe)
- MAFIA (Goil, Nagesh and Choudhary)
- DENCLUE (Hinneburg and Keim)
- STING (Wang, Yang, and Muntz)
- Bump Hunting (Friedman and Fisher)

collaboration with Daniel Neill <neill@cs.cmu.edu>

Density-based cluster detection

- Account for varying baseline?
- Are the hotspots significant?
- Is there a small rise over a large stripe?

—?—

- Kernel density based detection
- Spatial statistics
- Connected component approaches
- Density optima
- Linear scan approximations

- DBSCAN (Ester, Kriegel, Sander and Xu)
- CFF Clustering (Cuevas, Febrero and Fraiman)
- CLIQUE (Agrawal, Gehrke, Gunopulus, and Raghavan)
- Priebe's method (Priebe)
- MAFIA (Goil, Nagesh and Choudhary)
- DENCLUE (Hinneburg and Keim)
- STING (Wang, Yang, and Muntz)
- Bump Hunting (Friedman and Fisher)

collaboration with Daniel Neill <neill@cs.cmu.edu>

For more information and references to related work…

- http://www.autonlab.org/autonweb/14667.html

@inproceedings{neill-rectangles,
Howpublished = {Conference on Knowledge Discovery in Databases (KDD) 2004},
Month = {August},
Year = {2004},
Editor = {J. Guerke and W. DuMouchel},
Author = {Daniel Neill and Andrew Moore},
Title = {Rapid Detection of Significant Spatial Clusters}
}

- http://www.autonlab.org/autonweb/15868.html

@inproceedings{sabhnani-pharmacy,
Month = {August},
Year = {2005},
Booktitle = {Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection},
Author = {Robin Sabhnani and Daniel Neill and Andrew Moore},
Title = {Detecting Anomalous Patterns in Pharmacy Retail Data}
}

- Software: http://www.autonlab.org/autonweb/10474.html

Outline

Cached Sufficient Statistics

New searches over cached statistics

Biosurveillance and Epidemiology

Scan Statistics

Cached Scan Statistics

Branch-and-Bound Scan Statistics

Retail data monitoring

Brain monitoring

Entering Google

Asteroids

Multi (and I mean multi) object target tracking

Multiple-tree search

Entering Google

Outline

- Cached Sufficient Statistics
- New searches over cached statistics

- Biosurveillance and Epidemiology
- Scan Statistics
- Cached Scan Statistics
- Branch-and-Bound Scan Statistics
- Retail data monitoring
- Brain monitoring
- Entering Google

- Asteroids
- Multi (and I mean multi) object target tracking
- Multiple-tree search
- Entering Google

Asteroid Tracking

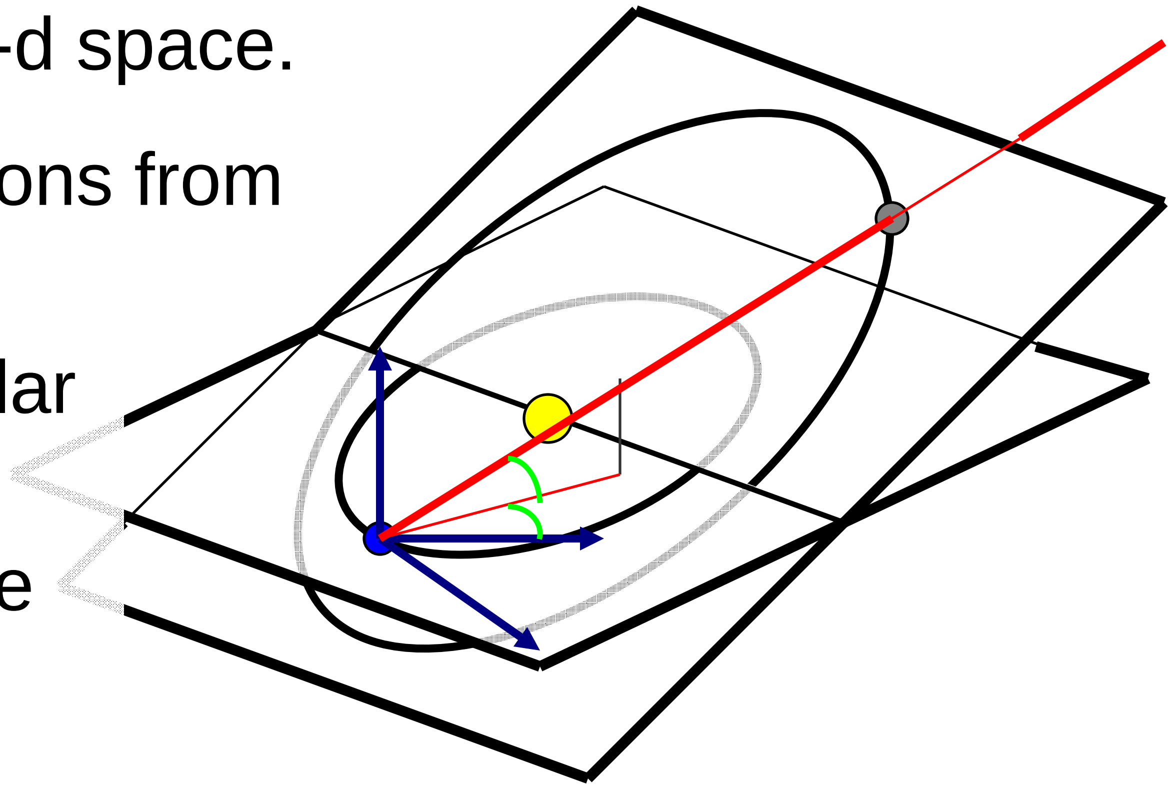
<u>Ultimate Goal</u>: Find all asteroids large enough to do significant damage, calculate their orbits, and determine risk.

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Why Is This Hard/Interesting?

Partial Observability:

- Positions are in 3-d space.
- We see observations from earth.
- We see two angular coordinates (α, δ)
- We do **not** see the distance (r).

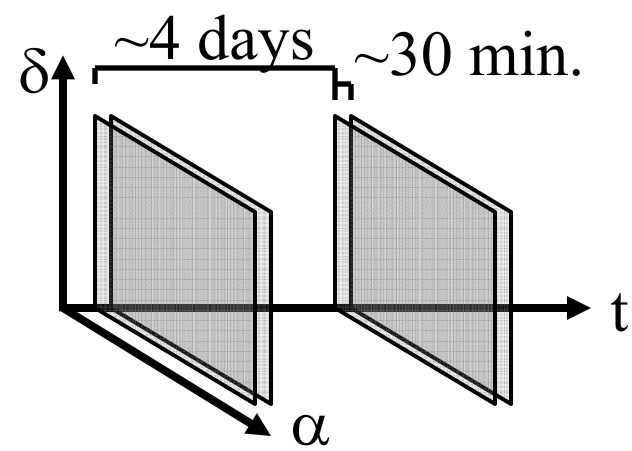


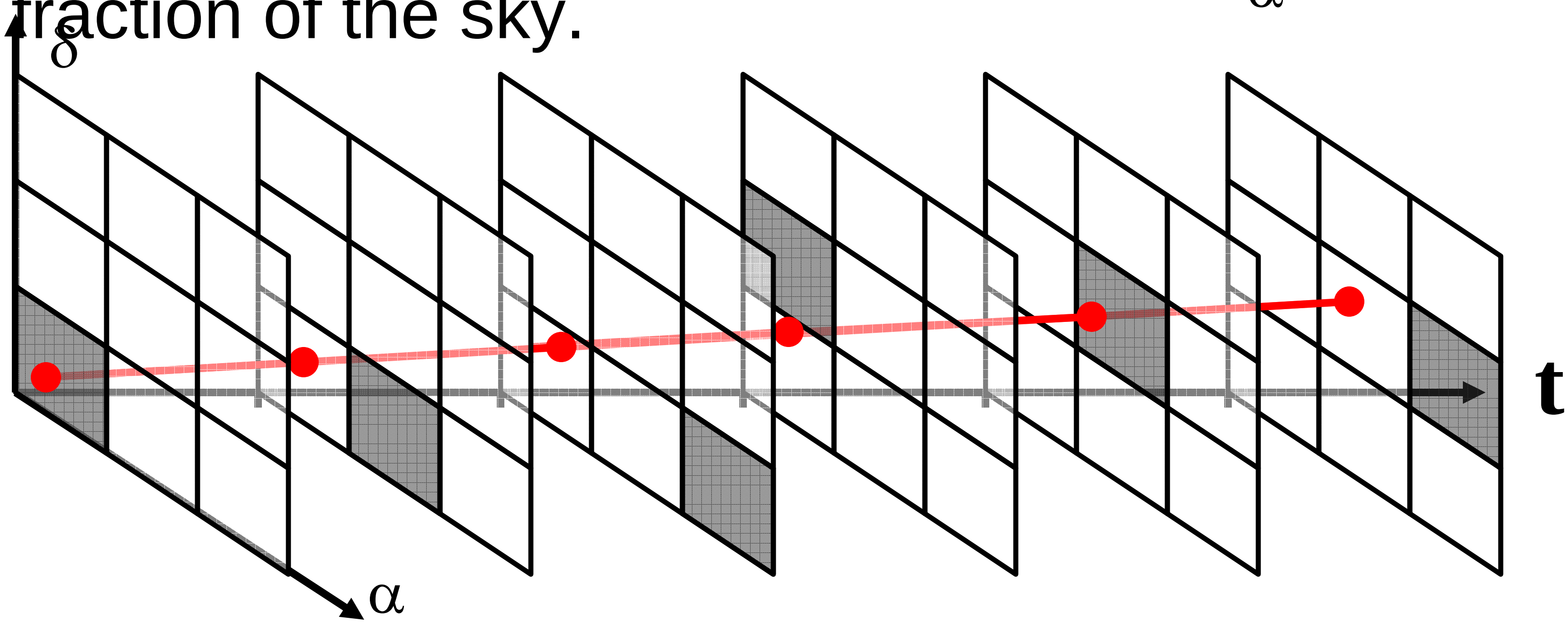
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Why Is This Hard/Interesting?

Temporally sparse:

- Each region viewed infrequently.
- Each viewing only covers a fraction of the sky.



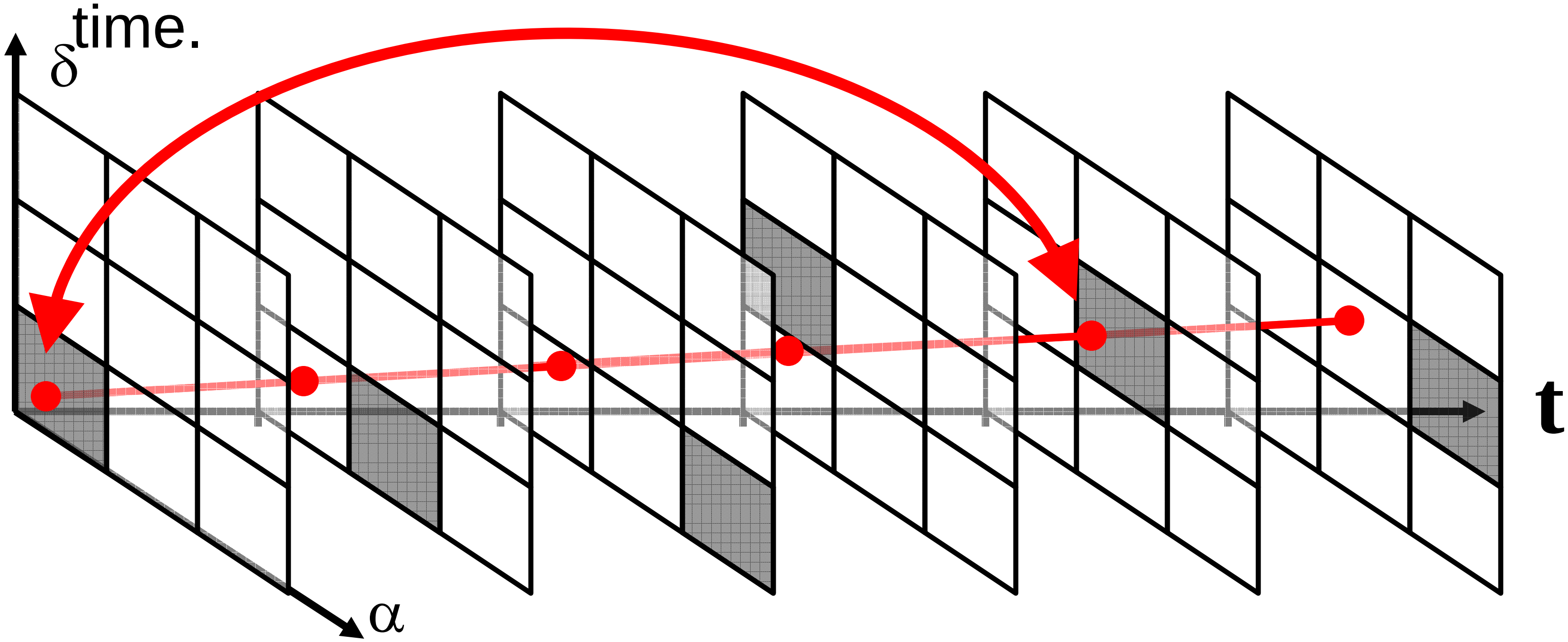


collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

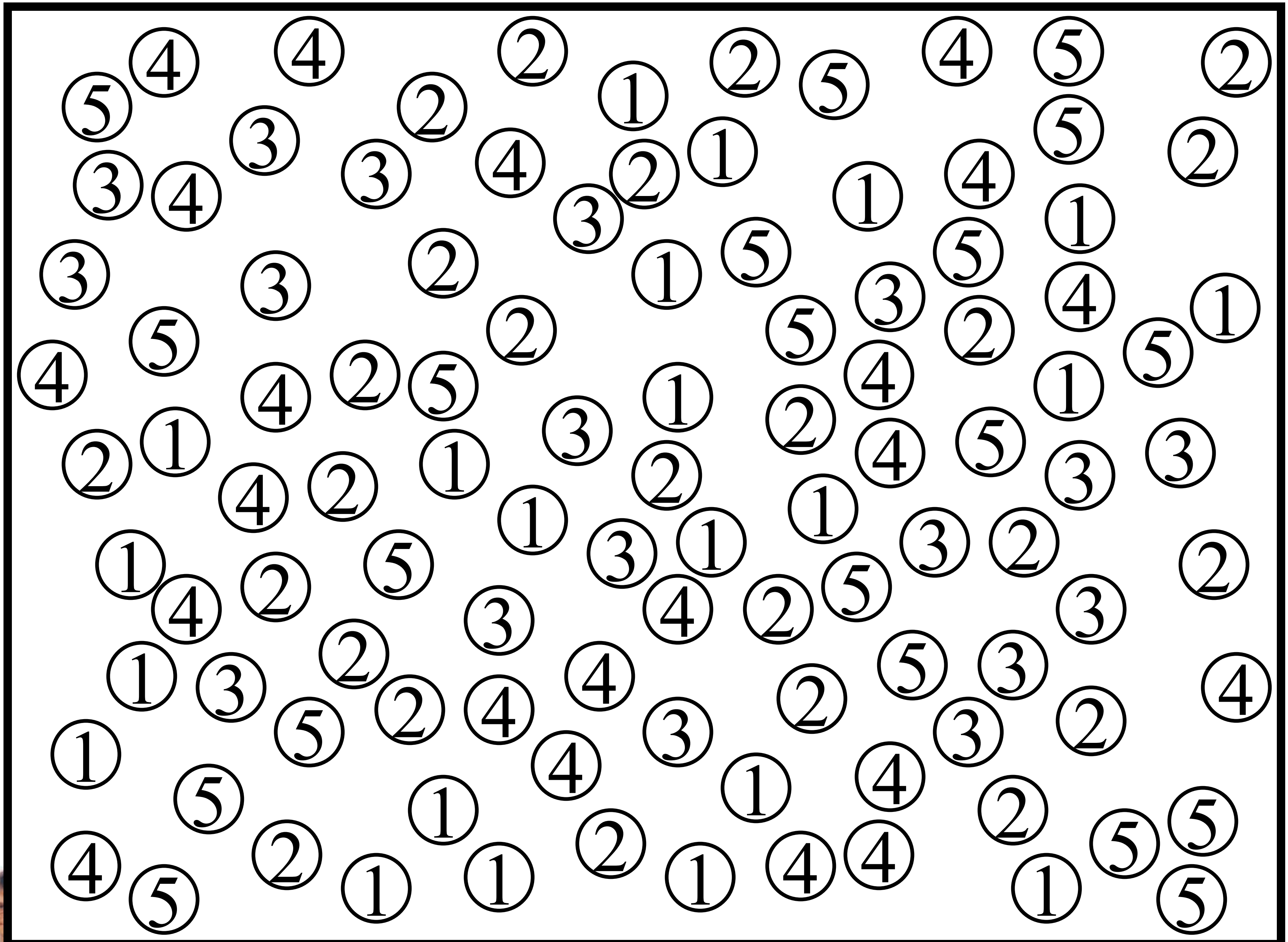
Why Is This Hard/Interesting?

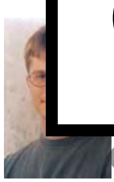
Lack of initial parameter information (and temporally sparse):

- We do not have initial estimates of all of the motion parameters.
- This becomes a significant problem for large gaps in time.

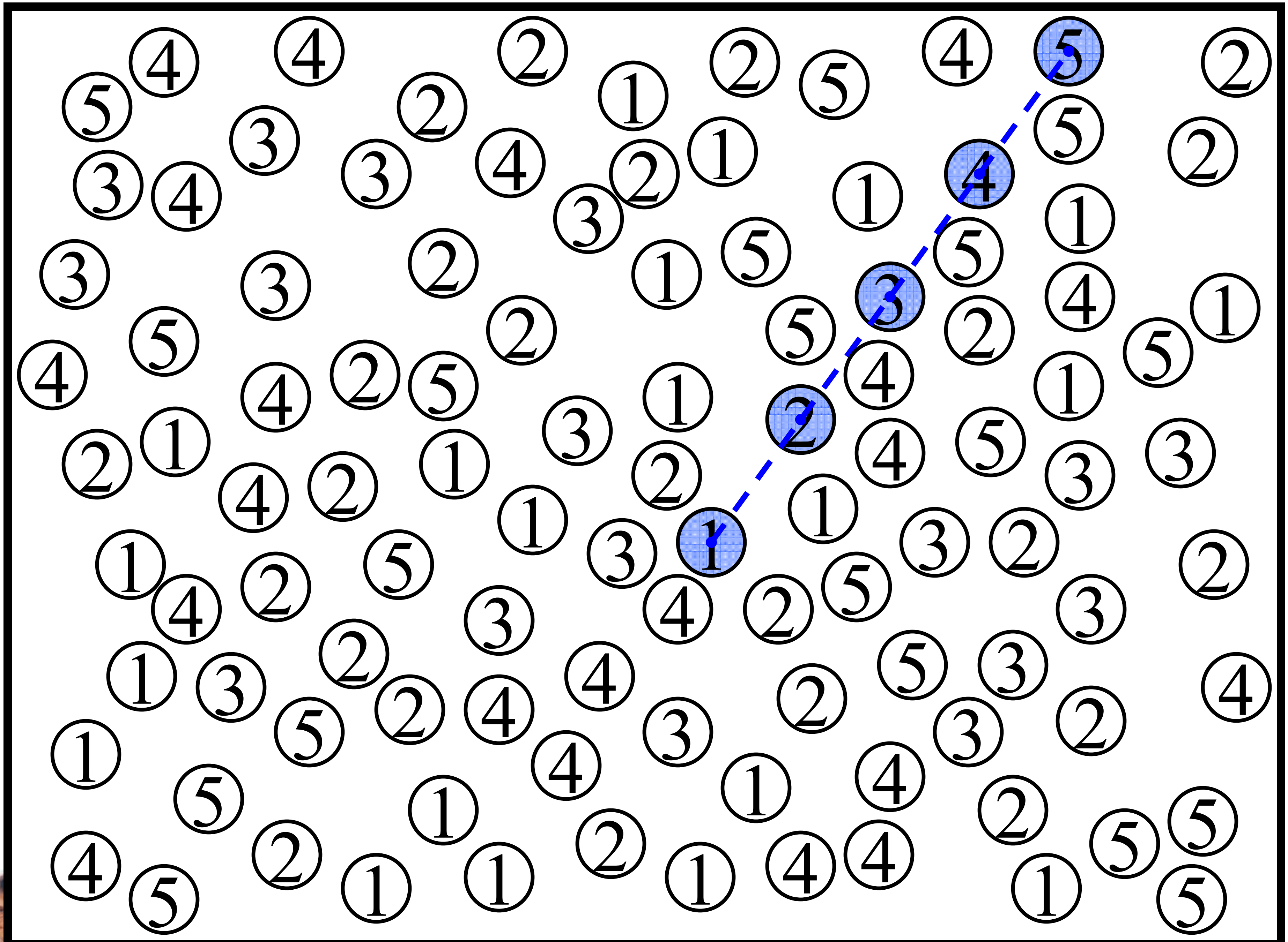


collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

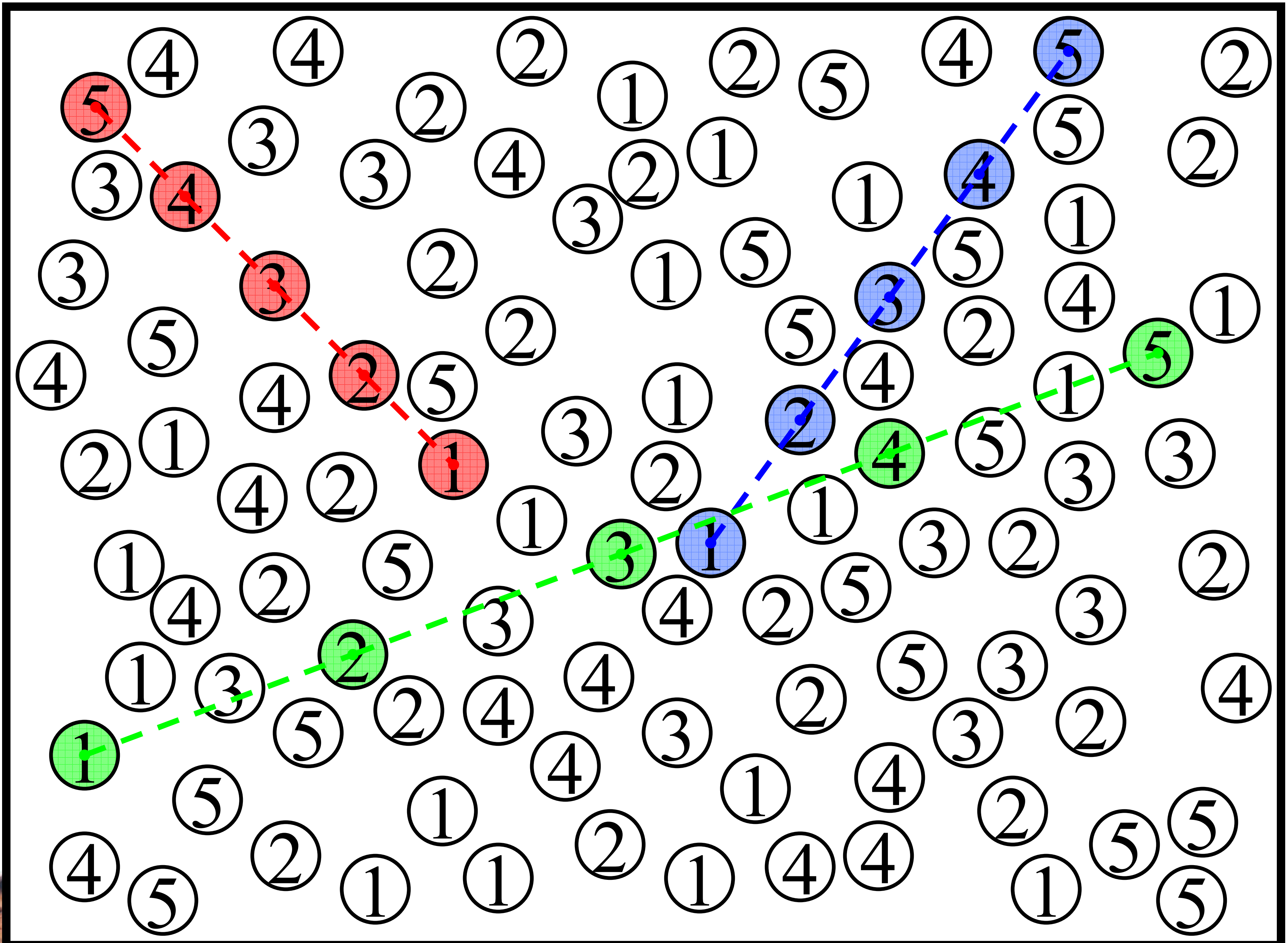




collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

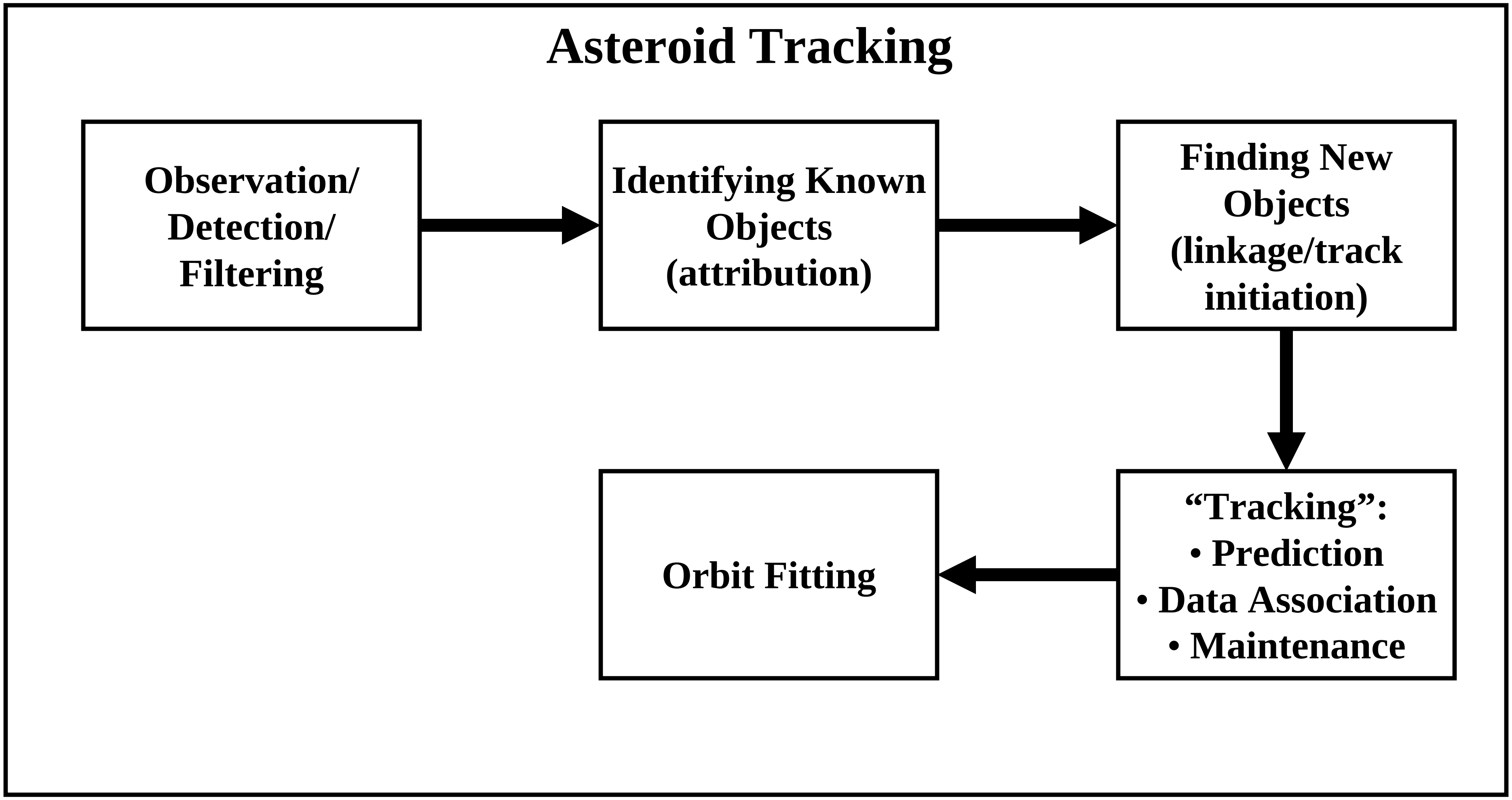


collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>



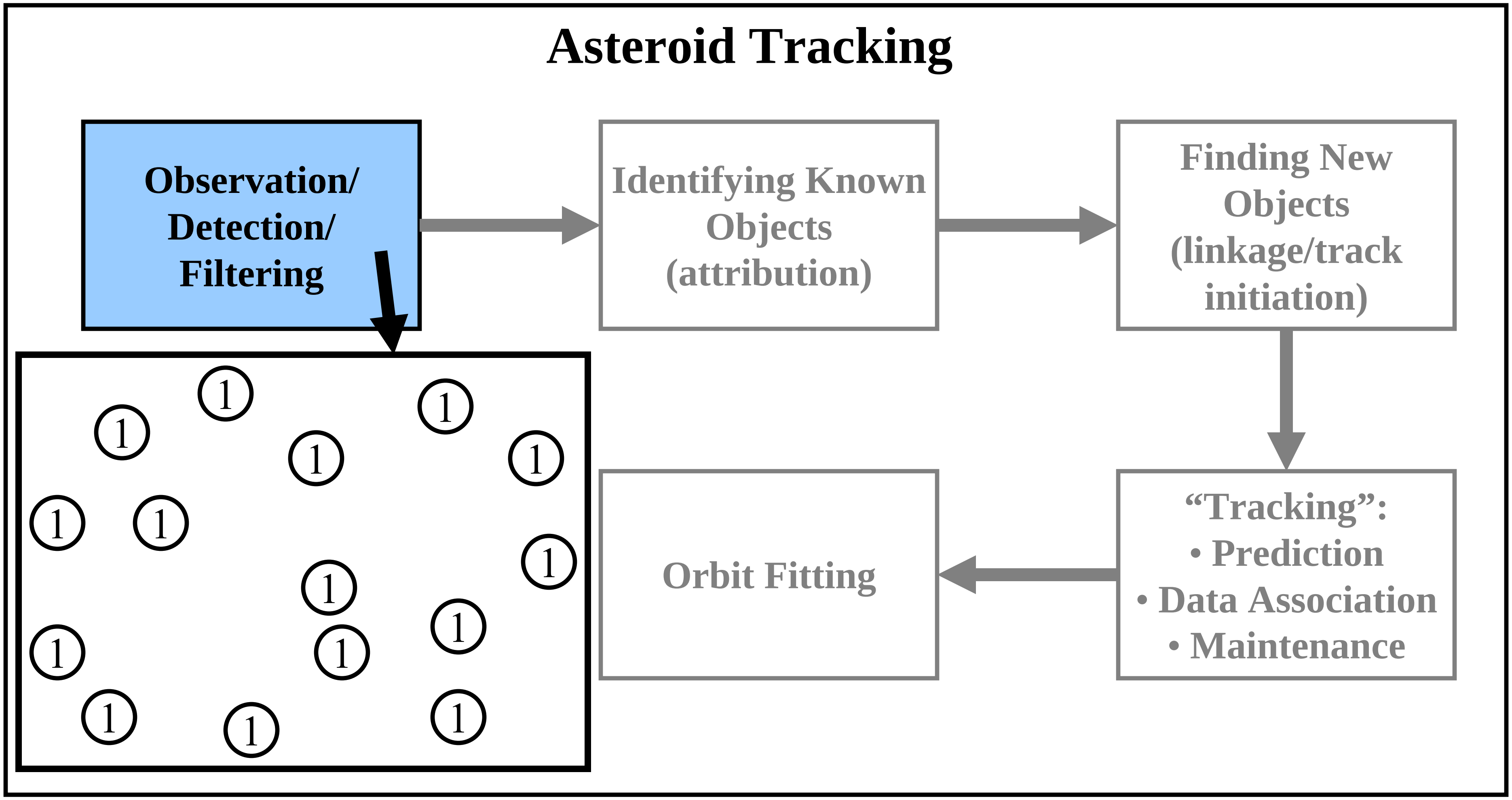

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Problem Overview



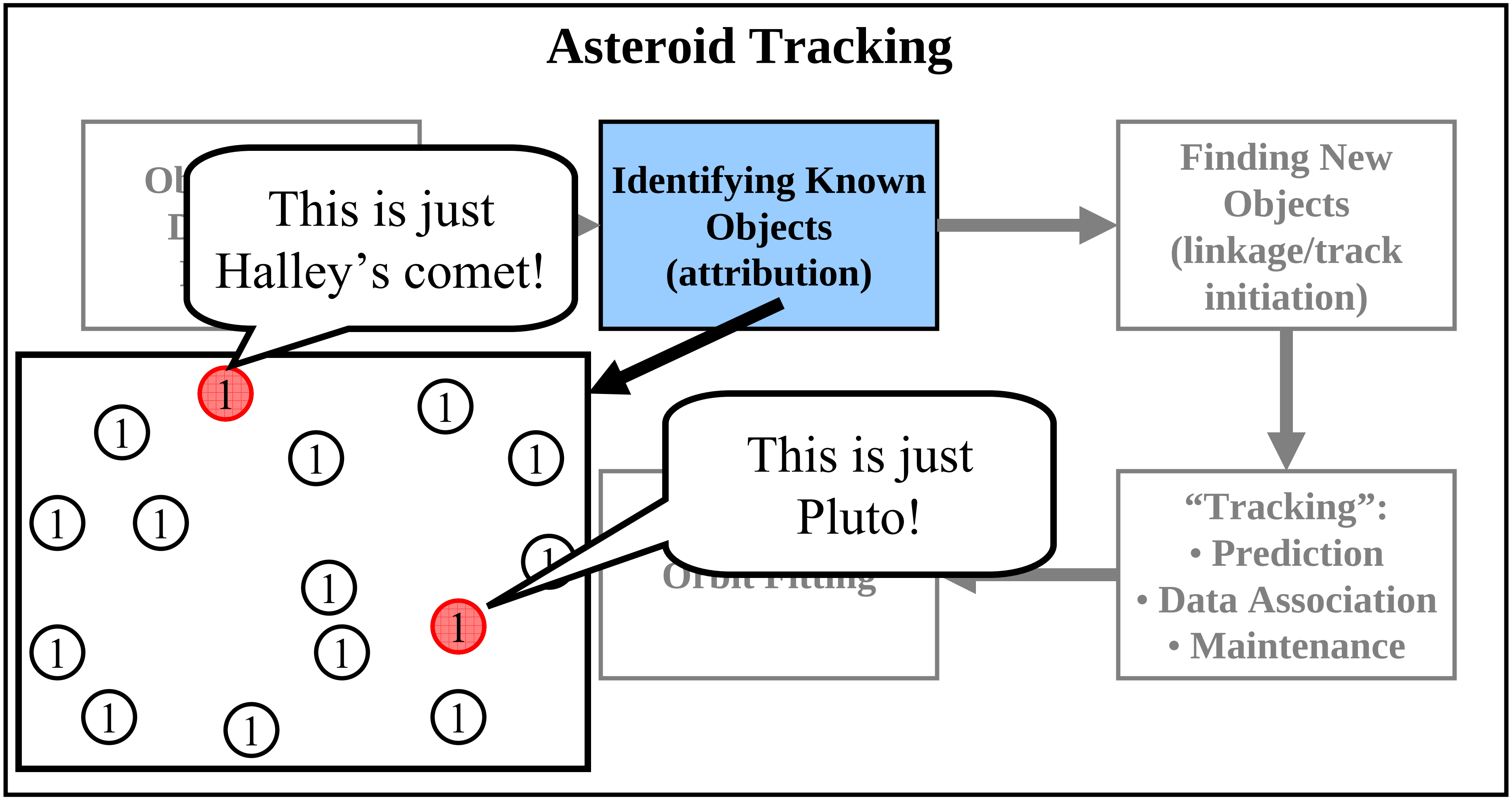
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Problem Overview



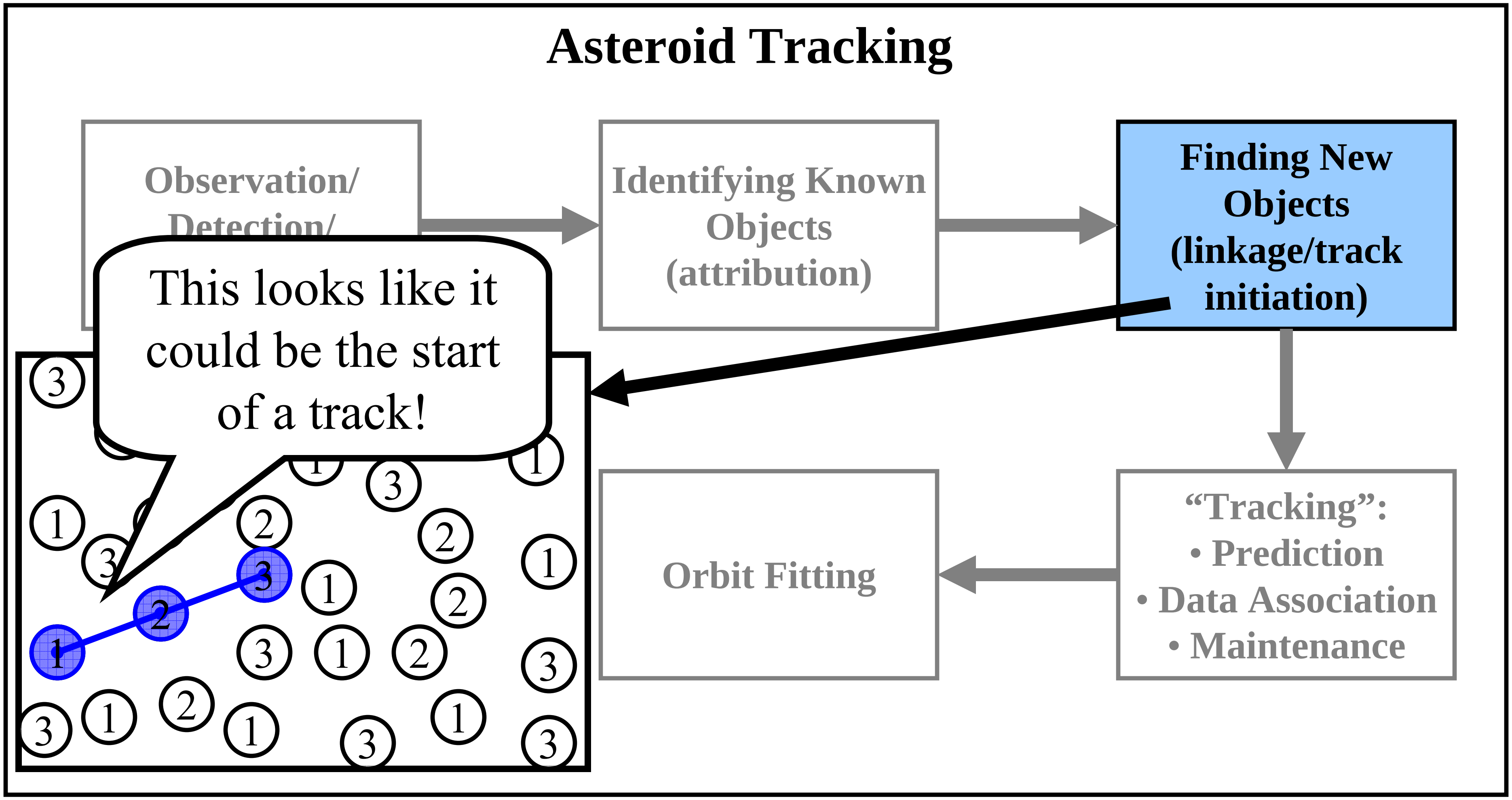
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Problem Overview



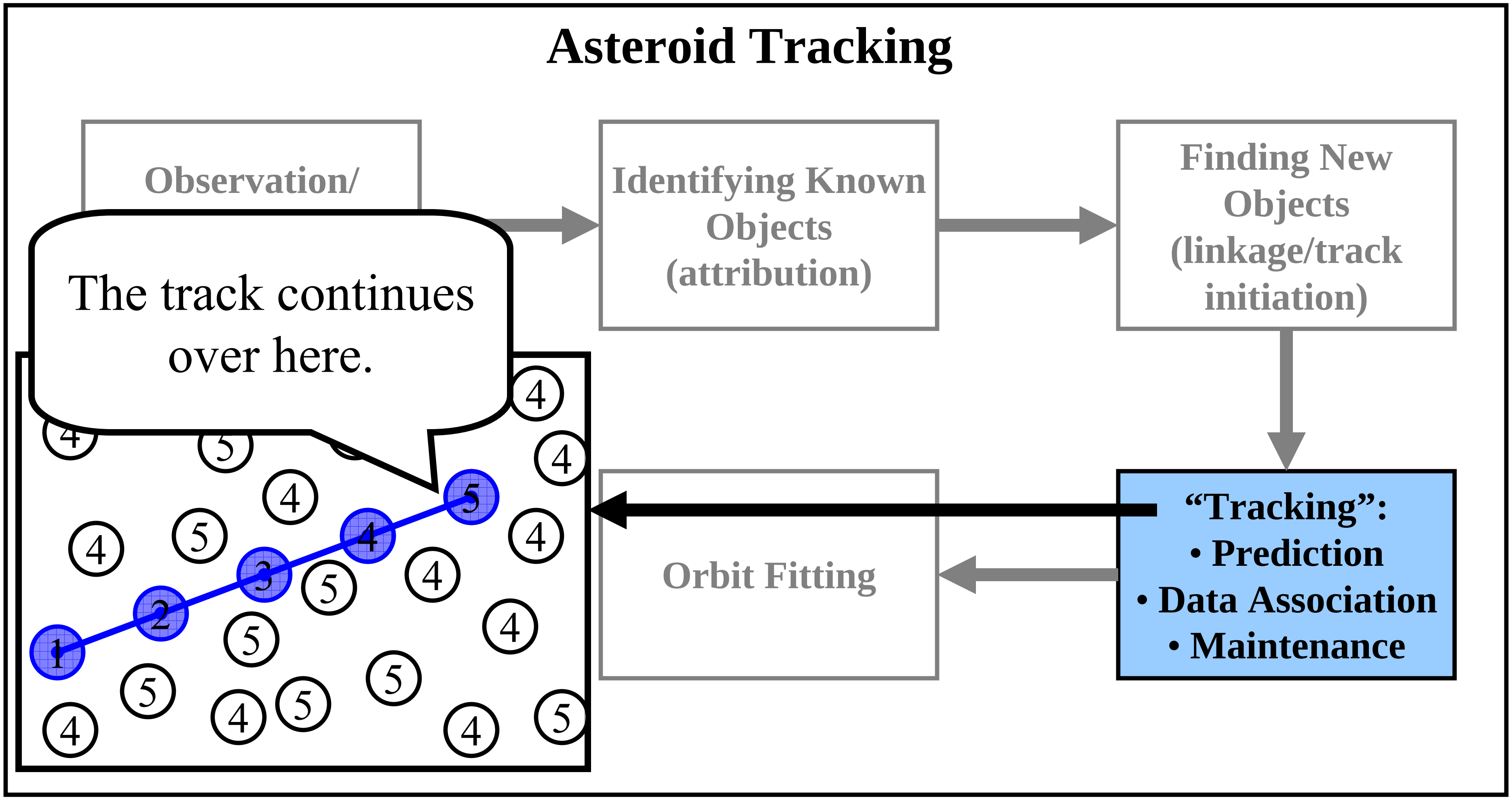
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Problem Overview



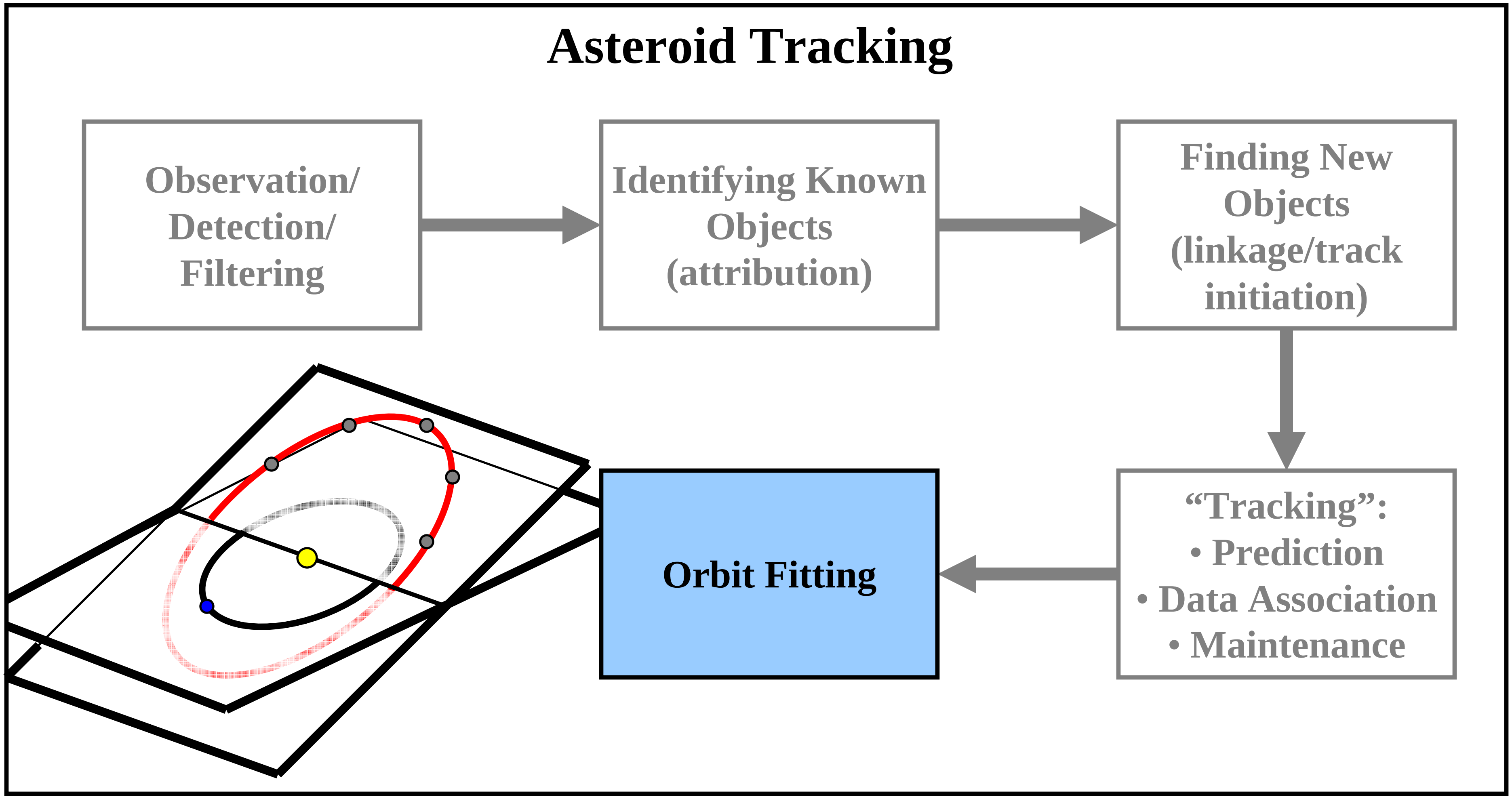
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Problem Overview



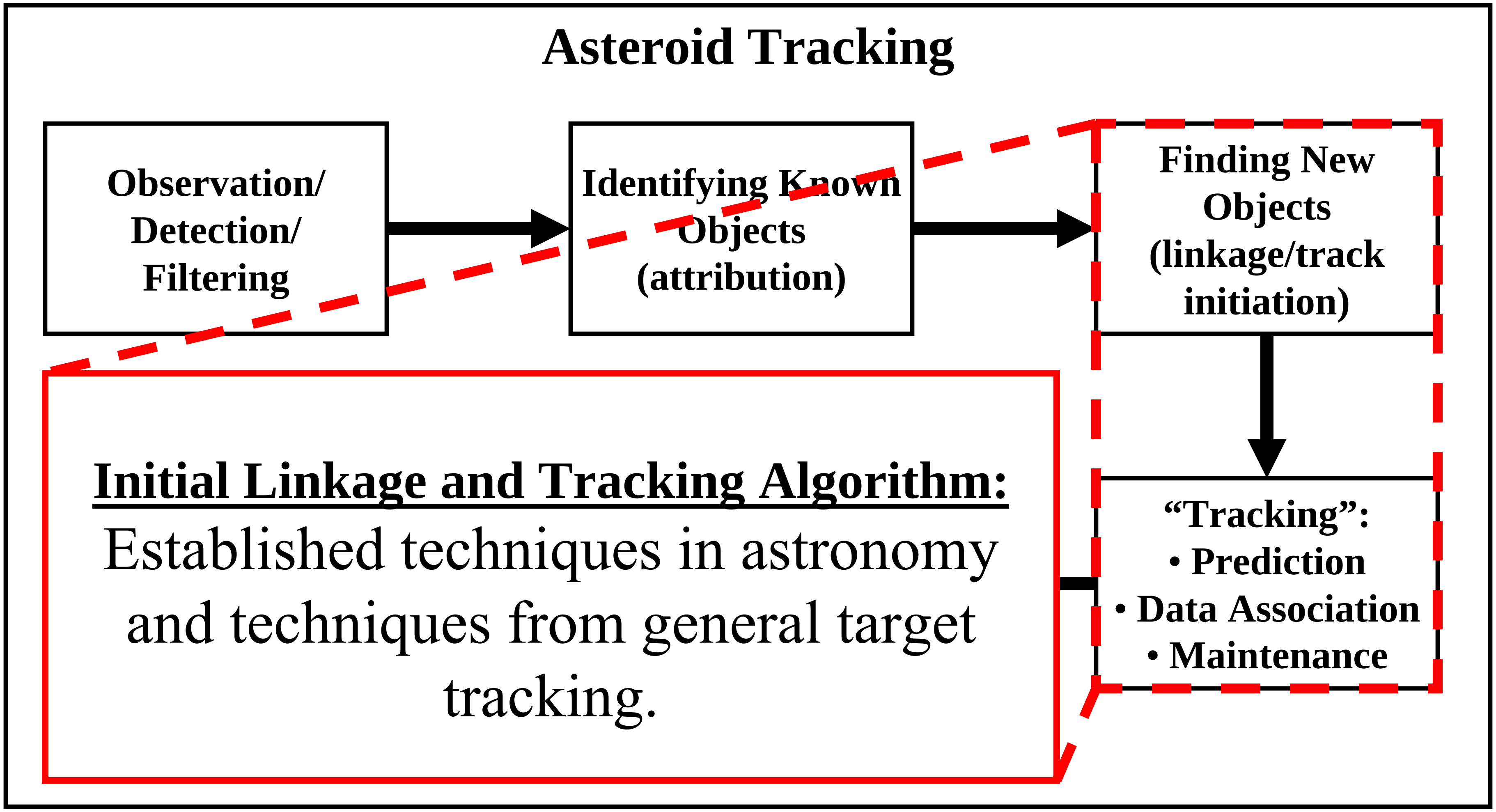
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Problem Overview



collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

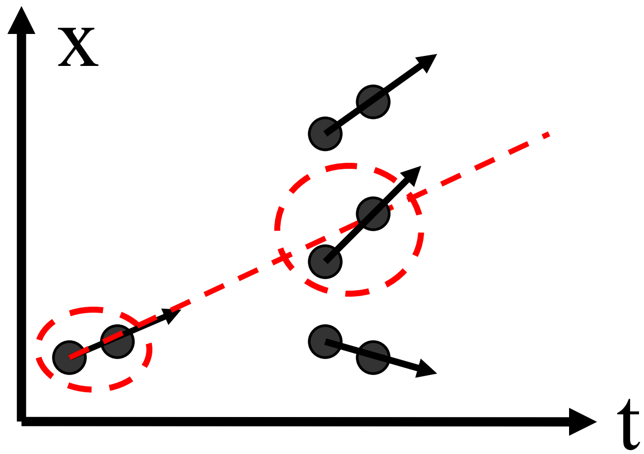
Problem Overview



collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Previous Approaches

- **Look for sets with linear movement over a short time span** (Kristensen 2003, Milani 2004).
- “Close” observations from *same night* linked and used to estimate line (Marsden 1991, Milani 2004).
- Asteroid is projected to later nights and associated with other observations.



- Proposed sets of observations are tested by fitting an orbit.

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Previous Approaches: Drawbacks

1. Linear projections will only be valid over a short time span.

   Accuracy

2. Checking every neighbor can be expensive.

   Cost

3. Orbit fitting is only applied after sets are found with linear approximation.
   – May need to fit many orbits to incorrect sets.

   Cost

   – May incorrectly reject true linkages based on linear model.

   Accuracy

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Initial Improvements

- We can improve accuracy and tractability by using techniques from general target tracking:
  - Sequential tracking,
  - Multiple hypothesis tracker,
  - Use of spatial structure via kd-trees, and
  - Quadratic track models.

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

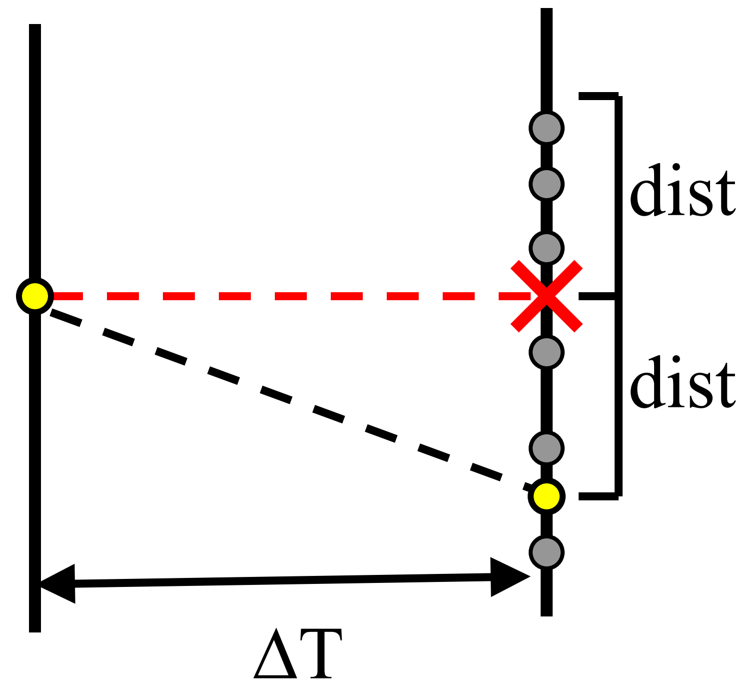
Evaluation

Model	kd-trees?	Time (sec)	Percent Found	Percent Correct
Linear	No	93	96.22	2.06
Linear	Yes	6	96.22	2.06
Quadratic	No	59	96.38	88.67
Quadratic	Yes	3	96.38	88.67

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Why “M-trees” method?

- Sequential approach is **heuristic**. We could end up doing a significant amount of work for “bad pairs”.
- Early associations may be done with incomplete and/or noisy parameters.
- Next observation may be far from predicted position.



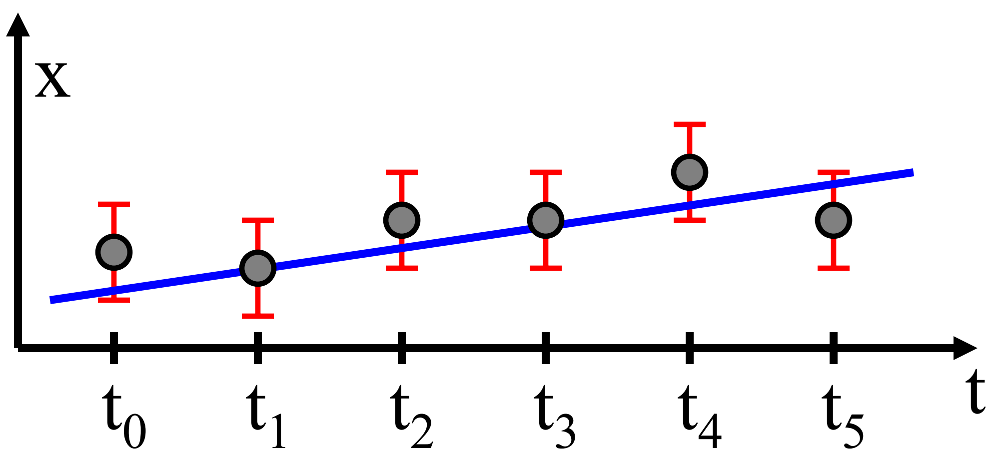
- Problem gets much worse as gap between observations increases.

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Motivation 2: Constrained Feasibility

- Find all tuples of observations such that:
  - We have exactly one observation per time, and
  - a track can exist that passes “near” the observations:

$$\boxed{\zeta^L[d] \le x_i[d] - g(t_i)[d] \le \zeta^H[d]}$$



Can phrase constraints in terms of **only** observation error!

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Feasibility

- "Can ***any*** track exist that is near all of the observations?"
- Each observation's bounds give constraints on track's position at that time:

$$\boxed{\begin{aligned} a[d]t_i^2 + v[d]t_i + p[d] &\geq x_i[d] - \varepsilon \\ a[d]t_i^2 + v[d]t_i + p[d] &\leq x_i[d] + \varepsilon \end{aligned}}$$

- We must either:
  – Find parameters satisfying these equations, OR
  – Prove that no such parameters exist.

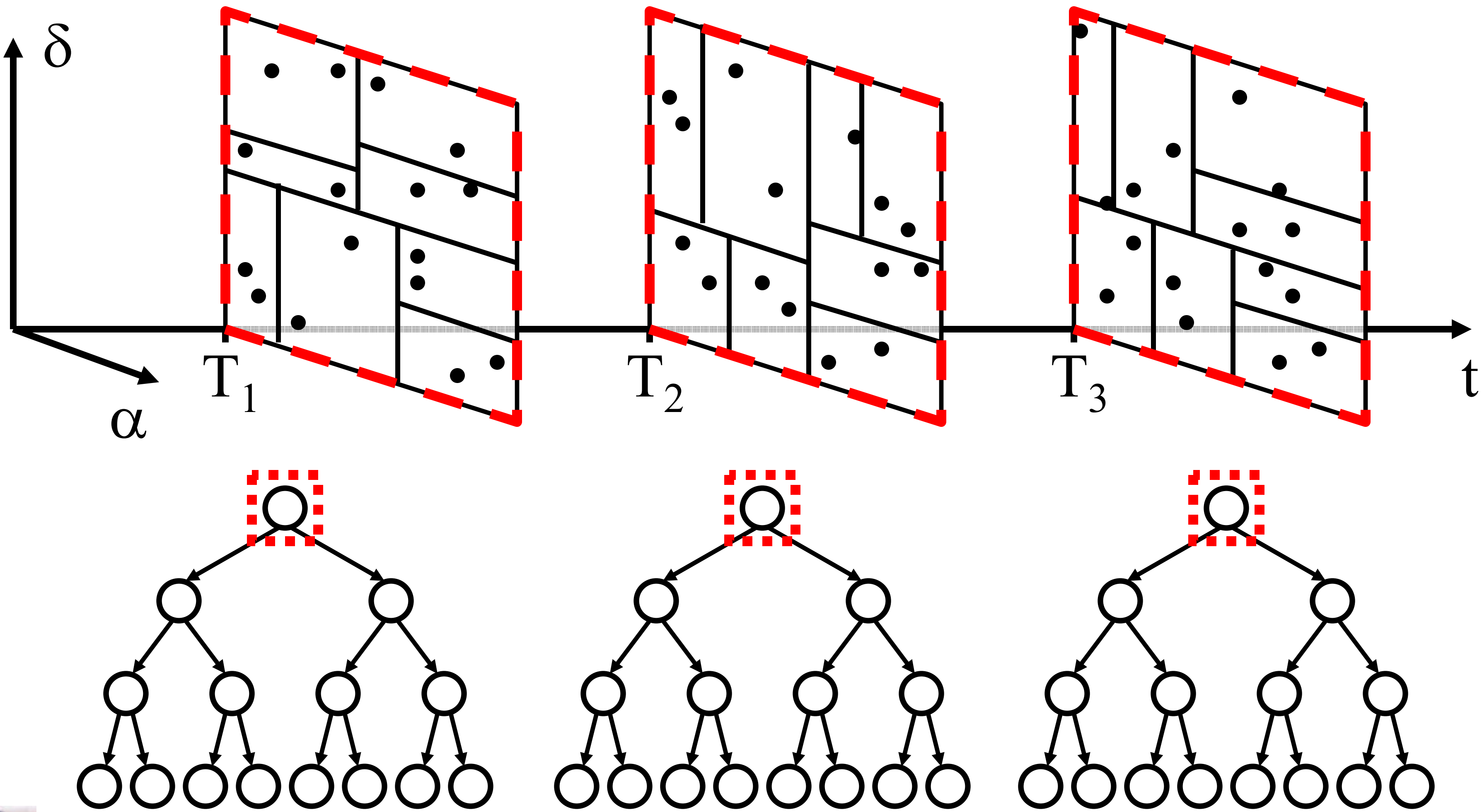
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Approach

- <u>Our approach</u>: Use a multi-tree algorithm (Gray and Moore 2001):
  - Build *multiple* kd-trees over observations.
  - Do a depth first search of *combinations* of tree nodes.

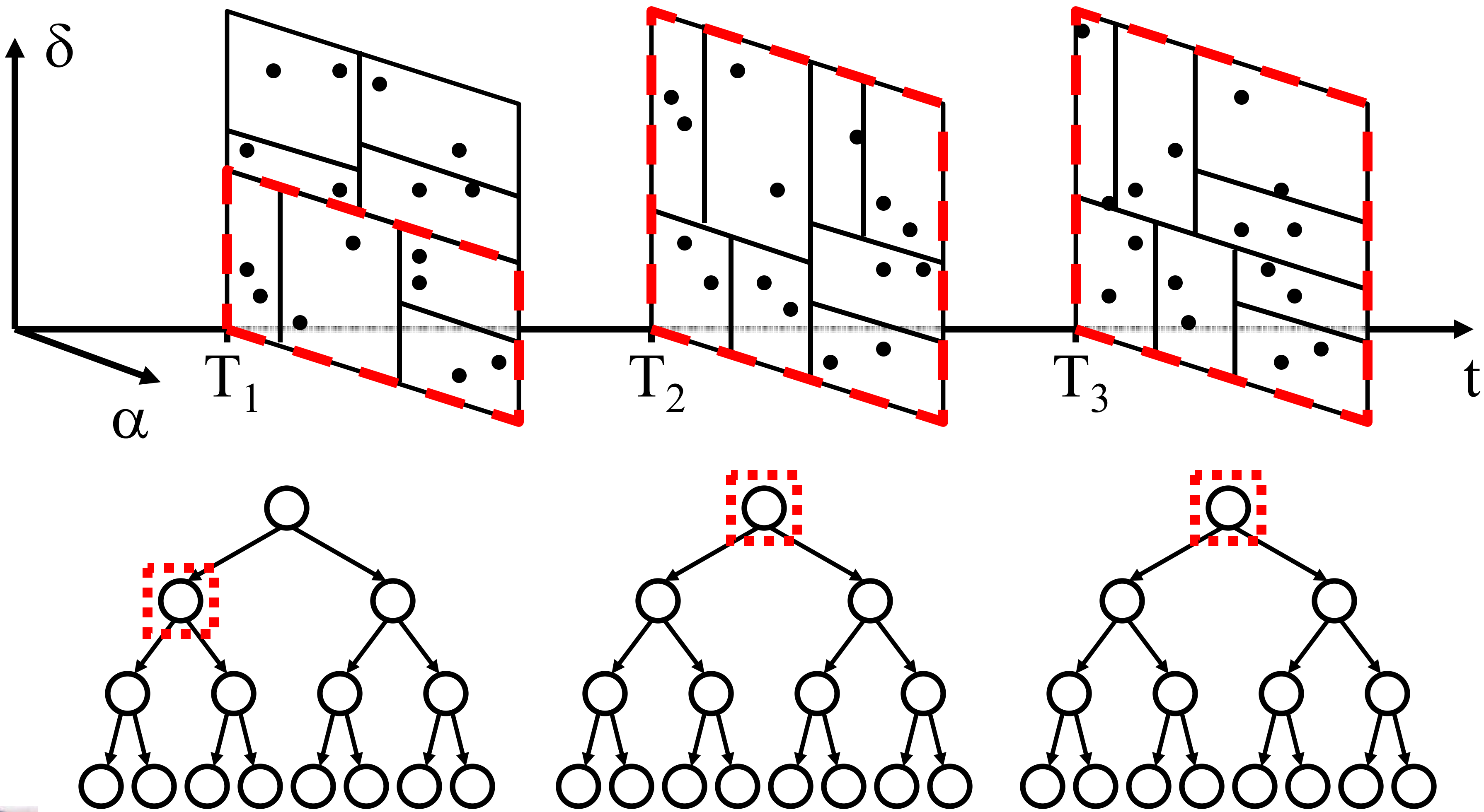
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Depth First Search



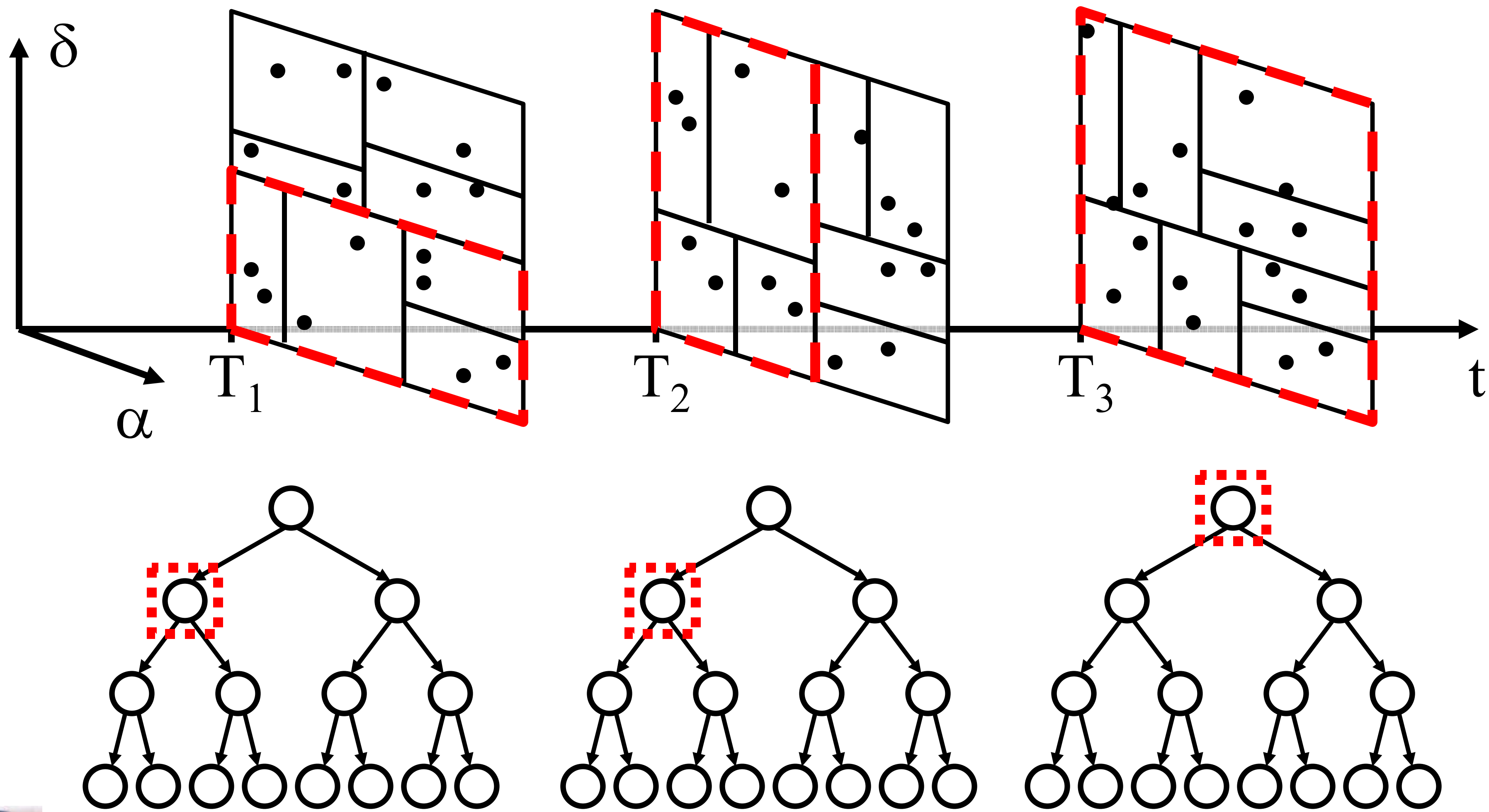
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Depth First Search



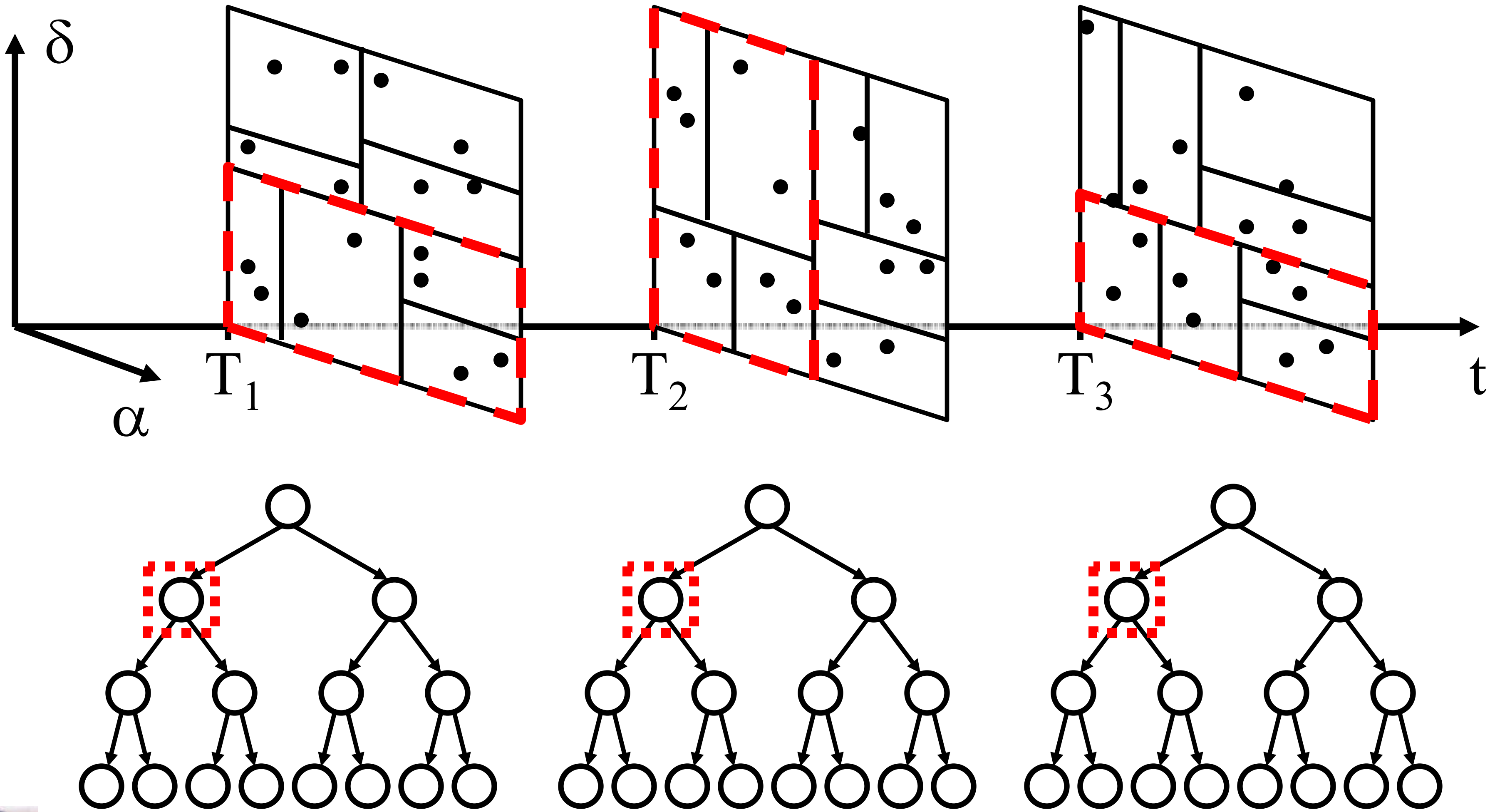
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Depth First Search



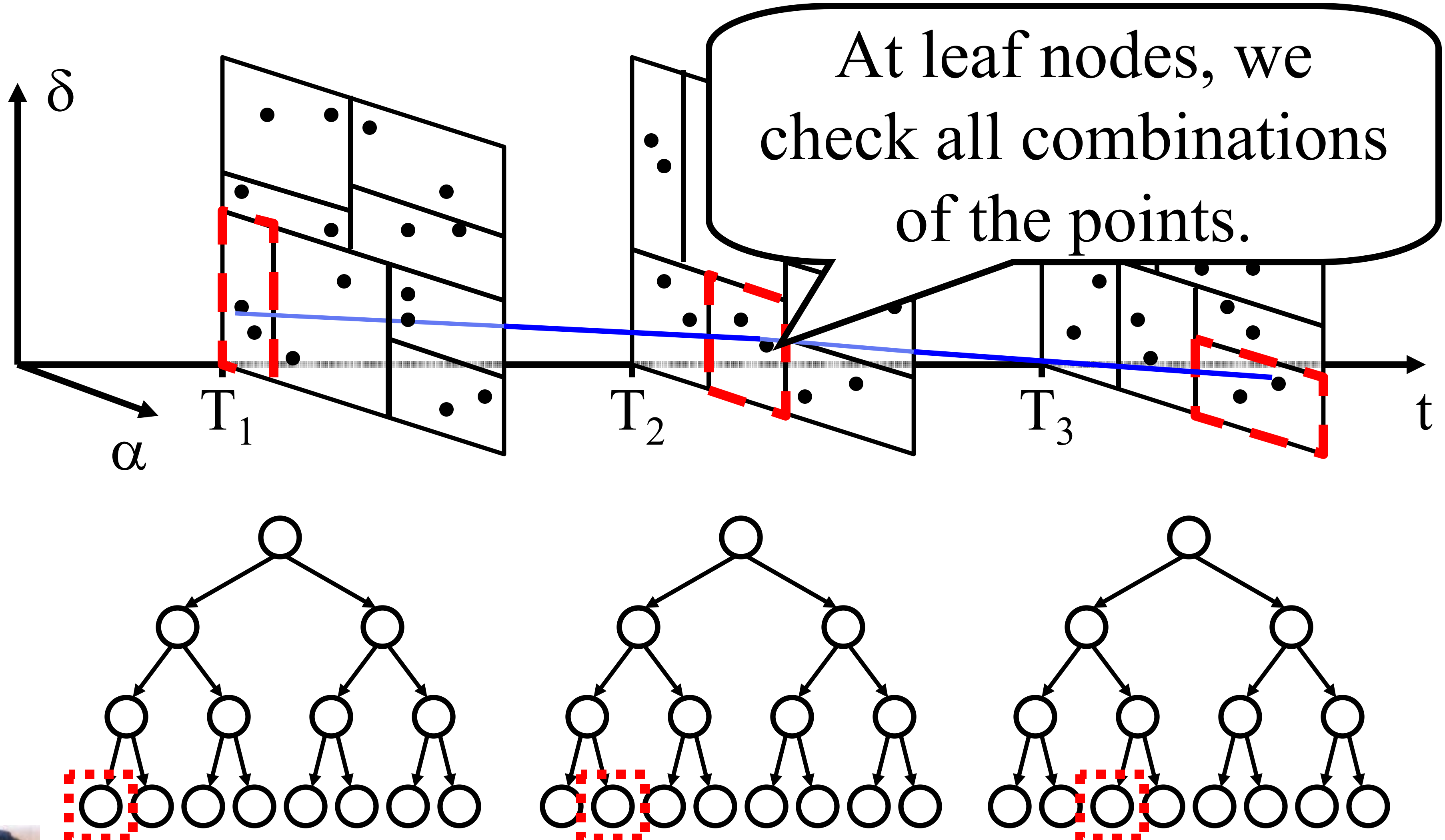
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Depth First Search



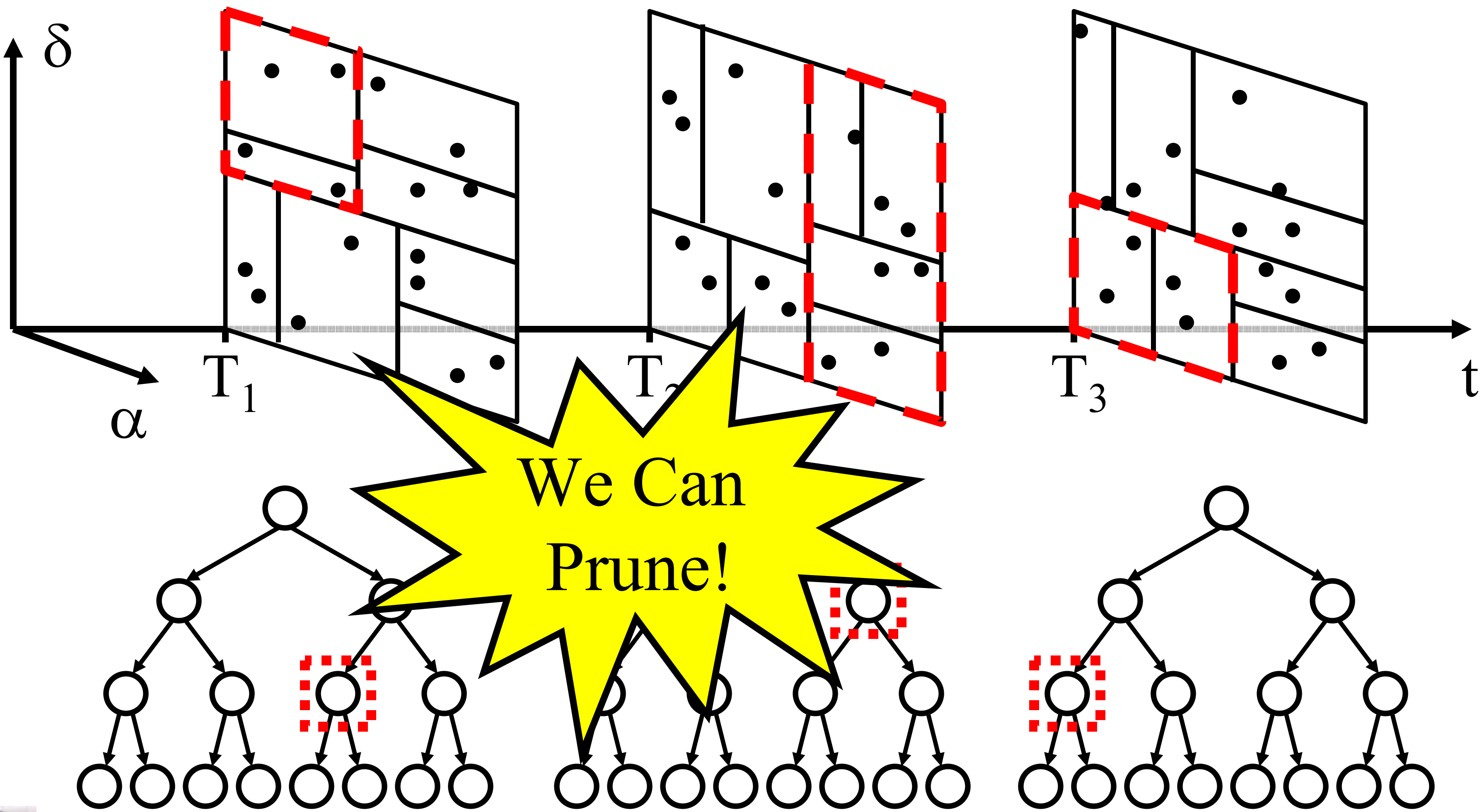
collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Depth First Search



collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Multiple Tree Depth First Search



collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Pruning

- "Can *any* track exist that hits all nodes?"

Given times $t_1, t_2, \ldots t_M$, and given kdtree bounding boxes $(L_1,H_1), (L_2,H_2), \ldots (L_M,H_M)$, at those times, we ask…

$$
\text{“}\ \exists \mathbf{a}, \mathbf{v}, \mathbf{p} . \forall i \in \{1,2,\cdots M\}, \forall d \in \{1,2 \cdots D\},
$$

$$
a[d]t_i^2 + v[d]t_i + p[d] \geq L_i[d] - \varepsilon
$$

$$
a[d]t_i^2 + v[d]t_i + p[d] \leq H_i[d] + \varepsilon \quad \text{”} ?
$$

- Pruning = proving that such parameters do not exist.

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Pruning: Independent Dimensions

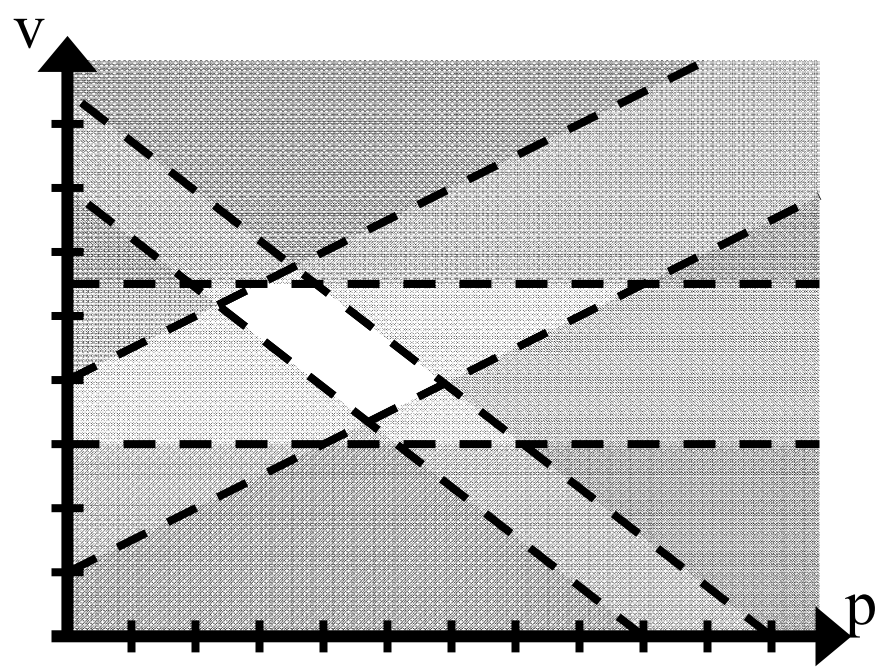
Theorem 1: *(a,v,p) is a feasible track if and only if (a[i],v[i],p[i]) satisfies the constraints in the i-th dimension for all i.*

- Allows us to check the dimensions separately.
- Breaks query on *2MD* constraints into *D* sub-queries of *MD* constraints.
- Each sub-query consists of significantly fewer variables.

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Constraints as Hyper-planes

- Each constraint specifies a C dimensional hyper-plane and half-space in parameter space:



$$H + \varepsilon < vt + p$$

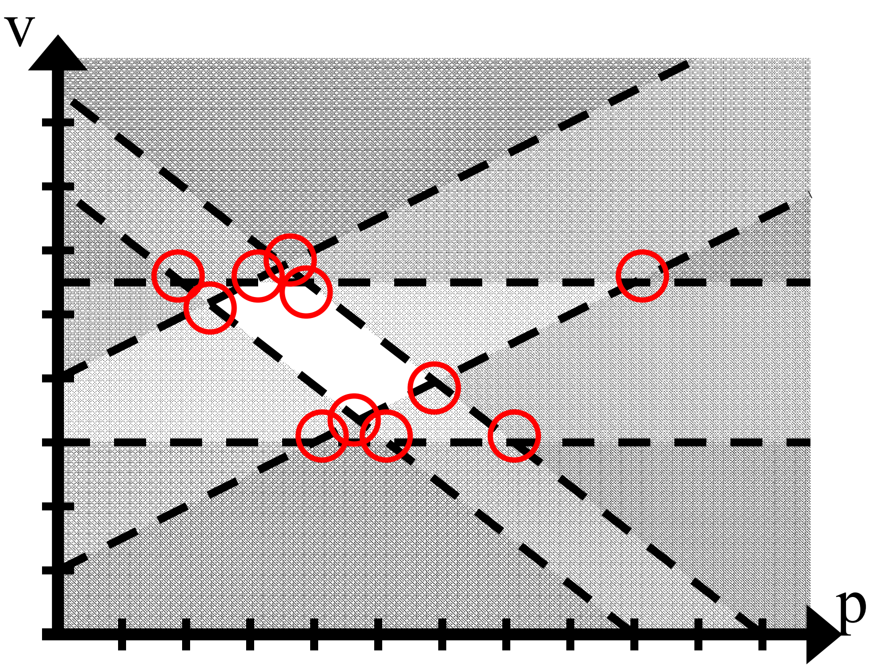
$$p > (-t)v + H + \varepsilon$$

- If the intersection of the feasible half-spaces is not empty, then there exists a track that satisfies all of the constraints.

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Smart Brute Force Search

- Search "corners" of constraint hyper-planes for feasible point.
- C nonparallel C-dimensional hyper-planes intersect at a point ("Corner").

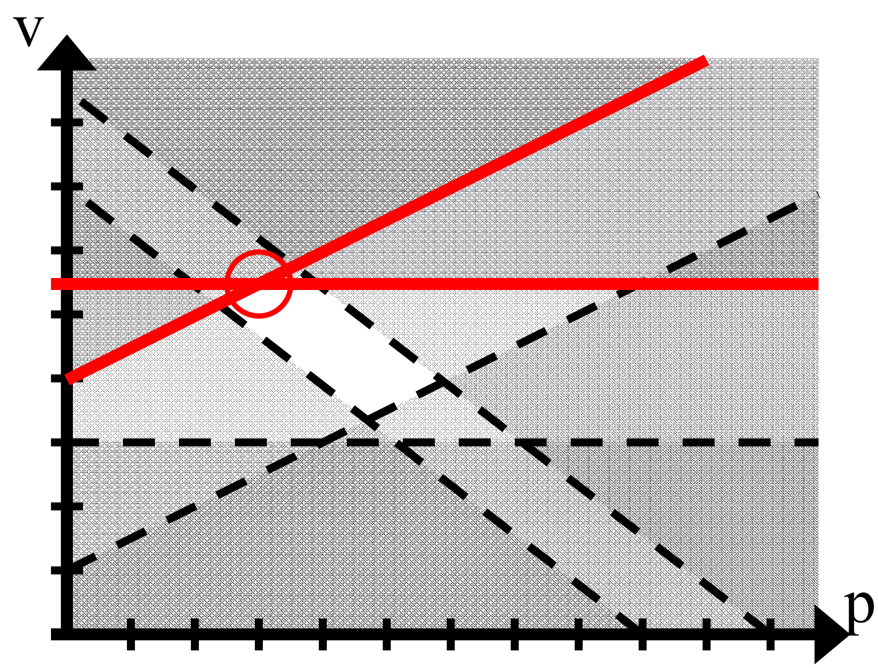


- **Theorem 2**: *The intersection of M half-spaces defined by at least C nonparallel C-dimensional hyper-planes is not empty if and only if there exists a point (a,v,p) such that (a,v,p) is feasible and lies on at least C hyper-planes.*

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Smart Brute Force Search

- For each set of C nonparallel hyper-planes:
  - Calculate the point of intersection.
  - Test point for feasibility against other constraints.



- Positives: Simple, exact
- Negatives: Painfully slow -> $O(DM^{(C+1)})$

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Using Structure In the Search

- The tree search provides a significant amount of structure that can be exploited:
  - At each level of the search, the constraints for all tree nodes except one are identical to the previous level.

We can save the feasible track from previous level and test it against new (tighter) constraints.

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Using Structure In the Search

- The tree search provides a significant amount of structure that can be exploited:
  – At each level of the search, the constraints for all tree nodes except one are identical to the previous level.
  – At each level of the search, the constraints for the one tree node that changed are *tighter* than at the pr

We can look for a new feasible point on hyper-planes from new constraints.

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Using Structure In the Search

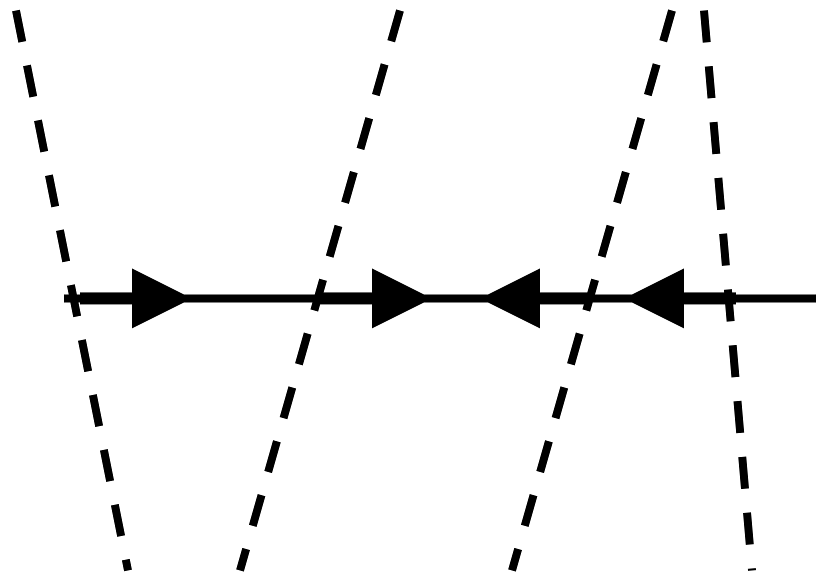
Theorem 3: *If the feasible track from the previous level is not compatible with a new constraint then either the new set of constraints is not compatible* or *a new feasible point lies on the plane defined by the new constraint.*

- Allows us to only check corners containing new constraints -> $O(DM^C)$
- Allows us to check new constraints one at a time.

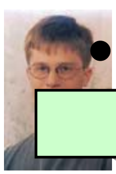
M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Using Structure In the Search

- We can combine search and test steps.
  - C-1 hyper-planes intersect at a line.
  - Remaining hyper-planes intersect the line at *signed* points.



  - There is feasible point on those C-1 constraints if and only if there is a feasible point on the line.



- Reduces cost to $O(DM^{(C-1)})$.

M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Additional Constraints

- This formulation of constraints allows us to add additional (non-node-based) constraints:

$$v_{\min[d]} \leq v[d] \leq v_{\max}[d]$$

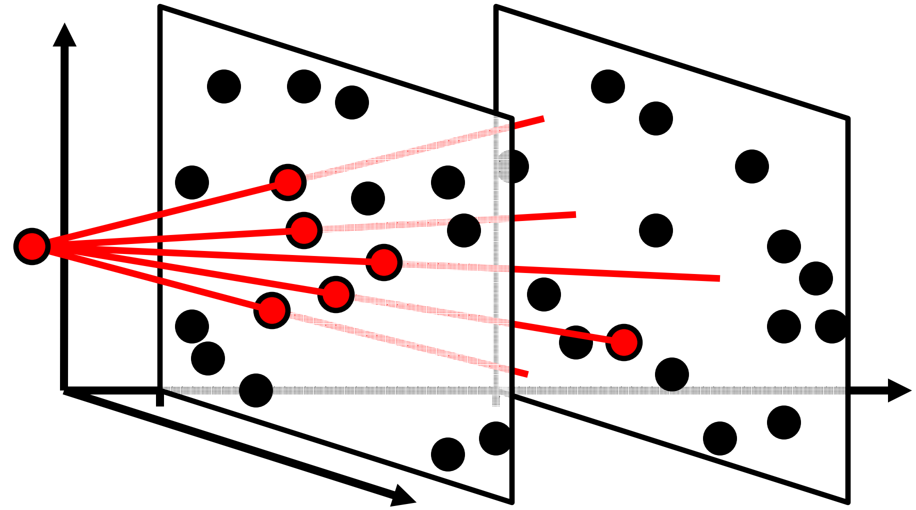
$$a_{\min[d]} \leq a[d] \leq a_{\max}[d]$$

- This allows us to encode additional domain knowledge!

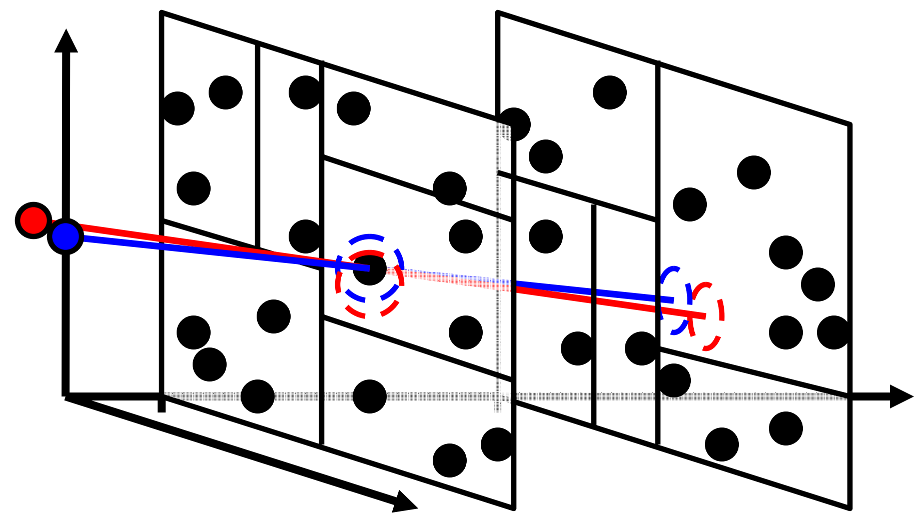
M = Number of timesteps (eg 4-6), D = Number of obs. dim'ns (eg 2), C = # Track params (eg 3)

Multiple Trees: Advantages

- Allows us to consider pruning opportunities resulting from future time-steps.



- Reduces work repeated over similar observations/initial tracks.



collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

Experiments

Experiment	Num Points	Seq secs	Seq P(C)	Singletree secs	Singletree P(C)	V-Tree secs	V-tree P(C)
BIGOBS	205424	66	0.18	31	0.46	15	0.46
Gap134	184016	31	0.07	24	0.83	6	0.90
Gap124	184016	28	0.10	12	0.69	6	0.69
61T.10.10	147244	102	0.3	5	0.77	2	0.77
61T.10.100	187178	451	0.22	7	0.76	7	0.76
61T.10.opp	179090	>2000	?	72	0.03	38	0.03
61T.1af	1433269	>2000	?	213	0.18	66	0.18

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>

For more information and references to related work…

- http://www.autonlab.org/autonweb/14667.html

```
@inproceedings{neill-rectangles,
    Howpublished = {Conference on Knowledge Discovery in Databases (KDD) 2004},
    Month = {August},
    Year = {2004},
    Editor = {J. Guerke and W. DuMouchel},
    Author = {Daniel Neill and Andrew Moore},
    Title = {Rapid Detection of Significant Spatial Clusters}
}
```

- http://www.autonlab.org/autonweb/15868.html

```
@inproceedings{sabhnani-pharmacy,
    Month = {August},
    Year = {2005},
    Booktitle = {Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection},
    Author = {Robin Sabhnani and Daniel Neill and Andrew Moore},
    Title = {Detecting Anomalous Patterns in Pharmacy Retail Data}
}
```

- Software: http://www.autonlab.org/autonweb/10474.html

For more information and references to related work…

- http://www.autonlab.org/autonweb/16063.html @inproceedings{kubicaNIPS05,
  Month = {December},
  Year = {2005},
  Booktitle = {Advances in Neural Information Processing Systems},
  Author = {Jeremy Kubica and Andrew Moore},
  Title = {Variable KD-Tree Algorithms for Spatial Pattern Search and Discovery}
  }

- http://www.autonlab.org/autonweb/14715.html
- @inproceedings{kubicaKDD2005,
  Month = {August},
  Year = {2005},
  Pages = {138-146},
  Publisher = {ACM Press},
  Booktitle = {The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining},
  Author = {Jeremy Kubica and Andrew Moore and Andrew Connolly and Robert Jedicke},
  Title = {A Multiple Tree Algorithm for the Efficient Association of Asteroid Observations}
  }

- http://www.autonlab.org/autonweb/14680.html
- @inproceedings{kubicaSPIE05,
  Month = {August},
  Year = {2005},
  Publisher = {SPIE},
  Booktitle = {Proc. SPIE Signal and Data Processing of Small Targets},
  Editor = {Oliver E. Drummond},
  Author = {Jeremy Kubica and Andrew Moore and Andrew Connolly and Robert Jedicke},
  Title = {Efficiently Identifying Close Track/Observation Pairs in Continuous Timed Data}
  }

Outline

Cached Sufficient Statistics

New searches over cached statistics

Biosurveillance and Epidemiology

Scan Statistics

Cached Scan Statistics

Branch-and-Bound Scan Statistics

Retail data monitoring

Brain monitoring

Entering Google

Asteroids

Multi (and I mean multi) object target tracking

Multiple-tree search

Entering Google

Justifiable Conclusions

Justifiable Conclusions

- Geometry can help tractability of Massive Statistical Data Analysis
- Cached sufficient statistics are one approach
- Not merely for simple friendly aggregates

Justifiable Conclusions

- Geometry can help tractability of Massive Statistical Data Analysis
- Cached sufficient statistics are one approach
- Not merely for simple friendly aggregates

Fluffy Conclusion

"Theorem of Statistical Computation Benevolence"

If Statistics thinks you're going the right way, it will throw in computational opportunities for you

Papers, Software, Example Datasets, Tutorials: www.autonlab.org

For more information and references to related work…

- http://www.autonlab.org/autonweb/14667.html

```
@inproceedings{neill-rectangles,
    Howpublished = {Conference on Knowledge Discovery in Databases (KDD) 2004},
    Month = {August}, Year = {2004},
    Editor = {J. Guerke and W. DuMouchel},
    Author = {Daniel Neill and Andrew Moore},
    Title = {Rapid Detection of Significant Spatial Clusters}
  }
```

- http://www.autonlab.org/autonweb/15868.html

```
@inproceedings{sabhnani-pharmacy,
    Month = {August}, Year = {2005},
    Booktitle = {Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection},
    Author = {Robin Sabhnani and Daniel Neill and Andrew Moore},
    Title = {Detecting Anomalous Patterns in Pharmacy Retail Data}
  }
```

- http://www.autonlab.org/autonweb/16063.html @inproceedings{kubicaNIPS05,

```
    Month = {December},  Year = {2005},
    Booktitle = {Advances in Neural Information Processing Systems},
    Author = {Jeremy Kubica and Andrew Moore},
    Title = {Variable KD-Tree Algorithms for Spatial Pattern Search and Discovery}
  }
```

- http://www.autonlab.org/autonweb/14715.html
- @inproceedings{kubicaKDD2005,

```
    Month = {August},  Year = {2005},
    Pages = {138-146},
    Publisher = {ACM Press},
    Booktitle = {The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining},
    Author = {Jeremy Kubica and Andrew Moore and Andrew Connolly and Robert Jedicke},
    Title = {A Multiple Tree Algorithm for the Efficient Association of Asteroid Observations}
  }
```

- http://www.autonlab.org/autonweb/14680.html
- @inproceedings{kubicaSPIE05,

```
    Month = {August},Year = {2005}, Publisher = {SPIE},
    Booktitle = {Proc. SPIE Signal and Data Processing of Small Targets},
    Editor = {Oliver E. Drummond},
    Author = {Jeremy Kubica and Andrew Moore and Andrew Connolly and Robert Jedicke},
    Title = {Efficiently Identifying Close Track/Observation Pairs in Continuous Timed Data}
  }
```


- Software: http://www.autonlab.org/autonweb/10474.html