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
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

<table>
<thead>
<tr>
<th>Size</th>
<th>Ellipticity</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>0.96</td>
<td>Red</td>
</tr>
<tr>
<td>33</td>
<td>0.55</td>
<td>Red</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>Green</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Analysis: The new days

Question

Seventeen months later…

Answer
Cached Sufficient Statistics

Question
Cached Sufficient Statistics

Question

Answer

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)
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
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
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 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
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 models:
  - 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 models:
  - of the null hypothesis \( H_0: \) no attacks.
  - of the alternative hypotheses \( H_1(S): \) attack in region \( S. \)

- Derive a score function \( \text{Score}(S) = \text{Score}(C, B). \)
  - 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: $Score(S) = \frac{L(Data \mid H_1(S))}{L(Data \mid H_0)}$
  – To find the most significant region: $S^* = \arg \max_S Score(S)$

Example: Kulldorf’s score

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

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

- To find the most significant region:
  
  $S^* = \arg \max_S Score(S)$

Example: Kulldorf’s score

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

$H_0$: $q = q_{all}$ 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

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%)

[Score = 9.3]

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

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...

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

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

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

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

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

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

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

Data is aggregated to a grid.

<table>
<thead>
<tr>
<th>B=25</th>
<th>B=18</th>
<th>B=22</th>
<th>B=14</th>
<th>B=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=27</td>
<td>C=14</td>
<td>C=22</td>
<td>C=15</td>
<td>C=5</td>
</tr>
<tr>
<td>B=25</td>
<td>B=20</td>
<td>B=6</td>
<td>B=20</td>
<td>B=5</td>
</tr>
<tr>
<td>C=26</td>
<td>C=17</td>
<td>C=9</td>
<td>C=12</td>
<td>C=4</td>
</tr>
<tr>
<td>B=25</td>
<td>B=25</td>
<td>B=20</td>
<td>B=15</td>
<td>B=20</td>
</tr>
<tr>
<td>C=19</td>
<td>C=26</td>
<td>C=43</td>
<td>C=37</td>
<td>C=20</td>
</tr>
<tr>
<td>B=24</td>
<td>B=24</td>
<td>B=19</td>
<td>B=15</td>
<td>B=19</td>
</tr>
<tr>
<td>C=18</td>
<td>C=20</td>
<td>C=40</td>
<td>C=32</td>
<td>C=16</td>
</tr>
<tr>
<td>B=23</td>
<td>B=15</td>
<td>B=14</td>
<td>B=10</td>
<td>B=2</td>
</tr>
<tr>
<td>C=20</td>
<td>C=17</td>
<td>C=8</td>
<td>C=10</td>
<td>C=3</td>
</tr>
</tbody>
</table>

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: $O(1)$

<table>
<thead>
<tr>
<th>B=25</th>
<th>B=18</th>
<th>B=22</th>
<th>B=14</th>
<th>B=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=27</td>
<td>C=14</td>
<td>C=22</td>
<td>C=15</td>
<td>C=5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B=25</th>
<th>B=20</th>
<th>B=6</th>
<th>B=20</th>
<th>B=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=26</td>
<td>C=17</td>
<td>C=9</td>
<td>C=12</td>
<td>C=4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B=25</th>
<th>B=25</th>
<th>B=20</th>
<th>B=15</th>
<th>B=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=19</td>
<td>C=26</td>
<td>C=43</td>
<td>C=37</td>
<td>C=20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B=24</th>
<th>B=24</th>
<th>B=19</th>
<th>B=15</th>
<th>B=19</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=18</td>
<td>C=20</td>
<td>C=40</td>
<td>C=32</td>
<td>C=16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B=23</th>
<th>B=15</th>
<th>B=14</th>
<th>B=10</th>
<th>B=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=20</td>
<td>C=17</td>
<td>C=8</td>
<td>C=10</td>
<td>C=3</td>
</tr>
</tbody>
</table>
Computational framework

Data is aggregated to a grid.

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

$n \times n$ grid has

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

rectangles to search

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

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

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%)

[Score = 9.3]

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

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 (e.g. Kulldorf’s score)

Collaboration with Daniel Neill <neill@cs.cmu.edu>
Many Steps 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)

Highest scoring region in search so far

[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.

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>
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'

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>
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>
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”

collaboration with Daniel Neill <neill@cs.cmu.edu>
Step 1: Gridded

Check a specific recursive overlapping set of regions called ‘Gridded Regions’

collaboration with Daniel Neill <neill@cs.cmu.edu>
Step 1: Gridded

Check a specific recursive overlapping set of regions called ‘Gridded Regions’

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

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.
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 right gridded child
- ...entirely inside my bottom 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

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.
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?
If $S'$ is a middle-containing subregion of $S$...

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

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

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

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

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

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

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

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

$$\text{Score}(S') = \text{Score}(\text{count}(S'), \text{baseline}(S'))$$

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 $C$

\[
\text{Score}(S') = \text{Score}(\text{count}(S'), \text{baseline}(S'))
\]

5. Is it possible that any “Type 5” subregion of “Gridded Region” could score better than best known score to date?
Properties of $D(S)$

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

$\text{Score}(S)$ increases with the total count of $S$, $C(S) = \sum S c_i$. 

\[\text{Pop} \ 1000 \quad \text{Count} \ 5\]

\[\text{Pop} \ 1000 \quad \text{Count} \ 500\]

collaboration with Daniel Neill <neill@cs.cmu.edu>
Properties of D(S)

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

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

collaboration with Daniel Neill  <neill@cs.cmu.edu>
Properties of $D(S)$

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

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

$$\text{Score}(S') = \text{Score}(\text{count}(S'), \text{baseline}(S'))$$

Possible that any “Type 5” subregion of a “Nested Region” could score better than best known score to date?

Bottom Line: all the above lets us put a good upper bound on $\text{Score}(S')$
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?
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?
Why the Scan Statistic speed obsession?

• Traditional Scan Statistics very expensive, especially with Randomization tests
• Going national
• A few hours could actually matter!
Which regions to search?

- We choose to search over the space of all rectangular regions.
- We typically expect clusters to be convex; thus inner/outer bounding boxes are reasonably close approximations to shape.
- We can find clusters with high aspect ratios.
  - Important in epidemiology since disease clusters are often elongated (e.g., from windborne 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 $2^d$ 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 $S_1\ldots S_{2^d}$.
  - $S'$ 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
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
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](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](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](http://www.autonlab.org/autonweb/10474.html)
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
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

Outline

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

**Ultimate Goal**: 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 ($\alpha$, $\delta$)
- We do not see the distance ($r$).
Why Is This Hard/Interesting?

Temporally sparse:
• Each region viewed infrequently.
• Each viewing only covers a fraction of the sky.

~4 days ~30 min.

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>
Problem Overview

Asteroid Tracking

Observation/ Detection/ Filtering → Identifying Known Objects (attribution) → Finding New Objects (linkage/track initiation)

Orbit Fitting

“Tracking”:
- Prediction
- Data Association
- Maintenance

Collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>
Problem Overview

Asteroid Tracking

Observation/ Detection/ Filtering

Identifying Known Objects (attribution)

Finding New Objects (linkage/track initiation)

Orbit Fitting

“Tracking”: • Prediction • Data Association • Maintenance

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

Asteroid Tracking

Identifying Known Objects (attribution)

Finding New Objects (linkage/track initiation)

"Tracking":
- Prediction
- Data Association
- Maintenance

This is just Halley’s comet!

This is just Pluto!

Orbit Fitting

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

Asteroid Tracking

Observation/Detection/

Identifying Known Objects (attribution)

Finding New Objects (linkage/track initiation)

Orbit Fitting

“Tracking”: • Prediction • Data Association • Maintenance

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!

This looks like it could be the start of a track!
Problem Overview

Asteroid Tracking

Observation/

Identifying Known Objects (attribution)

Finding New Objects (linkage/track initiation)

Orbit Fitting

“Tracking”: • Prediction • Data Association • Maintenance

The track continues over here.

The text is not clearly legible due to the image quality.
Problem Overview

Asteroid Tracking

Observation/Detection/Filtering → Identifying Known Objects (attribution) → Finding New Objects (linkage/track initiation)

Orbit Fitting

“Tracking”:
- Prediction
- Data Association
- Maintenance

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

Asteroid Tracking

Observation/Detection/Filtering → Identifying Known Objects (attribution) → Finding New Objects (linkage/track initiation)

Initial Linkage and Tracking Algorithm:
Established techniques in astronomy and techniques from general target tracking.

“Tracking”:
• Prediction
• Data Association
• Maintenance

(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.

2. Checking every neighbor can be expensive.

3. Orbit fitting is only applied after sets are found with linear approximation.
   - May need to fit many orbits to incorrect sets.
   - May incorrectly reject true linkages based on linear model.
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.
## Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>kd-trees?</th>
<th>Time (sec)</th>
<th>Percent Found</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>No</td>
<td>93</td>
<td>96.22</td>
<td>2.06</td>
</tr>
<tr>
<td>Linear</td>
<td>Yes</td>
<td>6</td>
<td>96.22</td>
<td>2.06</td>
</tr>
<tr>
<td>Quadratic</td>
<td>No</td>
<td>59</td>
<td>96.38</td>
<td>88.67</td>
</tr>
<tr>
<td>Quadratic</td>
<td>Yes</td>
<td>3</td>
<td>96.38</td>
<td>88.67</td>
</tr>
</tbody>
</table>
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:

\[
\zeta^L[d] \leq x_i[d] - g(t_i)[d] \leq \zeta^H[d]
\]

Can phrase constraints in terms of only observation error!
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:

\[
a[d]t_i^2 + v[d]t_i + p[d] \geq x_i[d] - \varepsilon
\]
\[
a[d]t_i^2 + v[d]t_i + p[d] \leq x_i[d] + \varepsilon
\]

• We must either:
  – Find parameters satisfying these equations, OR
  – Prove that no such parameters exist.
Multiple Tree Approach

Our approach: 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
Multiple Tree Depth First Search

At leaf nodes, we check all combinations of the points.

δ

α

t

T1 T2 T3

δ

α

t

T1 T2 T3

collaboration with Jeremy Kubica <jkubica@cs.cmu.edu>
Multiple Tree Depth First Search

We Can Prune!

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…

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

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

\[ a[d]t^2_i + v[d]t_i + p[d] \leq H_i[d] + \varepsilon \]

• 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.

\[M = \text{Number of timesteps (eg 4-6)}, \quad D = \text{Number of obs. dim’ns (eg 2)}, \quad C = \text{# Track params (eg 3)}\]
Constraints as Hyper-planes

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

\[
H + \varepsilon < vt + p \\
\downarrow \\
p > (-t)v + H + \varepsilon
\]

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

\[
M = \text{Number of timesteps (eg 4-6)}, \quad D = \text{Number of obs. dim’ns (eg 2)}, \quad C = \# \text{ Track params (eg 3)}
\]
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 hyperplanes:
  - Calculate the point of intersection.
  - Test point for feasibility against other constraints.
- **Positives:** Simple, exact
- **Negatives:** Painfully slow \( \rightarrow O(DM^{(C+1)}) \)

\[
M = \text{Number of timesteps (eg 4-6)}, \quad D = \text{Number of obs. dim’ns (eg 2)}, \quad C = \# \text{ Track params (eg 3)}
\]
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 previous level.

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

$M = \text{Number of timesteps (eg 4-6)}, \ D = \text{Number of obs. dim’ns (eg 2)}, \ C = \text{# Track params (eg 3)}$
Additional Constraints

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

\[ v_{\text{min}[d]} \leq v[d] \leq v_{\text{max}[d]} \]
\[ a_{\text{min}[d]} \leq a[d] \leq a_{\text{max}[d]} \]

• 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

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Num Points</th>
<th>Seq secs</th>
<th>Seq P(C)</th>
<th>Singletree secs</th>
<th>Singletree P(C)</th>
<th>V-Tree secs</th>
<th>V-tree P(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIGOBS</td>
<td>205424</td>
<td>66</td>
<td>0.18</td>
<td>31</td>
<td>0.46</td>
<td>15</td>
<td>0.46</td>
</tr>
<tr>
<td>Gap134</td>
<td>184016</td>
<td>31</td>
<td>0.07</td>
<td>24</td>
<td>0.83</td>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>Gap124</td>
<td>184016</td>
<td>28</td>
<td>0.10</td>
<td>12</td>
<td>0.69</td>
<td>6</td>
<td>0.69</td>
</tr>
<tr>
<td>61T.10.10</td>
<td>147244</td>
<td>102</td>
<td>0.3</td>
<td>5</td>
<td>0.77</td>
<td>2</td>
<td>0.77</td>
</tr>
<tr>
<td>61T.10.100</td>
<td>187178</td>
<td>451</td>
<td>0.22</td>
<td>7</td>
<td>0.76</td>
<td>7</td>
<td>0.76</td>
</tr>
<tr>
<td>61T.10.opp</td>
<td>&gt;2000</td>
<td>?</td>
<td>?</td>
<td>72</td>
<td>0.03</td>
<td>38</td>
<td>0.03</td>
</tr>
<tr>
<td>61T.1af</td>
<td>1433269</td>
<td>&gt;2000</td>
<td>?</td>
<td>213</td>
<td>0.18</td>
<td>66</td>
<td>0.18</td>
</tr>
</tbody>
</table>
For more information and references to related work…

- [http://www.autonlab.org/autonweb/14667.html](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](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](http://www.autonlab.org/autonweb/10474.html)
For more information and references to related work…

- [http://www.autonlab.org/autonweb/16063.html](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](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](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}
}
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](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](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](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](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](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](http://www.autonlab.org/autonweb/10474.html)