Fast Generalized Subset Scan for Anomalous Pattern Detection

Event & Pattern Detection Lab
H. John Heinz III College
Carnegie Mellon University
Edward McFowland III (mcfowland@cmu.edu)
Skyler Speakman (speakman@cmu.edu)
Daniel B. Neill (neill@cs.cmu.edu)

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330

Motivation

- Anomalous Pattern Detection
  - Detecting the data that were generated from an anomalous process
  - This data is self-similar and as a group different from rest of the data

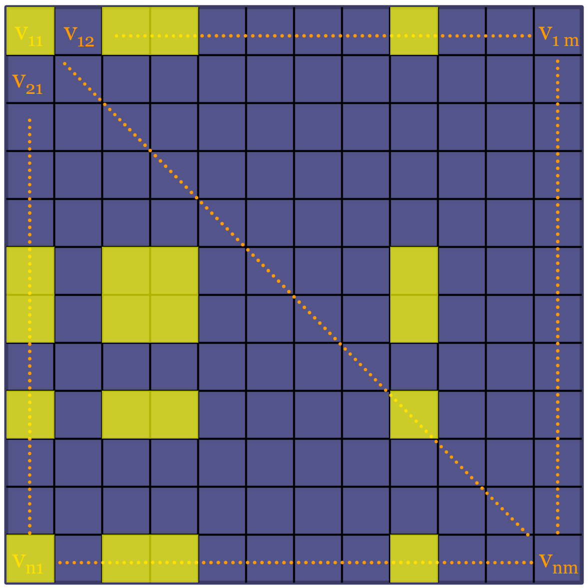
Pattern Detection in Health Domains

- Disease Surveillance
- Fraud Detection
- Anomalous Patterns of Care
- …And much more

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



I. Compute the anomalousness of each attribute value (for each record)

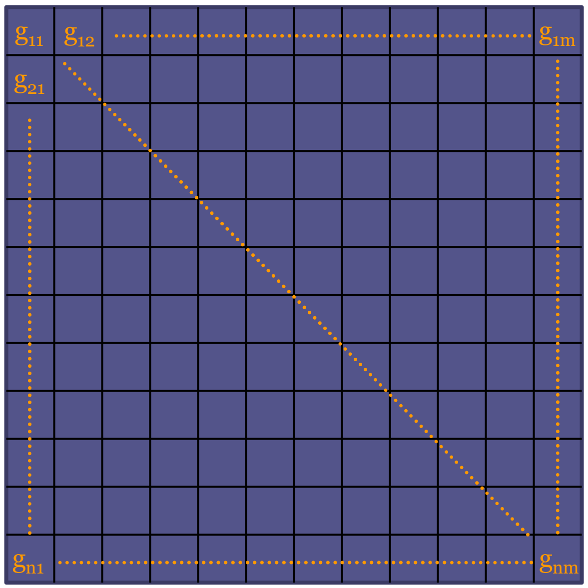
II. Discover subsets of records and attributes that are most anomalous

We propose a method, FGSS, for anomalous pattern detection in general datasets

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

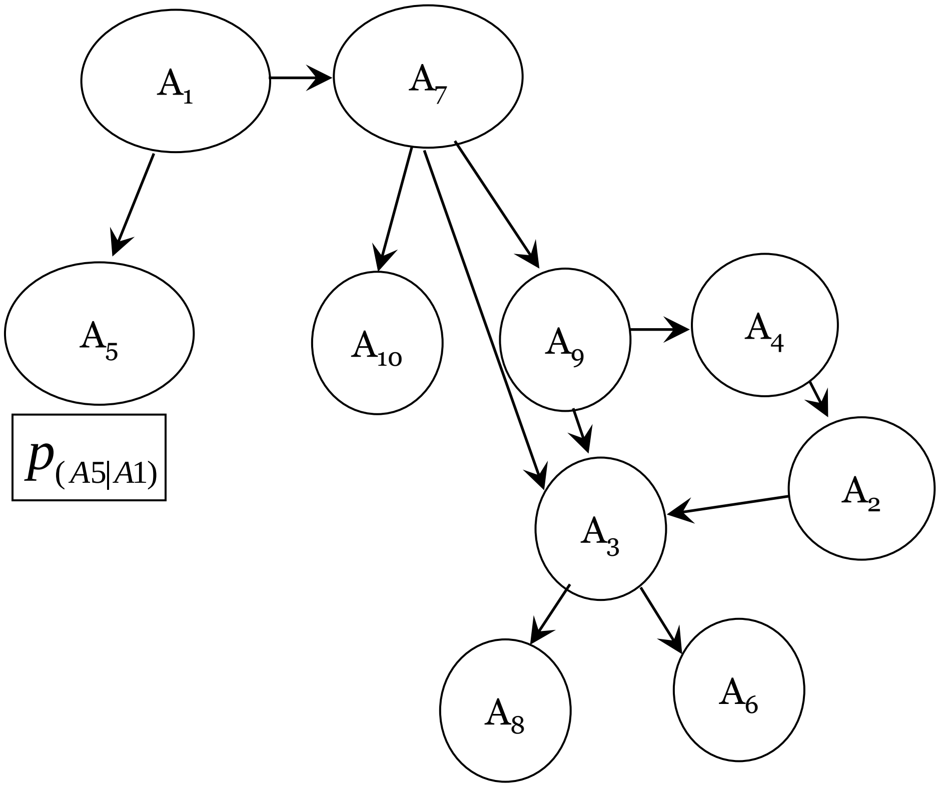
Records $R_1 \ldots R_N$



I. Compute the anomalousness of each attribute (for each record)

In order to compute the anomalousness of the data, we model the data distribution under expected system behavior

Fast Generalized Subset Scan (FGSS)

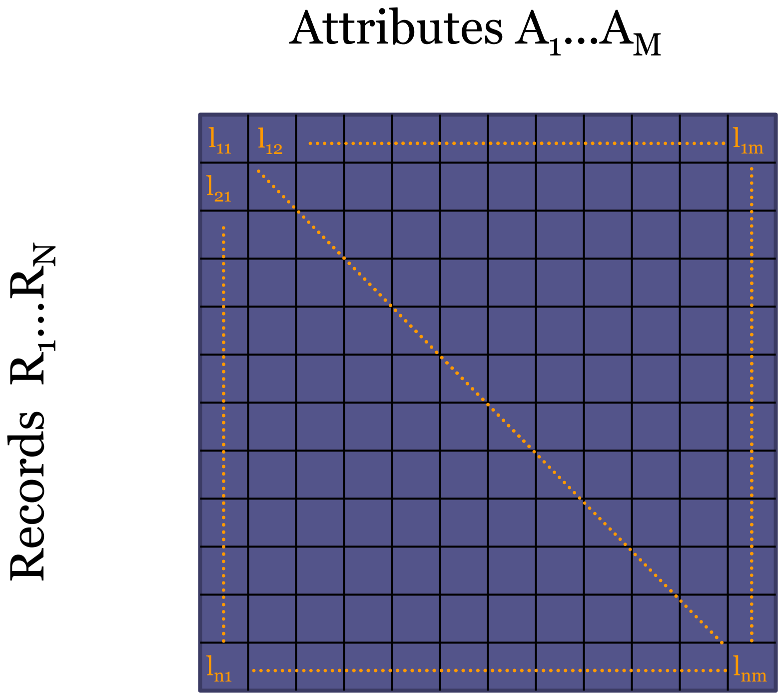


I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network

Learn a Bayesian Network representing the conditional probability distribution of each attribute (given the others) under the assumption that there are no events of interest

Fast Generalized Subset Scan (FGSS)



I. Compute the anomalousness of each attribute (for each record)

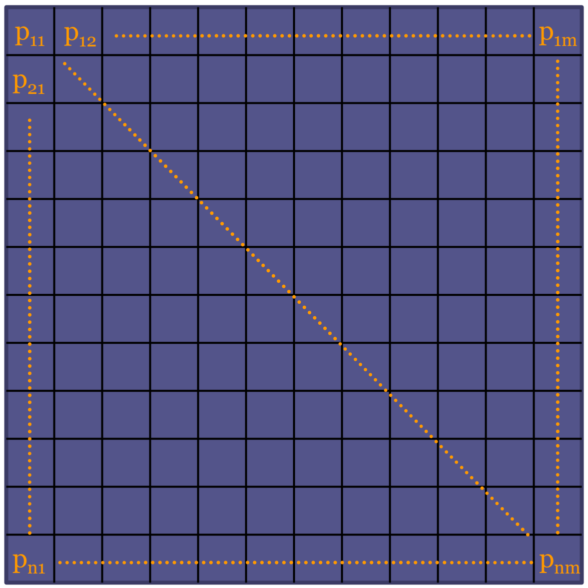
1. Learn Bayesian Network
2. Compute attribute value likelihoods

By performing inference on the Bayesian Network, for each record we can determine the likelihood of each of its attribute values

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$

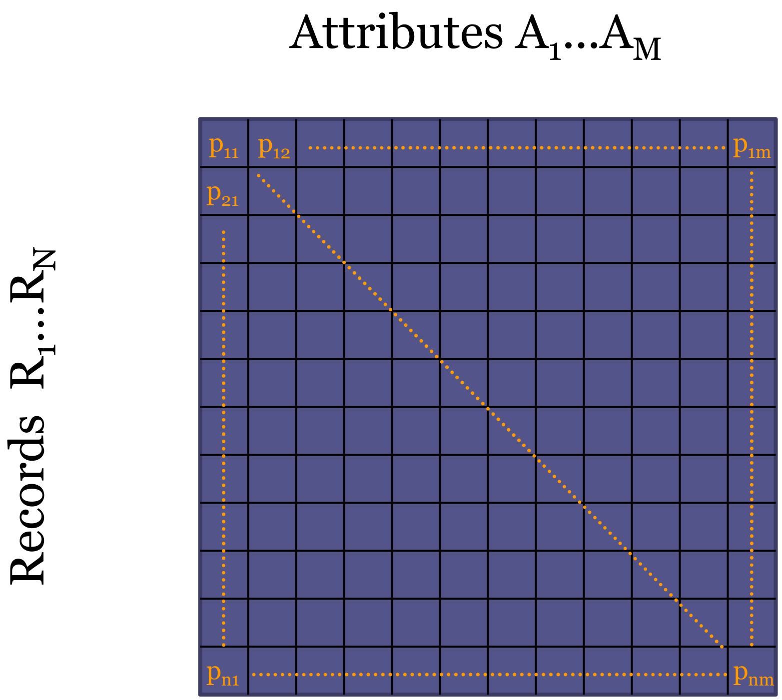


I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values
   - i. maps each attribute distribution to same space
   - ii. p_{ij} in S ~ Uniform(0,1) under H_0

Empirical p-values are a measure, mapped onto the interval [0,1] , of how surprising each attribute value is given the model of normal system behavior

Fast Generalized Subset Scan (FGSS)



I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

Subsets of data with a higher than expected quantities of significantly low p-values are possibly indicative of an anomalous process

Fast Generalized Subset Scan (FGSS)

Nonparametric Scan Statistic (NPSS)

$$F(S) = \max_{\alpha} F(S) = \max_{\alpha} F_{\alpha}(N_{\alpha}, N)$$

$$N_{\alpha} = |\{p_{ij} \in S : p_{ij} \leq \alpha\}|$$

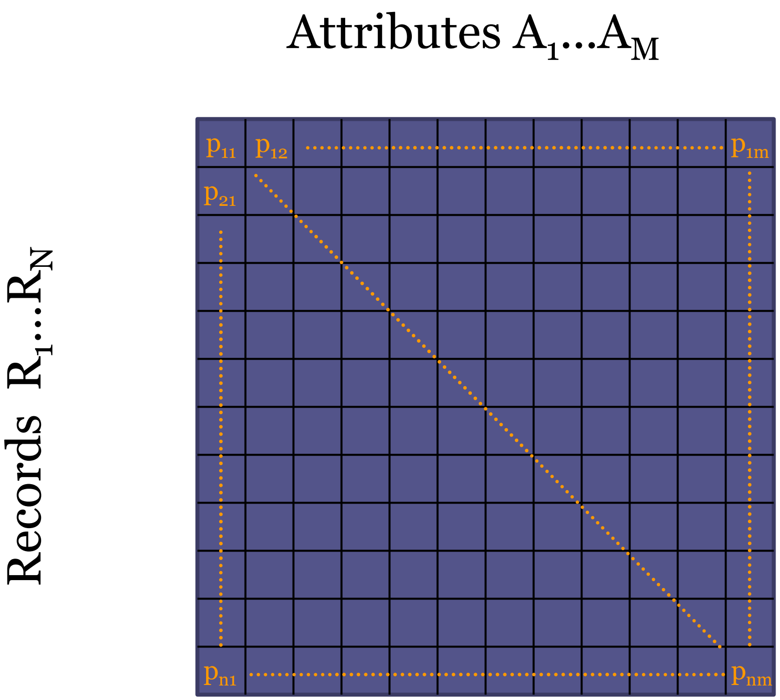
$$N_{tot} = |\{p_{ij} \in S\}|$$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   - Evaluate subsets with NPSS

NPSS quantifies how dissimilar the distribution of emperical p-values in S are from Uniform(0,1)

Fast Generalized Subset Scan (FGSS)



I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Naïve search is infeasible $O(2^{N+M})$

Search over all possible subsets of records' p-value ranges and find the maximizing F(S)

Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff :

$$\max_{S \subseteq D} F(S) = \max_{i=1 \ldots N} F\left(\left\{R_{(1)} \ldots R_{(i)}\right\}\right)$$

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Naïve search is infeasible $O(2^{N+M})$

We can reduce the search over records from $O(2^N)$ to $O(N \log N)$

Fast Generalized Subset Scan (FGSS)

<u>Linear Time Subset Scanning Property (LTSS)</u>

A F(S) satisfies LTSS iff :

$$\max_{S \subseteq D} F(S) = \max_{i=1 \ldots N} F\left(\left\{R_{(1)} \ldots R_{(i)}\right\}\right)$$

We only need to consider:

$\{R_{(1)}\}$
$\{R_{(1)}, R_{(2)}\}$
$\{R_{(1)}, R_{(2)}, R_{(3)}\}$
$\vdots$
$\{R_{(1)}, \ldots\ldots\ldots\ldots\ldots, R_{(n)}\}$

We can reduce the search over records from $O(2^N)$ to $O(N \log N)$

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Naïve search is infeasible $O(2^{N+M})$
   - NPSS satisfies LTSS with:

$$F(S) = \max_{\alpha} F_{\alpha}(N_{\alpha}, N_{tot})$$

Fast Generalized Subset Scan (FGSS)

<u>Linear Time Subset Scanning Property (LTSS)</u>

A F(S) satisfies LTSS iff :

$$\max_{S \subseteq D} F(S) = \max_{i=1 \ldots M} F\left(\left\{A_{(1)} \ldots A_{(i)}\right\}\right)$$

We only need to consider:

$$\{A_{(1)}\}$$
$$\{A_{(1)}, A_{(2)}\}$$
$$\{A_{(1)}, A_{(2)}, A_{(3)}\}$$
$$\vdots$$
$$\{A_{(1)}, \ldots\ldots\ldots\ldots, A_{(n)}\}$$

We want to maximize of subsets of records AND attributes; Observe F(S) is only a function of p_{ij}, thus we can use LTSS to also maximize over the attributes

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Naïve search is infeasible $O(2^{N+M})$
   - NPSS satisfies LTSS with:

$$F(S) = \max_{\alpha} F_{\alpha}(N_{\alpha}, N_{tot})$$

Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff :

$$\max_{S \subseteq D} F(S) = \max_{i=1 \ldots M} F\left(\left\{A_{(1)} \ldots A_{(i)}\right\}\right)$$

We only need to consider:

$$\{A_{(1)}\}$$
$$\{A_{(1)}, A_{(2)}\}$$
$$\{A_{(1)}, A_{(2)}, A_{(3)}\}$$
$$\vdots$$
$$\{A_{(1)}, \ldots\ldots\ldots\ldots\ldots, A_{(n)}\}$$

We can iterate between maximizing over the records and maximizing over the attributes

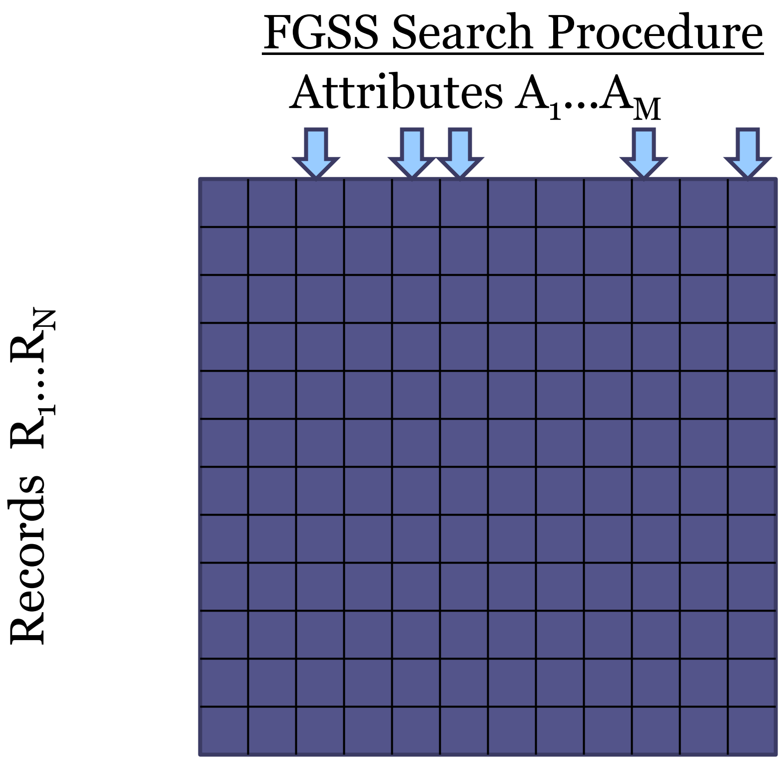
I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - LTSS over records O(N log N)
   - LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)



1. Start with a randomly chosen subset of attributes

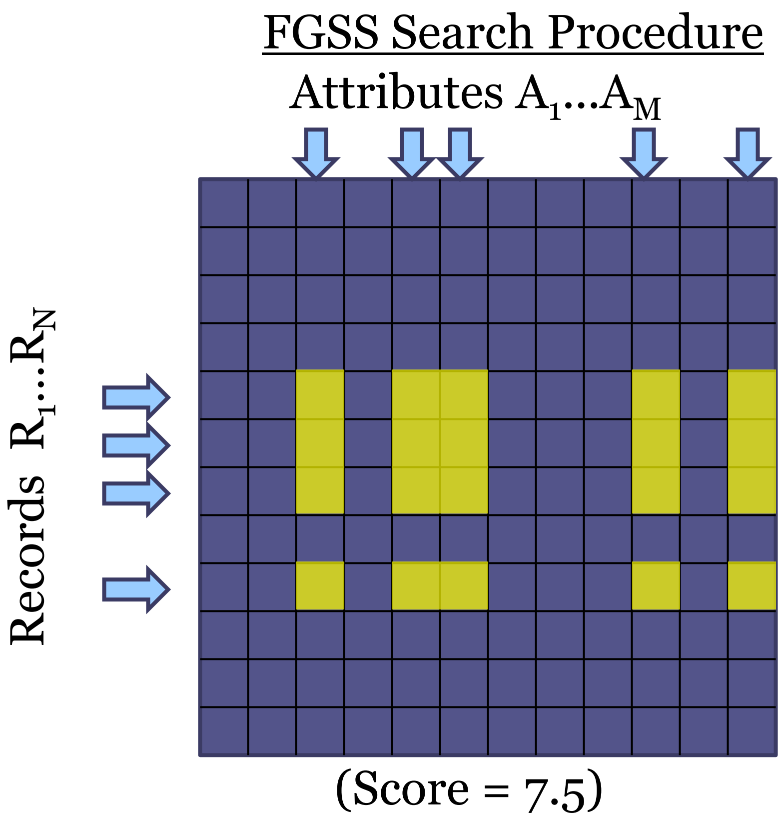
I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - LTSS over records O(N log N)
   - LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)



1. Start with a randomly chosen subset of attributes
2. Use LTSS to find the highest-scoring subset of recs for the given atts

I. Compute the anomalousness of each attribute (for each record)

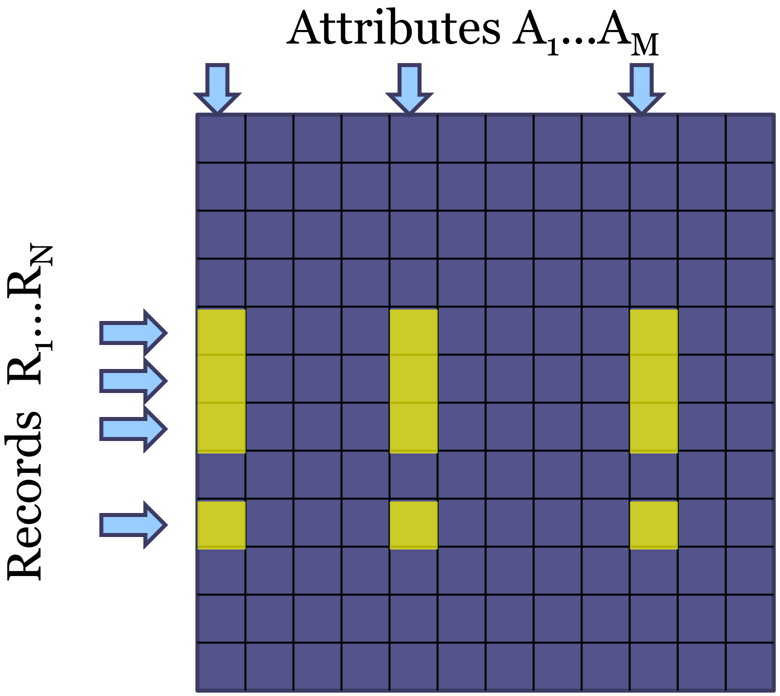
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - LTSS over records O(N log N)
   - LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure



(Score = 8.1)

2. Use LTSS to find the highest-scoring subset of recs for the given atts
3. Use LTSS to find the highest-scoring subset of atts for the given recs

I. Compute the anomalousness of each attribute (for each record)

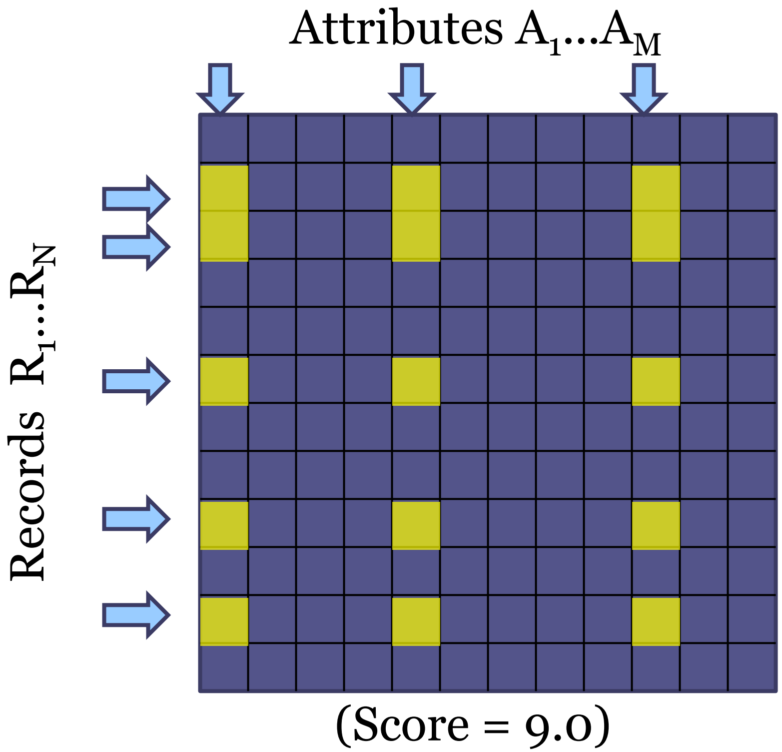
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - LTSS over records O(N log N)
   - LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure



3. Use LTSS to find the highest-scoring subset of atts for the given recs
4. Iterate steps 2-3 until convergence

I. Compute the anomalousness of each attribute (for each record)

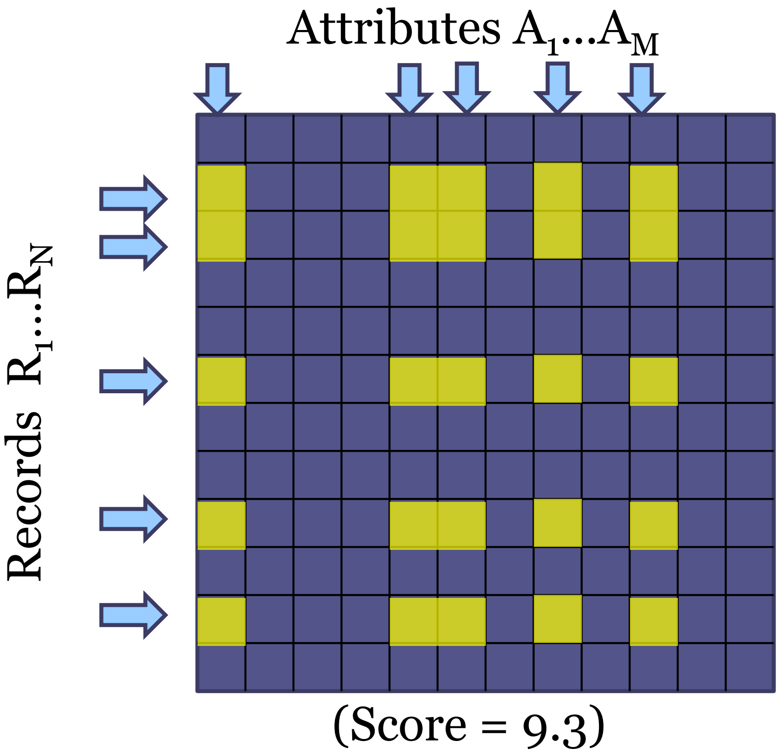
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
    - Iterate between following steps
        i. LTSS over records O(N log N)
        ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure



3. Use LTSS to find the highest-scoring subset of atts for the given recs
4. Iterate steps 2-3 until convergence

I. Compute the anomalousness of each attribute (for each record)

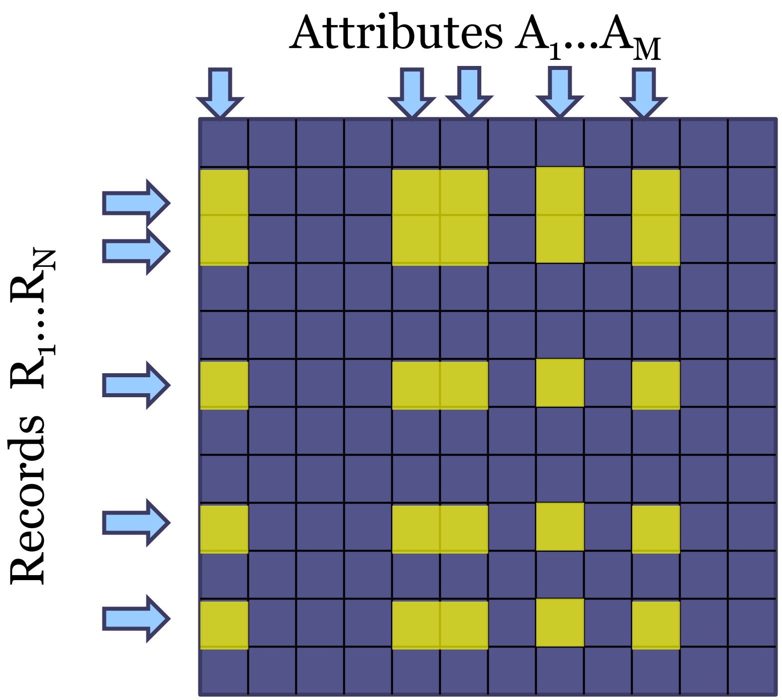
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

   1. Maximize F(S) over all subsets of S
      - Iterate between following steps
        - i. LTSS over records O(N log N)
        - ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure



(Score = 9.3)

Good News: Run time is (near) linear in number of recs & number of atts.

Bad News: Not guaranteed to find global maximum of the score function.

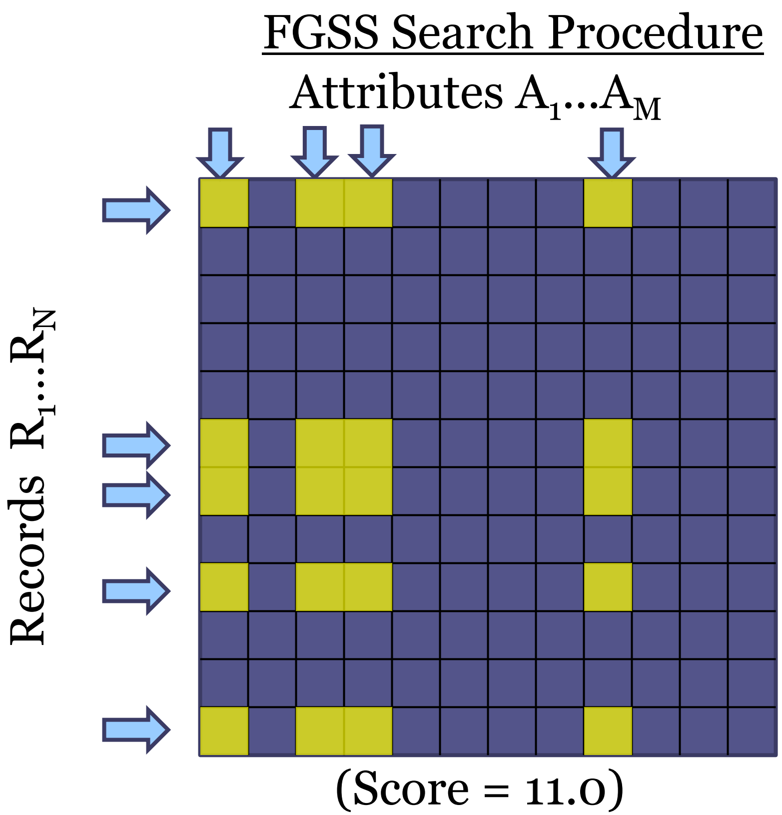
I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps
     - i. LTSS over records O(N log N)
     - ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)



5. Repeat steps 1-4 for 100 random restarts

I. Compute the anomalousness of each attribute (for each record)

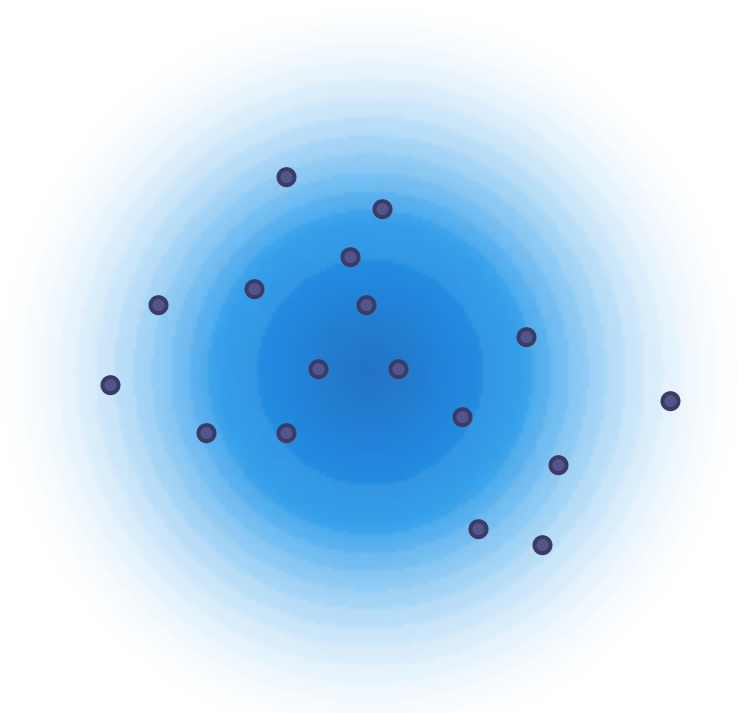
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps
     - i. LTSS over records O(N log N)
     - ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure



We want to enforce self-similarity, thus we create local neighborhoods.

I. Compute the anomalousness of each attribute (for each record)

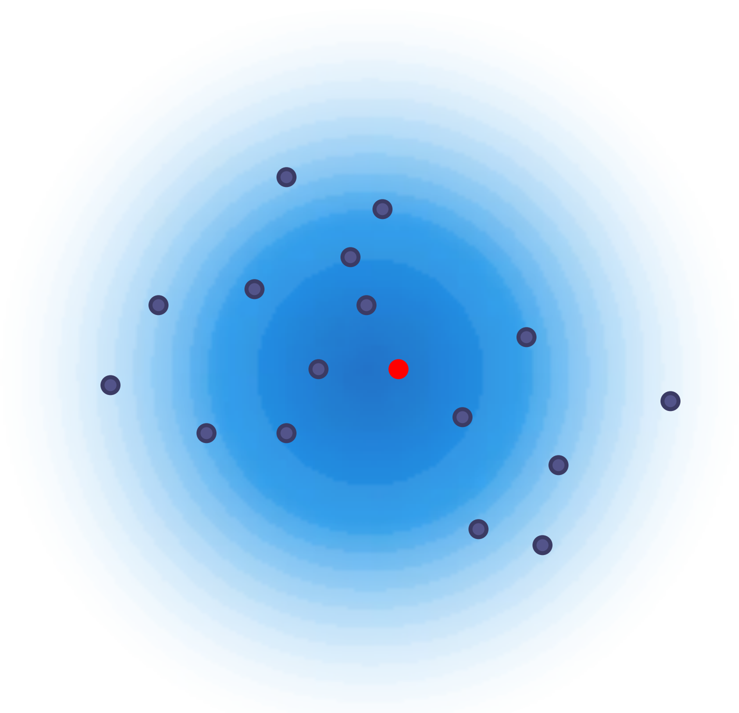
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps
     - i. LTSS over records O(N log N)
     - ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

<u>FGSS Constrained Search Procedure</u>



We want to enforce self-similarity, thus we create local neighborhoods defined by a center record

I. Compute the anomalousness of each attribute (for each record)

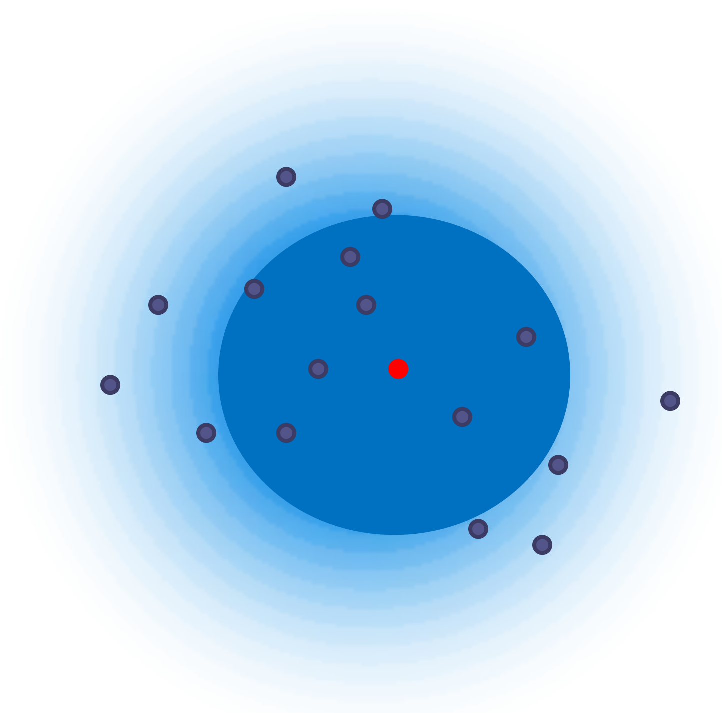
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps
     i. LTSS over records O(N log N)
     ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

<u>FGSS Constrained Search Procedure</u>



We want to enforce self-similarity, thus we create local neighborhoods defined by a center record and all other records within a max dissimilarity

I. Compute the anomalousness of each attribute (for each record)

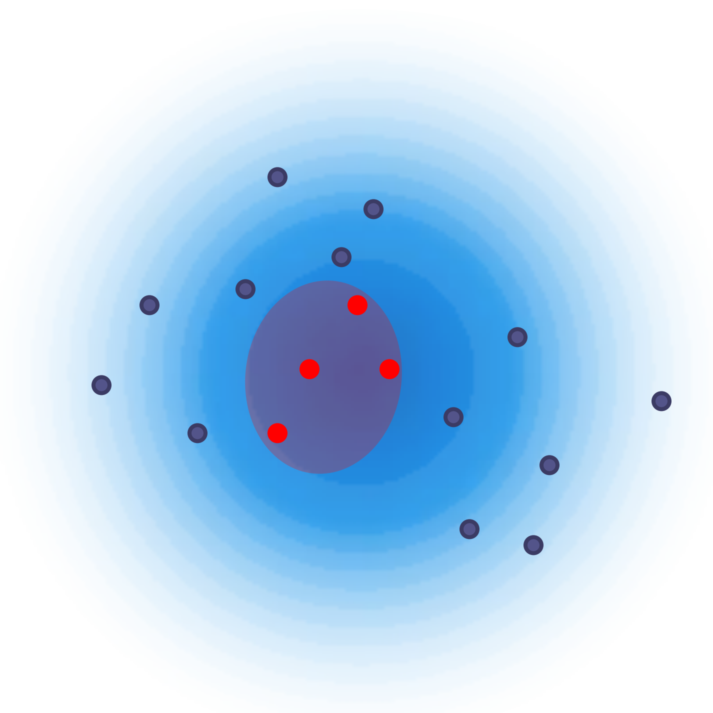
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
    - Iterate between following steps
        i. LTSS over records O(N log N)
        ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

<u>FGSS Constrained Search Procedure</u>



We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood

I. Compute the anomalousness of each attribute (for each record)

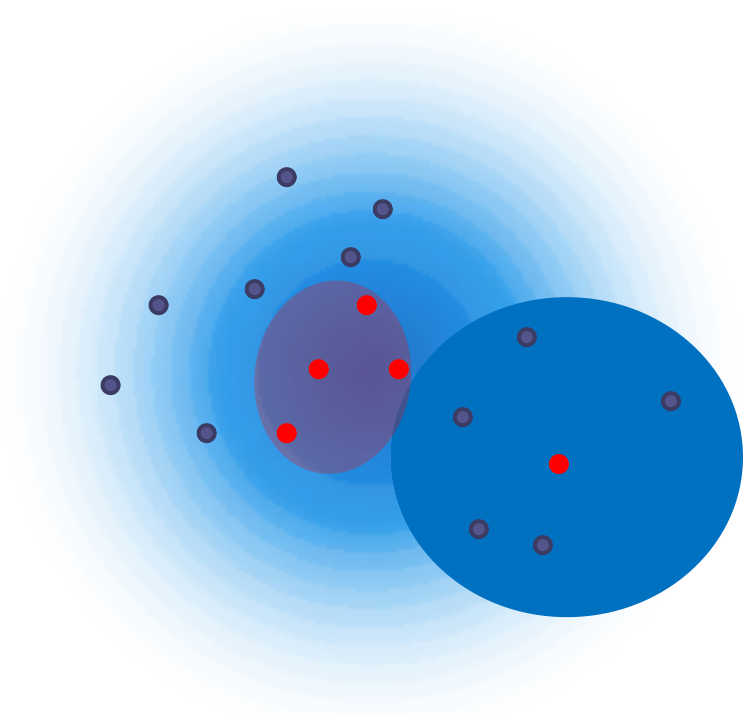
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps
     - i. LTSS over records O(N log N)
     - ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure



We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood

I. Compute the anomalousness of each attribute (for each record)

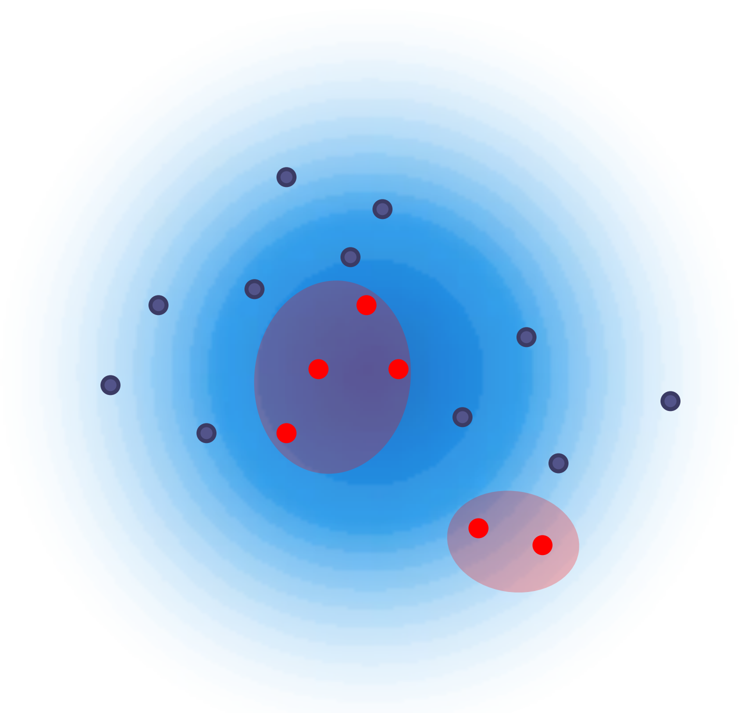
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps
     i. LTSS over records O(N log N)
     ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

<u>FGSS Constrained Search Procedure</u>



We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood, and maximize F(S) over all local neighborhoods

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - Iterate between following steps

     i. LTSS over records O(N log N)

     ii. LTSS over attributes O(M log M)

Emergency Department Dataset

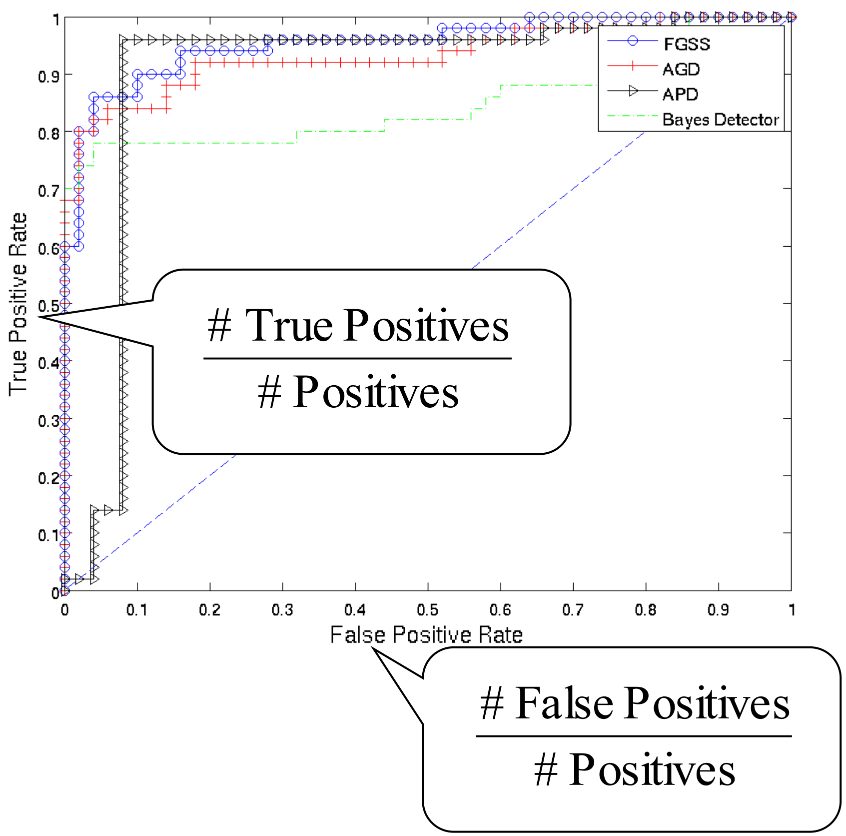
- Visits to ED in Allegheny County during 2004
  - Hopsital Id
  - Prodrome
  - Age Decile
  - Patient Home Zip-code
  - Chief Complaint

- Bayesian Aerosol Release Detector (BARD)
  - Injects simulated respiratory cases resembling an anthrax outbreak
  - Test data: First two days of the attack
  - Training data: Previous 90 days

- We compare FGGSS to other recently proposed methods
  - Bayes Anomaly Detector
  - Anomaly Pattern Detection (APD) (Das et al. 2008)
  - Anomalous Group Detection (AGD) (Das et al. 2009)

(BARD) Simulated Anthrax ED Dataset

Receiver Operator Characteristic

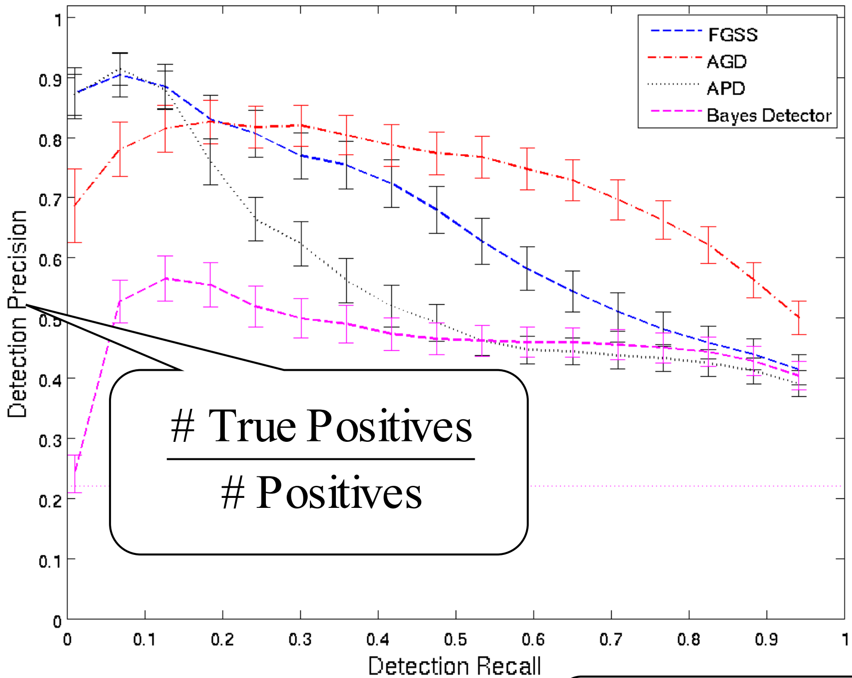


Evaluation Purpose

- Measures how well each methods can distinguish between datasets with anomalous patterns present

(BARD) Simulated Anthrax ED Dataset

Precision vs. Recall



$$\frac{\#\text{ True Positives}}{\#\text{ Positives}}$$

The proportion of true anomalies detected.

Evaluation Purpose

- Given a dataset affected by an anomalous process, measures how well each methods can identify the affected subsets

(BARD) Simulated Anthrax ED Dataset

Area Under the Curve (AUC)

Methods	ROC	Precision vs. Recall
FGSS	95.4±1.7	63.8±2.5
AGD	93.2±2.5	74.3±2.4
APD	90.0±2.0	52.0±2.0
Bayes Dectector	84.8±4.2	47.6±2.0

Conclusions & Future Work

- FGSS run significantly faster than methods with comparable detection power
- FGSS out performs other methods when patterns are:
  - a small portion of the data
  - subtle (not extremely individually anomalous)
- FGSS can characterize anomalous patterns
- What's Next?
  - Extend methods to handle mixed-value datasets
  - Extend methods to handle multiple models
  - Active Learning

Thank You…Questions/Comments?