Efficient Methods for Anomalous Pattern Detection in General Datasets

Event & Pattern Detection Lab
Carnegie Mellon University
Edward McFowland III (mcfowland@cmu.edu)
Daniel B. Neill (neill@cs.cmu.edu)

This work was partially supported by:
AT&T Research Labs Fellowship
NSF Graduate Research Fellowship
NSF grants IIS-0916345, IIS-0911032, and IIS-0953330

Anomalous Pattern Detection

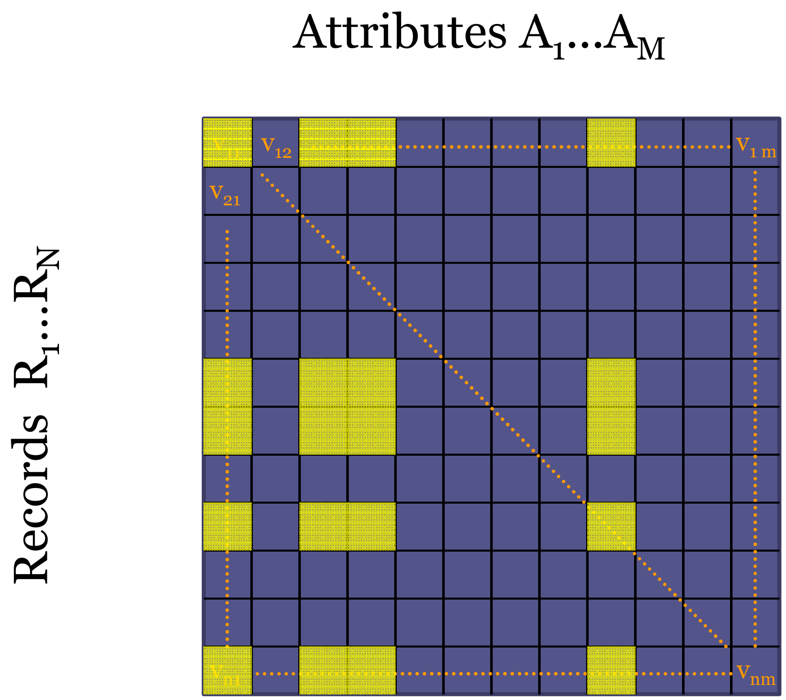
- Two set of processes generating data
  - Typical system behavior
  - Anomalous system behavior

- Discover and characterize the anomalous processes
  - Evaluating records in isolation may be insufficient
  - Find the subsets of data that correspond to anomalous system behavior
  - An anomalous subsets is self-similar and as a group different from rest of the data

Is it Useful?

- Fraud Detection
- Network Intrusion Detection
- Anomalous Patterns of Smuggling
- Disease Surveillance
- …And many more ways to make the world a better place

The Goal!



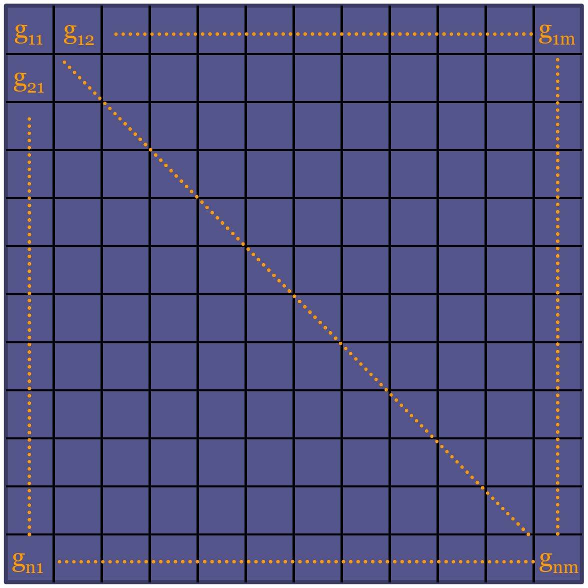
I. Compute the anomalousness of each attribute value (for each record)

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

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

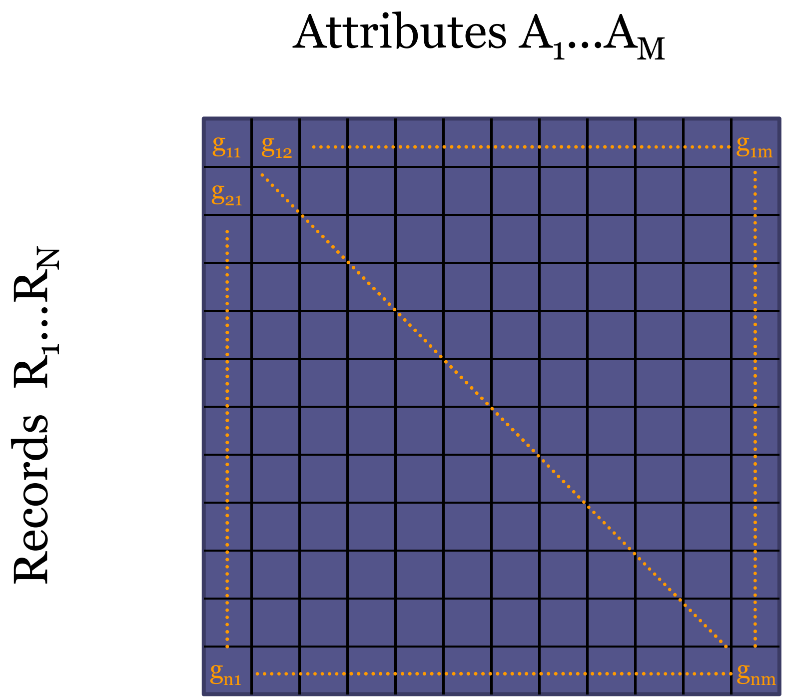
Records $R_1 \ldots R_N$



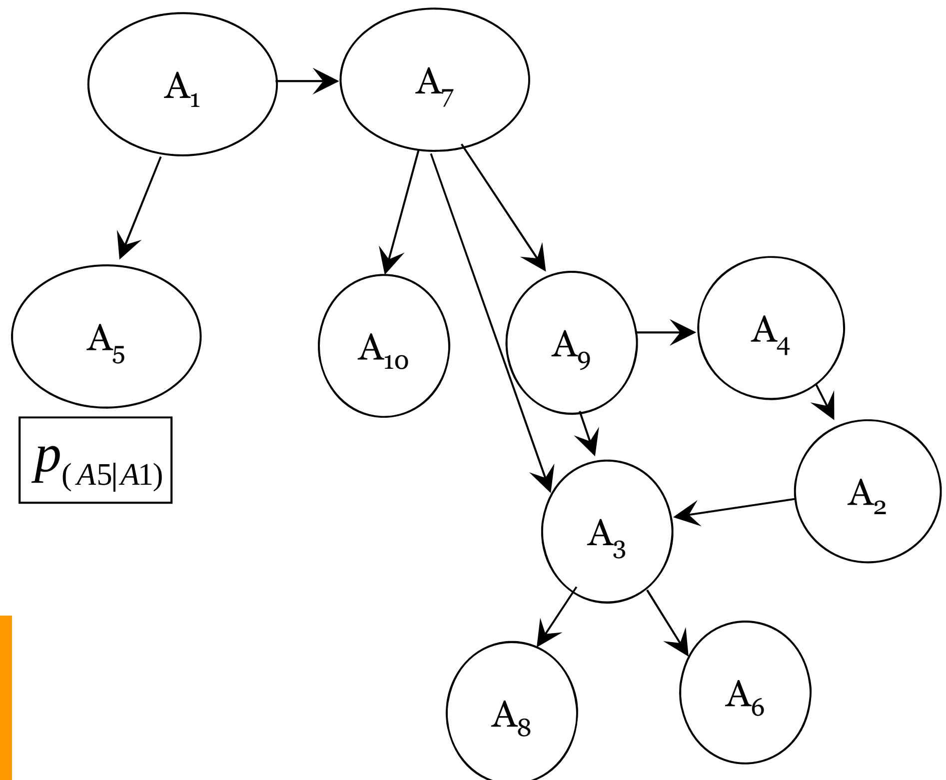
I. Compute the anomalousness of each attribute (for each record)

In order to compute the anomalousness of the data, FGSS models the data distribution under expected system behavior

Fast Generalized Subset Scan (FGSS)

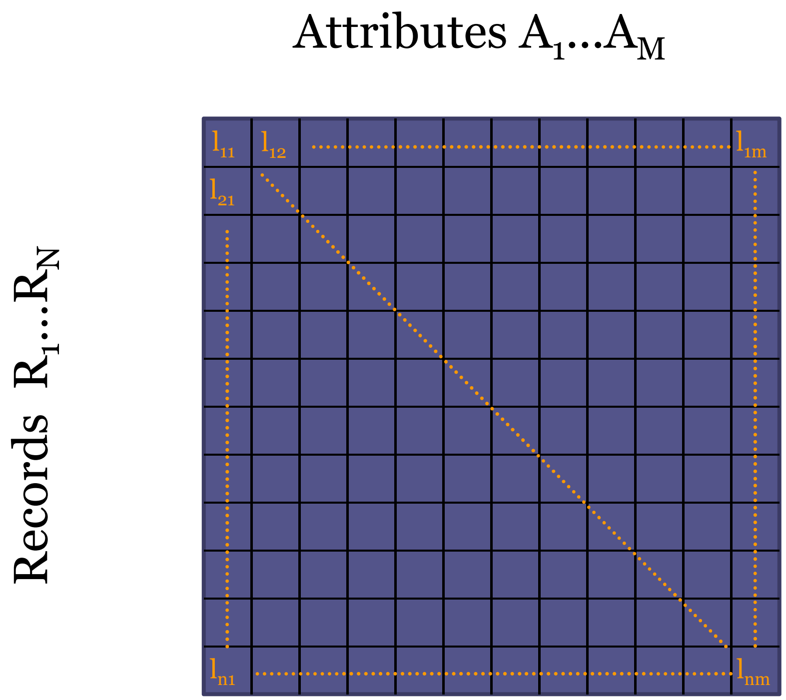


I. Compute the anomalousness of each attribute (for each record)



In order to compute the anomalousness of the data, FGSS models the data distribution under expected system behavior

Fast Generalized Subset Scan (FGSS)

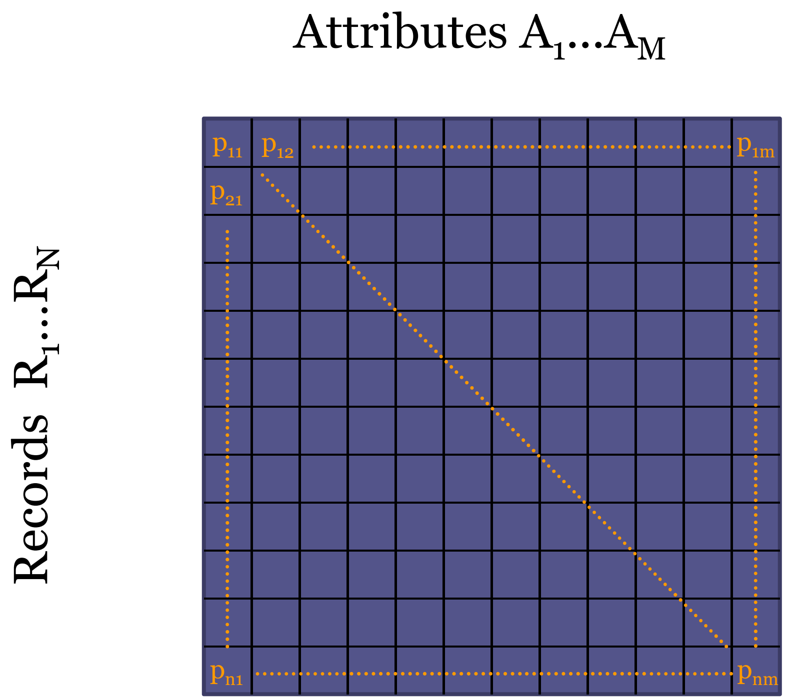


By performing inference on the Bayesian Network, for each record we can determine the likelihood of each of its attribute values

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

1. Learn Bayesian Network
2. Compute attribute value likelihoods

Fast Generalized Subset Scan (FGSS)

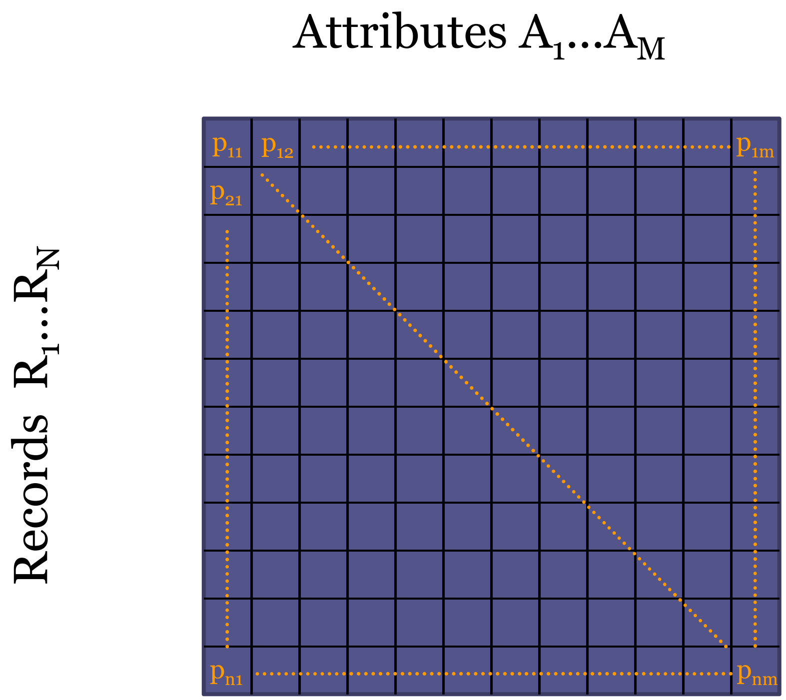


I. Compute the anomalousness of each attribute (for each record)

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



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

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

Fast Generalized Subset Scan (FGSS)

Nonparametric Scan Statistic (NPSS)

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

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

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

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

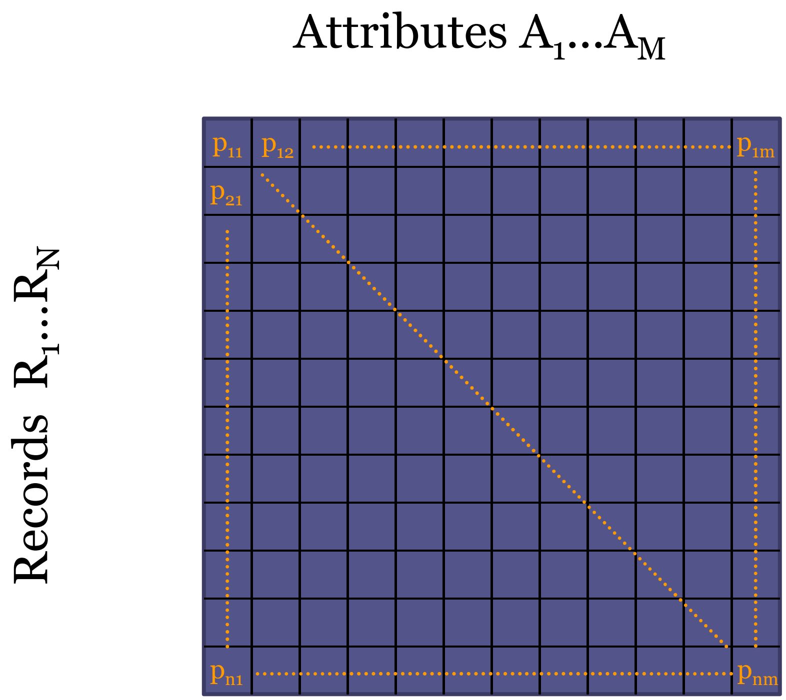
I. Compute the anomalousness of each attribute (for each record)

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

Fast Generalized Subset Scan (FGSS)



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

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

Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff :

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

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)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff :

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

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)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff :

$$\max_{S \subseteq D} F(S) = \max_{j=1...M} F\left(\left\{A_{(1)} \dots A_{(j)}\right\}\right)$$

We only need to consider:

$\{A_{(1)}\}$
$\{A_{(1)}, A_{(2)}\}$
$\{A_{(1)}, A_{(2)}, A_{(3)}\}$
$\vdots$
$\{A_{(1)}, \dots\dots\dots\dots, A_{(m)}\}$

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 :

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

We only need to consider:

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

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

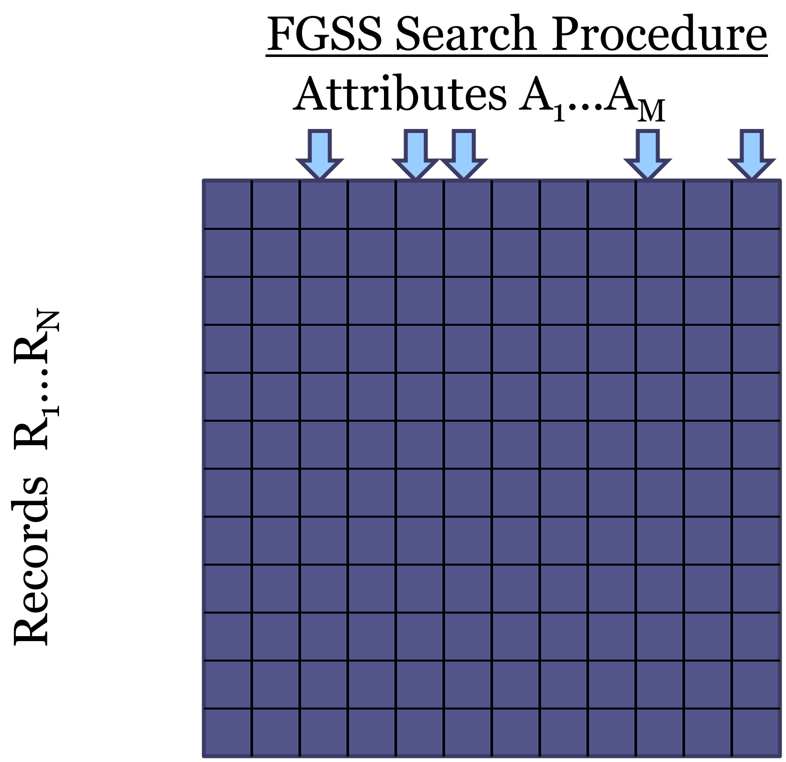
I. Compute the anomalousness of each attribute (for each record)

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

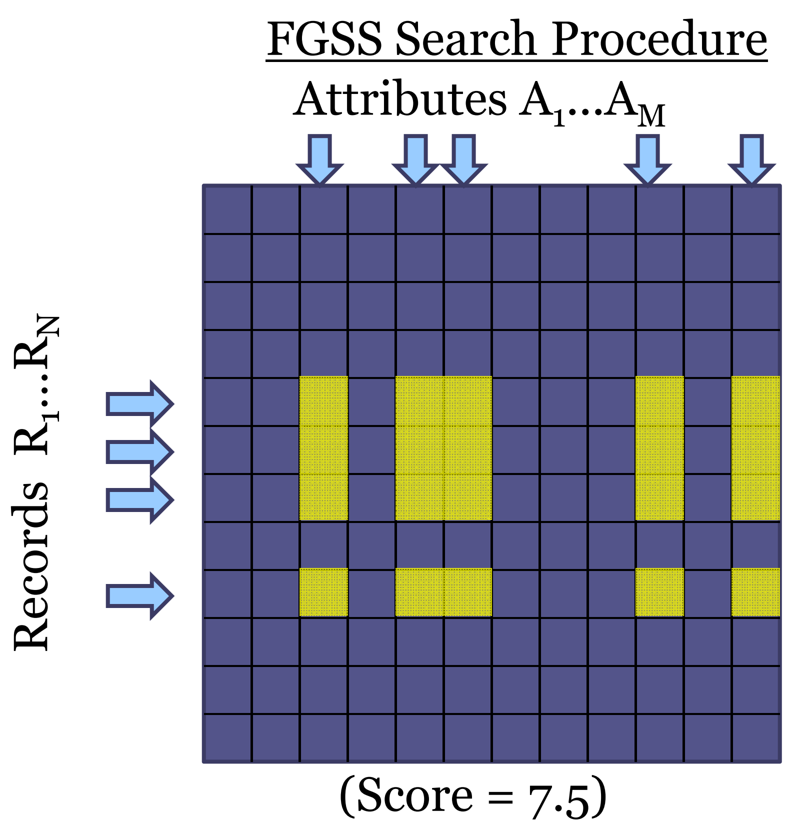
I. Compute the anomalousness of each attribute (for each record)

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

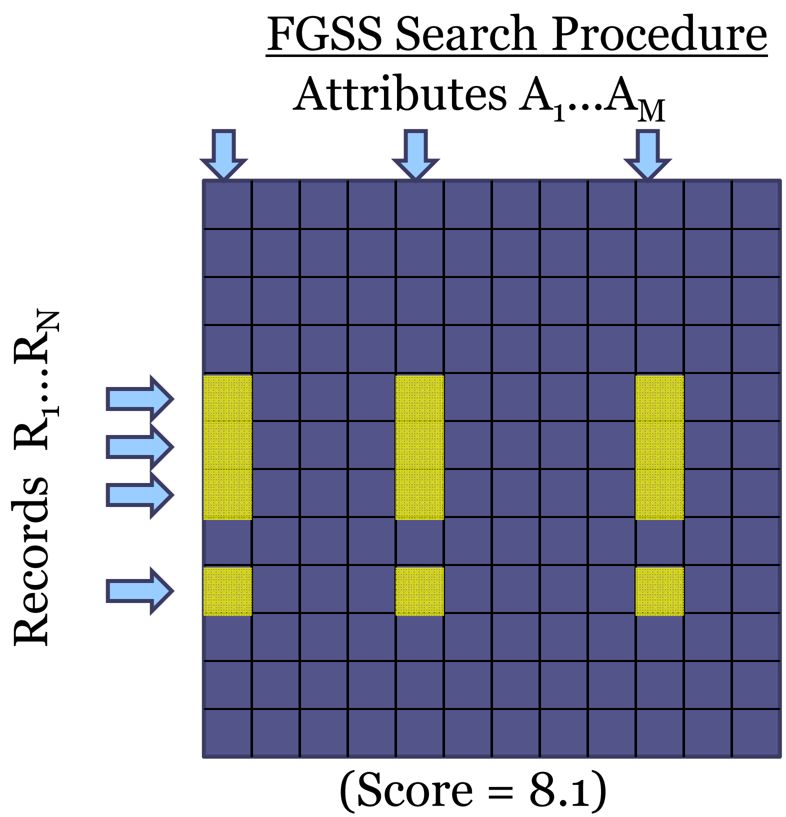
I. Compute the anomalousness of each attribute (for each record)

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)



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

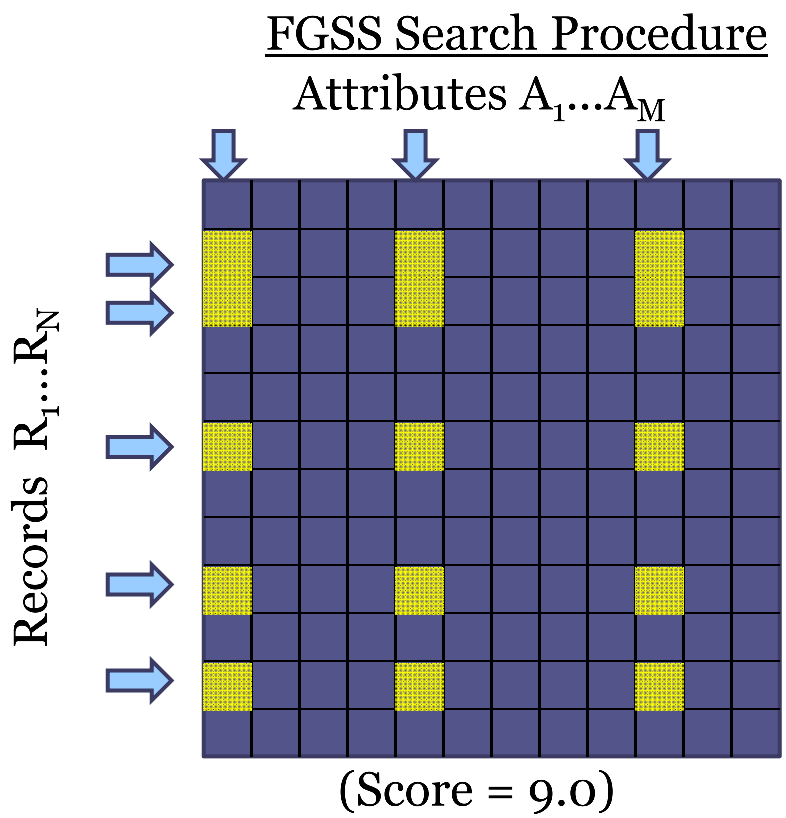
I. Compute the anomalousness of each attribute (for each record)

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)



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

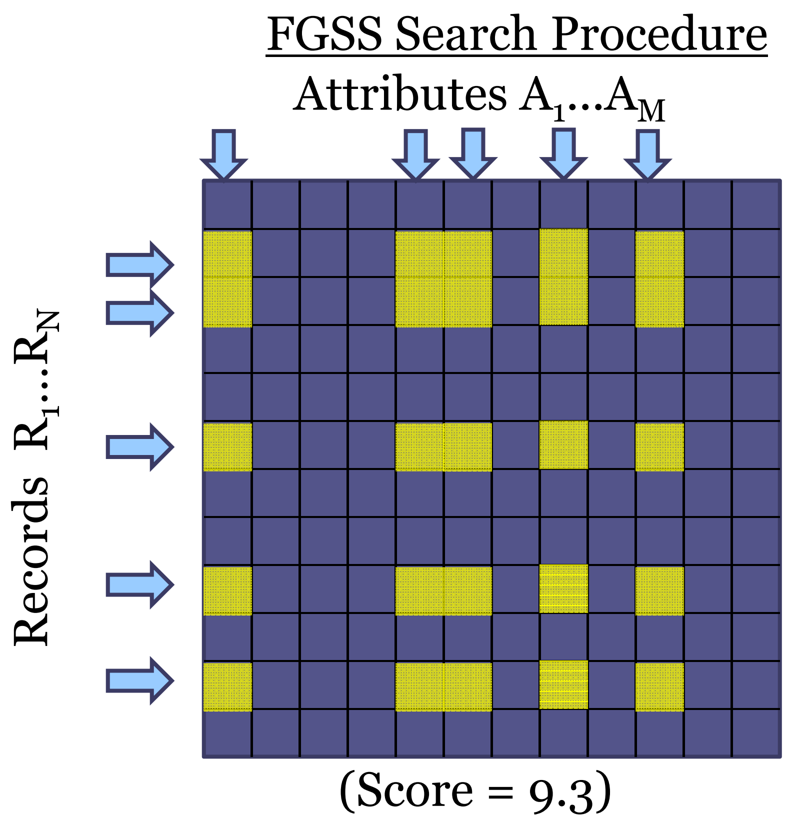
I. Compute the anomalousness of each attribute (for each record)

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)



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

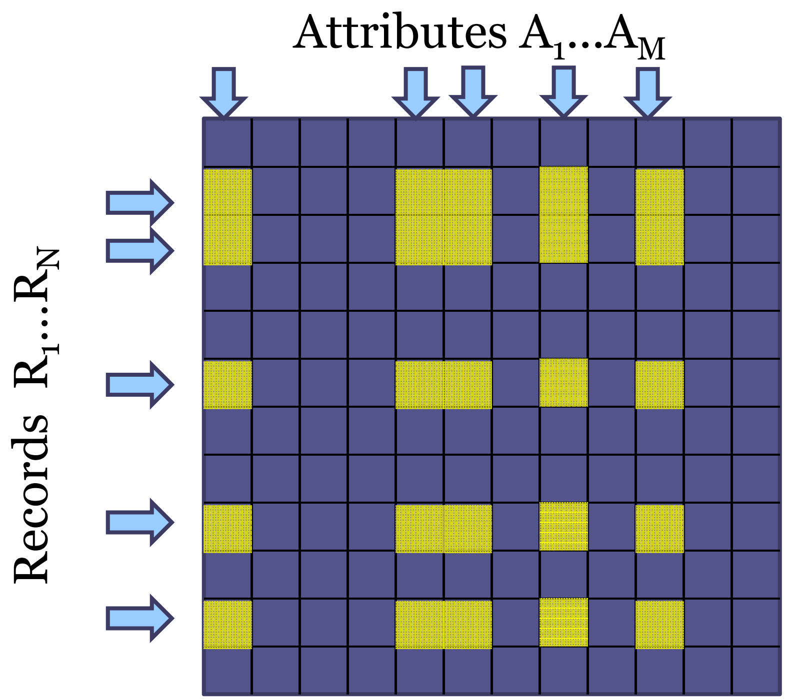
•Iterate between following steps

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)

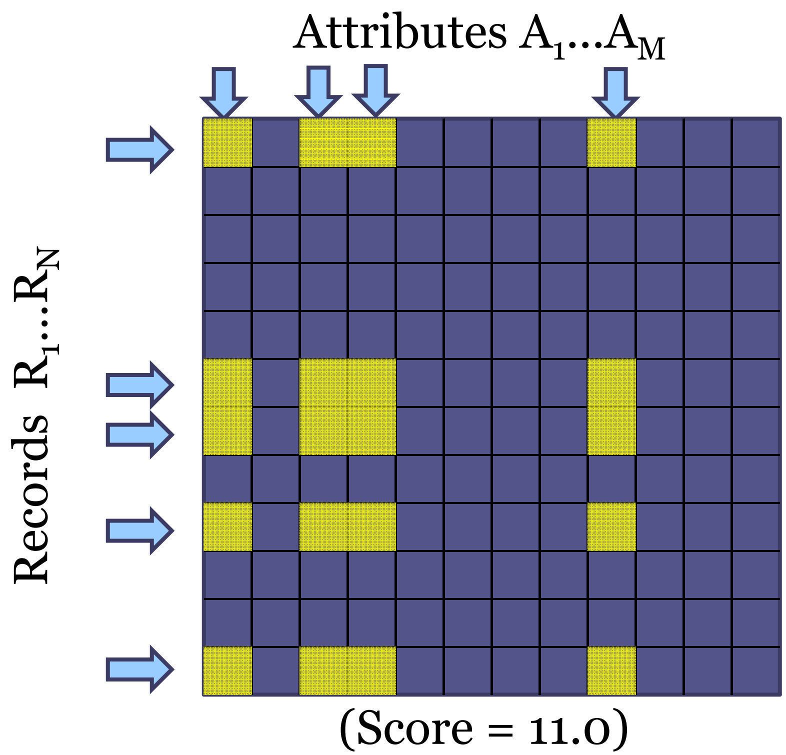
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

FGSS Search Procedure



5. Repeat steps 1-4 for 50 random restarts

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

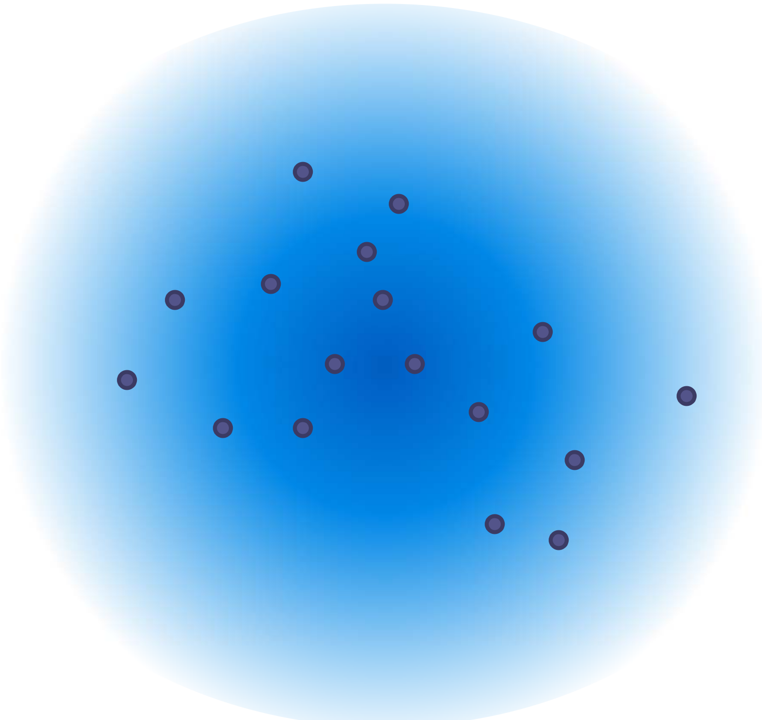
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

<u>FGSS Constrained Search Procedure</u>



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

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

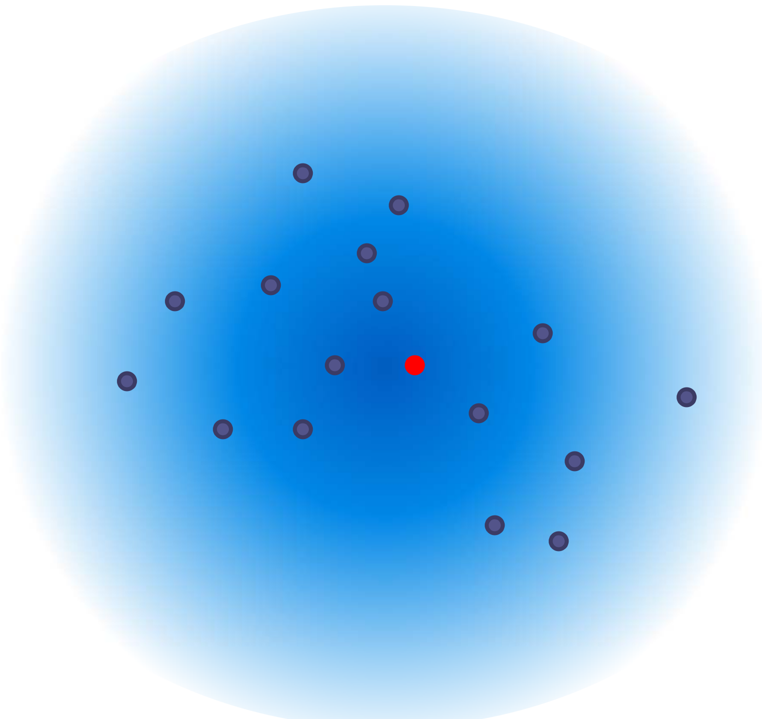
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

FGSS Constrained Search Procedure



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)

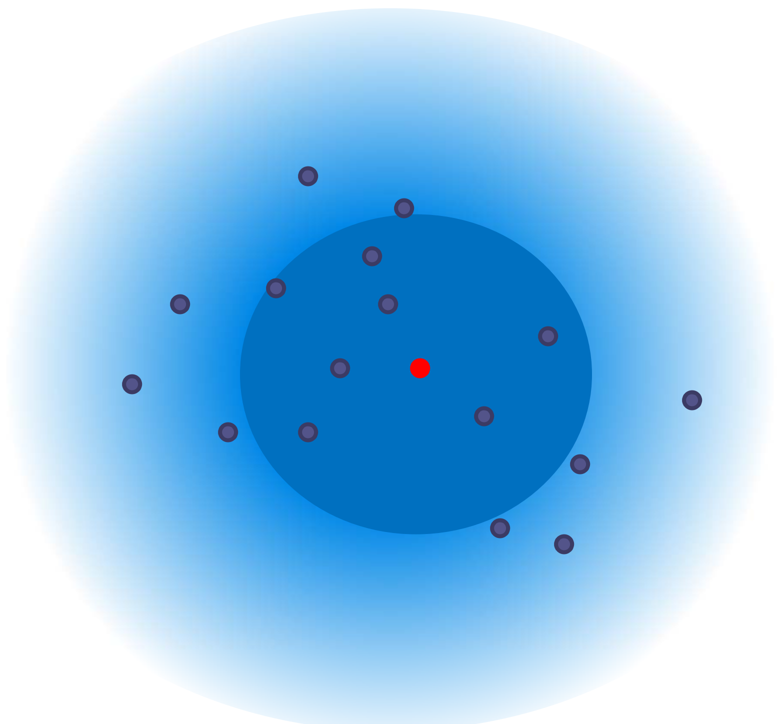
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

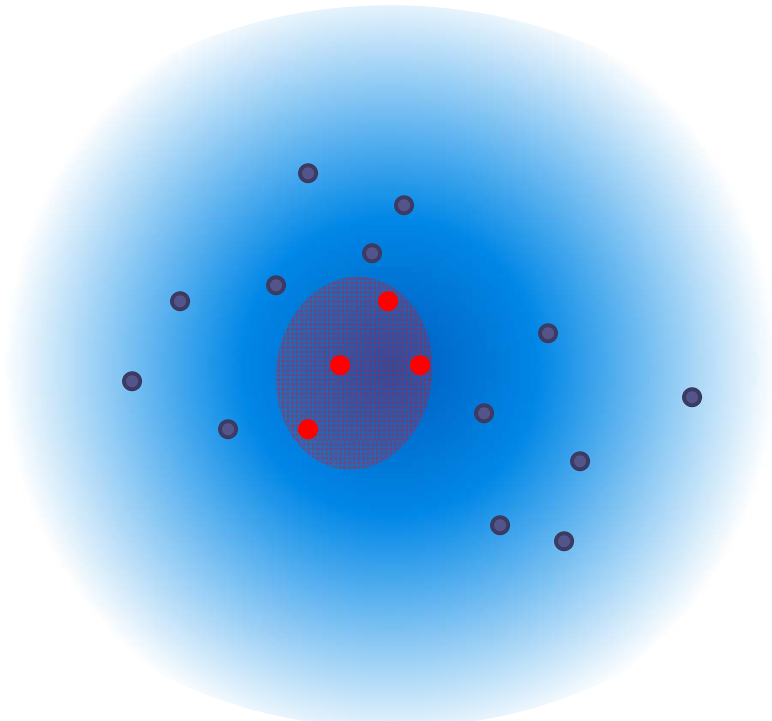
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

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)

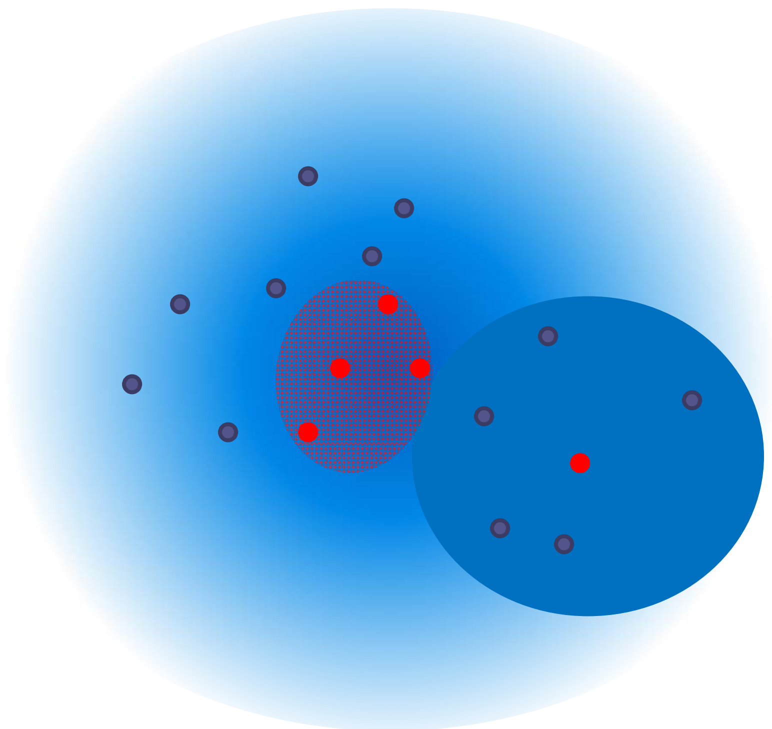
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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

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

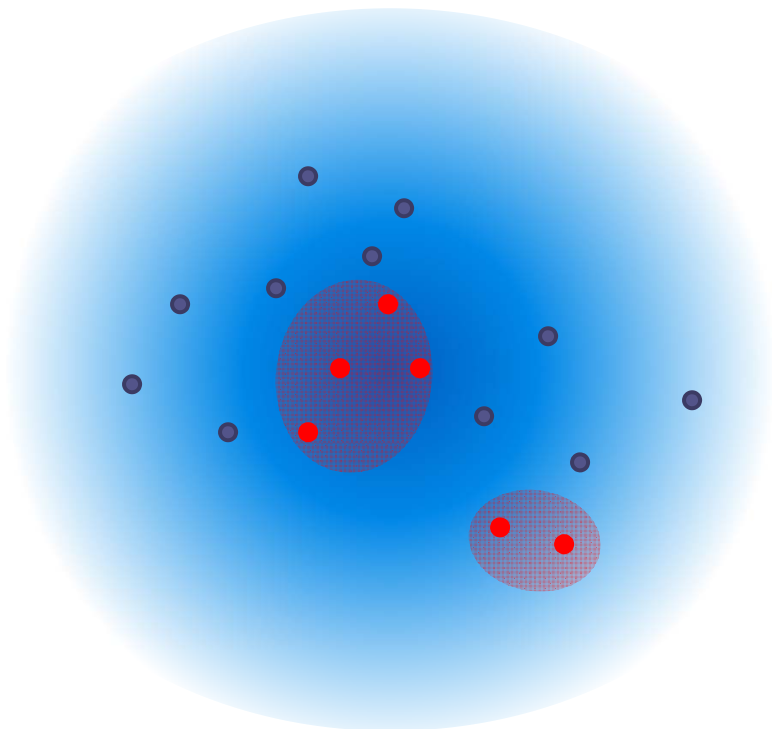
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

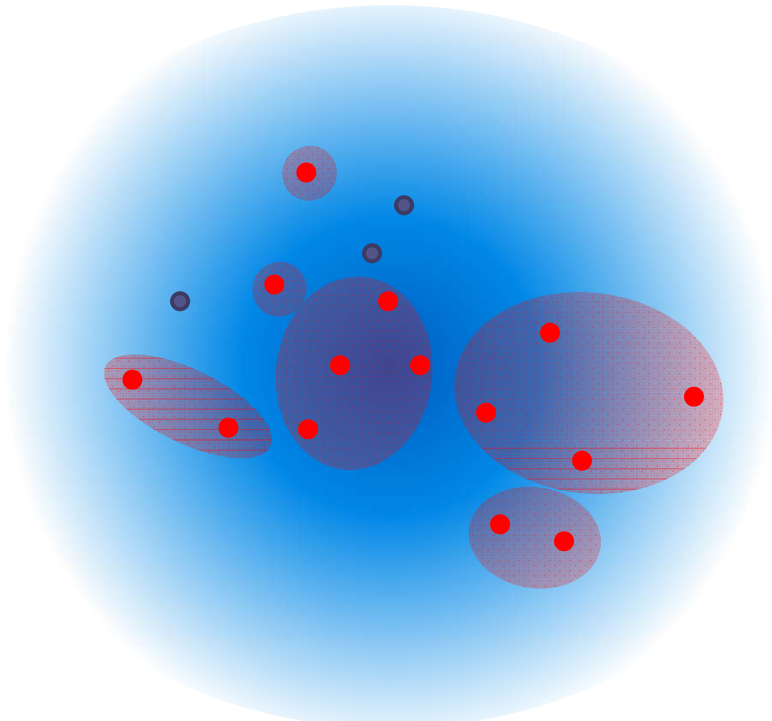
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

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)

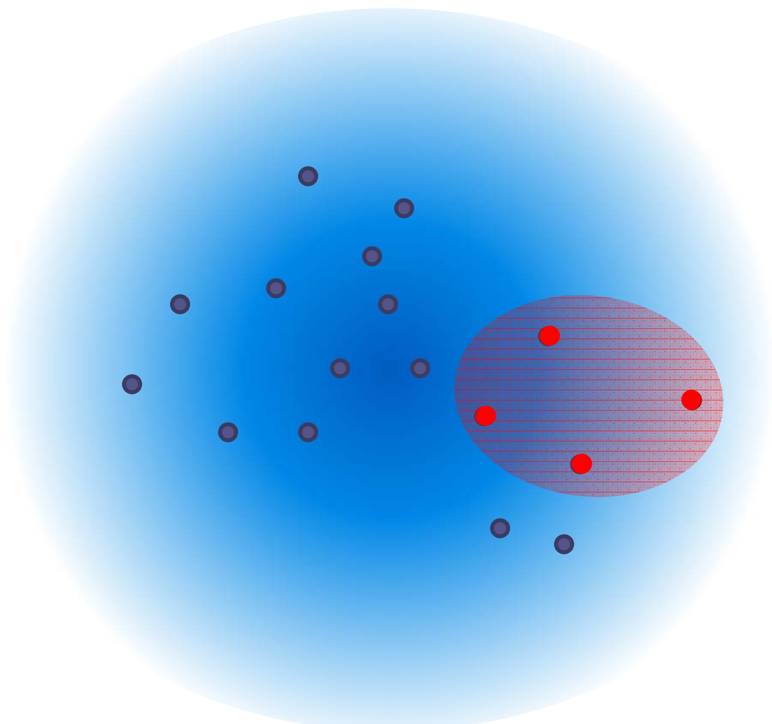
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

FGSS Constrained Search Procedure



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)

Experiments

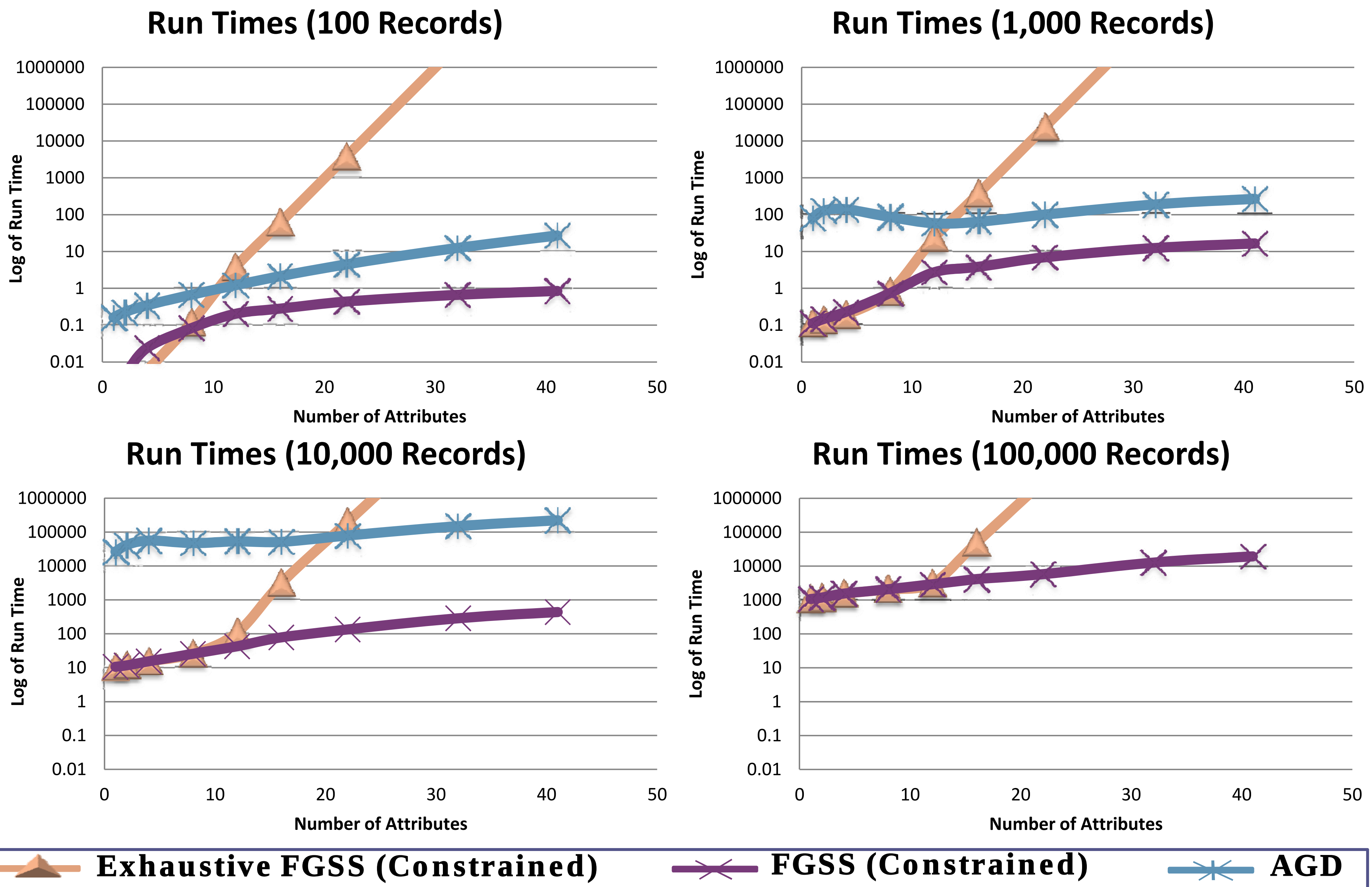
- Network Activity and Intrusion Data (KDDCUP '99)
  - 41 attributes representing extracted information from the raw data of the network connection

- BARD Simulated Anthrax Outbreak in ED visits
  - Hopsital Id
  - Prodrome
  - Age Decile
  - Patient Home Zip-code
  - Chief Complaint

- U.S. Customs and Boarder Patrol Data
  - Country of origin
  - Departing & Arriving ports, Shipping line
  - Shipper's & Vessel's name
  - Commodity being shipped

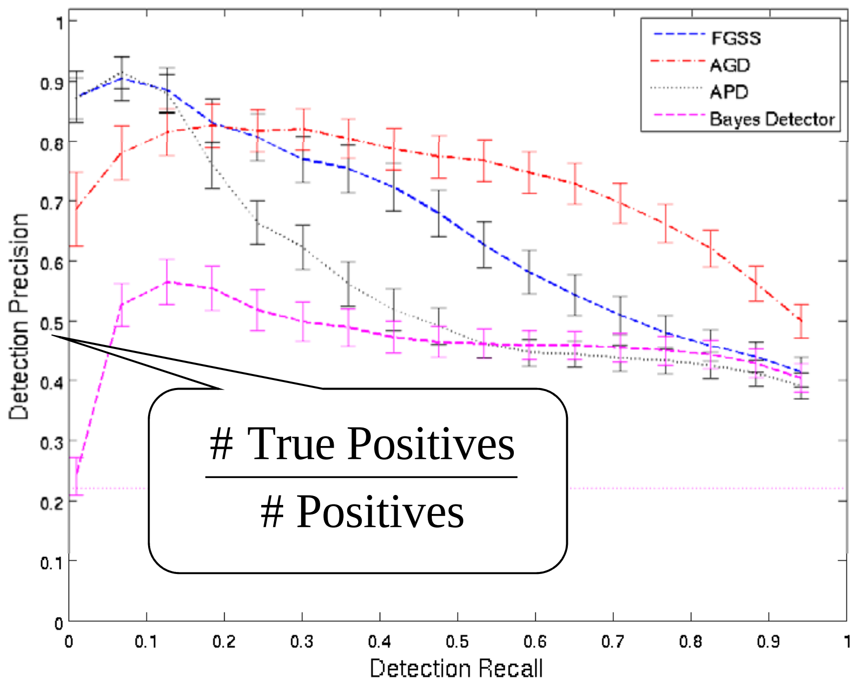
- We compare FGSS to other recently proposed methods
  - Bayesian Network Anomaly Detector
  - Anomaly Pattern Detection (APD) (Das et al. 2008)
  - Anomalous Group Detection (AGD) (Das et al. 2009)

Results


Run Times (100 Records)
Run Times (1,000 Records)
Run Times (10,000 Records)
Run Times (100,000 Records)
Log of Run Time
Number of Attributes
1000000
100000
10000
1000
100
10
1
0.1
0.01
0
10
20
30
40
50
Exhaustive FGSS (Constrained)
FGSS (Constrained)
AGD

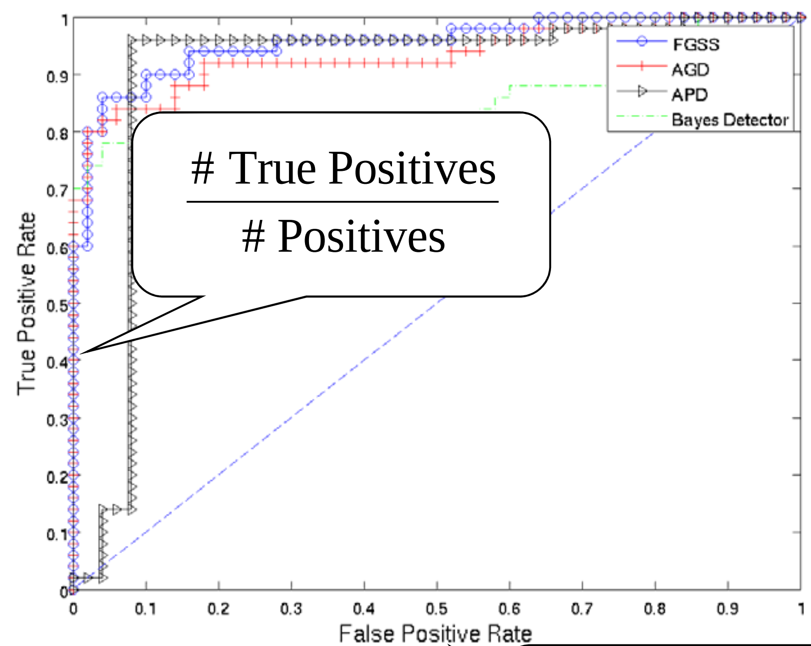
(BARD) Simulated Anthrax ED Dataset

Precision vs. Recall



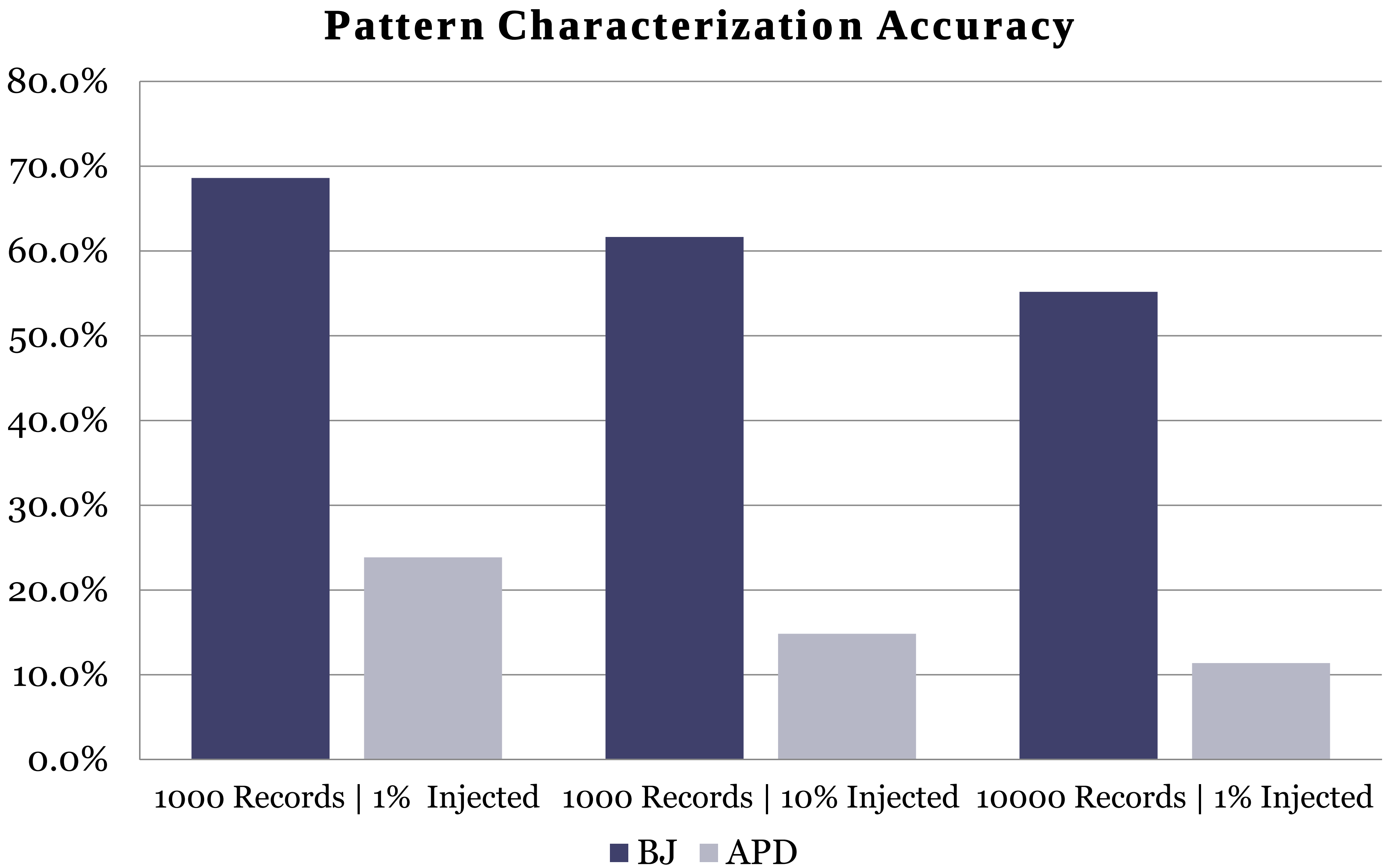
The proportion of true anomalies detected.

Receiver Operator Characteristic



False Positives
Positives

Results



Pattern Characterization Accuracy
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1000 Records | 1% Injected
1000 Records | 10% Injected
10000 Records | 1% Injected
BJ
APD

Conclusions

- FGSS run significantly faster than methods with comparable detection power
- FGSS out performs other methods when patters are:
  - a small portion of the data
  - subtle (not extremely individually anomalous)
- FGSS can characterize anomalous patterns
- Extensions
  - Extend method to handle multiple anomaly detectors
  - Extend method to handle multiple models
  - Active Learning

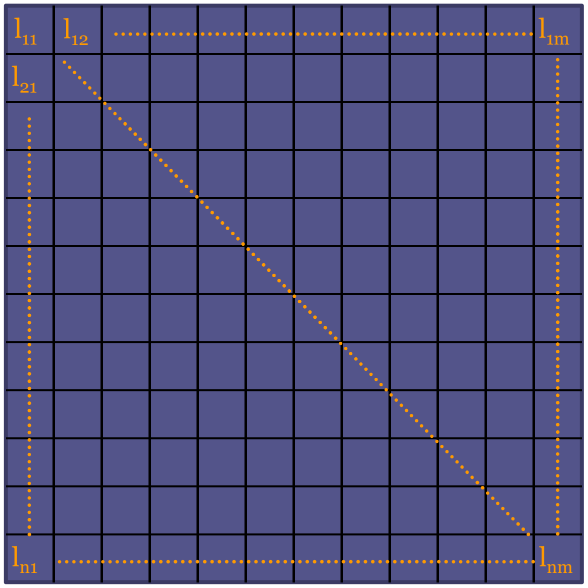
Extensions

(Preliminary)

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$

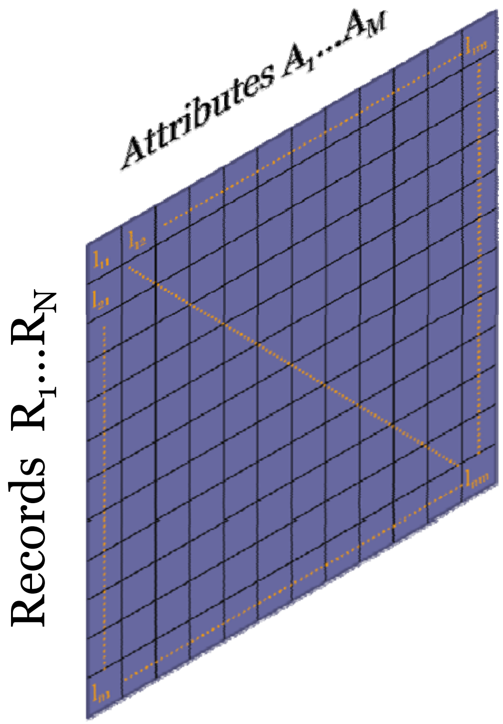


By performing inference on the Bayesian Network, for each record we can determine the likelihood of each of its attribute values

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

1. Learn Bayesian Network
2. Compute attribute value likelihood

Multiple Anomaly Detectors

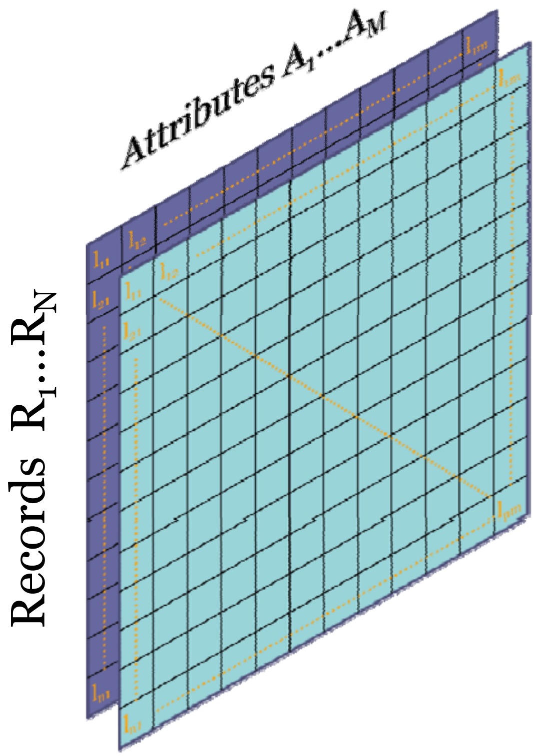


I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihood

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

Multiple Anomaly Detectors

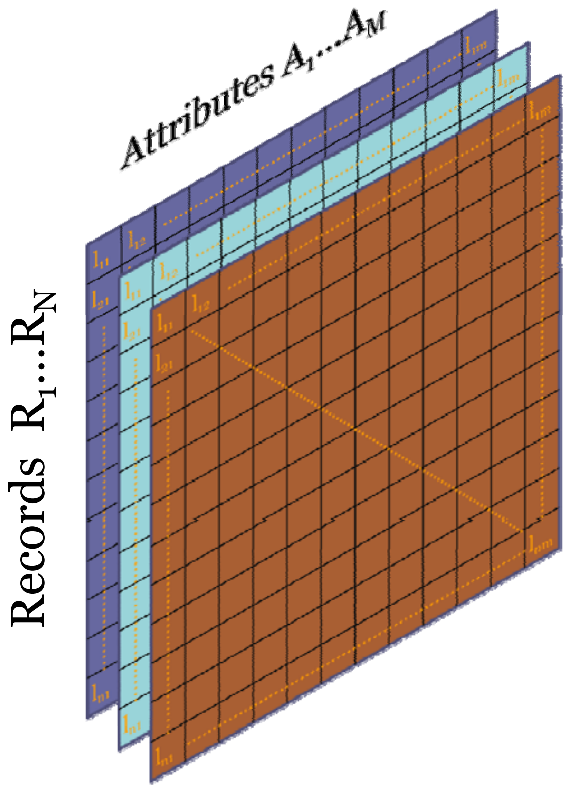


I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute outlier scores

Is the value sufficiently far away from the mean (outlier)?

Multiple Anomaly Detectors

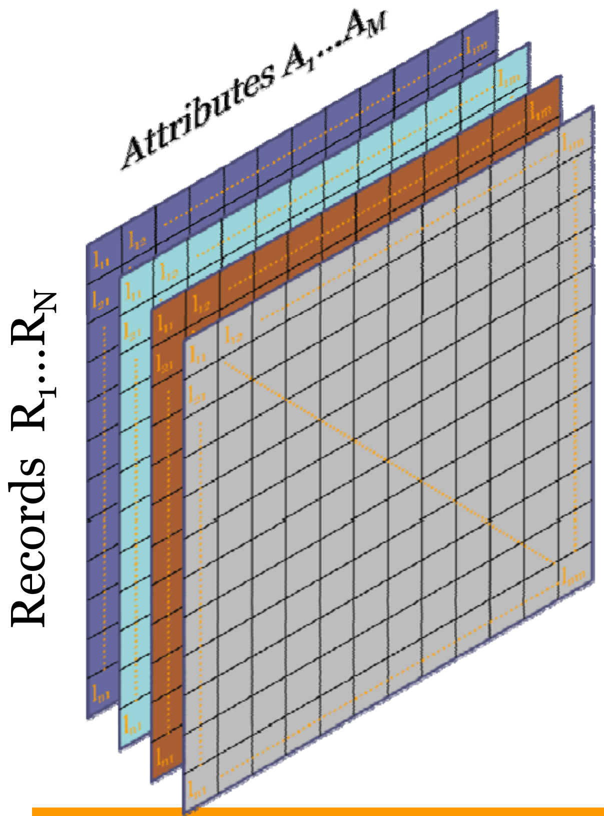


Is the value a duplicate?

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

1. Learn Bayesian Network
2. Compute duplicate scores

Multiple Anomaly Detectors

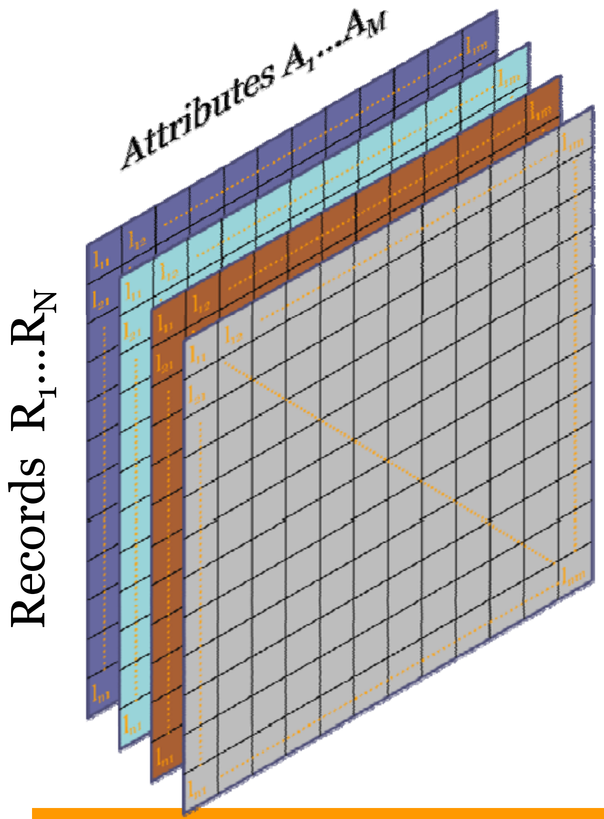


Is the value missing?

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

1. Learn Bayesian Network
2. Compute missing scores

Multiple Anomaly Detectors

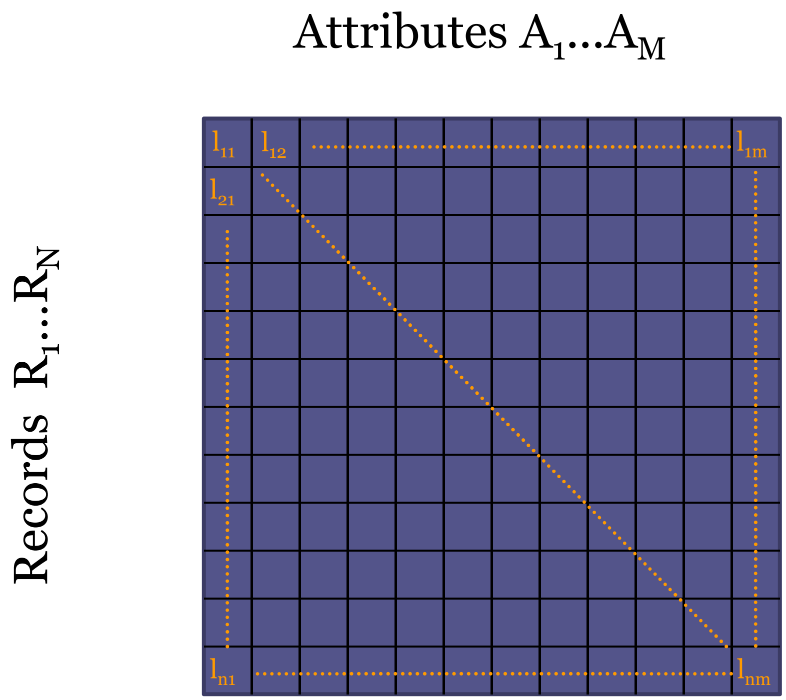


$$l_{ij} = \frac{\sum_K \alpha_k * I(isAnom(l_{ijk}))}{K}$$

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

1. Learn Bayesian Network
2. Compute anomaly scores

Multiple Anomaly Detectors



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

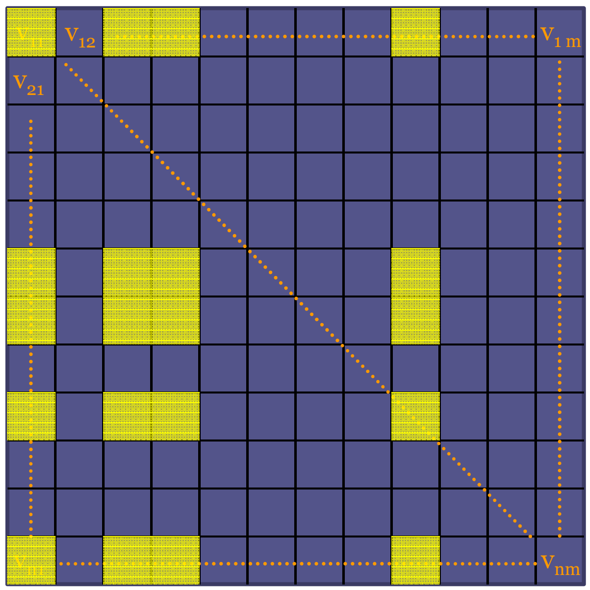
1. Learn Bayesian Network
2. Compute glitch scores

Now have new measure of the anomalous each record x attribute pair.

Multiple Anomaly Detectors

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



Search over all possible subsets of data and find the maximizing F(S)

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

1. Learn Bayesian Network
2. Compute glitch scores
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)

Thank You...Questions/Comments?