Fast Generalized Subset Scan for Anomalous Pattern Detection

Daniel B. Neill (neill@cs.cmu.edu)
Event & Pattern Detection Laboratory
Carnegie Mellon University
Joint work with Edward McFowland III and Skyler Speakman.

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

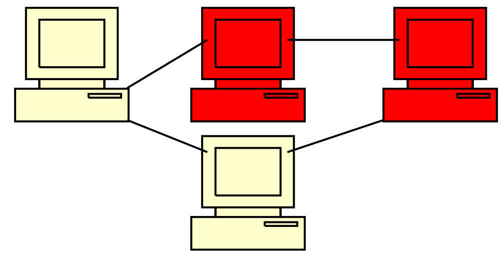
Anomalous Pattern Detection

- Two set of processes generating data
  - Typical system behavior
  - Anomalous system behavior
- **Discover** and **characterize** the anomalous processes (which records are anomalous, and how are they anomalous?)
  - Evaluating records in isolation may be insufficient to detect subtle patterns.
  - Find a **subset** of related data records that are anomalous when considered collectively.

Three Motivating Applications

1. Early detection of anthrax bio-attacks by monitoring Emergency Department visits

2. Intrusion detection in computer networks



3. Detecting patterns of illicit container shipments

FPORT	USPORT	COUNTRY	SLINE	VESSEL	SHIPPER NAME	F NAME	COMMODITY	SIZE	MTONS	VALUE
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	AMERICAN_TRI_NET_EXPRE	TRI_NET	EMPTY_RACK	0	5.6	27579
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRE	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRE	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	AMERICAN_TRI_NET_EXPRE	TRI_NET	CRUDE_IODINE_PURITY	1	17.68	251151
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRES	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	CHINA_OCEAN_SHPG	CHINA_OC	EMPTY_CONTAINERS	0	0	0
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	CHINA_OCEAN_SHPG	CHINA_OC	EMPTY_CONTAINERS	0	0	0

Three Motivating Applications

1. Early dete
monitorin

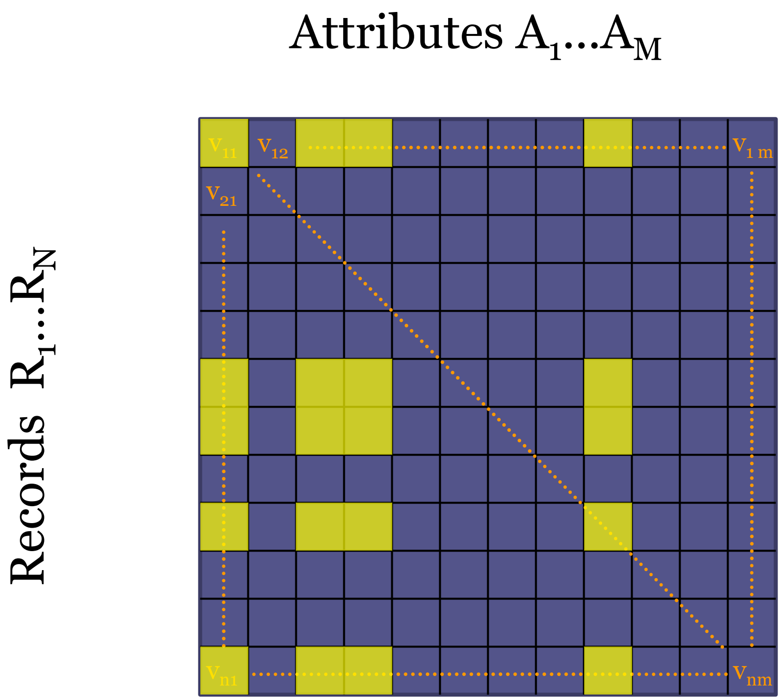
<u>Our solution to all three problems</u>: detect subsets of data records which are self-similar, and for which some subset of attributes are anomalous.

<u>Fundamental challenge</u>: N records and M attributes → $2^N \times 2^M$ subsets of records and attributes to consider!

3. Detecting patte

FPORT	USPORT	COUNTRY	SLINE	VESSEL	SHIPPER NAME	F NAME	COMMODITY	SIZE	MTONS	VALUE
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	AMERICAN_TRI_NET_EXPRE	TRI_NET	EMPTY_RACK	0	5.6	27579
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRE	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRE	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	AMERICAN_TRI_NET_EXPRE	TRI_NET	CRUDE_IODINE_PURITY	1	17.68	251151
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	NEW_WAVE_TRANSPORT	JIT	PANELS_F_MODEL_98	3	39.57	65169
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	ORDER	ORDER_C	USED_TIRES	2	13.43	9497
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	CHINA_OCEAN_SHPG	CHINA_OC	EMPTY_CONTAINERS	0	0	0
YOKOHAMA	SEATTLE	JAPAN	CSCO	LING_YUN_HE	CHINA_OCEAN_SHPG	CHINA_OC	EMPTY_CONTAINERS	0	0	0

Fast Generalized Subset Scan (FGSS)



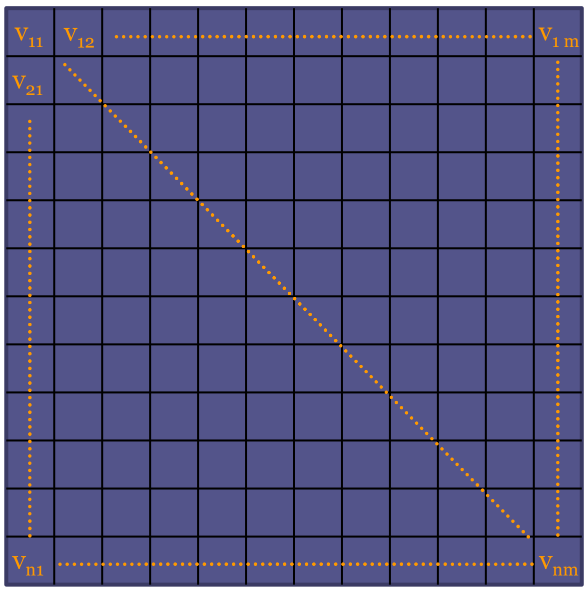
I. Compute the anomalousness of each attribute value (for each record)

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

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



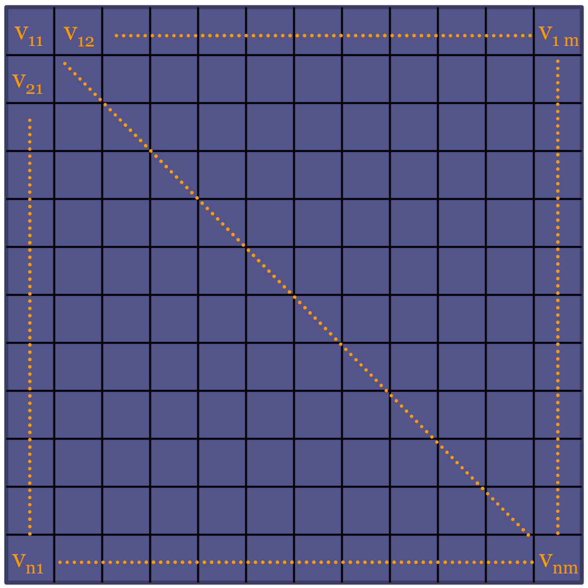
To compute the anomalousness of the data, FGSS models the data distribution given expected system behavior.

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

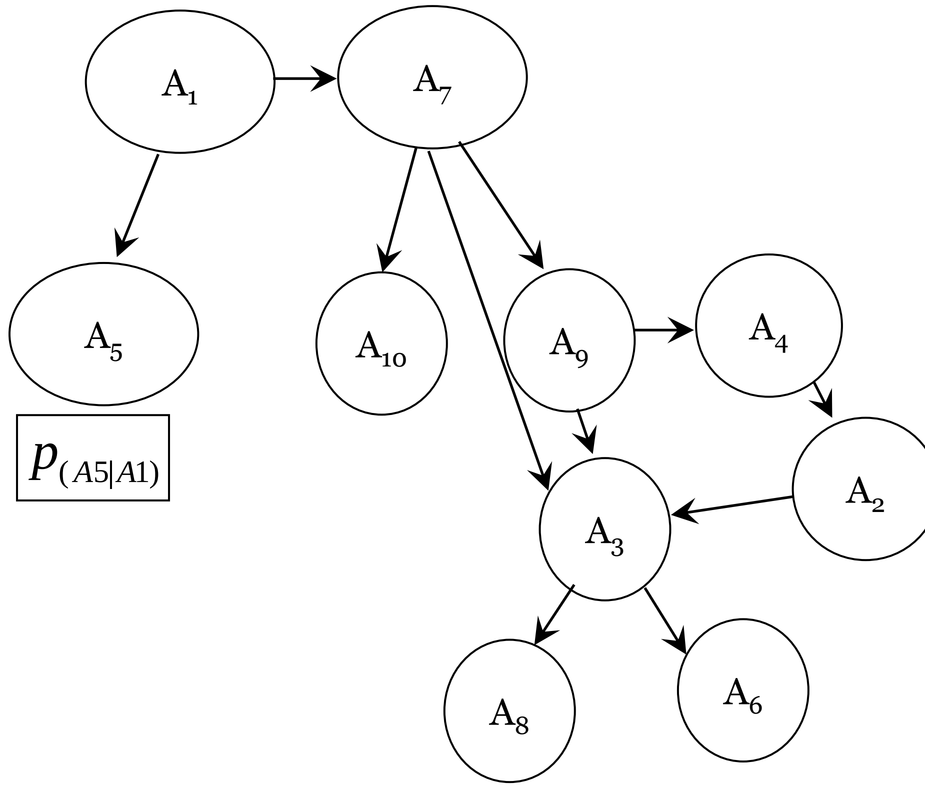
Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



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

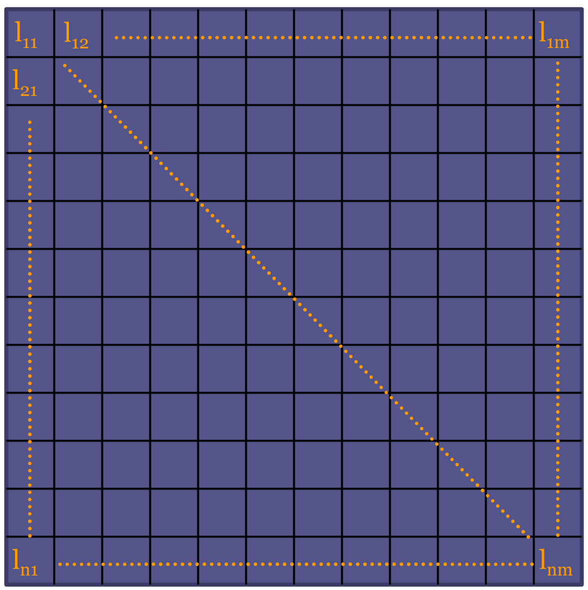


To compute the anomalousness of the data, FGSS models the data distribution given expected system behavior.

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



To compute the anomalousness of the data, FGSS models the data distribution given expected system behavior.

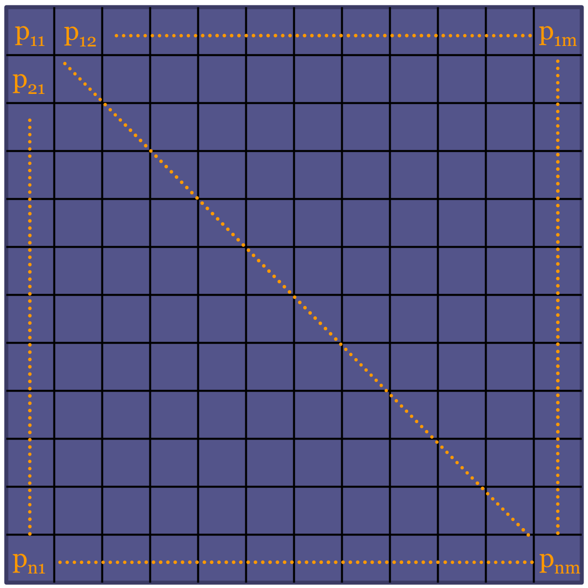
I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



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.

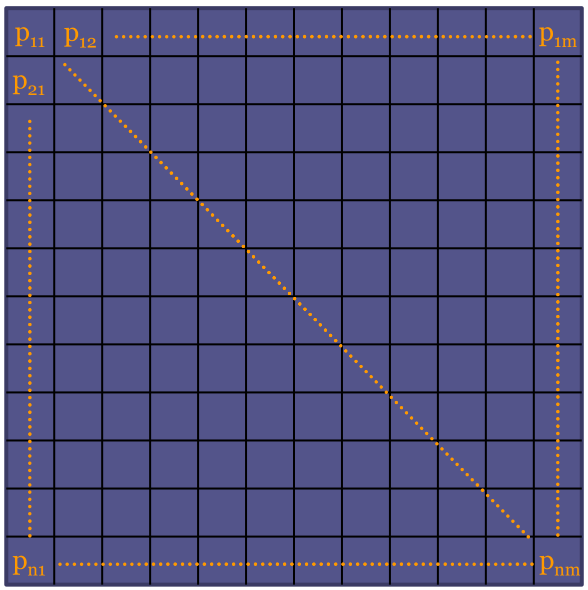
I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values
   i. Maps each attribute distribution to same space
   ii. $p_{ij} \sim$ Uniform(0,1) under H_0, so for a subset with N p-values, we expect $N\alpha$ to be less than α.

Fast Generalized Subset Scan (FGSS)

Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



A subset of data records and attributes with a higher than expected number of low (significant) p-values is 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
   i. Maps each attribute distribution to same space
   ii. $p_{ij} \sim \text{Uniform}(0,1)$ under H_0, so for a subset with N p-values, we expect $N\alpha$ to be less than α.

Fast Generalized Subset Scan (FGSS)

Nonparametric Scan Statistic (NPSS)

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

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

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

Example (Higher Criticism)

$$F(S) = \frac{N_{\alpha} - N_{tot}\alpha}{\sqrt{N_{tot}\alpha(1-\alpha)}}$$

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

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

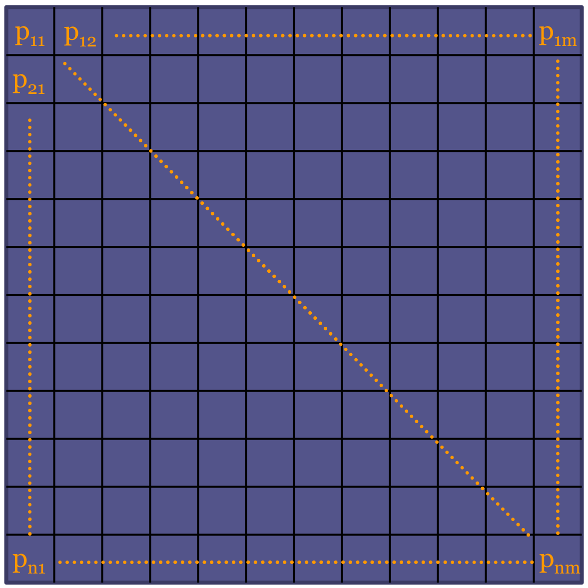
II. Discover subsets of records and attributes that are most anomalous

- Evaluate subsets with a **nonparametric scan statistic** (NPSS) to compare the actual and expected number of significant p-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

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 (Neill, 2012):

A score function F(S) satisfies LTSS if :

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

We only need to consider *N* subsets:

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

For a given subset of attributes, we can optimize $F_\alpha(S)$ over all 2^N subsets of records in 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 $O(2^{N+M})$
   - For NPSS, we can show that $F_\alpha(S)$ satisfies LTSS for each value of α.

Fast Generalized Subset Scan (FGSS)

Linear-Time Subset Scanning (Neill, 2012):

A score function F(S) satisfies LTSS if :

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

We only need to consider M subsets:

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

For a given subset of **records**, we can optimize $F_\alpha(S)$ over all 2^M subsets of **attributes** in O(M log M).

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 $O(2^{N+M})$
   - For NPSS, we can show that $F_\alpha(S)$ satisfies LTSS for each value of α.

Fast Generalized Subset Scan (FGSS)

Linear-Time Subset Scanning (Neill, 2012):

A score function F(S) satisfies LTSS if :

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

We only need to consider *M* subsets:

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

Thus we can **iterate** between optimizing over subsets of records and subsets of attributes.

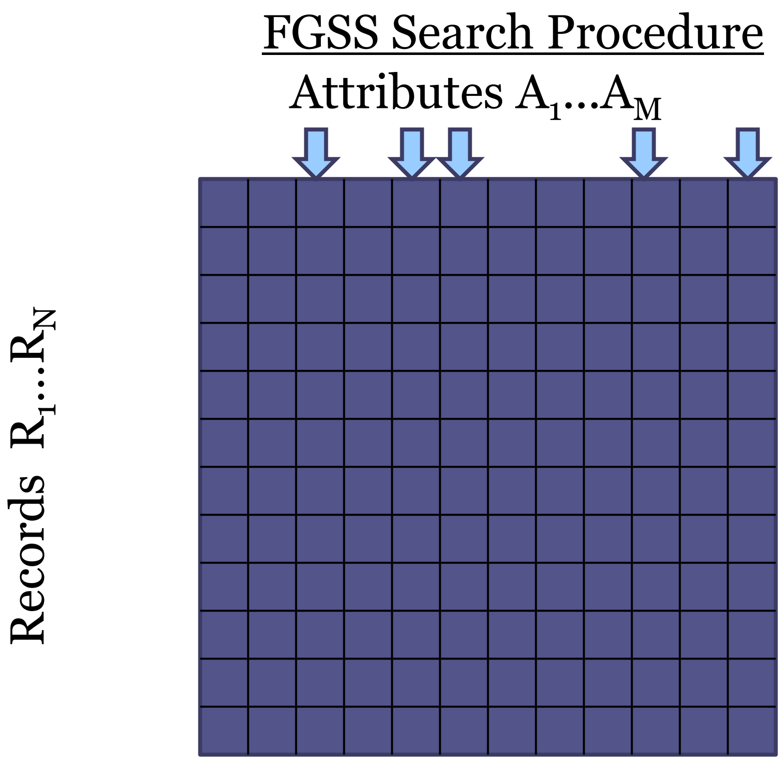
I. Compute the anomalousness of each attribute (for each record)

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
   - Naïve search is infeasible $O(2^{N+M})$
   - For NPSS, we can show that $F_\alpha(S)$ satisfies LTSS for each value of α.

Fast Generalized Subset Scan (FGSS)



1. Start with a randomly chosen subset of attributes

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

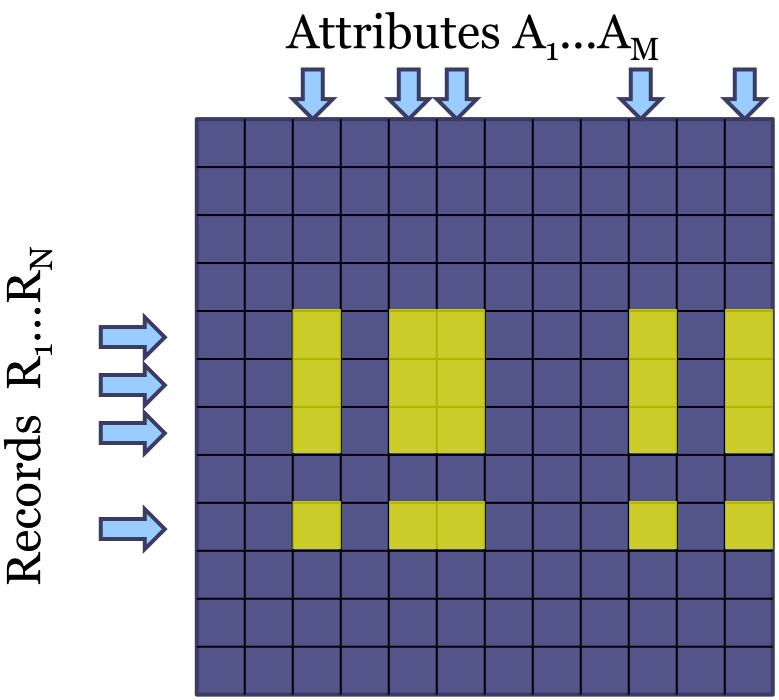
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

<u>FGSS Search Procedure</u>



(Score = 7.5)

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)

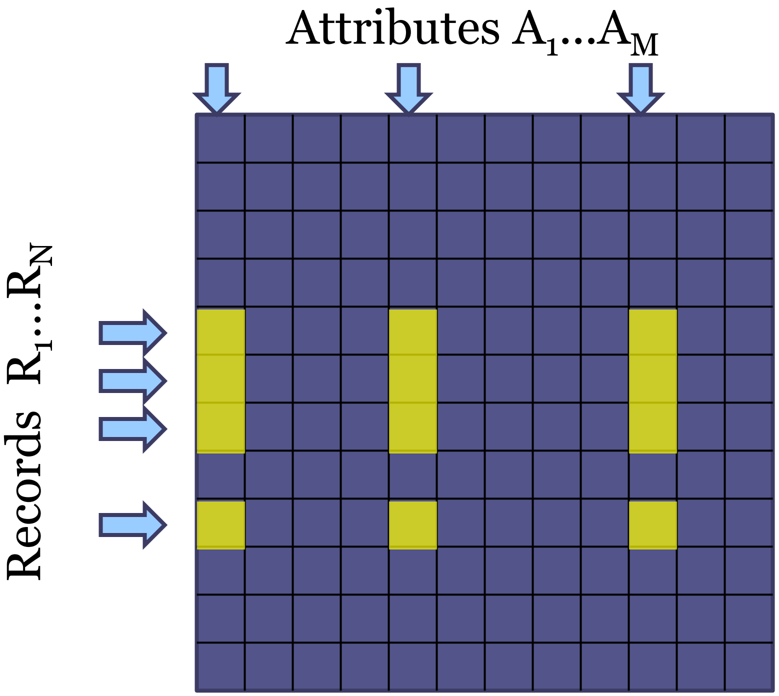
1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

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)

<u>FGSS Search Procedure</u>



(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

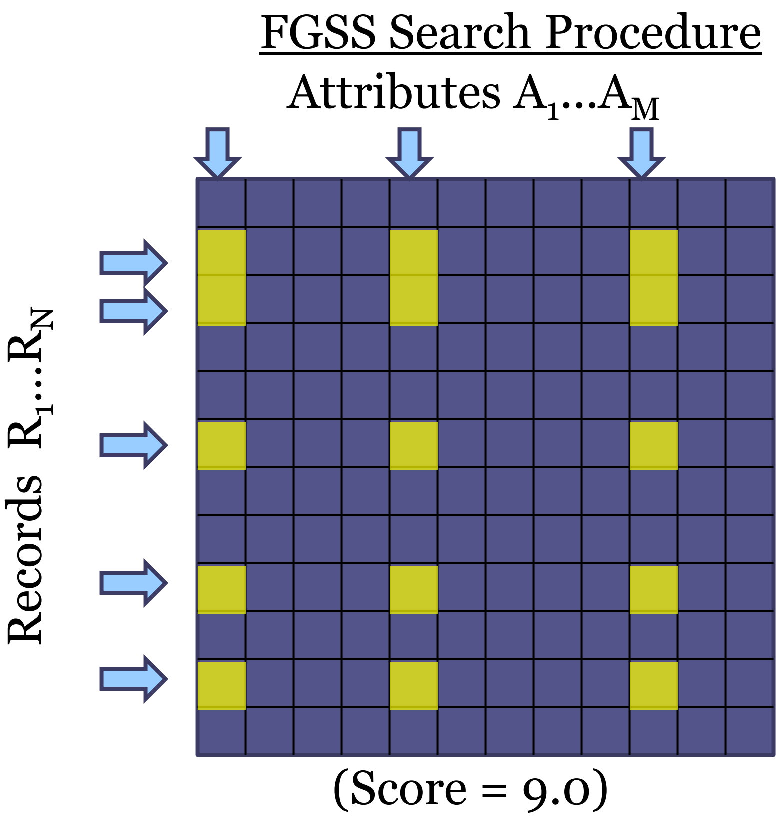
I. Compute the anomalousness of each attribute (for each record)

   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)



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

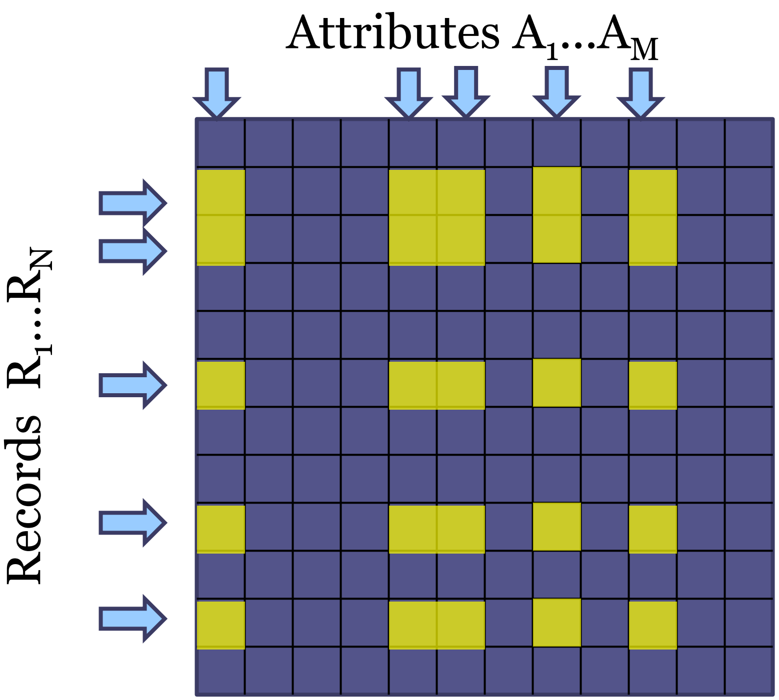
1. Maximize F(S) over all subsets of S
   - Iterate between following steps

   i. LTSS over records O(N log N)

   ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

<u>FGSS Search Procedure</u>



(Score = 9.3)

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)

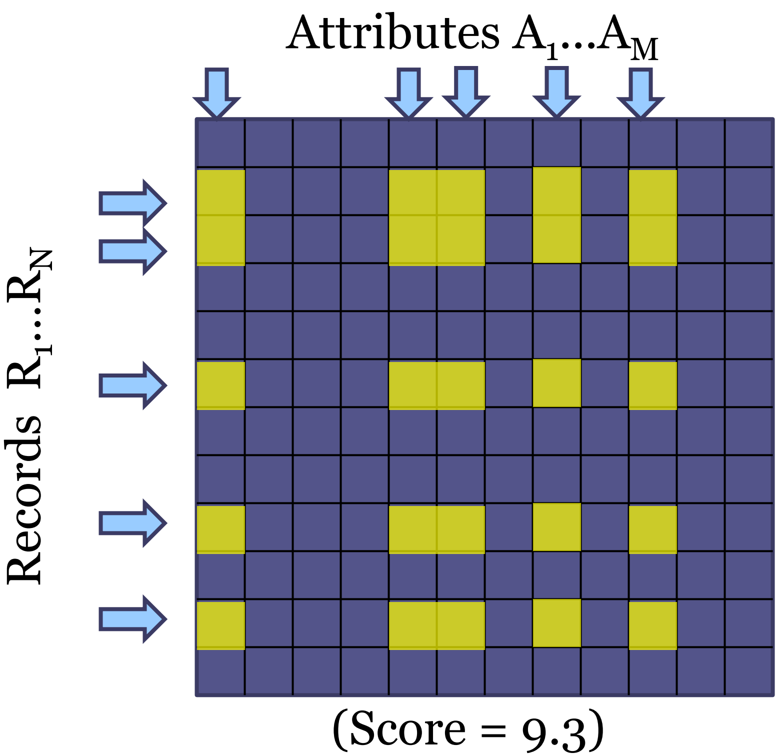
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

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)

<u>FGSS Search Procedure</u>



<u>Good News</u>: Run time is (near) linear in number of records & attributes.

<u>Bad News</u>: 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

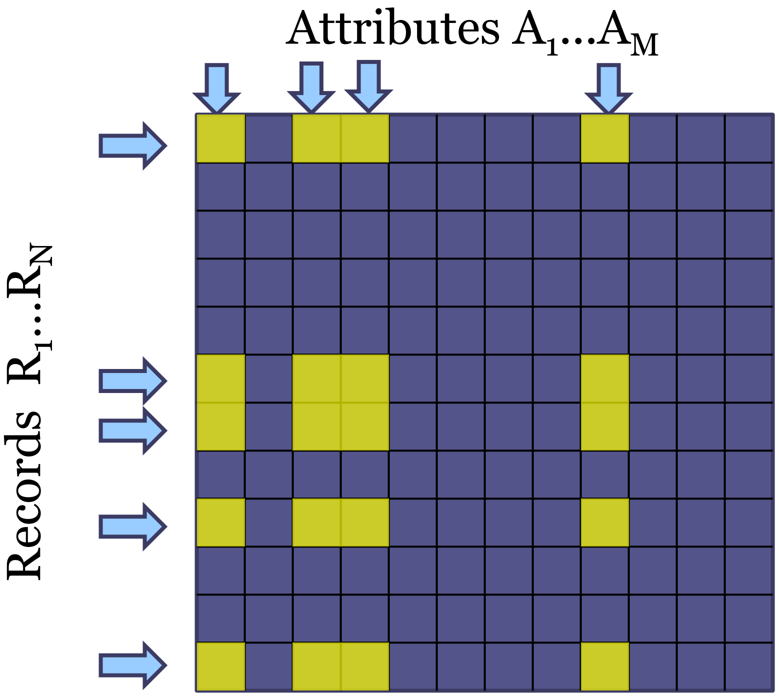
1. Maximize F(S) over all subsets of S
   - Iterate between following steps

     i. LTSS over records O(N log N)

     ii. LTSS over attributes O(M log M)

Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure



(Score = 11.0)

5. Repeat steps 1-4 for 50 random restarts

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

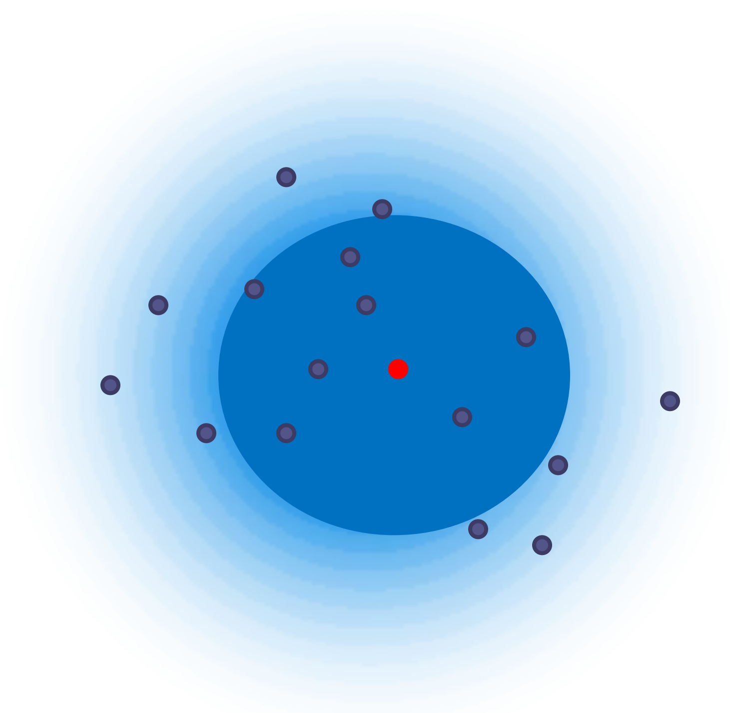
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

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, and thus we create local neighborhoods defined by a center record and all other records within a maximum 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

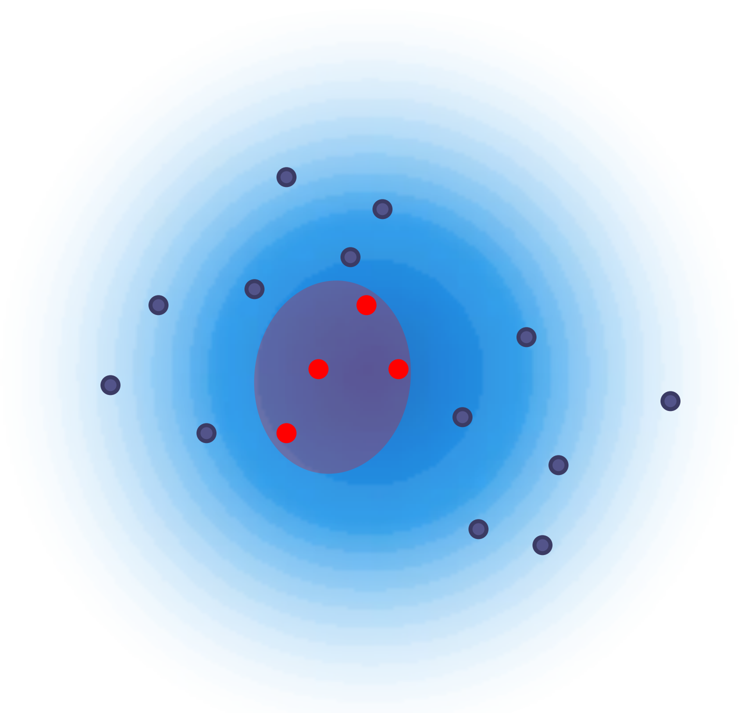
1. Maximize F(S) over all subsets of S
   - Iterate between following steps

   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 then perform the unconstrained scan over subsets of records and attributes within each 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

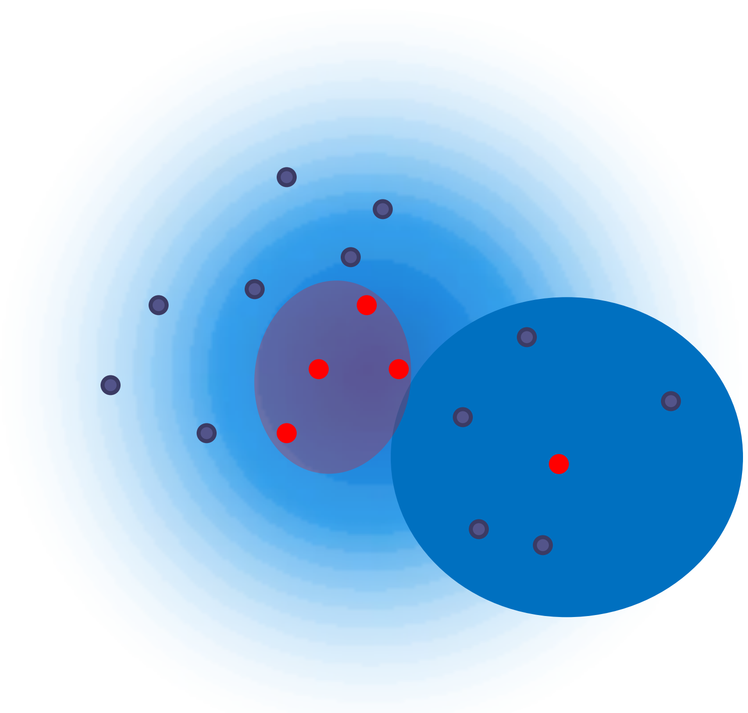
1. Maximize F(S) over all subsets of S
    - Iterate between following steps

    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 then perform the unconstrained scan over subsets of records and attributes within each 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

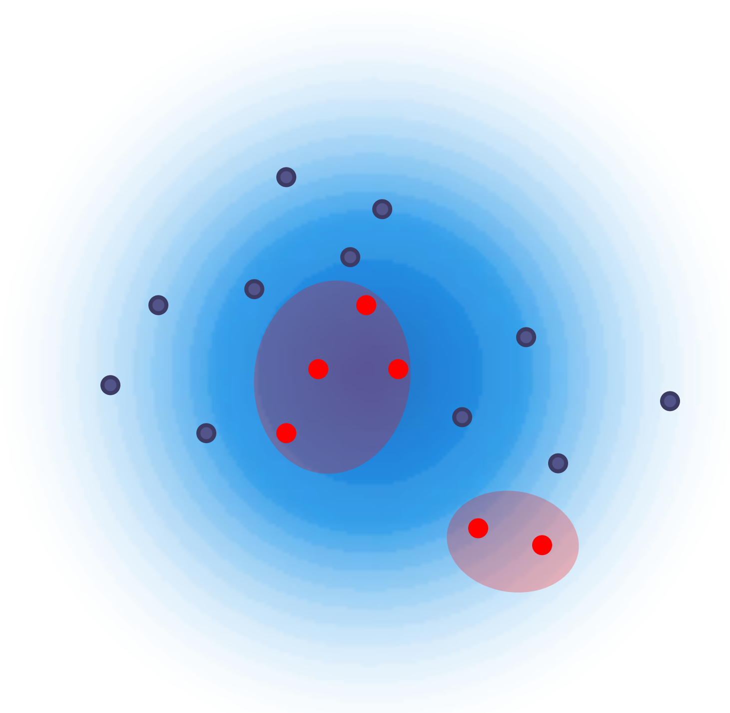
1. Maximize F(S) over all subsets of S
   - Iterate between following steps

   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 then perform the unconstrained scan over subsets of records and attributes within each 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

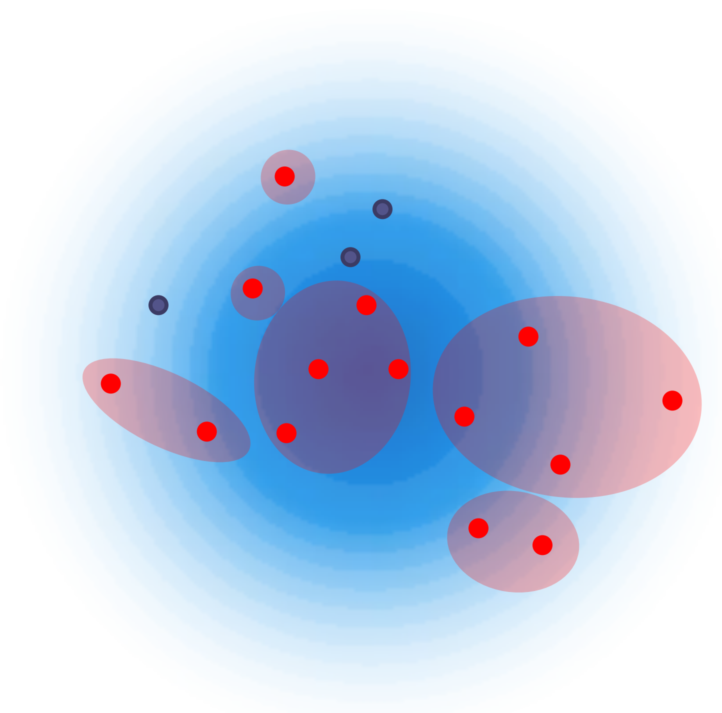
1. Maximize F(S) over all subsets of S
   - Iterate between following steps

   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>



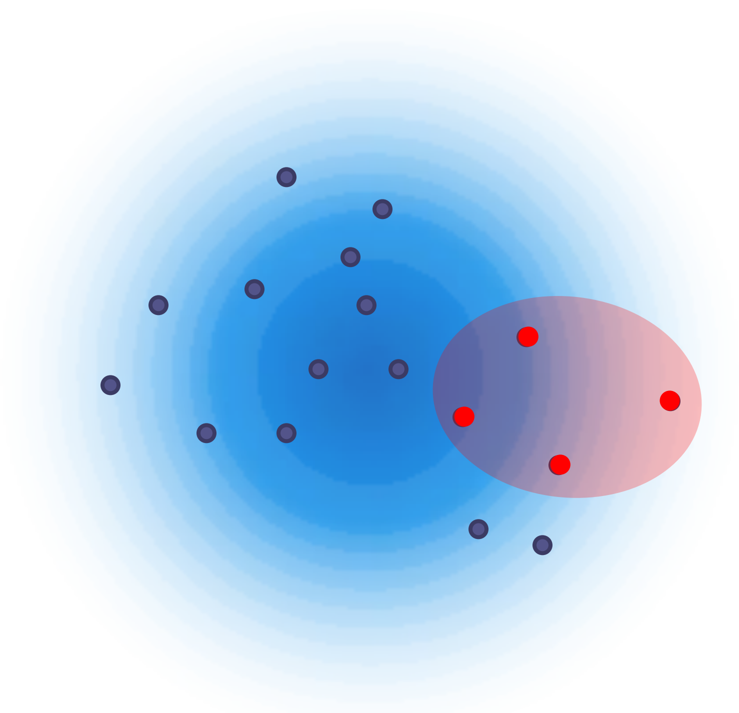
We then perform the unconstrained scan over subsets of records and attributes within each neighborhood.

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



Finally, we choose the neighborhood-constrained subset which maximizes F(S).

Optionally, we can compute statistical significance by randomization testing.

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

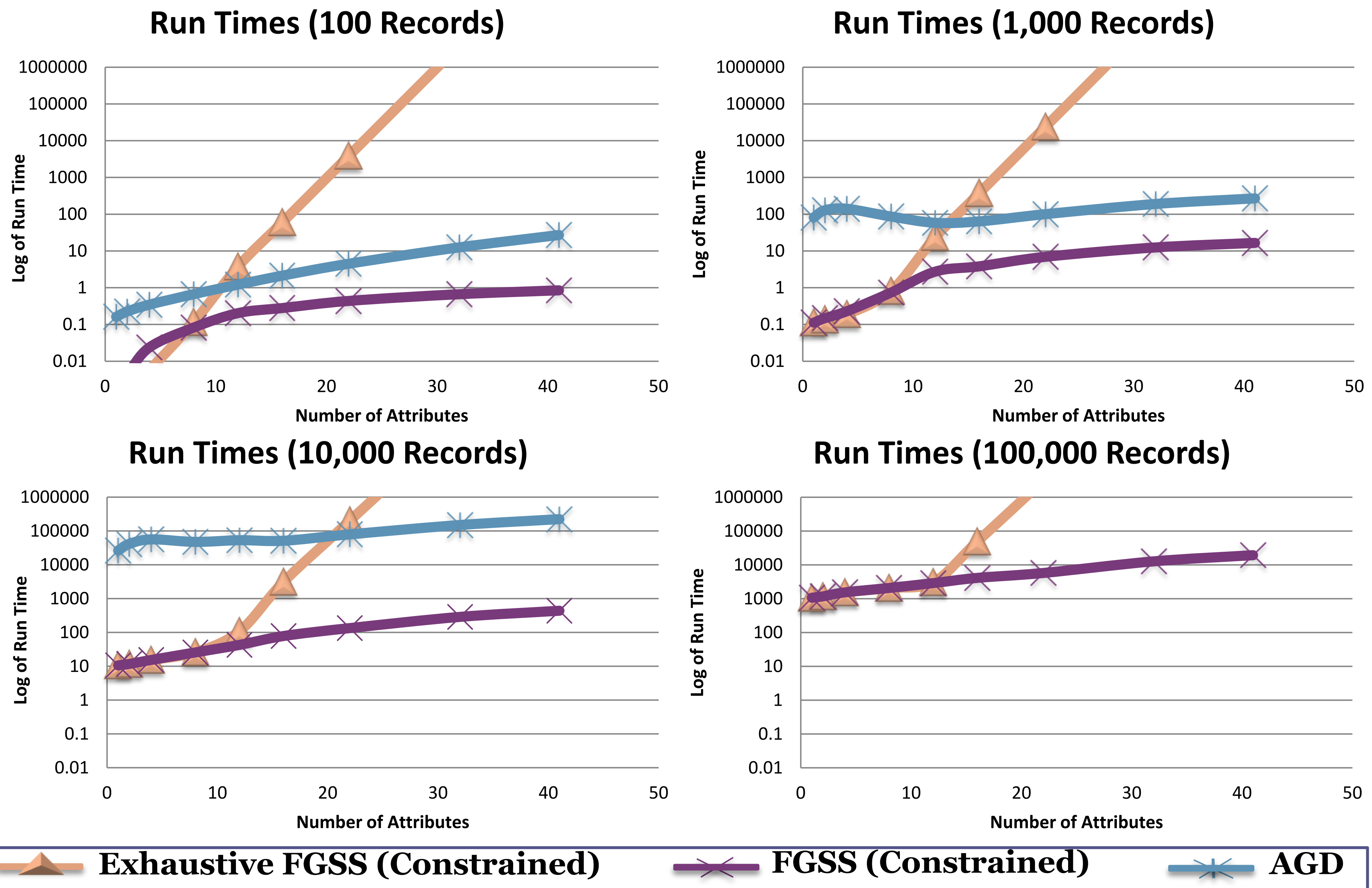
   i. LTSS over records O(N log N)

   ii. LTSS over attributes O(M log M)

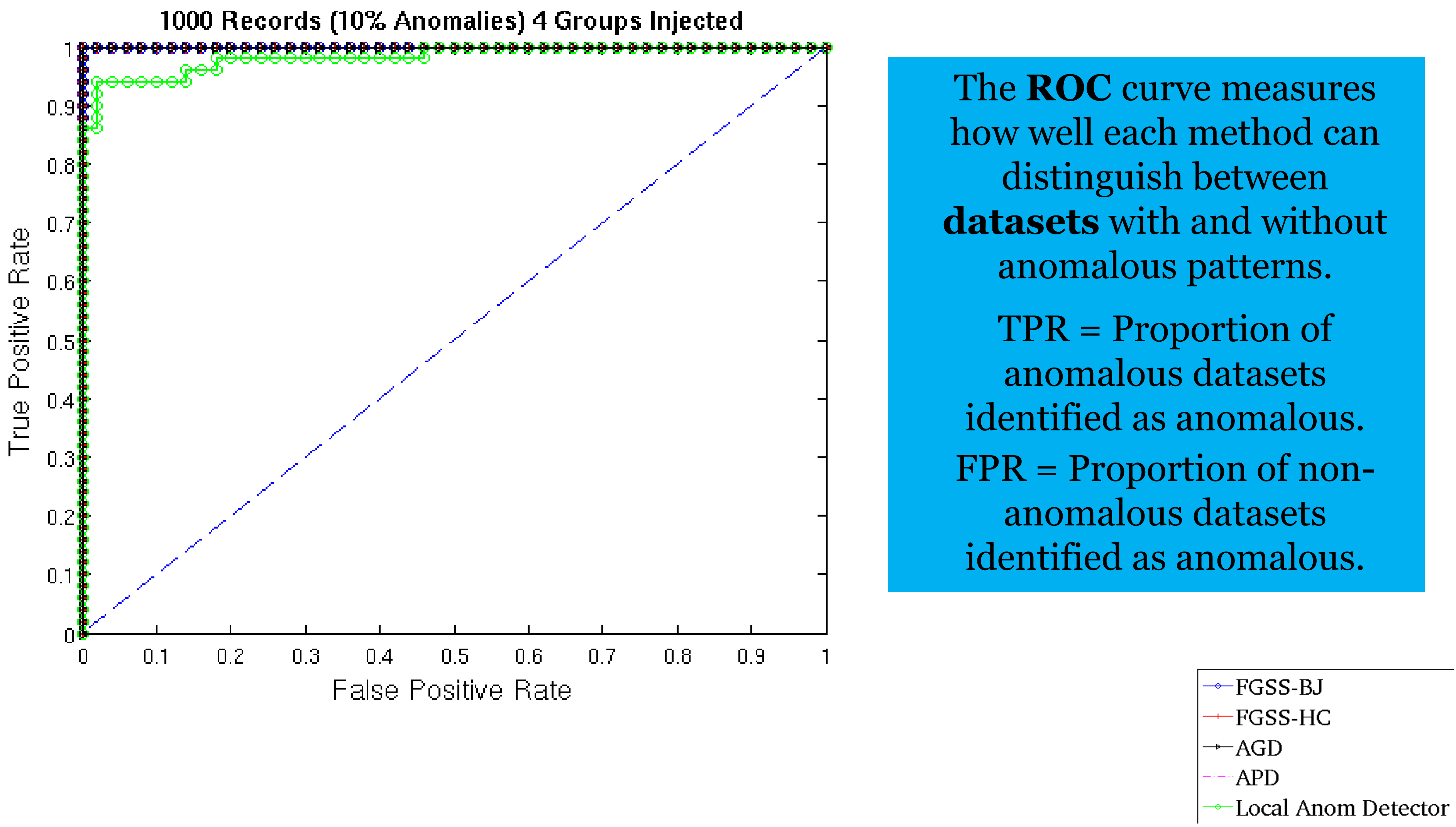
Experiments

- Network activity and intrusion data (KDDCUP ’99)
  - 41 attributes representing extracted information from the raw data of the network connection.
- Simulated anthrax outbreaks in Emergency Dept. visits
  - Hospital id
  - Prodrome (classification of free-text chief complaint)
  - Patient age decile
  - Patient home zip code
- U.S. Customs and Border 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-based anomaly detector (BN)
  - 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
1000000
100000
10000
1000
100
10
1
0.1
0.01
0
10
20
30
40
50
Number of Attributes
Exhaustive FGSS (Constrained)
FGSS (Constrained)
AGD

Receiver Operator Characteristic

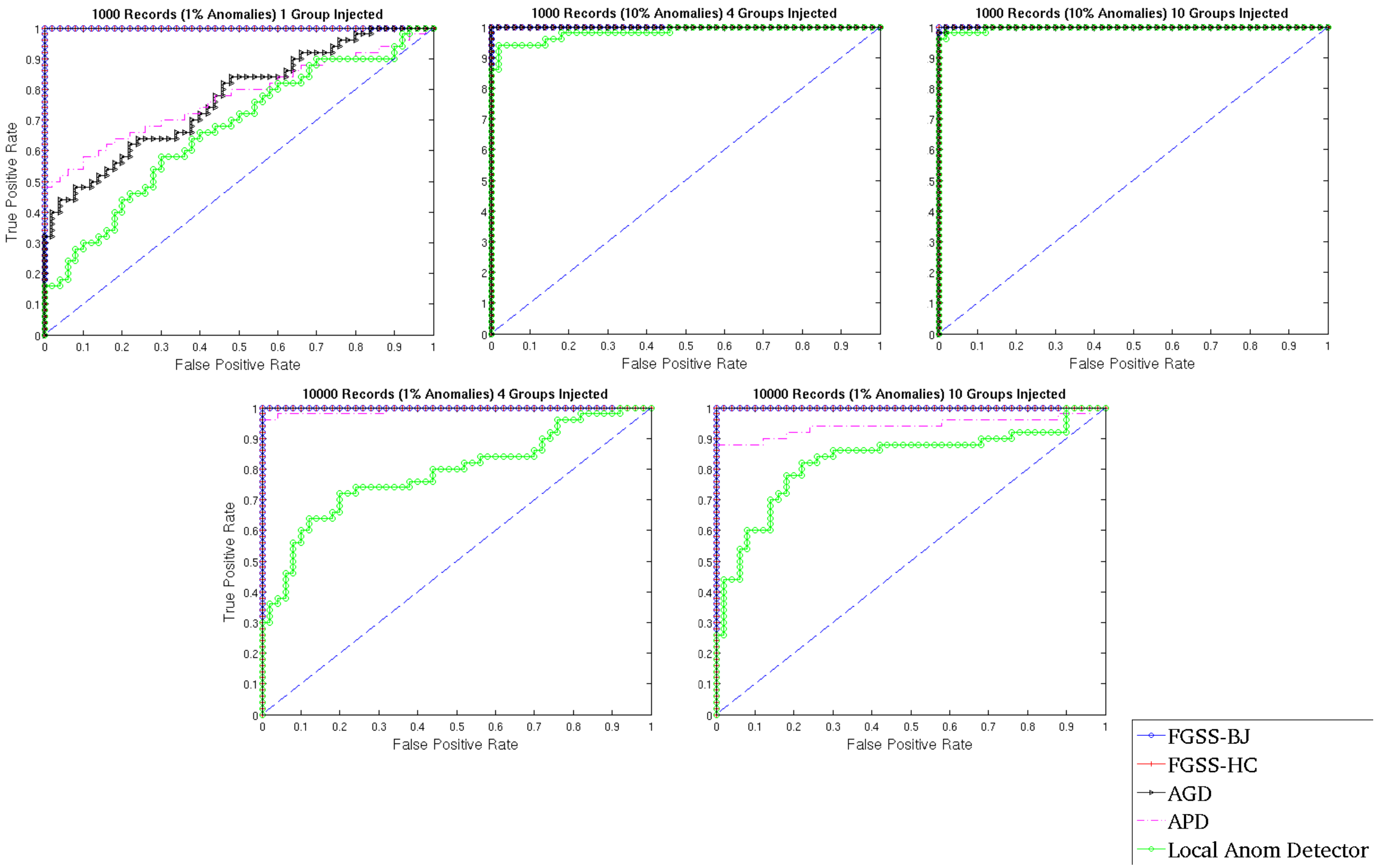


The **ROC** curve measures how well each method can distinguish between **datasets** with and without anomalous patterns.

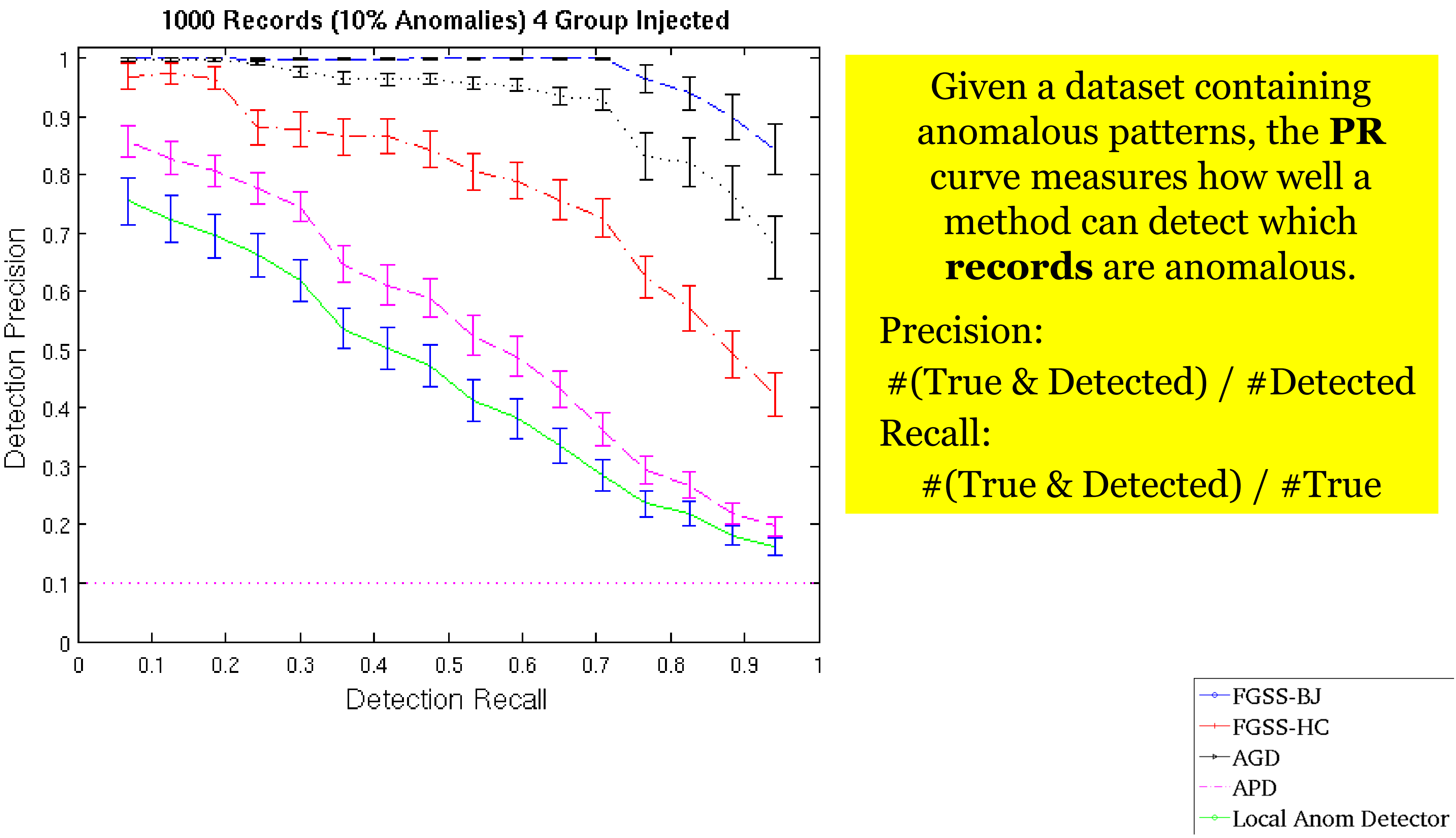
TPR = Proportion of anomalous datasets identified as anomalous.

FPR = Proportion of non-anomalous datasets identified as anomalous.

Receiver Operator Characteristic



Precision vs. Recall Curves



Given a dataset containing anomalous patterns, the **PR** curve measures how well a method can detect which **records** are anomalous.

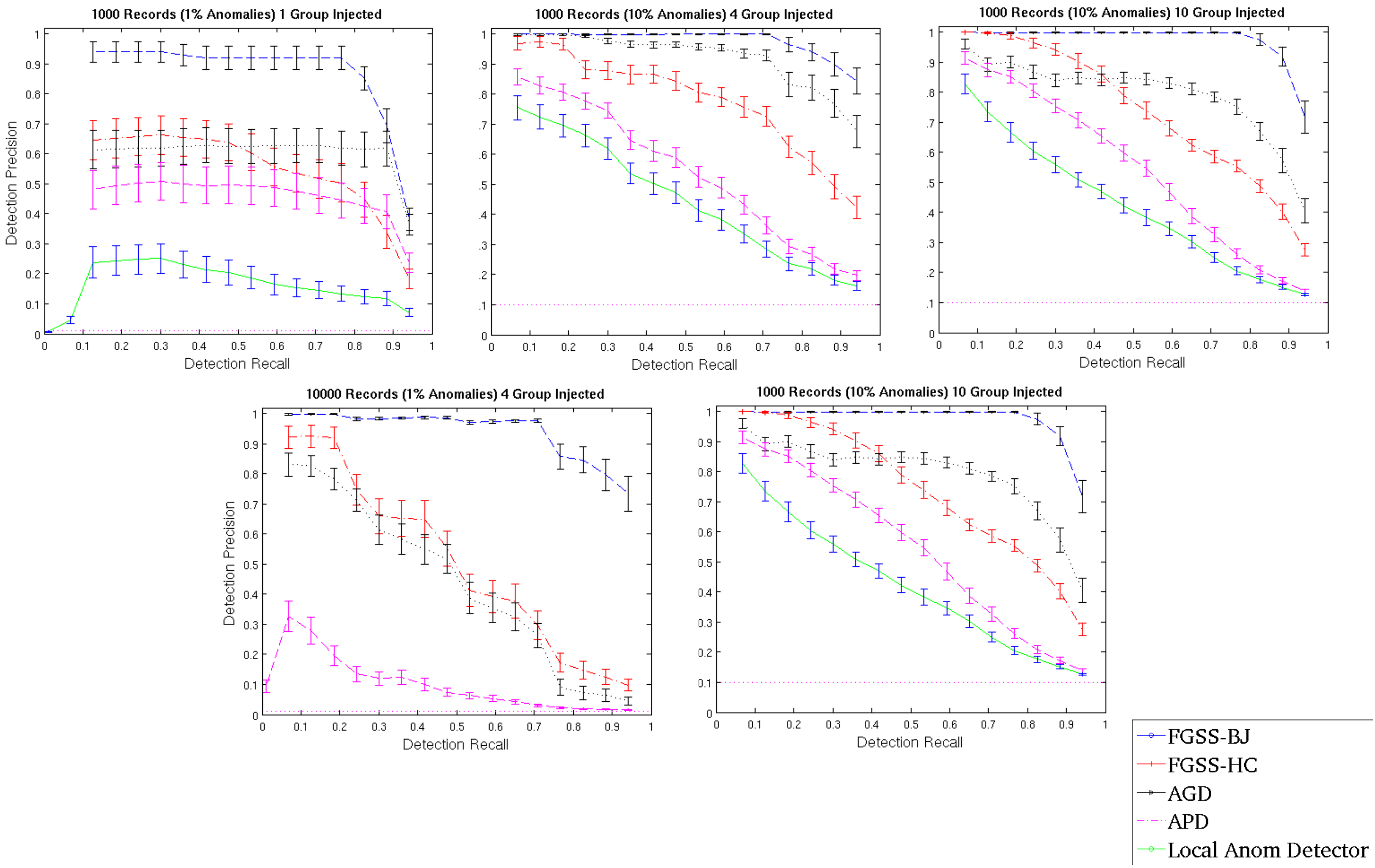
Precision:

#(True & Detected) / #Detected

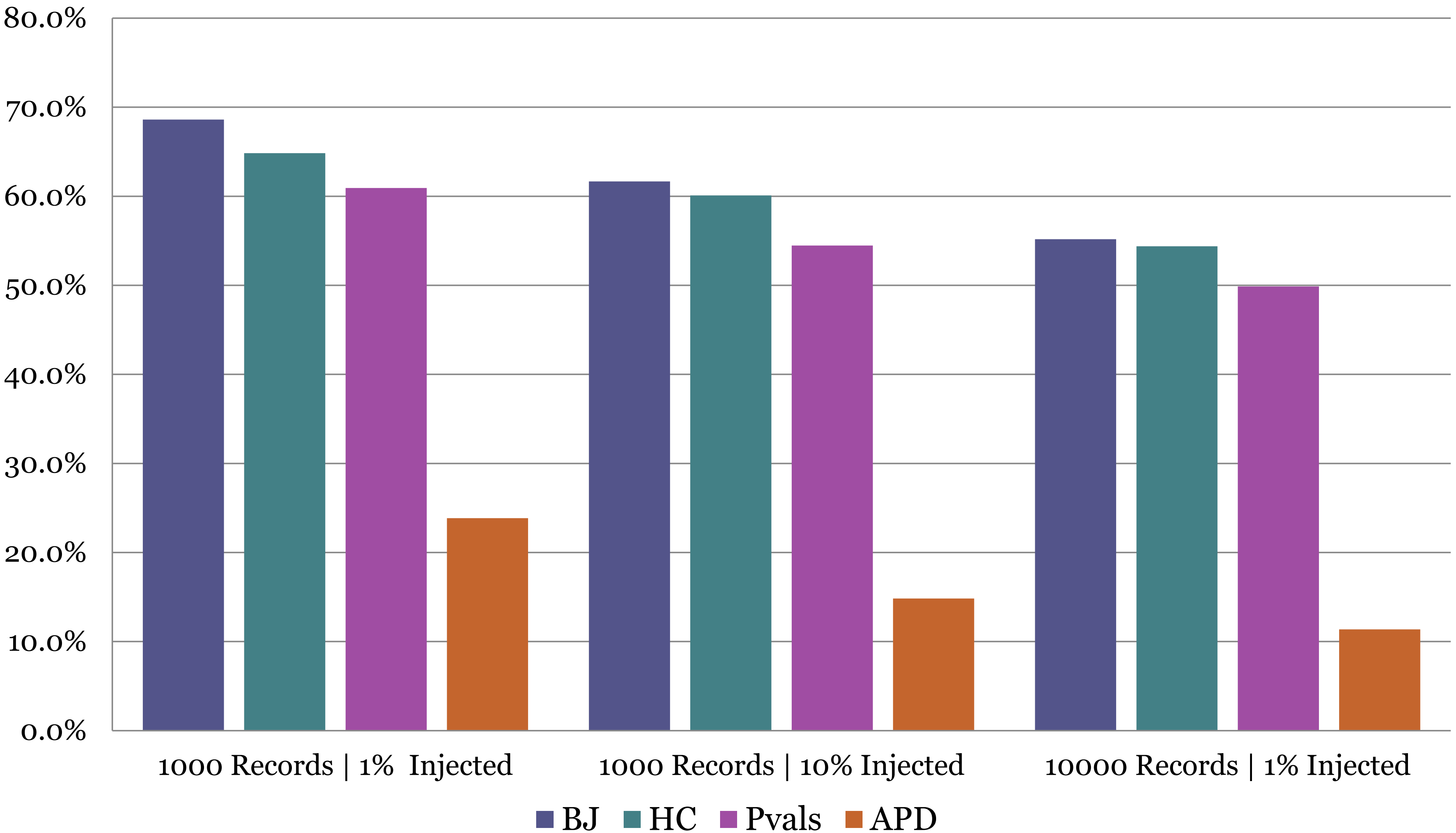
Recall:

#(True & Detected) / #True

Precision vs. Recall Curves



Pattern Characterization Accuracy



Conclusions

- FGSS is a general method for anomalous pattern detection which can be applied across many application domains.
- FGSS improves detection power and characterization accuracy as compared to competing methods, particularly when the patterns are:
  - a small portion of the data
  - subtle (not extremely individually anomalous)
- Extensions
  - Extend method to handle multiple anomaly detectors
  - Extend method to handle multiple models (find subsets not explained by any of the known patterns in the data)
  - Current applications include detection of anomalous patterns of patient care which influence health outcomes.