Discovering Novel Anomalous Patterns in General Data

Edward McFowland III (mcfowland@cmu.edu)
PhD Candidate
Event & Pattern Detection Lab
Carnegie Mellon University

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

A Day In The Life:
A Computer Systems Security Analyst

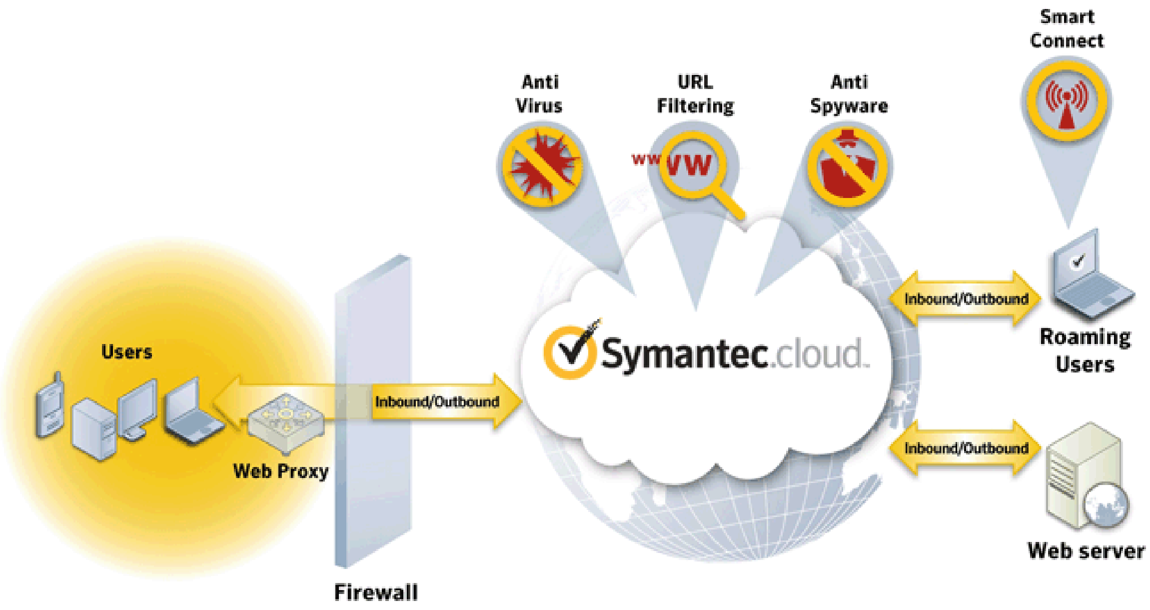
George

A Day In The Life:
A Computer Systems Security Analyst

George

A Day In The Life:
A Computer Systems Security Analyst

George



George

A Day In The Life:
A Computer Systems Security Analyst

- George's job is very challenging
  - 150,000 new malware strains are released each day[1]
  - Novel attacks can show up in the queues of analyst but may go unnoticed
    - Looking for a needle in a haystack, without really knowing the shape, size, or color of the needle.

- Ideal Scenario?
  - Continual discovery of novel security breaches
  - A fundamental approach to sort his queue
    - Priority: unknown, potentially pernicious attacks

[1]Morales, Jose. "Prioritizing Malware Analysis." *Software Engineering Institute Blog*. 4 November 2013.

Recent Advances In Network Security

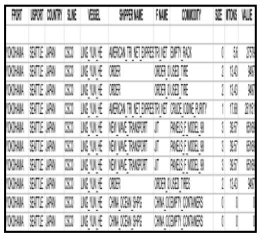
- Intrusion Detection System (IDS) (~ 1997)
  - Signature based detection (Virus Scanning)
    - Can only recognize things it already knows
      - (Essentially) NO power to detect novel attacks

- Anomaly Based-IDS (~ last 5-10 years)
  - Detect activity that falls out of normal system operation
    - Increased power to detect novel attacks
  - "Abnormal" activity may not be an attack
    - Raises alerts on non-attacks

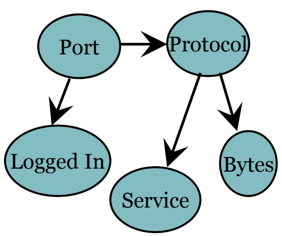
- Anomalous Pattern Detection (McFowland et al, 2013)
  - Fast Generalized Subset Scan (FGSS)
    - Groups of records that are collectively anomalous given normal system operation
    - Significantly increased power to detect novel attacks
    - Significantly decreased alarms on non-attacks

Anomalous Pattern Detection Procedure

Test Data

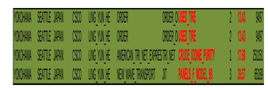


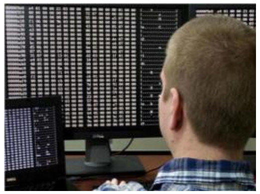
Normal Activity (M_o)



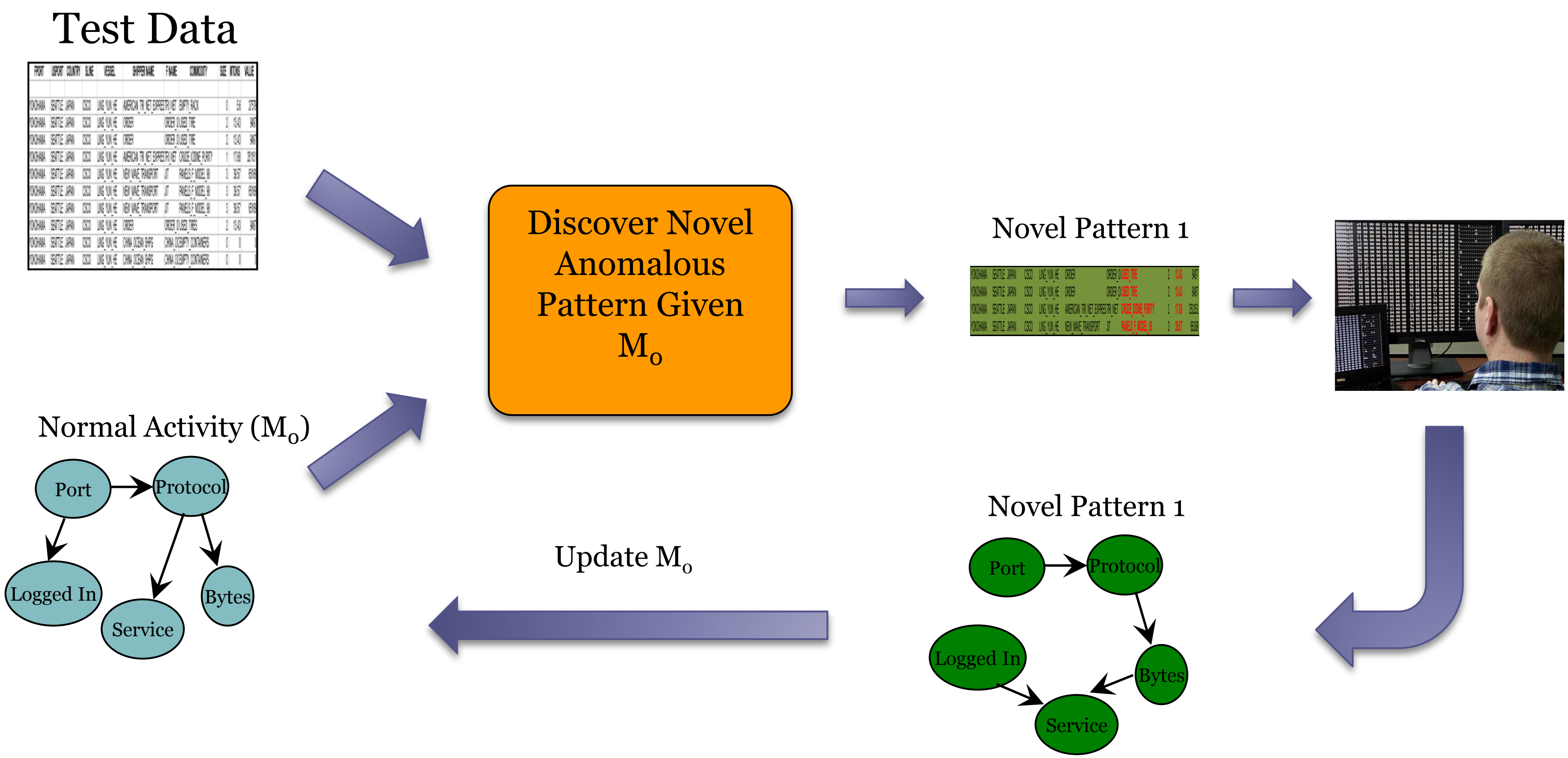
Discover Novel Anomalous Pattern Given M_o

Novel Pattern 1

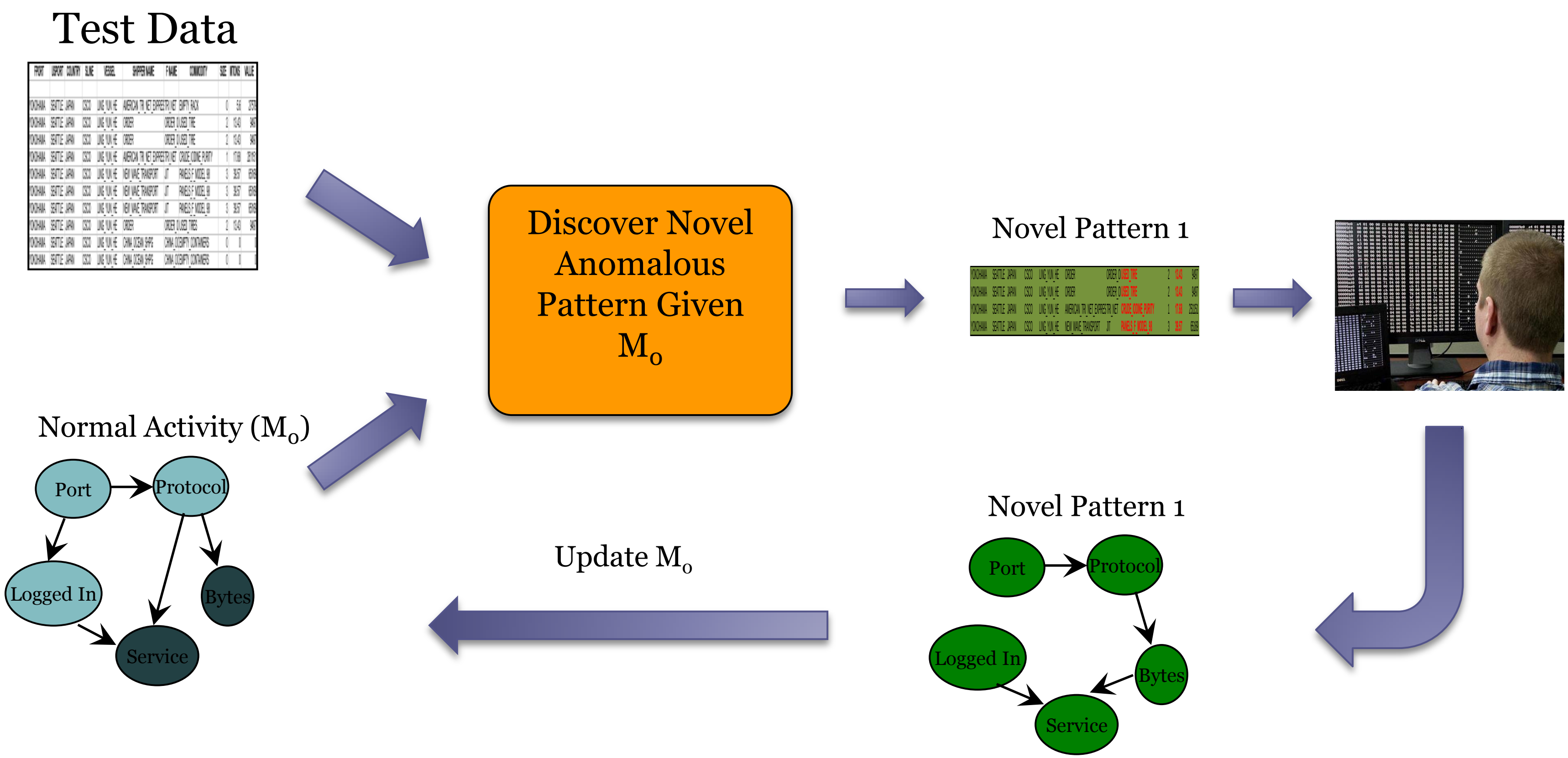




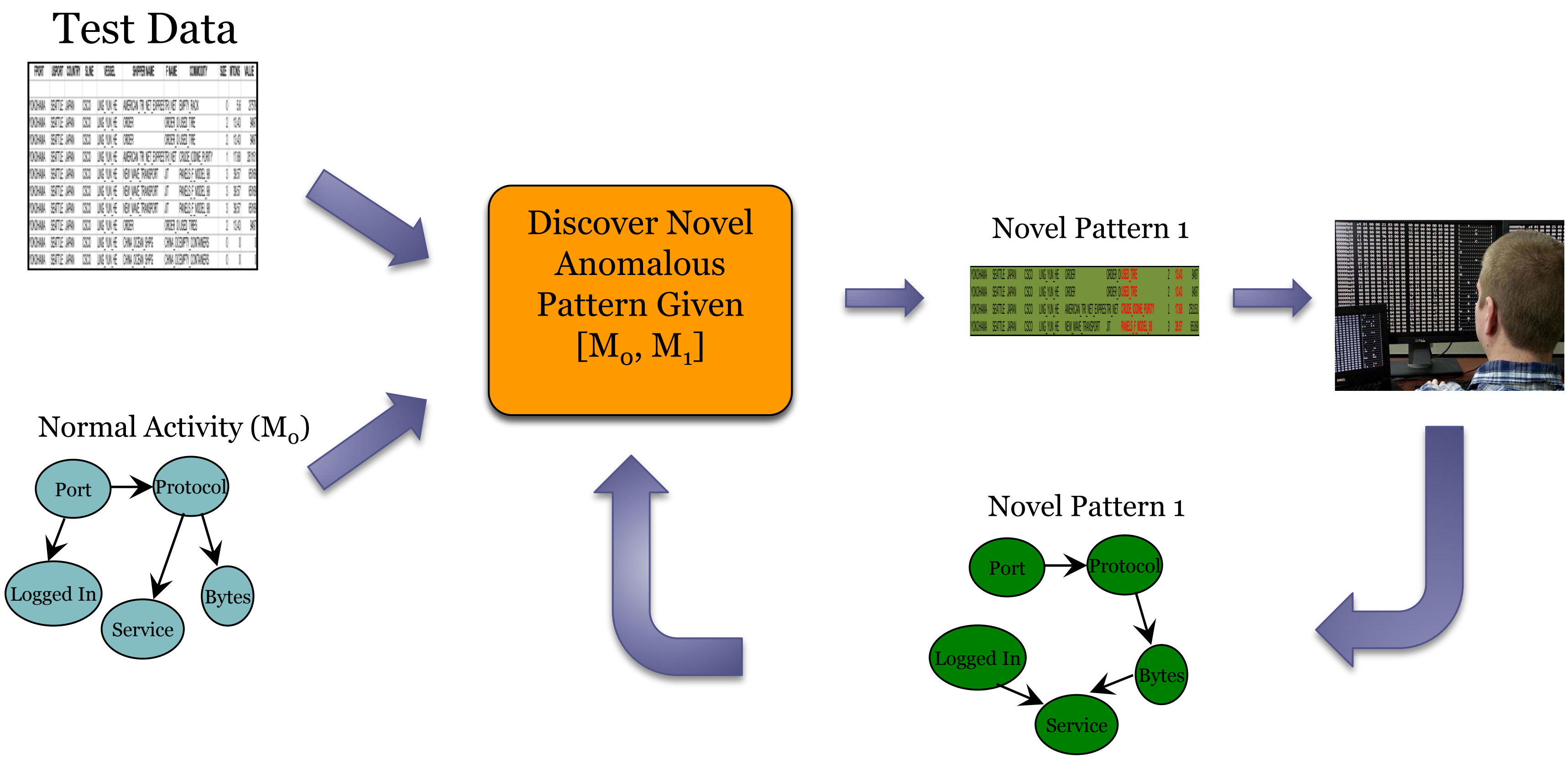
Anomalous Pattern Discovery Procedure



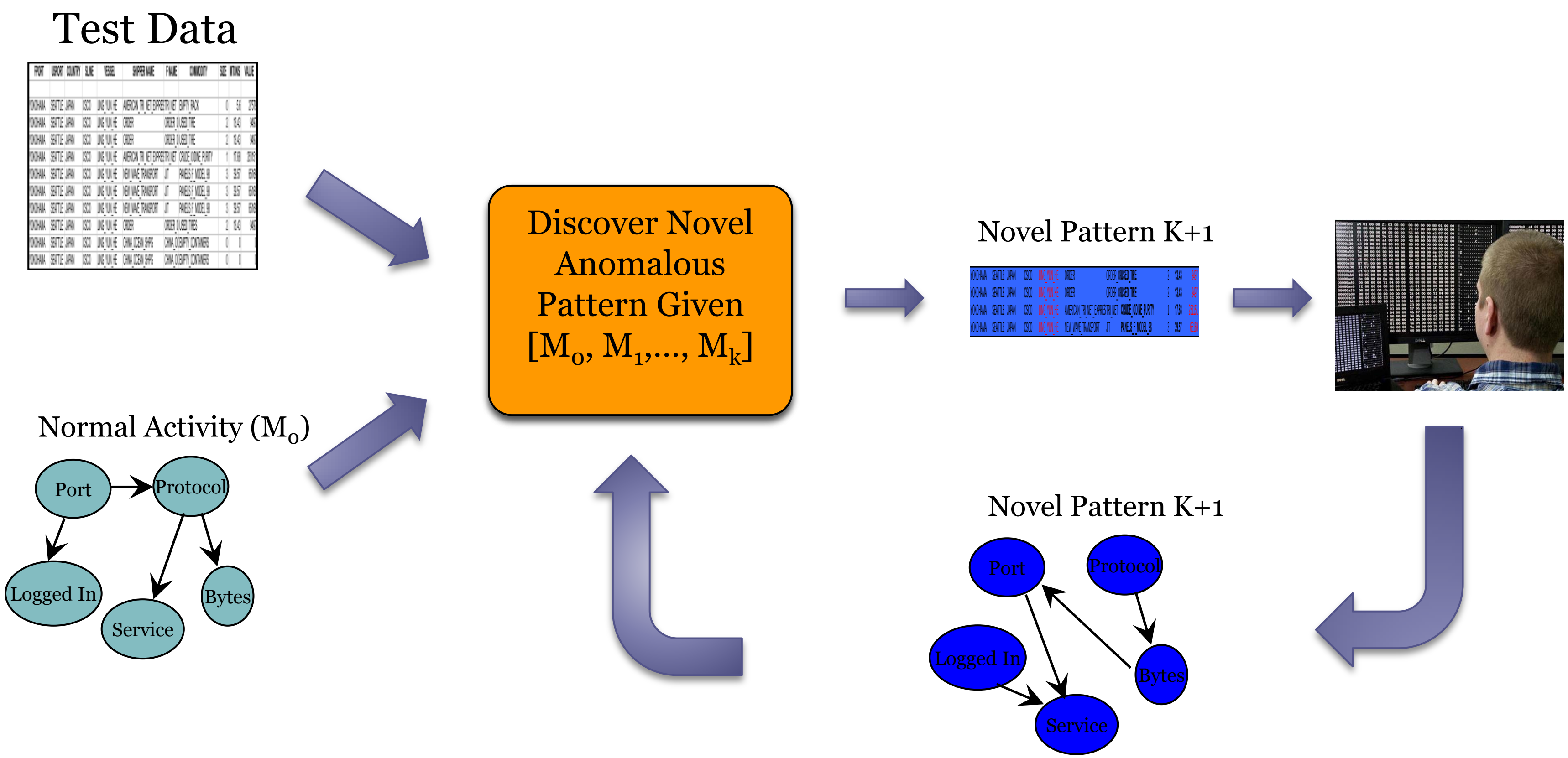
Anomalous Pattern Discovery Procedure



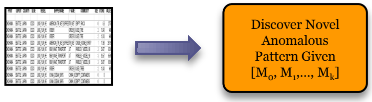
Anomalous Pattern Discovery Procedure

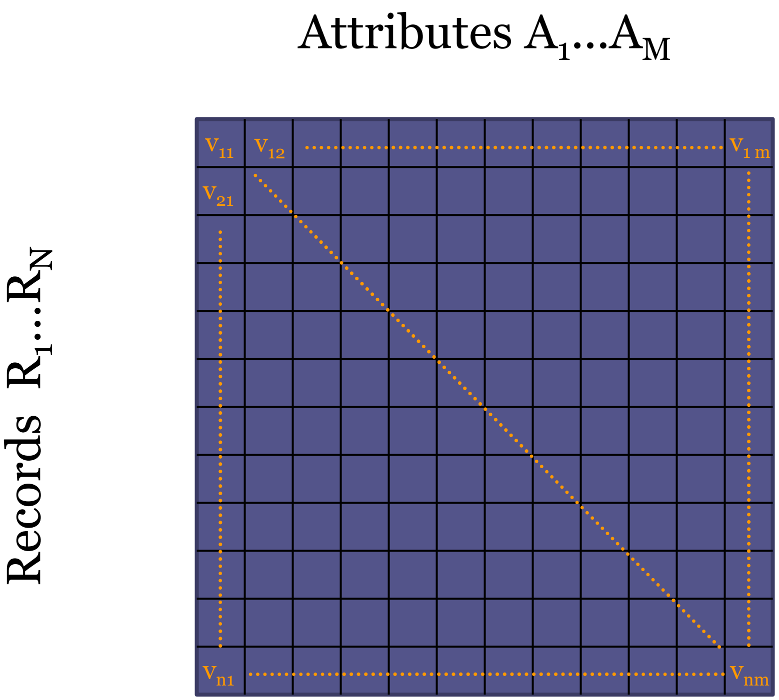


Anomalous Pattern Discovery Procedure

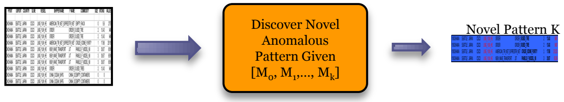


The Goal!


Discover Novel Anomalous Pattern Given [M_0, M_1,..., M_k]

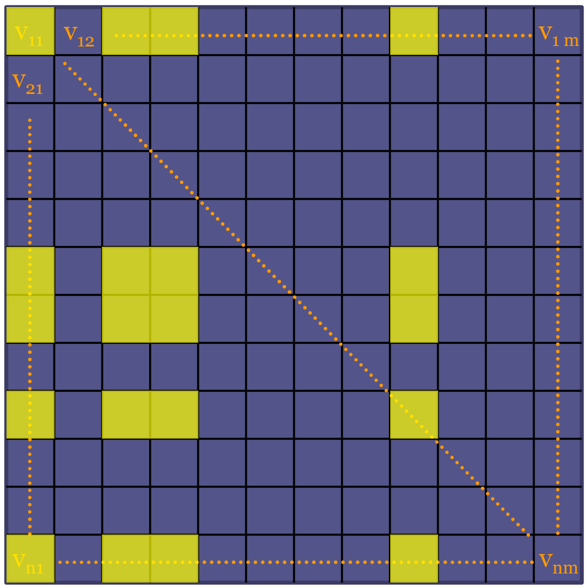

Attributes $A_1 \ldots A_M$
Records $R_1 \ldots R_N$
V_{11}
V_{12}
V_{1m}
V_{21}
V_{n1}
V_{nm}

The Goal!



Attributes $A_1 \ldots A_M$

Records $R_1 \ldots R_N$



Discover subsets of records, for which some subset of their attributes are the most anomalous!

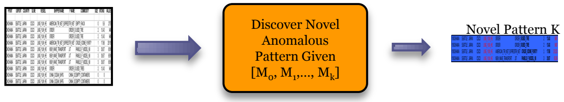
<u>The Optimization</u>

$$R \subseteq \{R_1 \ldots R_N\} \quad A \subseteq \{A_1 \ldots A_M\}$$

$$S = R \times A$$

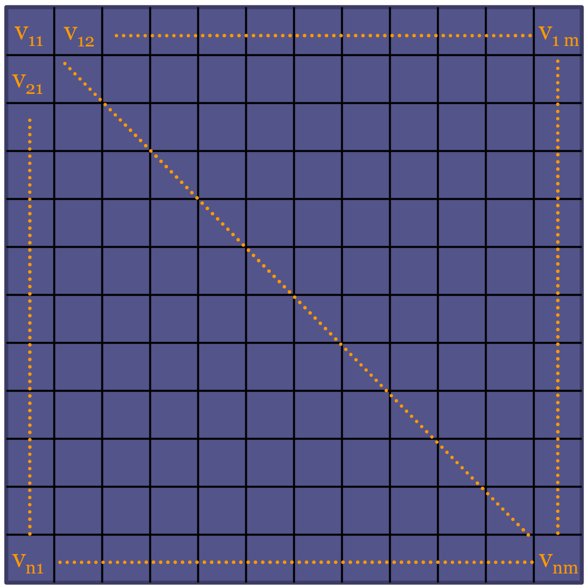
$$S^* = \mathit{argmax}_S \, F(S)$$

The DAP Algorithm



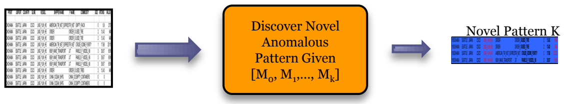
Attributes $A_1 \ldots A_M$

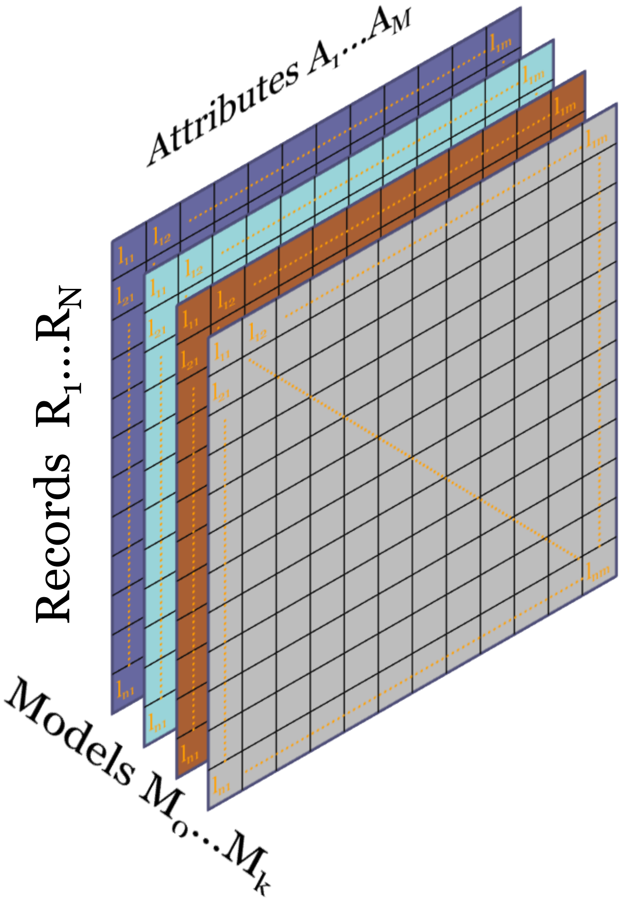
Records $R_1 \ldots R_N$



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

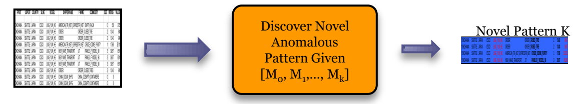
The DAP Algorithm

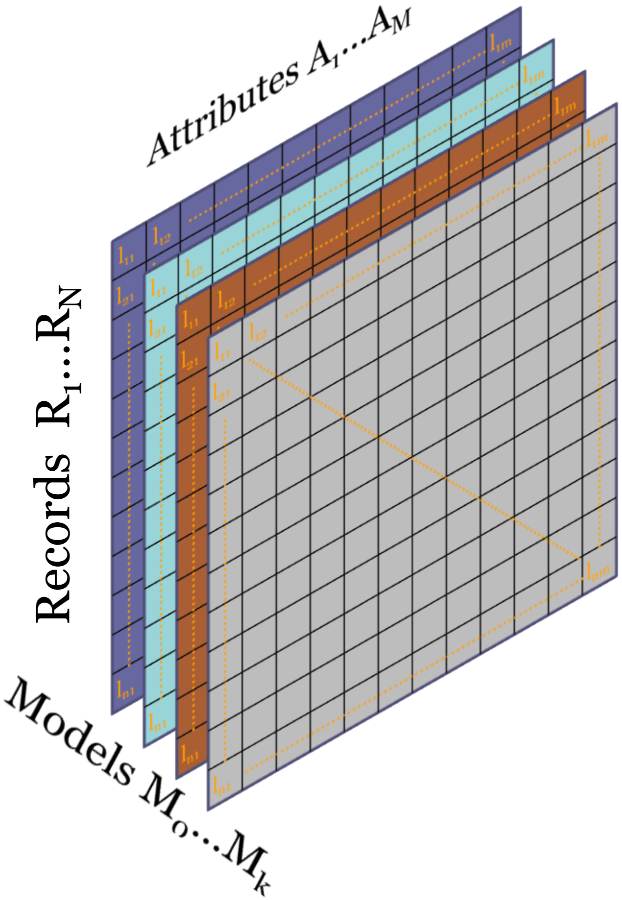




I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

The DAP Algorithm

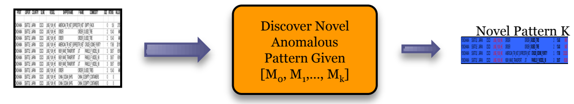


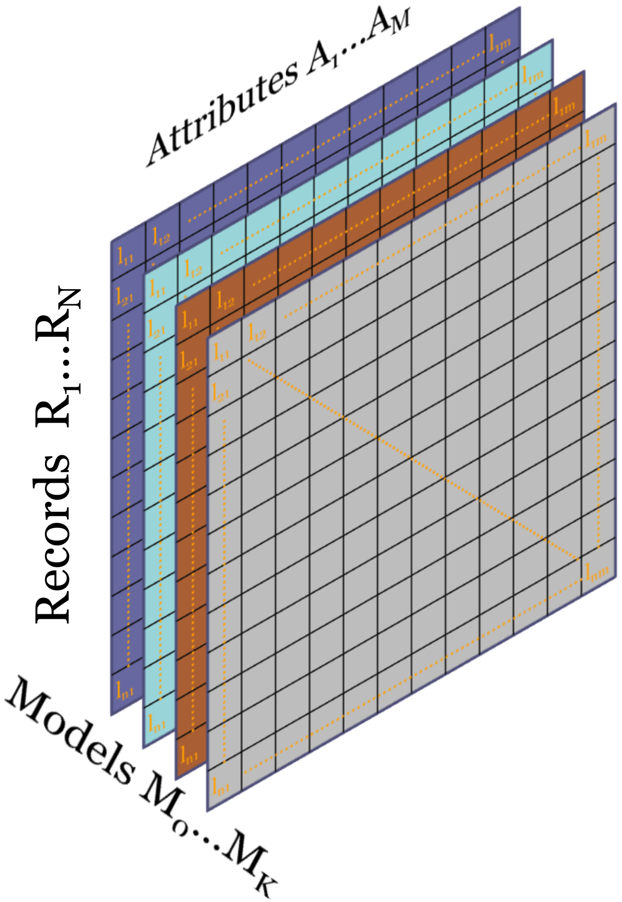


I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

The DAP Algorithm





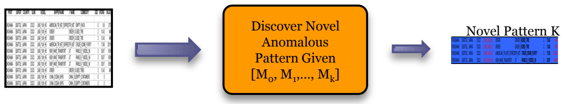
I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

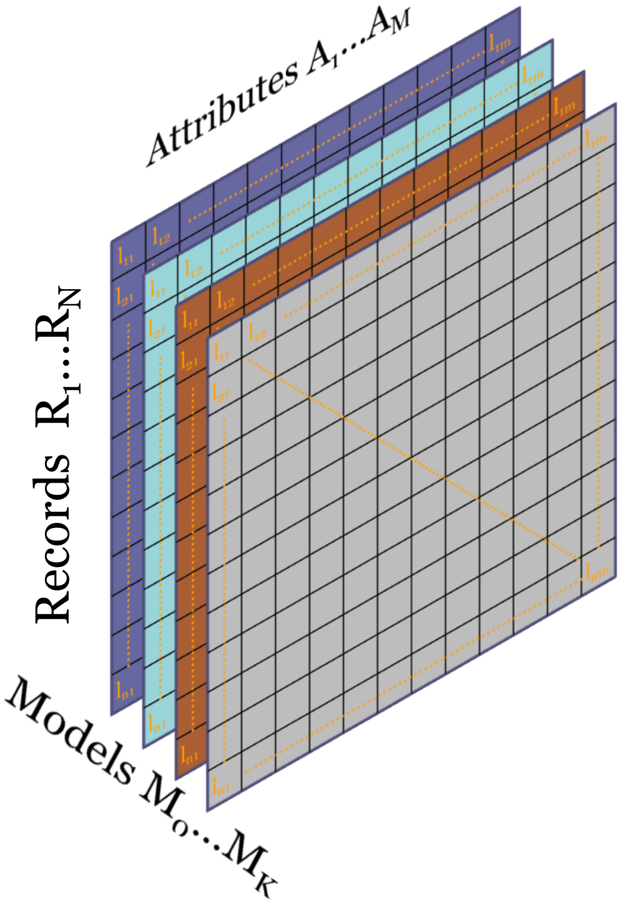
- Compute empirical p-values

  i. measures the interestingness of a v_{ij} under each M_k

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

The DAP Algorithm





I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

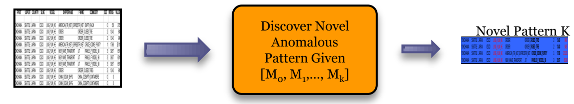
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

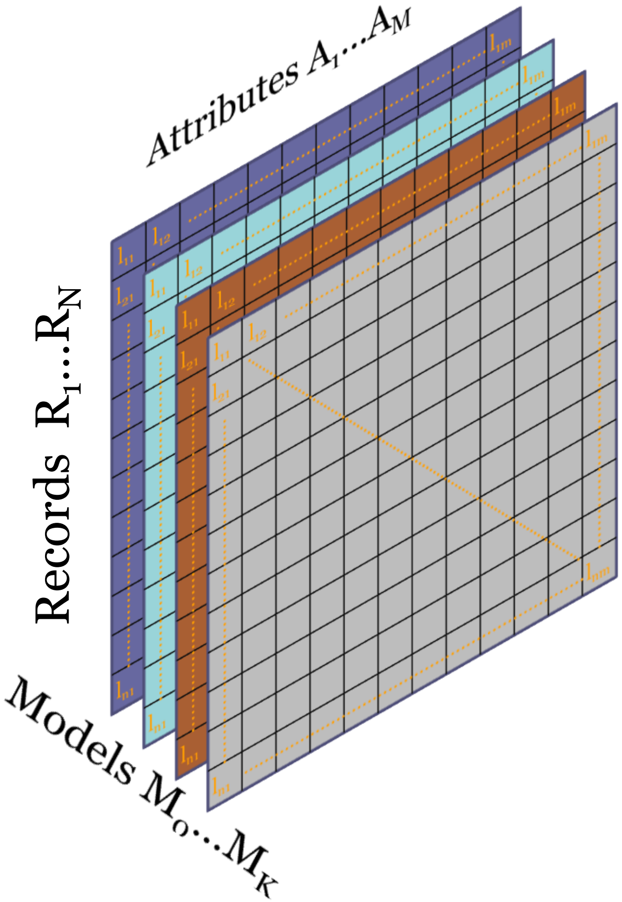
II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

- H_0: All records drawn from known models
- $H_A(S)$: Records in S drawn from unknown model

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

The DAP Algorithm





I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

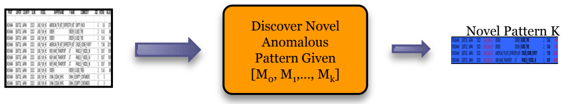
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

- H_0: All records drawn from known models
- $H_A(S)$: Records in S drawn from unknown model

More specifically we want subsets of data with significantly low p-values across all models.

The DAP Algorithm



Nonparametric Scan Statistic (NPSS)

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

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

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

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

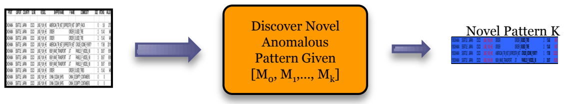
I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

- Evaluate subsets with NPSS

The DAP Algorithm



<u>Nonparametric Scan Statistic (NPSS)</u>

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

Higher Criticism:

$$F_{\alpha}(N_{\alpha}, N_{tot}) = \frac{N_{\alpha} - N_{tot}\alpha}{\sqrt{N_{tot}\alpha(1-\alpha)}}$$

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

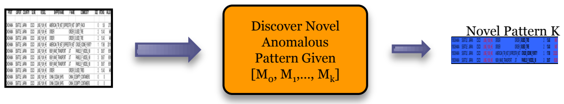
I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

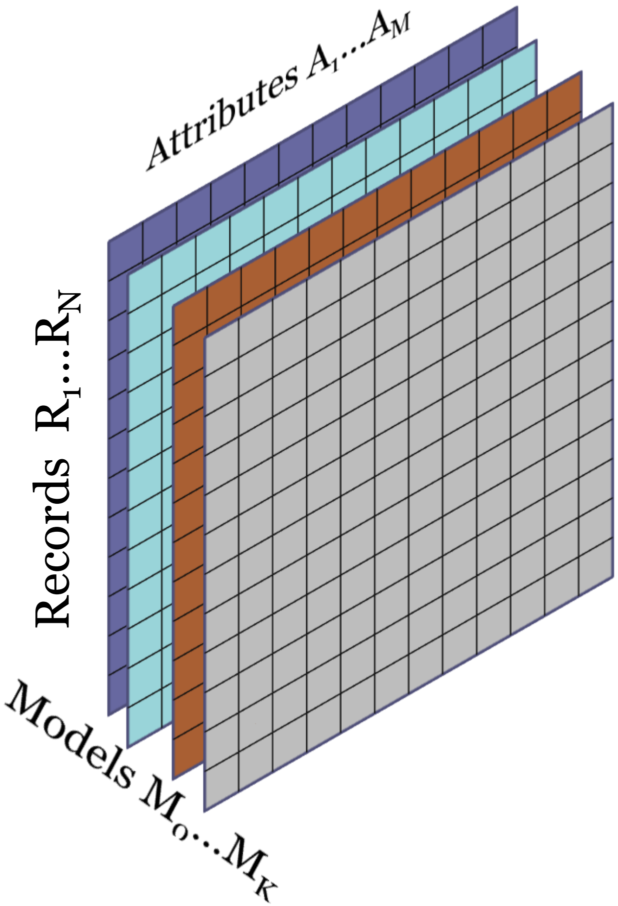
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

- Evaluate subsets with NPSS

The DAP Algorithm





I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

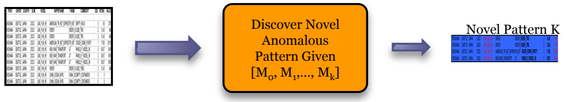
- Compute empirical p-values
  i. measures the interestingness of a v_{ij} under each M_k
  ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

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)

The DAP Algorithm



Linear Time Subset Scanning Property (LTSS)

I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values

    i. measures the interestingness of a v_{ij} under each M_k

    ii. p_{ijk} in S ~ Uniform(0,1) under H_0

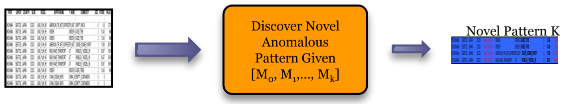
II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

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)

The DAP Algorithm



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

A F(S) and $G(R_i)$ 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 statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values

    i. measures the interestingness of a v_{ij} under each M_k

    ii. p_{ijk} in S ~ Uniform(0,1) under H_0

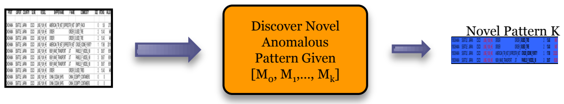
II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

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 and find the maximizing F(S)

The DAP Algorithm



Linear Time Subset Scanning Property (LTSS)

A F(S) and G(R_i) 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 statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values

i. measures the interestingness of a v_{ij} under each M_k

ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

1. Maximize F(S) over all subsets of S

•Naïve search is infeasible $O(2^{N+M})$

The DAP Algorithm

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

A F(S) and G(A_j) satisfies LTSS iff :

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

We only need to consider:

$\{A_{(1)}\}$
$\{A_{(1)}, A_{(2)}\}$
$\{A_{(1)}, A_{(2)}, A_{(3)}\}$

$\{A_{(1)}, \ldots\ldots\ldots\ldots\ldots, A_{(m)}\}$

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 statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values

   i. measures the interestingness of a v_{ij} under each M_k

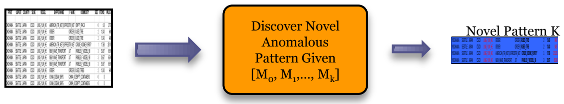
   ii. p_{ijk} in S ~ Uniform(0,1) under H_0

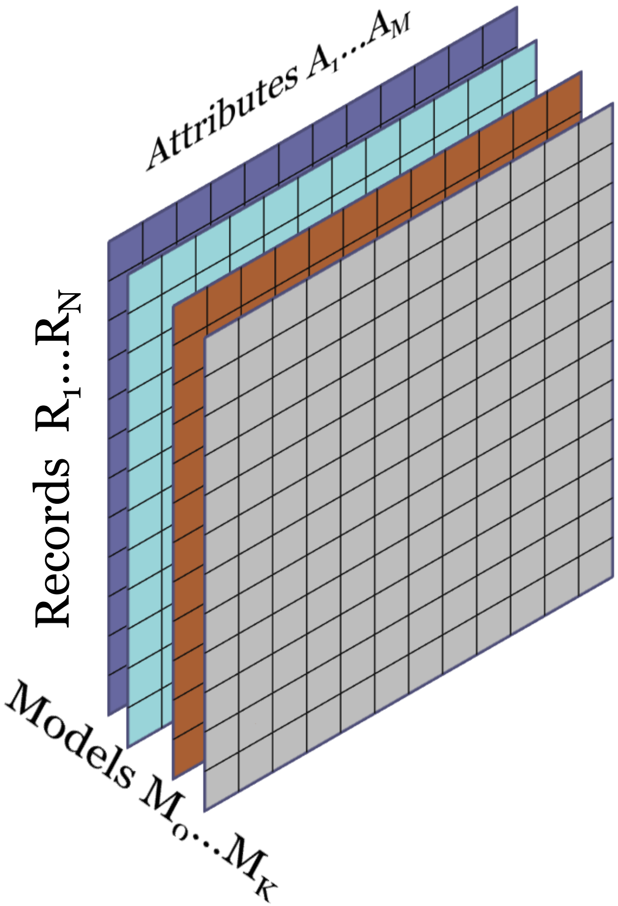
II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

   1. Maximize F(S) over all subsets of S

   •Naïve search is infeasible $O(2^{N+M})$

The DAP Algorithm





I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

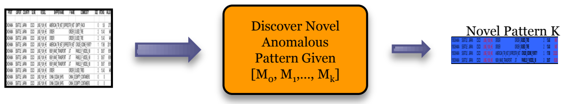
II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

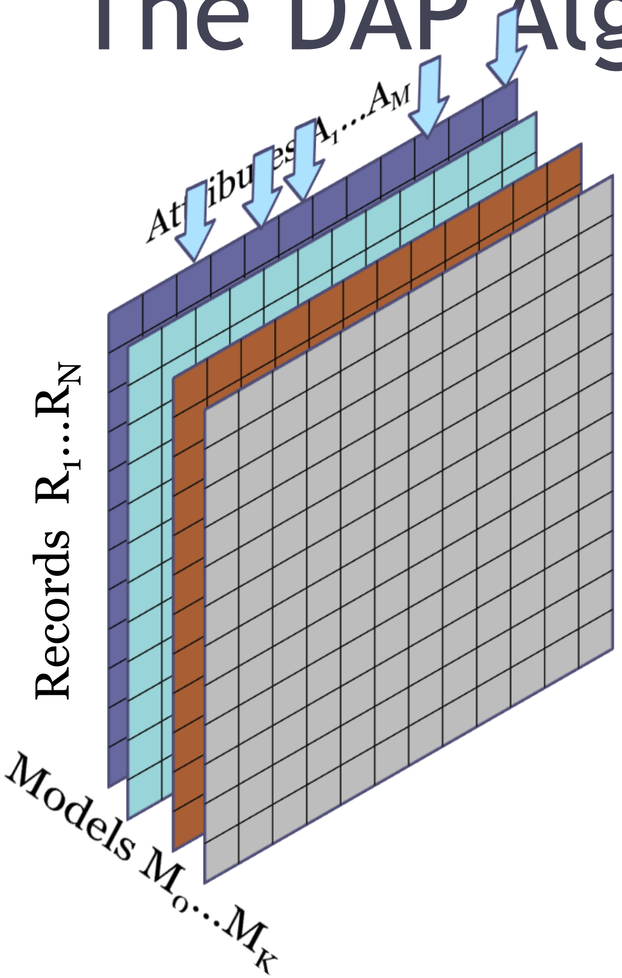
1. Maximize F(S) over all subsets of S

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

$$\max_{A} \min_{MAP} \max_{R} F(S = R\mathrm{x}A)$$

The DAP Algorithm





I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values
    - i. measures the interestingness of a v_{ij} under each M_k
    - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

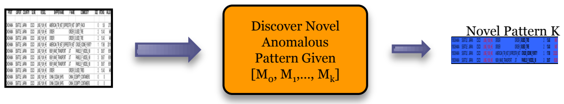
II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

1. Maximize F(S) over all subsets of S

1. Start with a randomly chosen subset of attributes

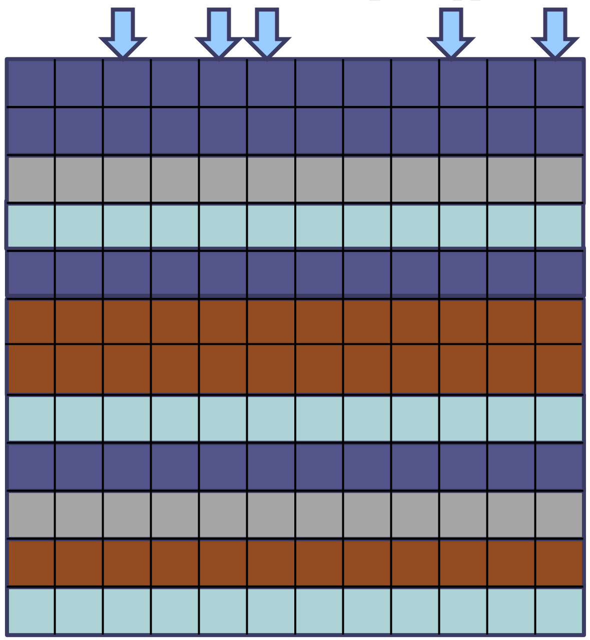
$$\max_{A}\min_{MAP}\max_{R} F(S = R\mathrm{x}A)$$

The DAP Algorithm



<u>DAP Search Procedure</u>

Attributes $A_1 \ldots A_M$



Records $R_1 \ldots R_N$

1. Start with a randomly chosen subset of attributes
2. Map each record to min M_k

I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

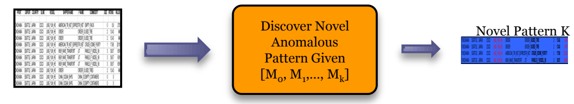
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

1. Maximize F(S) over all subsets of S

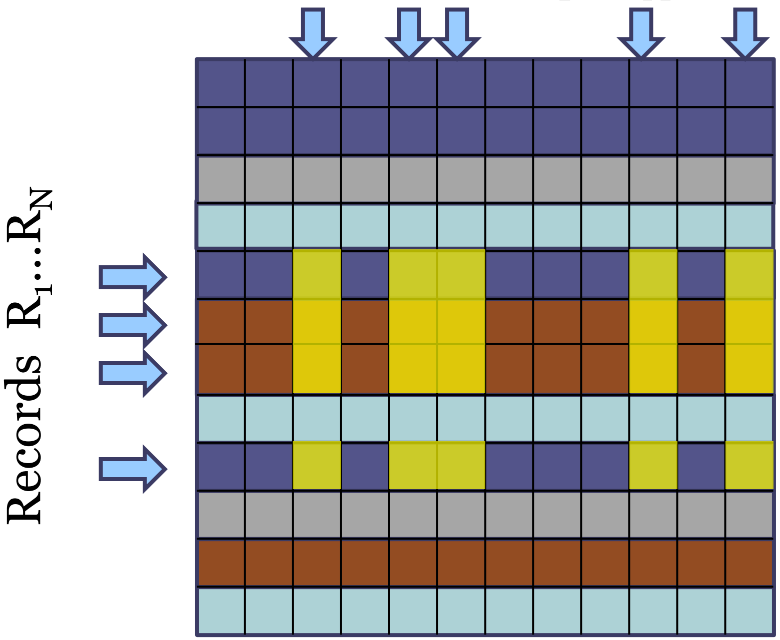
$$\max_{A} \min_{MAP} \max_{R} F(S = R\mathrm{x}A)$$

The DAP Algorithm



<u>DAP Search Procedure</u>

Attributes $A_1 \ldots A_M$



(Score = 7.5)

1. Start with a randomly chosen subset of attributes
2. Map each record to min M_k
3. Find the highest-scoring subset of recs for the given atts

I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

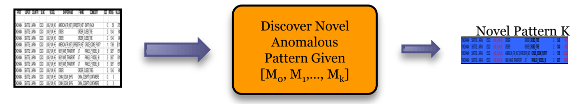
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

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

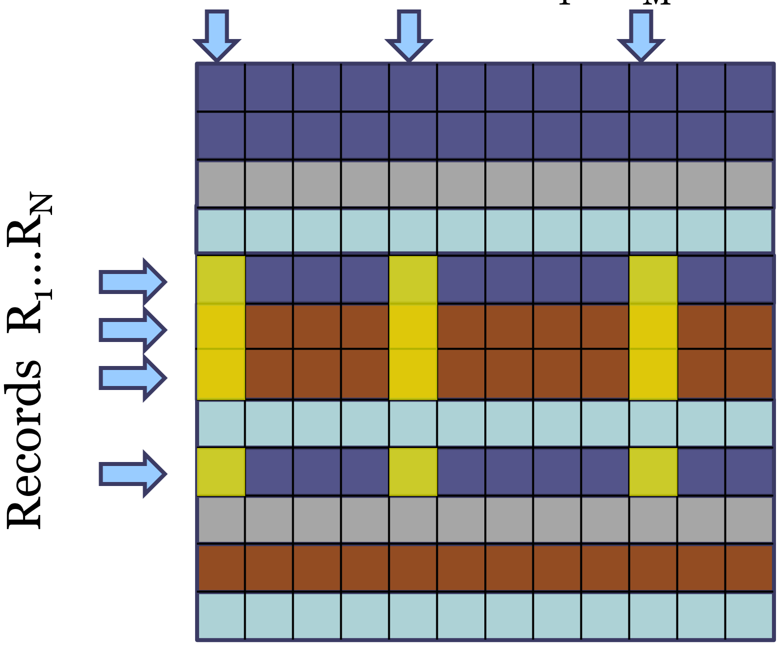
$$\max_{A} \min_{MAP} \max_{R} F(S = R \text{x} A)$$

The DAP Algorithm



DAP Search Procedure

Attributes $A_1 \ldots A_M$



(Score = 8.1)

2. Map each record to min M_k
3. Find the highest-scoring subset of recs for the given atts
4. Find the highest-scoring subset of atts for the given recs

I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

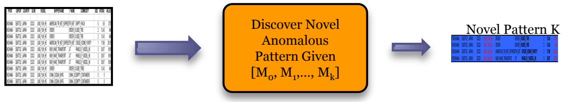
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

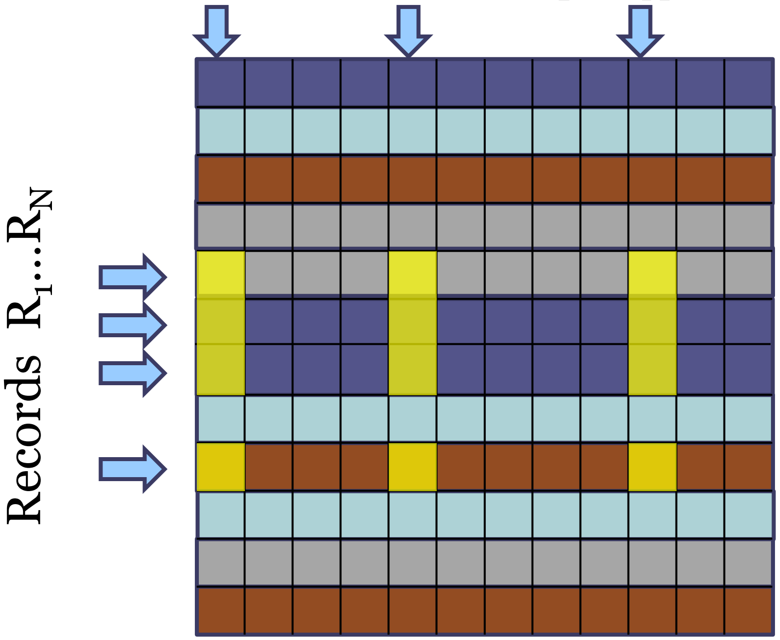
1. Maximize F(S) over all subsets of S
   - LTSS over records O(N log N)
   - LTSS over attributes O(M log M)

$$\max_{A} \min_{MAP} \max_{R} F(S = R\mathrm{x}A)$$

The DAP Algorithm



DAP Search Procedure



(Score = 8.1)

2. Map each record to min M_k
3. Find the highest-scoring subset of recs for the given atts
4. Find the highest-scoring subset of atts for the given recs

I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

- Compute empirical p-values

  i. measures the interestingness of a v_{ij} under each M_k

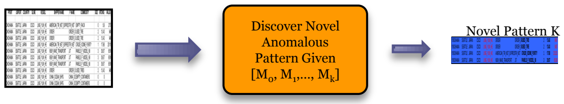
  ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

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

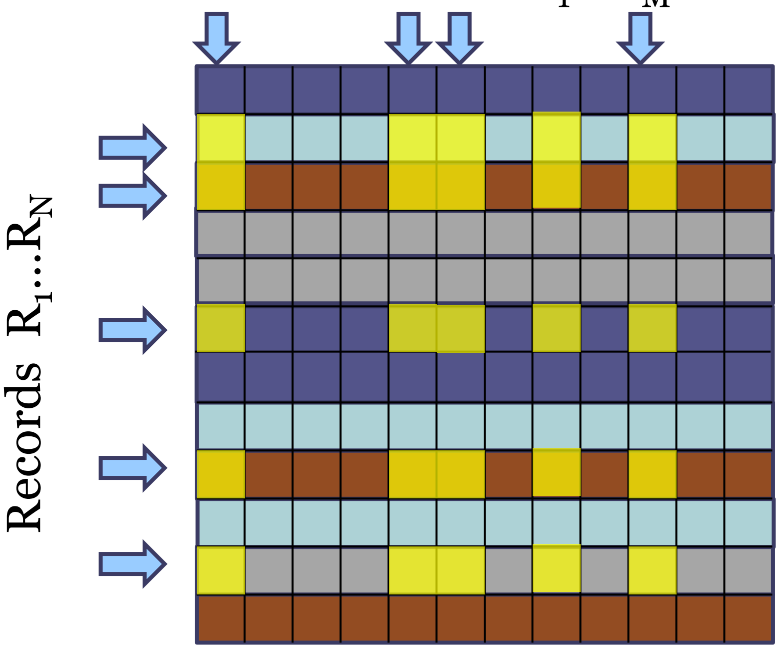
$$\max_{A} \min_{MAP} \max_{R} F(S = R\text{x}A)$$

The DAP Algorithm



DAP Search Procedure

Attributes $A_1 \ldots A_M$



(Score = 9.3)

5. Continue iterating until convergence (To Local Maximum)

I. Compute the statistical anomalousness of each attribute (for each record), under each known model.

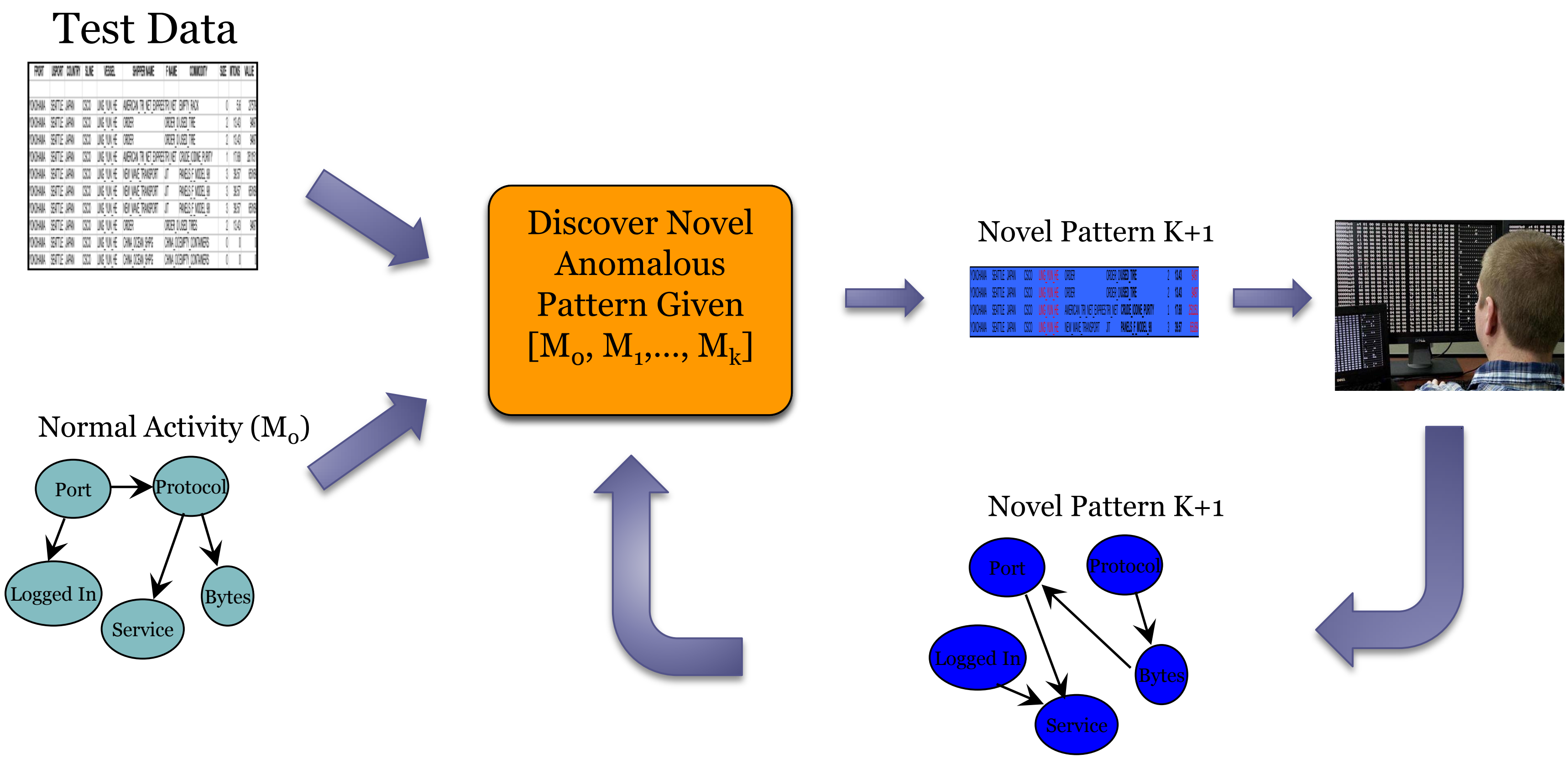
- Compute empirical p-values
  - i. measures the interestingness of a v_{ij} under each M_k
  - ii. p_{ijk} in S ~ Uniform(0,1) under H_0

II. Discover subsets of records and attributes that are anomalous under every mapping of records to models.

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

$$\max_{A} \min_{MAP} \max_{R} F(S = R \mathrm{x} A)$$

Anomalous Pattern Discovery Procedure



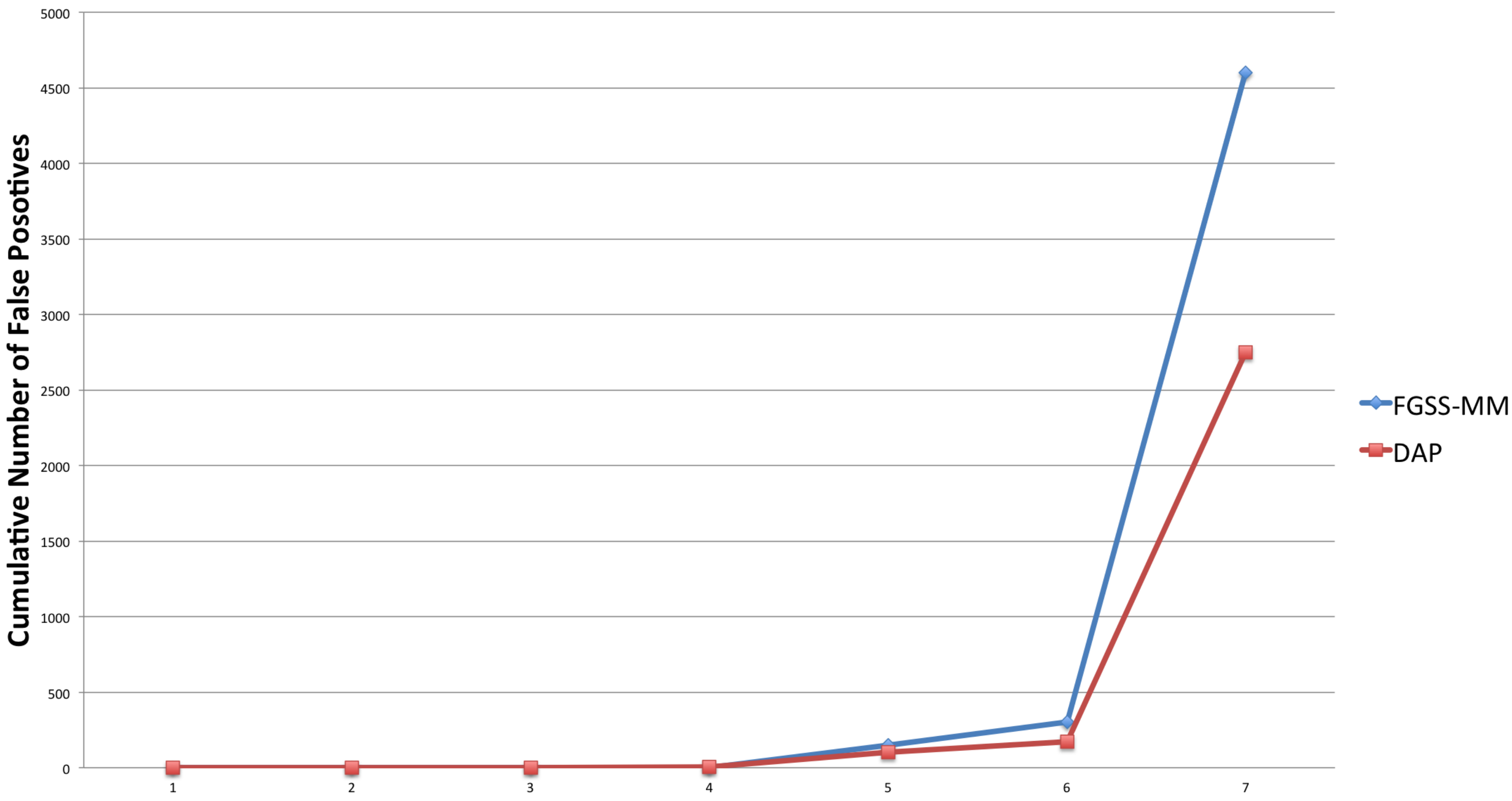
Experiments: Network Intrusion Detection

- Real-World dataset of sessions on a military network
  - Background Activity & 7 Intrusions

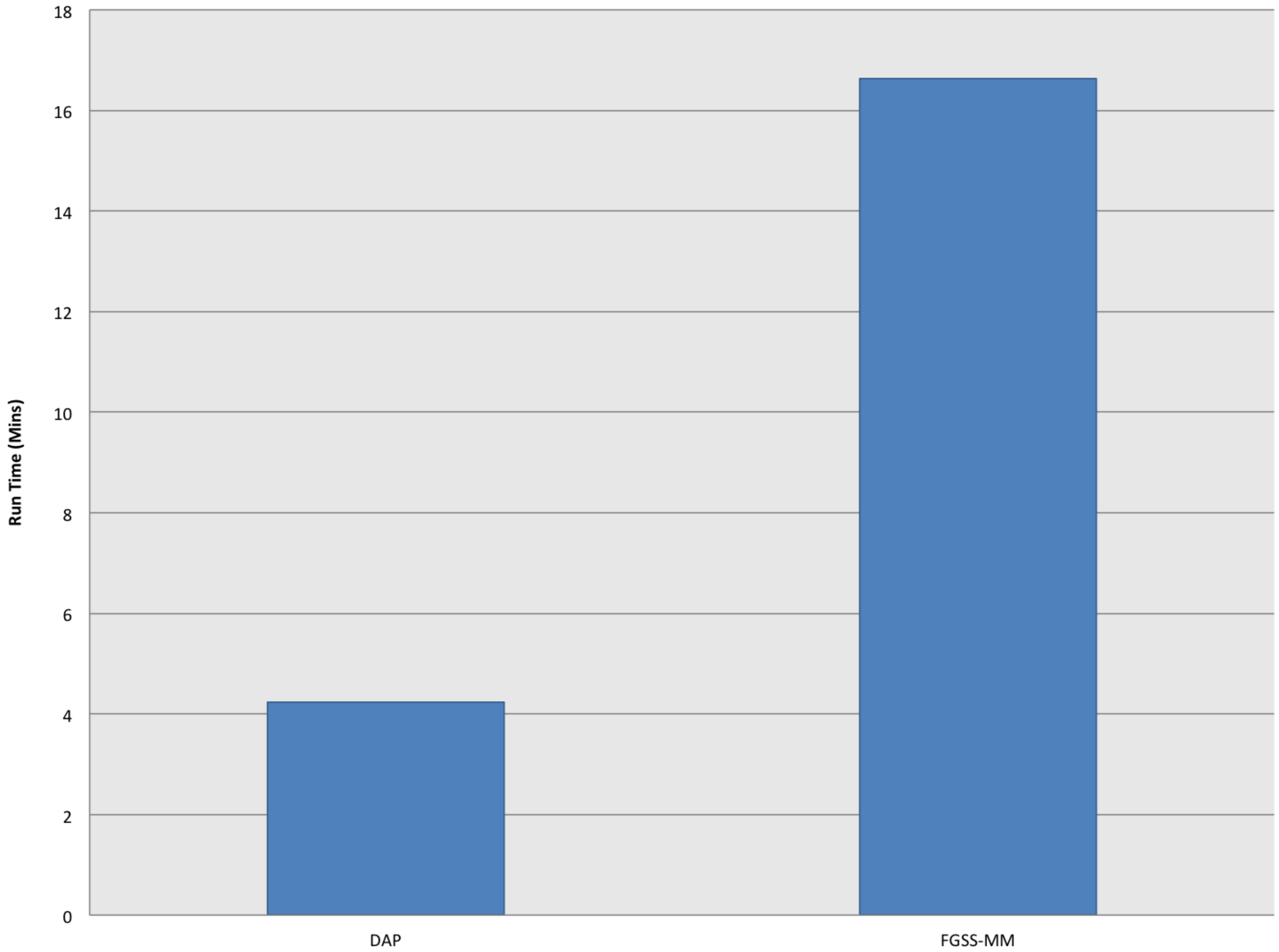
- Datasets Generated
  - Test Data: 10,000 records
    - 250 anomalies (2.5%) from each of the 7 intrusions
    - Remaining data is from the background activity
  - Training Data: up to 100,000 records

- Anomalous Pattern Discovery
  - Generate 50 Test Data Sets
    - Mix of intrusions and Background Activity
  - Generate Generate 50 Training Data Sets
    - For Background Activity
    - For each Intrusion
    - Start only with Background Training Data

Anomalous Pattern Discovery



Anomalous Pattern Discovery



Summary

- Outlined the challenging problem faced by many Computer Security firms
- Proposed the Anomalous Pattern Discovery task
  - Devised a method to solve the general discovery task
- Demonstrated the efficacy of this methodology for assisting security analyst in continual discovery of novel anomalous patterns
  - as compared to the current state of the art

Thank You