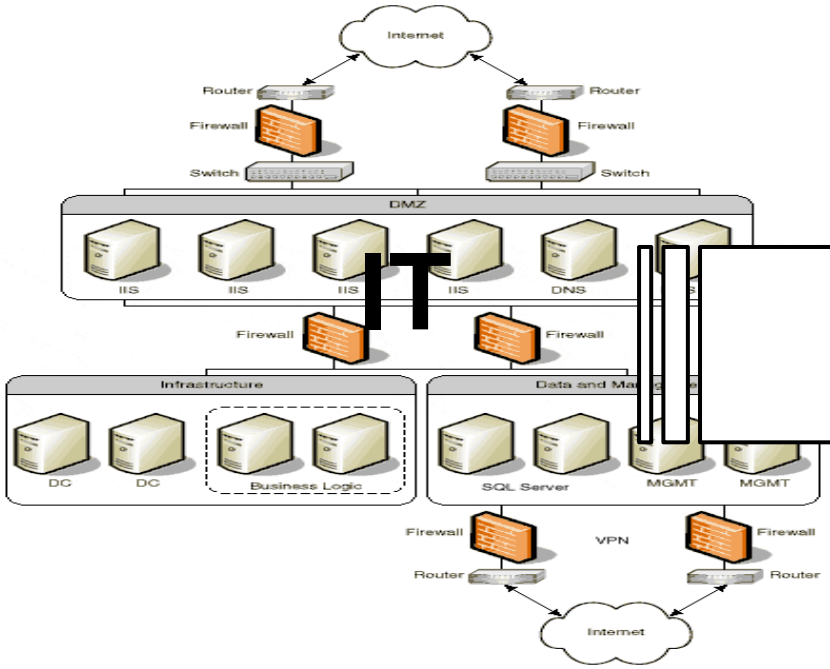
One Graph is worth a Thousand Logs

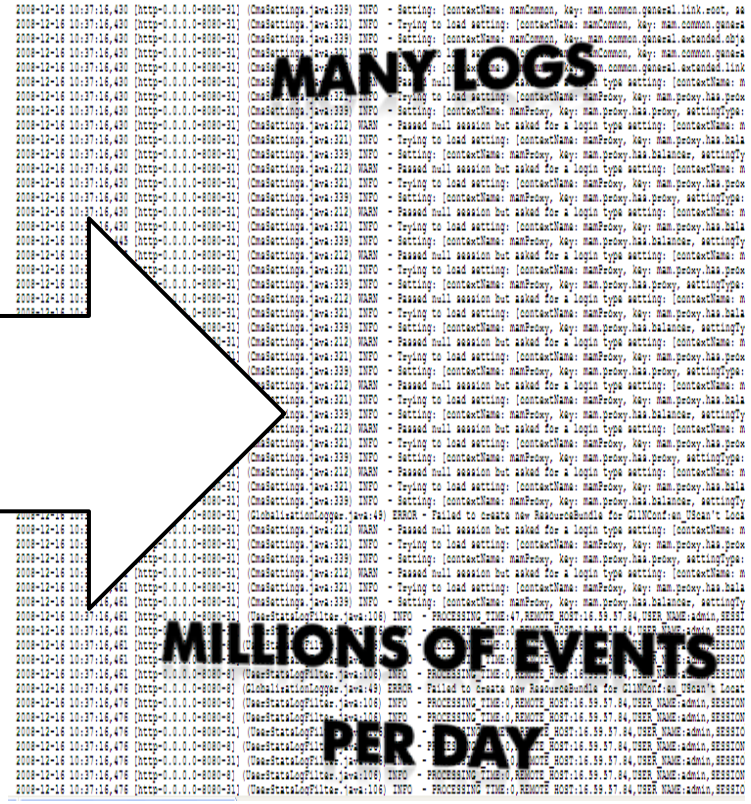
Uncovering Hidden Structures in Massive System Event Logs

Gilad Barash

Co-Authors: Ira Cohen, Michal Aharon, Eli Mordechai

Event Logs in the IT environment





- Each System component writes to its own log (events, errors)
- Logs are used to detect and troubleshoot problems in systems

Event Logs : The Problem

Huge volume of data is not amenable for human consumption

Semi-Structured text data not meant for automated consumption

Roadmap

Raw Logs

Template Discovery

Machine-Ready Data

PARIS

Human-readable Data

Logs in raw form

12/1/2008 12:34:03 failed to retrieve the meta data of project ‘null0’ the session auth has failed.
12/1/2008 12:35:03 failed to get licenses for project session the session auth has failed.
12/1/2008 12:40:31 error processing request from 192.111.22.33 data starts with 0 \00000023\0 conststr
12/1/2008 12:44:03 unexpected failure while trying to ping user session #44444 the session auth has failed
12/1/2008 12:50:03 failed to retrieve the meta data of project ‘null1’ the session authentication has failed.
12/1/2008 12:50:05 unexpected failure while trying to ping user session #33333 the session auth has failed
12/1/2008 12:50:23 failed to get licenses for project session the session auth has failed.
12/1/2008 12:55:09 failed to get licenses for project session the session auth has failed.
12/1/2008 12:56:22 error processing request from 192.222.22.55 data starts with 0 \00000014\0 conststr
12/1/2008 12:56:56 Failed to retrieve the meta data of project ‘null3’ the session auth has failed.
12/1/2008 12:57:03 error processing request from 193.111.26.33 data starts with 0 \00000512\0 conststr
12/1/2008 12:57:25 error processing request from 192.111.22.43 data starts with 0 \00000014\0 conststr

Let's Rearrange the Messages…

failed to retrieve the meta data of project **'null0'** the session auth has failed.
failed to retrieve the meta data of project **'null1'** the session auth has failed.
Failed to retrieve the meta data of project **'null3'** the session auth has failed.

Variable Word

failed to get licenses for project session the session auth has failed.
failed to get licenses for project session the session auth has failed.
failed to get licenses for project session the session auth has failed.

error processing request from **192.111.22.33** data starts with 0 **\00000023\0** conststr
error processing request from **192.222.22.55** data starts with 0 **\00000512\0** conststr
error processing request from **192.111.22.43** data starts with 0 **\00000014\0** conststr

Variable Word

Variable Word

unexpected failure while trying to ping user session **#44444** the session auth has failed
unexpected failure while trying to ping user session **#33333** the session auth has failed

Variable Word

11 messages

9 distinct messages

4 templates

Requirements for Template Discovery

1. Online
   - Produce immediate value
2. Consistent
   - Template assignment of a message should remain consistent over time
3. Efficient
   - Keep up with incoming message rates

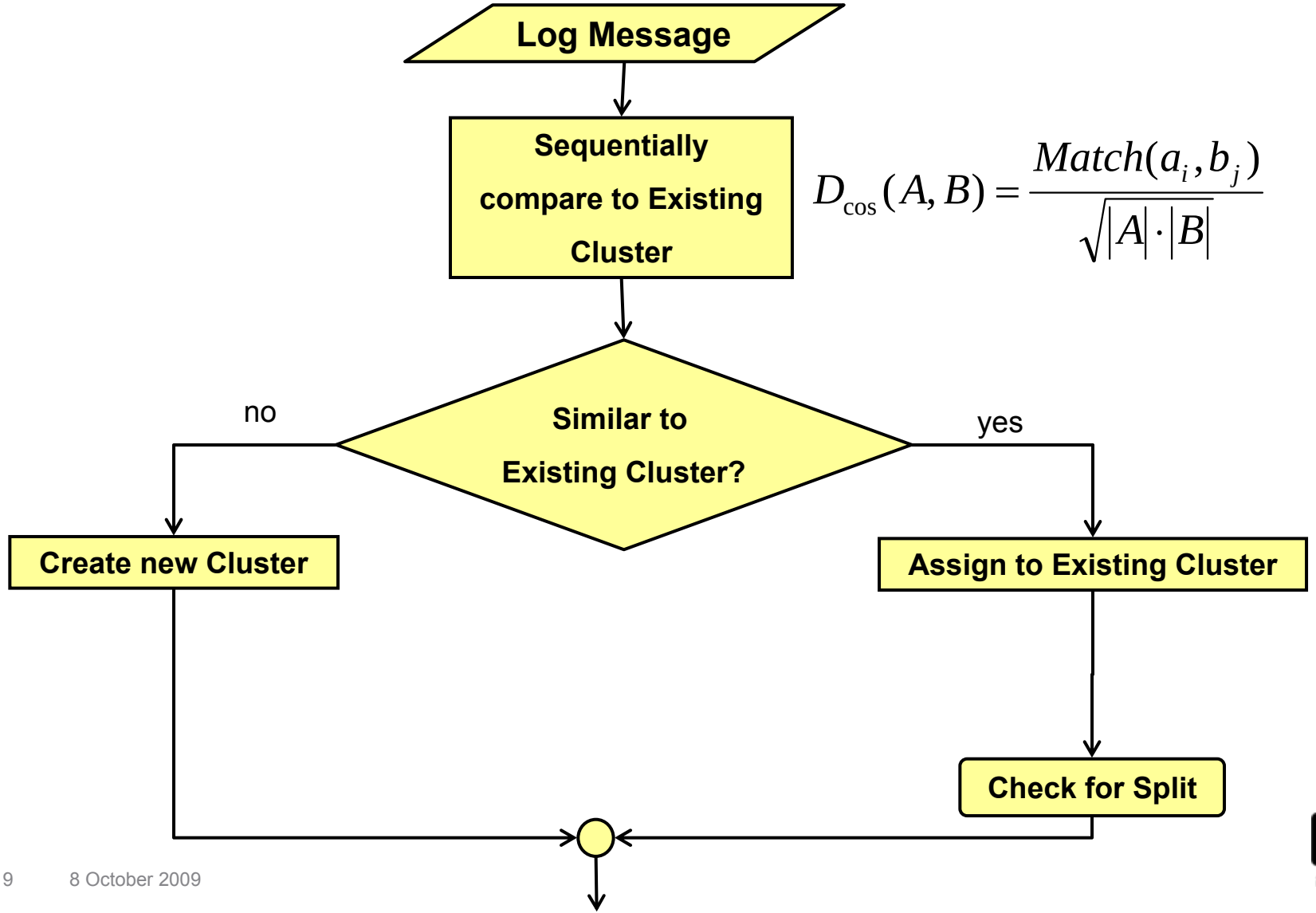
Template Discovery Algorithm:

Incremental Text Clustering

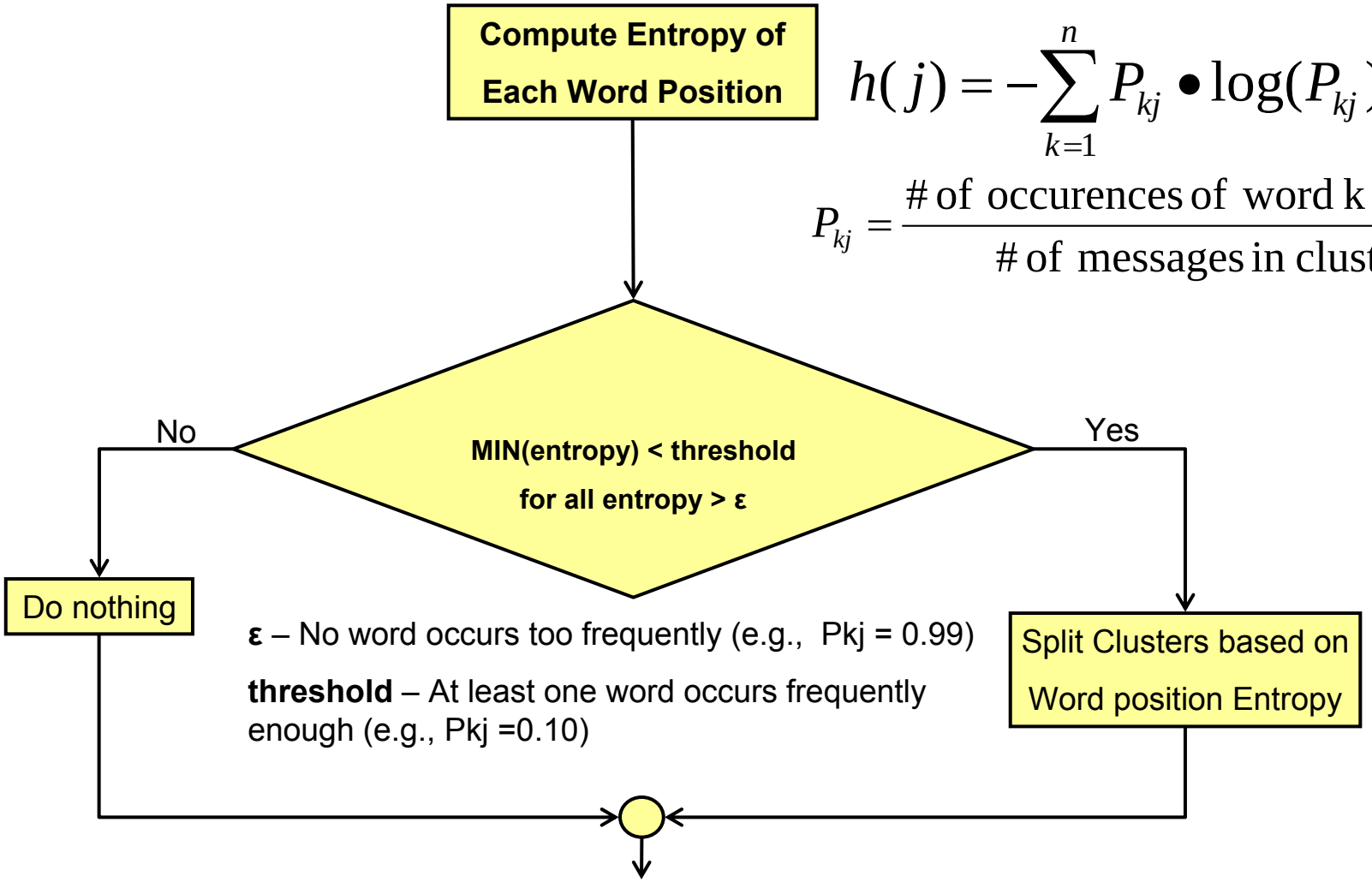
- Step 1: “Rough” clustering:
  - Creating/Assigning events to root clusters
- Step 2: Cluster refinement:
  - Splitting root clusters

Output: Forest of clusters

Step 1: “Rough” Clustering



Step 2: Cluster Refinement



Template discovery algorithm

$$D_{\cos}(A,B) = \frac{Match(a_i, b_j)}{\sqrt{|A| \cdot |B|}}$$

Similarity Threshold: 0.8

=0.83

Clustering example:

m1: B C D F A B

m2: B C D F A B J

m3: A C D F E K

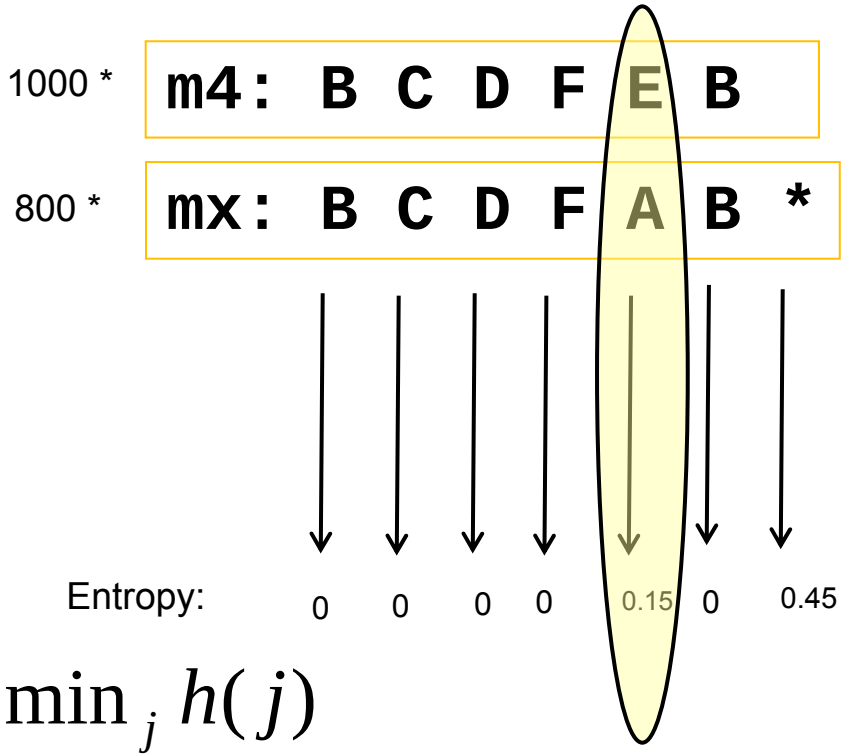
m4: B C D F E B

1000 appearances of m4

800 appearances of BCDFAB_

m3

Entropy Calculation for Split



$$h(j) = -\sum_{k=1}^{n} P_{kj} \bullet \log(P_{kj})$$

$$\arg\min_j h(j)$$

where $\varepsilon < h(j) < threshold$

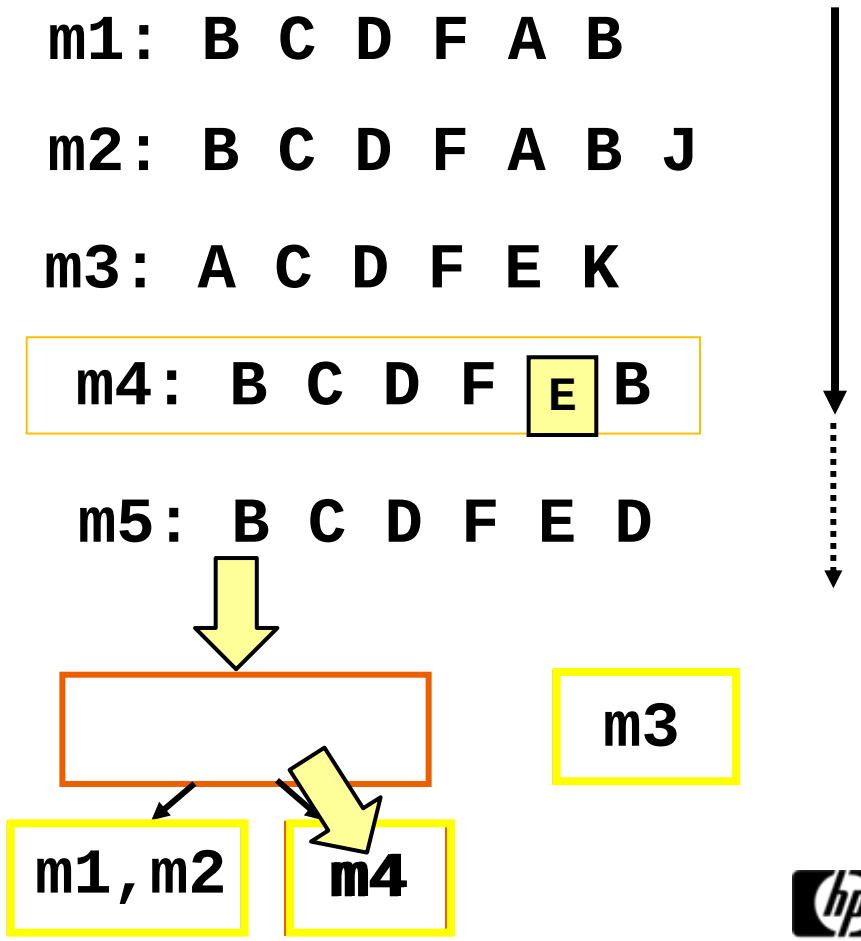
Template discovery algorithm

$$D_{\cos}(A,B) = \frac{Match(a_i, b_j)}{\sqrt{|A| \cdot |B|}}$$

Similarity Threshold: 0.8

=0.83

Clustering example:



Discovering System Process Patterns

Database Connection Startup

(1) JDBC3 getGeneratedKeys(): disabled
(3) Connection release mode: auto
(5) getConnectionURLs=tcp://websiteURL:2507
(6) Query translator: hql.ast.ASTQueryTranslatorFactory
(9) create connection. connectId
(10) mercury_db_loader_DB_Loader user=;pwd=;

Service Manager startup

(2) SH remote was null. Exported object monitor.
(4) Add task Main Flow
(7) Register provider class dataentry.loader.LoaderMain
(8) Service manager started
(3) Connection release mode: auto
(11) mercury_db_loader is up and running

Optional messages

Same message In different process

(1) JDBC3 getGeneratedKeys(): disabled
(2) SH remote was null. Exported object monitor.
(3) Connection release mode: auto
(4) Add task Main Flow
(5) getConnectionURLs=tcp://websiteURL:2507
(6) Query translator: hql.ast.ASTQueryTranslatorFactory
(7) Register provider class dataentry.loader.LoaderMain
(8) Service manager started
(9) create connection. connectId
(10) mercury_db_loader_DB_Loader user=;pwd=;
(3) Connection release mode: auto
(11) mercury_db_loader is up and running

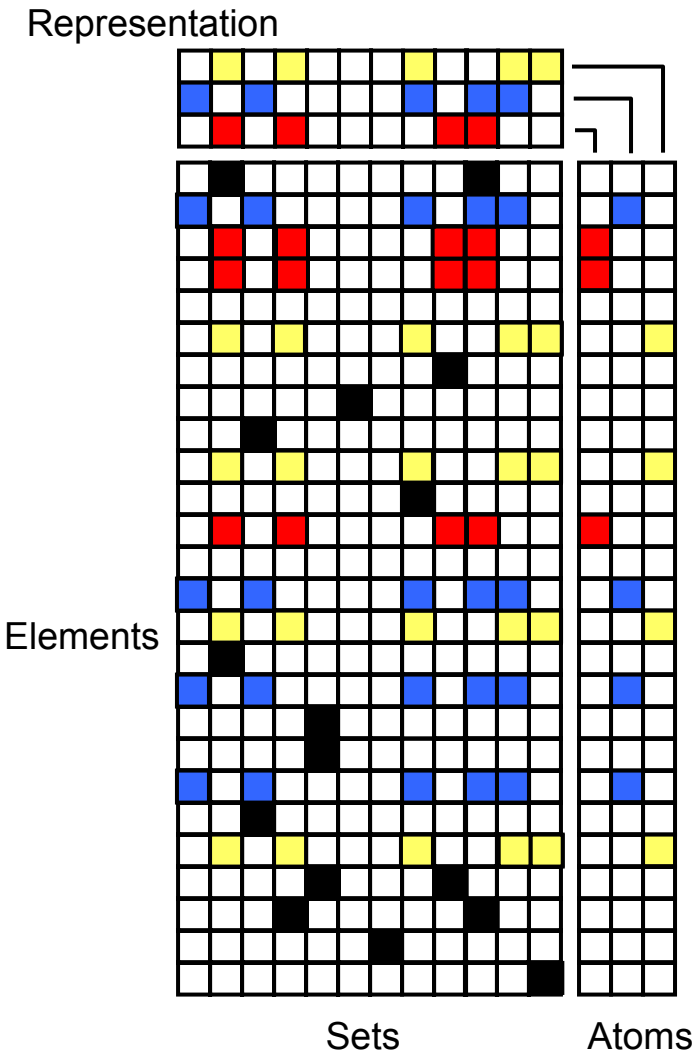
PARIS (Principal Atoms Recognition In Sets)

Identifies sets of events that tend to occur together with no innate knowledge of messages

Provides enhanced view of system behavior dynamics

hp invent

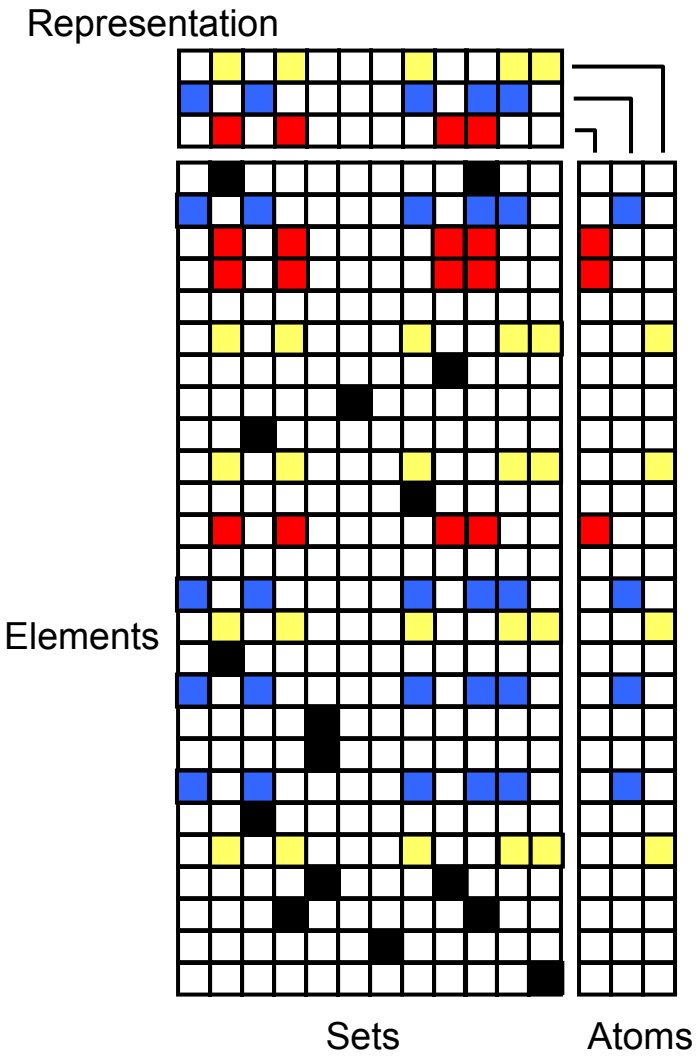
PARIS



- Gets as input a large number of sets, that are assumed to have some mutual characterization.
- Detects principal sets of elements that tend to appear together in the data.
- Overcomes non-exact repetitions
- Ignores additional noise
- Uses:
  - Analysis
  - Compression
  - Anomaly Detection

hp invent

PARIS



- Representation error must be small, but not necessarily zero.
- Representation should serve some sense of compression of the data (sparsity).
- Minimal number of atoms (K).

PARIS Cost Function

Minimize the representation error of the data.

$$PCF = \arg\min_{A,R}\left(\sum_{i=1}^{N} d_r\left(D_i, R(A, R_i)\right)\right) + \left(\sum_{i=1}^{N} \mu_i |R_i|\right) + \left(\tau |A|\right)$$

Minimize the size of the representation (compression).

Minimize the number of principal atoms.

- Representation error must be small, but not necessarily zero.
- Representation should serve some sense of compression of the data (sparsity).
- Minimal number of atoms (K).

Results

- Datasets

Source	Number of events	Number of distinct events
Business App 1	**4,210,513**	**153,619**
Printer Press	**11,204**	**5,631**
Windows Events	**66,102**	**25,340**
Business App 2	**483,768**	**70,102**

Results

- Template identification

Source	Number of events	Number of distinct events	Number of clusters (templates)
Business App 1	**4,210,513**	**153,619**	**4,193**
Printer Press	**11,204**	**5,631**	**204**
Windows Events	**66,102**	**25,340**	**476**
Business App 2	**483,768**	**70,102**	**1,115**

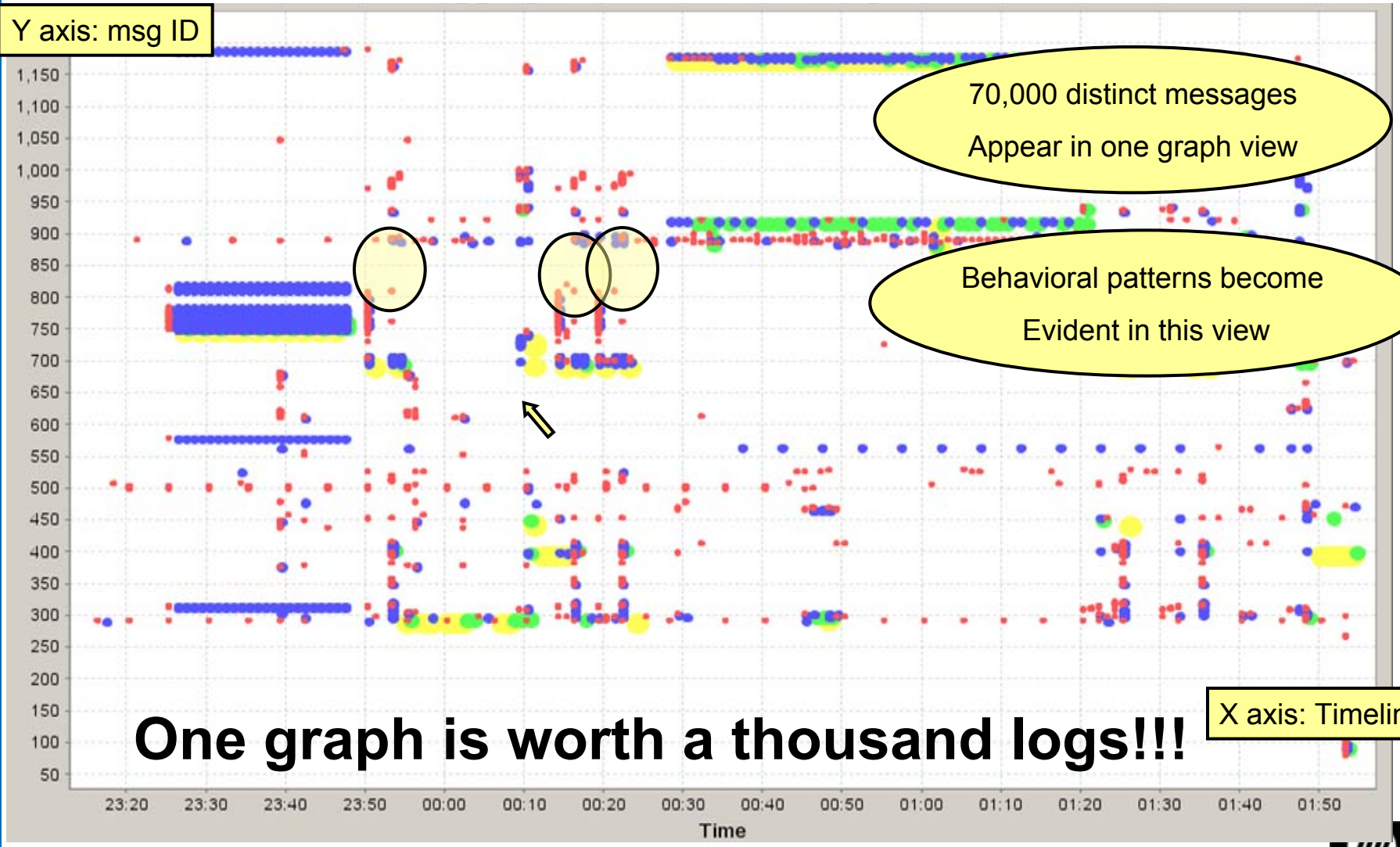
Representation Accuracy: 95%

Results

- Compression

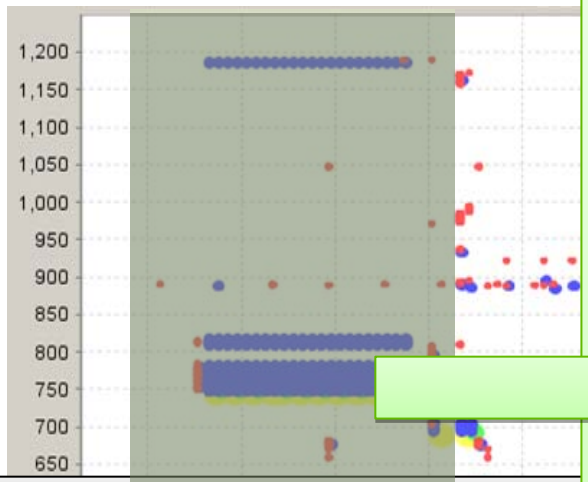
Source	Number of events	Number of distinct events	Number of clusters (templates)	Index size reduction (number of words saved)
Business App 1	**4,210,513**	**153,619**	**4,193**	**90%**
Printer Press	**11,204**	**5,631**	**204**	**50%**
Windows Events	**66,102**	**25,340**	**476**	**40%**
Business App 2	**483,768**	**70,102**	**1,115**	**90%**

Visualizing the logs: Business App 2 Event Timeline



PARIS Result: Correct Process Identification

Buss App 2 Logs



Atom ID: 12

- **890** Failed creating SiS sample
- **924** Failed processing http reques
  report_ss_samples, from remote
  Failed to acquire lock for publ
- **1183** Failed processing http reque
  report_transaction, from remotel
  Failed to acquire lock for publ

Atom ID: 27

- **734** User operation - stop nanny
- **748** message_broker STOPPED
- **753** Input main(String[] args:
- **754** Going to call WrapperManager.start(new Main(), args)
- **755** Initializing Spring files
- **757** Path for spring files is E:\HPBAC\conf\supervisor\spring
- **759** Loading spring file
- **764** NannyConfig
- Will load JMX
- No security f
- **767** Registering b
- **768** Autodetecting user-defined JMX MBeans
- **769** Bean with name 'nannyManager' has been autodetected for JMX exposure
- **770** HTTP adapter port is 11021
- **771** Succeeded adding html adapter
- **772** manager thread loop started.
- **773** Verifying time diff between cpp (local machine) and Java.
- **774** Log file of time diff is: E:\HPBAC\tools\TimeDiff\time_diff.log
- **776** Run java time diff
- **780** Trying to initialize Properties Manager**792** Config server check passed
- **793** Prerequisites have been met
- **794** start() Nanny Manager
- **795** Nanny Manager need to start all services?:true
- **796** Going to start all services.
- …

Service Restart

Summary

- Summarize Event Logs: Template Creation
  - Lossless Reduction in size of data
  - Machine-readable
- Process identification: PARIS
  - Strategic importance in managing IT environment
  - Human-readable

Q&A