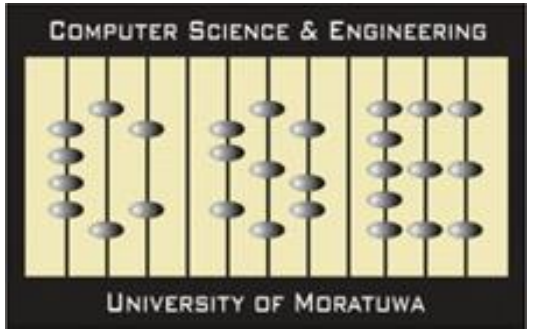


Shapelets and Parallel Coordinates Based Automated Query Generation for Complex Event Processing

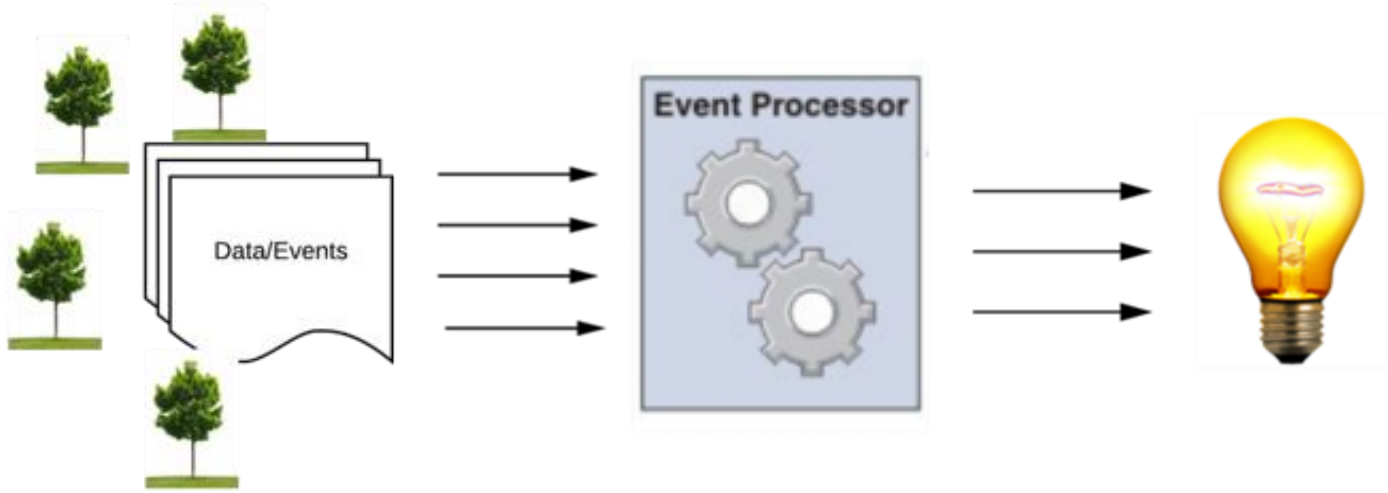
R.N. Navagamuwa, K.J.P.G. Perera, M.R.M.J. Sally, L.A.V.N. Prashan, and H.M.N. Dilum Bandara

(randika.12, pravinda.12, jaward.12, prashan.12, dilumb)@cse.mrt.ac.lk

Department of Computer Science and Engineering
University of Moratuwa
Katubedda, Sri Lanka

Motivation

Complex Event Processing



- Given a dataset consisting of temperature & smoke readings, how to generate a fire alert?

$$\textbf{SELECT} * \textbf{WHERE}\ TEMPERATURE \geq 40\ \textbf{AND}\ SMOKE > 800\ \textbf{WITHIN}\ 0 \leq TIME \leq 5$$

CEP Query Processing

$$\textbf{SELECT} * \textbf{WHERE}\, TEMPERATURE \geq 40\, \textbf{AND}\, SMOKE > 800\, \textbf{WITHIN}\, 0 \leq TIME \leq 5$$

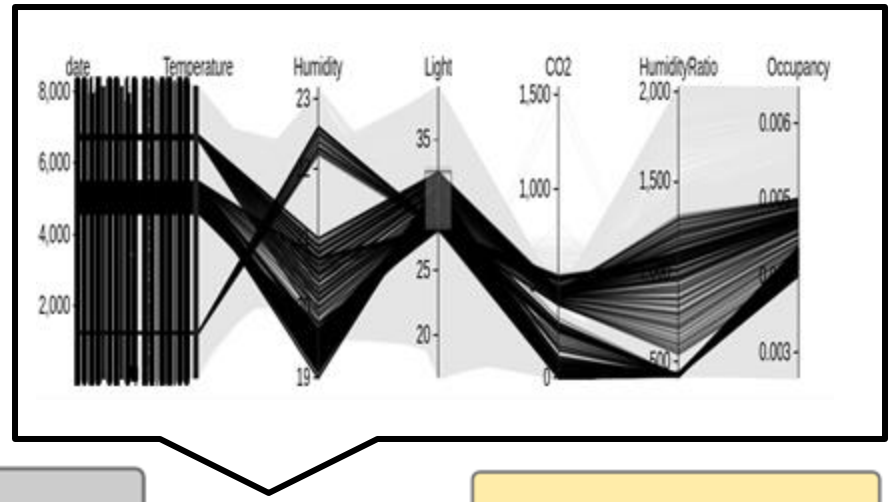
- Which values should attributes carry?
- Do they need to appear in a specific, temporal order?
- How query can be optimized?
- When we a prepare query, above should be answered

But Domain Experts have to be there to answer ...

We Automate This Process

Proposed Solution

Architecture



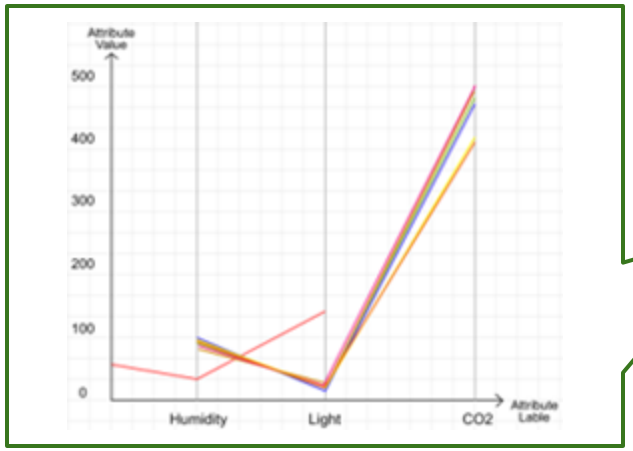
$$\textbf{SELECT} \quad \{*\}$$
$$\textbf{WHERE} \ \{attr_1 \geq a \ and \ attr_2 < b\}$$
$$\textbf{WITHIN} \ \{t_1 \leq time \leq t_2\}$$

Data Set

Data Processor

Shapelet Generator

Query Generator

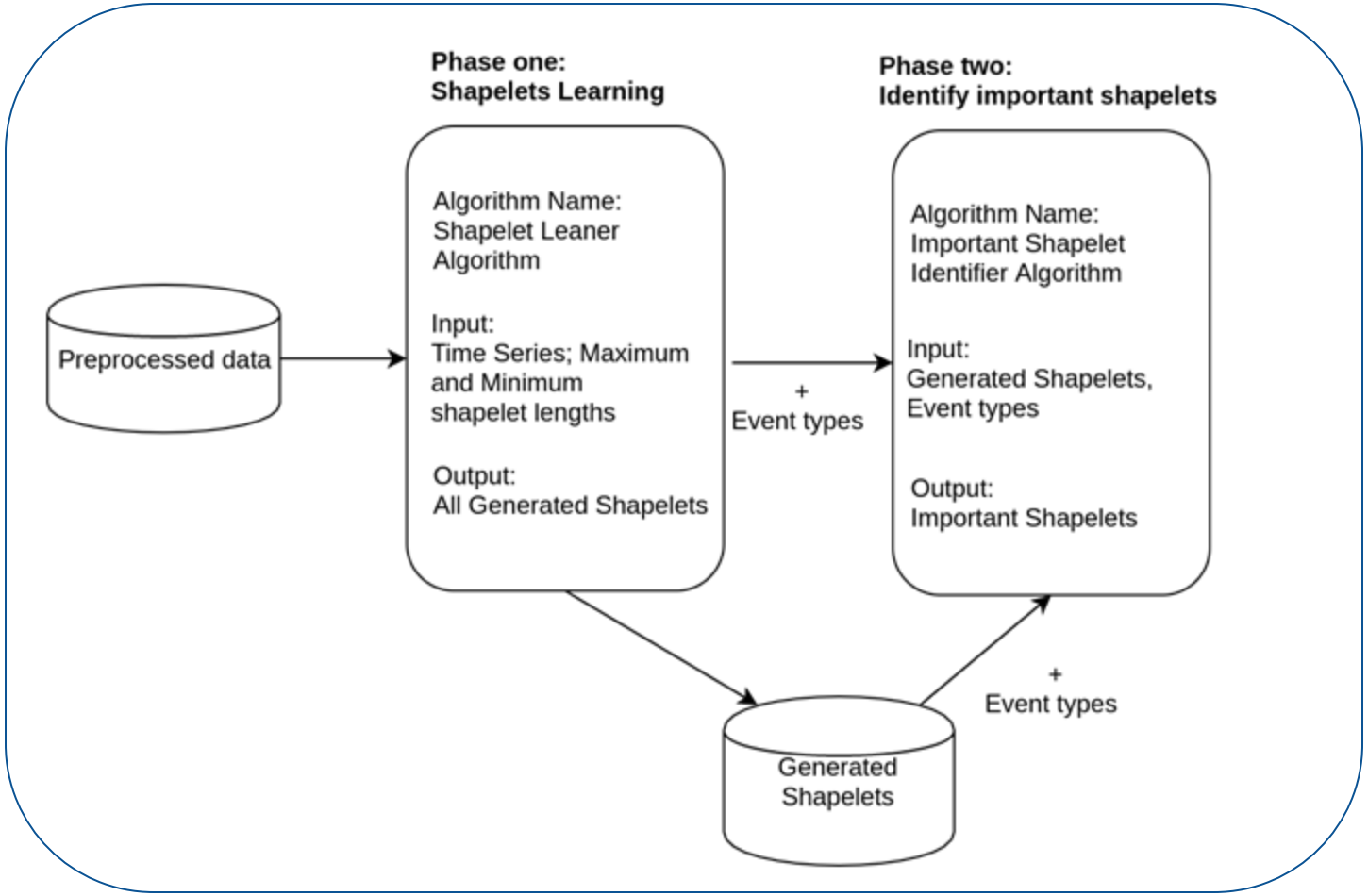


Visual Representation

User Approval/ Hints

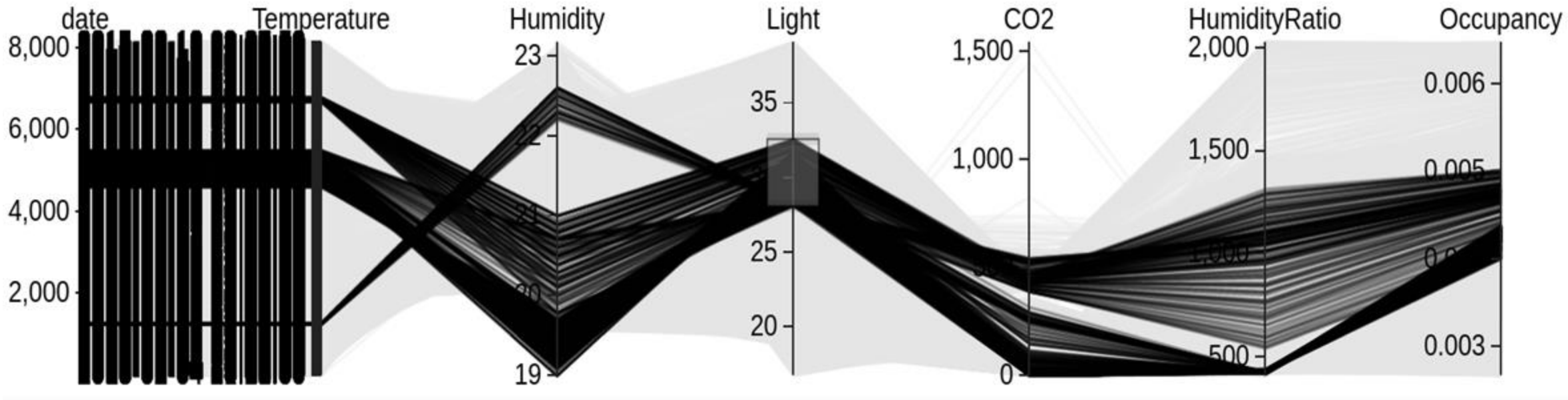
Research Methodology

Expanded Shapelet Generator

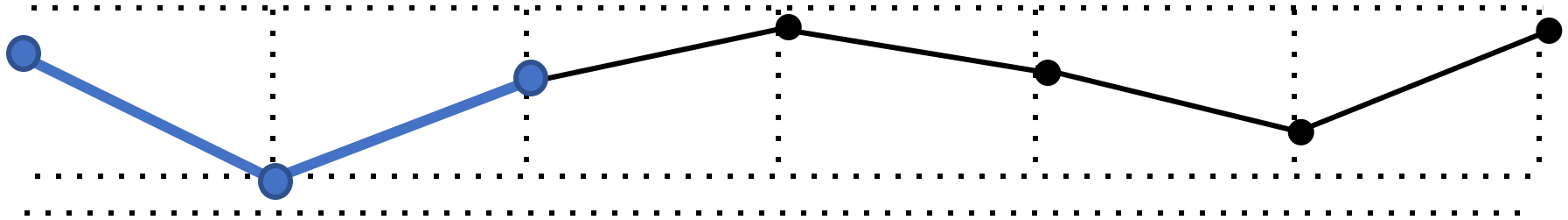


Shapelet Generation

- Parallel coordinates are used to generate all possible shapelets

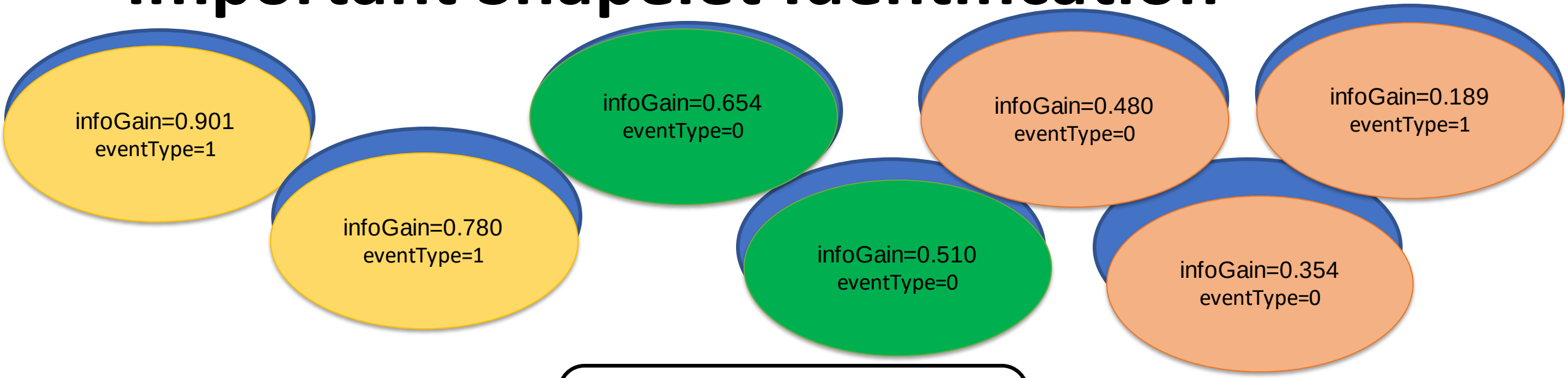


Shapelet Learner



```
"Shapelet" : {
                    "seriesID" = 4,
                    "startPos" = 1,
                    "content"  =[0.8,0.1, 0.7],
                    "informationGain" = 0.6546,
                    "eventType" = 2
                }
```

Important Shapelet Identification



Event distribution – Total Dataset
eventType 1 = 60%
eventType 0 = 40%

Probability of event 1 = 100%

Probability of event 0 = 0%

Probability of event 1 = 0%

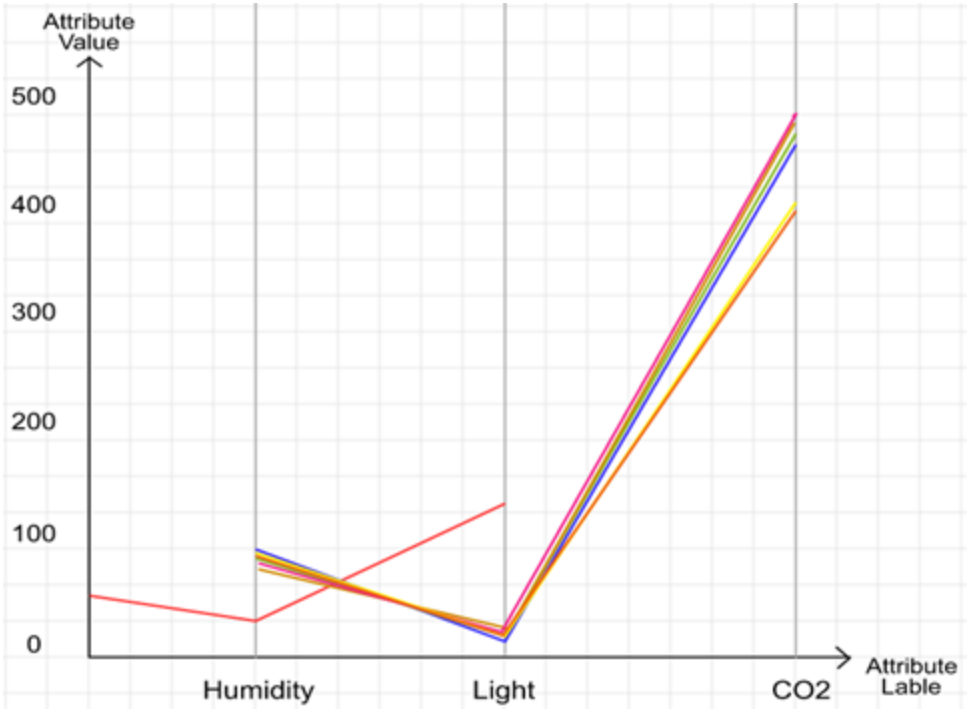
Probability of event 0 = 100%

Probability of event 1 = 33%

Probability of event 0 = 66%

Visual Representation & Query Generation

- Generated important shapelets will be shown to the user
- User can simply approve them or select another set of shapelets as the important ones



Dataset Description

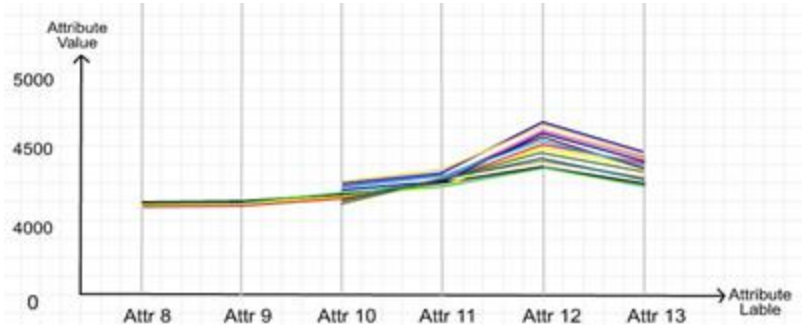
Name: Occupancy Detection Data Set
Number of instances: 20560
Number of attributes : 7
Data Set Characteristics: Multivariate, Time-Series
Source: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

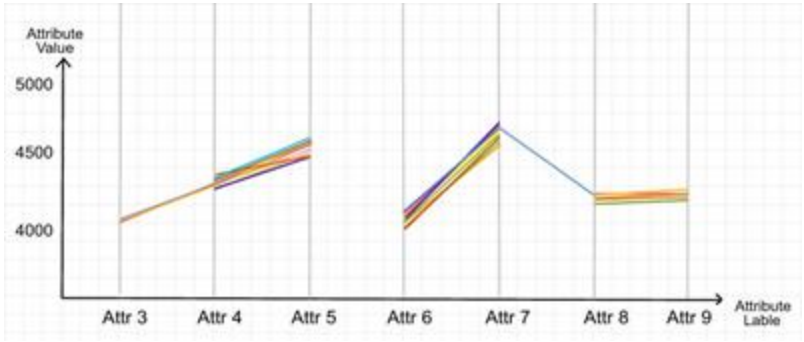
Results

Dataset 1

Dataset Description

Name: Occupancy Detection Data Set
No of instances: 20,560
No of attributes: 7
Data Set Characteristics: Multivariate, Time-Series
https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+





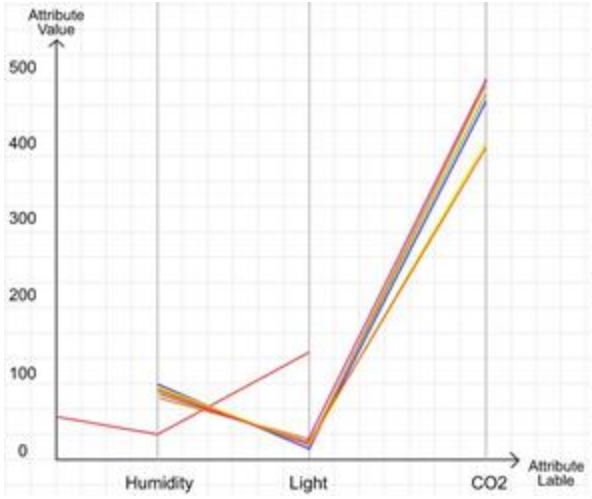
Event	Metric	Value
Not Occupied	No of events in dataset	291
	No of events detected using CEP query	286
	Recall	98.3%
	Precision	100.00%
	False positives	0
	False negatives	5 (1.7%)
Occupied	No of events in dataset	196
	No of events detected using CEP query	196
	Recall	100.00%
	Precision	84.5%
	False positives	36 (18.4%)
	False negatives	0

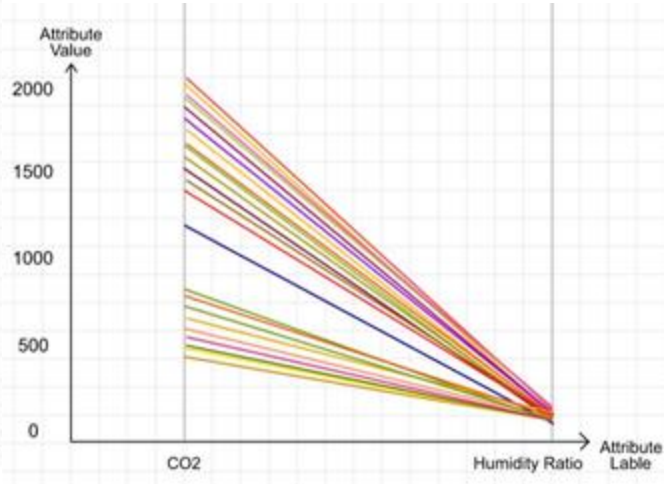
Dataset 2

Dataset Description

Name: EEG Eye State Data Set
No of instances: 14,980
No of attributes: 15
Data Set Characteristics: Multivariate, Sequential, Time-Series

http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State





Event	Metric	Value
Eye open	No of events in dataset	652
	No of events detected using CEP query	635
	Recall	97.39%
	Precision	100.00%
	False positives	0
	False negatives	17 (2.67%)
Eye closed	No of events in dataset	69
	No of events detected using CEP query	68
	Recall	98.55%
	Precision	100.00%
	False positives	0
	False negatives	1 (1.45%)

Computational Complexities

Algorithm	Time Complexity
Algorithm 1 - Shapelet Learner Algorithm	$O(nm^3)$
Algorithm 2 - Shapelet Merger Algorithm	$O(nm^2)$
Algorithm 3 - Important Shapelet Filter Algorithm	$O(n^{3/2}m^3)$

Summary

- A method to automate the query generation for CEP that combines
  - Parallel coordinates
  - Shapelets
  - Information gain
- Proposed methodology has high precision and recall
- It also has low computational and memory complexity
- In future we plan to extend proposed methodology to work with unannotated datasets

Q & A

Thank you!

Contact: pravinda.12@cse.mrt.ac.lk