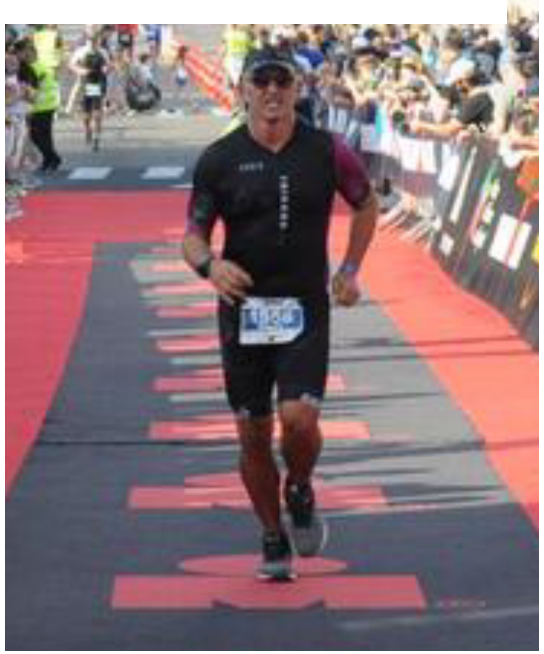
Sequential Pattern-based Anomaly Detection for Enhanced Interpretability

ASU

Omer Idgar, Prof. Robert Moskovitch
Software and Information Systems Engineering, BGU, IL
Ichan Medical School, Mount Sinai, NYC, USA

Robert Moskovitch, PhD

- **Post Doctoral**, Biomedical Informatics
  Columbia University, NYC
- **Head**, Complex Data Analytics Lab
  Software and Information Systems Engineering
  Ben Gurion University, Beer Sheva, Israel
- **Adjunct Faculty**, Ichan Medical School at Mount Sinai, NYC
- **Vice President**, International Artificial Intelligence in Medicine (AIME) Society
- Academic Professional Roles:
  - Academic Editor, PLOS ONE
  - Editorial Board member, Journal of Biomedical Informatics
  - S/PC member: ACM KDD, IJCAI, AAAI, UAI, AIME and more
  - Co-chair, AIME 2020
- Triathlete in the mornings ..



CDAL

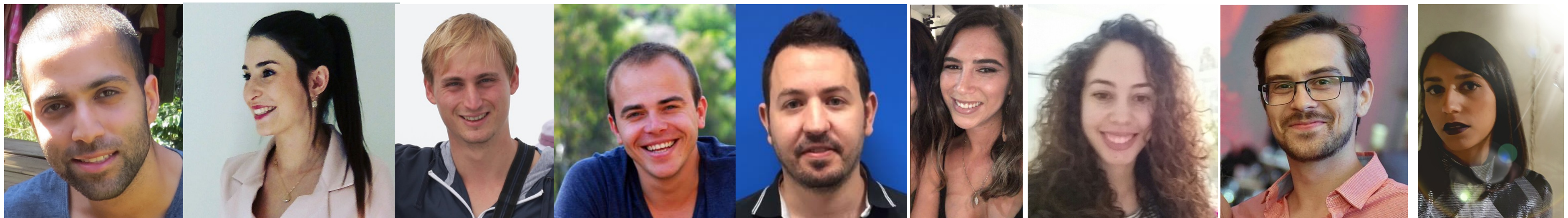
אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev
جامعة بن غوريون في النقب

- Funding: IBM, Amdocs, PCS, MoST, PMO Cyber Center, MAFAAT and more.

- Collaborations: Columbia University, Maccabi Healthcare Services, AIIMS/IIT New Delhi, Peking University, UTHealth, UPenn/CHOP and more

- Professional Activities (Organizer/SPC/PC): PLOS ONE – Editor, ACM KDD - PC, IJCAI – SPC, Editorial Board of Journal of Biomedical Informatics. AIME 2020 – Co-Chair.

- Students: Roni Mateless, Nofar Sarafian, Guy Danieli, Stav Sapir, Tal Ivshin, Maya Schvatz, Pavel Novitzki, Omer Harel, Amos Zamir, Noa Lemberger, Nevo Itzhak, Ofir Dvir, Guy Shitrit.



Funders and Collaborators

- Funding
- Collaborators

IBM

amdocs

PERSON CENTRED SOFTWARE
MOBILE CARE MONITORING

U.S.-Israel Energy Center

Prime Minister's Office
National Cyber Bureau

משרד המדע, הטכנולוגיה והחלל
Ministry of Science, Technology & Space

CENTER FOR VETERAN TRANSITION AND INTEGRATION
COLUMBIA UNIVERSITY

Maccabi
The Best Healthcare in Israel

MK&M

AIIMS

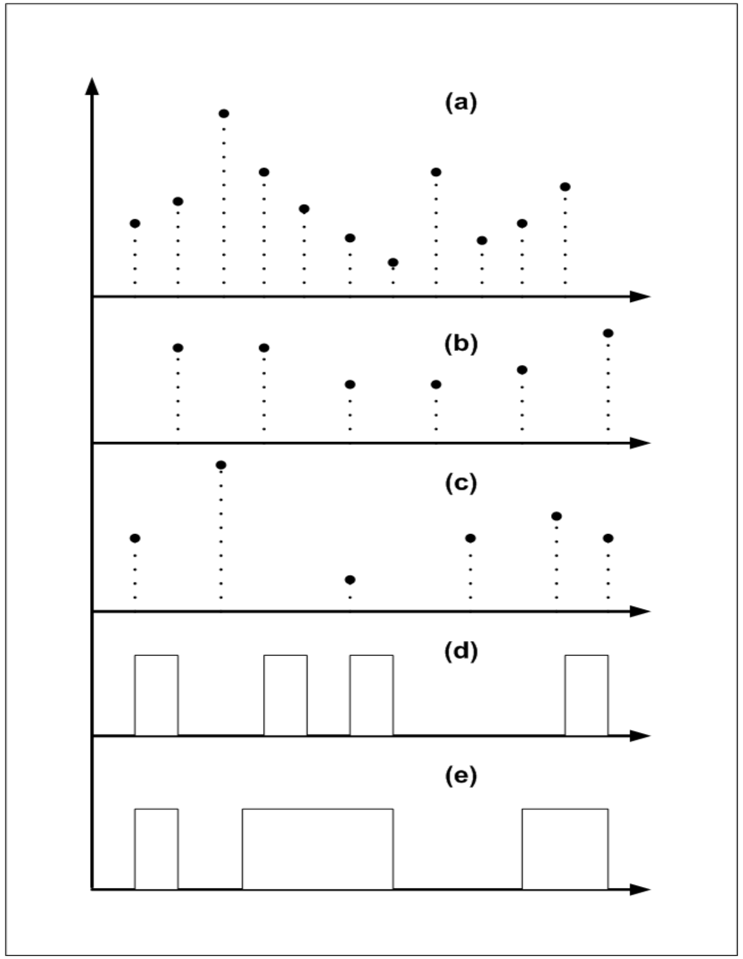
UCL

HOSPITAL
Assuta
RAISING HEALTH STANDARDS

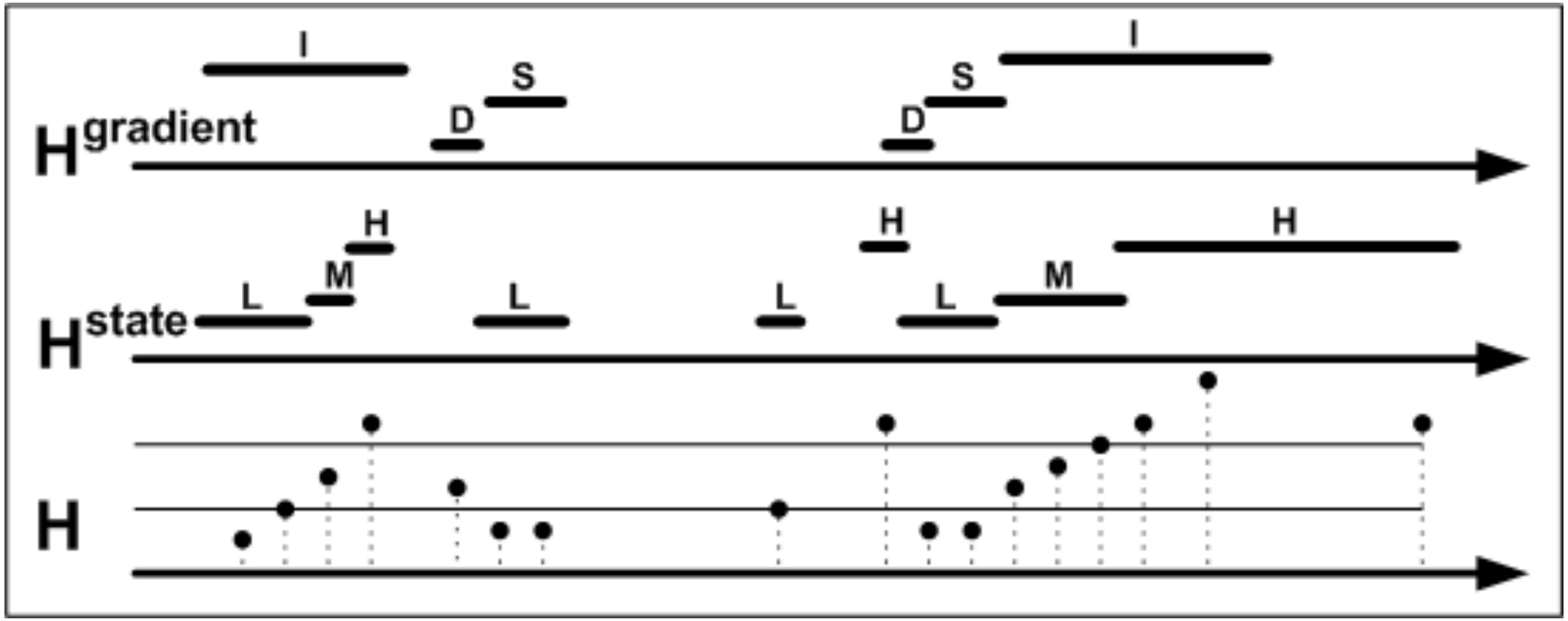
Mount Sinai

UTHealth
The University of Texas
Health Science Center at Houston

Temporal Raw Data



From Time Points to Time Intervals Series



Robert Moskovitch, Yuval Shahar, Classification Driven Temporal Discretization of Multivariate Time Series, *Data Mining and Knowledge Discovery*, 29, 4, 871-913, 2015.

Motivation

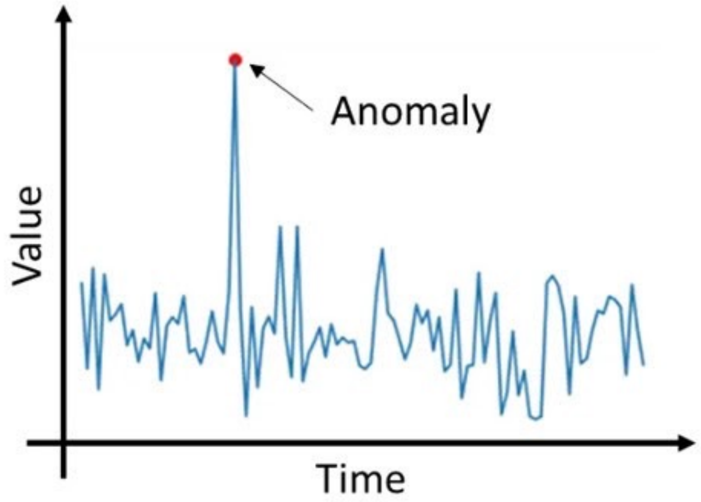
- **Industrial Control Systems** (ICS) are essential networks of hardware and software that manage industrial processes, such as water purification plants.

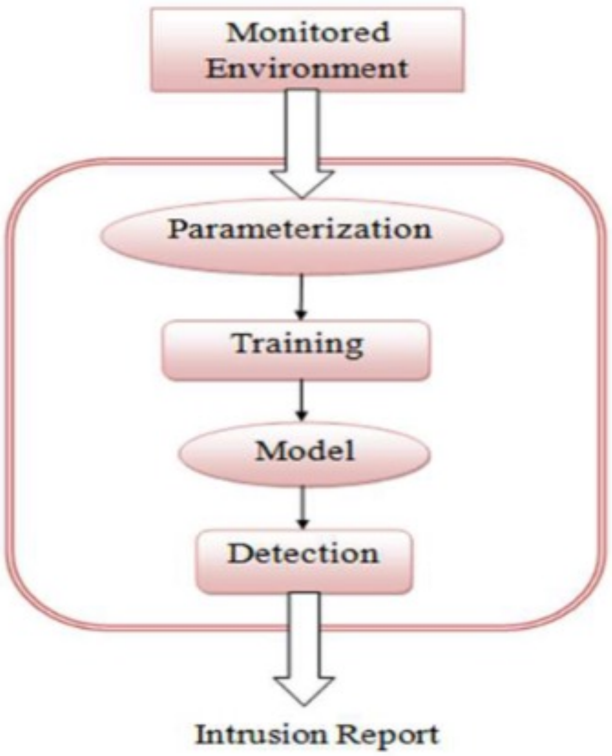
- Increased **internet connectivity** exposes ICS to cyber threats, highlighting the need for robust security measures and advanced anomaly detection techniques.

Anomaly Detection

Anomaly detection is a field of study that focuses on identifying unusual or abnormal behavior in data.

- **Statistical methods**
- **Machine learning algorithms**
- **Data mining techniques**
- **Rule-based methods**

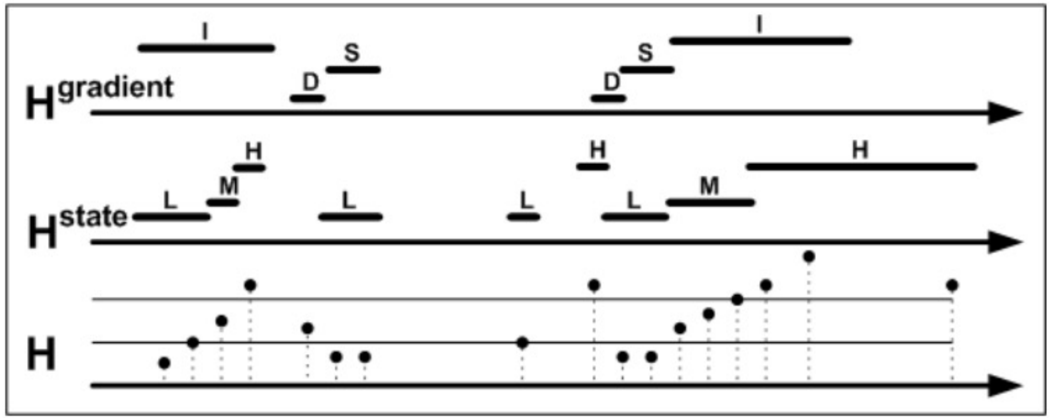


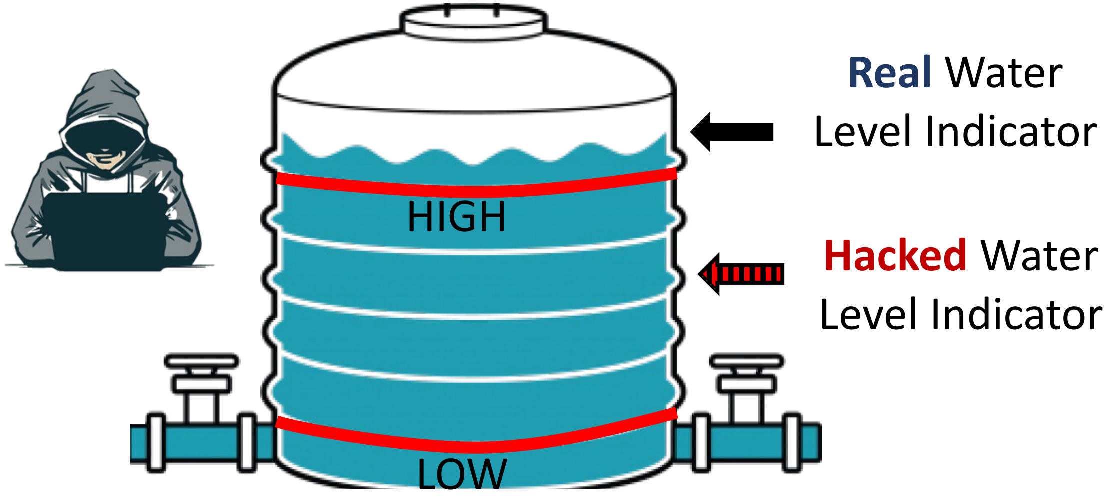


Research Goals

Objective 1: Develop an anomaly detection model using **sequential patterns**.

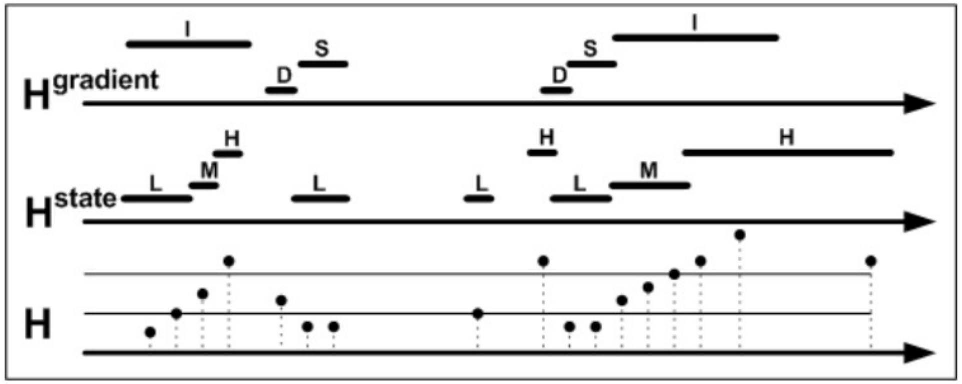
Objective 2: Develop an **interpretable** algorithm for cyber security experts in ICS.

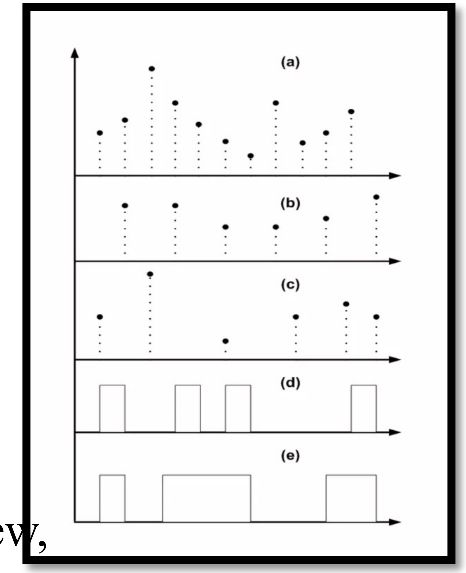


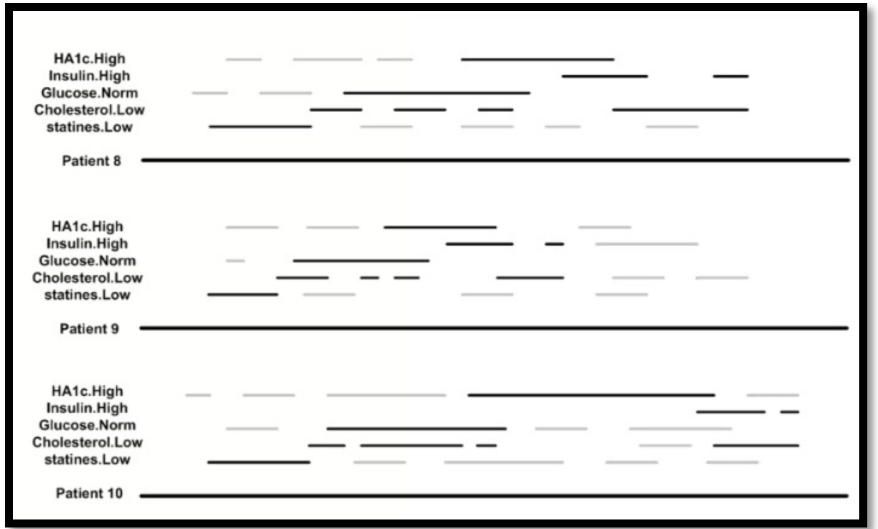


Pros of Temporal Abstraction

- **Handle missing values** – robust to gaps in data.
- **Handle irregular sampling** - unifies inconsistent data collection intervals.
- **Generalization** – detects patterns across varying time intervals.
- **Interpretability** - offer clearer interpretation than continuous data.







Robert Moskovitch, Temporal Data Analysis – a Review, *Data Mining and Knowledge Discovery*, 2022.

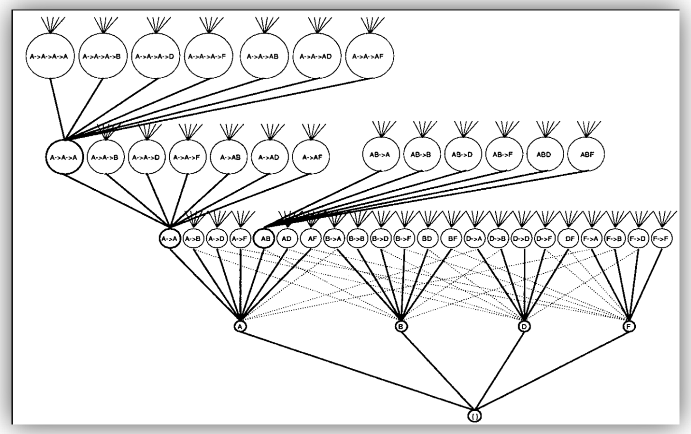
Sequential Pattern Mining

- Database consists of ordered events, in which each event is 1 time-unit long.

- Apriori property: If a sequence is infrequent, then all its super-sequences must also be infrequent.

Example: < e a > is infrequent, then
< e a b > must be infrequent as well.

SID	Sequences
10	< b a d c >
20	< (bd) a (de) c b >
30	< a d c b >
40	< a d (be) c >
50	< c d b c a b >



Sequential Pattern Mining - Example

SID	Sequences
10	< b a c >
20	< (bd) a (de) c b >
30	< a d c (bc) >
40	< a d (bc) c >
50	< c (bc) c a b >

minimum support = 4

Frequent Patterns	Support
< a >	5
< a b >	4
< a c >	4
< b >	5
< (bc) >	4
< b c >	4
< c >	5
< c c >	4

Sequential Pattern Mining

Real example of pattern:

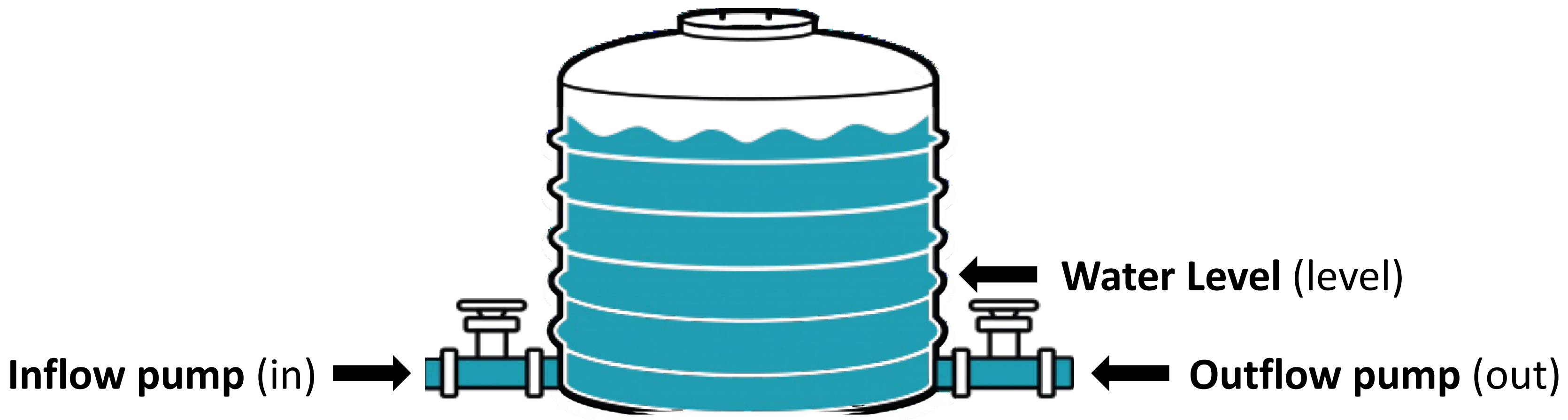
"Water tank fills, then empties"

3 sensors: **inflow pump** - in, **water level** - level, **outflow pump** – out

Abstractions Levels: **High – H**, **Low - L**

< (in_H, level_L, out_L) (in_H, level_H, out_L) (in_L, level_H, out_H) >

"Tank Filling Up" "Tank Full" "Tank Draining Down"



Sequential Pattern - definitions

Horizontal Support (HS)

HS of a pattern P in sequence S – The number of times pattern P is contained in S

SID	Sequences
10	< b a c >
20	< (**b**,d) a (d,e) c **b** >
30	< **a** d c b >
40	< a d (b,e) c >
50	< c b c a b c >

HS(< b >, 20) = 2

HS(< a >, 30) = 1

HS(< c c >, 50) = 3

< **c** b **c** a b c >
< c b **c** a b **c** >
< **c** b c a b **c** >

Sequential Pattern - definitions

Mean Duration (MD)

MD of a pattern P in sequence S – Sum of time spans for each P instance in S, divided by P's occurrences.

SID	Sequences
10	< b a c >
20	< (**b**,d) a (d,e) c **b** >
30	< **a** d c b >
40	< a d (b,e) c >
50	< c b c a b c >

$$MD(< b >, 20) = \frac{1+1}{2} = \frac{2}{2} = 1$$

$$MD(< c\ c >, 50) = \frac{3+4+6}{3} = \frac{13}{3} = 4.33$$

< **c** b **c** a b c >
< c b **c** a b **c** >
< **c** b c a b **c** >

Attack Periods Anomaly Detection Hypothesis

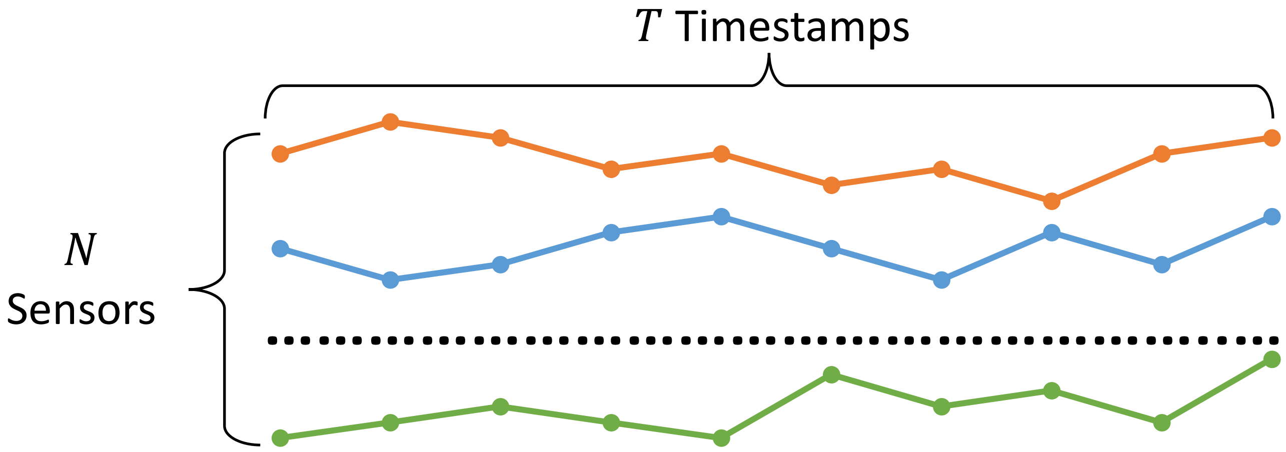
Since attacks can be seen as **anomalous periods** of sensor data, sequential patterns that are considered **frequent in mostly normal periods** will appear as **missing or have different properties**

- ✓ Motivation
- ✓ Background
- ❑ Proposed Method

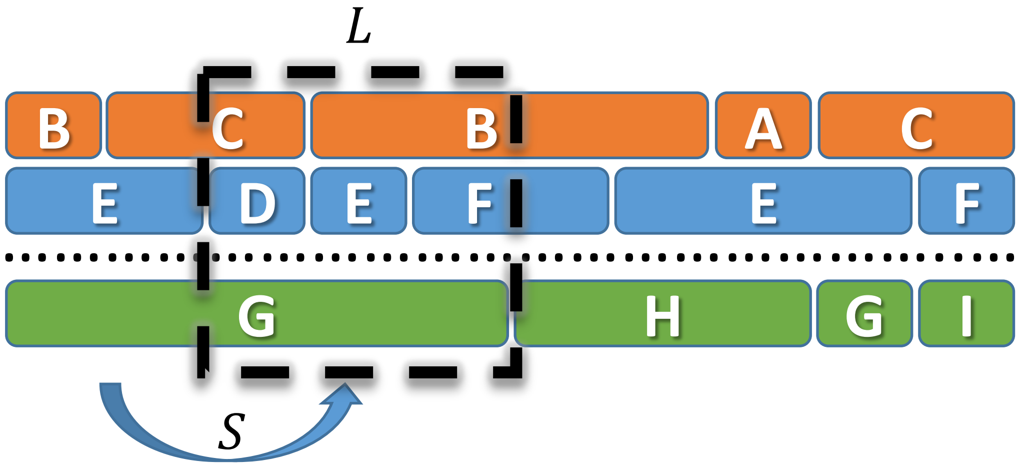
Proposed Method

Raw Time-Series to Sequence database

Raw Time-Series



Sliding Window

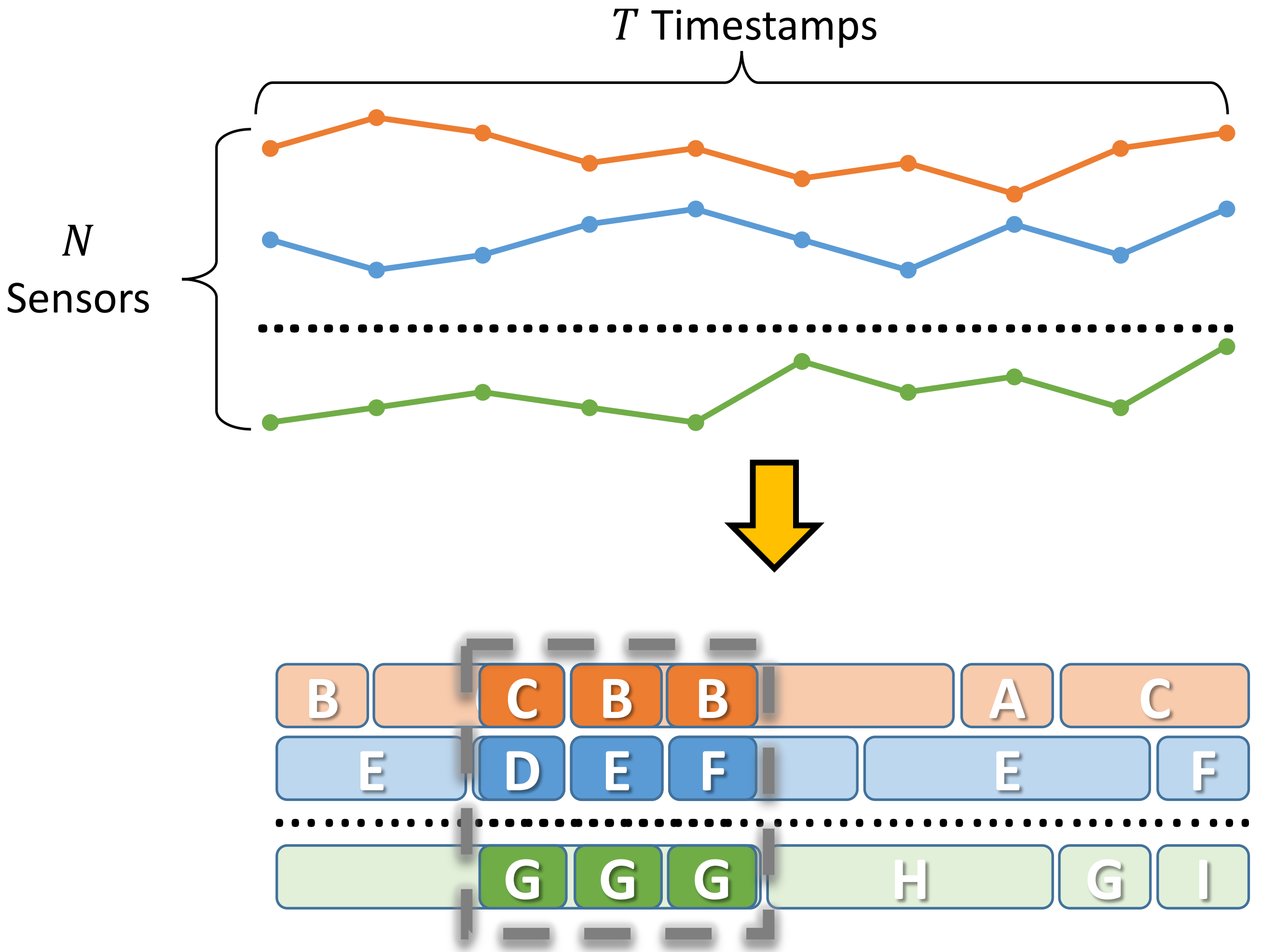


Sequence database

SID	Sequences
…	…
20	< (CD…G) (BE..G) (BF..G)>
…	…
…	…

Raw Time-Series to Sequence database

Raw Time-Series

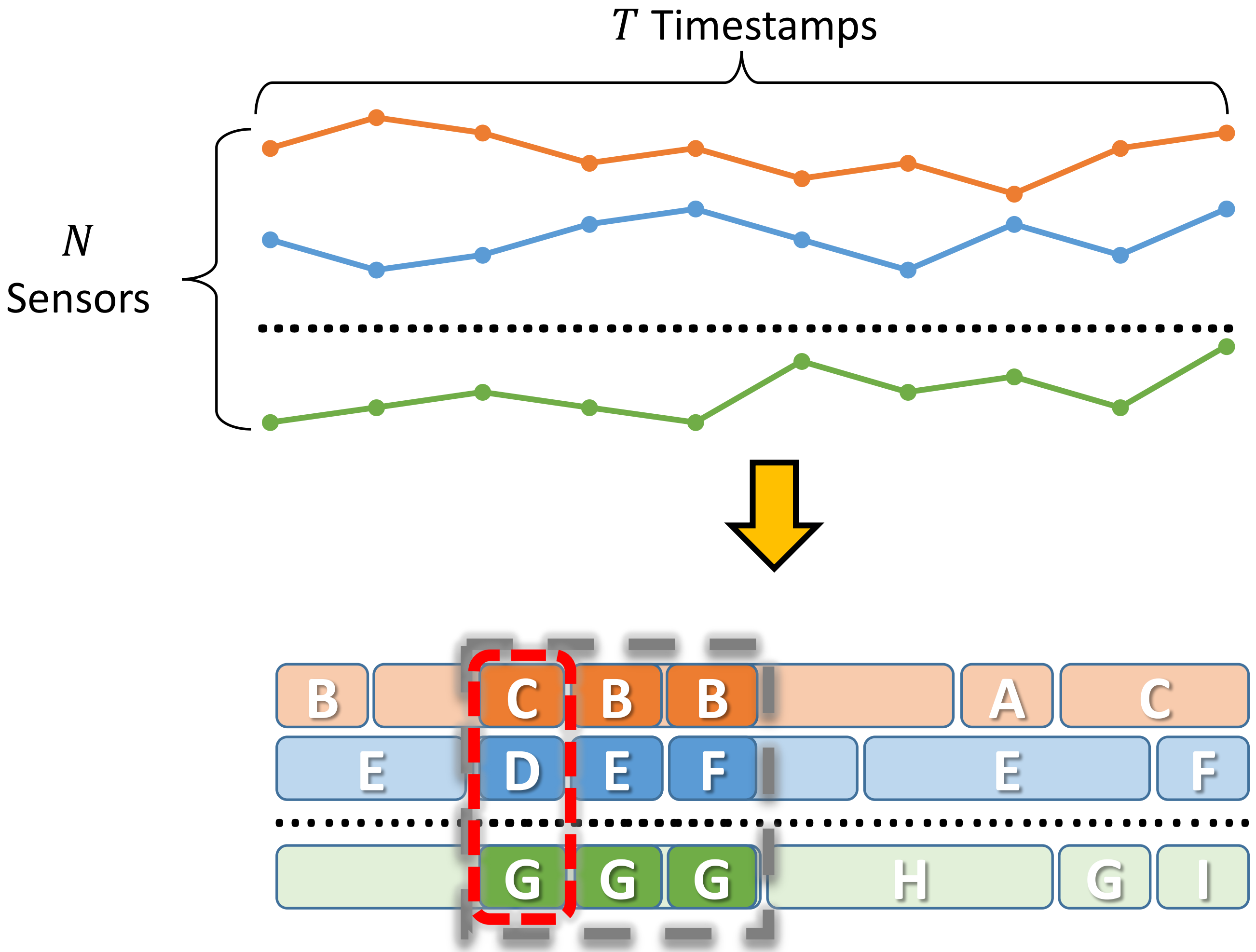


Sequence database

SID	Sequences
…	…
20	< (CD…G) (BE..G) (BF..G)>
…	…
…	…

Raw Time-Series to Sequence database

Raw Time-Series

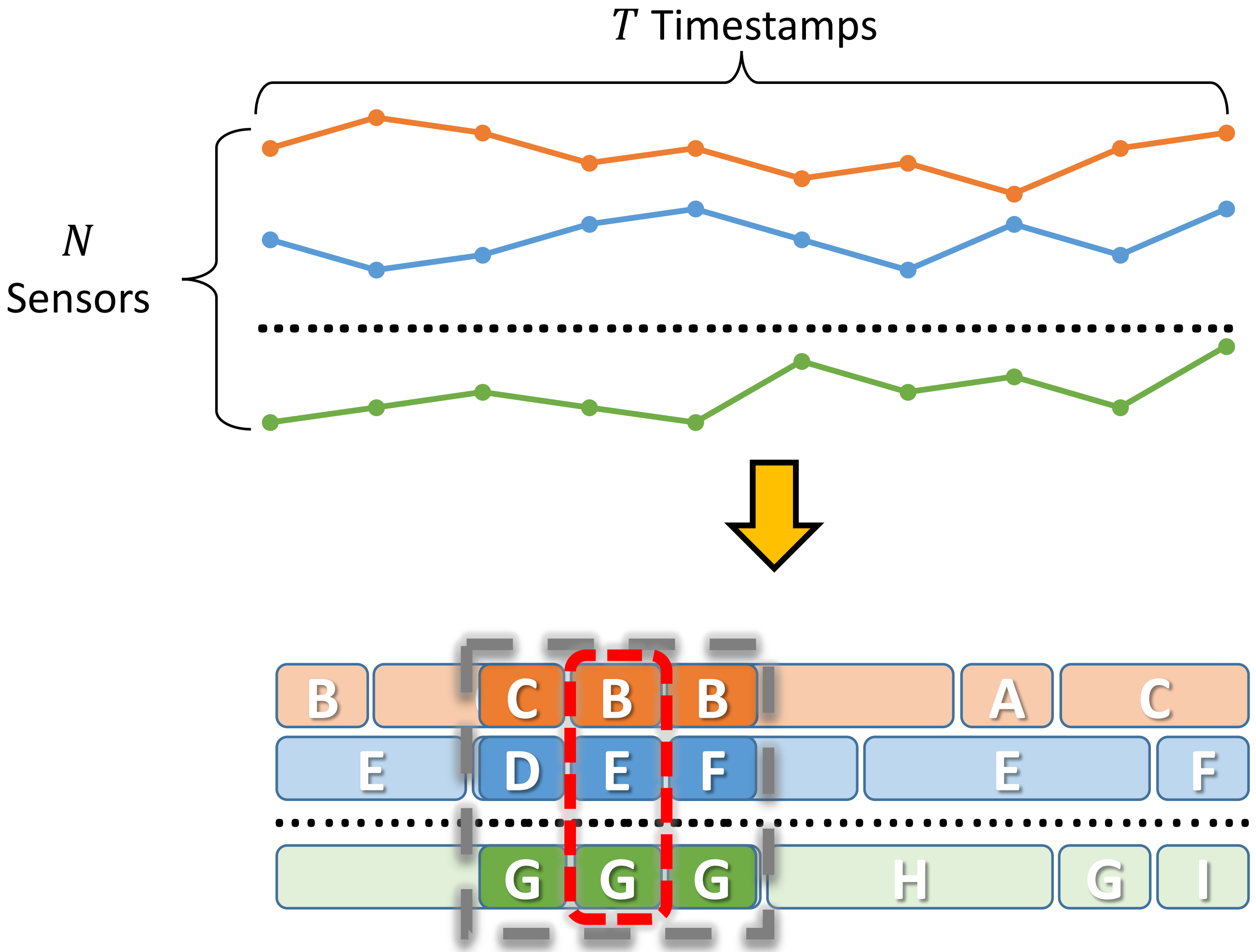


Sequence database

SID	Sequences
…	…
20	< **(CD…G)** (BE..G) (BF..G)>
…	…
…	…

Raw Time-Series to Sequence database

Raw Time-Series

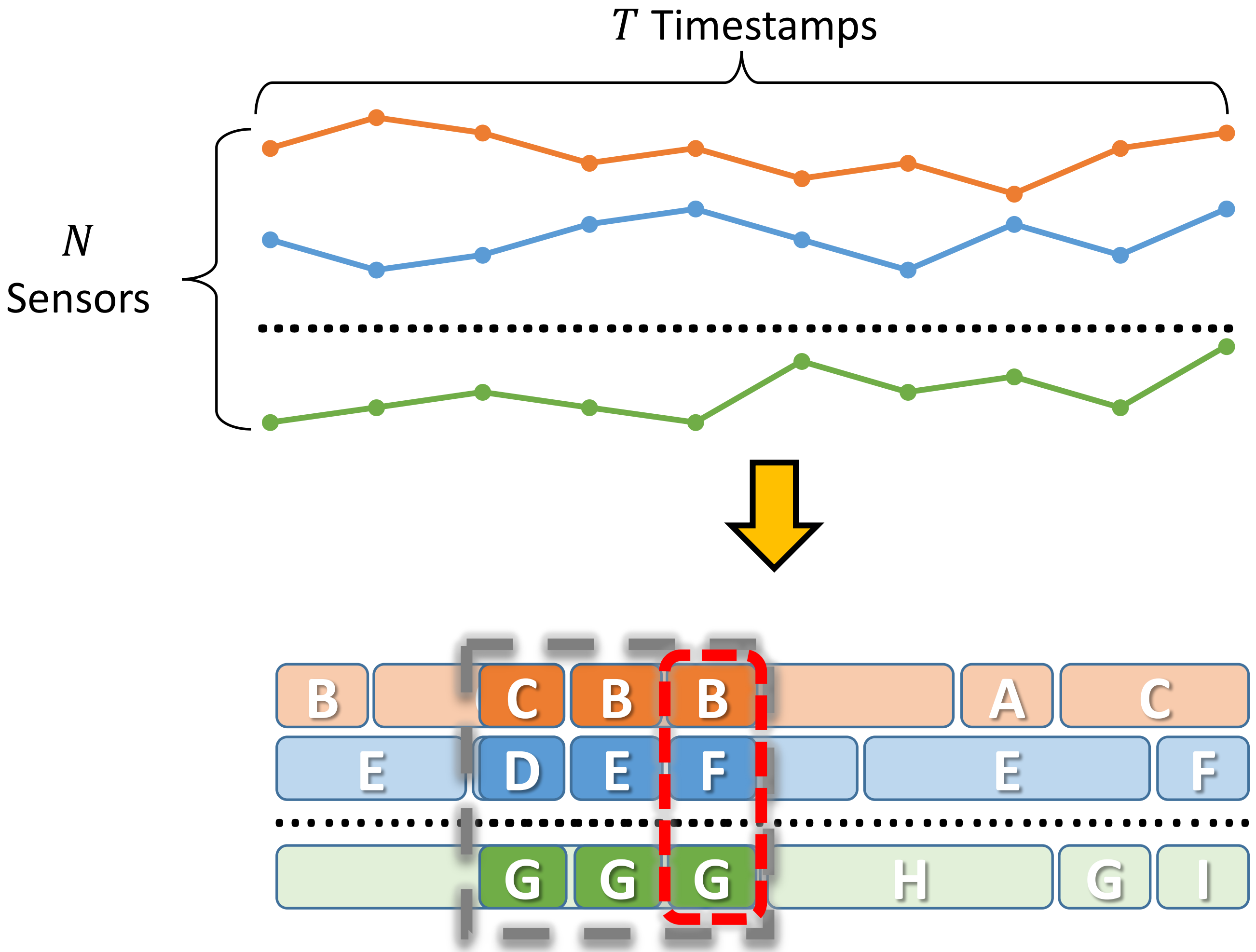


Sequence database

SID	Sequences
…	…
20	< (CD…G) **(BE..G)** (BF..G)>
…	…
…	…

Raw Time-Series to Sequence database

Raw Time-Series



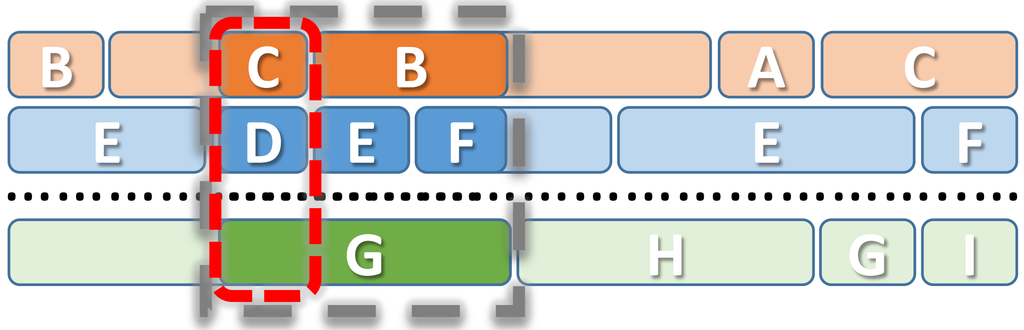
Sequence database

SID	Sequences
…	…
20	< (CD…G) (BE..G) **(BF..G)**>
…	…
…	…

Raw Time-Series to Sequence database - Tiep

We also introduce and evaluate a more efficient approach for sequence extraction - "Tiep" (Time Interval Endpoint).

- Aims to mitigate scalability issues in high-resolution data.
- Unlike traditional methods, focuses only on starting time points.
- Significantly reducing computational overhead and memory usage.



Full Original Sequence:

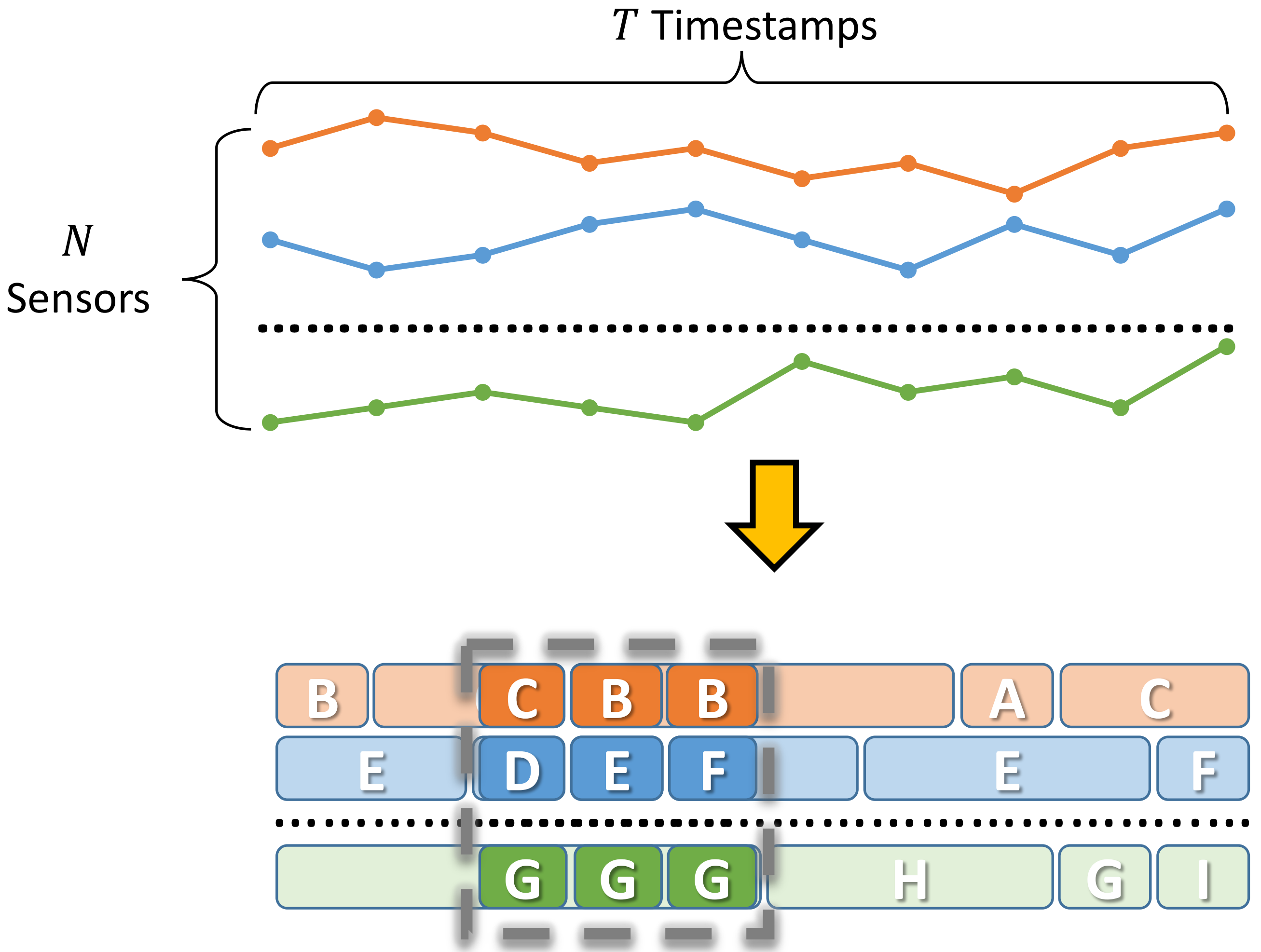
< (CDG) (BEG) (BFG)>

Tiep Sequence:

< (CDG)(BE) F >

Raw Time-Series to Sequence database

Raw Time-Series



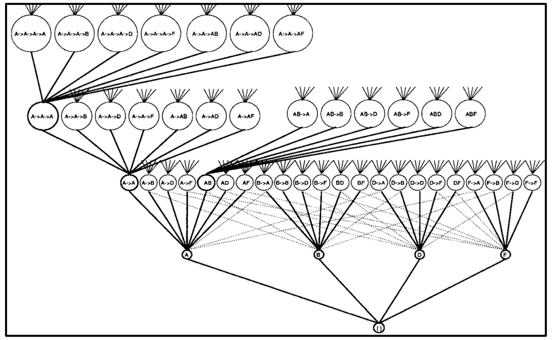
Sequence database

SID	Sequences
…	…
20	< (CD…G) (BE..G) (BF..G)>
…	…
…	…

Sequential Pattern Mining

Sequence database

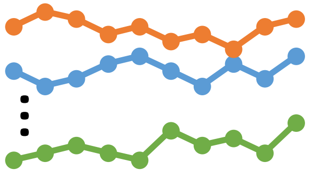
SID	Sequences
…	…
20	< (CD…G) (BE..G) (BF..G)>
…	…
…	…



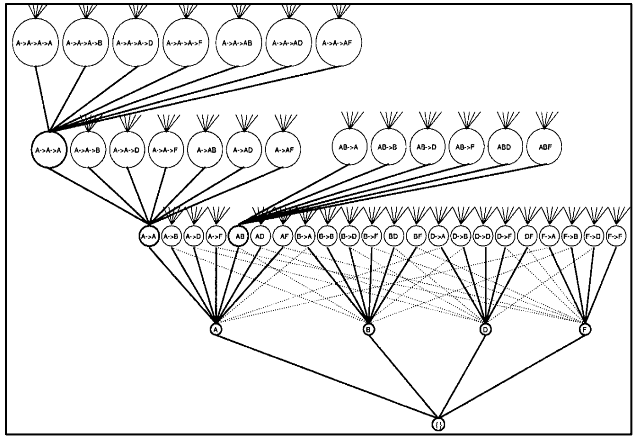
Frequent Patterns	Support
< C G>	1000
…	…
…	…
< (C G) B >	50
…	…
…	…
< (C G) B B>	20
…	…

Proposed Methodology

Normal Behavior

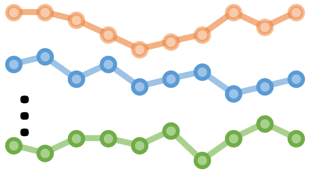


Sequences
Seq1
Seq2
…
SeqK



	FP_1	FP_2	…	FP_N
Seq1	1.3	4.6	…	1.7
Seq2	2	2.3	..	4
…	…	…	…	…
SeqK	3.5	5.2	…	4.25

New Observations



Sequences
SeqK+1
SeqK+2
…
SeqK+P

Pattern Detector

	FP_1	FP_2	…	FP_N
SeqK+2	3.7	0	…	1.25

Anomaly Detector

Normal

Anomalous

Evaluation

Dataset

Secure Water Treatment (SWaT)

- **Source**: iTrust, Singapore University of Technology and Design
- **URL**: https://itrust.sutd.edu.sg/itrust-labs_datasets/
- **Objective**: Evaluate anomaly detection in industrial control systems

Divided into two parts: (1) Training set and (2) Test set

- **Training set**: 7 days of normal operation data
- **Test set**: 4 days of data with 36 injected cyber-physical attacks

Timestamp	FIT101	LIT101	AIT201	MV101	P101	P102	AIT202
22/12/2015 4:00:00 PM	2.470294	261.5804	244.3284	2	2	1	8.19008
22/12/2015 4:00:01 PM	2.457163	261.1879	244.3284	2	2	1	8.19008
22/12/2015 4:00:02 PM	2.439548	260.9131	244.3284	2	2	1	8.19008
22/12/2015 4:00:03 PM	2.428338	260.285	244.3284	2	2	1	8.19008
22/12/2015 4:00:04 PM	2.424815	259.8925	244.4245	2	2	1	8.19008
22/12/2015 4:00:05 PM	2.425456	260.0495	244.5847	2	2	1	8.19008
22/12/2015 4:00:06 PM	2.472857	260.2065	244.5847	2	2	1	8.19008
22/12/2015 4:00:07 PM	2.513532	260.5991	244.5847	2	2	1	8.19008
22/12/2015 4:00:08 PM	2.559972	261.0309	244.5847	2	2	1	8.19008
22/12/2015 4:00:09 PM	2.598085	261.1093	244.809	2	2	1	8.19008
22/12/2015 4:00:10 PM	2.630753	261.7766	244.809	2	2	1	8.19008
22/12/2015 4:00:11 PM	2.649329	261.7766	244.809	2	2	1	8.19008
22/12/2015 4:00:12 PM	2.654133	261.8944	244.8731	2	2	1	8.19008
22/12/2015 4:00:13 PM	2.646446	261.6589	244.8731	2	2	1	8.19008
22/12/2015 4:00:14 PM	2.625949	261.2664	245.0333	2	2	1	8.19008
22/12/2015 4:00:15 PM	2.61602	260.8346	245.0333	2	2	1	8.19008
22/12/2015 4:00:16 PM	2.609935	261.0309	245.0333	2	2	1	8.19008
22/12/2015 4:00:17 PM	2.602889	261.1093	245.0333	2	2	1	8.19008
22/12/2015 4:00:18 PM	2.587516	260.9916	245.0333	2	2	1	8.19008
22/12/2015 4:00:19 PM	2.573103	261.3056	245.0333	2	2	1	8.19008
22/12/2015 4:00:20 PM	2.556769	261.6589	245.4499	2	2	1	8.19008

Research Questions

1. Which **number of bins** has the best performance in abstracting our data?

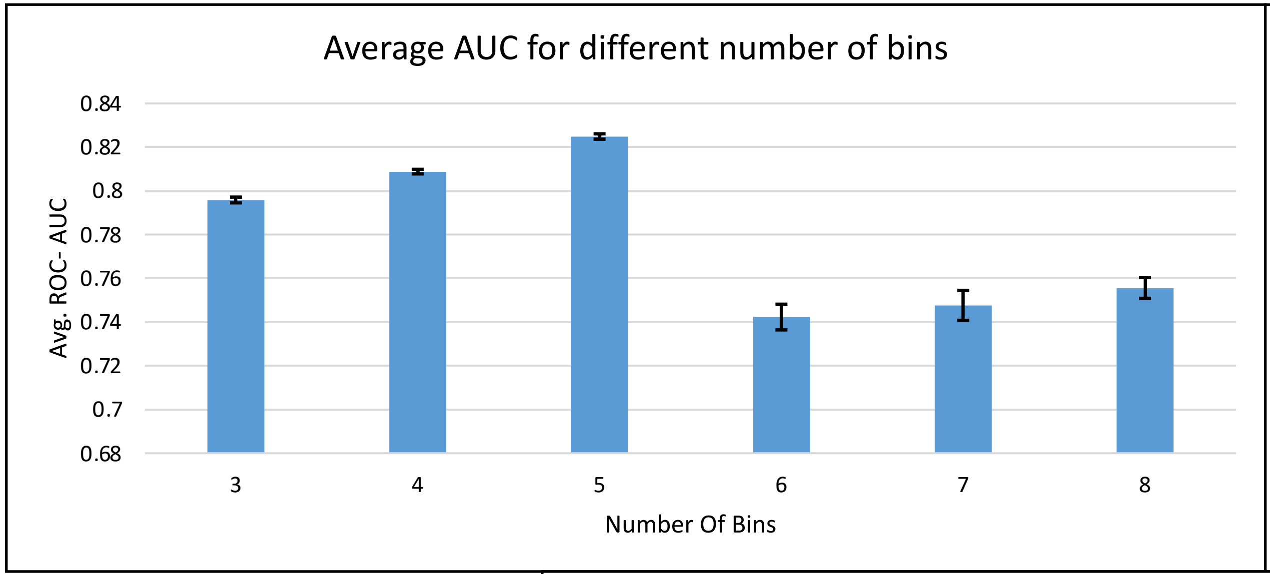
2. Which **metric** has the most effective performance between frequent patterns and transactions, such as Binary, Horizontal Support, and Mean Duration?

3. Which state-of-the-art **anomaly detector**, including OneClassSVM and SGDOneClassSVM, achieves optimal generalization on our data and has the best performance when utilizing them?

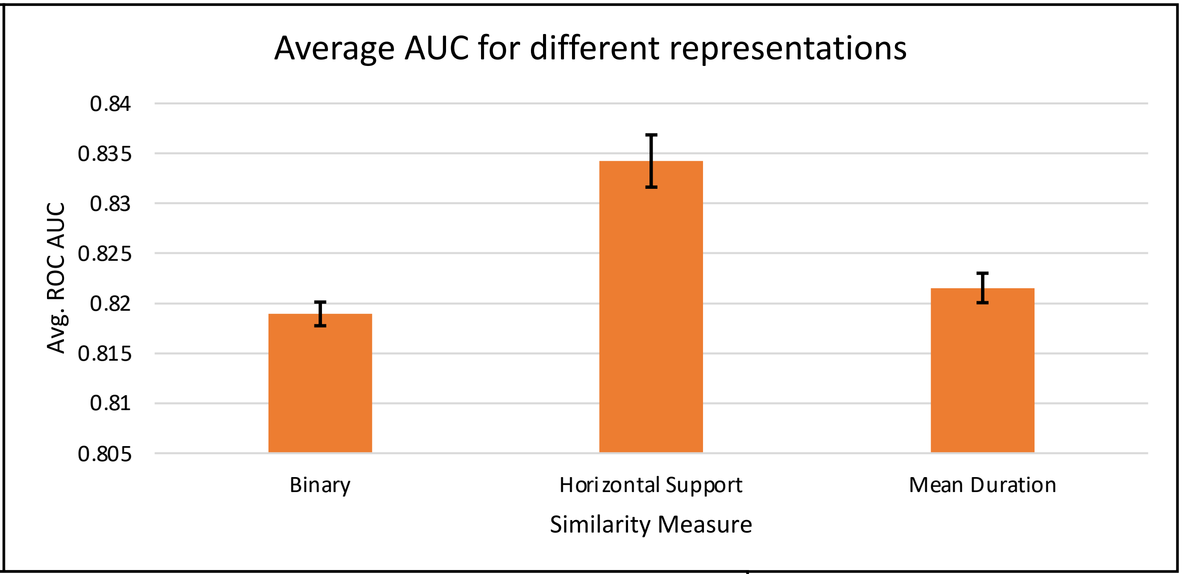
- ✓ Motivation
- ✓ Background
- ✓ Proposed Method
- ✓ Evaluation
- ❑ Initial Results

Initial Results

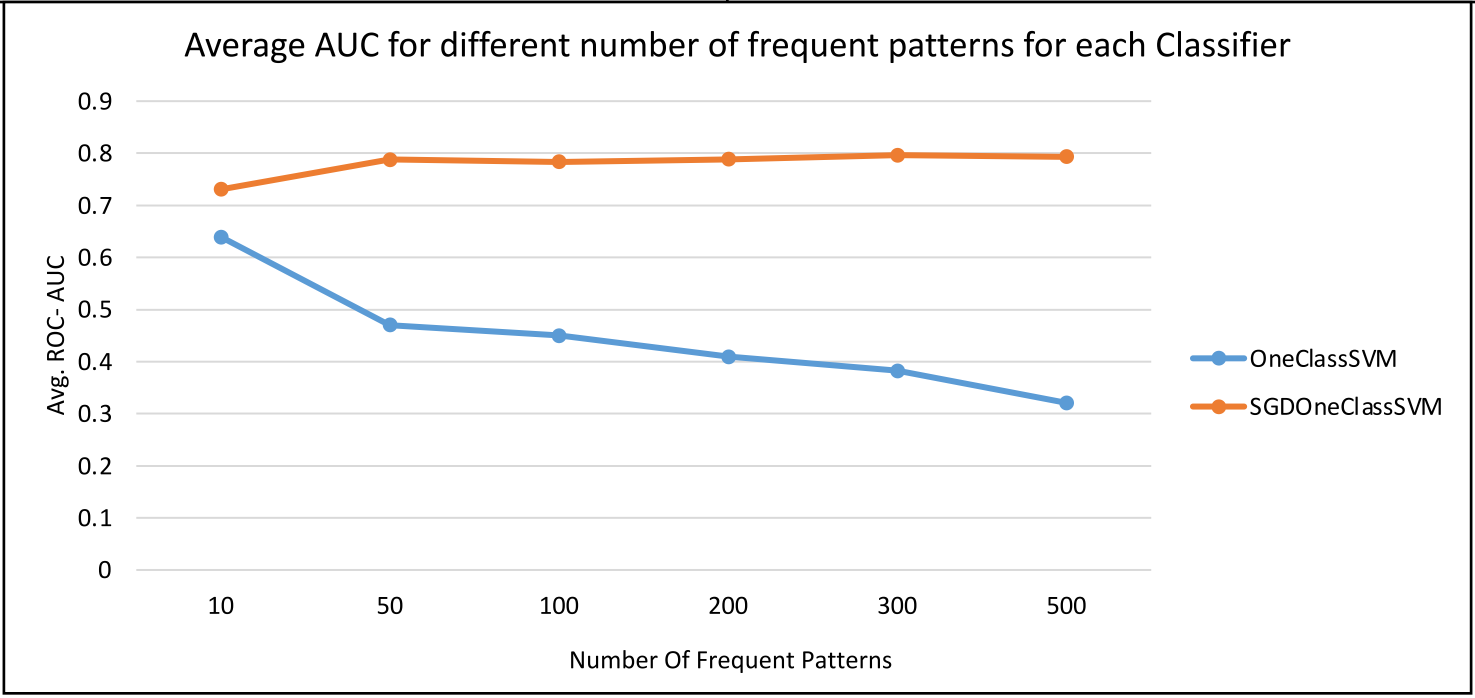
Key Results



Average AUC for different number of bins
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
Avg. ROC- AUC
3
4
5
6
7
8
Number Of Bins



Average AUC for different representations
0.84
0.835
0.83
0.825
0.82
0.815
0.81
0.805
Avg. ROC AUC
Binary
Horizontal Support
Mean Duration
Similarity Measure



Average AUC for different number of frequent patterns for each Classifier
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Avg. ROC- AUC
10
50
100
200
300
500
Number Of Frequent Patterns
OneClassSVM
SGDOneClassSVM

Overall Results

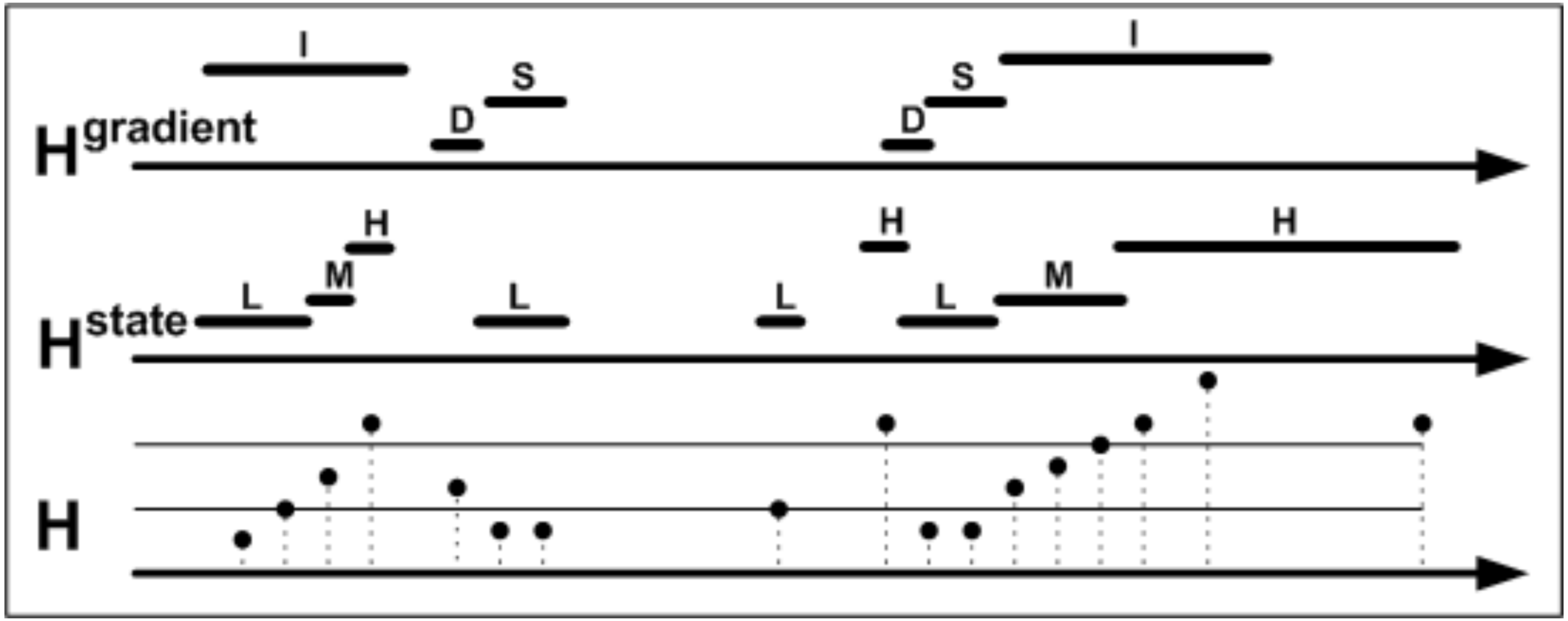
Method	SWaT		* P-Precision, R-Recall	
	P *	R *	AUC	F1
MERLIN	0.656	0.2547	0.6175	0.3669
LSTM-NDT	0.7778	0.5109	0.714	0.6167
DAGMM	0.9933	0.6879	0.8436	0.8128
OmniAnomaly	0.9782	0.6957	0.8467	0.8131
MAD-GAN	0.9593	0.6957	0.8463	0.8065
USAD	0.9977	0.6879	0.846	0.8143
MTAD-GAT	0.9718	0.6957	0.8464	0.8109
CAE-M	0.9697	0.6957	0.8464	0.8101
GDN	0.9591	0.6957	0.8462	0.8101
GRN-50	0.9972	0.5921	**0.8781**	0.7389
GRN-100	**0.9986**	0.5909	**0.8845**	0.7496
Our best algorithm:	**0.998**	0.583	0.8646	0.736

- ✓ Motivation
- ✓ Background
- ✓ Proposed Method
- ✓ Evaluation
- ✓ Initial Results
- ❑ Future Research Directions

Future Research Directions

More datasets
Evaluation with TIEPs
Explainability
Using TIRPs
Using TIRPs with Similarity

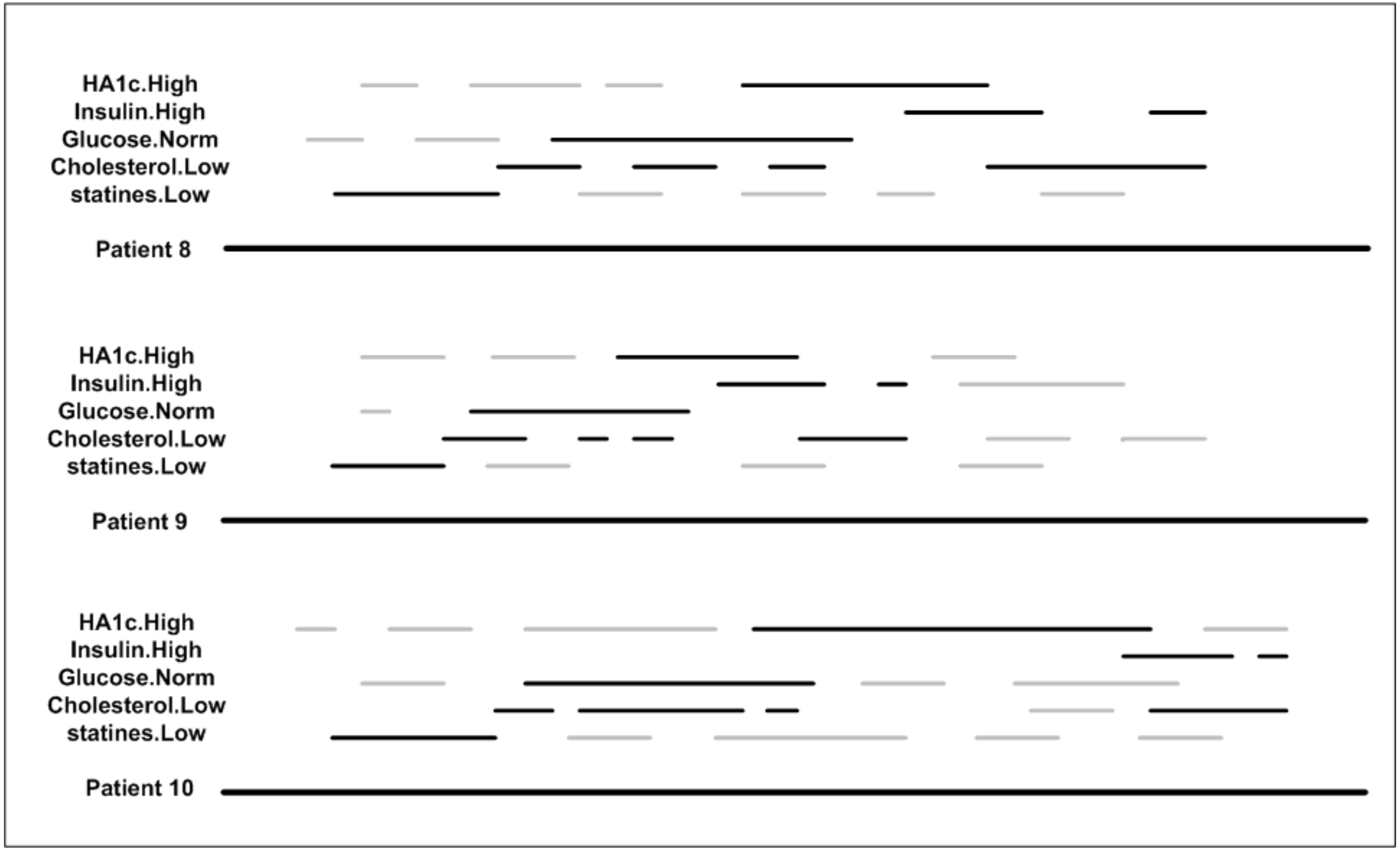
From Time Points to Time Intervals Series



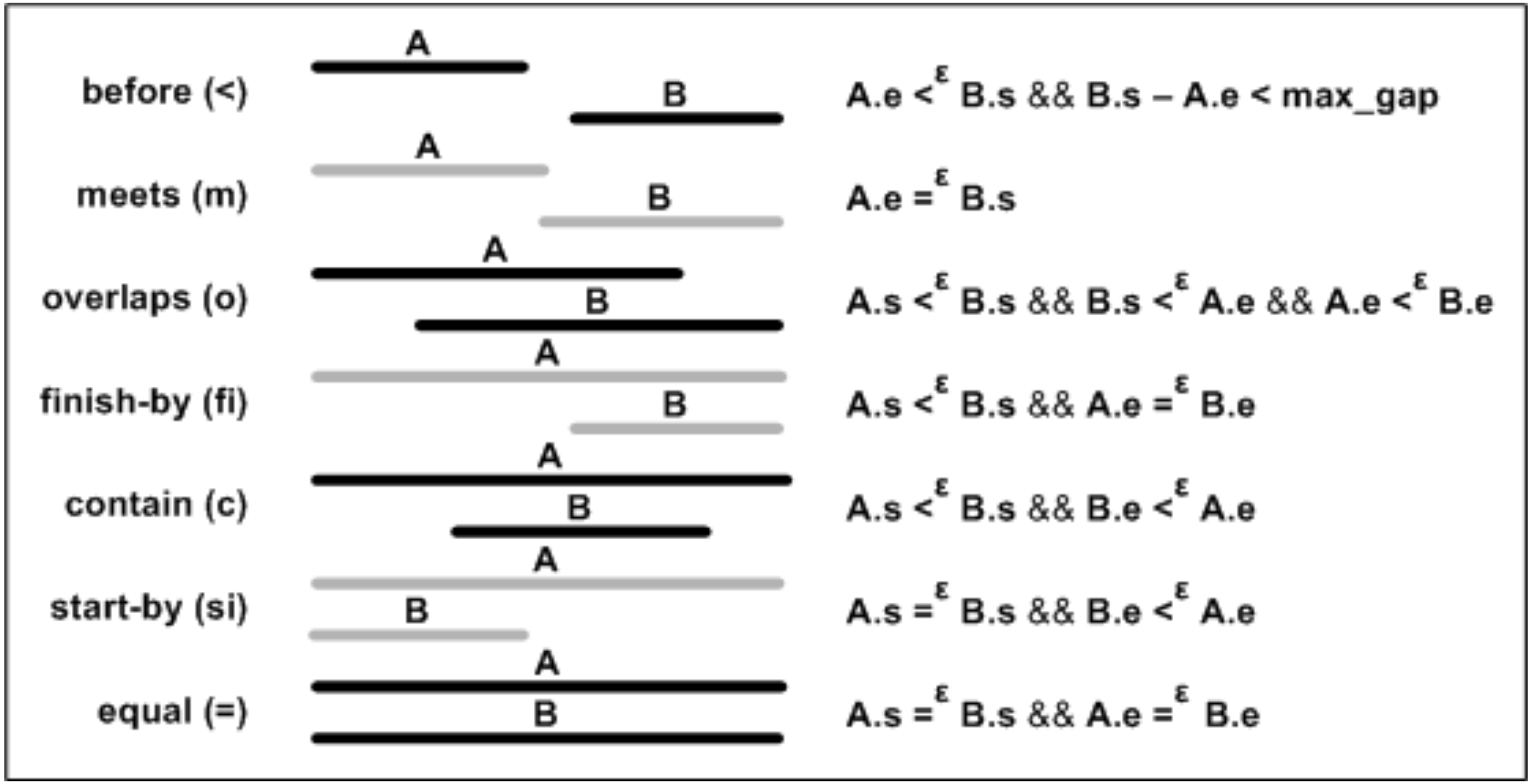
Robert Moskovitch, Yuval Shahar, Classification Driven Temporal Discretization of Multivariate Time Series, *Data Mining and Knowledge Discovery*, 29, 4, 871-913, 2015.

Time Intervals Related Patterns Discovery

Time Intervals Related Patterns Discovery

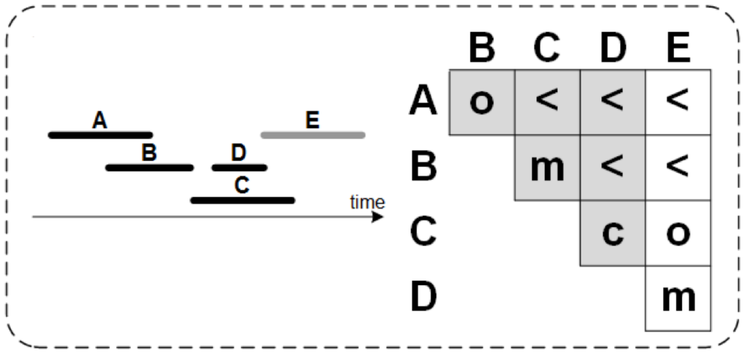


Allen’s (1983) Temporal Logic



Time Intervals Related Pattern - **TIRP**

A TIRP is a conjunction of pairwise temporal relations



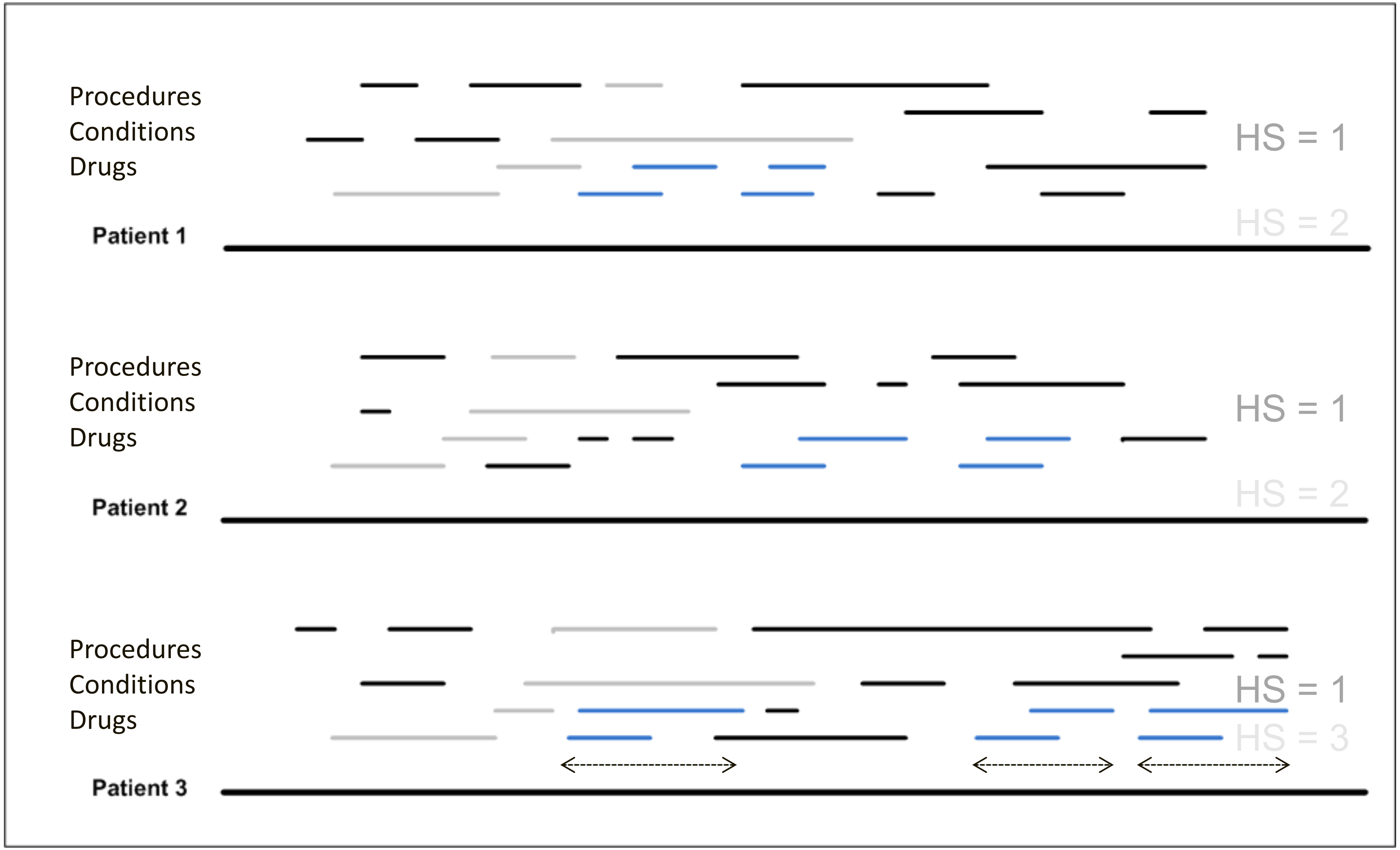
{A o B, A < C, As < D, A < E, B m C, B < D, B < E, C c D, C o E, D m E}

A k-sized TIRP includes $k(k-1)/2 = (k^2-k)/2$ temporal relations

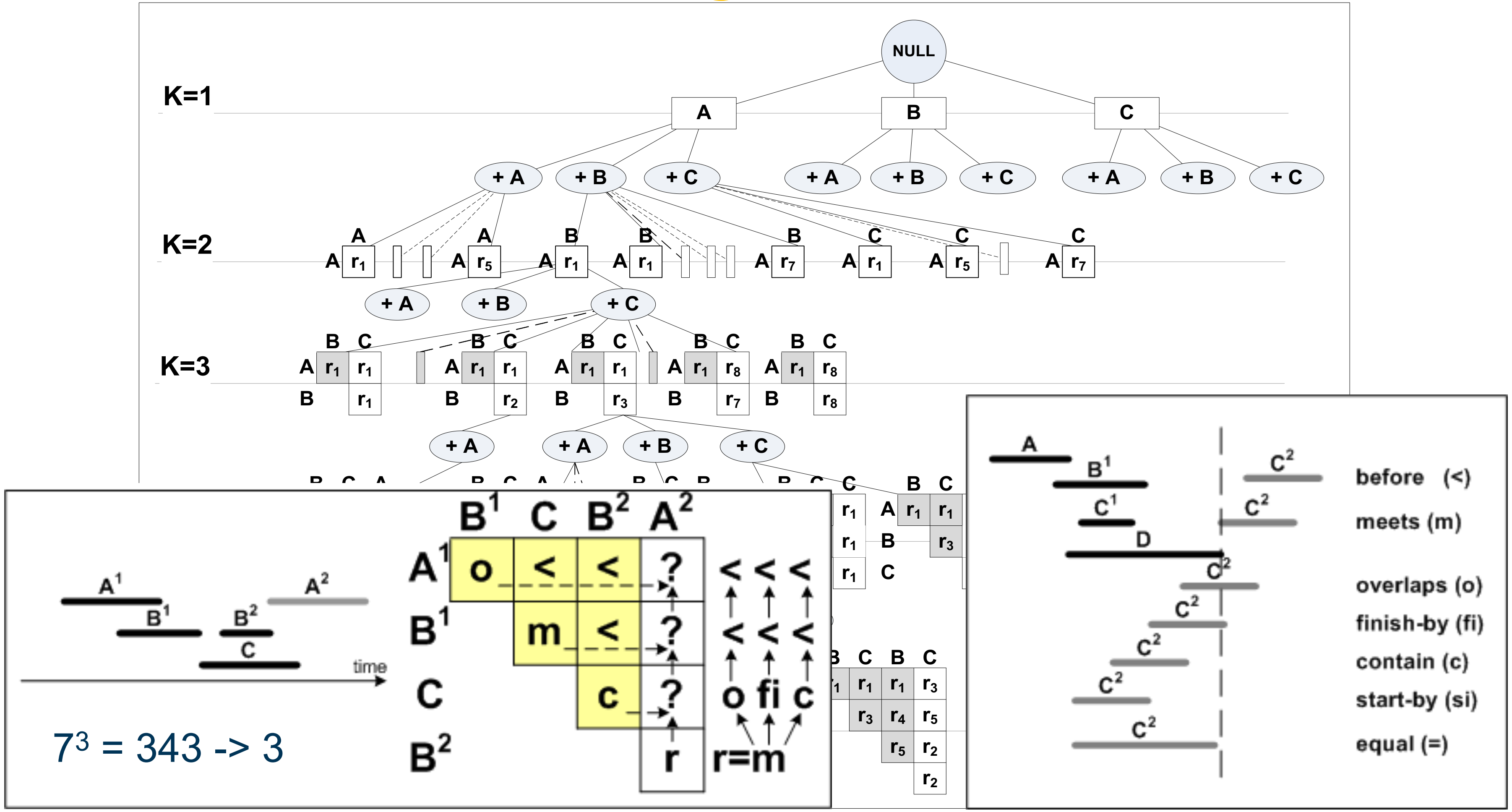
TIRPs have several metrics, which can be predictive

- Vertical Support – how many patients have the TIRP in the database
- Horizontal Support – how many instances (episodes) of the TIRP were in the past hours (or weeks)
- Mean Duration – what is the average duration of these instances

Time Intervals Related Patterns Discovery – an illustration

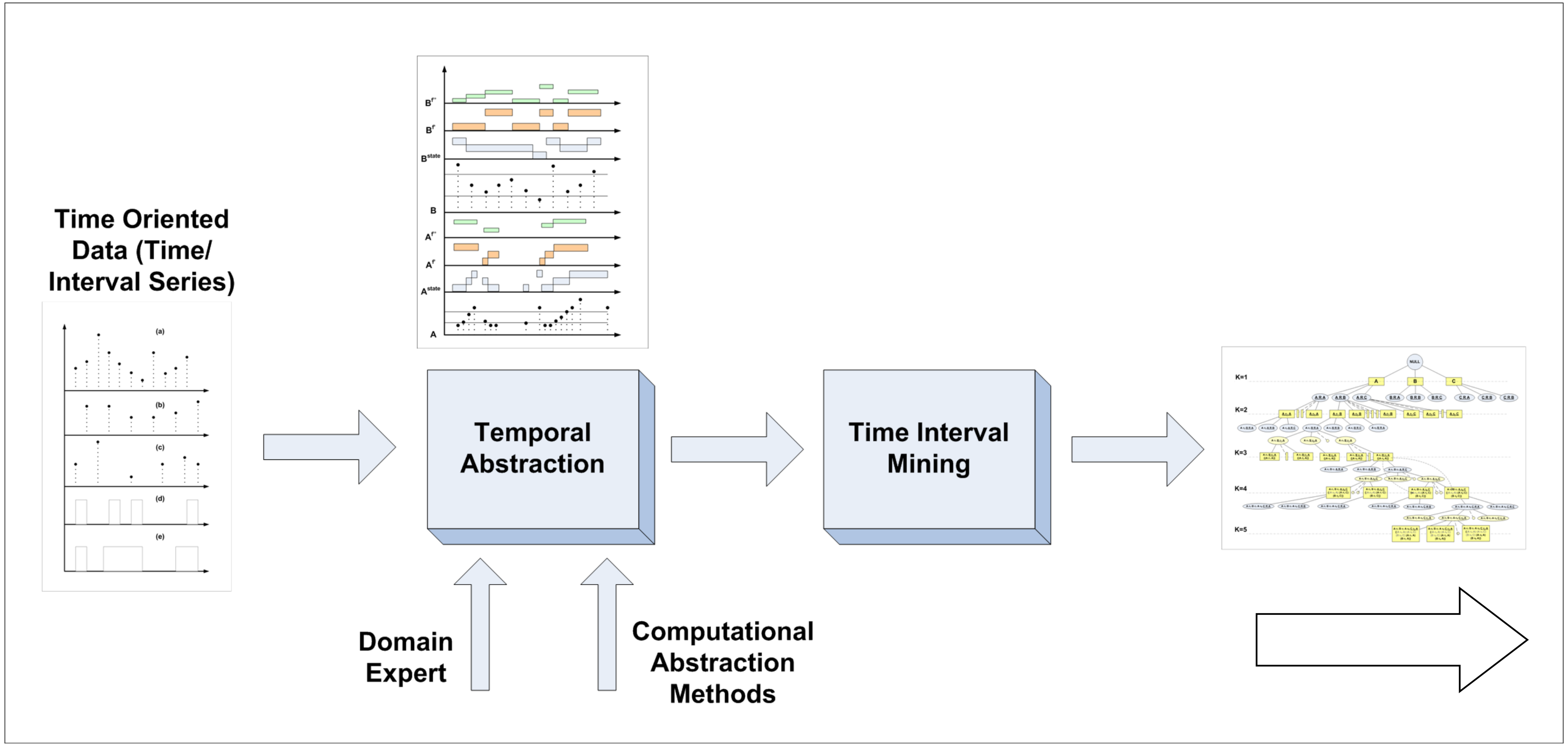


KarmaLego
Fast Time Intervals Mining



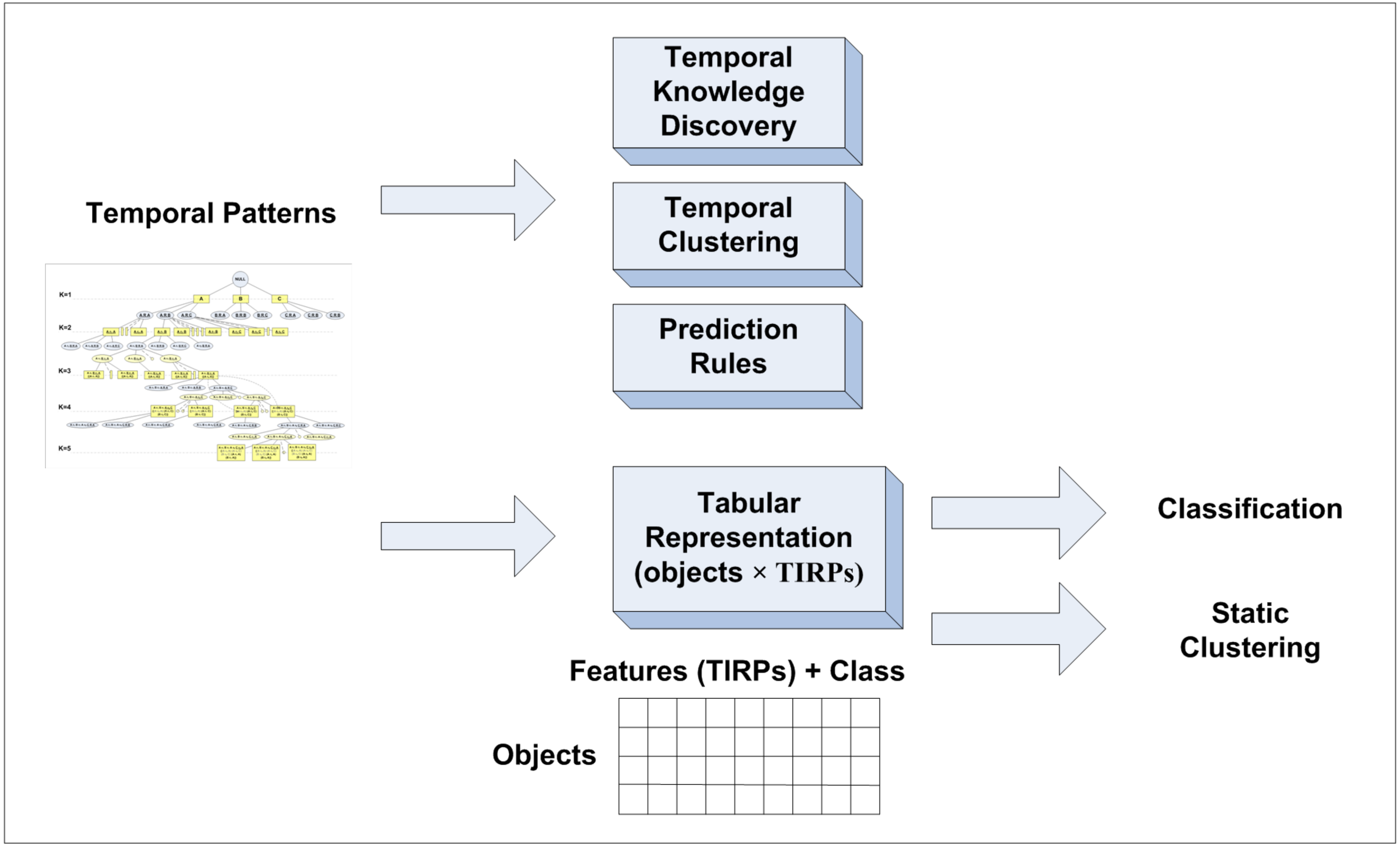
Moskovitch et al, Outcomes Prediction via Time Intervals Related Patterns, IEEE International Conference on Data Mining (ICDM), 2015

KarmaLego General Workflow

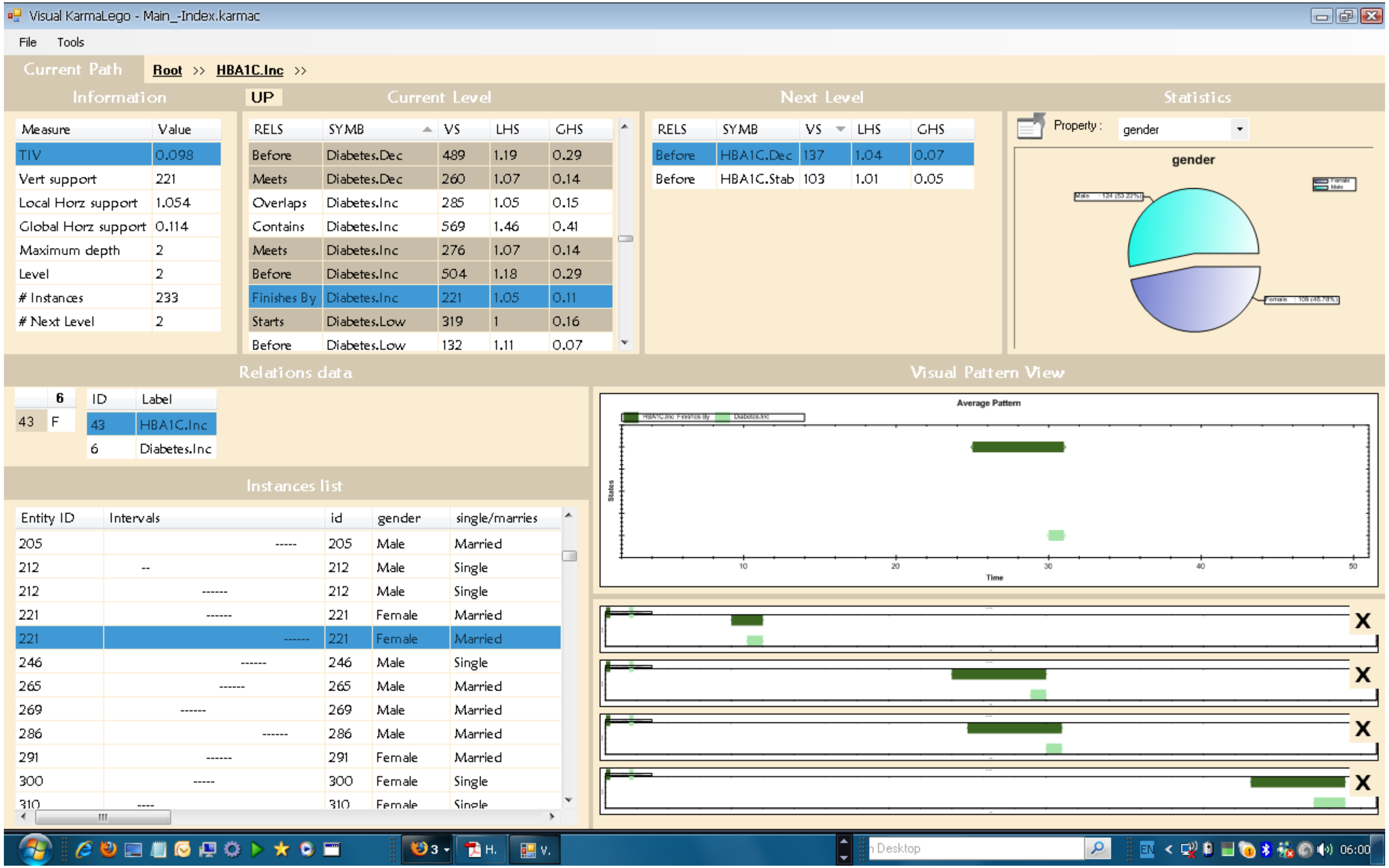


Moskovitch et al, Fast Time Intervals Mining by Exploiting the Transitivity of Temporal Relations, Knowledge and Information Systems, 2015.

Use of TIRPs from KarmaLego

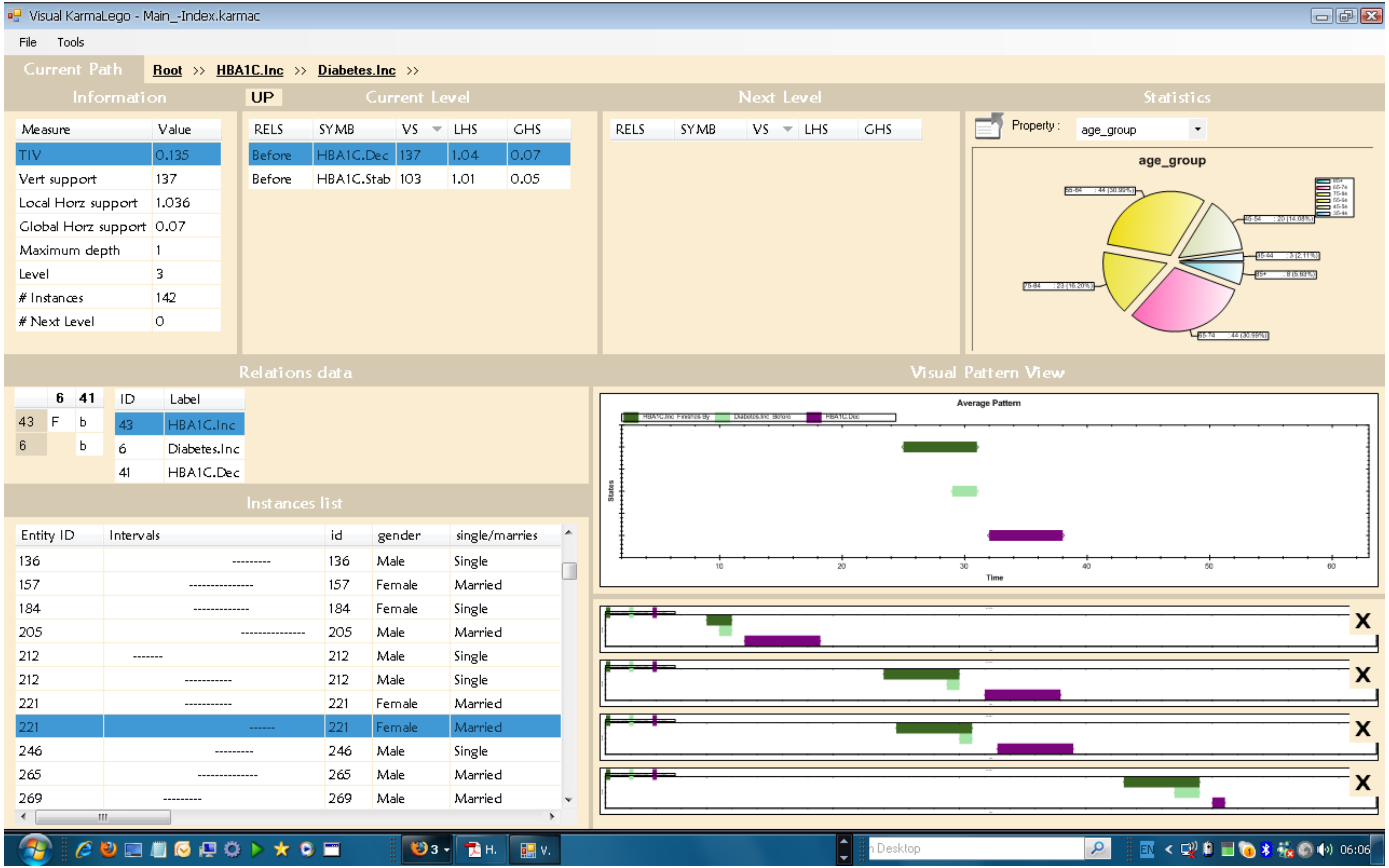


KarmaLegoV

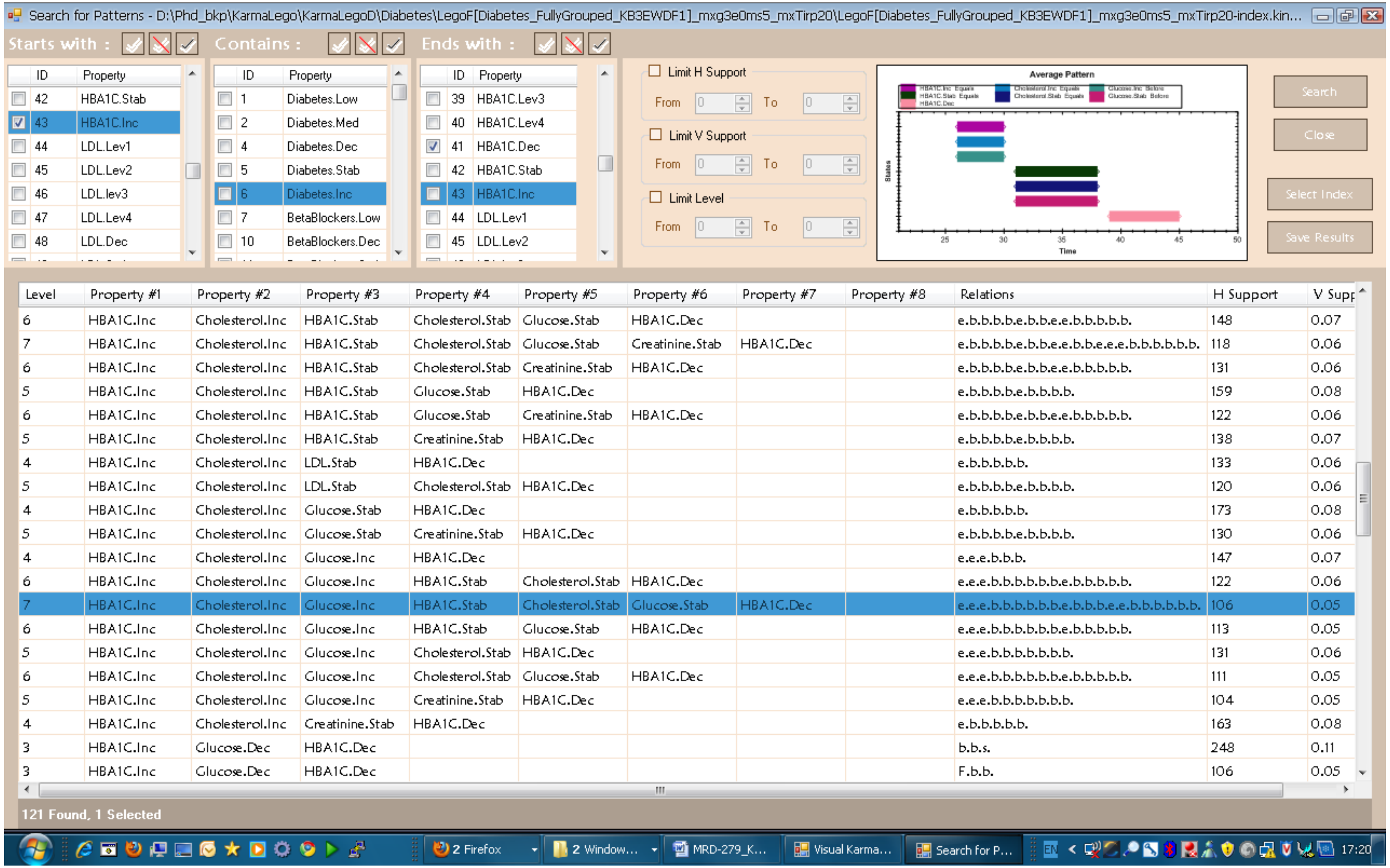


Visual KarmaLego - Main_-Index.karmac
File Tools
Current Path Root >> HBA1C.Inc >>
Information
UP
Current Level
Next Level
Statistics
Measure Value
TIV 0.098
Vert support 221
Local Horz support 1.054
Global Horz support 0.114
Maximum depth 2
Level 2
Instances 233
Next Level 2
RELS SYMB VS LHS GHS
Before Diabetes.Dec 489 1.19 0.29
Meets Diabetes.Dec 260 1.07 0.14
Overlaps Diabetes.Inc 285 1.05 0.15
Contains Diabetes.Inc 569 1.46 0.41
Meets Diabetes.Inc 276 1.07 0.14
Before Diabetes.Inc 504 1.18 0.29
Finishes By Diabetes.Inc 221 1.05 0.11
Starts Diabetes.Low 319 1 0.16
Before Diabetes.Low 132 1.11 0.07
RELS SYMB VS LHS GHS
Before HBA1C.Dec 137 1.04 0.07
Before HBA1C.Stab 103 1.01 0.05
Property : gender
gender
Relations data
Visual Pattern View
6
43 F
ID Label
43 HBA1C.Inc
6 Diabetes.Inc
Average Pattern
States
Time
Instances list
Entity ID Intervals id gender single/marries
205 205 Male Married
212 212 Male Single
212 212 Male Single
221 221 Female Married
221 221 Female Married
246 246 Male Single
265 265 Male Married
269 269 Male Married
286 286 Male Married
291 291 Female Married
300 300 Female Single
310 310 Female Single

KarmaLegoV

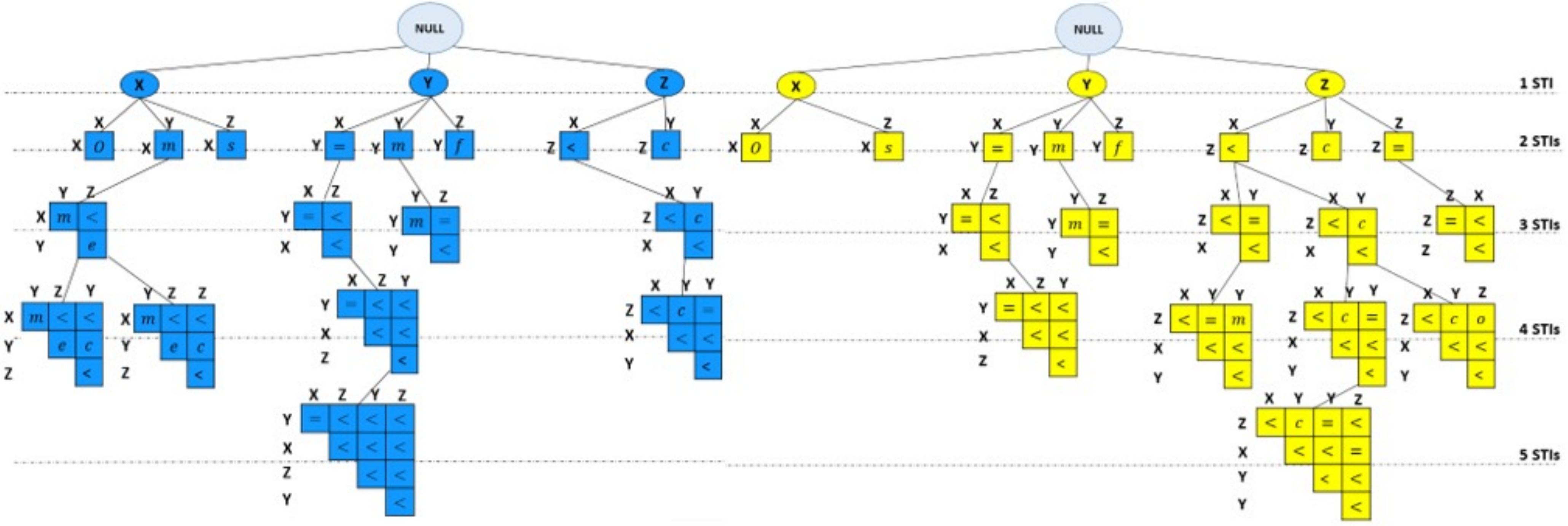

Visual KarmaLego - Main_-Index.karmac
File
Tools
Current Path
Root >> HBA1C.Inc >> Diabetes.Inc >>
Information
UP
Current Level
Next Level
Statistics
Measure
Value
TIV
0.135
Vert support
137
Local Horz support
1.036
Global Horz support
0.07
Maximum depth
1
Level
3
Instances
142
Next Level
0
RELS
SYMB
VS
LHS
GHS
Before
HBA1C.Dec
137
1.04
0.07
Before
HBA1C.Stab
103
1.01
0.05
Property :
age_group
Relations data
Visual Pattern View
ID
Label
43
HBA1C.Inc
6
Diabetes.Inc
41
HBA1C.Dec
Average Pattern
States
Time
Instances list
Entity ID
Intervals
id
gender
single/marries
136
Male
Single
157
Female
Married
184
Female
Single
205
Male
Married
212
Male
Single
221
Female
Married
246
Male
Single
265
Male
Married
269
Male
Married
Desktop
06:06

KarmaLegoV – Search Results



Level	Property #1	Property #2	Property #3	Property #4	Property #5	Property #6	Property #7	Property #8	Relations	H Support	V Supp
6	HBA1C.Inc	Cholesterol.Inc	HBA1C.Stab	Cholesterol.Stab	Glucose.Stab	HBA1C.Dec			e.b.b.b.b.e.b.b.e.e.b.b.b.b.b.	148	0.07
7	HBA1C.Inc	Cholesterol.Inc	HBA1C.Stab	Cholesterol.Stab	Glucose.Stab	Creatinine.Stab	HBA1C.Dec		e.b.b.b.b.e.b.b.e.e.b.b.e.e.e.b.b.b.b.b.b.	118	0.06
6	HBA1C.Inc	Cholesterol.Inc	HBA1C.Stab	Cholesterol.Stab	Creatinine.Stab	HBA1C.Dec			e.b.b.b.b.e.b.b.e.e.b.b.b.b.b.	131	0.06
5	HBA1C.Inc	Cholesterol.Inc	HBA1C.Stab	Glucose.Stab	HBA1C.Dec				e.b.b.b.b.e.b.b.b.b.	159	0.08
6	HBA1C.Inc	Cholesterol.Inc	HBA1C.Stab	Glucose.Stab	Creatinine.Stab	HBA1C.Dec			e.b.b.b.b.e.b.b.e.e.b.b.b.b.b.	122	0.06
5	HBA1C.Inc	Cholesterol.Inc	HBA1C.Stab	Creatinine.Stab	HBA1C.Dec				e.b.b.b.b.e.b.b.b.b.	138	0.07
4	HBA1C.Inc	Cholesterol.Inc	LDL.Stab	HBA1C.Dec					e.b.b.b.b.b.	133	0.06
5	HBA1C.Inc	Cholesterol.Inc	LDL.Stab	Cholesterol.Stab	HBA1C.Dec				e.b.b.b.b.e.b.b.b.b.	120	0.06
4	HBA1C.Inc	Cholesterol.Inc	Glucose.Stab	HBA1C.Dec					e.b.b.b.b.b.	173	0.08
5	HBA1C.Inc	Cholesterol.Inc	Glucose.Stab	Creatinine.Stab	HBA1C.Dec				e.b.b.b.b.b.b.b.b.b.	130	0.06
4	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	HBA1C.Dec					e.e.e.b.b.b.	147	0.07
6	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	HBA1C.Stab	Cholesterol.Stab	HBA1C.Dec			e.e.e.b.b.b.b.b.b.e.b.b.b.b.b.	122	0.06
7	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	HBA1C.Stab	Cholesterol.Stab	Glucose.Stab	HBA1C.Dec		e.e.e.b.b.b.b.b.b.e.b.b.b.e.e.b.b.b.b.b.b.	106	0.05
6	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	HBA1C.Stab	Glucose.Stab	HBA1C.Dec			e.e.e.b.b.b.b.b.b.e.b.b.b.b.b.	113	0.05
5	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	Cholesterol.Stab	HBA1C.Dec				e.e.e.b.b.b.b.b.b.b.	131	0.06
6	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	Cholesterol.Stab	Glucose.Stab	HBA1C.Dec			e.e.e.b.b.b.b.b.b.e.b.b.b.b.b.	111	0.05
5	HBA1C.Inc	Cholesterol.Inc	Glucose.Inc	Creatinine.Stab	HBA1C.Dec				e.e.e.b.b.b.b.b.b.b.	104	0.05
4	HBA1C.Inc	Cholesterol.Inc	Creatinine.Stab	HBA1C.Dec					e.b.b.b.b.b.	163	0.08
3	HBA1C.Inc	Glucose.Dec	HBA1C.Dec						b.b.s.	248	0.11
3	HBA1C.Inc	Glucose.Dec	HBA1C.Dec						F.b.b.	106	0.05

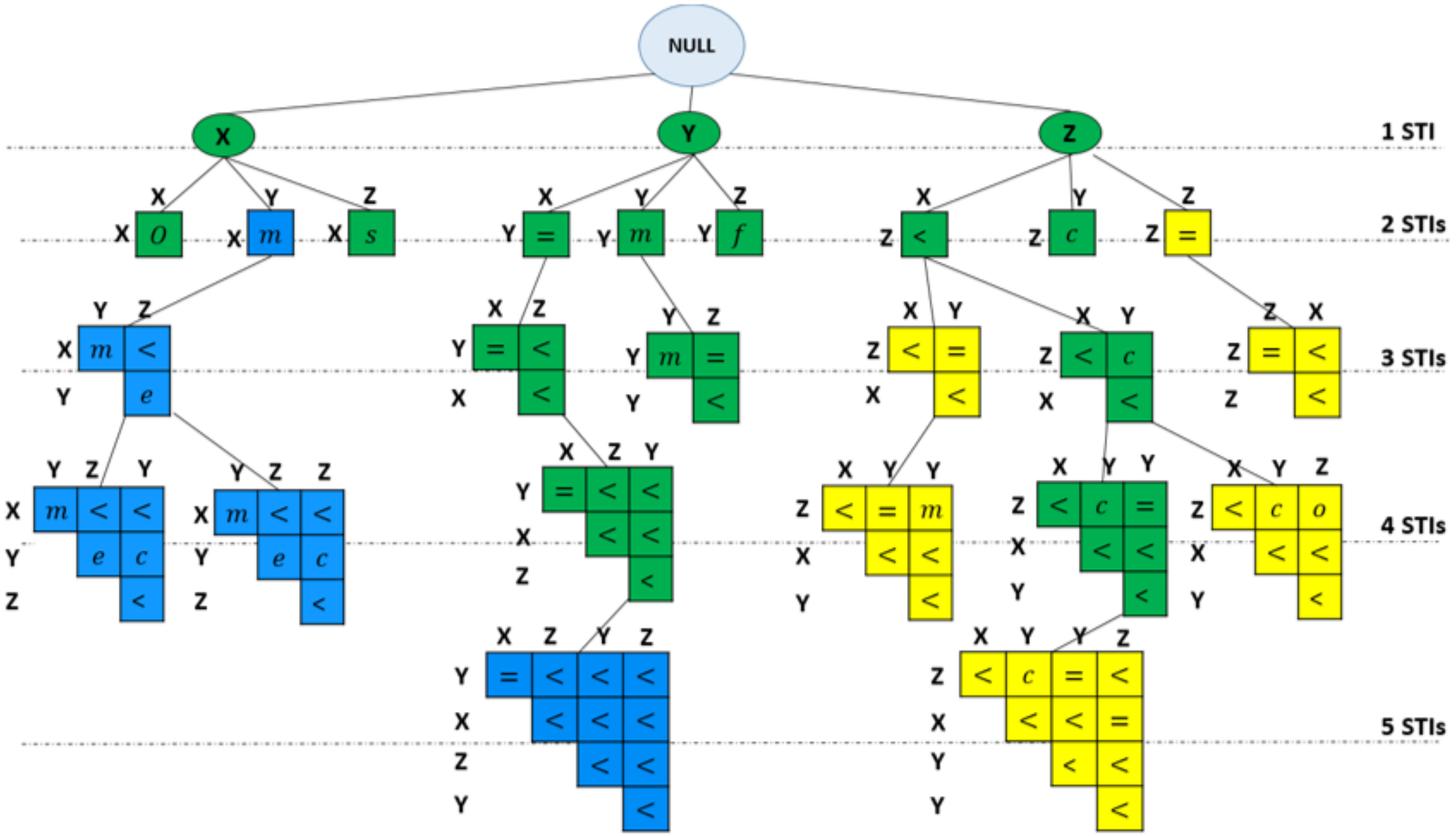
Visualization of Predictive TIRPs in Two Populations



a. **Population 1 TIRPs Tree**

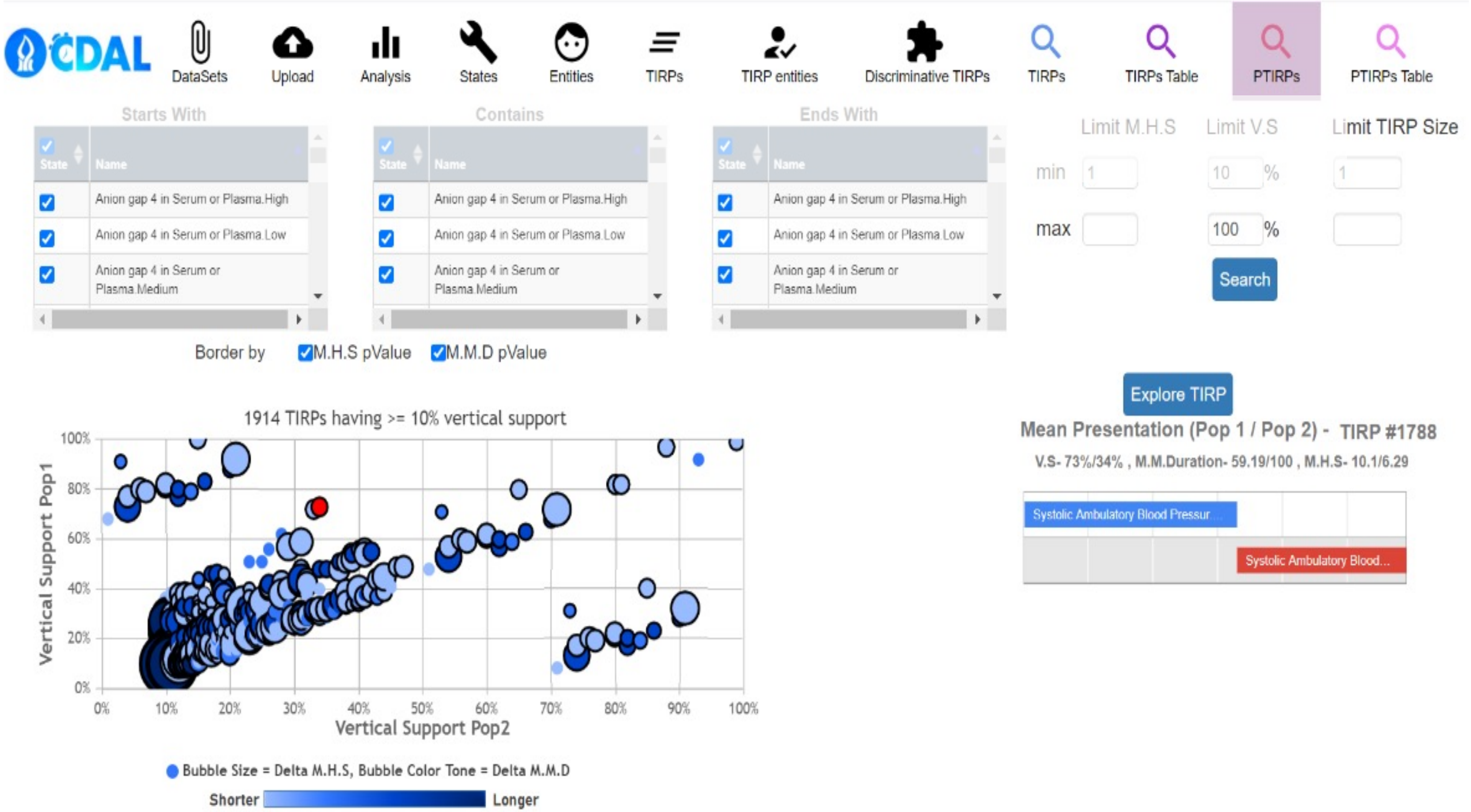
b. **Population 2 TIRPs Tree**

Visualization of Predictive TIRPs in Two Populations



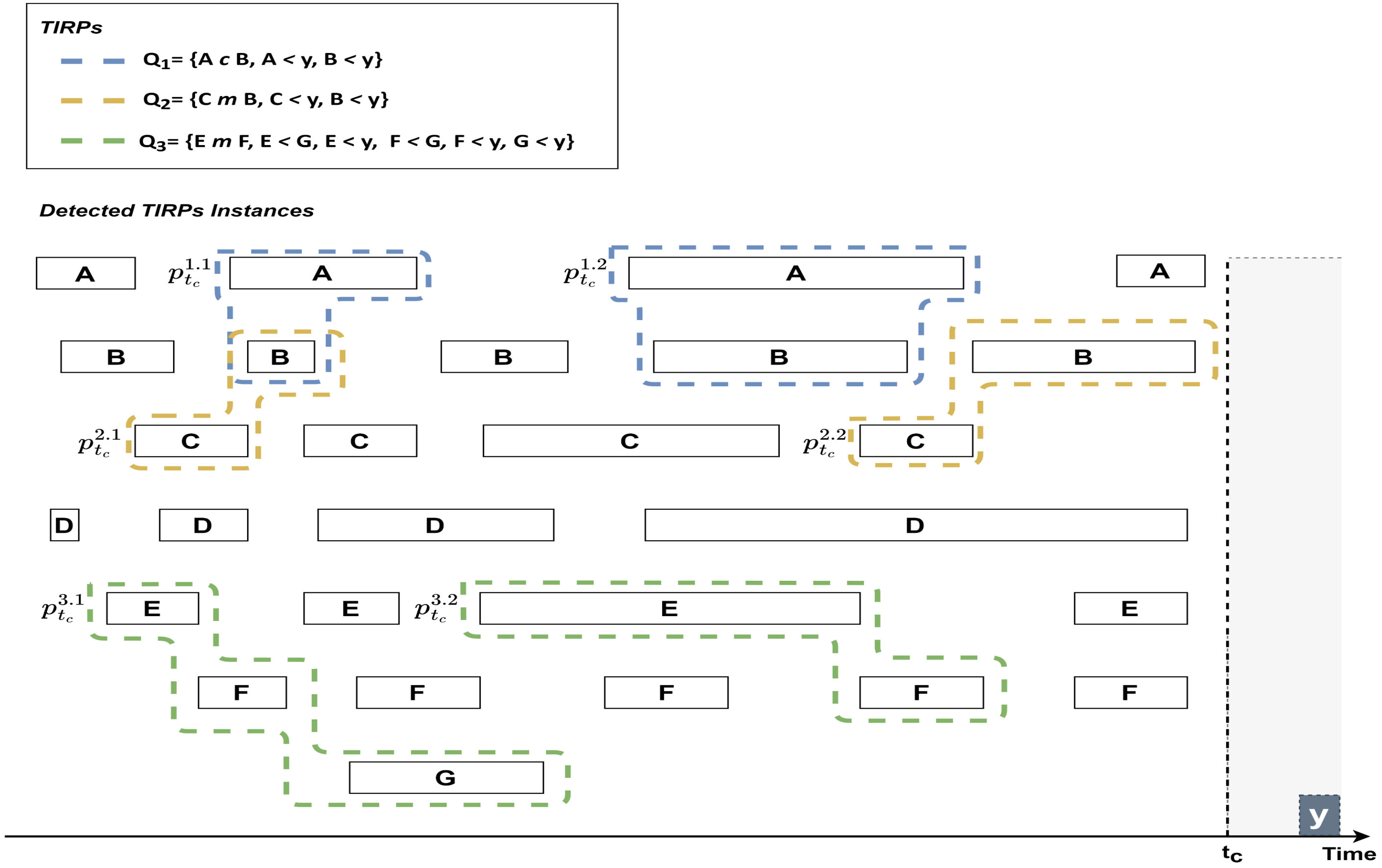
c. The Populations Unified TIRPs Tree

Visualization of Predictive TIRPs in Two Populations



Guy Shitrit, Noam Tractinsky, Robert Moskovitch, Visualization of Frequent Temporal Patterns in Single or Two Populations, *Journal of Biomedical Informatics*, 2022.

Continuous event's prediction via TIRPs



Itzhak et al, Continuous Prediction of Temporal Pattern Completion, PAKDD 2023

Next Directions with TIRPs Similarity

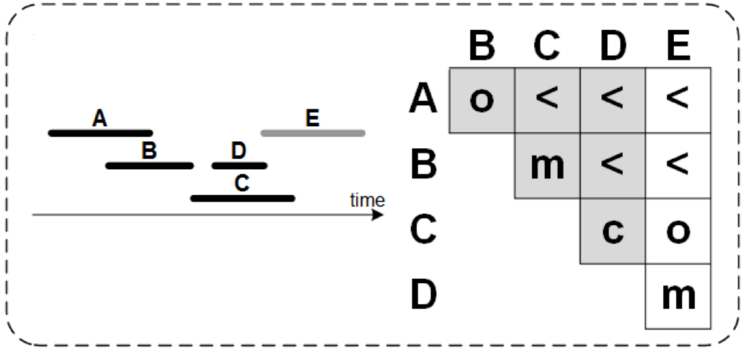
- The idea is to develop a similarity function for a TIRP, based on their time intervals duration, and temporal relations
- Use that for Anomaly Detection
- Use that for COPE identification

Time Intervals Related Pattern - **TIRP**

A TIRP is a conjunction of pairwise temporal relations

{A o B, A < C, As < D, A < E, B m C, B < D, B < E, C c D, C o E, D m E}

A k-sized TIRP includes $k(k-1)/2 = (k^2-k)/2$ temporal relations



TIRPs have several metrics, which can be predictive

- Vertical Support – how many patients have the TIRP in the database
- Horizontal Support – how many instances (episodes) of the TIRP were in the past hours (or weeks)
- Mean Duration – what is the average duration of these instances

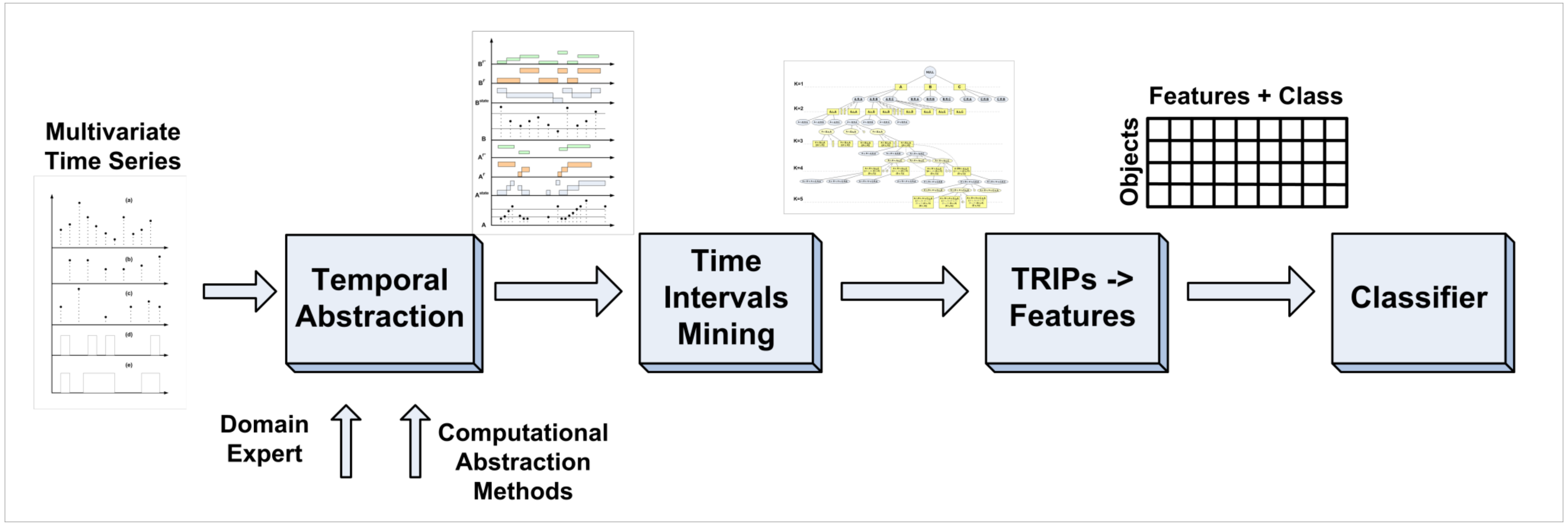
Using Similarity for Anaomaly Detection, or COPE ideantification

- Using the similarity function for Anomaly Detection
  - How the patterns in the data are similar or anomaluous?
  - Not similar -> anomalous, and how? Using a threshold
  - Include explainability – which patterns, and how different?

- Using the similarity function for COPE Identification
  - After having a TIRP that idenitifies a COPE (system situation)
  - How similar is it?
  - Or how different than common appearance
  - – will give much more granular detection framework

KarmaLegoSification

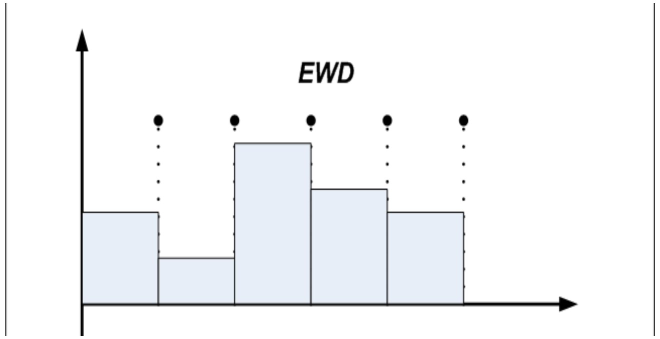
- A major problem in multivariate time series classification is the various types of raw temporal data
- TIRPs can be useful here as well, as classification features



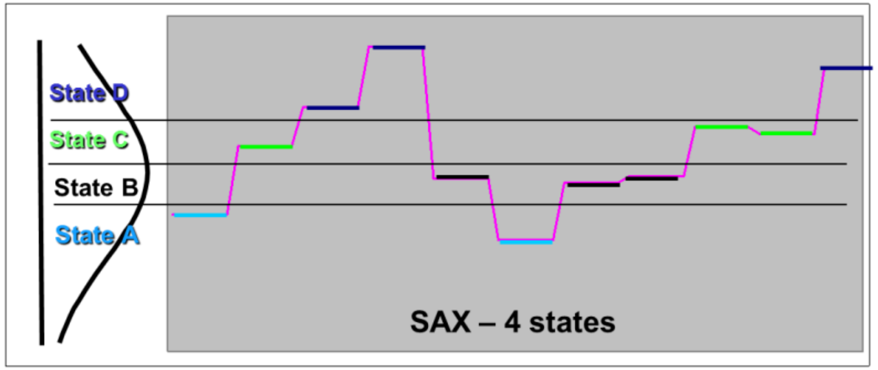
Moskovitch et al., *Data Mining and Knowledge Discovery*, 2015

Unsupervised Discretization: EWD and SAX

- :EWD the continuous values range is divided into k equal bins (states)



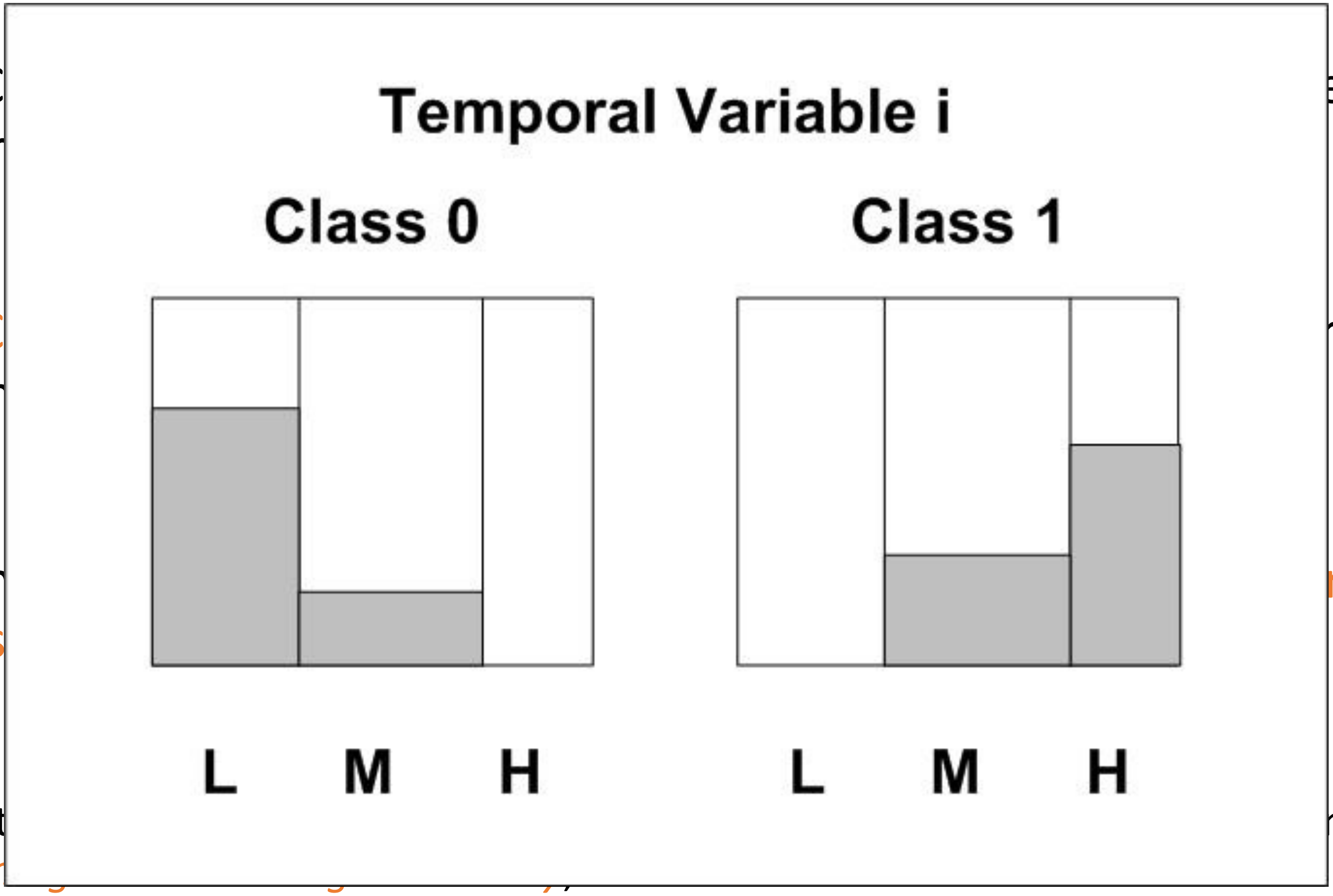
- Symbolic Aggregate approXimation Based on *PAA,* Piecewise Aggregate Approximation, A time series segmenting algorithm (Keogh et al.,2003).



- TD4C

TD4C - Temporal Discretization for Classification

- TD4C … e expected
  accur…
- TD4C … riable into
  mear…
- Havir … requent
  TIRPs…



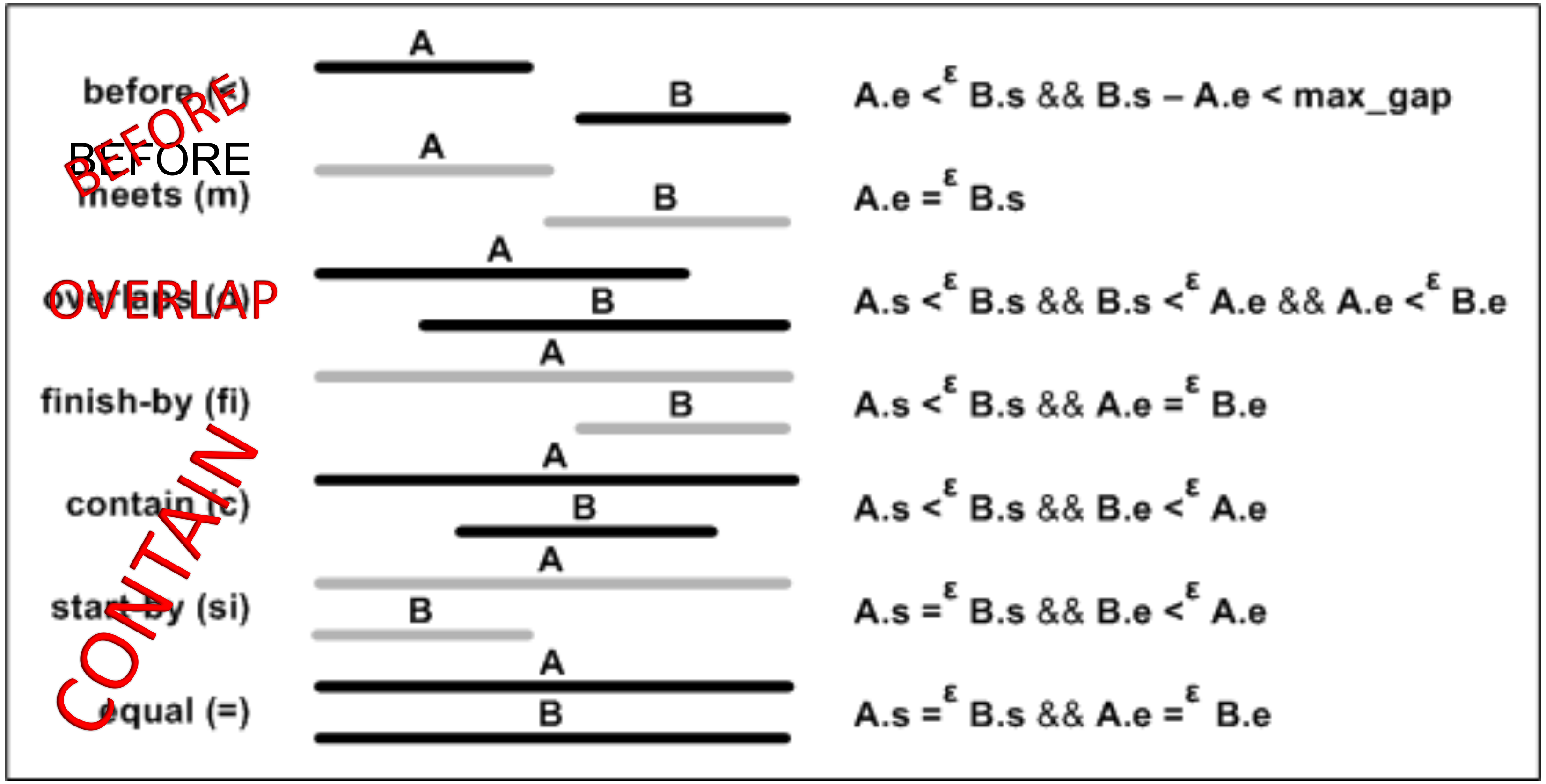
Moskovit… me series,
Data Mir…

TD4C Formulation

- Given C = $\{c_1, c_2, .. c_n\}$ classes, E entities divided into $\{E_1, E_2, .. E_c\}$ sets of entities per class and T = $\{t_1, t_2, .. t_m\}$ temporal variables, and A – a TD4C abstraction method.
- The problem is to find the set of cutoffs for each temporal variable t_i that increases the difference in the dominant states in each class.
- Thus, we want to measure the distribution of the states in each class entities, and to measure when these are most different.
- For that three measure were determined:
  - Entropy $$E(c) = -\sum_{i=1}^{k} p_i \cdot \log(p_i) \qquad D = \sum_{i=1}^{c} \sum_{j=i+1}^{c} |E(c_i) - E(c_j)|$$
  - Cosinus $$similarity(v,u) = \frac{v \cdot u}{|v||u|} \qquad D = \sum_{i=1}^{c} \sum_{j=1}^{c} similarity(c_i, c_j)$$
  - Kullback-Leibler $$KL(P,Q) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) \qquad D = \sum_{i=1}^{c} \sum_{j=i+1}^{c} KL(c_i, c_j)$$

Generalization of Allen's Temporal Relations

- KarmaLegoS enables to mine TIRPs with 7 (Allen's original) or 3 more general temporal relations



TIRPs as Features

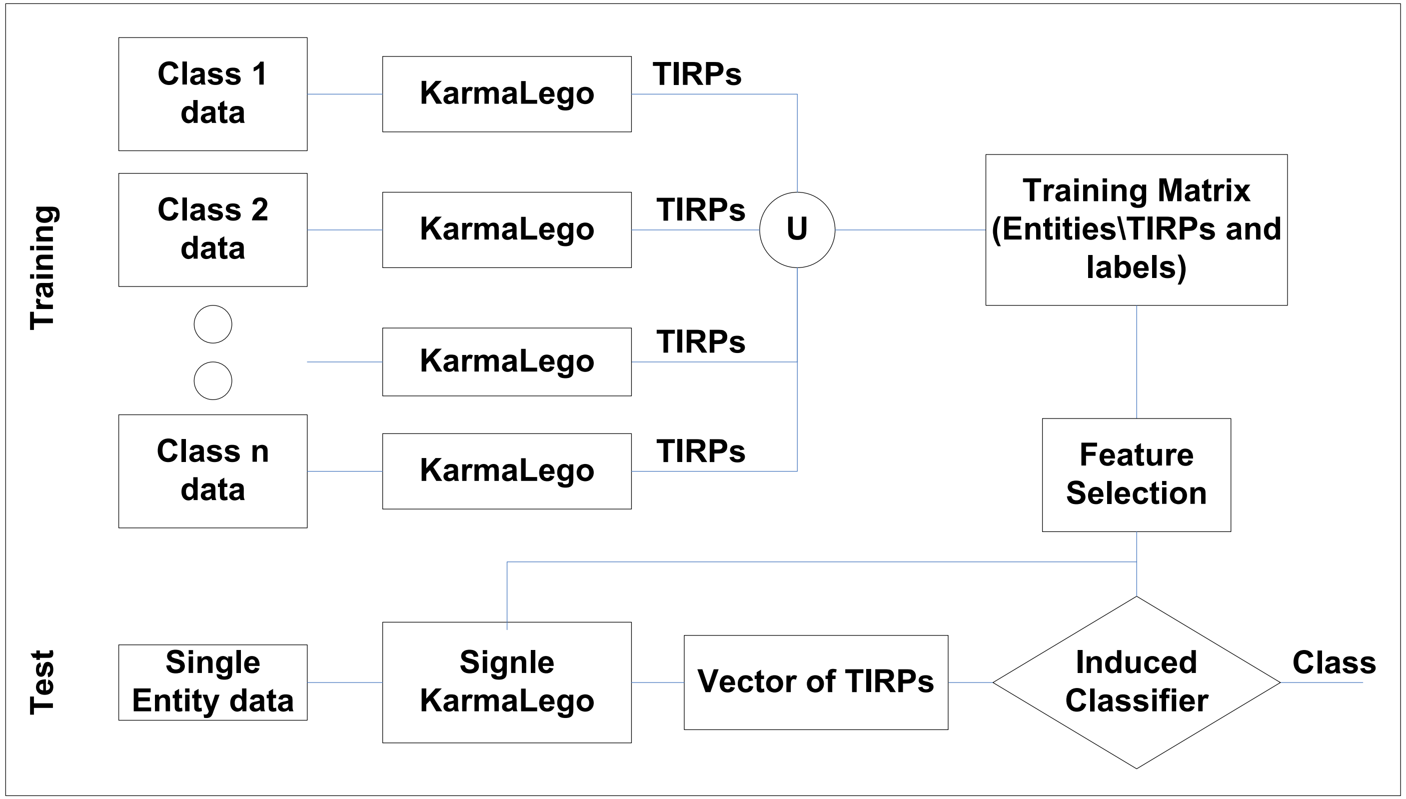
Horizontal Binary Mean Duration (Support)

entity	$Tirp_1$	$Tirp_2$	$Tirp_3$	$Tirp_4$	…	..	$Tirp_n$	Class
P_1	0	2.3	3.4	1.2	0	0	0	0
P_2	0	0	1.23	3.56	2.54	1.23	3.3	1
P_3	0	0	2.56	3.34	1	0	0	0
P_4	1.6	3.34	1.45	2.7	0	0	0	0
P_5	2.2	2.34	0	0	0	0	2.4	1
P_6	1.2	2.5	0	1.56	0	0	1.34	1
P_m	1.8	0	0	2.23	0	0	0	1

Datasets

- ICU Dataset - 645 patients who underwent cardiac surgery at the AMC in Amsterdam (2002-2004). Includes over 12 hours of *High* and *Low frequency*.
  196 patients were mechanically ventilated for more than 24 hours (70%), and the rest were 449.
- Diabetes- Contains 2038 diabetic patients data along 5 years (2002-2007) from Israeli HMO, measured monthly HbA1c, Glucose, Cholesterol values and medication purchased.
  992 males and 1012 females, having a quite balanced (~50%) dataset.
- Hepatitis – Laboratory measurements of Hepatitis B and C patients, admitted in Japan. Eleven temporal variables, having the top vertical support.
  204 Hepatitis B patients and 294 Hepatitis C patients (~60%).

KarmaLegoS Evaluation Setup

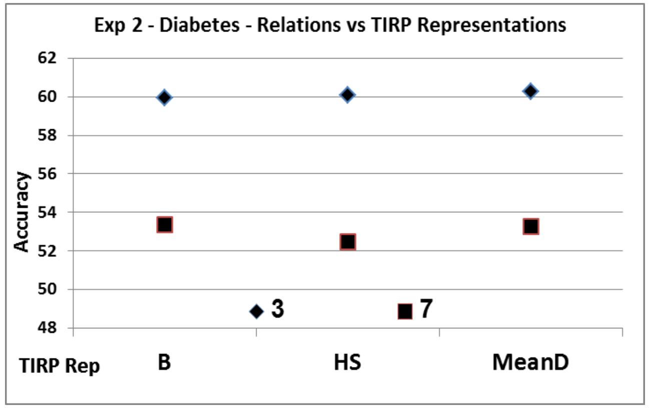


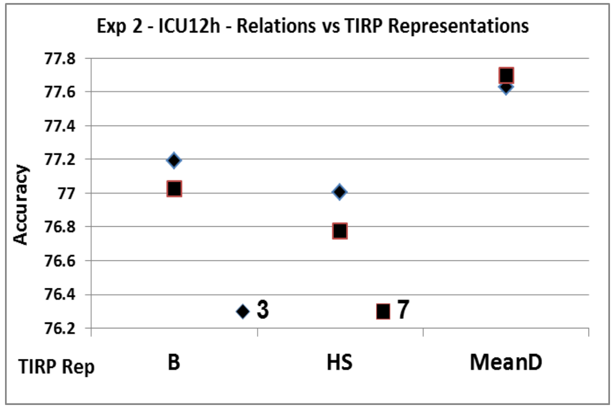
KarmaLegoS - Parameters

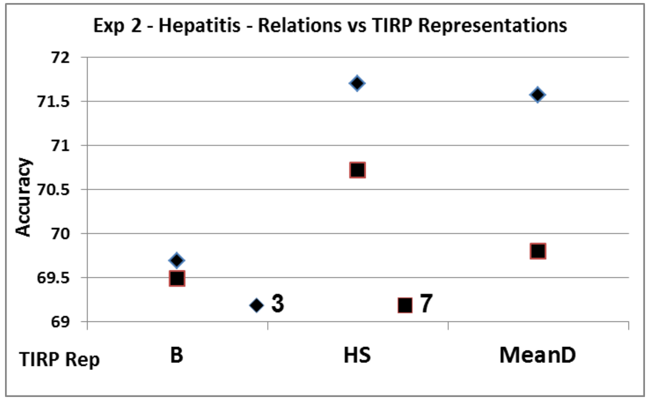
- To evaluate KarmaLegoS 5 experiments were designed for the various parameters in the KarmaLegoS framework, including:
  - Abstraction method (KB, EWD, SAX, TD4Cs)
  - Number of bins (3, 4)
  - Temporal relations set (3, 7)
  - TIRP Representation (Binary, HS, MeanD)

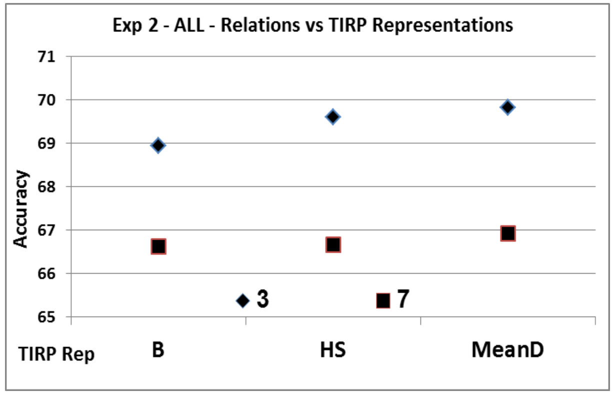
- For that experiments we designed and ran on three real datasets using 10 fold cross validation and RandomForest, compared according to the Accuracy measure.

Temporal Relations and TIRP representation



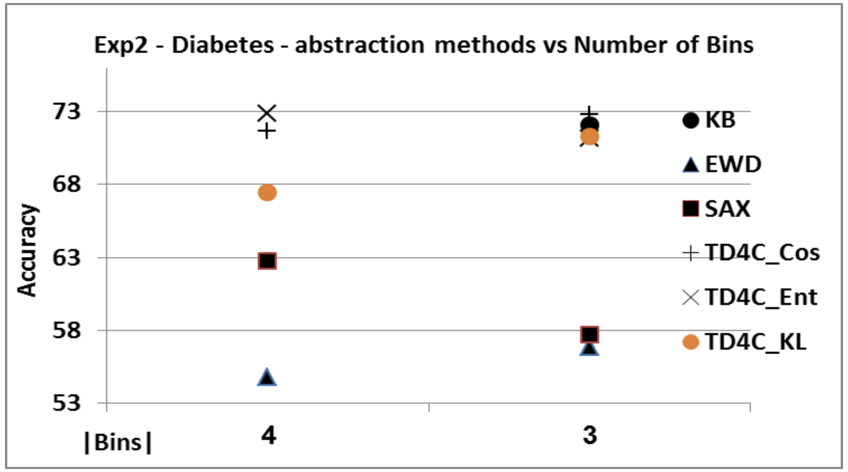


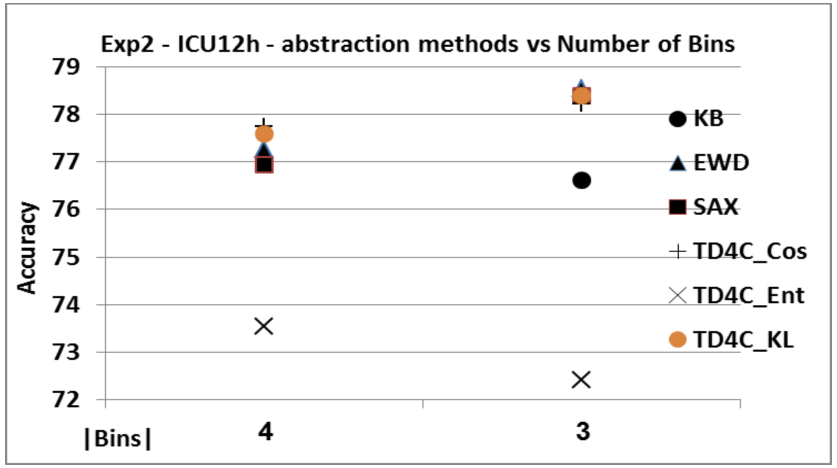


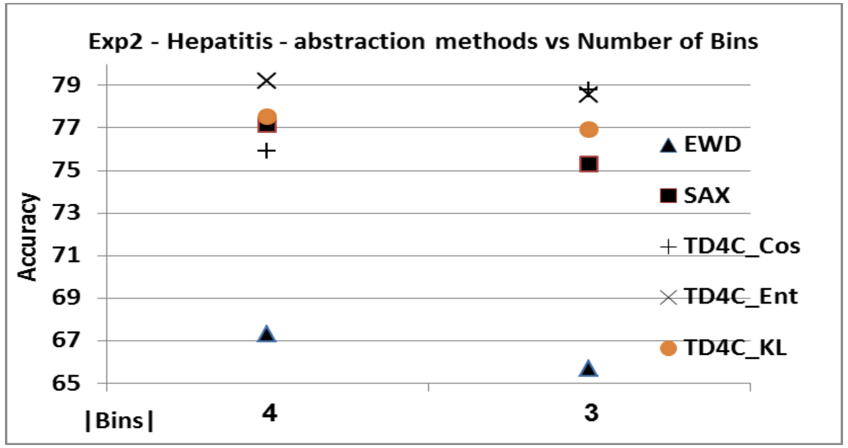


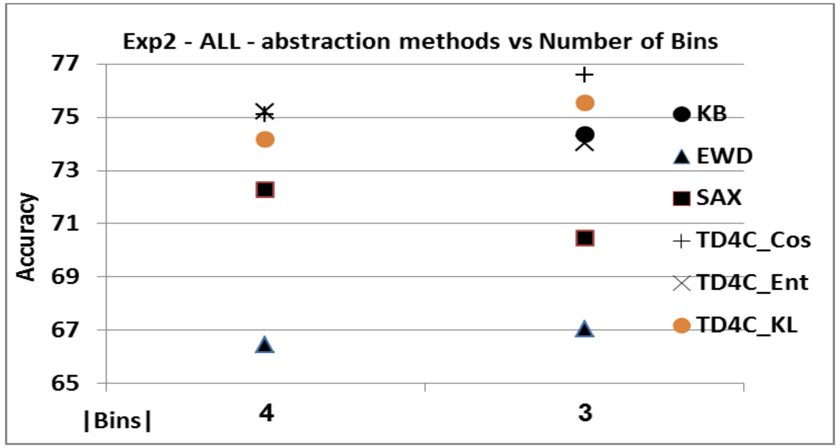
Abstraction Methods vs Bins Number

ε=0, 3 temporal relations, MeanD and NoFS

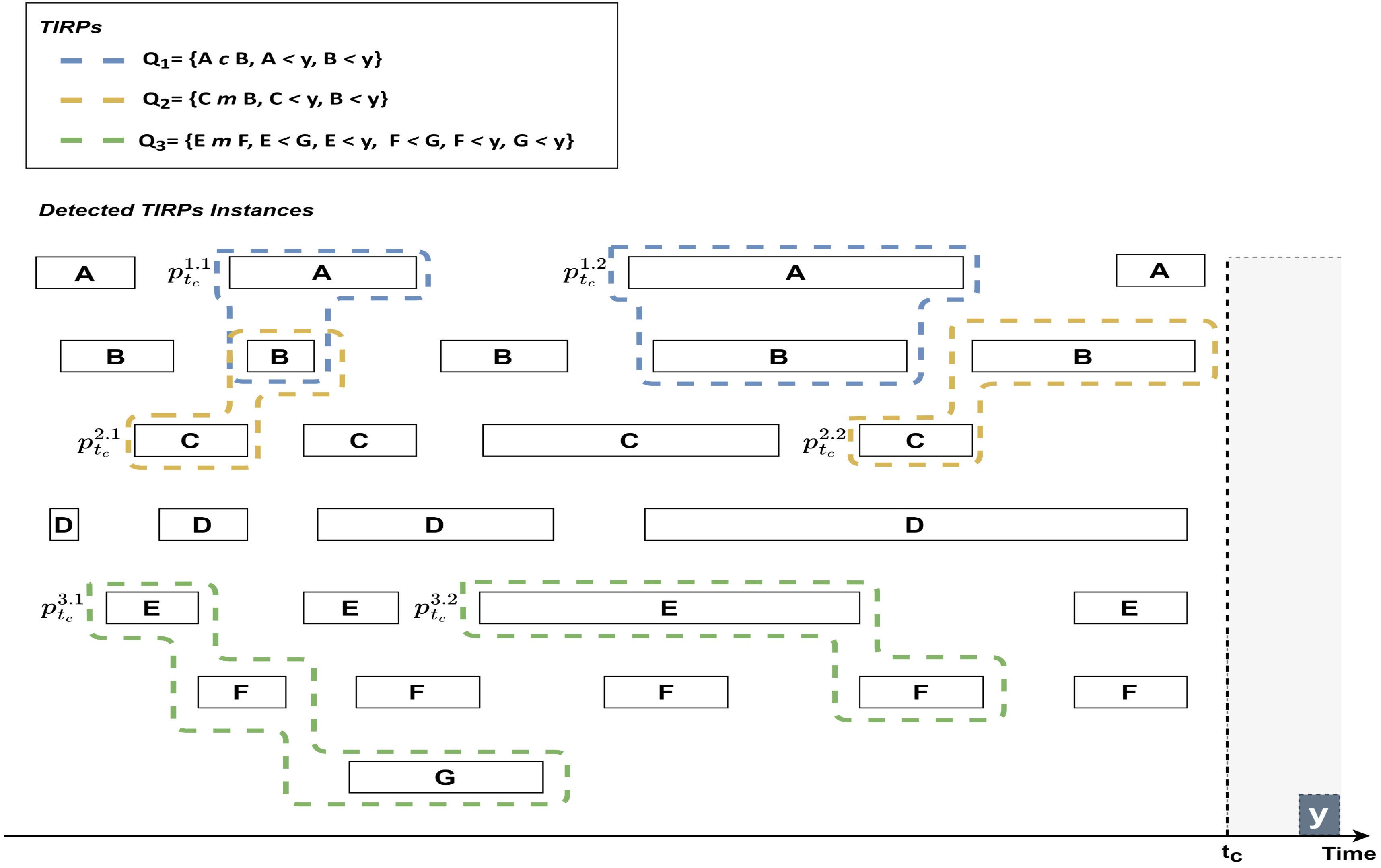








Continuous event's prediction via TIRPs



Continuous Prediction of TIRP's Completion

At any timepoint (e.g., t_c^1, t_c^2, t_c^3, and t_c^4), we aim to continuously estimate the probability of the TIRP's completion.

Nevo Itzhak, Szymon Jaroszewicz, Robert Moskovitch, Continuously Predicting a Time Intervals Based Pattern Completion Towards Event Prediction, PAKDD, **Osaka, Japan, 2023**.

Segmented CPM (SCPM)

Q – TIRP of interest

p_{t_c} - denote a prefix representing the observed part of Q at t_c

s_{t_c} - denote the remaining part of Q at t_c that is expected to occur.

Q:	B	C
A	o	<
B		<

$$Pr(Q|t_c) = Pr(s_{t_c}|p_{t_c}) = \frac{Pr(p_{t_c}|s_{t_c})Pr(s_{t_c})}{Pr(p_{t_c})} = \frac{\Pr(p_{t_c}, s_{t_c})}{\Pr(p_{t_c})} = \frac{\Pr(Q)}{\Pr(p_{t_c})} = \frac{\Pr(\{A\ o\ B, A < C, B < C\})}{\Pr(\{A\ o\ B\})}.$$

Q – TIRP of interest

p_{t_c} - denote a prefix representing the observed part of Q at t_c

s_{t_c} - denote the remaining part of Q at t_c that is expected to occur.

Unknown During The Prediction

$$Pr \ldots = \frac{\Pr(\{A\ o\ B, A < C, B < C\})}{\Pr(\{A\ o\ B\})}.$$

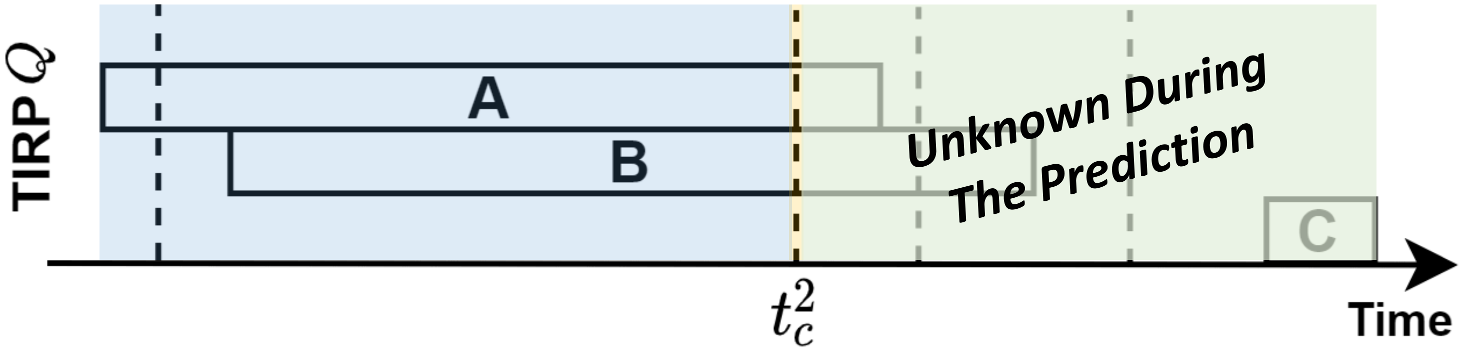
Segmented CPM (SCPM)

Q – TIRP of interest

p_{t_c} - denote a prefix representing the observed part of Q at t_c

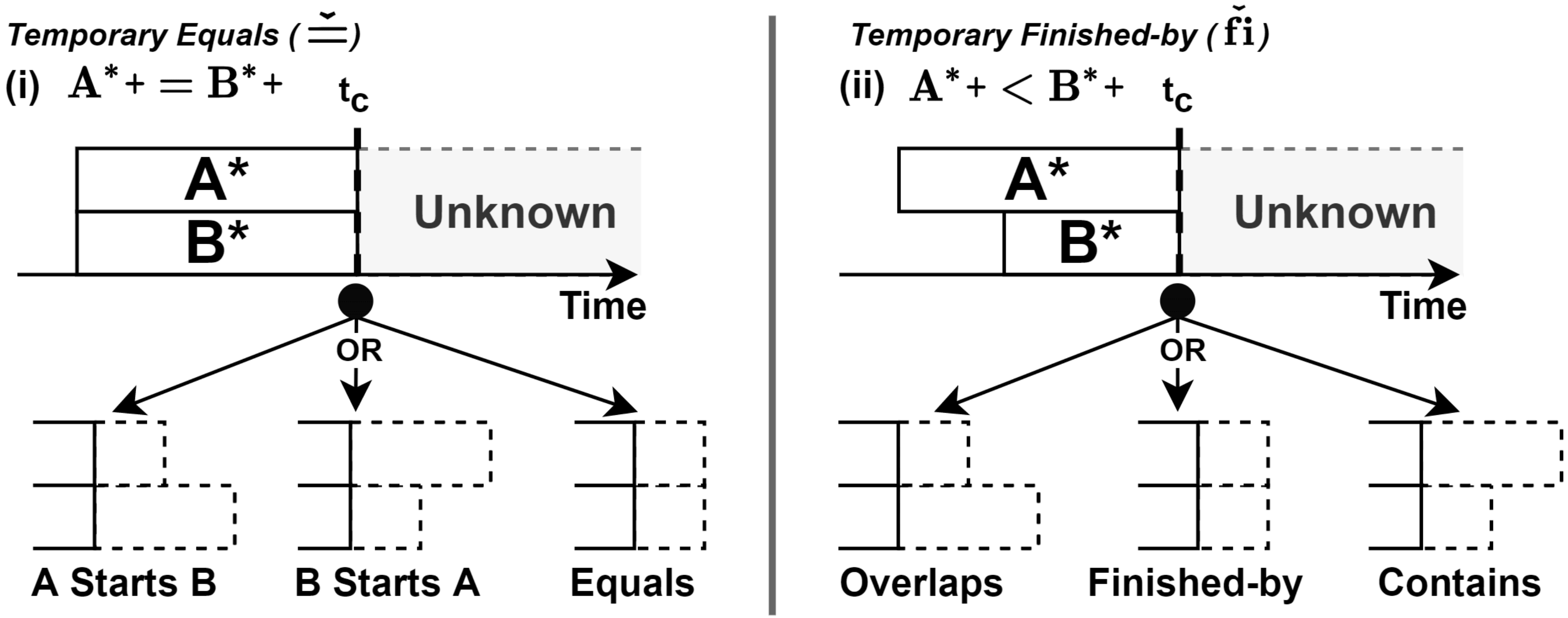
s_{t_c} - denote the remaining part of Q at t_c that is expected to occur.

$$Pr(Q|t_c) = Pr(s_{t_c}|p_{t_c}) = \frac{Pr(p_{t_c}|s_{t_c})Pr(s_{t_c})}{Pr(p_{t_c})} = \frac{\Pr(p_{t_c}, s_{t_c})}{\Pr(p_{t_c})} = \frac{\Pr(Q)}{\Pr(p_{t_c})} = \frac{\Pr(\{A\ o\ B, A < C, B < C\})}{\Pr(\_\_\_\_\_\_)}.$$



Introduction
Background
Objectives
Methods
Evaluation
Results
Summary

Unfinished Coinciding STIs

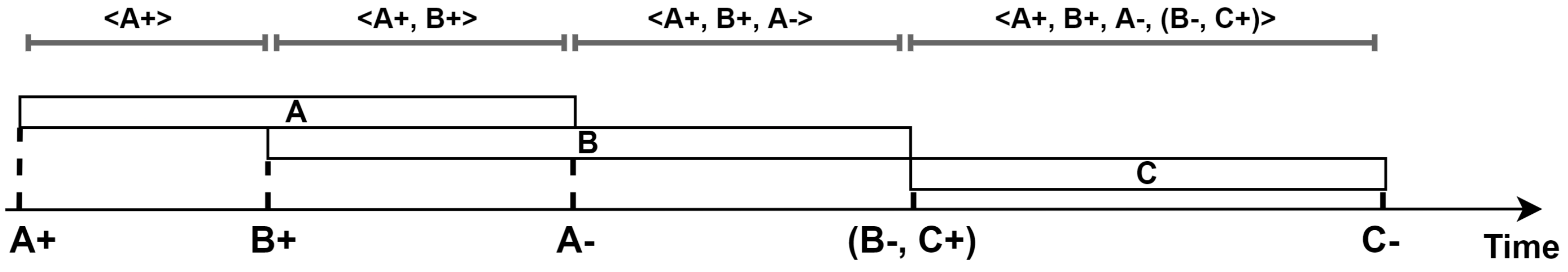


Possible evolving temporal relations, given that the start times of a pair of unfinished STIs are known.

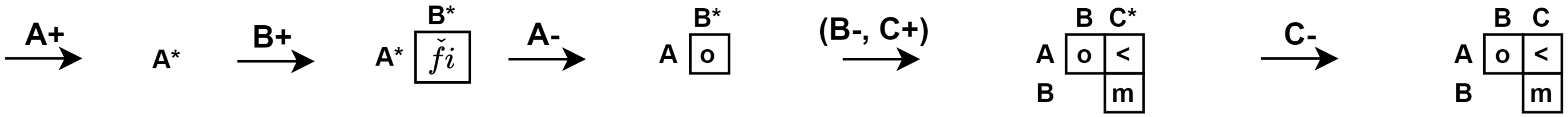
Introduction
Background
Objectives
Methods
Evaluation
Results
Summary

TIRP-Prefixes Representation

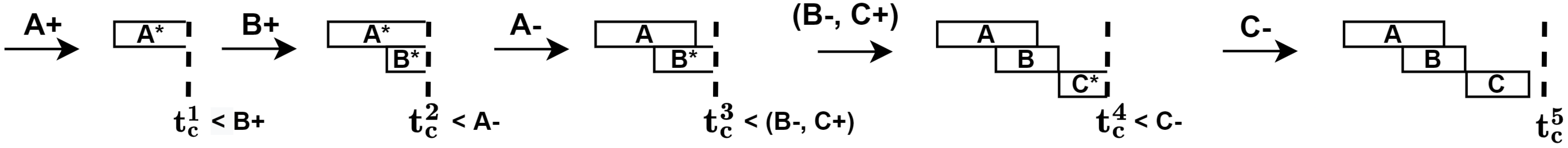
(i) TIRP Schematic Representation



(ii) TIRP-Prefixes Half Matrix Representation

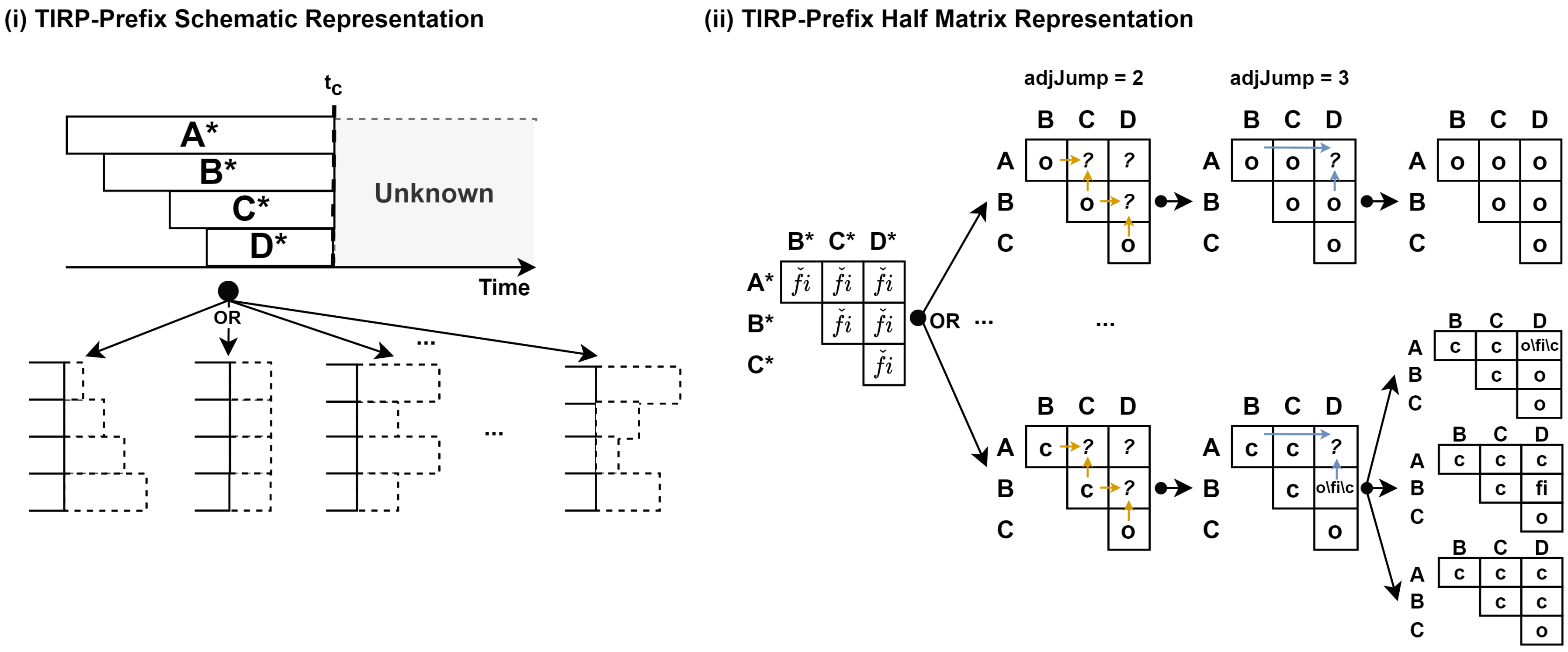


(iii) TIRP-Prefixes Schematic Representation



A TIRP of interest is divided into TIRP-prefixes that are part of the TIRP's evolving process

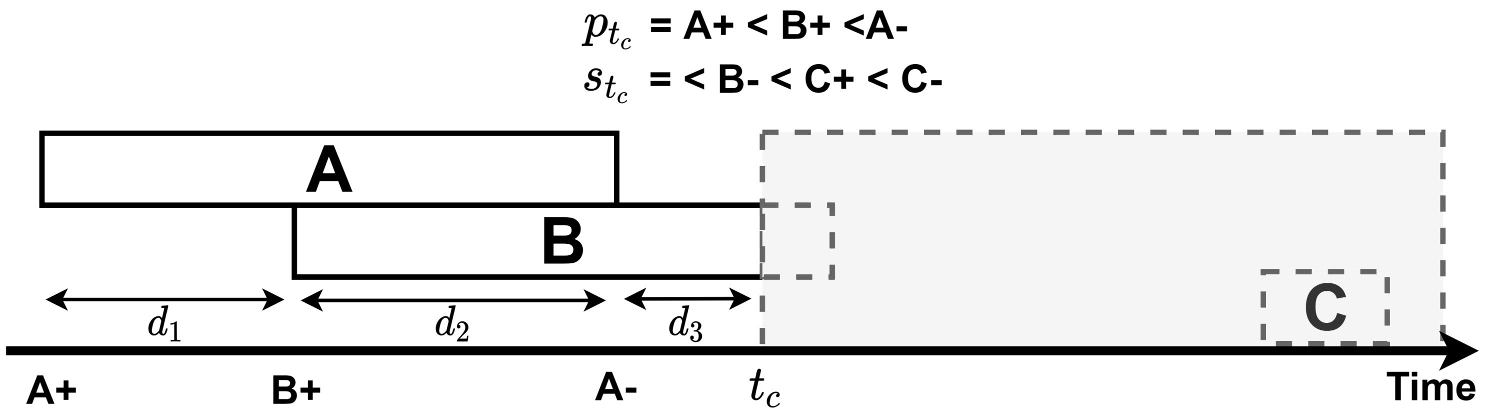
Unfinished Coinciding STIs



For four unfinished symbolic time intervals, a naive generation of all the suitable temporal relations among them requires generating up to $3^{(4^2-4)/2} = 3^6 = 729$ *patterns*

Introduction
Background
Objectives
Methods
Evaluation
Results
Summary

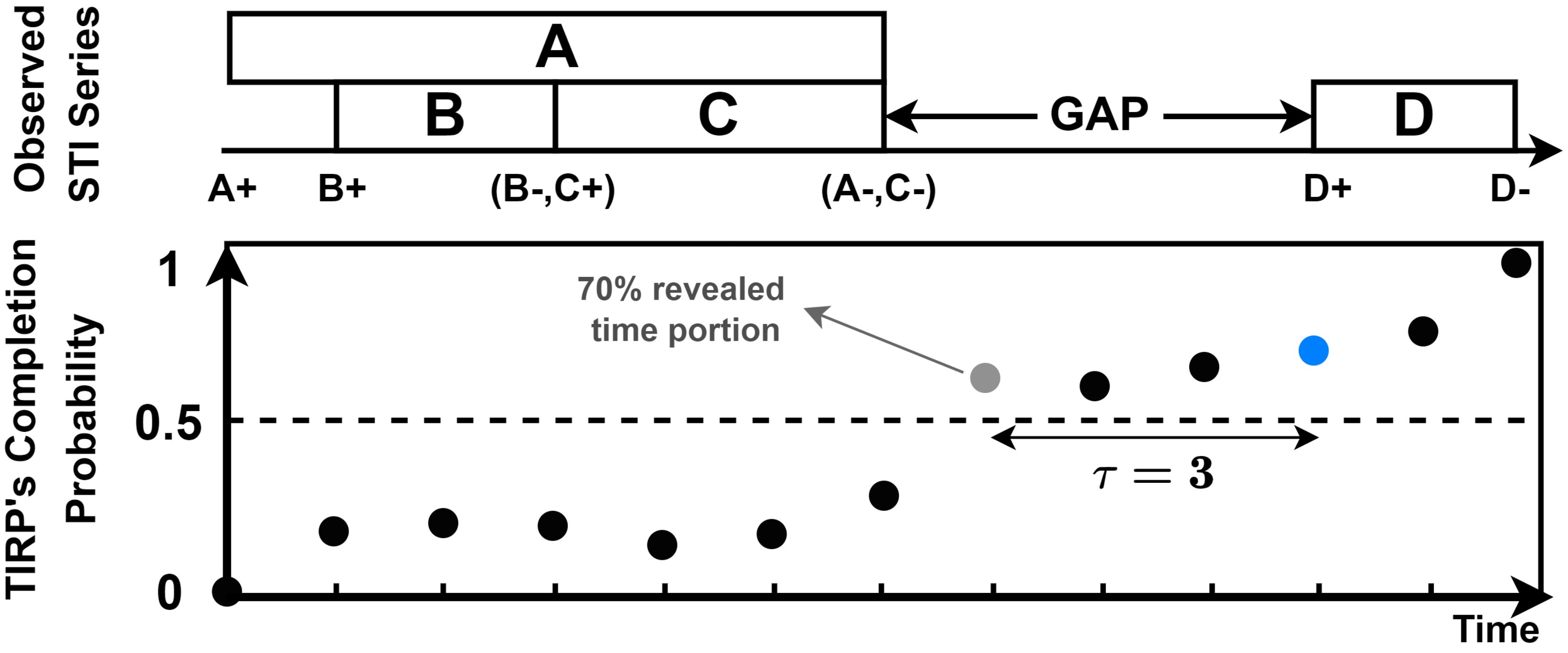
Machine-Learning-based CPM (CPML)



Time durations d1, d2, and d3 are based on tieps A+, B+, A-, and tc, which are used as features for the classifiers to perform the TIRP's completion prediction.

Instance	d1	d2	d3	...	dk	Class
1	1	2	12	...	0	**1**
2	9	12	1	...	0	**0**
...	...	...	...	...	..	...
n	1	2	11	...	0	**1**

Early Warning Strategies



An alert could be raised after the probability was consistently exceeded for some pre-defined decision time delay

Evaluation

Research Questions

A. Which CPM performs better, in terms of prediction performance and earliness, in predicting the completion of a TIRP?

B. Which value of τ performs best, in terms of prediction performance and earliness, in predicting the completion of a TIRP?

Evaluation

Datasets

Table 1. The evaluation datasets' parameters

Name	#Ent	#Var	#Timestamps	Granularity	#EntEvent	#TIRPs
CSP	329	13	720	minutes	115 (35%)	257
AHE	1,000	4	238	hours	500 (50%)	246
DBT	1,710	12	24	months	239 (14%)	256
EFIF	823	15	144	weeks	121 (15%)	529

We evaluated the proposed models using four real-life datasets:

1. CSP dataset – ICU patients, who underwent cardiac surgery
   Event: low cardiac index with values lower than 2.5 L/min/m

2. AHE dataset – ICU patients from multiple ICUs, in which half of the patients had acute hypertensive episodes
   Event: AHE onset

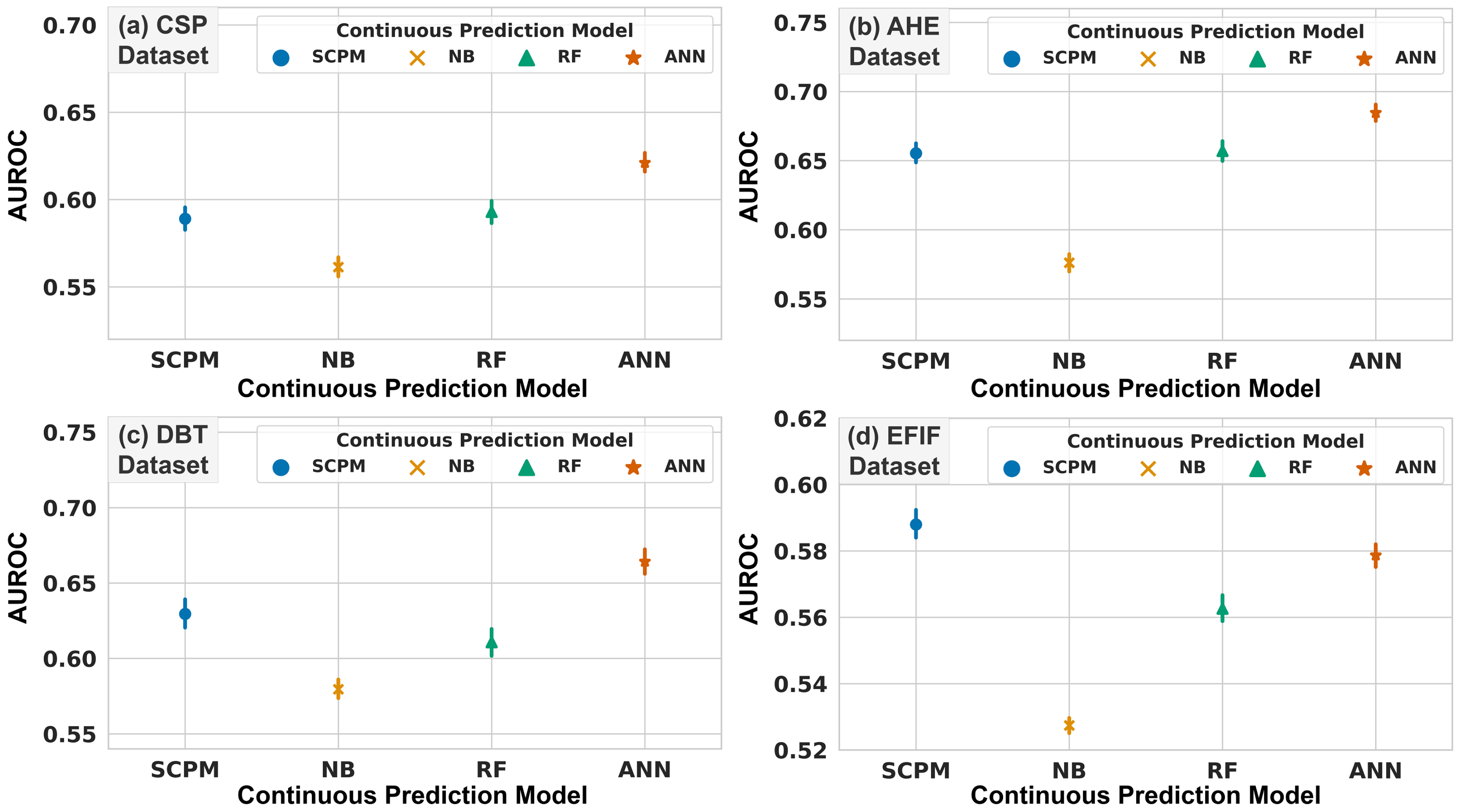
3. DBT dataset – Type II diabetes patients
   Event: high Hemoglobin A1C with values greater than 9%

4. EFIF dataset – elderly first fall of residents
   Event: First fall

Results

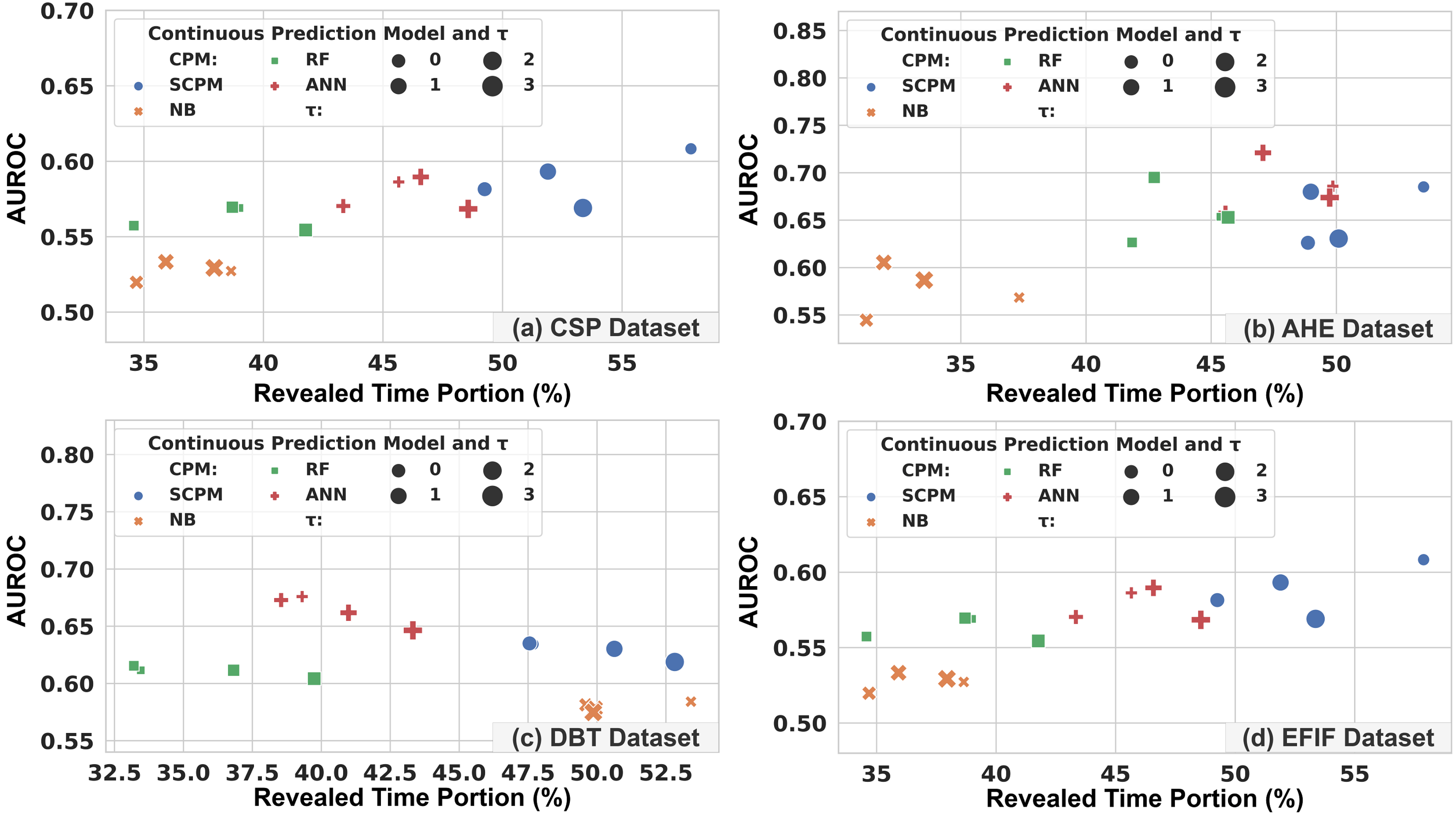
Preliminary Analysis

ANN performed better than the other models.



Results
Preliminary Analysis

More accurate models need more time to make decisions.
SCPM provided the latest predictions
NB and RF provided the earliest predictions.



Summary

1. Continuous TIRP's completion prediction

2. Uncertainty related to evolving temporal relations and our solution-- TIRP-prefix representation

3. CPML based on an ANN outperformed other models with 1.5% AUROC on average, while CPML based on NB or RF provided early TP predictions.

Thank You!

Prof. Robert Moskovitch
Head, Complex Data Analytics Lab
Software and Information Systems Engineering
Ben Gurion University
Israel

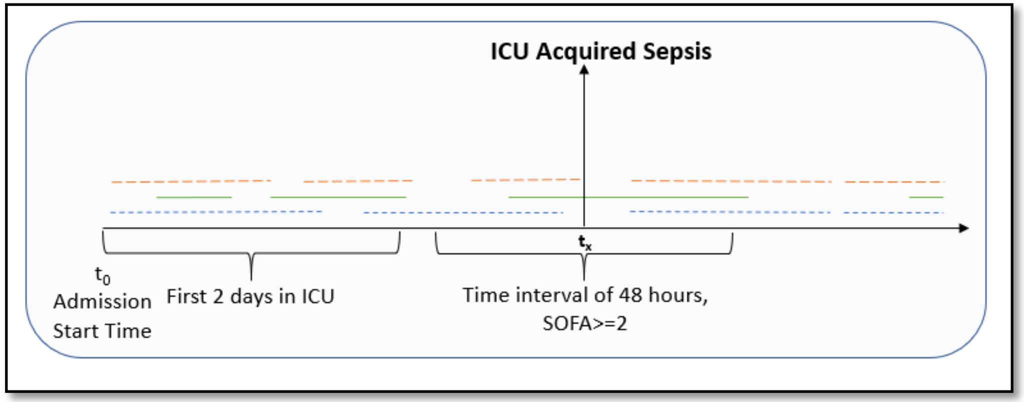
robertmo@bgu.ac.il

Relevant References

- Omer Harel, Robert Moskovitch, Complete Closed Time Intervals-Related Patterns Mining, The 35th AAAI Conference on Artificial Intelligence (AAAI 2021), Vancouver, Canada, 2021. Rank A*
- Nevo Itzhak, Szymon Jaroszewicz, Robert Moskovitch, Continuously Predicting a Time Intervals Based Pattern Completion Towards Event Prediction, PAKDD, Osaka, Japan, 2023.
- Nevo Itzhak, Maya Schvetz, Itay Pesach, Robert Moskovitch, Acute Hypertensive Episodes Prediction, *Artificial Intelligence in Medicine*, 2023.
- Maya Shwetz, Lior Fuchs, Victor Novack, Robert Moskovitch, Outcomes Prediction in Longitudinal Data: Study Designs Evaluation, use case in ICU Acquired Sepsis, *Journal of Biomedical Informatics*, 2021.

Sepsis Acquired During ICU Admission

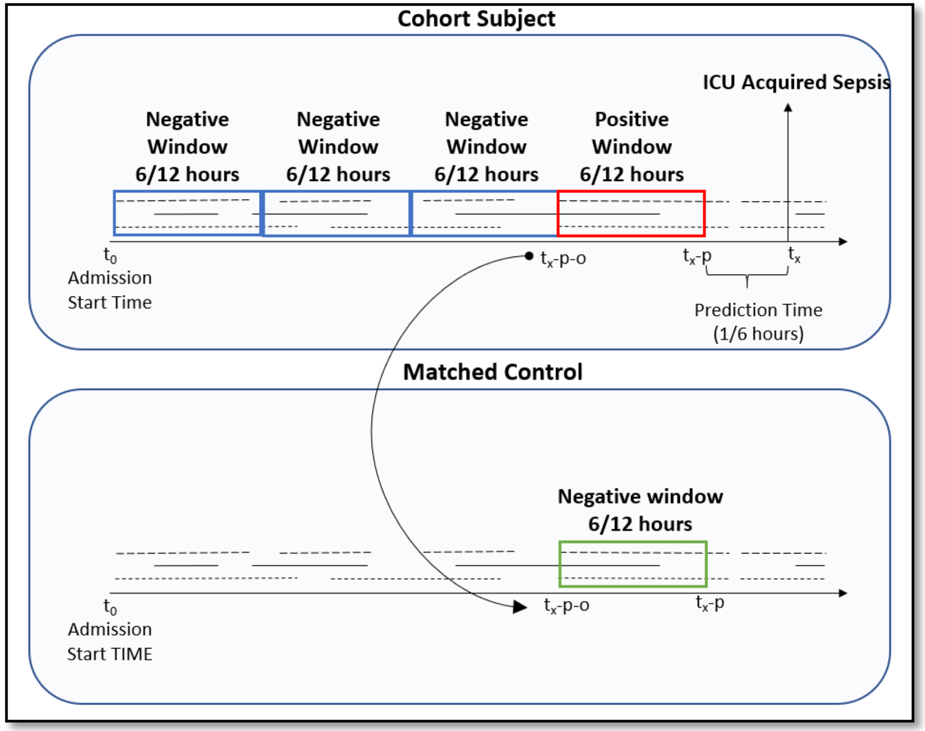
- Patient with suspected infection
  - Identified by administration of antibiotics and sampling of body fluid
- Patient with Sequential Organ Failure Assessment (SOFA) score >= 2
  - SOFA determines the extent of a person's organ rate of failure
- ICU-acquired Sepsis is defined as one that started at least 48 hours after admission



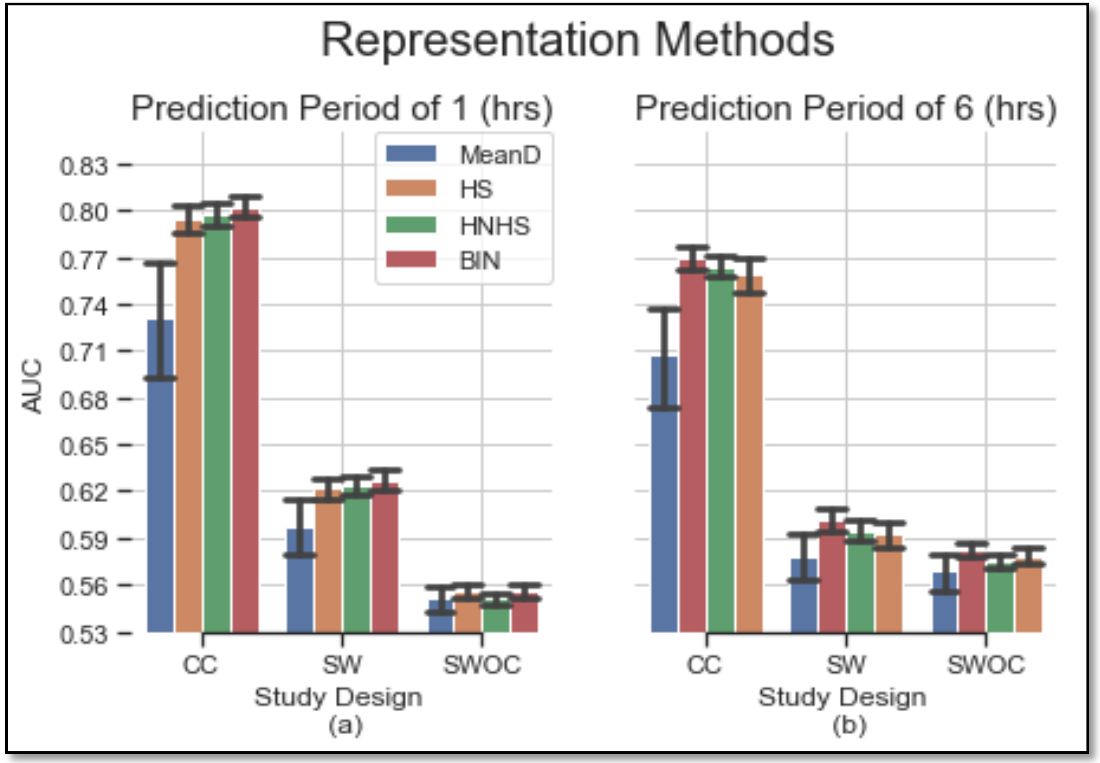
Maya Shwetz, Lior Fuchs, Victor Novack, Robert Moskovitch, Outcomes Prediction in Longitudinal Data: Study Designs Evaluation, use case in ICU Acquired Sepsis, *Journal of Biomedical Informatics*, 2021.

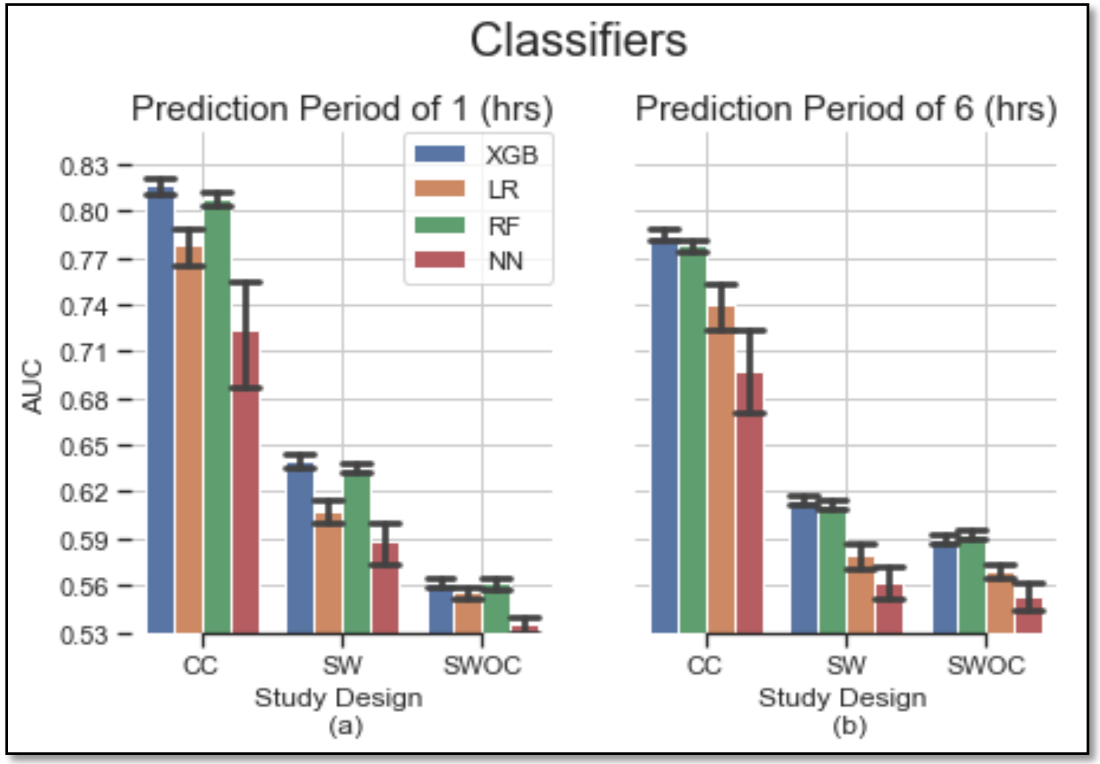
Exp 1- Predict the Occurrence of Sepsis Acquired in ICU Onset a Certain Time in Advance

- Case-Control design relative to the outcome
- Case-Crossover design with sliding window approach
- Sliding window (Case-Crossover-Control)

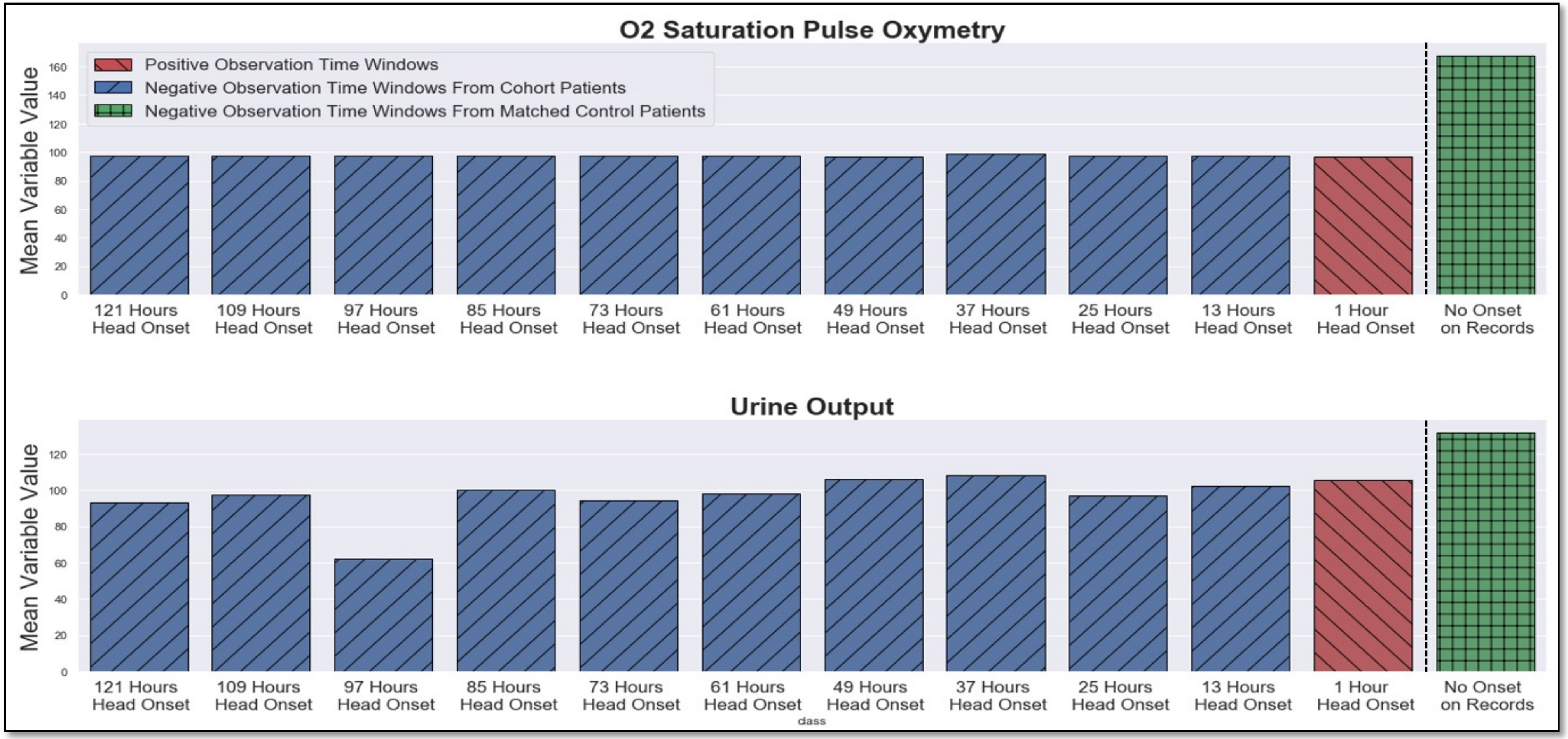


Representation Methods and Classifiers



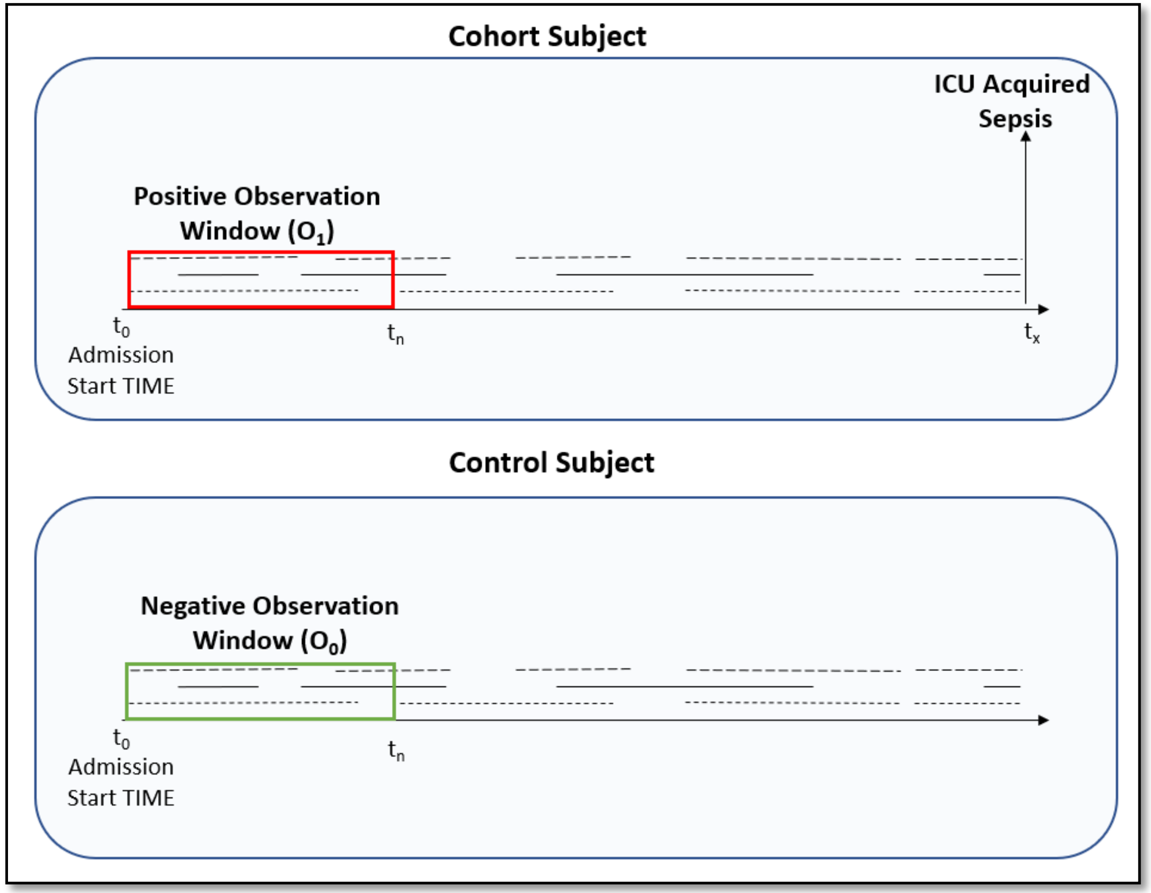


Low Performance Possible Reason

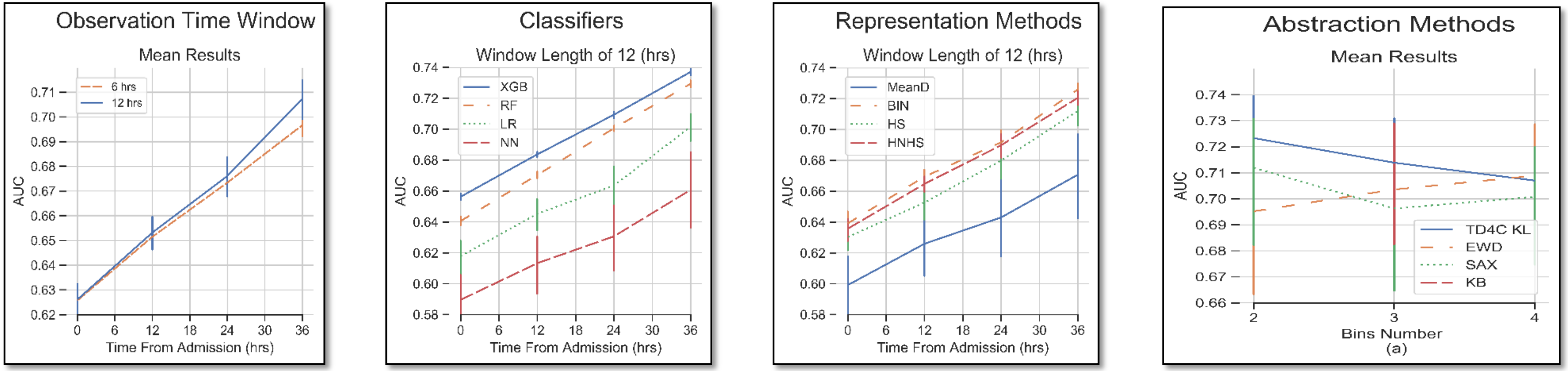


Exp 2 - Predict whether a patient will develop sepsis acquired in ICU at his stay

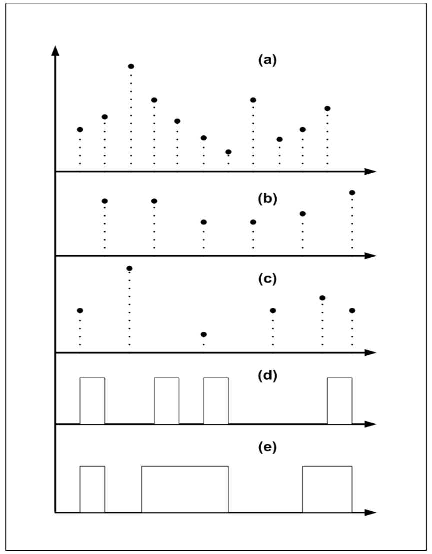
- Case-Control design **relative to prior event**, admission start time, to the outcome

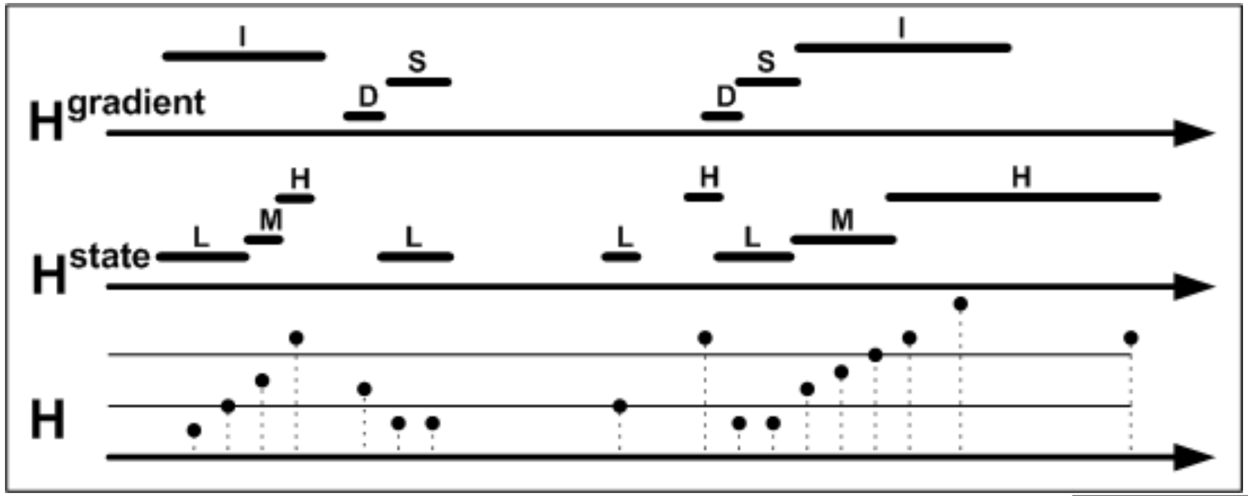


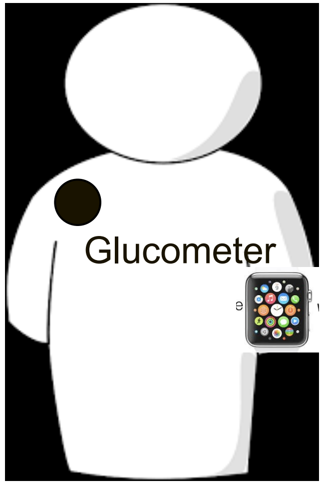
AUC results per hours from start ICU admission and observation period

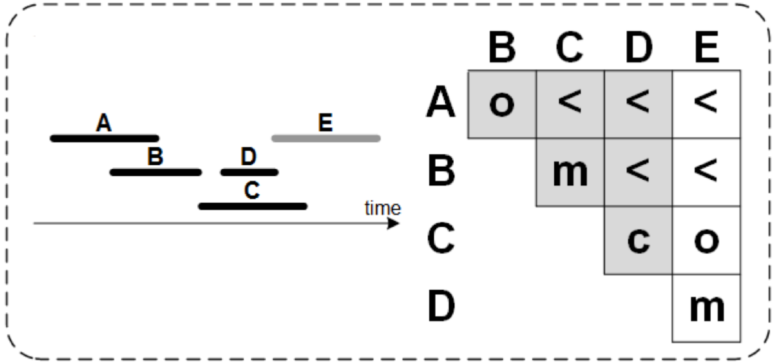


Temporal Data Analytics (IoT)









AI in Medicine

- Artificial Intelligence in Medicine (AIM) – one of the most important domains for human kind health
- Covers many research topics: text, image, temporal, and decision support
- In recent years with the breakthroughs in deep learning, there are tasks in medicine, such as image processing that benefited significantly. However, also longitudinal data, especially in Intense Care Units. AI was used in automating treatments, consisting on clinical guidelines, which were computerized.
- The speakers will cover ideally the following topics:
  Electronic Health Records Analytics
  Image Processing
  Longitudinal Data Analytics in medical data (outpatient, or inpatient data, or ICU).
  Generative AI in Medicine.

Order of preference	Country	Full name	Institution	Email address	Research field	Expertise/reason for choice
1	Israel	Dr. Mor Peleg	Haifa University	peleg.mor@gmail.com	AI for medicine	Chief Editor, Journal of Biomedical Informatics
2	Israel	Dr. Erez Shmueli	Tel Aviv University	shmueli@tau.ac.il	Breathing Complications Diagnosis, BIG DATA in Medicine	Had made meaningful contributions in that field
3	Israel	Dr. Eran Segal	Weitzman Institute		Diabetes data analytics	Well known studies around the world
1	Japan	Dr. Takanori Hasegawa	M&D Data Science Center, Tokyo Medical and Dental University	tk.hasegawa@gmail.com https://www.hase62.jp/	bioinformatics	Human genome, transcriptome, epigenome, metagenome analysis, and clinical data for personalized and preventive medicine through modelling, prediction and inference
2	Japan	Dr. Shuhei Kurita	RIKEN-AIP	Shuhei.kurita@riken.jp https://shuheikurita.github.io/	Vision and Language	Vision and Language are two most important and popular fields in generative AI.
3	Japan	Dr. Ryoma Bise	Kyushu University	bise@ait.kyushu-u.ac.jp	Computer Vision for bioinformatics	- Pathological Image Segmentation/Classification, Cell Tracking - Many papers published in top-tier conferences

Temporal Data Analytics

- While most of the works and methods developed in machine learning ignore time, dues to challenges and intention to simplify problems, longitudinal data is available more and more in recent years with the advancements of the Internet of Things.
- Thus, temporal data analytics is an important topic in many applicative domains, such as security in computers networks, intensive care unit and generally medicine, predictive maintenance and more.
- Other relevant tasks are forecasting in stocks, and other domains, in which sensory data are available. In recent decades there is a growing interest in discovery of frequent temporal patterns, such as sequential patterns, time intervals patterns and other, and their use for various tasks, such as classification, and more.

Order of preference	Country	Full name	Institution	Email address	Research field	Expertise/reason for choice
1	Israel	Mark Last	Ben Gurion University	mlast@bgu.ac.il	Data Mining	Had made several works in temporal data analysis
2	Israel	Assaf Shuster	Technion	assaf@cs.technion.ac.il	Data Mining	Is working in time series data
1	Japan	Koji Inoue	Kyoto University	Inoue.koji.3x@kyoto-ua.c.jp http://www.sap.ist.i.kyoto-u.ac.jp/members/inoue/	Spoken dialogue systems, Multimodal signal processing, Human-robot interaction	Highly evaluated young researcher in the field: http://www.sap.ist.i.kyoto-u.ac.jp/members/inoue/profile.html
2	Japan	Sho Yokoi	Tohoku University	yokoi@tohoku.ac.jp http://www.cl.ecei.tohoku.ac.jp/~yokoi/	Natural Language Processing	Active researcher who is also with RIKEN-AIP.