Pattern-Based Anomaly Detection in Mixed-Type Time Series

Vincent Vercruyssen[1], Len Feremans[2], Wannes Meert[1], Boris Cule[2], Bart Goethals[2,3]

[1] DTAI, KU Leuven, Belgium

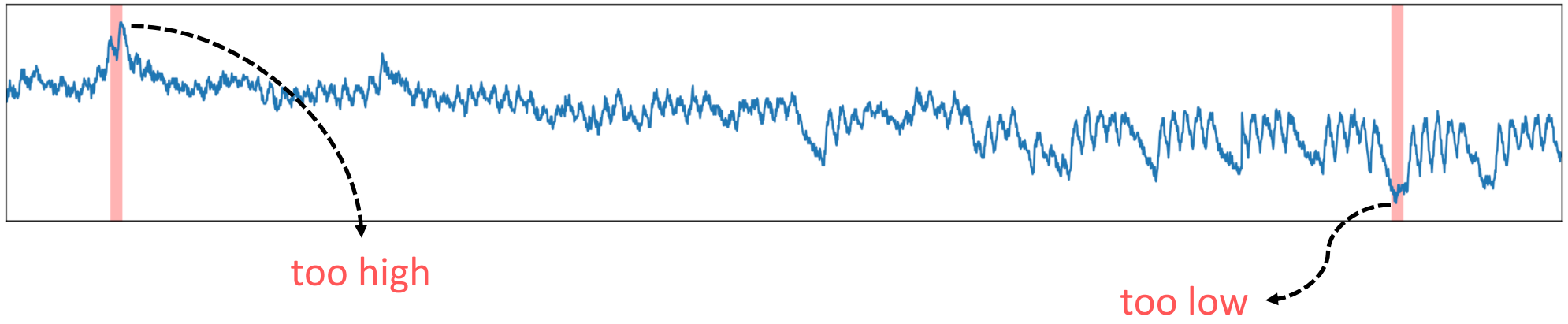
[2] Adrem Data Lab, University of Antwerp, Belgium

[3] Monash University, Melbourne Australia

Anomaly detection in univariate time series

univariate = a single continuous time series

Temperature readings in an office to detect abnormal temperature:

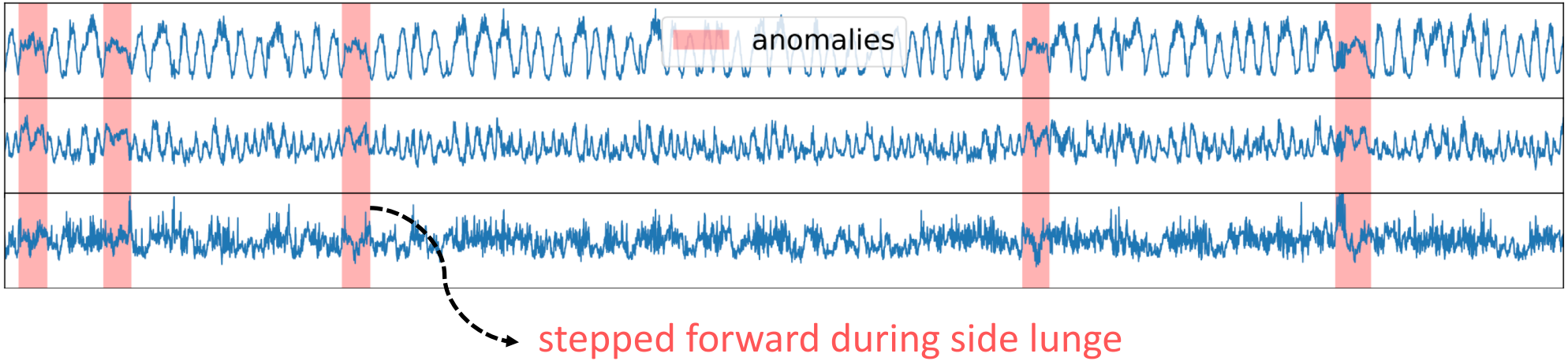


Existing algorithms that work out-of-the-box: FPOF, MATRIXPROFILE, PAV, MIFPOD

Anomaly detection in multivariate time series

multivariate = multiple continuous time series

Multiple sensors on an athlete to monitor incorrectly executed physical exercises:

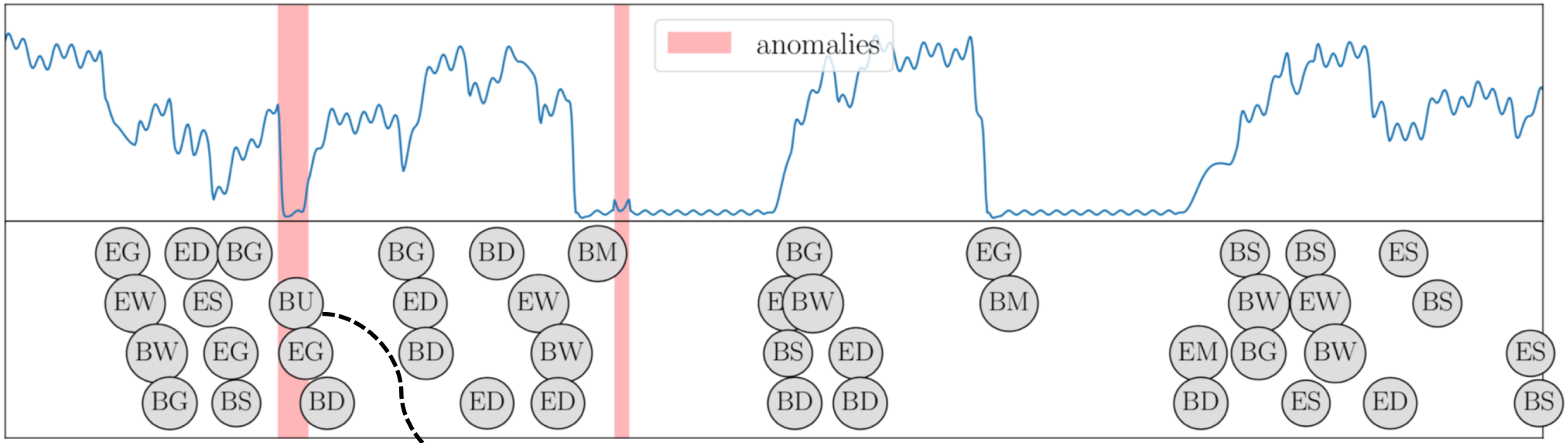


Existing algorithms that work after adaptation : FPOF, MATRIXPROFILE, PAV, MIFPOD

Anomaly detection in mixed-type time series

mixed-type = continuous time series & discrete event log

Power output of a small energy grid to detect unexpected energy production:



No existing algorithms that work for mixed-type data !

Real-world anomaly detection

GIVEN: univariate, multivariate, mixed type } time series data.

DO: detect periods of ***abnormal behaviour*** in the data by computing an anomaly score for each time series segment.

This is a ***challenging*** task:

1. How to deal with mixed-type time series?
2. How to assign an anomaly score to a segment based on its context?

Our approach: pattern-based anomaly detection

1. Extract patterns from the time series data
   … by mining frequent itemsets and sequential patterns

2. Map the time series segments to feature vectors
   … by computing the minimal distance between each segment and pattern

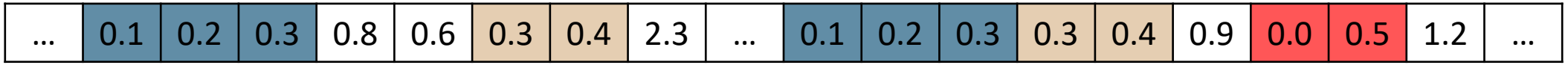
3. Apply anomaly detection on the feature-vector data
   … using the IsolationForest algorithm

1. Extract patterns from the time series data

Patterns are indicative of frequent behaviour

Anomalies are indicative of infrequent behaviour

Knowing the frequent patterns can help us detect anomalies



Frequent *normal* patterns **(0.1, 0.2, 0.3)** and **(0.3, 0.4)**

Infrequent *abnormal* pattern **(0.0, 0.5)**

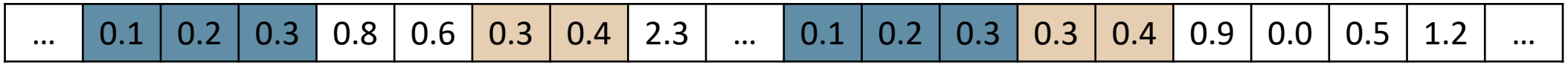
… by mining itemsets and sequential patterns

1. Frequent itemsets = patterns with unordered items

frequent itemsets **{a, b}** and **{c, d}** and …

Capture patterns where order of items is *not* important

2. Sequential patterns = patterns with ordered items

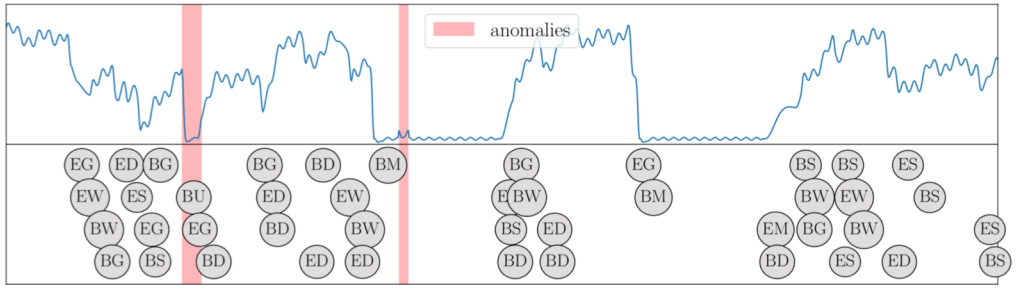


sequential patterns **(0.1, 0.2, 0.3)** and **(0.3, 0.4)** and …

Capture patterns where the order of items is important

2. Map the time series segments to feature vectors

Time series data



1. Extract

Frequent itemsets and sequential patterns

{a, b} {0.2, 0.4, 0.1} (c, d, e)

{a, d, f} (0.1, 0.2, 0.4)

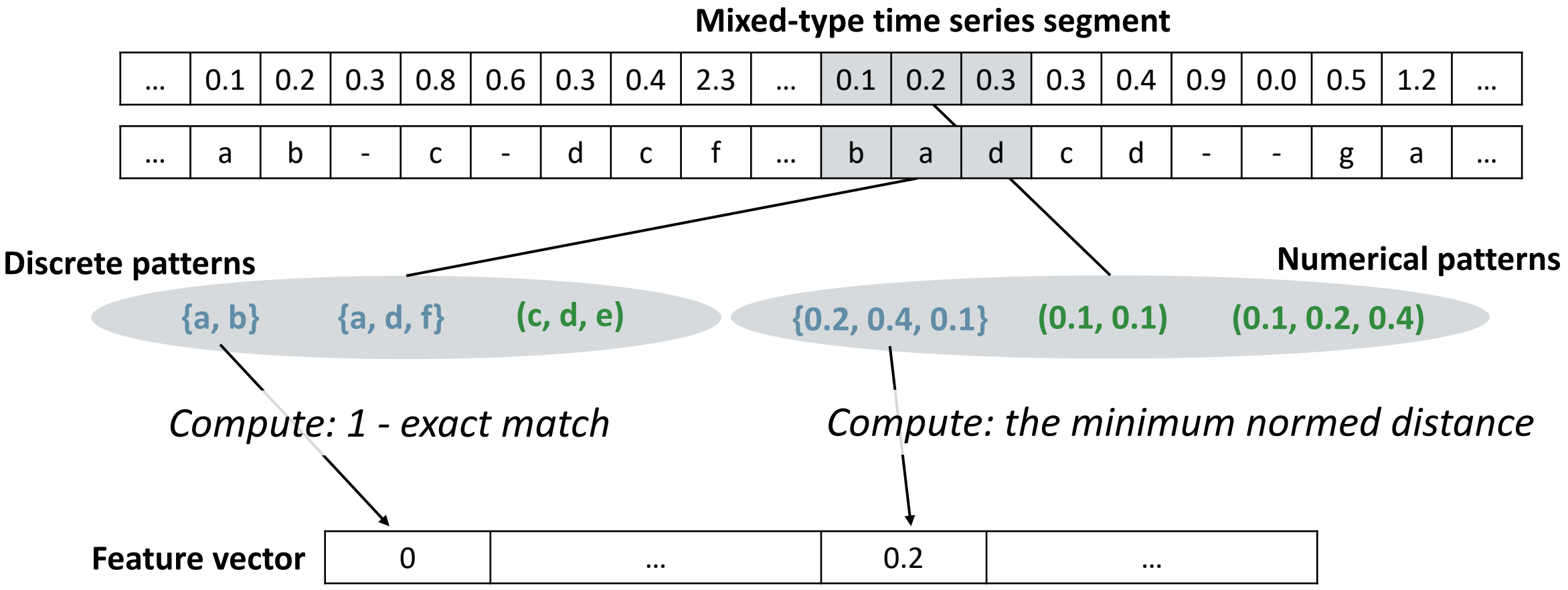
(0.1, 0.1)

2. Embed

Feature matrix

	$Pattern_1$	$Pattern_2$	$Pattern_3$	...	$Pattern_m$
$Segment_1$	0.1	0.6	0.1	...	0.1
$Segment_2$	0.2	0.4	0.2	...	0.5
$Segment_3$	...	...	...	...	...
$Segment_n$	0.3	0.5	0.3	...	0.5

… by computing the minimal distances to patterns

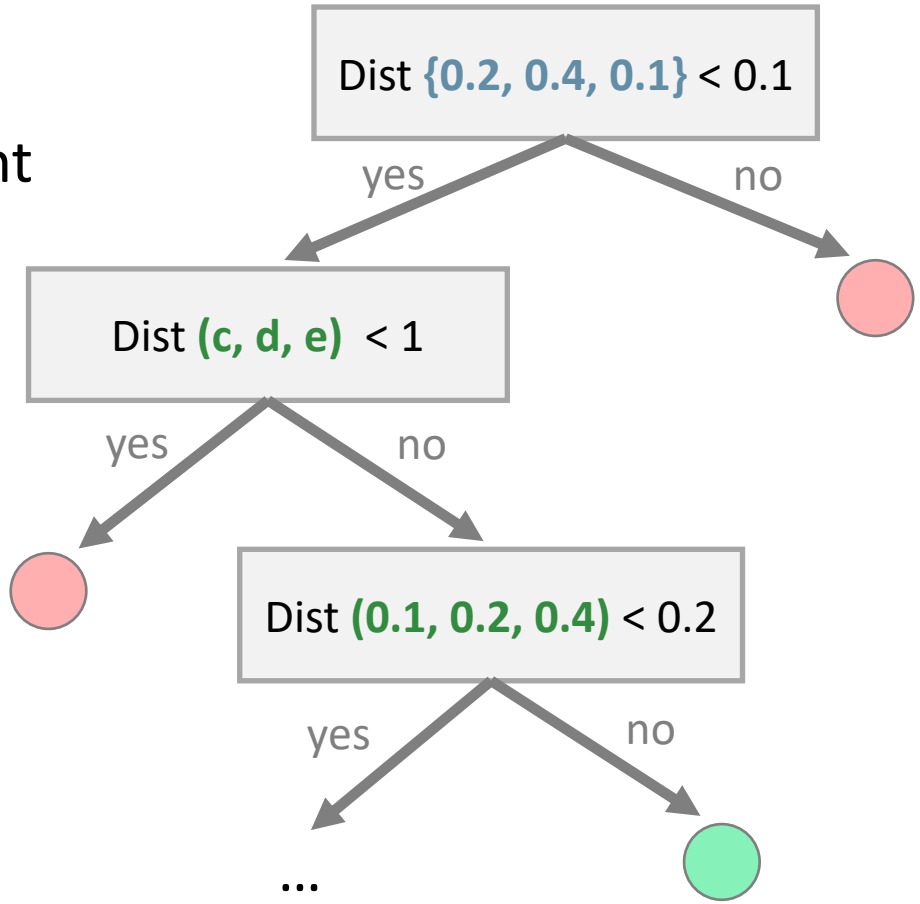


The distances are efficiently computed using dynamic programming!

Anomaly detection on the feature-vector data

ISOLATIONFOREST algorithm

- Computes an *isolation* score for each data point
  - Based on how difficult it is to separate a point
  - Averaged over an ensemble of trees
- Can deal with high-dimensional data
- Has low time complexity



What can we do with the framework?

Experimental evaluation

1. Real-world univariate time series data

2. Real-world multivariate time series data

3. Synthetic mixed-type time series data

Univariate time series benchmark

PBAD performs better than the baselines:

	AUROC				
dataset	MP	PAV	MIFPOD	FPOF	PBAD
Temp	0.240	0.590	0.997	**0.999**	0.998
Taxi	0.861	0.281	0.846	0.877	**0.879**
Latency	0.599	**0.608**	0.467	0.493	0.553
Water 1	0.656	0.482	0.514	0.825	**0.884**
Water 2	0.600	0.520	0.513	0.857	**0.945**
Water 3	0.536	0.457	0.544	**0.671**	0.605
Water 4	0.675	0.579	0.548	0.613	**0.721**
Water 5	0.444	0.581	0.455	0.790	**0.960**
Water 6	0.682	0.609	0.500	**0.874**	0.752
Average	0.588	0.523	0.598	0.778	**0.811**
Ranking	3.333	3.889	4.167	2	**1.611**

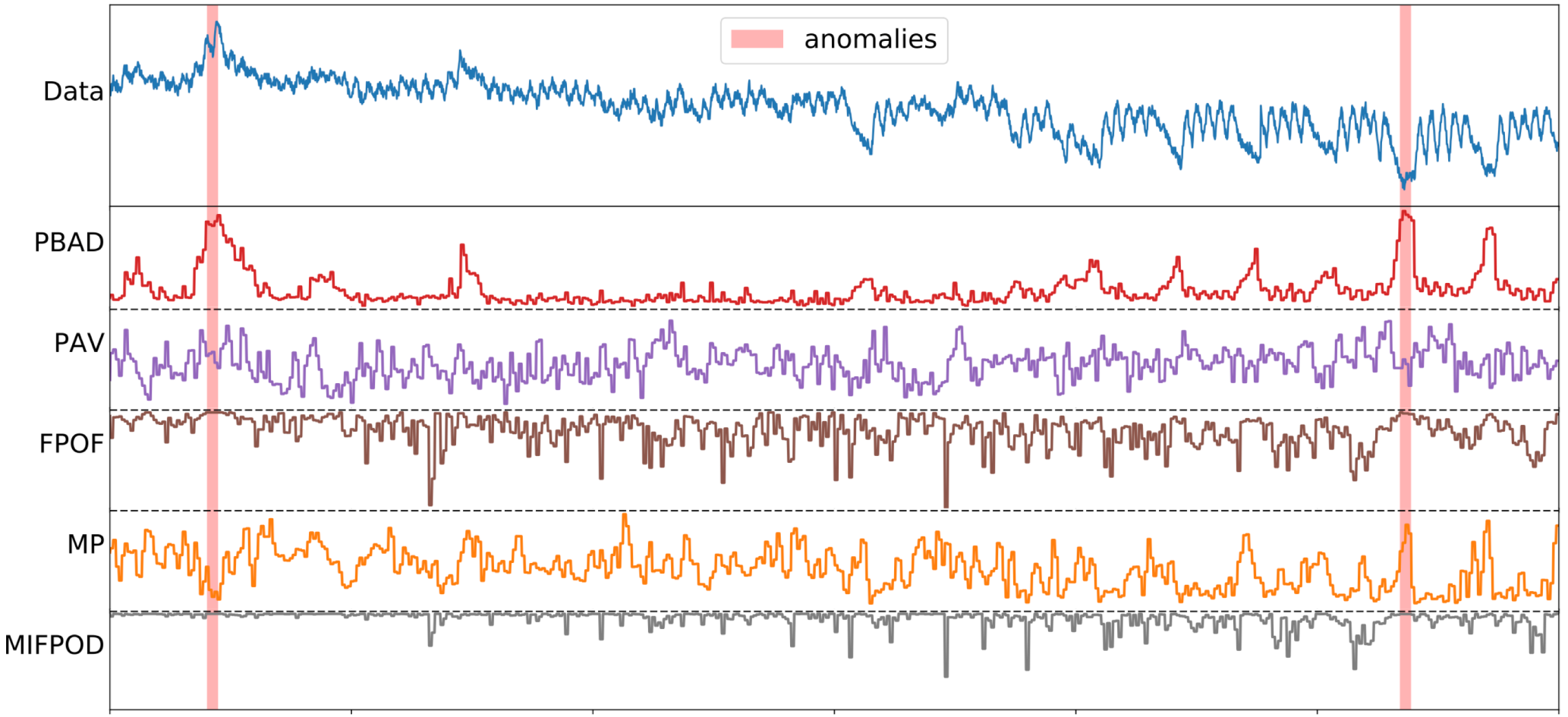
9 real-world datasets

Our method

- Vercruyssen et al. *Semi-supervised Anomaly Detection with an Application to Water Analytics.* IEEE ICDM 2018.

Univariate time series benchmark

Detecting abnormal temperature readings in an office:



Multivariate time series benchmark

PBAD performs better than the baselines:

4 real-world dataset problems

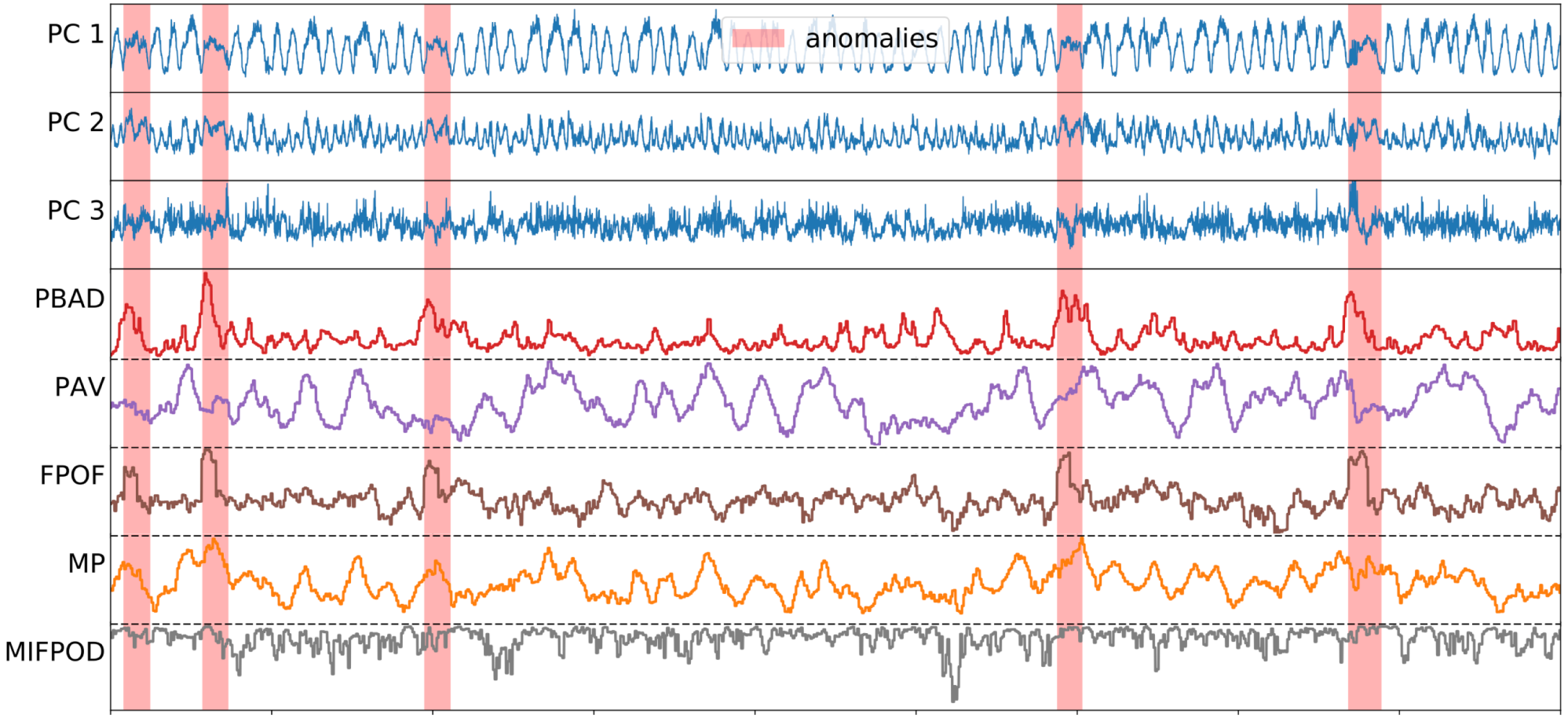
	AUROC				
Dataset	**MP**	**PAV**	**MIFPOD**	**FPOF**	**PBAD**
Lu+Si/Sq	0.472	0.571	0.819	0.966	**0.983**
Lu/Sq	0.604	0.671	0.775	**0.966**	0.940
Si/Lu	0.471	0.425	0.804	0.864	**0.907**
Sq/Si	0.484	0.504	0.482	0.903	**0.914**
Average	0.508	0.542	0.720	0.925	**0.936**
Ranking	4.5	4	3.5	1.75	**1.25**

Our method

- Decroos et al. *AMIE: Automatic Monitoring of Indoor Exercises.* ECMP PKDD 2018.

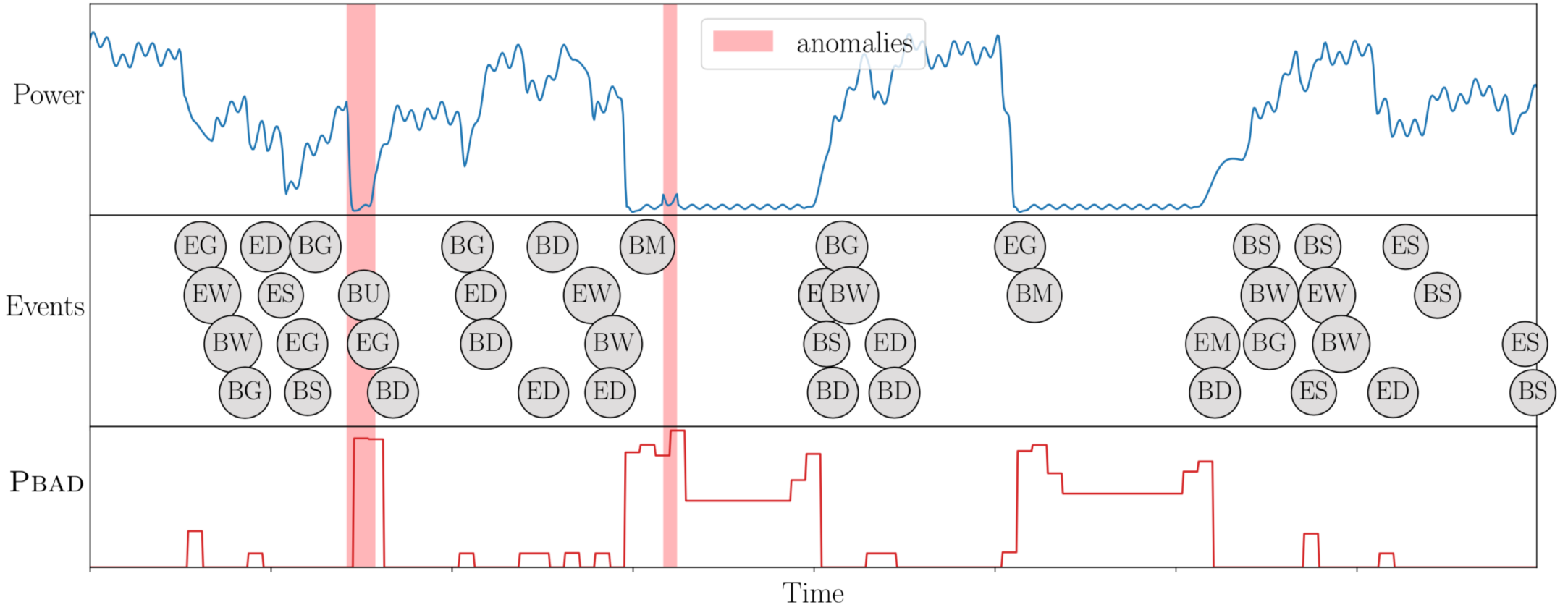
Multivariate time series benchmark

Detecting wrongly executed side lunge during an exercise session:



Mixed-type time series benchmark

Detecting abnormal power output in a small grid:



PBAD is the only method that tackles mixed-type time series data

Pattern-based anomaly detection

Vincent Vercruyssen, Len Feremans, Wannes Meert, Boris Cule, Bart Goethals

1. Discrete events are often recorded alongside continuous time series

2. Frequent patterns capture normal behaviour in time series data

3. By creating a pattern-based embedding of the time series data, we can use state-of-the-art anomaly detectors

CODE: *https://bitbucket.org/len_feremans/pbad*

WEBSITE: *https://people.cs.kuleuven.be/~vincent.vercruyssen/*