# Toward Interpretable Methods for Time Series Analytics

Félix Chavelli
felix.chavelli@inria.fr
Inria, École Normale Supérieure
Supervised by: Paul Boniol and Michael Thomazo

## ABSTRACT

Time series analytics is crucial for extracting meaningful patterns from ubiquitous time-varying data. While numerous methods exist, a significant gap persists in their evaluation and interpretation. For instance, current time series segmentation measures often fail to differentiate error types or offer clear interpretability. In this Ph.D., we tackle the problem of interpretability on time series analytical methods and evaluation measures. We first introduce two novel evaluation measures, WARI and SMS, designed to provide a more nuanced, insightful, and customizable assessment of segmentation quality. Beyond this initial focus on evaluation, the broader ambition of this PhD thesis is to develop interpretable methods and unified, meaningful representations for time series, potentially leveraging graph-based structures, to enable efficient and insightful execution of various downstream analytical tasks.

**VLDB Workshop Reference Format:**
Félix Chavelli. Toward Interpretable Methods for Time Series Analytics. VLDB 2025 Workshop: PhD Workshop.

## 1 INTRODUCTION

Time series data are prevalent across diverse domains, including environmental monitoring, energy management, and human activity recognition. The analysis of these temporal datasets supports various analytical tasks, such as classification [3], clustering [15], anomaly detection [16], motif discovery [18], and segmentation [8]. Time series segmentation, which involves identifying change points and underlying states, is crucial for discerning patterns and understanding process dynamics. Despite numerous algorithms from diverse methodological backgrounds—statistical [8], Markov models [14], auto-encoders [19], and symbolic representations [5]—evaluating their effectiveness remains challenging due to limitations in current evaluation and interpretation methods. The lack of robust and interpretable assessment frameworks impedes comprehensive performance understanding and analyses.

This Ph.D. research aims to develop interpretable methods for time series analytics. As an initial contribution, this paper addresses critical deficiencies in the evaluation of time series segmentation. Existing measures exhibit several limitations: change point-based metrics, while useful for localizing transitions, may not adequately reflect the overall quality of the identified segments; point-based measures, such as the Adjusted Rand Index (ARI), often treat all misclassified points uniformly, regardless of the error's nature or temporal context, thereby failing to differentiate between qualitatively distinct error types (e.g., minor boundary misalignments versus gross misclassification of entire segments); and current evaluation approaches generally lack mechanisms for categorizing error types, which limits interpretability.

To overcome these issues, we introduce two novel evaluation measures: **WARI** (**W**eighted **A**djusted **R**and **I**ndex) and **SMS** (**S**tate **M**atching **S**core). WARI extends the ARI by incorporating the temporal position of errors, penalizing errors differently based on their proximity to true segment boundaries. SMS provides a complementary approach by identifying, categorizing, and allowing differential weighting of four fundamental error types, thereby enhancing interpretability through a detailed performance breakdown. We describe in detail these measures, empirically demonstrates their advantages over existing metrics, and show how they offer a more nuanced and insightful assessment of segmentation quality. The findings reveal new perspectives on the performance of state-of-the-art methods. We conclude by outlining future research directions that build upon these contributions, aligning with the broader Ph.D. objective of advancing transparent and interpretable time series methods.

## 2 TOWARD INTERPRETABLE MEASURES

As an initial contribution of this Ph.D., we addresse critical deficiencies in the evaluation of time series segmentation. Therefore, this section delves into the specifics of time series segmentation, the challenges in its evaluation, and our proposed measures designed to offer more interpretable and nuanced assessments.

### 2.1 Background and Foundations

A **real-valued time series** of length $N$ and dimension $D$ is a time-ordered sequence denoted by $T = (t_1, \ldots, t_N)$, where each $t_i \in \mathbb{R}^D$ for $i = 1, \ldots, N$. We define a univariate time series as a time series with $D = 1$. Moreover, a subsequence of $T$ from index $i$ to $j$ (with $1 \leq i \leq j \leq N$) is denoted by $T_{[i,j]} = (t_i, t_{i+1}, \ldots, t_j)$.

A **state sequence** $S = (s_1, \ldots, s_N)$ associated with a time series $T = (t_1, \ldots, t_N)$ is a sequence of the same length, where each $s_i \in \mathcal{S}$ is a discrete label representing the latent state of the system at time step $i$, and $\mathcal{S}$ is a finite set of possible states. In a state sequence, $i$ (with $0 \leq i < N$) is a **change point** if $s_i \neq s_{i+1}$.

Time series segmentation divides a time series into meaningful, homogeneous segments. Two main approaches are **change point** and **state detection**, both aiming to capture behavioral shifts.

*Change Point Detection.* Change point detection identifies an increasing set of integers $\{c_1, \ldots, c_M\}$, where each change point $c_{1 \leq i \leq M}$ marks a transition between states. Existing change point detection methods include profile-based approaches like ClaSP [8],


Proceedings of the VLDB Endowment. ISSN 2150-8097.

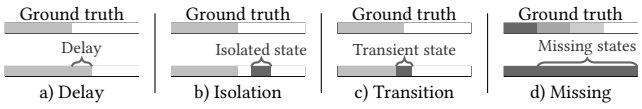


**Figure 1: Ground truth (top) with four error examples below: delay, isolation, transition, and missing.**

FLUSS [9], and ESPRESSO [7], statistical methods like PELT [11], and Bayesian techniques such as BOCD [2].

*State Detection.* State detection assumes an underlying sequence of latent states, outputting a predicted state sequence $P = (p_1, \ldots, p_N)$. Its goal is to identify recurring patterns by recognizing changes in the latent state. Approaches include encoder-based methods (e.g., Time2State [19], E2USD [13]), probabilistic models (e.g., HDP-HSMM [14]), and rule-based systems (e.g., PaTSS [5]).

Change point detection can be seen as a subproblem of state detection: change points partition the series, and clustering these segments assigns state labels (e.g., ClaSP [8] with kMeans clustering as in [13, 19]). Thus, state detection generalizes change point detection, and we will therefore focus on this more general task.

*2.1.1 Evaluating Time Series Segmentation: Typology of Errors and Desired Properties.* To evaluate segmentation quality against a ground truth, we define error types. Let $R = (r_1, \ldots, r_N)$ and $P = (p_1, \ldots, p_N)$ be the *real* and *predicted* state sequences, respectively, with states from a finite set $\mathcal{S}$. An *error block* is a maximal contiguous interval $[i, j]$ where $p_k = p_l \neq r_k$ for all $k, l \in [i, j]$. The *atomicity* of an error block $[i, j]$ is $A_{[i,j]} = \left|\{\, r_k : k \in [i, j] \,\}\right|$, the number of distinct true states in $R_{[i,j]}$. Based on $A_{[i,j]}$, we define a typology of errors (Fig. 1), where each error block is of exactly one type:

**Delay** ($A = 1$): True and predicted states in $[i, j]$ are constant. A neighbor matches the predicted state (e.g., $r_{i-1} = p_{i-1} = p_i$).
**Isolation** ($A = 1$): True and predicted states in $[i, j]$ are constant. Error occurs within a constant true state (i.e., $r_{i-1} = r_{j+1} = r_i$).
**Transition** ($A = 2$): Exactly two distinct true states in $R_{[i,j]}$.
**Missing** ($A \geq 3$): Three or more distinct true states in $R_{[i,j]}$.

This typology is crucial as error severity varies. Moreover, real-world transitions are often gradual, making sharply defined temporal boundaries a limited representation. Measures should therefore penalize errors near true boundaries (delays, transitions) less severely than those within homogeneous regions (missing, isolated).

**Desired Properties:** To rank error types and ensure meaningful evaluation, we propose properties for state detection measures.

**P1**: *The measure should be sensitive to the errors **length**, with larger errors leading to lower scores.*
**P2**: *The measure should account for the temporal structure, penalizing **positions** of errors differently.*
**P3**: *The measure should be sensitive to the **type** of error, with different penalties for different types.*
**P4**: *The measure should be **interpretable** and provide insights into the quality of the segmentation.*

These properties guide measure development but are not strict axioms; their relative importance can vary. For instance, error length (**P1**) might be contextualized by its position (**P2**) or type (**P3**). They therefore serve as principles, not rigid rules.

*2.1.2 Existing Measures and Limitations.* Several measures exist for evaluating segmentation, but encounter various limitations.

*Change Point Detection Measures.* The **F1 score**, common for change point detection, combines precision and recall. It matches predicted change points to ground-truth ones within a *margin*, avoiding double-counting. However, choosing the *margin* is difficult, and may result in scoring identically qualitatively different segmentations, violating **P1**. A margin relative to time series length (e.g., 1% as in [8]) is often used but remains parameter-dependent.

The **covering score** measures segment-level similarity using the average Intersection over Union (IoU) for each ground-truth segment. For real ($R$) and predicted ($P$) state sequences, it is:

$$C = \frac{1}{N} \sum_{r \in R} |r| \max_{p \in P} \frac{|r \cap p|}{|r \cup p|} \tag{1}$$

However, the covering score can also assign identical scores to qualitatively different segmentations (having same IoU), failing **P1**.

*State Detection Measures.* State detection is often evaluated with clustering-based measures like the **Adjusted Rand Index** (ARI), Normalized or Adjusted Mutual Information (NMI or AMI). We focus on ARI in this paper, although limitations and proposed solutions applies to NMI and AMI. ARI is derived from the Rand Index (RI), which measures the fraction of agreeing pairs in segmentations. Given true ($R$) and predicted ($P$) state sequences, and their unique states $U_R$ and $U_P$, a *contingency matrix* $C = [n_{ij}]$ is formed, where $n_{ij}$ is the count of observations in state $U_R[i]$ in $R$ and $U_P[j]$ in $P$. With $\mathbb{E}[\text{RI}]$ as the expected RI under randomness, ARI is:

$$\text{RI} = \frac{\sum_{i,j} \binom{n_{ij}}{2}}{\binom{n}{2}}, \quad \text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{1 - \mathbb{E}[\text{RI}]} \tag{2}$$

ARI is sensitive to the number of matching temporal points, satisfying **P1**. However, it is point-based and ignores error position or type. Therefore, two segmentations with different error patterns of same length (e.g. delay and isolation) yield the same ARI, thus failing **P2** and **P3**.

## 2.2 Proposed Measures: WARI and SMS

With existing measures having limitations (failing **P1**, **P2**, or **P3**) and lacking interpretability (**P4**), we propose two state detection measures: **WARI**, a *distance-to-boundary* weighted ARI, and **SMS** (State Matching Score), which maps predicted to true states and scores based on error types from Sec. 2.1.1.

*2.2.1 Toward Position-Sensitivity: WARI.* The Adjusted Rand Index (ARI) treats all segmentation errors equally, failing property **P2**. To address this, we introduce the *Weighted Adjusted Rand Index* (WARI). WARI weights observations based on their distance to true change points. For each time step $i$, let $d_i$ be the distance to the nearest ground truth change point. We define a weight $w_i = 1 + \alpha\, d_i$, where $\alpha \geq 0$ (default 0.1) is a user-configurable parameter. When $\alpha > 0$, observations far from true change points (i.e., large $d_i$) receive higher weights. Consequently, errors occurring in the interior of segments are more penalized than errors near segment boundaries.

To compute WARI, the standard contingency matrix (as described in Sec. 2.1.2) is adapted. Instead of using simple counts $n_{ij}$, WARI employs weighted sums: $\widetilde{n}_{ij} = \sum_{x_k \in U_i \cap V_j} w_k$, where $w_k$ is the weight of the $k$-th observation. WARI is then calculated using these

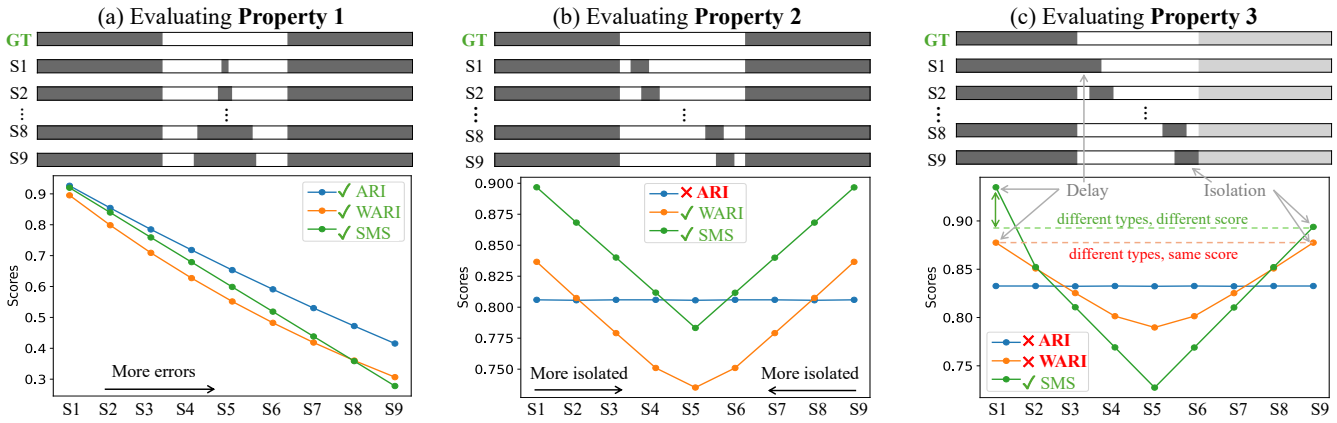


**Figure 2: Synthetic data examples illustrating various error types and measure responses.**

weighted $\widetilde{n}_{ij}$ values in the standard ARI formula (Equation 2). This weighting scheme can similarly be applied to other clustering-based measures like NMI and AMI.

*2.2.2 Enhancing Interpretability: SMS.* While WARI addresses the issue of error position (**P2**), it does not inherently handle sensitivity to error type (**P3**) or provide deep interpretability (**P4**). To overcome these remaining limitations, we introduce the State Matching Score (SMS), an interpretable and customizable measure.

The computation of SMS involves two main stages. First, an **Optimal State Mapping** procedure aligns the predicted state labels with the true state labels. This is typically achieved by constructing a cost matrix where entries represent the negative overlap (or a similar cost function) between each predicted state and each true state. An assignment algorithm, such as the Hungarian algorithm [12], can then be used to find the mapping that maximizes the total overlap (minimizes total cost). Special handling may be needed for predicted states that remain unassigned after this process, for instance, by mapping them to available true state labels or to new unique labels if the number of predicted states exceeds true states.

Second, a **Scoring** step evaluates the quality of the segmentation based on the mapped predicted sequence. Error blocks are identified by comparing the mapped predicted sequence to the true sequence, and each error block is classified according to the typology defined in Sec. 2.1.1 (*delay*, *transition*, *missing*, *isolation*). These error blocks are then penalized based on their length, type, and potentially their context (e.g., distance to true boundaries for certain error types). A key feature of SMS is its allowance for custom penalty weights for each error type, enabling practitioners to tailor the measure to specific application needs or error sensitivities. While flexible, SMS remains robust to the choice in these weights, with the total count and length of errors being the primary driver of the score. The final SMS score is typically normalized, for example, to a range of [0, 1].

To illustrate the differences between measures, Fig. 3 presents a qualitative comparison of segmentation results from SOTA algorithms E2USD and Time2State on a MoCap dataset time series. Traditional measures like ARI marginally favor E2USD, despite its segmentation exhibits several isolated errors and is qualitatively less accurate overall. In contrast, Time2State produces a more consistent segmentation, primarily with delay and transition errors. The proposed SMS, along with WARI, correctly identify Time2State's output as the better segmentation. Notably, SMS provides an interpretable diagnostic of error types, a capability absent in conventional measures.

## 2.3 Preliminary Results

We empirically evaluate our measures using 6 segmentation methods (E2USD [13], Time2State [19], HDP-HSMM [14], TICC [10], ClaSP [8] with kMeans, and PaTSS [5]), which are applied to 5 diverse datasets (PAMAP2 [17], USC-HAD [20], UCR-SEG [6], ActRecTut [4], MoCap [1]). We use a standard Intel Core i7 CPU with 32GB of RAM and set a time limit of 24h for each dataset. SMS weights for *delay*, *transition*, *missing* and *isolation* are set arbitrarily to 0.1, 0.3, 0.5 and 0.8 respectively. We then compare algorithm performance according to ARI, WARI and SMS using pairwise Wilcoxon sign rank tests (with $\alpha = 0.05$), treating each time series as a test instance.

*2.3.1 Evaluating the Evaluation Measures.* Complementing the qualitative example, a synthetic experiment (Fig. 2) systematically evaluates the sensitivity of ARI, WARI, and SMS to error **length**, **position**, and **type**. All measures satisfy **P1** (error length sensitivity), as scores decrease with growing error sizes (Fig. 2(a)). WARI and SMS address **P2** (position sensitivity) by penalizing isolated errors more heavily, unlike ARI which remains insensitive to position (Fig. 2(b)). Finally, only SMS fulfills **P3** (type sensitivity), by distinguishing between delay and transition errors, for instance (Fig. 2(c)).

*2.3.2 Impact on State of the Art.* Apart from ClaSP and PaTSS, that could not run under the time and memory limits for dataset PAMAP2 (comprising very long time series), algorithm rankings remain largely consistent across measures, with Time2State ranking first for multivariate data and ClaSP for univariate, while TICC and HDP-HSMM generally rank lower. SMS offers novel insights (Fig. 3), revealing that neural and probabilistic methods like Time2State, E2USD, and HDP-HSMM tend towards *isolated errors*, whereas ClaSP, TICC, and PaTSS show more *missing* and *delay errors*. These varied error patterns highlight diverse segmentation behaviors. Looking ahead, SMS's interpretability can guide model refinement;

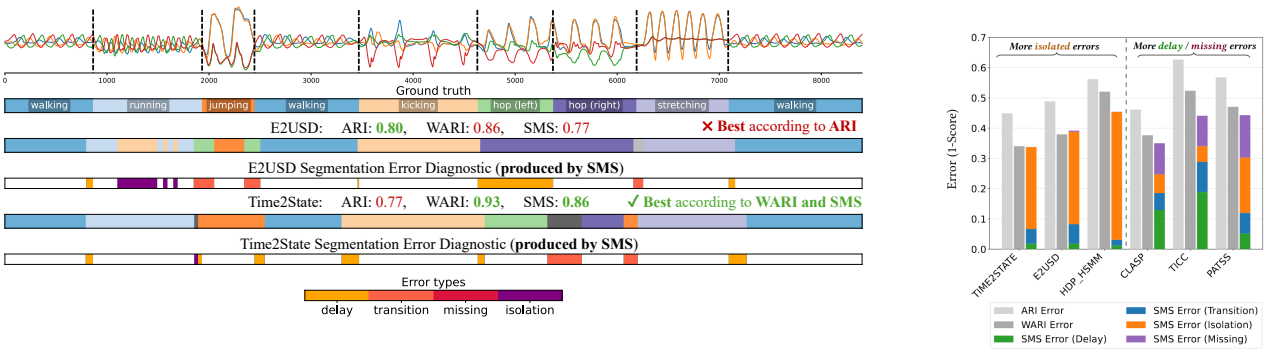


**Figure 3: Segmentation of a time series from the MoCap dataset using E2Usd and Time2State (left) and error rate and error type contribution across datasets (right).**

analyzing error types can inform parameter tuning (e.g., adjusting cluster parameters to mitigate specific errors) and algorithm development, thereby enhancing evaluation and design processes.

## 3 CONCLUSIONS AND FUTURE RESEARCH

This work lays a foundation for more nuanced time series segmentation evaluation by formalizing an error typology and proposing desirable properties for measures. The introduction of two new measures, **WARI** and **SMS**, addresses limitations of existing methods and offers novel insights into segmentation quality. Looking ahead, we aim to build upon this work, exploring two main ideas.

### 3.1 Tuning Encoding-based Methods with SMS

First, we will investigate the potential of SMS to guide hyperparameter tuning for encoding-based segmentation methods. The detailed error feedback from SMS can inform adjustments to encoder architectures, representation learning objectives, or clustering parameters to minimize specific, undesirable error types (e.g., reducing *missing* errors by encouraging finer-grained segmentations or penalizing *isolated* errors by promoting smoother transitions).

### 3.2 Interpretable Graph-based Representations

Second, we will broaden the scope to encompass other critical time series analysis tasks. A key direction will be the development of unified and semantically rich representations for time series data, potentially leveraging graph-based structures. The goal is to create representations that are not only effective for various downstream tasks (like segmentation, classification, anomaly detection) but are also inherently interpretable. This aligns with the overarching goal of this PhD thesis: to move toward more transparent and understandable time series analytical methods.

## REFERENCES


[1] [n.d.]. Carnegie Mellon University - CMU Graphics Lab - motion capture library. http://mocap.cs.cmu.edu/

[2] Ryan Prescott Adams and David J. C. MacKay. 2007. Bayesian Online Changepoint Detection. arXiv:0710.3742 [stat.ML] https://arxiv.org/abs/0710.3742

[3] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* 31, 3 (May 2017), 606–660. https://doi.org/10.1007/s10618-016-0483-9

[4] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *Comput. Surveys* 46, 3 (Jan. 2014), 1–33. https://doi.org/10.1145/2499621 Publisher: Association for Computing Machinery (ACM).

[5] Louis Carpentier, Len Feremans, Wannes Meert, and Mathias Verbeke. 2024. Pattern-based Time Series Semantic Segmentation with Gradual State Transitions. In *Proceedings of the 2024 SIAM International Conference on Data Mining, SDM 2024, Houston, TX, USA, April 18-20, 2024*. 316–324.

[6] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. 2018. The UCR Time Series Classification Archive.

[7] Shohreh Deldari, Daniel V. Smith, Amin Sadri, and Flora Salim. 2020. ESPRESSO: Entropy and ShaPe awaRe timE-Series SegmentatiOn for Processing Heterogeneous Sensor Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (Sept. 2020), 1–24. https://doi.org/10.1145/3411832 Publisher: Association for Computing Machinery (ACM).

[8] Arik Ermshaus, Patrick Schäfer, and Ulf Leser. 2023. ClaSP: parameter-free time series segmentation. *Data Mining and Knowledge Discovery* 37, 3 (May 2023), 1262–1300. https://doi.org/10.1007/s10618-023-00923-x Publisher: Springer Science and Business Media LLC.

[9] Shaghayegh Gharghabi, Yifei Ding, Chin-Chia Michael Yeh, Kaveh Kamgar, Liudmila Ulanova, and Eamonn Keogh. 2017. Matrix Profile VIII: Domain Agnostic Online Semantic Segmentation at Superhuman Performance Levels. In *2017 IEEE International Conference on Data Mining (ICDM)*. New Orleans, LA, 117–126.

[10] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. 2018. Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. https://doi.org/10.48550/arXiv.1706.03161 arXiv:1706.03161 [cs].

[11] R. Killick, P. Fearnhead, and I. A. Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* 107, 500 (Dec. 2012), 1590–1598. https://doi.org/10.1080/01621459.2012.737745 arXiv:1101.1438 [stat].

[12] H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955), 83–97.

[13] Zhichen Lai, Huan Li, Dalin Zhang, Yan Zhao, Weizhu Qian, and Christian S. Jensen. 2024. E2Usd: Efficient-yet-effective Unsupervised State Detection for Multivariate Time Series. In *Proceedings of the ACM Web Conference 2024*. ACM, Singapore Singapore, 3010–3021. https://doi.org/10.1145/3589334.3645593 E2USD.

[14] Masatoshi Nagano, Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, and Masahide Kaneko. 2018. Sequence Pattern Extraction by Segmenting Time Series Data Using GP-HSMM with Hierarchical Dirichlet Process. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Madrid, 4067–4074. https://doi.org/10.1109/iros.2018.8594029

[15] John Paparrizos and Luis Gravano. 2016. k-Shape: Efficient and Accurate Clustering of Time Series. *SIGMOD Rec.* 45, 1 (June 2016), 69–76.

[16] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. *Proc. VLDB Endow.* 15, 8 (April 2022), 1697–1711. https://doi.org/10.14778/3529337.3529354

[17] Attila Reiss. 2012. PAMAP2 Physical Activity Monitoring. https://doi.org/10.24432/C5NW2H

[18] Patrick Schäfer and Ulf Leser. 2022. Motiflets: Simple and Accurate Detection of Motifs in Time Series. *Proceedings of the VLDB Endowment* 16, 4 (2022), 725–737.

[19] Chengyu Wang, Kui Wu, Tongqing Zhou, and Zhiping Cai. 2023. Time2State: An Unsupervised Framework for Inferring the Latent States in Time Series Data. *Proceedings of the ACM on Management of Data* 1, 1 (May 2023), 1–18.

[20] Mi Zhang and Alexander A. Sawchuk. 2012. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Pittsburgh Pennsylvania, 1036–1043.