"What's Happening in This Line?" Eliciting Human Concepts and Mental Models in Time Series Interpretation

Matilde Silva
Fraunhofer Portugal AICOS
Porto, Portugal
matilde.silva@fraunhofer.pt

Ricardo Melo
Fraunhofer Portugal AICOS
Porto, Portugal
ricardo.melo@fraunhofer.pt

André V. Carreiro
Fraunhofer Portugal AICOS
Porto, Portugal
Comprehensive Health Research Center (CHRC)
Porto, Portugal
andre.carreiro@fraunhofer.pt

Duarte Folgado
Fraunhofer Portugal AICOS
Porto, Portugal
Comprehensive Health Research Center (CHRC)
Porto, Portugal
duarte.folgado@fraunhofer.pt

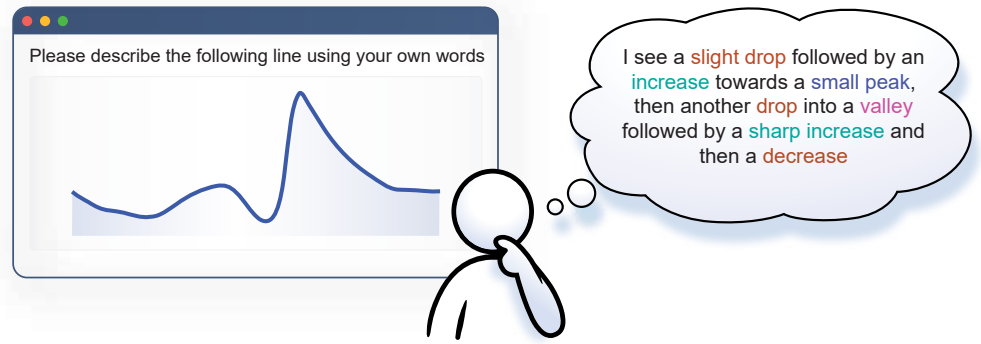

Figure 1: From lines to natural language: What do people describe when they look at a time series?

Abstract

People, regardless of formal training or domain expertise, interpret time series by inspecting them and describing what they see using intuitive concepts such as spikes, peaks, and stable periods. These informal descriptions shape data scientists' analyses, yet many time series representations and methods rely on concepts defined a priori by system designers rather than being empirically grounded in human reasoning. In this study, we ask a simple but underexplored question: "What concepts do people use when they describe time series?" We present preliminary results from an ongoing user study investigating how participants describe time series patterns in their own words. We describe the elicitation instrument, study design, and report preliminary observations from a pilot that highlight both shared and divergent conceptual interpretations across participants. Our long-term goal is to establish an empirical taxonomy of concepts used to interpret time series, informing the design of more human-centered analysis methods.

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798546

CCS Concepts

• **Human-centered computing** → **Empirical studies in visualization**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Time series, Concept elicitation, Sensemaking, Visual analysis, User studies, Data interpretation

ACM Reference Format:
Matilde Silva, Ricardo Melo, André V. Carreiro, and Duarte Folgado. 2026. "What's Happening in This Line?" Eliciting Human Concepts and Mental Models in Time Series Interpretation. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3772363.3798546

1 Introduction

Time series data are ubiquitous across domains, capturing how phenomena evolve over time. In analytical notebooks, dashboards, and monitoring tools, data scientists routinely inspect time series to understand system behavior, validate hypotheses, and surface unexpected patterns [32]. A single line can encode hours of physiological activity, years of change in pedestrian flows, or moments of disruption in economic markets. Before engaging in any formal

analysis, such as statistical tests, data mining, or machine learning pipelines, data scientists typically visually inspect the data. They describe what they see using intuitive, natural language *concepts*[1]: *several spikes*  , a *gradual rise*  , *two pronounced peaks*  , or a *decrease followed by a stable period*  . These informal descriptions emerge rapidly, often within seconds, and reflect the fundamental ways in which people make sense of temporal data before formalizing these intuitions into computational representations.

The time series research community has long recognized the value of moving beyond raw numbers toward higher-level representations. Yet, the concepts that operationalize these representations are typically defined by researchers and developers when specifying algorithms and models. As a result, algorithms that rely on such concepts encode implicit assumptions about which temporal patterns are meaningful and are often not grounded in empirical evidence of how people visually describe and interpret time series. Consequently, concept-based data representations, model explanations, and interfaces risk misalignment when these assumptions diverge from users' lived sensemaking practices. This concern is consistent with previous research showing that developers' cognitive biases shape software development processes [10] and that misalignments can arise between developers and end users in domains such as healthcare [8].

This work is motivated by a simple but underexplored question: "What concepts do people use when they describe time series?" We focus on three aspects of interpretation: (1) the vocabulary participants use to describe salient temporal patterns, (2) how they segment continuous signals into meaningful regions, and (3) where interpretations converge or diverge across individuals. We study this question through an elicitation study using a questionnaire designed to capture spontaneous descriptions rather than impose a predefined vocabulary or taxonomy.

Contribution. We introduce a structured elicitation protocol for systematically capturing how people describe time series patterns, providing an initial empirical characterization showing both shared patterns and individual differences in their descriptions.

The long-term goal of this research is to establish an empirical taxonomy of the concepts people use when interpreting time series. By foregrounding user interpretation, this work contributes to ongoing discussions in Human-Computer Interaction (HCI) on sensemaking and interpretability, and leverages that perspective to engage the time series analysis community in considering how such concepts emerge and how they might inform the design of more human-centered systems.

2 Related Work

Research on time series analysis has long sought to move beyond raw numerical representations toward higher-level abstractions that make temporal patterns more interpretable [1]. Early work in syntactic and symbolic pattern recognition treated time series as structured sequences of primitives or symbols, enabling pattern matching and querying based on shape rather than exact numeric values [3, 15, 34]. These ideas were later operationalized in scalable symbolic representations designed for large, noisy, real-world data [19]. Complementing these approaches, Rodrigues et al. [30] proposed a tool that converts time series into symbolic sequences and supports expressive, query-driven regular expression search. Later, Mannino and Abouzied [21] developed querying via hand-drawn sketches, while Imani et al. [16] enabled pattern search using natural language descriptions. This emphasis on abstraction and semantic description has also shaped recent work on concept-based explainability, which seeks to explain model predictions in terms of human-understandable attributes or abstractions, such as textual concepts or representative examples of distinctive characteristics of the input [27]. Rather than explaining low-level features (i.e., raw, fine-grained input parts, such as numerical values or pixels), these methods frame explanations in terms of higher-level concepts [17] intended to align with how people reason, making them among the most useful explanation approaches [18, 24]. We found that the concept vocabularies used by concept-based time series methods are rarely grounded in empirical evidence of how people actually perceive and describe temporal patterns, and instead largely reflect assumptions made by developers. This gap creates an opportunity to empirically elicit concepts from practitioners and use them to inform the design of more human-centered representations, explanations, and analytic interfaces for time series.

HCI and visualization research offer the theoretical grounding and methodological tools needed to address this challenge. Effective visualizations depend not only on perceptual accuracy, but also on alignment with users' mental models, interpretive strategies, and sensemaking practices [2, 5, 12]. Prior studies in discrete visualizations, such as graph and chart comprehension, have primarily focused on how users interpret visualizations to inform their design [13, 28]. Understanding what people attend to, describe, and label when interacting with visual representations is central to designing systems that support users' existing interpretive frameworks [14, 23]. In contrast, relatively little work has systematically examined how people perceive, segment, and label temporal patterns in time series, particularly with the goal of informing the design of time series representations and concept-based explanations. Time series present distinct interpretive challenges, since they encode temporal trajectories with ambiguous boundaries and variable granularity, requiring subjective judgments about structure over time. These properties make empirically elicited user concepts particularly important for representation and explanation design.

3 Methodology

3.1 Experimental Task

We designed a two-stage elicitation protocol using a questionnaire to capture how participants describe time series. First, participants freely described a time series using their own words. Second, they completed a small set of structured tasks that asked them to identify synonyms and to segment the time series into regions corresponding to the concepts they had described. This progression reflects best practices in survey design to help overcome potential limitations from question framing [35]. The protocol comprised three main tasks, each broken into short subtasks (see Appendix A for an overview of the questionnaire). This protocol was inspired

[1]Here, we use the term "concepts" to refer to morphological, shape-based descriptors of time series, specifically the qualitative patterns, perceptual groupings, and semantic distinctions that people spontaneously use when visually describing the form of a time series.

by Quadri et al. [28], which proposed a study combining think-aloud and written free-response elicitation methods to find patterns in common data visualizations.

Sensemaking and concept externalization (Task T1). Participants first describe a time series in their own words and then reflect on that description by explicitly listing the concepts they used. This separation between description and reflection supports elicitation by preserving spontaneous sensemaking while encouraging participants to externalize the semantic constructs underlying their interpretations. Finally, we included a segmentation task in which participants mark contiguous regions of the time series they perceived as conceptually distinct. The goal was to encourage more granular sensemaking and to externalize how participants partition continuous time series into meaningful units. This task supports comparison across participants by enabling analysis of where interpretations converge or diverge, including cases in which the same temporal segment is interpreted in different ways.

Semantic alignment and concept structuring (Task T2). To account for lexical variation, participants were asked to provide synonyms and, when relevant, broader or more specific concepts for their identified terms, enabling analysis of semantic convergence across participants.

Contrastive sensemaking (Task T3). To examine how participants reason about differences between time series, participants were asked to describe differences between two time series drawn from different behavioral categories, with each time series corresponding to a distinct behavior.

3.2 Stimuli

We collected time series examples from the UCR Time Series Archive [11], a widely used benchmark repository of time series datasets. We selected five representative samples according to the procedure described in Appendix B. The five stimuli were organized to vary in structural complexity, alternating between relatively simple time series and more structurally complex signals. The five stimuli were presented in the same order for all participants to reduce procedural variability across responses. Time series were presented as static dark line charts without titles or gridlines to minimize framing effects. The x-axis representing time was included to allow participants to reference specific regions of the series when describing patterns or attributing particular concepts and interpretations.

3.3 Participant Recruitment

We used purposive sampling to recruit 10 participants with diverse backgrounds and varying levels of experience with time series analysis. Four participants (P1, P2, P4, and P7) reported no prior familiarity working with time series. The remaining six participants reported between 1–2 years of experience (P3, P5, P6, and P8), 6–10 years of experience (P10), or more than 10 years of experience (P9) working with time series.

3.4 Procedure

The study protocol was implemented using an online questionnaire, LimeSurvey (version 6.16.4). Participants were invited via personalized email containing a link to access and complete the study. Participants answered in English and completed the three tasks (Section 3.1) for each of the five time series examples, entering their responses in text fields. Tasks were completed in a fixed sequence (T1 → T2 → T3). This ordering was intentionally designed to prioritize natural elicitation: participants first described the time series freely, without being provided with a definition of "concept," to avoid imposing interpretive constraints. Subsequent tasks encouraged reflection on previously used terms and closer analysis of specific regions of the signal. No tutorial examples were provided for the first description to avoid priming effects. Participants were explicitly encouraged to describe what they saw in their own words and were informed that there were no right or wrong answers, as the goal was to elicit natural, personally meaningful descriptions. After completing the questionnaire, participants took part in a brief post-questionnaire interview to gather feedback on the study protocol (see Appendix C for details).

3.5 Data Analysis

We analyzed participants' responses in two stages: (1) concept grouping and (2) segmentation agreement.

Concept grouping. We consolidated lexical variants into concept groups to account for synonymy. Grouping was informed by (i) the concept lists and segmentation labels from Task T1, and (ii) the synonyms provided in Task T2. We first created preliminary groups using the synonyms explicitly provided by participants in Task T2. Terms were assigned to the same group when participants identified them as equivalent. Due to the small number of participants (N = 10), some concepts and their variants did not naturally group together and required manual curation. In these cases, grouping decisions were made conservatively and only when semantic overlap was clear. The resulting structure preserves differences in granularity (e.g., *sharp drop* as a specific instance of *drop*).

Segmentation agreement. Segmentation agreement was assessed qualitatively. Participant annotations were visually compared to identify regions where temporal intervals overlapped. Two segments were considered overlapping when their annotated intervals covered substantially similar portions of the signal. We considered agreement when multiple participants marked overlapping intervals and assigned labels belonging to the same concept group. Because this study is exploratory and the sample size is limited, we did not perform formal agreement analysis. Instead, agreement patterns were used descriptively to identify shared and divergent interpretations.

4 Findings

We report preliminary observations from a pilot study conducted to assess and refine the questionnaire before deploying it with a broader audience. Given the exploratory nature and limited sample size, these observations aim to illustrate emerging patterns elicited by the current questionnaire, rather than to support generalizable or conclusive claims. We organize the findings into two purposes: (1) to assess how the questionnaire structure shapes concept elicitation, and (2) to surface preliminary patterns in how participants conceptualize time series. Additional analyses are provided in Appendix D.

4.1 Findings from the Elicitation Protocol

Task T1. When asked to list concepts after producing the description, some participants (P1, P2, P3, P5) tended to re-identify the concepts they had previously used and often collapsed detailed descriptive expressions into one or multiple general concepts. For example, qualifiers used in the description to distinguish specific variations (e.g., *steep positive slope* or *negative slope*) were frequently omitted in the concept list in favor of more general terms (e.g., *slope*) or kept as independent concepts. In contrast, the remaining participants retrieved and used the descriptive expressions from their description during the concept listing and segmentation subtasks.
Task T2. Participants used different linguistic expressions to refer to similar concepts, with vocabulary varying by domain expertise. For example, participants with experience in time series analysis used technical jargon such as *minimum* (P3, P5, P8, P9, and P10) and *slope* (P3 and P5), while P2 used *inclination* instead of *slope*. This reinforces the utility of this task in facilitating downstream analysis of convergence among the participants.
Task T3. Most participants focused on describing differences between the two classes. Some highlighted only the most discriminative concept, while others briefly described each class, highlighting similarities and differences. Participants with more experience in analyzing time series (e.g., P8 and P9) tended to provide more concise descriptions, whereas participants with no experience often described each class in more detail. Participants with 1–2 years of experience generally varied between conciseness and descriptiveness but still primarily focused on differences. This pattern illustrates that task framing shapes conceptual articulation, with participant expertise contributing to variability in response style.

4.2 Preliminary Patterns in Elicited Temporal Concepts

We analyzed responses from Tasks T1 and T2 by grouping recurring concepts together with the synonyms identified by participants across free-text initial descriptions, concept lists, and segmentation annotations (see Section 3.5).

Common concepts emerge in more general concept categories (see Figure 2). We observed convergence around participant descriptions, primarily in more general concept categories (i.e., *drop*, *wave*, *noise*, *peak*, *oscillating*, *rise*, and *plateauing*). The disagreement across participants was driven less by semantic conflict among concepts and more by differences in granularity, with participants describing the same temporal phenomena at different levels of abstraction. For instance, the common term *drop* was instantiated by P1, P3, P4, P6, and P8 as *dropping sharply*.

Differences in segmentation also reflected variation in granularity (see Figure 3). Some participants segmented a region into multiple smaller segments, while others treated it as a single segment. For example, in the *Wave* column (Figure 3), the third signal from the bottom shows a region that P3 labeled as *wave*, whereas P1 segmented the same temporal interval into smaller regions labeled *rise* and *plateauing*. Additionally, participants sometimes qualified common concepts (e.g., *rising sharply*), suggesting that even frequently mentioned concepts across participants might present multiple levels of granularity.

Ambiguity was most pronounced for plateau regions (see Figure 3). The same segment was sometimes categorized as a *plateau* by some participants and as a *rise* or *drop* by others (e.g., in the third signal from the top, the same temporal interval was labeled as *drop* by P2 and as *plateauing* by P3), reflecting uncertainty about whether gradual change should be treated as stability or directional movement. In contrast, when segments were labeled as *rise* or *drop*, participants tended to agree more closely on their overall structure.

5 Implications for Methods and Research Directions

These findings motivate the next step: deploying the proposed questionnaire with a larger, more diverse participant pool. A central research direction emerging from this work is the development of an empirically grounded taxonomy of concepts in time series data. A key challenge would be reconciling the convergence across participants with the systematic variation. Unlike a priori taxonomies grounded on fixed categories, an empirical taxonomy must accommodate concepts that are inherently fuzzy and context-dependent, raising open methodological questions about how to aggregate elicited data without collapsing meaningful ambiguity.

Taxonomies have been explored in scene and image understanding [9, 25, 26, 33], but to our knowledge, there is no work in time series. One plausible reason is perceptual stability: visual scenes contain relatively stable, spatially bounded structures that support consistent naming and aggregation across viewers, making taxonomy construction more tractable. Time series, by contrast, are continuous and temporal, with concepts that vary in granularity, have ambiguous boundaries, and often require contextual information [20], complicating efforts to derive shared conceptual structures.

An empirical taxonomy of time series concepts would have important implications for both data mining and concept-based model explainability. In query and retrieval tasks, such a taxonomy could support more principled and human-aligned abstractions. In concept-based explainability, misalignment between model-defined concepts and users' mental models can lead to an illusion of understanding, where explanations appear meaningful but fail to support genuine insight [4]. Prior work shows that users often find explanations involving many concepts overwhelming, instead preferring small, selective sets tailored to their goals and expertise [18, 24, 29]. Grounding concepts in empirical evidence of human interpretation is, therefore, critical for developing explanation methods that are both interpretable and useful in practice.

6 Conclusion and Future Work

This work presents an elicitation-based protocol for examining how people describe time series data, foregrounding the concepts that emerge during visual sensemaking rather than imposing predefined vocabularies. Through a pilot study, we observed that while participants frequently used a small set of common temporal concepts, they systematically varied in how they reflected the granularity of the concepts as they were introduced. These findings highlight both the promise and the challenge of grounding concept-based representations in empirical evidence.

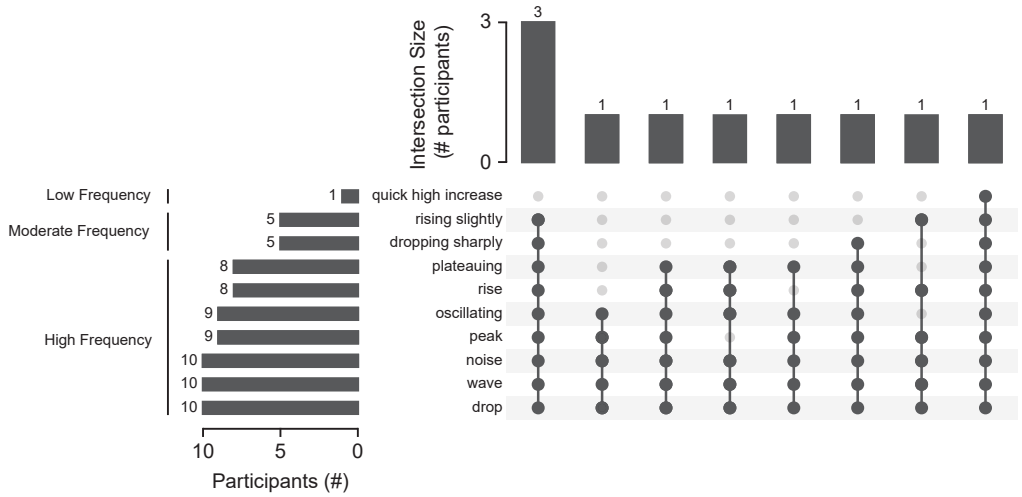

Figure 2: Concept usage and co-occurrence across participants (N = 10). Horizontal bars on the left show how many participants used each individual concept at least once. Each column in the matrix represents a specific combination of concepts (indicated by filled dots connected by vertical lines), and the bar above each column shows how many participants exhibited exactly that combination. The concepts selected for this figure are grouped into three frequency tiers: commonly used (8–10 participants), moderately frequent (3–7 participants), and rare (1–2 participants).

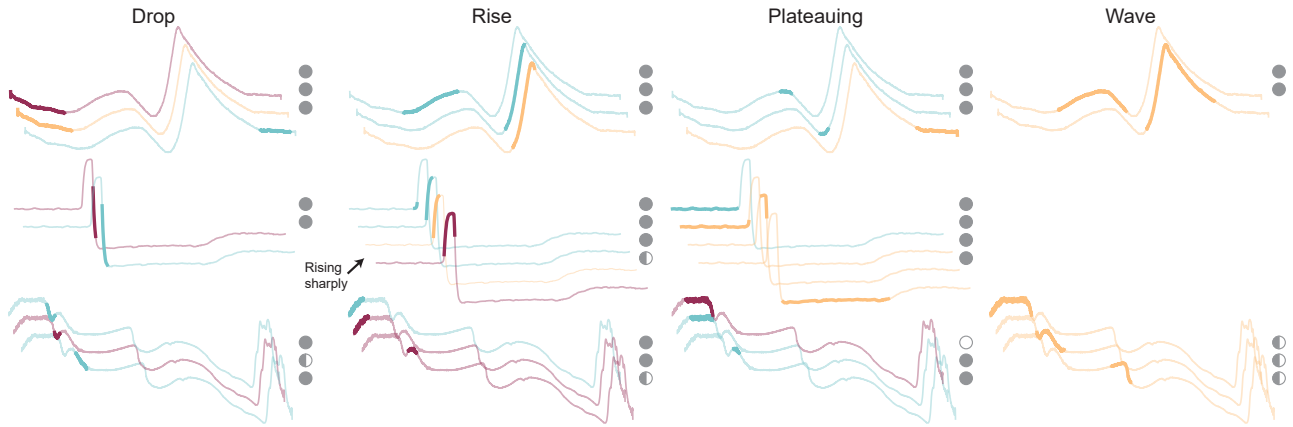

Figure 3: Example segmentation results for participants P1, P2, and P3. Each column corresponds to a concept label (shown at the top), and highlighted thicker regions indicate the portions of each time series that participants associated with that concept. Annotation confidence is self-reported by participants and indicated to the right of each time series as low (○), medium (◐), or high (●).

As a pilot study, these findings are necessarily limited in scope and generalizability. We used a fixed stimulus order to reduce procedural variability, which may have introduced sequencing or priming effects. Future deployments will randomize or counterbalance stimulus order. Participants also reported fatigue toward the end of the questionnaire, suggesting opportunities to improve the protocol.

The protocol can be adapted within a mixed-method study design that combines qualitative concept elicitation with quantitative analyses of convergence and segmentation agreement (e.g., intersection-over-union of annotated intervals). Statistical analysis of concept frequency and inter-participant agreement would help assess the generalizability of emerging categories.

Future work will scale the study to a larger and more diverse participant pool. Findings from a larger-scale deployment, complemented by follow-up expert focus groups, may inform the design and evaluation of systems that better align with how people interpret time series.

Acknowledgments

This work was supported by the European Union's Horizon Europe research and innovation programme (EU Horizon Europe), project "ACHILLES", project number 101189689. We thank Francisco Nunes

for his feedback on an earlier version of this manuscript, the contributors to the UCR Time Series Classification Archive for making their datasets publicly available, and all participants in our study.

References

[1] Wolfgang Aigner, Alexander Rind, and Silvia Hoffmann. 2012. Comparative Evaluation of an Interactive Time-Series Visualization that Combines Quantitative Data with Qualitative Abstractions. *Computer Graphics Forum* 31, 3 (2012), 995–1004. doi:10.1111/j.1467-8659.2012.03092.x

[2] Danielle Albers Szafir, Rita Borgo, Min Chen, Darren J. Edwards, Brian Fisher, and Lace Padilla (Eds.). 2023. *Visualization Psychology*. Springer International Publishing, Cham. doi:10.1007/978-3-031-34738-2

[3] King Sun Fu (Ed.). 1977. *Syntactic Pattern Recognition, Applications*. Communication and Cybernetics, Vol. 14. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-66438-0

[4] Maria Aslam, Diana Segura-Velandia, and Yee Mey Goh. 2023. A Conceptual Model Framework for XAI Requirement Elicitation of Application Domain System. *IEEE Access* 11 (2023), 108080–108091. doi:10.1109/ACCESS.2023.3315605

[5] Sriram Karthik Badam, Jieqiong Zhao, Shivalik Sen, Niklas Elmqvist, and David Ebert. 2016. TimeFork: Interactive Prediction of Time Series. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose California USA). Association for Computing Machinery, New York, NY, USA, 5409–5420. doi:10.1145/2858036.2858150

[6] 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 Mining and Knowledge Discovery* 31, 3 (2017), 606–660. doi:10.1007/s10618-016-0483-9

[7] Marília Barandas, Duarte Folgado, Letícia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. 2020. TSFEL: Time Series Feature Extraction Library. *SoftwareX* 11 (2020), 100456. doi:10.1016/j.softx.2020.100456

[8] Nadine Bienefeld, Jens Michael Boss, Rahel Lüthy, Dominique Brodbeck, Jan Azzati, Mirco Blaser, Jan Willms, and Emanuela Keller. 2023. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *npj Digital Medicine* 6, 1 (May 2023), 94. doi:10.1038/s41746-023-00837-4

[9] Bryan Burford, Pam Briggs, and John P. Eakins. 2003. A Taxonomy of the Image: On the Classification of Content for Image Retrieval. *Visual Communication* 2, 2 (2003), 123–161. doi:10.1177/1470357203002002001

[10] Souti Chattopadhyay, Nicholas Nelson, Audrey Au, Natalia Morales, Christopher Sanchez, Rahul Pandita, and Anita Sarma. 2020. A Tale from the Trenches: Cognitive Biases and Software Development. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (Seoul, South Korea) *(ICSE '20)*. Association for Computing Machinery, New York, NY, USA, 654–665. doi:10.1145/3377811.3380330

[11] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. 2019. The UCR Time Series Archive. *IEEE/CAA Journal of Automatica Sinica* 6, 6 (2019), 1293–1305. doi:10.1109/JAS.2019.1911747

[12] Jeffrey Heer and Ben Shneiderman. 2012. Interactive Dynamics for Visual Analysis. *Commun. ACM* 55, 4 (2012), 45–54. doi:10.1145/2133806.2133821

[13] Rafael Henkin and Cagatay Turkay. 2020. Words of Estimative Correlation: Studying Verbalizations of Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* PP (09 2020). doi:10.1109/TVCG.2020.3023537

[14] Donald D. Hoffman and Manish Singh. 1997. Salience of Visual Parts. *Cognition* 63, 1 (1997), 29–78. doi:10.1016/S0010-0277(96)00791-3

[15] Steven L. Horowitz. 1975. A Syntactic Algorithm for Peak Detection in Waveforms with Applications to Cardiography. *Commun. ACM* 18, 5 (1975), 281–285. doi:10.1145/360762.360810

[16] Shima Imani, Sara Alaee, and Eamonn Keogh. 2021. Qute: Query by Text Search for Time Series Data. In *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2*, Kohei Arai, Supriya Kapoor, and Rahul Bhatia (Eds.). Springer International Publishing, Cham, 412–427. doi:10.1007/978-3-030-63089-8_27

[17] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. Proceedings of Machine Learning Research, Online, 2668–2677.

[18] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human–AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581001

[19] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. Association for Computing Machinery, San Diego, CA, USA, 2–11. doi:10.1145/882082.882086

[20] Dongyu Liu, Sarah Alnegheimish, Alexandra Zytek, and Kalyan Veeramachaneni. 2022. MTV: Visual Analytics for Detecting, Investigating, and Annotating Anomalies in Multivariate Time Series. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30. doi:10.1145/3512950

[21] Miro Mannino and Azza Abouzied. 2018. Expressive Time Series Querying with Hand-Drawn Scale-Free Sketches. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173962

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* 3, 29 (2018), 861.

[23] Donald A. Norman. 2014. Some Observations on Mental Models. In *Mental Models*. Psychology Press, New York, NY, USA, 7–14.

[24] Christoph Obermair, Alexander Fuchs, Franz Pernkopf, Lukas Felsberger, Andrea Apollonio, and Daniel Wollmann. 2023. Example or Prototype? Learning Concept-Based Explanations in Time-Series. In *Proceedings of The 14th Asian Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 189)*, Emtiyaz Khan and Mehmet Gonen (Eds.). PMLR, Online, 816–831.

[25] Genevieve Patterson and James Hays. 2012. SUN attribute database: Discovering, annotating, and recognizing scene sttributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, RI, USA, 2751–2758. doi:10.1109/CVPR.2012.6247998

[26] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision* 108, 1 (2014), 59–81. doi:10.1007/s11263-013-0695-z

[27] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2025. Concept-based Explainable Artificial Intelligence: A Survey. *ACM Comput. Surv.* (2025). doi:10.1145/3774643

[28] Ghulam Jilani Quadri, Arran Zeyu Wang, Zhehao Wang, Jennifer Adorno, Paul Rosen, and Danielle Albers Szafir. 2024. Do You See What I See? A Qualitative Study Eliciting High-Level Visualization Comprehension. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Honolulu, HI, USA, 1–26. doi:10.1145/3613904.3642813

[29] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. 2023. Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Vancouver, BC, Canada, 10932–10941.

[30] João Rodrigues, Duarte Folgado, David Belo, and Hugo Gamboa. 2019. SSTS: A syntactic tool for pattern search on time series. *Information Processing & Management* 56, 1 (2019), 61–76. doi:10.1016/j.ipm.2018.09.001

[31] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, 1 (1978), 43–49. doi:10.1109/TASSP.1978.1163055

[32] Johanna Schmidt. 2021. Visual Data Science. In *Data Science, Data Visualization, and Digital Twins*, Sara Shirowzhan (Ed.). IntechOpen, London, UK, Chapter 6. doi:10.5772/intechopen.97750

[33] Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. 2012. Semantic hierarchies for image annotation: A survey. *Pattern Recognition* 45, 1 (2012), 333–345. doi:10.1016/j.patcog.2011.05.017

[34] P. Trahanias and E. Skordalakis. 1990. Syntactic Pattern Recognition of the ECG. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 7 (1990), 648–657. doi:10.1109/34.56207

[35] David L. Vannette and Jon A. Krosnick (Eds.). 2018. *The Palgrave Handbook of Survey Research*. Springer International Publishing, Cham. doi:10.1007/978-3-319-54395-6

A Protocol

This section provides an illustrative overview of the questionnaire protocol. The figures show how the experimental tasks described in Section 3.1 were presented to participants, with the aim of clarifying the structure and flow of the protocol. The figures illustrate the workflow for a single time series. Participants repeated this process across five representative time series.

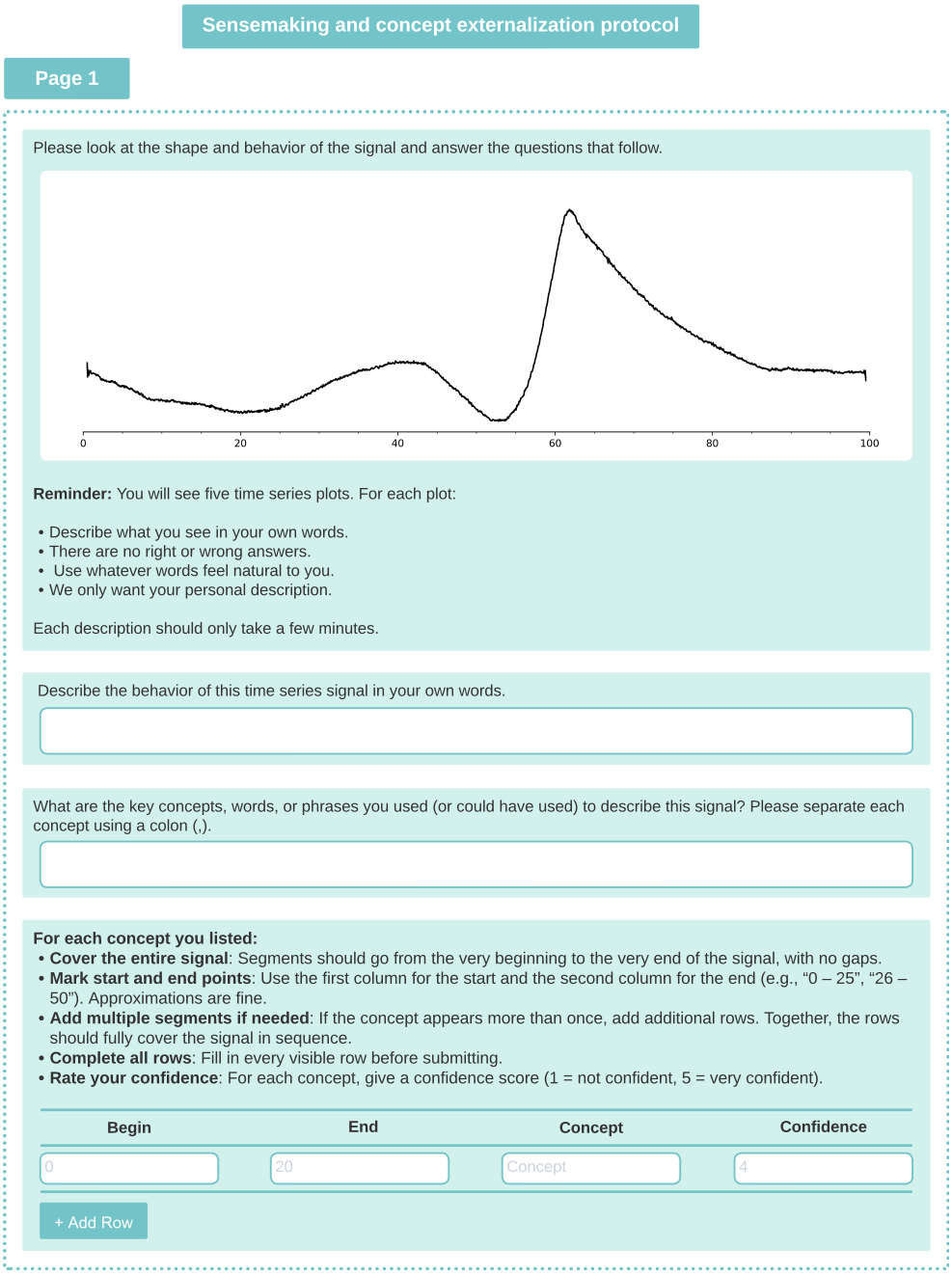

Figure 4: Illustrative overview of the questionnaire workflow for the sensemaking and concept externalization task (Task T1). Participants describe each time series in their own words, list the concepts used in their descriptions, segment the signal using those concepts, and self-report their confidence in each annotation.

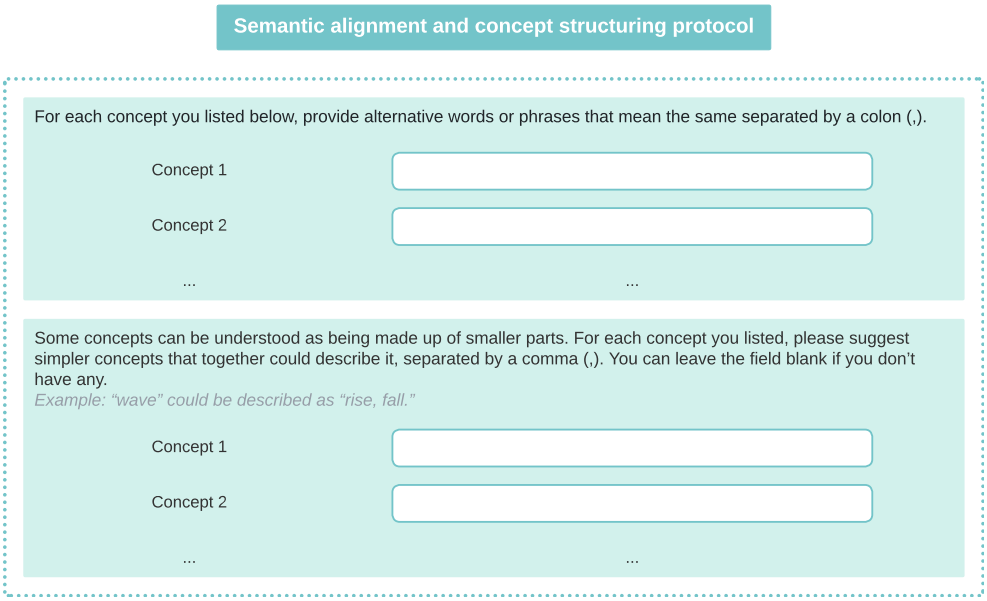

Figure 5: Illustrative overview of the questionnaire workflow for semantic alignment and concept structuring task (Task T2). Participants identify alternative expressions for previously listed concepts and, when applicable, decompose them into simpler terms.

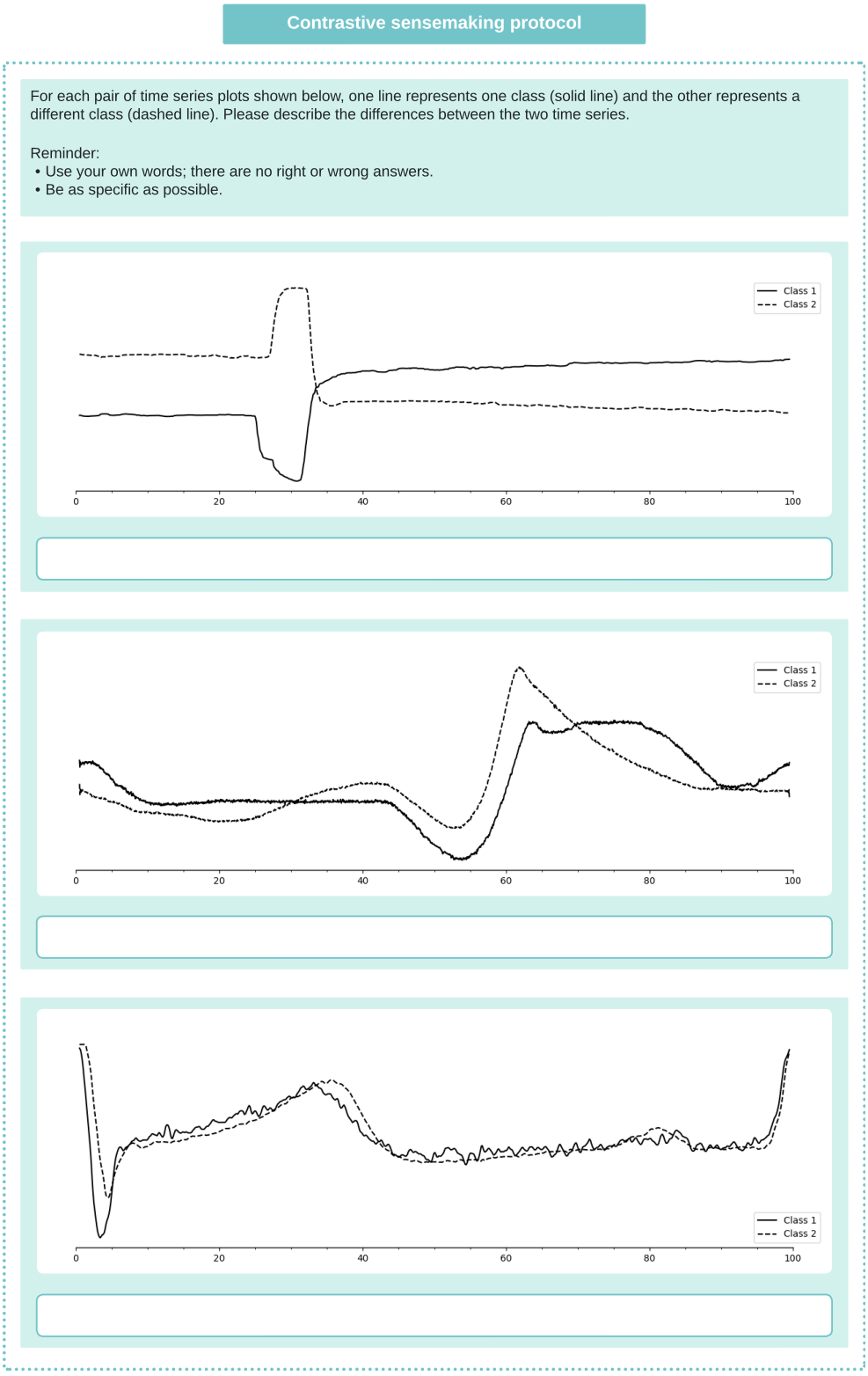

Figure 6: Illustrative overview of the questionnaire workflow for contrastive sensemaking task (Task T3). Participants describe the differences between two time series from different classes.

B Methodology: Representative Time Series Selection

We decided to select five representative samples from the UCR Time Series Archive [11]. This collection is a widely used benchmark repository for evaluating time series classification methods. It provides a large collection of datasets spanning many domains (e.g., motion, sensor readings, electrocardiogram, speech, among others) and is commonly used in the time series community as a shared reference point for algorithm benchmarking [6].

Careful sample selection is particularly important in this study, as the time series presented to participants may influence the number and granularity of concepts that can be observed. Time series that closely match canonical patterns (such as a sine wave or a clean step function) may elicit predictable, homogeneous descriptions, while atypical or irregular time series can draw attention to incidental features and obscure common interpretive patterns across participants.

To ensure a diverse set of representative time series, we followed the procedure below. By performing selection using clustering rather than relying solely on manual inspection, this approach reduces reliance on ad hoc or "cherry-picked" examples and provides a more systematic basis for sampling. While a final manual step remains, it is constrained by the clustering structure and intended to preserve diversity rather than optimize for specific conceptual patterns.

(1) Select one sample per class from each UCR univariate dataset by taking the first available instance of each class, and apply z-normalization to each selected time series. This choice was made for simplicity, as selecting the first or a randomly sampled instance is equivalent for the purposes of our analysis.
(2) We extracted a set of temporal features using TSFEL [7].
(3) The resulting feature vectors were embedded into a 7-dimensional space using UMAP [22]. We then clustered this representation into 30 clusters using the K-Means algorithm. The dimensionality and number of clusters were chosen empirically by inspecting representative samples from each cluster, with the goal of selecting a set of five samples that balanced distinct concept elicitation while avoiding cognitive overload.
(4) For each cluster, we plotted the cluster centroid in the time domain and manually selected representative signals. This manual step aimed to balance signal complexity and visual distinctiveness, ensuring a heterogeneous set of examples for participant analysis.

Figure 7 shows the five representative time series selected for inclusion in the questionnaire.

For the contrastive concept task, we selected pairs of time series from different UCR classes that were clearly distinguishable while remaining representative of their respective behaviors. To guide this selection, we computed dynamic time warping (DTW) [31] distances between candidate inter-class pairs and normalized these distances by the length of the series to allow comparison across samples. Pairs with larger normalized DTW distances were treated as more distinct and used to identify candidate examples. These candidates were then visually inspected, and a small number of pairs were manually selected to avoid trivial or visually uninformative cases and to ensure that the resulting comparisons were meaningful in a static line-plot setting.

C Post-Questionnaire Interview

A series of remote, semi-structured interviews were conducted with participants after completing the questionnaire. The interviews aimed to gather qualitative feedback on participants' experiences with the task, including their understanding of instructions, perceived difficulty, cognitive load, and the effect of specific task components (e.g., segmentation questions) on subsequent responses.

Task understanding and difficulty. One participant required clarification on how to annotate the signal and found the segmentation task challenging, while another was uncertain about how detailed or literal their descriptions should be, indicating a need for clearer instructions. Participants also noted challenges in describing signal characteristics without a vertical axis and in avoiding overly subjective descriptors. Some feedback highlighted potential biases from prior exercises and the repetition of similar concepts.

Impact of segmentation and task ordering. One participant reported that segmentation influenced the level of detail in subsequent descriptions.

Task design suggestions. Participants suggested:

- Providing an overview of phases, estimated completion time, and the ability to pause and resume the questionnaire.
- Adding a vertical axis or grid to time series because they had difficulty expressing signal characteristics and avoiding subjective descriptors without one.
- Modifying the segmentation question to focus on marking concept occurrences with segment boundaries and confidence, rather than framing the task as segmenting the entire signal.

Cognitive load. Participants reported fatigue, particularly toward the end of the questionnaire.

Overall, the interviews provided valuable insight into both task comprehension and practical challenges, informing recommendations for refining the questionnaire's design and instructions.

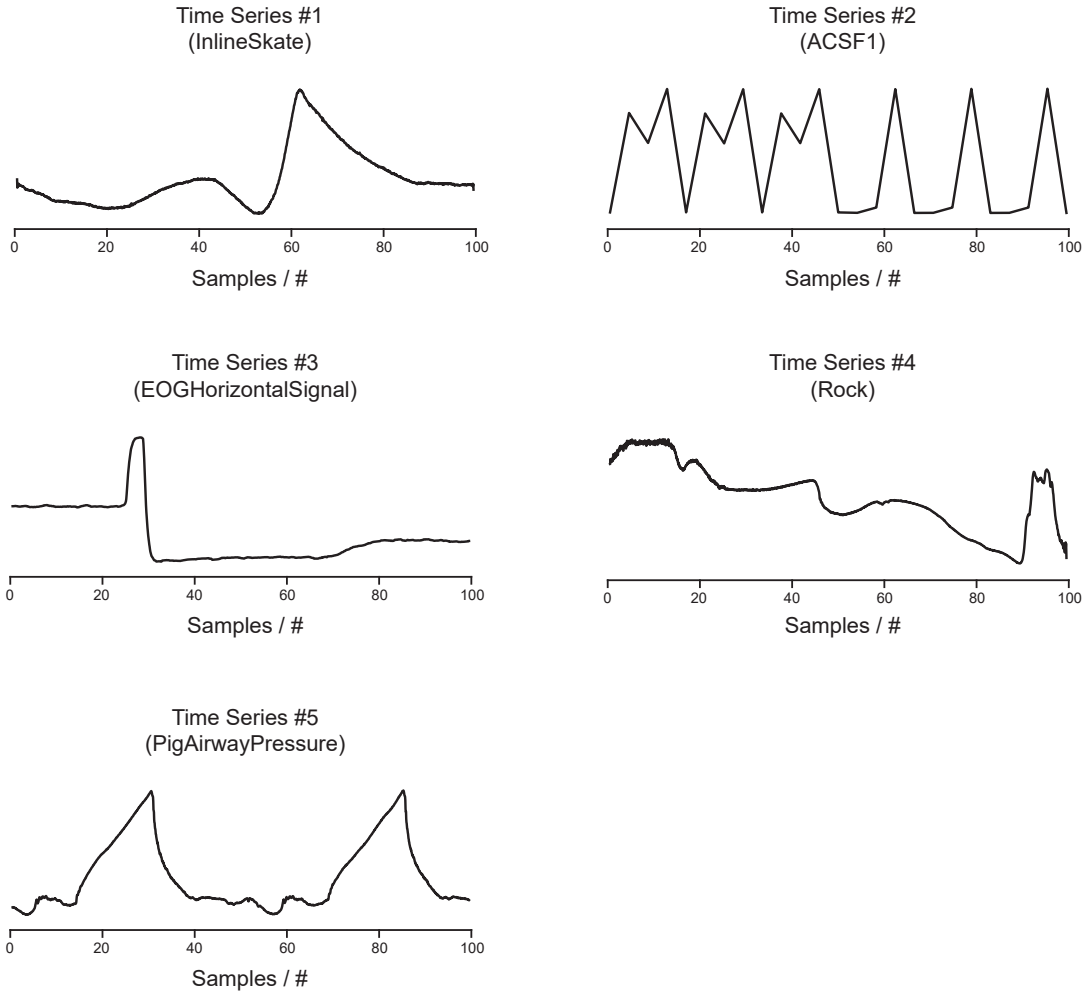

Figure 7: The five representative time series presented to participants. The number associated with each time series indicates the order in which it was presented. The dataset name is shown in parentheses. We selected a close-up of the ACSF1 time series, as the full-scale was not visually distinguishable. The x-axis is the same for all time series to avoid biasing participants by time series length and does not correspond to the actual length.

D Findings

As participants described time series using free-text responses, we followed a structured analysis to support comparison across responses while preserving individual variation.

For one participant, both the concept list and the table were relatively general; therefore, we mapped more specific concepts from the description onto the table entries.

D.1 Examples of Participant Sensemaking and Concept Externalization (T1)

The level of granularity in time series descriptions varied across participants, ranging from general descriptions of overall signal behavior (P3) to highly detailed descriptions of individual segments (P2). Below, we report example excerpts from participants' responses.

> ''The signal starts at zero in the X-axis, with around 25% value in the Y axis, starting with a gradual downwards slope, reaching it's lower value around the 20 mark, before starting to recover, reaching values similar to the start around the 40 mark, afterwards, followed by another descent, sharper this time, until around between x value 50 and 50, before a very sharp increase until the 60/65, reaching its peak, tripling highest Y value. Afterwards, it gradually descended until 100 x-axis value, where it stopped around the same y-axis value it started."
>
> P1, Time Series #1

> ''The time series shows 5 major changes in evolution, with other smaller fluctuations in between. In the first segment, between point zero and point 20 on the horizontal axis, there is a very tenuous decreasing trend, impossible to quantify because there is no vertical

axis. This trend is reversed between points 20 and 40 on the horizontal axis, rising, at the level of the vertical axis, to approximately the same value as the starting point. It remained stable for about 4 points on the horizontal axis. Then, between points 44 and 52 on the horizontal axis, there is a proportional depression/descent, remaining stable, that is, practically at the same values, between 52 and 55. From here, and up to 62 on the horizontal axis, there is an abrupt/steep/more than exponential rise in relation to the vertical axis, equivalent to about three times the previous decrease. From point 62 to 87, the curve descends, with some small discontinuities/variations in the decrease, showing some very small alternating rises and falls, but, in general, showing a decreasing trend until returning to the same initial starting point relative to the vertical axis. Between points 87 and 98, the values tend to remain stable, showing only slight variations, with almost immeasurable rises and falls, until a steep depression occurs relative to the vertical axis, less than 5 points compared to the spacing of the horizontal axis. The series ends here."

P2, Time Series #1

''Overall, the signal presents two main waves. It begins with a slight negative slope, followed by a positive slope forming the first wave, then decreases to a local minimum before rising to form the second wave, with a bigger positive slope. Finally, it decreases into a stationary segment (plateau-like) with low variability."

P3, Time Series #1

In particular, participant P3 showed a clear evolution in their descriptions from signal 1 to signal 4, illustrating the effect of explicitly asking participants to list concepts and segment the signal. This process appeared to encourage a more structured articulation of concepts, with hierarchical relationships emerging naturally from the time series.

''The signal begins with high-amplitude and a noisy segment characterized by numerous fluctuations. This segment starts by increasing and stabilizes into a plateau before decreasing to a local minimum around 15, forming a first wave. Between 15 and 20, a second wave is formed. Then, the signal decreases further and around 35 starts to form a third wave with lower amplitude. This is followed by a fourth wave, longer in duration and even lower amplitude, that persists until reaching a minimum near 90. Finally, the signal shows a wave (higher amplitude) with many fluctuations and noise towards the end."

P3, Time Series #4

D.2 Extended Segmentation Results

Figure 8 presents an extended version of Figure 3, demonstrating concept ambiguities, different granularities, and participants segment interpretations. Concepts were selected based on their frequency across participants and signals, as well as their ability to highlight similarities and differences in participants' interpretations.

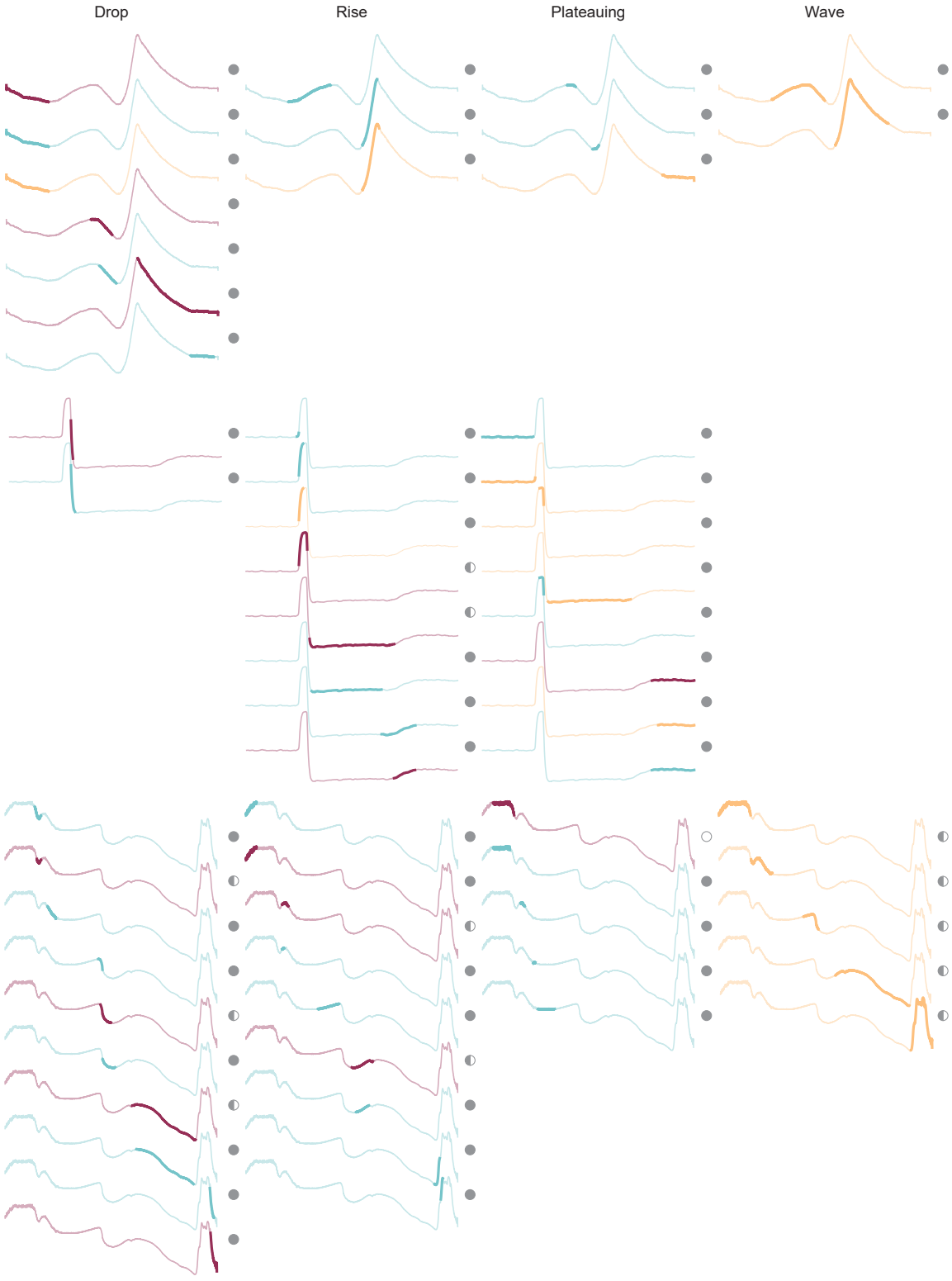

Figure 8: Segmentation results for participants P1, P2, and P3. Highlighted regions denote the segments each participant labeled with the concept shown at the top of each column. Annotation confidence is self-reported by participants and indicated to the right of each time series as low (○), medium (◐), or high (●).