Information Systems 137 (2026) 102644

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Nine years later: Reflecting on our article[☆]
A general process mining framework for correlating, predicting, and clustering dynamic behavior based on event logs

Massimiliano de Leoni [a,*], Wil M.P. van der Aalst [b], Marcus Dees [c]

[a] Department of Mathematics, University of Padova, Padua, Italy
[b] Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany
[c] Radboud University Medical Center (Radboudumc), Nijmegen, The Netherlands

ABSTRACT

This contribution revisits our article titled *"A General Process Mining Framework for Correlating, Predicting, and Clustering Dynamic Behavior Based on Event Logs"*, published in the *Information Systems* journal in 2016. It reflects on how the proposed general framework for process mining has grown in relevance with the rise of AI, emphasizing its value as a extensible approach to transforming event data into analytical and predictive insights. It also discusses how the framework relevance and the underlying message remains valid, including for emerging research directions such as prescriptive analytics, causal and/or object-centric process mining.

Process mining is an important enabler for applying Artificial Intelligence (AI) in enterprise settings. With the spectacular developments in AI, the relevance of our article has only increased. This manuscript revisits the contribution to Process Mining originally presented in our article *"A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs"*, published in *Information Systems* in 2016. It extended a previous conference paper by the same authors that was presented in 2014 at the 12th International Conference on Business Process Management (BPM 14) [1]. The primary trigger for this work was the observation that many researchers – including ourselves at the time – were repeatedly and independently designing similar conceptual frameworks. In these frameworks, particular features were combined and events were filtered to answer targeted questions using specific AI techniques. These problems ranged from everyday predictive tasks, such as forecasting the next activity or estimating process duration, to more specialized tasks, such as decision mining or identifying correlations between deviations and contextual features in conformance checking. These and other problems were – and unfortunately, still are – addressed by using an ad-hoc selection of features and often reimplementing mechanisms for feature extraction and event selection.

We proposed a unifying, extensible general framework that formalizes these analytical tasks, such as those mentioned above, as instances of a more general problem: how to generate predictive or analytical models from event logs through structured steps of event selection, feature enrichment, and identification of dependent and independent variables. The framework has served as a reusable template applicable to a broad spectrum of questions. The methodology was implemented in the FeaturePrediction package in ProM [2], whose modular plug-in architecture supported the framework's flexibility and extensibility.

While the framework is technique-independent, we demonstrate its usefulness using clustering, and classification and regression trees. These methods were chosen for their inherent explainability and for their wide adoption in research and practice during 2014–2016, when we published our framework.

While the first practical models of neural networks were deployed in the late 1950s [3], they became widely known only in the second half of the 2010s. Note that Java libraries for deep learning were not yet mature before that (e.g., Deeplearning4j was released in an alpha version at the end of 2014), while Python and its libraries were still niche. If we proposed the same framework today, we certainly would not overlook deep learning models, likely coupled with post-hoc explanation techniques or surrogate models.

In fact, we feel that the article has often been misinterpreted: many recent research works have focused on comparing their approaches to the specific techniques used in the original implementation, rather than recognizing that the main contribution was the framework itself, not the choice of algorithms.

The current BPM research is increasingly focusing on predictive and prescriptive process analytics, business process simulation, anomaly detection, conformance checking, and related topics. The challenges

☆ This article is part of a Special issue entitled: 'Information Systems 50th anniversary' published in Information Systems.

* Corresponding author.
E-mail addresses: deleoni@math.unipd.it (M. de Leoni), wvdaalst@pads.rwth-aachen.de (Wil M.P. van der Aalst), marcus.dees@radboudumc.nl (M. Dees).

https://doi.org/10.1016/j.is.2025.102644
Received 28 October 2025; Accepted 31 October 2025
Available online 4 November 2025
0306-4379/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

addressed in our article have naturally become even more urgent. It remains fundamental to provide a lens that turns raw event data into the input required by mainstream machine learning approaches. Many papers use experimental results based on a small number of event logs to show the superiority of the presented approach. However, such results highly depend on the data and assumptions, and the relevant features are naturally domain-dependent. It would be better to invest in more realistic scenarios that start from the actual data in source systems, rather than preprocessed event data.

We believe that the framework has strong ties to industrial practice. Many data-driven dashboards for process analytics, such as those provided by Aris, Celonis, and UiPath, have since integrated functionality that conceptually aligns with our framework's pipelines for event selection, feature extraction, and analysis. In fact, Marcus Dees, who was a data analyst at the Dutch Employee Insurance Agency (UWV) when we designed this framework, joined the research team because he was directly confronted with the same problems. At UWV, many information systems are custom-built, and the extraction of a proper data set for the analysis has been hard to achieve: a significant effort is necessary to integrate data across these systems. Moreover, because these systems are continuously evolving, the data format is subject to frequent changes. Our framework is designed to accommodate such settings: its extensible and configurable nature enables it to quickly adjust to dataset changes.

Once a business question is formulated and our framework's pipeline is applied to create a valid event log, existing AI techniques can be applied to find correlations between event-log features and an outcome of interest. This insight alone can be valuable, as it may either confirm existing hypotheses or challenge prevailing assumptions. However, these insights are not necessarily actionable when process issues are observed (unsatisfactory outcomes, costs, customer issues, delays, etc.) and need to be resolved. In [4], we reported on a real-life case study at UWV where we were able to accurately predict the cases yielding unsatisfactory outcomes, but failed to reduce these outcomes through interventions. The interventions were selected based on intuition rather than historical evidence, highlighting the necessity of prescriptive process analytics that couple predictions with evidence-based recommendations. We encourage future research to adopt the same principle when approaching prescriptive analytics and to use a more process-centric approach.

The framework has been operationalized to find correlations, rather than causation. Nonetheless, once the dataset is prepared through the pipeline, causal discovery methods can be applied. Recent research in causal discovery often relies on knowledge graphs, which naturally focus on categorical features and pose challenges when dealing with numerical features typical in process data (e.g., time, cost). Converting numerical variables into categorical bins can result in significant information loss. We advocate for research that leverages established causal discovery methods for continuous variables, as developed in fields such as statistics and AI, to better infer the cause-and-effect relations in processes.

The Object-Centric Process paradigm, and Object-Centric Process Mining (OCPM) is arguably the most significant innovation in the field since the publication of the article in 2016 [5]. In OCPM, there can be multiple types of objects, rather than a single case notion. Events and objects are related through Event-to-Object (E2O) relations and Object-to-Object (O2O) relations. This provides much more context and additional opportunities for feature extraction. Consider, for example, the question of whether a particular customer order will be delivered on time and in full. Using OCPM, the scope is not limited to the order itself and may include production, logistics, procurement, and sales. This leads to an explosion of possible features to be considered. Currently, there is no systematic approach and no comprehensive tool support for this problem. Due to the many-to-many and one-to-many relationships, this is a very challenging task. Therefore, we would like to encourage researchers to develop a generic infrastructure to support this, rather than "playing a numbers game" to justify incremental results. Source systems often provide a lot of contextual information that is missing from publicly available event logs. To successfully apply machine learning and AI, it is essential to extract object-centric event data that provides this context.

CRediT authorship contribution statement

Massimiliano de Leoni: Writing – original draft. **Wil M.P. van der Aalst:** Writing – original draft. **Marcus Dees:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. One of the co-authors, Marcus Dees, was employed by the Dutch Employee Insurance Agency (UWV) during this research. His involvement was not linked to any financial support or commercial interest that could have influenced the content of this work.

Data availability

No data was used for the research described in the article.

References

[1] Massimiliano de Leoni, Wil M.P. van der Aalst, Marcus Dees, A general framework for correlating business process characteristics, in: Shazia Sadiq, Pnina Soffer, Hagen Völzer (Eds.), Business Process Management, BPM 2014, in: Lecture Notes in Computer Science, vol. 8659, Springer, Cham, 2014, pp. 250–266.

[2] Massimiliano de Leoni, Wil M.P. van der Aalst, The FeaturePrediction Package in ProM: Correlating Business Process Characteristics, in: Lior Limonad, Barbara Weber (Eds.), Proceedings of the BPM Demo Sessions 2014 Co-Located with the 12th International Conference on Business Process Management, BPM 2014, Eindhoven, the Netherlands, September 10, 2014, in: CEUR Workshop Proceedings, vol. 1295, CEUR-WS.org, 2014, p. 26.

[3] Frank Rosenblatt, Perceptron simulation experiments, in: Proceedings of the IRE, vol. 48, (3) 1960, pp. 301–309.

[4] Marcus Dees, Massimiliano de Leoni, Wil M.P. van der Aalst, Hajo A. Reijers, Accurate predictions, invalid recommendations: Lessons learned at the dutch social security institute UWV, in: Jan vom Brocke, Jan Mendling, Michael Rosemann (Eds.), in: Business Process Management Cases, vol. 2, Springer, 2021, pp. 165–178.

[5] Wil M.P. van der Aalst, Object-Centric Process Mining: Unraveling the Fabric of Real Processes, Mathematics 11 (12) (2023) 2691.