Stochastic Process Mining: Characteristics and Challenges

Sander J.J. Leemans*¶, Xixi Lu*‡, David Chapela§, Claudio Di Ciccio‡, Izack Cohen‖, Benoît Depaire[xi,18], Stephan Fahrenkrog-Petersen[xix], Avigdor Gal[xvi], Massimiliano de Leoni**, Henrik Leopold††, Orlenys López Pintado§, Felix Mannhardt[xx], Fabrizio Maria Maggi[x], Niels Martin[xi,18], Marco Montali[x], Marco Pegoraro¶, Artem Polyvyanyy[xii], Arik Senderovich[xiii], Sam Verboven[xiv], Akio Watanabe[xv], Matthias Weidlich[xvii], Jan-Martijn E.M. van der Werf‡
¶RWTH Aachen University, Germany; ‡Utrecht University, the Netherlands; §University of Tartu, Estonia; ‖Bar-Ilan University, Israel; **University of Padua, Italy; ††Kühne Logistics University, Hamburg, Germany; [x]Free University of Bozen-Bolzano, Italy; [xi]Hasselt University, Belgium; [xviii]Hasselt University - Digital Future Lab, Belgium; [xii]The University of Melbourne, Australia; [xiii]York University, Toronto, Canada; [xiv]Vrije Universiteit Brussel, Belgium; [xv]NTT Corporation, Japan; [xvi]Technion – Israel Institute of Technology, Israel; [xvii]Hasso Plattner Institute, Germany; [xix]University of Liechtenstein, Liechtenstein; [xx]SINTEF Digital, Trondheim, Norway.

***Abstract*—Process mining aims to obtain insights from event logs through the automated analyses of recorded process data in information systems, with the ultimate aim to improve business processes running in organisations. However, real-life event logs are often incomplete, noisy, or ambiguous, such as missing timestamps or having ambiguous event labels, which traditional deterministic models cannot capture. Recent process mining developments have considered uncertainty in process mining artifacts more explicitly: in logs of recorded process behaviour, uncertainty may implicitly or explicitly influence process mining outcomes, while in process models, explicit uncertainty allows analysts to interpret and value outcomes. In this paper, we provide a conceptual foundation for uncertainty in process mining by introducing a four-level specification that separately addresses uncertainty in log attributes (e.g., activity labels of events, frequencies) and model elements (e.g., service times, read guards). For each type of uncertainty, we illustrate the levels with concrete examples to help understanding and application. We then provide a structured overview of the state of the art in stochastic process mining, classified using our specification, and present key open research challenges.**

***Index Terms*—stochastic process mining, process models, process discovery, levels of uncertainty**

I. Introduction

Process mining offers a plethora of (semi-)automated techniques to gather in-depth insights into business processes using event data extracted from information systems supporting such processes (e.g., Enterprise Resource Planning Systems and Health Information Systems). Leveraging these insights enables organisations to improve their processes in a data-driven way [1]. Typical process mining techniques include automatically discovering a process model that captures the process based on recorded event data, checking conformance between a normative process model and event data to highlight deviations between the desired and actual process flow, assessing the performance of a process to identify bottlenecks, and making various process-related predictions such as the next activity that will happen for, e.g., a particular customer or patient [2]. Process mining has also been applied in a wide range of domains such as finance, logistics, manufacturing, and healthcare [3], [4], [5].

While process mining techniques strive to provide organisations with reliable and actionable insights into their processes, they often do not give explicit consideration to *uncertainty* in the event data and the resulting process mining output. This is remarkable as various sources of uncertainty can emerge in reality and have been recognised in literature. Consider, for instance, the granularity of timestamps reflecting when an event took place. When the timestamps of events are, e.g., recorded at the day level, the order in which events occurred on a particular day is uncertain [6]. Similarly, it might be uncertain to which activity in the process a particular event is related [6].

Let us consider how this uncertainty affects conformance checking, which is the task of comparing observed behaviour (from event logs) against a normative process model. For example, if two events are logged on the same day without precise timestamps, it may be impossible to determine whether the process followed the correct sequence of steps, when the model prescribes strict ordering (e.g., "Activity A must happen before Activity B"). In such cases, traditional conformance checking techniques may incorrectly flag traces as non-conforming or overlook actual deviations. Others attempt to handle uncertainty but often do so implicitly by choosing a single interpretation of the ambiguous data, without acknowledging alternative possibilities.

This example highlights the need for reliable process mining techniques that explicitly capture and account for uncertainty both in the input event data and in the outputs, as overlooking it can compromise the reliability and validity of process mining insights. Recently, the increased awareness of the need to take uncertainty into account in process mining led to the emergence of *stochastic process mining*, a novel subdomain in process mining research. Explicit uncertainty may also be beneficial in the output of process mining techniques: the discovered model not only shows which steps may follow one

*Sander J.J. Leemans and Xixi Lu contributed equally to this work.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

another, but also how likely each path is. Such enriched models have increased expressiveness and enable practical benefits, for instance, focusing improvement efforts on frequently observed behaviour [7] and performing statistical tests on process variants [8].

Despite growing interest and efforts, existing stochastic process mining techniques remain fragmented. They address different sources of uncertainty (e.g., control-flow ambiguity, imprecise timestamps, and incomplete resource data), yet there is no unified conceptual foundation defining what "stochasticity" means in a process mining context. As a result, it is difficult for researchers and practitioners to survey and investigate the full landscape of stochasticity in process mining or to identify gaps where new methods are needed.

Against this background, this position paper aims to move research on stochastic process mining forward, by discussing state-of-the art algorithms, categorising them based on four levels of uncertainty, and thereby systematising previously fragmented research, making it easier to understand what research has been performed, identify gaps and compare different approaches based on the type of uncertainty they handle. It does so by delivering three key contributions: (i) conceptualising four different levels of stochastic information in a process mining context, (ii) providing a structured overview of the state of the art of research on stochastic process mining, and (iii) presenting an overview of the key open research challenges that need to be tackled. In this way, the paper provides a conceptual basis to reason about stochasticity and uncertainty in process mining, offers an overview of conducted research on stochastic process mining, and sets out a clear path on how to move forward for researchers of the field. For business analysts, advances in research on stochastic process mining will make the degree of uncertainty in event data and process mining outputs more explicit, leading to more nuanced process mining insights, which can support organisations in identifying appropriate process mining algorithms and tools to apply, interpreting the obtained results and, in turn, taking more appropriate process improvement actions.

The remainder of this paper is structured as follows. Section II introduces the basic concepts of event logs, process mining and stochastic process mining. Section III presents a conceptual framework of uncertainty in process mining. Section IV outlines several scenarios where considering stochastic information is necessary or useful. Section V discusses the state of the art in stochastic process mining. Section VI details high-level challenges of using stochastic information in process mining. Section VII concludes the paper.

II. Basic Concepts: Process Mining & Stochastic Process Mining

In process mining, analysts study records of cases in information systems to analyse the behaviour of a business process. Using process mining techniques, analysts aim to obtain insights into the inner functioning of a business process, such that managers may get inspired to change a process in order to optimise it [1]. Figure 1 shows the context of process mining schematically: a business process is supported by an information system, from which an analyst would extract an event log. Then, a variety of process mining techniques, often involving process models, is applied to gain insights into the process. We discuss several process mining techniques in Section V.

Process steps carried out in a business process, known as *activities*, are recorded in information systems. From these recordings, an event log can be extracted. Formally, an *event log* is a collection of *traces*, where each trace is an ordering of *events*, each representing the execution of a single process step. Logs, traces and events may be annotated with *attributes*, which, for instance, indicate timestamps of completion, amount of loan requested, cost of execution, by whom they were executed or tracking information [1].

A key step in many process mining projects [9] is to extract an event log from an information system [10]. Using an event log, common process mining tasks include discovering a process model from the log automatically (*process discovery*), verifying compliance of a log with respect to a process model or vice versa (*conformance checking*), analysing the performance of the process, compare different sub-processes or subgroups of traces with one another, or to make predictions about currently running cases [2].

Control-flow discovery, that is, the discovery of a process model representing the order of activities in a process, remains a prominent process mining use case, which involves inherent trade-offs. For instance, a process model designed for auditing purposes should be highly detailed and precise, capturing fine-grained activities, exceptions, and variations to ensure thorough compliance monitoring and accuracy. In contrast, a model built for a managerial dashboard should focus on the most frequent variations in performance indicators, bottlenecks, and overall process efficiency. Therefore, the model needs to be assessed based on the analysis questions at hand using a conformance checking technique [11]. Finally, the model can be enhanced with additional perspectives based on annotated attributes, such as performance, cost or resourcing [1].

A process model is a formal object, such as a BPMN model [12] or a Petri net [13]. A process model expresses which activities can be executed, in which order they can be executed, and what decision points are involved in an execution. The set of sequences of activities (*traces*) that a model expresses is its *language*.

Some process models express their behaviour in terms of variables. A *read guard* is a boolean expression over such variables, and restrict the choice they are attached to to only be enabled if the expression holds. Conversely, a *write guard* indicates how variables are to be updated on execution of an activity.

Process mining techniques may place strict assumptions on the event log. For instance, it is often assumed that the event log is a sample of behaviour that is possible in a business process; however it is also assumed that not all behaviour has been observed: "anything can happen". Nevertheless, it is typically assumed that the log contains only behaviour of the process, and if analysts are unsure about the log, they are expected to address that before applying further process mining

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

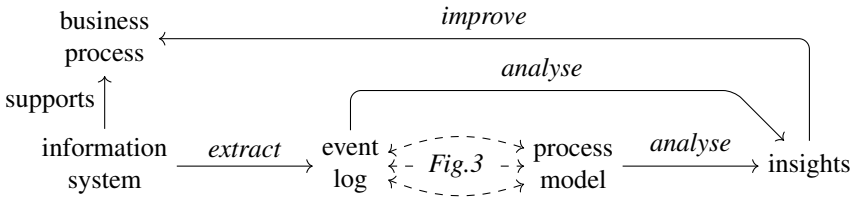


Fig. 1: The context of process mining. This paper discusses several log- and model-related techniques.

techniques [9]. As such, most techniques trust the event log fully: either an aspect is recorded and trusted, or it is missing completely. On the other hand, process models express aspects of process behaviour, such as which activities can happen, in what order, how likely certain pathways are, and whom executes which activity. Traditional process mining models express aspects either as deterministically – such as "this is the control flow of the model" – or non-deterministically – such as "here, a decision is taken"–. *Stochastic process mining* adds a third option: for logs, it allows explicit uncertainty, while for process models, it allows expression of likelihoods.

On the surface, stochastic process mining may seem quite similar to process mining: similar concepts, tasks and limitations can be distinguished. For event logs, one could be unsure about several aspects of the log, such about the reliability and precision of timestamps, the executed activities themselves, the order between activities, and whether traces are duplicated or missing. In stochastic process mining, techniques need to be aware of these uncertainties.

For process models, several aspects may be annotated with a stochastic perspective, such as the likelihood of choices in the model, the likelihood of certain values being written, and the distribution of durations of an activity. We refer to such models as *stochastic process models*. Examples include generalised stochastic Petri nets [14], stochastic labelled Petri nets [7], stochastic BPMN models [15] and simulation models [16]. Typically, such models consist of a control-flow basis, and express a likelihood to local steps to be executed, such that the likelihood of particular behaviour is the product of local decisions made to arrive at that behaviour.

A stochastic perspective allows for more expressiveness in process models. For instance, for control flow, a stochastic process model is a process model that not only expresses a language, but also a probability for each trace in the language. One can see this set, combined with the corresponding probabilities, as a distribution of traces, called a *stochastic language*.

III. Levels of Stochastic Information

In stochastic process mining, four levels of stochastic information can be distinguished: (i) certainty, which indicates that an aspect is known, (ii) stochastically known uncertainty, which indicates that an aspect is unknown but that there is a stochastic description for it, (iii) stochastically unknown uncertainty, for which an aspect can have one of multiple values, but a stochastic description is missing, and (iv) unknown, which indicates that any information on the aspect is absent beyond the mere existance of the aspect. Each level may apply to aspects of both event log data and process models, and may have a slightly different interpretation for logs and models. We provide some examples of aspects and their four-level specifications in Table I for logs and Table II for models.

A. Certain information

First, aspects of event data and models may be expressed deterministically by a model, or we may believe that the aspects are correctly recorded in a log (CER). For example, for an event log, we may be believe with 100% certainty that a particular trace has occurred, that a particular event was part of that trace, and that event involved executing a particular activity. For a process model, we may be certain of the traces it expresses, of the activities that appear within these traces, and the order in which the activities appear.

B. Stochastic information

Probabilistically known uncertainty (stochastic information; STO) in event data and process models means that uncertain aspects of event data and models can be characterised in terms of probability distributions (see the classification in [17]). For event logs, STO can be seen as epistemic uncertainty, that is, there is only one truth and the uncertainty is due to a lack of knowledge, while for process models, STO can be seen as a combination of epistemic and aleatoric uncertainty, that is, the variability in the model's instantiation is partly due to randomness, but also due to a lack of (modelled) knowledge [18].

For instance, for an event log, consider a scenario in which we are certain that an event has occurred, but we are not certain about which activity was performed at the event, though we do know an associated probability distribution: there is a 60% likelihood that the event was CUT TOMATO and a 40% likelihood that the event was CUT MELON. In this example, the activity labels are STO.

Another instance of STO in event logs arises in timestamps when the precise time of a given timestamp, with finite precision, is known to be uniformly distributed. For instance, a timestamp may have been recorded as 12:00, which implies that the event occurred sometime between 12:00 and 12:59:59.999.... Paired with a uniform distribution [19], we know a distribution over the actual timestamps, such that this aspect is STO. Probability distributions may either be explicitly provided (e.g. by a neural classification network that classifies activities [17]), or assumed or inferred from the context (e.g. distribution of timestamps [20]).

For a process model, an STO aspect is that the model expresses probabilistic information on its choices, such as: in

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

30% of cases, CUT TOMATO is chosen, while in 70%, CUT MELON is chosen. Such process models with STO frequency information include generalised stochastic Petri nets [21] and stochastic labelled Petri nets [22].

Another example of STO in a process model is if the arrival of traces into the process is modelled using, for instance, a Poisson distribution [23], or a probabilistic resource assignment [24], making the arrival process STO.

C. Unquantified information

Unquantified information (UNQ) in event data means that an aspect may be one of a set of multiple values, but there is no further quantification or distribution of this uncertainty.

For instance, if we are certain that two events have occurred in a trace, and the events have the same timestamp, but from domain knowledge we derive that their order of appearance in the trace is arbitrary, then their order is UNQ. Other examples of UNQ in event data include noise, for which we typically do not even know which aspects of which events are UNQ, and unquantified precision of timestamps.

For a process model, UNQ means that there are non-deterministic choices and no additional perspectives provide further information on the distribution of these aspects. For instance, which transition to fire next in a standard Petri net is UNQ.

D. Unknown (no information)

The last level of stochastic information is manifested by an aspect that is entirely unknown (UNK)—that is, not even the set of possibilities is known. Any aspect absent from an event log is UNK; any non-modelled aspect of a model is UNK as well.

E. Impact of uncertainty levels

Different aspects of event data and process models may have different levels of stochastic information, as exemplified in Tables I and II. Note that not all aspects and level combinations may yet have been described in literature; they simply follow from our definitions.

Given an event log, e.g., suppose that we know which traces occurred (CER-control flow), and for a particular event, we know that an event occurred in the trace (CER-control flow), on a given day (CER-time) at 12:31 (CER-minutes), but not at what precise moment (UNQ-seconds). This example shows that UNQ is not necessary a problem: depending on the analysis, minute-precision may suffice.

Aspects of stochastic information may interrelate and be dependent on the level of stochastic information of other aspects. For instance, assuming each event is atomic, the order of the events in a trace must be consistent with their timestamps, hence, the timestamp of each event provides a lower bound to the timestamp of its successor. Thus, unless one assumes the order of events in a trace to be CER, order depends on the level of stochastic information in the timestamps, and may therefore vary from CER to UNQ. As another example, consider that we know that a trace occurred (CER-control flow), we know that two events (e_1, e_2) occurred in the trace (CER-control flow), we know that e_1 happened before e_2 (CER-control flow), we know that e_1 happened at 12:31 (CER-time), and it is given that e_2 happened between 12:00 and 12:59 with a uniform distribution (STO-time). Then, we can increase the precision of e_2's timestamp, as e_2 could only have happened between 12:31 and 12:59.

For a process model, we may know which traces can occur (CER-control flow) and we may know how likely each trace is (STO-frequency). We may know for each activity which set of resources may execute the activity, but if no further resource assignment policy is known, the resources are UNQ-resources.

TABLE I: Examples of levels of uncertainty for event logs.

Aspect	level	example
control flow	CER	We are certain of the set of sequences of activities in the log.
	STO	We know the set of sequences of activities in the log stochastically, as the log contains partially ordered sequences of activities, with a known distribution of their total orders.
	UNQ	The log contains partially ordered sequences of activities without a distribution of their total orders, or the log contains events without explicit grouping into sequences.
	UNK	We know that the log contains incomplete traces.
frequency	CER	We are certain of the multiset of sequences of activities in the log.
	STO	We know a distribution on the multiset of sequences of activities in the log.
	UNQ	The log is given as a set of traces.
	UNK	The log does not contain traces.
activities	CER	We are certain that an event was the result of executing a particular activity.
	STO	A given event may have been the result of several activities, and we have a likelihood for each activity [25].
	UNQ	Some activities may have been erroneously recorded (noise), or we have a set of activities for each event.
	UNK	The log does not contain events.
timestamps	CER	For a given event, we are certain that the event happened precisely at the given timestamp.
	STO	We know the interval in which an event happened, and we know a distribution over that interval.
	UNQ	We know when a certain event happened in a certain interval, but we do not know the distribution within that interval.
	UNK	A given event has no timestamp.
resources	CER	For a given event, we know which resource executed it.
	STO	For a given event, we know the set of resources that could have executed it, and we know a likelihood for each resource in the set.
	UNQ	For a given event, we know a set of resources that could have executed it, but we do not know which resource executed the event with which likelihood.
	UNK	For a given event, we do not know the resource that executed it.
other data	CER	For a given trace, the value of a trace attribute is known exactly.
	STO	For a given trace, we know that the value of a trace attribute was taken from a known distribution.
	UNQ	For a given trace, we know that the value of a trace attribute came from an interval.
	UNK	For a given trace, we do not know the value of a trace attribute.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

TABLE II: Examples of levels of uncertainty in process models.

Aspect	level	example
control flow	CER	Given a sequence of activities, we can decide whether the sequence is captured by the model.
	STO	Given a sequence of activities, we can compute a probability that the sequence is captured by the model.
	UNQ	The model expresses a set of partially ordered sequences of activities, but for each partially ordered sequence, only one total order is captured.
	UNK	The model does not express a set of sequences of activities.
frequency	CER	The model is fully deterministic, including the number of traces it expresses.
	STO	The model expresses a likelihood for each trace that it supports.
	UNQ	The model specifies an interval likelihood for each trace it supports.
	UNK	The model does not express the likelihood of traces.
service time	CER	The model specifies the service time of each activity execution deterministically.
	STO	The model specifies a service time distribution for each activity execution.
	UNQ	The model specifies an interval of service time for each activity execution.
	UNK	The model does not specify the service time for each activity execution.
resources	CER	The model specifies the resource that can execute each activity deterministically.
	STO	The model specifies a distribution over resources that can execute each activity.
	UNQ	The model specifies a pool of resources that can execute each activity.
	UNK	The model does not specify the resources that can execute each activity.
read guards	CER	A set of read guards is non-overlapping, such that a decision is made deterministically.
	STO	A set of read guards overlap, but likelihoods for each choice are known [26].
	UNQ	A set of read guards overlap.
	UNK	The model does not express read guards.
write guards	CER	After executing the transition, all to-be-written variables are known deterministically [27].
	STO	After executing the transition, there is a distribution over the values of to-be-written variables.
	UNQ	After executing the transition, the model is non-deterministic in the values of the to-be-written variables.
	UNK	The model does not express write guards.

If the model expresses how long each activity takes with a time distribution, then the model is STO-service time.

One can "move" between these levels, as illustrated in Figure 2: one can move towards CER by making assumptions, and towards UNQ by introducing uncertainty or making it explicit. For instance, if explicit choice likelihoods are not known in a process model (UNQ-frequency), then one can assume equal probabilities (STO-frequency). However, this obviously introduces made-up information. One can also move towards UNQ by introducing uncertainty or making it explicit. For instance, if the order of events within a trace is known to be unreliable, one can consider the events as being partially ordered based on their timestamps, which moves the ordering from CER to UNQ.

Each process mining technique inherently assumes a level of uncertainty for each aspect of the event data. For instance,

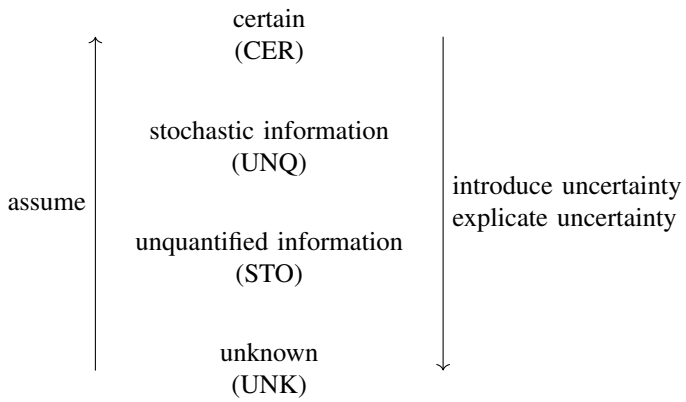


Fig. 2: Levels of stochastic information and transitions between them.

traditional process discovery techniques assume a fixed, known order for events in a trace (CER-control flow), although they attempt to deal with noise (UNQ-activities) by filtering out infrequently recorded behaviour [1]. If one applies a process mining technique to an event log in which these assumptions are not met (for instance, if the activities of events are STO while the technique assumes CER), the technique may induce information that is not there, and hence results may be arbitrary and unreliable.

IV. Motivating Scenarios

In this section, we describe several practical analysis scenarios in which stochastic information either is beneficial to consider, or must be considered. We link each scenario to the levels of stochastic information of Section III by explicating assumptions made, and by indicating how considering levels of stochastic information could assist with or enable analyses.

A. Log extraction

A key step of a process mining analysis is to create an event log of recorded process behaviour.

Event logs are typically extracted from business information systems that directly support processes. For instance, web-based process interactions can be logged with timestamps with plentiful precision.

However, if the process is partly manually recorded, or not supported by information systems at all, uncertainty in the recorded event logs may be inherent. With the rise of artificial intelligence, logs can nowadays be derived from semi-structured and unstructured data, such as movie recordings [17] and email [28],[29],[30],[31]. Artificial intelligence techniques, such as large language models, can be applied to such data to derive traces, events, activities, and other process aspects. These methods often produce multiple possible interpretations of events, each associated with a likelihood (STO).

Uncertainty is inherent to data recordings of real-life processes, and ignoring it may lead to a false sense of accuracy and reliability. For instance, timestamps may have been recorded with a precision of up to minutes. In such cases, analysing process performance at the level of milliseconds is

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

misleading. Additionally, if multiple events occur within the same minute, their recorded order in the event log may not reflect the actual sequence of execution. This can significantly impact the accuracy of process models obtained through discovery and conformance checking approaches [20]. As such, process analysts and process mining techniques need to be aware of the uncertainty in their input event logs, so that output insights and models can be appreciated for the certainty they provide.

B. Discovering a stochastic process model

As deriving information from an event log directly is challenging for human analysts, a key step in process mining projects is to obtain a process model that describes the event log well. Such a discovered process model can be analysed instead of the event log through manual inspection, or can be used as a stepping stone towards further use in a process mining pipeline for enhancement, prediction, simulation or queue analysis [32], [33] (see Figure 1).

Typically, a process discovery technique constructs a control-flow model (CER-control flow). Such a model is of limited use for manual inspection, as analyses tend to focus on high-traffic areas in process models, which can be identified with STO-frequency. In other words, insights based on the control flow alone do not include a consideration of the prevalence of the evidenced characteristics, thus, an assessment must be made afterwards on the cost-benefit trade-offs of using such insights for process improvement.

If the set of traces in the log is not certain (e.g. STO-control flow or UNQ-control flow), the discovery technique inherently has to resolve the uncertainty to construct a model that expresses a set of traces (CER-control flow). Thus, the discovery technique needs to be aware of the uncertainty in the log. Adding a stochastic perspective helps to indicate how likely certain behaviour is to occur in the process (STO-frequency); such a model is typically referred to as a *stochastic process model*. Automated stochastic process discovery involves discovering such a stochastic process model automatically.

C. Conformance checking

Conformance checking compares two event logs and/or process models with one another, for instance to prove or disprove the absence of deviations, to assess the quality of a discovered model, or to compare multiple similar processes with one another. Conformance checking can be used to compare different instances of processes with one another, such as different geographic regions fulfilling emergency maintenance orders [9], local authorities executing the same building permit process [34], or groups of students taking the same subject [35]. Conformance checking can reveal differences in the process execution between such groups, to obtain, for instance, best-practices and benchmarks.

Comparisons between processes are deepened by taking the stochastic perspective (CER-frequency in a log and STO-frequency in a model) into account: even if two given processes have the same control flow, it may be the case that some behaviour is likely in one process and rare in another, which would be a clear and potentially interesting difference between the two processes. Conversely, a control-flow difference that has a low occurrence in either models may not be of much interest.

The difference between a log and a process model can be expressed numerically, yielding a conformance measure. An example of a conformance measure is precision, which intuitively describes the amount of behaviour of the model that was also observed in the event log [36]. Precision is challenging to define conceptually for process models with loops if these models have no frequency information (UNQ-frequency). Then, the intuition dictates to measure the part of infinite behaviour (the model with loops) that was observed in a finite sample (the log) [37], [38]. This problem does not manifest for models with STO-frequency, as the infinite behaviour gets a probability. Therefore, stochastic conformance measures tend to be conceptually more elegant [39], [33]. Even when the event log contains UNQ aspects, these may still be leveraged to give upper and lower bounds to conformance measures [20], [40], [41].

D. Business process simulation

Business process simulation is a technique to assess the impact of changes to a process prior to their implementation. Simulation is typically used for ”what-if” analysis by, for example, estimating what would happen to the process cycle time if an activity is automated, or if 30% of the workforce goes on vacation for two weeks. Simulation entails building a simulation model representing the real-life process, specifying several scenarios, running these scenarios in a virtual environment to collect a variety of process performance measures such as waiting times and resource utilisation. The starting point of simulation is a stochastic process model enhanced with information such as the duration of each activity or the number of available resources and the activities they can perform (i.e., resource allocation), among other things [42], [43]. A simulation engine interprets this model and simulates the execution of the process under that specific scenario (modelling the process change), producing either a set of performance measures or an event log (from which the performance measures can later be derived).

To accurately estimate the impact of changes to a process, the simulation model must accurately represent each of the components involved in a process (e.g. the frequency of each control-flow path, the duration of each activity, the number of available resources and their working calendars, the rate of arrival of new process cases) [44]. Although these components could be modelled through CER approaches, they are typically represented with STO models to better reflect the uncertain nature of business processes. Another example is the allocation and availability of resources, such as human labour, machinery or software. Recognising the variability and unpredictability in resource behaviour is critical to closely align simulations with real-world dynamics [45]. For this reason, many proposals opt to follow STO approaches to model these aspects [45], [46], [47]. In fact, every aspect that is CER in the resulting event

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

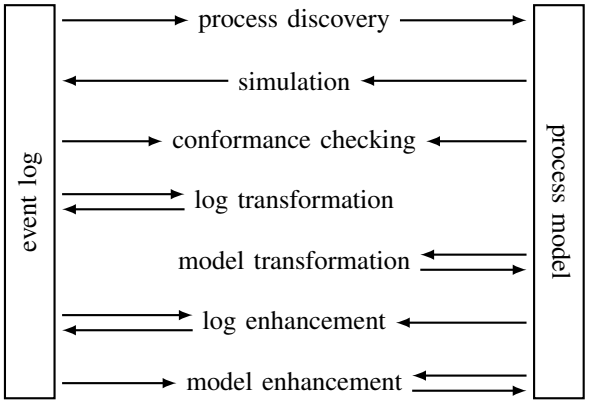


Fig. 3: Conceptual grouping of stochastic process mining techniques by their input and output.

log inherently must be STO or CER in the model: any UNQ aspect cannot be chosen from without inherently imposing a distribution.

E. Confirmatory analyses & causal inference

Many process mining projects have an exploratory nature, that is, the event data is typically visualised and an analyst attempts to gain insights into the inner workings of the process [48]. While the quality of models with respect to event logs can be assessed using conformance checking techniques, insights should be validated with domain experts [9].

However, in some cases event logs are too complex to be manually analysed (for instance, some well-known publicly available event logs contain over 400 activities [34], [49]), or access to domain experts is limited (for instance, in case of anonimised medical data [49], internet-of-things data [50] or study-progress data). In such cases, models may be of little use to derive or validate insights manually.

As an alternative, confirmatory analyses can be used to establish insights with a quantified uncertainty, such as hypothesis testing [8]. Such tests can for instance establish (i) that there is a significant difference in process behaviour between two logs (CER-control flow), or (ii) whether it can be concluded, despite uncertainty in recorded event logs (STO-activities), whether one variant of a process is different from another variant.

Furthermore, going beyond analysts manually inspecting process models, recent work has addressed the causal or associational influences of process behaviour (control flow and frequencies) on several other aspects. For instance, associational and causal relations have been identified between decisions in a process, between additional attributes to event log data and such decisions, and between such decisions and process outcomes [51], [8], [52], [53], [26], [54], [55], [56]. As associational and causal conclusions are inherently numeric, relevant aspects of involved logs and models need to be at least STO.

V. State of the Art Techniques

In this section, we discuss state-of-the-art stochastic process mining algorithms, and we link them to the levels of uncertainty of Section III. To structure the discussion, we group the techniques based on their input and output, that is, whether they accept logs and/or process models as input and/or output, as shown in Figure 3. Techniques that only have logs as input but produce neither logs nor process models as output are out of scope. Similarly, we did not encounter relevant techniques that take a log and a model as input, and produce a log and a model as output, nor techniques that take two models as input. We finish the section with a discussion on tool support (Section V-H).

A. Process discovery

Task overview: As described in Section IV-B, Process discovery techniques aim to automatically derive a process model from an event log, such that the model describes the event log well. Specifically, the basis of a process model is its control flow. To do so, process discovery algorithms can choose not to represent behaviour of the log in the model, or to represent behaviour in the model that is not in the event log. In doing so, the algorithms inherently classify behaviour as being noisy (the former), or include behaviour in the model that was not observed from the event log (the latter).

When considering stochastic information, a crucial, additional aspect is *frequency*, which requires that a discovered process model expresses the likelihood of behaviour, defined as assigning a probability to each possible trace; hence, the model's frequency perspective is STO, i.e., STO frequency. We refer to such models as *stochastic models* and the task of discovering them directly from an event log as *stochastic discovery*; please refer to Section V-G for techniques that take an existing model and enhance it with a stochastic perspective.

In case additional perspectives, such as time, resources, performance, quality or cost are discovered at the same time, we refer to the technique as a *simulation model discovery* technique. Hence, a simulation model is often CER on the control flow, STO on frequency, and STO on all other simulated aspects, such as service time and resources. Given that a simulation model is intended to be machine executable, every relevant aspect should be inherently CER or STO. Please note that STO indicates that the aspect is described stochastically, not that there is necessarily a known closed formula to compute a probability on it directly.

Existing techniques: The 'traditional' process discovery algorithm assumes that its input event log is CER for its control flow and UNK for every other aspect, and that the produced process model is CER for its control flow, as it only expresses a set of traces. Such techniques, such as [1], [57], [58], [59], are out of scope for this paper, as no uncertainty is involved.

We discuss five groups of stochastic discovery techniques: techniques that discover (i) models with a frequency perspective, (ii) a time perspective, (iii) a resource perspective or (iv) a data perspective, from CER event logs. The last group (iv) discovers models from non-CER event logs.

First, some techniques discover a STO frequency perspective from a CER event log; we refer to these techniques as stochastic process discovery techniques. Approaches include

reduction-based [60], optimisation [61], and genetic [62] approaches, and approaches that discover Markov models [63]

Second, next to an STO frequency perspective, some techniques also discover an STO time perspective. Approaches include [21], which discovers explicit distributions for waiting and service time (STO), and [63] leverages Markov decision processes that describe explicit waiting time distributions. Note the inherent relatedness of time and frequency: if two model choices are in a race, then their waiting time determines their likelihood of firing, and thus their frequency.

Third, discovery techniques have been proposed that result in models that express the resources that can execute transitions. Such techniques that use CER event data include [64], [43], [65], [66], as well as approaches that consider CER resource availability [23], or probabilistic resource availability (STO) through a probabilistic calendar [67]. Such models are often used in simulations, thus the discovery of these models is known as simulation model discovery.

Fourth, additional data perspectives can be annotated in models to guide routing decisions, either deterministically, non-deterministically or probabilistically (STO). Approaches such as [26], [16] discover such models. Note that if the probabilistic perspective depends on data that is UNQ, then that probabilistic perspective is UNQ, as we only know a probability distribution if the data is known.

Finally, a process model can be discovered directly from free-text, such as emails, where the emails contain just text, and thus control flow is UNQ while only activities and timestamps are CER [28], [29], [30], [31], however, these techniques do not yet result in a frequency perspective on the model (UNK).

Summary: A summary of stochastic process discovery and simulation model discovery approaches is given in Table IV of the Supplementary Material. Most techniques use solely CER information to discover models that are CER on control flow and UNQ on other aspects. Techniques that can accept logs in which the control flow is uncertain have been defined, however, given the plethora of techniques that extract uncertain event log data from unstructured sources (see Section IV-A), there seems to be a research gap in techniques that can handle such data.

B. Simulation

Task overview: As described in Section IV-D, simulation takes a process model and generates an event log, with the aim of, for instance, perform what-if analyses on key performance indicators. To this end, it is necessary that the model is machine-executable. Consequently, it is necessary that the model is CER on the control flow and at least STO on all other aspects that are relevant for the simulation, such as frequency and service time. For instance, a simulation model could express STO on waiting time as the waiting time depends only on CER and STO factors, even though we can only realistically sample waiting times.

Existing techniques: A previous literature review [68] summarises the key challenges; here, we focus on uncertainty and the stochastic perspective.

Typically, simulations are used to analyse the behaviour of cases in the context of resource scarcity. Resources can be modelled as a set of indistinguishable resources. In that case, it does not matter which resource is chosen and, hence, it suffices if the model expresses the class that must execute an activity (UNQ-resources) [43], [64], [65], [66].

In other cases, resources have distinguishing characteristics, such as a probabilistic availability, which necessitates that the simulator has a distribution or deterministic policy available to select an available resource (CER or STO resources) [23], [69], [70].

Summary: A summary of simulation techniques is shown in Table V of the Supplementary Material. The techniques surveyed focus on measuring performance based on resources that are either STO or UNQ, and generate event logs that are CER on all aspects. In the realm of uncertainty, it would be interesting to investigate techniques that deal with uncertainty by propagating it to the output logs.

C. Conformance checking

Task overview: Conformance checking techniques, as described in Section IV-C, compare an event log and a process model with one another, to obtain their commonalities, differences and confidence. That is, these techniques quantify the discrepancies by assessing behaviour prescribed by the model but absent in the log, and behaviour present in the log but not allowed by the model.

As discussed in Section IV-C, a scenario may arise where a process model and an event log specify the same control flow, but some trace variants have a higher probability than others. In such cases, *stochastic models*, where trace frequency is STO, are relevant. We call this problem where the log or the model is STO or UNQ for an aspect, *stochastic conformance checking*.

Existing techniques: Traditional conformance checking techniques [71] assume that both the event log and process model are CER for control flow. Approaches with this assumption include token-based replay [72] and alignments [71]. As these techniques do not consider uncertainty, they are out of the scope of this paper.

Conformance techniques can be defined for many combinations of perspectives. We discuss three stochastic conformance techniques in three groups: first, techniques that only take STO frequencies and CER control flow into account. Second, we discuss techniques that also consider other perspectives with uncertainty, but remain with CER control flow. Finally, we discuss conformance checking on event logs with uncertainty in control flow.

First, if the conformance of STO frequencies is to be compared, both event logs and/or models can be considered as distributions over traces. Corresponding measures of conformance include unit-Earth Movers' stochastic conformance, a.k.a. the inverted total variation distance [73], Jensen-Shannon conformance [74], as well as entropy-based [75], [76] and Markov-chain-based [77] approaches. These approaches are defined on log-log and log-model settings, though only result in a number. The Earth Movers' stochastic conformance, a.k.a.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

the inverted Wasserstein distance [73], [78] provides more detail though projection on log and model, though is currently only defined for log-model settings through sampling. An empirical study showed that these conformance measures can be conceptualised using three conformance dimensions [39]. Rather than a measure, another way to verify the conformance between logs and/or stochastic models with STO frequency is to perform hypothesis tests, in which, for instance, it can be established whether two logs have been derived from the same process [8], [79].

To include a time, resource or other data perspective, approaches such as [44], [79] have defined a corresponding distance measure and applied it using a sampled Earth Movers' measure. In principle, any perspective for which a distance measure can be defined can be checked for conformance, with the STO frequency as a baseline, however the curse of dimensionality awaits.

A notable other approach [33] uses a highly specific fork/join-based model that details the intended sequence for each case, the resources (CER) and the service time of activities (STO), and compares this to the actual process behaviour.

In [26], process models are extended to incorporate data aspects, in particular STO-read guards, as well as probabilities that depend on the data of a trace, thus possess frequency-UNQ. The technique calculates the probability of each trace occurring according to the model and derives a conformance score based on the log-likelihood.

Now, we examine techniques that assume classical process models (thus CER-control-flow and UNK for other aspects) while incorporating event logs with stochastic aspects (e.g., STO-control flow). Due to erroneous or coarse-grained recordings, event logs may exhibit UNQ control flow, frequency or other aspects. For example, if timestamps are recorded only at the day level, event ordering becomes UNQ. Various log formats have been proposed to capture such UNQ or STO characteristics while enabling conformance computation. In [80], [81], an event log containing traces of partially ordered events (represented using a directed acyclic graph) is proposed to capture uncertain event ordering (UNQ-control flow). A synchronous product between the partial order trace and the model is then used to compute partial order alignments. In [82], automatic planning is employed to compute the fitness over partial orders: the planner's solution imposes a total order of events, compatible with the events' partial order, so as to maximize the fitness. In [83], a sequence of sets of events is used to capture such uncertain ordering within the sets, with a probability distribution representing all feasible total orders of events for each trace. Based on these distributions, conformance checking results are derived.

As described in Section IV-A, STO data may be derived from sensors and prediction models that sample processes, and result in stochastically known logs, in which each event is a distribution of activities; for these logs, corresponding conformance checking techniques are proposed in [84]. A step further is if each event may be multiple activities, such that linear temporal logic over finite traces (LTLf) can be applied, allowing for expressing temporal rules that can handle different degrees of compliance, enabling conformance evaluation with diagnostic insights [85].

Uncertainty in event logs is further conceptualised in [86], where events are formalised in terms of: the probability that an event actually occurred and the probability of each activity that recorded this event (STO-control flow); interval of possible timestamps (STO-timestamps); range of possible data values (STO-other data). The authors proposed a SMT-based technique to check conformance between uncertain traces and a Data Petri Net. They do so by translating the DPN and the uncertain trace into constraints and formulating these constraints such that they capture the possible behaviours of the process model and the uncertainties in the event log. SMT solvers are then used to assess conformance and quantify alignment across multiple perspectives, including control flow and data.

Summary: A summary of the surveyed techniques is given in Table VI of the Supplementary Material. We surveyed a plethora of stochastic conformance checking techniques, particularly those considering stochastic models (STO aspects) or event logs with uncertain attributes. While traditional techniques assume CER for control flow and frequency, recent research has introduced methods that account for uncertainty and stochastic variations. These approaches enable more robust conformance assessments, particularly in domains where data uncertainty is inherent. We have discussed techniques that consider stochastic models as input and found that only a few conformance checking techniques consider STO aspects beyond frequency.

While out of scope for this paper, conformance checking techniques have been proposed for log-log, to which most techniques surveyed would be applicable, and, to a much more limited extent, model-model comparison. It would be interesting to extend the survey to these techniques.

D. Log transformation

Task overview: A *log transformation* takes an event log as input and returns another event log, in order to, for instance, improve its quality, group events into cases, abstract its activities, anonymise it, or reduce uncertainty.

Existing techniques: Existing log transformation techniques can be grouped according to the type of modification they perform. A first group focuses on constructing a structured event log from less structured event data. In particular, neural techniques have been proposed to derive an event log, including a case notion (CER-control flow), from a set of events without explicit information (UNQ-control flow) [87]. A second group abstracts or simplifies the information contained in a log [88]. For example, when activity labels are too detailed, documentation can be used to aggregate these activities to a level required by the business, yielding a distribution over these abstract activities (STO). A third group relies on simulation as an intermediate representation: a simulation model is discovered from an event log using time, resources, and other data perspectives (STO) that are CER in the event log, after which the simulation model is verified by comparing a simulated log to the original log [42], [43]. Finally, uncertainty-reduction techniques transform logs with

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

stochastically known activities (STO) to classical event logs (CER-control flow) by, for example, retaining the k most-likely activities [89].

Summary: A summary of log transformation techniques is shown in Table VII of the Supplementary Material. The works of [42], [43] can be used to anonimise event logs, which is closely related to the field of differential privacy and anonymisation [90]. In this context, one is faced with the aim of deliberately *decreasing* certainty, whereas other techniques typically aim to *increase* it.

E. Model transformation

Task overview: Model transformation techniques take a model as an input and return a model, for instance to add a stochastic perspective that is not based on an event log, to transform models into other formalisms, or to simplify models.

Existing techniques: Only a limited number of model transformation techniques have been proposed in literature: there exist techniques to add a trivial uniform stochastic perspective STO frequency to CER control-flow models [91], to transform between different formalisms of models such as generalised stochastic Petri nets, stochastic labelled Petri nets and stochastic deterministic finite automata [92] (CER control flow and STO frequency), and to simplify stochastic process models [93] with STO service time and CER resources.

Summary: A summary of model transformation techniques is included in Table VIII of the Supplementary Material. We believe that there may be more model transformation techniques implemented, for instance in proving the expressive power of formalisms to be equivalent, distinguishable or incomparable.

F. Log enhancement

Task overview: In log enhancement techniques, an event log and a process model are taken as input, and a log is returned; in other words: the log is repaired or enhanced based on information in a model, for instance to group events into traces, to decide on which activity was executed from a distribution, and to reduce uncertainty in event logs.

Existing techniques: Existing log enhancement techniques can be grouped according to the type of log information that is repaired or completed using a process model. A first group addresses the absence of a *case notion*. When events (CER) with timestamps (CER) are present in an event log, but no explicit case notion is present (UNQ control flow), several techniques can correlate events into cases using a process model and, thereby, enhance the event log with a CER control flow. These techniques differ mainly in the input information they exploit: some rely only on the events themselves [94], whereas others can also incorporate additional event attributes [95], [96].

A second group resolves mismatches between the activity labels recorded. In this setting, the event log contains recorded activities (CER), but the activities do not match those of a process model (CER). Using behavioural analysis, the log's activities can be mapped to the model's activities [97].

A third group of log enhancement techniques focuses on reducing uncertainty in stochastically known logs, that is, for each event (CER) there is a distribution of activities that may have happened at that event (STO) (see Section IV-A). Using a process model, a path that least deviates from the trace can be reconstructed, through which for each event, the executed activity is decided (CER) [98].

Summary: In Table IX of the Supplementary Material, we summarise several approaches. The surveyed techniques consider a plethora of uncertainty levels for their input log, but all only consider the control flow of their input model (CER control flow). It would be interesting to consider how stochastic models may assist here: it may be helpful for, for instance, case induction to leverage the likelihood of modelled behaviour (STO frequency). Furthermore, a technique that would be robust against STO frequency of logs would be suitable for streaming contexts.

G. Model enhancement

Task overview: Model enhancement focuses on improving an existing process model by incorporating insights from event log data. Specifically, the field comprises techniques to enhance models by incorporating stochastic information in the control flow dimensions, in the guards that govern the data access during process execution, or performance information. Moreover, model enhancement often also extracts information on temporal dependencies and resource allocations to include them in an existing process model.

Existing techniques: Existing enhancement techniques can be grouped according to the process perspective that is added or refined in the model.

A first group enriches the control-flow perspective with stochastic information. For instance, some techniques derive frequency annotations by overlaying the traces of the event log on a non-stochastic process model (control flow-CER) to count activities, thereby, deriving a stochastic process model (frequency-CER) [91]. Related approaches search for frequency annotations that maximise a conformance measure [61]. More advanced variants condition these annotations on additional behavioural context, such as the history of a trace [54], or use them to capture causal intra-trace dependencies [51].

A second group enhances models with data-aware behaviour. Here, the information in event logs is used to discover guards that constrain when activities may be executed. Depending on the assumptions made about the guards, the resulting process model may be enhanced with CER-read guards and UNQ-write guards for activities [99]. When overlap of guards is allowed, the enhancement typically results in a process model with UNQ-read guards and UNQ-write guards [100]. These techniques therefore extend the model beyond control flow by making data conditions explicit, while also introducing uncertainty when the observed behaviour does not determine unique guard conditions.

Another group of studies focuses on a performance perspective. One technique compares a predefined process schedule and an event log; for instance, a queueing network may be constructed [32], which learns distributions for service times and resource allocations (thus, STO). Other techniques use

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

event logs to train machine learning models for predicting waiting times and processing times [101] (STO-service time) or enhance process models by explicit extraneous waiting times, such as waiting for the client to respond [102] (STO-waiting time).

Finally, from the resource perspective, event logs with CER-resource information can be used to discover resource calendars that express the likelihood of resource availability at specific times, thereby enriching process models with stochastic availability information [23], [46].

Summary: The techniques mentioned above to handle uncertainty in process enhancement are summarised in Table X of the Supplementary Material. Here, we first note that most techniques actually combine CER information from both a process model and an event log to derive stochastic information for the enhancement of the model. Put differently, the perspectives to enhance process models that are already stochastic (even though STO models have been considered) as well as to leverage stochastic event logs for enhancements have not yet been explored comprehensively.

Moreover, first techniques that combine temporal information and resource information when incorporating stochasticity in process model enhancement. As widely established not only in business process management, but also, for instance, in operations research, both process perspectives may influence each other significantly, so that their interplay can be expected to influence the handling of uncertainty. However, the same holds true for the integration of data handling primitives in a process model, such as the guards mentioned above. Yet, so far, those have been considered solely in isolation.

H. Tool support

Regarding tool support, Table III provides an overview of the tools associated with the approaches surveyed in this section, showing the diversity of the current landscape of stochastic process mining tools. Approaches for which no tool or platform is mentioned, or for which no publicly available implementation or source code is provided, are excluded from this comparison. The selected tools are discussed across key dimensions, including modeling formalism, usability, and computational support.

Several observations can be drawn from this comparison. A first observation is the clear dominance of Petri net–based formalisms, particularly stochastic variants such as Stochastic Labelled Petri Nets (SLPN), Stochastic Petri Nets (SPN), and Stochastic Data Petri Nets (SDP), which are being used by a large number of the reported tools, especially those integrated in ProM. In contrast, BPMN-based stochastic models are mainly associated with simulation and discovery tools such as Prosimos and Simod, indicating a separation between business-oriented modeling and more formal stochastic analysis techniques.

Second, support for large-scale data remains limited across most tools. Notably, GPU acceleration is largely absent from the current landscape, with the exception of recent machine-learning-based approaches such as *DDTR* and matrix computations in *Ebi*.

Finally, many of the identified tools are research prototypes with limited integration, maintenance, and documentation. This highlights a gap between technical advances in stochastic process mining and the availability of robust, user-friendly tool support for practical applications.

VI. Challenges

In this section, we present an overview of research challenges with respect to uncertainty in event logs and process models. We describe the key issues, obstacles, and difficulties for the nine identified fundamental challenges.

The challenges discussed below are not meant to form a strict hierarchy, nor are they fully independent or mutually exclusive. Rather, they can be viewed as several related clusters that address different layers of stochastic-aware process mining. A first cluster concerns the foundations and representations of uncertainty: explicitly modeling stochasticity (in Sect. VI-A), integrating multi-perspective and intertwined uncertainty (Sect. VI-B), and extending these ideas to object-centric processes (Sect. VI-C). In this cluster, object-centric uncertainty can be seen as a special and more complex setting of multi-perspective uncertainty, where uncertainty not only spans activities, time, resources, and data, but also multiple interacting objects, processes, and their relations. A second cluster concerns modeling choices and usability: abstraction and expressiveness determine how detailed stochastic models should be (Sect. VI-D, while explainability addresses how such models and their results can be made understandable to users (Sect. VI-E). A third cluster concerns validation and adoption: trustworthiness focuses on whether stochastic-aware results are reliable and robust (Sect. VI-F); benchmarks and evaluation methods provide the shared infrastructure needed to assess and compare such reliability across techniques (Sect. VI-G); scalability concerns the computational feasibility of the proposed techniques (Sect. VI-H). Finally, causality concerns moving beyond associative stochastic relations toward intervention and counterfactual reasoning (Sect. VI-I).

A. Provide methods for explicitly modelling stochasticity

One of the fundamental challenges in stochastic process mining revolves around the philosophical foundations for modelling and profiling stochastic processes while distinguishing the *uncertainty*. This challenge focuses on the questions concerning how to precisely define and model stochasticity within a process. In this work, we distinguished four levels, CER, STO, UNQ and UNK, separately for event logs and process models. Most existing works assume CER-control flow and CER-frequency logs. Other aspects of logs are either UNQ or UNK, or assumed to be CER implicitly. For example, when an event log contains noisy behaviour, it is unclear which aspect of the log is affected and thus this aspect becomes UNQ: the log can be UNQ-cf, UNQ-activity, UNQ-freq, etc. Here, the challenges arise to explicitly define the aspects that are UNQ and the aspects that are STO.

Considering STO, stochasticity can be modelled in various forms. For example, given a task, one can model the probability that a particular resource will execute this task in two

ways: (i) one can model for each resource how likely it is that the particular resource starts executing the task, or (ii) one can model, for the task, the set of resources that may execute this task and assign each of them a probability. Similar modelling options may apply to other aspects of the log and the process model.

Considering UNQ, the challenge is to further study and identify more detailed levels of UNQ.

To summarise, stochasticity can manifest itself or be modelled in various forms, including random variations in execution times or behaviour, uncertain decision points, and probabilistic outcomes. It is essential to establish robust foundations for characterising and quantifying these uncertainties.

B. Model and integrate multi-perspective and intertwined uncertainty

In real-world processes, uncertainty manifests not only in one perspective (e.g., uncertain timestamps or control flow) but across *multiple, interdependent perspectives*, such as activities, time, resources, and data. This multifaceted nature introduces substantial complexity for stochastic process modelling and mining, especially when assumptions of independence between perspectives do not hold. We identify three stages at which this challenge must be addressed.

At the first stage, the need to model uncertainty for multiple perspectives should be made explicit. Uncertainty in process data and models can manifest in various forms and levels: activity labels may be ambiguous (UNQ), timestamps may be imprecise (UNQ) or missing (UNK), routing decisions may follow probabilistic logic (STO), and resource assignments or data values may vary stochastically (STO). Capturing and combining these distinct types of uncertainty, e.g., the UNQ, STO, and CER, and being able to combine these different perspectives into a coherent model is a non-trivial task. It requires not only sophisticated modeling approaches, but also clear semantics for how these uncertainties interact and affect the overall process behaviour. A practical barrier lies in the lack of standardized representations for stochastic logs and models. Currently, many such representations are stored in ad hoc or tool-specific formats, limiting reuse and interoperability. While Pegoraro et al. [25] propose a standardized data model to represent uncertainty in event logs, it assumes independence between uncertainties—an assumption that often does not hold in practice.

When the assumption about *independence between uncertain* does not hold, it brings us to the second stage: *interdependent uncertainty across perspectives*. In many cases, uncertainty and stochastic in one perspective depend on other perspectives [55], [103]. For instance, the data perspective (e.g., customer profile or risk rating) may influence both the likelihood of particular outcomes (control flow) and their processing time (time perspective). In such scenarios, simple independence assumptions do not hold, and capturing these dependencies is critical for building accurate and meaningful stochastic models.

However, these interdependencies complicate modelling, as classical independence assumptions are no longer valid. Accurately capturing such intertwined uncertainties requires expressive, data-aware modelling approaches that can condition stochastic behaviour on context. While some recent methods attempt to address this, e.g., using logistic regression to model conditional probabilities at branching points [103], they often require trade-offs between predictive accuracy, interpretability, and computational complexity. More advanced approaches, such as deep neural networks, may offer greater precision, but at the cost of transparency and conformance. These trade-offs are closely related to the challenges discussed in Sections VI-D and VI-E, which concern the balance between expressiveness, explainability, and expandability.

The final stage considers *quantifying uncertainty and stochasticity*. Beyond modeling, a fundamental challenge lies in *quantifying the uncertainty itself*. Expected values or conditional probabilities alone may be misleading. For example, methods from uncertainty quantification, such as those used in deep learning [18], [104], distinguish between *aleatoric uncertainty* (inherent variability in the system) and *epistemic uncertainty* (stemming from lack of knowledge). Developing techniques that can quantify, propagate, and reduce epistemic uncertainty in process models is crucial for supporting reliable predictions and informed decision-making.

C. Consider object-centric processes and their interactions

In real-world systems, processes are rarely isolated. Instead, they involve multiple interacting entities, such as orders, shipments, and invoices, evolving concurrently and influencing each other. Each of them can be a multiple-perspective process. These object-centric or interrelated (collaborative) processes introduce additional layers of complexity when extended to the stochastic setting. Capturing and analysing inter-process dependencies, synchronisation, and feedback loops becomes particularly challenging when uncertainty plays a role. Modelling the STO and UNQ across multiple interacting objects is still an open research problem. Unlike single-case process models, object-centric settings require reasoning about shared entities, asynchronous or synchronous interactions, and their compounded effects on the overall system-level behaviour.

Developing frameworks that can represent, analyse, and predict such multi-object stochastic dynamics remains an important, yet underexplored, challenge, which is different from the single-object or isolated-process stochastic modelling addressed in Section VI-B.

D. Investigate trade-offs between abstraction and expressiveness

Modelling stochastic processes involves making choices about the level of abstraction used in the process models. For example, when considering modelling the service time of an activity, one can choose between modelling the distribution of service time of each activity directly, or modelling it to depend on the number of resources available, the resource's experience, whether to include breaks, etc. Such design choices yield another challenge: on the one hand, highly abstract models may provide a simplified overview but may lack the detail needed for a comprehensive understanding and accuracy,

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

carrying the risk of under-fitting (low precision). On the other hand, detailed models may be overly complex and challenging to represent, explain or implement, and carry the risk of over-fitting. Explainability can be further hampered when using Deep-Learning models, which are naturally black box models. Techniques exist to explain black-box models, but they often assume linearity and additivity of the influence of the attributes used to produce the stochastic output (see, e.g., SHAP [105]). The choice of granularity in modelling stochasticity often depends on the specific objectives of the analysis. The challenge here is to derive guidelines, techniques, or metrics that support or guide users toward choosing the right abstraction level for modelling stochasticity for their specific analysis task while balancing between over-fitting and generalisation.

E. Design explainable techniques tailored for interpretability

Stochastic process mining techniques often produce models that include probabilities, distributions, and uncertainties in how events unfold. While these models can offer powerful insights, their complexity can make both the models themselves and the results they generate difficult to interpret, especially for non-technical stakeholders.

For example, a stochastic model may indicate that there is a 70% chance an invoice will be approved within 3 days and a 30% chance it will take over a week. However, if this variation is influenced by multiple hidden factors (e.g., customer type, time of day, or case load), it becomes difficult to explain why a particular case followed a certain path. Even more so, when these models include branching paths with probabilistic transitions or non-deterministic loops, the logic behind the model can be unintuitive and hard to communicate.

To improve adoption and trust, we need techniques that can translate stochastic behaviours or diagnostic results into explanations that are accessible and actionable, for example, by visualizing confidence intervals, summarizing typical vs. exceptional behaviours, tracing influential factors behind uncertainty, or providing textual explanation summaries. Designing these explainable techniques is especially important when the results are used for decision-making or compliance purposes.

This challenge is also related to others, such as evaluating stochastic-aware results and modeling object-centric processes, since interpretability affects how these models are validated, applied, and understood across different contexts.

F. Tackle trustworthiness

While explainability focuses on making models understandable to human users, *trustworthiness* addresses the foundational question: *Can we rely on the model and its outputs in the first place?* In stochastic process mining, ensuring the trustworthiness of models and results is particularly challenging due to the inherent uncertainty in the data and the probabilistic nature of the models themselves. This challenge involves developing algorithms and techniques that can assess the quality, reliability, and robustness of both the inputs and the resulting outputs (e.g., logs, models, diagnostic information). For example, how can we quantify the confidence in a discovered probability distribution over process paths? How can we identify whether uncertainty in the data significantly undermines the model's validity? Such as missing timestamps or conflicting traces.

Assessing the trustworthiness of stochastic process models may consider the following factors. The first key factor is to evaluate the completeness and consistency of the inputs (e.g., stochastic event logs). The second key factor is to also consider quantifying confidence, robustness, and stability of outputs, especially under noisy or partial observations. Even small variations in the input data or sampling strategy can lead to disproportionately different results in stochastic settings, raising concerns about robustness. Another critical factor involves identifying parts within the output (e.g., models) that are statistically weak or poorly supported by the data, as these regions may lead to misleading interpretations or unreliable decision support. Finally, applying sensitivity analysis is also an important factor that can help uncover how uncertainty in the input data propagates through the model and affects its outputs, offering a clearer picture of the model's reliability under various assumptions or scenarios.

Unlike interpretability, which supports user understanding, this challenge emphasizes technical correctness. It connects closely to topics like data quality, uncertainty quantification, and model validation. Addressing this challenge is essential for developing stochastic process mining techniques that are not only explainable but also reliable, robust, and usable in high-stakes applications such as compliance auditing or automated decision-making.

G. Establish methods and benchmarks to evaluate stochastic-aware results

While *trustworthiness* (Section VI-F) focuses on the reliability of individual models or results, this challenge addresses the need for general-purpose evaluation frameworks that enable systematic comparison and progress across the field.

Evaluating the effectiveness of models and algorithms in stochastic process mining presents unique challenges, particularly due to the lack of standardized evaluation frameworks. Unlike deterministic settings [57], [48], where ground truth is often available or can be generated, stochastic approaches require benchmarks that can account for probabilistic behaviour, uncertainty, and variability. Currently, there are almost no established reference datasets or evaluation criteria tailored to this context. While general-purpose event logs, such as those from the BPI Challenge, provide a starting point, they often lack the necessary annotations or stochastic ground-truth labels needed for thorough evaluations.

One promising direction is to generate stochastic models and the use of synthetic event logs generated from these stochastic models, which can offer controlled environments for testing and comparison. To advance the field, it is crucial to define evaluation metrics that reflect probabilistic accuracy, robustness under uncertainty, and computational performance. Developing such benchmarks will support fair and reproducible comparisons across methods and help drive the adoption of stochastic techniques in practice.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

H. Develop scalable methods for handling computational complexity

Another challenge is managing the computational complexity when considering uncertain or probabilistic event data and models. As uncertainty and stochastic aspects are introduced–such as imprecise timestamps, probabilistic paths, or stochastic attributes–the complexity of analysis often increases significantly, as it can result in very long traces with thousands of events, posing significant challenges for stochastic-aware process mining techniques.

This issue becomes even more pronounced in large-scale systems, where event logs may contain thousands of cases, and each case may consist of hundreds or even thousands of events. In particular, in domains such as smart manufacturing and video surveillance, sensors and video recordings generate rich but complex streams of data, that can be processed to uncertain event data (see Section IV-A). Processing such data using, for instance, EMD [73], is infeasible, which leads to stochastic process mining being infeasible in complex real-world scenarios. Typical conformance checking techniques consider a cartesian state space of log and model behaviour, which may lead to a state space explosion problem.

Efficient algorithms and scalable computational techniques are therefore important to support the practical application of stochastic-aware techniques. To analyse such complex settings, streaming and big-data approaches have been proposed for process mining, but have however, to the best of our knowledge, not yet been proposed for stochastic process mining settings.

While it was not within the scope of our survey, we noted that the feasibility of techniques discussed in our survey is typically evaluated empirically, using real-life event logs. As such, there is a need for larger and more complex publicly available data sets, such that techniques can be developed that are capable of handling them.

I. Model association and causality

Typical process mining techniques consider associative relations ("correlation"), such as "does the value of an order have anything to do with how it is processed?" [8]. Such a relation is associative, as the presence and direction of causality is unknown.

A step up in causal hierarchy are intervention questions: "what would happen if we would give this customer 10% rebate?" [106]. In process mining, such what-if questions are typically answered using simulation, though an open challenge is to decide which changes in the process will optimise process performance towards a particular goal [107]. Preliminary explicitly causal interventional analysis techniques have been proposed [51], [52], [108], [109], [56], [55], [110], however an end-user usable intuitive approach has not yet seen the light of day, and evaluation of such approaches is difficult due to the fundamental problem of causal inference [111], further complicating practical adoption.

The final level in the causal hierarchy is counterfactual reasoning, which attempts to answer questions of the type "what if I had acted differently?" [106]. Such questions, despite their relevance to the problems process mining aims to address, have not yet been proposed and remain an open challenge.

VII. Conclusion

Uncertainty in event logs on, for instance, whether an event actually happened, at what time, and what activity was performed, may impact conclusions, especially if the log is taken for certain. Process models typically express behaviour non-deterministically, that is, as a set. Adding explicit uncertainty to models therefore enriches the models to express the likelihoods, associations or causality of behaviour. In process mining, it is on the one hand important to take uncertainty in event logs into account: if uncertainty in event logs is recognised, acknowledged and addressed, drawn conclusions may be more reliable and their reliability can be assessed, which assists in jeopardise Business Process Management (BPM) efforts in organisations. On the other hand, a stochastic process model provides enables what-if analyses, supports confirmatory analyses, and allows BPM efforts to be focused on high-volume parts of the process. In stochastic process mining, both types of uncertainty are taken into account.

This paper conceptualised uncertainty in event logs and process models by adopting a framework of uncertainty and stochastic information, consisting of four levels: certainty, stochastic, unquantified and unknown. We illustrated each of the four levels of information uncertainty with concrete examples to help understanding and application. We provided a structured overview of the state of the art in stochastic process mining research, and classified each research work using our framework. Through this survey, we identified key open research challenges.

Next to the research challenges identified in Section VI, as future work, the survey could be extended with algorithmic complexity analyses of the surveyed techniques, as well as the development of comparable benchmarks to test their quality of outputs and feasibility empirically. Furthermore, the paper may serve as a positioning framework into structured literature reviews that deepen out one or several of the groups of techniques that we considered, or limit their scope based on the framework of levels of uncertainty. Finally, it may serve as a guide to researchers to develop and position their future stochastic process mining techniques, in particular in areas such as logistics and healthcare, where uncertain data may be omnipresent.

References

[1] W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.

[2] ——, "Process mining: A 360 degree overview," in *Process Mining Handbook*, ser. Lecture Notes in Business Information Processing, W. M. P. van der Aalst and J. Carmona, Eds. Springer, 2022, vol. 448, pp. 3–34. [Online]. Available: https://doi.org/10.1007/978-3-031-08848-3_1

[3] E. De Roock and N. Martin, "Process mining in healthcare – an updated perspective on the state of the art," *Journal of Biomedical Informatics*, vol. 127, p. 103995, 2022.

[4] L. Reinkemeyer, *Process mining in action: principles, use cases and outlook*. Heidelberg: Springer, 2020.

[5] ——, *Process intelligence in action: taking process mining to the next level*. Heidelberg: Springer, 2024.

TABLE III: Comparison of available stochastic process mining tools.

Tool	link	category	model	log	GUI	GPU
CPN Tools [27]	Tool	Simulation	Coloured Petri Nets	–	Yes	No
Prosimos [69], [23]	Repo	Simulation	BPMN (stochastic)	CSV	No	No
Simod [64], [69], [23], [67], [16], [43]	Repo	Discovery	BPMN (stochastic)	CSV	No	No
Ebi [7], [41], [73], [78], [74], [91], [92]	Tool	Disc./Conf./ Mod.Trans.	Stochastic Labelled Petri Nets/BPMN (stochastic)/...	XES/CSV	No	Yes
ProM: SDP Nets [26]	Repo	Disc./Conf.	Stochastic Data Petri Nets	XES/CSV	Yes	No
ProM: SLPN Miner [61]	Repo	Discovery	Stochastic Labelled Petri Nets	XES/CSV	Yes	No
ProM: SPD (Weight Estimation) [91]	Repo	Model Trans.	Generalized Stochastic Petri Nets	XES/CSV	Yes	No
ProM: Long-Distance Dependencies [54]	Repo	Disc./Conf.	Stochastic Labelled Petri Nets	XES/CSV	Yes	No
ProM: Partial Order Replay [80], [80]	Repo	Conformance	Labelled Petri Nets	XES/CSV	Yes	No
ProM: Planning Alignment [82]	Repo	Conformance	Labelled Petri Nets	XES/CSV	Yes	No
ProM: Partial Order Resolution [83]	Repo	Log Transformation	-	XES/CSV	Yes	No
ProM: StochasticPetriNet [21]	Repo	Discovery	Stochastic Petri Nets	XES/CSV	Yes	No
ProM: Earth Mov. Stoch. Conf. [73], [78]	Repo	Conformance	Stochastic Labelled Petri Nets	XES/CSV	Yes	No
ProM 5.2: sim. model integration [42]	Tool	Simulation	Colored Petri Nets	MXML	Yes	No
Fuzzy LTLf [85]	Repo	Conformance	Fuzzy LTL	XES	No	No
CoCoMoT SMT [86]	Repo	Conformance	Data Petri Net	XES	No	No
Entropia [112], [76]	Repo	Conformance	Stochastic Petri Nets, SDFA	XES	No	No
FoldA [113]	Repo	Conformance	Petri Nets	XES	No	No
SKTR [84], [98]	Repo	Conf./Log Enhancement	Probabilistic traces	XES	No	No
topK Realizations [89]	Repo	Log Transformation	Probabilistic traces	XES	No	No
DDTR [114]	Repo	Log Enhancement	Diffusion models	XES	No	Yes
Toothpaste [60]	Repo	Discovery	Stochastic Petri Net	XES	No	No
RL-(Semi)MDP [63]	Repo	Discovery	Semi-Markov Decision Process	XES	No	No
P3Folding [93]	Repo	Model transformation	GSPN	XES	Yes	No
ProSiT [103]	Tool	Simulation	Stochastic Data Petri Nets	XES	Yes	No

Abbreviations: CPN = Coloured Petri Nets, SLPN = Stochastic Labelled Petri Nets, GSPN = Generalized Stochastic Petri Nets, SDP = Stochastic Data Petri Nets, SPN = Stochastic Petri Nets, SDFA = Stochastic Deterministic Finite Automata.

[6] J. C. J. C. Bose, R. S. Mans, and W. M. P. van der Aalst, "Wanna improve process mining results?" in *CIDM*. IEEE, 2013, pp. 127–134.

[7] S. J. J. Leemans, F. M. Maggi, and M. Montali, "Enjoy the silence: Analysis of stochastic petri nets with silent transitions," *Inf. Syst.*, vol. 124, p. 102383, 2024.

[8] S. J. J. Leemans, J. M. McGree, A. Polyvyanyy, and A. H. M. ter Hofstede, "Statistical tests and association measures for business processes," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 7497–7511, 2023.

[9] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "PM ^2 : A process mining project methodology," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 9097. Springer, 2015, pp. 297–313.

[10] H. M. W. Verbeek, J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "Xes, xesame, and prom 6," in *CAiSE Forum (Selected Papers)*, ser. Lecture Notes in Business Information Processing, vol. 72. Springer, 2010, pp. 60–75.

[11] J. Carmona, B. F. van Dongen, A. Solti, and M. Weidlich, *Conformance Checking - Relating Processes and Models*. Springer, 2018.

[12] OMG, *Business Process Model and Notation (BPMN), Version 2.0*, Object Management Group Std., Rev. 2.0, January 2011. [Online]. Available: http://www.omg.org/spec/BPMN/2.0

[13] C. A. Petri and W. Reisig, "Petri net," *Scholarpedia*, vol. 3, no. 4, p. 6477, 2008.

[14] J. Campos, M. Marsan, G. Balbo, and G. Conte, "Generalized Stochastic Petri Nets: A Definition at the Net Level and Its Implications," *IEEE Transactions on Software Engineering*, vol. 19, no. 2, pp. 89–107, 1993.

[15] *SBAT: A Stochastic BPMN Analysis Tool*, ser. Engineering Systems Design and Analysis, vol. Volume 3: Engineering Systems; Heat Transfer and Thermal Engineering; Materials and Tribology; Mechatronics; Robotics, 07 2014. [Online]. Available: https://doi.org/10.1115/ESDA2014-20437

[16] O. López-Pintado, S. Murashko, and M. Dumas, "Discovery and simulation of data-aware business processes," in *ICPM*. IEEE, 2024, pp. 105–112.

[17] I. Cohen and A. Gal, "Uncertain process data with probabilistic knowledge: Problem characterization and challenges," in *Problems@BPM*, ser. CEUR Workshop Proceedings, vol. 2938. CEUR-WS.org, 2021, pp. 51–56.

[18] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction," *CoRR*, vol. abs/1910.09457, 2019.

[19] A. Gal, "Everything there is to know about stochastically known logs," in *ICPM*. IEEE, 2023, pp. xvii–xxiii.

[20] K. Goel, S. J. J. Leemans, N. Martin, and M. T. Wynn, "Quality-informed process mining: A case for standardised data quality annotations," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 5, pp. 97:1–97:47, 2022.

[21] A. Rogge-Solti, W. M. P. van der Aalst, and M. Weske, "Discovering stochastic petri nets with arbitrary delay distributions from event logs," in *Business Process Management Workshops*, ser. Lecture Notes in Business Information Processing, vol. 171. Springer, 2013, pp. 15–27.

[22] S. J. J. Leemans, F. M. Maggi, and M. Montali, "Finite-trace analysis of stochastic systems with silent transitions (extended abstract)," in *IJCAI*. ijcai.org, 2023, pp. 6452–6457.

[23] O. López-Pintado and M. Dumas, "Discovery and simulation of business processes with probabilistic resource availability calendars," in *ICPM*. IEEE, 2023, pp. 1–8.

[24] A. P. Freitas and J. L. M. Pereira, "Process simulation support in bpm tools: The case of bpmn," in *Proceedings of 2100 Projects Association Joint Conferences*. Scientific Association for Promotion of Technology and Management for Organizational and Social Transformative Change, 2015, pp. 148–156.

[25] M. Pegoraro, M. S. Uysal, and W. M. P. van der Aalst, "An XES extension for uncertain event data," *CoRR*, vol. abs/2204.04135, 2022.

[26] F. Mannhardt, S. J. J. Leemans, C. T. Schwanen, and M. de Leoni, "Modelling data-aware stochastic processes - discovery and conformance checking," in *Petri Nets*, ser. Lecture Notes in Computer Science, vol. 13929. Springer, 2023, pp. 77–98.

[27] M. Westergaard, "CPN tools 4: Multi-formalism and extensibility," in *Petri Nets*, ser. Lecture Notes in Computer Science, vol. 7927. Springer, 2013, pp. 400–409.

[28] D. Jlailaty, D. Grigori, and K. Belhajjame, "Mining business process information from email logs for business process models discovery," in *BDCSIntell*, ser. CEUR Workshop Proceedings, vol. 2622. CEUR-WS.org, 2019, pp. 1–4.

[29] A. J. Chambers, A. M. Stringfellow, B. B. Luo, S. J. Underwood, T. G. Allard, I. A. Johnston, S. Brockman, L. Shing, A. B. Wollaber, and C. VanDam, "Automated business process discovery from unstructured natural-language documents," in *Business Process Management Workshops*, ser. Lecture Notes in Business Information Processing, vol. 397. Springer, 2020, pp. 232–243.

[30] M. Pegoraro, M. S. Uysal, and W. M. P. van der Aalst, "Discovering process models from uncertain event data," in *Business Process Management Workshops*, ser. Lecture Notes in Business Information Processing, vol. 362. Springer, 2019, pp. 238–249.

[31] C. D. Ciccio and M. Mecella, "Mining artful processes from knowledge workers' emails," *IEEE Internet Comput.*, vol. 17, no. 5, pp. 10–20, 2013.

[32] A. Senderovich, M. Weidlich, A. Gal, A. Mandelbaum, S. Kadish, and C. A. Bunnell, "Discovery and validation of queueing networks in scheduled processes," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 9097. Springer, 2015, pp. 417–433.

[33] A. Senderovich, M. Weidlich, L. Yedidsion, A. Gal, A. Mandelbaum, S. Kadish, and C. A. Bunnell, "Conformance checking and performance improvement in scheduled processes: A queueing-network perspective," *Inf. Syst.*, vol. 62, pp. 185–206, 2016.

[34] J. C. A. M. Buijs and H. A. Reijers, "Comparing business process variants using models and event logs," in *BMMDS/EMMSAD*, ser. Lecture Notes in Business Information Processing, vol. 175. Springer, 2014, pp. 154–168.

[35] S. J. J. Leemans, S. Shabaninejad, K. Goel, H. Khosravi, S. W. Sadiq, and M. T. Wynn, "Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining," in *ER*, ser. Lecture Notes in Computer Science, vol. 12400. Springer, 2020, pp. 92–102.

[36] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "On the role of fitness, precision, generalization and simplicity in process discovery," in *OTM Conferences (1)*, ser. Lecture Notes in Computer Science, vol. 7565. Springer, 2012, pp. 305–322.

[37] N. Tax, X. Lu, N. Sidorova, D. Fahland, and W. M. P. van der Aalst, "The imprecisions of precision measures in process mining," *Inf. Process. Lett.*, vol. 135, pp. 1–8, 2018.

[38] A. F. Syring, N. Tax, and W. M. P. van der Aalst, "Evaluating conformance measures in process mining using conformance propositions," *Trans. Petri Nets Other Model. Concurr.*, vol. 14, pp. 192–221, 2019.

[39] A. T. Burke, S. J. J. Leemans, M. T. Wynn, W. M. P. van der Aalst, and A. H. M. ter Hofstede, "Stochastic process model-log quality dimensions: An experimental study," in *ICPM*. IEEE, 2022, pp. 80–87.

[40] M. Pegoraro, M. S. Uysal, and W. M. P. van der Aalst, "Conformance checking over uncertain event data," *Inf. Syst.*, vol. 102, p. 101810, 2021.

[41] S. J. Leemans, T. Brockhoff, W. M. van der Aalst, and A. Polyvyanyy, "Partially ordered stochastic conformance checking," *Knowledge and Information Systems*, pp. 1–29, 2024.

[42] A. Rozinat, R. S. Mans, M. Song, and W. M. P. van der Aalst, "Discovering simulation models," *Inf. Syst.*, vol. 34, no. 3, pp. 305–327, 2009.

[43] M. Camargo, M. Dumas, and O. González, "Automated discovery of business process simulation models from event logs," *Decis. Support Syst.*, vol. 134, p. 113284, 2020.

[44] D. Chapela-Campa, I. Benchekroun, O. Baron, M. Dumas, D. Krass, and A. Senderovich, "A framework for measuring the quality of business process simulation models," *Inf. Syst.*, vol. 127, p. 102447, 2025.

[45] L. Pufahl, F. Stiehle, S. Ihde, M. Weske, and I. Weber, "Resource allocation in business process executions - A systematic literature study," *Inf. Syst.*, vol. 132, p. 102541, 2025.

[46] O. López-Pintado, M. Dumas, and J. Berx, "Discovery, simulation, and optimization of business processes with differentiated resources," *Inf. Syst.*, vol. 120, p. 102289, 2024.

[47] L. Kirchdorfer, R. Blümel, T. Kampik, H. van der Aa, and H. Stuckenschmidt, "Agentsimulator: An agent-based approach for data-driven business process simulation," in *6th International Conference on Process Mining, ICPM 2024, Kgs. Lyngby, Denmark, October 14-18, 2024*. IEEE, 2024, pp. 97–104.

[48] J.-R. Rehse, S. J. Leemans, P. Fettke, and J. M. E. v. d. Werf, "On process discovery experimentation," *ACM Transactions on Software Engineering and Methodology*, 2024.

[49] J. Cremerius, L. Pufahl, F. Klessascheck, and M. Weske, "Event log generation in MIMIC-IV research paper," in *ICPM Workshops*, ser. Lecture Notes in Business Information Processing, vol. 468. Springer, 2022, pp. 302–314.

[50] M. L. van Eck, N. Sidorova, and W. M. P. van der Aalst, "Enabling process mining on sensor data from smart products," in *RCIS*. IEEE, 2016, pp. 1–12.

[51] S. J. J. Leemans and N. Tax, "Causal reasoning over control-flow decisions in process models," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 13295. Springer, 2022, pp. 183–200.

[52] Z. D. Bozorgi, I. Teinemaa, M. Dumas, M. L. Rosa, and A. Polyvyanyy, "Prescriptive process monitoring based on causal effect estimation," *Inf. Syst.*, vol. 116, p. 102198, 2023.

[53] Z. D. Bozorgi, M. Dumas, M. L. Rosa, A. Polyvyanyy, M. Shoush, and I. Teinemaa, "Learning when to treat business processes: Prescriptive process monitoring with causal inference and reinforcement learning," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 13901. Springer, 2023, pp. 364–380.

[54] S. J. J. Leemans, L. L. Mannel, and N. Sidorova, "Significant stochastic dependencies in process models," *Inf. Syst.*, vol. 118, p. 102223, 2023.

[55] J. J. Koorn, X. Lu, H. Leopold, N. Martin, S. Verboven, and H. A. Reijers, "Mining statistical relations for better decision making in healthcare processes," in *ICPM*. IEEE, 2022, pp. 32–39.

[56] J. J. Koorn, X. Lu, H. Leopold, and H. A. Reijers, "From action to response to effect: Mining statistical relations in work processes," *Inf. Syst.*, vol. 109, p. 102035, 2022.

[57] A. Augusto, R. Conforti, M. Dumas, M. L. Rosa, F. M. Maggi, A. Marrella, M. Mecella, and A. Soo, "Automated discovery of process models from event logs: Review and benchmark," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 686–705, 2019.

[58] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from event logs - A constructive approach," in *Petri Nets*, ser. Lecture Notes in Computer Science, vol. 7927. Springer, 2013, pp. 311–329.

[59] S. K. L. M. vanden Broucke and J. D. Weerdt, "Fodina: A robust and flexible heuristic process discovery technique," *Decis. Support Syst.*, vol. 100, pp. 109–118, 2017.

[60] A. Burke, S. J. J. Leemans, and M. T. Wynn, "Discovering stochastic process models by reduction and abstraction," in *Petri Nets*, ser. Lecture Notes in Computer Science, vol. 12734. Springer, 2021, pp. 312–336.

[61] S. J. J. Leemans, T. Li, M. Montali, and A. Polyvyanyy, "Stochastic process discovery: Can it be done optimally?" in *Advanced Information Systems Engineering - 36th International Conference, CAiSE 2024, Limassol, Cyprus, June 3-7, 2024, Proceedings*, ser. Lecture Notes in Computer Science, G. Guizzardi, F. M. Santoro, H. Mouratidis, and P. Soffer, Eds., vol. 14663. Springer, 2024, pp. 36–52. [Online]. Available: https://doi.org/10.1007/978-3-031-61057-8_3

[62] H. Alkhammash, A. Polyvyanyy, and A. Moffat, "Stochastic directly-follows process discovery using grammatical inference," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 14663. Springer, 2024, pp. 87–103.

[63] O. A. Hundogan, B. J. Verhoef, P. Theeven, H. A. Reijers, and X. Lu, "Reinforcement learning for optimizing responses in care processes," *Data Knowl. Eng.*, vol. 157, p. 102412, 2025.

[64] M. Camargo, M. Dumas, and O. G. Rojas, "Simod: A tool for automated discovery of business process simulation models," in *BPM Demonstration Track*, ser. CEUR Workshop Proceedings, vol. 2420. CEUR-WS.org, 2019, pp. 139–143.

[65] O. López-Pintado, I. Halenok, and M. Dumas, "Prosimos: Discovering and simulating business processes with differentiated resources," in *EDOC Demonstrations Track*, ser. Lecture Notes in Business Information Processing, vol. 466. Springer, 2022, pp. 346–352.

[66] A. Djedovic, A. Karabegovic, Z. Avdagic, and S. Omanovic, "Innovative approach in modeling business processes with a focus on improving the allocation of human resources," *Mathematical problems in engineering*, vol. 2018, 2018.

[67] O. López-Pintado and M. Dumas, "Business process simulation: Probabilistic modeling of intermittent resource availability and multitasking behavior," *Inf. Syst.*, vol. 127, p. 102471, 2025.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

[68] N. Martin, B. Depaire, and A. Caris, "The use of process mining in business process simulation model construction - structuring the field," *Bus. Inf. Syst. Eng.*, vol. 58, no. 1, pp. 73–87, 2016.

[69] O. López-Pintado and M. Dumas, "Business process simulation with differentiated resources: Does it make a difference?" in *BPM*, ser. Lecture Notes in Computer Science, vol. 13420. Springer, 2022, pp. 361–378.

[70] W. M. Van der Aalst, J. Nakatumba, A. Rozinat, and N. Russell, "Business process simulation," *Handbook on Business Process Management 1: Introduction, Methods, and Information Systems*, pp. 313–338, 2010.

[71] W. M. P. van der Aalst, A. Adriansyah, and B. F. van Dongen, "Replaying history on process models for conformance checking and performance analysis," *WIREs Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 182–192, 2012.

[72] A. Rozinat and W. M. P. van der Aalst, "Conformance checking of processes based on monitoring real behavior," *Inf. Syst.*, vol. 33, no. 1, pp. 64–95, 2008.

[73] S. J. J. Leemans, A. F. Syring, and W. M. P. van der Aalst, "Earth movers' stochastic conformance checking," in *BPM Forum*, ser. Lecture Notes in Business Information Processing, vol. 360. Springer, 2019, pp. 127–143.

[74] T. Li, S. J. J. Leemans, and A. Polyvyanyy, "The jensen-shannon distance metric for stochastic conformance checking," in *ICPM workshops*, ser. LNBIP, vol. to appear, 2024.

[75] S. J. J. Leemans and A. Polyvyanyy, "Stochastic-aware conformance checking: An entropy-based approach," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 12127. Springer, 2020, pp. 217–233.

[76] H. Alkhammash, A. Polyvyanyy, A. Moffat, and L. García-Bañuelos, "Entropic relevance: A mechanism for measuring stochastic process models discovered from event data," *Inf. Syst.*, vol. 107, p. 101922, 2022.

[77] G. Bergami, F. M. Maggi, M. Montali, and R. Peñaloza, "Probabilistic trace alignment," in *3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021*, C. D. Ciccio, C. D. Francescomarino, and P. Soffer, Eds. IEEE, 2021, pp. 9–16.

[78] S. J. J. Leemans, W. M. P. van der Aalst, T. Brockhoff, and A. Polyvyanyy, "Stochastic process mining: Earth movers' stochastic conformance," *Inf. Syst.*, vol. 102, p. 101724, 2021.

[79] C. Pitsch, T. Brockhoff, J. N. Adams, S. J. J. Leemans, L. A. Celi, and W. M. P. van der Aalst, "Hypothesis testing for processes," in *ICPM*. IEEE, 2025, pp. 1–8.

[80] X. Lu, D. Fahland, and W. M. P. van der Aalst, "Conformance checking based on partially ordered event data," in *Business Process Management Workshops - BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers*, ser. Lecture Notes in Business Information Processing, F. Fournier and J. Mendling, Eds., vol. 202. Springer, 2014, pp. 75–88.

[81] X. Lu, R. Mans, D. Fahland, and W. M. P. van der Aalst, "Conformance checking in healthcare based on partially ordered event data," in *ETFA*. IEEE, 2014, pp. 1–8.

[82] M. de Leoni, G. Lanciano, and A. Marrella, "Aligning partially-ordered process-execution traces and models using automated planning," in *ICAPS*. AAAI Press, 2018, pp. 321–329.

[83] H. van der Aa, H. Leopold, and M. Weidlich, "Partial order resolution of event logs for process conformance checking," *Decis. Support Syst.*, vol. 136, p. 113347, 2020.

[84] E. Bogdanov, I. Cohen, and A. Gal, "Conformance checking over stochastically known logs," in *International Conference on Business Process Management*. Springer, 2022, pp. 105–119.

[85] I. Donadello, P. Felli, C. Innes, F. M. Maggi, and M. Montali, "Conformance checking of fuzzy logs against declarative temporal specifications," in *BPM*, ser. Lecture Notes in Computer Science, vol. 14940. Springer, 2024, pp. 39–56.

[86] P. Felli, A. Gianola, M. Montali, A. Rivkin, and S. Winkler, "Multi-perspective conformance checking of uncertain process traces: An smt-based approach," *Eng. Appl. Artif. Intell.*, vol. 126, p. 106895, 2023.

[87] M. Pegoraro, M. S. Uysal, T. Hülsmann, and W. M. P. van der Aalst, "Uncertain case identifiers in process mining: A user study of the event-case correlation problem on click data," in *BPMDS/EMMSAD@CAiSE*, ser. Lecture Notes in Business Information Processing, vol. 450. Springer, 2022, pp. 173–187.

[88] T. Baier, J. Mendling, and M. Weske, "Bridging abstraction layers in process mining," *Inf. Syst.*, vol. 46, pp. 123–139, 2014.

[89] A. Lepsien, M. Pegoraro, F. Fonger, D. Langhammer, M. Aleknonytė-Resch, and A. Koschmider, "Ranking the top-k realizations of stochastically known event logs," 2024. [Online]. Available: https://arxiv.org/abs/2410.00067

[90] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. F. Sani, A. Koschmider, F. Mannhardt, S. N. von Voigt, M. Rafiei, and L. von Waldthausen, "Privacy and confidentiality in process mining: Threats and research challenges," *ACM Trans. Manag. Inf. Syst.*, vol. 13, no. 1, pp. 11:1–11:17, 2022.

[91] A. Burke, S. J. J. Leemans, and M. T. Wynn, "Stochastic process discovery by weight estimation," in *ICPM Workshops*, ser. Lecture Notes in Business Information Processing, vol. 406. Springer, 2020, pp. 260–272.

[92] S. J. J. Leemans, T. Li, and J. N. van Detten, "Ebi - a stochastic process mining framework," in *ICPM Doctoral Consortium / Demo*, ser. CEUR Workshop Proceedings, vol. 3783. CEUR-WS.org, 2024.

[93] A. Senderovich, A. Shleyfman, M. Weidlich, A. Gal, and A. Mandelbaum, "To aggregate or to eliminate? optimal model simplification for improved process performance prediction," *Inf. Syst.*, vol. 78, pp. 96–111, 2018.

[94] D. Bayomie, C. D. Ciccio, M. L. Rosa, and J. Mendling, "A probabilistic approach to event-case correlation for process mining," in *ER*, ser. Lecture Notes in Computer Science, vol. 11788. Springer, 2019, pp. 136–152.

[95] D. Bayomie, C. D. Ciccio, and J. Mendling, "Event-case correlation for process mining using probabilistic optimization," *Inf. Syst.*, vol. 114, p. 102167, 2023.

[96] D. Bayomie, K. Revoredo, C. D. Ciccio, and J. Mendling, "Improving accuracy and explainability in event-case correlation via rule mining," in *ICPM*. IEEE, 2022, pp. 24–31.

[97] T. Baier, C. D. Ciccio, J. Mendling, and M. Weske, "Matching events and activities by integrating behavioral aspects and label analysis," *Softw. Syst. Model.*, vol. 17, no. 2, pp. 573–598, 2018.

[98] E. Bogdanov, I. Cohen, and A. Gal, "Sktr: Trace recovery from stochastically known logs," in *2023 5th International Conference on Process Mining (ICPM)*. IEEE, 2023, pp. 49–56.

[99] M. de Leoni and W. M. P. van der Aalst, "Data-aware process mining: discovering decisions in processes using alignments," in *SAC*. ACM, 2013, pp. 1454–1461.

[100] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Decision mining revisited - discovering overlapping rules," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 9694. Springer, 2016, pp. 377–392.

[101] F. Meneghello, C. D. Francescomarino, C. Ghidini, and M. Ronzani, "Runtime integration of machine learning and simulation for business processes: Time and decision mining predictions," *Inf. Syst.*, vol. 128, p. 102472, 2025.

[102] D. Chapela-Campa and M. Dumas, "Enhancing business process simulation models with extraneous activity delays," *Information Systems*, vol. 122, p. 102346, 2024.

[103] M. de Leoni, F. Vinci, S. J. J. Leemans, and F. Mannhardt, "Investigating the influence of data-aware process states on activity probabilities in simulation models: Does accuracy improve?" in *BPM*, ser. Lecture Notes in Computer Science, vol. 14159. Springer, 2023, pp. 129–145.

[104] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley, "Uncertainty quantification using neural networks for molecular property prediction," *J. Chem. Inf. Model.*, vol. 60, no. 8, pp. 3770–3780, 2020.

[105] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *NIPS*, 2017, pp. 4765–4774.

[106] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Commun. ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[107] T. Fehrer, D. A. Fischer, S. J. J. Leemans, M. Röglinger, and M. T. Wynn, "An assisted approach to business process redesign," *Decis. Support Syst.*, vol. 156, p. 113749, 2022.

[108] Z. D. Bozorgi, I. Teinemaa, M. Dumas, M. L. Rosa, and A. Polyvyanyy, "Process mining meets causal machine learning: Discovering causal rules from event logs," in *ICPM*. IEEE, 2020, pp. 129–136.

[109] M. S. Qafari and W. M. P. van der Aalst, "Case level counterfactual reasoning in process mining," in *CAiSE Forum*, ser. Lecture Notes in Business Information Processing, vol. 424. Springer, 2021, pp. 55–63.

[110] A. J. Alaee, M. Weidlich, and A. Senderovich, "Data-driven decision support for business processes: Causal reasoning and discovery," in *BPM (Forum)*, ser. Lecture Notes in Business Information Processing, vol. 526. Springer, 2024, pp. 90–106.

[111] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

[112] A. Polyvyanyy, H. Alkhammash, C. D. Ciccio, L. García-Bañuelos, A. A. Kalenkova, S. J. J. Leemans, J. Mendling, A. Moffat, and M. Weidlich, "Entropia: A family of entropy-based conformance checking measures for process mining," in *ICPM Doctoral Consortium / Tools*, ser. CEUR Workshop Proceedings, vol. 2703. CEUR-WS.org, 2020, pp. 39–42.
[113] D. Geurtjens and X. Lu, "FoldA: Computing partial-order alignments using directed net unfoldings," in *BPM*, ser. Lecture Notes in Computer Science, vol. 16044. Springer, 2025, pp. 126–143.
[114] M. Matyash, A. Gal, and A. Senderovich, "DDTR: diffusion denoising trace recovery," in *7th International Conference on Process Mining, ICPM 2025, Montevideo, Uruguay, October 20-24, 2025*. IEEE, 2025, pp. 1–8.
[115] A. T. Burke, S. J. Leemans, M. T. Wynn, W. M. van der Aalst, and A. H. ter Hofstede, "A chance for models to show their quality: Stochastic process model-log dimensions," *Information Systems*, p. 102382, 2024.



Sander J.J. Leemans is a professor at the Rheinisch-Westfälische Technische Hochschule University in Aachen, Germany. His research interests include process mining, process discovery, conformance checking, stochastic process mining, and robotic process automation. In particular, he specialises in making solid academic techniques available to end-users, analysts and industry partners. He teaches business process management, business process modelling and business process improvement.



Xixi Lu is an assistant professor in the Information and Computing Sciences department at Utrecht University. Her research lies at the intersection of process mining and artificial intelligence. She develops methods and algorithms to analyze, predict, and optimize complex processes using event data. She has published more than 50 papers in peer-reviewed conferences and journals, with applications in domains including healthcare and auditing.



David Chapela obtained his PhD in Computer Science from the University of Santiago de Compostela in 2021. He is an Assistant Professor of Information Systems at the University of Tartu. His research focuses on process mining and business process management, specifically on data-driven techniques for business process simulation and optimization.



Claudio Di Ciccio is an associate professor at the Department of Information and Computing Science of Utrecht University (Netherlands). He received a PhD in Computer Science and Engineering in 2013 at Sapienza. His research interests include Process Mining, Formal Methods, Automated Reasoning in AI, and Blockchain Technologies. He is a member of the Steering Committee of the IEEE Task Force on Process Mining. PC member of numerous prestigious conferences in his field including IJCAI, ECAI, BPM and CAiSE, he served as a PC chair of ICPM in 2021, BPM in 2022, and of the Blockchain Forum at BPM in 2019 and 2024. Also, he was a General Chair of ICPM 2023. For more information and contact details, visit http://claudio.diciccio.net/



Izack Cohen is a researcher at Bar-Ilan University, specializing in operations research, optimization, and process mining. His work focuses on data-aware process modelling, analysis and optimization.



Benoît Depaire is a professor of Business Informatics at Hasselt University, Belgium and affiliated with the Digital Future Lab. His research interests focus on machine learning, statistics and other computational methods to gain valuable business insights from data, as well as on the philosophy and methodology of empirical research on algorithm design.



Stephan Fahrenkrog-Petersen is an Assistant Professor in Information Systems & Process Science at the University of Liechtenstein. His research focus is Data Engineering for Process Science and he published in venues such as AAAI, ICDCS, Information Systems, CAiSE and BPM. He was as a Research Group Lead at the Weizenbaum Institute and got his PhD from Humboldt-Universität zu Berlin.



Avigdor Gal is the Benjamin and Florence Free Chaired Professor of Data Science at the Faculty of Data & Decision Sciences, Technion - Israel Institute of Technology. Gal's research focuses on data integration and process management and mining with about 200 publications, including multiple best paper and test-of-time awards.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402



Massimiliano de Leoni is Associate Professor at the University of Padua. He is an active member of the IEEE Task Force on Process Mining, and among the coordinators of BPMItaly, the BPM-related working group within the Italian Association for Artificial Intelligence (AIxIA). His research spans Business Process Management, Process Mining, and AI-driven process improvement. He has received several awards, including the BPM 2023 Test-of-Time Award.



Henrik Leopold is a Full Professor at the Kühne Logistics University (KLU) in Hamburg, Germany. His research focuses on developing innovative methods for analyzing, automating, and improving business processes using Process Mining and Artificial Intelligence. He has authored over 100 publications in academic journals and conferences and collaborates closely with leading industry partners to bridge the gap between research and real-world applications.



Orlenys López Pintado is a Research Fellow in Information Systems at the University of Tartu, Estonia. His research interests include business process management (BPM), business process simulation, optimization, and blockchain. His current area of research is data-driven discovery, simulation, and optimization of business processes. He received a PhD in Computer Science in 2020 at the University of Tartu.



Felix Mannhardt received the Ph.D. degree from Eindhoven University of Technology, Eindhoven, The Netherlands, in 2018. He is currently an Assistant Professor with Eindhoven University of Technology and the Principal Data Scientist of KIT-AR Ltd. His research interests include process mining in various settings related to low-level event data, trust, fairness, and privacy concerns



Fabrizio Maria Maggi is an associate professor at the Free University of Bozen-Bolzano. His areas of research span process modeling and process analysis. He has coauthored more than 80 articles on process mining, automated revision of business process models, declarative business process modeling, monitoring of business constraints at runtime, service oriented architectures, service choreographies, and service composition.



Niels Martin is an Assistant Professor in the research group Business Informatics (Faculty of Business Economics, Hasselt University, Belgium). His research interests include process analytics in healthcare, work organisation mining, human resource mining, data-driven process simulation and process data quality. His research has been published in highly regarded journals such as Information & Management, Decision Support Systems and Business & Information Systems Engineering.



Marco Montali is a Full Professor in Computer Engineering at the Free University of Bozen-Bolzano, with a focus on artificial intelligence and process science. His research is in artificial intelligence and information systems for the modelling, analysis, and mining of processes and agents.



Marco Pegoraro Marco Pegoraro is a research assistant in the Chair of Process and Data Science, at RWTH Aachen University. His main research activity pertains the extension of process mining techniques to uncertain, non-deterministic and probabilistic data. His other research interests are predictive process monitoring and the application of process mining to clinical and medical data, optimizing operations in healthcare settings.



Artem Polyvyanyy is a Professor in the School of Computing and Information Systems, Faculty of Engineering and Information Technology, at the University of Melbourne (Australia). He leads the Process Science and Technology research group and serves as the Deputy Head of School (Academic). Artem's research and teaching interests include computing, distributed systems, process mining, process science, artificial intelligence, and algorithms.



Arik Senderovich is an Assistant Professor at the School of Information Technologies (ITEC) at York University in Toronto, Canada. Arik's work has been published in leading conferences including AAAI, BPM, ICAPS, ICPM, and the Winter Simulation Conference (WSC). He has also authored numerous papers in top-tier journals such as Information Systems and INFORMS Journal on Computing. Additionally, Arik has served as a Guest Editor for Information Systems and was the Program Chair for ICPM 2023.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402



Sam Verboven is assistant professor at the Vrije Universiteit Brussel at the Department of Business Technology and Operations (BUTO). He is a member of the Data Analytics Lab research group and cross-faculty member of the AI Lab. Research-wise, Sam focuses on developing innovative analytical tools based on causality, with applications in marketing, law, and healthcare processes



Akio Watanabe received his M.E. degree in Computer Science from the University of Electro-Communications in 2012. He is currently a researcher at NTT Network Service Systems Laboratories, Nippon Telegraph and Telephone Corporation (NTT). His research interests include process mining, machine learning, and system management.



Matthias Weidlich is professor and chair of databases and information systems with the Department of Computer Science, Humboldt-Universität zu Berlin. His research focuses on process-oriented and event-based information systems, and his results appear regularly in premier conferences (SIGMOD, VLDB, ICDE) and journals (IEEE Transactions on Knowledge and Data Engineering, Inf. Sys., VLDBJ) in the field. He serves as EiC for the Information Systems journal and on the Steering Committees of the ACM DEBS and BPM conference series.



Jan-Martijn E.M. van der Werf is an associate professor at the Department of Computer Science at Utrecht University, where he is also the director of education for the Bachelor of Information Science. His research focuses on modeling and analyzing behavior in software architecture of software products. Van der Werf received his PhD from two universities: Eindhoven University of Technology and Humboldt University of Berlin.

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

Pre-print copy of the manuscript published in **journal**
identified by DOI: 10.1109/TKDE.2026.3715402

This document is a pre-print copy of the manuscript
(Leemans et al. 2026)
published in **journal**.

The final version of the paper is identified by DOI: 10.1109/TKDE.2026.3715402

References

Leemans, Sander J. J., Xixi Lu, David Chapela-Campa, Claudio Di Ciccio, Izack Cohen, Benoît Depaire, Stephan A. Fahrenkrog-Petersen, Avigdor Gal, Massimiliano de Leoni, Henrik Leopold, Orlenys López-Pintado, Felix Mannhardt, Fabrizio Maria Maggi, Niels Martin, Marco Montali, Marco Pegoraro, Artem Polyvyanyy, Arik Senderovich, Sam Verboven, Akio Watanabe, Matthias Weidlich, and Jan Martijn E. M. van der Werf (2026). "Stochastic Process Mining: Characteristics and Challenges". In: *IEEE Trans. Knowl. Data Eng.* 38.10, pp. 7061–7079. DOI: 10.1109/TKDE.2026.3715402.

BibTeX

```
@Article{         Leemans.etal/TKDE2026:StochasticProcessMining,
  author        = {Leemans, Sander J. J. and Lu, Xixi and Chapela{-}Campa,
                   David and Di Ciccio, Claudio and Cohen, Izack and Depaire,
                   Beno{\^{i}}t and Fahrenkrog{-}Petersen, Stephan A. and Gal,
                   Avigdor and de Leoni, Massimiliano and Leopold, Henrik and
                   L{\'{o}}pez{-}Pintado, Orlenys and Mannhardt, Felix and
                   Maggi, Fabrizio Maria and Martin, Niels and Montali, Marco
                   and Pegoraro, Marco and Polyvyanyy, Artem and Senderovich,
                   Arik and Verboven, Sam and Watanabe, Akio and Weidlich,
                   Matthias and Jan Martijn E. M. van der Werf},
  journal       = {{IEEE} Trans. Knowl. Data Eng.},
  title         = {Stochastic Process Mining: Characteristics and
                   Challenges},
  year          = {2026},
  number        = {10},
  pages         = {7061--7079},
  volume        = {38},
  doi           = {10.1109/TKDE.2026.3715402},
  keywords      = {Stochastic process mining; Process models; Process
                   discovery; Levels of uncertainty}
}
```