

From analysis to findings: How do process mining analysts discover results?

Francesca Zerbato^a, Lisa Zimmermann^b, Katerina Vrotsou^c, Barbara Weber^b

^a Eindhoven University of Technology, Eindhoven, The Netherlands

^b University of St. Gallen, St. Gallen, Switzerland

^c Linköping University, Norrköping, Sweden

ARTICLE INFO

Keywords:

Process mining
User interactions
Visual analytics
Process mining results
User behavior

ABSTRACT

Process mining involves analyzing event data from business process executions to uncover valuable insights. Although obtaining meaningful results is crucial for any process mining initiative, there is still little understanding of how process analysts derive these insights. In this paper, we fill this gap by characterizing findings of process mining analysis, the processes that lead to these findings, and the role of process mining expertise in guiding these processes. To this end, we leverage empirical data from a study with process mining analysts, including user interactions from process mining tools and inference steps from think-aloud protocols. Our empirical insights provide a comprehensive understanding of how analysts interact with process mining tools, highlighting approaches that lead to valuable findings. The results of our analysis lay the groundwork for the design of tools and visualizations that can support process analysts in their analysis and reasoning processes.

1. Introduction

Information systems produce large volumes of data that are increasingly used by organizations to gain insights into their (business) processes and inform strategic decisions [1]. Among the many analytical techniques available, process mining (PM) techniques stand out for their capability to analyze and visualize data from a process-oriented perspective, capturing the temporal sequence of events for distinct instances [2]. By analyzing event logs with the help of various tools, process analysts can gain transparency into the actual flow and performance of processes, identify inefficiencies, and recommend improvements [3].

Like other analytical disciplines [4], PM inherently relies on the active engagement of users, such as *process analysts*, to derive insights from raw event data [5]. In goal-oriented initiatives, process analysts typically focus on finding answers to specific questions or aspects to improve [6]. Instead, in exploratory settings, analysts tend to follow an open-ended approach, often starting from the process model discovered from the event log and looking for “something interesting” that can reveal meaningful patterns or anomalies [7].

Although all of these *findings* of a PM analysis can be of great value to an organization, the journey from raw event data to actionable insights is far from straightforward [8]. To arrive at findings, process analysts must explore different scenarios, generate hypotheses that can explain their observations, and, iteratively, find evidence in the data that can confirm or disprove their hypotheses [9,10]. This so-called

process of process mining is complex and knowledge-intensive, as each finding that emerges during the analysis can trigger an “aha” moment that guides the analyst in their next analysis step.

Understanding these steps is crucial: it provides valuable insights into the findings themselves as well as into how process analysts arrive at these findings through their interactions with process mining tools and their reasoning [11]. Such insights can facilitate knowledge sharing among analysts regarding their strategies [12,13], and inform the design of process mining tools and visualizations that better support analytical reasoning [14].

In this paper, we aim to enhance our understanding of the process of process mining by focusing on how process analysts arrive at findings, i.e., the units of information that they identify as conclusive results of their analysis.

To this end, we define the following three research questions (RQs).

RQ1 How do findings manifest in process mining analyses?

With this question, we aim to characterize the types and quality of single findings, focusing on what process analysts find worth reporting as a result as well as how these findings develop during the analysis process.

RQ2 What are the analytical and inference processes leading to these findings?

With this question, we explore the processes that lead to findings, examining both the analytical steps performed within PM tools and the inference steps that guide the analysts’ reasoning flows.

* Corresponding author.

E-mail address: f.zerbato@tue.nl (F. Zerbato).

<https://doi.org/10.1016/j.is.2025.102596>

Received 31 July 2024; Received in revised form 18 July 2025; Accepted 23 July 2025

Available online 5 August 2025

0306-4379/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

To gain a complete picture of how analysts navigate and interpret data to arrive at (relevant) findings, we break this question into three parts. First, we examine how PM tools are used and how their use is linked to findings (RQ2.1). Next, we explore the analytical steps leading to findings, focusing on the sequences of actions carried out by analysts using these tools (RQ2.2). Finally, we explore how inference steps contribute to discovering findings during the analysis (RQ2.3).

RQ3 What is the role of process mining expertise in the discovery of findings?

With this question, we investigate whether and to what extent PM expertise influences the type and quality of the findings discovered and the analytical and inference steps leading to them.

To address RQ1–RQ3, we conducted a user study with 22 process analysts who have varying levels of PM expertise. Each participant analyzed a real-life event log using a process mining tool of their choice while thinking aloud [15]. This approach allowed us to collect multiple sources of empirical data, which we combined into a comprehensive *multi-perspective* dataset. This dataset includes both observable behavior, such as the analysts' high-level interactions with PM tools, and their inference processes, reflected in their verbalized reasoning. The richness of our dataset allowed us to investigate how analysts arrive at findings from different angles, combining tool interactions with their underlying thought processes.

Our contributions include a characterization of findings in PM analyses and an overview of the analytical steps that analysts take to derive these insights. We also explore the role of PM expertise in this discovery process. Previous research has primarily relied on interviews to study analysis strategies [10], challenges faced by individual analysts [8], and PM adoption in organizations [16]. To the best of our knowledge, this is the first study to use behavioral data to investigate a complete PM analysis, from data understanding to the discovery of findings. Moreover, our multi-perspective dataset allows us to study the behavior of process analysts on a fine-grained level, which cannot be achieved from retrospective reporting.

The results of this work can guide the development of interactive visualizations that align with the inference processes and analytical workflows of process analysts, potentially leading to more efficient decision-making. Moreover, by addressing the analytical challenges presented by such a rich but complex dataset, our research can inform joint efforts between the PM and visual analytics (VA) communities to develop tools that support the analytical capabilities of process analysts. This collaboration can lead to more effective, intuitive, and powerful VA tools for better insights and decision-making in PM.

2. Background and related work

We begin with a brief introduction to process mining in Section 2.1. We then define the term *finding* as used in our work in Section 2.2, and discuss relevant studies in both PM and VA in Section 2.3.

2.1. Process mining: A short introduction

Process mining is an interdisciplinary field at the intersection of data science and process science. Its main goal is “to improve operational processes through the systematic use of event data” [2]. As such, it provides a set of techniques that take event records of process executions as input and serve process discovery, conformance checking or process enhancement use cases. Such event records capture the execution time of relevant events for a specific instance of a process. Events can be linked with attributes related to the event itself (e.g., an amount paid, can be linked to a payment event) or to the instance to which the event belongs (e.g., the type of a car, linked to a traffic fine ID that is processed). In PM, a structured collection of events related to a set of process instances is called an event log. Event logs contain information

on all the events (also called “activities”) executed for a specific process instance (also called “case”) and are the main input for PM analyses.

To support process analysts in their work, there are many commercial and open-source tools available. PM tools are often interactive and provide many ways to analyze a process through visualizations such as process models, bar charts, and density plots [17]. One commonly used representation in PM is the directly-follows graph (DFG), which depicts the process flow in a graph-like structure [18] (see an example in Fig. 1). Many tools feature a Process Explorer, an interactive interface that visualizes the DFG, often enhanced with frequency or performance metrics. Analysts can interact with the DFG by filtering out infrequent behavior, highlighting specific events, or adjusting the level of detail. Many tools also include a Variant Explorer, which groups cases in the event log based on their execution sequence, allowing analysts to compare different trace variants. Many other (visual) analysis operations can be applied to event logs, depending on the tool used [19,20].

Although PM tools provide a variety of functions, it is up to the analyst to perform the “right” tasks and iterate through different analysis phases [21]. As such, the success of a PM analysis depends on the analyst's technical expertise, their understanding of the domain, and their active collaboration with process stakeholders [22].

2.2. Findings and insights

A finding is “information that is discovered as the result of researching into something or examining a specific problem”.¹ Many analytical disciplines, including PM and VA, aim to develop methods and tools that enable analysts to extract meaningful findings from data to address specific domain problems. VA, in particular, combines automated analysis techniques with interactive visualizations to help analysts synthesize information, discover unexpected patterns, and effectively communicate insights [23]. Given this strong emphasis on deriving insights from data using interactive tools and techniques, in this work we refer to the notion of *finding* from VA.

In this field, different definitions of *finding* exist, which, in some cases, overlap with the definition of *insight* [24]. For Sacha et al. [9] a finding is “an interesting observation made by an analyst using the VA system” that originates from exploration and is independent of the problem domain. Instead, an insight results from a verification loop and comes with “an interpretation in the problem domain” driven by domain knowledge. Saraiya et al. [25] define an insight as “an individual observation about the data”, while North [26] refers to insights as answers to questions about the data.

In this paper, we define “findings” as conclusive insights or potential answers to an analysis question that process analysts explicitly recognize as significant. Since our definition assumes that a finding is linked to the problem domain, it aligns well with the definition of insight by Sacha [9] and North [26]. The main difference is that we also consider potential answers as findings since, in our study, we document all the answers reported by participants, regardless of their correctness.

The findings that can be obtained with PM vary depending on the specific question being asked. This is well-reflected in case study reports. For example, Klinkmüller et al. [20] analyzed reports from Business Process Intelligence (BPI) Challenges² and found that process discovery is the most popular domain problem in PM. As a consequence, process models and the control flow patterns derived from them are among the most commonly reported findings. This aligns with the work of Stein Dani et al. [6], which highlights that common findings include discovered process models, but also high waiting times and data quality issues. However, as noted by Koorn et al. [27], there is still little

¹ Finding. In Oxford Learner's Dictionary <https://www.oxfordlearnersdictionaries.com/us/definition/english/finding>.

² BPI Challenges: <https://www.tf-pm.org/competitions-awards/bpi-challenge>.

agreement on what constitutes a meaningful outcome of a PM analysis, as this largely depends on the analysis question, the techniques used, and the needs of the target audience of these findings. This diversity in the reported results also reflects the variability of the spontaneous findings analysts gain during their analysis and the steps they follow to derive them.

2.3. Analyzing the behavior of process analysts

Our paper is related to work in the area of PM and, more in general, VA that investigates the behavior of (process) analysts.

Closely related to this paper, are works using interaction and think-aloud data to investigate the behavior of individual PM analysts. In our previous work, we investigated initial exploration practices using interaction and think-aloud data collected from 12 process analysts [7]. Our study focused on the processes followed by analysts in an open exploration setting, i.e., before an analysis question is posed. We identified three patterns of behavior among analysts: (i) those who focus on understanding data in a top-down manner, anchored in domain knowledge, (ii) those who start from the raw data and focus on descriptives, and (iii) those who compare different subsets of the log to make sense of the data. The work by Sorokina et al. [11] presents the PEM4PPM model, which captures the process of process mining from a cognitive perspective. The model is developed from error minimization theory and then refined and validated using interaction and think-aloud data collected from PM students and experienced analysts. The authors derived four cognitive strategies followed by analysts based on whether they engage in data interpretation, hypothesis making, and hypothesis testing steps. A key result of their analysis is that experienced analysts tend to go through all these steps, which leads to higher-quality results. Van Daele and Jannsenswillen [28] conducted an empirical study with 11 students completing 33 assignments using R,³ examining the steps and actions taken during exploratory data analysis. The authors identified 20 actions across preparation, analysis, and debugging steps, including “executing code”, “manipulating data” and “evaluating results”. Interestingly, the “evaluate results” action was often followed data summary or plot preparation/viewing actions, suggesting that analysts need to summarize and visualize the data before deriving results. The discovered strategies align well with strategies and tasks discovered in VA analysis reports, for example the strategies in [29] or the goals in [30], which further motivates the need for collaboration between researchers in these fields.

Although these works shed light on various aspects of process analysts’ behavior, they either focus on a specific analysis phase, such as initial exploration [7], or examine user behavior from a single perspective in isolation, i.e., either interactions [28] or cognitive steps [11]. Furthermore, they do not characterize analysis findings or investigate the full process leading to these findings, considering both the analytical and cognitive steps taken by analysts. At last, these studies report on user interactions within a single tool, e.g., Fluxicon Disco, which may provide a limited perspective on user behavior.

A recent stream of research has focused on studying PM practices from analysis reports or expert interviews. For example, Capitán-Agudo et al. [19] examined the BPI Challenge reports to uncover common operations used to answer performance questions. Interview studies have been conducted to explore analysis strategies [10] and the challenges faced by individual analysts [8] or by organizations adopting PM [16,31]. Although these works uncover unknown aspects of PM practices, they rely on retrospective reporting and, as such, do not allow one to study the real-time unfolding of interactions and reasoning steps.

Recently, Abb and Rehse proposed a model to describe user behavior from low-level user interactions captured in UI logs [32]. In a

similar vein, Rebmann et al. [33] have provided an approach to transform such UI logs into suitable format for PM analysis. Although these works fall under the broader umbrella of *user behavior mining* [34], the authors propose methods to record and analyze data from lower-level tasks, with the ultimate goal to improve tool usability or automate task execution [35]. Our work sets itself apart both in its objective and the data sources we used. Rather than proposing a method to capture interaction data, we aim to understand the analytical and reasoning processes of analysts from an empirical perspective. Moreover, we use a multi-perspective dataset that integrates interaction traces with think-aloud data, and allows us to abstract beyond specific UI actions and tools.

In visual analytics, extensive research has focused on analyzing user interaction logs to identify patterns and reasoning strategies from interaction data [36]. Analytic provenance, in particular, has become a cornerstone of VA, guiding the creation of tools and systems for capturing, encoding, recovering, and reusing user interactions [12]. However, most of these systems consider a one-dimensional interaction log reflecting the user’s actions in a specific system and focus on extracting interaction patterns and user profiles from low-level interactions (e.g., [37,38]). Our work diverges by adopting a multi-perspective approach to gain a deeper understanding of the analysis process. Instead, our work is more closely related to the paper of Yu et al. [39], who considers three complementary interaction sequences extracted from an interaction event log: interface events reflecting the creation/deletion of an interface element, interaction events reflecting the users’ interactions with these elements, and attribute change events reflecting the contextual details of these interactions and the content of the exploration. However, their approach does not explicitly capture the hypotheses made and tested during the analysis. We address this gap by incorporating think-aloud protocols, which allow us to enrich and validate the exploration strategies derived from system interaction logs. Guo et al. [40] also conducted a similar analysis, examining findings uncovered by data analysts using a VA system. They combined quantitative and qualitative analyses of high-level user interactions from screen-capture and think-aloud data, finding that exploration actions (“show me something else”) are positively correlated with insights, whereas filtering actions (“show me something conditionally”) are negatively correlated.

In this paper, we focus on different actions specific to the behavior and reasoning of PM analysts, enhancing the understanding of the analysis process through a comprehensive multi-perspective approach.

3. Method

In this section, we first describe the design and execution of our study with process analysts (Section 3.1) and, then, report on our coding and data preparation (Section 3.2).

3.1. Study design and execution

To address RQ1–RQ3, we designed a user study observing process analysts as they engaged in a PM analysis using a mixed-methods approach [41].

Study design. Our study included an **analysis session with think-aloud** [15] and a **follow-up questionnaire**. We aimed to involve participants with varying levels of PM expertise and hands-on experience with event log analysis. To participate, individuals needed to have (i) analyzed at least one real-life event log in the two years prior to the study; (ii) knowledge of at least one PM tool among bupaR, Celonis, Disco, Pm4Py, and ProM; (iii) experience with at least one PM project where third party data was analyzed. These requirements allowed us to exclude novices, but still include participants with different backgrounds and levels of PM expertise in order to gain a broad understanding of analyst behavior.

³ R: <https://www.r-project.org/about.html>.

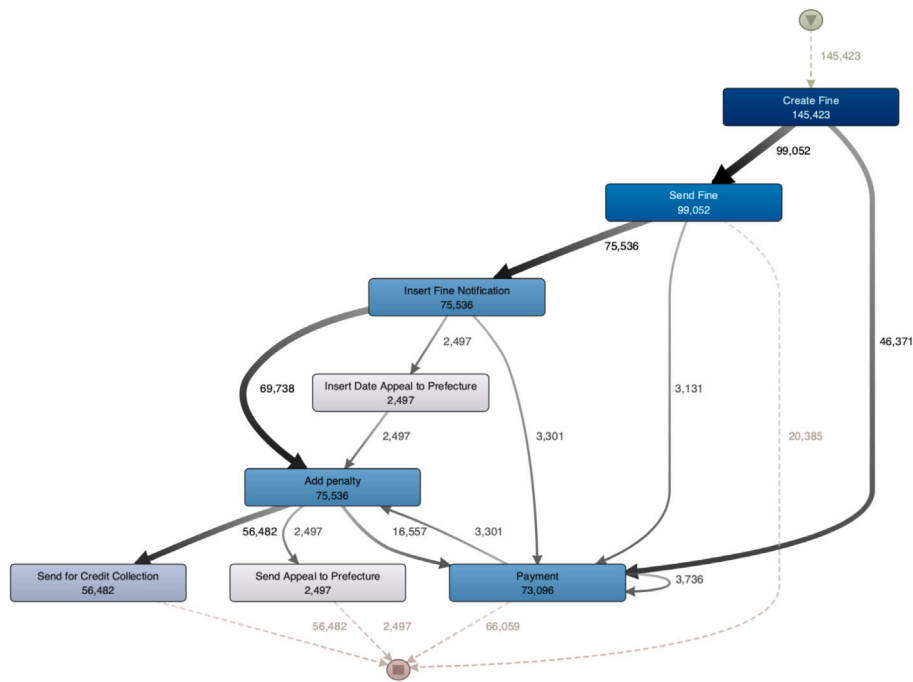


Fig. 1. Directly-follows graph derived using Fluxicon Disco after selecting the 8 most frequent variants of the RTFM log, which cover 145,423 cases out of the original 150,370.

For the analysis session, participants analyzed the road traffic fine management (RTFM) event log [42], a real-life log with 561,470 events and 150,370 cases that records the process of managing fines by the Italian police. We selected the RTFM log because it is large but relatively structured, which makes it accessible using PM tools without the need for extensive preprocessing. The (normative) process in the RTFM log begins with a *Create Fine* event, which specifies the amount of the fine to be paid. Then, event *Send Fine* triggers a notification letter to the offender, which incurs some postal expenses and whose receipt is documented by an *Insert Fine Notification* event. Offenders can pay their fine at any time, as indicated by the *Payment* event and its *paymentAmount* attribute. A *totalPaymentAmount* attribute associated with this event tracks the cumulative amount paid over all the *Payments*. If the payment is not received within 60 days, the offender receives a penalty (*Add Penalty*), which doubles the amount to be paid. Offenders can appeal against the fine to a prefecture or judge, which is recorded in the events *Insert Date Appeal to Prefecture*, *Send Appeal to Prefecture* and *Notify Result Appeal to Offender*, and *Appeal to Judge*. A *dismissal* attribute indicates if the fine is overturned. Finally, *Send for Credit Collection* indicates that unpaid fines are sent to a collection agency. An important temporal constraint in the process allows the offender to not pay if the fine is sent by the police later than 90 days after the offense. If the offender pays the full fine amount, including postal expenses and penalties, the process ends with a positive outcome. In all other cases, the fine is considered unpaid. Fig. 1 shows the directly-follows graph (DFG) discovered from the eight most frequent variants of the RTFM log.

In the RTFM log, non-payment is a central phenomenon of interest, and unpaid cases, including partial payments, constitute the majority of the data, i.e., more than 70% of the cases in the log. Therefore, we asked participants to answer the following analysis question (AQ):

“Can you identify at least three circumstances in which traffic fines are not paid in full? If possible from the data, can you provide reasons that explain such circumstances?”

The first part of the AQ was intentionally broad to allow many scenarios to fit as possible answers, while the second part aimed to encourage in-depth analysis. Some common scenarios we anticipated

as answers based on the literature [43] and our own analysis of the log included: (i) cases where offenders are exempt from paying their fine because it was sent later than 90 days, (ii) successful appeals to the prefecture or judge, (iii) partial payments due to overlooked postal expenses or penalties, and (iv) fines that remain unpaid and are sent for credit collection. From Fig. 1, it is possible to see that some of the described scenarios match frequent endpoints in the DFG.

For the analysis, participants could use a selection of commonly used PM tools: bupaR [44], Celonis,⁴ Disco,⁵ Pm4Py [45], and ProM [46]. These tools vary in terms of built-in functionality, customization options, and implementation of advanced PM techniques. We chose them because they were the most frequently used in reports from BPI Challenges. We considered BPI Challenges as a reference as they are a widely recognized event in the PM field, they appeal to both practitioners and academics, and include questions similar to our AQ.

Materials. Following the description of the RTFM log provided in [42], we created a document outlining the main activities, the data attributes, the temporal constraints and business rules to be respected by the process, and a conceptual data model of the log. This document could be referenced by participants throughout the session as a source of domain knowledge. Furthermore, we designed a follow-up questionnaire, in which participants could describe the circumstances and reasons they found in free text, skipping items if needed.

Study execution. The analysis sessions took place between May and July 2021 via virtual meetings. Before each session, we screened participants for participation requirements and collected their consent and demographic information. On the scheduled day, we introduced participants to the session and gave them access to a remote desktop environment with the study materials and tools. We instructed participants to verbalize their thoughts in a think-aloud manner as they performed the analysis, focusing on explaining *what* they were observing or doing, and *why*. Our goal was to gain insights into their reasoning steps, e.g., what they observed or what they hypothesized,

⁴ <https://www.celonis.com>.

⁵ <https://fluxicon.com/disco/>.

Table 1

Demographics of the 22 study participants. **ID** is a unique participant identifier; **Role** is the job role of the participant; **Experience** are the years of experience in process mining; **Expertise** is the self-reported process mining expertise; **# Projects** is the number of PM projects where the participant analyzed third party data; **Tool** is the tools used for the session; **Task time** is the total time spent in the analysis session. The gray background highlights participants with Basic or Average expertise, which we later refer to as non-experts.

ID	Role	Experience	Expertise	# Projects	Tool	Task time
1	Product lead	2	Good	1	Celonis	32.5
2	Research	2	Average	1	Disco, ProM	44.3
3	Professor	10	Advanced	6–10	Disco	42.2
4	Senior consultant	3.5	Good	>20	Celonis	33.3
5	Freelance consultant	8	Advanced	6–10	Disco	43.6
6	Assistant professor	8	Advanced	2–5	ProM	39.9
7	Partner relationships	1	Basic	1	Celonis	35.2
8	Product development	3	Good	10–20	Celonis	37.9
9	PhD candidate	4	Average	2–5	Disco	33.4
10	Senior consultant	4	Good	10–20	Celonis	36.8
11	Senior researcher	7	Advanced	6–10	bupaR	36.2
12	Assistant professor	8	Advanced	6–10	Disco, ProM	36.6
13	Partner relationships	0.5	Basic	1	Celonis	28.9
14	Digital consulting analyst	1	Average	2–5	Celonis	40.8
15	Product manager	3	Good	2–5	Celonis	34
16	Data analytics specialist	4	Good	2–5	bupaR	31.6
17	Digital project manager	7	Good	2–5	Disco	31.7
18	Professor	10	Average	2–5	Disco	30.9
19	Operational excellence	3	Advanced	6–10	Celonis	36.2
20	Operational excellence	5	Advanced	>20	Disco	36.2
21	Senior researcher	7	Advanced	10–20	ProM	43.4
22	Software engineer	7	Good	2–5	Pm4Py	39.4

and the decisions driving these steps, e.g., why they decided to pursue a specific analysis direction. The task was silently supervised by one author, who was available for questions about the set-up and reminded participants to think aloud but did not comment on the analysis or the provided answers. Participants were allowed to analyze the RTFM event log at their own pace, i.e., we did not limit the duration of the analysis. After the analysis, participants reported their answers in the follow-up questionnaire. For each session, we recorded the participant's screen showing the analysis in the PM tools and the synchronized audio capturing their think-aloud. This synchronization in time is essential to triangulate the different data sources and enable the combined analysis of both analytic and inference steps.

Participants. We involved 22 participants, 14 practitioners and 8 academics, from various companies and academic institutions around the world. Table 1 shows their demographics. The participants worked in a variety of industries, including software, research, insurance, food processing, and construction. Their experience with process mining ranged from less than a year to 10 years, with a median experience of 4 years.⁶ All participants had experience with PM projects, and half of them had conducted six or more projects. The majority (15/22) rated themselves as having “good” or “advanced” PM expertise.

3.2. Data preparation and coding

We recorded 805 min of analysis sessions, averaging 36.5 min per participant with a standard deviation of 4.4 min. First, we transcribed all the think-aloud utterances verbatim. Then, we prepared the data for the analysis following the four steps described below. Two authors and two external assistants conducted the coding in multiple iterations, reviewing and discussing each other's work to resolve disagreements. One author then consolidated the coding to ensure consistency across all participants.

Step 1: Creating multi-granular interaction logs. First, we created user interaction logs capturing how participants used PM tools during the analysis. To this end, we coded the **screen recordings**, i.e., we segmented them into units of interest for our analysis. Each segment has a start timestamp, an end timestamp, and a code name. This structure is similar to an event log, where each segment is an activity, and the code name describes the action performed by the analyst in the tool during a specific time frame. To identify these actions, we developed a coding scheme, inspired by Brehmer and Munzner's task typology [47]. In this work, we focus on codes derived from Munzner's HOW dimension, which describes how a task is performed. In our case, HOW describes the action of the participant from the perspective of the *tool function* they used within specific tools. Initially, we coded the videos at a fine granularity, using code names that were close to the functions available in different tools (e.g., one code name was “Follower Filter”, which is a filter for directly- and eventually-follows relations in Disco, another one was “Dotted Chart”, which refers to a specific plugin in ProM). We then proceeded in iterations, mapping functionalities across different process mining tools and creating “tool-agnostic” codes that preserved different levels of detail in their names. For example, the code “Filter/Applied Filter/Variant Filter” identifies the application of a variant filter to the event log. The hierarchy in the name of the code indicates that “Variant Filter” belongs to the category “Applied Filter”, which distinguishes between applying and removing a filter, and refers, more generally, to “Filter”. This structured approach allowed us to maintain multiple granularity levels for each segment, which we could adjust later in our analysis, depending on the patterns we observed [48]. Compared to the user interaction model in [32], we code tool functions as “UI groups” associated with “tasks”, while also reflecting multiple levels of UI hierarchy. Although our coding is more fine-grained compared to other studies [7,40], we refrained from coding low-level interactions [13] for two main reasons. First, our objective was to uncover analytical steps that align with the reasoning process. As such, we found that higher-level codes captured interaction intents better [49]. Second, since we allowed participants to use different PM tools, it would be impractical to manually code low-level interactions without prior instrumentation. However, since we did not aim to compare specific tool functions, this approach was not relevant to our goal.

⁶ At the time of data collection in 2021, having four years of industry experience was good considering the relatively young age of the discipline.

Table 2

(Top) Characteristics of a finding; Broad vs. deep, and Directed vs. Unexpected are adapted from [25]. (Bottom) Quality scale used to assess the quality of the findings and example findings from our participants.

	Value	Description
Characteristics	Broad vs. Deep	A broad finding represents an overview of the result without much detail, e.g., based on a DFG that is not examined for specific process variants or conditions. A deep finding presents a detailed, specific perspective of the result based on multiple observations that build upon each other to create depth.
	Directed vs. Unexpected	A directed finding is based on a (previously) stated intent and answers a specific question from the analyst. An unexpected finding is an insight derived from exploration or serendipitous activities that was not specifically looked for.
	Quantified	The finding is quantified, e.g., through the number of cases or events it covers.
Quality scale	Wrong	The finding reports only incorrect conclusions (e.g., misinterpretations of the data, fully invented, incorrect answers) and is not relevant. The explanation of reasons for the finding is speculative and not backed up by the data. Example: “No, notification is sent after 70... Probably after... Only 87 days afterward. So, it could be that in most cases, yes, that the offender does not know that they have received the fine.” (p13)
	Minimal	The finding reports a correct answer to the guiding question that includes a few relevant insights. Still, it is not well linked with the problem domain. The description is ambiguous and conclusions might not stem from the data. Reasons are missing or invented. No analysis of alternative results. Example: “Clearly one of the reasons why the fines are not being paid is because the offender has the intent to appeal to the prefecture and see if they can get their fine reversed” (p7)
	Adequate	The finding reports a correct answer to the guiding question that includes some relevant insights. It includes relevant details and is well-linked with the problem domain. Its description is clear and supported by evidence in the data. An explanation of reasons or alternative results are present but superficial. Example: “It seems people apply for an appeal and, if the outcome is in favor of the offender, they do not need to pay. And it’s important to note that the positive outcome seems not to be recorded in the event log. A hypothesis would be that people who have applied to an appeal did not receive any response. It’s not really easy to quantify, but this will be almost three thousand cases” (p21)
	Excellent	The finding reports a correct answer to the guiding question that includes relevant insights. It includes relevant details and is well-linked with the problem domain. The description is clear and goes in depth, and the conclusions supported by the data. An explanation of reasons and a thoughtful analysis of alternative results are present and go beyond the surface. Example: “If you look at the dismissal... we see that there are 1980 cases, which have been dismissed. So, among all cases that have been sent to the prefecture, those that have been dismissed are a little higher compared to the judge” (p11)

Step 2: Identifying the inference steps. To reconstruct the reasoning of our participants, we extracted inference steps in the form of verbal utterances from the **think-aloud** transcripts, focusing on *hypotheses made* and *hypotheses tested*. These steps allow analysts to structure their analytic process [9], often leading to a more systematic exploration of the data and higher analysis performance [11]. As such, they are critical to findings, as they indicate that analysts go beyond superficial analysis, engaging in systematic inquiry and testing steps. To identify inference steps, we mainly used the think-aloud data but we reviewed the video recordings for triangulation, especially to identify the testing steps. In this way, we could identify the object of interest for the participant and put their hypotheses into the broader context of their analysis. As part of this process, we also associated start and end timestamps to the verbal utterances we identified.

Step 3: Identifying the findings. Next we identified the *findings*, which we defined as answers to the AQ that participants explicitly recognized as significant (cf. Section 2). Here, we identified the “aha” moments when participants reported having found an answer to the question. Since such moments can be more or less explicit, depending on the participants’ ability to articulate their thoughts and confidence in their analytic approach, we triangulated the **think-aloud** transcripts, the **screen recordings**, and the answers from the **follow-up questionnaires**. In the think-aloud, we searched for expressions closely related to the AQ, including keywords like “answer”, “circumstance”, “reason”, and “not paying”, as well as phrases indicating conclusion. In the screen recordings, we focused on note-taking behavior, which often indicated that the participant had discovered something significant to report. Finally, we used the answers from the questionnaires as a source for validation, to check whether the findings we identified were indeed the “final” answers to the AQ. For example, in the think-aloud, p11 reported “Another reason are the appeals, which are approximately 3000. If we consider that we have a population of approximately 80 000, it’s five percent. So, an appeal is a reason why people are not paying”. In the questionnaire, the participant confirms that this is a finding by writing “People are appealing the fine”.

Subsequently, we grouped similar findings together, considering the expected answers we had identified (see 3.1). For example, we grouped the findings “So it is not paid because we have Insert Date Appeal to Prefecture and there is a dismissal so they don’t have to pay” (p1) and “... Insert Date Appeal to Prefecture. Yeah, this is an obvious reason, it’s not paid if the offender appeals” (p10) together, as they both identify appeals as the circumstance for not paying. Then, we further grouped answers related to appeals to the prefecture or the judge into “business rules” as appeals relate to one of business rules in the process. In this way, we identified groups of similar findings for our analysis.

Step 4: Assessing finding quality. Finally, we evaluated each finding along several characteristics that can be correlated with their quality, most of which we adapted from the characteristics of visualization insights proposed in [25]. These quantifiable characteristics are presented in the top part of Table 2. In line with existing literature, we rated the depth of the finding (broad or deep) and the relationship of the finding with the participant’s intent for their analysis (directed and unexpected), which we coded in the think-aloud. We also describe whether a finding is quantified by the number of cases or events in the log it refers to (quantification). Quantification is an aspect we added, as it often helps analysts assess the relevance of a finding based on how many cases it affects relative to the log size [50]. It also helps to assess the accuracy of the analysis and identify guessing, as analysts must interact with the data through filters to quantify their results. To assess the quality of the findings qualitatively, we designed the quality scale presented in the lower part of Table 2 together with examples drawn from our data. We created this scale to assess how well a finding addresses the AQ and to go beyond a binary definition of correctness [25], which we found impractical for exploratory questions. Our aim was to have a nuanced description of quality that considers the connection of a finding to the problem domain, its basis in the data, and whether the finding goes beyond simple observations and provides explanations and insights into possible reasons or root causes. Our quality scale is independent of whether an answer directly matches the

ID	Code (Multigranular)	Start time	End time	Type	FP	
1	HOW\PDF Reader\Process and Task Documentation\Process Documentation	0	5.5	HOW		
1	HOW\Variant Inspection\Variant Explorer	5.7	7.37	HOW		
1	Inference\Hypotheses Making and Testing\Hypothesis Making	6.28	6.47	HMAKE		
1	Inference\Hypotheses Making and Testing\Hypothesis Making	6.58	6.8	HMAKE		
1	HOW\Process Explorer\Process Model\DFG\Interactive DFG\Process Flow	7.78	8.38	HOW		
1	HOW\PDF Reader\Process and Task Documentation\Process Documentation	8.38	8.8	HOW		
1	HOW\PDF Reader\Analysis Question	8.9	10.3	HOW		
1	HOW\PDF Reader\Process and Task Documentation\Process Documentation	10.3	10.65	HOW		
1	HOW\PDF Reader\Analysis Question	10.65	10.87	HOW		
1	HOW\Variant Inspection\Variant Explorer	11.08	11.4	HOW		
1	HOW\Filter\Applied Filter\Activity Filter	11.4	11.7	HOW		
1	HOW\Variant Inspection\Variant Explorer	11.7	12	HOW		
1	HOW\Variant Inspection\Variant Explorer	12.2	12.73	HOW		
1	HOW\Variant Inspection\Variant Explorer	12.73	12.98	HOW		
1	HOW\Filter\Applied Filter\Activity Filter	12.98	13.13	HOW		
...	...	...	...	...		
1	Finding	23.12	23.15	FP	1	
...	...	...	...	...		

ID	FP	Finding (Follow-Up Questionnaire)	Reason (Follow-Up Questionnaire)	Finding (Think-aloud)	Broad or Deep	Related Intents	Directed or Unexpected	Quality Scale	Quantified	Finding Category	Unique
1	1	In the cases that go from Create Fine to Send Fine after 90 days passed from the notification, offender is not obliged to pay the fine. This is the Italian law and the event log includes cases with this circumstance.	90 days are passed and there no need for payment	All right, so this is actually one of the first findings. I mean, it's not paid in 52.7K cases because the time is basically more than 90 days between these two activities. So, the offender did not receive the notification and therefore there was no obligation to pay.	B	1) So, what I do now, I go and filter on cases that have this flow, Create Fine and Send Fine.	D	2	Yes	TC	No

Fig. 2. Extract from our multi-perspective dataset including think-aloud transcripts (highlighted in blue), video recordings (in purple), and answers from follow-up questionnaires (in green). The *Type* attribute allows us to focus on events related to analytical steps (HOW), inferential steps (HMAKE and HTEST), or findings (FP). For readability, we show start and end times in minutes. The lower part of the picture shows all the data collected for one finding.

expected scenarios for two reasons. First, we designed it to be reusable in different contexts, e.g., with other event logs. Second, we observed that participants could (i) arrive at some of the correct answers by simply guessing (which is considered *wrong* according to our scale) or (ii) report valid insights that were not in our predefined list.

Two authors independently assessed the findings by classifying them using the presented characteristics and rating them according to the described quality scale (cf. Table 2). The agreement between the two authors after the first assessment round was substantial (the weighted Cohen's Kappa using equal weights was 0.69) [51]. Disagreements were subsequently resolved through discussions among the authors and a re-examination of the findings in the video recordings.

These four preparatory steps resulted in a multi-perspective dataset that served as input for our analysis. An exemplary extract of this dataset is shown in Fig. 2. For each participant, we obtain a log in which events have a start and end timestamp and can denote user interactions with the PM tool (i.e., the tool function used for the analysis; event Type = HOW), the inference steps (i.e., hypothesis making and testing; Type = HMAKE or HTEST), and findings (Type = FP). The start and end timestamps allow us to retrieve events that overlap in time and analyze the interactions derived from screen recordings together with verbal utterances, as exemplified for the highlighted finding in Fig. 2.

The multi-granular event names (Code in Fig. 2) instead allow us to adjust the granularity level of the events during the analysis. In the lower part of the figure, we show all the attributes associated with findings, including the answer provided in the follow-up questionnaires and the characteristics and quality scale described in Table 2.

To address RQ1–RQ3, we used different parts of our dataset. For RQ1, we focus on findings, i.e., the events of type FP and their attributes. For RQ2, we combine different perspectives of our data. We first analyze events of type HOW and their relations with events of type FP for RQ2.1 and RQ2.2 and then relate these events with inference events (types HMAKE and HTEST) for RQ2.3. To address RQ3, we combine all the data shown in Fig. 2 with the demographics reported in Table 1.

4. An example analysis process

To illustrate how the analysis processes of our participants unfolded over time and led to findings, we describe the analysis of participant p10 in detail. Our goal is to offer a concrete example from our study, show concrete instances of the data we collected, and provide the necessary context for the reader to follow the results presented in the following sections.

Fig. 3 provides a visual representation of the analysis process followed by p10. Participant p10 is a senior consultant with good PM expertise, who chose Celonis for the task and engaged in the session for about 37 min. To identify reasons for unpaid fines in the RTFM log (see our AQ, in Section 3.1), p10 begins by (1) understanding the Task and Process Documentation for about 12 min. During this time, the participant alternates reading the documentation and using the Process Explorer to visualize the DFG of the process. Early in their exploration, p10 hypothesizes that the adding of the penalty is related to not paying (cf. H1 in Fig. 3). This hypothesis is closely related to (2) the AQ, which p10 reads again before starting with their analysis. First, p10 decides to (3) filter for unpaid cases as they are “interested in the ones that are not paid at all. So, I just click on Payment, and I look at everything without this activity”. In exploring the DFG of the filtered log, p10 notices that over half of the cases in the entire log do not contain a Payment activity. Their hypothesis (H2) proposes that these cases may be incomplete. To test H2, p10 interacts with the DFG and validates this hypothesis for cases that end with Send Fine, subsequently filtering them out. This leads to the first finding (F1 in Fig. 3), which the participant reports as “It looks like the offender is not doing anything. He’s basically not paying the fine at all and also not appealing to the judge or to the prefecture” and quantifies as “Almost 55 000 cases are affected by this”. The participant then annotates the finding in the Notepad. Then, p10 focuses on (4) the temporal constraints (TCs) listed in the Task and Process Documentation, and hypothesizes that violating this constraint may be a reason for not paying (H3). H3 is quickly tested by applying two filters that lead to F2: “11 000 cases. I mean, it’s one tenth of it...”

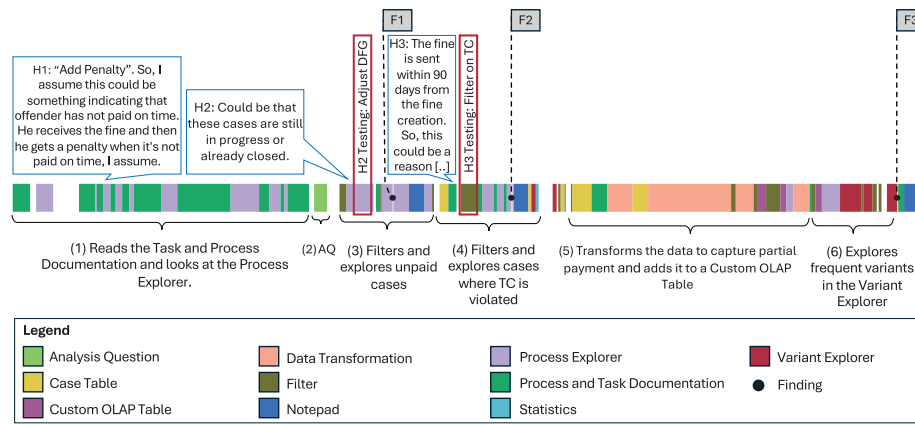


Fig. 3. Overview of p10's analysis process. The analytical steps are presented as the sequence of tool functions used over time, depicted as a scarf plot. A detailed description of the tool functions is available in Table 4. We also report the moments in which hypotheses are made and tested (H) as frames, and the findings (F), depicted as black dots on the scarf plot.

The offender is not paying because [the fine] *“hasn’t been sent within 90 days, so the date has expired”*. Despite being quick, their testing approach reflects a good understanding of the data. Afterward, p10 invests in exploring payment amount attributes to find cases in which the fine is partially paid. The participant creates a Custom OLAP Table in a new dashboard and writes a query to capture partial payments computing the difference between the *amount* to be paid and the latest *totalPaymentAmount* paid by the offender. After exploring the results in the Process Explorer, p10 switches to the Variant Explorer and finds *“...an obvious reason. It’s not paid if the offender appeals”* (cf. F3 in Fig. 3). The session ends with an explanation of how p10 would approach future analyses, which includes searching for correlations between the three findings and specific data attributes in the log.

For their analysis, p10 follows a structured approach, where findings emerge from the systematic combination of analytical steps to explore the data and inference steps to reason about it. Although our participants demonstrate a wide range of behaviors that reflect both how they work with PM tools and how they reason over the data, this example illustrates how the combination of interaction data and think-aloud is essential to interpret the user interactions and understand their behavior.

5. How do findings manifest in process mining analyses?

RQ1 aims to characterize findings of PM analysis focusing on their type, occurrence over time, and quality. Our participants reported a total of 59 findings, averaging about 2.7 findings per participant. This aligns with our analysis question, which asked for three circumstances related to non-payments.

5.1. Types of findings

We identified four types of findings, presented in Table 3. As we anticipated due to the focus of PM techniques and use cases on control flow aspects, 22 of the 59 findings are **control flow (CF)** patterns, including activities and their flow sequence. Participants highlighted frequent variants meeting specific criteria, such as those with a missing “Payment” activity or cases ending early. For example, p2 stated, *“I see that most cases that go through an Add Penalty activity then go to Send for Credit Collection. And then it ends without paying. Maybe, this is one reason why people do not pay”*. Other findings involve business rules and temporal properties or constraints, both included in the study materials. In the RTFM log, **business rules (BR)** concern the dismissal of the fine following an appeal to a prefecture or a judge. The 17 findings in this category were primarily the result of evaluating whether the known business rules justified missing payments. An example is *“For*

88 cases there is the judge’s dismissal, so the offender did not have to pay for it” (p1). **Temporal constraints (TC)** in PM often refer to rules governing expected time intervals between activities. Most findings in this category (7/12) focused on the 90-day constraint between Create Fine and Send Fine. Participants identified the violation of the constraint but, in most cases, they assumed that offenders did not pay without checking in depth. They also reported on long waiting times as something to improve and as a possible cause of missed payments. The remaining findings in this category included cases that are still in progress, e.g., *“you have received a fine and you have not paid it yet, but you are not over the due date”* (p5). The final category covers eight findings and includes patterns and correlations between **data attributes (DA)** alone or data attributes and the control flow. Within this category, findings include the identification of specific attribute values (5/8) that explain non-payment (e.g., the dismissal of the fine) or correlations between groups of cases with similar data attribute values (3/8). The participants compared the values of different *amount* attributes, discovering partial payments, e.g., *“if the last payment is less than the total payment amount, then it’s not paid in full”* (p4). They also reported unique correlations, such as *“there is a dependency between the amount and the payment rate, as the smaller amounts seem to be paid more often than the higher ones”* (p19) or *“some people pay and they’re still sent for credit collection [...] This happens for specific notification types”* (p6).

Our analysis shows that most participants identified findings of different types, suggesting that our AQ encouraged them to explore different process perspectives. The relatively low number of DA findings can be attributed to the emphasis that PM tools place on control flow analysis. Many tools provide algorithms and visualizations for control flow analysis but offer less direct support for working with event-level attributes. Still, correlations in data attributes can reveal unique patterns or unknown data quality issues that analysts find interesting. This is in contrast with normative business rules and temporal constraints, whose violations are desirable to check but often known by stakeholders.

5.2. Distribution of findings over time

Next, we analyzed how findings manifest throughout the analysis. Fig. 4 illustrates the distribution of findings over time, their type, and their quality on a scale from wrong to excellent (cf. Step 4 in Section 3.2).

On average, participants spent 36.5 min in their analysis, with a median of 36.2 min and a standard deviation of 4.4 min. However, although their sessions were comparable in length, the **occurrence of findings over time** shows significant variability. Some findings are discovered early in the analysis. For example, participants p2,

Table 3

Four main types of findings. The **Example** column reports findings taken from the think-aloud of our participants; the **#** column reports the number of findings in each category.

Finding type	Description	Example	#
Control flow (CF)	Refers to specific control flow patterns, including cases where a particular activity is (not) present or repeated, or relationships between multiple activities.	“One thing is this process flow where we sent a fine. It goes straight to process end without any Payment, without any Add Penalty. This is something I would definitely mention to the police department...” (p20)	22
Business rules (BR)	Refers to cases where the process meets or violates a business rule known to the analyst. Business rules can involve both the control flow and data attribute values.	“It is not paid because we have this period from Send Fine to Insert Date Appeal to Prefecture and there is a dismissal, so they don’t have to pay. So for 1.97K cases we know the payment is not done because of this dismissal.” (p1)	17
Temporal constraints (TC)	Refers to temporal properties of the process, including cases still running, or to violations of temporal constraints, such as deviations from the expected timing of the process or waiting times.	“It could be that the offender is aware of this constraint so he or she is not obliged to pay. So let me see here... Ok, 11000 cases. I mean, it’s one-tenth of it... The offender is not paying because it hasn’t been sent within 90 days, so the date has expired.” (p8)	12
Data attributes (DA)	Reports trends or patterns in the data attributes or relationship among data attribute values and the control flow.	“For some articles, the payment rate is lower, so that might be something like a driver. So I could not apply that kind of logic to other attributes. Yeah, so I found the article.” (p19)	8

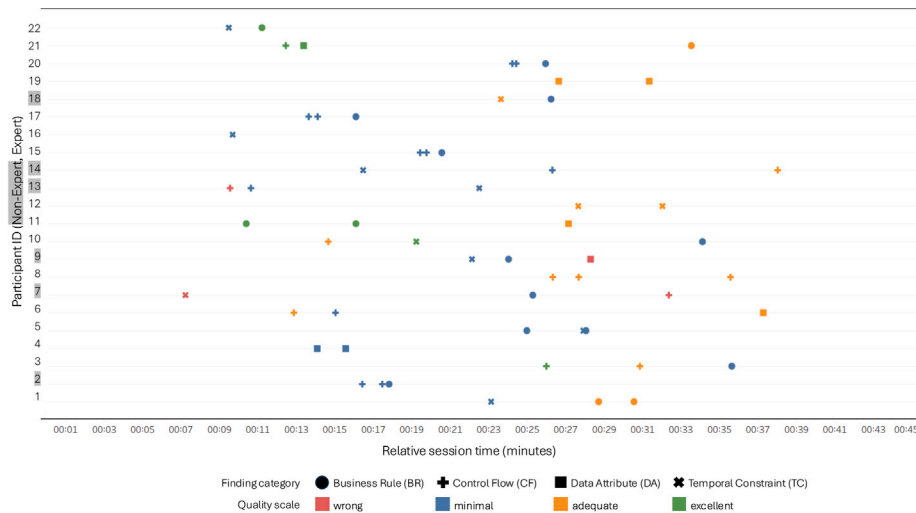


Fig. 4. Distribution of findings over the time of the session (in minutes) per participant. Findings are represented based on their Finding Category and Quality Scale.

p17, and p22 reported all their results within the first half of their analysis, specifically in less than 18 min of exploration. In contrast, other participants, such as p1, p8, p9, p12, p18, p19, and p20, reported their findings well after the midpoint of the analysis session. Overall, 22/59 findings emerged in the first half of the analysis session.

The **time to the first finding** also varies significantly across participants, with a mean of 17.77 min and a standard deviation of 6.75 min. Participants p7, p22, p13, and p16 all took less than 10 min to report their first finding, while p3, p5, p8, p19, p20 and p12 reported their first finding after 24 min of analysis. Participants p5 and p20 stand out in this group, as their first finding occurs late in the session, specifically in the last quartile of their total analysis time (76% and 80% of the duration of the session, respectively). Shorter times can suggest that analysts are able to familiarize themselves quickly with the data, but could also indicate speculative or erroneous conclusions. The four findings reported within 10 min of analysis, are all of low quality (wrong or minimal). However, the correlation between the time to the first finding and the quality is weak (Spearman correlation, $\rho = 0.37$, $p = 0.08$) and the quality of findings does not seem to improve during the analysis for these participants (p7, p22, p13, and p16). Their think-aloud further highlights this trend: p7 expressed uncertainty about their analysis, e.g., “*I don’t know if I’m doing this right*”, while p16 and p22 quickly concluded that the process ends with Send Fine due to incomplete data recordings: “*they are incomplete cases. So, they are not*

completed in the recordings” (p22). This is something that they did not check in the data and does not hold true for most of these cases.

When looking at the type of the first findings in Fig. 4, we see 10 findings of type control flow (CF), 8 of type temporal constraint (TC), 2 of type business rule (BR), and 2 of type data attribute (DA). The relatively high number of CF findings among the first ones being discovered can again be attributed to the way PM tools prioritize control flow analysis over, for example, data attributes. Instead, the high proportion of TC findings might be specific to the process we chose, as the temporal constraint was reported in the documentation and involves the first activity of the process. Still, in their think-aloud, participants reported focusing on temporal constraints to “*come up with a normative model that describes the process*” (p1) or because “*time is usually something on which I try to focus since we have constraints and that is something that customers also wants to see*” (p10).

When comparing the discovery of different findings by the same participant, we noticed patterns in their **temporal proximity**. Some participants (e.g., p2, p5, p15, p17, p20) report findings in close succession, sometimes only 40 s apart. Interestingly, p2, p15, p17, and p20 reported the same insights. Their CF findings were based on (i) the expected scenario in which the offender never pays, and (ii) the process that ends with activity *Send Fine*. However, all participants overlooked the temporal constraint that explains this endpoint. While they all identified the dismissal of fines by the prefecture as their third finding (type

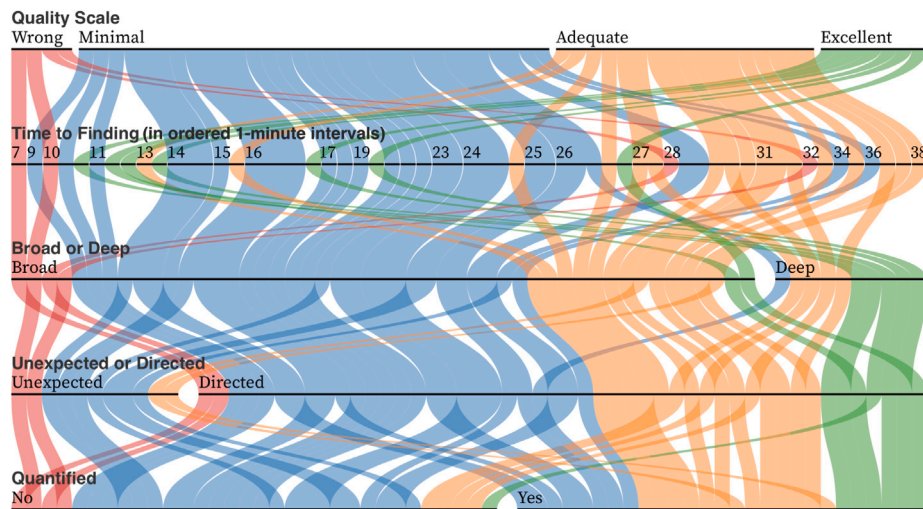


Fig. 5. Parallel sets representation [52] showing relations between the Quality Scale, the relative time of the finding quantized to 1 min intervals (Time to Finding) and measurable characteristics of a finding (Broad or Deep, Directed or Unexpected, Quantified).

BR), they failed to check the data attributes to confirm this, leading to findings of minimal quality. Other participants, like p3, p10, p11, and p14, reported findings regularly throughout the session, suggesting a systematic analysis of scenarios that manifest over different process perspectives. In the case of p10 (see Section 4), findings emerged as the result of the formulation and testing of hypotheses on different aspects of the log, ranging from the control flow to temporal constraints and business rules.

In general, while the time of a finding does not appear to correlate with its type or quality, we observed a tendency of findings reported within 10 min of analysis to be of low quality. The emergence of findings over time appears to be driven by the strategy adopted by individual participants: some perform superficial analyses, reporting similar findings quickly, while others seem to follow a structured approach. Participants often begin by exploring the control flow or testing temporal constraints. This reflects a strong focus on control flow analysis and an interest in assessing (time) performance and identifying violations of normative constraints, an approach that recalls compliance analysis.

5.3. Quality of findings

Fig. 4 also shows the quality of 59 findings identified by our participants, classified by their quality (cf. Step 4 in Section 3.2).

Overall, our participants discovered four wrong, 31 minimal, 17 adequate, and seven excellent findings. Of the 59 findings, 47 either fully or partially matched one of the expected answers. The remaining findings addressed either new aspects in the data that we had not considered or incorrect answers that cannot be confirmed based on the event log. Wrong findings included guesses or misinterpretations of the data, e.g., “We can also see this as a circumstance. For many cases we simply don’t have the data to know which fines were paid” (p9).

For all the findings, we also assessed their measurable characteristics, as shown in Fig. 5. These include whether the finding was broad or deep, directed or unexpected, and quantified. In general, our participants discovered mainly broad (48/59) and directed (48/59) findings, possibly due to the setting of our study, where the goal, although broad, was provided by the AQ. Directed findings were all preceded by clearly stated intents, such as “I would start by analyzing the most common trace. I want to focus on the main behavior as this is typically what interests the process owner. And, in the main behavior, I want to detect problematic cases. So, I will first check—‘is there a payment?’” (p21).

Half of the discovered findings were quantified (27/59), indicating that participants focused on the case frequency of their findings. Some

participants used the number of cases in their answer as a possible indication of its relevancy, e.g., “I’ll start by following the Pareto principle, where 20% of reasons likely account for 80% of cases. I’ll first quantify the most common reason and, if time allows, explore the specific cases” (p3), or to choose what to report “Ok, so I am going to discard this one because it’s only happening 3% of the cases” (p13).

Fig. 5 shows the relations between the quality scale, the relative time of the finding (quantized to 1 min intervals), and the measurable quality characteristics. Interestingly, the quality of the findings is not related to the timing of their discovery (Spearman correlation, $\rho = 0.08$, $p = 0.508$), although we found a weak correlation between the time to get to the first finding and its quality (see Section 5.2). Still, Fig. 5 reveals some trends. Excellent findings were reported early in the analysis, while adequate findings emerged later. Minimal findings appeared mainly in the early or middle phases of the analysis, whereas the four wrong findings were scattered either very early or late. We can also see that wrong findings are all broad and not quantified. The fact that excellent findings tend to occur early in the analysis is surprising, especially considering that they require examining the data in depth. This could indicate that analysts who arrive at these findings have a more focused approach (all of them are “directed”) and can quickly link the domain problem to the data. Many of these findings are also quantified, indicating that understanding their scale, such as the number of affected cases, helps analysts to connect them to the data. This approach also enables analysts to link and compare findings with each other (see the excellent finding of p11 in Table 2). Adequate findings are most often identified in the second part of the analysis, indicating that it generally takes time to find evidence in the data that can support a finding. These findings emerge primarily from a directed approach, but there is no clear correlation with depth or quantification. Minimal findings occur both early and later in the process and are usually broad and less connected with evidence in the data.

From this analysis we can conclude that *excellent* findings are often deep, well-quantified, and derived from a directed approach. Interestingly, they tend to emerge early in the analysis, but seem to result from structured thinking and clear intent. In contrast, *wrong* and *minimal* findings are often broad, lack quantification, and stem from a less focused approach.

6. What are the analytical and inference processes that lead to findings?

With RQ2, we investigate the processes that lead to findings and uncover patterns in the analytical and inference steps made by our

Table 4

Tool Functions used by our participants at the highest level of abstraction. Frequency (**Freq.**) denotes the number of times each function was used, **Duration** indicates the total time in minutes spent using each function by all participants, and **#P** is the count of participants (out of 22) who used the tool function at least once.

(Tool) Function	Description	Freq.	Duration	#P
 Filter	Definition of inclusion and exclusion criteria to select process instances (cases) or events (activities) and their application to the event log.	192	43.73	21
 Variant Explorer	Exploration of (individual) process variants, often annotated with frequency information.	147	123.5	16
 Process Explorer	(Interactive) DFG or process model (e.g., Petri net) resulting from process discovery. It displays the control flow possibly enhanced with performance or frequency information.	138	131.15	20
 Process and Task Documentation	Document containing information about the process and its domain (cf. Section 3.1).	135	118.08	20
 Data Transformation	Application of query or scripting languages (e.g., SQL, PQL) to manipulate or enrich the event log.	91	69.63	11
 Statistics	Statistics about the process, including statistics about activities, cases, and data attributes and how they are distributed over time.	80	40.2	16
 Analysis Question	Document describing the analysis question (AQ) (cf. Section 3.1).	58	25.62	20
 Custom OLAP Table	Table that analysts can tailor to their needs by selecting and transforming event data (e.g., to inspect a performance indicator over case attributes).	58	27.17	8
 Notepad	Text editor for note taking.	45	29.52	10
 Case Table/Explorer	List of cases that allows inspecting individual cases and case and event attributes in detail.	38	13.73	12
 Import/Export	Import/export of the event log or artifacts (e.g., process models) from/to the PM tool.	28	16.83	11
 Process Discovery Technique	Process discovery algorithm that can be chosen and configured by the analyst. This tool function is not available in Celonis or Disco at the time of our study.	23	20.37	6
 CSV Reader	Tool to inspect the raw event log (e.g., Excel).	17	21.3	7
 Plot or Chart	Visualization of event data (i.e., scatter plot, bar chart, histogram, dotted chart) that is not a process model.	14	13.62	4
 Conformance Overview	Built-in conformance analysis providing details of deviations from the normative process model.	11	4.07	5
		1075	698.5	

participants. To this end, in Section 6.1 we first analyze the general use of tool functions (RQ2.1). Next, in Sections 6.2 and 6.3, where we explore the *sequences* of analytical (RQ2.2) and inference steps (RQ2.3) followed to arrive at their findings.

6.1. Use of tool functions during the analysis task

To explore the use of tool functions, we analyzed the interaction logs derived from the video recordings (cf. Step 1 in Section 3.2). For this analysis, we chose the highest level of granularity and removed moments when participants switched screens or spoke without acting. Table 4 shows an overview of the 15 tool functions we obtained at this granularity level, their description, and use by our participants. The majority of the analysis was spent on (interactive) process models in the Process Explorer or the Variant Explorer, and on the Task and Process Documentation. Collectively, these three tool functions account for 53% of the total analysis time, which amounts to 372 min on a total of 698.5. In contrast, the Conformance Overview accounts for 0.6% of the total time (4 min), while Plots and Charts, including scatter plots or dotted charts take up 1.9% (13.62 min), similar to the Case Table/Explorer, which is also used for 1.9% of the total time (13.73 min).

After gaining a general overview of the tool functions used, we investigated their correlation with findings. The bar chart in Fig. 6 shows the number of findings discovered while using a specific tool function. Findings were discovered while using eight of the 15 tool functions. The majority was discovered using the Variant Explorer (25 findings) or the Process Explorer (18 findings). Four findings were derived using the Case Table/Explorer or the Process and Task Documentation, three

using the Notepad, two using the Statistics or Plots or Charts and one using the Custom OLAP Table.

In Fig. 6 we also show whether specific tool functions facilitate the discovery of findings compared to their frequency of use. In the figure, this relationship is shown by the line graph, where the number of findings discovered using each tool function is compared to the overall frequency of its use. The Variant Explorer has the highest ratio of findings over frequency of use, with findings being derived in 16% of its use (25 findings were reported over the 157 times the Variant Explorer was used). Interestingly, Plots and Charts also led to findings in 14% of their use, but their overall use in our study was low, as seen in Table 4.

Fig. 7 relates tool functions to finding types (Fig. 7(a)) and finding quality (Fig. 7(b)). From Fig. 7(a) we notice that almost all the findings of type BR were discovered while using Plots and Charts, as participants could quickly identify the presence or distribution of specific activities related to these rules in the log. For example, p21 used the dotted chart in ProM to identify fines dismissed by the prefecture (see their adequate finding in Table 2). They discovered from the visual pattern that the outcome of the appeal is not always recorded, which is indeed correct if the appeal is successful. The Case Table and the Custom OLAP Table mainly supported participants in revealing DA findings, as they allow to access data attributes and see them in detail (also in a non-aggregated form).

From Fig. 7(b), we see that wrong findings are generated when using the Variant or Process Explorer, or the Notepad. In the case of the Variant and Process Explorer, these errors may arise from potential misinterpretations of DFGs, while in the case of the Notepad, they likely result from speculative reasoning. Minimal findings are also associated with the Variant and Process Explorer or the Case Table,

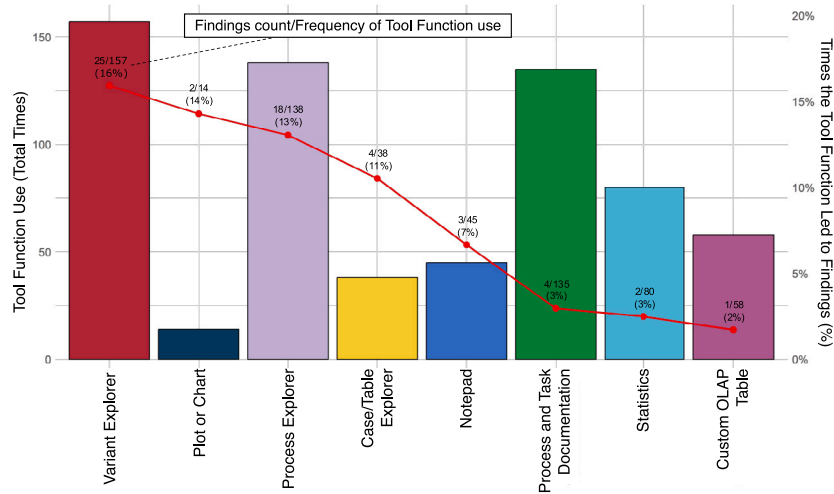


Fig. 6. Total times a tool function is used vs. percentage of times it led to a finding: The bar chart (primary y-axis) shows the absolute frequency of use for each tool function, labeled on the x-axis. The red line plot (secondary y-axis) represents the number of findings discovered while using a specific tool function over its frequency of use.

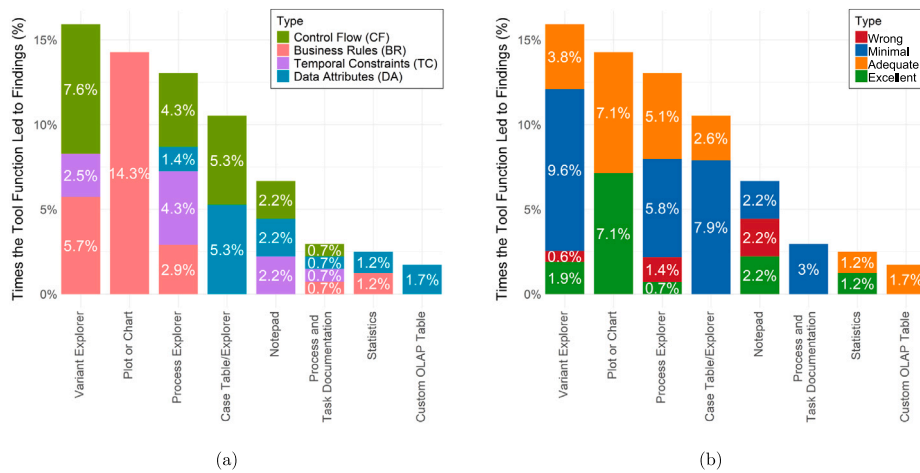


Fig. 7. Findings discovered using specific tool functions. (a) Ratio of findings by finding type; (b) Ratio of findings by quality scale.

indicating that these functions are useful for gaining initial insights or identifying example cases. Interestingly, the ratio of adequate or excellent findings is comparatively lower for both the Variant Explorer and the Case Table/Explorer, indicating that both tool functions might rather support initial hypothesis building than testing or verification. Instead, tool functions that require some customization, such as Plots or Charts and the Custom OLAP Table, seem to lead to good quality findings (adequate and excellent), likely because participants are more in control of the data they are visualizing. We observed such behavior in six participants, including p10 (cf. Section 4), who used the OLAP Table and transformed the data to discover their Finding F3. Participant p8 reflects on this approach and explains why using these specific tool functions was beneficial for them: “I used the Variant Explorer to see what the most common scenarios are. After finding those variants or scenarios, I then tried to answer the second part of the question. Why is the scenario like that? And I went into the raw data. For me, I really want to understand the data structure first, and once I’m happy with it in a table, that’s when I build components that allow me to filter and look into things more easily”.

Most findings are discovered while using interactive tool functions, like the Process and Variant Explorer. Although these functions are helpful for obtaining initial insights, they can also lead to misinterpretations. In contrast, customizable functions like Plots, Charts, and Custom OLAP Tables tend to yield higher-quality findings, as they offer greater control over the data. However, to use them effectively, analysts

need a good understanding of the data, which may also explain why higher-quality findings often result from their use.

6.2. Mining sequences of analytical steps

From the example analysis of p10 in Section 4, we see how findings emerge as a result of many, connected analytical steps. Thus, in this section, we explore sequences of analysis steps that lead to findings (cf. RQ2.2). To this end, we used the sequential pattern mining approach described in [53], which allows exploring sequential patterns in an interactive manner. Sequential patterns are defined as sub-sequences of events that meet a configurable support threshold and constraints. This VA method allows one to dynamically expand patterns using a tree-like visual interface and apply local constraints at each stage to direct the mining toward patterns of interest. For our analysis, we set the following constraints [54]: (i) a minimum support value (number of sequences, i.e., participants in our case) that exhibit a certain sub-sequence pattern, and (ii) a maximum gap, which constrains the maximum number of events allowed between any two events in the sub-sequence.

Fig. 8 shows the analysis processes of our 22 participants, using the 15 tool functions described in Table 4. The figure illustrates the variability in how participants interacted with the tools despite having the same task. At the beginning of the task, participants mainly used

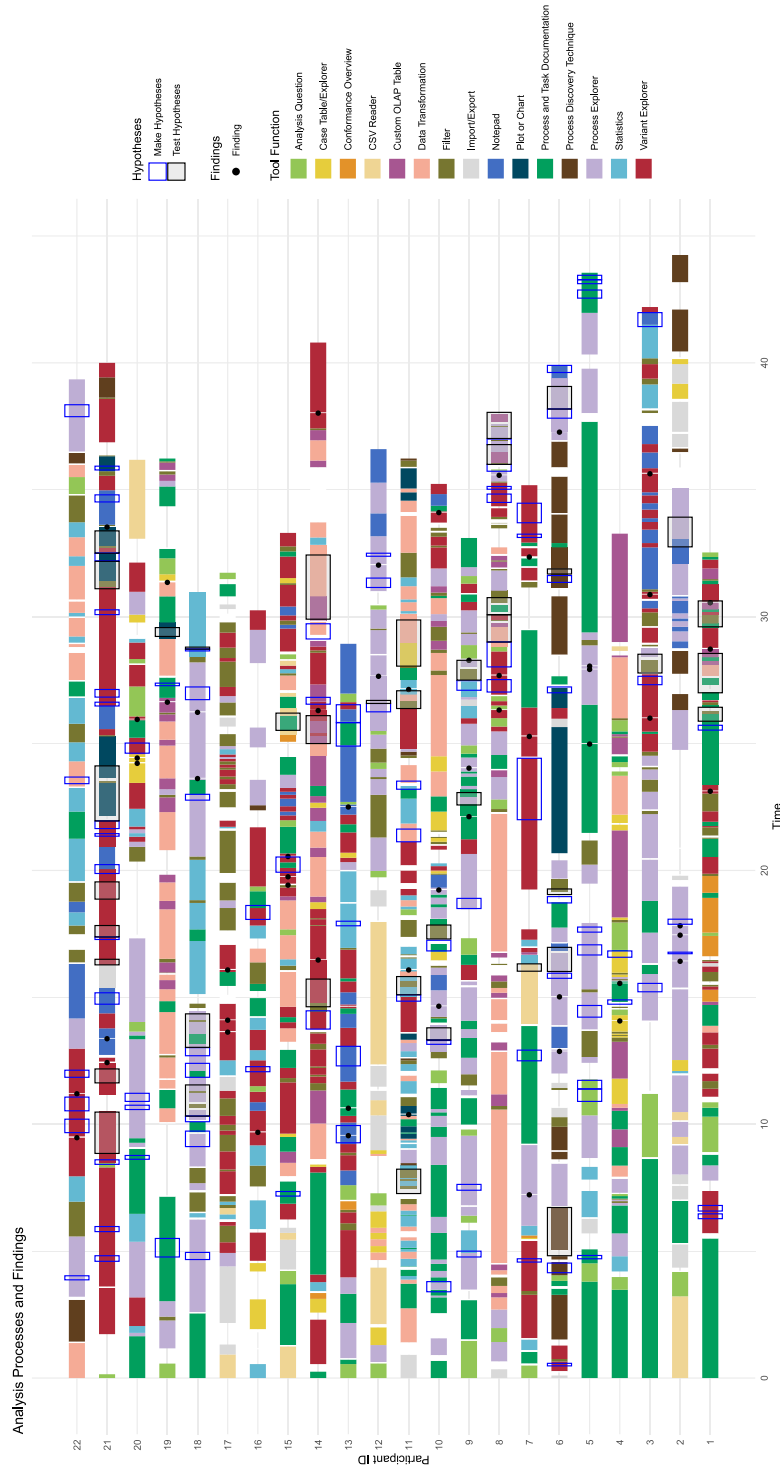
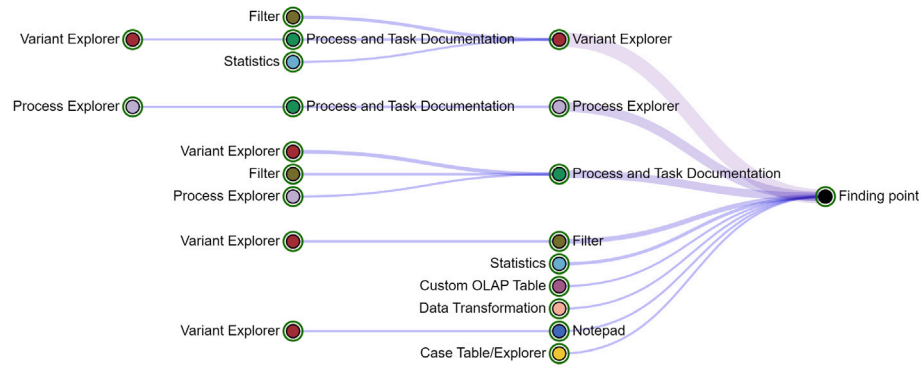


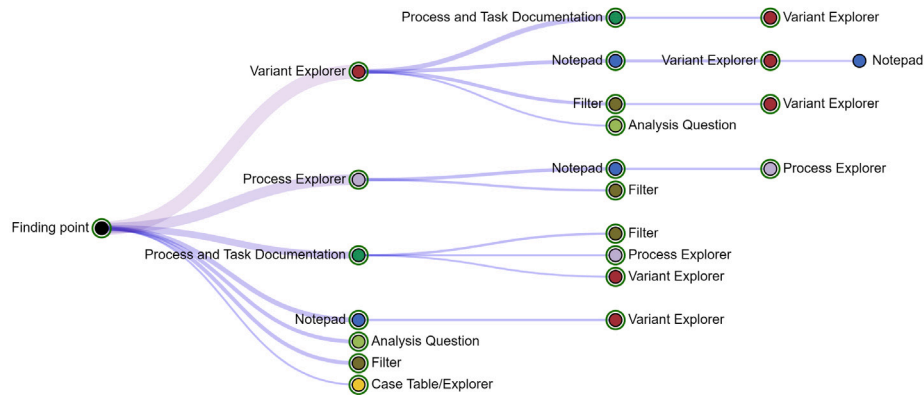
Fig. 8. Analysis processes and findings. For each participant, we show the sequence of analytical steps, represented by the **Tool Function** used, the inference steps **Make Hypotheses** and **Test Hypotheses** and the **Findings** reported during the analysis session.

the Process and Task Documentation and Analysis Question, along with the Process or Variant Explorer to understand the task and the process they were working with. This helped them to make early hypotheses, as done by p10 (cf. Section 4), p1, p9, p18, and p19 and, eventually, discover findings. As shown in Fig. 9(a), this alternation of Process and Task Documentation with the Variant or Process Explorer is a prominent pattern preceding Finding Points (FP) (Variant Explorer → FP: 27 findings, Process Explorer → FP: 18 findings, Process and Task Documentation → FP: 15 findings, Process and Task Documentation →

Variant Explorer/Process Explorer → FP: 5 findings). However, in the sequential patterns leading to findings, we can also observe that the Variant Explorer is often preceded or followed by some Filter (Filter → FP: 9 findings, Filter → Variant Explorer → FP: 6 findings Variant Explorer → Filter → FP: 5 findings), as our participants discovered findings while examining the most frequent variants in the log that they had filtered according to specific criteria (e.g., the absence of the *Payment* activity).



(a) Tool functions used before discovering a finding.



(b) Tool functions used after discovering a finding.

Fig. 9. Sequence patterns showing the analytical steps before and after a finding (Finding Point) obtained with a minimum support = 10% and maximum gap = 1.

A few participants started their analysis with the raw data (CSV Reader), and only two used it later in the analysis to add derived attributes (p12) or to discuss further analyses in a database (p20). However, it appears that the raw data did not produce findings, as confirmed by the patterns in Fig. 9(a).

Tool functions such as the Notepad and Plots and Charts were mainly used in the second part of the analysis to report and visualize results after the data had been substantially filtered and participants were confident about what to plot. The Notepad was used to report findings, as visible in the patterns shown in Fig. 9(b) (FP → Notepad: 10 findings). The figure also shows the presence of the Analysis Question among these patterns (FP → AQ: 8 findings). Participants used this as a reference to evaluate their findings and ensure that they had addressed it. Another observation is the use of the Process and Task Documentation (FP → Process and Task Documentation: 14 findings), which, in this case, helped participants validate findings.

From Fig. 8, we also see that many tool functions were used at any time during the analysis. This suggests that they can serve both exploratory and confirmatory analysis tasks, and it is up to the participant to combine them according to their goals. For example, the Process and Task Documentation serves as a source of domain knowledge, such as interpreting the process or guiding filtering at the initial stage, but also as a reference to validate findings. Similarly, the Case Table/Explorer is used to sample cases that can inspire hypotheses at the beginning, but it also supports the qualitative testing of assumptions before adding more complex filters. Also, in the figure, we see how the time spent using different tool functions varies, reflecting different analysis paces. For example, Data Transformation or Statistics are used for longer periods. This can be explained by the fact that activities that require data wrangling, (manual) querying, or computing statistics take longer compared to filtering, which is often done in quick interactions.

At last, participants showed notable differences in the number of tool functions they used and in the frequency of switching between them. Some participants, such as p10, p11, or p15 used many tool functions (11 for p10 and p11, 13 for p15) as opposed to p5 and p18 who used only 4 and 6, respectively. Participants p8, p10, and p11 are also among those who switched many times during their analysis, with p11 showing 141 switches and p8 and p10 more than 110. In contrast, p5 and p16 had less than 50 switches. Switches between different tool functions can indicate that analysts integrate different information sources or look at multiple process perspectives simultaneously. For example, p10 initially switched between Task Documentation and the Process Explorer to understand the process flow. Later, they combined Documentation with Filters and sample cases to check temporal constraints, and used the Process Explorer to visualize cases that meet their conditions. As another example, close to 13 min into their analysis, p4 explains their switch between the OLAP table to filtering the case table as follows: “I’m trying to make sense of the data and get to KPIs and add information here” [the Custom OLAP Table]. “Let’s just take one case to understand this better. So, I’m switching from like building something to identify on single cases why something is going wrong”.

The analytical steps of our participants show that findings emerge from alternating between process exploration and filtering. Tool functions like the Process or Variant Explorer are often used early to gather initial insights, while Filters are essential to refine the focus and arrive at findings. Interestingly, many tool functions support both the exploratory and confirmatory tasks, depending on the participants’ analysis styles.

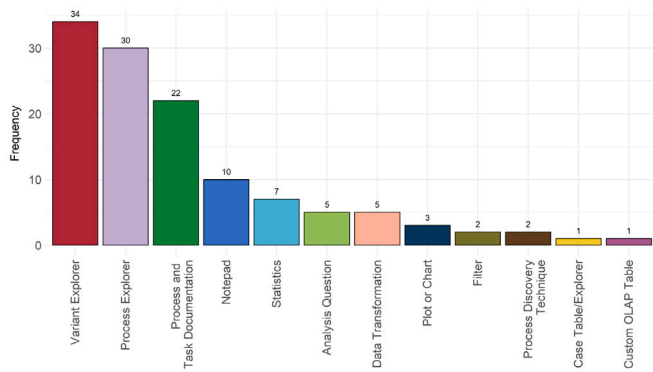


Fig. 10. Frequency of tool functions used while making hypotheses.

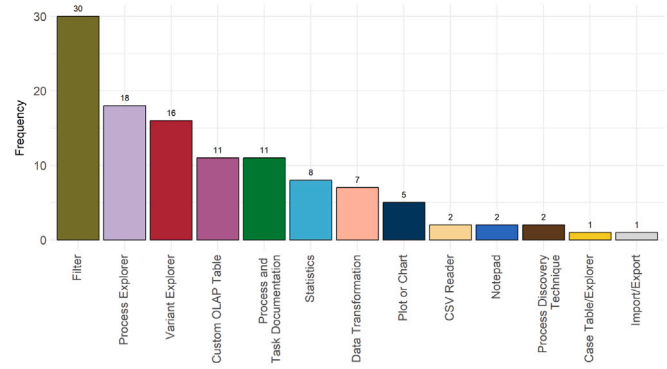


Fig. 11. Frequency of tool functions used while testing hypotheses.

6.3. Mining sequences of inference steps

In this section, we explore the participants' reasoning processes, focusing on inference steps (RQ2.3). To this end, we link inference steps with tool interactions, findings, and their quality (cf. Section 3.2).

Inference steps and tool functions. Fig. 8 displays the inference steps “Make Hypotheses” and “Test Hypotheses” made by our participants as overlays to the tool functions. All participants except p17 engaged in at least two hypothesis-making steps, while only 15 participants tested their hypotheses. Collectively, our participants performed 97 hypothesis-making steps and 40 hypothesis-testing steps. This discrepancy highlights different inference strategies employed by analysts. While some participants demonstrated a systematic approach to inference, others seemed to be less methodical in their reasoning. In particular, we observed that six participants (p4, p5, p13, p16, p20 and p22) made hypotheses but did not follow through with testing them, whereas others, such as p6, p8, and p21, engaged in thorough hypothesis-making and testing.

We observed a common trend where analysts generate more hypotheses than they test. This is because, early in the analysis, analysts make assumptions to comprehend the data (for example, H1 in Fig. 3), without necessarily testing them immediately. This often involves hypotheses that are not directly related to the AQ. Additionally, some hypotheses are rooted in the participants' experience or come from common sense, and are therefore not easily testable in the data, leading them to be accepted as true. An example of such hypothesis is “*In 526 days anything could have happened, a person could have moved*” (p7). Another example is “*I assume that they received some sort of notification, but it could be that it was online and, because the notification here [in the data] is only via post, they missed it or didn't receive it in their phones*” (p13).

In Fig. 10, we illustrate the tool functions our participants used while making hypotheses. If one hypothesis was made while using multiple tools functions, we counted it for all the tool functions used. Not surprisingly, different visualizations of the process control flow in the Variant and Process Explorer inspired most of the hypotheses, combined with the Process and Task Documentation. Some participants also made hypotheses while taking notes, which helped them organize their thoughts and observations.

When examining the tool functions used to test hypotheses, shown in the bar chart of Fig. 11, Filter stands out as the most frequently used testing approach. In the analysis, filters were crucial for checking activity presence, process endings, and temporal constraints. This indicates that effective hypothesis testing in PM often relies on a combination of filtering and visualizing the control flow of the process, e.g., through the Process or Variant Explorer. This combination of tool functions enables analysts to test and refine their hypotheses by examining specific process paths or variants. Often, the results of this filtering leads to

some finding, as previously seen in the patterns in Fig. 9(a). The Custom OLAP Table (Custom OLAP Table → FP: 4 findings), Statistics (Statistics → FP: 6 findings) and Data Transformations (Data Transformations → FP: 4 findings) are also used for testing hypotheses through custom querying and data aggregation, particularly for comparing groups of cases based on data attributes or their behavior over time.

Inference steps and findings. Of the 97 hypothesis-making steps, 52 are tied to findings, whereas all 40 hypothesis-testing steps are directly related to findings. We analyzed if the number of hypotheses made or tested *before* a finding affects its quality. The number of hypotheses made before a finding does not correlate with its quality (Spearman Correlation Coefficient $\rho = 0.18$, $p = 0.18$), but more hypothesis-testing steps are linked to higher-quality findings (Spearman Correlation Coefficient $\rho = 0.52$, $p < 0.001$). Indeed, of 33 hypotheses tested before a finding, 28 precede an adequate or excellent finding, aligning with our quality scale that expects these findings to be supported by data evidence.

From the relative occurrence of inference steps and findings, we identified two inference strategies followed by process analysts. The first strategy, which we name **systematic inference** approach, involves a step-by-step approach where analysts make hypotheses, then test them, and then derive findings based on the test results. 11/22 participants exhibit this behavior, including p6, p8, p10, p11, p14, and p21. These participants test their hypotheses within one minute of making them, using filters on the event log or sample cases to qualitatively check their hypotheses. Filters are often applied to test temporal constraints or business rules discovered in the Process and Task Documentation. A good example is p10 testing of H3 (cf. Fig. 3) with a follower and throughput time filter: “*So, I'm going to the throughput time selection and select Create Fine until Send Fine... This should be within 90 days*”. The second strategy, which we name **discovery-driven inference**, involves analysts starting with findings, often derived from data observations or previous testing, and then generating hypotheses to explain those findings. Overall, 10 participants made hypotheses immediately after reporting a finding (maximum within 30 s). These hypotheses helped contextualize findings and identify directions for further investigation. Examples include p6, who finds that “*for cases in which a value [of data attribute NotificationType] is recorded, either 'P' or 'C', the fine is not paid*”, and used a Petri net discovered through decision mining to further investigate, hypothesizing that “*the notification type might be only set for those that are actually sent for Credit Collection or for those that have a Penalty*”. Similarly, p8 reports findings after working with the Variant and Process Explorer and immediately hypothesized reasons, such as “*it could be they're just still young cases, so there still might be basically a buffer for people to pay*”.

These two strategies do not appear mutually exclusive, as some participants (p6, p8, p12, p14, and p18) applied both during their analysis. Still, they can explain why we see many hypotheses concerning findings that are made after the finding is discovered.

Our analysis reveals that a greater number of hypothesis-testing steps is associated with higher-quality findings. Additionally, we have identified two distinct strategies based how participants test hypotheses. While some participants make and test hypotheses systematically, others generate hypotheses after their findings, as a way to explain or contextualize them.

7. The role of expertise in the analysis process

In Sections 5 and 6, we investigated how findings manifest and the processes leading to these findings. In this section, we factor in the PM expertise of our participants and address RQ3.

To this end, we divide our participants into two groups based on their self-reported expertise. “Non-experts” are participants who indicated basic or average expertise, while “experts” have good and advanced expertise. In total, 16 experts and six non-experts participated in our study (see Table 1). To compare their analysis processes, we first normalized the duration and frequency of tool function use for both groups and, then, conducted a comparative analysis of the quality of the findings and the analytical and inference steps leading to them.

Interestingly, both groups found a similar number of findings. The 16 experts identified 42 findings, averaging 2.6 findings per analyst, while the six non-experts found 17 findings, averaging 2.8 per analyst. As shown in Fig. 12, all finding types (cf. Table 3) were observed in both groups. However, 7/8 findings related to data attribute patterns and correlations (DA) were discovered by experts. Generally, these findings are less apparent from the process and task documentation and are difficult to detect using only interactive tool functions. Moreover, when data attributes are recorded at the event level, it is necessary to apply aggregation and transformation operations to find these patterns on the case level, which requires some analysis expertise. Also, we see that experts tended to focus on “less obvious” findings, such as patterns related to fine amounts and partial payments. Examples are experts p5 and p19, who excluded the normative outcomes described in the Process and Task Documentation from being possible findings, explaining that “We need to exclude all allowed situations and really focus on the ones which are not paid” (p19). For p5, this resulted in a more reflective analysis characterized by fewer but long interactions with tool functions, during which they made hypotheses about the process but did not test them. Instead, p19 heavily engaged with Data Transformations and a Custom OLAP Table to derive unique correlations between unpaid cases and data attributes. Additionally, as shown in Fig. 12, experts tend to identify higher-quality findings (adequate and excellent), which is supported by the correlation between finding quality and participants’ expertise (Spearman Correlation Coefficient $\rho = 0.516$, $p < .001$).

Fig. 13 shows the differences in the sequential patterns that lead to findings for experts and non-experts. They were discovered using

the same approach as in Section 6.3 (cf. Fig. 9). Both groups primarily relied on the Variant Explorer and Process Explorer for their findings, as previously seen in Fig. 6. While experts exhibited more diverse patterns, they still derived nearly 50% of their findings (19 out of 42) from the Variant Explorer. This tool function was often immediately preceded by the Process and Task Documentation (for 4 findings: Process and Task Documentation → Variant Explorer → FP), Statistics (for 4 findings: Statistics → Variant Explorer → FP), or Filter (for 3 findings: Filter → Variant Explorer → FP), all of which allowed participants to identify and focus on variants of interest. Other tool functions leading experts to findings include the Process Explorer (10 findings), the Process and Task Documentation (10 findings), Filter (7 findings) Statistics (5 findings), Data Transformation (4 finding), Custom OLAP Table (3 findings), Case Table/Explorer (3 findings). Plots and Charts preceded a finding only on two occasions. This is reflected in the diverse analysis patterns that can be identified for experts (cf. Fig. 13(a)). In contrast, non-experts show less variation in their analysis approach as it can be seen in Fig. 13(b). They predominantly used the Process Explorer for nearly 50% of their findings (8 out of 17). This pattern repeated for multiple findings, one after the other, within the same analysis process, e.g., in the case of p2 and p18. p18 explains their approach being mostly based on the DFG: “So, I plotted the map and then I tried to see popular tasks and popular sequences or unexpected transitions”. Non-experts used the Variant Explorer for 8 findings that were preceded by a Filter operation on 3 occasions (Filter → Variant Explorer → FP). They also derived findings in conjunction with the Process and Task Documentation (5 findings). Participant p7 reported “I find that the Variant Explorer is the most accessible to novice learners” (p13). Instead, the use of filters (2 findings) and statistics (1 finding) was much less observed as directly leading to findings compared to experts.

Throughout their analysis, both groups used almost the full range of tool functions provided: experts used all the 15 tool functions listed in Table 4, while non-experts did not use Plots or Charts. For both groups, the three most used tool functions were the Variant Explorer, the Process Explorer, and the Process and Task Documentation, confirming our analysis in Section 6.1. Still, we observed some differences in the time spent on some tool functions, as well as in how often participants switched between different tool functions. A notable difference between the two groups is the use of Process Explorer, which was used more by non-experts (26% of total analysis time for non-experts compared to 16% for experts). Participants found the interactive DFG convenient to investigate control flow aspects, and described it as the “simplest way to analyze the process quickly” (p2) and “get a first impression fast” (p13). In contrast, experts engaged more with Data Transformations, using them to aggregate cases based on data attribute values 12% of the time, compared to only 4% for non-experts. They also used Filters 8% of the time, while non-experts used them just 2%. Additionally, Plots or Charts were employed exclusively by experts for about 3% of the total time. When looking at the frequency of tool functions, Filters were used twice as much by experts (10.4 times for experts vs. 4.3 for non-experts), which also explains why we see more filtering leading to findings for this group. Filtering often supports hypothesis testing (cf. Section 6.3).

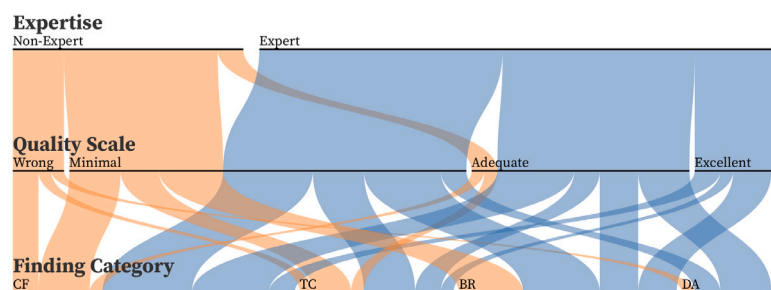


Fig. 12. Relations between expertise, finding quality (Quality Scale), and Finding Category.

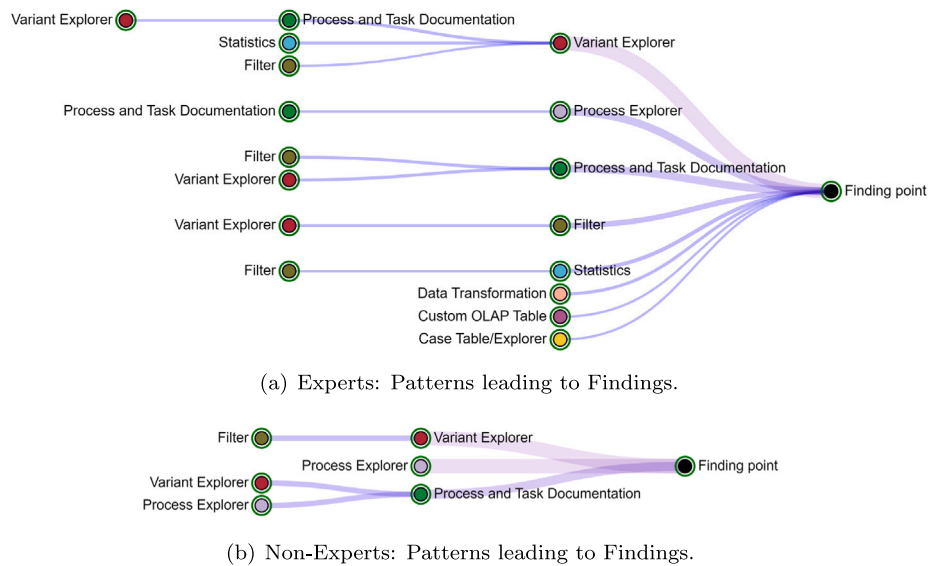


Fig. 13. Sequence patterns showing the analytical steps leading to a finding (Finding Point) for experts and non-experts. The patterns are obtained with a minimum support of two participants and maximum gap = 1.

When looking into inference steps, experts engage in testing for an average of 3.5 hypotheses compared to non-experts (who test about 1.6 hypotheses on average) with a comparable number of hypotheses made by the two groups (4.6 vs. 4.5 on average). However, there are also experts (5/16) who do not test the hypotheses they make, such as p5 and p22. As such, the correlation between the participants' expertise and the average number of hypothesis-testing steps is not significant: the Spearman Correlation Coefficient is $\rho = 0.314$, with $p = 0.167$.

Our findings highlight the important role that expertise plays in the analysis process. Experts dedicate more time to data transformation and filtering, whereas non-experts often rely on DFGs and do not always explore the data in depth. If it is true that experts seem to discover findings of higher quality, the correlation between expertise and the number of hypothesis-testing steps is weak and not statistically significant.

8. Discussion

In this paper, we have provided many rich empirical insights into how analysts work with tools and reason to arrive at findings. Our multi-perspective analysis explores the nature of findings as seen by process analysts (RQ1), the analytical and inference processes they engage in (RQ2), and how PM expertise shapes these processes (RQ3). In the remainder, we reflect on our results and methodology, highlighting how our approach can inspire the process mining (PM) and visual analytics (VA) communities to develop more effective methods for analyzing multi-perspective event logs.

Characterizing findings, analysis processes and the role of expertise. Our characterization of findings by type and quality in Section 5 shows what information process analysts consider relevant to report. Common findings are control flow patterns, violations of temporal constraints and business rules, and data attribute correlations. These finding types complement the existing body of knowledge by providing descriptions of findings from the perspective of process analysts, that are linked to information available in event logs. As such, we complement the description of PM results from case study reports that are often specific to a particular setting [27], and the framework in [6] that describes PM insights on a higher level, for example, data quality, a high amount of rework, or discovered process models. The quality characteristics and scale developed in this paper can be used to qualitatively evaluate answers, e.g., in educational settings, or to compare PM tools based

on their ability to facilitate the discovery of findings and the level of detail they offer [25]. For an exploratory analysis, broad findings might suffice, while confirmatory analyses may require tools capable of querying the data [7,9].

To arrive at findings, analysts used a variety of tool functions. We observed that some tool functions, such as the Variant and Process Explorer, help to gain initial insights, but are not sufficient to produce high-quality findings, as they may be subject to misinterpretations and do not allow analysts to test hypotheses. Still, our analysis in Section 6.1 shows that DFGs (available in the Process and Variant Explorer) are still the most commonly used tool functions by process analysts, a result that aligns with the PM literature [7,19,20,55]. Given the well-known limitations of DFGs [18] and the challenges they may pose for less experienced analysts [8], this raises the question of whether there are viable alternatives to visualize process data that are easy to understand and interact with. Also, our analysis indicates that visualization tools such as Plots or Charts are less frequently used by analysts, but when used, they often lead to findings (cf. Section 6.1, Fig. 6). This aligns with our previous work, where we found that specific visualizations, particularly sequence charts and scatter plots, are more effective than others in terms of number of observations made over time [55]. These insights underscore the need for further research to evaluate existing visualization tools for PM and to design interactive visualizations that effectively guide analysts to reach their findings. Additionally, our analysis of the analytical steps followed by process analysts showed that PM tool functions can serve different purposes over the course of an analysis. For example, the Process and Task Documentation, can inspire hypothesis-making at the beginning of the analysis (this resembles the knowledge-driven strategy in [7]) but can later be used to guide filtering and contextualize findings. In line with existing research [11], we found that hypothesis testing correlates with the quality of findings, as testing more hypotheses generally results in higher-quality results. Experts tend to produce higher-quality findings and focus more on data transformation and filtering, while non-experts often rely on DFGs. However, we found no significant correlation between hypothesis-testing steps and expertise.

An interesting insight from our analysis is that hypothesis testing is often performed through filters and analysts use external sources, like the Notepad, to organize their reasoning. While filters allow analysts to narrow down the cases, check for expected patterns, and refine their reasoning, they are not specifically designed to test hypotheses. This highlights the lack of robust support for hypothesis testing in

current PM tools and the need to design PM tools that align with the cognitive processes of analysts [14], making it easier for both experts and novices to generate and test hypotheses and externalize them as knowledge [56]. Such an improved tool design could also guide users in formulating testable, data-driven hypotheses, for example by drawing on expert strategies.

Reflections from our multi-perspective analysis. Our method of analyzing multi-perspective data offers a novel approach to understanding complex analytical workflows. By combining different sources of empirical data into a *multi-perspective dataset*, we gained insights into the work of process analysts from different perspectives, enhancing our understanding of how they interact with tools. Linking interactions with think-aloud data allowed us to reconstruct and explain the analysis processes, which would be impossible from logs of low-level interactions alone [14,34]. However, analyzing such complex and highly variable data poses some challenges, including (i) exploring multi-perspective events, (ii) managing the order of the events, (iii) detecting possibly long-term dependencies like cause–effect relationships, (iv) and adjusting the event granularity on-demand. For example, in some cases, we found that the order of the events is not significant to explain the behavior of analysts but rather causes high variability in the data. For instance, when analysts alternate between the Process and Task Documentation and the Process Explorer without making changes to the data, the order of such events is less relevant. Instead, when filters and data transformations are applied to the log, maintaining the order is crucial to understand the final result. In such settings, the ability to selectively “relax” the sequential order in specific parts of the sequence would have helped us reduce the variability in the data and go beyond strictly sequential patterns. Also, it was often hard to determine if all the steps preceding a certain event (e.g., a finding) are essential in leading to it or if events further along in the sequence also play a significant role in influencing its occurrence. A method for detecting and marking such long-term dependencies in the sequence, as well as identifying potential cause–effect relationships among events could have led to discover richer behavior patterns. At last, if we altered the granularity of events along any of the different dimensions, the dependencies among events could change completely and show different patterns of behavior [48]. Thus, to navigate complex data and enable the analysis of multi-perspective and multi-granular event data, we see clear opportunities to combine algorithmic processing with visualization. In particular, there is a need for approaches that allow analysts to explore this data using linked views that tightly incorporate the expertise of the human user both for directing the search for patterns and for meaningfully interpreting them. As resolving these challenges would benefit both PM and VA research, we call on the two communities to work together and solve some of these challenges.

Limitations. Our study has four key limitations related to both its design and our approach to data preparation and analysis. First, in terms of study design, a potential limitation is due to the selection bias of the participants. While we aimed to capture a diverse sample of analysts, we recruited more expert participants than non-experts. This imbalance could have biased the analysis of their behaviors. To mitigate this, we normalized the data by level of expertise before performing the comparative analysis. Still, the uneven distribution may affect the generalizability of our findings. The second design limitation pertains to the setup of the study itself. We aimed to create a setup that replicates a real PM analysis by using a real-life event log and an analysis task similar to BPI Challenges, and allowing participants to use the tools they were familiar with. However, despite these measures, the online environment and time constraints placed on participants pose limitations. Some participants expressed the need for more time and the opportunity for additional interactions with domain experts to confirm their findings. Although the setting may not perfectly mirror real-world projects, the recorded behaviors still provide valuable and novel insights.

Concerning data preparation and analysis, we identified two limitations. The first is related to the subjectivity involved in the coding of the screen recording and the think-aloud and the qualitative assessment of the findings. To mitigate such subjectivity, we built our coding schema based on an established task typology [47] and relied on deductive assessment with clear definitions. Also, multiple researchers were involved in the coding and assessment process, and regular meetings helped us reach a consensus, minimizing individual biases and increasing the reliability of our qualitative data analysis.

The final limitation concerns the tool-agnostic nature of our coding scheme. Since comparing tools was not the focus of our study, we designed the coding schema to be independent of specific tools. This was done by mapping the functionality across the five tools used by participants. However, it is worth noting that certain functions, such as the Custom OLAP Table (used in Celonis) and Data Transformation (available in ProM, bupaR, Celonis, and Pm4Py), were not accessible in all tools. This difference in tool affordances may have influenced the behavior of participants and could explain why some did not use those functions. However, we ensured that the groups of experts and non-experts were balanced in terms of tool use and that all participants had access to a CSV Reader, which can be used to represent the data in an OLAP table format. At last, in our study we focused on patterns of behavior that integrate multiple perspectives of the data, going beyond tool comparison.

9. Conclusion

In this paper, we have investigated how analysts arrive at findings, exploring the findings themselves and the processes leading to them. In addition, we have explored the influence of PM expertise on the analysis, discovering that experts seem to obtain findings of higher quality. Our results extend the existing body of knowledge on the work practices of process analysts, with behavioral observations. Such observations can inform the development of tools that better support analysts, such as enabling them to effectively track their hypothesis-testing process. Furthermore, strategies leading to high-quality findings could inform the creation of pattern-based guidelines to assist less experienced analysts. Finally, with our analysis, we have highlighted the potential of analyzing multi-perspective event data related to user interactions and reasoning. In the future, we plan to focus on resolving some of the ongoing challenges related to the analysis of multi-perspective data. This involves integrating algorithmic and visualization techniques to navigate these various perspectives in an interconnected way.

CRedit authorship contribution statement

Francesca Zerbato: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lisa Zimmermann:** Writing – original draft, Formal analysis, Data curation, Conceptualization. **Katerina Vrotsou:** Writing – original draft, Visualization, Methodology. **Barbara Weber:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is part of the ProMISE project, funded by the Swiss National Science Foundation, Switzerland under Grant No. : 2000 21_197032.

We would like to thank the organizers and participants of the Dagstuhl seminar 23271 “Human in the (Process) Mines” for the opportunity to collaborate and develop the ideas that inspired this work.

We also grateful to our study participants, whose time and insights made this research possible.

Data availability

Data will be made available on request.

References

- [1] M. Dumas, M. La Rosa, J. Mendling, H. Reijers, *Fundamentals of Business Process Management*, Springer, 2018, <http://dx.doi.org/10.1007/978-3-662-56509-4>.
- [2] W. Van Der Aalst, W. van der Aalst, *Data Science in Action*, Springer, 2016.
- [3] I. Ailenei, A. Rozinat, A. Eckert, W.M. van der Aalst, Definition and validation of process mining use cases, in: *Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I 9*, Springer, 2012, pp. 75–86.
- [4] S. Kandel, A. Paepcke, J.M. Hellerstein, J. Heer, Enterprise data analysis and visualization: An interview study, *IEEE Trans. Vis. Comput. Graphics* 18 (12) (2012) 2917–2926.
- [5] A. Doan, Human-in-the-loop data analysis: A personal perspective, in: *Proc. of the Workshop on Human-in-the-Loop Data Analytics, HILDA '18*, ACM, New York, NY, USA, 2018, <http://dx.doi.org/10.1145/3209900.3209913>.
- [6] V. Stein Dani, H. Leopold, J.M.E.M. van der Werf, I. Beerepoot, H.A. Reijers, From process mining insights to process improvement: All talk and no action? in: M. Sellami, M.-E. Vidal, B. van Dongen, W. Gaaloul, H. Panetto (Eds.), *Co-operative Information Systems*, Springer Nature Switzerland, Cham, 2024, pp. 275–292, http://dx.doi.org/10.1007/978-3-031-46846-9_15.
- [7] F. Zerbato, P. Soffer, B. Weber, Initial insights into exploratory process mining practices, in: *Business Process Management Forum: BPM Forum 2021, Rome, Italy, September 06–10, 2021, Proceedings 19*, Springer, 2021, pp. 145–161, http://dx.doi.org/10.1007/978-3-030-85440-9_9.
- [8] L. Zimmermann, F. Zerbato, B. Weber, What makes life for process mining analysts difficult? A reflection of challenges, *Softw. Syst. Model.* (2023) 1–29, <http://dx.doi.org/10.1007/s10270-023-01134-0>.
- [9] D. Sacha, A. Stoffel, F. Stoffel, B.C. Kwon, G. Ellis, D.A. Keim, Knowledge generation model for visual analytics, *IEEE Trans. Vis. Comput. Graphics* 20 (12) (2014) 1604–1613, <http://dx.doi.org/10.1109/TVCG.2014.2346481>.
- [10] F. Zerbato, P. Soffer, B. Weber, Process mining practices: Evidence from interviews, in: *Int. Conf. on Business Process Management, BPM*, in: LNCS, Springer, 2022, pp. 268–285, http://dx.doi.org/10.1007/978-3-031-16103-2_19.
- [11] E. Sorokina, P. Soffer, I. Hadar, U. Leron, F. Zerbato, B. Weber, PEM4PPM: A cognitive perspective on the process of process mining, in: *International Conference on Business Process Management*, Springer, 2023, pp. 465–481.
- [12] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, G. Fink, Analytic provenance: process+ interaction+ insight, in: *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 33–36.
- [13] E.T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, R. Chang, Finding waldo: Learning about users from their interactions, *IEEE Trans. Vis. Comput. Graphics* 20 (12) (2014) 1663–1672.
- [14] W. Dou, D.H. Jeong, F. Stukes, W. Ribarsky, H.R. Lipford, R. Chang, Recovering reasoning processes from user interactions, *IEEE Comput. Graph. Appl.* 29 (3) (2009) 52–61.
- [15] K.A. Ericsson, H.A. Simon, Verbal reports as data, *Psychol. Rev.* 87 (3) (1980) 215.
- [16] T. Grisold, J. Mendling, M. Otto, J. vom Brocke, Adoption, use and management of process mining in practice, *Bus. Process. Manag. J.* 27 (2) (2021) 369–387.
- [17] A. Yeshchenko, J. Mendling, A survey of approaches for event sequence analysis and visualization, *Inf. Syst.* (2023) 102283.
- [18] W.M. Van Der Aalst, A practitioner's guide to process mining: Limitations of the directly-follows graph, *Procedia Comput. Sci.* 164 (2019) 321–328.
- [19] C. Capitán-Agudo, M. Salas-Urbano, C. Cabanillas, M. Resinas, Analyzing how process mining reports answer time performance questions, in: *International Conference on Business Process Management*, Springer, 2022, pp. 234–250.
- [20] C. Klinkmüller, R. Müller, I. Weber, Mining process mining practices: an exploratory characterization of information needs in process analytics, in: *Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*, Springer, 2019, pp. 322–337.
- [21] M. van Eck, X. Lu, S. Leemans, W. van der Aalst, PM2: A process mining project methodology, in: J. Zdravkovic, M. Kirikova, P. Johannesson (Eds.), *Advanced Information Systems Engineering: CAISE 2015*, Springer, 2015, pp. 297–313, http://dx.doi.org/10.1007/978-3-319-19069-3_19.
- [22] A. Mamudu, W. Bandara, M.T. Wynn, S.J. Leemans, A process mining success factors model, in: *International Conference on Business Process Management*, Springer, 2022, pp. 143–160, http://dx.doi.org/10.1007/978-3-031-16103-2_12.
- [23] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon, Visual analytics: Definition, process, and challenges, in: *Information Visualization: Human-Centered Issues and Perspectives*, Springer, 2008, pp. 154–175.
- [24] L. Battle, A. Ottley, What do we mean when we say “insight”? A formal synthesis of existing theory, *IEEE Trans. Vis. Comput. Graphics* (2023) 1–14, <http://dx.doi.org/10.1109/TVCG.2023.3326698>.
- [25] P. Saraiya, C. North, K. Duca, An insight-based methodology for evaluating bioinformatics visualizations, *IEEE Trans. Vis. Comput. Graphics* 11 (4) (2005) 443–456.
- [26] C. North, Toward measuring visualization insight, *IEEE Comput. Graph. Appl.* 26 (3) (2006) 6–9.
- [27] J.J. Koorn, I. Beerepoot, V.S. Dani, X. Lu, I. Van de Weerd, H. Leopold, H.A. Reijers, Bringing rigor to the qualitative evaluation of process mining findings: an analysis and a proposal, in: *2021 3rd International Conference on Process Mining, ICPM, IEEE*, 2021, pp. 120–127.
- [28] S.V. Daele, G. Janssenswillen, Identifying the steps in an exploratory data analysis: A process-oriented approach, in: *International Conference on Process Mining*, Springer, 2022, pp. 526–538.
- [29] F. Du, B. Shneiderman, C. Plaisant, S. Malik, A. Perer, Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus, *IEEE Trans. Vis. Comput. Graphics* 23 (6) (2017) 1636–1649, <http://dx.doi.org/10.1109/TVCG.2016.2539960>.
- [30] H. Lam, M. Tory, T. Munzner, Bridging from goals to tasks with design study analysis reports, *IEEE Trans. Vis. Comput. Graphics* 24 (1) (2017) 435–445.
- [31] N. Martin, D.A. Fischer, G.D. Kerpedzhiev, K. Goel, S.J. Leemans, M. Röglinger, W.M. van der Aalst, M. Dumas, M. La Rosa, M.T. Wynn, Opportunities and challenges for process mining in organizations: results of a Delphi study, *Bus. Inf. Syst. Eng.* 63 (2021) 511–527.
- [32] L. Abb, J.-R. Rehse, Process-related user interaction logs: State of the art, reference model, and object-centric implementation, *Inf. Syst.* (2024) 102386.
- [33] A. Rebmann, H. van der Aa, Recognizing task-level events from user interaction data, *Inf. Syst.* 124 (2024) 102404, <http://dx.doi.org/10.1016/j.is.2024.102404>.
- [34] J.-R. Rehse, L. Abb, G. Berg, C. Bormann, T. Kampik, C. Warmuth, User behavior mining, *Bus. Inf. Syst. Eng.* (2024) 1–18.
- [35] V. Leno, A. Polyvyanyy, M. Dumas, M. La Rosa, F.M. Maggi, Robotic process mining: vision and challenges, *Bus. Inf. Syst. Eng.* 63 (2021) 301–314.
- [36] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, J. Wenskovitch, Survey on the analysis of user interactions and visualization provenance, in: *Computer Graphics Forum*, Vol. 39, Wiley Online Library, 2020, pp. 757–783.
- [37] P.H. Nguyen, R. Henkin, S. Chen, N. Andrienko, G. Andrienko, O. Thonnard, C. Turkay, VASABI: Hierarchical user profiles for interactive visual user behaviour analytics, *IEEE Trans. Vis. Comput. Graphics* 26 (1) (2020) 77–86, <http://dx.doi.org/10.1109/TVCG.2019.2934609>.
- [38] Y. Han, G.D. Abowd, J. Stasko, IntiVisor: A visual analytics system for interaction log analysis, *IEEE Trans. Vis. Comput. Graphics* (2024) 1–13, <http://dx.doi.org/10.1109/TVCG.2024.3370637>.
- [39] P. Yu, A. Nordman, M. Koc-Januchta, K. Schönborn, L. Besançon, K. Vrotsou, Revealing interaction dynamics: Multi-level visual exploration of user strategies with an interactive digital environment, *IEEE Trans. Vis. Comput. Graphics* 31 (1) (2025) 831–841, <http://dx.doi.org/10.1109/TVCG.2024.3456187>.
- [40] H. Guo, S.R. Gomez, C. Ziemkiewicz, D.H. Laidlaw, A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights, *IEEE Trans. Vis. Comput. Graphics* 22 (1) (2015) 51–60.
- [41] S. Easterbrook, J. Singer, M.-A. Storey, D. Damian, Selecting empirical methods for software engineering research, *Guid. Adv. Empir. Softw. Eng.* (2008) 285–311.
- [42] M. de Leoni, F. Mannhardt, Road traffic fine management process, *Eindh. Univ. Technol. Dataset* 284 (2015).
- [43] F. Mannhardt, Multi-perspective process mining, 2018.
- [44] G. Janssenswillen, B. Depaire, M. Swennen, M. Jans, K. Vanhoof, bupaR: Enabling reproducible business process analysis, *Knowl.-Based Syst.* 163 (2019) 927–930.
- [45] A. Berti, S. van Zelst, D. Schuster, PM4Py: A process mining library for Python, *Softw. Impacts* 17 (2023) 100556, <http://dx.doi.org/10.1016/j.simp.2023.100556>.
- [46] B.F. Van Dongen, A.K.A. de Medeiros, H. Verbeek, A. Weijters, W.M. van der Aalst, The ProM framework: A new era in process mining tool support, in: *Applications and Theory of Petri Nets 2005: 26th International Conference, ICATPN 2005, Miami, USA, June 20–25, 2005. Proceedings 26*, Springer, 2005, pp. 444–454.
- [47] M. Brehmer, T. Munzner, A multi-level typology of abstract visualization tasks, *IEEE Trans. Vis. Comput. Graphics* 19 (12) (2013) 2376–2385.
- [48] F. Zerbato, R. Seiger, G. Di Federico, A. Burattin, B. Weber, Granularity in process mining: Can we fix it? in: *International Workshop on BPM Problems to Solve before We Die, CEUR-WS*, 2021, pp. 40–44.
- [49] J.S. Yi, Y. ah Kang, J. Stasko, J.A. Jacko, Toward a deeper understanding of the role of interaction in information visualization, *IEEE Trans. Vis. Comput. Graphics* 13 (6) (2007) 1224–1231.
- [50] W.M. van der Aalst, On the Pareto principle in process mining, task mining, and robotic process automation, in: *DATA*, 2020, pp. 5–12.
- [51] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, *Biometrics* 33 (1) (1977) 159–174, URL <http://www.jstor.org/stable/2529310>.
- [52] R. Kosara, F. Bendix, H. Hauser, Parallel sets: Interactive exploration and visual analysis of categorical data, *IEEE Trans. Vis. Comput. Graphics* 12 (4) (2006) 558–568.

- [53] K. Vrotsou, A. Nordman, Exploratory visual sequence mining based on pattern-growth, *IEEE Trans. Vis. Comput. Graphics* 25 (8) (2019) 2597–2610, <http://dx.doi.org/10.1109/TVCG.2018.2848247>.
- [54] P. Fournier-Viger, J.C.-W. Lin, R.U. Kiran, Y.S. Koh, R. Thomas, A survey of sequential pattern mining, *Data Sci. Pattern Recognit.* 1 (1) (2017) 54–77.
- [55] L. Zimmermann, F. Zerbato, K. Vrotsou, B. Weber, A process-oriented approach to analyze analysts' use of visualizations: Revealing insights into the what, when, and how, in: *Computer Graphics Forum*, Vol. 44, Wiley Online Library, 2025, e70104, <http://dx.doi.org/10.1111/cgf.70104>.
- [56] P. Federico, M. Wagner, A. Rind, A. Amor-Amorós, S. Miksch, W. Aigner, The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics, in: *2017 IEEE Conference on Visual Analytics Science and Technology, VAST*, 2017, pp. 92–103, <http://dx.doi.org/10.1109/VAST.2017.8585498>.