

Initial Insights into Exploratory Process Mining Practices

Francesca Zerbato¹, Pnina Soffer², and Barbara Weber¹

¹ Institute of Computer Science, University of St. Gallen, Switzerland
{francesca.zerbato|barbara.weber}@unisg.ch

² University of Haifa, Mount Carmel, 3498838 Haifa, Israel
spnina@is.haifa.ac.il

Abstract. Process mining enables organizations to streamline and automate their business processes. The initial phases of process mining projects often include exploration activities aimed to familiarize with the data and understand the process. Despite being a crucial step of many analyses, exploration can be challenging and may demand targeted guidance and support. Still, little attention has been paid to understanding how process analysts approach this exploratory phase. With this goal in mind, in this paper, we report the results of an empirical study investigating exploration practices in process mining. Our study reveals that analysts follow different behavior patterns when exploring event logs and enact various strategies to understand the data and gain new insights. The results remark the need for a deeper understanding of process mining practices and inform future research directions to better support process analysts and explain the cognitive processes underlying the analysis.

Keywords: Process Mining · Empirical Study · Data Exploration

1 Introduction

Process mining enables the discovery, conformance checking, and enhancement of business processes through the analysis of event logs recorded by information systems supporting process execution [1]. Over the last decade, process mining has gained remarkable momentum in academia and the industry, leading to a wealth of techniques that empower organizations to streamline and automate their business processes. However, so far, process mining research has privileged the development of algorithms, approaches, and tools from a technical viewpoint, paying less attention to learning how process analysts work in practice [13].

The work of analysts is often characterized by manual and knowledge-intensive tasks [6,22]. In particular, in the initial phases of process mining projects, analysts engage in different *exploration* activities [27], i.e., they dedicate time to familiarize themselves with the data to develop an understanding of the process [13], generate or refine questions, and discover new insights [7].

Exploration is a crucial step of discovery-oriented applications across many data-science-related disciplines [11], including data analysis [2,10,14] and min-

ing [16], and process mining [13,20,22]. For example, in data mining, the CRISP-DM methodology includes an “explore data” task within the “data understanding” phase, while the SEMMA and Two Crows models comprise an explicit “explore” step [16]. In process mining, exploration occurs mainly during event log inspection [22], in combination with process discovery [1,13], or whenever research questions lead to unexpected findings that need to be explained [7].

Recent interviews with data analysts [2,28] have shed light on exploration activities and related challenges that may benefit from guidance and tool support, such as choosing what to analyze or finding new insights. Given the affinity between data science and process mining, some of these findings may well apply to the latter area, raising the question of whether process analysts conduct similar activities or face similar challenges when exploring event logs and, if so, how they could be supported. However, so far, little attention has been paid to understanding how process analysts approach exploration in practice.

In this paper, we take a step in this direction and report the results of an empirical study designed to observe analysts as they engage in an exploratory process mining task. In this work, we specifically focus on “initial exploration”, i.e., the phase in which analysts familiarize themselves with an event log before addressing specific analysis questions [28]. Mainly, we aim to understand (i) what activities analysts perform, (ii) what target artifacts they focus on, i.e., what are the objects of their analysis, and (iii) what are exploration goals and strategies. Accordingly, we formulate the following research questions.

(RQ1) *What are the patterns of behavior of the initial exploration?* With RQ1, we aim to discover patterns of behavior focusing on sequences of exploration activities performed on different target artifacts.

(RQ2) *What are the goals and strategies of the initial exploration? How do they relate to different patterns of behavior?* With RQ2, we aim to understand what exploration goals are and what strategies analysts implement to achieve them. Also, we look at the relationship between goals and patterns of behavior to investigate if different goals lead to different patterns of behavior.

To address RQ1 and RQ2, we analyze different kinds of *behavioral data*, namely the interactions with process mining software and target artifacts and subjective insights about exploration goals and strategies in the form of verbal data. Our study reveals that analysts follow different behavior patterns when exploring event logs and work with various goals in mind, enacting different strategies to familiarize themselves with the data and understand the process. Our findings contribute to an initial understanding of exploration in process mining and highlight activities that could benefit from practical guidance, e.g., choosing suitable analysis techniques based on event log characteristics. Starting from these preliminary results, we suggest avenues for future research investigating (exploratory) process mining practices to better meet the practical needs of process analysts through guidance and support and enhance our ability to explain empirical findings with the development of respective theories.

Our paper unfolds as follows. Sect. 2 reviews related work. Sect. 3 presents the method followed to design and conduct the empirical study. Sect. 4 reports

the results of our analysis. Sect. 5 discusses the lessons learned and limitations of this study. Sect. 6 closes the paper and presents an agenda for future research.

2 Related Work

With this paper, we aim to advance the understanding of exploratory process mining practices by analyzing process analysts’ behavior in the form of interaction and verbal data. Thus, our work is related to research on understanding exploration activities and goals in data science and, in particular, process mining.

Exploration has its roots in Exploratory Data Analysis (EDA), introduced by Tukey in the 1970s as the “foundation stone” of any analysis [27].

Nowadays, exploration is an essential step of many data-science-related endeavors [11], including data [16] and process mining methodologies [1,7]. For example, the L* lifecycle model [1] includes an activity “explore” in “Stage 2”, the step aimed to create a control-flow model of the process under analysis. The PM² methodology [7] mentions “data-driven exploration” as a way to refine abstract questions since coming up with good questions at the start of a project is not easy. The authors also emphasize that exploration can generate unexpected findings and lead to concrete questions to investigate these findings further.

Related to our work are approaches investigating exploration activities in process mining. A few studies focused on supporting the exploration of process data, e.g., by easing the mapping of information among event data sources [26] or recommending interesting sets of process instances in event logs [22]. However, these approaches did not investigate how analysts approach exploration in practice. Klinkmüller et al. [13] examined process mining reports to understand the information needs of process analysts. The paper shed light on process mining practices in general, uncovering, for example, that analysts often combine familiarization with process discovery. Still, the study did not tap into the potential of user behavior analysis to unravel the dynamics of exploration processes.

So far, process mining research has paid little attention to understanding the behavior of analysts as they engage in process mining and, in particular, in exploration activities. Still, the analysis of interaction and verbal data has contributed to advancing user behavior understanding in neighboring areas, e.g., data science [2,10,14,28]. A growing body of literature has recently focused on understanding data exploration practices from different angles, uncovering typical analysis activities, goals, and related challenges. For example, the interview studies by Alspaugh et al. [2] and Wongsuphasawat et al. [28] revealed that analysts explore data for profiling and discovery goals but have to deal with fragmented tool spaces, repetitive tasks, and limited access to stakeholders. Liu et al. [14] interviewed data workers to understand cognitive, artifact, and execution alternatives of data sense-making processes, emphasizing the need to support the navigation and linking of alternatives across abstraction levels and tools. Han et al. [10] used interaction and eye-tracking data to understand the behavior of data workers as they discover data quality issues, uncovering patterns of information usage and search strategies that can inform the design of

data curation platforms. Besides providing insightful results, these studies uncover many new questions, remarking the need for more research to reveal the actual underlying practices of analysts [14] and enhance tool support [2,10,28].

To our knowledge, this is the first paper looking into the behavior of *process* analysts. Indeed, although some of the findings in [2,28] may apply to process mining, we advocate that it is not unlikely that the complex dynamics hidden in event logs would lead to unique exploration patterns, goals, and challenges.

3 Research Methodology

This section describes the planning and execution of our study and outlines the key aspects of the analysis conducted to address RQ1 and RQ2 (cf. Sect. 1).

Participants. For our study, we target academics and practitioners with varying levels of process mining experience and expertise. We recruited participants by reaching out to people in our professional networks. Participation was voluntary and based on the following requirements: (i) having analyzed at least one real-life event log over the past three years and (ii) having sufficient knowledge of Disco³, the process mining tool chosen for our study.

Materials. The task was designed to observe participants as they use Disco to analyze the road traffic fine management event log [5], a real-life event log representing the process of managing fines by the Italian police. We chose Disco as it is a commercial, easy-to-use tool, often used for initial analyses [1]. Based on the details in [15], we prepared a document (referred to as “artifacts”) with a description of the temporal constraints that need to be respected by the normative process, the activities and the attributes recorded in the event log, and a conceptual data model. The artifacts were intended as a source of domain knowledge and were at the participants’ disposal for the whole session⁴.

Procedure. We organized the task into an initial exploration, a guided exploration, and a semi-structured interview. We instructed participants to verbalize their thoughts in a think-aloud manner [8] as they performed the analysis. For the initial exploration, we gave participants up to 10 minutes to familiarize themselves with the event log. For the guided exploration, we provided participants with a guiding question aimed to replicate a high-level business goal and asked them to explore the event log at their own pace with this question in mind. Afterward, we assessed their understanding of the circumstances related to the guiding question with a comprehension test. Then, we repeated the procedure for a second guiding question. Finally, we interviewed participants to gain subjective insights into analysis strategies, goals, and challenges.

Execution. The data collection took place between July and September 2020 via Zoom⁵ meetings during which we recorded the participants’ think-aloud audio and their interaction with Disco and the artifacts, captured through screen

³ Fluxicon Disco: <https://fluxicon.com/disco/>

⁴ Link to the material: <https://drive.switch.ch/index.php/s/wevV2gXmoLBXrSY>

⁵ Zoom: <https://zoom.us>

ID	GENDER	A/I	POSITION	PM EXPERTISE		PROCESS ANALYSIS		#TOOLS
				OVERALL	ANALYSIS	FREQUENCY	#LOGS	
P1	M	A	PhD student	good	good	twice a year	> 5	2
P2	M	A*	PhD student	good	good	monthly	> 5	5
P3	F	A	PhD student	good	good	daily	> 5	4
P4	M	I	analyst	advanced	good	twice a year	2-5	7
P5	M	A*	post-doc	average	average	monthly	2-5	2
P6	M	A*	professor	advanced	good	weekly	> 5	3
P7	F	A*	post-doc	good	good	twice a year	2-5	5
P8	M	I	programmer	good	average	twice a year	2-5	5
P9	M	A	PhD student	good	good	monthly	1	1
P10	F	A*	professor	good	good	twice a year	2-5	3
P11	F	A*	PhD student	good	good	weekly	2-5	6
P12	F	A	PhD student	average	average	monthly	> 5	2

Table 1: Information about participants. I/A is a position in industry resp. academia, * marks experience in process mining projects with customers for academics; PM EXPERTISE is a Likert scale with values “novice”, “basic”, “average”, “good”, “advanced”; PROCESS ANALYSIS refers to the past 3 years; #LOGS is the number logs analyzed in the past 3 years; #TOOLS is the number of known process mining tools.

sharing. Overall, 14 people participated in our study. Before the meetings, we provided instructions and administered a background questionnaire. During the meetings, we ensured that Disco and the artifacts were visible on the shared screen and reminded participants to speak.

Data Validation and Analysis. For each participant, we collected around 2 hours of audio and video recordings. Two participants showed difficulties using Disco, and thus, we excluded them from the study. Table 1 lists the demographics and other information about the 12 selected participants. Participants are employed by ten different academic institutions or companies located in 5 different countries in Europe, Israel and Chile.

For the analysis, we followed a qualitative approach, focusing on the videos, the think-aloud, and the interview parts referring to the initial exploration. Initially, we transcribed all the audio recordings. Then, building upon the principles of grounded theory [4], we coded both the videos and transcripts iteratively, with the support of the MAXQDA software⁶. First, one author coded all the data. The other authors then checked the coded data independently, discussing disagreements and iteratively revising the codes to ensure consistency.

To investigate behavior patterns (cf. RQ1), we analyzed the interaction traces derived from the videos, i.e., the sequences of activities performed in Disco on certain target artifacts. We defined a coding scheme by combining the functionalities available in Disco, e.g., *inspect map*, typical data processing and analysis activities of the PM² methodology [7], e.g., *filtering*, activities happening outside Disco, e.g., *consult artifacts*, and the different aspects of the event log that can be analyzed, e.g., *paths*. Later, we augmented interaction traces with concurrent

⁶ MAXQDA <https://www.maxqda.com>

think-aloud and interviews to achieve triangulation and refine the coding. After several coding iterations, we obtained a hierarchy of 23 distinct codes for activities and 22 codes for target artifacts⁷. Then, we focused on identifying behavior patterns from the coded interaction traces. We started by selecting six high-level activities from the highest levels of our code hierarchy and exploiting different visualizations in MAXQDA to inspect and cluster similar sequences of coded segments qualitatively. After identifying potential candidate patterns, we backed up our qualitative insights with sequence alignments [12]. To this end, we sampled each participant’s interaction trace into segments of 10-sec length to keep track of the unfolding of activities over time. Then, for each segment, we took the code corresponding to the high-level activity performed for the largest amount of time in that segment and extended shorter sequences with segments of no interaction. As a result, we obtained twelve activity sequences of equal length on which we computed global pairwise trace alignments and used the obtained scores as a backup for our qualitative analysis.

To investigate exploration goals and strategies (cf. RQ2), we examined think-aloud and interview data. Initially, we fragmented the text using open and in-vivo coding [21]. Then, we refined and aggregated codes into categories using axial coding [4], obtaining a hierarchy of 21 codes related to goals and strategies. During axial coding, we relied on the videos to understand the participants’ statements in the context of exploration activities. Indeed, since some participants were not always describing what activities they were doing while speaking about goals, strategies were not emerging clearly from the isolated analysis of verbal data. Finally, we used selective coding to find relationships among the inferred categories [21], ending up with three main categories.

4 Findings

In this section, we present the results of our analysis for RQ1 and RQ2.

4.1 Patterns of Behavior of the Initial Exploration (RQ1)

With RQ1, we focus on understanding what high-level activities analysts perform during the initial exploration, in which order, for how long, and on which target artifacts. The analysis of interaction traces and verbal utterances allowed us to identify three main patterns of behavior, henceforth BP₁–BP₃. In the following, we describe the main features of each behavior pattern, focusing on the two most representative participants, i.e., those showing the best pairwise alignment score (cf. Sect. 3). In parentheses, we show the number of participants for which a particular observation holds when not clear from the text. Also, we use (T) to label statements taken from the think-aloud transcripts and (I) for interviews.

⁷ Link to coding scheme: <https://drive.switch.ch/index.php/s/m6wud73z4ztL0ym>

BP₁: Artifacts as an initial focus, followed by attention to mainstream behavior and variants. This behavior pattern was observed in five participants, namely P3, P4, P7, P9, and P11. Fig. 1(a) shows the sequences of codes (not normalized) derived from the interaction traces of P7 and P9 and the average time spent on specific high-level activities by all the participants in BP₁. All participants started by reading the artifacts containing information about the event log thoroughly. P7 claimed that the artifacts were useful to “better understand what was happening in the process”(1), while P9 said that “it felt more important to understand this PDF file [the artifacts] than just to play around with the event log”(1). Then, everyone inspected the process map in frequency view, i.e., the default view showing the control-flow model generated by Disco after loading the event log. All participants looked at the control-flow and the frequency of events, focusing on the most frequent paths. P7 used a variation filter to isolate the most frequent variants before inspecting the map. She then repeated the filtering and map inspection focusing on the least common ones. Performance metrics were used limitedly (2/5). Most participants (4/5) inspected the statistics after having explored the map, sometimes (3/5) going through the data attributes to know

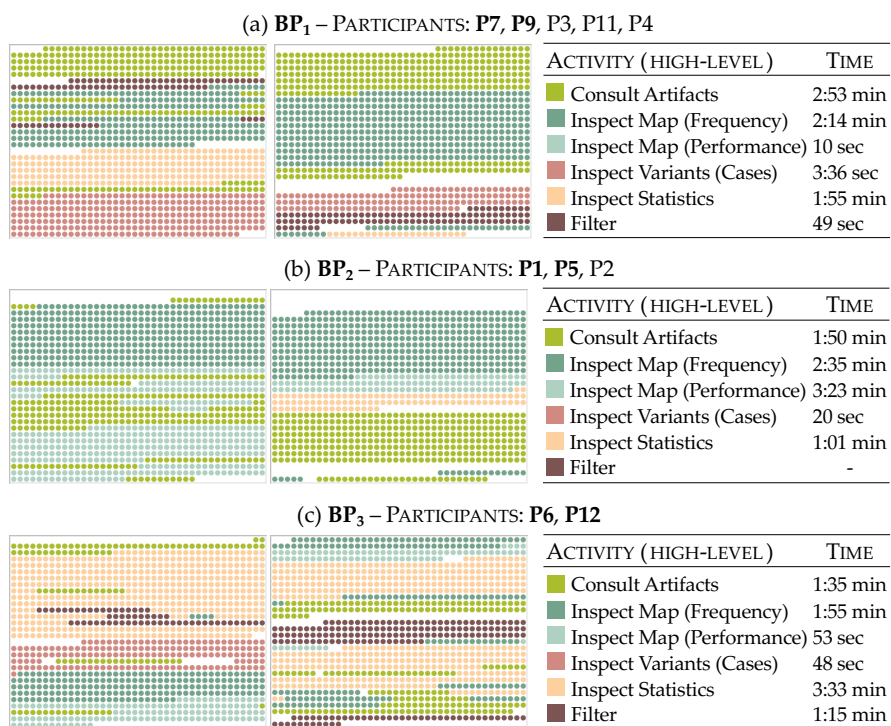


Fig. 1: Summary of behavior patterns (a) BP₁, (b) BP₂, and (c) BP₃. For each pattern: (Left) Rendering of coded interaction traces for representative participants (in bold); (Right) Average time spent on high-level activities by all participants.

“*what data we have there*”–P7 (T) and learn typical values. Finally, most participants (4/5) inspected the variants, focusing on the control-flow of the most frequent ones. P7 and P3 also inspected specific cases to find examples of parallel and repeated tasks and look for attribute correlation. Towards the end, P9 used variation and attribute filters to isolate “*the happy cases for the police*”(T).

BP₂: Attention to the shape of the data, performance metrics, and bottlenecks. This pattern was observed in P1, P2, and P5, with P1 and P5 being the most representative participants, as shown in Fig. 1(b). All participants started by focusing on the process map in frequency view, assuming activity labels to be self-explanatory. Indeed, artifacts were read at a later stage. All participants inspected the map at different levels of detail to check the structuredness or “*spaghetti-ness*” of the process, and two of them (2/3) focused on the most frequent paths. Participants focused on the shape of the data, as narrated by P2 “*The overall goal, I would say, is to get a feeling for the data [...] of how complex the data is*”(I) or P5 “*I was mostly driven by... by the dataset itself. I did some preliminary analysis from the structure, the time and the... attributes*”(I). Participants inspected the map in performance view at different stages of the exploration, focusing on paths exhibiting long duration. P1 examined the mean duration, alternating between the map and the temporal constraints provided in the artifacts to visually “check” them: “*I used them [the temporal constraints] at the beginning to see whether the process complies with these constraints*”–P1 (I). Instead, P5 used the performance metrics before reading the artifacts, searching for potential anomalies: “*I mostly focused on some anomalies and yeah, again, issues that could be present in the log*”–P5 (I). P2 and P5 also inspected the statistics after the map, focusing on the frequency of activities, variants’ distribution, and case duration. Then, they read the artifacts thoroughly, sometimes trying to explain the detected anomalies: “*If the constraint is not observed, the offender is not obliged to pay the fine. Ok, so in this case, the anomaly could be that we take too much to send it, and he doesn’t have to pay any longer*”–P5 (T).

BP₃: Attention to attribute distributions and patterns. This pattern was observed in P6 and P12, whose code sequences and average time spent on high-level activities are shown in Fig. 1(c). The common thread to the exploration of P6 and P12 was the tendency to observe distributions of cases and data attributes to spot patterns, mostly by using statistics and filters to create and compare different “scenarios”. Both participants inspected the statistics starting from the distribution of events over time and looking for patterns, e.g., “*I am now checking if there is any evident pattern. Like this is the 6th of April, this is the 6th of January...*”–P6 (T) and “*it looks like that in the middle of the month there are plenty of fines*”–P12 (T). When looking at attribute distributions, P6 decided “*to filter all the events without an article and check if... how is the distribution of those with respect to the other attributes*”(T). Similarly, P12 compared data attribute distributions in different scenarios that she created by filtering the log based on the temporal constraints described in the artifacts. P6 said that

attributes could help identify the causes underlying certain phenomena “*I would expect that if I can elicit the causes, I can do that from the attributes*”(I).

Comparing behavior patterns of the initial exploration. Fig. 2 sketches the behavior patterns that emerged from our analysis. While BP₁ and BP₂ capture opposite exploration approaches, BP₃ combines features of both of them.

Participants in BP₁ anchored their exploration to the domain knowledge embedded in the artifacts and explored the data in a top-down fashion through the lens of this knowledge, which helped them to guide their analysis, generate questions, and, sometimes, filter the data. P9 said “*I think most of my, um, knowledge came from just reading the artifacts [...] And I was thinking – Ok, why wouldn’t you pay a fine? Ok, maybe I don’t receive it by post or maybe I choose to ignore it...*” (I). They also spent a significant amount of time on process variants, looking into single cases better to understand the process and typical values of data attributes and their semantics.

In contrast, participants in BP₂ followed a bottom-up approach starting from the data. They focused on exploring the “problem space”, looking at the shape of the log, observing the frequency of activities and variants, and collecting the “hard numbers”–P2 (T). They identified the most frequent path of the process as the “happy path” and read the artifacts towards the end of the exploration as a way to check their understanding of the process from the data. They also examined performance metrics to spot bottlenecks or outliers, e.g., “*Once I found something that could be interesting, for instance, one activity lasting a huge amount of time, then I dug deeper [...] to understand why this occurred*”–P5 (I).

Despite spending comparable amounts of time on the map in frequency view, participants in BP₁ and BP₂ used it at different levels of detail. Those in BP₁ used it mainly at the default level of abstraction to visualize the activities that they had read in the artifacts. Two of them changed the detail level to assess the structuredness of the process, but the majority preferred to do so by looking at the variants. Instead, those in BP₂ worked at a lower level of abstraction, increasing the number of displayed paths to focus on repetitions and deviations.

Participants in BP₃ distinguished themselves for finding patterns in event and attribute distributions using statistics and filters to create and compare

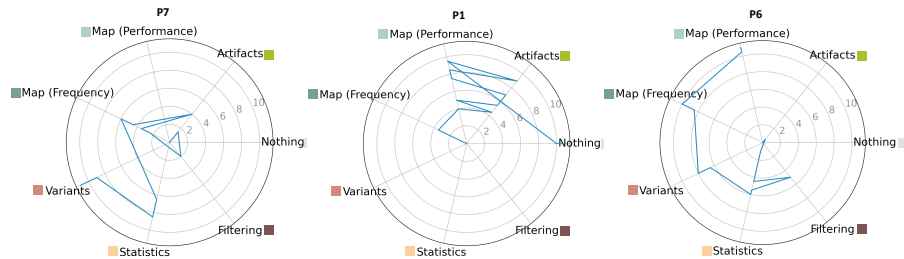


Fig. 2: Comparing the representative participants of BP₁, BP₂ and BP₃. Spider charts are obtained by sampling activity sequences in segments of 30-sec length and selecting the activity performed for the largest amount of time in that segment.

scenarios. They used the artifacts to derive the domain knowledge needed for filtering and understanding data attribute semantics, similar to BP₁, but, especially at the beginning, they focused on the shape of the data, similar to BP₂. e.g., *“it was more about understanding the structure of data, the shape of the log, rather than the meaning of it”*–P6 (1). Also, they looked at data attributes focusing on aggregated data rather than on specific values within cases (cf. BP₁).

When examining target artifacts, some commonalities among behavior patterns emerged. Indeed, all participants focused on the control-flow, particularly on the structuredness of the process (9/12) and the mainstream behavior (11/12). The interest for such process characteristics across behavior patterns can be explained by looking into exploration goals and strategies (cf. Sect. 4.2).

4.2 Goals and Strategies of the Initial Exploration (RQ2)

With RQ2, we aim to understand the goals of the initial exploration and what strategies analysts follow to achieve them. Also, we look at the relationships between goals and the behavior patterns introduced in Sect. 4.1. Thus, here we refer to the ten participants whose behavior is captured by BP₁–BP₃.

Goals emerged from the analysis of the think-aloud and interview data. Similar to [2,28], participants remarked their intention to *“become familiar with the data and the process before determining any direction”*–P3 (1). Indeed, their main goals were to (i) learn the main log characteristics, (ii) conceptualize the process, and (iii) identify interesting things to explore in further analyses (cf. Table 2). However, participants enacted different strategies to pursue these goals.

Learn the main log characteristics. All participants aimed to collect basic information about the log, e.g., general statistics such as log size and time covered, activity frequency and naming, number and kind of available attributes, and process structuredness. Learning about the event log was helpful to *“get familiar with the process”*–P4 (1) or *“know what information is essentially available”*–P6 (1). Despite having the same goal in mind, participants pursued enacted different strategies to achieve it. Some participants focused on *“understanding the log”*–P7 (1) and *“getting the background”*–P7 (1) from the artifacts. Instead, others concentrated on data shape and complexity, focusing mainly on quantitative information about the log. Participants also emphasized checking how structured the process was to choose what kind of analysis to apply and estimate

EXPLORATION GOAL	DESCRIPTION
LEARN MAIN LOG CHARACTERISTICS	Learn what the log contains and gather descriptive statistics about the log and the process
CONCEPTUALIZE THE PROCESS	Form an idea of the process and its context
IDENTIFY INTERESTING THINGS	Identify aspects that are deemed interesting and worth more in-depth analyses

Table 2: Main goals of the initial exploration emerged from our analysis.

GOAL: LEARN MAIN LOG CHARACTERISTICS			
Strategy	Quote	BP	#
Understand available information from the artifacts	“So like a little bit getting the background, so like, what are the tasks for what? And then also what are the attributes’ meanings for all the things you provided and the artifacts were really helpful. ”–P7 (1).	BP ₁ BP ₃	6
Assess data shape and complexity	“I think this is part of any data science project, not especially for process mining, but to get a feeling of how complex the data is and I think the first start it to... to look at the process map and see how spaghetti-like the model is and, um... then, also to get the hard numbers [...]”–P2 (1).	BP ₂ BP ₃	4
Assess process structuredness	“If the process would be extremely structured, I’d start guessing, for example, this is a xor-split [...] If it was extremely spaghetti, I would probably skip this inspection”–P6 (1)	All	8

Table 3: Strategies followed to learn the main event log characteristics.

the effort required. A few participants mentioned the further goal of splitting the log to reduce its complexity, e.g., “Why is it so complex? I would have looked into this and... potentially divided the log into sublogs”–P2 (1). Table 3 reports the described strategies, together with sample quotes, the behavior patterns (BP), and the number of participants (#) for which it was observed. Participants in BP₁ and one in BP₃ preferred to familiarize themselves with the semantics of activities and attributes over collecting quantitative data. Instead, most participants in BP₂ and one in BP₃ focused on the shape and complexity of the data. Most participants (8/10) across all behavior patterns assessed the structuredness of the process, looking at the map or the distribution of variants.

Conceptualize the process. Conceptualizing the process entails going beyond direct observation, bringing together domain knowledge, prior experience, and common sense to form an idea of how the process looks. Table 4 lists the strategies aimed to pursue this goal, which was deemed important by all participants. Unsurprisingly, all participants used common sense to conceptualize the process. However, those in BP₂ relied on common sense to also interpret activity and attribute names. Some participants (6/10) combined common sense with domain knowledge derived from the artifacts and prior experience. For example, P7 relied on her previous experience in using temporal constraints to “understand the payment culture”. All participants in BP₁ and one in BP₃ used the domain knowledge in the artifacts to learn about the context of the process and identify, for example, the “legal behavior”–P9 (T). Participants in BP₃ conceptualized the process also by establishing connections among different observations, e.g., scenarios created with the help of filters. All participants identified the mainstream behavior of the process based on event frequency. Besides being “the easiest thing to start with when you’re exploring a log that you don’t know”–P11 (T), the mainstream behavior helped participants to discern exceptional behavior, which, in turn, was one of the most interesting aspects to explore.

GOAL: CONCEPTUALIZE THE PROCESS			
Strategy	Quote	BP	#
Use common sense	<i>"I used common sense quite a lot [...] I had a vision of how the perfect process should look like"</i> -P2 (1)	All	10
Use domain knowledge and prior experience	<i>"These temporal constraints I know them from the medical field, it's always interesting to look at whether people follow this rule or don't follow this rule. "</i> -P7 (1)	BP ₁ BP ₃	6
Compare and combine observations	<i>"I inspected how those [frequent variants] look from a control-flow point of view. So, in this way, I constructed a bit of a mental model of the process. And once I got this mental model of which are actually the most frequent paths, I went back to the map and looked at these pairwise relationships with this knowledge of how to put the local relationships [...] in the larger spectrum of... a complete trace"</i> -P6 (1)	BP ₃	2
Identify mainstream behavior	<i>"We actually don't know what is exceptional [...] And that's why you focus on mainstream behavior first"</i> -P11 (1)	All	10

Table 4: Strategies followed to conceptualize the process.

Identify interesting things. A common exploration goal was to identify interesting aspects worth being further analyzed. All participants identified as interesting “unexpected things that do not meet their assumptions” or “cases showing infrequent behavior”, such as bottlenecks. To discover unexpected or exceptional and, thus, interesting things, participants followed the strategies described in Table 5. Some participants (6/10) relied on their knowledge of the mainstream behavior to build some “ground truth”, which helped them to establish what was exceptional. Others (3/10) used the temporal constraints to anchor to the normative process and check them “visually” from the process map or with the help of filters. Two participants (2/10) focused explicitly on detecting anomalies by looking for outliers in the data and bottlenecks in the process. All the par-

GOAL: IDENTIFY INTERESTING THINGS			
Strategy	Quote	BP	#
Use mainstream behavior as ground truth	<i>"I think that path [the most frequent] is the more interesting one because it would tell me that this is the most likely correct path and that the others would be outliers"</i> -P9 (1)	BP ₁ BP ₃	6
Check temporal constraints	<i>"The constraints helped me to, uh, try to retrieve the cases that are not in line with the constraints. [...] Because when you look at the process, you want to see unusual stuff. So, it's easy to go to the temporal constraints"</i> -P12 (1)	All	3
Detect bottlenecks or anomalies	<i>"I mostly focused on some anomalies and yeah, again, issues that could be present in the log; that could indicate that the process could be improved"</i> -P5 (1).	BP ₂	2

Table 5: Strategies followed to identify interesting things.

ticipants in BP₁ and one in BP₃ followed the first strategy. Three participants across all behavior patterns used the temporal constraints. Instead, the detection of anomalies was the main driver for participants in BP₂.

5 Discussion and Limitations

The objective of this paper was to investigate exploration practices in process mining, with a focus on patterns of behavior (cf. RQ1) and goals and strategies (cf. RQ2) of the initial exploration. We discovered that process analysts approach exploration in different ways, e.g., working top-down after gathering knowledge from the artifacts, or bottom-up, starting from the data and using the artifacts to check their understanding of the log, similar to open-ended analyses [2].

Analysts examined data attributes from different angles, either focusing on distributions or looking into single cases to explain how the process evolves. We also learned that the initial exploration has the main goals to (i) learn the main log characteristics, which resembles what is called *profiling* in the data science literature [28], (ii) conceptualize the process, and (iii) find interesting things that are relevant to analyze deeper (e.g., as part of concrete research questions), which is similar to *discovery*, i.e., gaining new insights or making hypotheses [28].

While (i) is a goal of “*any data science project, not especially for process mining*”-P2 (i) and the same holds for (iii), the structure and behavior of the underlying process require analysts to engage in specific exploration strategies.

For example, many participants assessed the structuredness of the process to decide which kind of analysis to conduct. To this end, they relied on the visual inspection of the process map and the variants and prior experience, having different perceptions of whether the process was “lasagna” or “spaghetti”. Although assessing process structuredness and, more in general, the structure and complexity of the data under investigation is typical of exploratory analyses [2,28], it is a nontrivial task that could tip the balance towards choosing an analysis technique over another. For example, two participants said they would split a spaghetti log into sub-logs before starting the analysis. Still, selecting the “right” pre-processing or analysis technique based on event log characteristics can be challenging [24], and we hypothesize that, especially for novices, it may be difficult to choose among analysis approaches and tools. This is particularly so when dealing with alternatives [14] or combining artifacts and techniques not specific to process mining [13]. Recent reviews of process discovery [3] and variants analysis [25] methods could help to compare different approaches, as well as the metrics in [24] could help to gain control-flow insights. However, practical guidance to support analysts in assessing event log characteristics (e.g., structuredness, presence, and quality of resource data) and, based on them, evaluating the applicability of pre-processing and analysis techniques is still missing.

Another insight emerging from our analysis is the attention paid to exceptional behavior and, particularly, negative deviations [18]. All participants identified exceptional behavior as the most valuable aspect to analyze, especially when seeking improvement opportunities. Improvement was perceived differently: par-

ticipants working top-down emphasized improvement opportunities for end-users (cf. BP₁), while those working bottom-up focused mainly on performance (cf. BP₂). Still, almost everyone tended to look for negative deviations, and only P9 focused on the “*positive cases for the police*” (τ). The focus on negative deviations is in line with the observation that inductive BPM approaches, including process mining, often put a strong focus on “responding to pain points” [20], while the exploration of opportunities is less emphasized. However, research has shown that processes can be improved by learning from positive deviations [23] or experience gained through past executions [9]. We find the tendency to look at negative deviations at the start of the analysis somewhat surprising. Indeed, the interviews in [2,28] did not mention interest for “negative” patterns, and the focus on deviations in the broader area of data science seems to be limited to spot outliers or data quality issues during profiling [28] or data curation tasks [10].

Overall, although some exploration goals overlap between data science and process mining, our analysis uncovered strategies that are not described in the data science literature, remarking the need for improving our understanding of process mining practices and develop targeted support for process analysts.

Limitations. Our study comes with some limitations. First, since we invited participants in our professional network, our sample was not drawn randomly from the overall process analysts population and is biased towards academics. Still, six academic participants were involved in process mining projects where the main goal was to analyze data for a customer (cf. Table 1) and, when interviewed, they provided insights into these experiences. Second, our study is subject to a limited number of participants. Still, twelve is considered an appropriate sample size for think-aloud studies [19], given the richness of the verbal data that we also augmented with interaction traces. Third, in this paper, we focused on exploratory process mining, i.e., the analysis phase aimed to familiarize with the data and refine abstract research questions. Thus, our findings cannot be generalized to other analysis phases, e.g., answering concrete questions. Finally, since all our subjects used Disco, we cannot claim that our findings can be generalized across tools. Indeed, Disco presents the map view after loading the event log and, thus, it would not be surprising if participants used the map at the beginning. Also, the functionality provided by the tool limits the analysis that our participants could perform and, as a consequence, the behavior we could observe in our data.

6 Conclusion and Outlook

In this paper, we have presented the findings of an empirical study investigating how process analysts approach exploration in practice, focusing on understanding common patterns of behavior, goals, and strategies of the initial exploration phase. Our findings revealed that analysts exhibit different behavior patterns when exploring an event log and work with different goals in mind, enacting various strategies to familiarize themselves with the data and the process. Besides providing novel insights into exploratory process mining practices, our findings can inform different directions for future research.

An obvious direction for future work is to improve the generalizability of our findings by conducting additional empirical studies to gain a comprehensive understanding of process mining across different analysis phases and tools. To this end, we will conduct quantitative and focused studies involving more practitioners and observing how analysts approach different analysis phases (e.g., familiarization, question answering) with one or more process mining tools. This will allow us to look into typical behavior patterns, goals, strategies, and challenges of process mining and, potentially, discern effective and non-effective behavior.

Gaining an in-depth understanding of (exploratory) process mining practices will provide empirical evidence for developing guidance and support for process analysts, for example, by enabling knowledge transfer from experienced analysts to novices. For instance, existing process mining methodologies could be complemented with practical recommendations, e.g., to choose analysis techniques based on event log characteristics or to foster the exploration of opportunity points [20], and tool support, e.g., to ease the comparison of “scenarios” (cf. BP₃) or process variants [25] along multiple process perspectives.

Another possible avenue for future research is the development of theories. In neighboring areas such as process modeling, cognitive theories have been used to explain empirical observations (e.g., [29]) and inform advancements, for example, to reduce the cognitive load when performing relevant tasks [17]. We suggest that the development of similarly relevant theories for explaining the cognitive processes involved in process mining, in general, and the exploratory phase, in particular, will contribute to advancement in this area.

Acknowledgment. This work is part of the Process Mining Support for End-users (ProMiSE) project, funded by the Swiss National Science Foundation (SNSF) under Grant No.: 200021_197032. We sincerely thank all our participants and Dr. Anne Rozinat, who provided us with Disco licenses.

References

1. van der Aalst, W.: *Process Mining—Data Science in Action*. Springer, 2 edn. (2016)
2. Alspaugh, S., Zokaei, N., Liu, A., Jin, C., Hearst, M.A.: Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Trans. Vis. Comput. Graphics* **25**(1), 22–31 (Jan 2019)
3. Augusto, A., Conforti, R., Dumas, M., Rosa, M.L., Maggi, F.M., Marrella, A., Mecella, M., Soo, A.: Automated discovery of process models from event logs: Review and benchmark. *IEEE Trans. Knowl. Data Eng.* **31**(4), 686–705 (2019)
4. Corbin, J., Strauss, A.: *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications (2014)
5. De Leoni, M., Mannhardt, F.: Road traffic fine management process. Eindhoven University of Technology. Dataset (2015)
6. Diba, K., Batoulis, K., Weidlich, M., Weske, M.: Extraction, correlation, and abstraction of event data for process mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **10**(3), e1346 (2019)
7. van Eck, M.L., Lu, X., Leemans, S.J., Van Der Aalst, W.M.: PM²: A process mining project methodology. In: *Int. Conf. on Advanced Information Systems Engineering (CAiSE)*. LNCS, vol. 9097, pp. 297–313. Springer (2015)

8. Ericsson, K.A., Simon, H.A.: Protocol analysis: Verbal reports as data. the MIT Press (1984)
9. Ghattas, J., Soffer, P., Peleg, M.: Improving business process decision making based on past experience. *Decision Support Systems* **59**, 93–107 (2014)
10. Han, L., Chen, T., Demartini, G., Indulska, M., Sadiq, S.W.: On understanding data worker interaction behaviors. In: *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. pp. 269–278. ACM (2020)
11. Idreos, S., Papaemmanouil, O., Chaudhuri, S.: Overview of data exploration techniques. In: *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*. pp. 277–281. ACM (2015)
12. Jagadeesh Chandra Bose, R.P., van der Aalst, W.: Trace alignment in process mining: Opportunities for process diagnostics. In: *Int. Conf. on Business Process Management*. LNCS, vol. 6336, pp. 227–242. Springer (2010)
13. Klinkmüller, C., Müller, R., Weber, I.: Mining process mining practices: An exploratory characterization of information needs in process analytics. In: *Int. Conf. on Business Process Management (BPM)*. pp. 322–337. Springer (2019)
14. Liu, J., Boukhelifa, N., Eagan, J.R.: Understanding the role of alternatives in data analysis practices. *IEEE Trans. Vis. Comput. Graphics* **26**(1), 66–76 (2019)
15. Mannhardt, F.: Multi-perspective process mining. Ph.D. thesis, Technische Universiteit Eindhoven (2 2018)
16. Mariscal, G., Marban, O., Fernandez, C.: A survey of data mining and knowledge discovery process models and methodologies. *Knowl. Eng. Rev.* **25**(2), 137 (2010)
17. Moody, D.: The “physics” of notations: toward a scientific basis for constructing visual notations in software engineering. *IEEE Trans. Softw. Eng.* **35**(6), 756–779 (2009)
18. Nguyen, H., Dumas, M., La Rosa, M., Maggi, F.M., Suriadi, S.: Mining business process deviance: A quest for accuracy. In: *On the Move to Meaningful Internet Systems: OTM 2014 Conferences*. pp. 436–445. Springer (2014)
19. Nielsen, J.: Estimating the number of subjects needed for a thinking aloud test. *Int. J. Hum. Comput. Stud.* **41**(3), 385–397 (1994)
20. Rosemann, M.: Explorative process design patterns. In: *Int. Conf. on Business Process Management*. pp. 349–367. Springer (2020)
21. Saldaña, J.: *The coding manual for qualitative researchers*. Sage (2015)
22. Seeliger, A., Sánchez Guinea, A., Nolle, T., Mühlhäuser, M.: ProcessExplorer: Intelligent Process Mining Guidance. In: *Int. Conf. on Business Process Management (BPM)*. LNCS, vol. 11675, pp. 216–231. Springer (2019)
23. Setiawan, M.A., Sadiq, S.: A methodology for improving business process performance through positive deviance. *Int. J. Inf. Syst. Model. Des.* **4**(2), 1–22 (2013)
24. Swennen, M., Janssenswillen, G., Jans, M., Depaire, B., Vanhoof, K.: Capturing process behavior with log-based process metrics. In: *SIMPDA*. pp. 141–144 (2015)
25. Taymouri, F., Rosa, M.L., Dumas, M., Maggi, F.M.: Business process variant analysis: Survey and classification. *Knowl. Based Syst.* **211**, 106557 (2021)
26. Tsoury, A., Soffer, P., Reinhartz-Berger, I.: A Conceptual Framework for Supporting Deep Exploration of Business Process Behavior. In: *Int. Conf. on Conceptual Modeling (ER)*. LNCS, vol. 11157, pp. 58–71. Springer (2018)
27. Tukey, J.W.: *Exploratory data analysis*, vol. 2. Reading, Mass. (1977)
28. Wongsuphasawat, K., Liu, Y., Heer, J.: Goals, process, and challenges of exploratory data analysis: An interview study. arXiv:1911.00568 (2019)
29. Zugal, S., Soffer, P., Haisjackl, C., Pinggera, J., Reichert, M., Weber, B.: Investigating expressiveness and understandability of hierarchy in declarative business process models. *Software & Systems Modeling* **14**(3), 1081–1103 (2015)