NOVA IMS Information Management School

MGI

Master Degree Program in
Statistics and Information Management

CHALLENGES, FRAMEWORKS AND FUTURE DIRECTIONS IN PROCESS DISCOVERY

Navigating the Evolution of Process Discovery

Lilit Yezekyan

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CHALLENGES, FRAMEWORKS AND FUTURE DIRECTIONS IN PROCESS DISCOVERY
NAVIGATING THE EVOLUTION OF PROCESS DISCOVERY

By

Lilit Yezekyan

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, Specialization in Information Analysis and Management

Supervisor: Professor Doutor Vítor Duarte dos Santos

November 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[student signature]

[place, date]

DEDICATION

This thesis is dedicated to my cherished mother, whose love, wisdom, and unwavering faith in me laid the foundation for this journey. Her dreams for me have shaped the path I took to become the scientist she always envisioned. Though she is no longer physically with us, her spirit continues to guide me every day. This achievement is as much hers as it is mine, and I hope her soul is proud of the person and scientist I strive to be.

This work is also dedicated to my devoted father, whose constant support and belief in my abilities have been my inspiration at every turn. Dad, you always dreamed of me becoming a scientist, and as I stand at the pinnacle of this journey, I carry your dream with me. Your influence echoes in every discovery and accomplishment I make. This thesis is a tribute to your vision, love, and sacrifices that kept me going. I hope I continue to make you proud as I explore the world with the same passion and curiosity you instilled in me.

Lastly, this work is dedicated to my beloved sister, whose encouragement, laughter, and unwavering belief in me have been a constant source of strength. Through every challenge, your support has reminded me of the power of family and the bond we share. Your love has been a guiding force, and I am deeply grateful to have you by my side as I embark on this new chapter of my life.

ACKNOWLEDGEMENTS

This thesis is deeply indebted to my supervisor, Professor Santos, whose unwavering support and mentorship have been pivotal throughout my academic journey. Your consistent encouragement, especially in moments of uncertainty, and your patience during the more challenging phases, have made this accomplishment possible. You fostered my intellectual growth, constantly motivating me to strive for excellence while honoring my independence. I am profoundly grateful for the invaluable guidance you have offered, which has not only made this achievement possible but also imbued it with significant personal meaning. Thank you for your dedication and for being a constant source of inspiration.

I would also like to express my gratitude to CRRC-Armenia for providing a supportive environment to test and refine the theoretical concepts of my thesis. The organization's resources and collaborative atmosphere allowed me to evaluate my ideas critically and gain invaluable insights. I am deeply appreciative of the intellectual exchange and opportunities for growth that CRRC-Armenia has provided throughout this journey.

ABSTRACT

Process mining, a discipline that evolved from Business Process Management (BPM), has seen significant advancements, particularly in contrasting approaches such as **process mining vs. task mining**, **case-centric vs. object-centric process mining**, and the distinction between **process models and process architectures**. Core components of process mining include **process discovery**, **conformance checking**, and **process enhancement**. Among these, **process discovery** has become a focal point for automation efforts. Currently, automated process discovery incorporates **domain knowledge**, and future developments aim to deepen this integration, refining algorithms to enhance automation.

To address the challenges in this domain, a **Systematic Literature Review (SLR)** was conducted within the framework of a thesis. Based on the findings, a **recommendation framework for process discovery** was proposed, emphasizing the critical roles of **event log pre-processing** and **domain knowledge utilization** in achieving reliable outcomes. The study aligns with existing literature to move closer to **fully automated process discovery** and provides a comprehensive mapping of algorithm families, detailing the specific challenges each addresses. This framework serves as a guideline for advancing automated and domain-informed process discovery methodologies.

KEYWORDS

Process model discovery; Process mining; Event logs; Automation of process discovery; Process discovery algorithms; Domain-knowledge; Quality databases; Incomplete data; Noise in event logs; Process architecture; Interrelated processes; Design artifact; High-quality process models

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AA	Alpha Algorithm
APD	Automatic Process Discovery
B2C	Business-to-Consumer
BPM	Business Process Management
CA	Cluster Analysis
GA	Genetic Algorithms
GAA	General Algorithmic Approaches
HM	Heuristic Miner
IIoT	Industrial Internet of Things
IM	Inductive Miner
IPD	Interactive Process Discovery
LLM	Large Language Model
MA	Markovian Approaches
ML	Machine Learning
NN	Neural Networks
OCPM	Object-Centric Process Mining
PCPM	Process-Centric Process Mining
PDA	Process Discovery Algorithm
PM	Process Mining

1. INTRODUCTION

1.1 BACKGROUND AND PROBLEM IDENTIFICATION

In 2011, the Process Mining Manifesto highlighted a key shift from traditional Business Process Management (BPM) by emphasizing the importance of analyzing and improving organizational processes through data (Brock, 2024). As businesses increasingly depend on data for efficiency, process mining has become a crucial link between data mining and process modeling. Unlike traditional data mining, it focuses on process structures and effectively handles complex workflows. This shift reflects the rapid digital growth over the past 50 years, enabling organizations to capture diverse event data, such as ATM transactions or medical appointments. However, much of this data's potential for driving sustainable progress remains unexplored (Beerepoot et al., 2023). Process mining plays a critical role in advancing Industry 4.0 by focusing on process-centric approaches to address the growing challenges posed by unstructured data in modern organizations. The core aim of process mining is to use event data to uncover insights, resolve bottlenecks, ensure compliance, and suggest improvements (Brock, 2024). It includes three main tasks: process discovery, conformance checking, and process optimization (Beerepoot et al., 2023). Process mining provides a valuable tool for uncovering process models embedded within databases, enabling improvements through comparative analysis with actual operations (Sedlakova, 2023).

Process mining has recently evolved to integrate task mining, complementing traditional process mining approaches. This expansion acknowledges that individual process models often fail to capture the full complexity of organizational systems. To address this, process mining introduced the concept of process architecture, a comprehensive framework consisting of interconnected process models. This shift enables a transition from case-centric to object-centric process mining, providing a broader, more integrated perspective of organizational workflows. By combining case-centric and object-centric approaches, process mining now offers a more balanced view that captures both the overall process flow and detailed insights into individual objects within the process (van der Aalst, 2023). In the contemporary landscape of process mining, key techniques like process discovery, conformance checking, and process enhancement are essential for addressing inefficiencies, ensuring compliance, and optimizing operations (Berti, 2023). Additionally, the distinction between process models and process architectures signifies a shift from mere workflow visualization to the design of resilient, scalable systems capable of adapting to complex organizational demands (Oldenburg, 2024). Figure 1.1 illustrates the interdisciplinary nature of process mining, showcasing how contributions from various fields, such as data mining and BPM, have enriched its methodologies and expanded its scope. It also highlights the core areas of interest within process mining, including task mining, object-centric process mining and process architecture, emphasizing their alignment with the evolving demands of modern organizational workflows.

Figure 1 - Process mining connections with other fields (inspired by Reinkemeyer, 2022)

Figure 1.2 shows emerging trends in process mining include intelligent process execution, proactive and predictive solutions, as well as the impact on digital workforce and data democratization. The integration of cloud technology and IIoT platforms is also at the forefront. These trends lay the foundation for mid-term developments in process mining, such as self-learning and self-optimizing systems, artificial intelligence, and benchmarking. In the long term, process mining is expected to enhance inter-company processes, sustainability, and focus on B2C models. Central to this evolution is the improvement of event log quality, which is crucial for mining comprehensive process models. (Reinkemeyer, 2022).

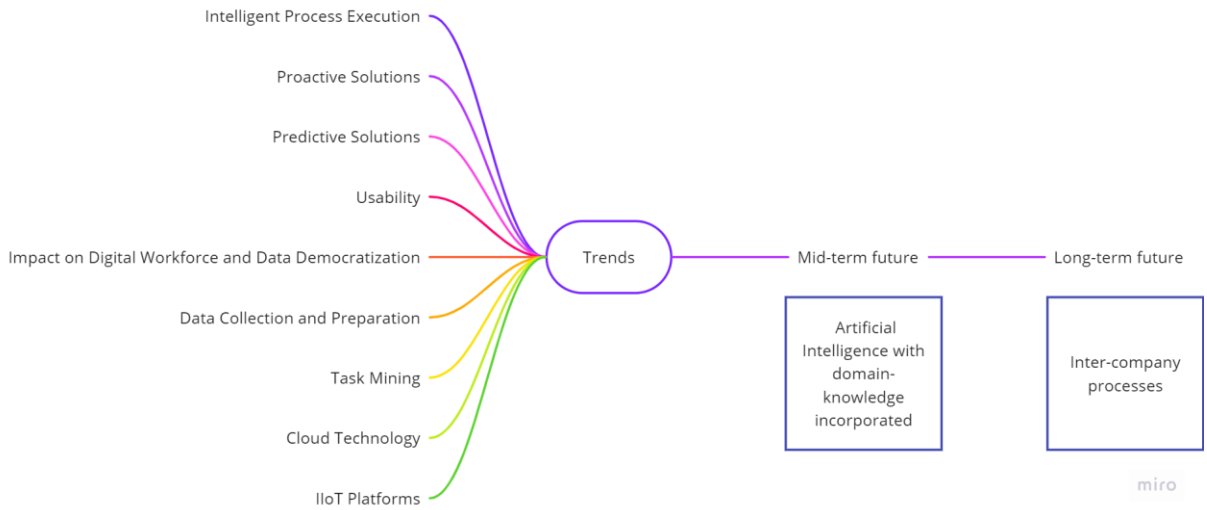


Figure 2 - Process mining trends and future directions (Reinkemeyer, 2022)

Challenges such as poor-quality event logs, which may be incomplete, noisy, or excessively detailed, can significantly affect the accuracy of process mining results. Additionally, factors like concept drift and representational bias add further complexity. Despite over 100 process discovery algorithms, researchers are continually assessing their effectiveness and suitability in different scenarios (Brock et al., 2024; Beerepoot et al., 2023). Traditional data mining methods also struggle with issues like concurrency, prompting the need for innovative algorithms and new data representations (Monti, 2024). The core objective of process mining is to leverage event data for extracting actionable insights, such as automatically discovering process models that capture dynamic behaviors. More than just analyzing performance metrics, process mining also identifies causal relationships between activities and facilitates knowledge extraction from historical data, thereby enriching ongoing case analysis (Monti, 2024).

Process mining has found applications across several industries, helping to improve operational efficiency and process visibility. Key sectors include:

- **Finance and Banking**: Enhancing operations, ensuring compliance, and improving customer service.
- **Healthcare**: Streamlining patient flow, reducing waiting times, and optimizing resource allocation.
- **Manufacturing**: Monitoring production, eliminating bottlenecks, and ensuring quality control.
- **Logistics**: Optimizing supply chains and improving delivery times.
- **Insurance**: Streamlining claims processing and refining risk assessments.
- **Education**: Analyzing student engagement and improving administrative workflows.
- **Retail**: Enhancing customer experience and optimizing inventory management.
- **Information Technology**: Improving software development and IT service management.

1.2 Event Logs and Challenges in Process discovery

> *"Definition: A process discovery algorithm is a function that maps event log onto a process model such that the model is "representative" for the behavior seen in the event log. The challenge is to find such an algorithm."*

The pursuit of fully automated process model discovery presents significant challenges. Ongoing research focuses on enhancing process mining algorithms, addressing key concerns such as improving model accuracy, ensuring the quality of event logs, and developing tools that are user-friendly for non-experts (Dakic, 2023; Maggi et al., 2018). As process mining becomes increasingly prevalent in organizational settings, the demand for standardized methodologies is growing to ensure consistent and reliable process discovery and analysis (Celonis, 2023; Maggi et al., 2018). Despite its critical role in the success of process mining initiatives, the process discovery phase is often impeded by challenges in efficiently and accurately deriving process models from event logs (Dakic, 2023; Maggi et al., 2018). These challenges can be broadly categorized into several problem areas:

- **Noise in Data:** Logged data often contain inaccuracies or missing information, which complicates the mining process. Noise in the data can lead to incorrect or incomplete process models, reducing the reliability of the discovered processes ().
- **Hidden Tasks:** Some tasks may exist within a process but are not captured in the event logs, making them "hidden" and difficult to detect during process discovery. This leads to incomplete process models that fail to represent the true workflow.
- **Duplicate Tasks:** The presence of duplicate tasks, where two process nodes refer to the same process step, can create confusion and redundancy in the process model, complicating the interpretation and analysis.

- **Non-Free Choice Constructs:** These constructs represent controlled choices within a process that depend on decisions made elsewhere in the process model. This dependency makes it difficult to accurately capture and model the process flow.
- **Mining Loops:** Processes often include loops, where certain steps are repeated multiple times. Mining these loops can be challenging, especially when the loops are complex and involve multiple events.
- **Different Perspectives:** Process events may be recorded with additional contextual information for mining purposes. However, the varying perspectives captured in the event logs can complicate the process discovery, as they may introduce inconsistencies or additional noise.
- **Delta Analysis:** Comparing the discovered process model with a reference model to identify similarities and disparities is crucial for validating the accuracy of the process discovery. However, delta analysis can be difficult when the models are complex or when there are significant deviations.
- **Visualizing Results:** Presenting the results of process mining in a clear and understandable graphical form is essential for decision-making. However, visualizing complex process models in a way that is accessible to non-experts remains a significant challenge.
- **Heterogeneous Results:** Process mining often involves accessing and analyzing data from information systems based on different platforms. This heterogeneity can complicate the integration of data and the discovery of coherent process models.
- **Concurrent Processes:** When multiple processes occur simultaneously, mining them accurately without losing critical information or creating overlaps in the process models is a complex task.
- **Local/Global Search Strategies:** Local search strategies limit the search space, reducing complexity but also the likelihood of finding the optimal solution. Conversely, global search strategies, while more comprehensive, are computationally expensive and difficult to implement effectively.
- **Process Re-discovery:** Selecting a mining algorithm that can accurately rediscover a class of process models from a complete workflow log is a non-trivial task, requiring deep knowledge of both the algorithms and the specific context of the data.

These challenges underscore the fact that process discovery is more of an art than a fully automated process. The success of process discovery depends heavily on the expertise of the individual conducting it, particularly in terms of selecting and combining the right algorithms, as well as effectively cleaning and preparing the database. Although the field has made strides towards automating process discovery, significant progress is still needed. The primary obstacles to automation include the poor quality and inconsistencies within existing databases, as well as the inherent complexity of real-world processes. Several algorithms have been developed to address these challenges, each with its strengths and limitations. Among the primary algorithms are the Alpha algorithm, Heuristic Miner, Inductive Miner, and Genetic Algorithms.

- **Alpha Algorithm**: The Alpha algorithm reconstructs process models by identifying causality relationships between events based on event logs. It is one of the earliest process mining algorithms, effective in discovering simple, structured workflows. However, it struggles with noise, loops, and non-free choice constructs (Porouhan et al, 2014).

- **Heuristic Miner**: This algorithm improves on the Alpha algorithm by addressing noise and incomplete data issues. It uses frequency-based heuristics to detect the most likely relationships between events, making it suitable for real-world logs with noise or irregularities. Heuristic Miner produces more robust models for unstructured processes but requires careful parameter tuning (Bousdekis et al., 2023).
- **Inductive Miner**: The Inductive Miner creates process models by recursively dividing event logs into smaller parts, ensuring soundness and high precision in the discovered models. It is particularly adept at handling complex loops, concurrency, and nested processes. This algorithm is well-suited for creating hierarchically structured models that are easier to validate (AlQaheri et al., 2022).
- **Genetic Algorithms:** Inspired by Darwinian natural selection, genetic algorithms (Alves de Medeiros et al., 2004) are designed to evolve solutions over time, making them suitable for reducing noise and mining hidden tasks in event logs. Despite their adaptability, these algorithms can be computationally expensive and may struggle with large, complex datasets.
- **General Algorithmic Approaches:** These are custom algorithms created by individual researchers to mine specific processes (van der Aalst and Song, 2004). While they can be tailored to particular contexts, their specificity often limits their generalizability across different types of processes and datasets.
- **Markovian Approaches:** Markovian algorithms examine both past and future behavior to determine potential current states (Cook and Wolf, 1998). These approaches are effective for modeling processes with probabilistic transitions but may become less accurate in environments with significant noise or when data is incomplete.
- **Neural Networks:** Neural networks mimic the human brain's ability to learn and identify patterns in data (Cook and Wolf, 1998). While powerful, these models require extensive training data and computational resources, and their "black box" nature can make the discovered process models difficult to interpret.
- **Cluster Analysis:** This approach groups similar solutions into homogeneous subgroups (Schimm, 2004), which can help identify patterns and dependencies within the data. However, clustering is sensitive to noise and the choice of parameters, which can significantly affect the quality of the discovered process models.

Despite the availability of these diverse algorithms, several persistent problems make process model discovery challenging:

- **Dependency Patterns:** Identifying and accurately modeling dependencies between different process steps is complex, particularly in processes with intricate interactions.
- **Noise in Event Logs:** Event logs often contain errors, missing data, and irrelevant information, which can distort the process model. Noise treatment is one of the most frequently addressed issues in process mining.
- **Complexity in Event Logs:** The complexity of real-world event logs, with their multiple layers of interactions and varied data types, makes it difficult to extract clear and accurate process models.
- **Mining Loops:** Processes that involve repeated cycles or loops pose a challenge, especially when these loops vary in complexity and involve multiple events.

- **Concurrent Processes:** Mining processes that occur simultaneously require sophisticated techniques to distinguish and accurately model overlapping activities.

The primary motivation for this thesis is to advance the field of process mining by addressing the technical and methodological challenges inherent in process model discovery. Specifically, the research aims to develop a recommendation framework designed to improve the selection of appropriate process discovery algorithms. This framework seeks to ensure that the chosen algorithms align with specific contexts and requirements, thereby enhancing the effectiveness and accuracy of process mining initiatives (Jalonen, 2023; Di Federico, 2023)

1.3 RESEARCH QUESTIONS

This situation led to the formulation of the following research questions:
RQ1: How can event log pre-processing methods be improved to address challenges in automated process discovery, and what role does domain knowledge play in this enhancement?
RQ2: What are the key differences between existing frameworks for process discovery, and how do they address issues such as noise, hidden tasks, and duplicate tasks in event logs?
RQ3: What are the future directions for integrating process discovery with advanced machine learning algorithms to support fully automated, domain-informed methodologies?

1.4 GOAL AND FOCUS

The main goal of this investigation is to address the research question: **"How to advance the field of process discovery by addressing foundational challenges in event log pre-processing, understanding the comparative effectiveness of discovery frameworks, and integrating cutting-edge machine learning techniques for more automated, domain-informed process discovery systems"**

To answer this, the investigation will focus on:

1. Investigate how to improve the accuracy and completeness of event log data by mitigating challenges such as noise, missing data, and task duplication.
2. Compare existing process discovery frameworks to identify their unique strengths and weaknesses in addressing noise, hidden tasks, and duplicate tasks.
3. Proposing a structured framework that systematically addresses these challenges to enable greater automation in process model discovery.

1.5 OBJECTIVES

The study's objectives are to advance the field of automated process discovery by:

1. Improving event log pre-processing methods to address challenges such as noise, hidden tasks, and duplicate activities through the integration of domain knowledge.

2. Comparing and evaluating existing frameworks to identify their relative effectiveness in mitigating common event log issues.
3. Proposing future directions for merging process discovery techniques with advanced machine learning algorithms to create fully automated and domain-informed discovery methodologies.

2. LITERATURE REVIEW

The primary objective of this dissertation is to develop a **process discovery recommendation framework** that synthesizes findings from existing literature to address key challenges inherent in event logs. This framework systematically maps process discovery algorithms to the specific log-related issues they tackle, such as noise, hidden tasks, and concurrency, as well as their alignment with various process mining requirements.

To achieve this, the study conducts a **comprehensive evaluation of process discovery approaches**, examining their effectiveness against critical performance metrics such as fitness, precision, and generalization. Emerging trends and future directions in process discovery are also analyzed to ensure the framework remains relevant in evolving process mining landscapes.

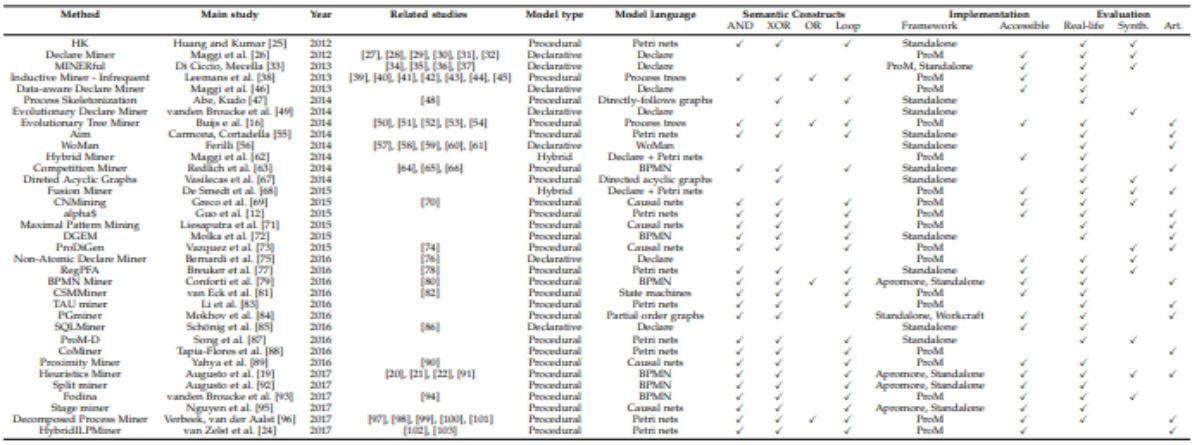
- **Section 2.1** outlines the evaluation criteria for comparing process discovery algorithms, providing a detailed analysis of their strengths, limitations, and applicability to different process discovery scenarios.
- **Section 2.2** demonstrates the practical application of these algorithms, emphasizing how they resolve specific challenges in process model derivation, including non-free choice constructs, complex loops, and heterogeneous event data.
- **Section 2.3** focuses on **event log repair methodologies**, detailing strategies for handling noise, missing data, and other quality issues. These techniques are integrated into process discovery workflows to enhance the reliability of mined models and support real-time operational optimization.
- **Section 2.4** highlights the role of **domain knowledge integration** in improving process discovery accuracy, particularly for context-specific processes. Additionally, it explores emerging innovations, such as hybrid discovery approaches and the incorporation of AI-driven enhancements, which shape the future of process discovery.
- **Section 2.5** delves into the process and significance of conducting a **systematic literature review (SLR)**. A systematic literature review is a structured approach to reviewing and synthesizing existing research on a specific topic, ensuring that all relevant studies are considered. Unlike traditional literature reviews, which may be more narrative in nature, an SLR adheres to strict methodologies to minimize bias and ensure comprehensive coverage of the topic.

This structured approach ensures the recommendation framework is robust, adaptable, and capable of addressing the diverse complexities of real-world event logs.

2.1 MAPPING EXISTING PROCESS MINING ALGORITHMS BY THEIR CHARACTERISTICS

The rapid development of the process mining field has led to a development of algorithms, each made to address specific challenges and requirements. Periodic assessments of these algorithms are essential for maintaining an up-to-date understanding of their capabilities and limitations (Huang & Kumar, 2012). Metrics for assessing the complexity of business process models play a pivotal role in development of the process mining field, as they can predict potential sources of errors in processes (van der Aalst, 2016). Quantifying the complexity of a process model helps researchers and practitioners to determine the suitability of different algorithms for its discovery. This makes

complexity metrics a valuable tool for evaluating and benchmarking process discovery algorithms (Rozinat & van der Aalst, 2008).



Method	Main study	Year	Related studies	Model type	Model language	Semantic Constructs				Implementation		Evaluation		
						AND	XOR	OR	Loop	Framework	Accessible	Real-life	Synth.	Art.
HK	Huang and Kumar [25]	2012		Procedural	Petri nets	✓	✓		✓	Standalone		✓	✓	
Declare Miner	Maggi et al. [26]	2012	[27], [28], [29], [30], [31], [32]	Declarative	Declare					ProM	✓	✓	✓	
MINERful	Di Ciccio, Mecella [33]	2013	[34], [35], [36], [37]	Declarative	Declare					ProM, Standalone	✓	✓	✓	
Inductive Miner - Infrequent	Leemans et al. [38]	2013	[39], [40], [41], [42], [43], [44], [45]	Procedural	Process trees	✓	✓	✓	✓	ProM	✓	✓		
Data-aware Declare Miner	Maggi et al. [46]	2013		Declarative	Declare					ProM	✓	✓		
Process Skeletonization	Abe, Kudo [47]	2014	[48]	Procedural	Directly-follows graphs		✓		✓	Standalone		✓		
Evolutionary Declare Miner	vanden Broucke et al. [49]	2014		Declarative	Declare					Standalone			✓	
Evolutionary Tree Miner	Buijs e al. [16]	2014	[50], [51], [52], [53], [54]	Procedural	Process trees	✓	✓	✓	✓	ProM	✓	✓		✓
Aim	Carmona, Cortadella [55]	2014		Procedural	Petri nets	✓	✓		✓	Standalone		✓		✓
WoMan	Ferilli [56]	2014	[57], [58], [59], [60], [61]	Declarative	WoMan					Standalone		✓		✓
Hybrid Miner	Maggi et al. [62]	2014		Hybrid	Declare + Petri nets					ProM	✓	✓		
Competition Miner	Redlich et al. [63]	2014	[64], [65], [66]	Procedural	BPMN	✓	✓		✓	Standalone		✓		✓
Direted Acyclic Graphs	Vasilecas et al. [67]	2014		Procedural	Directed acyclic graphs		✓			Standalone		✓	✓	
Fusion Miner	De Smedt et al. [68]	2015		Hybrid	Declare + Petri nets					ProM	✓	✓	✓	✓
CNMining	Greco et al. [69]	2015	[70]	Procedural	Causal nets	✓	✓		✓	ProM	✓	✓	✓	
alpha$	Guo et al. [12]	2015		Procedural	Petri nets	✓	✓		✓	ProM	✓	✓		✓
Maximal Pattern Mining	Lieuwpatra et al. [71]	2015		Procedural	Causal nets	✓	✓		✓	ProM		✓		✓
DGEM	Molka et al. [72]	2015		Procedural	BPMN	✓	✓		✓	Standalone		✓		✓
ProDiGen	Vazquez et al. [73]	2015	[74]	Procedural	Causal nets	✓	✓		✓	ProM			✓	✓
Non-Atomic Declare Miner	Bernardi et al. [75]	2016	[76]	Declarative	Declare					ProM	✓	✓	✓	
RegPFA	Breuker et al. [77]	2016	[78]	Procedural	Petri nets	✓	✓		✓	Standalone	✓	✓	✓	
BPMN Miner	Conforti et al. [79]	2016	[80]	Procedural	BPMN	✓	✓	✓	✓	Apromore, Standalone	✓	✓		✓
CSMMiner	van Eck et al. [81]	2016	[82]	Procedural	State machines	✓	✓		✓	ProM	✓	✓		
TAU miner	Li et al. [83]	2016		Procedural	Petri nets	✓	✓		✓	ProM		✓		✓
PGminer	Mokhov et al. [84]	2016		Procedural	Partial order graphs	✓	✓			Standalone, Workcraft	✓	✓		✓
SQLMiner	Schönig et al. [85]	2016	[86]	Declarative	Declare					Standalone	✓	✓		
ProM-D	Song et al. [87]	2016		Procedural	Petri nets	✓	✓		✓	Standalone		✓	✓	
CoMiner	Tapia-Flores et al. [88]	2016		Procedural	Petri nets	✓	✓		✓	ProM				✓
Proximity Miner	Yahya et al. [89]	2016	[90]	Procedural	Causal nets	✓	✓		✓	ProM	✓	✓		
Heuristics Miner	Augusto et al. [19]	2017	[20], [21], [22], [91]	Procedural	BPMN	✓	✓		✓	Apromore, Standalone	✓	✓	✓	✓
Split miner	Augusto et al. [92]	2017		Procedural	BPMN	✓	✓		✓	Apromore, Standalone	✓	✓		
Fodina	vanden Broucke et al. [93]	2017	[94]	Procedural	BPMN	✓	✓		✓	ProM	✓	✓	✓	
Stage miner	Nguyen et al. [95]	2017		Procedural	Causal nets	✓	✓		✓	Apromore, Standalone	✓	✓		
Decomposed Process Miner	Verbeek, van der Aalst [96]	2017	[97], [98], [99], [100], [101]	Procedural	Petri nets	✓	✓	✓	✓	ProM	✓	✓		✓
HybridILPMiner	van Zelst et al. [24]	2017	[102], [103]	Procedural	Petri nets	✓	✓		✓	ProM	✓			✓

Figure 3 - Mapping of studies on process mining tools (Huang et al., 2012)

Lee and Yoon made significant contributions to defining metrics for Petri net process models by classifying them into structural and dynamic categories. Structural metrics focus on elements such as the number of places, transitions, arcs, and the cyclomatic complexity of the control graph, providing insights into the model's architecture. In contrast, dynamic metrics assess behavior by counting the number of markings and measuring the maximum and average number of tokens in both original and reduced process models (Lee & Yoon, 2004).

Building on this, Nissen integrated design heuristics and knowledge-based systems to evaluate process model quality. Metrics proposed by Nissen include the count of paths, hierarchy levels, nodes, cycles, the diameter of the model, and parallelism, calculated as the ratio of nodes to the diameter (Nissen, 2008). These metrics offer a robust framework for assessing the complexity and quality of process models, contributing to more effective algorithm selection and process optimization.

Tjaden, Narasimhan, and Gupta addressed four critical characteristics of business processes—simplicity, flexibility, integration, and efficiency—focusing on balancing these dimensions. **Simplicity** was quantified by basic process complexity, measured as the sum of nodes, arcs, and roles. The overall simplicity was calculated as the ratio of average activity complexity to maximum activity complexity. **Flexibility** and **integration** were assessed using a scoring methodology similar to function point analysis (Tjaden et al., 1998).

Morasca proposed a robust set of metrics for Petri nets grounded in a theoretical framework. These metrics are designed to fulfill specific axiomatic properties, including size, length, structural complexity, and coupling, to evaluate Petri net processes effectively (Morasca, 1999).

Additionally, Augusto et al. provided a comprehensive framework for evaluating process discovery algorithms using a variety of metrics, detailed in their study. These metrics informed the development of mapping approaches (Table 2.1) for various tools utilized in process mining. Evaluations of process discovery algorithms in the literature were conducted based on these metrics, offering insights into their performance and applicability (Augusto et al., 2019).

Table 1 - Mapping existing process mining algorithms (Augusto et al., 2019)

Algorithms	Model	Metrics									
		Completeness	Constructs							Abstraction	Fitness
			seq	par	cho	lo	nfc	it	dt		
Abstraction based algorithms											
alpha	Petri nets	DS+	+	+	+	+/-	-	-	-	1:1	90% overfit
alpha+	Petri nets	DS+	+	+	+	+	-	-	-	1:1	90% overfit
tsinghua-alpha	Petri nets	CD+	+	+	+	+	-	-	-	1:2	90% overfit
alpha++	Petri nets	DS+	+	+	+	+	+	-	-	1:1	90% overfit
alpha#	Petri nets	DS+	+	+	+	+	-	+	-	1:0 1	90% overfit
alpha*	Petri nets	DS+	+	+	+	+/-	-	-	+/-	1..n:1	90% overfit
alpha-FL	Petri nets	DS+				+					90% overfit
Heuristic based algorithms											
Heuristic miner	Petri nets	ES+	+	+	+	+	+/-	+	-	1:0 1	30% underfit
Search based algorithms											
Genetic	Petri nets	TS+	+	+	+	+	+	+	-	1:0 1	tend to overfit
Duplicates GA	Petri nets	TS+	+	+	+	+	+	+	+	1..n:0 ...1	tend to overfit
Language based algorithms											
LangReg Basis	Petri nets	GC+	+	+	+	+	+	-	-	1:1	100% overfit
LangReg Sep	Petri nets	GC+	+	+	+	+	+	-	+	1:1	99% overfit
LangReg LP	Petri nets	none	+	+	+	+	+	-	-	1:1	tend to underfit
State discovery algorithms											
State discovery	Petri nets	none	+	+	+	+	+	+	+	1..*:0 ...*	flexible
Other											
Integer linear programming (ILP)	Petri nets or transition systems										overfit minimally
Inductive miner	Petri nets or directly-follows graphs										low overfit
Inductive miner incomplete (IMin)	Petri nets										less pronbe to overfit
fuzzy miner	Fuzzy models										moderate overfitting
multi-phase miner	Petri nets or workflow graphs										low overfit
Artificially Generated Negative Events (AGNEs) Miner	Petri nets										low overfit
DVS	Petri nets or state-based models										low overfit
Maximal Pattern Mining (MPM)	Pattern-based sequence models or directly-follows graphs										tend to overfit
Beta-algorithm	Petri nets or Causal nets										tend to overfit

2.2 DISCOVER PROCESS MINING MODELS USING DIFFERENT ALGORITHMS

As mentioned earlier, both the literature and practical applications have led to the widespread adoption of nearly 100 process discovery algorithms. Among the prominent algorithms are the α algorithm, GA algorithm, Heuristic miner, and Region miner. The α-algorithm, a key technique in process discovery, is designed to infer causality from event sequences. Variants such as α+, α++, and α# are extensions within the α-algorithm family. The GA algorithm, part of the genetic process mining family, focuses on achieving a high fitness model. The Heuristic miner builds on the α-algorithm by integrating trace frequencies into log analysis (Gupta, 2014). Gupta notes that the Heuristic miner effectively addresses "noise" by leveraging frequency and parameterization, while the GA algorithm can be resource-intensive, particularly with complex or noisy logs. She suggests novel techniques involving clustering and abstraction to handle noisy or complex models, recommending that clustering be done at the trace or activity level (Gupta, 2014).

When selecting a process discovery algorithm, it's crucial to consider the inherent properties of different algorithms. Some are better suited to handling models with invisible tasks, while others may struggle. Wen stresses the importance of choosing an algorithm that produces models semantically similar to the original, with structural equivalence or improvement (Wen, 2009). Recent studies and software developments have worked towards creating evaluation frameworks tailored to specific

organizational databases (Rozinat et al., 2007; Wang et al., 2012). However, evaluating all available algorithms against a business's models is computationally costly and time-consuming. Wen proposes an alternative: selecting models for evaluation and estimating their similarities through a regression model, without needing prior empirical evaluation (Wen, 2014). This approach contrasts with methods that evaluate model similarity using existing models. Wen's research compared process models generated by major algorithms (α algorithm, GA algorithm, Heuristic miner, and Region miner), using model similarity to assess algorithm quality. His findings support the idea that only a subset of process models needs empirical evaluation, as most can be recommended based on a regression model (Wen, 2009).

2.3 REPAIRING NOISY EVENT LOGS, INCOMPLETE LOGS AND LOOPS IN PROCESS MODELS

In theory, event logs should capture all activities of a process, whether performed by a human or a machine, and be stored in a standard database. However, in practice, this is often not the case. Participants may only manually track a portion of their work, sometimes leaving activities unrecorded, leading to incomplete or "noisy" event logs.

- Noise refers to the presence of rare or infrequent behaviors in the data, which do not reflect the typical process behavior, and it can result from errors or inconsistencies in event recording (Suriadi et al., 2017; Breuker et al., 2016).
- Incomplete logs may lack sufficient data to uncover certain control-flow structures (Suriadi et al., 2017).
- Process loops—where an event repeats or returns to a previous activity—can complicate process discovery.

These issues present significant challenges for process mining methods, leading to suboptimal results if not properly addressed. Ongoing research in the field focuses on improving event log quality by tackling these problems. Approaches include designing algorithms that are resilient to noise, as well as strategies for dealing with missing or incomplete data. Techniques such as Heuristics Miner and AGNEs miner address noise by analyzing behavior patterns and ensuring that only frequently occurring patterns are used in the process model (Goel et al., 2021). Additionally, probabilistic techniques, such as hidden Markov models, have been found effective in dealing with noisy or incomplete data, although they must be carefully managed to avoid overfitting (Breuker et al., 2016)

2.3.1 Process discovery from noisy event logs

In the field of process mining, noise is a significant challenge for event log analysis, particularly because its definition and quantification are not universally agreed upon. Noise in event logs generally refers to rare or infrequent behaviors that deviate from the typical process flow, often leading to suboptimal mining outcomes. Several methods have been proposed to address this issue:

➔ **Cheng and Kumar's Method**: This technique focuses on cleansing noisy logs by applying data mining strategies. They construct a classifier on a subset of the event log and then use this

classifier to filter out the noisy traces, thus enhancing the quality of the logs (Cheng & Kumar, 2023).

- ➔ **Nolle et al.'s Neural Network Approach**: This approach uses neural networks to detect and eliminate anomalies in event logs. It is based on identifying behavior that does not fit the expected patterns, effectively uncovering the true underlying process model (Nolle et al., 2022).
- ➔ **Weber et al.'s Probability Distribution Approach**: Weber and colleagues apply probability distributions over event traces to represent the underlying process, sampling these distributions to create a "true" process model while isolating instances of noise. Their approach, demonstrated with the Heuristics Miner algorithm, shows robustness against various types of noise (Weber et al., 2021).
- ➔ **Li et al.'s Distance Calculation Method**: This method involves calculating distances between event traces to identify outliers. By constructing a Petri net representation of the process model, Li and colleagues use clustering to group traces, treating those outside the largest cluster as noise (Li et al., 2020).
- ➔ **Rembert and Omokpo's Bayesian Approach**: They integrate prior knowledge from domain experts through Bayesian statistics to guide the process discovery using the α-algorithm, which helps handle uncertainty and noise (Rembert & Omokpo, 2021).
- ➔ **Folino et al.'s WFMiner Algorithm**: An enhanced version of WFMiner is proposed to address noise, duplicate tasks, and non-free choice in process discovery. This modification improves the algorithm's ability to manage noisy data while maintaining process model accuracy (Folino et al., 2020).

These methods contribute to reducing the impact of noise in process mining, although the lack of a standardized definition or metric for noise continues to complicate the reliable extraction of process insights from event logs.

2.3.2 Process discovery from incomplete even logs

In their work, Zareh Farkhady et al. (2013) define incomplete logs as event logs that lack sufficient information to accurately derive the underlying process model. To address this issue, they introduce a probabilistic approach using time Petri nets for process model discovery from incomplete logs. This method aims to recover process models despite the missing or incomplete data (Farkhady et al., 2013).

Leemans et al. (2013) highlight the importance of handling probabilistic behavioral relations to discover models from incomplete event logs. Their proposed solution, the Inductive Miner - Incompleteness (IMin) algorithm, is designed to partition the event log by selecting an operator that maximizes the probability of relations between activities that cross the partition. This method enhances the discovery of process models even when the logs are incomplete (Leemans et al., 2013).

Van der Werf et al. (2010) present a different approach by utilizing Integer Linear Programming (ILP) combined with the theory of regions. This method handles incomplete event logs by constructing process models through mathematical optimization techniques, ensuring the recovery of valid models from incomplete data (Van der Werf et al., 2010).

These approaches represent significant advancements in the field of process discovery, focusing on overcoming the challenges posed by incomplete event logs and improving the accuracy of the mined models.

2.3.3 Dealing with process loops

In the field of process mining, handling loops in process models remains a critical challenge. Medeiros et al. (2008) recommend the use of the α+ algorithm, an enhanced version of the original α-algorithm, specifically designed to address short loops in event logs. This algorithm improves upon its predecessor by offering more accurate discovery of process models that include small repetitive cycles (Medeiros et al., 2008).

To manage free loops, He et al. (2009) propose the α-FL algorithm. This approach is particularly adept at discovering free loop structures within a process model. The α-FL algorithm extracts repeated activities from incomplete event logs, analyzes the relationships among these activities, and applies specialized techniques to uncover loop structures, even when the data is sparse (He et al., 2009).

Weijters et al. (2007) suggest that the Heuristic Miner can enhance the robustness of the α-algorithm when dealing with short loops. The Heuristic Miner incorporates trace frequencies into the mining process, which helps in improving the model's accuracy, especially when short loops are present in the data (Weijters et al., 2007).

In a similar vein, Liesaputra et al. (2014) introduce Maximal Pattern Mining (MPM) as a technique for process model discovery. MPM constructs patterns based on entire event sequences, ensuring the soundness of the mined models, particularly in the presence of loops (Liesaputra et al., 2014).

These advancements contribute to the growing body of research focused on improving process mining techniques for more accurate and efficient modeling, even in the presence of loops or incomplete event logs.

2.3.4 Dealing with parallelism

To address parallelism in process models, Wen et al. (2020) introduce the betta-algorithm, a method designed to effectively handle activities occurring simultaneously with distinct start and end timestamps. This algorithm utilizes the temporal aspect of tasks, allowing for the explicit detection of parallel activities. By analyzing event logs that contain two types of events—START and COMPLETE—the betta-algorithm identifies overlaps in task occurrences, thus detecting parallelism more accurately (Wen et al., 2020).

On the other hand, Sahu et al. (2015) examine the performance of α-algorithms in dealing with task-based parallelism. Their study explores the use of the Message Passing Interface (MPI) model for parallel computing across multiple nodes, which enables better scalability and efficiency when applying the α-algorithm to process mining tasks involving parallel activities (Sahu et al., 2015).

These contributions highlight ongoing efforts to improve process mining algorithms' ability to handle parallelism and enhance their computational efficiency.

2.3.5 Approaches dealing with several problems in event logs simultaneously

Some authors have proposed methods to repair event logs without specifying the particular issues within the logs. Rogge-Solti et al. (2012) advocate for a stochastic approach that combines stochastic Petri nets, alignments, and Bayesian networks to model process behavior and repair event logs, addressing issues like missing timestamps and structural anomalies. This method decomposes the problem into two sub-problems: repairing the time and repairing the structure for each trace, utilizing observed data for efficient estimations (Rogge-Solti et al., 2012)

On the other hand, Zheng et al. (2012) introduce logical Petri nets for repairing models, suggesting an algorithm that identifies deviations and adds directed arcs to repair models using logic Petri nets and transitions based on logical functions (Zheng et al., 2012). Xu et al. (2013) also focus on repairing logical Petri net-based models, proposing a technique to add new activities by identifying logical concurrent and causal relations between new and original activities (Xu et al., 2013).

Liu et al. (2014) take a different approach by addressing the problem from the perspective of trace tracking. They propose a method for dealing with incomplete logs by clustering event traces, assigning missing traces to similar clusters, and supplementing the missing data to mine sub-process models (Liu et al., 2014).

Goedertier et al. (2007) present the AGNEsMiner method, which addresses expressiveness, noise, incomplete event logs, and the inclusion of prior knowledge by framing process discovery as a multi-relational classification problem, supplemented with Artificially Generated Negative Events (AGNEs) (Goedertier et al., 2007). Lastly, De Medeiros et al. (2006) propose the DT Genetic Miner to tackle typical challenges like noise, duplicate tasks, hidden tasks, and loops, although it is computationally intensive due to the genetic algorithm's requirements (De Medeiros et al., 2006).

These various techniques highlight the growing focus on improving event log quality in process mining, each contributing a unique approach to repairing and refining logs for more accurate process model discovery.

2.4 DOMAIN-KNOWLEDGE USE IN PROCESS MINING

Domain knowledge plays an important role in enhancing the accuracy and relevance of process discovery, particularly in context-specific processes. By integrating expert knowledge, organizations can guide the discovery algorithms to focus on the most relevant activities, ensuring that the resulting models are not only accurate but also tailored to the unique business needs and constraints of the environment. Domain knowledge integration helps address challenges such as variability in process execution, unexpected exceptions, or complex dependencies that cannot be fully captured by automated mining techniques alone (van der Aalst et al., 2011; Werr et al., 2013). For instance, domain experts can provide insight into key process steps or potential bottlenecks, improving the interpretability and usability of the discovered models (Dumas et al., 2018).

Moreover, the field of process discovery is witnessing the rise of hybrid approaches that combine the strengths of different techniques to tackle limitations inherent in individual methods. For example, hybrid models that merge classical process mining algorithms with machine learning or artificial

intelligence (AI) methods are emerging. These innovations allow process discovery systems to become more adaptive and dynamic, making them capable of handling noisy, incomplete, or evolving data (Bose et al., 2020; van der Aalst et al., 2020). AI-driven enhancements, particularly those leveraging deep learning and reinforcement learning, show promise in automating the refinement of process models based on continuous feedback and real-time data, which can lead to more efficient and accurate process discovery outcomes (Cappelli et al., 2020).

As these innovations continue to evolve, the future of process discovery is likely to be shaped by increased automation, context awareness, and adaptability, providing businesses with powerful tools for process optimization and continuous improvement.

2.5 SYSTEMATIC LITERATURE REVIEW

In response to the gaps identified in existing literature and practice, this thesis employs a systematic literature review (SLR) approach to explore the challenges and advancements in process model discovery. The methodology involves defining a research question, translating it into Scopus's advanced search language, and retrieving relevant articles. The process includes manual refinement of search results to focus on a core set of approximately ten articles. Additionally, the use of AI tools like ChatGPT's Web Access feature allows for the validation of traditional SLR processes, enhancing the comprehensiveness of the literature review.

Table 2 - SLR research questions

N	Research question
RQ1	What are the process discovery algorithms?
RQ2	Which part of the process discovery can be automated and semi-automated?
RQ3	What methods are used for repairing the noisy event logs?

The research questions served as a guiding mechanism to selectively filter the relevant articles, which will subsequently be used to build the recommendation framework for this thesis. This framework is grounded in the principles outlined by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

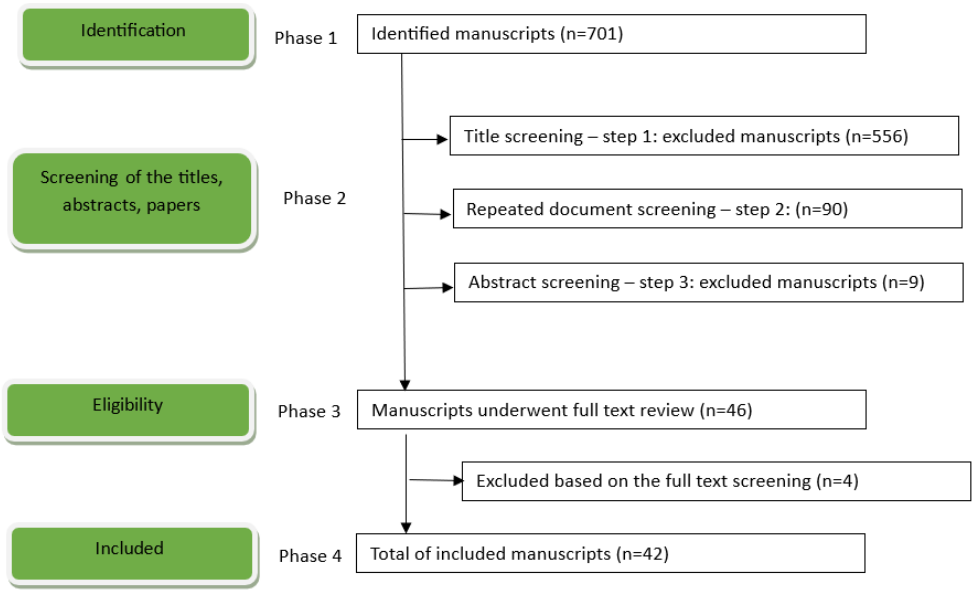


Figure 4 - PRISMA Flowchart (David et al., 2015)

In Phase 1, the research string was applied across all relevant electronic repositories, aiming to identify papers published between 2004 and 2024, resulting in a total of 701 publications.

Phase 2 followed a structured 3-step approach:

1. In Step 1, 556 papers were excluded based on title screening, leaving 145 papers for further assessment.
2. In Step 2, duplicate documents (n=90) were removed.
3. In Step 3, an abstract screening led to the exclusion of 9 manuscripts, narrowing the pool to 46 papers for deeper analysis.

In Phase 3, a thorough full-text reading and evaluation resulted in 4 additional exclusions. As a result, 42 articles were retained for integration into the study.

3. METHODOLOGY

3.1 RESEARCH METHODS

Research steps based on the challenges and goals of automatic process model discovery are the following:

1. **Literature Review**: Conduct a systematic literature review to identify existing process discovery algorithms, focusing on noise handling, data variability, and real-world application.
2. **Algorithm Comparison**: Map current algorithms, highlighting their strengths and limitations in handling incomplete or noisy event logs.
3. **Database Quality Evaluation**: Assess the impact of low-quality databases on process discovery, identifying common issues like noise, hidden tasks, and repeated activities.
4. **Recommendation Framework Development (Artifact)**: Propose a structured framework for automating process model discovery, integrating key steps like database cleaning, algorithm selection, and process architecture analysis.
5. **Refinement and Validation**: Iteratively refine the framework based on feedback from field specialists.

3.2 DESIGN SCIENCE RESEARCH

Design science research (DSR) is a methodology that focuses on the creation of artifacts to address real-world challenges, particularly in disciplines such as engineering, computer science, and architecture. Unlike analytical sciences, which aim to explain existing phenomena, DSR's primary goal is to develop and assess innovative solutions, typically artifacts or models, and test them in diverse settings to see how they perform under different conditions. This problem-solving approach contrasts with analytical sciences, such as physics and biology, which seek to understand natural laws and existing structures (Gregory, 2018; Simon, 1996).

In the context of Information Systems, research is guided by two main paradigms: behavioral science and design science. Behavioral science focuses on understanding and predicting human behavior within organizational contexts. In contrast, design science is applied when the objective is to create new artifacts that can solve specific, often complex, problems. The aim here is not just to predict or explain, but to innovate and improve practical systems (Hevner et al., 2004) (Figure 3.1).

These approaches reflect different ways of advancing knowledge and practice. Behavioral science theorizes about human actions, while design science builds tangible solutions to problems, grounded in real-world applicability. Therefore, design science is considered essential in creating tools, systems, or processes that are not just theoretical but capable of implementation and improvement in organizational contexts (Peffers et al., 2007).

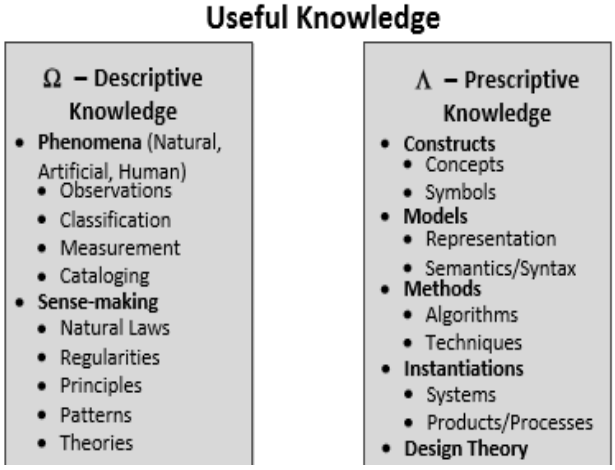


Figure 5 - Design Science Research Knowledge Base (Hevner et al., 2004)

Design thinking integrates two primary dimensions of design: as an artifact and as a process. As an artifact, design includes constructs, models, methods (processes), and their tangible instantiations, which represent the outcomes or innovations. Conversely, design as a verb focuses on the actual creation, evaluation, and refinement of these artifacts (Brown, 2009; Cross, 2011). In the context of Information Systems (IS) research, this dual focus emphasizes the importance of integrating relevance and rigor. Relevance is shaped by the environment, which encompasses people, organizations, and technology, ensuring that the design aligns with business needs. Rigor, on the other hand, is grounded in a solid knowledge base, including theoretical foundations and methodologies that guide the generation of applicable knowledge (Hevner et al., 2004; Peffers et al., 2007).

In Figure 3.2 the justification for these developed artifacts often arises from empirical methods such as case studies, simulations, and experiments, which test and validate the artifacts in real-world scenarios (March & Smith, 1995; Walls et al., 2004). These methods serve as essential tools for demonstrating the practical utility and scientific validity of the solutions designed within the field of Information Systems.

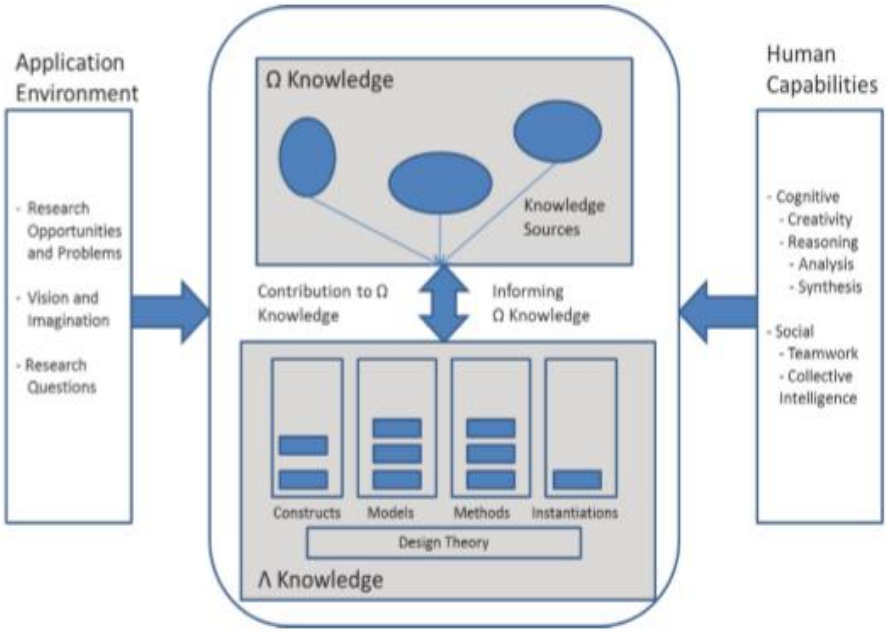


Figure 6 - The Roles of Knowledge in Design Science Research (Gregor & Hevner, 2013)

High-quality design science research is grounded in several key guidelines that ensure both its rigor and relevance to the field. These guidelines emphasize the importance of producing innovative artifacts, solving significant business problems, and contributing to existing knowledge.

1. **Design as an Artifact**: The research must create a novel artifact that advances the field and provides a tangible solution (Hevner et al., 2004; Peffers et al., 2007).
2. **Problem Relevance**: A central objective is to develop a technology-based solution that addresses a relevant and impactful business problem, ensuring real-world applicability (March & Smith, 1995).
3. **Design Evaluation**: The artifact must undergo thorough evaluation, with evidence of its novelty and positive impact on the intended business problem (Gregor & Hevner, 2013).
4. **Research Contributions**: The study should clearly contribute to both the existing body of knowledge and the field's development, demonstrating the artifact's significance (Walls et al., 2004).
5. **Research Rigor**: The research must be grounded in established theories, ensuring that experimental designs and methodologies are robust and scientifically sound (Hevner et al., 2004; Gregor & Hevner, 2013).
6. **Design as a Search Process**: The process of designing the artifact should involve a systematic search for solutions, utilizing available resources to meet the stakeholders' needs and achieve the desired outcomes (Peffers et al., 2007).
7. **Communication of Research**: Finally, the research and its findings must be clearly communicated to stakeholders, ensuring the value of the artifact is effectively conveyed and understood (March & Smith, 1995).

These guidelines help shape the structure of design science research, ensuring that the developed artifacts are not only theoretically grounded but also practically impactful.

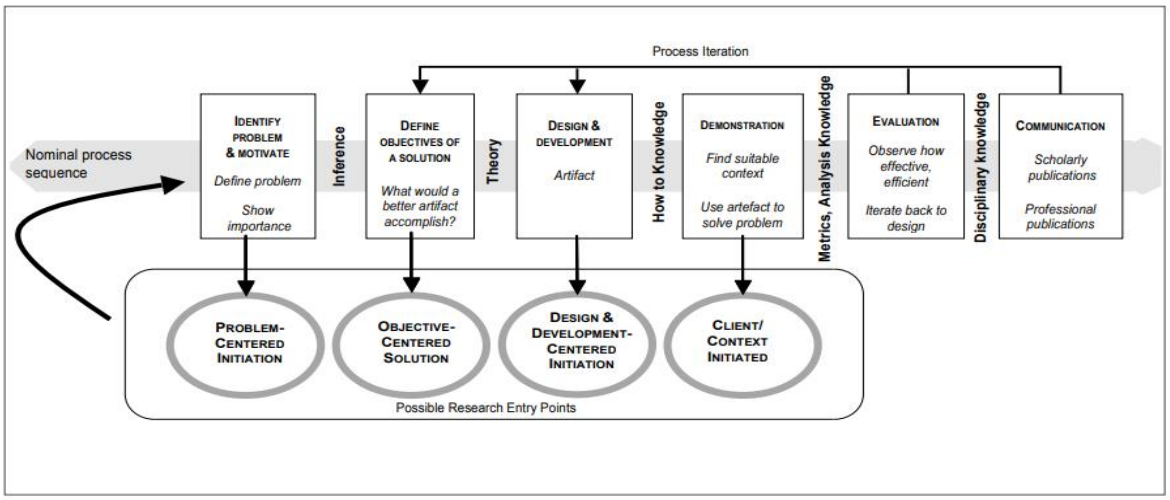


Figure 7 - Design Science Research Methodology (DSRM) Process Model (Peffers et al., 2014)

Assessing the impact of design within the field of Information Systems requires a robust framework that considers both the novelty and utility of the artifact. Gregor and Hevner (2013) emphasize several key areas where design science research (DSR) can provide significant contributions:

- **Artifacts:** The focus should extend beyond isolated artifacts created during individual phases to encompass the holistic process outcomes.
- **Evaluation:** Researchers need to clarify the methods used for assessing artifacts, ensuring these evaluations substantiate both their quality and practical utility.
- **Rigor:** DSR must prioritize rigor in both the construction and evaluation of artifacts, which directly influences the quality and reliability of innovations.
- **Search Processes:** The methodologies used to explore complex solution spaces must be critically examined for their alignment with solution design objectives.
- **Contribution and Value:** Clear distinctions should be made between the artifact's contribution to the knowledge base and its practical value to businesses, both of which are essential for impactful innovation.

These considerations highlight the necessity of structured frameworks and evaluation techniques to elevate the quality and relevance of research outputs in the design science domain.

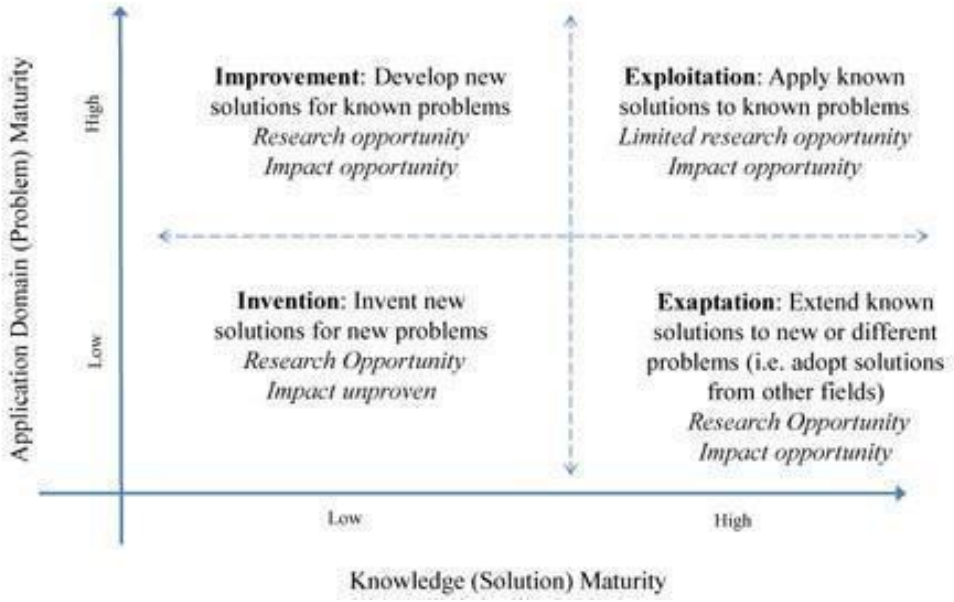


Figure 8 - DSR Knowledge Innovation Matrix (KIM) (Gregor & Hevner, 2013)

3.3 RESEARCH STRATEGY

The research strategy is structured into a systematic framework, encompassing problem identification, literature review, objective formulation, artifact development, rigorous evaluation, process implementation, metrics development, and critical reflection. This ensures a holistic approach to addressing challenges in process mining and design science research.

1. **Identification of Problematic Areas:** The initial step involves analyzing the landscape of process mining to identify deficiencies, particularly in process discovery with noisy event logs. This aligns with the diagnostic phase in design science research, as described by Peffers et al. (2007).
2. **Literature Review:** A thorough review of literature in Information Systems and Design Science Research (DSR) is essential to understanding existing theories, models, and methodologies, as emphasized by March and Smith (1995). This ensures that the research builds upon a robust knowledge foundation.
3. **Formulation of Research Objectives:** Clearly defined research objectives are necessary for high-quality DSR, as they provide direction and ensure alignment with the principles of creating innovative and valuable artifacts (Gregor & Hevner, 2013).
4. **Artifact Development:** The development phase includes designing a foundational process model schema and supplementary models tailored to process discovery in noisy datasets. This reflects the emphasis on innovative artifact creation highlighted by Hevner et al. (2004).
5. **Research Rigor and Contributions:** Ensuring rigor involves grounding the research in established theories and using sound experimental design. Contributions to the field should be clearly articulated, whether in theoretical advancements or practical applications (Gregor & Hevner, 2013).

6. **Design Process Implementation:** The implementation phase focuses on integrating constructs, models, methods, and instantiations into the artifact development process, following best practices in design process methodologies (Peffers et al., 2007).
7. **Evaluation Metrics:** A structured evaluation matrix is developed to assess the artifact's impact, considering aspects such as novelty, utility, and relevance to both theoretical and practical domains (Gregor & Hevner, 2013).
8. **Critical Reflection and Refinement:** The final stage involves reflecting on the entire process, identifying areas for improvement, and refining the artifact to address limitations and enhance contributions. This iterative approach is central to achieving high-quality design outcomes (March & Smith, 1995).

3.4 EXPERT INTERVIEW

Qualitative research methods, particularly expert interviews, are often used in design research to validate artifacts and gain insights into the practical implications of a design. Expert interviews provide a valuable means of evaluating the functionality, usability, and effectiveness of a design artifact from those with specialized knowledge and experience in the field (Patton, 2002). This method involves structured or semi-structured interviews with professionals who possess deep understanding of the subject matter, allowing researchers to capture expert perspectives that are critical for the refinement and validation of design artifacts.

Expert Interviews as a Validation Tool: Expert interviews serve as a means to gather in-depth insights from professionals who have significant expertise or experience related to the design artifact. These interviews are essential for validating the design's relevance, addressing potential gaps, and understanding the real-world challenges the artifact may face (Mason, 2002). For example, in process mining, experts can evaluate how well a proposed algorithm addresses the challenges of event log repair, noise handling, and dealing with incomplete or corrupted data, thus providing feedback that ensures the design aligns with industry needs (Seidman, 2013).

Advantages of Expert Interviews: One of the main advantages of expert interviews is their ability to provide deep, context-rich data that goes beyond the surface-level features of the artifact. Experts can offer feedback on how the design artifact might be applied in practice, identify potential shortcomings, and propose improvements based on real-world experience (Eisenhardt, 1989). This process helps ensure that the design is both theoretically sound and practically viable. Additionally, expert interviews allow for flexibility in probing deeper into specific areas of concern, which can lead to more nuanced findings (Bryman, 2016).

Pros:

1. **Depth of Insight**: Expert interviews provide deep, nuanced insights into specific aspects of a design artifact, allowing researchers to explore complex issues from those with specialized knowledge (Patton, 2002). This is especially valuable when validating the relevance, functionality, and practical implications of a design in real-world contexts.

2. **Contextual Understanding**: Experts bring a wealth of contextual knowledge, which can enhance the interpretation of findings. For example, when validating a process discovery algorithm, expert interviews can help explain how particular features or challenges might manifest in practice (Mason, 2002). This can be particularly important in fields like process mining, where the complexities of real data (e.g., noise, loops, incomplete logs) require domain expertise for proper understanding.
3. **Flexibility and Adaptability**: Qualitative methods, including expert interviews, allow for a flexible approach where questions can be adjusted during the interview to probe deeper based on expert responses (Bryman, 2016). This flexibility makes expert interviews useful for exploring new or emerging issues that were not initially anticipated in the research design.
4. **Rich, Detailed Data**: Expert interviews can provide detailed descriptions, anecdotes, and examples that enrich the research findings. This can be crucial for validating a design artifact, as experts often offer specific recommendations or insights that can refine the artifact's features or design (Eisenhardt, 1989).

Challenges and Limitations: Despite their advantages, expert interviews have certain limitations. The subjectivity of the expert's perspective can introduce bias into the validation process, especially if the expert has a vested interest in a particular outcome (Seidman, 2013). Furthermore, the small sample size typical of expert interviews limits the generalizability of findings. Since expert input is often specific to particular industries or use cases, the feedback may not always be applicable to other contexts (Marshall & Rossman, 2016). Another challenge is the time-consuming nature of arranging and conducting these interviews, which can extend the research timeline (Bryman, 2016).

Cons:

1. **Subjectivity and Bias**: Expert interviews can be highly subjective, depending on the individual expert's experience and perspective. This introduces the potential for bias, which may skew validation results (Seidman, 2013). For instance, an expert from a particular industry might emphasize certain aspects of the design that are not relevant in other sectors, which limits the generalizability of the findings.
2. **Limited Generalizability**: While expert interviews provide in-depth insights, the findings may not always be generalizable to a broader population. The sample size is typically small, and experts may represent only a specific viewpoint or subset of the industry (Marshall & Rossman, 2016). This limitation can affect the overall validity of the design artifact if the expert sample is not diverse enough.
3. **Time-Consuming**: Conducting expert interviews can be time-consuming, both in terms of preparation and analysis. Interviews often require careful selection of participants, developing a rapport, and interpreting qualitative data, all of which can significantly extend the research timeline (Bryman, 2016).
4. **Reliance on Expert Availability**: Expert interviews are contingent on the availability and willingness of the experts, which can sometimes be a limitation, especially if the research requires input from multiple experts across various fields (Seidman, 2013). Additionally, experts may be hesitant to critique a design artifact if they are not directly involved in its development, potentially leading to less critical feedback.

Integration with Other Research Methods: To enhance the validity and reliability of the findings, expert interviews are often combined with other qualitative or quantitative research methods, such as case studies, surveys, or user testing (Eisenhardt, 1989). This triangulation approach strengthens the conclusions drawn from the interviews, providing a more comprehensive validation process for the design artifact.

4. EMPIRICAL STUDY

4.1 DESCRIPTION OF THE PROCESS MODEL DISCOVERY

Process model discovery is a vital component of process mining that focuses on analyzing event logs through algorithms to deduce how processes are executed in real time. A key advantage of this technique lies in its ability to uncover actual process flows without relying on predefined assumptions, making it an indispensable tool for enhancing organizational workflows (van der Aalst, 2016).

Effectively addressing the challenges associated with process model discovery is critical, as the accuracy and utility of subsequent analysis depend on the quality of the discovered model. Common challenges include:

1. **Unstructured or "Spaghetti-Like" Models**
   Complex and loosely structured processes often result in unmanageable "spaghetti-like" models, hindering clarity and usability. Techniques such as model simplification and hierarchical structuring have been proposed to address this issue (Rozinat & van der Aalst, 2008; Mans et al., 2008).
2. **Concurrency Detection**
   Identifying concurrent tasks, or activities executed simultaneously, is essential for creating accurate process models. Algorithms like the Alpha Miner and heuristic mining methods have been refined to manage concurrency effectively (Weijters & Ribeiro, 2011; van der Aalst, 2016).
3. **Noise and Outlier Filtering**
   Noise and outliers in event logs can distort process discovery results. Preprocessing strategies, including filtering and the application of robust algorithms, are vital for ensuring high-quality outputs (Buijs et al., 2014; Augusto et al., 2018).

Addressing these challenges ensures that process model discovery yields actionable and reliable representations of operational processes. Suggested steps for addressing these challenges and improving process discovery are illustrated in Figure 4.1. When these challenges are systematically managed, process mining contributes significantly to process analysis, optimization, and decision-making.

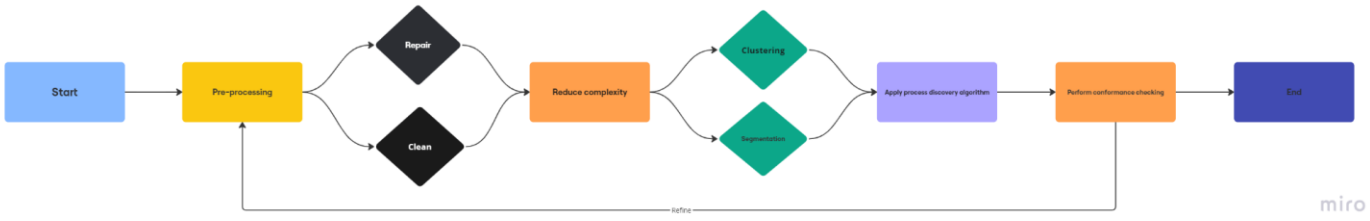


Figure 9 - Process model discovery in steps (inspired by Marin-Castro et al., 2021)

Preprocessing: The preprocessing phase is critical for preparing raw event log data for process mining. It encompasses tasks such as cleaning, repairing, and transforming event logs to ensure they align with the requirements of process mining algorithms. Key activities include removing noise, addressing

missing data, and standardizing formats. Techniques like event concatenation help group related events, while filtering eliminates irrelevant or incomplete data, thus enhancing data quality and algorithmic performance. Effective preprocessing ensures that event logs are structured and accurate, ultimately enabling reliable and efficient process discovery (Augusto et al., 2018; Buijs et al., 2014, Marin-Castro et al., 2021; Liu, 2023).

Reducing Complexity: This phase simplifies event log data to enhance the efficiency and precision of process discovery. Dimensionality reduction techniques, such as clustering and event log abstraction, consolidate similar activities and remove redundant or less relevant events. These methods minimize the number of cases or activities, reducing data complexity and making analyses more manageable. By streamlining the data, this step fosters targeted insights into business processes, ensuring the effectiveness of subsequent discovery processes (van der Aalst, 2016; Dongen et al., 2012, Marin-Castro et al., 2021).

Applying Process Discovery Algorithms: Following preprocessing and simplification, process discovery algorithms are employed to derive actionable process models. Algorithms like Alpha Miner, Inductive Miner, and Heuristic Miner analyze event logs to identify patterns, dependencies, and activity relationships. The choice of algorithm depends on factors such as data complexity and the desired abstraction level. While many algorithms can function autonomously, incorporating domain expertise or manual refinement often improves the accuracy and relevance of the models to reflect real-world processes (van der Aalst, 2011; Weijters & Ribeiro, 2011,van der Aalst, 2016; Marin-Castro et al., 2021).

Performing Conformance Checking: Conformance checking compares the discovered process models with actual event logs to identify discrepancies between observed and expected behavior. This step is essential for validating the model's accuracy and detecting deviations in real-world operations. By identifying and addressing inconsistencies, conformance checking refines the process model, making it a more accurate representation of operational processes. These insights are instrumental in driving process improvement and optimization (Carmona et al., 2018; Munoz-Gama & Carmona, 2010, Marin-Castro et al., 2021).

4.2 ASSUMPTIONS

Building on the insights derived from the literature review on process mining, process model discovery, and the integration of domain knowledge, several assumptions have been defined for the development of a recommendation framework for process model discovery. These assumptions are based on the findings of previous research and aim to address key challenges and optimize the process mining workflow:

1. **Diversity of Datasets**: Process mining works with both synthetic and real-life datasets, each presenting unique challenges. Synthetic datasets offer control over the complexity of the data, while real-life datasets introduce noise, variability, and human error (van der Aalst, 2016; Weijters et al., 2018).
2. **Event Log Quality**: The quality of event logs significantly impacts the accuracy of the discovered process models. Noisy, incomplete, or inconsistent event logs can lead to less accurate process models (Li et al., 2017; He et al., 2020).

3. **Need for Preprocessing**: Preprocessing is essential for transforming raw event logs into a suitable format for process mining. This phase includes noise reduction, missing data handling, and data normalization (Appice & Malerba, 2016; van der Aalst et al., 2016).
4. **Handling Noise**: Noise in real-life event logs is a critical challenge. Algorithms that account for noise and outliers are essential to generate reliable process models (Li et al., 2017; He et al., 2020).
5. **Complexity of Process Models**: Process models derived from real-life logs often exhibit complexity, such as "spaghetti-like" structures, due to the irregularities in data. A framework that simplifies these models is necessary for better interpretability and usability (Leemans et al., 2013; Weijters et al., 2018).
6. **Role of Clustering**: Trace clustering can be a key preprocessing step to reduce complexity by grouping similar event traces. This helps to handle variability and can improve the clarity of the discovered models (Appice & Malerba, 2016).
7. **Data Incompleteness**: Missing data is a common issue in event logs. Techniques like trace clustering and the assignment of missing traces to the most similar clusters are essential to minimize the impact of incomplete data on the model (Xu & Liu, 2018).
8. **Impact of Process Discovery Algorithms**: Different process mining algorithms, such as Inductive Miner, Heuristic Miner, and the alpha-algorithms, are suitable for different types of data. A recommendation framework should consider the specific characteristics of the data when selecting an algorithm (Leemans et al., 2013; Weijters et al., 2018).
9. **Data Format Standardization**: The use of standardized formats like XES for event logs is critical for ensuring compatibility across process mining tools. A recommendation framework must consider the format of input data to guide the selection of tools and algorithms (van der Aalst et al., 2016).
10. **Scalability of Algorithms**: As datasets grow in size and complexity, the scalability of the process mining algorithms becomes a significant concern. The recommendation framework should incorporate algorithms that can scale effectively with large datasets (van der Aalst, 2016).
11. **Integration of Domain Knowledge**: A process mining framework can be enhanced by incorporating domain knowledge to refine process models. Domain experts can guide the selection of relevant events or activities, making the process model more meaningful and contextually accurate (van der Aalst et al., 2016).
12. **Iterative Model Refinement**: Process models need to be iteratively refined, especially when discrepancies are found between the discovered models and the real process execution. A recommendation system should provide a way to identify discrepancies and suggest corrective actions (He et al., 2020).
13. **Real-time Analysis**: In dynamic business environments, real-time process discovery and analysis can provide immediate insights. A recommendation framework should support real-time event log processing and model adaptation (Weijters et al., 2018).
14. **Interpretability of Models**: Process models should not only be accurate but also interpretable by business users. A recommendation system should focus on algorithms and techniques that produce models that are understandable and actionable (Leemans et al., 2013).
15. **Continuous Learning**: A process mining framework should have the capacity for continuous learning, adapting as more event logs are processed over time. The recommendation system should be able to update its suggestions based on feedback and new data (Nguyen et al., 2019).

4.3 RECOMMENDATION FRAMEWORK FOR CHOOSING PROCESS DISCOVERY ALGORITHM

The existing recommendation framework for process discovery addresses common challenges in process discovery, guiding users in selecting the most suitable algorithm for extracting process models from event logs. It considers factors such as event log quality, structural complexity, concurrency, and specific analysis objectives. By evaluating each algorithm's strengths and limitations, the framework suggests the best match based on data characteristics and process requirements. This systematic approach enables organizations to efficiently pinpoint the ideal discovery technique, which is particularly useful in data-heavy settings where manual selection would be time-consuming and potentially less precise.

The provided diagram illustrates a **Recommendation Framework for Process Discovery**, emphasizing structured steps for refining and generating process models. Each stage integrates both automated techniques and domain knowledge, ensuring effective handling of challenges commonly found in process mining. The framework provides a comprehensive roadmap for addressing challenges in process discovery through systematic steps. By blending automated discovery methods with domain knowledge, it ensures accurate and adaptable process models that meet real-world organizational needs.

1. Data Collection and Importing: The process begins with the collection of event log data from organizational systems. This data is then imported for preprocessing, a crucial step in ensuring data suitability for further analysis. Preprocessing includes handling noise, identifying inconsistencies, and validating event structures (Al-Absi et al., 2023). This stage ensures that raw data is ready for accurate process discovery.

2. Event Log Preprocessing: Event log preprocessing focuses on noise filtering and log sanitization, using techniques such as task mapping, timestamp refinement, and redundancy reduction. These steps aim to resolve data quality issues that could hinder process model accuracy. This phase is vital for addressing challenges like hidden tasks, duplicate entries, and silent activities (Al-Absi et al., 2023).

3. Automated Process Discovery (APD): Next, APD algorithms are employed to generate initial process models. Tools such as the Alpha family algorithms, Heuristic Miner, Fuzzy Miner, Inductive Miner, Genetic Miner and Split Miner are utilized to extract processes. These methods address challenges including concurrency, event noise, and incomplete traces. The choice of algorithm depends on the characteristics of the event log and the desired outcome (Dymora et al., 2019).

4. Interactive Process Discovery (IPD): In this phase, user feedback and domain knowledge play a critical role. These inputs refine the process models by resolving coarse-grained timestamps, handling duplicate events, and identifying silent activities. The inclusion of domain expertise allows the iterative improvement of process models, bridging the gap between raw data and practical insights (Schuster et al., 2022).

5. Advanced Refinement: The final stage incorporates large language models (LLM) and reinforcement learning to address unstructured data and optimize the process model. These advanced techniques create a "Full APD" by integrating all components and refining the process iteratively. This ensures

robustness and accuracy, especially when working with complex, real-world data. The feedback loop remains critical throughout the framework, ensuring continuous improvement (Karn et al., 2021; Norouzifar et al., 2024).

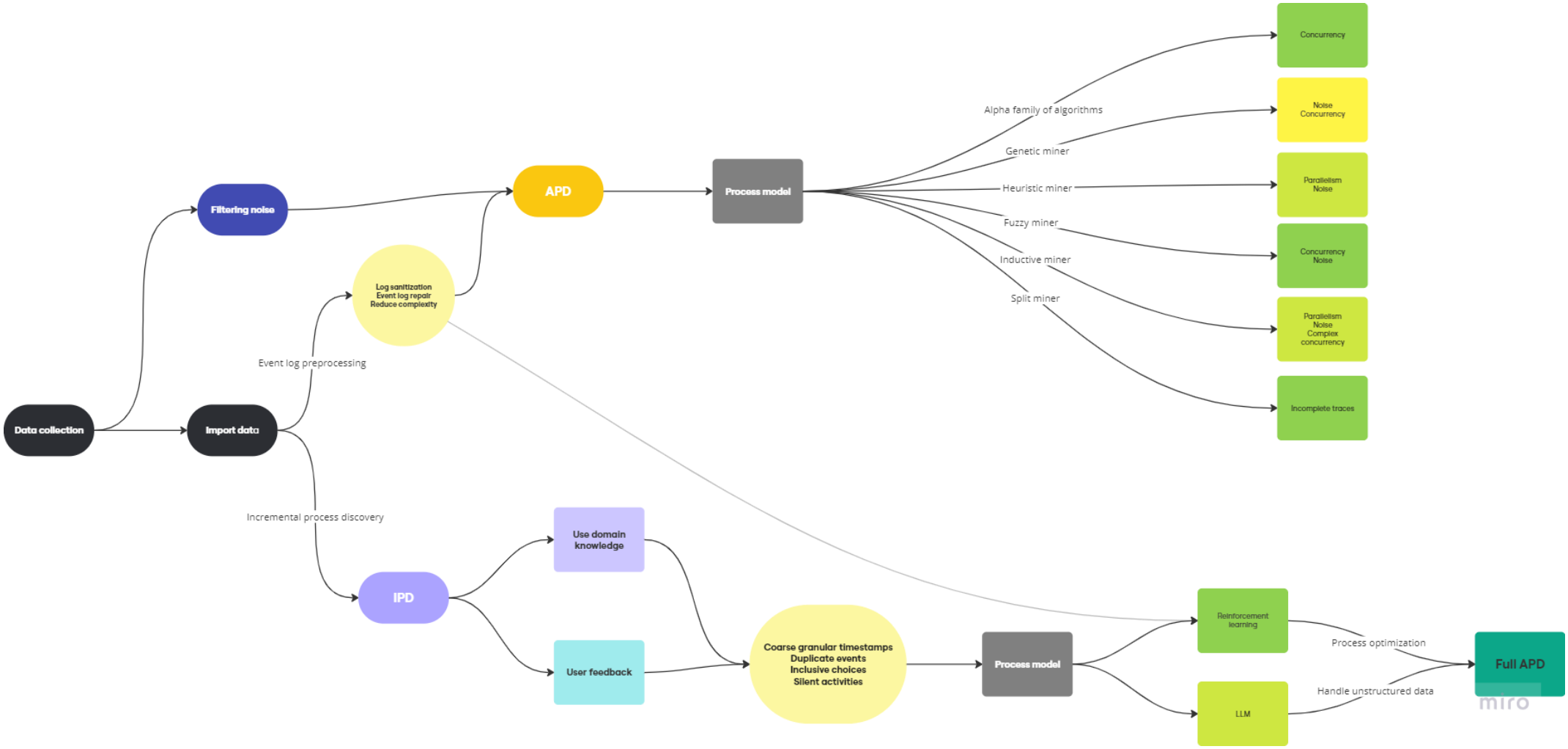


Figure 10 - Recommendation framework for process discovery

The Alpha algorithm (α-algorithm), one of the earliest and most foundational methods in process mining, is designed to identify process models by analyzing event logs. It uncovers causal relationships and event order, making it a core tool in process discovery. However, despite its historical significance, the Alpha algorithm has several limitations that can be addressed by more advanced methods (van der Aalst, 2011; Weijters et al., 2006).

1. **Concurrency Handling**: The Alpha algorithm can handle concurrency in simple cases but struggles with more complex, concurrent events, especially in cases involving multiple parallel activities (van der Aalst et al., 2003). More sophisticated algorithms, such as Heuristic Mining (Weijters et al., 2007) and Genetic Miner (Weber et al., 2012), use statistical and probabilistic approaches to detect and manage concurrency more effectively.

2. **Noise and Incomplete Logs**: The Alpha algorithm assumes that the event logs are perfect, which makes it ill-equipped to handle noise—unexpected or incomplete event data that often occurs in real-life processes. Modified versions, like the Fuzzy Miner (Günther & van der Aalst, 2007) and Interactive Process Discovery (Leemans et al., 2013), address this limitation by using techniques that filter out noise or adapt to noisy data, improving the robustness and accuracy of the resulting process models.

3. **Loops**: The Alpha algorithm has limited ability to model loops, particularly nested or complex loops of varying lengths (Weijters et al., 2006). Algorithms like Heuristic Miner and the improved Alpha++, which refines the Alpha algorithm, offer better loop detection by leveraging more advanced heuristics and statistical methods (Weijters et al., 2013).

4. **Scalability**: One of the key challenges with the Alpha algorithm is its poor scalability, particularly when dealing with large datasets. As process logs grow in size, the Alpha algorithm can become computationally expensive and less efficient. More scalable approaches, such as the Inductive Miner (Leemans et al., 2013) and Genetic Miner (Weber et al., 2012), have been developed to handle large logs with better computational efficiency, often producing more scalable and interpretable process models.

In summary, while the Alpha algorithm laid the groundwork for process model discovery, its limitations in handling concurrency, noise, loops, and scalability have led to the development of several advanced algorithms. These improved methods, including Heuristic Miner, Fuzzy Miner, Inductive Miner, and Genetic Miner, build upon the foundations of Alpha to address its weaknesses, enabling more accurate and robust process discovery (Günther & van der Aalst, 2007; Leemans et al., 2013; Weber et al., 2012).

4.4 RESULTS AND DISCUSSION

The proposed framework significantly advances the field of process discovery by addressing longstanding challenges. By incorporating advanced preprocessing techniques, the framework ensures that raw data is transformed into clean, structured event logs, which are essential for accurate and meaningful process discovery. The integration of automated algorithms for model generation introduces scalability, enabling the efficient handling of large and complex datasets. Furthermore, the use of incremental discovery ensures that domain expertise remains central to the process, mitigating the risk of producing models that lack practical relevance.

It is worth mentioning that advanced refinement methods, such as Large Language Models (LLMs) and reinforcement learning, enhance the framework's robustness, enabling it to navigate complex scenarios, including unstructured and heterogeneous data. While the ultimate goal of fully automated process discovery remains achievable, incorporating domain knowledge into algorithms poses a significant challenge due to the difficulty of standardizing and formalizing such expertise. This highlights the need for ongoing research into methodologies that seamlessly embed domain knowledge into automated discovery processes.

A persistent challenge in process discovery lies in the quality of event logs. The framework raises an important question: should efforts prioritize improving database quality, or should the focus shift to developing techniques capable of mining accurate models from suboptimal event logs? Despite the framework's robustness, its performance may be constrained when initial data collection is incomplete or poorly structured. This underscores the need for better practices in data management and collection at the organizational level.

Additionally, while the iterative nature of the framework is a strength, it necessitates sufficient domain expertise at various stages. Organizations with limited access to experienced analysts may face difficulties in leveraging the framework to its full potential. Addressing this gap through training programs or developing user-friendly tools could further enhance the framework's accessibility and applicability.

In conclusion, the proposed framework represents a significant step forward, striking a balance between automation and human expertise. It provides a flexible and effective solution for real-world process discovery, while its scalability and adaptability make it well-suited to diverse organizational contexts. However, continuous improvement in data quality and integration of domain knowledge remain key priorities for the advancement of fully automated and reliable process discovery.

4.5 MID-TERM VALIDATION

This interview, conducted midway through the thesis process, aimed to validate and refine the recommendation framework being developed. The process mining expert from Celonis (referred to as "the Expert") provided insightful feedback on the application and improvement of process discovery algorithms. Specifically, the Expert highlighted key challenges in event log repair and emphasized the importance of validating the proposed framework using real-world data, particularly in complex cases such as loops and incomplete logs. Furthermore, she recommended integrating qualitative insights, such as expert knowledge, to enhance the interpretation of event log data, offering a richer understanding of process mining. Her feedback combined both theoretical and practical perspectives, underscoring the need for further exploration and testing to ensure the framework’s relevance and applicability in industry contexts. These insights were incorporated into the development of the recommendation system, advancing the thesis toward a more robust and industry-aligned solution.

The following questions were discussed during the interview, with a summary of the responses provided afterward:

Q1: From a research perspective, how do you perceive the usefulness of the proposed recommendation system for process discovery algorithms? Do you think it addresses key challenges in event log repair and algorithm selection?

Q2: What recommendations or suggestions do you have for improving the effectiveness of the recommendation system, particularly in terms of validation on real-world data and handling complex scenarios such as loops and incomplete logs?

Q3: Are there any additional comments or potential concerns you have about the proposed analysis or methodology? For example, do you see any areas where further refinement or expert insights could enhance the results or applicability in industry?

Summary of Interview Responses

1. **Usefulness of the Recommendation System (Q1)**: From a research perspective, the Expert views the concept of the proposed recommendation system as highly valuable, particularly the focus on repairing event logs, which aligns with growing research interest. The introduction of probabilistic methods, such as stochastic event blocks, is gaining traction in both academic and industry settings. While the system presents a promising theoretical framework, the Expert emphasizes the importance of further validation with real-world data to assess its practical application. She notes that event log repair, particularly in the context of incomplete logs and loops, remains a key challenge, which the proposed recommendation system addresses effectively.
2. **Suggestions for Improving the System (Q2)**: The Expert suggests enhancing the system by validating it with actual data, particularly in scenarios involving loops and incomplete logs. She recommends conducting a comparative analysis of algorithm performance across various real-world datasets, including industry-specific cases such as purchase-to-pay or order-to-cash processes. By testing the algorithms on real data, the system can better assess how these challenges manifest in different sectors, ensuring its relevance and effectiveness. Additionally, she advocates for a broader exploration of industry-specific use cases to refine the algorithm selection recommendations further.
3. **Additional Comments and Potential Issues (Q3)**: The Expert commends the systematic and thorough nature of the analysis, particularly the integration of both qualitative and quantitative methods. She highlights the value of expert insights in interpreting event log data, suggesting that expert interviews could complement the quantitative analysis and provide deeper contextual understanding of the data. Incorporating qualitative insights into future stages of the study, especially when testing real-world scenarios, would enhance the practical applicability and robustness of the findings. Angela also emphasizes the importance of qualitative data in understanding the narrative behind the logs, which could lead to richer, more actionable results.

5. CONCLUSIONS AND FUTURE WORKS

5.1. SYNTHESIS OF DEVELOPED WORK

The proposed framework for process mining effectively integrates both automated techniques and domain knowledge, addressing the complexities typically encountered in real-world applications. It is structured into several distinct stages that collectively ensure the production of robust and adaptable process models. Starting with the data collection and preprocessing phase, the framework addresses common challenges such as noise, inconsistencies, and redundant data, ensuring that raw data is transformed into a suitable format for analysis. Automated Process Discovery (APD) techniques provide an initial process model by extracting meaningful patterns from the event logs while handling challenges like concurrency, missing data, and noise in traces. Building upon this, Iterative Process Discovery (IPD) further refines the model by incorporating domain knowledge, enhancing the model's relevance and providing actionable insights for the business context. Finally, the advanced refinement stage employs machine learning techniques to further enhance the model's adaptability and accuracy, ensuring it can adjust to dynamic changes in organizational processes. This multi-stage approach makes the framework highly adaptable, providing not just accurate models but also ones that evolve with changing business needs.

The framework's strength lies in its ability to merge automated discovery with expert-driven insights, ensuring the resulting process models are not only accurate but also tailored to the specific operational context of the organization. This methodology, while grounded in existing process mining techniques, pushes the boundaries by refining models iteratively and leveraging advanced technologies such as machine learning, ensuring they remain relevant over time.

5.2. LIMITATIONS

Despite its strengths, the proposed framework is not without limitations. One of the key challenges is its reliance on high-quality input data. While the framework is effective at preprocessing data and mitigating issues such as noise and inconsistencies, the success of the process discovery is still dependent on the quality of the initial event logs. In practice, data quality issues such as missing or incomplete traces could still undermine the accuracy of the discovered models, especially in more complex scenarios.

Another limitation is the computational cost associated with the more advanced stages of the framework, such as the machine learning-based refinement. These techniques can be resource-intensive, requiring significant computational power, which may be a barrier for organizations with limited technical infrastructure. Additionally, the iterative nature of IPD can be time-consuming, especially in larger or more complex processes, potentially delaying the time to insight for organizations seeking rapid process optimization.

Lastly, while the framework incorporates domain knowledge, its success heavily depends on the expertise available to provide meaningful input during the iterative discovery and refinement stages. This dependence on human input may limit the scalability of the framework across industries where

domain expertise may not be readily available or where processes are too variable to define clear, consistent models.

5.3. FUTURE WORK

Future research should focus on enhancing the framework's scalability and applicability across various industries. One promising avenue is the integration of emerging technologies such as advanced AI and natural language processing (NLP). These technologies could improve the interpretability of process models, making them more accessible to business users without deep technical knowledge. NLP could, for instance, help bridge the gap between event log data and human-readable process models, facilitating better communication between technical and non-technical stakeholders.

Additionally, developing automated mechanisms that can detect and adapt to changes in business processes will be crucial for improving the framework's robustness and real-time capabilities. The dynamic nature of business environments demands agile process mining solutions that can update process models as new data becomes available. This feature would be particularly beneficial in industries where processes are highly dynamic, and agility is critical.

Scalability studies are another important direction for future work. Testing the framework in diverse industrial settings will help validate its applicability across different domains and ensure that it can handle the unique challenges posed by each. This could include experimenting with larger datasets or more complex process types to evaluate the framework's performance and efficiency under varying conditions.

Finally, exploring the potential for real-time process discovery would be a valuable next step. This involves continuously updating the process model as event logs are generated, allowing businesses to monitor and optimize their operations in real-time. Given the growing demand for operational agility, aligning the framework with the real-time needs of organizations could significantly enhance its value.

REFERENCES

Adriansyah, A., & van Dongen, B. F. (2013). Measuring process model similarity using an alignment-based approach. *Business Process Management Journal, 19*(6), 888-918. https://upcommons.upc.edu/bitstream/handle/2117/26715/pre-print%20isem2013.pdf

Al-Absi, M. A., & R'bigui, H. (2023). Process Discovery Techniques Recommendation Framework. *Electronics, 12*(14), 3108. https://doi.org/10.3390/electronics12143108

Alves de Medeiros, A. K., van Dongen, B. F., van der Aalst, W. M. P., & Weijters, A. J. M. M. (2004). Process mining: Extending the alpha-algorithm to mine short loops. *BETA publicatie: Working Papers, Vol 113*. https://research.tue.nl/en/publications/process-mining-extending-the-alpha-algorithm-to-mine-short-loops

Appice, A., & Malerba, D. (2015). A co-training strategy for multiple view clustering in process mining. *IEEE Transactions on Services Computing, 9*(6), 832–845. https://doi.org/10.1109/TSC.2015.2430327

Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F. M., Marrella, A., Mecella, M., & Soo, A. (2018). Automated discovery of process models from event logs: Review and benchmark. *IEEE Transactions on Knowledge and Data Engineering, 31*(4), 686-705. https://doi.org/10.1109/TKDE.2018.2841877

Augusto, A., Carmona, J., & Verbeek, E. (2022). Advanced process discovery techniques. In W. M. P. van der Aalst & J. Carmona (Eds.), *Process mining handbook* (Vol. 448, pp. 85–113). Lecture Notes in Business Information Processing. Springer. https://doi.org/10.1007/978-3-031-08848-3_3

Augusto, A., Conforti, R., Dumas, M., La Rosa, M., & Maggi, F. M. (2019). Automated discovery of process models from event logs: Review and benchmark. *Journal of Data Science and Analytics, 49*(6), 270–329. https://doi.org/10.1007/s41060-019-00103-9

Badakhshan, P., Wurm, B., Grisold, T., Geyer-Klingeberg, J., Mendling, J., & Brocke, J. v. (2022). Creating business value with process mining. *The Journal of Strategic Information Systems, 31*, 2023. https://doi.org/10.1016/j.jsis.2022.101745

Bakır, Ç., Yuzkat, M., & Sebag, M. (2022). Process mining algorithms performance according to new Bayes conformance function. *IEEE Conference Proceedings*, 1-6. https://doi.org/10.1109/IEEECONF55059.2022.9810387

Benevento, E., Aloini, D., & van der Aalst, W. M. P. (2022). How can interactive process discovery address data quality issues in real business settings? Evidence from a case study in healthcare. *Business Process Management Journal, 28*(3), 705–722. https://doi.org/10.1108/BPMJ-02-2022-0411

Bose, R. P. J. C., & van der Aalst, W. M. P. (2009). Trace clustering based on conserved patterns: Towards achieving better process models. *Business Process Management Workshops*, 170-181. https://doi.org/10.1007/978-3-642-01042-0_19

Bottrighi, A., Canensi, L., Leonardi, G., Montani, S., & Terenziani, P. (2015). Interactive mining and retrieval from process traces. *Expert Systems with Applications, 110*, 62–79. https://doi.org/10.1016/j.eswa.2018.05.041

Brock, J., Brennig, K., Löhr, B., et al. (2024). Improving process mining maturity – From intentions to actions. *Business & Information Systems Engineering, 66*(5), 585–605. https://doi.org/10.1007/s12599-024-00882-7

Brown, T. (2009). *Change by design: How design thinking creates new alternatives for business and society*. HarperBusiness. https://doi.org/10.23860/MGDR-2019-04-02-08

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press. https://doi.org/10.1093/he/9780199689453.001.0001

Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P. (2012). On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. In: Meersman, R., et al. On the Move to Meaningful Internet Systems: OTM 2012. OTM 2012. *Lecture Notes in Computer Science, vol 7565. Springer, Berlin, Heidelberg*. https://doi.org/10.1007/978-3-642-33606-5_19

Burattin, A., Cimitile, M., Maggi, F. M., & Sperduti, A. (2015). Online discovery of declarative process models from event streams. *IEEE Transactions on Services Computing, 8*(6), 833–846. https://doi.org/10.1109/TSC.2015.2459703

Burattin, A., Maggi, F. M., & Sperduti, A. (2016). Conformance checking based on multi-perspective declarative process models. *Expert Systems with Applications, 65*, 194–211. https://doi.org/10.1016/j.eswa.2016.08.040

Carmona, J., & Cortadella, J. (2014). Process discovery algorithms using numerical abstract domains. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 3064–3076. https://doi.org/10.1109/TKDE.2013.156

Carmona, J., Cortadella, J., & Kishinevsky, M. (2008). A region-based algorithm for discovering Petri nets from event logs. In *Business Process Management. BPM 2008. Lecture Notes in Computer Science* (Vol. 5240, pp. 176-190). https://doi.org/10.1007/978-3-540-85758-7_26

Celik, U., & Akçetin, E. (2018). Process mining tools comparison. *AJIT-e: Online Academic Journal of Information Technology, 9*(34). https://doi.org/10.5824/1309-1581.2018.4.007.x

Cheng, H. J., & Kumar, A. (2015). Process mining on noisy logs — Can log sanitization help to improve performance? *Decision Support Systems, 79*, 138–149. https://doi.org/10.1016/j.dss.2015.08.003

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 1*(13), 15-42. https://doi.org/10.1207/s15327809jls1301_2

Conforti, R., La Rosa, M., & ter Hofstede, A. H. M. (2017). Filtering out infrequent behavior from business process event logs. *IEEE Transactions on Knowledge and Data Engineering, 29*(2), 300–314. https://doi.org/10.1109/TKDE.2016.2614680

Cross, N. (2011). *Design thinking: Understanding how designers think and work*. Berg Publishers. https://doi.org/10.5040/9781474293884

de Leoni, M., & Marrella, A. (2017). Aligning real process executions and prescriptive process models through automated planning. *Expert Systems with Applications, 82*, 162–183. https://doi.org/10.1016/j.eswa.2017.03.047

de Medeiros, A. K. A., & van der Aalst, W. M. P. (2008). Towards multi-perspective process mining. In *CAiSE Forum* (pp. 87-92).

de Medeiros, A. K. A., Weijters, A. J. M. M., & van der Aalst, W. M. P. (2007). Genetic process mining: An experimental evaluation. *Data Mining and Knowledge Discovery, 14*(3), 245-304. https://doi.org/10.1007/s10618-006-0061-7

De Weerdt, J., & van der Aalst, W. M. P. (2013). Process mining: A guide to process discovery techniques. *Decision Support Systems, 54*(1), 138-149. https://doi.org/10.1016/j.dss.2012.06.003

Diamantini, C., Genga, L., Potena, D., & van der Aalst, W. M. P. (2016). Building instance graphs for highly variable processes. *Expert Systems with Applications, 59*, 101–118. https://doi.org/10.1016/j.eswa.2016.04.021

Dixit, P., Buijs, J., van der Aalst, W., Hompes, B., & Buurman, J. (2017). Using domain knowledge to enhance process mining results. In *Lecture Notes in Business Information Processing* (Vol. 244, pp. 76–104). Springer. https://doi.org/10.1007/978-3-319-53435-0_4

dos Santos Garcia, C., Meincheim, A., Faria Junior, E. R., Rosano Dallagassa, M., Vecino Sato, D. M., Carvalho, D. R., Portela Santos, E. A., & Scalabrin, E. E. (2019). Process mining techniques and applications: A systematic mapping study. *Expert Systems with Applications, 133*, 260-295. https://doi.org/10.1016/j.eswa.2019.05.003

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). *Fundamentals of business process management*. Springer. https://doi.org/10.1007/978-3-662-56509-4

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532–550. https://doi.org/10.5465/amr.1989.4308385

Essmita, A. & Gupta, E. (2014). Process mining algorithms. *International Journal of Advance Research in Science and Engineering, 3*(11). https://www.researchgate.net/publication/274248099_PROCESS_MINING_ALGORITHMS

Evermann, J. (2016). Scalable process discovery using Map-Reduce. *IEEE Transactions on Services Computing, 9*(3), 469–481. https://doi.org/10.1109/TSC.2014.2367525

Fahland, D., & van der Aalst, W. M. P. (2012). Repairing process models to reflect reality. In A. Barros, A. Gal, & E. Kindler (Eds.), *Business process management. BPM 2012. Lecture Notes in Computer Science* (Vol. 7481, pp. 288–303). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32885-5_19

Folino, F., Greco, G., Guzzo, A., & Pontieri, L. (2009). Discovering expressive process models from noised log data. In *Proceedings of the 2009 International Database Engineering & Applications Symposium on - IDEAS '09* (pp. 72-79). https://doi.org/10.1145/1620432.1620449

García-Bañuelos, L., van Beest, N. R. T. P., Dumas, M., La Rosa, M., & Mertens, W. (2018). Complete and interpretable conformance checking of business processes. *IEEE Transactions on Software Engineering, 44*(3), 262–290. https://doi.org/10.1109/TSE.2017.2668418

Goedertier, S., Martens, D., Vanthienen, J., & Baesens, B. (2009). Robust process discovery with artificial negative events. *The Journal of Machine Learning Research, 10*, 1305-1340. https://doi.org/10.1145/1577069.1577113

Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*, 337-356. https://doi.org/10.25300/MISQ/2013/37.2.01

Günther, C. W., & Rozinat, A. (2012). Disco: discover your processes. In N. Lohmann, & S. Moser (Eds.), *Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012)* (pp. 40-44). (CEUR Workshop Proceedings; Vol. 940). CEUR-WS.org. http://ceur-ws.org/Vol-940/

Günther, C. W., & van der Aalst, W. M. P. (2007). Fuzzy mining: Adaptive process simplification based on multi-perspective metrics. In *International Conference on Business Process Management* (pp. 328-343). Springer. https://doi.org/10.1007/978-3-540-75183-0_28

Gunther, C. W., & Verbeek, H. M. W. (2014). XES - standard definition. (BPM reports; Vol. 1409). BPMcenter. org. https://research.tue.nl/en/publications/e3791a42-4f32-4ff1-8dc1-338ab8f34afb?

He, Z., Du, Y., Wang, L., Qi, L., & Sun, H. (2018). An Alpha-FL algorithm for discovering free loop structures from incomplete event logs. *IEEE Access, 6*, 27885–27901. https://doi.org/10.1109/ACCESS.2018.2840818

Hevner, A. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems, 19*. https://www.researchgate.net/publication/254804390_A_Three_Cycle_View_of_Design_Science_Research

Hevner, A. R., March, S. T., & Park, J. (2004). Design research in information systems research. *MIS Quarterly, 28*(1), 75-105. https://doi.org/10.1007/978-1-4419-5653-8

Hommes, B.-J., & van Reijswoud, V. (2000). Assessing the quality of business process modelling techniques. *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Vol. 1, 10 pages. https://doi.org/10.1109/HICSS.2000.926591

HSPI Management Consulting. (2017). Process mining: A database of applications, 2017 edition. Retrieved from https://www.hspi.it/wp-content/uploads/2017/11/HSPI_Process_Mining_Database_v1.1-Nov_17.pdf

Huang, Z., & Kumar, A. (2012). A study of quality and accuracy tradeoffs in process mining. *INFORMS Journal on Computing, 24*(2), 211–230. https://doi.org/10.1287/ijoc.1100.0444

Karn, A., & Acharya, A. (2021). Incorporation of deep neural network & reinforcement learning with domain knowledge. *arXiv preprint*. https://doi.org/10.48550/arXiv.2107.14613

Kreimeyer, M., König, C., & Braun, T. (2008). Structural metrics to assess processes. In *Process Modeling and Analysis* (pp. 245–258). Presented at the 10th International DSM Conference, Stockholm. https://www.researchgate.net/publication/259461256_Structural_Metrics_to_Assess_Processes

Lee, G. S., & Yoon, J.-M. (1992). An empirical study on the complexity metrics of Petri nets. *Microelectronics and Reliability, 32*(3), 323–329. https://doi.org/10.1016/0026-2714(92)90061-O

Lee, T., & Yoon, S. (2004). Metrics for process model complexity in Petri nets. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 34*(4), 423–430. https://doi.org/10.1109/TSMCA.2004.830674

Leemans, S. J. J., Fahland, D., & van der Aalst, W. M. P. (2014). Discovering block-structured process models from event logs containing infrequent behaviour. *International Conference on Business Process Management - Business Process Management Workshops, 171*, 66–78. https://doi.org/10.1007/978-3-319-06257-0_6

Li, W., Zhu, H., Liu, W., Chen, D., Jiang, J., & Jin, Q. (2018). An anti-noise process mining algorithm based on minimum spanning tree clustering. *IEEE Access, 6*, 48756–48764. https://doi.org/10.1109/ACCESS.2018.2865540

Liesaputra, V., Yongchareon, S., & Chaisiri, S. (2015). Efficient process model discovery using maximal pattern mining. In *Business Process Management. BPM 2016. Lecture Notes in Computer Science (Vol. 9253, pp. 355-368)*. https://doi.org/10.1007/978-3-319-23063-4_29

Maaradji, A., Dumas, M., La Rosa, M., & Ostovar, A. (2017). Detecting sudden and gradual drifts in business processes from execution traces. *IEEE Transactions on Knowledge and Data Engineering, 29*(10), 2140–2154. https://doi.org/10.1109/TKDE.2017.2720601

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251-266. https://doi.org/10.1016/0167-9236(95)00041-2

Marin-Castro, H., & Tello-Leal, E. (2021). Event log preprocessing for process mining: A review. *Applied Sciences, 11*(22), 10556. https://doi.org/10.3390/app112210556

Marshall, C., & Rossman, G. B. (2016). *Designing qualitative research* (6th ed.). Sage Publications. https://www.proquest.com/scholarly-journals/designing-qualitative-research-sixth-edition/docview/1722656115/se-2

Mason, J. (2002). *Qualitative researching* (2nd ed.). Sage Publications. https://uk.sagepub.com/en-gb/eur/qualitative-researching/book244365#contents

Mendling, J. (2008). *Metrics for process models: Empirical foundations of verification, error prediction, and guidelines for correctness.* Springer. https://doi.org/10.1007/978-3-540-89224-3

Montasser, R. K., & Helal, I. M. A. (2023). Process discovery automation: Benefits and limitations. In *2023 Intelligent Methods, Systems, and Applications (IMSA)* (pp. 496-501). IEEE. https://doi.org/10.1109/IMSA58542.2023.10217621

Morasca, S. (1999). Measuring attributes of concurrent software specifications in Petri nets. In *METRICS '99: Proceedings of the 6th International Symposium on Software Metrics* (pp. 100–110). IEEE Computer Society Press.

Morasca, S. (1999). Measuring the complexity of Petri nets: A theoretical approach. *Journal of Systems and Software, 48*(3–4), 227–241. https://doi.org/10.1016/S0164-1212(99)00038-1

Munoz-Gama, J., & Carmona, J. (2010). A fresh look at precision in process conformance. In *International Conference on Business Process Management* (pp. 211-226). Springer. https://doi.org/10.1007/978-3-642-15394-6_16

Nguyen, H. T. C., Lee, S., Kim, J., Ko, J., & Comuzzi, M. (2018). Autoencoders for improving quality of process event logs. *Expert Systems with Applications, 131*, 132–147. https://doi.org/10.1016/j.eswa.2019.04.052

Nissen, M. E. (1994). Valuing it through virtual process measurement. In *Proc. 15th International Conference on Information Systems* (pp. 309–323).

Nissen, V. (2008). Using design heuristics and metrics to assess process model quality. *Information Systems Research, 19*(2), 133–146. https://doi.org/10.1287/isre.1070.0134

Nolle, T., Seeliger, A., & Mühlhäuser, M. (2016). Unsupervised anomaly detection in noisy business process event logs using denoising autoencoders. In *Discovery Science: Lecture Notes in Computer Science, 9956*, 442-456. https://doi.org/10.1007/978-3-319-46307-0_28

Norouzifar, A., Kourani, H., Dees, M., & van der Aalst, W. M. P. (2024). Bridging domain knowledge and process discovery using large language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2408.17316

Okoye, K., Naeem, U., & Islam, S. (2017). Semantic fuzzy mining: Enhancement of process models and event logs analysis from syntactic to conceptual level. *International Journal of Hybrid Intelligent Systems, 14*(1-2), 67-98. https://doi.org/10.3233/HIS-170243

Osman, C. C., & Ghirana, A. M. (2019). When Industry 4.0 meets process mining. *Procedia Computer Science, 159*, 2130-2136. https://doi.org/10.1016/j.procs.2019.09.386

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Sage Publications. https://journals.sagepub.com/doi/10.1177/1035719X0300300213

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77. https://doi.org/10.2753/MIS0742-1222240303

Pérez, D., Fundora-Ramírez, O., Lazo-Cortés, M., & Roche-Escobar, R. (2015). Recommendation of process discovery algorithms through event log classification. In *Proceedings of the 13th International Conference on Business Process Management* (pp. 3–12). Springer. https://doi.org/10.1007/978-3-319-19264-2_1

Rabbi, F., Banik, D., Hossain, N. U. I., & Sokolov, A. (2024). Using process mining algorithms for process improvement in healthcare. *Computers in Industry, 134*, 103567. https://doi.org/10.1016/j.compind.2024.103567

Reijers, H. A., & van der Aalst, W. M. P. (2005). The effectiveness of workflow management systems: Predictions and lessons learned. *Information and Software Technology, 47*(5), 315-330. https://doi.org/10.1016/j.infsof.2004.10.005

Reinkemeyer, L. (2022). Status and Future of Process Mining: From Process Discovery to Process Execution. In van der Aalst, W. M. P., & Carmona, J. (Eds.), *Process Mining Handbook (Vol. 448, pp. xx-xx)*. Springer, Cham. https://doi.org/10.1007/978-3-031-08848-3_13

Rembert, A. J., Omokpo, A., Mazzoleni, P., & Goodwin, R. T. (2013). Process discovery using prior knowledge. In *Service-Oriented Computing. ICSOC 2013. Lecture Notes in Computer Science (Vol. 8274, pp. 173-186)*. https://doi.org/10.1007/978-3-642-45005-1_23

Ribeiro, J., Carmona, J., Mısır, M., & Sebag, M. (2022). A recommender system for process discovery. In *Proceedings of the 2022 IEEE Conference* (pp. 1-6). IEEE. https://doi.org/10.1109/IEEECONF55059.2022.9810387

Rocha Silva, F. A. (2018). Analytical intelligence in processes: Data science for business. *IEEE Latin America Transactions, 16*(8), 2240–2247. https://doi.org/10.1109/TLA.2018.8528241

Rogge-Solti, A., Mans, R. S., van der Aalst, W. M. P., & Weske, M. (2013). Improving documentation by repairing event logs. In *The Practice of Enterprise Modeling: PoEM 2013. Lecture Notes in Business Information Processing, Vol 165*. https://doi.org/10.1007/978-3-642-41641-5_10

Rozinat, A., de Medeiros, A. K. A., Gunther, C. W., Weijters, A. J. M. M., & van der Aalst, W. M. P. (2007). Towards an evaluation framework for process mining algorithms. *BPM Center Report BPM-07-06*. BPMcenter.org. https://www.researchgate.net/publication/228621875_Towards_an_evaluation_framework_for_process_mining_algorithms

Rozinat, A., & van der Aalst, W. M. P. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems, 33*(1), 64-95. https://doi.org/10.1016/j.is.2007.07.002

Sahu, M., Chakraborty, R., & Nayak, G. (2018). A task-level parallelism approach for process discovery. *International Journal of Engineering & Technology, 7*, 2446-2452. https://doi.org/10.14419/ijet.v7i4.14748

Schuster, D., van Zelst, S. J., & van der Aalst, W. M. P. (2022). Utilizing domain knowledge in data-driven process discovery: A literature review. *Computers in Industry, 139*, 103612. https://doi.org/10.1016/j.compind.2022.103612

Seidman, I. (2013). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (4th ed.). Teachers College Press. https://www.scirp.org/reference/referencespapers?referenceid=1508079

Song, W., Jacobsen, H. A., Ye, C., & Ma, X. (2016). Process discovery from dependence-complete event logs. *IEEE Transactions on Services Computing, 9*(5), 714-727. https://doi.org/10.1109/TSC.2015.2426181

Song, W., Xia, X., Jacobsen, H. A., Zhang, P., & Hu, H. (2017). Efficient alignment between event logs and process models. *IEEE Transactions on Services Computing, 10*(1), 136-149. https://doi.org/10.1109/TSC.2016.2601094

van den Broucke, S. K. L. M., De Weerdt, J., Vanthienen, J., & Baesens, B. (2014). Determining process model precision and generalization with weighted artificial negative events. *IEEE Transactions on Knowledge and Data Engineering, 26*(8), 1877–1889. https://doi.org/10.1109/TKDE.2013.130

van der Aalst, W. M. P. (2018). Process mining and simulation: A match made in heaven! *In Proceedings of the 50th Computer Simulation Conference (SummerSim '18). Society for Computer Simulation International, San Diego, CA, USA, Article 4, 1–12.* https://doi.org/10.22360/summersim.2018.scsc.005

van der Aalst, W. M. P. (2016). Process mining: Data science in action (2nd ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

van der Aalst, W. M. P. (2011). Process mining: Discovery, conformance and enhancement of business processes. Springer-Verlag. https://doi.org/10.1007/978-3-642-19345-3

van der Aalst, W. (2012). Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems, 3*(7), 1-17. https://doi.org/10.1145/2229156.2229157

van der Aalst, W. M. P., Adriansyah, A., Alves De Medeiros, A. K., Arcieri, F., Baier, T., Blickle, T., ... & Wynn, M. (2012). Process mining manifesto. In F. Daniel, K. Barkaoui, & S. Dustdar (Eds.), *Business Process Management Workshops* (pp. 169-194). Lecture Notes in Business Information Processing, 99. https://doi.org/10.1007/978-3-642-28108-2_19

van der Aalst, W. M. P., & de Medeiros, A. K. A. (2005). Process mining and security: Detecting anomalous process executions and checking process conformance. *Electronic Notes in Theoretical Computer Science, 121. 3-21.* https://doi.org/ 10.1016/j.entcs.2004.10.013

van der Aalst, W. M. P., & Song, M. (2005). Mining social networks: Uncovering interaction patterns in business processes. *Business Process Management Journal, 11*(3), 224-240. https://doi.org/10.1108/14637150510605262

van der Aalst, W. M. P., & Weijters, A. J. M. M. (2004). Process mining: A research agenda. *Computers in Industry, 53*(3), 231-244. https://doi.org/10.1016/j.compind.2003.10.001

van der Aalst, W. M. P., & Weijters, T. (2006). Process mining: A research agenda. *Computers in Industry, 53*(3), 231-244. https://doi.org/10.1016/j.compind.2003.10.001

van der Werf, J. M. E. M., van Dongen, B. F., Hurkens, C. A. J., & Serebrenik, A. (2008). Process discovery using integer linear programming. In *Applications and Theory of Petri Nets. PETRI NETS 2008. Lecture Notes in Computer Science (Vol. 5062, pp. 170-185)*. https://doi.org/10.1007/978-3-540-68746-7_24

van Dongen, B. F., Alves de Medeiros, A. K., & Wen, L. (2009). Process mining: Overview and outlook of Petri net discovery algorithms. In *Transactions on Petri Nets and Other Models of Concurrency II: Lecture Notes in Computer Science, Vol 5460*. https://doi.org/10.1007/978-3-642-00899-3_13

van Dongen, B. F., & van der Aalst, W. M. P. (2005). A meta-model for process mining data. *Lecture Notes in Computer Science, 3540*, 309-320. https://doi.org/10.1007/11539188_30

van Eck, M. L., & van der Aalst, W. M. P. (2016). Artifact-centric process mining: Analyzing object-centric behavior in event logs. *Business Process Management Journal, 22*(2), 11-26. https://doi.org/10.1108/BPMJ-04-2015-0106

van Zelst, S. J., & van der Aalst, W. M. P. (2018). Process mining in practice: Comparative study of tools and techniques. *Business Process Management Workshops*, 513-524. https://doi.org/10.1007/978-3-319-72142-0_49

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36–59. https://doi.org/10.1287/isre.3.1.36

Wang, J., Tan, S., Wen, L., Wong, R. K., & Guo, Q. (2012). An empirical evaluation of process mining algorithms based on structural and behavioral similarities. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 211–213). https://doi.org/10.1145/2245276.2245316

Wang, J., Wong, R. K., Ding, J., Guo, Q., & Wen, L. (2012). Efficient selection of process mining algorithms. *IEEE Transactions on Services Computing, 6*(4), 484-496. https://doi.org/10.1109/TSC.2012.20

Weber, P., Bordbar, B., & Tinô, P. (2013). A principled approach to mining from noisy logs using Heuristics Miner. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. https://doi.org/10.1109/CIDM.2013.6597226

Weijters, A. J., & van der Aalst, W. V. (2001). Process mining: Discovering workflow models from event-based data. In *Proceedings of the ECAI Workshop on Knowledge Discovery and Spatial Data* (pp. 1-11). Springer. http://www.padsweb.rwth-aachen.de/wvdaalst/publications/p128.pdf

Weijters, A. J. M. M., van der Aalst, W. M. P., & Alves de Medeiros, A. K. (2006). Process mining with the heuristics miner-algorithm. *Department of Technology Management, Eindhoven University of Technology*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.8288&rep=rep1&type=pdf

Wen, L. (2014). A universal significant reference model set for process mining evaluation framework. In *Conference: AP-BPM 2014*. https://www.researchgate.net/publication/269572281_A_Universal_Significant_Reference_Model_Set_for_Process_Mining_Evaluation_Framework

Wen, L. (2009). Process mining: Overview and outlook of Petri net discovery algorithms. *International Journal of Computer Science & Applications, 6*(3), 33-56. https://doi.org/10.24255/ijcsa.2009.6.3.33

Wen, L., Wang, J., van der Aalst, W. M. P., Huang, B., & Sun, J. (2009). A novel approach for process mining based on event types. *Journal of Intelligent Information Systems, 32*(2), 163-190. https://doi.org/10.1007/s10844-007-0052-1

Xu, Y., Du, Y., Luan, W., Qi, L., & Sun, H. (2018). Repairing process models with logical concurrent and causal relations via logical Petri nets. *IEEE Access, 6*, 56340-56355. https://doi.org/10.1109/ACCESS.2018.2872640

Xu, J., & Liu, J. (2019). A profile clustering based event logs repairing approach for process mining. *IEEE Access, 7*, 17872–17881. https://doi.org/10.1109/ACCESS.2019.2894905

Yahya, B. N., Song, M., Bae, H., Sul, S., & Wu, J. Z. (2016). Domain-driven actionable process model discovery. Available online, May 7, 2016. Version of record published September 3, 2016. https://doi.org/10.1016/j.cie.2016.05.010

ZarehFarkhady, R., Aali, S. H., & Branch, B. (2012). A two-phase approach for process mining in incomplete and noisy logs. *International Journal of Computer Science, 9*, 160-165. https://www.ijcsi.org/papers/IJCSI-9-1-2-160-165.pdf

Zhang, X., Du, Y., Qi, L., & Sun, H. (2018). An approach for repairing process models based on logic Petri nets. *IEEE Access, 6*, 29926–29939. https://doi.org/10.1109/ACCESS.2018.2843137

Zhang, Z., Johnson, C., Venkatasubramanian, N., & Ren, S. (2022). Process scenario discovery from event logs based on activity and timing information. *Journal Name, Volume(Issue)*, page numbers. https://doi.org/10.1016/j.somej.2022.02.003

Zheng, W., Du, Y., Qi, L., & Wang, L. (2019). A method for repairing process models containing a choice with concurrency structure by using logic Petri nets. *IEEE Access, 7*, 13106–13120. https://doi.org/10.1109/ACCESS.2019.2893327

zur Muehlen, M. (2002). Workflow-based process controlling. Foundation design and Application of Workflow-driven Process Information Systems. https://lia.disi.unibo.it/Staff/DarioBottazzi/1.pdf

zur Muehlen, M., & Rosemann, M. (2000). Workflow-based process monitoring and controlling-technical and organizational issues. *Proceedings of the 33rd Hawaii international conference on system science (HICSS-33). 10 pp. vol.2*. https://doi.org/10.1109/HICSS.2000.926853

APPENDIX A. MID-TERM VALIDATION INTERVIEW TRANSCRIPT

Researcher: Thanks. Let me just open this. So, this master's thesis is dedicated to process discovery algorithms, focusing on qualitative research based on an extensive literature review, with some parts involving a systematic literature review. I used the Scopus database for the articles, leveraging its repository for thorough analysis. The thesis title addresses key challenges in discovering process models from repaired event logs. Often, our databases are not very "clean," meaning there are various types of noise and other issues like loops or incomplete logs that make process model discovery challenging.

Different algorithms have been identified and grouped through literature review. I aimed to map and categorize them, also providing recommendations on how to choose specific algorithms for different scenarios, such as incomplete logs or loops. My research question is dedicated to mapping these algorithms and offering recommendations for selecting them based on distinct situations. This will involve schemes and steps for choosing algorithms effectively.

In the thesis, I present three main types of process mining: process discovery, conformance checking, and enhancement. My focus is primarily on process discovery within process mining. I divided the research design into four stages, followed by a conclusion. First, I investigated the various types of algorithms through a literature review, identifying as many process discovery algorithms as possible using Scopus with relevant keyword combinations. This information will also be included in the dissertation.

I validated some aspects of my findings with ChatGPT, checking specific formulations to ensure I had captured all relevant articles, as ChatGPT can serve as an additional literature review tool with web access. My primary sources remain Scopus, applying filtering and systematic literature review steps. ChatGPT helped validate certain aspects, ensuring comprehensive coverage of the algorithms.

Initially, I intended to conduct quantitative analysis and test the algorithms on different datasets. However, accessing databases presented challenges, partly due to privacy concerns and limited database availability, especially in Armenia. Consequently, I focused on qualitative research, performing comparative analyses to assess the algorithms. I also assigned scores to the algorithms based on different criteria, evaluating their utility in specific scenarios.

In the end, I created four simplified models, or steps, to provide recommendations for algorithm selection across various situations. The conclusion synthesizes all findings, grouping algorithms into four categories and discussing challenges across databases with differing issues. These recommendations suggest a possible automation system, which could be implemented through a script—something I plan to discuss as a future research avenue.

I reviewed around 400 articles from Scopus, filtering them to approximately 220-250 relevant ones, and ultimately used 32 articles for the core thesis. The entire list is available in the annex. The research steps involved filtering, grouping algorithms, assigning scores based on specific criteria, and providing a recommendation system illustrated through simple models. I used around 11 criteria for scoring, supported by literature that also informs the criteria.

This study can also apply to conformance checking and reinforcement methods. I plan to validate some aspects further through expert interviews. In the second phase, I developed decision trees to depict the recommendation system, evaluating algorithms based on metrics. The process discovery basics include filtering event logs from databases, identifying algorithms, and exploring elements such as control flow and resources.

For cases with incomplete event logs, I included specific algorithms to assist in mining models. Similarly, in scenarios with loops, I categorized free loops and short loops and identified algorithms suitable for these conditions. Additionally, I highlighted the noise, incompleteness, and loops as key challenges. I also discussed alternative process discovery approaches, such as time-based and stochastic process models.

That's an overview of my thesis. I'm open to feedback on its usefulness and any recommendations for further analysis, which could strengthen the validation section.

Shall we go through the questions one by one?

The Expert: Yes, let's proceed.

Researcher: Perfect. First of all, looking at this from a research perspective, is it useful for researchers?

The Expert: I think, in general, the concept of a repaired event block and understanding how to optimally repair one is incredibly valuable. There's a lot being done in this field right now. For instance, are you familiar with the concept of a stochastic event block? This involves a probabilistic understanding of an event block, where you consider whether something occurs or doesn't occur with a certain level of certainty. This approach aligns with current research directions and has been receiving a lot of attention in the past couple of years. Research interest in this area has picked up recently, and although it's still primarily in the research phase, I believe it will be relevant in industry within the next few years.

To give some insights, we at Celonis, for example, are already having initial discussions on what can be done regarding event block repair and finding optimal methods for doing so. Currently, our software does not have extensive functionality for this; while it allows some preliminary data adjustments before loading into Celonis, there's no direct feature for event block repair at the moment. However, it's something that interests us greatly, and we're beginning to investigate it further. We're looking at the latest research to understand the best practices for this area, as we believe it's already a trending topic in research and will soon gain traction in industry as well. I anticipate that in a year or two, we'll see implementations of these methods in software.

Researcher: All right, onto the next question: What recommendations or suggestions would you consider for improving the proposed recommendation system?

The Expert: It hasn't been extensively tested on real datasets, so I'm not certain if what I found in the literature is directly applicable. For example, there might be an article about a specific algorithm, but I couldn't find relevant studies on its usage. So, I think what you did in terms of comparing different algorithms and the mechanisms behind them is solid. At this point, it's mostly theoretical. I think the

next step—if you were to continue this as a project post-thesis—would be to test it on actual data. This could involve finding datasets that include the exact challenges you've presented, such as loops and other complex cases.

I would love to see a comparative analysis of how these different algorithms perform on real data, including all the challenges you mentioned. You might also conduct an industry check-in to confirm the relevance of these challenges in different sectors, like analyzing how often loops occur in a purchase-to-pay or an order-to-cash process. Despite the unique characteristics of each dataset, certain challenges, like maverick buying and price changes in purchase-to-pay processes, are common across organizations. By examining these scenarios with theoretical concepts, we can narrow down likely outcomes in these processes and see how different algorithms handle event log repair.

Another direction could be to apply these insights across core process mining areas, like purchase-to-pay, order-to-cash, accounts payable, and accounts receivable, to determine how they apply to each field on real data. I believe this would be a valuable next step.

Researcher: Great, thank you. And finally, do you have any additional comments on the proposed analysis, besides what you've mentioned? Any potential issues, perhaps?

The Expert: No, I think the approach is systematic, thorough, and well-structured. One aspect I particularly liked about your thesis is the incorporation of qualitative insights alongside quantitative ones, which I find practically relevant. When it comes to data repair, it's often about understanding what actually happened in the process. This is where qualitative analysis can complement process mining, which is typically more quantitative. Combining process mining with interviews or discussions can enrich the data insights. You could bring process mining event logs into interviews and ask about specific instances to get a narrative behind the data. This approach of integrating quantitative findings with expert knowledge could be expanded upon.

Having both quantitative and qualitative perspectives enhances the analysis. So, I'd suggest further consideration of this step, either as an integral part of the study or in later stages when testing the data, as qualitative insights provide valuable context.

Researcher: Thank you, Angela. That was very helpful.

The Expert: Perfect. I'm glad it helped.

Researcher: Sometimes shorter, targeted discussions can be more beneficial than longer meetings. Testing this on real data and scenarios would definitely add value. I envision two stages for testing: first, on fabricated datasets containing loops, noise, and other challenges, and then on real-world scenarios relevant to fields like purchase-to-pay and accounts receivable. This can also be combined with qualitative insights and expert knowledge to identify patterns and best practices in commonly used cases. This is a very promising direction. Thank you very much.

The Expert: You're welcome! Please let me know if you need anything else. I'll send over the recording shortly. Thank you again.