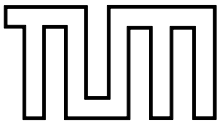
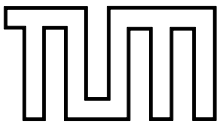


DEPARTMENT OF INFORMATICS
TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Information Systems

Simulation of Continuous Business Process Data for Process Mining in Teaching

Andreas Heckl



DEPARTMENT OF INFORMATICS
TECHNICHAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Information Systems

Simulation von kontinuierlichen Geschäftsprozessdaten für Process Mining in der Lehre

Simulation of Continuous Business Process Data for Process Mining in Teaching

Author:	Andreas Heckl
Examiner:	Prof. Dr. Helmut Krcmar
Supervisor:	Clemens Drieschner
Submission Date:	25.02.2022

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

I assure the single-handed composition of this bachelor's thesis only supported by declared resources.

Garching, Munich, 25.02.2022 A. Heckl

Place, Date Signature

Abstract

Process mining is a promising technology that sits between Business Process Management and Data Science. Among other benefits, process mining enables companies to reduce costs (e.g., by identifying bottlenecks or automation potentials). Digital business process data, which is stored in so-called event logs, allows the visualization and analysis of business processes. The SAP University Competence Center, in cooperation with the process mining software vendor Celonis SE, offers teaching materials for professors, students, and pupils. However, the existing training data is static and does not change with specific user interventions, such as a release of invoices. This work contributes to a more realistic teaching approach by providing a simulation software for continuous business process data. First, the application scope of process mining projects in the industry is determined. Then, a prototypical implementation is presented that allows the continuous simulation of business process data. Finally, a feedback mechanism is developed that allows specific user interventions to affect the future simulation of the process data. A database is set up and connected to the Celonis Execution Management System. The simulation of the data is performed with a Python script. The order-to-cash process of a fictitious bike rental company was used as an example process.

Keywords: Business Process, Celonis, Continuous Simulation, Event Log, Order-to-Cash, Process Data, Process Mining, Simulation, Teaching.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BPM	Business Process Management
BPMN	Business Process Model and Notation
DFG	directly-follows graph
ERP	Enterprise Resource Planning
GBS	Global Bike Share
IT	Information Technology
O2C	order-to-cash
PQL	Process Query Language
RQ1	Research Question 1
RQ2	Research Question 2
RQ3	Research Question 3
TFPM	Task Force on Process Mining
UCC	University Competence Center

1 Introduction

1.1 Motivation

For most companies, digital transformation not only comprises products, services and business models, but also business processes (Handelsblatt Research Institute, 2020). Process complexity is higher than ever and this confronts companies with major challenges (Daniels, 2022). As more and more business processes are being digitized, business process data is generated and stored in Enterprise Resource Planning (ERP) systems or Customer Relationship Management systems, for example. One technology that aims to take advantage of this data is process mining. It is used to discover, monitor, and improve processes (van der Aalst, Adriansyah, et al., 2011). Process mining can be seen as the missing link between traditional Business Process Management (BPM) and Big Data (van der Aalst, 2012a). Traditional approaches to process discovery (e.g., user interviews) are prone to disagreement and subjectivity, whereas process mining techniques promote objectivity (Koplowitz, Mines, Vizgaitis, & Reese, 2019). Examples of benefits that companies expect from the adoption of process mining initiatives are improving business agility, reducing process cycle times, and increasing transparency of process flows (Kong, 2021). Process mining can even be used to predict completion times of running processes (van der Aalst, Schonenberg, & Song, 2011).

Process mining is a hot topic both in academia and industry. The number of publications on process mining has been growing at a high rate since the mid-2000 (Reinkemeyer, 2020), and research companies expect the market size to grow considerably more than 40% annually in the coming years (Galic & Wolf, 2021; Kerremans, Srivastava, & Choudhary, 2021; Kong, 2021; Mehar, 2020). Especially against today's background of widely used information systems and phenomena such as complex supply chains, it seems plausible that process mining will continue to grow strongly in importance. This forecast is supported by the fact that traditional interview-based process discovery and modeling is costly and time-consuming (Kerremans et al., 2021). Today, about 40 vendors of process mining software exist in the market (Kerremans et al., 2021) .

Process mining is often not a trivial task, but introduces challenges such as finding process loops, hidden processes, or duplicates (van der Aalst & Weijters, 2004). It is therefore essential that, for the purposes of teaching, the underlying process data is as close to reality as possible. The problem, however, is that no real process data from industries that continuously supply data can be used in teaching; instead, the data that is used is static and simulated. This means that no specific user actions, such as a change of supplier or the release of an invoice, can be evaluated, as no new data is simulated that reflects these changes.

This thesis aims to contribute to improving teaching in the field process mining by creating a prototypical software artifact that continuously simulates the business process data of a fictitious bike rental company. Through continuity and a feedback mechanism between specific user actions and the simulated data, it is possible to evaluate these actions. In addition to the high practical relevance of this work, it also makes scientific contributions. First, a literature study was carried out to investigate how process mining is applied in the industry. Not only scientific literature was considered, but also the latest reports from research companies. Sec-

ond, the considerations and challenges that arose during the implementation of the simulation software are generalized to fit a variety of business processes. Thus, this thesis provides students with a more realistic, hands-on learning experience. Furthermore, companies will also benefit from employees who are better informed about process mining.

1.2 Research Questions and Thesis Organization

This thesis is structured as follows. Chapter 2 presents the methodological approach. It explains the general framework (namely, design science research) and presents the literature review approach that was taken for the first research question. Chapters 4, 5, and 6 investigate the following three research questions:

Research Question 1 (RQ1): *In which application areas and use cases is process mining utilized and what are its challenges and limitations?*

This research question is addressed in Chapter 4. A review of the scientific literature was conducted and enriched with reports and surveys from research companies to illuminate the application of process mining in the industry.

Research Question 2 (RQ2): *What aspects must a continuous simulation of business process data for teaching consider?*

This research question is addressed in Chapter 5. A prototypical simulation software was developed and is presented in this chapter.

Research Question 3 (RQ3): *What influence do operational changes have on continuously simulated data and how is a feedback mechanism to be translated technically?*

This research question is addressed in Chapter 6. It directly builds upon RQ2, outlining how the software prototype presented in Chapter 5 is extended to meet the requirements of RQ3.

Chapter 7 provides a conclusion to this work. This chapter also discusses the limitations that emerged while working on this thesis and gives an outlook on future work.

2 Approach and Methodology

2.1 Framework for This Thesis

This thesis adheres to the methodological framework for design science research developed by Hevner et al. (2004). In contrast to other common research paradigms in the field of Information Systems, such as the behavioral science paradigm, which sets out to explain human or organizational behavior, design science is fundamentally about problem solving and aims at creating new artifacts that solve real existing problems (Hevner et al., 2004). As the purpose of this thesis is to create a simulation software artifact for business process data, design science was chosen as a research methodology. In their framework, Hevner et al. propose seven guidelines that a design science work should follow. The rest of this section describes how this thesis complies with these guidelines.

Guideline 1 - Design as an Artifact: By developing a prototype of a simulation software for continuous business process data, a viable artifact is created.

Guideline 2 - Problem Relevance: This thesis addresses an existing problem in the real world, namely the lack of continuous data in process mining education. For teaching fundamental process mining techniques, static event data is not as suitable as continuous event data, as because static data is inherently not immutable and therefore does not allow the evaluation of changes to evaluate changes in a process or certain actions that were taken to improve a process. This problem can be resolved by continuously simulating event data.

Guideline 3 - Design Evaluation: The simulated data of the prototype was compared to an existing data set. The feedback mechanism, which is outlined in Chapter 6, was evaluated together with the help of Celonis employees.

Guideline 4 - Research Contributions: The simulation prototype contributes to a more realistic teaching approach in the field of process mining. Students will be able to apply process mining techniques and best practices to the continuously simulated data. This helps them to evaluate the impact of specific actions.

Guideline 5 - Research Rigor: By adhering to the principles of prototyping proposed by Naumann and Jenkins (1982), a rigorous method for designing the software artifact was used. This method is described in more detail in Chapter 5.

Guideline 6 - Design as Search Process: To create an effective artifact, the output of the continuous simulation is designed to be similar to an existing data set of 10,000 cases. This data set is provided by the SAP University Competence Center (UCC).

Guideline 7 - Communication of Research: This work is presented effectively, because the thesis explains the software artifact in a general way for management-oriented readers and, for technically oriented readers, further details can be found in the appendix and the documentation.

2.2 Literature Review Approach

To create a solid foundation in academic projects, it is essential to review relevant literature in the respective field (Webster & Watson, 2002). Thus, to investigate RQ1, a literature review was conducted. As readers should be able to assess the quality of a literature review, it is vital to document the review process properly (vom Brocke et al., 2009). In the field of Information Systems, a literature review is typically a complicated task, as a large and increasing number of articles are published every year in a broad spectrum of journals, conference proceedings, and many other sources (vom Brocke et al., 2009). Hence, it makes sense to use a framework that serves as a guide for writing a review. Vom Brocke et al. (2009) proposed a common framework for this purpose consisting of five phases. The rest of this section outlines how these phases were approached.

Phase I - definition of review scope: To define the scope of the review, a taxonomy proposed by Cooper (1988) was used (Table 1). Six characteristics and the respective categories of embodiment are listed. The green cells show which category applies to the respective characteristic in the review. For example, the audience for the review was defined as general and specialized scholars. The review was not written for the general public, who is typically not interested in process mining, nor was it written for practitioners, as they are already familiar with process mining and do not need to simulate data, but rather work with real data.

Characteristic	Categories			
Focus	Research Outcomes	Research Methods	Theories	Applications
Goal	Integration	Criticism	Central Issues	
Organization	Historical	Conceptual	Methodological	
Perspective	Neutral Representation		Espousal of Position	
Audience	Specialized Scholars	General Scholars	Practitioners/Politicians	General Public
Coverage	Exhaustive	Exhaustive & Selective	Representative	Central/Pivotal

Table 1: Definition of Review Scope

Source: Own representation, based on Cooper (1988)

Phase II - conceptualization of topic: In this phase, definitions of key terms should be provided (vom Brocke et al., 2009). An overview of the topic was gained by reading review articles on process mining and a book by the Dutch scientist Wil van der Aalst.

Phase III - literature search: To find relevant articles, Webster and Watson (2002) suggest starting a literature search in the leading journals. One set of highly reputed journals in the Information Systems research community is the *Senior Scholars' Basket of Journals of the Association for Information Systems* (AIS, 2011), also known as the *Basket of 8*. These eight journals were supposed to serve as an entry point for the literature review. Three scientific databases (Scopus, WebOfScience and IEEE Xplore) were searched for the term "process

mining” and filtered by the eight journals. However, this search yielded only one result, and the topic of the resulting article was considered irrelevant for this thesis. A second search approach was used, which focused on keywords rather than on journals. Figure 1 shows the searched keywords.

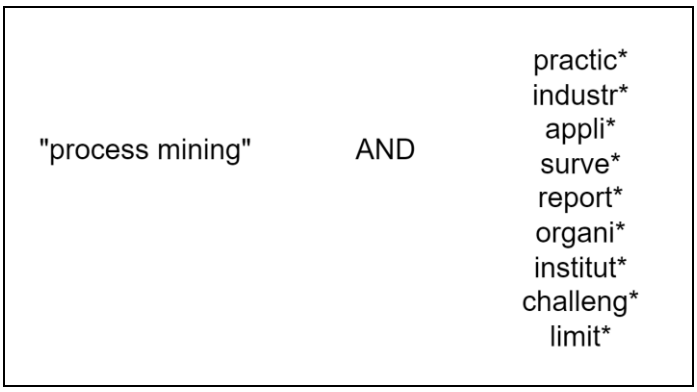


Figure 1: Combined Search Terms

Source: Own representation

The resulting articles were analyzed based on their titles and abstracts. A forward and backward search was conducted. Thirteen of the articles are referenced in the literature review. To complete the review and emphasize the practical relevance of process mining, studies, reports, and surveys from research firms such as Gartner, Deloitte, and Forrester were included. Such documents give information such as estimates of the market size for process mining and how companies use process mining. This type of information is usually not found in scientific articles and therefore complements the traditional literature search. The collection of these documents did not adhere to a structured pattern, but was rather conducted as a general internet search using the search engine Google. Search terms such as “process mining report 2021” and “process mining survey” were used. Nine of the reports found are referenced in the literature review.

The literature search was conducted in January and February 2022.

Phase IV - literature analysis and synthesis: To structure the selected documents, a concept matrix (Webster & Watson, 2002) was established. It is provided in Appendix A.

Phase V - research agenda: A comment on future work is given in Chapter 7.

3 Overview of Process Mining

This chapter gives an overview of the history, important terms and concepts, and market figures of process mining. It contributes to a better understanding of the following chapters, in which the three interrelated research questions are addressed.

Process mining settles in between BPM and Data Science. Figure 2 illustrates the interdisciplinary nature of process mining. In many Data Science approaches and technologies, the process perspective is absent. Traditional BPM approaches heavily rely on models instead of extracting knowledge from process data. Process mining overcomes these issues by combining perspectives and technologies from both disciplines. (van der Aalst, 2016)

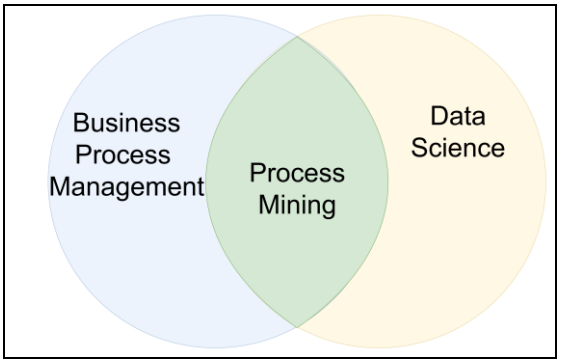


Figure 2: Interdisciplinary Nature of Process Mining

Source: Own representation, based on van der Aalst (2016)

The *IEEE Task Force on Process Mining*[1] (TFPM) defines process mining as "techniques, tools, and methods to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs commonly available in today's (information) systems" (van der Aalst, Adriansyah, et al., 2011). This definition shows that process mining is a comprehensive approach rather than a single specific technology or tool. The three types of process mining (discovery, conformance, and enhancement) also emerge from this definition. These are addressed in Section 3.5. The concept of event logs, which is fundamental in process mining, is elaborated in Section 3.4. Finally, the definition makes it clear that process mining is explicitly not about processes as they should be, but as they happen in the real world. Before the basic concepts of process mining are discussed further, a brief overview is given of the milestones in its history, mainly based on Reinkemeyer (2020).

3.1 A Brief History of Process Mining

Research on process mining started in 1999 at Technical University of Eindhoven. The Dutch scientist Wil van der Aalst coined the term and is a leading figure in this field. In the beginning, the first process discovery algorithms were developed, namely the alpha algorithm (van der Aalst, Weijters, & Maruster, 2004) and heuristic miner (Weijters, van der Aalst, & Alves de Medeiros, 2006). In 2005, researchers from Technical University of Eindhoven published

[1] https://www.tf-pm.org/

the ProM framework[2] (van Dongen, Alvares de Medeiros, Verbeek, Weijters, & van der Aalst, 2005). It supports a broad spectrum of process mining techniques and algorithms and is extensible via plug-ins. The open-source platform is still actively used and being further developed. In 2009, the TFPM was established. In 2011 they published the *Process Mining Manifesto*, which includes six guiding principles and eleven challenges of process mining (van der Aalst, Adriansyah, et al., 2011). In the same year, Wil van der Aalst published the first book solely on process mining (*Process Mining: Discovery, Conformance Checking and Enhancement*). In 2016, the second edition of the book followed, entitled *Process Mining – Data Science in Action*. In 2018, Gartner Inc. published its first *Market Guide for Process Mining* (Kerremans, 2018). In the same year, Celonis SE[3] became the first process mining company to reach a valuation of over one billion dollars (Steger, 2018). In the following year, the first conference specifically dedicated to process mining, namely the International Conference on Process Mining,[4] was established by the TFPM, which has been held annually since. PM4PY,[5] a Python framework for process mining, was also released in 2019, supporting process discovery and conformance checking. In addition, Celonis was awarded the German Future Award (Deutscher Zukunftspreis) this year (Höpner & Kerkmann, 2019). In 2021, Celonis became the first process mining software vendor to reach a valuation of 10 billion dollars (Holzki, 2021). As of 2021, there are about 40 vendors of process mining software available (Kerremans et al., 2021).

The next section gives market figures on process mining. They demonstrate that process mining has emerged from academia and entered industry, as it is shown that the market for process mining is expected to grow strongly in the coming years.

3.2 Market Figures on Process Mining

Estimates of market size for process mining software vary. Since 2018, Gartner Inc. has released annual market guides that include size estimations. Table 2 shows their results in the respective years. In these reports, *market size* is defined as product license and maintenance revenue. In their latest 2021 report, they forecast the market to grow between 40% and 50%, passing $1 billion in 2022.

Year of report	Reference year for market size	Market size in million USD	Growth rate compared to previous year
2018	2017	120	-
2019	2018	160	33%
2020	2019	320	100%
2021	2020	550	72%

Table 2: Market Size Estimations for Process Mining Software

Source: Own representation, based on Kerremans et al. (2018-2021)

[2] https://www.promtools.org/doku.php
[3] https://www.celonis.com/
[4] https://icpmconference.org/
[5] https://pm4py.fit.fraunhofer.de/

Nelson Hall (2021) estimates the 2021 process mining market to be 670 million USD, and forecasts it to be 4.4 billion USD by 2025, having a compounded average annual growth rate of 45%.

Moreover, Quadrant Knowledge Solutions expects the process mining market to have a compounded average annual growth rate of 75.3 % from 2020 to 2025. They estimated that the size of the process mining market would rise from 245 million USD in 2019 to more than 7 billion in 2025. Their results are illustrated in Figure 3, which shows the (expected) market size for process mining from 2019 to 2025 in million USD. (Mehar, 2020)

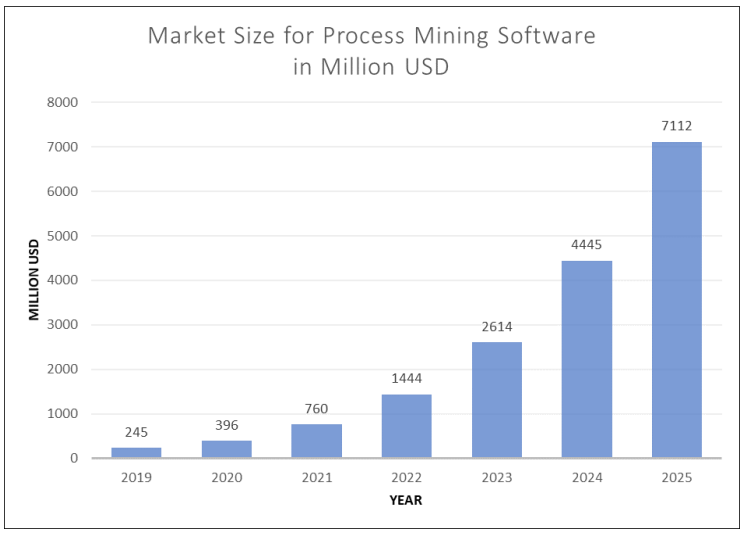


Figure 3: Market Size for Process Mining Software

Source: Own representation, based on Mehar (2020)

One indication that the market still has considerable growth ahead is the fact that the U.S. market is not yet highly developed compared to Europe's, especially Germany's and the Netherlands'. Other regions, such as Russia, South America, and South East Asia, are also expected to catch up. (Galic & Wolf, 2021)

HFS also gives a geographical assessment. Regarding the adoption of process mining software, they estimate that Europe holds a share of 52%, followed by North America with 26%. The rest of the world accounts for 22%. (Fleming & Duncan, 2020)

3.3 Terminology

This section explains important technical terms related to process mining. They are particularly relevant for the practical part of this thesis, which is addressed by RQ2 and RQ3 in Chapters 5 and 6.

Process, Activity, and Process Step

In the literature, no uniform definition exists for the term *process* (Lindsay, Downs, & Lunn, 2003). A comparatively broad definition is given by Davenport (1993), describing a process as a structured collection of **activities** that aim to generate a particular output. Other defini-

tions assign further properties to the term. For example Dumas et al. (2013) state that processes encompass not only activities, but also events and decisions. For this thesis, Davenport's definition is sufficient, as the important insight for the remaining chapters is that a process consists of activities. The terms *process* and *business process* are used interchangeably in this work.

Activities can also be referred to as **process steps**. There are publications that differentiate between process steps and activities, defining *activities* as process steps that are well-defined (van der Aalst, 2016). For the sake of simplicity, the terms *activity* and *process step* are used interchangeably in this thesis.

Process Model

Processes can be visualized with **process models**. In process mining, typical modeling techniques are Petri nets, Business Process Model and Notation (BPMN), Event-driven Process Chains and UML Activity Diagrams. These models focus on the control-flow of a process. However, process mining can also be used to investigate other perspectives, such as the organizational perspective. (van der Aalst, Adriansyah, et al., 2011)

Figure 4 shows an example BPMN process model of the order-to-cash (O2C) process of a bike-rental company. An O2C process is a process that refers to the ordering, delivery, invoicing, and payment reception of goods or services (Dumas et al., 2013).

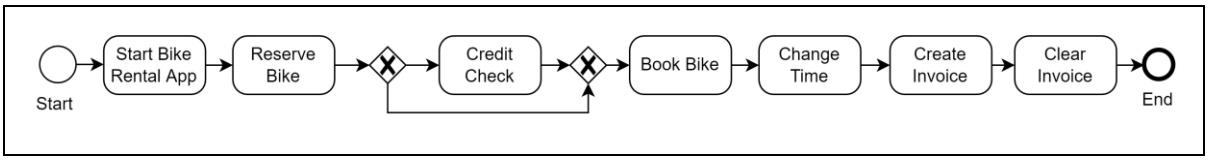


Figure 4: BPMN Model of the O2C Process of a Bike-Rental Company

Source: Own representation

Each node in the model, except the start and end node and the two XOR gateways, corresponds to a process step (i.e., an activity).

In addition to the modeling techniques mentioned previously, directly-follows graphs (DFGs) are another way to model business processes. They are widely used in process mining software solutions. A DFG is a graph in which nodes represent activities and directed edges represent "directly follows" relationships. It is a simpler modeling technique than BPMN or Petri nets, as it does not allow for concurrency. (van der Aalst, 2019)

Figure 5 shows a DFG that models the same process as the BPMN model in Figure 4. The activity "Reserve Bike" now has two outgoing edges, as it is followed by either "Credit Check" or "Book Bike," but not both in parallel.

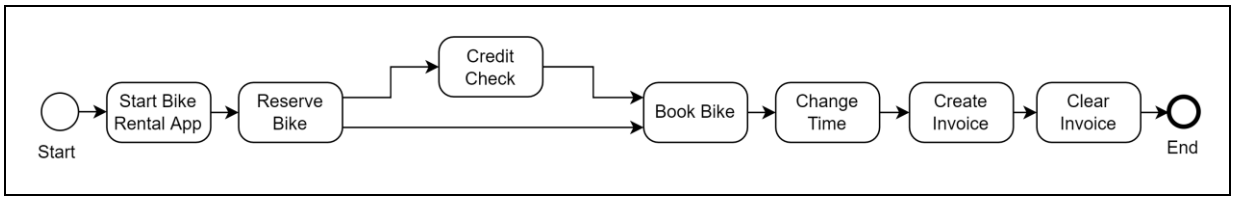


Figure 5: DFG of the O2C Process in Figure 4

Source: Own representation

Path and Process Variant

A **path** is a specific sequence of process steps in a process model. As process models usually allow for more than one path from a start node to an end node, they model more than one **process variant**. For example, activities in a process model can be optional, or the model can include a loop. Figure 4 models exactly two process variants, as the activity "Credit Check" is optional. These variants can be denoted as follows:

Variant 1: ⟨Start Bike Rental App, Reserve Bike, Credit Check, Book Bike, Change Time, Create Invoice, Clear Invoice⟩

Variant 2: ⟨Start Bike Rental App, Reserve Bike, Book Bike, Change Time, Create Invoice, Clear Invoice⟩

Process Instance and Case

A **process instance** is one execution of a specific process. For example, consider a customer who rents a bike and takes the steps of Variant 1. Another customer also orders a bike and takes the same steps in the same order as the first customer. In this scenario, two process instances have taken place, yet only one process variant occurred.

A process instance is also referred to as a **case** (van der Aalst, 2016), especially in the context of event logs, which are discussed in Section 3.4.

Happy Path

The so-called **happy path** is the path or process variant that is executed if the process is performed as desired and no exceptional situations occur (Bollen, 2010).

To-Be and As-Is Process

Processes can be classified into **To-Be** and **As-Is** processes. To-Be processes describe how processes should work, whereas As-Is processes describe how the processes actually work in reality. Modeling To-Be processes is a typical technique of BPM, and visualizing and analyzing As-Is processes is a feature of process mining. (Reinkemeyer, 2020)

3.4 Event Log and Case Table

Modern information system, such as ERP systems, generate and store vast amounts of data. Examples include the booking of a goods receipt or the creation of a customer account. In the context of process mining, this type of data is referred to as event data. Process mining tech-

nologies use this data as an input from which processes are reconstructed. Therefore, a central artifact in process mining is the so-called **event log**. (van der Aalst, 2016)

An **event** in an event log relates to an activity (i.e., process step) and is associated with a specific case (i.e., a process instance) (van der Aalst, 2016). The events of a case must be ordered (van der Aalst, 2012b). If they are not ordered, an event must have an additional attribute that allows the events to be sorted (e.g., a timestamp that is unique within a case). The ordering of events is crucial, as it allows dependencies between activities to be determined.

The notion of an event log as a single table is a simplification of reality, as such information is mostly stored across multiple tables of an information system's database (van der Aalst, Adriansyah, et al., 2011). Hence, it is necessary to extract the data in a proper way, which is a fundamental task in real-world process mining projects (van der Aalst, 2016). Additionally, in reality, further information is stored in event logs, such as the user who conducted the activity and their role (van der Aalst, 2016).

For the sake of simplicity, in this thesis it is assumed that one event log contains only the data of exactly one process and does not mix up several processes. Table 3 shows an event log of the O2C process modeled in Figure 4. Each row corresponds to one event. This event log contains three cases (001, 002, and 003). For readability, the events that refer to the same case are grouped into the same shade of blue.

Such an event table is sufficient to derive variants. Cases 001 and 003 share the same set of activities in the same order. Thus, they belong to the same variant; that is, Variant 1 introduced in Section 3.4. Case 002 belongs to Variant 2, as the optional activity "Credit Check" occurred.

Case ID	Activity	Timestamp
001	Start Bike Rental App	2021-12-11 15:07:32
001	Reserve Bike	2021-12-11 15:09:23
001	Book Bike	2021-12-11 15:10:44
001	Change Time	2021-12-11 15:12:45
001	Create Invoice	2021-12-12 17:15:50
001	Clear Invoice	2021-12-14 08:24:45
002	Start Bike Rental App	2021-12-11 13:22:26
002	Reserve Bike	2021-12-11 13:23:31
002	Credit Check	2021-12-11 13:24:11
002	Book Bike	2021-12-11 13:26:35
002	Change Time	2021-12-11 13:28:52
002	Create Invoice	2021-12-11 14:05:54
002	Clear Invoice	2021-12-15 15:21:14
003	Start Bike Rental App	2021-12-19 09:31:05
003	Reserve Bike	2021-12-19 09:32:02
003	Book Bike	2021-12-19 09:33:22
003	Change Time	2021-12-19 09:33:43
003	Create Invoice	2021-12-19 09:55:25
003	Clear Invoice	2021-12-21 09:06:14

Table 3: Example of an Event Log

Source: Own representation

Event logs can be described as a tree structure. An event log is a collection of cases, a case consists of events, and an event has any number of attributes (van der Aalst, 2016). Figure 6 illustrates this structure.

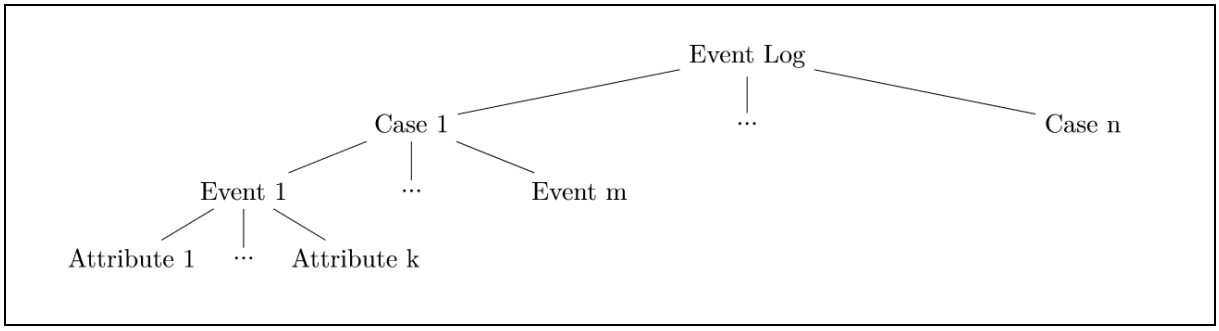


Figure 6: Tree Structure of an Event Log

Source: Own representation, based on van der Aalst (2016)

Case Table

It is not only events that can have additional attributes, but also cases. Such information could be stored in the event log, but this would be redundant, because in the event log there are typically multiple rows for each case. Thus, storing case-level information in a separate table and linking this table to the event log via the case ID is an important concept for this thesis. This concept is especially relevant in Chapters 5 and 6. Table 4 shows an example of such a case table. It contains additional information for each case from the event log above (Table 3), namely where a bike was rented and by whom.

Case ID	Location	Customer
001	Garching	Jennifer Smith
002	Munich	Alex Foles
003	Garching	John Doe

Table 4: Example of a Case Table

Source: Own representation

3.5 Discovery, Conformance, and Enhancement

Regarding the goal and application of process mining initiatives, three main types can be distinguished: process discovery, conformance checking and process enhancement (van der Aalst, Adriansyah, et al., 2011). They differ in terms of the input and output, as shown in Figure 7.

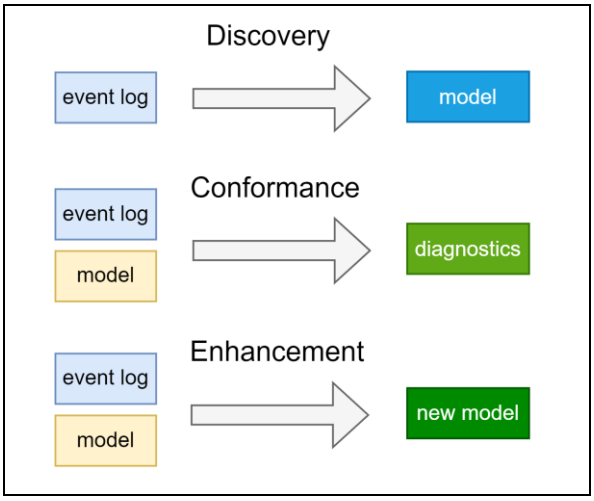


Figure 7: The Three Types of Process Mining

Source: Own representation, based on van der Aalst (2012)

Process discovery takes an event log as an input and uses it to reconstruct a process model. Examples of discovery algorithms are the alpha algorithm (van der Aalst et al., 2004), heuristic miner (Weijters et al., 2006), genetic process mining (van der Aalst, Alves de Medeiros, & Weijters, 2005), and inductive miner (Leemans, Fahland, & van der Aalst, 2013). Take the event log shown in Table 3 as an example. It consists of three cases, of which two (001 and 003) follow the same process variant (Variant 2 from Section 3.3). Case 002 is different, as it lists the additional activity “Credit Check” (Variant 1 from Section 3.3). A discovery algorithm would yield a model like the one from Figure 4, modeling the activity “Credit Check” as optional because it did not occur in all cases.

Conformance checking is used to compare an already existing model to reality. The model is checked to determine whether it matches the data of the corresponding event log. (van der Aalst, 2016)

Process enhancement aims at improving or changing an already existing process model by additionally considering event log data that belongs to the same process (van der Aalst, Adriansyah, et al., 2011). There are two types of enhancement:

- Repair: This type of enhancement changes the existing process model to better match reality; for example, by reordering activities or modelling an activity as optional (van der Aalst, 2016).
- Extension: This type of enhancement adds new information from the event log to the existing process model. For example, a process model that shows only the individual process steps and the sequence can be enriched by adding throughput times or frequencies. (van der Aalst, 2016)

In industry, companies mostly apply process discovery (Celonis, 2021; Kerremans et al., 2021). Since 2017, however, discovery tends to be used less, enhancement is being expanded, and conformance has remained at a relatively stable level (Kerremans et al., 2021). This trend is illustrated in Figure 8.

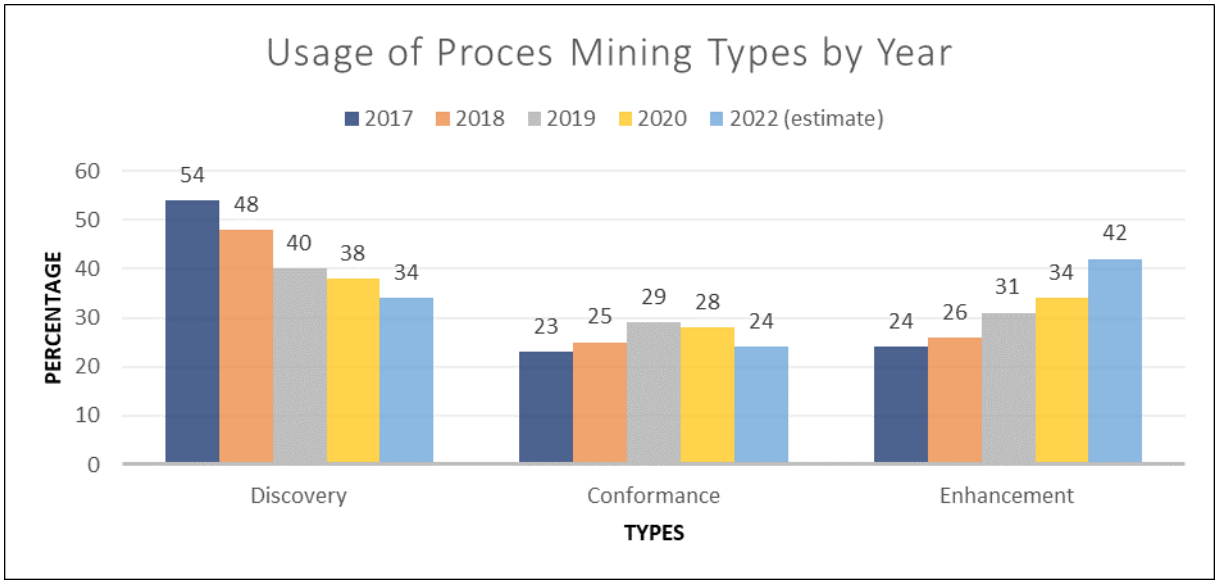


Figure 8: Usage of Process Mining Types by Year

Source: Own representation, based on Kerremans et al. (2021)

Now that the key terms and concepts of process mining have been introduced, the following chapters investigate the three interrelated research questions.

4 Process Mining in Practice

As shown in Section 3.2, process mining is an important topic in the industry and is expected to become even more relevant in the coming years. Expectations of process mining projects are manifold: Examples are process improvement, transparency, monitoring, cost reduction, automation identification and process compliance (Galic & Wolf, 2021). Given this background, this chapter investigates **RQ1**:

In which application areas and use cases is process mining utilized and what are its challenges and limitations?

For this purpose, both the scientific literature and the latest reports from research companies were consulted. Regarding the application areas of process mining, a top-down approach is taken. First, the economic sectors in which process mining is applied are outlined. Second, the functional areas involved in process mining initiatives are described. Third, use cases are investigated. After the challenges and limitations are discussed, a conclusion is given.

4.1 Application Areas and Use Cases

To obtain an overview of how process mining is applied in industry, important economic sectors are first identified. In 2021, Nelson Hall surveyed and analyzed process mining vendors, revealing that banking, financial services and insurance are the major sector for process mining initiatives, while logistics, manufacturing and healthcare are also important and growing sectors (Kong, 2021). In scientific literature, a different approach to investigate key sectors for process mining exists. For example, Dakic et al. (2018) conducted a literature review of the application of process mining and grouped the investigated articles by sectors. Their results indicate that healthcare is the most important industry, followed by information technology (IT) and finance. Thiede et al. (2018) followed a similar approach in their literature review, and yet they reached a different classification of sectors, indicating that public administration was the most represented sector. Similar to Dakic et al.'s (2018) classification, financial services and healthcare were ranked in second and third place, respectively, followed by manufacturing.

Regarding the adoption of process mining within a company, a distinction between functional areas can be made. According to a company survey by HFS, the three most important areas are finance and accounting, procurement, and customer service (Fleming & Duncan, 2020). Detailed results can be found in Figure 9.

Deloitte found similar results. They also grouped process mining into functional areas, but not in exactly the same way. In their survey, purchase and procurement, sales, and accounting and payment were the areas where process mining is applied most. Note that IT was placed fourth, in contrast to the findings of HFS. Areas where process mining was not applied as often included human resources, warehousing, and marketing. The results are shown in Figure 10. (Galic & Wolf, 2021)

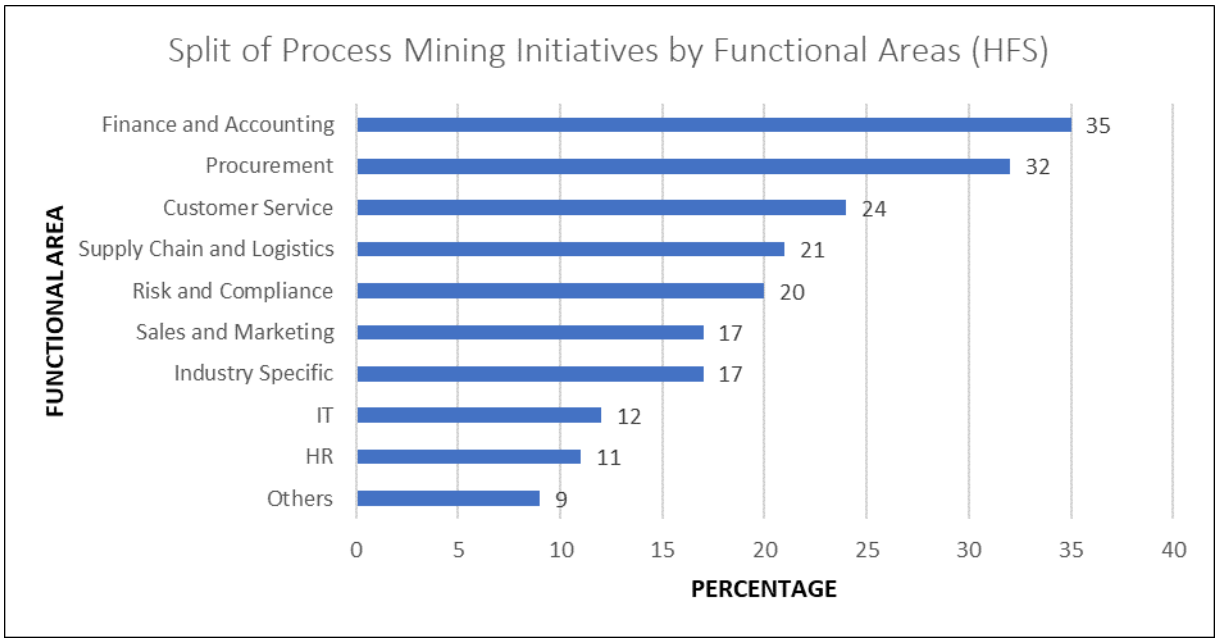


Figure 9: Process Mining Initiatives by Functional Area (HFS)

Source: Own representation, based on Fleming & Duncan (2020)

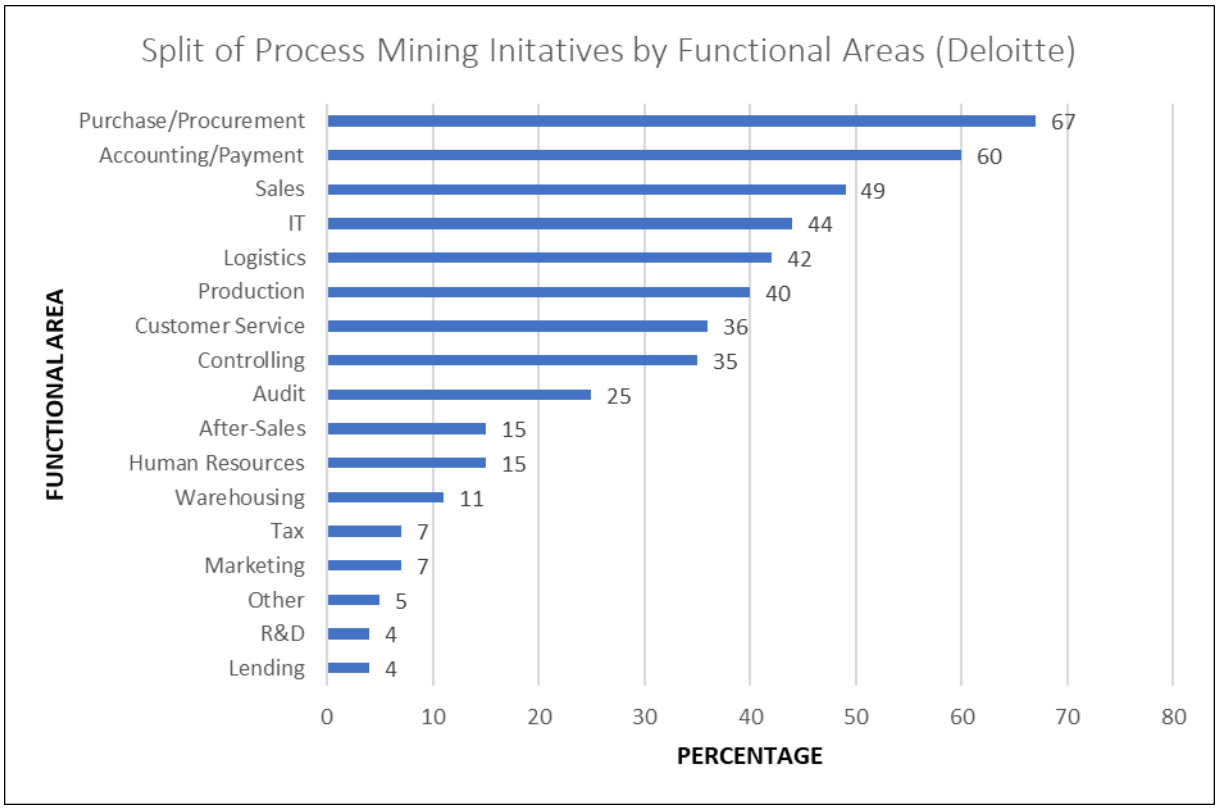


Figure 10: Process Mining Initiatives by Functional Area (Deloitte)

Source: Own representation, based on Galic & Wolf (2021)

No common classification exists of use cases of process mining initiatives. For example, in their annual market guides on process mining, Gartner defines the following five use cases:

- Business process improvement
- Auditing and compliance
- Process automation
- Digital transformation
- IT operations

According to their market survey, the main use case for process mining is process improvement. However, the percentage of process improvements measured against all other process mining initiatives is expected to decrease in the coming years. Process automation and digital transformation are expected to play a greater role. The use cases and their proportions in the respective years are shown in Figure 11. (Kerremans et al., 2021)

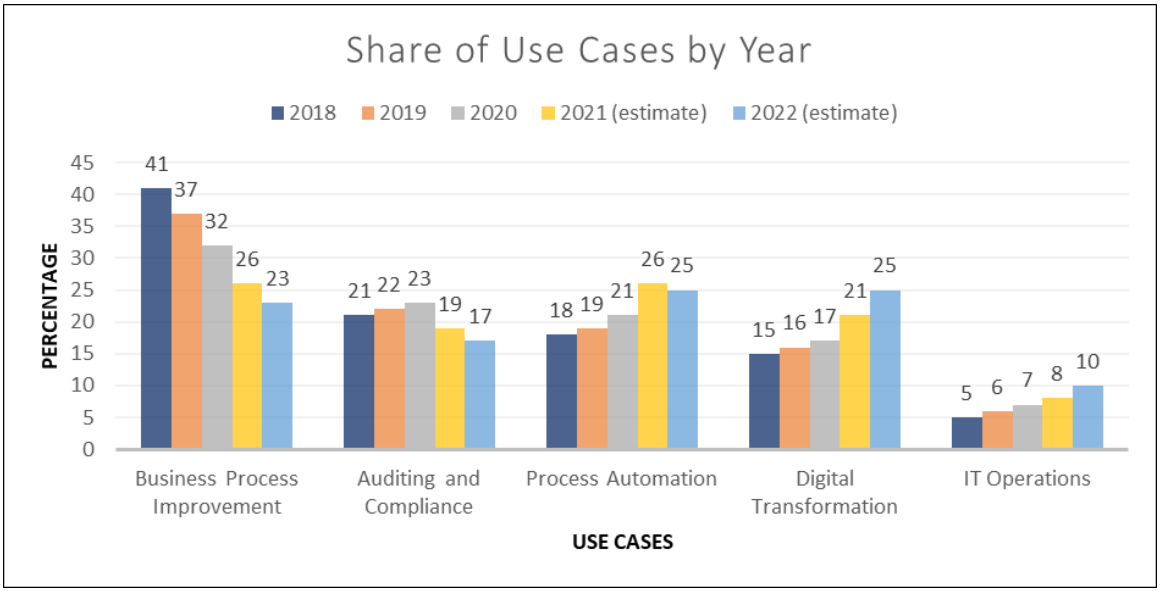


Figure 11: Share of Process Mining Use Cases by Year

Source: Own representation, based on Kerremans et al. (2021)

Quadrant Knowledge Solutions also lists process automation, digital transformation, and auditing and compliance as primary use cases. Furthermore, they view quality management, service management, enterprise software implementation, business process management, and operational excellence as vital. (Mehar, 2020)

In contrast, the PEX annual report for 2020 classifies use cases of process mining initiatives differently. They identified operational efficiency, intelligent automation, IT transformation, compliance, and customer experience as the main drivers of process mining initiatives of the surveyed companies. The results are shown in Figure 12. (Hawkins, 2020)

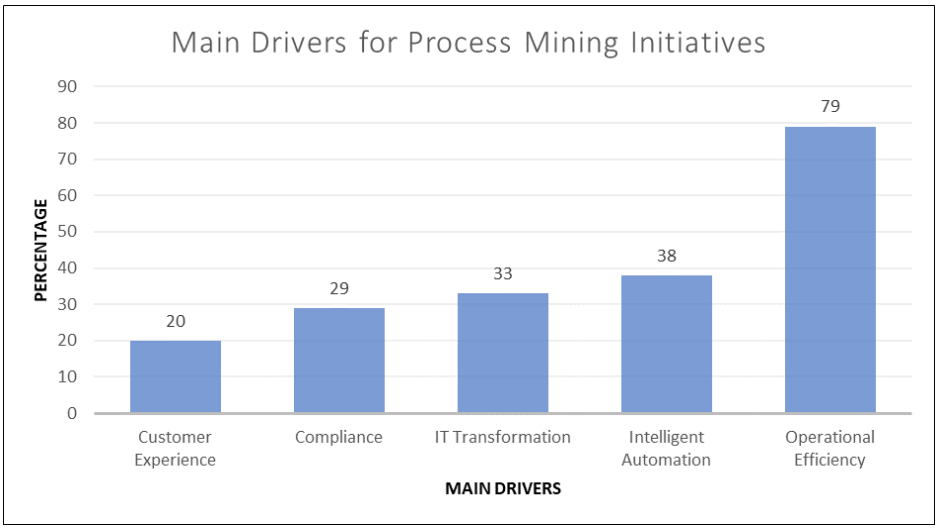


Figure 12: Main Drivers of Process Mining Initiatives

Source: Own representation, based on Hawkins (2020)

An entirely different approach was taken by Ailenei et al. (2011). They defined use cases and grouped them by the three types of process mining (discovery, conformance, and enhancement). They validated their definitions with expert interviews, which led to a classification by the role of the respective expert (e.g., researcher or consultant). They proposed the following four use cases that were relevant for all roles:

- Structure of the process
- Most frequent path in the process
- Distribution of cases over paths
- Compliance with the explicit model

Among the less important use cases were longest waiting times, work handovers, and resources per task. It is notable that the use cases from this study are more specific than those previously mentioned in the reports and surveys from research companies.

4.2 Challenges and Limitations

In scientific the literature, there is no common understanding of process mining challenges. For example, the *Process Mining Manifesto* (van der Aalst, Adriansyah, et al., 2011) lists 11 challenges. Reinkemeyer (2020) postulates seven novel challenges. Martin et al. (2021) published a set of 32 challenges as a result of a study with process mining experts from academia and industry. These diverse perceptions are inherent to the field, as over time, old challenges may be overcome, or new challenges may arise (van der Aalst, Adriansyah, et al., 2011). Thus, this section gives an overview of selected challenges and limitations. The focus is on scientific publications from recent years and current reports from research companies. The results are grouped into challenges that companies face before they start process mining initiatives, challenges that they face when applying process mining and general limitations of current process mining tools and approaches.

Challenges Before Process Mining Initiatives Are Started

The reasons companies consciously or unconsciously do not use process mining are manifold. For example, other projects are given higher priority, or the management simply does not pay attention to process mining. Budget restrictions are also a factor. (Galic & Wolf, 2021)

This is consistent with the findings from Gartner's market guides. Their reports contain interviews with Wil van der Aalst, in which he states that managers may be afraid to use process mining because it potentially reveals faulty and inefficient management. It is also outlined that many managers and consultants are simply unaware of the advantages of process mining. (Kerremans, 2019; Kerremans, Searle, Srivastava, & Iijima, 2020)

Finding a suitable provider of process mining software can also be a challenge for companies. Turner et al. (2012) give advice on which aspects to consider when looking for a process mining solution.

Before process mining can be used in companies, it must be decided to which processes it should be applied. Grisold et al. (2020) found that managers do not differentiate by process type, such as O2C or purchase-to-pay, but rather examine the processes for their characteristics. In their survey, managers stated that the most important criterion was how much data was generated in a process. Another criterion was the number of people involved in a process. It was argued that more people involved also means that more knowledge can be gained. Furthermore, managers agreed that process mining should be used only for processes that are repeated many times. (Grisold et al., 2020)

Challenges During Process Mining Initiatives

Difficulties may arise when process mining is applied in companies. It is often hard to quantify the financial outcome of process mining projects (Celonis, 2021; Galic & Wolf, 2021). Companies state that process knowledge is the most important skill for successful process mining projects (Galic & Wolf, 2021). To a certain extent, however, this can be viewed as a paradox, since process mining aims at generating process knowledge. This knowledge is typically not available beforehand. Furthermore, the lack of skills in critical areas such as analytics or data engineering is one of the biggest obstacles companies face in the application of process mining (Galic & Wolf, 2021).

In a Delphi Study, Martin et al. (2021) found that two of the three challenges that were rated as extremely relevant by the participants related to process data (poor data quality and complex data preparation). Often in process mining initiatives, 80% of the effort and time is devoted to extracting and transforming the data into the right format, as the data is usually erroneous, incomplete, or unstructured; only 20% of the time and effort is given to the actual application of process mining (e.g., process analysis) (Kerremans, 2019; Kerremans et al., 2020). One common issue related to data quality and preparation is that events are not always associated with a human-readable activity (Martin et al., 2021) . This is inline with Andres et al. (2020), who describe how the construction of the event log is especially challenging when different processes interweave with one another and multiple organizations are involved. Suriadi et al. (2017) provide a pattern based approach to prepare event logs.

Limitations

One limitation of modern process mining tools is that they heavily rely on filtered DFGs for process discovery. These graphs are limited in terms of their analysis capabilities, as they are not able to express concurrency and filtering them is prone to misleading results. It is also possible that a DFG suggests loops where in reality there are none. Although vendors are aware of these issues, they often decide to continue using DFGs, because they guarantee high performance and are easier for clients to understand than Petri nets or other alternatives. (Reinkemeyer, 2020; van der Aalst, 2019)

Another limitation in practice is that conformance checking is still not well supported and rarely found in current process mining software solutions (Reinkemeyer, 2020). If it is implemented, either it performs poorly on large data sets or simplified best-effort algorithms are carried out (Lee, Verbeek, Munoz-Gama, van der Aalst, & Sepúlveda, 2018).

Privacy issues can also be a limiting factor in process mining analyses. Analyses must respect legal restrictions, such as the European General Data Protection Regulation. In certain use cases where personalized data is collected, such as sensor data from employees' wearables, legal restrictions may limit the insights that can be gained compared to a situation when there are no regulations to adhere to. (Mannhardt, Petersen, & Oliveira, 2018)

4.3 Conclusion of Research Question 1

This chapter has shown that process mining is applied in many different sectors, functional areas, and use cases. Even if no clear classification of the criteria mentioned above exists, important areas have been identified. It has been shown that process mining still faces a variety of challenges and that current tools are limited in certain respects. In process mining research, the focus is mostly on technical aspects such as algorithms (Martin et al., 2021). Thus, the investigation of RQ1 contributes to a holistic view of process mining projects in industrial practice.

5 Continuous Simulation of Business Process Data

This chapter investigates **RQ2**:

What aspects must a continuous simulation of business process data for teaching consider?

For this purpose, a prototypical implementation of a simulation program is presented. The data of this simulation will serve as an input for process mining analyses in teaching. The focus is on the O2C process of a fictitious bike rental company.

5.1 Problem Situation and Relevance

The previous chapters showed that process mining plays an important role in the industry and offers a wide range of opportunities for companies to optimize their processes. It was also shown that the process mining market is expected to grow strongly in the coming years. Hence, it is beneficial for students to learn about process mining during their education. They acquire skills that will be increasingly in demand in the labor market in the future. Moreover, companies benefit from well-trained personnel, which is especially valuable given the shortage of skilled IT workers in Germany (Berg, 2019; Pauly & Holdampf-Wendel, 2022).

The SAP UCC, in cooperation with Celonis SE, provides teaching materials for process mining. Celonis is the world's leading provider of process mining software (Kerremans et al., 2021; Kong, 2021; Modi, Makan, & Kumar, 2021). These teaching materials include presentations and case studies that can be worked through in Celonis's software solution. For the remainder of this thesis, the name "Celonis" will refer not only to the company itself, but also to their cloud-based software solution Celonis Execution Management System.

The data set provided by the SAP UCC to conduct the case studies comprises 10,000 cases of the O2C process of the fictitious company Global Bike Share (GBS). This data set was generated once and uploaded to Celonis. However, process mining is optimally used in a continuous way to deliver the most value (van der Aalst, Adriansyah, et al., 2011). In industry, process mining with Celonis is commonly conducted in the following way. An enterprise information system, such as SAP or Salesforce, is connected to Celonis. The process data generated by the information system is then transferred to Celonis continuously (e.g., daily or hourly). The data is analyzed in Celonis, and certain actions are taken (e.g., the change of a supplier or the release of invoices). When sufficient time has elapsed, the changes can be assessed to see if they have had the desired effect. However, such an analysis is not possible with the static data set provided by the SAP UCC, as no new data is loaded into Celonis. The objective of investigating RQ2 is to identify the key aspects of a continuous simulation of business process data for teaching and to implement the simulation technically.

Simulation is the imitation of the operation of a real process or system over a period of time (Banks, 1998). In the literature, different approaches and frameworks for simulation exist (Maria, 1997). The term *continuous simulation* in the context of this thesis might be misleading at first glance. In the simulation literature, this term typically refers to systems that allow

variables to change continuously (Özgün & Barlas, 2009). In contrast, so-called discrete event simulation is applied when variables change at discrete time steps (Özgün & Barlas, 2009). Thus, this distinction refers to the properties of simulated variables. In this thesis. however, the term continuous refers to the temporal application of the simulation program itself, meaning that the program is not run at one point in time and then finishes the execution, but rather runs continuously and simulates new data at certain intervals (e.g., daily or hourly). As the simulated variables themselves are discrete (e.g., timestamps and revenues), the prototype uses principles of discrete-event simulation. It also applies principles of stochastic simulation (Asmussen & Glynn, 2007), as assigning values to certain variables is subject to randomness. This is described in more detail in Section 5.3.

Before the most important concepts of the implementation are outlined, a brief overview of the relevant components of Celonis is given.

5.2 Overview of Celonis

Celonis offers a cloud-based process mining software solution that includes various components: for example, the Action Engine, Event Collection, Machine Learning, Process Analytics, and Studio. In the context of RQ2, the relevant components are the Event Collection and Process Analytics. For RQ3, which is discussed in Chapter 6, the Action Engine is also relevant.

Event Collection

The Event Collection component allows process data from a source system to be loaded into Celonis in various ways. There are pre-built connectors for common information systems including Oracle, SAP, Salesforce, and many others. It is also possible to connect to databases (e.g., Microsoft SQL Server, PostgreSQL, MySQL) directly. This is the option that was chosen for the software prototype of this thesis. A one-time file upload (e.g., in the form of a CSV file) is also possible in the Event Collection.

When a data source is to be connected via the Event Collection, a hierarchical structure is set up. On the top level is the Data Pool. It can contain any number of Data Connections. A Data Connection contains the necessary information to connect a data source, such as the IP address of the database server and authentication information. On the next level of the hierarchy are the Data Jobs. A Data Connection can be associated with any number of Data Jobs, but each Data Job must be associated with exactly one Data Connection. A Data Job includes any number of Extractions, Transformations, and Data Model Loads. An Extraction specifies which tables that are available through the Connection are to be loaded into Celonis. A transformation is used to modify the extracted data (e.g., by filtering or changing data types). This is achieved with the help of SQL statements. A Data Model Load specifies which previously defined Data Model is to be applied to the extracted data. A Data Model relates the extracted tables from the source system to one another. For every Data Model, it is necessary to specify the event log. Note that in the context of Celonis and the prototype implementation, the event log is referred to as an **activity table**. The case table can be specified optionally. Foreign key relationships between tables can also be established in a Data Model. Figure 13 illustrates the hierarchical structure within the Event Collection.

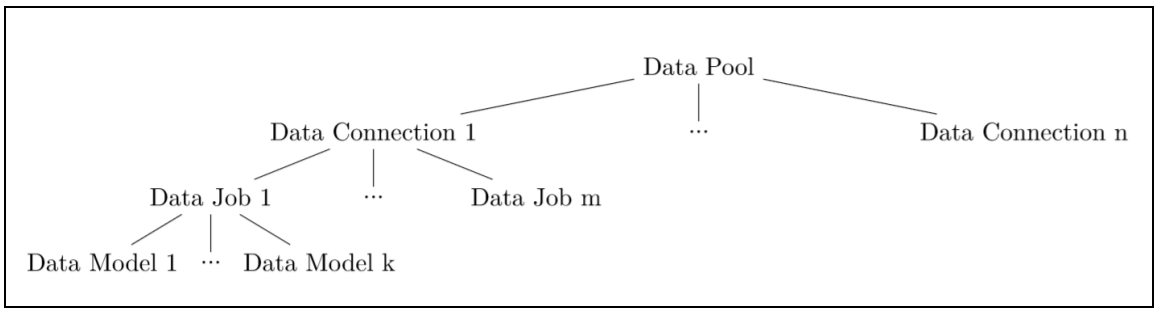


Figure 13: Hierarchical Structure Within the Event Collection

Source: Own representation

Process Analytics

Once the data to be extracted has been defined in the Event Collection, it is possible to analyze the data in the Process Analytics component. This is also the component that serves as a starting point when students work with the teaching materials provided by the SAP UCC. For the creation of analyses, Celonis provides prefabricated templates and also allows users to create templates themselves. Examples of prefabricated templates are the dashboard view, the Process Explorer, and the Variant Explorer. The dashboard view shows mainly key performance indicators and charts related to a process. The Process Explorer and the Variant Explorer provide DFGs for analyzing the process.

Figure 14 shows an example of the Process Explorer. Activities are represented as nodes in the DFG in the white area on the left. In the gray area on the right, activities and connections between them can be added or removed. The connections can be labeled with additional information, such as the median duration of activities. The Process Explorer is intended to give an overview of a process. In combination with other templates, detailed insights can be obtained.

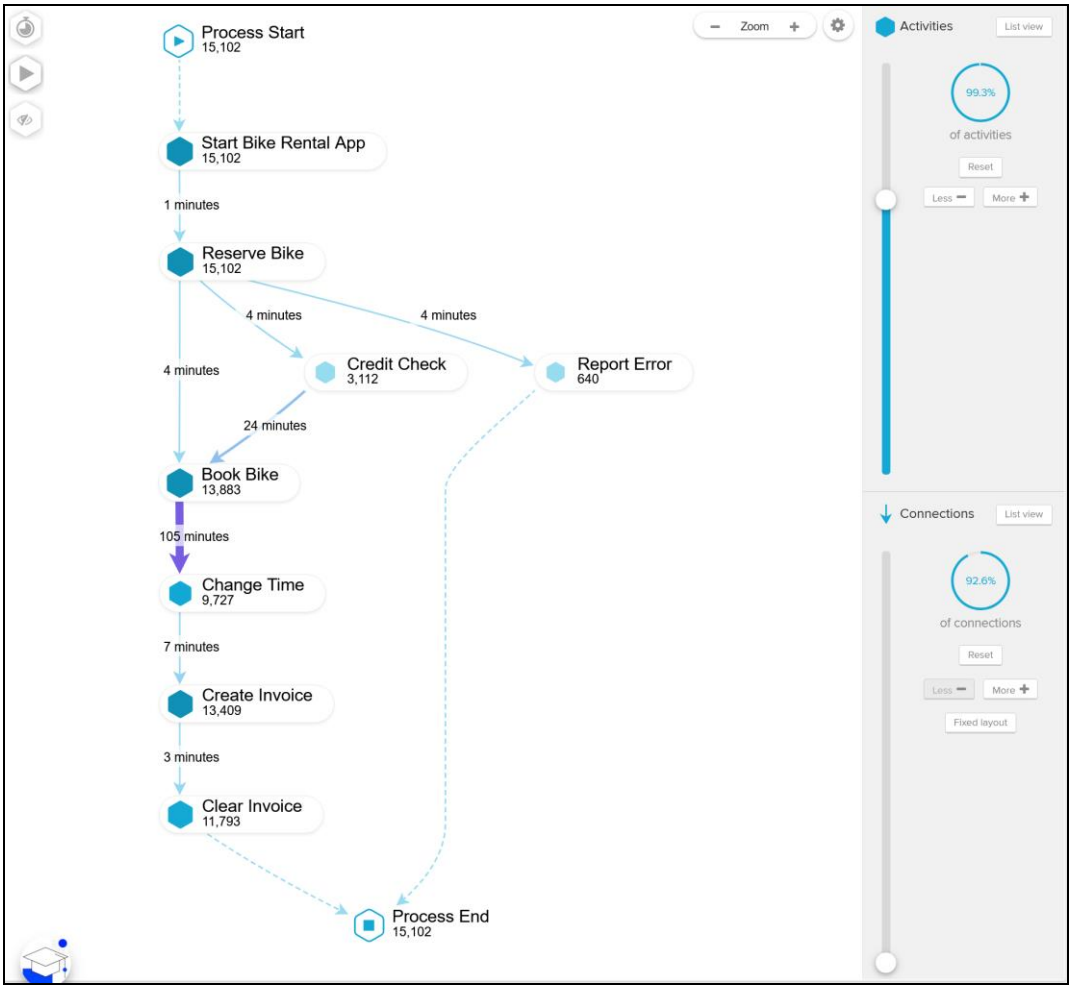


Figure 14: Example of the Process Explorer in Celonis

Source: Own project in demo version of Celonis

Now that an overview of the two relevant Celonis components has been given, details of the continuous simulation can be elaborated.

5.3 Sample Process, Procedure, and Components of the Continuous Simulation

This section gives a conceptual overview of the technical components of the prototype and the procedure for its development. First, it is necessary to elaborate on the process that has been chosen for the simulation.

For the teaching materials provided by the SAP UCC in cooperation with Celonis, the fictitious company GBS was created. GBS's business model is to rent bicycles to end customers via an app and collect a fee for this service. It was decided to focus on this O2C process as the underlying process of the prototypical simulation software for three main reasons. First, this process is easy to understand, as many students are already familiar with the concept of renting a bicycle in the real world. Second, there are already case studies and other teaching materials on this process. If the teaching materials are adopted to fit the data of the continuous simulation, less effort will be needed than if a whole new process were simulated. Third, a

data set of 10,000 cases already exists for this process. It was decided that this data set should serve as an orientation for the continuous simulation.

Due to the time and cost constraints of this work, a full implementation was not possible, but a prototype was created. In the information systems literature, there is no universal definition of the term *prototype* (Carr & Verner, 1997). A broad definition is given by Nauman and Jenkins (1982), who refer to a prototype as a system that encapsulates the main characteristics of a subsequent system.

The process of prototyping is an iterative procedure consisting of four steps: first, basic needs are identified; second, a working implementation is developed; third, the prototype is used; and finally, the prototype is revised and optimized. (Naumann & Jenkins, 1982)

In software development, different types of prototyping have emerged. For instance, throwaway prototyping involves building a prototype as soon as possible and implementing requirements that are not clearly understood. The goal is to gain a better understanding of the requirements during prototyping. When these are fully understood, the prototype is discarded, and a full system is developed. Another approach is evolutionary prototyping. In this approach, only well understood requirements are implemented at first. Then, the prototype is used in an experimental way to gain knowledge of new requirements that have not been thought of yet. Evolutionary prototyping is well suited for systems in which the core functionality is understood beforehand. (Davis, 1992)

This is the case for this thesis, as there already exists a data set of 10,000 cases that serves as an entry point for the prototype. Thus, evolutionary prototyping was used as a methodological approach to the technical implementation.

Conceptually, the development of the simulation software proceeded in the following eight steps:

1. **Analyzing the existing data set for activities, variants and attributes:** The existing data set was analyzed. It consists of 10,000 cases (i.e., process instances) of the O2C process of GBS, which are described in a case table and an activity table. In total, there are nine different activities and 39 process variants. The attributes included in the case table are:

    - the time at which the case started
    - the customer who rented the bike
    - the place where the bike was rented
    - the type of the bike
    - the duration of the rental
    - the revenue of the rental

2. **Deciding which attributes and values to adapt, omit or add:** It was decided that all the attributes above should be used for the new simulation. The possible values of the attributes were also adopted, with the following exceptions: 14 of the 39 variants and 10 of the 50 customers were adopted to simplify the implementation. It was also de-

cided not to add any new attributes at this point. The process and its 14 possible variants are represented by the BPMN model in Figure 15. One path from the start node to the end node corresponds to one process variant. The happy path is highlighted by thick arrows.

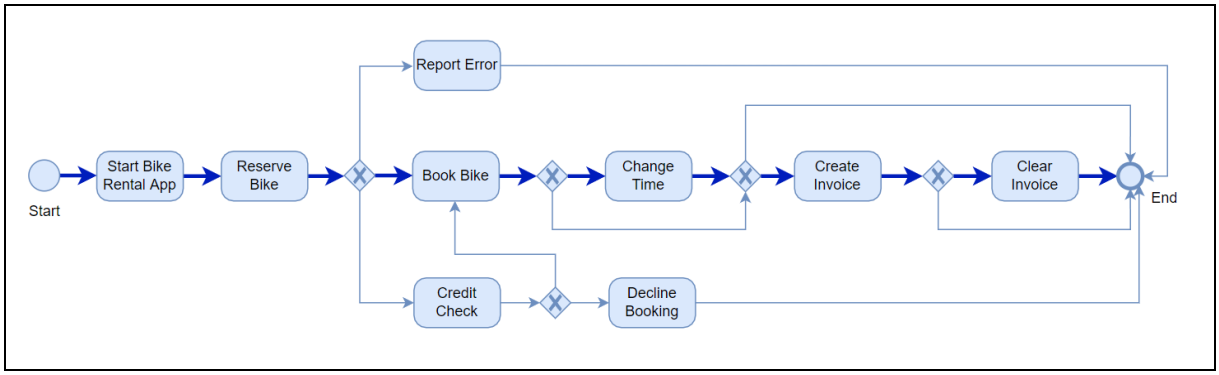


Figure 15: BPMN Model of the O2C Process

Source: Own representation

3. **Analyzing the existing data set for attribute distributions:** The values of the attributes in the existing data set are not uniformly distributed. For example, more bikes are rented in the summer than in the winter, and more rentals take place in certain locations than in others. All attributes were analyzed for their distributions, and it was decided to implement the distributions similarly in the new simulation.

4. **Developing a first version of the simulation software:** A Python script was written that is able to simulate a specified number of cases for an entire year. The result of this script was compared to the existing data set (e.g., for the similarity of distributions).

5. **Running the script continuously:** The script was modified so that not only could it simulate an entire year, but it could also be started once and simulate new cases every hour.

6. **Setting up a database server:** A Linux server was set up and MySQL Server was installed on it. A database was created that imitates GBS's information system. In reality, this could be an SAP ERP system, for example. The database holds master data such as customers, locations, and bicycle types, and it stores the case table and activity table. Appendix B contains a model of the database schema

7. **Deploying the Python script on the server:** The Python script was modified so that in each new simulation round it pulls the necessary master data from the database and simulates new cases based on this data. Before this step was introduced, this type of data was hard-coded in the script itself. Subsequently, the Python script was placed on the same server as the MySQL database, where it can run continuously.

8. **Connecting the database to Celonis:** In the Celonis Event Collection, a Data Pool with a Data Connection to the MySQL database, a Data Job and a Data Model were set up. To continuously load recently simulated data into Celonis, an hourly schedule was set up. Finally, an analysis was created in the Process Analytics component.

Figure 16 illustrates the architecture of the components of the continuous simulation. The blue boxes represent the involved components. The arrows represent data transfers from one component to another. The numbered grey boxes indicate the order in which the data transfers happen. In every new simulation run, the Python script pulls master data from the database, such as the existing customers or locations. Then, based on this data, new cases are simulated and written to the activity table and case table on the database. Celonis then extracts the entire data from the database. These three steps are executed every hour. In this way, a continuous simulation of business process data and the possibility to analyze it in Celonis is achieved.

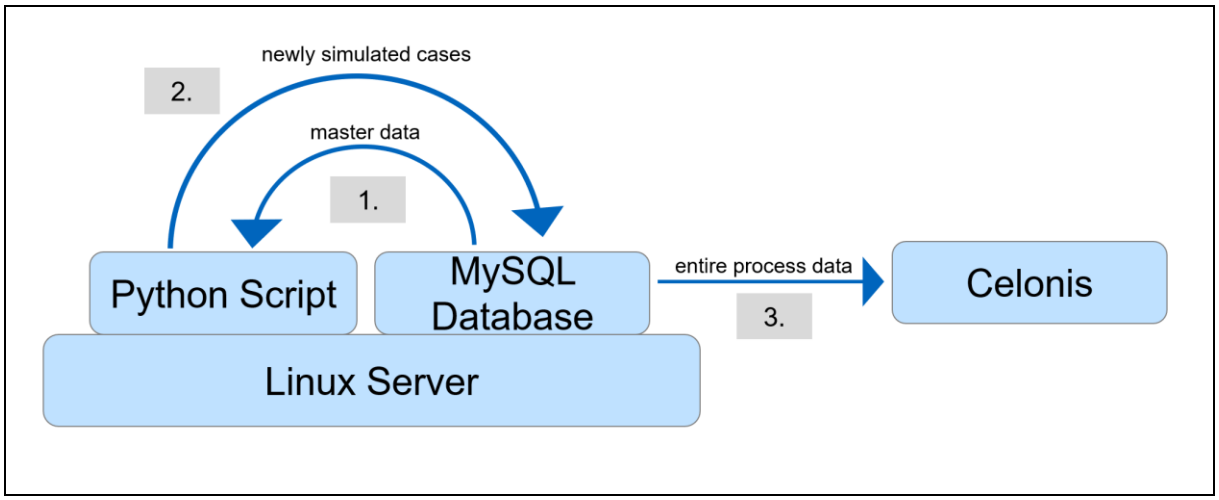


Figure 16: Components of the Continuous Simulation Prototype

Source: Own representation

5.4 Challenges of Simulating Continuous Process Data

In this section, an overview is given of the main challenges that arose during the implementation of the simulation software. The concepts and challenges presented here concern basic issues in the continuous simulation of business process data and are not restricted to a certain process.

Challenge 1: Defining the Interval Between the Simulation Runs

From the research question it is clear that the process data should be simulated continuously. There are different ways of achieving this; for example, daily, hourly, or even every second. With regard to RQ3, which is about building a feedback mechanism between actions in Celonis and the simulation, the hourly simulation was chosen. An hourly simulation is more beneficial for teaching than a daily simulation, because the effects of actions in Celonis can be analyzed by students on the same day. Moreover, it does not put as much load on the database and Celonis as a more tightly timed simulation.

Challenge 2: Determining How Many Cases Should Be Simulated in Each Simulation Run

Now that the interval between the runs is defined, the first decision in each run is how many cases are to be simulated in the run. The easiest way would be to define a fixed value. To create a more realistic model, the number of simulated cases follows certain probability distribu-

tion, depending on the current month and hour. This models the fact that more bicycles are rented in the summer months than in the winter months and that more rentals take place in the morning hours than at night. The target number for an entire year was defined as 10,000 cases, which corresponds to an average of about 1.14 cases per hour. Thus, in many runs, no cases are simulated at all (e.g., at midnight in a winter month), and several cases are simulated in the morning of a summer month.

Challenge 3: Assigning Probability Distributions to Attributes of the Case Table

Depending on the application domain, a case is associated with several attributes. For this thesis, these attributes are described in Section 5.3. For every attribute, a certain probability distribution for its values must be chosen. The simplest method is to assign a uniform distribution to each attribute (i.e., every value of an attribute has the same probability of being simulated as the others). However, this method is not an adequate representation of reality. Therefore, distributions were derived from the existing data set. For example, in most of the cases, a certain type of bicycle is rented, whereas another type is rarely rented. If an attribute has only a few possible values, a probability can be assigned to each value manually. In the simulation for this thesis, this was done for the month and hour of a timestamp and for types of bicycles, locations, and variants. Customers and the remaining elements of a timestamp (day, minute, and second) were uniformly distributed to reduce complexity. For the attribute that has too many possible values to assign probabilities manually, namely the duration of a rental, another approach was taken. The values were grouped into intervals, and each interval was assigned a probability manually. Within each interval, the values were distributed uniformly.

Challenge 4: Determining Activity Durations

The problem of simulating activity durations is similar to the problem of simulating rental durations described in Challenge 3. There are too many possible values to assign probabilities manually. Thus, the values were also grouped into intervals, and each interval was assigned a probability manually. Again, within each interval, the values were distributed uniformly.

Challenge 5: Dependencies Between Attributes of the Case Table

With regard to the relationships of attributes within a case, various dependencies are conceivable. For example, a certain customer might always rent the same type of bicycle, or the customer might only rent bicycles on workdays. These types of dependencies can be designed to be arbitrarily complex. Nevertheless, they are not taken into account in this simulation, because they significantly increase the complexity of the implementation and have only little influence on the quality of the teaching.

5.5 Conclusion of Research Question 2

This chapter investigated RQ2 (What aspects must a continuous simulation of business process data for teaching consider?). The design science methodology (Hevner et al., 2004) was adhered to in developing a software artifact. This artifact contributes to a more realistic teaching experience, as process data is continuously simulated and transferred to Celonis, where

the data can be analyzed using process mining techniques. It was stated that it is necessary to simulate both an activity table and a case table. For the activity table, it is necessary to define how many and which variants should be simulated. For the case table, it is necessary to define the attributes and their corresponding values and probability distributions that should be part of the later analysis in process mining software such as Celonis. The main challenges lie in the area of conflict between realistic representation and the complexity of implementation. The more realistic the simulation, the more complex the implementation. For teaching, a suitable compromise must be found (e.g., by defining attributes for which certain probability distributions are useful and others where an even distribution is sufficient). These considerations are not restricted to the specific process that was implemented, but hold for other types of business processes.

6 Feedback Mechanism Between Celonis and the Simulation

This chapter investigates **RQ3**:

> *What influence do operational changes have on continuously simulated data and how is a feedback mechanism to be translated technically?*

This research question builds directly upon RQ2 from the previous chapter. Thus, the prototypical implementation presented in Chapter 5 will be modified and extended.

6.1 Problem Situation and Relevance

Chapter 5 investigated which aspects must be considered when business process data is to be simulated continuously. It was also shown how such a simulation can be implemented technically. In this chapter, the focus is not on generating process data, but on changes to the simulation from within Celonis. In industry, process mining software is used not only for analysis of processes, but also for the execution of actions (e.g., bookings or transactions) in a connected information system. Therefore, is valuable to also facilitate this procedure in teaching with simulated data. This enables students not only to analyze processes, but also to modify processes and test the effectiveness of these modifications. The prototype developed for this research question thus contributes to an improvement in the quality of teaching and makes it more realistic.

6.2 Target Architecture

In systems theory, the term *feedback* describes a concept in which two or more systems are interconnected in such a way that they influence each other (Åström & Murray, 2006). Figure 17 shows the technical components from Figure 16 extended by the classification into two systems. System 1 influences System 2, but not vice versa, which is why there is no feedback.

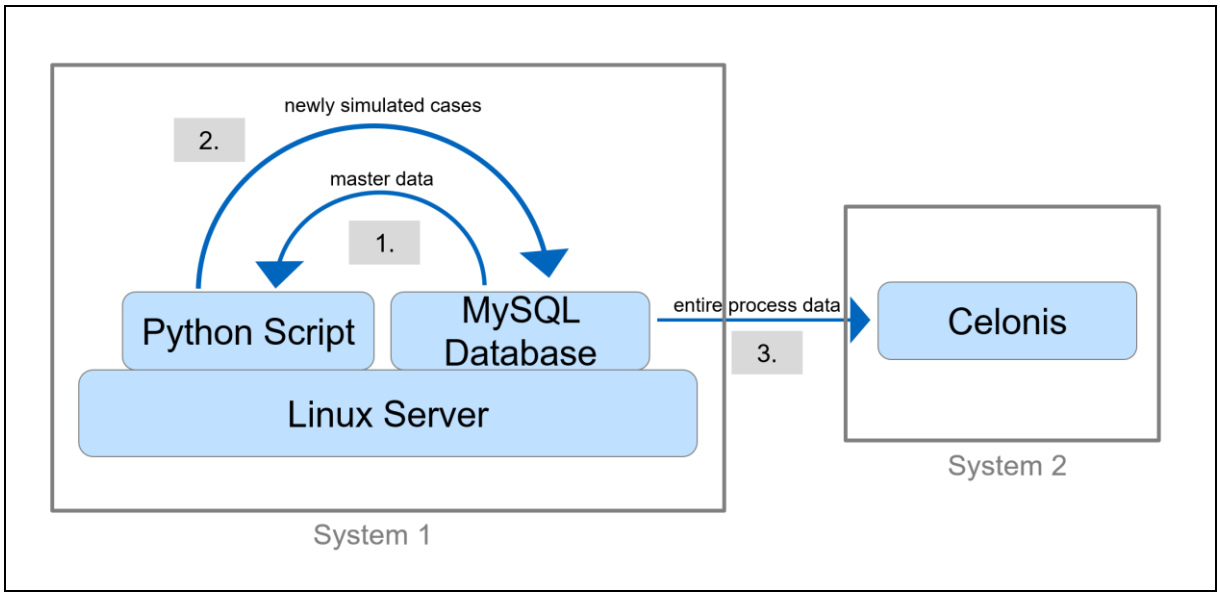


Figure 17: Classification of the Technical Components Into Two Systems

Source: Own representation

The goal of this Chapter is to establish a connection from System 2 to System 1 and investigate which influence this has on the technical implementation of System 1 and its output.

While investigating RQ3, the question arose of how changes can be made in the simulation software itself (i.e., in the Python script) without manually terminating, modifying, and then restarting the program. As shown in Figure 17, there is no direct connection between the script and Celonis, but rather an indirect one via the database. Such a direct connection to a Python script cannot be set up in Celonis. Therefore, if changes in Celonis are to affect the simulation, the simulated data must be made dependent on certain data in the database, and this data must be queried before each simulation run. For example, a flag could be added to the customer table, and if it is set, the simulation will behave differently than if it is not set. This target architecture is depicted in Figure 18. Compared to Figure 17, a data transfer from Celonis to the database has been added as a fourth step. If actions are performed in Celonis, this fourth step terminates the simulation run and a new run starts from the first step, pulling the modified data.

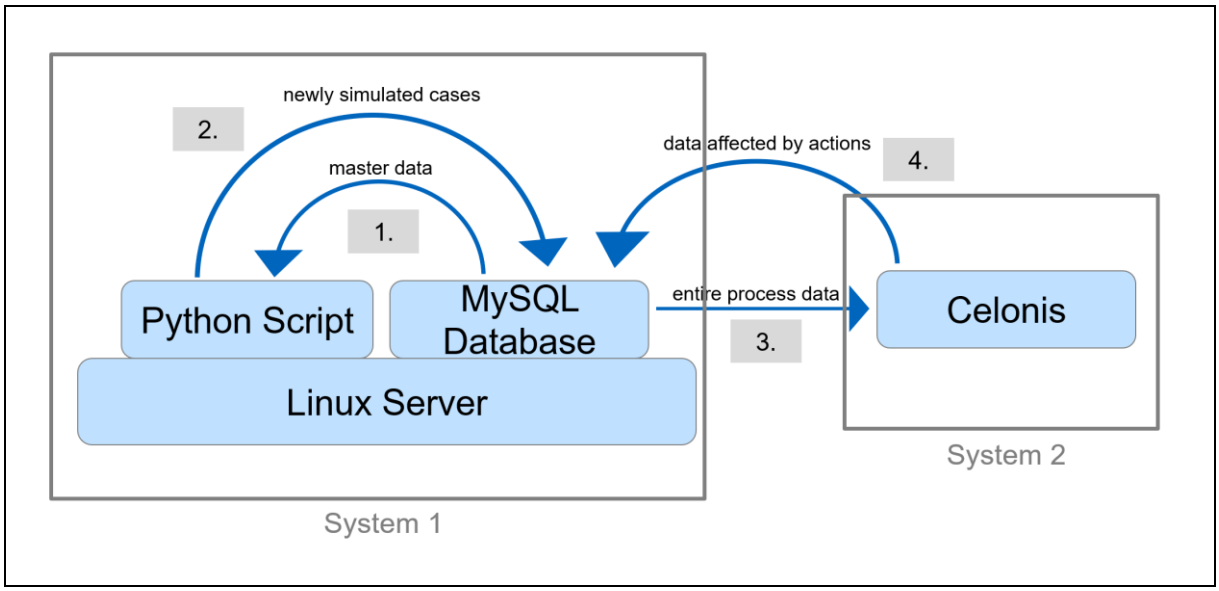


Figure 18: Feedback Mechanism Between the Two Systems

Source: Own representation

In this way, a feedback mechanism is established between the two systems. In systems theory, such a construct is also referred to as a *closed-loop* (Åström & Murray, 2006).

Next, a brief introduction to Celonis's Action Engine is necessary, as this is the component that allows actions to be performed that affect the future simulation.

6.3 The Celonis Action Engine

In Section 5.2, the Event Collection and the Process Analytics component of Celonis were introduced. For RQ3, the Action Engine plays a central role. Within the Action Engine, so-called signals can be set up. As soon as cases are loaded in the Event Collection that trigger these signals, actions that have been defined previously are suggested to the user. Examples of such actions include performing a transaction in a connected information system or sending

an email to a supplier. A signal can have any number of associated actions. Celonis offers both pre-built actions and the opportunity to create custom actions.

To illustrate the relationship between signals, actions, and cases, consider the following scenario: In Celonis, a signal is set up that is triggered when a case is loaded into the Event Collection that contains the activity “Create Invoice” but not the activity “Clear Invoice”. The desired action associated with this signal is to send the corresponding customer an email reminder, asking them to pay the invoice.

Now that a general understanding of the Action Engine has been reached, the specific signals and actions that were implemented for the feedback mechanism can be explained.

6.4 Defining Sample Signals and Actions in Celonis

For the implementation of actions, a variety of options are available in Celonis. A distinction can be made between two types of actions: those that actually change the underlying data and those that do not. Examples of the former include the execution of a transaction in a connected information system (e.g., releasing an invoice); examples of the latter include sending emails, displaying information, or opening links and analyses. Since RQ3 investigates the influence of actions on the simulation, it was decided to select actions that belong to the first group. Since the prototype is to be used in teaching, signals and actions were set up that are easily understandable. Moreover, it was decided to limit the number of signals to three, because this was rated as sufficient to achieve learning effects. The three signals and their associated actions are shown in Table 5. The explicit effect on the simulation is described for every signal.

The criteria that a case must meet to trigger a signal can be defined in Celonis using the Process Query Language (PQL) invented by Celonis. It is a language inspired by SQL and specifically built for querying process data (Klenk, 2021). Its ability to filter on aggregated process data is useful for setting up a signal. The following PQL statement exemplifies this functionality. This statement was used to set up Signal 2 from Table 5 (i.e., finding the customers that have three or more unpaid invoices).

```
FILTER PU_COUNT ("customer",
                 "case_table"."case_id",
                 "case_table"."variant_id" IN (2,5,6,8)) >= 3
```

The PU_COUNT function inside the filter statement is an aggregate function. For every customer, it counts the number of their cases that have the variant ID’s 2, 5, 6, or 8. These are the variants in which an invoice is created but not cleared. The filter then selects only those customers who have three or more such cases. The advantage of PQL over SQL in this example is the amount of code needed to achieve the same goal. Here is the corresponding SQL statement to achieve the same result:

```
SELECT c.*
FROM case_table ct
LEFT JOIN customer c ON ct.customer_id = c.customer_id
WHERE ct.variant_id in (2,5,6,8)
GROUP BY c.customer_id
HAVING COUNT(ct.case_id) >= 3;
```

Signal ID	Signal Name	Signal Description	Associated Action(s)	Effect on the Simulation
1	20% Discount for Next Rental	Every customer that rented a certain type of bike at least 120 minutes receives 20% discount on the next rental.	(1) inform customer via email (2) set discount on next rental in database	The revenue of the next rental will be 20% less than without the action.
2	Block Customer	Every customer that has three or more unpaid invoices will be blocked from future rentals.	(1) inform customer via email (2) set blocking flag on customer in database	The customer will not be simulated any more.
3	Set Credit Check on Location	Every location for which 30 or more cases with unpaid invoices exist will require a mandatory credit check for every future case.	(1) set credit check flag on location in database	Only variants that include the activity “Credit Check” will be simulated for this location.

Table 5: Signals and Actions Established in Celonis

Source: Own representation

Now that it has been explained which signals and actions were created in Celonis, the modifications that they require in the simulation can be discussed.

6.5 Influences of the Actions on the Simulation and Technical Realization

For the implementation of the first signal, a new table was created in the database that assigns a discount rate to each customer. The attribute in the case table that is affected by this signal is the revenue of the rental. It is calculated as before, when the signal did not exist, and additionally multiplied by (1 − *discount rate*). Thus, in every simulation run, the discount rates from the new table must be queried by the Python script.

For the second signal, the customer table in the database was extended by a column indicating whether the customer is blocked. At the start of each new simulation run, the script checks

which customers are blocked according to the flag and considers only those that are not blocked.

The implementation of the third signal works similarly to the second one. The location table in the database was extended by a column that indicates whether the location is subject to a credit check. Before the script simulates the location of a case, it checks if the flag is set in this location, and if so, it simulates a variant that contains the activity "Credit Check." The third signal also introduced an additional dependency in the simulation. The variant that is to be simulated is dependent on the selected location. This means that, in the script, the location must be simulated first, and only then can the variants be determined. In the previous version of the script, which was discussed in Chapter 5, these two attributes were independent of each other.

For the second and third signal, it was necessary to extend the case table with the attribute variant id. This was previously not an explicit part of the case table, but the variants were computed by Celonis implicitly during the data import. Since for both signals it is crucial to distinguish between variants in which the customer pays their invoice and variants in which they do not (i.e., if the activity "Create Invoice" is present but not the subsequent activity "Clear Invoice"), it is useful for this information to be included in the case table to enable filtering by means of PQL statements in the Action Engine. If this information was not included, it would be necessary to compute it via PQL, which would increase complexity significantly.

Appendix C contains a model of the modified database schema, which includes the above mentioned adjustments.

Generalizing these insights yields the following influences of operational changes from within a process mining tool on a continuous simulation of business process data:

- Operational changes may require new tables or columns in an existing database schema. In particular, the introduction of flags (i.e., Boolean values) often appears to be a simple yet useful method.

- Operational changes may introduce new dependencies of function parameters or change existing ones. This may also lead to a reordering of the execution of functions. When two functions are not dependent on each other, the order of execution plays no role. As soon as one function is dependent on another's value, the execution must occur in the right order.

- Operational changes may change certain calculations (e.g., because new attributes such as a discount must be considered).

- Any data that underlies operational changes and serves as an input for the simulation must be queried by the simulation script. Ideally, this happens in every simulation run, such that the simulation is based on the latest data.

- Operational changes may require making implicit information explicit. For example, introducing the variant of a process instance explicitly in a new table column can lead to a significantly simpler solution. This is especially the case if this information has to be used in the process mining system (e.g., for filtering).

6.6 Conclusion of Research Question 3

This chapter investigated RQ3 (What influence do operational changes have on continuously simulated data and how is a feedback mechanism to be translated technically?). It was shown that such a feedback mechanism, implemented using the Celonis Action Engine, extends the interaction of the simulation's technical components by one step, connecting two systems via a closed-loop. An indirect connection was established between the component containing the simulation logic and the process mining software, namely Celonis, via a database. Technically, this was implemented by new tables, columns, flags, and dependencies of simulation functions. The specific modifications of the implementation of the prototype were generalized, contributing to a better understanding of the effects that operational changes can have on a continuous simulation of business process data.

7 Limitations, Conclusion, and Outlook

7.1 Limitations of This Thesis

Three academic databases were searched to conduct the literature review for RQ1. Other databases may contain further relevant articles. Due to the large number of search results, the sources used in this thesis were selected based on their title and abstract. The result of this procedure depends on the assessment of a document's relevance and may exclude more relevant documents and include less relevant ones. Furthermore, other search terms may also yield relevant articles. However, due to the limited amount of time for this thesis, the approach used for the literature search is justified.

The reports from research companies, which include surveys and estimations, investigate the adoption of process mining in the industry. However, surveys can be biased (e.g., by a non-representative selection of respondent companies), and estimates can be error-prone (e.g., due to a wrong choice or weighting of parameters). Even if surveys and estimates from different companies do not match exactly, however, they do show tendencies. These tendencies should be taken seriously, as reports from different companies are created independently of each other, making the data more robust.

The simulation software developed in this thesis is also subject to limitations. The ISO/IEC 25010 standard (ISO/IEC, 2011) lists eight characteristics of software quality. These are functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability and portability. Some of these characteristics could not be fully realized due to time and costs constraints. For example, the simulation script queries the entire database in every new simulation run. This may cause performance issues for large datasets. However, the core functionality has been implemented in a satisfactory manner, allowing for future enhancements.

Only data of an O2C process was simulated. Simulation of other processes would increase the variety for students and thus lead to a more comprehensive education. Nevertheless, key concepts and takeaways can be taught with a single process, as it simplifies the learning experience compared to a simulation of multiple processes.

The simulated data was stored in a database that was created specifically for this thesis. It would have been desirable to feed the data directly into an existing SAP ERP system of the SAP UCC and to connect this with Celonis. In this way, the built-in functionality provided by Celonis for SAP systems (e.g., in the Action Engine) could have been taken advantage of. However, due to time constraints and limited knowledge of the internal table structures of SAP, this approach was not taken. Nevertheless, the concepts used for the continuous simulation can be adapted in the future to build such a connection to SAP.

7.2 Conclusion

This thesis sets out to enhance teaching in the field of process mining. Process mining sits between traditional model-based process analysis and data-centric analysis techniques such as data mining and machine learning (van der Aalst, 2016). In cooperation with Celonis SE, the

SAP UCC offers process mining training material for professors, students, and pupils. This material comprises not only tutorial documents, but also event data. Since no real data from the industry can be used in teaching, currently, a static data set of 10,000 cases is available for learners to practice with. However, this data is not mutable (i.e., it does not change accordingly when an action is performed, such as the release of invoices). This thesis improves the learning experience by making it more realistic and relevant to practical use. This improvement was achieved in three steps.

First, a theoretical foundation of process mining was established by conducting a structured literature review. The review not only refers to process mining as an academic discipline, but also covers its application and importance in business practice. Second, a software prototype was developed that adheres to the design science principle. It transitions from statically to continuously simulated business process data. As an example process, the O2C process of the fictitious company GBS was simulated. The main aspects and challenges that arose during the implementation were outlined and generalized to fit any simulation of business process data. Finally, the software prototype was extended to include a feedback mechanism between the simulated data and user actions in the process mining software Celonis. This feedback mechanism allows users to evaluate their process modifications. The influences of the feedback loop on the simulation software were demonstrated and generalized to fit a variety of business processes. In conclusion, the objective of this work has been fulfilled as it contributes to more realistic teaching of concepts and practices of process mining.

7.3 Outlook

Process mining is a promising technology that can be applied in many industries. Especially against the backdrop of trends such as ongoing digital transformation, process mining may become even more important in the future. This thesis can serve as a starting point for improvements in the education of the next generation of process mining experts. The transition from statically to continuously simulated process data including a feedback loop is a first step toward improvement, but it is not the end of the road. One further step would be the creation of an appropriate teaching curriculum; for example, a case study that guides learners to discover how actions in Celonis affect the simulated data. Another option would be to customize the software prototype to embed the simulated data into a specific target system such as SAP. In this way, learners would not be restricted to the actions they can take in Celonis; they could also make changes directly in the system that stores the process data. For the implementation of the feedback mechanism, a workaround was established that uses a an email account that was set up specifically for this thesis to connect Celonis to the MySQL database. This approach was chosen because at the time of implementation, there was no knowledge of any functionality directly available in Celonis that stores the desired data in the database. However, the Celonis Execution Management System is continuously being enhanced. In addition to the Action Engine, it offers the Studio component, in which Action Flows can be set up. These Action Flows enable users to take more complex actions and offer a greater automation potential than the action engine. Thus, integrating Action Flows into the simulation could make the above mentioned workaround obsolete and lead to better teaching.

It is important that teaching in the field of process mining progresses, as the market is expected to grow at a high rate (Kerremans et al., 2021). Thus, industry needs a generation of well-educated employees to meet the future demands of process mining.

Bibliography

Ailenei, I., Rozinat, A., Eckert, A., & van der Aalst, W. (2011). *Definition and validation of process mining use cases*. Paper presented at the International Conference on Business Process Management.

AIS. (2011). Senior Scholars' Basket of Journals. Retrieved 25.02.2022 from https://aisnet.org/page/SeniorScholarBasket

Andrews, R., Wynn, M., Vallmuur, K., Ter Hofstede, A., & Bosley, E. (2020). A comparative process mining analysis of road trauma patient pathways. *International journal of environmental research and public health, 17*(10), 3426.

Asmussen, S., & Glynn, P. (2007). *Stochastic simulation: algorithms and analysis* (Vol. 57): Springer Science & Business Media.

Åström, K., & Murray, R. (2006). Feedback Systems: An Introduction for Scientists and Engineers.

Banks, J. (1998). *Handbook of simulation: principles, methodology, advances, applications, and practice*: John Wiley & Sons.

Berg, A. (2019). *Der Arbeitsmarkt für IT-Fachkräfte*. Bitkom.

Bollen, P. (2010). *BPMN: A meta model for the Happy Path*: Citeseer.

Carr, M., & Verner, J. (1997). Prototyping and software development approaches. *Department of Information Systems, City University of Hong Kong, Hong Kong*, 319-338.

Celonis. (2021). *The State of Process Excellence*. Celonis.

Cooper, H. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in society, 1*(1), 104-126.

Dakic, D., Stefanovic, D., Cosic, I., Lolic, T., & Medojevic, M. (2018). Business Process Mining Application: a Literature Review. *Annals of DAAAM & Proceedings, 29*.

Daniels, J. (2022). *Trends In Process Improvement And Data Execution*. Forrester Research.

Davenport, T. (1993). *Process innovation: reengineering work through information technology*: Harvard Business Press.

Davis, A. (1992). Operational prototyping: A new development approach. *IEEE software, 9*(5), 70-78.

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). *Fundamentals of business process management* (Vol. 1): Springer.

Fleming, R., & Duncan, S. (2020). *HFS Top 10 Process Intelligence Products*. HFS Research.

Galic, G., & Wolf, M. (2021). *Global Process Mining Survey 2021*. Deloitte.

Grisold, T., Mendling, J., Otto, M., & vom Brocke, J. (2020). Adoption, use and management of process mining in practice. *Business process management journal*.

Handelsblatt Research Institute. (2020). *Die Rolle der internen Kommunikation bei der Transformation*. Handelsblatt Research Institute.

Hawkins, I. (2020). *The PEX Report 2020: Global State of Process Excellence*. Process Excellence Network.

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Holzki, L. (2021). Bewertung übersteigt zehn Milliarden Dollar: Münchener Start-up Celonis ist erstes deutsches Decacorn. Retrieved 25.02.2022 from https://www.handelsblatt.com/technik/it-internet/mega-finanzierungsrunde-bewertung-uebersteigt-zehn-milliarden-dollar-muenchener-start-up-celonis-ist-erstes-deutsches-decacorn/27247478.html?ticket=ST-3979771-zVbqUaSsEPheHPKE9dsg-cas01.example.org

Höpner, A., & Kerkmann, C. (2019). „Technologie mit einzigartigem Potenzial“: Einhorn Celonis erhält Deutschen Zukunftspreis. Retrieved 25.02.2022 from https://www.handelsblatt.com/technik/it-internet/process-mining-spezialist-technologie-mit-einzigartigem-potenzial-einhorn-celonis-erhaelt-deutschen-zukunftspreis/25279612.html?ticket=ST-4133878-53FJ1XqK6TP1i2oDh0fp-ap3

ISO/IEC. (2011). ISO/IEC 25010: Systems and software engineering-Systems and software Quality Requirements and Evaluation (SQuaRE)-System and software quality models. In: ISO/IEC.

Kerremans, M. (2018). *Market Guide for Process Mining*. Gartner.

Kerremans, M. (2019). *Market Guide for Process Mining*. Gartner.

Kerremans, M., Searle, S., Srivastava, T., & Iijima, K. (2020). *Market Guide for Process Mining*. Gartner.

Kerremans, M., Srivastava, T., & Choudhary, F. (2021). *Market Guide for Process Mining*. Gartner.

Klenk, M. (2021). The Celonis Process Query Language (PQL) Turned 5 in February. Retrieved 25.02.2022 from https://www.celonis.com/blog/the-celonis-process-query-language-turns-5/

Kong, B. (2021). *Neat Evaluation for Celonis: Process Discovery & Mining*. Nelson Hall.

Koplowitz, R., Mines, C., Vizgaitis, A., & Reese, A. (2019). *Process Mining: Your Compass For Digital Transformation*. Forrester Research.

Lee, W., Verbeek, H., Munoz-Gama, J., van der Aalst, W., & Sepúlveda, M. (2018). Recomposing conformance: Closing the circle on decomposed alignment-based conformance checking in process mining. *Information Sciences, 466*, 55-91.

Leemans, S., Fahland, D., & van der Aalst, W. (2013). *Discovering block-structured process models from event logs-a constructive approach.* Paper presented at the International conference on applications and theory of Petri nets and concurrency.

Lindsay, A., Downs, D., & Lunn, K. (2003). Business processes—attempts to find a definition. *Information and software technology, 45*(15), 1015-1019.

Mannhardt, F., Petersen, S., & Oliveira, M. (2018). *Privacy challenges for process mining in human-centered industrial environments.* Paper presented at the 2018 14th International Conference on Intelligent Environments (IE).

Maria, A. (1997). *Introduction to modeling and simulation.* Paper presented at the Proceedings of the 29th conference on Winter simulation.

Martin, N., Fischer, D., Kerpedzhiev, G., Goel, K., Leemans, S., Röglinger, M., . . . Wynn, M. (2021). Opportunities and challenges for process mining in organizations: results of a Delphi study. *Business & Information Systems Engineering, 63*(5), 511-527.

Mehar, R. (2020). *Spark Matrix: Process Mining 2020*. Quadrant Knowledge Solutions.

Modi, A., Makan, H., & Kumar, S. (2021). *Everest Group Peak Matrix for Process Mining Technology Vendors 2021*. Everest Group.

Naumann, J., & Jenkins, A. (1982). Prototyping: the new paradigm for systems development. *MIS quarterly*, 29-44.

Özgün, O., & Barlas, Y. (2009). *Discrete vs. continuous simulation: When does it matter.* Paper presented at the Proceedings of the 27th international conference of the system dynamics society.

Pauly, B., & Holdampf-Wendel, A. (2022). IT-Fachkräftelücke wird größer: 96.000 offene Jobs. Retrieved 25.02.2022 from https://www.bitkom.org/Presse/Presseinformation/IT-Fachkraefteluecke-wird-groesser

Reinkemeyer, L. (2020). *Process mining in action*. Cham: Springer.

Steger, J. (2018). Deutschland hat ein neues Milliarden-Start-up. Retrieved 25.02.2022 from https://www.handelsblatt.com/unternehmen/mittelstand/celonis-ist-jetzt-ein-einhorn-deutschland-hat-ein-neues-milliarden-start-up/22735410.html?ticket=ST-3249679-rfdF3Cz9uGUEgshqkN0W-ap3

Suriadi, S., Andrews, R., ter Hofstede, A., & Wynn, M. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information systems, 64*, 132-150.

Thiede, M., Fuerstenau, D., & Barquet, A. (2018). How is process mining technology used by organizations? A systematic literature review of empirical studies. *Business process management journal*.

Turner, C., Tiwari, A., Olaiya, R., & Xu, Y. (2012). Process mining: from theory to practice. *Business process management journal*.

van der Aalst, W. (2012a). Process mining. *Communications of the ACM, 55*(8), 76-83.

van der Aalst, W. (2012b). Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems (TMIS), 3*(2), 1-17.

van der Aalst, W. (2016). *Process Mining - Data Science in Action*. Berlin: Springer.

van der Aalst, W. (2019). A practitioner's guide to process mining: Limitations of the directly-follows graph. In (Vol. 164, pp. 321-328): Elsevier.

van der Aalst, W., Adriansyah, A., Alvarez de Medeiros, A., Arcieri, F., Baier, T., Blickle, T., . . . Buijs, J. (2011). *Process mining manifesto.* Paper presented at the International conference on business process management.

van der Aalst, W., Alves de Medeiros, A., & Weijters, A. (2005). *Genetic process mining.* Paper presented at the International conference on application and theory of petri nets.

van der Aalst, W., Schonenberg, M., & Song, M. (2011). Time prediction based on process mining. *Information systems, 36*(2), 450-475.

van der Aalst, W., & Weijters, A. J. (2004). Process mining: a research agenda. In: Elsevier.

van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE transactions on knowledge and data engineering, 16*(9), 1128-1142.

van Dongen, B., Alvares de Medeiros, A., Verbeek, H., Weijters, A., & van der Aalst, W. (2005). *The ProM framework: A new era in process mining tool support.* Paper presented at the International conference on application and theory of petri nets.

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). *Reconstructing the giant: On the importance of rigour in documenting the literature search process.* Paper presented at the 17th European Conference on Information Systems, ECIS 2009.

Webster, J., & Watson, R. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii-xxiii.

Weijters, A., van der Aalst, W., & Alves de Medeiros, A. (2006). Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP, 166*, 1-34.

Appendix

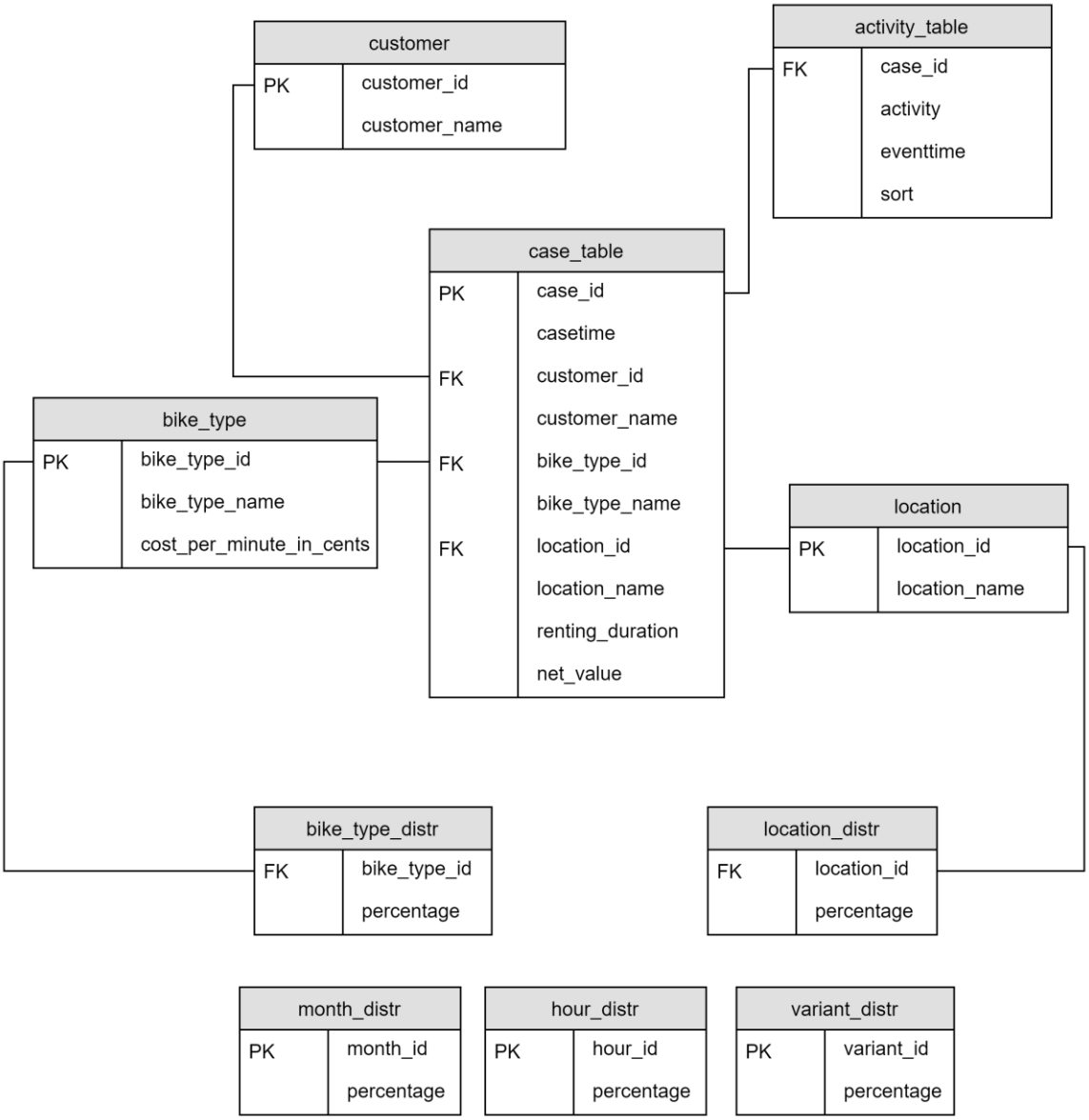
Appendix A: Concept Matrix of the Literature Review

Author(s)	Title	Year	Type	Concept Group A: Market			Concept Group B: Goals		
				Market Size	Geographical Assessment	Market Leader	Drivers/ Goals	Expecta-tions	Opportu-nities
Ailenei, Irina Rozinat, Anne Eckert, Albert van der Aalst, Will	Definition and validation of process mining use cases	2011	Scientific Article						
van der Aalst et al.	Process mining manifesto	2011	Scientific Article						x
Turner, Chris Tiwari, Ashutosh Olaiya, Richard Xu, Yuchun	Process mining: from theory to practice	2012	Scientific Article						
Suriadi, Suriadi Andrews, Robert ter Hofstede, Arthur Wynn, Moe	Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs	2017	Scientific Article						
Dakic, Dusanka Stefanovic, Darko Cosic, Ilija Lolic, Teodora Medojevic, Milovan	Business Process Mining Application: a Literature Review	2018	Scientific Article						
Lee, Wai Verbeek, H Munoz-Gama, Jorge van der Aalst, Wil Sepúlveda, Marcos	Recomposing conformance: Closing the circle on decomposed alignment-based conformance checking in process mining	2018	Scientific Article						
Mannhardt, Felix Petersen, Sobah Oliveira, Manuel	Privacy challenges for process mining in human-centered industrial environments	2018	Scientific Article						
Thiede, Malte Fuerstenau, Daniel Barquet, Ana	How is process mining technology used by organizations? A systematic literature review of empirical studies	2018	Scientific Article						
Kerremans, Marc	Gartner Market Guide for Process Mining	2019	Industry Report/Survey	x		x	x		x
van der Aalst, Will	A practitioner's guide to process mining: Limitations of the directly-follows graph	2019	Scientific Article						
Kerremans, Marc Searle, Samantha Srivastava, Tushar Iijima, Kimihiko	Gartner Market Guide for Process Mining	2020	Industry Report/Survey	x		x	x		x
Fleming, Reetika Duncan, Sam	HFS Top 10 Process Intelligence Products	2020	Industry Report/Survey		x	x			
Mehar, Riya	Quadrant Knowledge Solutions SPARK Matrix	2020	Industry Report/Survey	x	x	x			
Hawkins, Ian	The PEX Report 2020: Global State of Process Excellence	2020	Industry Report/Survey				x		
Reinkemeyer, Lars	Process mining in action	2020	Book						x
Andrews, Robert Wynn, Moe Vallmuur, Kirsten Ter Hofstede, Arthur Bosley, Emma	A comparative process mining analysis of road trauma patient pathways	2020	Scientific Article						
Grisold, Thomas Mendling, Jan Otto, Markus vom Brocke, Jan	Adoption, use and management of process mining in practice	2020	Scientific Article						
Kerremans, Marc Srivastava, Tushar Choudhary, Farhan	Gartner Market Guide for Process Mining	2021	Industry Report/Survey	x		x	x		x
Galic, Gabriela Wolf, Marcel	Deloitte Global Process Mining Survey 2021	2021	Industry Report/Survey		x		x	x	x
Kong, Bailey	Nelson Hall Neat Evaluation for Celonis: Process Discovery & Mining	2021	Industry Report/Survey	x	x	x	x	x	
Celonis	The State of Process Excellence	2021	Industry Report/Survey						
Martin et al.	Opportunities and challenges for process mining in organizations: results of a Delphi study	2021	Scientific Article						x

Author(s)	Title	Year	Type	Concept Group C	Concept Group D: Application Areas				
				Success Factors	Sectors	Functional Areas (Logistics, Finance ...)	Use Cases (Complicance, Automation, IT Operations ...)	Application Areas	Which Processes
Ailenei, Irina Rozinat, Anne Eckert, Albert van der Aalst, Will	Definition and validation of process mining use cases	2011	Scientific Article				x		
van der Aalst et al.	Process mining manifesto	2011	Scientific Article						
Turner, Chris Tiwari, Ashutosh Olaiya, Richard Xu, Yuchun	Process mining: from theory to practice	2012	Scientific Article						
Suriadi, Suriadi Andrews, Robert ter Hofstede, Arthur Wynn, Moe	Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs	2017	Scientific Article						
Dakic, Dusanka Stefanovic, Darko Cosic, Ilija Lolic, Teodora Medojevic, Milovan	Business Process Mining Application: a Literature Review	2018	Scientific Article		x				
Lee, Wai Verbeek, H Munoz-Gama, Jorge van der Aalst, Wil Sepúlveda, Marcos	Recomposing conformance: Closing the circle on decomposed alignment-based conformance checking in process mining	2018	Scientific Article						
Mannhardt, Felix Petersen, Sobah Oliveira, Manuel	Privacy challenges for process mining in human-centered industrial environments	2018	Scientific Article						
Thiede, Malte Fuerstenau, Daniel Barquet, Ana	How is process mining technology used by organizations? A systematic literature review of empirical studies	2018	Scientific Article		x				
Kerremans, Marc	Gartner Market Guide for Process Mining	2019	Industry Report/Survey				x		
van der Aalst, Will	A practitioner's guide to process mining: Limitations of the directly-follows graph	2019	Scientific Article						
Kerremans, Marc Searle, Samantha Srivastava, Tushar Iijima, Kimihiko	Gartner Market Guide for Process Mining	2020	Industry Report/Survey	x			x		
Fleming, Reetika Duncan, Sam	HFS Top 10 Process Intelligence Products	2020	Industry Report/Survey			x			
Mehar, Riya	Quadrant Knowledge Solutions SPARK Matrix	2020	Industry Report/Survey	x			x		
Hawkins, Ian	The PEX Report 2020: Global State of Process Excellence	2020	Industry Report/Survey				x		
Reinkemeyer, Lars	Process mining in action	2020	Book				x		
Andrews, Robert Wynn, Moe Vallmuur, Kirsten Ter Hofstede, Arthur Bosley, Emma	A comparative process mining analysis of road trauma patient pathways	2020	Scientific Article						
Grisold, Thomas Mendling, Jan Otto, Markus vom Brocke, Jan	Adoption, use and management of process mining in practice	2020	Scientific Article						x
Kerremans, Marc Srivastava, Tushar Choudhary, Farhan	Gartner Market Guide for Process Mining	2021	Industry Report/Survey				x		
Galic, Gabriela Wolf, Marcel	Deloitte Global Process Mining Survey 2021	2021	Industry Report/Survey	x		x			
Kong, Bailey	Nelson Hall Neat Evaluation for Celonis: Process Discovery & Mining	2021	Industry Report/Survey	x	x				
Celonis	The State of Process Excellence	2021	Industry Report/Survey						
Martin et al.	Opportunities and challenges for process mining in organizations: results of a Delphi study	2021	Scientific Article						

Author(s)	Title	Year	Type	Concept E	Concept F	Concept Group G: Challenges			Concept H
				Adoption by Type (3 Types)	Status (Company wide/Island/ Prototype ...)	Main Hurdles/ Reasons Not To Use	Challenges	Limitations	Covid Impact
Ailenei, Irina Rozinat, Anne Eckert, Albert van der Aalst, Will	Definition and validation of process mining use cases	2011	Scientific Article	x					
van der Aalst et al.	Process mining manifesto	2011	Scientific Article				x		
Turner, Chris Tiwari, Ashutosh Olaiya, Richard Xu, Yuchun	Process mining: from theory to practice	2012	Scientific Article				x		
Suriadi, Suriadi Andrews, Robert ter Hofstede, Arthur Wynn, Moe	Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs	2017	Scientific Article				x		
Dakic, Dusanka Stefanovic, Darko Cosic, Ilija Lolic, Teodora Medojevic, Milovan	Business Process Mining Application: a Literature Review	2018	Scientific Article	x					
Lee, Wai Verbeek, H Munoz-Gama, Jorge van der Aalst, Wil Sepúlveda, Marcos	Recomposing conformance: Closing the circle on decomposed alignment-based conformance checking in process mining	2018	Scientific Article					x	
Mannhardt, Felix Petersen, Sobah Oliveira, Manuel	Privacy challenges for process mining in human-centered industrial environments	2018	Scientific Article					x	
Thiede, Malte Fuerstenau, Daniel Barquet, Ana	How is process mining technology used by organizations? A systematic literature review of empirical studies	2018	Scientific Article						
Kerremans, Marc	Gartner Market Guide for Process Mining	2019	Industry Report/Survey	x		x	x		
van der Aalst, Will	A practitioner's guide to process mining: Limitations of the directly-follows graph	2019	Scientific Article					x	
Kerremans, Marc Searle, Samantha Srivastava, Tushar Iijima, Kimihiko	Gartner Market Guide for Process Mining	2020	Industry Report/Survey	x		x	x		
Fleming, Reetika Duncan, Sam	HFS Top 10 Process Intelligence Products	2020	Industry Report/Survey		x				
Mehar, Riya	Quadrant Knowledge Solutions SPARK Matrix	2020	Industry Report/Survey				x		
Hawkins, Ian	The PEX Report 2020: Global State of Process Excellence	2020	Industry Report/Survey		x		x		x
Reinkemeyer, Lars	Process mining in action	2020	Book				x	x	
Andrews, Robert Wynn, Moe Vallmuur, Kirsten Ter Hofstede, Arthur Bosley, Emma	A comparative process mining analysis of road trauma patient pathways	2020	Scientific Article				x		
Grisold, Thomas Mendling, Jan Otto, Markus vom Brocke, Jan	Adoption, use and management of process mining in practice	2020	Scientific Article				x		
Kerremans, Marc Srivastava, Tushar Choudhary, Farhan	Gartner Market Guide for Process Mining	2021	Industry Report/Survey	x		x	x		
Galic, Gabriela Wolf, Marcel	Deloitte Global Process Mining Survey 2021	2021	Industry Report/Survey		x	x	x		
Kong, Bailey	Nelson Hall Neat Evaluation for Celonis: Process Discovery & Mining	2021	Industry Report/Survey						
Celonis	The State of Process Excellence	2021	Industry Report/Survey				x		x
Martin et al.	Opportunities and challenges for process mining in organizations: results of a Delphi study	2021	Scientific Article				x		

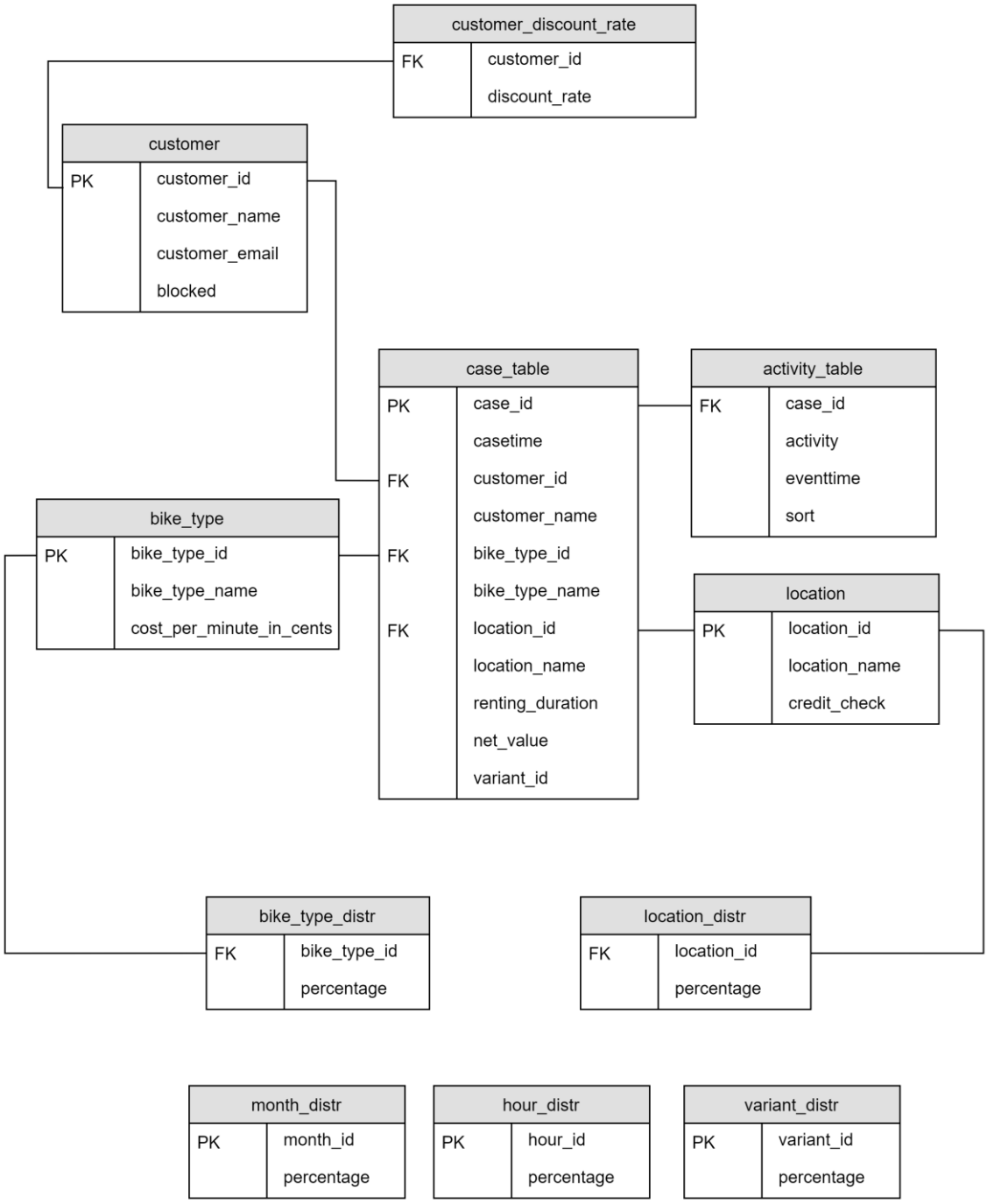
Appendix B: Database Schema for Research Question 2



Legend

PK indicates a primary key column.
FK indicates a foreign key column.
Connections between columns of different tables indicate foreign key relationships.

Appendix C: Database Schema for Research Question 3



Legend

PK indicates a primary key column.
FK indicates a foreign key column.
Connections between columns of different tables indicate foreign key relationships.