PHD CONFIRMATION OF CANDIDATURE

Queensland University of Technology, School of Information Systems

---

Proposed Thesis Title

Process Mining with Exogenous Data

Traditional Thesis by Monograph (Chapters)

Candidate: *Adam Peter Banham*

Contact: adam.banham@hdr.qut.edu.au, **Identification:** n7176546

Submission Date: 2024-05-07

Principal Supervisor: *Prof Moe Wynn*

Associate Supervisor: *Dr Robert Andrews*

External Supervisor: *Prof Sander Leemans*

Contents

Chapter 1

Supervisory team

1.1 Principal Supervisor: Prof Moe Wynn

Professor Moe Wynn leads the Process Science academic program at Queensland Univerity of Technology and is a leading researcher in the Business Process Management (BPM) field, as noted by her 100+ refereed research papers on the topics of process automation, data quality and process mining. She is also a co-leader of the Data for Discovery research theme within QUT's Centre for Data Science and is a steering committee member of the IEEE Task Force on Process Mining. Professor Moe Wynn will contribute to this thesis's theoretical efforts and act as a guide for the PhD candidature as a senior member of the School of Information Systems. She was the mentoring associate supervisor of Adam until March 2022. She takes over as the principal supervisor upon the departure of Dr Sander Leemans.

1.2 Associate Supervisor: Dr Robert Andrews

Dr Robert Andrews is a senior research fellow in the BPM Group of the School of Information Systems, Queensland University of Technology. His research interests include process mining, data mining and machine learning with a particular focus on data quality. He has published over 30 refereed research papers, with his work appearing in well-known journals including Information Systems, Decision Support Systems, Knowledge-based Systems and Neurocomputing. He has considerable experience in conducting process mining analyses in the healthcare and insurance domains. He has recently embarked on a research program around root-cause analysis of event data quality issues. Dr Robert Andrews will contribute to this thesis's theoretical efforts and act as a guide to understanding the medical domain from which our datasets originate.

1.3 External Supervisor: Prof Sander Leemans

Born in Boxtel, the Netherlands, Prof Sander J.J. Leemans was a senior lecturer within the School of Information Systems at the Queensland University of Technology until March 2022 and was the

principal supervisor. He has recently taken up a new position as a Professor at RWTH Aachen University in Germany. He has over 30 refereed research papers that focus on process mining, process discovery, conformance checking, stochastic process mining, and robotic process automation. In particular, he specialises in making solid academic techniques available to end-users, analysts and industry partners. He teaches business process management, business process modelling and business process improvement. Prof Sander Leemans will contribute to the theoretical efforts, implementation of theory and provide his understanding mechanics of process mining techniques.

Chapter 2

Introduction

In this project, we use process mining and time series data to study the influence of contextual factors on behaviour seen in processes. Using process mining, we can analyse how a collection of activities/events can be structured to present an end-to-end understanding of process behaviour. However, this analysis is usually driven by a state-based understanding, meaning that studying temporal changes in the context surrounding a process is challenging. To address this challenge, our project will create algorithms and techniques to find and understand contextual factors that processes adapt to.

Contextual data that can influence processes come in many forms, and we refer to this data as exogenous data. The exogenous data we focus on is time-series data, consisting of numerical, categorical, or ordinal measurements that can be used to present the context surrounding a process. Using this data, we investigate decision points within a process and if trends/artefacts within time series can explain why a choice was chosen. Identifying trends/artefacts will allow for an understanding of the context at decision points, which otherwise would not be considered in existing techniques.

In this section, we present the contextual backdrop of our discipline and present an overview of the PhD project. We start by explaining the subject domain for this work, process mining. We then discuss our motivation and the goals of this study. Finally, we conclude with the structure of the remaining report.

2.1 Process mining

Process mining is a discipline that uses historical data extracted from an organisation about a business process, to better understand how process behaviour occurs and its performance [1]. The benefit of using process mining techniques is that they provide evidence based insights and clarification on the reality of the process, based on historical records. Process mining techniques bridge the gap between the historical data, the presumed process and the reality, as highlighted in Figure 2.1.

Process mining techniques can usually be classified as either process *discovery*, *conformance* or *enhancement*, examples of each can be seen in Figure 2.2. Process discovery techniques [2–5] exploit historical process executions, to recreate the structure of a process and present a process model.

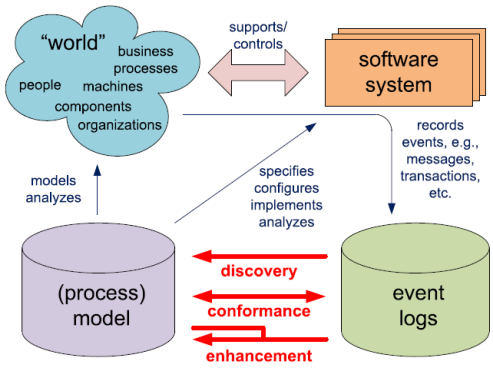
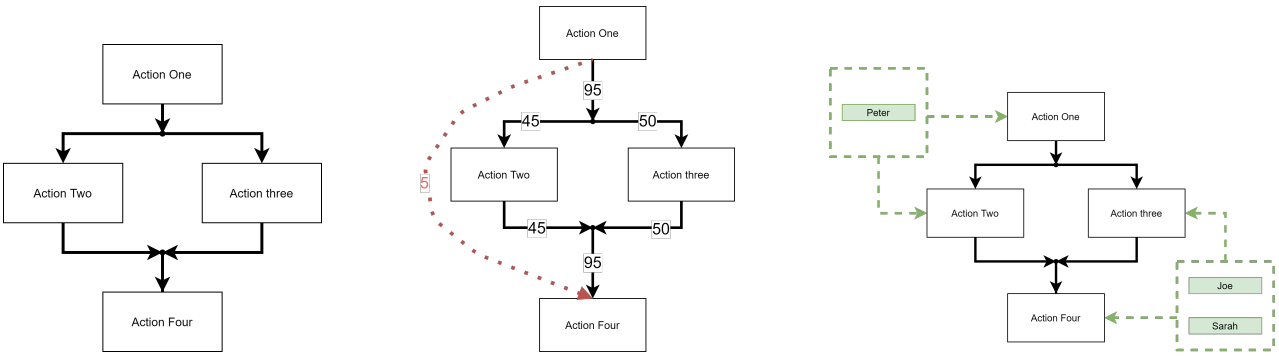


Figure 2.1: In [1, fig. 2.5, p32], the author expresses the link between the real world, historical data (event logs) and process mining techniques.

Process conformance techniques [6–8] use a process model and process executions, to understand the agreement between the behaviour in these elements, usually as a mapping or a measurement. Process enhancement techniques [1, 9] enrich a process model with an additional perspective of the process, such as the performance or the resource utilisation for activities. Typical process mining outcomes focus on presenting an objective view of a business process, identifying bottlenecks that slow down progress, under- and over-used resources, or highlighting inefficient behaviour.



(a) Discovery (b) Conformance (c) Enhancement

Figure 2.2: Discovery (a) finds an ordered set of actions. Conformance (b) outlines when process instances do not follow such an order (highlighted by the red dotted line). Enhancement (c) shows another influence; an action is completed by a set of people (highlighted by green boxes and names).

Enabling these process mining techniques in one form or another is an *event log* [1]. Contained within event logs are the *events* related to a single process. A sequence of events is called a *trace*. A trace describes an instance of a process. Extracting these events is often a challenge of its own[1, 10–12], as they may be stored across several information systems or document management systems at various levels of data quality. The purpose of these events is to describe a single moment of change, for example: what happened (activity), when it happened (time) and may have more descriptive attributes to comment on how it happened (resource/cost/data objects touched or created). These

elements, events and event logs, represent the behaviour of a process (process behaviour), which we refer to as process or *endogenous* data.

When analysing the endogenous data in an event log, process mining techniques will present an analysis on a single perspective of the process. The majority of the research focus has been on techniques that focus on the control-flow perspective [13], i.e. presenting the structure of a process, such as the Petri net model. Other notable perspectives to analyse are the data perspective (the flow of data objects across a process execution), the resource perspective (the groups of employees related to activities) or the time perspective (the queuing of jobs waiting to start or statistical distributions for activity completion). However, recent work in the literature has shown that considering aggregation of perspectives can provide additional benefits [14]. Such an approach is often called a **multi-perspective** approach, where at least two perspectives are considered in combination to provide insights.

2.2 Exogenous data

In this project, we take a multi-perspective approach, combining the data perspective and the control-flow perspective. To distinguish between different types of data sources, we define the following terms; process or*endogenous* data and *exogenous* data. We define *endogenous* process data as data internal to a process, meaning data that is directly produced by process activities, or that describes the moment of activity. Examples of endogenous data that describe a moment of a (healthcare) process include the name of the procedure being performed, the nurse performing the procedure, and the patient's level of discomfort during the procedure. In contrast, we define *exogenous* data as data external to a process, meaning data that is not tied directly to a process' activities, but which provides the context in which process instances execute. Examples of exogenous data sources include availabilities of surgeons, ICU bed availability, and patient vital sign measurements.

This project will revolve around how process mining techniques can use exogenous data. Specifically, we focus on using exogenous data to understand how decisions in processes are informed by contextual data and what reasoning is used to derive information from this data. By including exogenous data for decisions, analysts will be able to use process mining to understand how processes adapt to broader environment effects and derive behaviour that are reactions to external factors. These new insights will allow analysts to understand a process beyond the endogenous performance or conformance, leading to more modes of engagement with process owners. To achieve these outcomes, this project will show the benefits of including exogenous data within process mining techniques, and answer our research aim of: *How can analysts study the influence of exogenous data on decisions in processes?* Chapter 4 elaborates on our research aim, and questions.

2.3 Report structure

The remainder of this report is structured as follows. In Chapter 3, we outline the related work. In Chapter 4, we present our research aim and questions that will be addressed by this project. In Chapter

5, the research design for this project is outlined. Finally, in Chapter 6, we describe the progress to date and future research contributions.

Chapter 3

Related work

In this section, we present the related work for our project. We start by presenting previous multi-perspective studies that consider both the control-flow and the data perspectives. Next, we present work around the usage and presentation of data for processes, in terms of data sources and time series data. Then, we discuss work around the investigation of decisions in processes and how the reasoning for a choice is presented. Finally, we conclude with a summary of the research gaps.

3.1 Multi-perspective process mining

Methodologies that encourage contextual data collection and log enrichment are few. However, some recent studies have focused on the enrichment of an event log with new types of data. In [15], the authors present a framework for intra- and inter-trace predictive monitoring and introduce the notion of *bi-dimensional coding* to deal with intra- and inter-trace dependencies. In [16], the author's proposed approach extracts the implicit nature of the process from an event log, creating temporal process variables that describe the process quantitatively, then using these variables they generate high-level simulation models but do not consider exogenous data.In [17], the authors suggest that not all events within an event log are about the control flow, and are instead, about the data flow of a process. They use the concept of *context events* to deal with the two types of events and show how distinguishing between the two can lead to discovering less complex models. However, this approach would incorporate exogenous context into the control flow perspective instead of clarifying whether the context influences process execution.

The benefits of additional data attributes can be seen in the recent evolution of techniques that use such data [18, 19]. In [18], the authors present a discovery algorithm that uses data attributes to create a hierarchical process model to improve the simplicity of outcomes. Another approach, in [19], was to create a constraint operator for process trees notation, whereby data semantics can be expressed. While these techniques can create/discover control flow sequences based on data attributes, no extensions have been proposed to use exogenous sources outside what can be found in the events within an event log.

The previously mentioned studies often focused on the imperative representation of models (strictly specifying how a process executes); however, a more flexible description for process execution exists in techniques that are focused on declarative modelling [20]. One such example of a declarative modelling approach is DECLARE, which has developed using a constraint-based system grounded in temporal logic [20].

Studies such as those in [21] and [22] use this approach, which incorporates multiple process perspectives. In [21], the authors present a framework for multi-perspective declarative process discovery, whereby different process perspectives (time, resource, or data) could be used to generate a set of constraints. In [22], the authors present declarative process discovery, focused on finding constraints based on correlated data attributes within constraint instances, explicitly focusing on the control-flow and data perspective of a process. While declarative modelling may offer greater flexibility by using constraints, without an explicit understanding of behaviour for a process, finding decision points and reasoning for a choice is unclear. Thus, for this project, we will be considering imperative representations like Petri nets instead of a declarative approach.

3.2 Process data in context

The categorisation of the types of data sources used to describe processes has been discussed in several studies. The 'onion skin' model in [23] conceptualises the relationship between data and process through layers as the viewpoint is moved further away from a process. The authors present four layers: where an immediate layer, contains the process perspectives (e.g. control-flow or data), an internal layer contains business plans and process models, an external layer contains competitors or workforce pressures, and an environmental layer contains economic or social pressures. This conceptualisation is then applied to process mining in [24], where data is categorised in layers according to the likelihood of cause and effect between variables with the process, with a greater focus on the business. In this categorisation, the authors start with a process executions context (instance context), then a process context where the amount of resources allocated to a process are considered, then in the social context the internal pressures from within a business are considered, finally in the external context we consider environmental pressures. The difference between these categorisations, is the emphasis on process-relevant contexts rather than a systematic approach to identifying all possible pressures that processes face; our study focuses on finding process-relevant contexts.

In recent years, research has focused on providing a more integrated consideration of contextual factors that processes face [25]. In [25], the authors motivate for a better understanding of various contextual factors and analysis of informed decisions using these factors, which was expanded upon in [26].

The benefits of including a variety of data categories are discussed in [14] which (i) motivates the use of data attributes for distinguishing between noise and conditional behaviour, (ii) considers if data attributes influence decision points by creating an internal state as the process executes through boolean expressions and decision trees, and (iii) studied how alignments [6] can be extended such that they balance both the control flow perspective and the data perspective.

In [27], the authors present recommendations for future work to stimulate a more widespread use of process mining, such as recommending more multi-perspective studies that use multi-organisational data sources in analysis. This recommendation is not new, as it can be seen in recent discussion papers [28] and past recommendations [29–31]. As a constant recommendation, the need for more contextual data within process mining activities has been emphasised several times, and these studies highlight that it has not been addressed. Furthermore, the contextual component remains an optional consideration during event data extraction based on recent systematic reviews [32–34].

In this work, we propose to support separate entities for endogenous and exogenous data sources, such that they can studied in combination.

3.3 Time series data for processes

In process mining, including time series data (such as exogenous data) is often challenging as techniques are not designed with temporal data in mind [35]. In [35], the author motivates for process mining techniques that consider a collection of events with arbitrary size and a temporal understanding of data availability. Another approach to studying processes over time, are process conformance techniques around concept drift, where the structure of process behaviour changes over time [36–38]. Instead of focusing on the macro-level for temporal changes, the process may experience a high velocity of atomic changes within temporal windows, where techniques like [39–41] can be applied to ensure process mining techniques can provide insights. Using these approaches, we can consider how a collection of process executions differs from others in a later temporal window or how high velocity events can be abstracted into higher-level activities. However, work in this area has focused on internal changes in processes rather than external factors impacting process behaviour.

Little consideration has been given to time series techniques such as those presented in [42]. Where techniques like dynamic time warping [43], Granger's concept of causality [44], or signal estimation using Fourier transforms [45], are considered to identify external factors affecting process conformance or structure. In many cases, studies have set aside the temporal information and instead preferred a simple representation for analysis to suit process mining activities, rather than making time series data a first-class citizen. Therefore, to have exogenous data be included in analysis which leads to meaningful insights, we need to ensure that time series data can be treated as a first-class citizen.

3.4 Decision mining

Some process enhancement techniques look into the data perspective of an event log and decision points in a process model to provide insights about decision rules/reasoning. These techniques are often referred to as decision mining. Decision mining was first present in [46], where the goal of decision mining is to find rules, explaining what circumstances lead to the execution of certain paths in processes. In order to find rules, each decision point (a place with multiple outgoing arcs in a Petri net) is turned into a classification problem, and typically a decision tree is learnt, creating a human-readable rule/expression.

This work was expanded upon in [47], where the authors showed that by using alignments [6], event attributes could be considered instead of only trace attributes when analysing decision points. The introduction of event attributes into the classification meant alternative algorithms could be used to solve the classification problem. In [48], the authors present an alternative solution which can discover linear equalities with 2 or 3 variables and an optional constant within rules, e.g. length + age ≤ 70 where length and age are event attributes. In [49], a solution that does not focus on finding exclusive rules for decision points was presented, which previously had been not considered. In [50], the authors presented an approach to consider rules that distinguished between the read (current) and written (next) values of variables, e.g. $ok^{read} = \text{false} \wedge \text{granted}^{written} \leq \text{requsted}^{read}$.

Decision mining techniques like [46, 48–51] are considered local techniques, as they consider localised decision points within a process. In contrast, proposals like [52, 53] are global techniques that present natural language statements about overreaching decision-making of the process. Both types of decision mining techniques are constrained to endogenous data, more specifically event/trace attributes. Without a representation of data in the form of an attribute, these techniques (local or global) cannot find rule/expressions using any external data sources like exogenous data. Thus, in our project, we consider how local techniques can use exogenous data through a representative attribute or use the raw time series data.

3.5 Research gaps

Research to date has focused on process mining techniques that make use of endogenous data, rather than exogenous data available around the process. This focus means that even when we have access to exogenous data, we enrich event logs only by encoding this data directly into the event log. Following this approach limits the types of exogenous data used in process mining, as many data sources would not aggregate without a loss of information or are not suited to aggregation. However, without this representation as an event/trace attribute, the inclusion of exogenous data in existing process mining activities is not possible, meaning the benefits of this data have not been studied in detail. This highlights that a research gap exists, as existing process mining techniques do not allow for the broad inclusion of exogenous data.

Furthermore, while context has been recognised as influencing process behaviours, only limited progress has been made in including contextual data for process mining. To the best of our knowledge, there are no techniques capable of exploiting exogenous data to discern its influence on process behaviour or decisions in processes. The lack of process mining techniques that use exogenous data or contextual factors in analysis constitutes a research gap that we address in our project.

Chapter 4

Research aim and problems

The research gap identified in our literature search showed a need for multi-perspective techniques that consider data beyond event/case attributes in an event log. In order to better understand the role that exogenous data plays at a decision point within a process using a multi-perspective approach, our project will consider the interplay between exogenous data (time series data) and decision rules informing a choice in a process model. By including trends/artefacts within exogenous data as part of the process mining analysis, analysts will be able to study contextually-appropriate reasoning being applied at decision points. Expanding upon this work will allow for a broader understanding of the influences that affect processes.

The overarching aim of this PhD project is: *How can analysts study the influence of exogenous data on decisions in processes?* To achieve this aim, we propose **three** research questions that will provide a robust understanding of the interplay between exogenous data and process decisions.

As mentioned in Chapter 3, the literature review does not uncover studies that define exogenous data, or describe a practical inclusion of contextual factors (beyond representing static elements as trace or event attributes) in an event log. Thus, the inclusion of exogenous data for process mining activities, or its benefits for process mining outcomes, has not been studied in detail. We emphasise that exogenous data should be a first class citizen and will require a systematic approach to provide meaningful insights and benefits. This leads to research question one:

RQ One: How can we represent and use exogenous data in existing process mining techniques?

Without an understanding of time series data and how it changes, it will not be possible to fully understand the influence of exogenous data on decision points within a process. In practical settings where time series data is used, domain experts consider at least three aspects of the data: trends evident in the longitudinal data of the time series, occurrence of domain-relevant artefacts/patterns, or some sudden change/anomaly in the time series shape. Automatically selecting an appropriate aspect and a period of time that is relevant to a process decision point, is as of yet unresolved. This leads us to research question two:

RQ Two: How can patterns/artefacts in exogenous data be identified and represented for decision-making?

It is frequently the case, that decision-making informing a choice within a process model exhibits ambiguities and inconsistencies. That is, when considering a collection of traces and their endogenous data, it is not possible to derive an expression that consistently/accurately describes the trajectory of each trace through the decision point. We cannot study this inconsistent decision-making at a decision point using existing approaches. To clarify/reduce this variability, we will consider how exogenous data can be used to build context-inclusive expressions that provide greater coverage of process decision variance. Our intention is to move from indeterministic to probabilistic decision-making through inclusion of exogenous data within an expression. This leads to research question three: **RQ Three: How can analysts study the influence of exogenous data on decision-making variance in processes?**

The specifics of how this project will address each of these research questions are discussed in Chapter 6.

Chapter 5

Project design

This chapter outlines the research methodology we will use to conduct our research for the remainder of this project's timeline. Then we outline the considerations we have for resources that will enable this project, concluding with the projected timeline for contributions in this project.

5.1 Research methodology

Our approach to research is generally based on Design Science Research (DSR). Design Science Research (DSR) is a standard methodology used when an iterative approach to development is needed and is presented in several forms by [54–57]. DSR has an iterative nature (seen in Figure 5.1) as such it will integrate well with process mining projects and allows for more agile project management processes. Furthermore, the principles outlined by [57] are of particular note as they would ensure that contributions to the field could exist before testing any artefact in practice.

To apply a general DSR approach for our contributions, we look to the stages presented in [58] for proposing robust (discovery) process mining techniques. In the rigour phase, a literature review alongside a benchmarking of existing techniques should be conducted to understand the gap that needs to be addressed. In the design phase, we consider the gap and the guarantees that a process mining algorithm should have to fill it, design a prototype with these guarantees, and then test that our prototype provides such guarantees in a controlled setting. In the relevance cycle, we work with domain experts, investigate an unknown process, test that guarantees are kept, and evaluate a contribution's usefulness.

5.2 Resources

As we have outlined several evaluation states during the project, we will use publicly available data sets to test if our proposed techniques are feasible and are of quality. A collection of commonly used datasets can be found in 4TU.ResearchData repositories[1]. However, this collection may not provide our research with exogenous data and, as such, it will be required to find such data. Some simple weather or

[1] https://data.4tu.nl/search?q=:keyword:%20%22real%20life%20event%20logs%22

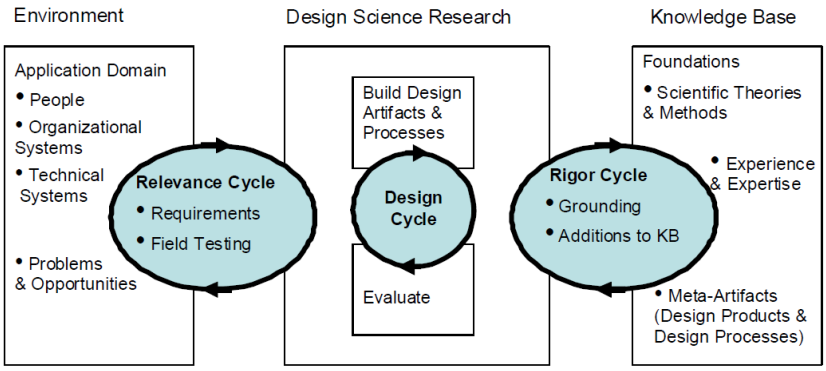


Figure 5.1: An example of how the DSR methodology is applied to research contributions [56, pp. 7].

general measures of the environment around businesses and their cultural context may be extractable from free APIs. However, another dataset of particular note is MIMIC-III [59], which provides "highly granular information about patient journeys for over forty thousand patients that stayed at a medical centre between 2001 and 2012" [59]. This dataset provides detailed monitoring information about patients, which would be considered exogenous data of interest; however, the anonymisation used on the data to make the dataset publicly available may limit how we can use this data within our study.

As such, we will start with publicly available datasets, as outlined above. However, we may also engage with stakeholders to collect new datasets as a contribution to provide the closest representation of exogenous data for process mining studies. Handling and computing any outcomes using an arbitrary amount of exogenous data could mean that QUT's supplied generic desktops are not suitable to conduct research effectively. In this case, we will need access to more appropriate computer resources, such as QUT's HPC. However, it may be of beneficial to see how general computing platforms like Google cloud or Amazon web services perform could be used to share our project outcomes.

5.3 Project timeline

In Figure 5.2, we outline the proposed plan for this project. We also highlight the key milestone points, which are required by QUT and when mandatory training will occur. We also outline the research focus of our work and highlight how these could convert into high quality journal and conference papers. Over the duration of the PhD we will aim for several high-quality publications in Q1/D1 journals and publications at the following high-quality conferences:

- International Conference for Process Mining (ICPM),
- International Conference on Business Process Management (BPM),
- International Conference on Advanced Information Systems Engineering (CAiSE) and
- International Conference on Application and Theory of Petri Nets and Concurrency (Petri Nets)

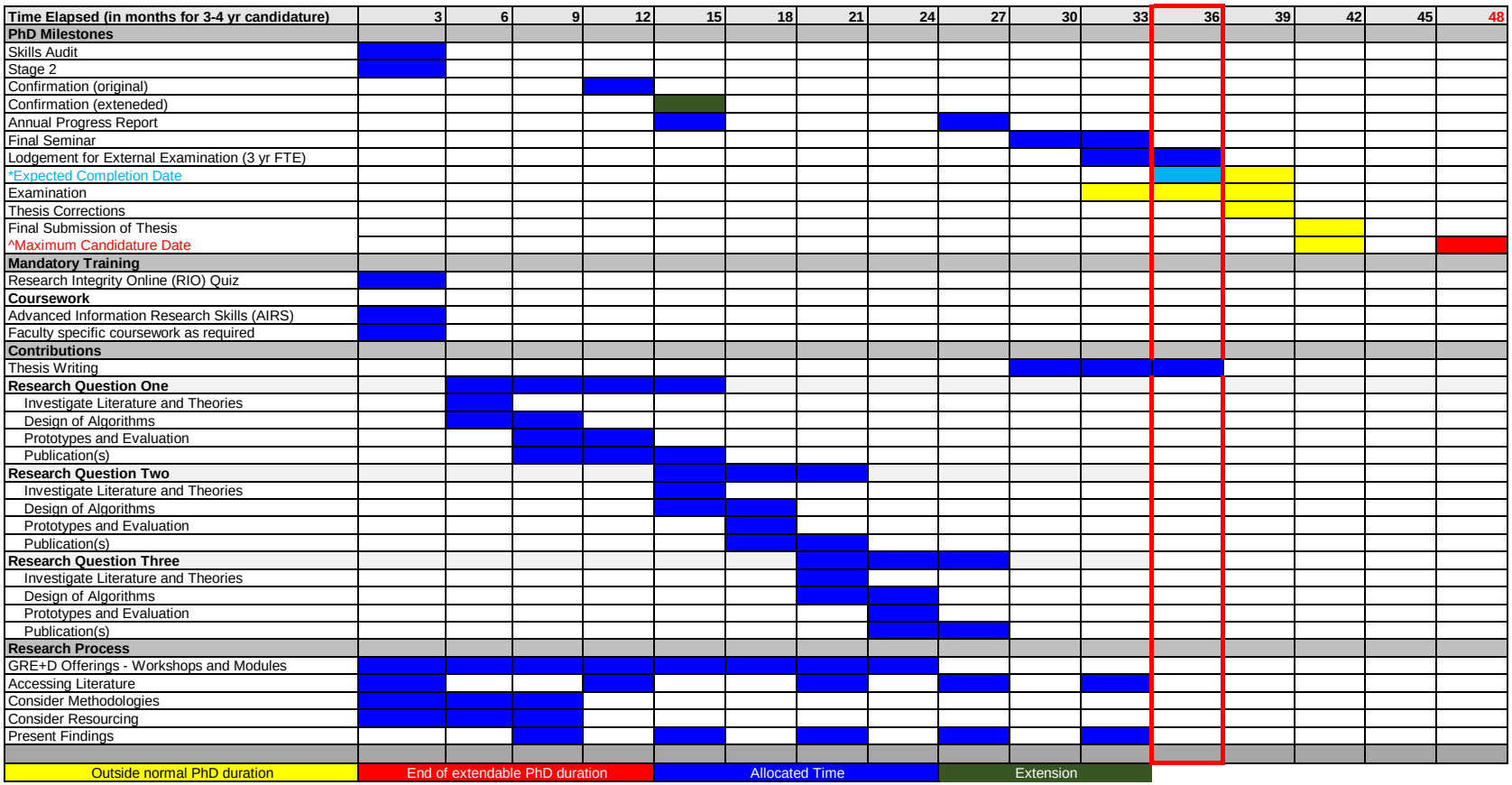


Figure 5.2: Timeline for this Project.

Chapter 6

Research contributions

In this section, we motivate the problem that this project addresses. Then, we discuss the completed contributions for research question one. Concluding the report with the planned contributions to address research questions two and three.

To date, we have addressed our first research question by proposing a novel framework for process mining with exogenous data [60]. We have presented our findings in two articles. First, our initial findings were presented at EBDA2021, a co-located workshop at ICPM2021, 'xPM: A Framework for Process Mining with Exogenous Data' [60]. Later, we extended the workshop paper for consideration in a special issue in the Journal of Artificial Intelligence for Medicine (Q1 since 2018 on SJR) (see Appendix B, under review - submitted April 2022), addressing the need for temporal analysis in moments of action for a (healthcare) process in close collaboration from healthcare co-authors.

6.1 Motivating Problem

In this section, we provide formalisms for data and process models used in our research questions. Then provide a motivating example for using exogenous data and process mining.

6.1.1 Process formalisms

In this section, we present the formalisms that are used in literature for events, event logs and Petri nets. We also present our formalism for exogenous data, such that it can be broken down to a single measurement.

Event logs. The execution of each step in a process can be recorded as an event. Let Γ be the set of all timestamps. An event e occurred at some timestamp $t \in \Gamma$, noted as e^t. An end-to-end execution of a process is called a *trace*. A trace is a sequence of events $\langle e_1^{t_1}, \dots e_n^{t_n} \rangle$. An event log is a collection of traces. Both traces and events can have attributes attached to store data. We denote $e_{\#\mathrm{A}}$ as being the value of attribute A for event e.

Exogenous Data. An *exogenous measurement* (exo–measurement) m^t is a single measurement at timestamp $t \in \Gamma$. For instance, $37^{18-01-202211:55:23}$ may represent the result of a nurse measuring the temperature of a particular patient at the given timestamp. An *exogenous series* (exo–series)

$\langle m_1^{t_1}, \ldots m_n^{t_n} \rangle^A$ is a sequence of exo–measurement, annotated with a set of attributes A. For instance, $\langle 37^{18-01-202211:55:23}, 37.5^{18-01-202212:13:58}, 38^{18-01-202213:14:32} \rangle^{\text{patient\_id:1034847}}$ may represent the repeated measuring of body temperature by a nurse for a particular patient.

An *exogenous panel* (exo–panel) is a collection of exo–series. For instance, $\{\langle 37^{18-01-202211:55:23}, 38^{18-01-202213:14:32} \rangle^{\text{patient\_id:1034847}}, \langle 37.8^{18-01-202215:47:39}, 37.8^{18-01-202220:15:41} \rangle^{\text{patient\_id:1034882}}\}$. An *exogenous description* (exo–description) is a collection of exo–panels. Typically, an exo–panel would contain all related exo–measurements; for instance: all blood pressure measurements for all patients. An exo–description then contains several exo–panels; for instance: one for blood pressure, one for heart rate, and one for body temperature.

Labelled Petri Nets. A *Petri net* is a triple $N = (P, T, F)$, where P is a finite set of places, T is a finite set of transitions such that $P \cap T = \emptyset$ and $F \subset (P \times T) \cup (T \times P)$ is a set of directed arcs, called a flow relation [1]. A *labelled Petri net* is a quintuple, $(P, T, F, \Sigma, \lambda)$, where (P, T, F) is a Petri net, Σ is a set of activity labels and λ is an event labelling function $T \to \Sigma$ [1]. Places may hold tokens, which are produced and consumed when transitions fire according to the flow relation. A transition is *enabled* if each input place contains a token.

The state of a Petri net is a *marking*, which records what places have tokens and how many. An enabled transition l can *fire*, which updates the marking according to the flow relation F and, if l is labelled by λ, denotes the execution of activity $\lambda(l)$. An *initial marking* denotes the initial state of a Petri net before the first transition is fired. Elements of $P \cup T$ are called nodes. A node a is an input node of another node b if and only if there is a directed arc from a to b. A node a is an output node of another b if and only if there is a directed arc from b to a. For any $a \in P \cup T$, $\bullet a = \{b \mid (b, a) \in F\}$ and $a\bullet = \{b \mid (a, b) \in F\}$.

6.1.2 Exogenous data and decision-making

In this exemplar scenario, we consider a process containing all the behaviour and individual actions needed to see an operation conducted successfully by the fire and emergency services. The endogenous data that inform how this process is conducted include the types of actions conducted by field members and their recorded observations of an area, and the actions of members at a base camp that monitor field members and their findings. The external factors affecting this process include the following exogenous data: the fire risk rating for the day and local wind speed measurements to ensure the operation's overall safety. The decisions occurring within the process by various members include field member's choosing the most appropriate action given the exogenous data and decisions by the home camp based on the exogenous data to ensure that the operation is not endangering the public or field members. Improving how decisions are made by the home camp or the field members could increase the chances of a positive outcome occurring or saving a life.

In order to understand how operations are conducted, we used the processed events to create an event log (see a snippet in Table 6.1), considering all operations that have ended. By grouping events, we can replay an operation (a trace) and we can begin to study the structure of the operations (process). To start our study, we need a clearer understanding of how these operations are conducted, so we apply a process discovery algorithm and find the Petri net shown in Figure 6.1. Investigating

Table 6.1: Exemplar event log for search and rescue operations conducted by a fire and rescue service. Data presented was not based on any real data. Scenario is only for illustrative purposes. Placeholders for activity labels are used instead of meaningful activity labels (T_x).

trace	e_{i}	$e_{\#\text{activity}}$	$e_{\#\text{teamName}}$	$e_{\#\text{searchGrid}}$	$e_{\#\text{areaDisturbed}}$
1	1	T_1	home		
1	2	T_2	home		
1	3	T_3	home		
1	4	T_4	alpha	1	No
1	5	T_9	alpha	1	No
1	6	T_{13}	beta	2	Yes
1	7	T_6	home		
1	8	T_{14}	beta	2	No
1	9	T_5	gamma	3	Yes
1	10	T_9	gamma	3	No
1	11	T_{16}	alpha	1	Yes
1	12	T_{10}	home		
1	13	T_{18}	home		
2	14	T_1			
2	15	T_2			
2	16	T_3			
2	17	T_4	alpha	1	No
2	18	T_8	home		
2	19	T_{12}	home		
2	20	T_{17}	home		
2	21	T_9	alpha	1	No
2	22	T_{16}	alpha	1	No
2	23	T_{18}			

the discovery outcome further, we can see the separated behaviour between home camp and field members, as well as the many decision points. However, to understand how contextual factors influence decision-making we need to consider exogenous data.

The exogenous data we consider in this scenario is an exo–description containing two exo–panels: the fire risk rating for the day and localised measurements of wind speed for an area. In Figure 6.1, we present examples of exo–series from the different exo–panels. Applying existing decision mining techniques, we could only find if the endogenous data (observations of an area) can represent decision-making at decision points. Furthermore, we cannot find any evidence-based insights for the temporal understanding used by home camp or the field members to interpret this exogenous data. Without techniques that consider exogenous data, we can understand the process structure but little about the decision-making occurring.

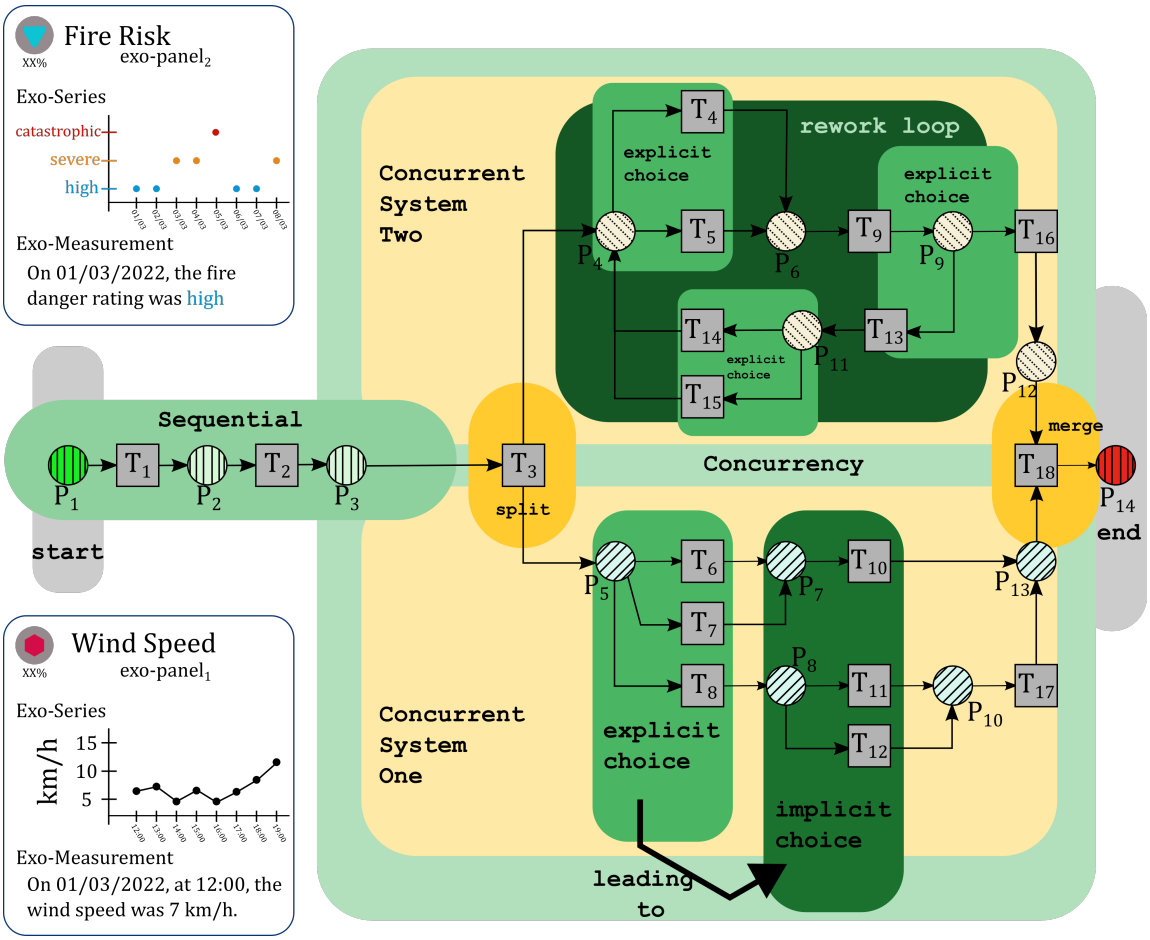


Figure 6.1: A process model, using a Petri net notation, that may be discovered from a process discovery technique. In this process model, we have highlighted several behavioural types that can be noted by studying the process structure. The process starts at P_1 and would end once we get to P_{14}. It starts off by sequentially executing T_1, T_2 and T_3. After T_3, we have concurrency between two concurrent Petri nets, where they may execute in an interleaving manner. In system one, the process is faced with an explicit choice and then an implicit choice based on the first activity that is executed. As T_{10} can only occur if we execute either $T_6|T_7$, similarly, the explicit choice of either $T_{11}|T_{12}$ can only occur if T_8 is executed. In system two, we have a rework loop, where all behaviour can be repeated many times until we see T_{16} executed. Once we see either T_{10} or T_{17} in system one and T_{16} in system two executed, the concurrency ends as well as the process. Included on the left side, is the exogenous data that is being considered for this process, the fire risk rating for the day and the recorded wind speed. The breakdown of exogenous data from exo–panel, to exo–series, and to exo–measurement is demonstrated.

6.2 RQ One: How can we represent and use exogenous data in existing process mining techniques?

The research gap identified in our literature search showed a need for analysis that considers contextual factors of process variance and a comprehensive range of multi-perspective process mining techniques. To fill the gap on contextual analysis in the literature, we described our understanding of contextual factors through our definition of exogenous data. We did not find any previous studies that used a similar definition of exogenous data through our literature search, or considered a practical inclusion of exogenous data (outside trace or event attributes) in an event log. To emphasise the benefits of exogenous data, we presented two contributions, a framework for process mining with exogenous data (xPM) and explorative exogenous series analysis (EESA). These contributions address the challenges of including exogenous data within process mining activities, such as the need to criticise how data is used within the process.

We contrasted our contributions against pre-processing methods in this phase, as enriching events with new data attributes would allow exogenous data to be present within event logs. Existing pre-processing methods can consider information implicitly available within event logs (such as workload utilisation); however, they do not analyse if external sources have influenced the process. Without this consideration, analysts can run into situations where events are overloaded with attributes that do not contribute additional insights or may mislead the analyst. Furthermore, pre-processing data requires some understanding of accurately representing data as a single value and that this understanding does not change as processes execute. To resolve these drawbacks, in our proposed approach, xPM, analysts consider determinations, where temporal sequences and transformed attributes are associated with events.

In xPM (see Figure 6.2), we apply several *determinations*, to annotate events with exo–series and transformed attributes to describe the possible influences affecting process behaviour. Determinations are tuples of $(x, \mathcal{L}, \mathcal{S}, \mathcal{T})$, consisting of: an exo–panel (x), a linking function to connect a trace and an exo–panel $(\mathcal{L})$, a slicing function to create a subset of an exo–series (*slice*) for an event $(\mathcal{S})$, and a transformation function to create transformed attributes for an event $(\mathcal{T})$. These determinations allow for a more flexible understanding of how data is interpreted as processes execute, as opposed to pre-processing methods, by allowing analysts to trace back to the original time series. The next two steps of xPM, discovery $(\mathcal{D})$ and enhance $(\mathcal{E})$, emphasis the need to test if a determination can show evidence of an influence on the process.

We emphasise testing determinations' ability to provide evidence of influence through our visualisation technique, explorative exogenous series analysis (EESA), and demonstrate the advantages of xPM. This contribution visualises slices from determinations for a given point of the process, highlighting the general trends of slices leading to the same process outcome (as seen in Figure 6.3). While this technique does not provide empirical evidence of an influence, it shows the potential of understanding temporal changes in the context around processes. Furthermore, while manually investigating EESAs is possible, it may not be possible to find clear insights across visualisations if a process model is sufficiently large. To address this concern, an automated analysis (a ranking procedure) was presented

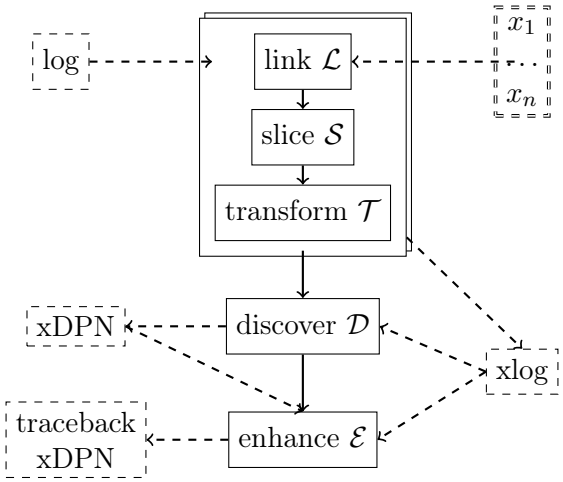


Figure 6.2: Our proposed framework (xPM) for process mining with exogenous data. In the first steps of the framework, given an event log and a exo–series, several determinations are applied. After all determinations are applied to an event log, we have an *exogenous-annotated* log (xlog). Then in the discovery step, we use the xlog to discover a process structure informed by exogenous data, producing a xDPN (a Petri net with exogenous data). Finally, in the enhance step, we use the xlog and xDPN, to provide evidence that determinations can describe an influence on the process.

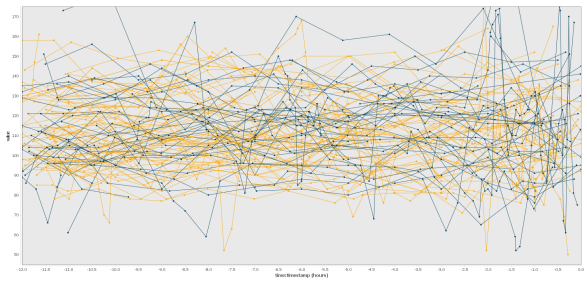
to identify EESAs with the most or least amount of difference between temporal changes presented in the visualisation.

Research contributions:

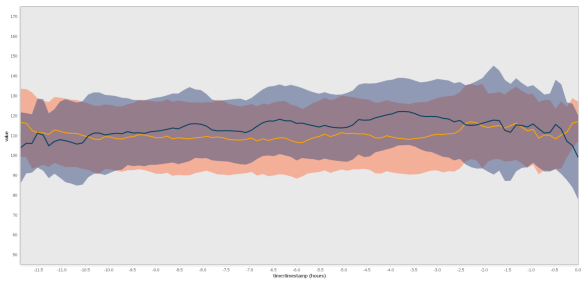
- A formalisation of exogenous data within event log theory using set and automata theory.
- A framework (xPM) that allows for exogenous data in event logs, creating an exogenous aware log.
- A new analysis for studying if exogenous data can be used to discriminate between process outcomes.
- An evaluation of how the algorithms can be used to show evidence that data can be associated with a process.
- An evaluation of existing process mining techniques using xPM.
- An evaluation of the implementations in a simulated setting and against public datasets.

6.3 RQ Two: How can patterns/artefacts in exogenous data be identified and represented for decision-making?

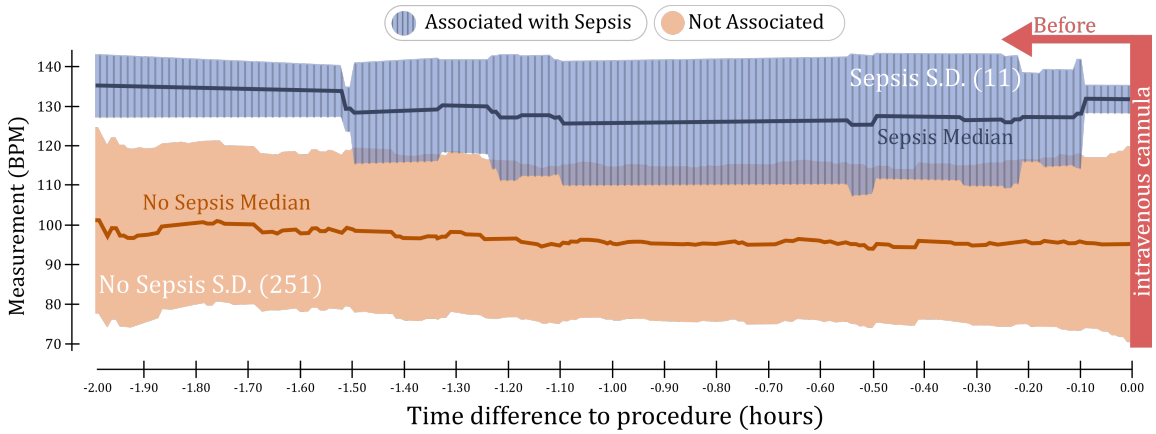
In order to understand how changes within exogenous data influence decision-making for processes, we will reconsider the set of logical operators used in making expressions for decisions. In previous



(a) A visualisation of raw slices from a determination, using all events associated with a transition in a Petri net.



(b) An overlapping EESA Visualisation using the same slices from Figure 6.3a.

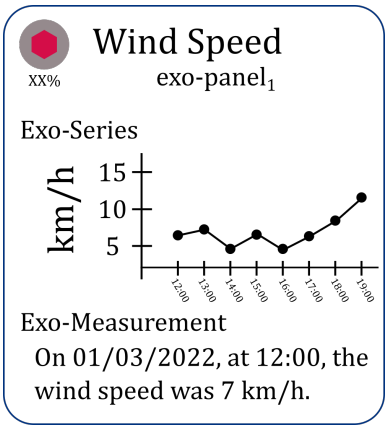


(c) A stylised EESA identified from our ranking procedure for an evaluation in a medical domain showing possible discrimination power.

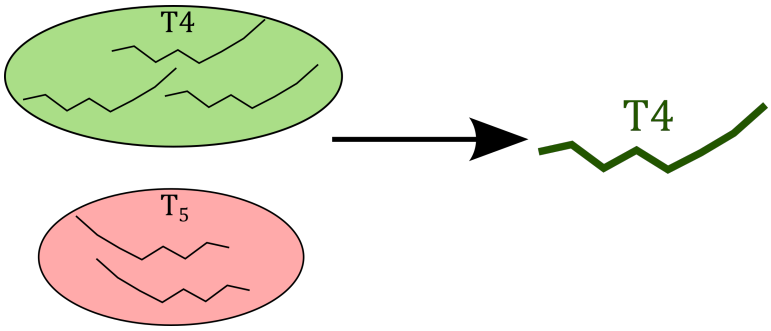
Figure 6.3: A comparison of visualisation techniques, one using the exogenous slices as-is (a), and another showing the generalisation of slices using explorative exogenous series analysis or EESA (b). Using our automated analysis, a large collection of EESAs can be considered and analysts can find where a determination and EESA may provide discrimination, such as (c) being identified from 2480 candidates. The usage of colour in each figure, is to highlight that a slice is associated with a particular process outcome.

work [46, 48–50], the logical operators that have been used to create an expression for association with a transition, have focused on a variable's absolute value, i.e. $i = 10, \text{if } i > 15 \rightarrow T_4$. Thus, the set of operators that have been considered so far in decision mining techniques consists of the following: $\{=, >, <, \neq\}$. However, it is not clear what is the absolute value of a time series, meaning we need temporal operators that can express changes across many values within expressions.

An important aspect that can represent change within time series (such as exogenous data) is if the longitude trends follow some shape. This phase will investigate presenting a shape operator, which given a collection of time series, will find a representative time series for a target class (t_x or T_4). An exemplar procedure to do so is shown in Figure 6.4. In this procedure, we would cluster time series by their similarity (using dynamic time warping [43]), find a cluster with high coverage of t_x, and then



(a) Samples of time series are taken from exo–panel$_1$, and are associated with target class of either T_4 or T_5 for process executions considered in Figure 6.7c.



(b) Time series are clustered based on similarity. Then the coverage for each cluster is considered. In this case each cluster has complete coverage over a single target class t_x. Thus, we find a representative time series for T_4 on the right.

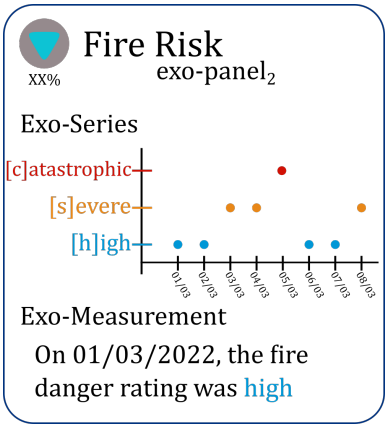
Figure 6.4: Breakdown of how a procedure might find a shape for a given t_x and collection of time series. Then, using the representative time series found, a shape operator can be introduced into an expression. The shape operator logically expresses that some time series must be similar to be satisfied.

find a representative time series from the samples in this cluster. After which, the shape operator expresses how similar a given time series must be to the representative to be satisfied.

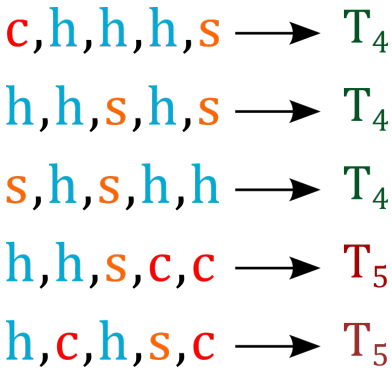
An alternative way to look for changes within time series is to look for patterns within values. We will investigate a sequential operator, which given a collection of sequential values, finds a (repeatable) sequence that predicts a target class (t_x). An example of an evolutionary algorithm that could find such a sequential pattern can be seen in Figure 6.5. In this example, we create candidates from a collection of sequential values, mutate them via promotion (adding a repetition or new element) or simplify (removing a repetition or element), and find the shortest sequence with the highest entropy for a given t_x. Thus, a sequential operator would use the found sequence and would be satisfied if a given time series contained the sequence.

For this phase, the final consideration of a temporal operator would focus on longitude statistical measurements, i.e. is the overall trend line increasing or decreasing. Compared to the other temporal operators in this phase, this is a simpler approach that aims to find a straightforward understanding. A procedure for this operator could consider if the gradient or acceleration of a time series has a cut-off value that can predict t_x. For example, if the velocity summation is greater than 50 for a given time series, we can confidently predict T_4. However, this operator could only apply to numerical or ordinal categorical values for time series.

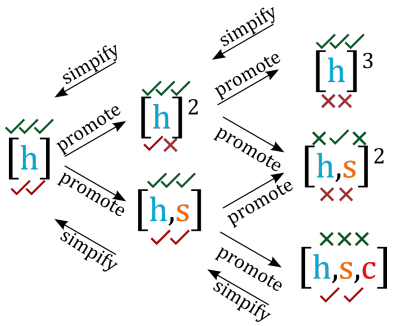
Without these temporal operators, analysts would not be able to fully understand the influence of exogenous data on decisions. By introducing these operators, this phase will highlight the importance of temporal changes in exogenous data. A new decision mining technique will be required to enable these operators within expressions and will be the focus of the publication in this phase. Given the



(a) Samples of time series are taken from exo–panel$_2$, considering the same process executions as Figure 6.7c.



(b) A collection of samples, each associated with some target class t_4. Samples are a sequence of fire risk ratings, either catastrophic (c), severe (s) or high (h).



(c) Breakdown of a single candidate either being promoted or simpified. A tick means that agreement was found with a sample. A cross means that disagreement was found with a sample. A tick/cross at the top of sequence is related to T_4, likewise at the bottom it is related to T_5. The right-most column shows two sequences with clear target seperation.

Figure 6.5: Breakdown of a procedure to find a sequential operator for a t_x.

technical focus of this work, we will target either ICPM 2023 or CAiSE 2023 as a conference venue or directly submit to a Q1 journal outlet.

Proposed Research Contributions:

- A new operator for expressions, called a shape operator and an algorithm for finding a shape across observations that can predict a target class (t_x).
- A new operator for expressions, called a sequential operator and an algorithm for finding a sequence (one or more times) in a time series that can predict a target class (t_x).
- A new operator for expressions, called a trend operator and an algorithm for finding a cut-off threshold for a statistical measurement (i.e. gradient, velocity, acceleration) in a time series that can predict a target class (t_x).
- A new decision mining technique that can use these operators and through various construction strategies (e.g. one vs all apposed to mutliclass classification strategies previously used).
- A new form of analysis to understand where exogenous data has enabled decision-making to occur within a process, such that we can identify the shortest and longest portion of exogenous data required, what temporal understanding (shape/sequential/trend) should be considered and when the decision outcome is apparent (at the decision point or was it already foreseen).

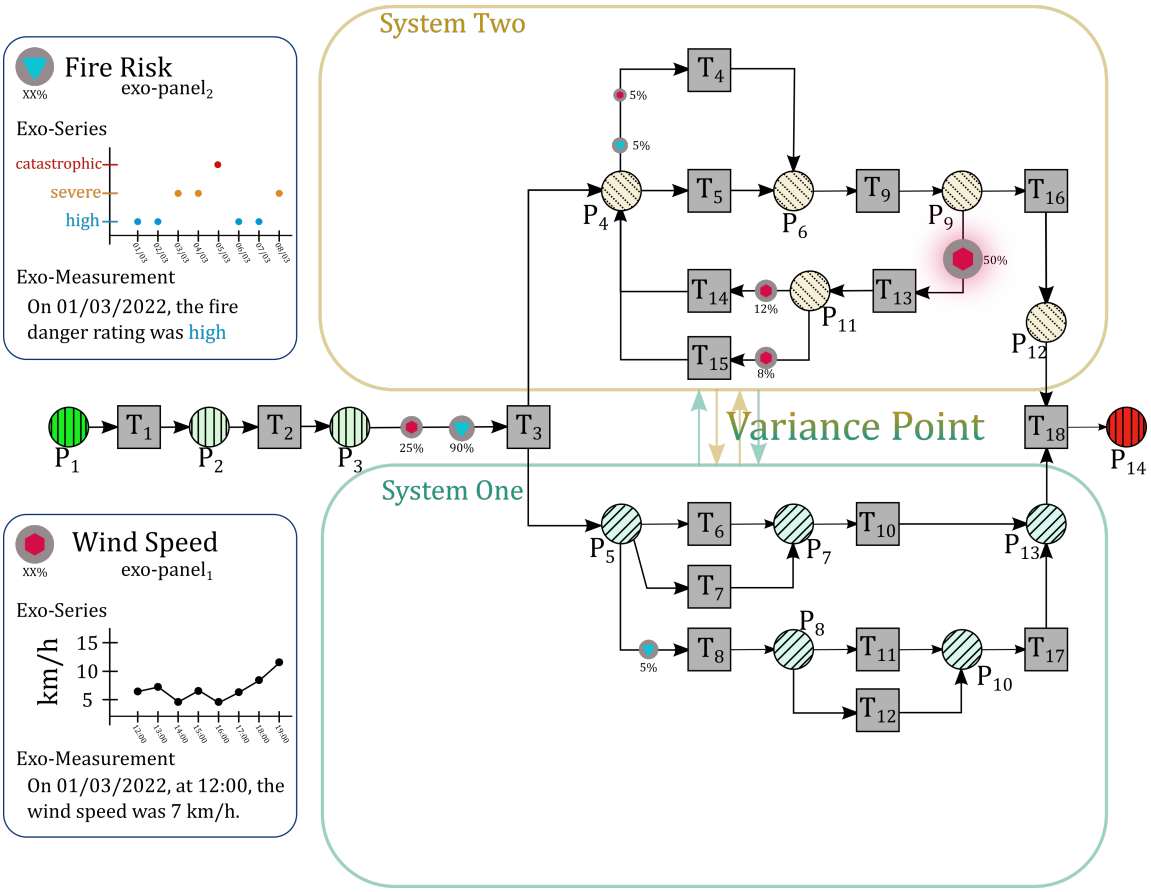


Figure 6.6: Revisiting our exemplar scenario, we can see that two concurrent systems are alive in this process. After T_3 is executed, these two systems should occur in an interleaving manner, where the order of between systems is not encoded into the process structure. In this research question, we would find these systems as a variance point and investigate if contextual factors will influence the decision-making between these systems. For example, system one may have a policy in place to act if the fire risk reported for the next day is catastrophic, where T_8 is immediately executed if possible, triggering the end of the fire fighting operation.

6.4 RQ Three: How can analysts study the influence of exogenous data on decision-making variance in processes?

In order to understand how decision-making occurs within processes, we need to be able (1) to find when the decision occurs within a Petri Net (process model) and (2) to describe the thought process being applied at the decision point (an expression for the next activity). In previous techniques [46, 48–50], finding decision points has consisted of finding a place in a Petri net with multiple outgoing arcs. After finding a decision point, a collection of observations is created for this decision point, where each observation is a feature vector ($< f_1, f_2, \ldots, f_n >$) and a target class (next activity, t_x). These techniques will use the observations, with differences in terms of implementation or the usage of observations; however, they will all create a logical expression for a target class (t_x). This expression

is then associated with a transition (t_x), which restricts executions to those that agree with this expression before it can fire and brings a deterministic nature to an otherwise indeterministic point.

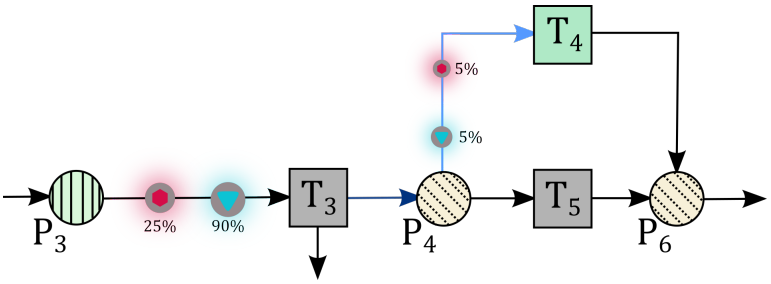
In our project, we reconsider the finding of 'decision points', more specifically the focus on exclusive choices, and instead present a broader definition. Our definition instead considers, 'variance points', where given a marking of a Petri net and the likelihood of a trace execution, we find points where the variance in future markings can be expressed. This definition still considers places with multiple outgoing arcs (decision points) but extends to consider discrimination between concurrent markings (i.e. if a marking in system one or system two in Figure 6.6 should occur next). Furthermore, we can consider if multiple variance points exist for P_4 (see Figure 6.7) and determine if T_4 or T_5 should occur next, For example, we would consider if the first occurrence of a choice ($P_4 \rightarrow T_4|T_5$) a different variance point from the first three repeated markings and every repeated marking after the third repeat. However, for each variance point, we would still find a collection of observations to find expressions to describe a given choice.

We also reconsider how the expression informing a choice should be associated with behaviour in a process. The association of an expression with a transition should be more explicit, as seen in Figure 6.7d. Firstly, the expression should have some notion of coverage, the number or proportion of observations in agreement, i.e. an expression covers 30% of observations for target class t. Secondly, there should be an unknown option for a transition, i.e. an intractable expression covers 60% of observations for target class t. Finally, there should be expressions that consider if exogenous data sources are available, i.e. an expression including exo-panel$_x$ covers 10% of observations for target class t. We argue that the availability of exogenous data triggers a unique decision-making process, and thus observations should be treated differently as well as the resulting expressions.

Combining these elements, variance points and a conditional set of expressions associated with a transition, will allow for a new analysis that has not been previously considered. This analysis will allow analysts to study if decisions will change, either becoming more or less restrictive, if the decision is repeatedly considered during a process execution. We refer to this phenomenon as 'decision decay', where different variance points for the same target class will experience a change in the likelihoods of executions, exogenous data or expression coverage when considering longer or shorter marking sequences. Such as the repeated marking of P_4, in the segments shown in Figures 6.7a & 6.7c. Therefore, analysis of decision decay will be a major focus for the publication in this research phase, targeting either ICPM2023, BPM2023 or the Journal of Decision Support Systems (Q1 on SJR since 2005).

Planned Research Contributions:

- An algorithm for enhancing a control-flow of process model with entry and exit points of exogenous-panels.
- An algorithm for finding variance points within a control-flow model of a process, creating a set of observations with a likelihood and target class for decision mining techniques.
- A notation for a conditional set of expressions being associated with a transition in a Petri net with Data.

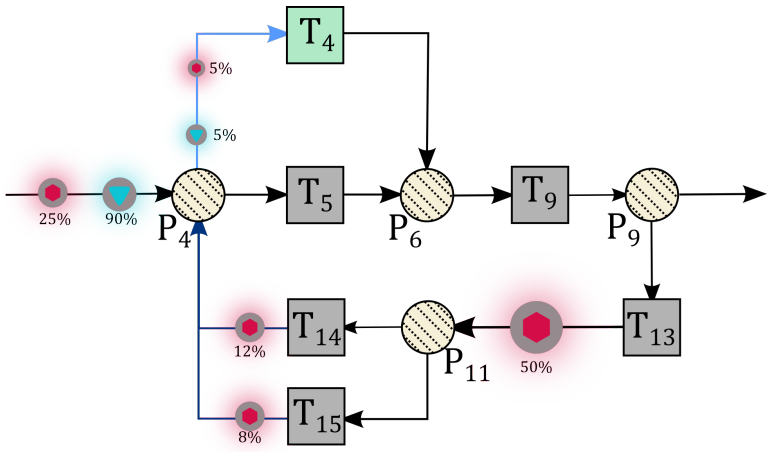


(a) A segement of the Petri net model from Figure 6.1. In this segement, we are finding variance points for target class, transition T_4, highlighted in green. Variance points being found are only considering process executions that follow: $\{P_3\} \rightarrow T_3 \rightarrow \{P_4, P5\} \rightarrow T_4 \rightarrow \{P_6, P_5\}$. Notated on arcs, are the symbols for two exo-panels, with the likelihood of seeing this exo-panel for the first time.

T_4 - First Occurence

	P(VP)	Name
VP_1	70%	Control flow
VP_2	X%	intractable
VP_3	21%	exo-panel$_1$
VP_4	66%	exo-panel$_2$
VP_5	20%	exo-panel$_{1,2}$

(b) Based on the target class (T_4), and the likely path of process executions, we find several possible variance points (VP_x), with likelihoods (P(VP)) and different collections of observations. The intractable expression, is named as a variance point. However until a decision mining technique is applied, we are unable to describe it.



(c) In this segement, we consider process executions that have followed the rework path in the model, for the same target class, T_4. Variance points being found are only considering process executions that follow, the previous execution in Figure 6.7a and the following extension: $\{P_5, P_6\} \rightarrow T_9 \rightarrow \{P_5, P_9\} \rightarrow T_{13} \rightarrow \{P_5, P_{11}\} \rightarrow T_{14}|T_{15} \rightarrow \{P_5, P_4\} \rightarrow T_4 \rightarrow \{P_5, P_6\}$.

T_4 - First Rework

	P(VP)	Name	C(exp)
VP_6	50%	Control flow	C%
VP_7	X%	intractable	X%
VP_8	44%	exo-panel$_1$	C%
VP_9	47%	exo-panel$_2$	C%
VP_{10}	42%	exo-panel$_{1,2}$	C%

(d) Again a set of variance points for T_4 are found. This time we find expressions for each VP, except for the intractable. After which these expresions can be associated to T_4 alonside the coverage (C(exp)) of the found expression.

Figure 6.7: Breakdown of our proposed contributions for this research phase. Finding multiple variance points for a single target class and then presenting each one in consideration of different trace execution likelihoods.

- The process execution semantics for one or more sets of conditional expressions being associated with a transition, in a Petri net with data.
- A new form of analysis to study how ‘decision decay‘ can affect decision-making considering exogenous data within a process, such that we can identify expressions that change over a process execution, how they change (more/less restrictive) or if an expression is still relevant (active/disabled after some point).

Bibliography

[1] W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016, ISBN: 978-3-662-49850-7. DOI: 10.1007/978-3-662-49851-4.

[2] S. Leemans, E. Poppe, and M. Wynn, "Directly follows-based process mining: Exploration and a case study," in *2019 International Conference on Process Mining, ICPM 2019*, IEEE, 2019, pp. 25–32.

[3] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from incomplete event logs," in *Application and Theory of Petri Nets and Concurrency*, Springer International Publishing, 2014, pp. 91–110. DOI: 10.1007/978-3-319-07734-5_6.

[4] W. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, Sep. 2004.

[5] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from event logs containing infrequent behaviour," in *BPM Workshops*, ser. Lecture Notes in Bus. Inf. Process. Vol. 171, Springer, 2013, pp. 66–78.

[6] A. Adriansyah, "Aligning observed and modeled behavior," en, Ph.D. dissertation, Mathematics and Computer Science, 2014. DOI: 10.6100/IR770080.

[7] J. M. E. M. van der Werf, H. M. W. Verbeek, and W. M. P. van der Aalst, "Context-aware compliance checking," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2012, pp. 98–113.

[8] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Balanced multi-perspective checking of process conformance," *Computing*, vol. 98, no. 4, pp. 407–437, Feb. 2015.

[9] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Data-driven process discovery - revealing conditional infrequent behavior from event logs," in *Advanced Information Systems Engineering*, Springer International Publishing, 2017, pp. 545–560.

[10] S. Suriadi, R. Andrews, A. H. M. ter Hofstede, and M. T. Wynn, "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Inf. Syst.*, vol. 64, pp. 132–150, 2017.

[11] F. Emamjome, R. Andrews, and A. H. M. ter Hofstede, "A case study lens on process mining in practice," in *On the Move to Meaningful Internet Systems: OTM 2019 Conferences - Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21-25, 2019, Proceedings*, ser. Lecture Notes in Computer Science, vol. 11877, Springer, 2019, pp. 127–145.

[12] R. Andrews, F. Emamjome, A. H. ter Hofstede, and H. A. Reijers, "An expert lens on data quality in process mining," in *2020 2nd International Conference on Process Mining (ICPM)*, IEEE, Oct. 2020. DOI: 10.1109/icpm49681.2020.00018.

[13] G. Adamo, C. Ghidini, and C. Di Francescomarino, "What is a process model composed of?" *Software and Systems Modeling*, Jan. 2021. DOI: 10.1007/s10270-020-00847-w.

[14] F. Mannhardt, "Multi-perspective process mining," English, Ph.D. dissertation, Technische Universiteit Eindhoven, 2018, ISBN: 978-90-386-4438-7.

[15] A. Senderovich, C. Di Francescomarino, and F. M. Maggi, "From knowledge-driven to data-driven inter-case feature encoding in predictive process monitoring," *Inf. Systems*, vol. 84, pp. 255–264, 2019.

[16] M. Pourbafrani and W. M. P. van der Aalst, "Extracting process features from event logs to learn coarse-grained simulation models," in *Advanced Information Systems Engineering - 33rd International Conference, CAiSE 2021, Melbourne, VIC, Australia, June 28 - July 2, 2021, Proceedings*, ser. Lecture Notes in Computer Science, vol. 12751, Springer, 2021, pp. 125–140. DOI: 10.1007/978-3-030-79382-1_8.

[17] M. Dees, B. Hompes, and W. M. P. van der Aalst, "Events put into context (EPiC)," in *ICPM*, IEEE, 2020, pp. 65–72.

[18] S. J. Leemans, K. Goel, and S. J. van Zelst, "Using multi-level information in hierarchical process mining: Balancing behavioural quality and model complexity," in *2020 2nd International Conference on Process Mining (ICPM)*, IEEE, Oct. 2020, pp. 137–144.

[19] R. Shraga, A. Gal, D. Schumacher, A. Senderovich, and M. Weidlich, "Inductive context-aware process discovery," in *1st International Conference on Process Mining (ICPM)*, IEEE, Jun. 2019, pp. 33–40.

[20] M. Pesic, H. Schonenberg, and W. M. van der Aalst, "DECLARE: Full support for loosely-structured processes," in *11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007)*, IEEE, Oct. 2007, pp. 287–287.

[21] S. Schönig, C. D. Ciccio, F. M. Maggi, and J. Mendling, "Discovery of multi-perspective declarative process models," in *Service-Oriented Computing*, Springer International Publishing, 2016, pp. 87–103.

[22] V. Leno, M. Dumas, F. M. Maggi, M. L. Rosa, and A. Polyvyanyy, "Automated discovery of declarative process models with correlated data conditions," *Information Systems*, vol. 89, p. 101 482, Mar. 2020.

[23] M. Rosemann, J. Recker, and C. Flender, "Contextualisation of business processes," *Inter. J. of Business Process Integration and Management*, vol. 3, no. 1, pp. 47–60, 2008.

[24] W. M. P. van der Aalst and S. Dustdar, "Process mining put into context," *IEEE Internet Computing*, vol. 16, no. 1, pp. 82–86, 2012.

[25] J. vom Brocke, S. Zelt, and T. Schmiedel, "On the role of context in business process management," *International Journal of Information Management*, vol. 36, no. 3, pp. 486–495, Jun. 2016. DOI: 10.1016/j.ijinfomgt.2015.10.002.

[26] J. vom Brocke, M.-S. Baier, T. Schmiedel, K. Stelzl, M. Röglinger, and C. Wehking, "Context-aware business process management," *Business & Information Systems Engineering*, Mar. 2021. DOI: 10.1007/s12599-021-00685-0.

[27] N. Martin *et al.*, "Recommendations for enhancing the usability and understandability of process mining in healthcare," *Artificial Intelligence in Medicine*, vol. 109, p. 101 962, Sep. 2020.

[28] J. Munoz-Gama and et. at. al., "Process mining for healthcare: Characteristics and challenges," *Journal of Biomedical Informatics*, 2021.

[29] R. Gatta *et al.*, "Clinical guidelines: A crossroad of many research areas. challenges and opportunities in process mining for healthcare," in *Business Process Management Workshops*, Springer, 2019, pp. 545–556. DOI: 10.1007/978-3-030-37453-2_44.

[30] E. Batista and A. Solanas, "Process mining in healthcare: A systematic review," in *9th International Conference on Information, Intelligence, Systems and Applications, IISA 2018, Zakynthos, Greece, July 23-25, 2018*, IEEE, Jul. 2018, pp. 1–6.

[31] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of Biomedical Informatics*, vol. 61, pp. 224–236, Jun. 2016.

[32] J. De Smedt, S. K. L. M. van den Broucke, J. Obregon, A. Kim, J.-Y. Jung, and J. Vanthienen, "Decision mining in a broader context: An overview of the current landscape and future directions," in *BPM Workshops*, Springer, 2017, pp. 197–207.

[33] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, "Extraction, correlation, and abstraction of event data for process mining," *WIREs Data Mining and Knowl. Discovery*, vol. 10, no. 3, 2019.

[34] A. E. Marquez-Chamorro, K. Revoredo, M. Resinas, A. Del-Rio-Ortega, F. M. Santoro, and A. Ruiz-Cortes, "Context-aware process performance indicator prediction," *IEEE Access*, vol. 8, pp. 222 050–222 063, 2020.

[35] S. van Zelst, "Process mining with streaming data," English, Proefschrift, Ph.D. dissertation, Mathematics and Computer Science, Mar. 2019, ISBN: 978-90-386-4699-2.

[36] F. Stertz and S. Rinderle-Ma, "Process histories - detecting and representing concept drifts based on event streams," in Springer International Publishing, 2018, pp. 318–335. DOI: 10.1007/978-3-030-02610-3_18.

[37] T. Brockhoff, M. S. Uysal, and W. M. van der Aalst, "Time-aware concept drift detection using the earth mover's distance," in *2020 2nd International Conference on Process Mining (ICPM)*, IEEE, Oct. 2020. DOI: 10.1109/icpm49681.2020.00016.

[38] F. Stertz, S. Rinderle-Ma, and J. Mangler, "Analyzing process concept drifts based on sensor event streams during runtime," in *Lecture Notes in Computer Science*, Springer International Publishing, 2020, pp. 202–219. DOI: 10.1007/978-3-030-58666-9_12.

[39] J. D. Smedt, A. Yeshchenko, A. Polyvyanyy, J. D. Weerdt, and J. Mendling, "Process model forecasting using time series analysis of event sequence data," in Springer International Publishing, 2021, pp. 47–61. DOI: 10.1007/978-3-030-89022-3_5.

[40] F. Mannhardt and N. Tax, "Unsupervised event abstraction using pattern abstraction and local process models," CEUR Workshop Proceedings, vol. 1859, pp. 55–63, 2017. [Online]. Available: http://ceur-ws.org/Vol-1859/bpmds-06-paper.pdf.

[41] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, and P. J. Toussaint, "Guided process discovery – a pattern-based approach," *Information Systems*, vol. 76, pp. 1–18, Jul. 2018. DOI: 10.1016/j.is.2018.01.009.

[42] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering – a decade review," *Information Systems*, vol. 53, pp. 16–38, Oct. 2015. DOI: 10.1016/j.is.2015.04.007.

[43] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb. 1978. DOI: 10.1109/tassp.1978.1163055.

[44] C. Granger, "Some recent development in a concept of causality," *Journal of Econometrics*, vol. 39, no. 1-2, pp. 199–211, Sep. 1988. DOI: 10.1016/0304-4076(88)90045-0.

[45] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984. DOI: 10.1109/tassp.1984.1164317.

[46] Rozinat, A (Anne), "Process mining:conformance and extension," en, Ph.D. dissertation, 2010. DOI: 10.6100/IR690060.

[47] M. de Leoni and W. M. P. van der Aalst, "Data-aware process mining," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13*, ACM Press, 2013. DOI: 10.1145/2480362.2480633.

[48] M. de Leoni, M. Dumas, and L. Garcia-Bañuelos, "Discovering branching conditions from business process execution logs," in Springer Berlin Heidelberg, 2013, pp. 114–129. DOI: 10.1007/978-3-642-37057-1_9.

[49] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Decision Mining Revisited - Discovering Overlapping Rules," in *Advanced Information Systems Engineering*, Springer International Publishing, 2016, pp. 377–392.

[50] P. Felli, M. de Leoni, and M. Montali, "Soundness verification of decision-aware process models with variable-to-variable conditions," in *2019 19th International Conference on Application of Concurrency to System Design (ACSD)*, IEEE, Jun. 2019, pp. 82–91. DOI: 10.1109/acsd.2019.00013.

[51] J. D. Smedt, F. Hasić, S. K. vanden Broucke, and J. Vanthienen, "Holistic discovery of decision models from process execution data," *Knowledge-Based Systems*, vol. 183, p. 104 866, Nov. 2019. DOI: 10.1016/j.knosys.2019.104866.

[52] K. Winter and S. Rinderle-Ma, "Discovering instance-spanning constraints from process execution logs based on classification techniques," in *2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC)*, IEEE, Oct. 2017. DOI: 10.1109/edoc.2017.20.

[53] K. Winter, F. Stertz, and S. Rinderle-Ma, "Discovering instance and process spanning constraints from process execution logs," *Information Systems*, vol. 89, p. 101 484, Mar. 2020. DOI: 10.1016/j.is.2019.101484.

[54] V. Vaishnavi, B. Kuechler, and S. Petter. "Design science research in information systems," Association for Information Systems. (Jun. 2004), [Online]. Available: http://www.desrist.org/design-research-in-information-systems/.

[55] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *Management Information Systems Quarterly*, vol. 28, no. 1, p. 75, 2004. DOI: 10.2307/25148625.

[56] A. R. Hevner, "A three cycle view of design science research," *Scandinavian Journal of Information Systems*, vol. 19, p. 4, 2007.

[57] C. Sonnenberg and J. vom Brocke, "Evaluations in the science of the artificial – reconsidering the build-evaluate pattern in design science research," in *Lecture Notes in Computer Science*, Springer, 2012, pp. 381–397. DOI: 10.1007/978-3-642-29863-9_28.

[58] J. M. E. M. van der Werf, A. Polyvyanyy, B. R. van Wensveen, M. Brinkhuis, and H. A. Reijers, "All that glitters is not gold," in *Advanced Information Systems Engineering*, Springer International Publishing, 2021, pp. 141–157. DOI: 10.1007/978-3-030-79382-1_9.

[59] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, May 2016. DOI: 10.1038/sdata.2016.35.

[60] A. Banham, S. J. J. Leemans, M. T. Wynn, and R. Andrews, "xPM: A framework for process mining with exogenous data," in *Process Mining Workshops - ICPM 2021 International Workshops*, ser. Lecture Notes in Business Information Processing, vol. 433, Springer, 2021, pp. 85–97. DOI: 10.1007/978-3-030-98581-3_7.

Appendix A

Additional coursework requirements

A.1 Research Integrity Online

I have completed the research integrity online (RIO) course required by the PhD coursework and understand my responsibilities as a researcher for QUT and the greater community. See evidence of the course completion in Figure A.1 below.

Figure A.1: Evidence of completion of RIO at QUT.

A.2 IFN001 AIRS

I have completed the AIRS library course and submitted my assessment (19/03/2021) and have received a passing mark. See evidence of marked submission below in Figure A.2.

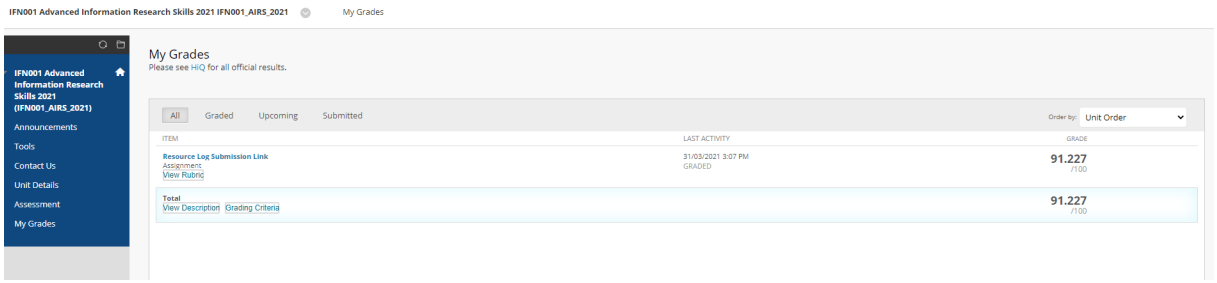


Figure A.2: Evidence of marked submission completion of AIRS course.

A.3 Coursework

Besides AIRS, no further coursework is required by this PhD. All required coursework units for the PhD course have been completed through the previous honours program at QUT. See evidence below of the completion of INN700 and INN701 required for the PhD in Figure A.3.

ID Number : 7176546 **Name : Adam Peter Banham** **Page 4 of 5**

Bachelor of Information Technology (Honours)

Units of Study

Unit Code	Unit Title	Grade	Description	Credit Points
Semester 1, 2020				
CAB402.2	Programming Paradigms	7	High Distinction	12
IFN403-1.1	IT Honours Research Project-1	6	Distinction	12
INN700.4	Introduction to Research	7	High Distinction	12
INN701.3	Advanced Research Topics	7	High Distinction	12
Semester 2, 2020				
CAB401.2	High Performance and Parallel Computing	6	Distinction	12
IFN403-2.1	IT Honours Research Project-2	6	Distinction	12
IFN403-3.1	IT Honours Research Project-3	6	Distinction	12
IFN403-4.1	IT Honours Research Project-4	6	Distinction	12
Prizes Awarded				
2020	Dean's List Award - Semester 1 - Science and Engineering			

Course Grade Point Average (GPA): **6.375**

Bachelor of Information Technology (Honours) Second Class Honours - Division A
Course requirements completed on **25/11/2020**
Conferred on **22 January 2021**

Figure A.3: Evidence of completed required coursework units in previous study.

Appendix B

Special Issue Article