DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
31

VERONIKA PLOTNIKOVA

FIN-DM: A Data Mining Process for the Financial Services

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

31

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
31

VERONIKA PLOTNIKOVA

FIN-DM: A Data Mining Process for the Financial Services

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on 12 November, 2021 by the Council of the Institute of Computer Science, University of Tartu.

Supervisors

Prof.	Marlon Dumas University of Tartu Estonia
Assoc. Prof.	Fredrik P. Milani University of Tartu Estonia

Opponents

Prof.	Longbing Cao University of Technology Sydney Australia
Assoc. Prof.	Jennifer Horkoff Chalmers — University of Gothenburg Sweden

The public defense will take place on 23 December, 2021 at 09:00 online.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906
ISBN 978-9949-03-769-8 (print)
ISBN 978-9949-03-770-4 (PDF)

University of Tartu Press
`http://www.tyk.ee/`

To my family and friends

ABSTRACT

The use of data mining and advanced analytics to support decision-making has grown considerably in the past decades. The development and implementation of complex data mining and analytics projects requires well-defined methodologies and processes. A number of guidelines and standard process models to conduct and manage data mining projects have emerged in the past two decades. Among the existing standard process models for data mining, the most widely adopted one is CRISP-DM – the Cross-Industry Standard Process for Data Mining. CRISP-DM is industry-agnostic. It provides generic guidelines applicable to data mining projects across any industry, but it does not take into account specific requirements or constraints that arise in specific industry sectors. Accordingly, CRISP-DM is often adapted to meet sector-specific requirements. Industry-specific adaptations of CRISP-DM have been proposed across a number of domains, including healthcare, education, industrial and software engineering, logistics, etc. However, to the best of our knowledge, there is no adaptation of CRISP-DM for guiding and structuring data mining projects in the financial services industry, which has its own set of domain-specific requirements. This PhD Thesis addresses the gap by designing, developing, and evaluating a sector-specific data mining process model for financial services.

To this end, we adopted a Design Science Research Methodology (DSRM) in combination with a funnel approach. Initially, we investigated, by reviewing a broad range of academic and professional literature, how standard data mining process models are used across various industry sectors and in the financial services. From this study, we elucidated a number of adaptation patterns of the standard data mining processes. This examination suggested that these approaches do not pay sufficient attention to deployment issues, which play a prominent role when turning data mining models into software products integrated into the IT architectures and business processes of organizations. It was also concluded that refinement of existing guidelines aimed at combining data, technological, and organizational aspects, could help to mitigate these gaps. In the same vein, in the financial services domain the main discovered adaptation scenarios concerned technology-centric aspects (scalability), business-centric aspects (actionability) and human-centric aspects (mitigating discriminatory effects) of data mining.

Next, we conducted a case study in a financial services organization to elucidate how standardized data mining processes are adapted by practitioners. The results revealed 18 perceived gaps in CRISP-DM alongside their perceived impact and mechanisms employed by practitioners to address these gaps.

Using the data and results of the literature review and case study, we designed and developed the Financial Industry Process for Data Mining (FIN-DM). FIN-DM adapts and extends CRISP-DM to support privacy-compliant data mining, to tackle AI ethics risks, to fulfill model risk management requirements, and to embed quality assurance as an integral part of the data mining project life-cycle.

The utility of the framework has been ensured via design iterations conducted in collaboration with data mining and IT practitioners in an actual financial services organization.

FIN-DM is intended to support practitioners working with data mining projects in the financial services and provides a guide for scaling and industrializing data mining functions in this sector. Beyond its applicability in the financial services sector, the thesis contends that some elements in FIN-DM are applicable in the context of other sectors where similar challenges arise, such as the telecommunications sector.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI Artificial Intelligence
BI Business Intelligence
BU Business Understanding phase of CRISP-DM
COBIT Control Objectives for Information Technologies
CRISP-DM Cross-industry Standard Process for Data Mining
CRM Customer Relationship Management
DevOps Development and Operations
DF Design Features
DM Data Mining
DSRM Design Science Research Methodology
DU Data Understanding phase of CRISP-DM
EEA European Economic Area
EU European Union
FIN-DM Financial Industry Process for Data Mining
GDPR General Data Protection Directive
IIA Institute of Internal Auditors
IS Information Systems
ISACA Information Systems Audit and Control Association®
IT Information Technology
ITIL Information Technology Infrastructure Library
ITSM Information Technology Service Management
KDD Knowledge Discovery in Databases
ModelOps Model Operations
MR Meta-requirements
PSD Payment Service Directive
QA Quality Assurance
R Requirements
RACI Responsible, Accountable, Consulted, Informed
RO Research objective
RQ Research question
SEMMA Sample, Explore, Modify, Model, Assess
SLR Systematic Literature Review

1. INTRODUCTION

1.1. Research Motivation

Over the past decades, data mining practices have been widely adopted among organizations seeking to maintain and enhance their competitiveness and business value [DH17]. This trend has led a number of large organizations to manage a rich portfolio of data mining projects [DH17]. The successful development, implementation, and management of data mining projects in such organizations requires a structured and repeatable approach. Accordingly, academic and industry practitioners have proposed several guidelines and standard processes for conducting data mining projects [MMF10], most notably CRISP-DM[1] - a standard process that captures a wide range of recurrent data mining tasks and deliverables structured around a project's life-cycle [MMS09].

CRISP-DM is industry-agnostic. Organizations that wish to use CRISP-DM often need to adapt it to meet their domain-specific requirements [PDM19]. Accordingly, adaptations of CRISP-DM have been developed in the fields of healthcare [Nia15], education [TVP17], industrial engineering [Sol02], [Hub+19], software engineering [Mar+07], [Mar+09], logistics [RDW11], supply chains [Xia09], and e-commerce [HF09], [BKC11].

Despite the vast number of adaptations of CRISP-DM, to the best of our knowledge, there is no adaptation of CRISP-DM for guiding and structuring data mining projects in the financial services industry. Yet, just like the aforementioned industry sectors, the financial services sector has its own set of domain-specific requirements, as well as challenges (cf. [Cao21], [CYY21]). Apart from addressing sector-specific business problems, a broad set of challenges inherent in financial services relates to governance, risk and compliance, regulatory requirements and ethics (cf. [Cao21], [Cao20]). Governance, risk and compliance demands assurance for integrity, efficiency, fairness and adequate risk management in any financial services organization. In turn, ethics associated challenges include addressing privacy and ethic concerns (cf. [Cao21], [Cao20]).

1.2. Research Objectives

This PhD Thesis addresses the lack of data mining process models adapted for the specific needs of the financial services sector. To this end, we implement a funnel approach, starting at the 'macro' level and gradually adopting a 'micro' perspective.

In this setting, the objectives of this PhD Thesis are:

1. to shed light into how standardized data mining process models are used

[1] Cross-industry standard process for data mining

in industry settings and, specifically, in the financial services domain[2] - 'macro-level' - Research Objective 1 (RO1).

2. to identify perceived gaps and the subsequent workaround mechanisms used by practitioners to address these gaps when applying standardized data mining processes - 'micro-level' - Research Objective 2 (RO2).
3. to propose an adapted sector-specific data mining process model (artifact) for the financial services industry - FIN-DM[3] - Research Objective 3 (RO3), and
4. to evaluate FIN-DM with potential users to confirm its practical utility, relevance, and users' acceptance - Research Objective 4 (RO4).

To address the research objectives, three overarching research questions are formulated in this PhD work. Initially, to investigate how standardized data mining models are used in industry settings (RO1), the research question is defined as:

RQ1: how are data mining methodologies applied by researchers and practitioners, both in their generic (standardized) form and in specialized settings? The research question is tackled on 'macro' level– reviewing the reported findings across a broad range of academic and professional literature.

As a next step, to improve the understanding of how standardized data mining processes are extended and adapted in actual practice (RO2), the research question is formulated as follows:

RQ2: What are the perceived gaps in the standardized data mining process, and what adaptations and workaround mechanisms are used by practitioners to address these gaps?

This research question is tackled on 'micro' level, and the case study is conducted in an actual financial services organization.

Next, to fill in the research gap and to propose a sector-specific data mining process model for the financial services industry (RO3, RO4), the following research question is tackled:

RQ3: How can the identified gaps be cohesively addressed within an extension of an existing standardized data mining process, namely CRISP-DM?

[2]The term *financial services* and *banking* are used interchangeably throughout this Thesis. They refer to: (1) traditional businesses providing universal banking and insurance products and services (e.g. lending, transactions, capital markets, asset management, etc.) to all types of clientele (private, corporate, financial institutions and firms), and (2) niche players, disruptors (FinTech, monoline banks etc.) specialized in specific banking, insurance products and services

[3]Financial Industry Process for Data Mining

1.3. Research Methodology

1.3.1. Design Science Research and Methods

Substantial part of Information Systems research is conducted based on two complementary paradigms - behavioral and designs science (cf. [Hev+04], [HC10]). The behavioral paradigm has a strong orientation towards theory development, while design science is an inherently problem-solving paradigm [HC10]. For instance, IS behavioral research frequently studies an IT artifact[4], implemented in an organizational context, predicting or explaining the artifact use, its usefulness, impacts on individuals and organizations, and similar factors based on theories [Hev+04]. In contrast, design science research concentrates on the development and evaluation of IT artifacts, which are usually intended to solve identified organizational problems [Hev+04]. Further, artifact utility and performance are put at the core of the evaluation. The knowledge gained in such research process highlights, for instance, how the artifact can more efficiently solve the problem (cf. [Ven06], [HC10]).

There are many approaches in design science research (cf. [OH11], [Ost14] [PTN18]) that support various research objectives (as presented in Figure 1). The diverse pool of methods, design science research standards, and guidelines has not been systematized nor consolidated except for a limited number of studies such as [OH11], [Ost14], [Alt+12], [PTN18] and partially [VPB17]. To this end, there is lack of validated and widely accepted uniform methodology to carry out design science research [Alt+12], [Ost14]. Hence, the execution of design science projects is regarded as difficult and costly [PTN18]. The issue can be mitigated by providing adequate rationale for the method selection and research design [PTN18].

[PTN18], [Ost14] distinguished three broad directions in design science guidelines and methods, and they are briefly reviewed here. Figure 1 below provides a complete overview of the design science research methodologies starting from 1970s (retrieved from [PTN18], [Ost14]). In the figure, the respective author and associated publication reference are mapped along the timeline. We will focus on the works representing the key directions distinguished by [PTN18], [Ost14]; these are marked by red arrows and numbered. The first is Design Theory direction, as represented by IS Design Theory (ISDT) (point 1 on Figure 1). It gravitates towards the behavioral paradigm and is concerned with developing IS design theories, while its instantiation[5] is optional [PTN18]. The other well-known approach in this domain is Explanatory Design Theory (EDT) (point 2 in Figure 1). It is a subset of design theory concerned with investigating alternative designs and their

[4] IT artifacts are not limited to computer-based IS, but also encompass designed artifacts (designated as solution technologies [Ven06]. These might include IS development methods, techniques, tools, IS planning methods, IS management methods, IS/IT security, and risk management practices, algorithms, and the like [Ven06]

[5] Design theory instantiation, in the form of expository or representational tool, is considered to serve communicative purpose in illustrating design principles [PTN18]

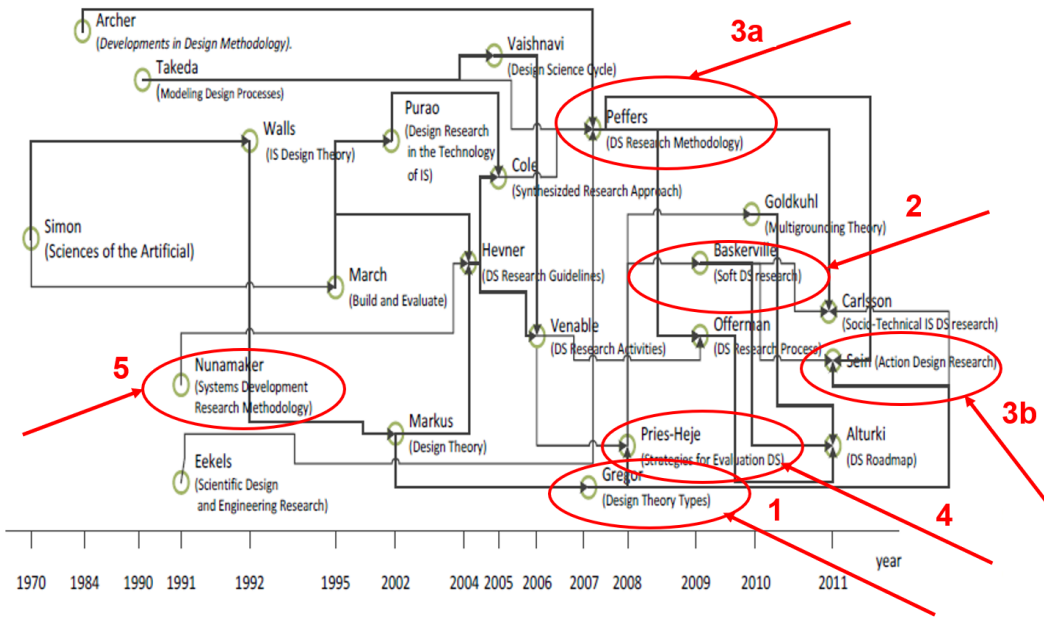


Figure 1. Design Science research methodologies [OH11], [Ost14]

impacts [PTN18] and then developing corresponding design theory. Here, artifact instantiation is not required, yet it is typically needed for manipulation of design alternatives, and artifact evaluation [PTN18].

In contrast, the alternative Methods direction, represented by Design Science Research Methodology (DSRM) (point 3a in Figure 1), puts forward the design and development of artifacts applicable and useful in practical settings of organizational IS [PTN18]. Another well-known approach in Methods research space is Action Design Research (ADR), represented by Sein, et al. (point 3b in the Figure 1). Similarly to DSRM, it is concerned with artifacts and how organizational perspective impacts their construction and use [PTN18]. ADR integrates action research with design science and focuses on evaluating artifacts in practical organizational settings [PTN18]. [Ost14] distinguished such combinations of design science and other research guidelines as a separate research direction.

[PTN18] also differentiated Design-oriented research (DOIS) concerned with designing well-performing artifacts and their utility. As a result, artifact evaluation methods are emphasized in this research direction (point 4 in Figure 1). To this end, [Ost14] also distinguished one more research direction associated with the System Development life-cycle (point 5 in Figure 1). It later converged into Methods direction, and DSRM methodology in particular (point 3a in Figure 1).

Given that the research objectives of this PhD work is to design and develop an artifact, the guidelines and approaches belonging to the Methods design science research area are the most appropriate.

1.3.2. Selection of Design Science Research Method

As mentioned, there is little practical guidance on the selection of the most suitable design science method, except for [VPB17]. The latter provides practical guidance (in the form of technological rules) on the choice of the most suitable design science approach given the research objectives and settings. [VPB17] differentiated and compared the most common design science research methods based on a broad IS research paradigm categorization as Objectivist/ Positivist or Subjectivist/ Interpretative (cf. [HK89], [Ada14]), and other key research characteristics like objective, domain, scope, etc. The key principle of the Objectivist/ Positivist research paradigm, so far dominating behavioral IS research, is that the researcher is independent of the phenomenon that is being investigated [Ada14] and the purpose of the research is to design an artifact that achieves objectively defined results [VPB17]. In contrast, in the interpretivist paradigm, researchers engage in the social setting investigated and learn how the interaction takes place from the participants' perspective [Ada14]. This research approach attempts to understand phenomena through the meanings and interpretations that people assign to them [Ada14].

[VPB17] recommends Objectivist/ Positivist-based methods as the preferred choice to Subjectivist/ Interpretivist approaches if: (1) the research goal is to achieve the best and most suitable artifact, and (2) research results have to be objective [VPB17]. Overall, three design science research methods are classified into the Objectivist/Positivist class. Then, based on the set of technological rules, [VPB17] recommends to:

- apply Systems Development Research Methodology (SDRM) if the artifact outcome of the research should be an IT system, or
- adopt Design Science Research Methodology (DSRM) if extensive adaptation of artifact to daily use is needed, or
- choose DSR Process Model (DSRPM) if the goal of the research to develop design theory.

This PhD research aims not only to develop an artifact, but to assess and ensure its utility, applicability, and relevance in application domain. This implies an extensive adaptation of the artifact to daily use in practical settings. Hence, DSRM [Pef+08] was chosen as the most suitable research method (presented in Figure 2).

DSRM is a six-step process model allowing for iterations. It starts with *Problem Identification and Motivation*, which aims to define the research problem and rationalize the value and significance of the solution [Pef+08]. As a next step, *Define Objectives of a Solution*, the objectives for the artifact, either quantitative (e.g. improvements over existing solutions) or qualitative (e.g. how the artifact is expected to provide solutions to problems), are derived. Objectives have to be calibrated to what is possible and feasible to achieve [Pef+08]. Then, in the *Design and Development* step, the artifact is constructed. These can be any design objects, for example constructs, models, methods, instantiations, or new properties

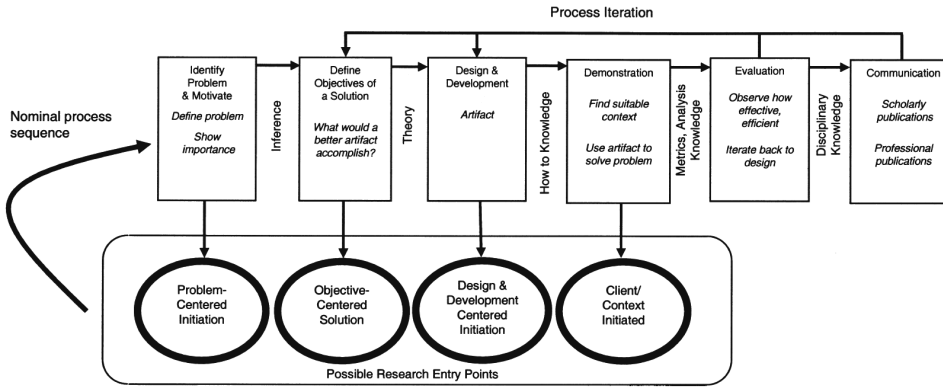


Figure 2. Design Science Research Methodology (as in [Pef+08])

thereof [Pef+08]. In this step, initially, the desired functionality and architecture of the artifact are determined. Then, its prototype is created [Pef+08]. Then, in *Demonstration* step, the artifact's use and how the problem is solved are presented. Typically, a demonstration is done by means of experiment, simulation, case study, or any other applicable method [Pef+08]. Next, the formal *Evaluation* is executed– it aims to gauge how well the artifact assists in solving the problem. Similarly to the *Demonstration*, it can take many appropriate forms, e.g. comparison of artifact functionalities to solution objectives, performance measures, etc. [Pef+08] Based on the evaluation results, the artifact is either reiterated and improved, or *Communication* step follows, while any other potential improvements are left for future research projects [Pef+08].

To contextualize the choice of DSRM with the methods used in related works, it was identified that the majority of related studies, including sector-specific adaptations (cf. [Nia15], [Mar+07]), follow a common step-by-step approach to design and develop novel data mining processes. Initially, based on the literature and in-depth domain reviews, improvement needs in the standard data mining process are identified and then solutions are proposed. Typically, studies neither evaluate newly proposed process models with the potential users nor calibrate to the users' needs. If the evaluation is executed, it is primarily focused on discussing data mining results and project outcome, but not on the proposed process itself. Rarely, studies follow general design science principles (cf. [Hub+19]), and only one domain adaptation followed a specialized design science research method [TVP17]. In this context, using a concrete design science method could provide a number of benefits and extend existing practices to derive sector-specific adaptations. In this PhD research setting, using DSRM assists in: (1) supporting artifact design and development in a structured manner (RO3), and (2) providing for its evaluation with the potential users (RO4) to confirm the artifact's utility, relevance and acceptance.

1.3.3. DSRM to Design and Develop FIN-DM Artifact

In applying DSRM, we defined and executed four iterative cycles, as presented graphically in Figure 3 below.

Cycle 1 - Problem and Objectives Formulation In this cycle, a complete nominal sequential order of DSRM was followed, starting with the very first activity (A1.1) of a *Problem-centered initiation* [Pef+08]. This is the recommended procedure when a research problem is already observed or suggested being examined in prior research, but the artifact does not yet exist [Pef+08]. In *Problem-centered initiation* Research Questions 1 and 2 were addressed by three separate studies. RQ1 was tackled by conducting two Systematic Literature Reviews (SLR I, SLR II). The first cross-domain SLR investigated standard data mining approaches usage across a number of industries (SLR I - [PDM20]), while the second sector-specific one concentrated on the financial services industry (SLR II - [PDM19]). RQ2 was addressed in the third study - an industrial case study that identified and reported perceived 'gaps' in the CRISP-DM process when conducting data mining projects within the actual financial services organization [PDM21].

Then, we proceeded with data collection and processing activity (A1.2) using data originating from these works as *Input 1* (Figure 3). The data and results of all the three studies were combined and triangulated, and a consolidated catalog of the standardized data mining process gaps was constructed (*Consolidate Outcome 1* in Figure 3).

Cycle 2 - Artifact Design and Development (Figure 3) For this cycle, we have adopted a blended approach by extending DSRM with the artifact requirements concept, and such research approach is described and motivated in detail in section *5.2 FIN-DM Conceptualization*. An explicit elicitation of requirements and design principles was conducted as activity 2.1 (A2.1). To this end, based on the analysis of consolidated 'gaps' (*Consolidated Input 2*), we specified the *Research Outcome 2* containing three items: (1) artifact objectives, (2) artifact (meta)-requirements, and (3) design principles to address the given requirements. To ensure the consistency and preciseness of the requirements and design principles, the template proposed in [CSG15] was used. Then, based on *Outcome 2*, we determined the desired functionalities of the artifact and developed the first prototype as activity 2.2 (A2.2) with *Outcome 3* (see Figure 3).

Cycle 3 - Evaluation Next, the artifact prototype was subjected to an *Evaluation cycle*, which comprised three distinct activities (see Figure 3). First, the evaluation approach was designed and planned as activity 3.1 (A3.1). Subsequently, it was executed as activity 3.2 (A3.2), and collected data was analyzed in activity 3.3 (A3.3).

In activity A3.1, the design of evaluation was detailed in *Research Outcome 4* (see Figure 3). Key evaluation design items concerning the criteria, the method, the organizational settings, and the instruments are specified in detail in the section *7.2 Evaluation Design*.

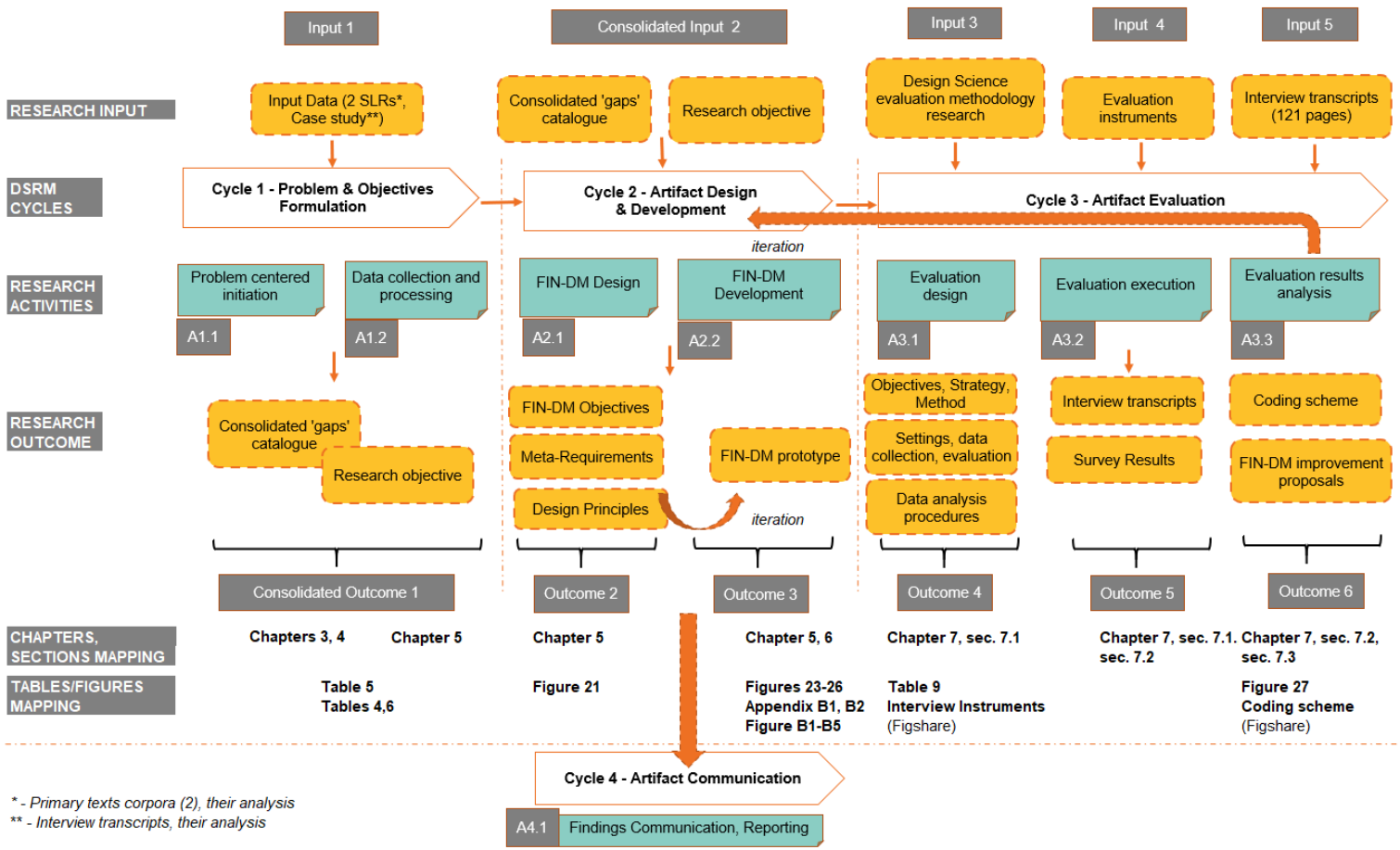


Figure 3. Research process

Artifact evaluation focused on examining it from users' perspective. Initially, a common set of evaluation criteria was derived based on the relevant artifact evaluation methods and models. The criteria set was subsequently broken down into lower-level constructs. *Ex-Ante, Naturalistic evaluation* [CSG15] was conducted by the means of two methods, starting with the combined individual demonstration session followed by *semi-structured interviews*, and a *questionnaire*. An interview guide and questionnaire instruments were constructed as *Input 4* available at link[6] towards Activity 3.2 (A3.2 *Conducting evaluation*).

As activity 3.3. (*Evaluation results analysis*), interviews were transcribed and evaluated jointly with the questionnaire results as *Input 5* to activity 3.3 (A3.3). Interviews' transcripts were coded iteratively based on the methods proposed by [Sal15][7]. The coded interview responses were categorized into four distinct categories (*Minor, Medium, Major* and *Critical*) based on the suggested improvements' complexity and impact. The detailed taxonomy for users' proposals is presented and discussed in detail in section *7.3 Suggested Improvements to FIN-DM*. The suggested improvements were iterated back into the ***Artifact design and development cycle***, which was repeated to produce its final, improved version demonstrated in ***Cycle 4 - Artifact Communication***.

1.4. Outline

The thesis is structured as follows. In Chapter 2, the key concepts are introduced and a general overview of the data mining process models and methodologies is presented. In Chapter 3 and Chapter 4, the first key phase of DSRM approach, Problem Centered Initiation, is provided. In particular, Chapter 3 presents two 'macro' level studies - Systematic Literature Reviews (SLR I and SLR II), their research design, implementation, and results. This chapter addresses RQ1, *How are data mining methodologies applied?*. It is followed by the 'micro' perspective– the industrial case study in the financial services organization (Chapter 4) presented in the same manner as SLRs. This chapter tackles RQ2, *What are perceived gaps in the standard data mining process?*, and covers the case study design, its implementation, and results. In Chapter 5, the design and development of the artifact, FIN-DM, is presented, focusing on key method activities - data triangulation, constructing gaps catalog and deriving requirements towards new artifact. Chapter 6 follows with presentation of the proposed FIN-DM Process Model emphasizing solutions to gaps, while Chapter 7 is devoted to FIN-DM Ex-Ante Evaluation with the potential users, its design, implementation and results. In this manner, Chapter 5-7 tackles RQ3, *How to address the identified gaps with extension of CRISP-DM?*. Lastly, Chapter 8 concludes by considering contributions, limitations, and future research.

[6] `https://figshare.com/s/1bda7ccadaa254fcabe1`

[7] The initial and final coding schemes are available on `https://figshare.com/s/d2bf5084f3e6cfc1af80`

2. BACKGROUND

This chapter introduces main data mining concepts and provides an overview of existing data mining process models and methodologies and their evolution.

2.1. Key Concepts

Data mining is defined as a set of rules, processes, and algorithms that are designed to generate actionable insights, extract patterns, and identify relationships from large data sets [Mor16]. Data mining incorporates automated data extraction, processing, and modeling by means of a range of methods and techniques. In contrast, data analytics refers to techniques used to analyze and acquire intelligence from data (including 'big data') [GH15]. It is positioned as a broader field, encompassing a wider spectrum of methods that includes both statistical and data mining [CCS12]. A number of algorithms have been developed in statistics, machine learning, and artificial intelligence domains to support and enable data mining.

Data mining projects commonly follow a structured process or methodology, as exemplified by [MMF10], [MMS09]. A process model is defined as the set of framework activities and tasks to be performed for developing the particular elements, inputs, and outputs of each task. Methodology is a process model instance specifying tasks, inputs, and outputs, as well as the way in which the tasks must be carried out [MMF10]. Thus, data mining methodology provides a set of guidelines for executing a set of tasks to achieve the objectives of a data mining project [MMF10].

In the past decade we witness the emergence of data science as a broader discipline (cf. [Cao17], [Cao16]), which focuses on a more systematic and complex view, transforming data to intelligence for decision-making in the settings of complex systems and organizations [Cao17]. This PhD Thesis takes a narrower perspective of data mining as a set of techniques to generate actionable insights, extract patterns, etc. i.e. acting as a tactical tool to address concrete business problems.

2.2. Data Mining Methodologies and Process Models

The foundations of structured data mining methodologies were first proposed by [FPS96a], [FPS96c], [FPS96b], and were initially related to Knowledge Discovery in Databases (KDD). KDD presents a conceptual process model of computational theories and tools that support information extraction (knowledge) with data [FPS96a]. In KDD, the overall approach to knowledge discovery includes data mining as a specific step. As such, KDD, with its nine main steps (exhibited in Figure 4 below), has the advantage of considering data storage and access, algorithm scaling, interpretation and visualization of results, and human computer interaction [FPS96a], [FPS96b].

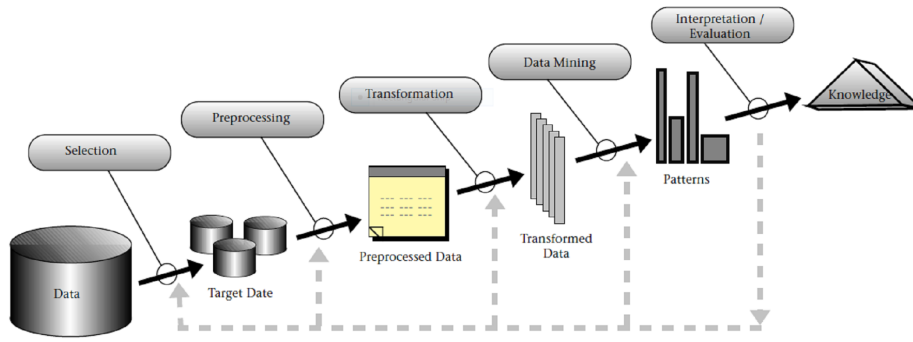


Figure 4. An overview of the steps composing the KDD process, as presented in [FPS96a], [FPS96b]

The main steps of KDD are as follows:

- Step 1 - Learning application domain: the first step involves developing understanding of the application domain and relevant prior knowledge; it is followed by identifying the goal of the KDD process from the customer's viewpoint.
- Step 2 - Data set creation: the second step involves selecting a data set and focusing on a subset of variables or data samples to perform discovery.
- Step 3 - Data cleaning and processing: in the third step, basic operations to remove noise or outliers are conducted. Necessary information to model or account for noise, decisions on strategies for handling missing data fields, data types, schema, and mapping of missing and unknown values are also considered.
- Step 4 - Data reduction and projection: here, the work of finding useful features to represent the data depending on the goal is performed. Also, transformation methods are used to find the optimal features set for the data.
- Step 5 - Choosing the function of data mining: in the fifth step, the target outcome (e.g., summarization, classification, regression, clustering) are defined.
- Step 6 - Choosing data mining algorithm: the sixth step concerns selecting method(s) to search for patterns in the data. Decisions on which appropriate models and parameters are taken.
- Step 7 - Data mining: In the seventh step, the work of mining the data i.e., searching for patterns of interest in a particular representational form or a set of such representations (classification rules or trees, regression, clustering, etc.) is conducted.
- Step 8 - Interpretation: In this step, the redundant and irrelevant patterns

are removed while relevant ones are interpreted and visualized to make the result understandable to the users.

- Step 9 - Using discovered knowledge: In the last step, the results are incorporated with the performance system, documented and reported to stakeholders, and used as a basis for decisions.

The KDD process became dominant in industrial and academic domains [KM06], [MMS09]. As the timeline-based evolution of data mining methodologies and process models shows (Figure 5 below), the original KDD model served as basis for other methodologies and process models, which addressed its various gaps and deficiencies. These approaches extended the initial KDD framework, yet the extension degree has evolved, ranging from process restructuring to complete change in focus. For example, [BA96] and further [GD04] (in the form of a case study) introduced practical adjustments to the process based on iterative nature of process as well as interactivity. The complete KDD process, in their view, was enhanced with supplementary tasks and the focus was changed to the user's point of view (human-centered approach), highlighting decisions that need to be made by the relevant stakeholders in the course of a data mining project. In contrast, [Cab+97] proposed a different number of steps emphasizing and detailing data processing and discovery tasks. Similarly, in a series of works [AB98], [Ana+98], [Buc+99] presented additional data mining process steps by concentrating on the adaptation of data mining process to practical settings. They focused on cross-sales (entire life-cycles of online customer), with further incorporation of internet data discovery processes (web-based mining). Furthermore, the Two Crows data mining process model is a consultancy originated framework that has defined the steps differently, but is still close to the original KDD. Finally, SEMMA (Sample, Explore, Modify, Model and Assess) based on KDD, was developed by the SAS institute in 2005 [SAS17]. It is a logical organization of the functional toolset of SAS Enterprise Miner for carrying out the core tasks of data mining. Compared to KDD, this is a vendor-specific process model, which limits its application in different environments. Also, it skips two steps of the original KDD process ('Learning Application Domain' and 'Using of Discovered Knowledge'), which are regarded as essential for the success of a data mining project [MMF10]. In terms of adoption, new KDD-based proposals received limited attention across academia and industry [KM06], [MMS09]. Subsequently, most of these methodologies converged into the CRISP-DM methodology.

Additionally, there have only been two non-KDD based approaches proposed alongside extensions to KDD. The first one is 5A's approach presented by [Pis03] and used by an SPSS vendor. The key contribution of this approach is related to adding an 'Automate' step, while the disadvantages were associated with omitting the 'Data Understanding' step. The second approach was 6-Sigma, which is an industry originated method to improve quality and customer's satisfaction [PK03]. It has been successfully applied to data mining projects. In 2000, as a response to

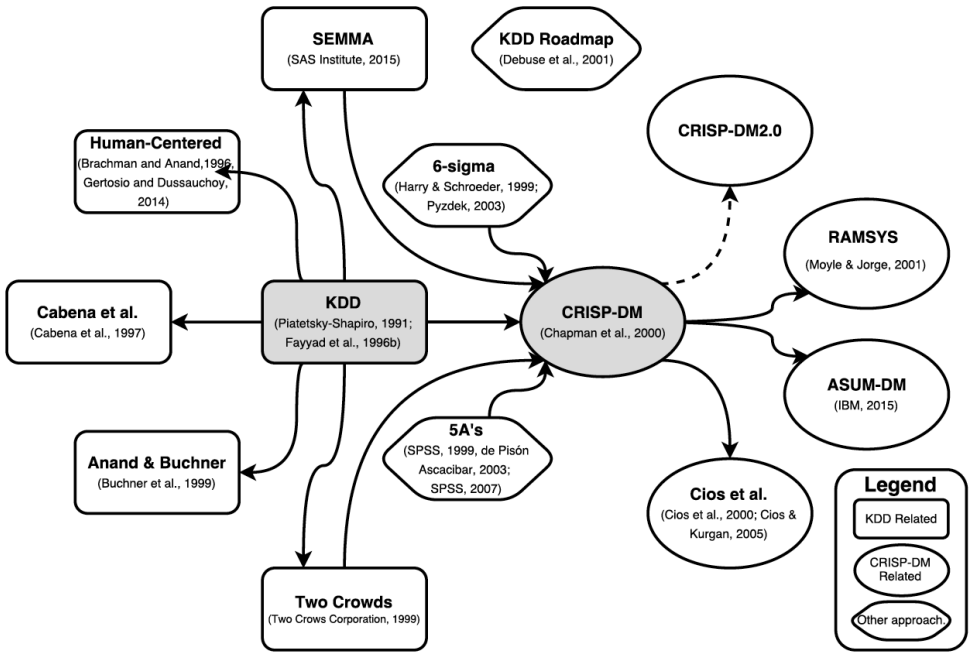


Figure 5. Evolution of data mining process and methodologies, as presented in [Mar+17]

common issues and needs [MMS09], an industry-driven methodology called the Cross-Industry Standard Process for Data Mining (CRISP-DM) was introduced as an alternative to KDD. It also consolidated the original KDD model and its various extensions. While CRISP-DM builds upon KDD, it consists of six phases that are executed in iterations [MMS09]. The iterative executions of CRISP-DM stand as CRISP-DM's most distinguishing feature. It contrasts with initial KDD that assumes a sequential execution of its steps. CRISP-DM, much like KDD, aims at providing practitioners with guidelines to perform data mining on large data sets. However, CRISP-DM, with its six main steps with a total of 24 tasks and outputs, is more refined than KDD. The main steps of CRISP-DM, as depicted in Figure 6 below, are as follows:

- Phase 1- Business understanding: The focus of the first step is to gain an understanding of the project's objectives and requirements from a business perspective. These are then converted into data mining problem definitions. A presentation of a preliminary plan to achieve the objectives is also included in this first step.
- Phase 2 - Data understanding: This step begins with an initial data collection and proceeds with activities in order to get familiar with the data, identify data quality issues, discover first insights into the data, and potentially detect and form hypotheses.
- Phase 3 - Data preparation: The third step covers activities required to

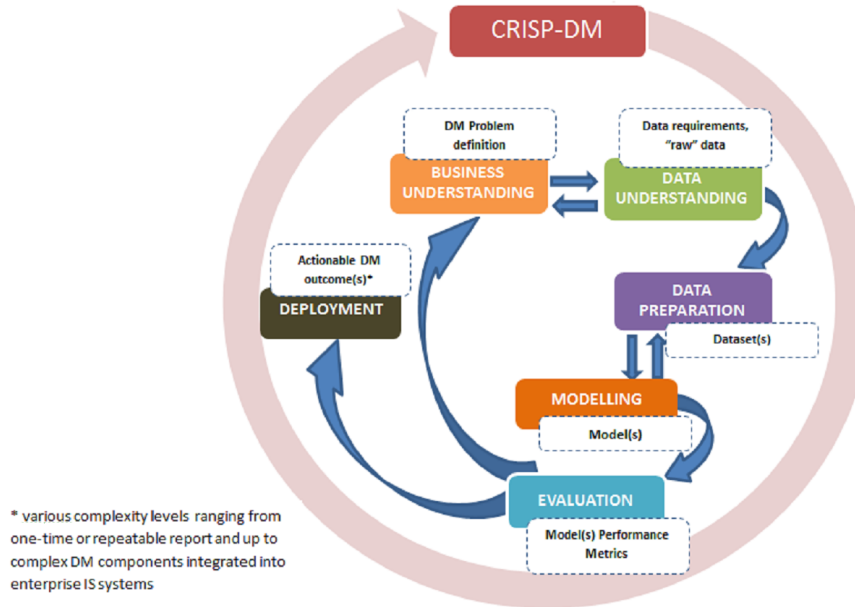


Figure 6. CRISP-DM phases and key outputs (adapted from [Cha+00])

construct the final data set from the initial raw data. Data preparation tasks are performed repeatedly.

- Phase 4 - Modeling phase: In this step, various modeling techniques are selected and applied, followed by a calibration of their parameters. Typically, several techniques are used for the same data mining problem.
- Phase 5 - Evaluation of the model(s): The fifth step begins with the quality perspective and then, before proceeding to final model deployment, ascertains that the model(s) achieves the business objectives. At the end of this phase, a decision should be reached on how to use data mining results.
- Phase 6 - Deployment phase: In the final step, the models are deployed to enable end-customers to use the data as basis for decisions or for support in the business process. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized, presented, and distributed in a way that the end-user can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

The development of CRISP-DM was led by an industry consortium. It is designed to be domain-agnostic [MMF10] and, as such, CRISP-DM is now widely used by industry and research communities [MMS09]. These distinctive characteristics have made CRISP-DM to be considered as the 'de-facto' standard of data mining methodology and as a reference framework to which other methodologies

are bench-marked [MMF10].

Similarly to KDD, a number of refinements and extensions of the CRISP-DM methodology have been proposed with the two main directions - extensions of the process model itself and adaptations and a merger with the process models and methodologies in other domains. Extensions in the direction of process models could be exemplified by [CK05], who have proposed the integrated DMKD (Data Mining and Knowledge Discovery) process model. It contains several explicit feedback mechanisms and a modification of the last step to incorporate discovered knowledge and insights. Also, DMKD relies on technologies for results deployment (see Table 1). In the same vein, [MJ01], [BM02] proposed the Rapid Collaborative Data Mining System (RAMSYS) framework - this is both a data mining methodology and a system for remote collaborative data mining projects (presented in Table 1). RAMSYS attempted to combine a problem-solving methodology, with knowledge sharing, and ease of communication. It intended to promote collaborative work. In particular, remotely placed data miners collaborate in a disciplined manner as regards information flow while the free flow of ideas occurs for problem-solving [MJ01]. CRISP-DM modifications and integrations with other specific domains were proposed in Industrial Engineering (Data Mining for Industrial Engineering by [Sol02]) and Software Engineering by [Mar+07], [Mar+09]. Both approaches enhanced CRISP-DM and contributed with additional phases, activities, and tasks typical for engineering processes and addressing ongoing support [Sol02], and project management, organizational, and quality assurance tasks [Mar+09].

Finally, a limited number of attempts to create independent or semi-dependent data mining frameworks was undertaken after CRISP-DM creation (see Table 1). These efforts were driven by industry players and comprised the KDD Roadmap by [Deb+01] for a proprietary predictive toolkit (Lanner Group). There is also the recent effort by IBM with Analytics Solutions Unified Method for Data Mining (ASUM-DM) in 2015 [IBM16][1]. Both frameworks contributed with additional tasks, e.g. resourcing in KDD Roadmap, or a hybrid approach assumed in ASUM, e.g., a combination of agile and traditional implementation principles.

The Table 1 below summarizes the reviewed data mining process models and methodologies by their origin, basis, and key concepts.

[1] `https://developer.ibm.com/technologies/artificial-intelligence/articles/architectural-thinking-in-the-wild-west-of-data-science/`

Table 1. Key aspects of existing Data Mining process models and methodologies

Name	Origin	Basis	Key concept	Year
Human-Centered	Academy	KDD	Iterative process and interactivity (user's point of view and needed decisions)	1996, 2004
Cabena et al.	Academy	KDD	Focus on data processing and discovery tasks	1997
Anand and Buchner	Academy	KDD	Supplementary steps and integration of web-mining	1998, 1999
Two Crows	Industry	KDD	Modified definitions of steps	1998
SEMMA	Industry	KDD	Tool-specific (SAS Institute), elimination of some steps	2005
5A's	Industry	Independent	Supplementary steps	2003
6 Sigmas	Industry	Independent	6 Sigma quality improvement paradigm in conjunction with DMAIC performance improvement model	2003
CRISP-DM	Joint industry and academy	KDD	Iterative execution of steps, significant refinements to tasks and outputs	2000
Cios et al.	Academy	CRISP-DM	Integration of data mining and knowledge discovery, feedback mechanisms, usage of received insights supported by technologies	2005
RAMSYS	Academy	CRISP-DM	Integration of collaborative work aspects	2001-2002
DMIE	Academy	CRISP-DM	Integration and adaptation to Industrial Engineering domain	2001
Marban	Academy	CRISP-DM	Integration and adaptation to Software Engineering domain	2007
KDD roadmap	Joint industry and academy	Independent	Tool-specific, resourcing task	2001
ASUM	Industry	CRISP-DM	Tool-specific, combination of traditional CRISP-DM and agile implementation approach	2015

3. PROBLEM INITIATION - SYSTEMATIC LITERATURE REVIEWS

In this Chapter, the first activity in the DSRM approach, *Problem Centered Initiation* is presented and discussed (Figure 3 in Chapter 1). This Chapter covers two Systematic Literature Reviews. For each of these two studies, the research design and findings are described. Then, threats to validity for each SLR study are discussed and addressed.

3.1. Cross Domain Systematic Literature Review: Adaptations of Data Mining Process Models (SLR I)

3.1.1. SLR I: Research Design

This part of the PhD thesis research investigates *how are data mining methodologies applied by researchers and practitioners (RQ1)*. To this end, the systematic literature review (SLR) method is used as scientific method for two reasons. Firstly, the systematic review method is based on trustworthy, rigorous, and auditable methodology. Secondly, SLR supports structured synthesis of existing evidence, identification of research gaps, and provides framework to position new research activities [KBB15]. For this SLR, the guidelines proposed by [KBB15] are followed. All SLR details have been documented in the separate, peer-reviewed SLR protocol (available at link[1]).

SLR I Research Questions As suggested by [KBB15], as a first step, research questions are formulated, and we motivate them as follows. In the preliminary phase of research, an exploratory (free-form) search of the literature led us to conclude that there are a number of studies on how data mining methodologies are adapted to fit requirements found in different industry verticals. At the same time, there is no clear overview of the scope of these different adaptations of data mining methodologies. This research gap led to the following question: 'How data mining methodologies are applied ('as-is' vs adapted) (RQI.I)? Further, the nature of the adaptations is different depending on the application domain and requirements, thus giving rise to RQI.II, 'How have existing data mining methodologies been adapted?'. Finally, if adaptions are made, it is beneficial to explore what the associated reasons and purposes are, which in turn led to RQI.III, 'For what purposes are data mining methodologies adapted?'.

Thus, for this SLR review, there are three research questions defined[2]:

[1] `https://figshare.com/articles/Systematic-Literature-Review-Protocol/10315961`

[2] Hereinafter, research questions within respective input studies are tagged with Roman numerals, whereby the first part of the index represents the study number and the second part of the index refers to the question number

- **Research Question I.I: How data mining methodologies are applied ('as-is' versus adapted)?** - this question aims to identify data mining methodologies application and usage patterns and trends.
- **Research Question I.II: How have existing data mining methodologies been adapted?** - this questions aims to identify and classify data mining methodologies adaptation patterns and scenarios.
- **Research Question I.III: For what purposes have existing data mining methodologies been adapted?** - this question aims to identify, explain, classify and produce insights on what are the reasons and what benefits are achieved by adaptations of existing data mining methodologies. Specifically, what gaps do these adaptations seek to fill and what have been the benefits of these adaptations. Such systematic evidence and insights will be valuable input to potentially new, refined data mining methodology. Insights will be of interest to practitioners and researchers.

SLR I Data Collection Strategy The data collection and search strategy followed the guidelines proposed by [KBB15]. It defined the scope of the search, selection of literature and electronic databases, search terms and strings as well as screening procedures.

Primary Search The primary search aimed to identify an initial set of papers. To this end, the search strings were derived from the research objective and research questions. The term 'data mining' was the key term, but 'data analytics' term was also included to be consistent with observed research practices. The terms 'methodology' and 'framework' were also included. Thus, the following search strings were developed and validated in accordance with the guidelines suggested by [KBB15]:

> ('data mining methodology') OR ('data mining framework') OR ('data analytics methodology') OR ('data analytics framework')

The search strings were applied to the indexed scientific databases Scopus, Web of Science (for 'peer-reviewed', academic literature) and to the non-indexed Google Scholar (for non-peer-reviewed, so-called 'grey' literature). The decision to cover 'grey' literature in this research was motivated as follows. As proposed in a number of information systems and software engineering domain publications (e.g. [GFM19], [Net+19]), SLR as a stand-alone method may not provide sufficient insight into 'state of practice'. It was also identified (e.g. in [GFM16]) that 'grey' literature can give substantial benefits in certain areas of software engineering, in particular, when the topic of research is related to industrial and practical settings. Taking into consideration the research objectives, which is investigating data mining methodologies application practices, we have opted for inclusion of elements of Multivocal Literature Review (MLR) [3] in our study. Also, [KBB15]

[3] Multivocal Literature Review (MLR), as in [GFM19], is a form of an SLR which includes the 'grey' literature (e.g., blog posts, videos and white papers) in addition to the published (formal or 'white') literature (e.g., journal and conference papers).

recommends including 'grey' literature to minimize publication bias, as positive results and research outcomes are more likely to be published than negative ones. Following MLR practices, inclusion criteria for types of 'grey' literature were designed, and they are reported below.

The selection of databases is motivated as follows. In case of peer-reviewed literature sources, we concentrated to avoid potential omission bias. The latter is discussed in IS research (e.g. [LE06]) in case research is concentrated in limited disciplinary data sources. Thus, broad selection of data sources including multidisciplinary-oriented (Scopus, Web of Science, Wiley Online Library) and domain-oriented (ACM Digital Library, IEEE Xplorer Digital Library) scientific electronic databases was evaluated. Multidisciplinary databases have been selected due to wider domain coverage, and it was validated and confirmed that they do include publications originating from domain-oriented databases, such as ACM and IEEE. From multi-disciplinary databases as such, Scopus was selected due to the widest possible coverage (it is the world's largest database, covering app. 80% of all international peer-reviewed journals) while Web of Science was selected due to its longer temporal range. Thus, both databases complement each other. The selected non-indexed database source for 'grey' literature is Google Scholar, as it is a comprehensive source of both academic and 'grey' literature publications and referred as such extensively (e.g. [GFM19], [Net+19]).

Further, [GFM19] presented a three-tier categorization framework for types of 'grey literature'. In this study, we restricted scope to the 1st tier 'grey' literature publications of the limited number of 'grey' literature producers. In particular, from the list of producers [Net+19], we focused on government departments and agencies, non-profit economic, trade organizations ('think-tanks') and professional associations, academic and research institutions, businesses and corporations (consultancy companies and established private companies). The 1st tier 'grey' literature selected items include: (1) government, academic, and private sector consultancy reports[4], (2) theses (not lower than Master level) and PhD Dissertations, (3) research reports, (4) working papers, (5) preprints. With inclusion of the 1st tier 'grey' literature criteria, quality assessment challenge especially relevant and reported for it was mitigated (cf. [GFM19], [Net+19]).

Scope and Domains Inclusion As recommended by [KBB15] it is necessary to initially define research scope. To clarify the scope, we have defined what is not included and is out of scope of this research. The following aspects are not included in the scope of the study:

1. context of technology and infrastructure for data mining/data analytics tasks and projects.
2. granular methods application in data mining process itself or their application for data mining tasks, e.g. constructing business queries or applying regression or neural networks modeling techniques to solve classification problems.

[4]Including white papers, market reports, industry overviews and similar

Studies with granular methods are included in primary texts corpus as long as method application is part of overall methodological approach.

3. technological aspects in data mining e.g. data engineering, dataflows and workflows.
4. traditional statistical methods not associated with data mining directly, including statistical control methods.

Similarly to [Bud+06], [LE06], initial piloting revealed that search engines retrieved literature available for all major scientific domains, including ones outside the author's area of expertise (e.g. medicine). Even though such studies could be retrieved, it would be impossible for to analyze and correctly interpret literature published outside the possessed area of expertise. The adjustments toward search strategy were undertaken by retaining domains closely associated with Information Systems, Software Engineering research. Thus, for the Scopus database the final set of inclusive domains was limited to nine and included Computer Science, Engineering, Mathematics, Business, Management and Accounting, Decision Science, Economics, Econometrics and Finance, and Multidisciplinary as well as Undefined studies. Excluded domains covered 11.5% or 106 out of 925 publications; it was confirmed in validation process that they primarily focused on specific case studies in fundamental sciences and medicine [5]. The included domains from Scopus database were mapped to Web of Science to ensure consistent approach across databases, and the correctness of mapping was validated.

Screening Criteria and Procedures Based on the SLR practices (as in [KBB15], [Bre+07]) and defined SLR scope, multi-step screening procedures (quality and relevancy) with associated set of *Screening Criteria* and *Scoring System* were designed. The purpose of relevancy screening is to find relevant primary studies in an unbiased way [Van+11]. Quality screening, on the other hand, aims to assess primary relevant studies in terms of quality in an unbiased way.

Screening Criteria consisted of two subsets - *Exclusion Criteria* applied for initial filtering and *Relevance Criteria*, also known as *Inclusion Criteria*.

Exclusion Criteria were initial threshold quality controls aiming at eliminating studies with limited or no scientific contribution. The exclusion criteria also address issues of understandability, accessibility, and availability. The *Exclusion Criteria* were as follows:

1. Quality 1 - the publication item is not in English (understandability).
2. Quality 2 - publication item duplicates, which can occur when:
    - either the same document retrieved from two or all three databases.
    - or different versions of the same publication are retrieved (i.e. the

[5] Excluded domains were Medicine, Biochemistry, Genetics and Molecular Biology, Environmental Science, Earth and Planetary Science, Physics and Astronomy, Energy and Material Science, Agricultural and Biological Science, Chemistry and Chemical Engineering, Pharmacology, Toxicology and Pharmaceuticals, Arts and Humanities, Neuroscience, Immunology and Microbiology, Health Professions and Nursing

same study published in different sources) - based on best practices, the decision rule is that the most recent paper is retained as well as the one with the highest score [Kof14].

- if a publication is published both as conference proceeding and as journal article with the same name and same authors or as an extended version of the conference paper, the latter is selected.

3. Quality 3 - length of the publication is less than 6 pages - short papers do not have the space to expand and discuss presented ideas in sufficient depth to examine for us.
4. Quality 4 - the paper is not accessible in full length online through the university subscription of databases and via Google Scholar - not full availability prevents us from assessing and analyzing the text.

The initially retrieved list of papers was filtered based on *Exclusion Criteria*. Only papers that passed all criteria were retained in the final studies' corpus. Mapping of criteria towards screening steps is exhibited in Figure 7 below.

Relevance Criteria were designed to identify relevant publications and are presented in Table 2 below while mapping to respective process steps is presented in Figure 7 below. These criteria were applied iteratively.

Table 2. *Relevance Criteria* mapping to screening process steps

Relevance Criteria	Criteria Definition	Criteria Justification
Relevance 1	Is the study about data mining or data analytics approach and is within designated list of domains?	Exclude studies conducted outside the designated domain list. Exclude studies not directly describing and/or discussing data mining and data analytics
Relevance 2	Is the study introducing/describing data mining or data analytics methodology/framework, or modifying existing approaches?	Exclude texts considering only specific, granular data mining and data analytics techniques, methods or traditional statistical methods. Exclude publications focusing on specific, granular data mining and data analytics process/subprocess aspects. Exclude texts where description and discussion of data mining methodologies or frameworks is manifestly missing

As a final SLR step, the full texts quality assessment was performed with constructed *Scoring Metrics* (in line with [KC07]). It is presented in the Table 3 below.

Table 3. *Scoring Metrics*

Score	Criteria Definition
3	Data mining methodology or framework is presented in full. All steps described and explained, tests performed, results compared and evaluated. There is clear proposal on usage, application, deployment of solution in organization's business process(es) and IT/IS system, and/or prototype or full solution implementation is discussed. Success factors described and presented
2	Data mining methodology or framework is presented, some process steps are missing, but they do not impact the holistic view and understanding of the performed work. Data mining process is clearly presented and described, tests performed, results compared and evaluated. There is proposal on usage, application, deployment of solution in organization's business process(es) and IT/IS system(s)
1	Data mining methodology or framework is not presented in full, some key phases and process steps are missing. Publication focuses on one or some aspects (e.g. method, technique)
0	Data mining methodology or framework not presented as holistic approach, but on fragmented basis, study limited to some aspects (e.g. method or technique discussion, etc.)

Data Extraction and Screening Process The conducted data extraction and screening process is presented in Figure 7 below. In Step 1 initial publications list were retrieved from pre-defined databases - Scopus, Web of Science, Google Scholar. The lists were merged and duplicates eliminated in Step 2. Afterwards, texts being less than 6 pages were excluded (Step 3). Steps 1-3 were guided by *Exclusion Criteria*. In the next stage (Step 4), publications were screened by Title based on pre-defined *Relevance Criteria*. The ones which passed were evaluated by their availability (Step 5). As long as the study was available, it was evaluated again by the same pre-defined *Relevance Criteria* applied to Abstract, Conclusion and if necessary Introduction (Step 6). The ones which passed this threshold formed primary publications corpus extracted from databases in full. These primary texts were evaluated again based on full text (Step 7) applying *Relevance Criteria* first and then *Scoring Metrics*.

The extraction and screening of the texts has been performed by one author strictly adhering to detailed research design (as per SLR protocol) developed, validated and confirmed by all authors. To mitigate potential bias stemming from the subjective screening and rating of studies, especially, when the studies are

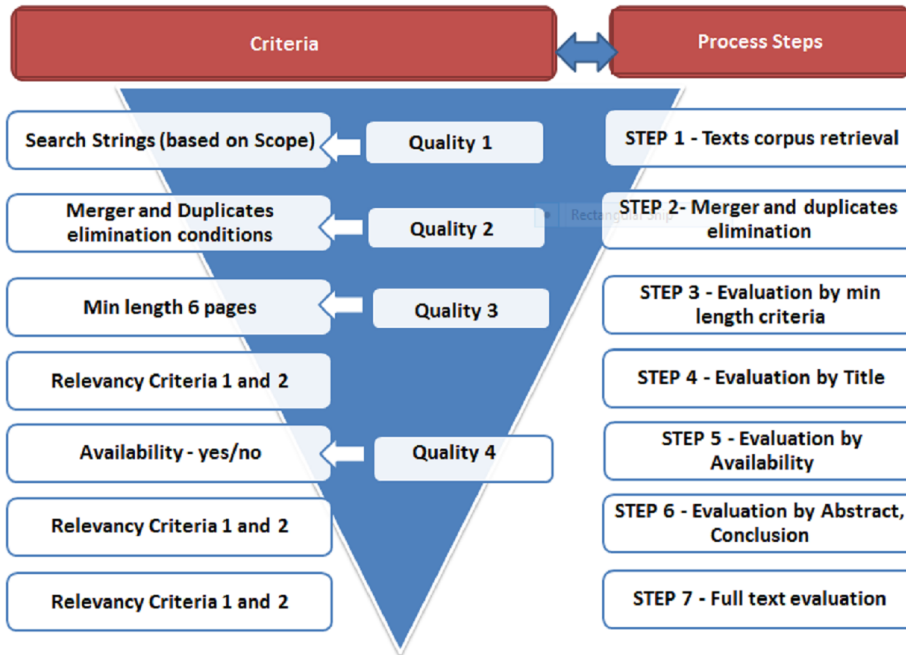


Figure 7. Relevance and quality screening steps with criteria

assessed based on relevance and quality criteria, the sample of the texts have been screened and evaluated by the other author independently, and results have been compared for consistency. In case of deviations, results have been discussed and approach agreed.

Results and Quantitative Analysis In Step 1, 1 715 publications[6] were extracted from relevant databases with the following composition - Scopus (819), Web of Science (489), Google Scholar (407). In terms of scientific publication domains, Computer Science (42.4%), Engineering (20.6%), Mathematics (11.1%) accounted for app. 74% of Scopus originated texts. The same applies to Web of Science harvest. *Exclusion Criteria* application produced the following results. In Step 2, after eliminating duplicates, 1 186 texts were passed for minimum length evaluation, and 767 reached assessment by *Relevancy Criteria*.

As mentioned, *Relevance Criteria* were applied iteratively (Step 4-Step 6) and in conjunction with availability assessment. As a result, only 298 texts were retained for full evaluation, with 241 originating from scientific databases while 57 were 'grey'. These studies formed a primary texts corpus which was extracted, read in full and evaluated by *Relevance Criteria* combined with *Scoring Metrics*. The decision rule was set as follows. Studies that scored '1' or '0' were rejected, while texts with '3' and '2' evaluation were admitted as final primary studies corpus. To

[6]All texts corpus is available at `https://figshare.com/articles/dataset/Cross-domain_Systematic_Literature_Review_dataset/15169593`

this end, as an outcome of SLR-based, broad, cross-domain publications collection and screening we identified 207 relevant publications from peer-reviewed (156 texts) and 'grey' literature (51 texts). Figure 8 below exhibits yearly published research numbers with the breakdown by 'peer-reviewed' and 'grey' literature starting from 1997.

In terms of composition, 'peer-reviewed' studies corpus is well-balanced with 72 journal articles and 82 conference papers, while book chapters account for 4 instances only. In contrast, in 'grey' literature subset, articles in moderated and non-peer reviewed journals are dominant (n=34) compared to overall number of conference papers (n=13), followed by small number of technical reports and pre-prints (n=4).

Temporal analysis of texts corpus (as per Figure 8 below) resulted in two observations. Firstly, we note that stable and significant research interest (in terms of numbers) on data mining methodologies application has started around a decade ago - in 2007. Research efforts made prior to 2007 were relatively limited, with the number of publications below 10. Secondly, we note that research on data mining methodologies has grown substantially since 2007, an observation supported by the 3-year and 10-year constructed mean trend lines. In particular, the number of publications have roughly tripled over the past decade, hitting an all-time high of 24 texts published in 2017.

Further, there are also two distinct spike sub-periods in the years 2007-2009 and 2014-2017 followed by stable pattern with overall higher number of released publications on annual basis. This observation is in line with the trend of increased penetration of methodologies, tools, cross-industry applications and academic research of data mining.

3.1.2. SLR I: Findings and Discussion

In this section, the research questions are addressed. Initially, as part of RQI.I, overview of data mining methodologies 'as-is' and adaptation trends is presented. In addressing RQI.II, the identified adaptations are classified further. Then, as part of RQI.III subsection, each category identified under RQI.II is analyzed with particular focus on the goals of adaptations.

RQI.I: How data mining methodologies are applied ('as-is' vs adapted)?

The first research question examines the extent to which data mining methodologies are used 'as-is' versus adapted. Our review based on 207 publications identified two distinct paradigms on how data mining methodologies are applied. The first is "as-is" where the data mining methodologies are applied as stipulated. The second is with 'adaptations', i.e., methodologies are modified by introducing various changes to the standard process model when applied.

We have aggregated research by decades to differentiate application pattern between two time periods - 1997-2007 with limited vs 2008-2018 with more intensive data mining application. The given cut has not only been guided by

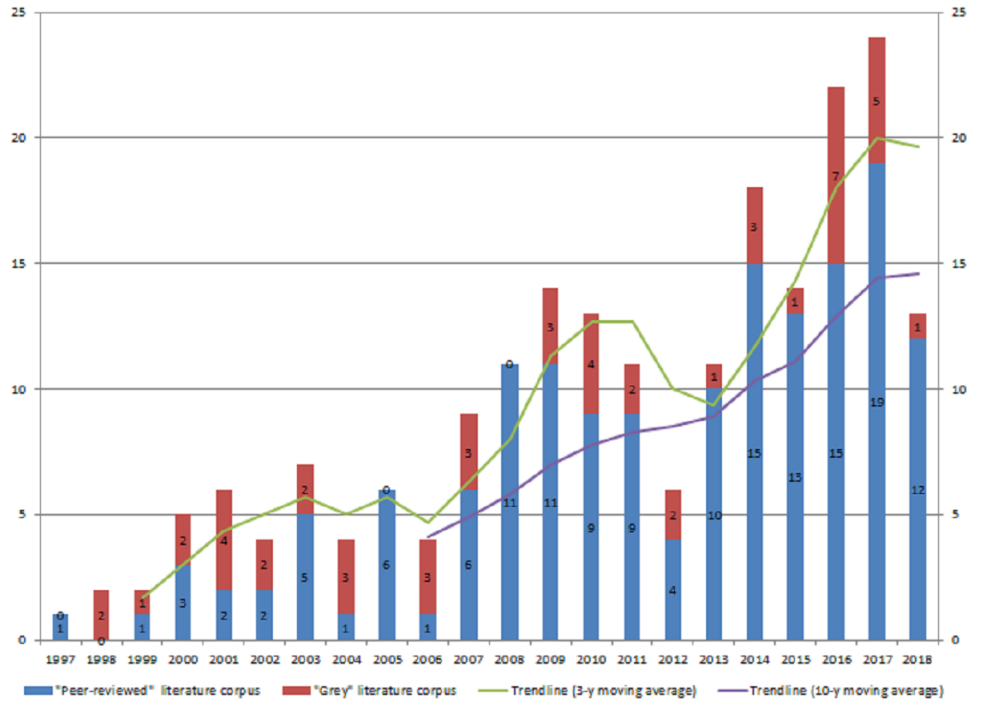


Figure 8. SLR derived relevant texts' corpus - data mining methodologies 'peer-reviewed' research and 'grey' for period 1997-2018 (no. of publications).

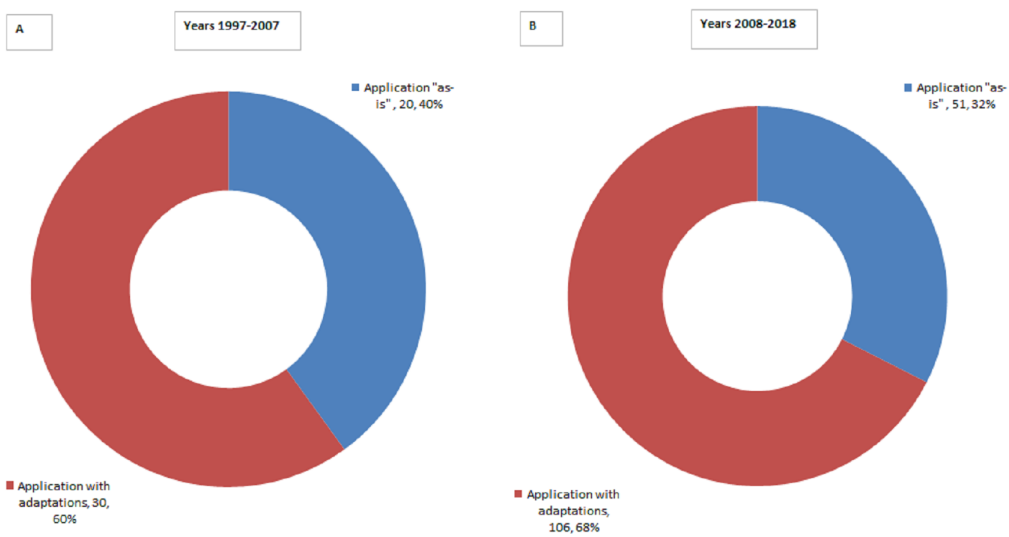


Figure 9. Applications of data mining methodologies: A) breakdown by 'as-is' vs adaptions for 1997-2007 period; B) breakdown by 'as-is' vs adaptions for 2008-2018 period

extracted publications corpus, but also by earlier surveys. In particular, during the pre-2007 research, there were ten new methodologies proposed, but since then, only two new methodologies have been proposed. Thus, there is a distinct trend observed over the last decade of large number of extensions and adaptations proposed vs entirely new methodologies.

We note that during the first decade of our time scope (1997-2007), the ratio of data mining methodologies applied 'as-is' was 40% (as presented in Figure 9A). However, the same ratio for the following decade is 32% (Figure 9B). Thus, in terms of relative shares, we note a clear decrease in using data mining methodologies 'as-is' in favor of adapting them to cater to specific needs. The trend is even more pronounced when comparing numbers - adaptations more than tripled (from 30 to 106) while 'as-is' scenario has increased modestly (from 20 to 51). Given this finding, we continue with analyzing how data mining methodologies have been adapted under RQ2. Our review led us to identify three distinct adaptation scenarios namely 'Modification', 'Extension', and 'Integration' and each of them is discussed in depth.

RQI.II: How have existing data mining methodologies been adapted?

We identified that data mining methodologies have been adapted to cater to specific needs. In order to categorized adaptations scenarios, we applied a two-level dichotomy, specifically, by applying the following decision tree:

1. Level 1 Decision - Has the methodology been combined with another methodology? - If yes, the resulting methodology was classified in the 'integration' category. Otherwise, we posed the next question.
2. Level 2 Decision - Are any new elements (phases, tasks, deliverables) added to the methodology? - If yes, we designate the resulting methodology as an 'extension' of the original one. Otherwise, we classify the resulting methodology as a modification of the original one.

Thus, when adapted, three distinct types of adaptation scenarios can be distinguished:

- Scenario 'Modification' - introduces specialized sub-tasks and deliverables in order to address specific use cases or business problems. Modifications typically concentrate on granular adjustments to the methodology at the level of sub-phases, tasks or deliverables within the existing reference frameworks (e.g. CRISP-DM or KDD) stages. For example, [Che+14], in the study of mobile network domain, proposed automated decision-making enhancement in the deployment phase. In addition, the evaluation phase was modified by using both conventional and own-developed performance metrics. Further, in a study performed within the financial services domain, [Yan+16] presents feature transformation and feature selection as sub-phases, thereby enhancing the data mining modeling stage.
- Scenario 'Extension' - primarily proposes significant extensions to reference data mining methodologies. Such extensions result in either integrated data

mining solutions, data mining frameworks serving as a component or tool for automated IS systems, or their transformations to fit specialized environments. The main purposes of extensions are to integrate fully-scaled data mining solutions into IS/IT systems and business processes and provide broader context with useful architectures, algorithms, etc. Adaptations, where extensions have been made, elicit and explicitly present various artifacts in the form of system and model architectures, process views, workflows, and implementation aspects. A number of soft goals are also achieved, providing holistic perspective on data mining process, and contextualizing with organizational needs. Also, there are extensions to this scenario where data mining process methodologies are substantially changed and extended in all key phases to enable execution of data mining life-cycle with the new (Big) Data technologies, tools and in new prototyping and deployment environments (e.g. Hadoop platforms or real-time customer interfaces). For example, [KKR13] presented extensions to traditional CRISP-DM data mining outcomes with fully fledged Decision Support System (DSS) for hotel brokerage business. Authors [KKR13] have introduced spatial/non-spatial data management (extending data preparation), analytical and spatial modeling capabilities (extending modeling phase), provided spatial display and reporting capabilities (enhancing deployment phase). In the same work, domain knowledge was introduced in all phases of data mining process, and usability and ease of use were also addressed.

- Scenario 'Integration' - combines reference methodology, e.g. CRISP-DM with: (1) data mining methodologies originated from other domains (e.g. Software engineering development methodologies), (2) organizational frameworks (Balanced Scorecard, Analytics Canvass, etc.), or (3) adjustments to accommodate Big Data technologies and tools. Also, adaptations in the form of 'Integration' typically introduce various types of ontologies and ontology-based tools, domain knowledge, software engineering, and BI-driven framework elements. Fundamental data mining process adjustments to new types of data, IS architectures (e.g. real-time data, multi-layer IS) are also presented. Key gaps addressed with such adjustments are prescriptive nature and low degree of formalization in CRISP-DM, obsolete nature of CRISP-DM with respect to tools, and lack of CRISP-DM integration with other organizational frameworks. For example, [BC08] developed KEOPS data mining methodology (CRISP-DM based) centered on domain knowledge integration. Ontology-driven information system has been proposed with integration and enhancements to all steps of data mining process. Further, an integrated expert knowledge used in all data mining phases was proved to produce value in data mining process.

To examine how the application scenario of each data mining methodology usage has developed over time, we mapped peer-reviewed texts and 'grey' literature

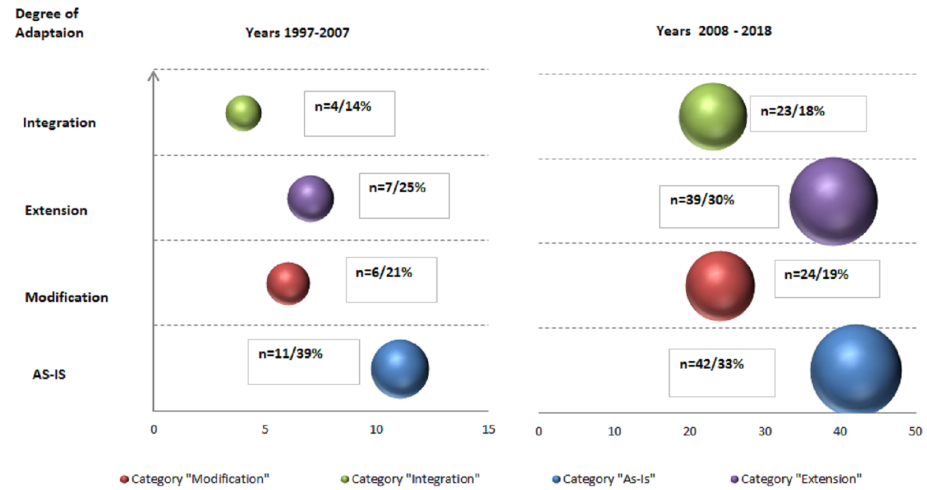


Figure 10. Data Mining methodologies application research - primary 'peer-reviewed' texts classification by types of scenarios aggregated by decades (with numbers and relative proportions)

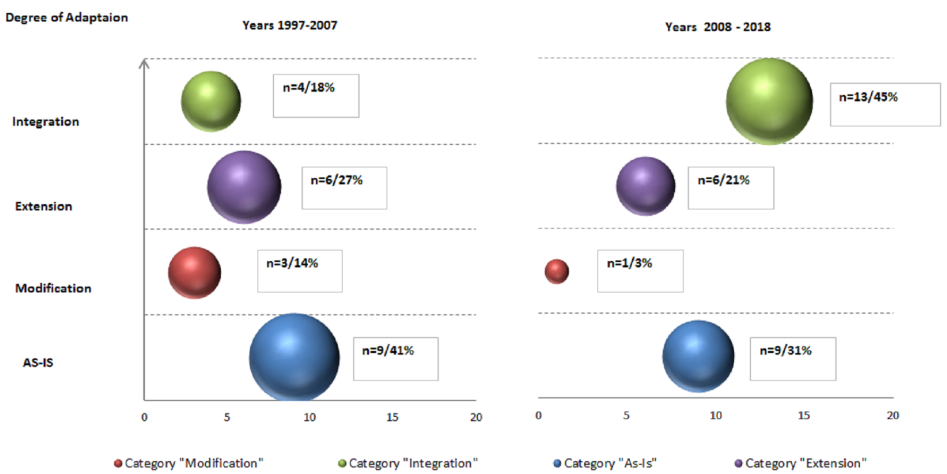


Figure 11. Data Mining methodologies application research - primary 'grey' texts classification by types of scenarios aggregated by decades (with numbers and relative proportions)

to respective adaptation scenarios, aggregated by decades (as presented in the Figure 10 for peer-reviewed and Figure 11 for 'grey'). There, the number of the texts classified into respective category is represented by the size of the ball in a bubble chart. For each research decade a distinct chart is presented.

For peer-reviewed research, such temporal analysis resulted in three observations. Firstly, research efforts in each adaptation scenario has been growing and number of publication more than quadrupled (128 vs 28). Secondly, as noted above, relative proportion of 'as-is' studies is diluted (from 39% to 33%) and primarily replaced with 'Extension' paradigm (from 25% to 30%). In contrast, in relative terms 'Modification' and 'Integration' paradigms gains are modest. Further, this finding is reinforced with other observation - most notable gaps in terms of modest number of publications remain in 'Integration' category where excluding 2008-2009 spike, research efforts are limited and number of texts is just 13. This is in stark contrast with prolific research in 'Extension' category, though concentrated in the recent years. We can hypothesize that existing reference methodologies do not accommodate and support increasing complexity of data mining projects and IS/IT infrastructure, as well as certain domains specifics and as such need to be adapted.

In 'grey' literature, in contrast to peer-reviewed research, growth in number of publications is less profound - 29 vs 22 publications or 32% comparing across two decade (as per Figure 11). The growth is solely driven by 'Integration' scenarios application (13 vs 4 publications) while both 'as-is' and other adaptations scenarios are stagnating or in decline.

RQI.III: For what purposes have existing data mining methodologies been adapted?

The third research question is addressed by analyzing what gaps the data mining methodology adaptations seek to fill and the benefits of such adaptations. There have been three adaptation scenarios identified, namely 'Modification', 'Extension', and 'Integration'. Here, we analyze each of them.

'Modification'. Modifications of data mining methodologies are present in 30 peer-reviewed and 4 'grey' literature studies. The analysis shows that modifications overwhelmingly consist of specific case studies. However, the major differentiating point compared to 'as-is' case studies is clear presence of specific adjustments towards standard data mining process methodologies. Yet, the proposed modifications and their purposes do not go beyond traditional data mining methodologies phases. They are granular, specialized and executed on tasks, sub-tasks, and at deliverables level. With modifications, authors describe potential business applications and deployment scenarios at a conceptual level, but typically do not report or present real implementations in the IS/IT systems and business processes.

Further, this research subcategory can be the best classified based on domains where case studies were performed and data mining methodologies modification scenarios executed. Four distinct domain-driven applications have been identified, and they are presented in the Figure 12 below:

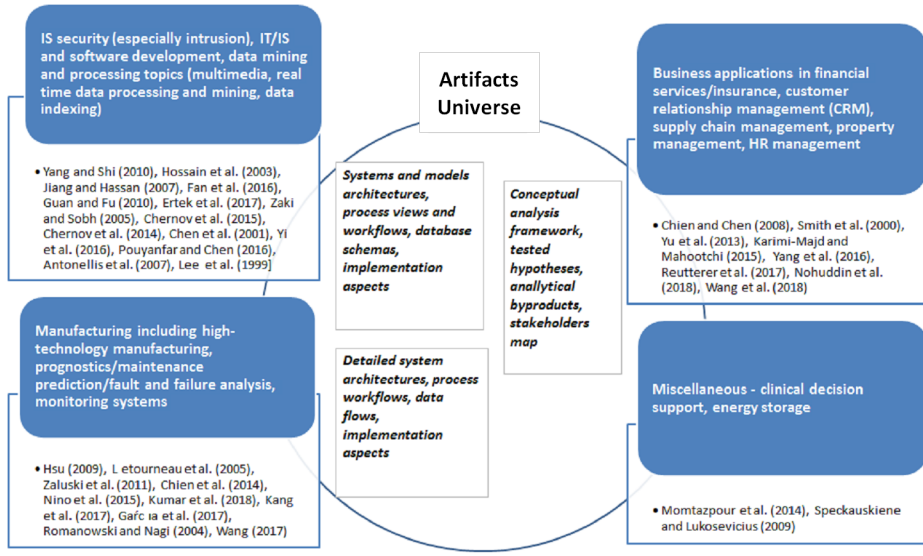


Figure 12. 'Modification' paradigm application studies for period 1997-2018 - mapping to domains

- **IT, IS Domain** The largest number of publications (14 or app. 40%), was performed on IT, IS security, software development, specific data mining and processing topics. Authors address intrusion detection problem in [HBJ03], [FYC16], [LSM99], specialized algorithms for variety of data types processing in [YS10], [Che+01], [YTX16], [PC16], effective and efficient computer and mobile networks management in [GF10], [ECZ17], [ZS05], [CPR15], [Che+14].
- **Manufacturing and Engineering** The next most popular research area is manufacturing/engineering with 10 case studies. The central topic here is high-technology manufacturing, e.g. semi-conductors associated - study of [CDL14], and various complex prognostics case studies in rail, aerospace domains [Lét+05], [Zal+11] concentrated on failure predictions. These are complemented by studies on equipment fault and failure predictions and maintenance [KST18], [Kan+17], [Wan17] as well as monitoring system [Gar+17].
- **Sales and Services, incl. Financial Industry** The third category is presented by 7 business application papers concerning customer service, targeting and advertising [KM15], [Reu+17], [Wan17], financial services credit risk assessments [SWB00], supply chain management [Noh+18], and property management [YFH13], and similar.

As a consequence of specialization, these studies concentrate on developing 'state-of-the art' solution to the respective domain-specific problem.

'Extension'. 'Extension' scenario was identified in 46 peer-reviewed and 12 'grey' publications; 'Extension' to existing data mining methodologies were executed with four major purposes:

1. Purpose 1 - **To implement fully scaled, integrated data mining solution and regular, repeatable knowledge discovery process** - address model, algorithm deployment, implementation design (including architecture, workflows and corresponding IS integration). Also, a complementary goal is to tackle changes to business process to incorporate data mining into organization activities.
2. Purpose 2 - **To implement complex, specifically designed systems and integrated business applications with data mining model/solution as component or tool.** Typically, this adaptation is also oriented towards Big Data specifics, and is complemented by proposed artifacts such as Big Data architectures, system models, workflows, and data flows.
3. Purpose 3 - **To implement data mining as part of integrated/combined specialized infrastructure, data environments and types (e.g. IoT, cloud, mobile networks)**.
4. Purpose 4 - **To incorporate context-awareness aspects.**

The specific list of studies mapped to each of the given purposes presented in the Appendix A.1, Table A1. Main purposes of adaptations, associated gaps and/or benefits along with observations and artifacts are documented in the Figure 13 below.

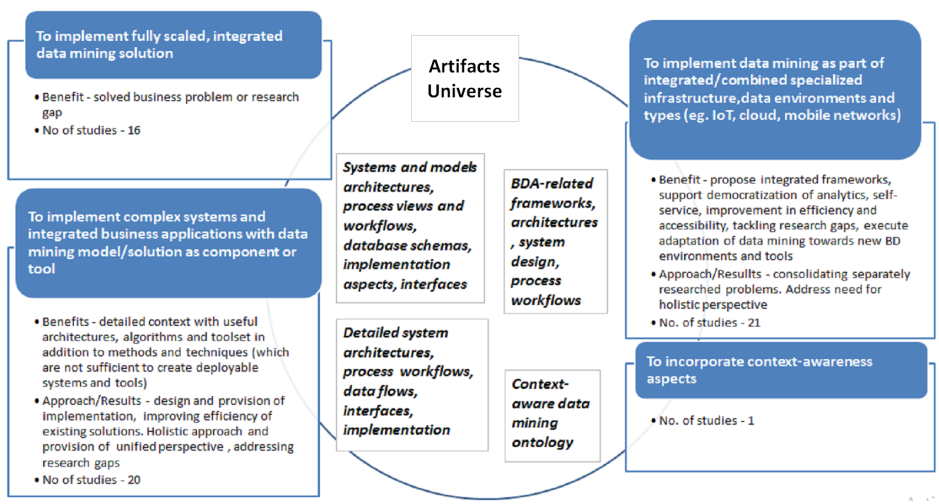


Figure 13. 'Extension' scenario adaptations goals, benefits, artifacts and number of publications for period 1997-2018

In 'Extension' category, studies executed with the Purpose 1 propose fully scaled, integrated data mining solutions of specific data mining models, associated

frameworks and processes. The distinctive trait of this research subclass is that it ensures repeatability and reproducibility of delivered data mining solution in different organizational and industry settings. Both the results of data mining use case and deployment and integration into IS/IT systems and associated business process(es) are presented explicitly. Thus, 'Extension' subclass is geared towards specific solution design, tackling concrete business or industrial setting problem or addressing specific research gaps thus resembling comprehensive case study.

This direction can be well exemplified by expert finder system in research social network services proposed by [Sun+15], data mining solution for functional test content optimization by [Wan15] and time-series mining framework to conduct estimation of unobservable time-series by [Hu+10]. Similarly, [Du+17] tackle online log anomalies detection, automated association rule mining is addressed by [Çin+11], software effort estimation by [DPP11], network patterns visual discovery by [SG08]. A number of studies address solutions in IS security [SJ05], manufacturing [Güd+14], [CBK16], materials engineering domains [Dor08], and business domains [XQ08], [DD07].

In contrast, 'Extension' studies executed for the Purpose 2 concentrate on design of complex, multi-component information systems and architectures. These are holistic, complex systems and integrated business applications with data mining framework serving as component or tool. Moreover, data mining methodology in these studies is extended with systems integration phases.

For example, [Mob07] presents data mining application in Web personalization system and associated process; here, data mining cycle is extended in all phases with utmost goal of leveraging multiple data sources and using discovered models and corresponding algorithms in an automatic personalization system. Authors comprehensively address data processing, algorithm, design adjustments and respective integration into automated system. Similarly, [HSC04] tackle improvement of Webpage recommender system by presenting extended data mining methodology including design and implementation of data mining model. Holistic view on web-mining with support of all data sources, data warehousing and data mining techniques integration, as well as multiple problem-oriented analytical outcomes with rich business application scenarios (personalization, adaptation, profiling, and recommendations) in e-commerce domain was proposed and discussed by [BM98]. Further, [Sin+14] tackled scalable implementation of Network Threat Intrusion Detection System. In this study, the data mining methodology and resulting model are extended, scaled and deployed as a module of a quasi-real-time system for capturing Peer-to-Peer Botnet attacks. A similar complex solution was presented in a series of publications by [Lee+00] and [Lee+01] who designed real-time data mining-based Intrusion Detection System (IDS). These works are complemented by comprehensive study of [Bar+01] who constructed experimental testbed for intrusion detection with data mining methods. Detection model combining data fusion and mining and respective components for Botnets identification was developed by [Kia+09] too. A similar approach is presented

in [Ala+11] who proposed and implemented zero-day malware detection system with associated machine-learning based framework. Finally, [ARA11] presented a multi-layer framework for fuzzy attack in 3G cellular IP networks.

A number of authors have considered data mining methodologies in the context of Decision Support Systems and other systems that generate information for decision-making, across a variety of domains. For example, [KKR13] executed significant extension of data mining methodology by designing and presenting integrated Decision Support System (DSS) with six components acting as supporting tool for hotel brokerage business to increase deal profitability. Similar approach is undertaken by [Cap+17] focusing on improving energy management of properties by provision of occupancy pattern information and reconfiguration framework. [Kab16] presented data mining information service providing improved sales forecasting that supported solution of under/over-stocking problem, while [LZX18] addressed sales forecasting with sentiment analysis on Big Data. [KRG01] proposed GA-based Intelligent Diagnosis system for fault diagnostics in manufacturing domain. The latter was tackled further in [Sha+10] with a complex, integrated data mining system for diagnosing and solving manufacturing problems in real time.

Lenz et al. [LWW18] propose a framework for capturing data analytics objectives and creating holistic, cross-departmental data mining systems in the manufacturing domain. This work is representative of a cohort of studies that aim at extending data mining methodologies in order to support the design and implementation of enterprise-wide data mining systems. In this same research cohort, we classify [LCR17], which presents a data mining toolset integrated into the Moodle learning management system, with the aim of supporting university-wide learning analytics.

One study addresses Multi-Agent based data mining concept. [KMB13] have developed a unified theoretical framework for data mining by formulating a unified data mining theory. The framework is tested by means of agent programming proposing integration into Multi-Agent System which is useful due to scalability, robustness and simplicity.

The subcategory of 'Extension' research executed with Purpose 3 is devoted to data mining methodologies and solutions in specialized IT/IS, data and process environments which emerged recently as consequence of Big Data associated technologies and tools development. Exemplary studies include IoT associated environment research, for example, Smart City application in IoT presented by [Str+15]. In the same domain, [BG16] addressed IoT-enabled smart buildings with the additional challenge of large amount of high-speed real time data and requirements of real-time analytics. Authors proposed an integrated IoT Big Data Analytics framework. This research is complemented by interdisciplinary study of [Zho+17] where IoT and wireless technologies are used to create RFID-enabled environment, producing analysis of KPIs to improve logistics.

A significant number of studies addresses various mobile environments, some-

times complemented by cloud-based environments or cloud-based environments as stand-alone. [GPK13] addressed mobile data mining with execution on mobile device itself; the framework proposes innovative approach addressing extensions of all aspects of data mining including contextual data, end-user privacy preservation, data management and scalability. [YHE14] and [YH14] introduced cloud-based mobile data analytics framework with application case study for smart home based monitoring system. [CPT16] have presented innovative FollowMe suite which implements data mining framework for mobile social media analytics with several tools with respective architecture and functionalities. Interesting paper was presented by [Tor+17] who addressed data mining methodology and its implementation for congestion prediction in mobile LTE networks, tackling also feedback reaction with network reconfiguration trigger.

Further, [Bil+14] presented cloud-based Future Internet Enabler - automated social data analytics solution which also addresses Social Network Interoperability aspect supporting enterprises to interconnect and utilize social networks for collaboration. Real-time social media streamed data and resulting data mining methodology and application was extensively discussed by [ZLL14]. Authors proposed design of comprehensive ABIGDAD framework with seven main components implementing data mining based deceptive review identification. An interdisciplinary study tackling both these topics was developed by [Put+16] who proposed integrated framework and architecture of disaster management system based on streamed data in cloud environment, ensuring end-to-end security. Additionally, key extensions to data mining framework have been proposed merging variety of data sources and types, security verification and data flow access controls. Finally, cloud-based manufacturing was addressed in the context of fault diagnostics by [Kum+16].

Also, [Mah+13] tackled Wireless Sensor Networks and associated data mining framework required extensions. Interesting work is executed by [NJ03] addressing rare topic of data mining solutions integration within traditional data warehouses and active mining of data repositories themselves.

Supported by new generation of visualization technologies (including Virtual Reality environments), [WLM11] proposed and implemented CAVE-SOM (3D visual data mining framework) which offers interactive, immersive visual data mining with multiple visualization modes supported by plethora of methods. Earlier version of visual data mining framework was successfully developed and presented by [Gan+96] as early as in 1996.

Large-scale social media data is successfully tackled by [Lem16] with comprehensive framework accompanied by set of data mining tools and interface. Real-time data analytics was addressed by [SP17] in the domain of enterprise service ecosystem. Images data was addressed in [Hua+02] by proposing multimedia data mining framework and its implementation with user relevance feedback integration and instance learning. Further, exploded data diversity and associated need to extend standard data mining is addressed by [Sin+16] in the study devoted to

object detection in video surveillance systems supporting real time video analysis.

Finally, there is also limited number of studies which addresses context awareness (Purpose 4) and extends data mining methodology with context elements and adjustments. In comparison with 'Integration' category research, here, the studies are at lower abstraction level, capturing and presenting a list of adjustments. [SVL03] generate taxonomy of context factors, develop extended data mining framework and propose deployment including detailed IS architecture. Context-awareness aspect is also addressed in the papers reviewed above, e.g. [LWW18], [KKR13], [Sun+15], and other studies.

'Integration'. 'Integration' of data mining methodologies scenario was identified in 27 'peer-reviewed' and 17 'grey' studies. Analysis revealed that this adaptation scenario at a higher abstraction level is typically executed with the 5 key purposes:

1. Purpose 1 - **to integrate/combine with various ontologies existing in organization**.
2. Purpose 2 - **to introduce context-awareness and incorporate domain knowledge**.
3. Purpose 3 - **to integrate/combine with other research or industry domains framework, process methodologies and concepts**.
4. Purpose 4 - **to integrate/combine with other well-known organizational governance frameworks, process methodologies and concepts**.
5. Purpose 5 - **to accommodate and/or leverage upon newly available Big Data technologies, tools and methods.**

The specific list of studies mapped to each of the given purposes presented in the Appendix A.1, Table A2. Main purposes of adaptations, associated gaps and/or benefits along with observations and artifacts are documented in Figure 14 below.

As mentioned, number of studies concentrates on proposing ontology-based Integrated data mining frameworks accompanies by various types of ontologies (Purpose 1). For example, [SO08] focus on ontology-based organizational view with Actors, Goals and Objectives which supports execution of Business Understanding Phase. [BC08] propose KEOPS framework which is CRISP-DM compliant and integrates a knowledge base and ontology with the purpose to build OIS (ontology-driven information system) for business and data understanding phases while knowledge base is used for post-processing step of model interpretation. [Par+17] propose and design comprehensive ontology-based data analytics tool IRIS with the purpose to align analytics and business. IRIS is based on concept to connect dots, analytics methods or transforming insights into business value, and supports standardized process for applying ontology to match business problems and solutions.

Further, [Yin+14] propose domain-specific data mining framework oriented to business problem of customer demand discovery. They construct ontology for customer demand and customer demand discovery task which allows executing

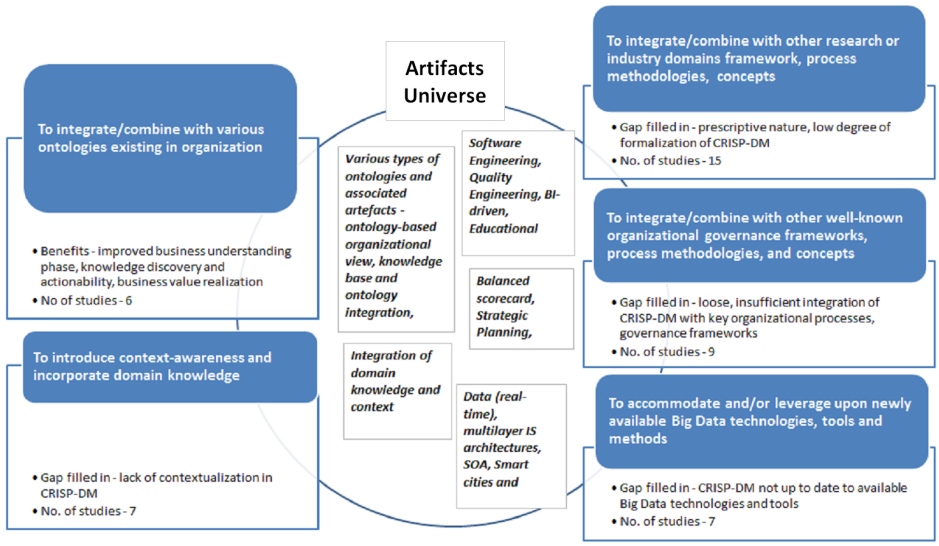


Figure 14. 'Integration' scenario adaptations goals, benefits, artifacts and number of publications for period 1997-2018

structured knowledge extraction in the form of knowledge patterns and rules. Here, the purpose is to facilitate business value realization and support actionability of extracted knowledge via marketing strategies and tactics. In the same vein, [CC03] presented ontology for the Data Mining domain, which main goal is to simplify the development of distributed knowledge discovery applications. Authors offered to a domain expert a reference model for different kind of data mining tasks, methodologies, and software capable to solve the given business problem and find the most appropriate solution.

Apart from ontologies, [SO09] in another study propose IS inspired, driven by Input-Output model data mining methodology which supports formal implementation of Business Understanding Phase. This research exemplifies studies executed with Purpose 2. The goal of the paper is to tackle the prescriptive nature of CRISP-DM and address how the entire process can be implemented. [CSZ05] study is also exemplary in terms of aggregating and introducing several fundamental concepts into traditional CRISP-DM data mining cycle - context awareness, in-depth pattern mining, human-machine cooperative knowledge discovery (in essence, following human-centricity paradigm in data mining), loop-closed iterative refinement process (similar to Agile-based methodologies in Software Development). There are also several concepts, like data, domain, interestingness, rules which are proposed to tackle a number of fundamental constrains identified in CRISP-DM. They have been discussed and further extended by [CZ07], [CZ08], [Cao10] into integrated domain driven data mining concept resulting in fully fledged D3M (domain-driven) data mining framework. Interestingly, the same concepts, but on individual basis are investigated and presented by other authors, e.g. context-aware data mining

methodology is tackled by [Xia09a], [Xia09b] in the context of financial sector. [Pou+16] attempted very crucial privacy-preservation topic in the context of achieving effective data analytics methodology. Authors introduced metrics and self-regulatory (reconfigurable) information sharing mechanism, providing customers with controls for information disclosure.

A number of studies have proposed CRISP-DM adjustments based on existing frameworks, process models or concepts originating in other domains (Purpose 3), for example, software engineering [Mar+07], [Mar+09], [MMS09] and industrial engineering [Sol02], [Zha+05].

Meanwhile, [MMF10] proposed a new refined data mining process based on a global comparative analysis of existing frameworks while [Ang14] outlined a data analytics framework based on statistical concepts. Following a similar approach, some researchers suggest explicit integration with other areas and organizational functions, for example, BI-driven Data Mining by [HF09]. Similarly, [CKH16] developed an architecture-centric agile Big Data analytics methodology, and an architecture-centric agile analytics and DevOps model. Alternatively, several authors tackled data mining methodology adaptations in other domains, e.g. educational data mining by [TVP17], decision support in learning management systems [MH11], and in accounting systems [AF17].

Other studies are concerned with actionability of data mining and closer integration with business processes and organizational management frameworks (Purpose 4). In particular, there is a recurrent focus on embedding data mining solutions into knowledge-based decision-making processes in organizations, and supporting fast and effective knowledge discovery [MB17].

Examples of adaptations made for this purpose include: (1) integration of CRISP-DM with the Balanced Scorecard framework used for strategic performance management in organizations [YWY14]; (2) integration with a strategic decision-making framework for revenue management [Seg+16]; (3) integration with a strategic analytics methodology [RS08], and (4) integration with a so-called 'Analytics Canvas' for management of portfolios of data analytics projects [Küh+18]. Finally, [AP15] explored methodological attributes important for adoption of data mining methodology by novice users. This latter study uncovered factors that could support the reduction of resistance to the use of data mining methodologies. Conversely, [LJ17] comprehensively evaluated factors that may increase the benefits of Big Data Analytics projects in an organization.

Lastly, a number of studies have proposed data mining frameworks, (e.g. CRISP-DM) adaptations to cater for new technological architectures, new types of datasets and applications (Purpose 5). For example, [Lu+17] proposed a data mining system based on a Service-Oriented Architecture (SOA), [ZAS13] developed a concept of self-service data analytics, [OEB17] blended CRISP-DM into a Big Data Analytics framework for Smart Cities, and [Nie+16] proposed a data-driven risk management framework for Industry 4.0 applications.

Analysis of RQ3, regarding the purposes of existing data mining methodologies

adaptations, revealed the following key findings. Firstly, adaptations of type 'Modification' are predominantly targeted at addressing problems that are specific to a given case study. The majority of modifications were made within the domain of IS security, followed by case studies in the domains of manufacturing and financial services. This is in clear contrast with adaptations of type 'Extension', which are primarily aimed at customizing the methodology to take into account specialized development environments and deployment infrastructures, and to incorporate context-awareness aspects. Thirdly, a recurrent purpose of adaptations of type 'Integration' is to combine a data mining methodology with either existing ontologies in an organization or with other domain frameworks, methodologies, and concepts. 'Integration' is also used to instill context-awareness and domain knowledge into a data mining methodology, or to adapt it to specialized methods and tools, such as Big Data. The distinctive outcome and value (gaps filled in) of 'Integrations' stems from improved knowledge discovery, better actionability of results, improved combination with key organizational processes and domain-specific methodologies, and improved usage of Big Data technologies.

Summary It was discovered that the adaptations of existing data mining methodologies found in the literature can be classified into three categories: modification, extension, or integration. They are executed either to address deficiencies and lack of important elements or aspects in the reference methodology (chiefly CRISP-DM). Furthermore, adaptations are also made to improve certain phases, deliverables or process outcomes.

In short, adaptations are made to:

- improve key reference data mining methodologies phases - for example, in case of CRISP-DM these are primarily business understanding and deployment phases.
- support knowledge discovery and actionability.
- introduce context-awareness and higher degree of formalization.
- integrate closer data mining solution with key organizational processes and frameworks.
- significantly update CRISP-DM with respect to Big Data technologies, tools, environments.
- incorporate broader, explicit context of architectures, algorithms and toolsets as integral deliverables or supporting tools to execute data mining process.
- expand and accommodate broader unified perspective for incorporating and implementing data mining solutions in organization, IT infrastructure and business processes.

3.1.3. SLR I: Threats to Validity

Systematic literature reviews have inherent limitations that must be acknowledged. These threats to validity include subjective bias (internal validity) and incompleteness of search results (external validity).

The internal validity threat stems from the subjective screening and rating of studies, particularly when assessing the studies with respect to relevance and quality criteria. To mitigate potential bias stemming from the subjective screening and rating of studies executed by one author, the sample of the texts have been screened and evaluated by the other author independently, and results have been compared for consistency. In case of deviations, results have been discussed and approach agreed. Other internal validity threats were addressed by documenting the survey protocol (SLR Protocol), strictly adhering to the inclusion criteria, and performing significant validation procedures, as documented in the Protocol.

The external validity threat relates to the extent to which the findings of the SLR reflect the actual state of the art in the field of data mining methodologies, given that the SLR only considers published studies that can be retrieved using specific search strings and databases. We have addressed this threat to validity by conducting trial searches to validate our search strings in terms of their ability to identify relevant papers that we knew about beforehand. Also, the fact that the searches led to 1700 hits overall suggests that a significant portion of the relevant literature has been covered.

3.2. Single Domain Systematic Literature Review: Data Mining Methodologies in the Banking Domain (SLR II)

3.2.1. SLR II: Research Design

This part of PhD research investigates RQ1 in the context of specific industry settings, tackling *how are data mining methodologies applied in the banking domain*. Systematic literature review (SLR) method was applied for this purpose as well. Also, this SLR followed the guidelines proposed by [KC07].

To formulate the research questions, the traditional set of 'W' questions was formulated, specifically 'Why?', 'What?' and 'How?'. The 'Why' question led to RQII.I, 'For what purposes are data mining methodologies used in the banking domain?'. Then, 'What' question was raised, 'What data mining methodologies are used in the banking domain?', but it was discarded after a preliminary analysis - it was found that all major data mining methodologies (e.g. CRISP-DM, SEMMA, etc.) are used in this domain and there are little insights to be derived from analyzing this question further. Next, the 'How?' question was raised, which led to RQII.II, 'are data mining methodologies in the banking domain used 'as-is' or are they adapted?'. An initial exploration of this question led to the preliminary conclusion that indeed data mining methodologies are sometimes adapted when applied in banking domain, which in turn led to pose a third research question: 'With what goals are data mining methodologies adapted for the banking domain' (RQII.III)?

According to the guidelines for conducting SLR [KC07] search terms and strings were derived and validated, types of literature in study scope were identified,

then, electronic databases were selected, and screening procedures were defined.

The search string were derived from the research questions and included the terms 'data mining' and 'data analytics' as these are often used interchangeably. The terms 'methodology', 'framework' and 'banking' were added, resulting in the search string being defined as ('data mining methodology') OR ('data mining framework') OR ('data analytics methodology') OR ('data analytics framework') AND ('banking'). Validation of the search string according to [KC07], led to adding the search string of ('CRISP-DM') OR ('SEMMA') OR ('ASUM') AND ('banking') in order to capture case study papers. Also, substituting term 'banking' with the broader term 'finance' was also checked. Piloting the search with the latter term retrieved a significant number of studies related to the financial markets and economy, while the number of the works related to the financial services has dropped significantly and was underrepresented. As the research purpose is to examine usage of data mining methodologies in the concrete industry settings, the final term 'banking' has been chosen as part of the search string. The search strings were applied to Scopus, Web of Science, and Google Scholar databases. Multidisciplinary indexed/non-indexed electronic databases were selected to ensure wide data sources coverage, and to include studies from both academic (peer-reviewed) and practitioners communities ('grey' literature). Specifically, 'grey' literature search covered industry reports, white papers, technical reports, and research works not indexed by Scopus or Web of Science.

Based on the SLR best practices [Kit04], [Bre+07] a multi-step screening procedures (*relevancy* and *quality*) with associated set of *Screening Criteria* (*Exclusion and Inclusion Criteria*), and *Scoring System* were designed. The *Exclusion Criteria* served to eliminate studies in languages other than English, duplicating texts, as well as publications shorter than 6 pages, or the ones not accessible (by University subscriptions). Papers that passed all *Exclusion Criteria* were retained and assessed according to *Relevance Criteria*. Each paper was considered relevant if it was: (1) about a data mining approach within the banking domain, and (2) introduced or described a data mining methodology/framework or modification of existing approaches. Finally, quality screening was conducted for full texts evaluation. For that we developed a *Scoring Metrics* as proposed in [KC07]. Papers were given the score of 3 if all steps of the data mining process were clearly presented and explained. Further, to merit a score of 3, the paper must have also presented a proposal on usage, application, or deployment of the solution in an organization's business process(es) and IT/IS system, and/or discuss prototype or full solution implementation. If description of some process steps were missing, but without impacting the holistic view and understanding of the work performed, the paper was given a score of 2. Only papers scoring '2' or '3' were included in the final primary studies corpus.

The extraction and screening of the texts has been performed by one author strictly adhering to detailed research design developed, validated and confirmed by all authors. To mitigate potential bias when the studies are evaluated based

on relevance and quality criteria, the sample of the texts have been screened and evaluated by the other author independently, and results have been compared for consistency. In case of deviations, results have been discussed and approach agreed.

The initial number of studies retrieved amounted to 693 of which 167 were academic and 526 'grey' literature. Having performed the screening based on *Exclusion Criteria*, 509 studies remained and were subject to relevance screening. 141 papers were finally identified as relevant and moved into quality assessment phase, and 41 peer-reviewed papers and 61 studies from 'grey' literature received a score of 2 or higher. By means of SLR, we identified primary texts' corpus with 102 relevant studies. Figure 15 below exhibits yearly published research numbers with the breakdown by peer-reviewed and 'grey' literature starting from 1997.

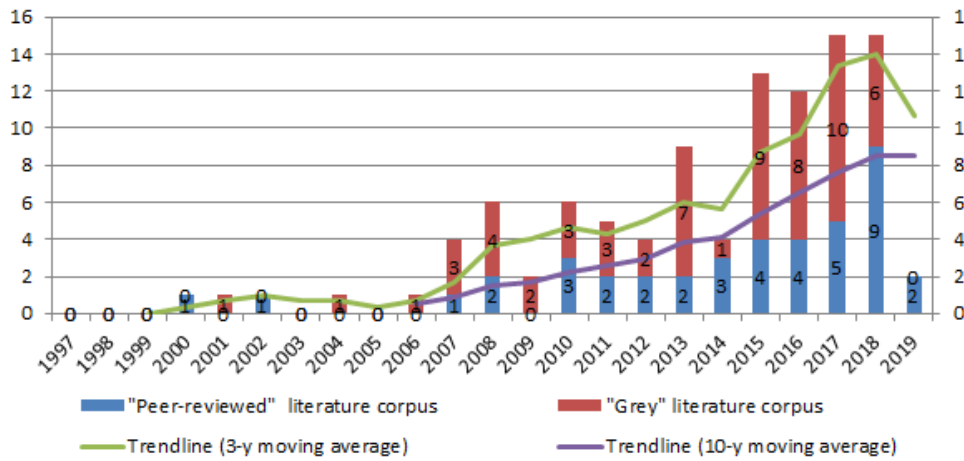


Figure 15. SLR derived texts corpus - data mining methodologies peer-reviewed research and 'grey' literature for period 1997-2019 (no. of publications).

Temporal analysis of texts corpus resulted in two observations. Firstly, it was noted that research on application of data mining methodologies within the banking domain began more than a decade ago - in 2007. Research efforts made prior to 2007 were infrequent and irregular, with 3-4 years gap periods between publications. Secondly, it was observed that research on data mining methodologies within the banking domain has grown since 2007, an observation supported by the 3-year and 10-year constructed mean trend lines. In particular, it was noted that the number of publications have roughly tripled over the last decade, hitting an all-time high in 2018 with 22 texts released.

Comparing SLR 2 primary texts' corpus with the one obtained in cross-domain study (SLR 1), we note the following. SLR 1 is not a strict subset of the cross-domain SLR 2 study, as they have been conducted independently and concatenated in this PhD Thesis. As a result, there is a portion of studies retrieved and present in both SLR 1 and SLR 2, and a number of studies that have been found in SLR 2, but

are not present in cross-domain research. The primary reason for discrepancy is usage of the specific search string of ('CRISP-DM') OR ('SEMMA') OR ('ASUM') AND ('banking') in order to capture case study papers in SLR 2. The research string was added to investigate in-depth how data mining methodologies are applied in the settings of the banking domain. This research string was not used in cross-domain study which focused on broadly researching data mining frameworks usage across 9 different domains. To illustrate, the case studies are typically falling into 'as-is' category, and for instance [KSB18] tackling credit risk assessment or [KR08] predicting churn are present in SLR 2 primary texts' corpus, but not in the cross-domain study. In contrast, [Deb07] which extends data mining methodologies for organizational factors, or [Cao10] research addressing domain-driven data mining are 'Extension' studies and are retrieved in both SLR 1 and SLR 2.

3.2.2. SLR II: Findings

In this section, results of publications analysis are presented and research questions addressed. Lastly, threats to validity are discussed.

RQII.I - For what purposes are data mining methodologies used in the banking domain? In-depth analysis of text corpus revealed that data mining methodologies are predominantly being employed in the banking domain for two main purposes - customer-oriented and risk-oriented (see Figure 16A below).

47 customer-oriented studies which address various aspects related to customer behavior modelling were identified. A typical example is profiling according to usage pattern of different digital channels, [Man+10][7] authors profiled Internet bank users, while [EBN17] focuses on patterns of electronic transactions based on demographic and behavioral features. In the field of Customer Relationship Management (CRM), the most common business problem analyzed relate to identifying and predicting customers who are likely to churn [KR08], customer loyalty and retention [BE15], customer segmentation [TC11], and customer value identification [MA16]. Further, smart and improved customer targeting in sales campaigns [NSG15] and improved targeting and customer prioritization decision support are also popular business problem [Gho+18]. A few studies consider efficiency aspects of bank's infrastructure such as Automated Teller Machines (ATMs) and branch networks (e.g. [Met+17]).

The second most commonly analyzed area is Risk Management, predominantly, credit risk. 34 studies focused on modelling tasks for supporting a variety of risk management processes including credit risk scoring and default prediction [KSB18], prediction of financial distress [GBC15], and credit decisions for private and corporate customers (especially, small and medium enterprises as in [GK19]). Further, identification and prevention of fraud behavior [Ade+11] and AML (anti-money laundering) risks [CR17] are addressed as well. Finally, other

[7]Examples of key texts are presented throughout the analysis. All texts corpus is available at `https://figshare.com/articles/dataset/MasterList_xlsx/8206604`

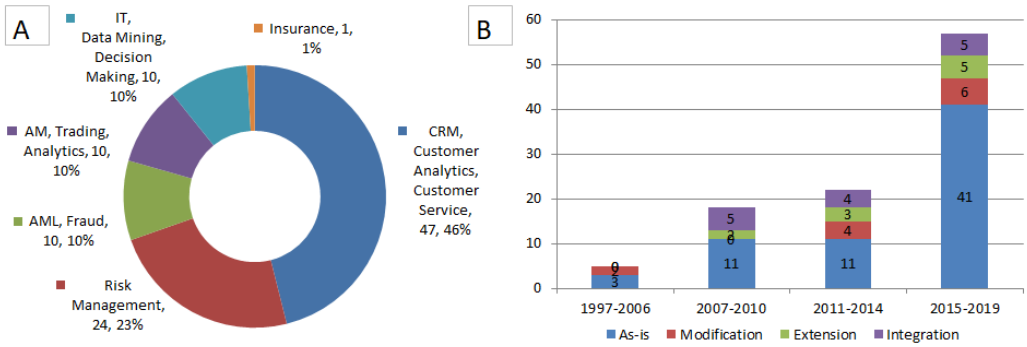


Figure 16. Applications of data mining methodologies in banking: A) breakdown by purposes; B) breakdown by adaptation paradigms

risk management topics, such as market risk, as well as asset management [LY16], trading strategies [AAA13], overall economic analysis and predictions [BD18] are also addressed.

RQII.II - How are data mining methodologies applied ('as-is' vs adapted)? The second research question addresses the extent to which data mining methodologies are used 'as-is' versus adapted. Our review identified two distinct paradigms on how data mining methodologies are applied. The first is 'as-is' where the data mining methodologies are applied as stipulated. The second is with 'adaptations', i.e., methodologies are modified by introducing various changes to the standard process model when applied. Furthermore, our review led us to identify three distinct adaptation scenarios namely 'Modification', 'Extension', and 'Integration':

Scenario 'Modification' - introduces specialized sub-tasks and deliverables in order to address a specific use cases or business problems. Modifications typically concentrate on granular adjustments to the methodology at the level of sub-phases, tasks or deliverables within the existing CRISP-DM or KDD stages.

Scenario 'Extension' - primarily proposes significant extensions to CRISP-DM resulting in either fully-scaled and integrated data mining solutions, data mining frameworks as a component or tool for automated IS systems or adapted to specialized environments. Adaptations where extensions have been made elicit and explicitly presents various artifacts in the form of system and model architectures, process views, workflows, and implementation aspects. Key benefits achieved are deployment, implementation and leveraging of data mining solutions as integral components of IS systems and business processes. Also, data mining process methodology is substantially changed and extended in all key phases to accommodate new Big Data technologies, tools and environments [Pen+11], [BD18].

Scenario 'Integration' - 'Integration' primarily concentrates on either combining CRISP-DM with data mining methodologies originated from other do-

mains (e.g. Business Information Management, Business Process Management, BI [Piv+13]), adjusting to specific organizational aspects [CZ07], and discrimination-awareness with respect to customers [BP14]. Adaptations in the form of integration typically introduces various types of ontologies and ontology-based tools, business processes, business information, and BI-driven framework elements. Key benefits are improved at the deployment phase, improved usage of data and discovered knowledge, higher business processes effectiveness and efficiency. Key gap filled in is lack of CRISP-DM integration with other organizational and domain frameworks.

It was also noted that publications discussing 'as-is' implementations have grown strongly but at the same time, adaptations are also gaining ground (as exhibited in Figure 16B above). Further, there is balanced development and distribution of the research among 'Modification', Extension' and 'Integration' paradigms. We can hypothesize that existing reference methodologies do not accommodate and support increasing complexity of data mining projects and IS/IT infrastructure, as well as banking domain specific requirements and as such need to be adapted.

RQII.III - What are the goals of adaptations? We address the third research question by analyzing each of the adaptation scenarios in depth.

Modification This adaptation scenario was identified in 12 publications, where modifications overwhelmingly consist of specific case studies. However, the major differentiating point compared to 'as-is' case studies is clearly the presence of specific adjustments towards standard data mining process methodologies. Yet, the proposed modifications and their purposes do not go beyond traditional CRISP-DM phases. They are granular, specific and executed on tasks, sub-tasks, and at the level of deliverables. This is in clear contrast to 'extensions' where one of the key proposals are new phases, such as including a new IS/IT systems implementation and integration phase. Also, with modifications, authors describe potential business applications and deployment scenarios at a conceptual level, but typically do not report or present real implementations to the IS/IT systems and business processes.

Further, in the context of banking domain, this research subcategory can be classified with respect to business problems addressed (presented in Appendix A.2, Table A3 [8]).

Extension 'Extension' scenario was identified in 10 publications, and we noted that it was executed for the two major purposes:

1. To implement fully scaled, integrated data mining solution and regular, repeatable knowledge discovery process - address model, algorithm deployment, implementation design (including architecture, workflows and corresponding IS integration). Also, a complementary goal is to address

[8] Two most cited texts references are presented for each subcategory, if number of texts per category exceed two, all texts corpus is available in the full texts' corpus at link `https://figshare.com/articles/dataset/MasterList_xlsx/8206604`

changes to business process to incorporate data mining into organization activities.

2. To implement complex, specifically designed systems and integrated business applications with data mining model/solution as component or tool. Typically, this adaptation is also oriented towards Big Data specifics, and is complemented by proposed artifacts such as Big Data architectures, system models, workflows, and data flows.

It was also concluded that the first purpose focuses on implementation of specific data mining models and associated frameworks and processes. For example, apart from classification model and evaluation framework, [Pen+11] proposes a knowledge-rich financial risk management process while [Cla18] introduces framework for machine-learning audits. [LMK10] presented data mining-based solution for AML implemented as a tool with respective IS architecture and investigative process. [SR13] focused on combined data mining concept introducing multiple data sources, methods and features, all incorporated in the real-time prototyped solution. [Yan09] focused on actionable data mining by presenting post-processing data mining framework which enables automated actions generation. In the similar vein, [Yua+18] presented large-scale data mining framework extended to incorporate social media data, including adaptions to parallel processing. The major benefit achieved by these adaptations, apart from a resolved business problem or research gap, is the usefulness of results produced in the decision-making process.

In contrast, the second purpose concentrates on design of complex, multi-component information systems and architectures. For instance, [BD18] have constructed a framework that considers socio-economic data, its processing methods, a new data life-cycle model, and presented an architecture for Big Data systems to integrate, process and analyze data for forecasting purposes. [Ang+18] proposed refinements of reference data mining methodology to address Big Data analytics, applications prototyping and its evaluation, project management and results communication. Finally, [DTA12] proposed cross-border market monitoring and surveillance system with 3 subsystem components, system and data flows. In this research, authors discuss and present useful architectures, algorithms and tool sets in addition to methods and techniques which alone are not sufficient to create deployable systems and tools. The key benefits provided are broad context enabling practical implementations of complex, integrated data mining solutions. The specific list of studies mapped to each of the given purposes is presented in the Appendix A.2, Table A4 while adaptation goals with key artifacts are visualized in Figure 17 below.

Integration Integration of data mining methodologies were found in 14 publications. Analysis shows that these adaptations are at the highest abstraction level and typically executed with the goals to (1) introduce discrimination-awareness in data mining, (2) integrate/combine with other organizational frameworks, and (3) integrate/combine with other well-known frameworks, process methodologies and

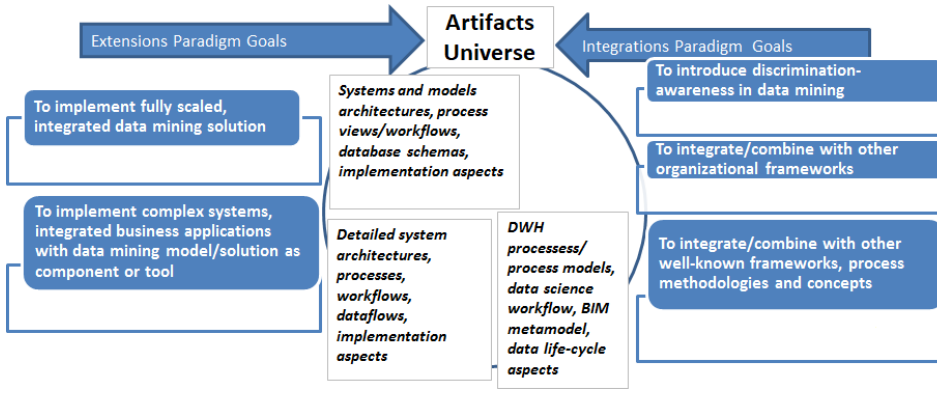


Figure 17. Data Mining Methodologies in Banking - 'Extension' and 'Integration' scenarios adaptation goals, their artifacts and example texts mapping

concepts. Example list of studies is available in the Appendix A.2, Table A5 while artifacts are presented in Figure 17[9] and further discussed.

Discrimination-aware data mining (DADM), as proposed by [BP14], includes tool support for 'correct' decision process. The major benefit is increased correctness and usefulness of results in the decision-making process, monitoring, avoidance of discrimination and transparency.

[Deb07] the author combined data mining methodology with organizational context to instill and improve data-driven decision-making. Further, [Piv+13] integrated data mining with business process frameworks and models (also proposed by [LLV10]). [PM15] integrated data mining with BIM (Business Information Modelling) while [BG11] merged data mining with BI. All with the purpose to improve usage of data, business processes effectiveness and deployment of data mining solutions. These works are complemented by a number of publications [CZ07], [Cao10] specifically tackling actionability of data mining results, which aim to reduce the likelihood of data mining project producing high quality knowledge with limited or no business benefit. Authors propose shift to domain-driven data mining paradigm by integrating such new key component as domain intelligence, human-machine cooperation, in-depth mining, actionability enhancement, and iterative refinement process. Emphasis on data-mining business requirements, model sharing and reuse from business user perspective is also tackled by introducing ontology-based data mining model management approach [LTO17]. Identical problems are addressed from organizational point of view by [LJ17], which focused on Big Data Analytics governance framework. Finally, number of innovative research papers focused on integrating data mining with technical concepts and frameworks from other domains, for example, relational (symbolic) data mining methods [KV08] and game theory [Qin+15].

[9] All texts corpus with complete mappings is available at `https://figshare.com/articles/dataset/MasterList_xlsx/8206604`

To summarize, from 'extension' and 'integration' research, three important banking domain specific factors were identified. They require adjustments of existing data mining process frameworks and models. Firstly, potential discrimination in the context of credit decision-making requires financial services companies to adapt data mining to achieve transparency. Secondly, large number of accumulated data and associated complex IS/IT architectures, require adapting data mining process to address complex data mining models deployment patterns and implement them as component of complex systems and business applications. Thirdly, actionability of data mining results, adaptation of analytics outcomes to end, business-user needs are of utmost importance to achieve business value realization. We can hypothesize that in banking domain as the leading adopter of data mining solutions with significant investments, failures of realizing full business value of data mining projects are more explicit and observable and need to be addressed.

Summary In this research part, it was discovered that data mining methodologies are applied regularly since 2007 and their usage has tripled. Further, data mining in financial services domain is primarily used for two main purposes - to address Customer Relationship Management and Risk Management related business problems.

We have also identified that over the last decade, data mining methodologies have been primarily applied 'as-is' without modifications. Yet, we have also discovered an emerging and persistent trend of using data mining methodologies in banking with adaptations. Further, we have distinguished three adaptations scenarios ranging from granular modifications on tasks, subtask and deliverables level and ending up with merging standard data mining methodologies with other frameworks.

Data mining process adaptations are made to address banking domains specific factors, in particular:

- to tackle discriminatory awareness and transparent decision-making - *human-centric aspect*.
- to address actionability of data mining results - *business-centric aspect* - which plays a central role in the banking domain.
- to tackle lack of deployment and implementation aspects in the standard data mining methodologies - *technology-centric aspects* - which are required to scale and transform data mining models into software products and components integrated into Big Data Architectures. Therefore, adaptations are used to integrate data mining models and solutions in complex IT/IS systems and business processes of the banking industry.

Table 4. Key characteristics of methodologies adaptations discovered in SLR 1, SLR 2

Study - Adaptation	Purpose of Adaptation	Benefits
SLR 1-Modification	Specific domain-driven case studies in IS/IT, Manufacturing and Engineering, Sales and Services (incl. Financial industry)	'State-of-the-art' solutions to the specific business problem
SLR 2-Modification	Specific case studies tackling Customer Relationship Management, data-driven decision-making, Risks and AML-related business problems	'State-of-the-art' solutions to the specific business problem
SLR 1-Extension	- To implement fully scaled, integrated data mining solution, repeatable knowledge discovery, - To implement complex, specifically designed systems and integrated business applications, - To implement data mining as part of integrated/combined specialized environments, - **To incorporate context-awareness aspects**.	Practical IS/IT implementations, addressing complex deployment patterns
SLR 2-Extension	- To implement fully scaled, integrated data mining solution, repeatable knowledge discovery, - To implement complex, specifically designed systems and integrated business applications.	Practical IS/IT implementations, addressing complex deployment patterns
SLR 1-Integration	- To integrate/combine with various ontologies existing in organization, - To introduce context-awareness and incorporate domain knowledge, - To integrate/combine with other research/industry frameworks, - To integrate/combine with other well-known organizational governance frameworks, - To accommodate/leverage upon newly available Big Data technologies, tools, methods.	Improve actionability, address complex deployment patterns
SLR 2-Integration	- **To introduce discrimination-awareness in data mining**, - To integrate/combine with other organizational frameworks, - To integrate/combine with other research/industry frameworks.	Achieve decision-making transparency, improve actionability

Further, comparing key findings of the cross-domain (SLR 1) and banking domain (SLR 2) reviews (Table 4 below), the following is observed. The 'Modification' type of studies in both literature reviews are case studies providing solutions to the domain-specific problem. In cross-domain review (SLR 1), these are studies concerning IS/IT, Manufacturing and Engineering, as well as Sales and Services including Financial Services. In SLR 2, these are banking domain studies addressing business problems in Customer Relationship Management, Risks, and similar. The key benefits achieved by this adaptation type relate to identifying 'state-of-the-art' solutions to the specific business problems.

For 'Extension' category adaptations, the identical *technology-centric purposes* for adaptation are identified both in SLR 1 and SLR 2. They concern IS/IT implementation of data mining solutions and complex business applications and systems with data mining components. One adaptation reason for context-awareness aspects is present only in cross-domain review. The beneficial outcomes from this adaptation pattern relate to practical IS/IT implementations of data mining solutions and addressing complex deployment patterns.

Lastly, for 'Integration' category adaptations, the purpose to combine data mining process with other organizational and industry frameworks is identified in both the banking domain and in cross-domain review. In contrast, adaptations to tackle *human-centric aspects* associated with discriminatory awareness and transparent decision-making is the distinct purpose identified for the banking domain only. At the same time, adaptations for *context awareness and domain knowledge* are present only in the cross-domain review. The benefits associated with this adaptation pattern include improved actionability, addressing complex deployment scenarios, and achieving decision-making transparency to address potential discrimination.

Thus, *technology-centric aspects* and adapting data mining to *organizational aspects and other frameworks* are rather domain-agnostics, and have been demonstrated as important part of data mining process in the context of other sectors beyond financial services. In contrast, adapting the data mining process to incorporate *discriminatory awareness* is a distinguished feature of executing data mining in the financial sector. Also, while many case-by-case adaptations of standard data mining frameworks (e.g. CRISP-DM) in the banking are reported and their number is growing, there is no sector-specific data mining process proposed for the financial services.

3.2.3. SLR II: Threats to Validity

This research part has inherent threats to validity and limitations associated with the selected research method (SLR). The validity threats include incompleteness of search results (internal validity[10]) and general publication bias (external valid-

[10]The internal validity stems from subjective screening and rating of studies when applying relevancy and quality criteria

ity[11]). Internal validity was mitigated by strictly adhering to inclusion criteria, and performing significant validation procedures. With respect to external validity, we conducted trial searches to ensure validity of search strings and proper identification of potential papers. Our initial publications harvest size reached almost 700 texts originated from indexed peer-review research and 'grey' literature, thus mitigating external validity risk. Further, the key limitation of the SLR method for this study is that banking industry internal practices are not frequently disclosed in academic literature. We mitigated the negative impact by inclusion of 'grey' literature, where reporting on existing industry practices by professionals is common.

[11]The threats to external validity relate to the extent by which the results can be generalized beyond the scope of this study

4. PROBLEM INITIATION - CASE STUDY

In this Chapter, the last study, industrial case study, in *Problem Centered Initiation* activity (Figure 3 in Chapter 1) is presented. Initially, the case study design and its execution are described, followed by findings and discussion. Then, threats to validity are addressed.

4.1. Case Study Design

This part of PhD research investigates how standard data mining processes are extended and adapted in actual practice. Assuming 'micro' perspective, we investigate *what are the perceived gaps in the standardized data mining process (such as CRISP-DM), and what adaptations and workaround mechanisms are used by practitioners to address these gaps? (RQ2)*. This research question is tackled by conducting study in actual financial services organization. A case study is an empirical research method aimed at investigating a specific reality within its real-life context [Run+12], and it is particularly suitable when the defining boundaries between what is studied and its context are unclear [Yin17]. Case studies are commonly used for exploratory purposes and, therefore, applicable for addressing RQ2.

The detailed protocol followed in this case study is available at link[1]. In this protocol, we provide comprehensive information about the case study design including more detailed information about the interview questions, the steps taken to validate these questions, coding procedure of the data, responses from the interviews, etc. Here, we provide a summary of the case study design.

The first step in the case study design is to define its objective and research questions. We decomposed our research objectives into three components: perceived gaps, their respective impact, and the adopted workarounds. Accordingly, we defined three research questions:

- **Research Question III.I - What gaps in CRISP-DM practitioners perceive in the financial services industry?**
- **Research Question III.II - Why do practitioners perceive these gaps, i.e. what is the perceived impact of the identified gaps?**
- **Research Question III.III - How is CRISP-DM adapted to address these gaps**.

The second step of the case study design was to define the organizational context of the case study and the scope of the study. We sought an organization that fulfills the requirements of (1) operating within the financial service industry, (2) has systematically engaged with data mining over the last 3 years, i.e. has extensive experience with data mining projects of varying scopes and sizes, (3)

[1] `http://figshare.com/s/33c42eda3b19784e8b21`

uses CRISP-DM as guiding process for their data mining projects, and finally, (4) grants access to domain experts and documentation.

The setting for the case study is the data mining department of a bank operating mainly in Northern Europe. This department acts as a centralized data mining function (Center of Excellence), responsible for executing of data mining projects across the organization. The department's portfolio of data mining projects spans over several years and covers several geographical areas and business lines (cf. Table 5).

Table 5. Projects Characteristics.

No.	Project Definition	Geography	Project Type	Time span	No.of interviews	Participants
1	Product propensity model	1	Business-driven	2018	2	Data Scientist, Project Manager
2	Retail customers micro-segmentation	2,3,4	Business-driven	2017-2019	2	Data Scientist, Project Manager
3	Product propensity model	2,3,4	Business-driven	2018	1	Data Scientist
4	Lending process mining	2,3,4	POC (Proof of Concept)	2019	1	Data Scientist
5	Payments categorization model	2,3,4	Model rebuild	2019	1	Data Scientist
6	Graph analytics library	1	Capabilities development	2019	1	Data Scientist

In this organization, data mining is organized as project activity. We selected a representative subset of such projects that accounted for about 15% of the total portfolio. The selection covers four project types. The first is Business Delivery, i.e., the development of models for different banking products or complex algorithms for analysis of a bank's customers, such as private customers micro-segmentation. The second type is Model Rebuild. These projects share the commonality of rebuilding, retraining, and re-deploying existing models and algorithms. The third is a 'Proof of Concept' project that explored the use of new analytics techniques, namely process mining, for discovering improvement opportunities in lending processes. The fourth and last category is Capability Development, i.e., projects aimed at the development of competencies and tools for repeatable usage in other data mining projects. The selected project in this category concerns exploration

of advanced graph analytic methods and development of visualization algorithm library. It is intended to be used to support discovery of customer behavior patterns in other data mining projects. All projects relied on the CRISP-DM process, but with adaptations.

The third step of the case study was data collection. We approached this step in a two-pronged manner. First, we collected documentation about each project to familiarize ourselves with their objectives and requirements. Second, we conducted semi-structured interviews with data scientists and project managers of the data mining projects. The interview questions were derived from the research questions and literature review.

The interviews were transcribed (total of 115 pages) and encoded following the method proposed by [Sal15]. The first level coding scheme was derived and refined in iterations. It resulted in combining a set of initial codes (based on reviews and research questions) and codes elicited during the coding process. Second level coding, also obtained by an iterative approach, was based on themes that emerged from the analysis. The coding was performed by one author strictly adhering to research design (as in case study protocol) developed, validated and confirmed by all authors. The samples of interviews were coded independently by the two other authors, and results were compared for consistency and completeness. In case of deviations, results were discussed and approach agreed. The final coding scheme is available in the case study protocol.

4.2. Case Study Results

In this section, the results of the case study are presented. We address the case study research questions of perceived gaps (RQIII.I), their perceived impact (RQIII. II), and how the gaps have been addressed (RQIII.III). We have structured the results according to the main components of ITIL framework (Information Technology Infrastructure Library). ITIL is industry-agnostic, an accepted approach for management of IT services, and adopted across different business domains [MRL10]. It consists of three main elements; process inputs and outputs, process controls, and process enablers. We view data mining projects as instances of IT delivery and, thereby, are encompassed by the scope of ITIL. Therefore, the results for each of the five phases of CRISP-DM correspond to the main process according to ITIL, while aspects concerning process controls and enablers are related to CRISP-DM life-cycle as a whole. Thus, we first present the results for each of the phases of CRISP-DM. Then, we present results concerning CRISP-DM as a whole rather than specific phases thereof.

4.2.1. Phase 1: Business Understanding (BU)

The business understanding (BU) phase focuses on identifying business objectives and requirements of the project. Our study shows a significant inter-dependency between BU and the other phases. All interviewees noted 'numerous' iterations

and reversals back to the BU phase during the project. One participant expressed that BU *". . . had a lot of back and forth with business. It is basically spread over the whole duration of the project I would say."* As expressed, there was a *". . . lot of going back to the beginning, and a lot of finding something in the data, and then needing to understand what it means in the business terms. And then, based on that, changing the business requirements."* It seems that although such iterations are time-consuming, they enable adequate elicitation and management of business requirements.

The number and degree of iterations vary across projects. Projects with multiple stakeholders reported higher degree of iterations. One participant stated that *". . . the CRISP-DM process, when it is applied to use cases which are unsupervised, especially when there is some kind of segmentation exercise with a lot of different interested business counterparties, it is little bit more difficult to apply [. . .] because there's [sic] lots of going back to the business discussion, and scoping and Business Understanding part"*. More complex data mining solutions, such as project 2 that required layers of multidimensional calculations, reported more extensive iterations. Exploratory projects, such as project 4, required iterations when the obtained results were first applied to end-users. The introduction of new data types, and the discovery of previously unknown data limitations, necessitated reverting to the BU phase for continuous updating and understanding of the requirements. In this particular case, the BU was essentially intertwined with the data understanding phase.

Projects that deliver a model as a product (projects 1 and 3) reported fewer iterations. However, the BU phase was both demanding and crucial for delivery of the right product. One participant underscored BU's significance when expressing that the BU phase is *"one of the most important [. . .] just a little mistake on the focus and not understanding well what you are targeting [. . .] you have to start all over again"*. Another participant emphasized the necessity of the BU phase, and its iterations, because if *". . . you don't really exactly know the scope [...] you might have an idea and you need to present that, but then it can go back and forth a couple of times before you even know the actual population and what kind of products are we looking at ..."*. Unexpected iterations are also necessitated by the introduction of new regulations and compliance requirements (projects 2 and 3).

CRISP-DM does not fully reflect the interdependence between BU and the other phases. The main gap (RQIII.I) of BU is the lack of specific tasks and activities to capture, validate, and refine business requirements. This can cause a (RQIII.II) mismatch between a business' needs and the outputs of data mining projects. Furthermore, it can lead to missed insights and incorrect inferences. Practitioners commonly address this gap (RQIII.III) by iterating back to the BU phase in order to align the project outputs with the business needs, regularly eliciting new requirements, and validating existing ones.

4.2.2. Phases 2-3: Data Understanding (DU) and Data Preparation (DP)

The Data Understanding (DU) and Data Preparation (DP) phases concentrate on data collection, dataset construction, and data exploration. Here, the subjects highlighted a recurrent need to iterate between DU and DP, as well as between DU, DP and BU. The need for these iterations was highlighted specifically in the more complex project (project 2) and in both proof of concept projects (projects 4 and 6). In one of these latter projects (project 4), the three phases DU, DP and BU, had been merged altogether. The subjects indicated that the reason for iterating between DU, DP, and BU is that business requirements (identified in BU) often give rise to new data requirements or refinement of existing ones, and reciprocally, insights derived during DU give rise to observations that are relevant from a business perspective and thus affect the BU phase. Meanwhile, data limitations identified during DP may require stakeholders to refine the scope of the questions raised during BU. For instance, for one of the exploratory projects, it was reported that *"...basically to understand what kind of data we are going to work with, what dimensions it has and yes, how flexible we can be with respect to the final implementations, when we look at the data..."*

Data quality issues were continuously detected when working with new data types, methods, techniques, and tools. Such issues required referring back to the DU phase. Furthermore, modelling, analysis, and interpretation of results prompted replacing certain data points or enhancing the initial dataset with new data points. Such changes required an iterative process between DU and other phases. In Project 5, though it aimed at rebuilding and releasing updated version of already deployed model, data scientists had to redo the entire process. For instance, one interviewee expressed that *"I would say that from one side, we have this Data Understanding from first version, but due to different data preparation tools planned to use, it kind of required pretty much to start from scratch.... Data Understanding, obviously, we had some, and from the feedback we had some, but then [. . .] it's kind of requires completely different data sources. So, it's also a bit different Data Understanding, so it was not that straight."*

We also observed an important adjustment in regard to data privacy. CRISP-DM includes privacy as a sub-activity to the 'Assess Situation' task in the context of project requirements elicitation. However, GDPR, a recently introduced legislation to safeguard customer data, strictly regulates processing of personal data. Institutions can implement privacy preserving tools to reduce efforts and secure compliance. However, if such tools are lacking, the data mining projects have to include evaluation of data falling under GDPR and consider how to manage it (anonymize or remove). In our study, the interviewees underscored the GDPR requirements (discussed in *Data Mining Process Enablers*, subsection 4.2.6 Life-Cycle Gaps).

In summary, our findings indicate that DU and DP phases have inter- dependencies, data requirement elicitation, and privacy compliance gaps (RQIII.I). In

particular, the inter-dependencies between DU/DP and other phases are not catered for in the CRISP-DM process. Furthermore, CRISP-DM does not provide specific tasks for capturing, validating, and refining data requirements throughout its life-cycle. Tasks to ensure compliant data processing are also lacking. Such gaps prolong the projects' execution (RQIII.II). These gaps are addressed with iterations between the phases. In some cases, the frequency of the iterations causes such an intertwining between DU/DP and BU that the phases are practically merged into one. The iterations between DU/DP and BU are also employed when confronted with new data requirements and validation of existing ones, in particular, in regard to data privacy (RQIII.III).

4.2.3. Phase 3: Modelling

In CRISP-DM, the Modelling phase focuses on constructing the model after selecting suitable method and technique. The case study showed that the Modelling phase was not limited to prototyping only, as stipulated by CRISP-DM. Rather, models were developed in iterations (especially, in projects 1, 2, 4). The iterative approach to developing models required iterations between the modelling and the other phases, especially DU/DP, BU, and Deployment. We observed that iterations were born of the need to improve the models. For instance, the requirements discovered during the BU and deployment phases influenced which technique to use and how to design the models. One interviewee expressed that *"...there is one quite new dependency or requirement for our side, this is actually latency, because we need to classify or scoring part should happen very fast. ....even here in the Modelling phase, we kinda consider, that at least kept in mind this latency thing, or essentially what it means is how fast the scoring will happen."*

For models to be accepted, their outcomes have to satisfy pre-defined performance criteria as measured with evaluation metrics. In contrast to standard CRISP-DM, we observed that model performance metrics and requirements have been adjusted and adapted to business stakeholders' requirements, such as acceptable level of false positives, accuracy, and other criteria, to make the model fit real business settings and needs.

Projects with complex modelling tasks (projects 1, 3, and 5) adopted a distinct step-up modelling approach. These projects were characterized by first creating a baseline model (benchmark) followed by a set of experiments to identify how to best improve the models, i.e., satisfy specific performance metrics. *"... I think we just started off the model, any model just to get start, to get some sort of results to incorporate that in a pipeline [. . .] once we got one model up and running, then we started to incorporate several other models just to make any comparisons. [. . .] So, that's what we tried to, a lot of different models and, and we, we wanted to, we wanted the model to be suitable for amount of data that we had, the skewed data, the number of rows and the number of the attributes. And since the data was very skewed and we didn't have that many targets, so to say, then we, then we didn't*

want that many features and that's, that limited our dataset and in turn that limited which model we would use [. . .] So we compare these models also by different measures, and the one we ended up with stood out quite significantly. I believe that this was the way to go." Also, the Modelling phase explicitly incorporated elements of software development approaches resembling agile processes (project 1), specifically a Test-Driven Development approach (project 6). *" [. . .] we tried to then develop the actual function, and it could only pass that test if the criteria was met. So, it was a test-based or test driven implementation what we did [. . .] So even in the code we have all the test cases available....".*

Practitioners commented on the restrictive notion that the outcome of the modelling phase should be a model. They discussed situations where the results of the modelling phase were various interpretations of the model and different analytical metrics (projects 2, 4, 6). To this end, interviewees reported on both applying actual modelling techniques and executing algorithm-based data processing (e.g. using Natural Language Processing techniques) or experimenting with various process representations (in case of project 4). For one of the proof of concept projects, a practitioner noted that *" [. . .] it can be quite questionable what we consider as the modelling here ... process map, or the more formal process model in the process model language but as next steps in more advanced process mining projects, there could also be, additional models, for prediction and detection and so on. So, the process mining project can end up as a quite big project where many different types of modelling are involved."* Therefore, the Modelling phase can be defined as 'multi-modelling' with the set of unsupervised and supervised modelling outcomes.

In summary, the Modelling phase of CRISP-DM does not cater to needs born of the process of developing, improving, and refining models. Furthermore, explicit guidelines how to iterate between phases, in particular, the BU, DU/DP, and Deployment, are lacking. Refinement of existing requirements and capturing new requirements, which originate from the Modelling phase and other phases, is not supported. Finally, CRISP-DM is restrictive with respect to modelling outcomes in that they do not cater to 'multi-modelling', unsupervised modelling, and specialized modelling techniques (RQIII.I). These gaps can prolong data mining projects and increase the risk of mismatch between business need and outcome (RQIII.II). Commonly, practitioners address these gaps by employing an iterative and metric-driven modelling process, frequent iterations with other phases, and calibration with requirements from other phases. Also, tasks and activities are introduced to deliver various analytical outcomes ('multi-modelling') and to accommodate use of various techniques (RQIII.III).

4.2.4. Phase 4: Evaluation

The Evaluation phase is concerned with quality assessment and confirming that the business objectives of the projects are met. Practitioners underscored the

importance of validating and testing the models in a real usage scenario setting. While CRISP-DM prescribes assessing if the models meet business objectives, the 'how' is not discussed. As noted, *"... Crisp-DM should be updated specifically on the step of Evaluation to include how to test the model in business industry. I mean taking into account real scenarios, and there should be a list of steps in there. Which actually we have figured, figured out these steps [...] in an empirical way."* CRISP-DM prescribes a two-step validation. The first is a technical model validation, which is conducted in the Modelling phase and considers metrics such as accuracy. The second step assesses if the models meet the business objectives, which is conducted in the Evaluation phase. However, practitioners conducted these validations concurrently (projects 1-4). Stakeholders evaluated the models by considering the technical aspects, such as accuracy, along with assessing if the models are meaningful in a business setting. As one participant noted, the *"... important thing is that we like to think the evaluation through and really measure the thing that we want to measure and, and also not rely on only one measure, but can see the results from different angles."*

For unsupervised models (project 2 and 4), we noted that the evaluation was primarily subjective. The consideration was given to how meaningful the results were for the business, how the results could be interpreted, and to what extent actions could be taken based on the results. Thus, suitability and model usage, i.e., business sensibility, were the basis for model evaluation. For instance, one participant noted that *"... it is difficult to define some sort of quality measure for this kind of unsupervised result other than, well, actionability and future usage because we could have a quality measures for the clustering itself that just means that the clustering, cluster is distinct, but they don't mean that clusters are actionable for business and there can be non-distinct groups which on the other hand are interesting for business. So, there was, in this case it's kind of [...] technical quality measures are not necessarily suitable for a practical, practical quality."*

Our findings show that the Evaluation phase of CRISP-DM does not specify how models can be assessed to determine if they meet the business objectives. Specifically, there are gaps related to assessing and interpreting the models in their business context. Furthermore, the separation of technical and business evaluation, as outlined by CRISP-DM, can be problematic (RQIII.I). These gaps can lead to poor model performance in real settings and reduce actionability (RQIII.II). Practitioners address these deficiencies by piloting models in actual business settings (RQIII.III).

4.2.5. Phase 5: Deployment

The Deployment phase is concerned with implementing data mining project outcomes, so they are available and serve business needs of end-users. In CRISP-DM, tasks and activities concerning deployment are first considered in this phase. How-

ever, we observed the necessity to consider deployment strategy and elicitation of deployment requirements earlier (project 1, 3, and 5). As one participant noted, *"... when we develop a model, we think about what's important to us..and the business side, it could be interested in to see the results in a different way, or to include different columns or some things... So I understood after that process that one should have the Deployment phase already on your mind when [making] up the model, also, more or less from the very start, ...and to see the actual data that the business will pick up, and in the way they will pick it up...."* Furthermore, CRISP-DM does not cover deployment requirements well, thus forcing projects to cover particular deployment requirements, especially those concerned with the format of the deployed solution and its end-usage in business contexts (projects 1, 2, and 4). For instance, one of the participants stated that *"... the results were meant to be used on a daily basis by frontline people ... so, in this sense there are different levels of results that are needed, the more sort of complex ones, but then...there have to be some very simple KPIs and some very simple visualizations that don't need this more advanced process knowledge and understandable for everyone. So, that was something that we didn't know at the beginning that actually we need to report it not only to the Business Development department and process managers, but also to really frontline people ..."* Thus, the deployment phase can involve calibrating requirements to adapt models for their ongoing end-usage, for example, simplifying customer segment algorithms, so they can be implemented as rulesets that are easier to maintain and more stable compared to a pure clustering solution (project 2).

We also observed that the practitioners adopted a different deployment process from that of CRISP-DM. In CRISP-DM, the focus is on the deployment plan rather than implementation. Practitioners, however, reported using a wider range of deployment formats. For instance, projects based on unsupervised models might not require deployment at all, as their purpose is discovery of features and interpreting said features within the context of a specific business problem (project 4). One interviewee expressed, for project 6, that *"... we developed this as a like a library in Python, so that everyone can simply import this. So basically, we wanted to make sure that this is easy to use, and therefore the structure of the code was that you just need to download this package, and put it where in your project, [...] then you just in the beginning of your code import [...] and online functionalities would be available. And this is deployed on Git, so everyone can just clone and have this locally on their machine and use it. We were also planning to make it open source."*

Our main finding from the Deployment phase is that CRISP-DM begins the elicitation of requirements for deployment too late. The often-needed calibration of deployment requirements elicited in earlier phases of CRISP-DM, is not covered. Furthermore, this phase, as stipulated by CRISP-DM, assumes a restrictive stance and, as such, is not open to different deployment strategies used by practitioners. Lastly, CRISP-DM focuses on producing a deployment plan that does not address

implementation itself (RQIII.I). These gaps can cause projects to require more time than necessary to complete and, concurrently, gaps increase the risk of a mismatch between the project outcome and the intended end-usage, i.e., the business need (RQIII.II). Practitioners address these gaps by discussing deployment scenarios and eliciting deployment requirements early on, as well as extending the Deployment phase to include implementation tasks (RQIII.III).

4.2.6. Life-Cycle Gaps

We identified gaps that concern the whole CRISP-DM life-cycle rather than a specific phase thereof. Below, we present these gaps, organized according to two key pillars of the ITIL framework: process controls and enablers.

Data Mining Process Controls. ITIL identifies five process controls: process documentation, process owners, policy, objectives, and feedback. Furthermore, in the context of IT delivery, it specifies process owners, process quality measurement, and reporting as key controls. In our case study, practitioners highlighted three main aspects of data mining project controls– governance, quality, and compliance– which are in line with ITIL.

Our analysis shows that the practitioners have adopted elements of agile practices into their data mining life-cycles. This is explicitly visible in recent projects where 2-week sprints have been used, requirements are captured in epics, teams have daily stand-ups, sprint planning, and retrospectives. In the words of one interviewee, *"...it is very good that you have these two weeks sprints. So, I would say that you need to have independence for two weeks where you basically receive some goal and you try to achieve it. But in this sense, if you fail to achieve it, it's also kind of you arrive at the goal."* Practitioners also noted that CRISP-DM does not support the agility required for some data mining projects. There is a, as one practitioner stated, *"... flaw in this methodology... it just like tells you that's dynamic, but it does not tell you what to do, [. . .] when you have to go back to step 1, and how to do it faster, but yes, you have to improvise on the way."*

Another aspect, mentioned in relation to governance, is roles and responsibilities of both internal and external stakeholders. The importance of stakeholder management was emphasized. When stakeholders were not identified early on, *" [. . .] there was [sic] a lot additional tasks because not all stakeholders where identified at the beginning and secondly, each part of the segmentation, individual parts [. . .] ...could have different stakeholders....so we had many different stakeholders and we had lots of different stakeholders for different parts of the project."* The internal stakeholders' in-depth understanding of the business problem, and what the team delivering data mining projects can achieve, matter. For instance, in project 4, it was noted that *"...it can be two ways like either we present to the stakeholder a solution for a potential problem they could have, meaning we have to, to sell an idea. Or the other way around, they already have a problem that they have identified very clearly, and then they come looking for a solution, that we will*

provide. So I think as stakeholders understand more what we do it's more than the second one because they know we can help as they data scientists." External stakeholders (customers) are not included in the validation of the data mining solutions. Nevertheless, external stakeholders can, potentially, contribute to improving the quality of the results. As one practitioner expressed it, in relation to project 5, *"... another very important part would actually be the external stakeholders, the end customer, the client -I would be so very interested to get their point of view. The only thing we can measure currently is if they accept or not accept a product or service, but there will be a very interesting [to] involve,...to have them ask why they took a consumer loan or why didn't they, and in what way? [. . .] get a lot more information about the end user. [. . .] just to understand what, what is the actual driver. We can only read black and white on data, but we don't know if something else motivates them to do certain choices."*

Another aspect observed is that of quality. Specifically, we noted the evolution of adopting quality assurance mechanisms in the data mining process. These measures are expressed in the implementation of a formal peer-review process that is integrated in the project execution. Such quality assurances are visible as checkpoints– both in the daily work routines, and via review-based checklists. These quality assurance measures serve to validate five key aspects of data mining projects: (1) privacy-compliant data processing, (2) project scope, business goals, and data mining target, (3) input dataset quality and usage, (4) modelling method application, and (5) code quality and compliance (software development controls).

Lastly, participants highlighted compliance, in particular in regard to GDPR, as an example of external requirements that impact data mining projects. GDPR, in particular, has introduced a set of privacy-compliance tasks that must be considered. Furthermore, GDPR has limited the usage of certain data types and how final results can be used. For instance, in project 5, the company required customers to express GDPR consent, resulting in a limited number of customers using the data mining-based solutions.

Data Mining Process Enablers. The data mining process enablers, in this context, refer to capabilities required for the organization to be able to execute data mining projects. These enablers concern aspects that support projects that follow, to a different extent, the CRISP-DM process. The capabilities discussed are related to data, data mining code, tools, infrastructure, technology, and organizational factors.

Data quality, understood as reliability, persistence, and stability, was reported as crucial for all projects. Practitioners expressed that more important than tools is *". . . it's about the data because you have great tools, but if the quality of the data is not good enough, then, it doesn't matter, so to me this is like the most important thing".* Another practitioner stated that *" [. . .] I've heard the discussion of many times and for many years on the quality of data, and the thing people often are referring to is if they have like a lot of nulls maybe, or like missing fields in the data. So that would be one side of the quality, but that's just according to me lack*

of data, that wouldn't be, that wouldn't be really a concern in my mind. Quality of data would be that it's reliable and that the sources are stable and not changing. So that would be quite important, and I guess you just have to incorporate a lot of sanity checks in order to trust the sources. And it's also hard to discover if you have corrupted data..." For instance, in project 6, which tackled capability development (graph analytics tool), the crucial data characteristic identified was data richness and variety. As stated by a practitioner, *"...in a way it was relevant because we needed to make sure that we try the implementations with different sorts of datasets. And as I mentioned, we tried different topologies as an input. So basically, that was the quality of the dataset because if we had only one type of graphs, like only random graphs for example, and then the functions work well that wasn't enough. We wanted to have a random graph, we wanted to have a chain graph, we wanted to have clustered graphs and so on, the basically different topologies and making sure that the algorithm is meeting all the non-functional and functional requirements for all those input datasets. So basically, to ensure data quality, we just tried different topologies as an input..."*

We also observed a consensus that data should be made readily available for (self-service) usage and, as underscored by one interviewee, *"good databases are the key"*. Data consistency and completeness across various data sources, is another critical aspect reported in, for instance, project 4 (specialized process mining project). In this case, setting up correct workflow registration in source systems was a prerequisite to obtain acceptable data. In addition, it was emphasized that self-reported data was subject to biases and interpretations, thus it may be less reliable and, therefore, should be used with caution. The practitioners also referred to the quality of data mining project code. Its importance was chiefly noted for projects 5 and 6, in the context of scalability and optimization.

Available tools and infrastructure that ensure adequate prototyping, scaling, and deployment are regarded as necessary capabilities for all projects. Limitations to operate with large datasets and difficulties in applying methods and algorithms were cited as consequences of tools and infrastructure limitations in projects 2 and 4. *" [. . .] we had a lot of impediments on the technology side, in the sense that we were using quite big amount of data, and we were doing the analysis on local computer so there were some restrictions or issues with data size, sometimes the data size actually didn't allow to compute clustering quality measures, for example, because often the methods for that were not meant for big data in, well, implementation we have local machines and this data size also put restrictions on the algorithms that You can use, for example, it was not possible to use many different clustering methods because they just do not scale so we were somewhat limited in choice of algorithms....No hierarchical clustering....."*

Furthermore, a critical requirement mentioned, in regard to tools and infrastructure, was the ability to support automated, repeatable, and reproducible data mining deployments, such as regular automated dashboards. *"I would say it's important that we have, that entire pipeline, the infrastructure for the entire pipeline*

would be prioritized. So, so like going from start to end, maybe in a very thin or narrow manner in the sense that we might not have that many different systems or programs to use, but that we can deploy going also fast to, to market. That would be much more important, important to me than going wide and say that you have a lot of tools and infrastructure, but you cannot deliver in the end. So, it will be much more important to decrease the time to market, than have the availability for many different tools or languages or whatever that could be." Also, interviewees reported that available tools, platforms, and infrastructure had an impact on the choice of model design and language used. For instance, for project 6 *"....the main concern was to create a library that is usable inside our team. .....We needed something to work on the existing platforms [. . .] And we even considered a different programming language like Scala, as it could be more efficient. But since most of the end users, which are basically our team members were Python programmers, we decided to go for Python library [. . .]... So basically...the tools and technologies were defined by the criteria we had in the team, available to the team members. ...Because for the Python library, of course we could just do Python, but we wanted the solution to be also scalable for, for large graphs. And that's why we chose Spark...that depended on the...business requirements, and business requirements indicated that we need to work on large datasets.... And that also, also affected our choice of Spark."*

Finally, organizational factors, such as data-driven decision-making culture and maturity, have been referred to as crucial elements enabling adoption of data mining solutions in business practice (project 2). Interviewees referred to 'push-pull' paradigm whereby stakeholders actively 'pushed' for a solution initially, and with active participation have converted to 'pull'. Further, education of stakeholders to support data-driven decision-making culture thus transforming organization towards 'pull' paradigm has been emphasized and reported.

To summarize, we found that the CRISP-DM life-cycle has gaps related to governance, quality assurance, and external compliance management. Also, we found that CRISP-DM has gaps associated with data quality and stakeholders' management (RQIII.I). These gaps cause protracted project timeframes, a higher risk of mismatch between project outputs and business needs, and a negative impact on business value realization (RQIII.II). We found that these gaps are filled by adopting agile software development practices, specifically Test-Driven Development (TDD), Scrum ceremonies and Kanban boards, via regular interaction with business stakeholders across all CRISP-DM phases, and via integration of regulatory compliance requirements into the data mining process (RQIII.III).

4.3. Discussion

In this section, we present the identified gaps captured as six categories of CRISP-DM when applied to data mining projects within the financial industry. We address the research questions by presenting each category of gaps (RQIII.I), discussing

their perceived impact (RQIII.II), and how these gaps are mitigated by practitioners (RQIII.III).

A detailed summary of gaps is provided in Figure 18 below. It presents identified gap classes mapped to specific data mining life-cycle phases. In the context of each phase or entire data mining process, each gap, its impact and identified workaround(s) are specified. For example, Inter-dependencies gaps class consists of gaps 1,3,6 which belong to and are present in Business Understanding, Data Understanding/Data Preparation and Modelling phases. In each of the given phases, impact(s) of the respective gap and their mitigations are detailed.

The first category of gaps, Inter-dependency gaps, comprising gaps 1,3,6 concerns the lack of iterations between different phases of CRISP-DM. The practitioners expressed that such a gap leads to missed insights, skewed interpretations, and an increased risk of incorrect inferences. If the Interdependency gaps are not addressed, an increased effort in the form of re-work and repeated activity cycles is required, resulting in prolonging the project time. Practitioners address this gap by making numerous iterations between the CRISP-DM phases, and if needed, merging the two phases.

The Requirement gaps (Gaps 2,4,7,8,13) relate to the lack of tasks for validation and modification of existing requirements and elicitation of new ones. These requirement gaps are present in all CRISP-DM phases except for the Evaluation phase. These gaps increase the risk of a mismatch between the outputs of the data mining project and the business needs. Practitioners reduce the impact of these gaps by adding validation and calibration steps and by iteratively eliciting new requirements. Furthermore, practitioners also adopt software development support tools and incorporate elements from agile practices in their data mining projects.

The Inter-dependencies and Requirements gaps constitute the lion share of the gaps. These gaps stem from the largely sequential structure of CRISP-DM. Although iterations between the phases are possible, the procedural structure of CRISP-DM prescribes a linear approach where each phase is dependent on deliverables from the previous phase. Such a structure does fit projects where common understanding of deliverables evolve through collaboration, as is the case of exploratory projects.

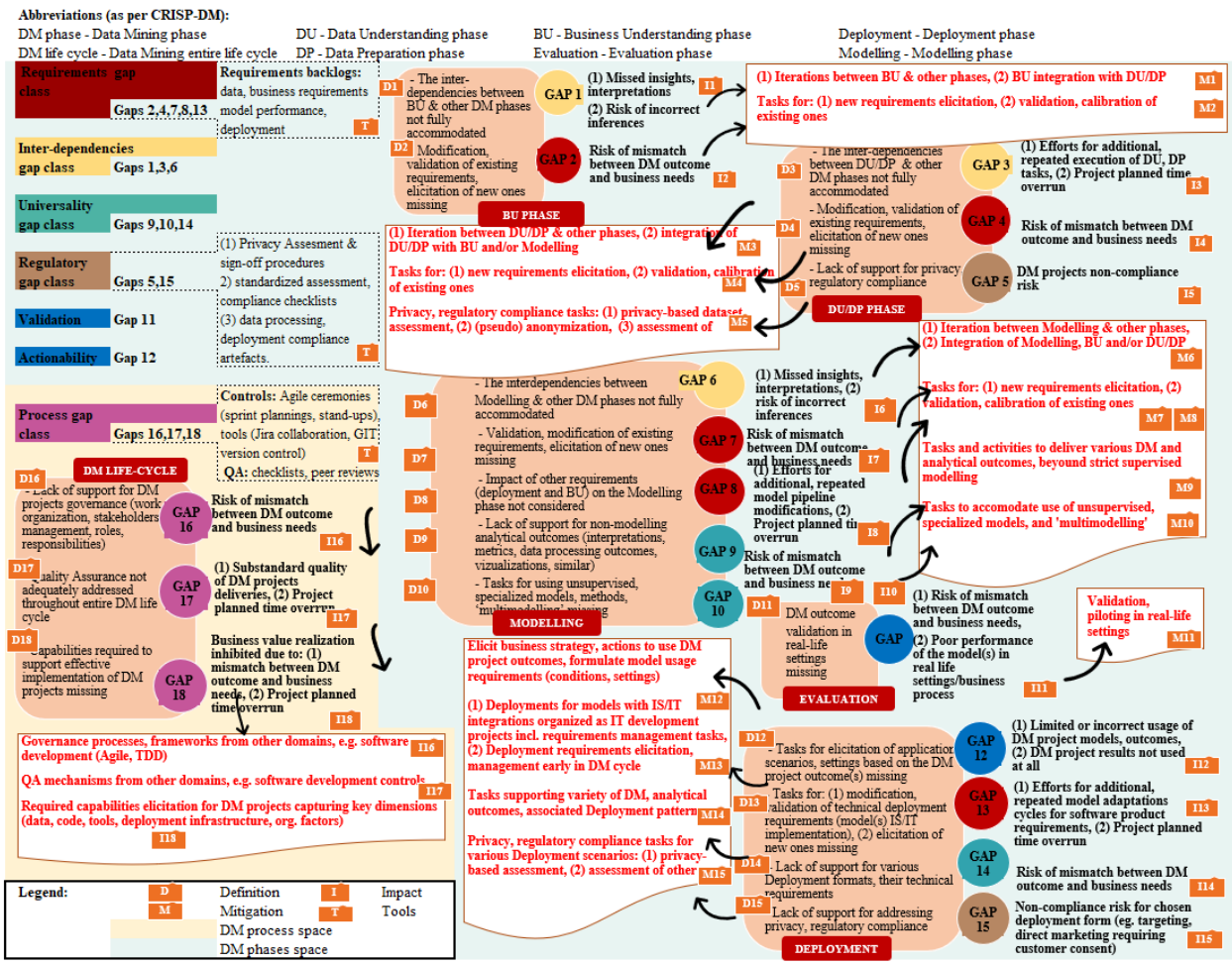


Figure 18. Identified Gaps Mindmap

The third category, Universality gaps (gaps 9,10,14) concerns a lack of support for various analytical outcomes, unsupervised and specialized techniques, as well as deployment formats. This category has been discovered for the Modelling and Deployment phases. Our results indicate that the standard CRISP-DM is, at times, overly specialized. In the case of the Modelling phase, it is restrictive in supporting standard, supervised, modelling techniques and associated data mining outcomes. For deployment, CRISP-DM does not provide tasks for implementation and associated technical requirements. These gaps lead to an increased risk of mismatch between data mining outcomes and business needs. Practitioners address these gaps by adding tasks to support unsupervised, specialized models' development and the delivery of various non-modelling analytical outcomes (multi-modelling) as well as different deployment formats.

Further, we discovered Validation gap and Actionability gap, which concern the Evaluation and Deployment phases respectively. These gaps refer to a lack of support for piloting models in real-life settings. Furthermore, CRISP-DM does not address elicitation of scenarios for model application, too. Thus, if models are not validated in practical settings, they are likely to exhibit poor performance when deployed. The lack of a model usage strategy, and an improper understanding of application settings, leads to limited or incorrect usage of the models, or result in models not being used at all ('producing models to the shelf' scenario). These gaps were filled by extensively using pilots in real-life settings, as well as addressing the actionability of the created analytical and model assets. These gaps Universality, Validation, and Actionability stem from CRISP-DM over-emphasis on classical data mining and supervised machine learning modelling. Data mining itself is regarded as mostly a modelling exercise, rather than about addressing business problems or opportunities using data.

The sixth category of gaps, Privacy and regulatory compliance gaps, deals with externally imposed restrictions. These gaps are related to the DU, DP, and deployment phases. CRISP-DM does not, generally, cater for privacy and compliance issues and, in particular, lacks tasks to address the processing of customer data. The impact of these gaps can result in non-compliance. Thus, practitioners have established standardized privacy risk assessments and adopted compliance procedures and checklists. These gaps stem from the fact that CRISP-DM was developed over two decades ago, when a different regulatory environment existed. Furthermore, CRISP-DM is intended for cross industry application. The financial service industry is extensively regulated, in particular in the usage of customer data.

We also identified Process gaps which do not concern a specific phase but, rather, the entire data mining life-cycle. These gaps encompass data mining process controls, quality assurance, and critical process enablers required for the effective execution of data mining projects. We note that CRISP-DM does not address projects governance aspects such as work organization, stakeholders, roles, and responsibilities. Furthermore, procedures for quality assurance are not provided

for. Also, required key capabilities, i.e., for data, code, tools, infrastructure and organizational factors, are not taken into consideration. These gaps can hamper organizational learning, reduce project effectiveness, and inhibit business value realization of data mining projects. Practitioners, in addressing these gaps, have incorporated agile practices into the execution of data mining projects, such as peer-reviews for quality assurance. These gaps appear because CRISP-DM only partially incorporates project management activities. CRISP-DM does not take broader organizational and technical aspects of project management into consideration. Thus, process controls and enablers needed to support multiple data mining projects on organizational level continuously are not addressed.

4.4. Threats to Validity

When conducting case study research, there are threats to validity that should be considered. In particular, these are construct validity, external validity, and reliability [Run+12]. Construct validity refers to the extent to which what is studied corresponds to what is intended and defined to be studied by the research questions. In our study, the interview method can be a source of construct validity risk. We mitigated this threat by including internal validity checkpoints for the purpose of verifying the interviewee's understanding of the questions. We also confirmed the content with participants via the interview transcripts. External validity concerns the extent by which the findings can be generalized. Case study approach has inherent limitation of generalizability, and further studies will be required to assert the generalizability of our findings. Finally, reliability concerns the level of dependency between the researcher and study results. We have tackled this risk by logging coding procedures and interviews and adopting an iterative research process with regular validations within our research group. We have also enhanced reliability of the findings by using triangulation of projects documentation and interviews. We also maintained appropriate chain of evidence, keeping track of the research materials and process and in that way ensuring replicability of the research steps and results.

5. FIN-DM DESIGN AND DEVELOPMENT

This chapter presents key method activities to tackle RQ3, *how the identified CRISP-DM gaps can be cohesively addressed by its extension in the settings of financial services domain?* The chapter describes the FIN-DM (Financial Industry Process for Data Mining) design and development cycle (Cycle 2 in Figure 3 in Chapter 1). We start with the inputs to this cycle– focusing on the derived catalog of gaps. Next, the design activity (A2.1 in Figure 3) is described– highlighting (meta)-requirements, design principles, and design features.

5.1. Inputs to FIN-DM Design - Gaps Catalog

Following a complete nominal sequential order of DSRM Methodology, all data originated from Problem-centered initiation was triangulated and consolidated as Activity 2.1 (A2.1 in Figure 3). Then, the gaps catalog was constructed (*Consolidated Input 2* in Figure 3), and used to design FIN-DM.

The findings of the two Systematic Literature Reviews described in sections 3.1 (SLR I) and 3.2 (SLR II) of Chapter 3, and industrial case study (Chapter 4) were consolidated. The data was in the form of gaps lists in the reference data mining processes (CRISP-DM) and they were available with various granularity and in various formats (as presented in Figure 19 below). In the case of SLR II, focusing on financial services, each *Gap* has a contained *Definition/Description* and key attributes, such as the *Adaptation scenario* it was associated with, and the *Publication* it was reported in. In the case of cross-domain SLR (SLR I) each gap contained, also, one more refined characteristic *Adaptations purpose/domain* with more granular sub-classification of executed adaptation.

For the industrial case study, the dataset contained a reported *Gap* with the five attributes. Besides *Name* and *Definition*, the *Origin* attribute was retrieved, which included reference to the CRISP-DM element where the gaps were identified - either at task, phase, or at the life-cycle level. As well, the *Impact* for each *Gap*– and respective *Mitigation* activities adopted by practitioners to alleviate it– was collected and included into the dataset. *Gaps* identified in the industrial case study have been classified into 7 gap classes based on the taxonomy developed in the study. Based on the *Gap Classes'* characteristics and descriptions in the study, *Gaps* discovered in both SLRs studies were classified and matched towards a common taxonomy proposed in the industrial case study (presented in Figure 18 in Chapter 4). In this manner, a consolidated *Gaps* catalog was constructed (*Consolidated Outcome 1*), and *Gaps class* has emerged as a unit of analysis.

Overall, there are seven gaps' classes in the gaps catalog (as specified in the Table 6 below). The majority of the gaps (6 out of 7) have been identified across at least two input studies. Over half of the gaps (4 out of 7) are not specific to the financial services sector and were reported in cross-domain research. This, thus,

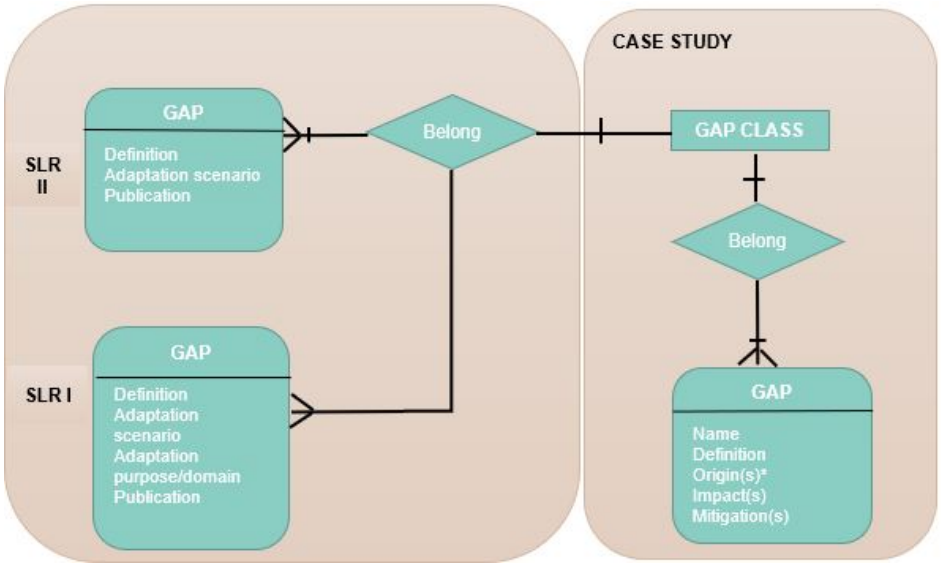


Figure 19. Triangulation of the data obtained in Problem-centered initiation

implies they are generic and universal by nature. Each gaps class is summarized below.

The *Requirements management and elicitation Gap class* (as per gap taxonomy in Figure 18 in Chapter 4) encompasses all phases of the standard data mining life-cycle, both in the context of financial services sector [PDM19], and in other sectors [PDM20]. This gap class has been reported as the most critical in the standard CRISP-DM model. It is related to a lack of structured and explicit tasks for requirements elicitation, development, and management. To this end, a vast spectrum of requirements is concerned, including business requirements, technological aspects (tools, platforms and implementation and deployment processes), and data mining itself – e.g., data, modelling and evaluation requirements. This gap primarily causes: (1) the risk of a mismatch between data mining outcomes, and (2) additional efforts (especially, for model deployment) and project time overruns.

The *Universality Gap class* is discovered to be present in the *Modelling and Deployment phases* of the standard data mining process. This gap class is related to a lack of support for non-modelling analytical outcomes apart from traditional supervised modelling, such as interpretations, metrics, visualizations, and the like. Another related aspect is a lack of support for 'multi-modelling' (applying such techniques in combination) and for various deployment formats and their associated technical requirements. This gap results in the risk of a mismatch between the delivered data mining outcomes and business needs.

The *Validation Gap class* has been identified in the *Evaluation phase*. This gap class relates to a lack of validation in data mining outcomes in real business settings causing a risk of: (1) poor model(s) performance when used in actual

Table 6. Consolidated 'gaps' classes catalog

Gap Class Name	Definition	Data Source (input study)
G1 - Requirements management & elicitation	a lack of tasks for validation and modification of existing requirements, elicitation of new ones	Case study, SLR I[1], SLR II[2]
G2 - Interdependencies	a lack of iterations between different phases of CRISP-DM	Case study
G3 - Universality	a lack of support for various analytical outcomes, unsupervised and specialized techniques, deployment formats	the Case study, SLR II
G4 - Regulatory & Compliance	a lack of tasks to address regulatory compliance (in particular, GDPR)	the Case study, SLR I
G5 - Validation	a lack of support for piloting models in real-life settings	the Case study, SLR II
G6 - Actionability	a lack of support for piloting models in real-life settings	the Case study, SLR I, SLR II
G7 - Process	data mining process controls, quality assurance, and critical process enablers (data, code, tools, infrastructure and organizational factors, are not taken into consideration) required for the effective execution of data mining projects	the Case study, SLR I

business activities and processes, and (2) a mismatch of the data mining solution to real business needs.

The *Actionability Gap class* has been identified as missing tasks for discovering application scenarios and business settings in which the data mining project outcome would be used. This gap class, while narrow in scope, is most likely to produce a disproportionately high negative impact by inhibiting the data mining project's business value realization.

The *Regulatory and Compliance Gap class* has been broadly identified as a lack of support for privacy and regulatory compliance in the standard data mining process. This gap especially impacts the *Data Understanding and Preparation phases* and project *Deployment* due to the risk of non-compliant processing of private customer data as stipulated by the GDPR[3]. Though GDPR is a general regulation, its highest impact is on industries collecting and processing customer data. Hence, the financial services sector is among the most affected sectors.

Finally, the identified *Process Gaps class* refer to a lack of three key components in the standard data mining process required for the effective execution of data min-

[3]General Data Protection Directive (2016/679) - an EU Regulation on data protection and privacy in EU and EEA countries

ing projects. In particular, a lack of quality assurance (i.e. lack of internal controls and quality assurance in data mining development), governance mechanisms, and capabilities. To this end, capabilities refer to a lack of critical data mining process enablers including data, code, tools, infrastructure, and organizational factors. As well, the given aspects lack in the context of a stand-alone data mining life-cycle. They are also considered as prerequisites to enable the continuous, industrialized execution of data mining projects at scale.

On a higher abstraction level, the standard data mining process model, to some extent, misses *agility* via a lack of iterativeness and recognition of some dependencies. As well, it is perceived as not *complete*, with regard to content, to cover regulatory and practitioners' needs. Lastly, the *organizational perspective* is not completely addressed– as the standard data mining process model does not support repeatable and reproducible data mining. Hence, from a design science perspective, *Improvement research* is needed to create a *new solution for a known problem* within a known application context [GH13]. This would tackle the discovered gaps in the standard data mining process. Thus, the objective of FIN-DM (*Artifact Objective*) is to provide practical solutions and mechanisms to mitigate or eliminate the identified gaps.

It should be noted, that the given gaps have been identified primarily in the context of CRISP-DM, however, examining their relevancy for other data mining frameworks is a point of interest. We have selected a sample of methodologies for such review, taking KDD, SEMMA as examples of well-known and widely adopted frameworks besides CRISP-DM. Also, we have considered CASP-DM (please refer to Table 1) as an example of new generation of recently proposed frameworks. We have analyzed whether the identified gaps (as per *Gaps catalog*), are also present in the selected frameworks. We have identified that, for instance, neither KDD process nor SEMMA address any privacy related aspects or regulatory concerns. Requirements management and elicitation are not explicitly tackled, and there are actionability gaps. Lastly, gaps related to lack of quality assurance controls and governance, as well as IT deployment and associated enablers are present too. We have also concluded that CASP-DM being recently proposed methodology addressed some identified gaps, e.g. CASP-DM tackled actionability, versatility and context-awareness. It has also proposed a number of quality assurance tasks and mechanisms focused on data quality. model results and performance. However, it has not addressed privacy aspects, and necessary IT controls and governance mechanisms.

5.2. FIN-DM Conceptualization

The design science as Information Systems research paradigm has emerged to address design problems characterized by unstable requirements [Hev+04]. Thus, design science approaches have incorporated the requirements' concept on a higher abstraction level, cf. [Hev+04], [Hev07]. In earlier works, cf. [Hev+04]), require-

ments are referred on the general level in the context of design artifact completeness and effectiveness characteristics, which are achieved if the requirements are satisfied. Later, one of the design science key frameworks, known as three-cycle view (as presented in Figure 20 below), explicitly included a requirements concept, cf. [Hev07]. There, it has been integrated in the Relevance Cycle, which initiates design science research as input for the research and relates them with acceptance criteria for the ultimate evaluation of research results. Iteration towards requirements in artifact design and evaluation cycles are suggested too. At the same time, these key design science frameworks have incorporated requirements on a higher abstraction level- such as research requirements - but not explicitly specifying them as requirements for the artifact.

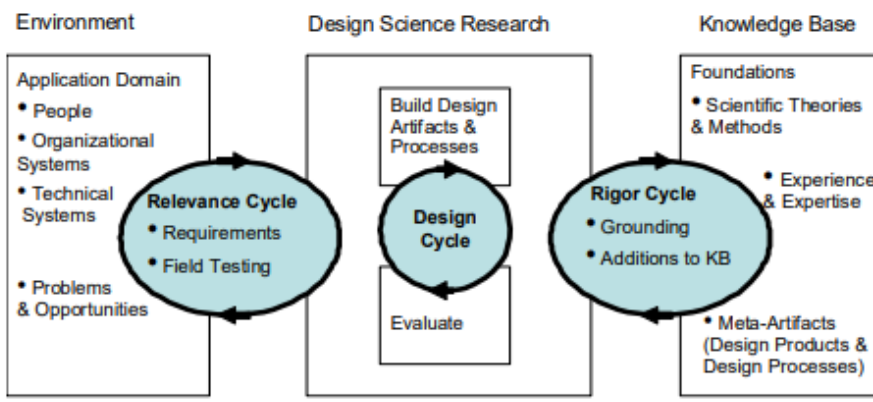


Figure 20. Three-cycle view on design science research, as in [Hev07]

The concept of artifact requirements in design science has emerged explicitly when different design science methodologies have been specified. One of the most known is [BPV09], which proposed a Soft Design Science approach combining Design Science and the soft systems methodology. In this work, concepts of general and specific requirements, including system requirements and their translation into artifact design requirements, have been emphasized. In the same vein, DSRM methodology used in this PhD research refers to the possibility to transform the problem into system objectives by means of meta-requirements or requirements [VPB17].

So far, there have been a limited number of works which proposed or adopted blended approaches by extending established design science frameworks with artifact requirements concepts. The most recent blended approach is suggested by [Bra+15], and refers to Requirements-driven design science research in the form of ontology, where requirements are integrated into all phases of design science methodology. In this way, the authors provide a tool to address the insufficient design support inherent to the majority of Design Science frameworks and models. A similar approach, but more in the settings of particular artifact design, has been

proposed by [Sil+14]– emphasizing usability-driven requirements. Lastly, recent works on using requirements engineering concepts for designing sector-specific data mining methodology have emerged [TVP17].

While such blended research has been proposed and reported on a limited basis, it provides a number of benefits relevant for the design and development of FIN-DM. Firstly, it is recommended as necessary step when designing socio-technical IS artifacts[4] , i.e. decision-support systems, modelling tools and alike, which comprise and embed IT, people, and processes simultaneously [Dre15]. FIN-DM has the features of a complex socio-technical artifact comprised of various components. For the complex design of such artifacts, due to their size and the number of components, the explicit formulation of design principles is recommended too [GH13]. Secondly, using explicit requirements and design principles allows us to define specific attributes and expected behaviors of the artifact when users interact with it. Then, the evaluation with the users in the form of interviews can be structured and defined precisely, based on a set of the concrete FIN-DM characteristics. That, in turn, supports and facilitates the processing of potential users' feedback and suggestions, and addressing it within the final artifact design. Therefore, we have adopted a blended approach, by augmenting DSRM method with requirements development.

Requirements for FIN-DM are derived based on the artifact objective, insights from the *Problem Formulation* and guided by related works practice highlighted above. Then, they are translated into design principles and features, hence satisfying the requirements (as presented in Figure 21 below). To this end, requirements (as in [MMM15]) relate to generic requirements that any artifact instantiated from this design should meet. They define *what* needs to be satisfied. In turn, design principles are generic capabilities through which requirements are addressed [MMM15], they define *how* (meta)-requirements are to be satisfied. Lastly, as defined by [MMM15], design features are specific ways to implement a design principle in an actual artifact, they are the key attributes of the artifact.

The common requirements' classification into functional and non-functional is used, cf. [PR14], [WB13]. In the context of the research, functional requirements define the content of FIN-DM – i.e., content attributes which enable users to execute data mining projects in the financial sector. These attributes make FIN-DM useful and relevant in this application domain. In turn, the non-functional requirements define FIN-DM from a user acceptance perspective, with importance placed on usability, flexibility, etc. [WB13]. As mentioned, design principles are defined following the pattern suggested by [CSG15]. Accordingly, it contains three types of information: (1) actions possible through the artifact, (2) the artifact's properties which facilitate the given actions, and (3) conditions under which such design will work (if applicable) [CSG15].

[4]Defined as artifacts possessing a combination of technical elements of information technology (concepts, models, methods, etc.) and/or the social elements (humans, roles, work processes, teams, groups, etc.) of organizations [Dre15]

To satisfy the *Artifact objective* of FIN-DM, the first meta-requirements (MR1) and design principles (DP1) are as follows:

MR1. *FIN-DM provides customization and user support for data mining execution in financial services -* ***DP1.*** *FIN-DM solves standard data mining process' 'gaps'– enabling users to address specifics of data mining in the financial domain effectively, given that users have general data mining knowledge and experience*

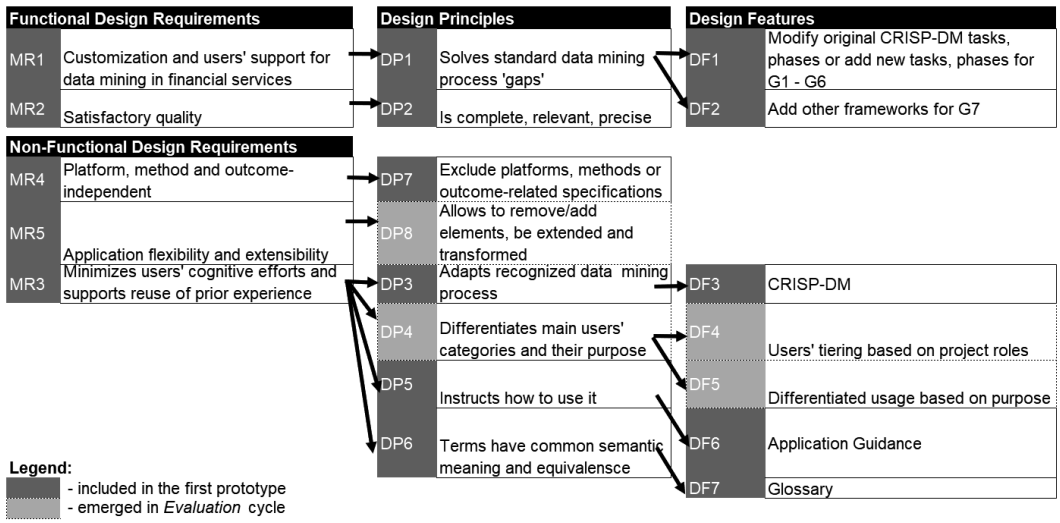


Figure 21. FIN-DM design requirements and translation into design principles and features

Next, given that FIN-DM is a user-oriented solution, it should address the identified 'gaps' and satisfy key quality characteristics (MR2). In particular, FIN-DM should be precise and understandable by users. Also, it should be complete in each element so that it is useful and serves the purpose well.

MR2. *The quality of FIN-DM and its elements is satisfactory -* ***DP2.*** *FIN-DM and its elements are complete, relevant, and precise, enabling users to improve data mining projects structuring, execution experience, and delivery time, given users have basic knowledge on applying structured approaches.*

FIN-DM is designed for practitioners; therefore, it should be easy to use– promoting a wider adoption and usage of FIN-DM. In the context of data mining projects, 'ease of use' could imply: (1) minimizing the potential user's cognitive efforts to get acquainted with FIN-DM, (2) supporting the reuse of prior knowledge of the data mining execution experience when applying FIN-DM. To satisfy such requirements, we should not oblige the user to learn a conceptually new process model, but rather to minimize familiarization efforts. Also, users need to be supported to reproduce the existing experience. To this end, FIN-DM should be based on a recognized standard data mining process. Hence:

MR3. *FIN-DM minimizes users' cognitive efforts to familiarize and apply*

FIN-DM by supporting the reuse of existing data mining process experience and knowledge. - ***DP3.*** *FIN-DM adapts a recognized standard process, thus enabling users to apply FIN-DM effortlessly without specialized, prerequisite training, given that users have general knowledge about most common data mining approaches (e.g., CRISP-DM, KDD)*

Further, to minimize users' efforts: ***DP4.*** *FIN-DM differentiates the roles of users, thus enabling users to apply FIN-DM effectively depending on their roles in data mining projects. This is under the assumption that the data mining project adheres to most common roles*

DP5. *FIN-DM includes instructions, thus enabling users to apply FIN-DM effortlessly. This is under the assumption that users have general knowledge about the most common data mining approaches (e.g., CRISP-DM)*

DP6. *FIN-DM terms have common semantic meaning and equivalence for users, thus enabling users to interpret FIN-DM's elements correctly and consistently. This is under the assumption that users have basic knowledge of most common data mining and IT concepts*

Next, to extensively support users and financial services organizations in data mining, FIN-DM possesses a number of flexibility and adaptability characteristics across broad ranges of contexts.

Firstly, while adapted towards financial services specifics, FIN-DM supports and guides potential users across a great variety of data mining projects and execution capabilities (flexibility in technical execution). Therefore, FIN-DM is independent of:

- environments - allows application in projects executed across a broad range of data mining environments, platforms, and tools,
- methods - is applicable for projects utilizing a broad range of data mining and modeling methods,
- outcomes - allows for a broad range of data mining and modeling outcomes.

Accordingly, this results in the following meta-requirement:

MR4. *FIN-DM is platform, method, and outcome-independent -* ***DP7.*** *FIN-DM and its elements exclude environment, methods, or outcomes-related specifications, thus enabling users to apply FIN-DM in any environment, with any methods, and for a broad range of data mining purposes within the financial services industry*

Secondly, to achieve complete flexibility in the project and organizational execution, FIN-DM needs to address two other flexibility dimensions, in particular: it needs (1) to be easily adaptable to the specifics of the particular project or organization (context adaptability), (2) to be extensible, so that practitioners and researchers can easily integrate new aspects into the existing artifact (general adaptability).

Thirdly, to achieve flexibility on a higher abstraction level, FIN-DM needs to possess dynamic adaptation mechanisms to stay relevant and adapt to internal and external environmental changes (general agility). Consequently, the following

meta-requirement is formulated:

MR5. *FIN-DM is flexible in application, extensible, and adaptable to internal and external environments changes -* ***DP8.*** *FIN-DM allows users to remove or add elements, accommodate extensions, make changes, and be permanently transformed. Thus, users are able to adapt FIN-DM without significant effort, given that users define the scope and purposes of such transformations*

As a final step of FIN-DM Design Activity 2.1 (A2.1 in Figure 3), the design features to satisfy design principles (as in Figure 21) were specified.

5.3. FIN-DM Development

FIN-DM Design Activity A2.1 was subsequently followed by *FIN-DM Development* Activity A2.2 (as in Figure 3). Following the design features DF1, DF2, DF3 (as in Figure 21), where feasible original CRISP-DM tasks and phases were modified to address the gaps. If, inherently, elements were missing in the original CRISP-DM framework, then, new tasks and phases were added. Finally, other relevant frameworks originating from other domains (e.g. IT domain) were considered and added when appropriate to address Process Gaps (G7) in CRISP-DM, which is transversal and encompasses all its phases.

Next, user-oriented design features DF6 (Application Guidance) and DF7 (Glossary) (in Figure 21) were added to FIN-DM documentation. Two other user-oriented design features were discovered in the evaluation process when interviewing potential future users (DF4, DF5). These design features have specifically concerned differentiating user's categories and the usage of FIN-DM. They have been implemented in the final iteration of FIN-DM Development.

In this manner, the initial artifact prototype (*Outcome 3* in Figure 3) was developed. Then, the fulfillment of requirements and design principles was implicitly incorporated in the *Evaluation design* and validated when conducting FIN-DM *Evaluation* with the future potential users. The details and evaluation are reviewed in 7 *Ex-Ante Evaluation* of the Thesis (Chapter 7). The feedback received from users triggered the iteration of the *FIN-DM Development* Activity (A2.2 in Figure 3) to produce a final artifact. The final version of FIN-DM and associated functionalities are reviewed in the following chapter in conjunction with the design decisions.

6. FIN-DM PROCESS MODEL

In this Chapter, the outcome of the PhD research, sector-specific data mining process model FIN-DM is presented and discussed. We start with *concepts* section where relevant definitions and concepts are provided. Then, we introduce FIN-DM structure, and it is followed by a section presenting how earlier identified *gaps* are addressed in FIN-DM. There, RQ3, *how the CRISP-DM gaps can be cohesively address by its extension*, is tackled by describing and discussing the solutions to gaps and the respective design decisions in detail.

All FIN-DM materials are available at links[1], and relevant selection of key components is presented below and in Appendix B.

6.1. Background for FIN-DM

FIN-DM retains key CRISP-DM terminology and elements structure, therefore, we review them here. In particular, apart from conceptual representation (introduced in Figure 6, *Chapter 2*) CRISP-DM is a *hierarchical process model* with four levels of abstraction (general to specific) consisting of *phases, generic tasks, specialized tasks, and process instances* [Cha+00] (hierarchical view is reproduced in Figure 22 below).

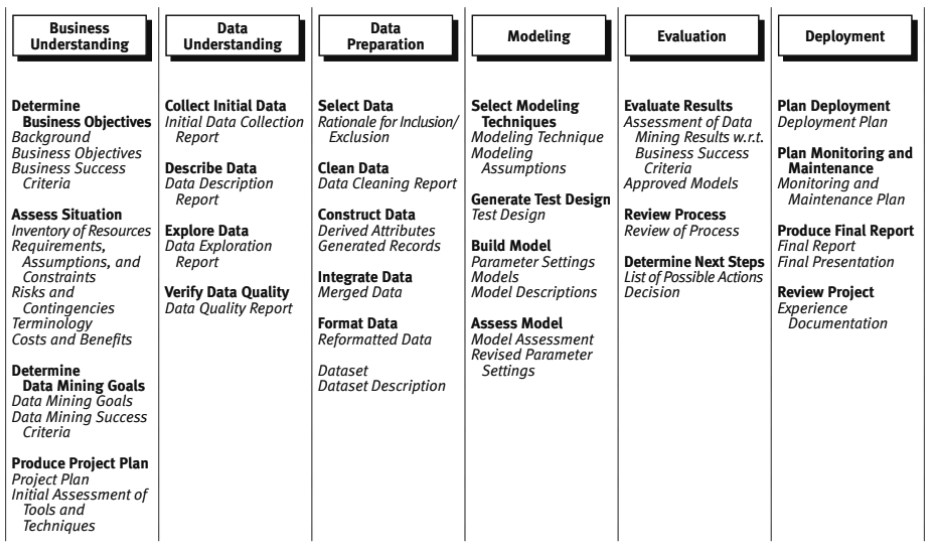


Figure 22. CRISP-DM Hierarchical Process representation [Cha+00]

[1]Permanent FIN-DM materials repository is placed at `https://figshare.com/s/e9fede5237a8577c8d0c`; website for user-friendly navigation of the FIN-DM is available at `https://fin-dm.info`

At the top level, the process is structured into six *phases*, each phase consisting of several second-level generic *tasks* with respective *outputs*. *Generic tasks* are designed to be general enough to cover all data mining situations [Cha+00]. Also, by design, both *phases and generic tasks* have two characteristics, namely, completeness (to cover whole process and all possible data mining applications) and stability (to be valid and flexible enough to accommodate new developments, e.g. new techniques, not embedded into process model design). These two abstraction levels constitute CRISP-DM Reference Model [Cha+00]. Data mining project lifecycle according to CRISP-DM starts with *Business Understanding* phase, which focus on determining projects objectives and business requirements. It is followed by *Data Understanding* (initial data collection and exploration), and *Data Preparation* with the dataset construction from raw data. In *Modelling* phase, various techniques are applied to built model(s), which further are evaluated in *Evaluation* phase. Lastly, if model(s) are assessed as meeting objectives and requirements and their performance is satisfactory, they are put into use in *Deployment* phase.

In CRISP-DM, there are also third-level *specialized tasks*, which are particular to data mining problem or situation or project specific as well as tool specific. They are further complemented by fourth-level *process instance* level with account of actual activities within concrete data mining projects. CRISP-DM itself focuses on generic *phases and tasks* level, while detailing specialized level tasks and instances are left to Reference Model users. Instead, mapping guidance with examples is provided to instruct users how to arrive from generic level to specialized [Cha+00]. Likewise, FIN-DM covers *phases and generic tasks* based on CRISP-DM definitions, while its application and scoping at specialized level is left to users.

Lastly, for FIN-DM to solve gaps associated with IT capabilities, and governance and controls, we propose to extend FIN-DM with elements of other frameworks, such as ITIL and COBIT. Extension approach is in detail presented and discussed in the sections 6.2 and 6.3 below, while here, we briefly introduce both concepts. ITIL[2] framework is the most widely adopted approach for the management of the IT services in organizations. ITIL covers the whole life cycle of IT services [Alm+20], and focuses on defining a comprehensive set of best practice processes for IT service management and support [Gul+10]. In turn, COBIT[3] is a comprehensive control and management framework aimed to ensure holistic IT governance and management throughout the organization. COBIT does not include process steps and tasks and due to such broad scope, COBIT is referred as 'integrator' establishing link between various IT practices and business requirements [Gul+10]. COBIT operates from the viewpoint of the entire enterprise, while ITIL focuses entirely on IT and associated service management practices. ITIL can be adapted and used in conjunction with other practices and extensively with COBIT,

[2]Information Technology Infrastructure Library

[3]Control Objectives for Information and Related Technologies

and both practices are viewed as complementary [Alm+20].

Reference Model of CRISP-DM provides overview of phases, tasks, outputs, and key activities in a data mining project. Apart from that, CRISP-DM contains user guide describing how to carry out a data mining project [Cha+00]. Likewise, FIN-DM contains supplements including Application Guidance and Glossary. Application Guidance describes to potential FIN-DM users its components and guides users in applying the model with practical examples, while Glossary provides explanations on relevant terminology and definitions used throughout FIN-DM.

Key distinction of FIN-DM are additional users' supporting elements and tools in the form of Enablers List and Checklists. In FIN-DM, Enablers[4] is preselected set of strategic IT Management practices and activities (based on ITIL), and governance and control practices and mechanisms (based on COBIT). It is intended to support FIN-DM users in establishing repeatable and reproducible data mining. The key Enablers are especially useful to be considered in the context of establishing industrialized data mining practices at scale and developing a portfolio of data mining services and projects. The list is also beneficial to evaluate in the context of individual data mining projects. In turn, Checklists (lists of tasks and questions) provide hands-on support and input for project participants to execute the data mining project in structured manner, keep track of its progress, scope and deliverables (e.g. manage backlog of the tasks effectively).

6.2. FIN-DM Structure - Conceptual and Hierarchical Views

FIN-DM consists of five items (Figure 23 below) structured into two distinct parts: (1) two representations of the proposed process, in the form of a Conceptual View and a Hierarchical Process View (including Enablers list), and accompanying Checklists, as well as (2) supplementary materials - the Application Guidance and the Glossary.

The FIN-DM components interact as follows. FIN-DM Conceptual View provides 'helicopter' view and assists with hands-on understanding on key phases of any data mining project. Also, it specifies concise view on key pre-requisites required for such project's execution. Especially, it would assist project managers in explaining data mining project(s) execution to the strategic leaders and functional managers. For the latter, it also gives a concise overview of the key pre-requisites/enablers required to manage and execute a portfolio of data mining projects and initiatives and perform data mining at scale. FIN-DM Hierarchical View equips data mining project participants with the detailed understanding of each data mining project phase, the sequence of activities within each phase and

[4]Broadly defined as anything that can help to achieve the objectives of the enterprise. Typically, enablers contain principles, policies, and frameworks, processes, organizational structures, culture, ethics and behavior, information, services, infrastructure and applications, people, skills and competencies [Nyi15]

outputs. Lastly, FIN-DM Checklists provides hands-on support in concrete tasks execution, progress tracking, etc.

The FIN-DM Conceptual Representation consists of the three components (visualized left to right in Figure 24): (1) ITIL-based wing - the foundation capabilities, (2) the data mining process (FIN-DM), and (3) COBIT-based wing - governance and controls.

The FIN-DM process model is the central component; it is a hierarchical process model. The FIN-DM process consists of three cyclic sub-processes visualized as rings - the inner, the middle, and the outer rings. The inner ring contains seven *phases*, adapted and extended from CRISP-DM. Acknowledging numerous iterations in between any phases, the process is represented as fully recurring. The middle ring is the *Requirements phase*, while the outer ring is the *Compliance phase* - they complement the data mining process (inner ring) end-to-end. Each phase consists of several generic *tasks* with respective *outputs* (see Figures 25,26 below). By design, both *phases and generic tasks* have two characteristics, namely, independence (DP7) and flexibility (DP8). As noted, FIN-DM, itself, is based on a recognized standard data mining process, CRISP-DM (DP3, DF3), hence FIN-DM retains key CRISP-DM terminology and structures (in inner ring).

The ITIL-based wing in the form of *Foundation capabilities* element is positioned as a specialized *framework* extension of FIN-DM– it is relevant and applicable to the whole FIN-DM life-cycle. Therefore, it is placed on the same hierarchical level as the process model itself. The *Foundation capabilities* are based on frameworks or elements originating from ITIL. The COBIT-based wing in the form of *Governance and Controls* element is also a specialized *framework* extension of FIN-DM. These two higher-level supplementary extensions provide guidance towards FIN-DM process elements and their execution (as visualized by two input arrows in Figure 24).

To this end, to satisfy FIN-DM's key research objective – to solve standard data mining process' gaps– the proposed solutions to 'gaps' (DP1, DP2) are embedded top-down at three levels of FIN-DM with decreasing abstraction, such as:

1. extensions with *frameworks* - in *Conceptual View* in Figure 24 above,
2. extensions with *phases* additions and modifications of existing *phases* - in *Conceptual View* in Figure 24 above and in *Hierarchical Process Model* view (as visualized in Figures 25, 26 below),
3. additions of new *generic tasks* and modifications of existing ones - in *Hierarchical View* in Figure 25, 26) below.

Lastly, FIN-DM phases, tasks, and other frameworks elements are to be evaluated in the context of the data mining projects. Only relevant elements are to be picked up, while irrelevant can be freely omitted. Also, all four accompanying checklists can be evaluated, and relevant items used (entirely or some parts). FIN-DM allows iterating between all phases in any sequence. Further, users are not prescribed to start with Business Understanding, but encouraged to evaluate

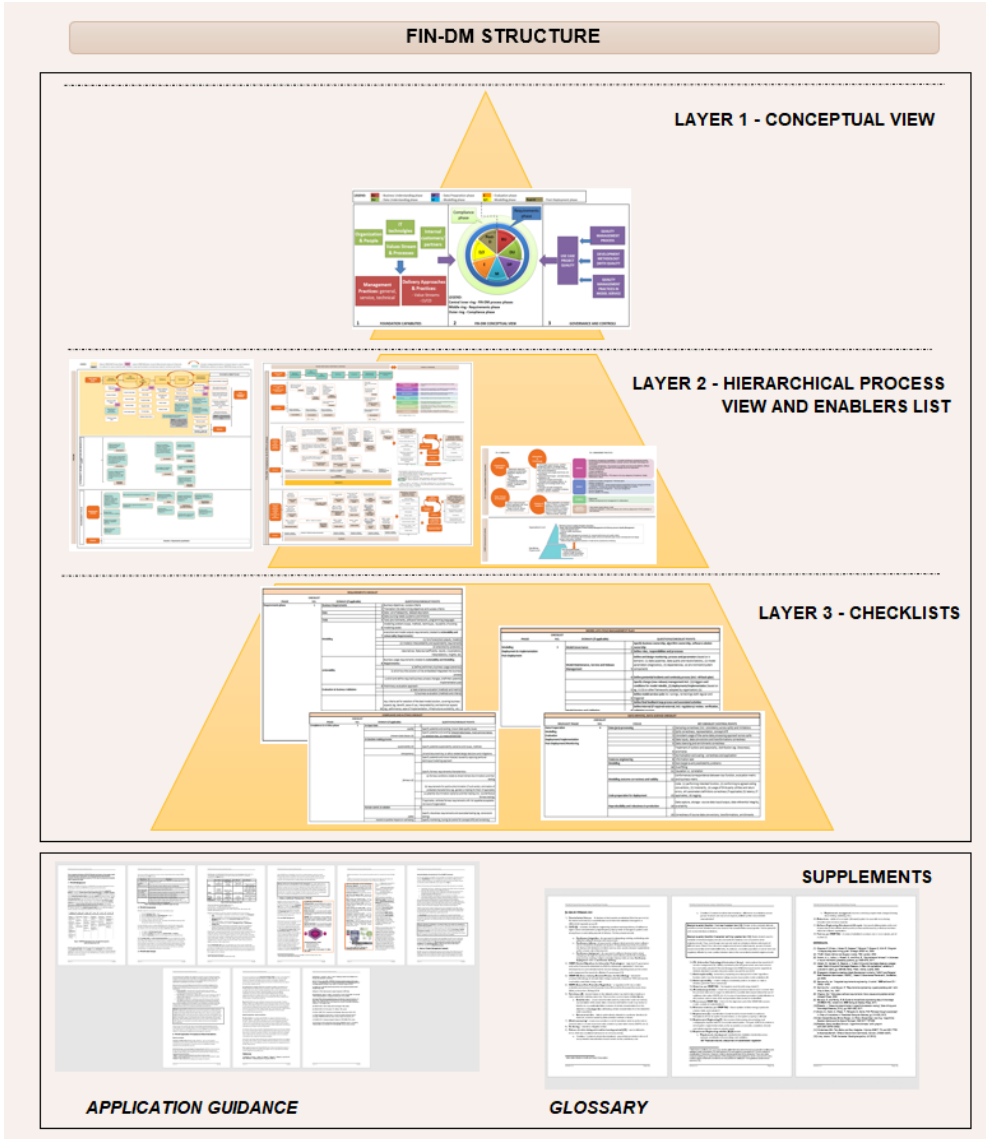


Figure 23. FIN-DM Structure

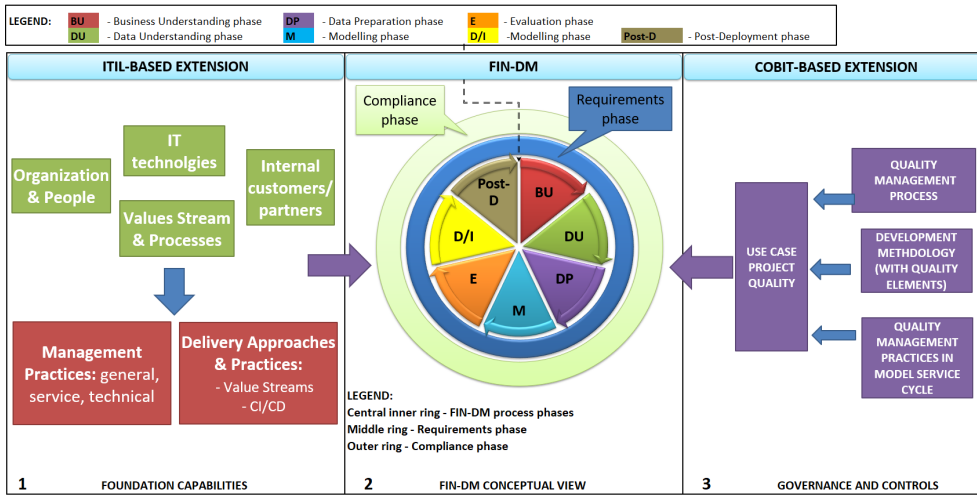


Figure 24. FIN-DM Conceptual Representation

depending on the project which phase to start with, e.g. with Data Understanding or combining Data and Business Understanding.

6.3. FIN-DM Solutions to Gaps

This section reviews FIN-DM (based on a detailed Hierarchical View in Figure 25, 26 below), focusing on the gaps solutions. We review them in a 'bottom-up' manner, starting with the new *tasks*, *phases*, and finally *domain extensions*.

In developing solutions to gaps, we have followed a structured, five-step approach. Initially, we have reviewed the financial service industry's general regulatory and governance best practices, recommendations, and requirements from a domain specific perspective. Then, from a data mining process perspective, we drew an analogy between IT delivery projects, software development practices, and data mining projects. We reviewed if any solutions were reported in the context of these domains. Next, we formulated a set of constructs or aspects for the respective gap and conceptualized the relationships between them and the characteristics of the data mining process. In this way, conceptualizations were derived for the *Compliance and AI Ethics phase*, which are presented in *Concepts universe* in Figure 26 below. As a final step, we addressed each of the required attributes or characteristics via the actions-tasks presented in *FIN-DM tasks and phases universe* in Figures 25, 26 below. Each action-task is tagged with the respective attribute and characteristics it tackles.

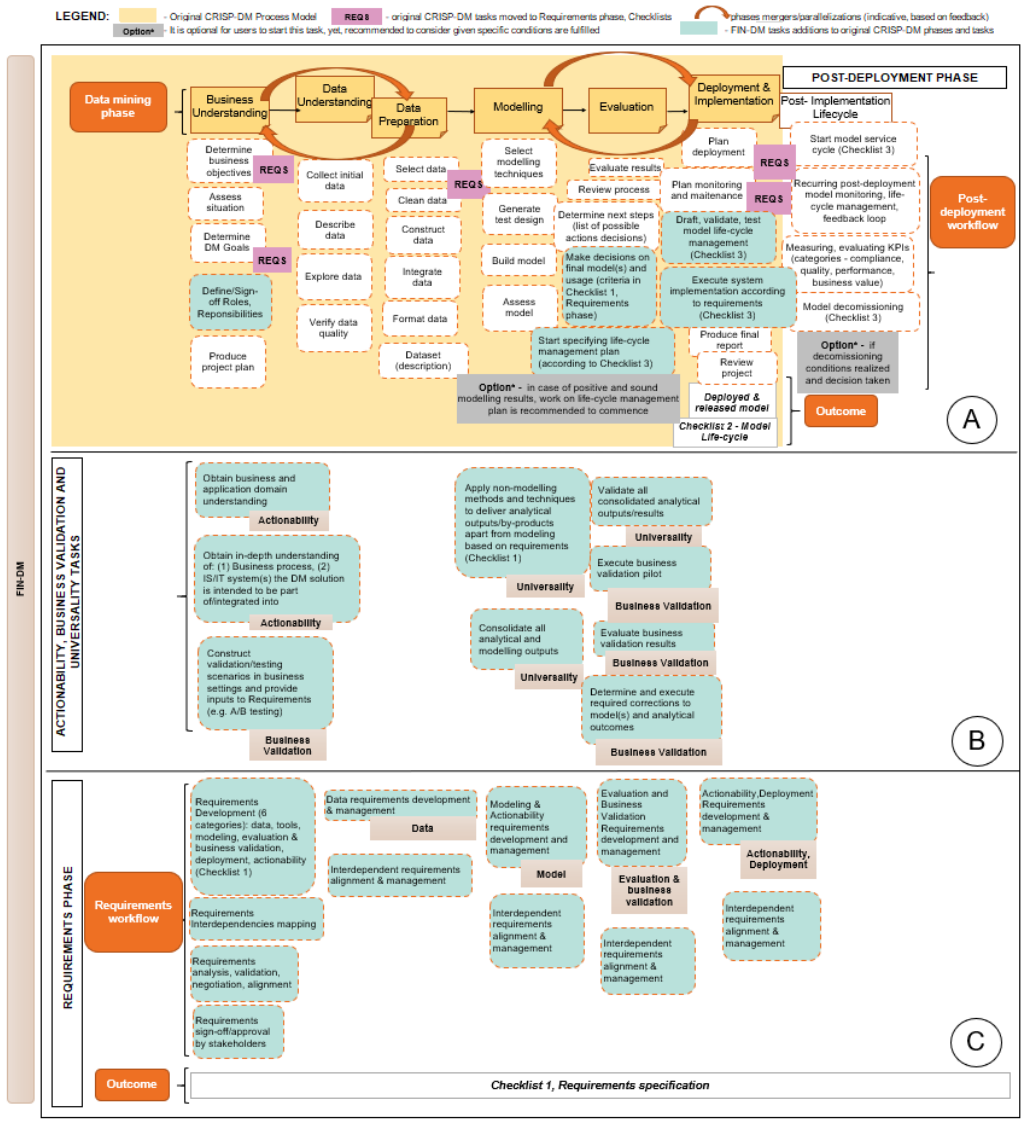


Figure 25. FIN-DM Hierarchical View - Additional Tasks, Requirements and Post-Deployment phases

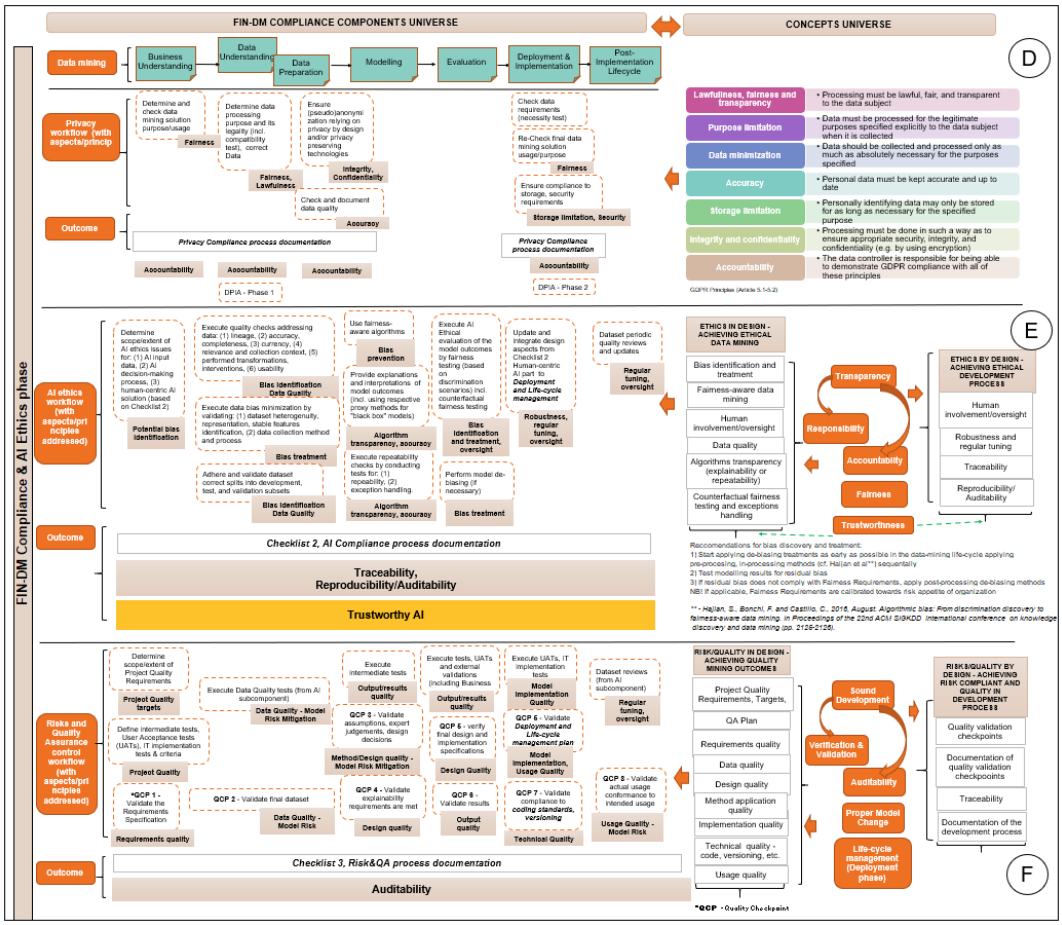


Figure 26. FIN-DM Hierarchical View - Compliance phase

6.3.1. Addressing Universality Gap

This gap is mitigated by: (1) including an explicit requirements elicitation and development for analytical by-products and their associated deployments in separate *Requirements phases* (reviewed below), and (2) augmenting the original CRISP-DM phases with respective tasks of *Applying non-modelling methods and techniques*, *Consolidating all analytical and modelling outputs*, and *Validating all outputs with stakeholders*. These additional tasks have been integrated into FIN-DM's *Modeling* and *Validation phases*, respectively (section B in Figure 25).

6.3.2. Addressing Validation Gap

This gap is addressed by embedding explicit *Business validation tasks* (section B in Figure 25). In particular, in the *Business Understanding phase*, we formulated *Construct validation/testing scenarios in business settings*, for example A/B testing as a common approach. Next, the respective scenarios are implemented in the *Evaluation phase* by *Executing* and *Evaluating business validation pilot*. Importantly, as a feedback loop, validation results are fed into the model's final design through the *Determine and execute required corrections* task.

6.3.3. Addressing Requirements and Actionability Gap

In proposing a solution to this fundamental gap, we have considered various existing requirements engineering methods and approaches, including the well-known viewpoint-oriented approaches for elicitation and analysis, formal mathematical methods, and goal-oriented approaches. While adopting the methods' perspective could be beneficial, it might not support satisfying *Independence* (MR4, DP7) and *Application flexibility and extensibility*, (MR5, DP8) meta-requirements, and design principles. Furthermore, specific method implementation might require the involvement of potential users' in mastering the particular requirements engineering method, which contradicts *Minimization of users' cognitive efforts* (MR3) meta-requirement of FIN-DM. Therefore, we opted to incorporate *Requirement activities* which are agnostic and well-recognized.

Hence, as the solution to the gap, we have proposed a new *Requirements phase* (section C in Figure 25) to augment the FIN-DM process and run in parallel to all process phases. Based on [Som05], we complement the *Business Understanding phase* with key requirements tasks including: (1) *Requirements Development*, (2) *Requirements analysis and validation*, (3) *Requirements inter-dependencies mapping*, and (4) *Requirements negotiation and sign-off/approval with stakeholders*. Throughout the rest of the data mining life-cycle, there are two *Requirement tasks* undertaken in each phase, namely– *Requirements development and management*, which is focused on the respective phase's scope (data, modeling, evaluation, deployment, respectively), and *Requirements alignment*. Aligning to CRISP-DM's standard process, the *Requirements phase and tasks* have a specific output, which is the *Requirements Specification*. As well, this phase has a supporting Checklist 1

(Appendix B.1, Figure B2) to assist in documenting and managing requirements from the inception of data mining life-cycle up until its end. The checklist is structured based on a high-level area of a data mining project, to which respective questions/points are mapped. Overall, we have identified six key areas - data, tools, modeling, actionability, evaluation and business validation, and deployment. To summarize on the *Requirements phase* proposal, FIN-DM is explicitly systematic and structured on the requirements elicitation and management throughout all data mining life-cycle compared to CRISP-DM. Further, requirements are elicited and managed for all key areas of the data mining project with Checklist 1 support.

We also propose to address the *Actionability* problem as part of the *Requirements phase*, eliciting *Business usage requirements* which cover four interrelated items: (1) preliminary business usage scenario(s), (2) how the solution would be embedded/integrated into the business process, (3) what the prerequisite business process changes are, and (4) their potential implementation plan. This is reinforced by two additional tasks (section B of Figure 25) aimed at obtaining an in-depth understanding of the business domain and business process, as well as how the respective data mining solutions will be used. Finally, as a means to address *Actionability* we propose a separate *Implementation/Post-Deployment phase* (section A in Figure 25), which explicitly formulates model deployment associated tasks and a data mining solution life-cycle management plan and activities (both not present in the original CRISP-DM model).

The proposed practical tasks follow and resonate with the *Domain-driven knowledge discovery* concept and associated Domain-Driven in-depth Pattern Discovery (DDID-PD) framework (cf. series of works [CZ06a], [Cao+07], [CZ07], [CZ06b]).

Domain-driven actionable knowledge discovery focuses on domain-driven discovery of knowledge, acknowledging existence and necessity to accommodate a number of constraints [CZ06a]. The four key constraints relate to: (1) domain-specific business rules and processes (so-called *domain constraints*), (2) quantities, complexity and source variety of data (so-called *data constraints*), (3) interestingness gap emerging as discovered patterns or insights while interesting for research are not actionable to business (*interstingness constraint*), and (4) limitations to real-life deployments if not integrated with business rules, process, etc. (*deployment constraint*) (cf. [CZ06a], [CZ06b]). To overcome the given constraints, in-depth investigation of the business context constraints, domain-associated data, involvement and cooperation with business users, and iterative feedback loop are recommended (cf. [CZ06a], [CZ06b]).

In line with these suggestions, the tasks for *Business usage requirements* elicitation, including close interaction with business users throughout all data mining life-cycle, as well as tasks focused on gaining in-depth understanding of business domain support understanding of *domain constraints*. *Data requirements elicitation and management* support tackling *data constraints*. In turn, iterative validation of data mining outcomes with business users (including final sign-off), testing in

real-life settings address *interstingness constraint*. Lastly, *Implementation/Post-deployment phase* support technical implementation of the data mining solutions in IS/IT systems and in the business process.

6.3.4. Addressing Regulatory and Compliance gap

Relevant financial services regulatory and governance frameworks As noted, this gap, in the context of financial services, is primarily associated with privacy regulation, GDPR in particular. At the same time, GDPR is not the only legislation applicable in the context of financial services, which traditionally have been heavily supervised and regulated. Therefore, we enhanced the regulatory and compliance scope of FIN-DM to tackle other core regulations applicable beyond privacy (fulfillment of MR1, MR2 and DP1, DP2), such as *AI ethics* and *Risk Management*.

In particular, many emerging policy initiatives and regulatory frameworks concern *AI ethics* and the associated risks inherent in the data mining process. Similar to privacy legislation, AI ethics guidelines are general and impact businesses and sectors developing and using AI-based solutions and products, including the financial services sector. FIN-DM tackles also financial sector specific regulations. In particular, following the aftermaths of the 2007-2008 financial crisis, key regulatory requirements, implemented over the last decade, focused on the systemic stability and management of risks in the financial industry [KI20]. These regulations concern enhancing risk governance in financial institutions, which in the context of data mining relates to governing and managing *model risk*[5] emerging when data mining models are developed and used. Risk governance requirements have been addressed and implemented in the financial services sector via the generally accepted 3LoD (three Lines of Defense model) organizational model. Accordingly, data mining experts in financial services organizations belong to the first line of defense; they have to manage and control model risks as part of the model development process. Drawing on an analogy to the IT domain, a software developer is an example of a front-line expert in the first line of defense, and managing risks in a software development life-cycle as part of daily operations.

To tackle the compliance gap and to address *privacy, AI ethics, and risk management* in data mining, we introduced the *Compliance phase* (visualized in Figure 26), consisting of three distinct sub-components for each area (areas D, E, F respectively in Figure 26). Similarly to the *Requirements phase*, this new phase covers the data mining life-cycle end-to-end.

Tackling Privacy In designing the *Privacy sub-component* (section D in Figure 26), we were guided by conceptualized relationship between GDPR principles[6] and data mining privacy risks (see *Concepts Universe* in section D of Figure 26).

[5]Model risk is defined as a sub-type of operational risk and refers to the potential for adverse financial, reputational, or regulatory consequences of decisions based on incorrect or misused model outputs and reports

[6]GDPR stipulates 9 core principles - Fairness, Lawfulness, Minimization, Integrity, Confidentiality, Accuracy, Data Storage Limitation and Security, Accountability

To this end, GDPR has been selected as the most notable and expansive privacy regulation currently in force. The regulation applies to the processing of personal data of individuals located in the European Economic Area (EEA), and therefore applicable to any enterprise doing such processing inside the EEA. Also, GDPR core principles have been and are continued to be adopted in the new privacy regulations worldwide. Concerning risks, there are two key risks[7] of transgressing privacy regulations in data mining. The first one concerns illegitimate personal data processing while developing the model or executing analytical tasks (impact on lawfulness and fairness principles). The second risk relates to the deployment and usage of data mining models in direct marketing, automated decision-making, and profiling without appropriate legal grounds (affecting lawfulness). In earlier works, these risks have been addressed with Privacy-Preserving Data Mining (PPDM) (cf. [MV17]), which suggests specific methods and techniques for safeguarding privacy of data collection (lawfulness), data mining output (accuracy), data distribution, and data publishing (integrity, confidentiality, security).

Combining both perspectives, in the *Privacy sub-component* (see section D of Figure 26), we initially propose to determine the purpose of the data mining solution and personal data processing. By assessing legitimacy, it is possible to prevent the risk of illegal data processing by updating and removing respective data requirements or, alternatively, ensuring legal grounds. As a next step, Data requirements are checked for adherence to Data Minimization principle [8]. In the Data Preparation phase, explicit tasks for (pseudo)anonymization are proposed, either by applying respective techniques or through privacy-preserving technologies (ensure privacy by design). That step is complemented by data quality checks to comply with accuracy principles. The legality of the final data mining results or model usage for the intended purpose is checked, along with security and storage requirements. The Accountability principle is followed throughout the data mining life-cycle by documenting the privacy compliance process and outcomes.

Tackling AI ethics Addressing issues represented by common *AI ethics* constructs is the first step to establish a framework to build ethical AI solutions [SW20]. The four constructs - *Transparency, Accountability, Responsibility, and Fairness* - are regarded as central in AI ethics[9]. These constructs have been adopted when tackling *AI ethics* issues in IS development (cf.[VKA19]). In a similar vein, [Dig17] proposed a so-called ART principles framework combining *Accountability, Responsibility, and Transparency (ART)* as means to design and engineer ethical automated decision-making mechanisms in AI systems. ART principles serve as

[7] The other interrelated aspect is personal data collection without legitimate grounds, but since data is used as input in the data mining life-cycle it is out of scope

[8] GDPR requirement to collect and process only as much data as absolutely necessary for the purposes specified

[9] There is no, yet, universal agreement achieved on these values being the core of AI ethics [VKA19], though there are alternative principles proposed (cf. [Mor+20]) with Beneficence, Non-maleficence, Autonomy, Justice, and Explicability constructs

basis for EU legislative discussions, augmented with Trustworthiness. Similarly, they are also covered in the IEEE EAD guidelines[10], while *Fairness* is considered by professional organizations (e.g. ACM) and in other key geographies (e.g. US) [VKA19]. In our solution we addressed a combination of *ART principles*, *Trustworthiness*, and *Fairness*.

In the context of this research, *Transparency* is determined for:

1. the data used in data mining and data mining algorithms [Dig17], and
2. the data mining development process - here, it refers to providing understanding why the model acts in certain way, and who made what decisions during development process (cf. [Vak+19]).

Transparency is the central element in ART framework as it is prerequisite and requirement to implement two other inter-related principles - *Accountability* and *Responsibility*. *Accountability* is externally motivated and defined as being liable for the decisions made by AI-based system or algorithm [Dig17], [Vak+19]. It is achieved, for instance, by explaining and justifying AI-based system decisions and actions to the relevant stakeholders. In contrast, *Responsibility* is internally motivated, and it is defined in actionable form as moral obligation to act responsibly [Vak+19]. In the settings of data mining, *Responsibility* refers to data scientists' conscious understanding of the impact of their model design choices and other decisions in the data mining process. This interpretation is in line with the perception of responsibility the developers have for their actions [VKA19]. Finally, we approach *Trustworthiness* as prescribed by EU Ethics Guidelines for Trustworthy AI [Hig20], which defines it as the ultimate goal of the system. These are subjective feelings of the user, and it is not possible to implement *Trustworthiness* directly into systems design [Vak+19]. However, the EU Guidelines [Hig20] specifies seven key requirements that AI systems should meet to be deemed trustworthy, including *Fairness*.

Furthermore, guided by [Dig18], from an AI ethics design perspective, we selected to tackle two categories of the AI ethics life-cycle, *Ethics by Design*, which refers to achieving ethical data mining outcomes, and *Ethics in Design*, which refers to enhancing the FIN-DM process with tasks supporting the implementation of ethics. The third category of *Ethics for Design* is out of scope since it is not an integral part of the data mining process, but rather a part of regulating data mining experts' behavior. We propose to govern that dimension in the form of a code of conducts, or similar principles within corporate or professional organizations (e.g., associations), to regulate ethical and compliance behaviors of employees or association members. Selected AI ethics categories and constructs are presented in *Concepts Universe* in section D of Figure 26.

In the *AI ethics sub-component* (section E of Figure 26) and the *Business Understanding phase*, we initially propose identifying the potential bias and risks of AI ethics issues for the data mining project by evaluating key areas where

[10]Guidelines on Ethically Aligned Design

they might appear. In particular, AI ethics might be transgressed when using biased or incorrect input data, making discriminating decisions, and delivering unsafe solutions with a negative impact on well-being (ignoring human-centricity). This evaluation task is supported by Checklist 2 (Appendix B.1, Figure B3), (cf. [Mad+20] for similar solutions), which specifies what key aspects to consider. For instance, for input data, these are quality and potential inherent bias (with the most common being selection and measurement biases). For the decision-making process, one of the key aspects to specify is potential discrimination scenarios, which can be tested when evaluating data mining results. As a next step, in the *Data Understanding and Preparation phases*, a number of tasks to check and ensure data quality (incl. possible bias) are proposed. In the *Modelling phase* we suggest tackling AI ethics either with Fairness Aware Data Mining[11] applying adjusted techniques and algorithms (cf. [Fri+19]), or with traditional techniques. In the case of the latter, rigorous fairness testing (incl. counterfactual fairness testing[12]) needs to be applied based on earlier defined scenarios. Then, concurrently identified bias and discrimination require mitigation by applying model de-biasing techniques (cf. [HBC16]). Further, in the *Modelling phase*, we propose explanation and interpretation tasks to achieve algorithm transparency (via ensuring explainability) and providing accuracy. As a next step, in the *Deployment and Implementation phase* human-centric design aspects are revisited, focusing on technical robustness, safety, and planning for the regular tuning and oversight of the data mining solution (as part of *Life-cycle management plan*). Finally, as part of the *Post-implementation life-cycle*, periodic quality reviews of datasets are proposed while data mining solution monitoring is part of the service cycle (cf. [Haa+21]). Traceability, reproducibility and auditability aspects are ensured throughout the FIN-DM life-cycle by documenting the AI compliance process and outcomes that, in turn, underpin data mining's *Trustworthiness*.

Tackling Model Risk While AI risks are heavily discussed and debated, *risk management* practices in the data mining process are under-investigated, practical guidance in the form of frameworks is scarce and fragmented, and at a higher abstraction level (tools and standards) are largely unavailable [Bra20]. The risk management of data mining models is currently considered from a model governance perspective (cf. [KSC20], [Haa+21]), a project management perspective (cf. [Bra20]), a specialized perspective, for instance, documentation [Ric+20], or an automation perspective of model life-cycle conceptualized as ModelOps[13] (cf. [Hum+19], [Arn+20]). Such conceptualization is related to the adaptation of the software development and deployment life-cycle to data mining projects; however,

[11] Developing data mining models and systems which are discrimination-conscious by design [HBC16]

[12] AI decision is counter-factually fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group [Kus+17].

[13] Model Operations

this conceptualization primarily focuses on enabling technologies and automation. In contrast, *quality assurance* in data mining, so far, has received more attention compared to risk management. In particular, guidelines for quality assurance in machine-learning-based applications have been proposed in [Ham+20b] and further extended in [Ham+20a]. Ishikawa et al. defined key quality evaluation aspects and proposed a development process model for ML systems. In a similar vein, [Stu+20] expanded CRISP-DM for the development of machine-learning applications, which covered the entire project life-cycle addressing two key deficiencies of CRISP-DM: (1) lack of application (in our case deployment) tasks where a machine-learning model is maintained as the application, and (2) lack of guidance on quality assurance [Stu+20].

Guided by software development practices in our design, we opted to combine *risk management* and *quality assurance* tasks as one component, thereby addressing both *Regulatory and Compliance* as well as portion of *Process gaps* related to quality assurance in CRISP-DM. Furthermore, by embedding and relating both elements, we expect to mitigate some model risk aspects with quality assurance tasks. Based on common model risk management frameworks (cf. [Bla+18]), data mining *model risk* management is structured around three interrelated sub-processes, such as *model development*, *model validation*, and *model use*. In these settings, we propose tasks to ensure *sound development methodology*, *intermediate testing*, *user acceptance tests*, and *proper model change policies and practices*. However, compared to original CRISP-DM, we replace the *testing* task with a much broader *verification and validation*, with a more extensive scope (e.g. assurance of requirements) and broader techniques. This is complemented by *auditability* requirements for data mining project artifacts and the data mining process itself, and *model monitoring* when a model is used. This conceptualization is presented in *Concepts Universe* (section F of Figure 26), while solutions are specified in the *Risk Management and QA sub-component* in Figure 26 (section F). There, we have made a distinction between two sub-processes - one workflow is inherent to the *model development* life-cycle, and constitutes tasks executed by project participants to ensure consistent quality in the development process. To improve development quality, the primary activity is testing based on respective plans on project quality targets and metrics. The other workflow addresses *model validation* and is executed by peer reviewers who are external towards a given data mining project– i.e., they are not part of a development team. They provide independent quality assurance via *Quality Checkpoints* and aim to validate all key data mining project artifacts and outputs, including the requirements, data used, methods and design decisions, the model or solution prototype, and the model's implementation in systems. In the case of the latter, software verification and validation methods should be used. Lastly, *model usage* is independently validated to mitigate the model risk and ensure the model's conformance to model change procedures. Regular data oversight and tuning is ensured to keep control of model risk (as part of the *AI ethics sub-component* routines), complemented by regular model

life-cycle management activities (Checklist 3 in Appendix B.1, Figure B4), and based on the plan defined in the *Deployment phase*. Deployed techniques include inspections, technical reviews, walkthroughs, conformance checks (adherence to processes), design and product release reviews, etc., while the output is checked for the '3Cs' (correctness, consistency, and completeness) [PR14].

6.3.5. Addressing Process Gaps

We addressed *Process gaps* in two contexts. The lack of *controls and quality assurance mechanisms* for stand-alone data mining projects have been tackled in the *Risk Management and Quality Assurance sub-component*. Here, we present the mechanisms of the solution to address prerequisites for repeatable data mining at scale.

Overview of ITIL and COBIT The most common approach to systematically address any business processes in the organization is to consider it via the prism of the value chain or value delivery (cf. [Dum+18]). This paradigm is industry-agnostic and equally applicable to the creation of tangible products (e.g., manufactured goods), intangible assets (e.g., software), and services. Therefore, this methodology has been successfully adopted across various industries, including the IT domain, and widely incorporated into IT practices. Notably, ITSM[14] is the most well-known and widely adopted approach to IT service delivery and management; it is process-centered (in contrast to technology-oriented) and focused on value delivery [Cus20]. One of the ITSM frameworks, ITIL[15], is the most widely adopted approach for the management of IT services. ITIL defines IT delivery as a value delivery process with three core components - process inputs and outputs, process controls, and process enablers. Notably, it is successfully adopted and fits modern software development practices, like DevOps[16] [GDQ20], and has incorporated Agile and Lean delivery methods. Drawing an analogy to IT and software delivery, we can view data mining projects as similar IT delivery instances, which can be encompassed and supported by ITIL.

In the same vein, one more notable ITSM framework is COBIT, which provides a comprehensive foundation for IT governance and management (cf. [HGD13]) in organizations. One of the key features of COBIT is its focus on governance and controls, including building comprehensive quality assurance governance and management.

Tackling process gaps with ITIL elements In the context of achieving effective data mining at scale, and tackling *Process Gaps* in CRISP-DM, we have selected the number of prerequisites and enablers *Foundation Capabilities* from the ITIL framework. Based on ITIL typology, the following attributes across 4 ITIL dimensions (left wing part 1 in Figure 24) were chosen:

[14] Information Technology Service Management
[15] Information Technology Infrastructure Library
[16] Development and Operations

- Organization and People - stakeholders management and data mining competencies (via knowledge management and codification systems)
- Information and Technology - data management, model development, deployment and self-service technologies encompassing the whole data mining life-cycle including sharing the results with users
- Partners and Suppliers - planning of resources and internal coordination
- Value Stream and Processes - delivery models (horizontal vs. virtual cross-functional teams) and the design of service, delivery, and improvement processes

The relevant dimensions and attributes were complemented by sets of selected ITIL Management practices defined as key enablers (upper section in Figure B1 in Appendix B.1). They cover general management (portfolio, project, relationships, etc.), service (primarily related to implemented data mining models service management), technical (deployment and software development practices), and delivery approaches (including delivery models, and continuous integration, delivery and deployment practices).

Tackling process gaps with COBIT elements These prerequisites are well-complemented by COBIT elements (right wing, part 3 in Figure 24), which refer to quality management at two levels: (1) data mining process' quality principles and policy (incl. quality management practices of data mining models), (2) the project's quality - quality management plans for respective data mining projects based on stakeholders' quality requirements. This approach with respective attributes as key enablers is detailed in Figure B1 in Appendix B.1 (lower section).

As shown in Figure 26 above, ITIL and COBIT elements at higher abstraction levels are cascaded down and natively integrated, where applicable, to the FIN-DM hierarchical process via respective phases, sub-components and tasks. For COBIT elements, these are data mining project quality management tasks reflected in the *Risk Management and Quality Assurance sub-component* of the *Compliance phase* (section F in Figure 26). For ITIL elements, these are data mining requirements reflected in the *Requirements phase* (section C in Figure 25).

6.3.6. Addressing the Gaps - Summary

We have summarized the solutions to gaps along with potential benefits in Table 7 below. We briefly discuss them here, with more extensive discussion provided in Chapter 8 *Conclusion*.

Table 7. FIN-DM proposed solutions to mitigate the gaps

Gap Class Name	Solution	Contribution/Benefits
G1- Requirements Management & Elicitation	- New **Requirements phase** with Requirements development, interdependencies mapping, sign-off tasks, iterative Requirements' management activities, - **Requirement Checklist** to assist in documenting, managing requirements throughout all data mining life-cycle	Ensures structured data mining life-cycle requirements elicitation, management end-to-end; improved data mining project outcomes
G2 - Inter - dependencies	Iterations, parallelization of phases, recommendations on the merged phases	Improves efficiency, effectiveness of data mining projects
G3 - Universality	- Explicit requirements' elicitation, development tasks for analytical by-products, associated deployments, - New tasks for applying non-modelling methods, their validations	Supports development, deployment of broad range of analytical insights, outcomes, improves actionability
G4- Regulatory & Compliance	- **Privacy sub-component** addressing all 9 core principles stipulated by GDPR, - **AI Ethics sub-component** addressing 5 most common ethical principles to identify and mitigate biases, ethical risks, - **Risk Management and Quality Assurance sub-component** to manage and mitigate model risks with verification, validation tasks, auditability requirements, model monitoring; model quality assurance tasks, - **Compliance and AI Ethics Checklist** to document, assess regulatory requirements and risks.	Supports regulatory compliance - ensures privacy compliant data mining life-cycle, AI ethical risks mitigation; ensures effective risk management (especially for financial services), effective quality assurance in data mining
G5-Validation	- New tasks for constructing validation scenarios in actual settings, - New tasks for execution of business validation pilots, feedback loop	Improves actionability, quality of data mining outcomes
G6- Actionability	- New tasks for Business usage requirements elicitation, - New **Implementation/Post-Deployment phase** with deployment tasks, data-mining solution life-cycle management plan, activities	Improves actionability, support conversion of data mining models into integrated software products, their life-cycle management
G7-Process	- Controls, quality assurance mechanisms for stand-alone projects, and at the level of governance frameworks and principles, - List of concrete prerequisites. enablers to support repeatable, reproducible data mining at scale (based on ITIL), - List of quality management elements/practices (based on COBIT).	Adequate technological support, governance of data mining projects

The solution in the form of *Requirement phase* supports a structured approach to requirements elicitation and management in data mining projects, in this manner improving data mining outcomes. Specialized *Requirements Checklist* is a practical tool assisting in documenting and management requirements. Significant degree of iterations and recommendations for phases mergers and parallelization supports better control and management of interdependencies in data mining life-cycle. This fosters improved efficiency and effectiveness of such projects. Specific tasks supporting non-modelling analytical outcomes, their development and usage improve actionability and use of broad range of analytical insights.

A number of new sub-components (*Privacy, AI Ethics, and Risk Management and Quality Assurance*) support regulatory compliance and addresses financial service's inherent challenges related to governance, risk management and ethics. In particular, *Privacy Sub-component tackles* all 9 core principles stipulated by GDPR[17]. The tasks for assessing legitimacy of data processing and adherence to Data Minimization principles are incorporated into data mining life-cycle. Also, activities for (pseudo)-anonymization, accuracy checks, and checks for legality of data mining results and their use are introduced. In this manner, risks for illegitimate data processing and use of data mining results without adequate legal grounds are mitigated and prevented.

In turn, *AI Ethics sub-component* tackles 5 most common ethical principles of *Accountability, Responsibility, Transparency (ART), Trustworthiness and Fairness*. To address them, tasks for potential bias and ethical issues identification, outlining potential discrimination scenarios, data quality checks, use of specialized techniques for modelling or fairness testing, and regular procedures for control over deployed data mining solutions are introduced. In this manner, potential biases and AI ethical risks (e.g. discrimination) are mitigated and prevented. *Compliance and AI Ethics Checklist* is a practical tool to document and evaluate AI ethics requirements in each stage of data mining projects.

As complementary to *AI ethics sub-component*, *Risk Management and QA sub-component* validation tasks are introduced as part of data mining model development and as independent peer-review. Validations of model usage support effective control and management of model risk. Further, specific tasks are designed to ensure sound development methodology, user acceptance testings, proper model change policies and practices, along with auditability requirements for model artifacts and model monitoring. *Model Life-cycle and Data Mining Checklists* provide practical support in documenting and assessing key quality assurance and risk management requirements. In this manner, sound risk management and quality assurance are incorporated as integral part of data mining life-cycle, and they are in line with modern internal control operating models of the financial services sector, such as *3LoD (three Lines of Defense)* model.

[17]Fairness, Lawfulness, Minimization, Integrity, Confidentiality, Accuracy, Data Storage Limitation and Security, Accountability

Interrelated tasks for business validations in real-life settings, business usage requirements elicitation along with significant support for data mining models deployment and life-cycle management improve actionability and quality of data mining project outcomes. Lastly, ITIL and COBIT elements in the form of enablers, prerequisites and recommended practices and governance mechanisms ensure adequate technological and governance support to execute data mining projects at scale.

6.4. FIN-DM Supplements

Most widely adopted data mining process models including CRISP-DM are supplemented with well-developed, comprehensive application guidance and glossaries. Likewise, we have developed *Application Guidance* (Appendix B.2) and *Glossary* as integral components of FIN-DM. They serve as design features to fulfill meta-requirement on users' support *MR3. Ensure minimized user's cognitive efforts*, and corresponding design principles of *DP5. Instruct how to use* and *DP6. Common terminology*.

To fulfill *DP4. Differentiate main users' categories*, FIN-DM is constructed and adapted for 3 distinct users' groups. Customization is achieved by users' tiering (DF4) and differentiating FIN-DM usage scenarios (DF5). The tiering has been derived based on common roles and responsibilities on the data mining projects. For this purpose, we have applied RACI framework[18] widely adopted and recommended for software development, IT delivery and service management in ITIL and COBIT guidelines. To this end, RACI principles have been recently adopted and tested for applicability in data mining projects (cf. [CA20]). Based on [Spa18], we broadly divided potential users into 3 groups depending on the overall role and primary activities on the data mining project (as presented in Table 8 below). We have differentiated 3 user categories: (1) User Group 1 - management stakeholders, (2) User Group 2 - domain business users and experts, and (3) Users Group 3 - technical delivery team (data mining experts, data analysts, data scientists, data engineers, etc.) and project manager(s). FIN-DM representation layers are matched to the respective Users Group as follows:

1. Layer 1 representation - intended for all User Groups, contains FIN-DM *Conceptual View*.
2. Layer 2 representation - intended for User Groups 2,3, contains FIN-DM *Hierarchical Process View* and *Key Enablers list*.
3. Layer 3 representation - primarily intended for User Group 3, contains all FIN-DM *Checklists (1-4)*.

It should be noted that user might migrate and/or be involved in more than one primary group and assume additional activities beyond primary. For instance, business users (User Group 2) might get involved into testing activities and be

[18]RACI - Responsible, Accountable, Consulted, Informed [Spa18]

closely related to technical development activities at certain times in the project, thus, being involved into User Group 3 too. However, they are likely not to be required to use proposed reference process most specific Layer 3 checklists in these activities. Therefore, Layer 2 representation will remain most suitable and adequate for User Group 2 given their primary role.

Table 8. FIN-DM User Groups categorization by roles, activities and RACI [Spa18] mapping

Dimensions	User Group 1	User Group 2	User Group 3	User Group 3
Role	(Top) Management stakeholders	Project members with business domain knowledge	Project members with project management role	Project members with development role
Profile	Functional Managers (business and tech domain)	Business users, business domain experts	Project manager(s)	Technical project delivery team - data mining experts, data /business analysts, data scientists, data engineers, software developers,etc.
Primary Activity	Overall oversight	Business input (domain knowledge, validation, etc.)	Project management	Technical Development and Deployment end-to-end
RACI mapping	I, C, R	C, R	R, A	R, A

Application Guidance specifies guidelines for each User Group on how the respective process model layer can support it. For example, User Group 2 by using *Hierarchical Process Model* representation will understand which are key phases of data mining project and what is the sequence of activities. This view also assists project managers to explain business and domain experts key data mining project activities where their engagement is required.

Each representation layer is self-sufficient when used in conjunction with *Glossary*. In turn, *Glossary* provides basic terminology of definitions and items used in FIN-DM and in *Application Guidance*.

Apart from User's group differentiation and targeted usage explanations, *Application Guidance* presents common background information including: (1) concise CRISP-DM overview, (2) purpose of FIN-DM and how it relates to CRISP-DM, 3) FIN-DM structure, and (4) explanation of key modifications proposed in FIN-DM along with their goals.

7. FIN-DM EX-ANTE EVALUATION

This Chapter tackles Research Objective 4 - *to evaluate the artifact with potential users to confirm its practical utility, relevance and users' acceptance*. It describes the FIN-DM (Financial Industry Process for Data Mining) *Evaluation cycle* (Cycle 3 in Figure 3 in Chapter 1). Initially, the evaluation design and its execution are described. Then, potential users' feedback and findings are presented and discussed. We highlight how the suggested improvements were iterated back to final FIN-DM design. Subsequently, threats to validity are outlined.

7.1. FIN-DM Evaluation Design

Here, we present and motivate the selection and design of FIN-DM evaluation criteria, the method, the organizational settings, and the instruments.

An artifact can be evaluated in terms of validity, utility, quality, and efficacy as key criteria [GH13]. To address our research goal of developing an artifact which would be useful and applicable in practical settings and to identifying means for its improvement, we focused on examining it from a user perspective. To achieve this we first broke the aforementioned criteria down to their lower level constructs, for example, related to *Perceived quality* we considered FIN-DM's *complexity, completeness, the existence of overlapping elements and presentation quality*. In addition, we also assessed the potential users' acceptance of FIN-DM utilizing the Technology Acceptance Model (TAM) (cf.[LKL03]). We thus consequently focused on *Perceived Usefulness (PU)*, *Perceived Ease of Use (PEOU)*, and *Behavioral Intention*. Additionally, guided by [Bha01], we included *Satisfaction* as a factor, which has been widely used and confirmed as the determinant for continued IS use (so-called Post-Acceptance Model).

Based on the Framework for Evaluation in Design Science (FEDS) (cf. [VPB16]), we have selected and executed *Ex-Ante, Naturalistic evaluation* [CSG15]. This approach is aligned towards our research objective - developing a user-oriented solution with a practical utility/benefit and identifying means for its improvement. Further, when executing the evaluation, we applied two methods: (1) a combined individual demonstration session followed by *semi-structured interviews*, and (2) a *questionnaire*.

We have opted to conduct the evaluation in one organization to eliminate the impact of differing cultures, business processes, and similarities, which might interfere with the study results. The selected financial institution is the same company where case study (presented in *Chapter 4*) was conducted. To balance the concentration in one organizational setting, and given that the framework is intended for use by different experts and cross-functional teams, we aimed to diversify the interviewees' cohort within this setting (interviewees list detailed in Table 9 below). In this manner, we also mitigated the potential bias as some of the interviewed experts (primarily, data scientists) have also participated in the initial case study.

We subsequently constructed an interview guide and questionnaire instruments[1]. Interview questions were formulated based on lower-level evaluation constructs, for example, *Quality perception* questions considered FIN-DM functional quality, completeness, accuracy, relevancy, clarity, etc. *Acceptance perception* discussion was structured to investigate users views on FIN-DM ease of use and usage intentions. Interviews were transcribed verbatim (121 pages of transcripts in total), and were evaluated jointly with the questionnaire results.

Then, interviews were coded iteratively, guided by [Sal15] (the initial and final coding schemes are available at link[2]). The initial coding scheme was derived from research objectives and corresponding evaluation criteria and constructs. For example, *Quality perception* initial codes contained FIN-DM completeness, existence of overlaps and complexity. It was further refined during analysis when additional codes were discovered, for instance, in case of *Quality perception* the *Agility* code emerged. At this stage, the coding was validated by collaboratively coding and evaluating on the sample of interviews. The responses of the questionnaire were aggregated and calculated per construct. Due to the low number of questionnaires (8 in total), we did not analyze them in a quantitative way. We rather used them to provide additional context to interviews.

7.2. FIN-DM Ex-Ante Evaluation Results

In this section, we will report on the results of our evaluation of FIN-DM. We will particularly focus on the participants' perception of FIN-DM's quality (completeness, existence of overlaps, complexity and presentation quality) as well as their perception of its usefulness, ease of use, future use intentions and satisfaction. In addition, we will also discuss improvement suggestions by participants and elaborate on how we integrated them. The profiles of the participants are summarized in the Table 9 below.

[1] Available at `https://figshare.com/s/1bda7ccadaa254fcabe1`

[2] `https://figshare.com/s/d2bf5084f3e6cfc1af80`

Table 9. Experts-participants in evaluation - profiles and key characteristics

No.	Interviewees profile	Primary area of expertise (* advanced academic degree in the field)	Current role	Experience in the field (years)
1	Data Scientist 1 (DS1)	Data science, advanced analytics, quant finance	Senior expert in data science	8+
2	Data Scientist 2 (DS2)	Data science, advanced analytics	Lead expert in data science	8+
3	Data Scientist 3 (DS3)	Data science, advanced analytics*	Senior expert in data science	10+
4	Data Scientist 4 (DS4)	Data science, advanced analytics, quant finance*	Senior expert in data science	10+
5	Data Scientist 5 (DS5)	Data science, advanced analytics, quant finance	Senior expert in data science	10+
6	Project Manager 1 (PM1)	Project management, data science, advanced analytics	Lead project manager in data science	8+
7	Project Manager 2 (PM2)	Project management, data science, advanced analytics	Lead project manager in data science	8+
8	Expert 1 (Exp1)	IT, Software development	Agile Coach, Operational leader in IT	8+
9	Expert 2 (Exp2)	IT, Software development, Risks in IT	Operational leader in IT	20+
10	Expert 3 (Exp3)	Project management, data science, advanced analytics, tech	Analytics consultancy operational leader	10+
11	Expert 4 (Exp4)	Data science, advanced analytics, technology, strategy	Strategic leader in Data Science, advanced analytics	20+

7.2.1. Ex-Ante Quality Evaluation - Design Aspects

In the following we will elaborate on our findings related to the quality perception of FIN-DM starting with its completeness, existence of overlaps, complexity and presentation quality before discussing specifics about how it can be implemented in practice and connected to existing frameworks such as ITIL.

Completeness - 'Nothing is missing for data mining...organization and technology needs attention' The participants acknowledged the general completeness of FIN-DM in covering data mining aspects, especially the ones missing in the CRISP-DM model, as noted by one interviewee, *"It seems that most [...] of the things that were missing in CRISP-DM have been added here"* (DS2). In particular, the solutions addressing privacy, AI ethics 'gaps', and enhancements to deployment phase were very well-received, as one interviewee noted *" I think it was very good that you captured the ethics and [...] the GDPR [...] or a AI and general perspective"* (DS5). At the same time, some organizational and technology aspects were acknowledged as missing, as one interviewee reflected *"As I mentioned, maybe these change management and real business implementations is missing, otherwise, [it] looks quite OK, at least all the bullet points and topics that need to be cared of [...]"* (Exp3).

Overlaps - 'Iterations and Dependencies' Interviewees did not detect significant overlaps in the FIN-DM life-cycle; however, some interviewees classified such intersections as dependencies or necessary iterations intrinsic to data mining development as such. Some felt that *"no, I, couldn't say now if anything overlaps, I mean, to me this is a very iterative process, did you jump back and forth all the time during a project"* (DS5), while others considered, *"maybe there are some overlaps in the quality management and evaluation or validation part, and a little bit with a new deployment and monitoring parts. But it's like maybe not so much overlap, but more like dependencies"* (DS2).

FIN-DM complexity - 'Generally adequate...needs to be simplified a bit' The majority of the interviewees perceived FIN-DM as balanced with respect to its complexity and the details it includes, as one interviewee put it *"I think it's quite balanced because there doesn't seem to be anything that is not needed [and] I immediately don't even see that many options, how it could be simplified that much further without losing important information."* (DS2). However, the comment below suggests the need for some simplification to increase usability and foster easier adoption, *"I think that there are a lot of details that have been put into this new framework, and they are all very much needed [...] but, it is also if it should be really helpful and usable for the different organizations and easy to implement, then it needs to be a bit simplified maybe. But then, if you simplify it too much then I'm scared that we will lose out of a lot of very, very important aspects. So, yeah, somewhere in the middle"* (PM2).

FIN-DM presentation quality - 'Complicated but could be used with adequate support' Interviewees provided valuable remarks on the presentation quality

of the initial FIN-DM prototype, focusing on four key presentation aspects:

- terminology and definitions of elements - more clarity required (e.g. differentiation between functional and non-functional requirements, robustness of algorithms were vaguely defined) - *those non-functional requirements that we talked about again availability, robustness I don't know, also something redundancy of the data and so on, but these are for the IT system and the software system mainly.* (DS3)
- application guidance - too short and requires expansion.
- communication design - difficult to grasp when looking - *"....from framework perspective, You should think on the way how to simplify the common content like presentation, means of presenting the content, showing the content, because otherwise it deserves too much time for explanation. And a lot of efforts of understanding."* (Exp 3)
- conceptual design - interrelations between elements (e.g. ITIL elements not natively integrated) - *Just the question is how to easily incorporate it in the overall framework so that it's clear that You need to pay attention to this part. I would say that maybe at this point it's not natively integrated to each other.* (Exp3)

As commented by one interviewee, *"I would say it's maybe from presentation perspective over-complicated, from content perspective not [it is] adequate, that's for sure, everything is needed and it should be checklisted for [the] project manager. But from framework perspective, you should think [about] how to simplify the common content like presentation, means of presenting the content, showing the content, because otherwise it deserves too much time for explanation. [...] And a lot of effort [for] understanding."* (Exp3).

Reference Process Instantiation[3] **- 'Requires adequate technology and documentation support'** Interviewees also mentioned how the instantiation of FIN-DM would benefit from the support of adequate tools, technologies, and process documentation. In reference to this issue, an interviewee said, *"Well, those I don't remember exactly about those architectural and business value verification documents, or checklists, or agreements, but I trust this can be done even in JIRA somehow or something like this by checkpoints. So, I think it can be realized by technology. So, I think this process has lots of underwater artifacts that are not mentioned here. Maybe you can even later visualize it somehow by BPM and diagram and adjust show what may happen and what may be the output"* (Exp1).

FIN-DM integration into existing frameworks Interviewees also proposed improving and distinctly positioning FIN-DM by specifying the purpose, and the issues it intends to solve, more clearly. In the same vein, interviewees also suggested differentiating FIN-DM to other software development frameworks, as noted *"The difference between this framework and software frameworks could be a*

[3] General approach for mapping the generic process model involves specializing (or instantiating) generic contents according to concrete characteristics of the data mining project context [Cha+00]

bit more emphasized explicit, and also I think explainability, slash interpretability could be also a more explicit and more bold, more visible in the framework." (DS3).

7.2.2. Ex-Ante Acceptance Evaluation

Related to the acceptance of FIN-DM, we will subsequently report on the participants' perception of its ease of use and usefulness, as well as their intentions to continue using it and their overall satisfaction.

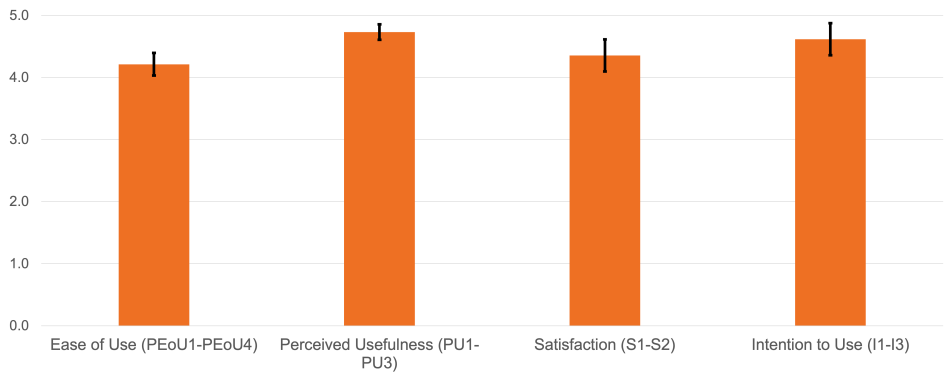


Figure 27. Results of the questionnaire. All responses were given on a 5-point scale which were anchored between *strongly disagree (1)* and *strongly agree (5)*. The bars indicate the mean (m) and standard deviation (SD) for each scale.

FIN-DM ease of use The questionnaire results indicate that the participants generally perceived FIN-DM to be easy to use, as evident by a mean value above 4 (as presented in Figure 27 above). This score however ranks the lowest of all the questionnaire scores. Taking this together with the previously discussed findings related to the participants' perception of the presentation quality of FIN-DM however suggests that the understandability can still be improved, e.g. by reducing complexity or providing adequate guidance and support.

FIN-DM usefulness Our participants generally perceived FIN-DM as useful, as evident by it receiving the highest score and lowest standard deviation among the aspects in the questionnaire (Figure 27). This finding is underpinned by interviewees stating that, *"[...] of course there is no doubt that [the new elements] will be useful. They will help to achieve more transparency, clearance and maybe more mistake-proof environment. And of course, understanding between counterparties, and no missing processes or no gaps and disagreements as well."* (Exp1). This perception is also supported by other interviewee, *"Absolutely, undoubtedly, the artifacts that you have brought they add value to the whole process, and make it more manageable. For the further industrialization and scaling, it's absolutely must"* (Exp2).

In the following, we will go into more details regarding the interviewees per-

ception of the usefulness of FIN-DM to address existing gaps before discussing its different components both from a micro (individual FIN-DM component or data mining project) and macro (life-cycle or repeatable data mining settings) perspective.

FIN-DM addresses existing gaps Participants also extensively discussed how FIN-DM solves CRISP-DM 'gaps'. The central issue for data mining projects in data scientists' view was business value realization. They, therefore, perceive the positive impact and potential of FIN-DM to lie in increasing the business value realization of data mining projects. As noted by one interviewee, *"as I understand this CRISP-DM, this [FIN-DM] is more integral. Basically, it's like trying to concentrate all of the required things from the beginning. And, I think that the main problem now is that we built an impressive analytic solution and that sometimes does not add business value. So, defining like this definitely helps to bring this business value on the table."* (DS4). Furthermore, FIN-DM was characterized as an up-to-date, modern, and reusable data mining process that solves typical data mining process problems by design.

FIN-DM Components View - 'contribution and positive impact per component' In terms of individual FIN-DM components, interviewees also emphasized the importance and necessity of the proposed *Requirements* elements in conjunction with *Business validation* later in the data mining life-cycle. *"I would like to emphasize that the relationship to business, like a business understanding, business validation on the outputs, et cetera, that was really very valuable to see."* (DS5). The participants perceived the *Requirements* as an integral part of the data mining project to timely discover, capture, and specify key data mining project prerequisites, such as data, tooling, deployment patterns, and infrastructure, as commented: *"I think it's something that should be investigated [...] at the very beginning any data mining project [...] and there have been many examples in the past for us where we start with a project [...] and we discover half-way that we lack something"* (DS5). Another interviewee concurred: *"I liked that idea that immediately in the beginning requirements not only towards business but deployment, data [...] identified since the start [...] that's valuable, that is frequently missing in business, that business is starting to [deploy] requirements quite lately, and they witness themselves a lot of problems because in many cases their infrastructure is not ready to deployment so that is definitely valuable."* (Exp3).

As an aspect closely related to the *Requirements*, participants reflected on the *Deployment and Post-deployment* additions introduced in FIN-DM. In particular, the model governance and maintenance aspects and follow-up activities of FIN-DM were overlooked in other standard processes as the following comment shows *"So, I would say that this is extremely valuable part, especially for companies that are starting their analytical journey. Because when they start, they mostly focus on modelling, on results, and just delivering the model that works and predict something that they expect. But they don't think about the long-term maintenance of [...] their solution. So, I would say that if in the beginning they would start*

thinking about the ownership of the model, like governance, about the principles, how their maintenance assured, model is giving good results, it would result in better ROI for their models. Because the frequent problem that analytics is done, but it's not fully implemented in the business process. And even if it's implemented, in many cases, nobody follows-up on the model performance. So, that's very important part which needs to be discussed for each project somewhere in the beginning even, like owner of the model." (Exp3). In the same way, the importance of the *Post-deployment phase*, especially to ensure the data mining solution has been properly used, has been emphasized by other participants: *"And yeah, I think that this is a very nice way of thinking about data mining projects.[...] I especially like the last phase, [...] post-deployment, because without that, what's the point of all the products. So, it's like I'm creating something that will never be properly used [...]"* (DS3).

Another critical element discussed extensively by interviewees has been the *AI Ethics and Compliance phase*. Interviewees especially emphasized the AI ethics context, which is very new in data mining. Also, lack of guidelines and practices in traditional process models. As noted by one interviewee, *"I think this is quite necessary to get the CRISP-[DM] or the methodology up to date. Since both the GDPR and ethics are very important at the times today. I absolutely think that we should incorporate things like this...Absolutely."* (DS5).

The critical importance of AI ethics for the financial services industry, in conjunction with GDPR compliance and personal data handling, has been highlighted as well, *"...we should [...] have this in mind throughout like all of our cases."* (DS5). Furthermore, the imminent need for larger organizations and industries with a lot of data to control and tackle specific AI ethics risks, e.g. biases, were emphasized. As one interviewee put it, *"I think this is super well-appreciated, welcome, welcome component here, because I, as I mentioned earlier, I don't think this part has ever been covered, or at least not extensively in these traditional frameworks. [...] it makes the approach more solid both with regards to ethics as such, but also with regards to, as you highlighted here compliance. Because those are becoming imminently more important in our work, especially if you consider [this] from a banking perspective - banking is about trust and these two components directly address underpinning a bank's trust with data and customers."* (Exp4).

Lastly, participants recognized and appreciated *Quality management* as a required activity. Concurrently, the critical role of quality management in two different contexts was emphasized, (1) as a stand-alone data mining project and (2) as an integral part of overall organizational practices. For example, *"Yeah, and that is absolutely needed. Yes, I mean, just both on the high level, as you mentioned, the more quality management plan, more if you have these more regulatory models that need to be in place, we have to ensure certain qualities and there are different departments involved. But also once the model is built, that we have these peer review processes, making sure that we are always having four eyes on each step."* (PM2).

Macro Settings - Industrialized Data Mining at Scale The usefulness of FIN-DM in broader settings for scaling and industrializing data mining as an internal service has been proposed. In particular, FIN-DM would help establish repeatable data mining and thus resemble IT established processes of IT services' portfolio management. On this point, the interviewee commented, *"Because I see that [the] data mining role is becoming more and more essential in each organization, and it's not anymore hidden somewhere in the corner like it was before, just some kind of R&D.[...] And of course the services, because your data mining, it's closely related with using the technology, and maybe later it will become some kind of a service, which will also require some SLA. And SLAs mean that, of course it has, has to be managed by ITIL, [...] And of course, technical things, I think they also are based on good practices and they are relevant. So, basically it states everything that is needed, and it will be a guide for people who are unsure what they are doing or want to establish. So, this would serve as a framework even for those who would like to establish something like this. It's my opinion, because you have all the components, it's like the establishment guide."* (Exp1). The other participant resonated, *"absolutely, undoubtedly, the artifacts that you have brought adds value to the whole process, and makes it more manageable. For the further industrialization and scaling, it's absolutely must."* (Exp2).

Micro Settings - Stand-alone Project The project participants extensively discussed FIN-DM's usefulness in terms of goals and benefits achieved in stand-alone data mining projects. They particularly mentioned the integrated, holistic approach for data mining projects from an organizational perspective. In particular, crucial alignment between different functions, stakeholders, and required capabilities to support end-to-end project execution was emphasized and perceived well. For example, one interviewee said *"[...] for me, this is just my point of view, CRISP-DM lacks the bridge of three things - business, IT, and data science, because we are doing the models, but if the model gets too complex, then there is nobody else that will ensure that it would be operationalized. And this is where IT should jump in, but [...] there's no link [...] or somehow it gets lost [...]. And on the other hand, sometimes you develop some models, whatever, but the link with business is lost as well. [...] there is a lack of clarity in the business goal, or that either we do not understand the business objective, or that we want to minimize overhead and leap into the interesting bit of the project, analyzing the data and so on, and to [open] these results in very interesting models that don't meet a real business need. So, a methodology that connects these three parts, [...] the real end-to-end process that has to touch all these three parts....."* (DS4).

FIN-DM intention to use and satisfaction Compared to the *Perceived Usefulness* results, the interviews and questionnaire exhibited a higher variability in answers on other *Acceptance Perception* aspects, such as the *Intention to use* and *Satisfaction* of FIN-DM. Evaluation participants expressed firm *Usage intentions* both in interviews and in the questionnaire (Figure 27). In the latter case, the *Usage intention* is preceded by the overall *Satisfaction* of potential users with FIN-DM -

it is on average in the *Very satisfied* category. Lower *Satisfaction* relative to the *Intention to use* of FIN-DM can be explained by, and was captured in, the *Ease of use* measure previously discussed in conjunction with *Quality perception*.

Related to future usage intentions the participants included the overall logic and concept of the FIN-DM, as noted *"I think, anyways, later it will be visualized more and more. I take this as a draft material, so I have no objection, objections towards this format and looking absolutely logically..."* (Exp1). *"[...] it is so detailed but it is good. [...] but I understand the main picture, the main big picture, which is this - to connect these three main parts, and to make it like full end to end process. This part I understand well, maybe just the particularities, the specific artifacts, and so on.... and I want to emphasize that, that is very good."* (DS4).

Also, interviewees further contextualized potential usage scenarios and proposed the usage of FIN-DM by various data mining projects participants and stakeholders, as noted: *"I think it's also very important that this framework is not only used by data scientists, but by the whole cross-functional team, including product owners and, and so on. And, and even maybe, perhaps, stakeholders and so on. So, I think it's like very important that everyone understands it."* (DS2). Lastly, participants also expressed conditional usage– provided concrete improvements suggestions are fulfilled– for example, *"If it were simplified, I will definitely take it to my next project."* (Exp3).

7.3. Suggested Improvements to FIN-DM

Here, we present and discuss key improvement proposals received in FIN-DM's evaluation (*Outcome 6* in Figure 3 in Chapter 1). We group them by design aspects, missing elements, components, and prerequisites. A complete list of all suggested improvements as formulated by interviewees and how they were addressed in the final FIN-DM version is available at link[4]. Overall, we have received 55 improvement proposals, with 51 unique suggestions. We classified them into four distinct categories - *Minor, Medium, Major and Critical* - based on the suggested improvements' complexity and impact. Proposals in the *Minor* category concerned improvements of some existing elements concerning their explicit differentiation, presentation, or conceptual design. *Medium* improvements concerned the use of terminology and the necessity of defining elements clearer. In the case of the *Major* improvement proposals, a significant rework of some content elements was required. In contrast, the *Critical* category included suggestions for adding and specifying new elements and the process model's overall conceptual design. Close to 80% (40 suggestions out of 51) of all the proposals have been categorized as *Minor and Medium*. The remaining 11 items were *Major and Critical*. There have been no conflicting suggestions except one requiring substantial simplification, which originated from the strategic leader. We have resolved this conflicting

[4] `https://figshare.com/s/37eaab47cb024e3a2023`

suggestion by introducing a new design principle (DP4), and integrating clear-cut process model adaptations to differentiate users' categories based on data mining project roles (DF4). Below, we proceed with the analysis of the *Major and Critical* suggestions, their potential adverse impacts, and how we addressed or mitigated them for the final version of FIN-DM.

Design Aspects Some interviewees reflected on FIN-DM's complexity, which in turn could hamper its adoption. Talking about these issues, an interviewee noted, *"CRISP-DM [is] extremely simple, so here you have a combination, so you really need to think how to put it together to make it simple, because otherwise all the content is correct, but it's complex. And if it's complex, it means people will not understand, will not adopt, and will not stick to it, so that's the risk."* (Exp3).

Furthermore, interviewees underscored two key trade-offs and balances to achieve in FIN-DM. One was associated with equalizing data mining-related elements and other elements from the IT domain, which some interviewees perceived to be over-emphasized: *"It was a lot on the requirements in the beginning, and it was a lot on the deployment phase, but those steps in between, I think those are the main tasks of the data scientist, because of course the data science project needs all kinds of skills"* (DS3).

The other trade-off we found was between simplicity and details. As pinpointed by one interviewee, *"[...] then it needs to be a bit simplified maybe. But with, then, if you simplify it too much, then I'm scared that we will lose out of a lot of very, very important aspects. So, yeah, somewhere in the middle [...]"* (PM2).

Another aspect proposed for improvement relates to FIN-DM's flexibility and its ability to accommodate external changes, i.e., to stay dynamic. To this end, changes in IT domain-related elements have been emphasized as the ones requiring an update when the underlying ITIL and COBIT frameworks change. As one interviewee put it, *"couple of years back, we didn't even consider to include general IT processes into this, and, and nor did we, You know, we were not talking about CI/CD and software development practices. So, so there is a bit of a so to say in, there is a bit of so to say trends that come and influence the formations of these frameworks. The question is now, that process to also re-engineer or revitalize or update the framework itself, for a use case, for models portfolio, how does that process looks like? And how do You make sure that, because now You refer to, You know, different frameworks and practices and so, how do You make sure that You always stay on top and make sure You bring in what is most relevant into Your processes and frameworks and, and omit the ones that are not. It's more of a philosophical question, it's more of a meta question, then rather than the more of a specific use case question."* (Exp4).

Interviewees also recognized the need for change management approaches towards FIN-DM, e.g. based on external requirements. The interviewees perceived the emergence of external disruptive technologies and business models as a core source for such changes. As one interviewee commented, *"[...] one is that how do we ensure [...] in the processes and stages stated here capture for example, all*

the new technology enablement, that brings times, reduce times to market, and it includes efficiencies. So, how do you include that, those components into this? How do you include the new frameworks around Open Banking, APIs that, that so to say, and cloud as such, the technology enablement that opens up new ways of consuming information. And with that comes also certain level of controls and qualities and process checks that need to be introduced, which traditional approaches maybe perhaps or not really taking into account." (Exp4).

Also, the required adaptation of FIN-DM towards new business models can be related to the emergence of new business ecosystems[5], and associated collaboration between participants. This is especially relevant for the financial services sector with the active emergence of open banking ecosystems fostered by respective regulations, most notably the PSD2[6]. Talking about the issues, as one interviewee noted, *"The other thing is that in, I guess this is more of a business understanding stages [...] and data readiness. Maybe this is how to say, this is to what extent your frameworks take into account both the internal and external perspective. And with external, probably more external perspective, I mean, if you are operating in an open banking environment, for example in our case, and you will need to adhere to bunch of other companies connecting to your data and exchanging data with other partners and so on. How, what sort of requirements that external environment will have on your use cases? And, and as part of your business understanding, you need to have steps, activities that capture that and make sure that further requirements internalized. This is, I feel, probably all of these frameworks can improve upon, or at lease clarify the standpoints on."* (Exp4).

Missing Elements and Pre-requisites Interviewees also identified missing elements in FIN-DM. In particular, a necessity to transform business processes (via adequate change management) to embed and integrate data mining solutions has been emphasized, *"And actually one important things what is maybe a bit missing, [...], it's regarding deployment. What we observe quite frequently, this is change management because in many cases, business might have put understanding why they need like product owner has a good vision. The model is designed like according to the best practice. (The) model is good enough, it's, it deployed, it works even maybe on a real-time, but all organization is not adapted. Then, they're just not using in their daily processes, they follow old, old methods, old reports, old practices, and et cetera. So, I would say (a) crucial part after deployment, or in parallel with deployment should be business change management. [...] to make it happen You need really to make a change inside the proposed system, not just make mathematics."* (Exp3).

[5]Defined as the network of organizations— including suppliers, distributors, customers, competitors, government agencies, etc.— involved in the delivery of a specific product or service through both competition and cooperation (cf. [Pal+20], [Adn17])

[6]Payment Services Directive 2015/2366, promotes a harmonized payments services market open to operate for payment providers beyond traditional incumbent banks, requiring close cooperation between payment services providers in the execution of payments and data exchange (cf. [CRN16])

The other aspect is associated with ITIL and COBIT's more native integration with the data mining life-cycle, and embedding concrete ITIL and COBIT practices into the FIN-DM process. This aspect is complementary to balancing and specifying data mining and IT perspectives discussed above. As noted by one interviewee, *"it would be nice to direct like these both COBIT and ITIL frameworks towards data mining, because this, for me [...] are [...] too general, which applies to every project of course, but I can't right away see what the specific for data mining project in when we are making use of this COBIT and ITIL. So, it would be interesting to see what are the best practices in ITIL or COBIT that we can specifically use for data mining projects."* (DS3).

Some interviewees expressed the necessity for detailing roles and responsibilities and governance aspects in the proposed reference process. As one interviewee commented, *"And maybe responsibility [...] who is responsible for doing what."* (Exp1), while other noted *"with regards to more governance of the different stages that you might have some more questions. In regards to the roles and responsibilities needed in various stages to execute these tasks."* (Exp1).

Further, interviewees emphasized risks and aspects which would hamper the straightforward application and adoption of FIN-DM. In particular, the risk of moving far away from the lean process and being too restrictive was pinpointed, *"Well, I think t this is a very important step, but at the same time [...] the threat is that we move far away from a lean process, like it's if we just add lots of checklists and documents, and lots of meetings, and lots of discussions about around this. If it's too much, and it can easily be, get too much, because you want to make sure you cover all the requirements for [...] functional requirements, what they expect, the nonfunctional requirements, the deployment plan requirements, the maintenance requirements, and this can get a lot."* (DS3).

Also, it was emphasized that it might be easier to apply FIN-DM in organizations with established IT foundations and practices, as one interviewee considered, *"Yeah, I would not say that it's too much details, on the contrary, under each bullet point that you have provided there is like much deeper area of activities and knowledge. And then basically if you imagine starting from the scratch, I would say that it would be massive work even to you know, get ready the framework as such and the practices, et cetera. If we already apply this thing on the existing for example, organization, where (the) IT foundation is quite established, then it would fly much easier."* (Exp2).

7.4. Threats to Validity

This research part also has a number of inherent limitations, primarily associated with conducted evaluation and techniques. In particular, the evaluation has been performed in one financial institution with a small set of selected participants, which limits the ability for the results to be generalized beyond the context of the study (external validity risk as per [Run+12]). We mitigated this concern by

selecting interviewees with different backgrounds and roles in data mining projects. We also invited external experts from outside the organization. However, more studies are recommended in other financial companies. Besides, the evaluation was executed in the same institution where the requirements were derived, which also represents a threat to external validity. It was mitigated by diversifying the interviewees' cohort with inclusion of experts external to the organization or function where the study was hosted. In this manner, we also mitigated the potential bias as some interviewed experts (primarily, data scientists) have also participated in the initial case study presented in *Chapter 4*. Next, FIN-DM has undergone an initial evaluation, but it requires a complete evaluation by applying FIN-DM in actual data mining projects.

Apart from external validity, our study has construct validity risks and reliability risks. Construct validity refers to the extent to which what is studied corresponds to what is intended and defined to be studied. In our research, the interview method can be a source of construct validity risk. We mitigated this concern by verifying the interviewees' understandings of the questions and reconfirming the responses we received. Finally, reliability risks concern the researcher's biases in generating and interpreting study results. We have tackled this risk by logging coding procedures and interviews and by adopting an iterative research process with regular validations within our research group. We also ensured the replicability of the research steps and results by keeping evidence-tracking the research materials and process.

8. CONCLUSION

8.1. Summary

This PhD research addressed the lack of the data mining process model adapted for the specific needs of the financial services sector. The work has investigated three research questions.

Initially, the PhD thesis tackled *how are data mining methodologies applied by researchers and practitioners (RQ1)* in cross-domain and in specialized settings. In two studies, we have examined the use of data mining methodologies by means of a systematic literature review covering both peer-reviewed and 'grey' literature. In the case of cross-domain study, it was found that the use of data mining methodologies, as reported in the literature, has grown substantially since 2007 (four-fold increase relative to the previous decade). Also, it was observed that data mining methodologies were predominantly applied 'as-is' from 1997 to 2007. This trend was reversed from 2008 onward, when the use of adapted data mining methodologies gradually started to replace 'as-is' usage.

The most frequent adaptations have been in the 'Extension' category. This category refers to adaptations that imply significant changes to key phases of the reference methodology (chiefly CRISP-DM). These adaptations particularly target the business understanding, deployment, and implementation phases of CRISP-DM (or other methodologies). Moreover, it was discovered that the most frequent purposes of adaptions are: (1) adaptations to handle Big Data technologies, tools, and environments (technological adaptations); and (2) adaptations for context-awareness and for integrating data mining solutions into business processes and IT systems (organizational adaptations). A key finding is that standard data mining methodologies do not pay sufficient attention to deployment aspects required to scale and transform data mining models into software products integrated into large IT/IS systems and business processes.

Apart from the adaptations in the 'Extension' category, we have also identified an increasing number of studies focusing on the 'Integration' of data mining methodologies with other domain-specific and organizational methodologies, frameworks, and concepts. These adaptions are aimed at embedding the data mining methodology into broader organizational aspects.

Overall, the findings of the study highlighted the need to develop refinements of existing data mining methodologies that would allow them to seamlessly interact with IT development platforms, processes (technological adaptation), and with organizational management frameworks (organizational adaptation). In other words, there is a need to frame existing data mining methodologies as being part of a broader ecosystem of methodologies, as opposed to the traditional view where data mining methodologies are defined in isolation from broader IT systems engineering and organizational management methodologies.

Similar results were obtained when examining data mining methodologies usage

in the banking domain. By means of a Systematic Literature Review, 102 relevant studies of peer-reviewed and 'grey' literature were identified. We discovered that data mining methodologies are applied regularly since 2007 and their usage has tripled. Furthermore, data mining in the financial services domain is primarily used for two main purposes - to address Customer Relationship Management and Risk Management related business problems.

It was also discovered that over the last decade, data mining methodologies have been primarily applied 'as-is' without modifications. Yet, the emerging and persistent trend of using data mining methodologies in banking with adaptations was also noted. Further, we have distinguished three adaptation scenarios ranging from granular modifications on tasks, sub-tasks, and deliverables levels and ending up with merging standard data mining methodologies with other frameworks.

We have also examined the adaptation objectives, banking domain specific factors behind such adaptations, and, as a result, we have identified three such aspects. Firstly, discriminatory awareness and transparent decision-making (human-centric aspect) require the adaptation of data mining processes. Secondly, the actionability of data mining results (business-centric aspect) plays a central role in the banking domain. Thirdly, we have also identified that standard data mining methodologies lack deployment and implementation aspects (technology-centric aspects) required to scale and transform data mining models into software products and components integrated into Big Data Architectures. Therefore, adaptations are used to integrate data mining models and solutions in complex IT/IS systems and business processes of the banking industry. This study highlighted the need to develop adaptations of existing data mining methodologies for the banking domain, which would address three aforementioned concerns.

As a next step, we took a 'micro' perspective, and investigated how standardized data mining processes are applied in actual practice - we examined *what are the perceived gaps in the standardized data mining process (CRISP-DM), and what are adaptations and workaround mechanisms used by practitioners to address them (RQ2)*. For this purpose, we conducted an industrial case study in an actual financial services organization.

The case study involved a representative subset of 6 projects within the selected company. Data was collected from project documentation and via semi-structured interviews with project participants. By combining these data sources, 18 gaps in the CRISP-DM data mining process were identified, as perceived by projects stakeholders. For each gap, the study elicited its potential impact and the adaptations that the interviewed project participants have made to the CRISP-DM process in order to address them.

The identified gaps are spread across all phases of the CRISP-DM life-cycle, which were classified into seven classes based on the management concerns they relate to. About half of the gaps relate to *Requirements* management or the insufficient recognition of *Inter-dependencies* between CRISP-DM phases. These gaps primarily emerge from the semi-waterfall nature of CRISP-DM. These findings

confirm those discussed in [Mar+19], which highlighted that, in practice, there are many pathways for navigating across the tasks and phases of the CRISP-DM life-cycle. Other gaps emerge from the overspecialization of CRISP-DM (cf. Universality gaps). The case study also highlighted that CRISP-DM does not explicitly address *Privacy and Regulatory Compliance* issues, which are omnipresent in the financial services sector, and that it does not explicitly tackle *Validation and Actionability* concerns. Finally, a category of gaps (*Process Gaps*) emerged. It arises from the fact that CRISP-DM only partially incorporates project management activities and does not fully consider the wider organizational and technical context of a data mining project.

The case study also identified five adaptations: (1) the inclusion of explicit iterations between phases or merging of phases, (2) the addition of tasks to address requirements elicitation and management concerns, (3) the addition of 'piloting' tasks for validation, (4) the combination of CRISP-DM with IT development project management practices, and (5) the addition of quality assurance mechanisms.

Lastly, this PhD research has proposed a data mining reference process for the financial services domain – FIN-DM, by *cohesively addressing the identified gaps with extension of CRISP-DM (RQ3)*. Guided by DSRM methodology, the design artifact was developed to tackle all the CRISP-DM deficiencies discovered by the input studies. First, data from systematic literature reviews and industrial case study on the gaps of CRISP-DM was consolidated. Then, we proceeded with formulating the requirements and design aspects of the new data mining process, FIN-DM, and developing its prototype. Next, the prototype was evaluated by conducting demo sessions and semi-structured interviews with the experienced data mining and IT practitioners actively engaged in data mining projects in the financial services industry. We also constructed and distributed a qualitative questionnaire among the pre-selected group of data scientists. Finally, the feedback received from the evaluation was integrated to improve the final version of FIN-DM.

A number of new design principles were implemented in FIN-DM. Also, the somewhat restrictive nature of CRISP-DM was addressed by incorporating full iterativeness among all life-cycle elements. We also embedded higher flexibility and adaptability than in the standard process (CRISP-DM). Dynamic adaptation mechanisms were proposed that would allow FIN-DM to stay relevant and adapt to its composing frameworks (ITIL, COBIT) and external environment changes, such as disruptive technologies and the emergence of new business models.

8.2. Contribution

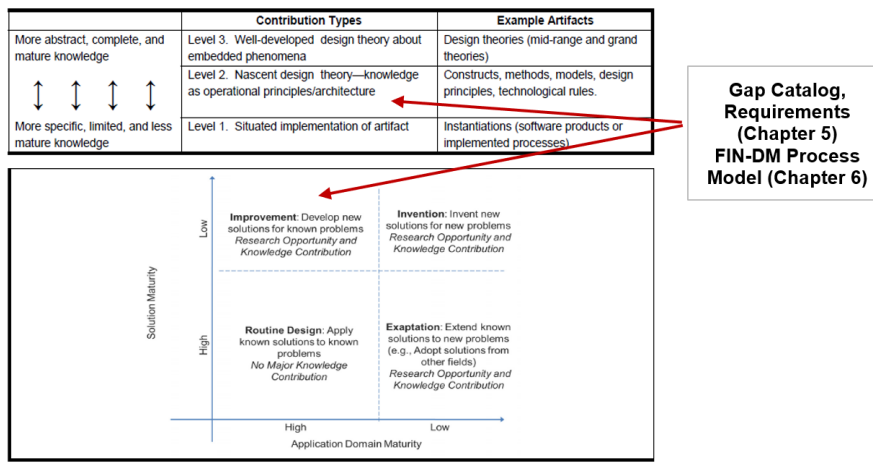
In this research, the non-existence of a specialized data mining process for the financial service sector was addressed. As a result, for financial industry practitioners, the research outcome is a domain-specific process model– FIN-DM– which supports the execution of data mining. FIN-DM is contextualized and is contemporaneous with the main technological and regulatory developments

impacting financial services, and thus addresses challenges of governance, risk, compliance, and privacy and ethical concerns, which are inherent for financial services [CYY21], [Cao21]. To this end, FIN-DM captures and provides data mining experts with concrete phases, tasks, and activities to tackle a number of critically important regulatory requirements and recommendations that emerged over the last decade. Some given regulations are not financial-services specific, but financial services have been one of the key sectors impacted. In particular, FIN-DM supports a systematic approach to effectively address *privacy regulations* (e.g. GDPR) and execute fully *privacy compliant data mining* life-cycles. It also provides a practical approach to execute the best practices of *AI ethics and AI ethical risks mitigation* in all key phases of the data mining project. The proposed tasks support the effective tackling of the key, and most widely recognized, AI ethics principles and concerns - *Transparency, Responsibility, Accountability, Fairness, and Trustworthiness*. Also, FIN-DM will be helpful to practitioners in the financial services industry to design the operations of data mining function in line with modern *internal control operating models* (3LoD - three Lines of Defense Model) and practically fulfill *model risk management requirements* in data mining *model development, validation, and use*. Lastly, FIN-DM supports practitioners with practical solutions to embed *quality assurance* as an integral component of the whole data mining life-cycle– introducing both internal project *verification* tasks and external project *quality controls (validations)*.

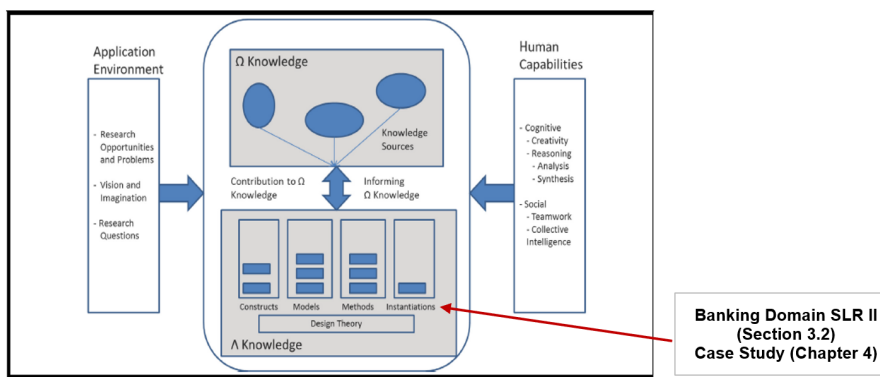
Additionally, FIN-DM supports improving the *business validation, actionability* of data mining results and provides an improved and structured data mining life-cycle *requirements elicitation and management* end-to-end. Furthermore, two other key CRISP-DM gaps - the lack of actual data mining model deployment/ the transition of data mining models into software products and model life-cycle management– have been specified explicitly. To ensure adequate technological support and governance to data mining projects, we have incorporated a number of ITIL and COBIT elements in life-cycle practices and as a broader enablers' set. Also, COBIT elements address the challenges of sound governance and enhanced risk management relevant for financial services, possessing a comprehensive set of control and quality assurance mechanisms. Finally, FIN-DM has multiple advantages and limitations when applied. It can serve as an establishment guide or blueprint for scaling and industrializing data mining functions; nevertheless, it is likely to be more effective in organizations that have already established IT foundations and operate within standard IT industry frameworks (e.g., ITIL). Moreover, knowledge of the CRISP-DM life-cycle and prior experience would be beneficial to foster more effective organizational adoption. Some proposed solutions to gaps might not be entirely financial-sector specifics, but rather applicable in the context of other sectors tackling similar challenges (e.g. telecom). A good example of such generic solutions are *privacy* and *AI ethics* components.

From the design science perspective, based on Gregor's categorization [GH13], FIN-DM classifies as *improvement research*, as we have attempted to create a

solution for the pre-discovered gaps in CRISP-DM within the application context (financial services sector) (see Figure 28a below, with mapping and references to the respective Thesis chapters). Additionally, the research has also contributed to theory in the form of *nascent design theory* [GH13]. Concurrently, by creating FIN-DM, Lambda knowledge space[1] (as per Figure 28b) was enhanced with a more in-depth understanding of data mining models, methods, and their instantiation specifics in the context of financial services. Figure 28b) provides mappings and references to the relevant PhD study parts. Lastly, FIN-DM design and development is also one of the first works where formal design science methodology was applied to derive a process model.



(a) Design Science Research contributions.



(b) Design Science knowledge spaces.

Figure 28. Design Science contributions and knowledge categorization [GH13].

[1]Consists of constructs, models, methods, and instantiations [GH13]

8.3. Limitations and Future Research

This PhD research has a number of limitations. In particular, FIN-DM has undergone an initial evaluation, but it still needs to mature by being tested in practice. At the same time, as underscored in [CD17], the evaluation of reference models constitutes a problem and research challenge. In particular, the reference model, by nature, is able to be generalized, thus its development is decoupled from its application. Hence, the evaluation in practical settings of data mining projects will be focused on FIN-DM's applicability. To this end, as pinpointed in [CD17], each application of reference process increases their chance of adoption. Therefore, the immediate direction for future research could be testing and evaluating FIN-DM in the settings of an actual data mining project(s). Further, apart from comprehensive evaluation, there are promising methods proposed to measure concrete aspects of the data mining frameworks, such as *Actionability*, for example, by evaluating *interestingness* (cf. [CZ07]). While out of scope of this PhD work, conducting such evaluation of the specific FIN-DM aspects, could be another direction for future research.

The other limitation of this PhD study is that FIN-DM was evaluated and considered in the context of financial services. However, the FIN-DM process model and its solutions address several generic and universal CRISP-DM gaps, that could be applied equally to other fields of studies. Therefore, another direction for future research could be associated with the broader scope of data mining problems and solutions embedded in FIN-DM, which go beyond the specifics of the financial sector (e.g., *privacy* and *AI ethics* components). Thus, the potential application of FIN-DM in a broader solution space, i.e., other industries (telecom) and organizations, could also be investigated.

This PhD research focused on identifying gaps in the most widely adopted CRISP-DM methodology. However, the discovered gaps are relevant for other data mining frameworks and methodologies, e.g. KDD, SEMMA or recently proposed ones, such as CASP-DM. Therefore, transposing the FIN-DM adaptation, and using COBIT and ITIL-based extensions in combination with other data mining methodologies beyond CRISP-DM is another potential direction for the future work.

Also, this study took a narrower perspective of data mining as tactical tool, and FIN-DM process model was designed for data mining projects. Another future research avenue is to take a broader view of the more complex data science discipline, which plays a strategic role in decision-making and data-driven transformations of organizations. An open research question is what type of frameworks, processes are necessary in order to drive successful adoption of data science in the financial services. Design and development of data science frameworks, process models could be a practical step in answering this question. In this setting, FIN-DM can be used as starting point, yet, significant extension might be required given the complexity of data science, its strategic role and organizational impact.

BIBLIOGRAPHY

[AAA13] Qasem A Al-Radaideh, Adel Abu Assaf, and Eman Alnagi. "Predicting stock prices using data mining techniques". In: *The International Arab Conference on Information Technology (ACIT'2013)*. 2013.

[AB98] Sarabjot S Anand and Alex G Büchner. *Decision support using data mining*. Financial Times Management, 1998.

[Ada14] Ibrahim Osman Adam. "The ontological, epistemological and methodological debates in information systems research: a partial review". In: *Epistemological and Methodological Debates in Information Systems Research: A Partial Review (March 2014)* (2014).

[Ade+11] JA Adeyiga et al. "A neural network based model for detecting irregularities in e-Banking transactions". In: *African Journal of Computer and ICTs* 4.2 (2011), pp. 7–14.

[Adn17] Ron Adner. "Ecosystem as structure: An actionable construct for strategy". In: *Journal of management* 43.1 (2017), pp. 39–58.

[Adr+04] C. Adrian et al. "Big Data Analytics Implementation for Value Discovery: A Systematic Literature Review." In: *Journal of Theoretical and Applied Information Technology* 93.2 (2004), pp. 385–393.

[AF17] Farzaneh Amani and Adam Fadlalla. "Data mining applications in accounting: A review of the literature and organizing framework". In: *International Journal of Accounting Information Systems* 24 (2017), pp. 32–58. DOI: 10.1016/j.accinf.2016.12.004. URL: https://doi.org/10.1016/j.accinf.2016.12.004.

[Ala+11] Mamoun Alazab et al. "Zero-day Malware Detection based on Supervised Learning Algorithms of API call Signatures". In: *Ninth Australasian Data Mining Conference, AusDM 2011, Ballarat, Australia, December 2011*. 2011, pp. 171–182. URL: http://crpit.com/abstracts/CRPITV121Alazab.html.

[Ale+19] Rohan Alexander et al. "Workshop on barriers to data science adoption: why existing frameworks aren't working". In: *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*. 2019, pp. 384–385.

[Alm+20] Rafael Almeida et al. "Integrating COBIT 5 PAM and TIPA for ITIL Using an Ontology Matching System". In: *Int. J. Hum. Cap. Inf. Technol. Prof.* 11.3 (2020), pp. 74–93. DOI: 10.4018/IJHCITP.2020070105. URL: https://doi.org/10.4018/IJHCITP.2020070105.

[Alt+12] Ahmad Alturki et al. "Validating The Design Science Research Road–map: Through The Lens Of "The Idealised Model For Theory Development"". In: *16th Pacific Asia Conference on Information*

Systems, PACIS 2012, Ho Chi Minh City, Vietnam, 11-15 July 2012. Ed. by Shan L. Pan and Tru H. Cao. 2012, p. 2. URL: http://aisel.aisnet.org/pacis2012/2.

[Ana+98] Sarabjot S Anand et al. "A data mining methodology for cross-sales". In: *Knowledge-based systems* 10.7 (1998), pp. 449–461.

[Ang+18] Santiago Angée et al. "Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects". In: *Knowledge Management in Organizations - 13th International Conference, KMO 2018, Žilina, Slovakia, August 6-10, 2018, Proceedings*. Ed. by Lorna Uden, Branislav Hadzima, and I-Hsien Ting. Vol. 877. Communications in Computer and Information Science. Springer, 2018, pp. 613–624. DOI: 10.1007/978-3-319-95204-8_51. URL: https://doi.org/10.1007/978-3-319-95204-8%5C_51.

[Ang14] Plamen Angelov. "Outside the box: an alternative data analytics framework". In: *Journal of Automation Mobile Robotics and Intelligent Systems* 8.2 (2014), pp. 29–35.

[Ant+07] P Antonellis et al. "A data mining methodology for evaluating maintainability according to ISO/IEC-9126 software engineering–product quality standard". In: *Special Session on System Quality and Maintainability - SQM2007* (2007).

[AP15] Supunmali Ahangama and Danny Chiang Choon Poo. "What Methodological Attributes Are Essential for Novice Users to Analytics? - An Empirical Study". In: *Human Interface and the Management of Information. Information and Knowledge in Context - 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II*. 2015, pp. 77–88. DOI: 10.1007/978-3-319-20618-9_8. URL: https://doi.org/10.1007/978-3-319-20618-9%5C_8.

[ARA11] Faraz Ahmed, M. Zubair Rafique, and Muhammad Abulaish. "A Data Mining Framework for Securing 3G Core Network from GTP Fuzzing Attacks". In: *Information Systems Security - 7th International Conference, ICISS 2011, Kolkata, India, December 15-19, 2011, Procedings*. 2011, pp. 280–293. DOI: 10.1007/978-3-642-25560-1_19. URL: https://doi.org/10.1007/978-3-642-25560-1%5C_19.

[Arn+20] Matthew Arnold et al. "Towards Automating the AI Operations Lifecycle". In: *CoRR* abs/2003.12808 (2020). arXiv: 2003.12808. URL: https://arxiv.org/abs/2003.12808.

[BA96] Ronald J. Brachman and Tej Anand. "The Process of Knowledge Discovery in Databases". In: *Advances in Knowledge Discovery and*

Data Mining. American Association for Artificial Intelligence, 1996, pp. 37–57.

[Bar+01] Daniel Barbará et al. "ADAM: A Testbed for Exploring the Use of Data Mining in Intrusion Detection". In: *SIGMOD Record* 30.4 (2001), pp. 15–24. DOI: 10.1145/604264.604268. URL: https://doi.org/10.1145/604264.604268.

[Bas99] Richard L. Baskerville. "Investigating Information Systems with Action Research". In: *Commun. Assoc. Inf. Syst.* 2 (1999), p. 19. URL: http://aisel.aisnet.org/cais/vol2/iss1/19.

[BC08] Laurent Brisson and Martine Collard. "How to Semantically Enhance a Data Mining Process?" In: *Enterprise Information Systems, 10th International Conference, ICEIS 2008, Barcelona, Spain, June 12-16, 2008, Revised Selected Papers*. 2008, pp. 103–116. DOI: 10.1007/978-3-642-00670-8_8. URL: https://doi.org/10.1007/978-3-642-00670-8%5C_8.

[BC14a] Zhuming Bi and David Cochran. "Big data analytics with applications". In: *Journal of Management Analytics* 1.4 (2014), pp. 249–265. DOI: 10.1080/23270012.2014.992985. eprint: https://doi.org/10.1080/23270012.2014.992985. URL: https://doi.org/10.1080/23270012.2014.992985.

[BC14b] Zhuming Bi and David Cochran. "Big data analytics with applications". In: *Journal of Management Analytics* 1.4 (2014), pp. 249–265.

[BD18] Desamparados Blazquez and Josep Domenech. "Big Data sources and methods for social and economic analyses". In: *Technological Forecasting and Social Change* 130 (2018), pp. 99–113.

[BE15] T. Femina Bahari and M. Sudheep Elayidom. "An efficient CRM-data mining framework for the prediction of customer behaviour". In: *Procedia computer science* 46 (2015), pp. 725–731.

[BG11] Sule Balkan and Michael Goul. "A Portfolio Theoretic Approach to Administering Advanced Analytics: The Case of Multi-Stage Campaign Management". In: *44th Hawaii International International Conference on Systems Science (HICSS-44 2011), Proceedings, 4-7 January 2011, Koloa, Kauai, HI, USA*. IEEE Computer Society, 2011, pp. 1–10. DOI: 10.1109/HICSS.2011.22. URL: https://doi.org/10.1109/HICSS.2011.22.

[BG16] Muhammad Rizwan Bashir and Asif Qumer Gill. "Towards an IoT Big Data Analytics Framework: Smart Buildings Systems". In: *18th IEEE International Conference on High Performance Computing and Communications; 14th IEEE International Conference on Smart City; 2nd IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2016, Sydney, Australia, December 12-*

14, 2016. 2016, pp. 1325–1332. DOI: 10.1109/HPCC-SmartCity-DSS.2016.0188. URL: https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0188.

[Bha01] Anol Bhattacherjee. "Understanding Information Systems Continuance: An Expectation-Confirmation Model". In: *MIS Quarterly* 25.3 (2001), pp. 351–370.

[Bil+14] Evmorfia Biliri et al. "Infusing social data analytics into Future Internet applications for manufacturing". In: *11th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2014, Doha, Qatar, November 10-13, 2014*. 2014, pp. 515–522. DOI: 10.1109/AICCSA.2014.7073242. URL: https://doi.org/10.1109/AICCSA.2014.7073242.

[BKC11] M Pesaran Behbahani, S Khaddaj, and I Choudhury. "A multilayer data mining approach to an optimized ebusiness analytics framework". In: *International Proceedings of Economics Development and Research* (2011), pp. 66–71.

[Bla+18] Rob Black et al. "Model risk: illuminating the black box". In: *British Actuarial Journal* 23 (2018).

[BM01] Indranil Bose and Radha K. Mahapatra. "Business data mining - a machine learning perspective". In: *Information & Management* 39.3 (2001), pp. 211–225. DOI: 10.1016/S0378-7206(01)00091-X. URL: https://doi.org/10.1016/S0378-7206(01)00091-X.

[BM02] Hendrik Blockeel and Steve Moyle. "Collaborative data mining needs centralised model evaluation". In: *Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned*. 2002, pp. 21–28.

[BM98] Alex G. Büchner and Maurice D. Mulvenna. "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining". In: *SIGMOD Record* 27.4 (1998), pp. 54–61. DOI: 10.1145/306101.306124. URL: https://doi.org/10.1145/306101.306124.

[BP14] Bettina Berendt and Sören Preibusch. "Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence". In: *Artif. Intell. Law* 22.2 (2014), pp. 175–209. DOI: 10.1007/s10506-013-9152-0. URL: https://doi.org/10.1007/s10506-013-9152-0.

[BPV09] Richard L. Baskerville, Jan Pries-Heje, and John R. Venable. "Soft design science methodology". In: *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2009, Philadelphia, Pennsylvania, USA, May 7-8, 2009*. Ed. by Vijay K. Vaishnavi and Sandeep Pu-

rao. ACM, 2009. DOI: 10.1145/1555619.1555631. URL: https://doi.org/10.1145/1555619.1555631.

[Bra+15] Richard Braun et al. "Proposal for Requirements Driven Design Science Research". In: *New Horizons in Design Science: Broadening the Research Agenda - 10th International Conference, DESRIST 2015, Dublin, Ireland, May 20-22, 2015, Proceedings*. Ed. by Brian Donnellan et al. Vol. 9073. Lecture Notes in Computer Science. Springer, 2015, pp. 135–151. DOI: 10.1007/978-3-319-18714-3_9. URL: https://doi.org/10.1007/978-3-319-18714-3%5C_9.

[Bra20] Patrick Bradley. "Risk management standards and the active management of malicious intent in artificial superintelligence". In: *AI Soc.* 35.2 (2020), pp. 319–328. DOI: 10.1007/s00146-019-00890-2. URL: https://doi.org/10.1007/s00146-019-00890-2.

[Bre+07] Pearl Brereton et al. "Lessons from applying the systematic literature review process within the software engineering domain". In: *Journal of Systems and Software* 80.4 (2007), pp. 571–583. DOI: 10.1016/j.jss.2006.07.009. URL: https://doi.org/10.1016/j.jss.2006.07.009.

[Buc+99] Alex G Buchner et al. "An internet-enabled knowledge discovery process". In: *Proceedings of the 9th international database conference, Hong Kong*. Vol. 1999. 1999, pp. 13–27.

[Bud+06] David Budgen et al. "Investigating the applicability of the evidence-based paradigm to software engineering". In: *Proceedings of the 2006 international workshop on Workshop on interdisciplinary software engineering research, WISER 2006, Shanghai, China, May 20, 2006*. 2006, pp. 7–14. DOI: 10.1145/1137661.1137665. URL: https://doi.org/10.1145/1137661.1137665.

[BW18] T. Bellof and C.S. Wehn. "On the treatment of model risk in the internal capital adequacy assessment process". In: *Journal of Applied Finance and Banking* 8(4) (2018), pp. 1–15.

[BW98] R. Baskerville and A. T Wood-Harper. "Diversity in Information Systems Action Research Methods". In: *Eur. J. Inf. Syst.* 7.2 (1998), pp. 90–107.

[CA20] C.J. Costa and J.T. Aparicio. "OST-DS: A Methodology to Boost Data Science". In: *In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2020, pp. 1–6.

[Cab+97] Peter Cabena et al. *Discovering data mining: from concept to implementation*. Prentice Hall PTR New Jersey, 1997.

[Cao+07] Longbing Cao et al. "Domain-Driven, Actionable Knowledge Discovery". In: *IEEE Intell. Syst.* 22.4 (2007), pp. 78–88. DOI: 10.

1109/MIS.2007.67. URL: https://doi.org/10.1109/MIS.2007.67.

[Cao10] Longbing Cao. "Domain-Driven Data Mining: Challenges and Prospects". In: *IEEE Trans. Knowl. Data Eng.* 22.6 (2010), pp. 755–769. DOI: 10.1109/TKDE.2010.32. URL: https://doi.org/10.1109/TKDE.2010.32.

[Cao16] Longbing Cao. "Data Science: Nature and Pitfalls". In: *IEEE Intell. Syst.* 31.5 (2016), pp. 66–75. DOI: 10.1109/MIS.2016.86. URL: https://doi.org/10.1109/MIS.2016.86.

[Cao17] Longbing Cao. "Data science: challenges and directions". In: *Commun. ACM* 60.8 (2017), pp. 59–68. DOI: 10.1145/3015456. URL: https://doi.org/10.1145/3015456.

[Cao20] Longbing Cao. "AI in Finance: A Review". In: *Available at SSRN 3647625* (2020).

[Cao21] Longbing Cao. "AI in Finance: Challenges, Techniques and Opportunities". In: *CoRR* abs/2107.09051 (2021). arXiv: 2107.09051. URL: https://arxiv.org/abs/2107.09051.

[Cap+17] Alfonso Capozzoli et al. "Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings". In: *Sustainable cities and society* 35 (2017), pp. 191–208.

[CBK16] Thai Chuan Chee, Ahmad Suhaimi Baharudin, and Kamal Karkonasasi. "Data Mining Framework for Test Time Optimization in Industrial Electronics Manufacturing Enterprise". In: *International Journal of Applied Engineering Research* (2016).

[CC03] Mario Cannataro and Carmela Comito. "A data mining ontology for grid programming". In: *Proc. 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing*. Citeseer. 2003, pp. 113–134.

[CC08] Chen-Fu Chien and Li-Fei Chen. "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry". In: *Expert Syst. Appl.* 34.1 (2008), pp. 280–290. DOI: 10.1016/j.eswa.2006.09.003. URL: https://doi.org/10.1016/j.eswa.2006.09.003.

[CCS12] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4 (2012), pp. 1165–1188. URL: http://www.misq.org/skin/frontend/default/misq/pdf/V36I4/SI%5C_ChenIntroduction.pdf.

[CD17] Christian Czarnecki and Christian Dietze. "Domain-Specific Reference Modeling in the Telecommunications Industry". In: *Designing the Digital Transformation - 12th International Conference, DESRIST 2017, Karlsruhe, Germany, May 30 - June 1, 2017, Pro-*

ceedings. Ed. by Alexander Maedche, Jan vom Brocke, and Alan R. Hevner. Vol. 10243. Lecture Notes in Computer Science. Springer, 2017, pp. 313–329. DOI: 10.1007/978-3-319-59144-5_19. URL: https://doi.org/10.1007/978-3-319-59144-5%5C_19.

[CDL14] Chen-Fu Chien, Alejandra Campero Diaz, and Yu-Bin Lan. "A data mining approach for analyzing semiconductor MES and FDC data to enhance overall usage effectiveness (OUE)". In: *Int. J. Comput. Intell. Syst.* 7.sup2 (2014), pp. 52–65. DOI: 10.1080/18756891.2014.947114. URL: https://doi.org/10.1080/18756891.2014.947114.

[Cha+00] P. Chapman et al. "CRISP-DM 1.0 Step-by-step data mining guide". In: *SPSS Inc.* (2000).

[Che+01] Yiqiang Chen et al. "Mining audio/visual database for speech driven face animation". In: *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics: "e-Systems and e-Man for Cybernetics in Cyberspace", Tucson, Arizona, USA, 7-10 October 2001*. 2001, pp. 2638–2643. DOI: 10.1109/ICSMC.2001.972962. URL: https://doi.org/10.1109/ICSMC.2001.972962.

[Che+14] Sergey Chernov et al. "Data mining framework for random access failure detection in LTE networks". In: *25th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communication, PIMRC 2014, Washington DC, USA, September 2-5, 2014*. 2014, pp. 1321–1326. DOI: 10.1109/PIMRC.2014.7136373. URL: https://doi.org/10.1109/PIMRC.2014.7136373.

[Çin+11] Esma Nur Çinicioğlu et al. "A framework for automated association mining over multiple databases". In: *2011 International Symposium on Innovations in Intelligent Systems and Applications*. IEEE. 2011, pp. 79–85.

[CK05] Krzysztof J Cios and Lukasz A Kurgan. "Trends in data mining and knowledge discovery". In: *Advanced techniques in knowledge discovery and data mining*. Springer, 2005, pp. 1–26.

[CKA13] George Chatzikonstantinou, Kostas Kontogiannis, and Ioanna-Maria Attarian. "A Goal Driven Framework for Software Project Data Analytics". In: *Advanced Information Systems Engineering - 25th International Conference, CAiSE 2013, Valencia, Spain, June 17-21, 2013. Proceedings*. 2013, pp. 546–561. DOI: 10.1007/978-3-642-38709-8_35. URL: https://doi.org/10.1007/978-3-642-38709-8%5C_35.

[CKH16a] Hong-Mei Chen, Rick Kazman, and Serge Haziyev. "Agile Big Data Analytics Development: An Architecture-Centric Approach". In: *49th Hawaii International Conference on System Sciences, HICSS*

2016, Koloa, HI, USA, January 5-8, 2016. 2016, pp. 5378–5387. DOI: 10.1109/HICSS.2016.665. URL: https://doi.org/10.1109/HICSS.2016.665.

[CKH16b] Hong-Mei Chen, Rick Kazman, and Serge Haziyev. "Agile big data analytics development: An architecture-centric approach". In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE. 2016, pp. 5378–5387.

[Cla18] Andrew Clark. "The machine learning audit-CRISP-DM Framework". In: *ISACA Journal* 1 (2018), pp. 42–47.

[Col17] L. Columbus. *53% Of Companies Are Adopting Big Data Analytics*. 2017. URL: https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/ (visited on 02/24/2020).

[CPR15] Sergey Chernov, Dmitry Petrov, and Tapani Ristaniemi. "Location accuracy impact on cell outage detection in LTE-A networks". In: *International Wireless Communications and Mobile Computing Conference, IWCMC 2015, Dubrovnik, Croatia, August 24-28, 2015*. 2015, pp. 1162–1167. DOI: 10.1109/IWCMC.2015.7289247. URL: https://doi.org/10.1109/IWCMC.2015.7289247.

[CPT16] Alfredo Cuzzocrea, Giuseppe Psaila, and Maurizio Toccu. "An Innovative Framework for Effectively and Efficiently Supporting Big Data Analytics over Geo-Located Mobile Social Media". In: *Proceedings of the 20th International Database Engineering & Applications Symposium, IDEAS 2016, Montreal, QC, Canada, July 11-13, 2016*. 2016, pp. 62–69. DOI: 10.1145/2938503.2938517. URL: https://doi.org/10.1145/2938503.2938517.

[CR17] Andrea Fronzetti Colladon and Elisa Remondi. "Using social network analysis to prevent money laundering". In: *Expert Syst. Appl.* 67 (2017), pp. 49–58. DOI: 10.1016/j.eswa.2016.09.029. URL: https://doi.org/10.1016/j.eswa.2016.09.029.

[CRN16] Mounaim Cortet, Tom Rijks, and Shikko Nijland. "PSD2: The digital transformation accelerator for banks". In: *Journal of Payments Strategy & Systems* 10.1 (2016), pp. 13–27.

[CSG15] Leona Chandra, Stefan Seidel, and Shirley Gregor. "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions". In: *48th Hawaii International Conference on System Sciences, HICSS 2015, Kauai, Hawaii, USA, January 5-8, 2015*. Ed. by Tung X. Bui and Ralph H. Sprague Jr. IEEE Computer Society, 2015, pp. 4039–4048. DOI: 10.1109/HICSS.2015.485. URL: https://doi.org/10.1109/HICSS.2015.485.

[CSZ05] Longbing Cao, R Schurmann, and Chengqi Zhang. “Domain-driven in-depth pattern discovery: a practical methodology”. In: *Australian Data Mining Conference*. The University of Technology, Sydney. 2005.

[Cus20] James J. Cusick. “Business Value of ITSM. Requirement or Mirage?” In: *CoRR* abs/2001.00219 (2020). arXiv: 2001.00219. URL: http://arxiv.org/abs/2001.00219.

[CYY21] Longbing Cao, Qiang Yang, and Philip S. Yu. “Data science and AI in FinTech: an overview”. In: *Int. J. Data Sci. Anal.* 12.2 (2021), pp. 81–99. DOI: 10.1007/s41060-021-00278-w. URL: https://doi.org/10.1007/s41060-021-00278-w.

[CZ06a] Longbing Cao and Chengqi Zhang. “Domain-Driven Actionable Knowledge Discovery in the Real World”. In: *Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006, Proceedings*. Ed. by Wee Keong Ng et al. Vol. 3918. Lecture Notes in Computer Science. Springer, 2006, pp. 821–830. DOI: 10.1007/11731139_96. URL: https://doi.org/10.1007/11731139%5C_96.

[CZ06b] Longbing Cao and Chengqi Zhang. “Domain-Driven Data Mining: A Practical Methodology”. In: *Int. J. Data Warehous. Min.* 2.4 (2006), pp. 49–65. DOI: 10.4018/jdwm.2006100103. URL: https://doi.org/10.4018/jdwm.2006100103.

[CZ07] Longbing Cao and Chengqi Zhang. “The Evolution of KDD: towards Domain-Driven Data Mining”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 21.4 (2007), pp. 677–692. DOI: 10.1142/S0218001407005612. URL: https://doi.org/10.1142/S0218001407005612.

[CZ08] Longbing Cao and Chengqi Zhang. “Domain driven data mining”. In: *Data Mining and Knowledge Discovery Technologies*. IGI Global, 2008, pp. 196–223.

[DD07] Qin Ding and Charles Daniel. “Multimedia Data Mining Framework for Banner Images”. In: *Multimedia Data Mining and Knowledge Discovery*. Springer, 2007, pp. 448–457.

[Deb+01] JCW Debuse et al. “Building the KDD Roadmap”. In: *Industrial Knowledge Management*. Springer, 2001, pp. 179–196.

[Deb07] JCW Debuse. “Extending data mining methodologies to encompass organizational factors”. In: *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research* 24.2 (2007), pp. 183–190.

[DGG09] Xiong Deng, Moustafa Ghanem, and Yike Guo. “Real-Time Data Mining Methodology and a Supporting Framework”. In: *Third International Conference on Network and System Security, NSS 2009,*

Gold Coast, Queensland, Australia, October 19-21, 2009. 2009, pp. 522–527. DOI: 10.1109/NSS.2009.49. URL: https://doi.org/10.1109/NSS.2009.49.

[DH17] Thomas Davenport and Jeanne Harris. *Competing on analytics: The new science of winning*. Harvard Business Press, 2017.

[Dig17] Virginia Dignum. "Responsible Autonomy". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. Ed. by Carles Sierra. ijcai.org, 2017, pp. 4698–4704. DOI: 10.24963/ijcai.2017/655. URL: https://doi.org/10.24963/ijcai.2017/655.

[Dig18] Virginia Dignum. "Ethics in artificial intelligence: introduction to the special issue". In: *Ethics Inf. Technol.* 20.1 (2018), pp. 1–3. DOI: 10.1007/s10676-018-9450-z. URL: https://doi.org/10.1007/s10676-018-9450-z.

[Dor08] Doreswamy. "A Survey For Data Mining Frame Work For Polymer Matrix Composite Engineering Materials Design Applications". In: *Int. J. Comput. Intell. Syst.* 1.4 (2008), pp. 313–328. DOI: 10.1080/18756891.2008.9727628. URL: https://doi.org/10.1080/18756891.2008.9727628.

[Dou11] K. Doughty. "Guest editorial: the three lines of defence related to risk governance." In: 5. ISACA (Information Systems Audit and Control Association), 2011, p. 6.

[DPP11] Jeremiah D. Deng, Martin K. Purvis, and Maryam Purvis. "Software Effort Estimation: Harmonizing Algorithms and Domain Knowledge in an Integrated Data Mining Approach". In: *IJIIT* 7.3 (2011), pp. 41–53. DOI: 10.4018/jiit.2011070104. URL: https://doi.org/10.4018/jiit.2011070104.

[Dre15] Andreas Drechsler. "Designing to Inform: Toward Conceptualizing Practitioner Audiences for Socio-technical Artifacts in Design Science Research in the Information Systems Discipline". In: *Informing Sci. Int. J. an Emerg. Transdiscipl.* 18 (2015), pp. 31–47. DOI: 10.28945/2288. URL: https://doi.org/10.28945/2288.

[Dre17] Christian Dremel. "Barriers to the adoption of big data analytics in the automotive sector". In: (2017).

[DTA12a] David Diaz, Babis Theodoulidis, and Eliza Abioye. "Cross-Border Challenges in Financial Markets Monitoring and Surveillance: A Case Study of Customer-Driven Service Value Networks". In: *2012 Annual SRII Global Conference, San Jose, CA, USA, July 24-27, 2012*. IEEE Computer Society, 2012, pp. 146–157. DOI: 10.1109/SRII.2012.26. URL: https://doi.org/10.1109/SRII.2012.26.

[DTA12b] David Diaz, Babis Theodoulidis, and Eliza Abioye. “Cross-border challenges in financial markets monitoring and surveillance: a case study of customer-driven service value networks”. In: *2012 Annual SRII Global Conference*. IEEE. 2012, pp. 146–157.

[Du+17] Min Du et al. “DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning”. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. 2017, pp. 1285–1298. DOI: 10.1145/3133956.3134015. URL: https://doi.org/10.1145/3133956.3134015.

[Dum+18] Marlon Dumas et al. *Fundamentals of Business Process Management, Second Edition*. Springer, 2018. ISBN: 978-3-662-56508-7. DOI: 10.1007/978-3-662-56509-4. URL: https://doi.org/10.1007/978-3-662-56509-4.

[EBN17] Wael Etaiwi, Mariam Biltawi, and Ghazi Naymat. “Evaluation of classification algorithms for banking customer’s behavior under Apache Spark Data Processing System”. In: *The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops, September 18-20, 2017, Lund, Sweden*. Ed. by Elhadi M. Shakshuki. Vol. 113. Procedia Computer Science. Elsevier, 2017, pp. 559–564. DOI: 10.1016/j.procs.2017.08.280. URL: https://doi.org/10.1016/j.procs.2017.08.280.

[EC21] EC. *Artificial Intelligence - Important milestones of the AI Strategy*. 2021. URL: https://ec.europa.eu/digital-single-market/en/artificial-intelligence (visited on 01/08/2021).

[ECZ17] Gürdal Ertek, Xu Chi, and Allan N. Zhang. “A Framework for Mining RFID Data From Schedule-Based Systems”. In: *IEEE Trans. Systems, Man, and Cybernetics: Systems* 47.11 (2017), pp. 2967–2984. DOI: 10.1109/TSMC.2016.2557762. URL: https://doi.org/10.1109/TSMC.2016.2557762.

[FPS96a] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “The KDD process for extracting useful knowledge from volumes of data”. In: *Communications of the ACM* 39.11 (1996), pp. 27–34.

[FPS96b] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “From Data Mining to Knowledge Discovery in Databases”. In: *AI Magazine* 17.3 (1996), pp. 37–54. URL: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230.

[FPS96c] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “Knowledge Discovery and Data Mining: Towards a Unifying Frame-

work". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. 1996, pp. 82–88. URL: http://www.aaai.org/Library/KDD/1996/kdd96-014.php.

[FPS96d] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD Process for Extracting Useful Knowledge from Volumes of Data". In: *Commun. ACM* 39.11 (1996), pp. 27–34. DOI: 10.1145/240455.240464. URL: https://doi.org/10.1145/240455.240464.

[Fra08] Damien François. "Methodology and standards for data analysis with machine learning tools". In: *ESANN 2008, 16th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 23-25, 2008, Proceedings*. 2008, pp. 239–246. URL: https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2008-4.pdf.

[Fri+19] Sorelle A. Friedler et al. "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, 2019, pp. 329–338. DOI: 10.1145/3287560.3287589. URL: https://doi.org/10.1145/3287560.3287589.

[FYC16] Yujie Fan, Yanfang Ye, and Lifei Chen. "Malicious sequential pattern mining for automatic malware detection". In: *Expert Syst. Appl.* 52 (2016), pp. 16–25. DOI: 10.1016/j.eswa.2016.01.002. URL: https://doi.org/10.1016/j.eswa.2016.01.002.

[Gan+96] M Ganesh et al. "Visual data mining: Framework and algorithm development". In: *Department of Computing and Information Sciences, University of Minnesota, MN, USA* (1996).

[Gar+17] D. Garcia et al. "Data analytics methodology for monitoring quality sensors and events in the Barcelona drinking water network". In: *Journal of Hydroinformatics* 19.1 (2017), pp. 123–137.

[GBC15] Ruibin Geng, Indranil Bose, and Xi Chen. "Prediction of financial distress: An empirical study of listed Chinese companies using data mining". In: *Eur. J. Oper. Res.* 241.1 (2015), pp. 236–247. DOI: 10.1016/j.ejor.2014.08.016. URL: https://doi.org/10.1016/j.ejor.2014.08.016.

[GD04] Christine Gertosio and Alain Dussauchoy. "Knowledge discovery from industrial databases". In: *J. Intelligent Manufacturing* 15.1 (2004), pp. 29–37. DOI: 10.1023/B\%3AJIMS.0000010073.54241.e7. URL: https://doi.org/10.1023/B%5C%3AJIMS.0000010073.54241.e7.

[GDQ20] Stuart D. Galup, Ronald Dattero, and Jing Quan. “What do agile, lean, and ITIL mean to DevOps?” In: *Commun. ACM* 63.10 (2020), pp. 48–53. DOI: 10.1145/3372114. URL: https://doi.org/10.1145/3372114.

[GF10] Qiang Guan and Song Fu. “auto-AID: A data mining framework for autonomic anomaly identification in networked computer systems”. In: *29th International Performance Computing and Communications Conference, IPCCC 2010, 9-11 December 2010, Albuquerque, NM, USA*. 2010, pp. 73–80. DOI: 10.1109/PCCC.2010.5682334. URL: https://doi.org/10.1109/PCCC.2010.5682334.

[GFM16] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. “The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature”. In: *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, EASE 2016, Limerick, Ireland, June 01 - 03, 2016*. 2016, 26:1–26:6. DOI: 10.1145/2915970.2916008. URL: https://doi.org/10.1145/2915970.2916008.

[GFM19] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. “Guidelines for including grey literature and conducting multivocal literature reviews in software engineering”. In: *Information & Software Technology* 106 (2019), pp. 101–121. DOI: 10.1016/j.infsof.2018.09.006. URL: https://doi.org/10.1016/j.infsof.2018.09.006.

[GH13] Shirley Gregor and Alan R. Hevner. “Positioning and Presenting Design Science Research for Maximum Impact”. In: *MIS Q.* 37.2 (2013), pp. 337–355. URL: http://misq.org/positioning-and-presenting-design-science-research-for-maximum-impact.html.

[GH15] Amir Gandomi and Murtaza Haider. “Beyond the hype: Big data concepts, methods, and analytics”. In: *Int J. Information Management* 35.2 (2015), pp. 137–144. DOI: 10.1016/j.ijinfomgt.2014.10.007. URL: https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

[Gho+18] Soumadip Ghosh et al. “A Comparative Study to the Bank Market Prediction”. In: *Machine Learning and Data Mining in Pattern Recognition - 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part I*. Ed. by Petra Perner. Vol. 10934. Lecture Notes in Computer Science. Springer, 2018, pp. 259–268. DOI: 10.1007/978-3-319-96136-1_21. URL: https://doi.org/10.1007/978-3-319-96136-1%5C_21.

[GK19] Nihan Gulsoy and Sinem Kulluk. “A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers”. In: *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 9.3 (2019). DOI: 10.1002/widm.1299. URL: https://doi.org/10.1002/widm.1299.

[GPK13] João Bártolo Gomes, Clifton Phua, and Shonali Krishnaswamy. “Where Will You Go? Mobile Data Mining for Next Place Prediction”. In: *Data Warehousing and Knowledge Discovery - 15th International Conference, DaWaK 2013, Prague, Czech Republic, August 26-29, 2013. Proceedings*. 2013, pp. 146–158. DOI: 10.1007/978-3-642-40131-2_13. URL: https://doi.org/10.1007/978-3-642-40131-2%5C_13.

[Güd+14] Mennan Güder et al. “Data mining framework for power quality event characterization of iron and steel plants”. In: *2014 IEEE Industry Application Society Annual Meeting, Vancouver, BC, Canada, October 5-9, 2014*. 2014, pp. 1–11. DOI: 10.1109/IAS.2014.6978449. URL: https://doi.org/10.1109/IAS.2014.6978449.

[Gul+10] Erik Guldentops et al. *Aligning Cobit, ITIL and ISO 17799 for Business Benefit: A Management Briefing from ITGI und OGC*. 2010.

[Haa+21] Mark Haakman et al. “AI lifecycle models need to be revised”. In: *Empir. Softw. Eng.* 26.5 (2021), p. 95. DOI: 10.1007/s10664-021-09993-1. URL: https://doi.org/10.1007/s10664-021-09993-1.

[Ham+20a] Koichi Hamada et al. “Guidelines for Quality Assurance of Machine Learning-Based Artificial Intelligence”. In: *International Journal of Software Engineering and Knowledge Engineering, Special Issue: Best Paper from SEKE 2020* 11.12 (2020), pp. 1589–1606.

[Ham+20b] Koichi Hamada et al. “Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence”. In: *The 32nd International Conference on Software Engineering and Knowledge Engineering, SEKE 2020, KSIR Virtual Conference Center, USA, July 9-19, 2020*. Ed. by Raúl Garcia-Castro. KSI Research Inc., 2020, pp. 335–341. DOI: 10.18293/SEKE2020-094. URL: https://doi.org/10.18293/SEKE2020-094.

[Har15] S.E. Harpe. “How to analyze Likert and other rating scale data”. In: *Currents in Pharmacy Teaching and Learning* 7.6 (2015), pp. 836–850.

[HBC16] Sara Hajian, Francesco Bonchi, and Carlos Castillo. “Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco,*

CA, USA, August 13-17, 2016. Ed. by Balaji Krishnapuram et al. ACM, 2016, pp. 2125–2126. DOI: 10.1145/2939672.2945386. URL: https://doi.org/10.1145/2939672.2945386.

[HBJ03] Mahmood Hossain, Susan M. Bridges, and Rayford B. Vaughn Jr. "Adaptive intrusion detection with data mining". In: *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics: Washington, D.C., USA, 5-8 October 2003*. 2003, pp. 3097–3103. DOI: 10.1109/ICSMC.2003.1244366. URL: https://doi.org/10.1109/ICSMC.2003.1244366.

[HC10] Alan Hevner and Samir Chatterjee. "Design science research in information systems". In: *Design research in information systems*. Springer, 2010, pp. 9–22.

[Hev+04a] Alan R Hevner et al. "Design science in information systems research". In: *MIS quarterly* (2004), pp. 75–105.

[Hev+04b] Alan R. Hevner et al. "Design Science in Information Systems Research". In: *MIS Q.* 28.1 (2004), pp. 75–105. URL: http://misq.org/design-science-in-information-systems-research.html.

[Hev07] Alan R Hevner. "A three cycle view of design science research". In: *Scandinavian journal of information systems* 19.2 (2007), p. 4.

[HF09a] Yang Hang and Simon Fong. "A Framework of Business Intelligence-Driven Data Mining for E-business". In: *International Conference on Networked Computing and Advanced Information Management, NCM 2009, Fifth International Joint Conference on INC, IMS and IDC: INC 2009: International Conference on Networked Computing, IMS 2009: International Conference on Advanced Information Management and Service, IDC 2009: International Conference on Digital Content, Multimedia Technology and its Applications, Seoul, Korea, August 25-27, 2009*. 2009, pp. 1964–1970. DOI: 10.1109/NCM.2009.403. URL: https://doi.org/10.1109/NCM.2009.403.

[HF09b] Yang Hang and Simon Fong. "A framework of business intelligence-driven data mining for e-business". In: *2009 Fifth International Joint Conference on INC, IMS and IDC*. IEEE. 2009, pp. 1964–1970.

[HGD13] Steven De Haes, Wim Van Grembergen, and Roger S. Debreceny. "COBIT 5 and Enterprise Governance of Information Technology: Building Blocks and Research Opportunities". In: *J. Inf. Syst.* 27.1 (2013), pp. 307–324. DOI: 10.2308/isys-50422. URL: https://doi.org/10.2308/isys-50422.

[HHS18a] Hossein Hassani, Xu Huang, and Emmanuel Silva. "Digitalisation and big data mining in banking". In: *Big Data and Cognitive Computing* 2.3 (2018), p. 18.

[HHS18b] Hossein Hassani, Xu Huang, and Emmanuel Sirimal Silva. "Digitalisation and Big Data Mining in Banking". In: *Big Data Cogn. Comput.* 2.3 (2018), p. 18. DOI: 10.3390/bdcc2030018. URL: https://doi.org/10.3390/bdcc2030018.

[Hig20] EC High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy Artificial Intelligence*. Brussels: High-Level Expert Group on AI, European Commission, 2020.

[HK89] Rudy Hirschheim and Heinz K. Klein. "Four Paradigms of Information Systems Development". In: *Commun. ACM* 32.10 (1989), pp. 1199–1216. DOI: 10.1145/67933.67937. URL: https://doi.org/10.1145/67933.67937.

[HSC04] Choochart Haruechaiyasak, Mei-Ling Shyu, and Shu-Ching Chen. "A Data Mining Framework for Building A Web-Page Recommender System". In: *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, IRI - 2004, November 8-10, 2004, Las Vegas Hilton, Las Vegas, NV, USA*. 2004, pp. 357–362. DOI: 10.1109/IRI.2004.1431487. URL: https://doi.org/10.1109/IRI.2004.1431487.

[Hsu09] Chih-Hung Hsu. "Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry". In: *Expert Syst. Appl.* 36.3 (2009), pp. 4185–4191. DOI: 10.1016/j.eswa.2008.04.009. URL: https://doi.org/10.1016/j.eswa.2008.04.009.

[Hu+10] Xiao Hu et al. "A data mining framework for time series estimation". In: *Journal of Biomedical Informatics* 43.2 (2010), pp. 190–199. DOI: 10.1016/j.jbi.2009.11.002. URL: https://doi.org/10.1016/j.jbi.2009.11.002.

[Hua+02] Xin Huang et al. "Mining High-Level User Concepts with Multiple Instance Learning and Relevance Feedback for Content-Based Image Retrieval". In: *Mining Multimedia and Complex Data, KDD Workshop MDM/KDD 2002, PAKDD Workshop KDMCD 2002, Revised Papers*. 2002, pp. 50–67. DOI: 10.1007/978-3-540-39666-6_4. URL: https://doi.org/10.1007/978-3-540-39666-6%5C_4.

[Hub+19] Steffen Huber et al. "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model". In: *Procedia CIRP* 79 (2019), pp. 403–408.

[Hum+19] Waldemar Hummer et al. "ModelOps: Cloud-Based Lifecycle Management for Reliable and Trusted AI". In: *IEEE International Conference on Cloud Engineering, IC2E 2019, Prague, Czech Republic, June 24-27, 2019*. IEEE, 2019, pp. 113–120. DOI: 10.1109/IC2E.2019.00025. URL: https://doi.org/10.1109/IC2E.2019.00025.

[IBM16] Corporation IBM. *Analytics Solutions Unified Method*. IBM Corporation, New Orchard Road Armonk, NY 10504., 2016.

[IIA13] IIA. "IIA position paper: the three lines of defense in effective risk management and control." In: Institute of Internal Auditors, 2013.

[J05] Tian J. *Software quality engineering: testing, quality assurance, and quantifiable improvement*. 2nd ed. New Jersey: John Wiley & Sons, 2005.

[JH07] Zhen Ming Jiang and Ahmed E. Hassan. "A Framework for Studying Clones In Large Software Systems". In: *Seventh IEEE International Workshop on Source Code Analysis and Manipulation (SCAM 2007), September 30 - October 1, 2007, Paris, France*. 2007, pp. 203–212. DOI: 10.1109/SCAM.2007.19. URL: https://doi.org/10.1109/SCAM.2007.19.

[Kab16] Md Humayun Kabir. "Data mining framework for generating sales decision making information using association rules." In: *International Journal of Advanced Computer Science and Applications* 7.5 (2016), pp. 378–385.

[Kad13] Aziz Kaddouri. "Why human expertise is critical for data mining". In: *International Journal of Computer and Information Technology* 2.1 (2013), pp. 99–108.

[Kan+17] Seokho Kang et al. "Mining the relationship between production and customer service data for failure analysis of industrial products". In: *Computers & Industrial Engineering* 106 (2017), pp. 137–146. DOI: 10.1016/j.cie.2017.01.028. URL: https://doi.org/10.1016/j.cie.2017.01.028.

[KBB15] Barbara A. Kitchenham, David Budgen, and Pearl Brereton. *Evidence-based software engineering and systematic reviews*. CRC press., 2015.

[KC07] Barbara Kitchenham and Stuart Charters. "Guidelines for performing systematic literature reviews in software engineering". In: *EBSE Technical Report No. EBSE-2007-01* (2007).

[Kee+07] Staffs Keele et al. *Guidelines for performing systematic literature reviews in software engineering*. Tech. rep. Citeseer, 2007.

[KI20] Shruti Kashyap and Einar Iveroth. "Transparency and accountability influences of regulation on risk control: the case of a Swedish bank." In: *Journal of Management and Governance* (2020). DOI: 10.1007/s10997-020-09550-w. URL: https://doi.org/10.1007/s10997-020-09550-w.

[Kia+09] Aggelos Kiayias et al. "A combined fusion and data mining framework for the detection of botnets". In: *2009 Cybersecurity Applications & Technology Conference for Homeland Security*. IEEE. 2009, pp. 273–284.

[Kit04a] B. Kitchenham. “Procedures for Performing Systematic Reviews.” In: *Keele University Technical Report TR/SE-0401,ISSN:1353-7776; NICTA Technical Report 0400011T.1* (2004), pp. 1–28.

[Kit04b] Barbara Kitchenham. “Procedures for performing systematic reviews”. In: *Keele, UK, Keele University* 33.2004 (2004), pp. 1–26.

[KKR13] Slava Kisilevich, Daniel A. Keim, and Lior Rokach. “A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers’ context”. In: *Decision Support Systems* 54.2 (2013), pp. 1119–1133. DOI: 10.1016/j.dss.2012.10.038. URL: https://doi.org/10.1016/j.dss.2012.10.038.

[KM06] Lukasz A. Kurgan and Petr Musilek. “A survey of Knowledge Discovery and Data Mining process models”. In: *Knowledge Engineering Review* 21.1 (2006), pp. 1–24. DOI: 10.1017/S0269888906000737. URL: https://doi.org/10.1017/S0269888906000737.

[KM15] Amir-Mohsen Karimi-Majd and Masoud Mahootchi. “A new data mining methodology for generating new service ideas”. In: *Inf. Syst. E-Business Management* 13.3 (2015), pp. 421–443. DOI: 10.1007/s10257-014-0267-y. URL: https://doi.org/10.1007/s10257-014-0267-y.

[KMB13] Dost Muhammad Khan, Nawaz Mohamudally, and D. K. R. Babajee. “A Unified Theoretical Framework for Data Mining”. In: *Proceedings of the First International Conference on Information Technology and Quantitative Management, ITQM 2013, Dushu Lake Hotel, Sushou, China, 16-18 May, 2013*. 2013, pp. 104–113. DOI: 10.1016/j.procs.2013.05.015. URL: https://doi.org/10.1016/j.procs.2013.05.015.

[Kof14] A Kofod-Petersen. “How to do a structured literature review in computer science (version 0.2)”. In: *Copenhagen: Alexandra Institute* (2014).

[KR08] Dudyala Anil Kumar and Vadlamani Ravi. “Predicting credit card customer churn in banks using data mining”. In: *Int. J. Data Anal. Tech. Strateg.* 1.1 (2008), pp. 4–28. DOI: 10.1504/IJDATS.2008.020020. URL: https://doi.org/10.1504/IJDATS.2008.020020.

[KRG01] Ali Kamrani, Wang Rong, and Ricardo Gonzalez. “A genetic algorithm methodology for data mining and intelligent knowledge acquisition”. In: *Computers & Industrial Engineering* 40.4 (2001), pp. 361–377.

[KSB18] Sihem Khemakhem, Fatma Ben Said, and Younes Boujelbene. “Credit risk assessment for unbalanced datasets based on data

mining, artificial neural network and support vector machines". In: *Journal of Modelling in Management* (2018).

[KSC20] Eren Kurshan, Hongda Shen, and Jiahao Chen. "Towards Self–Regulating AI: Challenges and Opportunities of AI Model Governance in Financial Services". In: 2020, p. 8.

[KST18] Ajay Kumar, Ravi Shankar, and Lakshman S. Thakur. "A big data driven sustainable manufacturing framework for condition-based maintenance prediction". In: *J. Comput. Science* 27 (2018), pp. 428–439. DOI: 10.1016/j.jocs.2017.06.006. URL: https://doi.org/10.1016/j.jocs.2017.06.006.

[Küh+18] Arno Kühn et al. "Analytics Canvas–A Framework for the Design and Specification of Data Analytics Projects". In: *Procedia CIRP* 70 (2018), pp. 162–167.

[Kum+16] Ajay Kumar et al. "A big data MapReduce framework for fault diagnosis in cloud-based manufacturing". In: *International Journal of Production Research* 54.23 (2016), pp. 7060–7073.

[Kus+17] Matt J. Kusner et al. "Counterfactual Fairness". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 4066–4076. URL: https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[KV08] Boris Kovalerchuk and Evgenii Vityaev. "Symbolic methodology for numeric data mining". In: *Intelligent Data Analysis* 12.2 (2008), pp. 165–188. URL: 'http://content.iospress.com/articles/intelligent-data-analysis/ida00322'.

[LCH12] Shu-Hsien Liao, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications - A decade review from 2000 to 2011". In: *Expert Syst. Appl.* 39.12 (2012), pp. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063. URL: https://doi.org/10.1016/j.eswa.2012.02.063.

[LCR17] José Maria Luna, Cristobal Castro, and Cristóbal Romero. "MDM tool: A data mining framework integrated into Moodle". In: *Comp. Applic. in Engineering Education* 25.1 (2017), pp. 90–102. DOI: 10.1002/cae.21782. URL: https://doi.org/10.1002/cae.21782.

[LE06] Yair Levy and Timothy J. Ellis. "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research". In: *InformingSciJ* 9 (2006), pp. 181–212. URL: http://www.inform.nu/Articles/Vol9/V9p181-212Levy99.pdf.

[Lee+00] Wenke Lee et al. "A Data Mining and CIDF Based Approach for Detecting Novel and Distributed Intrusions". In: *Recent Advances*

in Intrusion Detection, Third International Workshop, RAID 2000, Toulouse, France, October 2-4, 2000, Proceedings. 2000, pp. 49–65. DOI: 10.1007/3-540-39945-3_4. URL: https://doi.org/10.1007/3-540-39945-3%5C_4.

[Lee+01] Wenke Lee et al. "Real time data mining-based intrusion detection". In: *Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01*. Vol. 1. IEEE. 2001, pp. 89–100.

[Lem16] Victoria L. Lemieux. "Innovating Good Regulatory Practice Using Mixed-Initiative Social Media Analytics and Visualization". In: *2016 Conference for E-Democracy and Open Government, CeDEM 2016, Krems, Austria, May 18-20, 2016*. 2016, pp. 207–212. DOI: 10.1109/CeDEM.2016.38. URL: https://doi.org/10.1109/CeDEM.2016.38.

[Lét+05] Sylvain Létourneau et al. "A domain independent data mining methodology for prognostics". In: *Essential technologies for successful prognostics: proceedings of the 59th Meeting of the Society for Machinery Failure Prevention Technology, Virginia Beach, Virginia, April 18-21, 2005*. 2005.

[Liu+18] Fengbo Liu et al. "Data analytics approach for train timetable performance measures using automatic train supervision data". In: *IET Intelligent Transport Systems* 12.7 (2018), pp. 568–577.

[LJ17] James Lawler and Anthony Joseph. "Big Data Analytics Methodology in the Financial Industry". In: *Information Systems Education Journal* 15.4 (2017), p. 38.

[LKL03] Younghwa Lee, Kenneth A. Kozar, and Kai R. T. Larsen. "The Technology Acceptance Model: Past, Present, and Future". In: *Commun. Assoc. Inf. Syst.* 12 (2003), p. 50. URL: http://aisel.aisnet.org/cais/vol12/iss1/50.

[LLV10] Stefan Lessmann, Mariana Listiani, and Stefan Voß. "Decision Support in Car Leasing: a Forecasting Model for Residual Value Estimation". In: *Proceedings of the International Conference on Information Systems, ICIS 2010, Saint Louis, Missouri, USA, December 12-15, 2010*. Ed. by Rajiv Sabherwal and Mary Sumner. Association for Information Systems, 2010, p. 17. URL: http://aisel.aisnet.org/icis2010%5C_submissions/17.

[LMK10a] Nhien An Le Khac, Sammer Markos, and M-Tahar Kechadi. "A data mining-based solution for detecting suspicious money laundering cases in an investment bank". In: *2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications*. IEEE. 2010, pp. 235–240.

[LMK10b] Nhien-An Le-Khac, Sammer Markos, and M. Tahar Kechadi. "A Data Mining-Based Solution for Detecting Suspicious Money Laun-

dering Cases in an Investment Bank". In: *The Second International Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA 2010, Menuires, France, 11-16 April 2010*. Ed. by Fritz Laux and Lena Strömbäck. IEEE Computer Society, 2010, pp. 235–240. DOI: 10.1109/DBKDA.2010.27. URL: https://doi.org/10.1109/DBKDA.2010.27.

[LSM99] Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok. "Mining in a Data-Flow Environment: Experience in Network Intrusion Detection". In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 15-18, 1999*. 1999, pp. 114–124. DOI: 10.1145/312129.312212. URL: https://doi.org/10.1145/312129.312212.

[LTO17] Yan Li, Manoj A. Thomas, and Kweku-Muata Osei-Bryson. "Ontology-based data mining model management for self-service knowledge discovery". In: *Inf. Syst. Frontiers* 19.4 (2017), pp. 925–943. DOI: 10.1007/s10796-016-9637-y. URL: https://doi.org/10.1007/s10796-016-9637-y.

[Lu+17] Qi Lu et al. "Research on data mining service and its application case in complex industrial process". In: *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*. IEEE. 2017, pp. 1124–1129.

[Luo14] Xingrong Luo. "Suspicious transaction detection for anti-money laundering". In: *International Journal of Security and Its Applications* 8.2 (2014), pp. 157–166.

[LWW18] Juergen Lenz, Thorsten Wuest, and Engelbert Westkämper. "Holistic approach to machine tool data analytics". In: *Journal of manufacturing systems* 48 (2018), pp. 180–191.

[LY16] Xianwei Liu and Qiang Ye. "The different impacts of news-driven and self-initiated search volume on stock prices". In: *Inf. Manag.* 53.8 (2016), pp. 997–1005. DOI: 10.1016/j.im.2016.05.009. URL: https://doi.org/10.1016/j.im.2016.05.009.

[LZX18] Raymond Yiu Keung Lau, Wenping Zhang, and Wei Xu. "Parallel aspect-oriented sentiment analysis for sales forecasting with big data". In: *Production and Operations Management* 27.10 (2018), pp. 1775–1794.

[MA16] M Moeini and SH Alizadeh. "Proposing a new model for determining the customer value using RFM model and its developments (case study on the Alborz insurance company)". In: *J. Eng. Appl. Sci* 100.4 (2016), pp. 828–836.

[Mad+20] Michael A. Madaio et al. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI". In: *CHI '20: CHI Conference on Human Factors in Computing*

Systems, Honolulu, HI, USA, April 25-30, 2020. Ed. by Regina Bernhaupt et al. ACM, 2020, pp. 1–14. DOI: 10.1145/3313831.3376445. URL: https://doi.org/10.1145/3313831.3376445.

[Mah+13] Azhar Mahmood et al. "Data Mining Techniques for Wireless Sensor Networks: A Survey". In: *IJDSN* 9 (2013). DOI: 10.1155/2013/406316. URL: https://doi.org/10.1155/2013/406316.

[Man+10] Gunjan Mansingh et al. "Application of a Data Mining Process Model: A Case Study- Profiling Internet Banking Users in Jamaica". In: *Sustainable IT Collaboration Around the Globe. 16th Americas Conference on Information Systems, AMCIS 2010, Lima, Peru, August 12-15, 2010*. Ed. by Martin Santana, Jerry N. Luftman, and Ajay S. Vinze. Association for Information Systems, 2010, p. 439. URL: http://aisel.aisnet.org/amcis2010/439.

[Mar+07a] Oscar Marban et al. "An engineering approach to data mining projects". In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2007, pp. 578–588.

[Mar+07b] Oscar Marbán et al. "An Engineering Approach to Data Mining Projects". In: *Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th International Conference, Birmingham, UK, December 16-19, 2007, Proceedings*. 2007, pp. 578–588. DOI: 10.1007/978-3-540-77226-2_59. URL: https://doi.org/10.1007/978-3-540-77226-2%5C_59.

[Mar+09] Oscar Marbán et al. "Toward data mining engineering: A software engineering approach". In: *Information systems* 34.1 (2009), pp. 87–107.

[Mar+17] Fernando Martinez-Plumed et al. "CASP-DM: Context Aware Standard Process for Data Mining". In: *CoRR* abs/1709.09003 (2017). arXiv: 1709.09003. URL: http://arxiv.org/abs/1709.09003.

[Mar+18a] M. Mariani et al. "Business intelligence and big data in hospitality and tourism: a systematic literature review." In: *International Journal of Contemporary Hospitality Management* 30.12 (2018), pp. 3514–3554. DOI: 10.1108/IJCHM-07-2017-0461. eprint: https://doi.org/10.1108/IJCHM-07-2017-0461. URL: https://doi.org/10.1108/IJCHM-07-2017-0461.

[Mar+18b] Marcello Mariani et al. "Business intelligence and big data in hospitality and tourism: a systematic literature review". In: *International Journal of Contemporary Hospitality Management* (2018).

[Mar+19] Fernando Martinez-Plumed et al. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories". In: *IEEE Transactions on Knowledge and Data Engineering* (2019).

[MAS17] Hussain Ahmad Madni, Zahid Anwar, and Munam Ali Shah. "Data mining techniques and applications - A decade review". In: *23rd*

International Conference on Automation and Computing, ICAC 2017, Huddersfield, United Kingdom, September 7-8, 2017. 2017, pp. 1–7. DOI: 10.23919/IConAC.2017.8082090. URL: https://doi.org/10.23919/IConAC.2017.8082090.

[Mat+18] Artem Mateush et al. "Building Payment Classification Models from Rules and Crowdsourced Labels: A Case Study". In: *Advanced Information Systems Engineering Workshops - CAiSE 2018 International Workshops, Tallinn, Estonia, June 11-15, 2018, Proceedings*. Ed. by Raimundas Matulevicius and Remco M. Dijkman. Vol. 316. Lecture Notes in Business Information Processing. Springer, 2018, pp. 85–97. DOI: 10.1007/978-3-319-92898-2_7. URL: https://doi.org/10.1007/978-3-319-92898-2%5C_7.

[MB17] Marko Robnik-Sikonja Marko Bohanec and Mirjana Kljajic Borstnar. "Decision-making framework with double-loop learning through interpretable black-box machine learning models". In: *Industrial Management and Data Systems* 117.7 (2017), pp. 1389–1406. DOI: 10.1108/IMDS-09-2016-0409. URL: https://doi.org/10.1108/IMDS-09-2016-0409.

[MBA17] Jacob Montiel, Albert Bifet, and Talel Abdessalem. "Predicting over-indebtedness on batch and streaming data". In: *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*. Ed. by Jian-Yun Nie et al. IEEE Computer Society, 2017, pp. 1504–1513. DOI: 10.1109/BigData.2017.8258084. URL: https://doi.org/10.1109/BigData.2017.8258084.

[Met+17] İlker Met et al. "Branch efficiency and location forecasting: application of Ziraat bank". In: *Journal of Applied Finance and Banking* 7.4 (2017), p. 1.

[MH11] Phelim Murnion and Markus Helfert. "A framework for decision support for learning management systems". In: *10th European Conference on e-Learning ECEL-2011, Brighton, UK*. 2011.

[MJ01] Steve Moyle and Alipio Jorge. "RAMSYS-A methodology for supporting rapid remote collaborative data mining projects". In: *ECML/PKDD01 Workshop: Integrating Aspects of Data Mining, Decision Support and Meta-learning (IDDM-2001)*. 2001.

[MMF10a] Gonzalo Mariscal, Oscar Marban, and Covadonga Fernandez. "A survey of data mining and knowledge discovery process models and methodologies". In: *The Knowledge Engineering Review* 25.2 (2010), pp. 137–166.

[MMF10b] Gonzalo Mariscal, Óscar Marbán, and Covadonga Fernández. "A survey of data mining and knowledge discovery process models and methodologies". In: *Knowledge Eng. Review* 25.2 (2010), pp. 137–

166. DOI: 10.1017/S0269888910000032. URL: https://doi.org/10.1017/S0269888910000032.

[MMM15] Hendrik Meth, Benjamin Müller, and Alexander Maedche. "Designing a Requirement Mining System". In: *J. Assoc. Inf. Syst.* 16.9 (2015), p. 2. URL: http://aisel.aisnet.org/jais/vol16/iss9/2.

[MMS09a] O. Marban, G. Mariscal, and J. Segovia. "A data mining and knowledge discovery process model". In: *Data Mining and Knowledge Discovery in Real Life Applications, edited by P. Julio and K. Adem, Paris, I-Tech, Vienna, Austria* (2009), pp. 438–453.

[MMS09b] Óscar Marbán, Gonzalo Mariscal, and Javier Segovia. *A data mining & knowledge discovery process model*. IntechOpen, 2009.

[Mob07] Bamshad Mobasher. "Data Mining for Web Personalization". In: *The Adaptive Web, Methods and Strategies of Web Personalization*. 2007, pp. 90–135. DOI: 10.1007/978-3-540-72079-9_3. URL: https://doi.org/10.1007/978-3-540-72079-9%5C_3.

[Mor+20] Jessica Morley et al. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices". In: *Sci. Eng. Ethics* 26.4 (2020), pp. 2141–2168. DOI: 10.1007/s11948-019-00165-5. URL: https://doi.org/10.1007/s11948-019-00165-5.

[Mor16] Vincenzo Morabito. *The future of digital business innovation: Trends and practices*. Springer, 2016.

[MRL10] Blake McNaughton, Pradeep Ray, and Lundy Lewis. "Designing an evaluation framework for IT service management". In: *Information & Management* 47.4 (2010), pp. 219–225.

[MSR14] Marjan Momtazpour, Ratnesh Sharma, and Naren Ramakrishnan. "An integrated data mining framework for analysis and prediction of battery characteristics". In: *2014 IEEE Innovative Smart Grid Technologies-Asia (ISGT ASIA)*. IEEE. 2014, pp. 774–779.

[MV17] Ricardo Mendes and João P. Vilela. "Privacy-Preserving Data Mining: Methods, Metrics, and Applications". In: *IEEE Access* 5 (2017), pp. 10562–10582. DOI: 10.1109/ACCESS.2017.2706947. URL: https://doi.org/10.1109/ACCESS.2017.2706947.

[NBI15] Mikel Nino, José Miguel Blanco, and Arantza Illarramendi. "Business understanding, challenges and issues of Big Data Analytics for the servitization of a capital equipment manufacturer". In: *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015*. 2015, pp. 1368–1377. DOI: 10.1109/BigData.2015.7363897. URL: https://doi.org/10.1109/BigData.2015.7363897.

[Net+19] Geraldo Torres G. Neto et al. “Multivocal literature reviews in software engineering: Preliminary findings from a tertiary study”. In: *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2019, Porto de Galinhas, Recife, Brazil, September 19-20, 2019*. 2019, pp. 1–6. DOI: 10.1109/ESEM.2019.8870142. URL: https://doi.org/10.1109/ESEM.2019.8870142.

[Nia15] Olegas Niaksu. “CRISP data mining methodology extension for medical domain”. In: *Baltic Journal of Modern Computing* 3.2 (2015), p. 92.

[Nie+16] Tim Niesen et al. “Towards an Integrative Big Data Analysis Framework for Data-Driven Risk Management in Industry 4.0”. In: *49th Hawaii International Conference on System Sciences, HICSS 2016, Koloa, HI, USA, January 5-8, 2016*. 2016, pp. 5065–5074. DOI: 10.1109/HICSS.2016.627. URL: https://doi.org/10.1109/HICSS.2016.627.

[NJ03] Svetlozar Nestorov and Nenad Jukic. “Ad-Hoc Association-Rule Mining within the Data Warehouse”. In: *36th Hawaii International Conference on System Sciences (HICSS-36 2003), CD-ROM / Abstracts Proceedings, January 6-9, 2003, Big Island, HI, USA*. 2003, p. 232. DOI: 10.1109/HICSS.2003.1174605. URL: https://doi.org/10.1109/HICSS.2003.1174605.

[Noh+18] Puteri Nohuddin et al. “A case study in knowledge acquisition for logistic cargo distribution data mining framework”. In: *International Journal of Advanced and Applied Sciences* 5.1 (2018), pp. 8–14.

[NSG15] Behzad Soleimani Neysiani, Nasim Soltani, and Shima Ghezelbash. “A framework for improving find best marketing targets using a hybrid genetic algorithm and neural networks”. In: *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*. IEEE. 2015, pp. 733–738.

[Nyi15] Abraham Nyirongo. *Auditing Information Systems: Enhancing Performance of the Enterprise*. Trafford Publishing, 2015.

[OEB17] Ahmed M Shahat Osman, Ahmed Elragal, and Birgitta Bergvall-Kåreborn. “Big Data Analytics and Smart Cities: A Loose or Tight Couple?” In: *10th International Conference on Connected Smart Cities 2017 (CSC 2017), Lisbon, 20-22 July 2017*. IADIS. 2017, pp. 157–168.

[OH11] Lukasz Ostrowski and Markus Helfert. “Commonality in Various Design Science Methodologies”. In: *Federated Conference on Computer Science and Information Systems - FedCSIS 2011, Szczecin, Poland, 18-21 September 2011, Proceedings*. Ed. by Maria Ganzha,

Leszek A. Maciaszek, and Marcin Paprzycki. 2011, pp. 317–320. URL: http://ieeexplore.ieee.org/document/6078172/.

[Ort+15] Joaquin Pérez Ortega et al. "A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases". In: *New Contributions in Information Systems and Technologies - Volume 1 [WorldCIST'15, Azores, Portugal, April 1-3, 2015]*. 2015, pp. 1173–1182. DOI: 10.1007/978-3-319-16486-1_116. URL: https://doi.org/10.1007/978-3-319-16486-1%5C_116.

[Ost14] Lukasz Ostrowski. "Design and evaluation of activities and reference model for the meta-design phase of design science - demonstrated on business process model artefacts". PhD thesis. Dublin City University, 2014.

[Pal+20] Maximilian Palmié et al. "The evolution of the financial technology ecosystem: an introduction and agenda for future research on disruptive innovations in ecosystems". In: (2020).

[Par+17] Grace Park et al. "A Goal-Oriented Big Data Analytics Framework for Aligning with Business". In: *Third IEEE International Conference on Big Data Computing Service and Applications, BigDataService 2017, Redwood City, CA, USA, April 6-9, 2017*. 2017, pp. 31–40. DOI: 10.1109/BigDataService.2017.29. URL: https://doi.org/10.1109/BigDataService.2017.29.

[PC13] European Parliament and Council. *Capital Requriements Derective IV (CRR IV), 2013/36/EU*. Brussels: European Parliament, 2013.

[PC16] Samira Pouyanfar and Shu-Ching Chen. "Semantic Concept Detection Using Weighted Discretization Multiple Correspondence Analysis for Disaster Information Management". In: *17th IEEE International Conference on Information Reuse and Integration, IRI 2016, Pittsburgh, PA, USA, July 28-30, 2016*. 2016, pp. 556–564. DOI: 10.1109/IRI.2016.82. URL: https://doi.org/10.1109/IRI.2016.82.

[PDM19] Veronika Plotnikova, Marlon Dumas, and Fredrik Milani. "Data Mining Methodologies in the Banking Domain: A Systematic Literature Review". In: *Perspectives in Business Informatics Research - 18th International Conference, BIR 2019, Katowice, Poland, September 23-25, 2019, Proceedings*. Ed. by Malgorzata Pankowska and Kurt Sandkuhl. Vol. 365. Lecture Notes in Business Information Processing. Springer, 2019, pp. 104–118. DOI: 10.1007/978-3-030-31143-8_8. URL: https://doi.org/10.1007/978-3-030-31143-8%5C_8.

[PDM20] Veronika Plotnikova, Marlon Dumas, and Fredrik Milani. "Adaptations of data mining methodologies: a systematic literature review".

In: *PeerJ Comput. Sci.* 6 (2020), e267. DOI: 10.7717/peerj-cs.267. URL: https://doi.org/10.7717/peerj-cs.267.

[PDM21] Veronika Plotnikova, Marlon Dumas, and Fredrik Milani. "Adapting the CRISP-DM Data Mining Process: A Case Study in the Financial Services Domain". In: *Research Challenges in Information Science - 15th International Conference, RCIS 2021, Limassol, Cyprus, May 11-14, 2021, Proceedings*. Ed. by Samira Si-Said Cherfi, Anna Perini, and Selmin Nurcan. Vol. 415. Lecture Notes in Business Information Processing. Springer, 2021, pp. 55–71. DOI: 10.1007/978-3-030-75018-3_4. URL: https://doi.org/10.1007/978-3-030-75018-3%5C_4.

[Pef+08] Ken Peffers et al. "A Design Science Research Methodology for Information Systems Research". In: *J. Manag. Inf. Syst.* 24.3 (2008), pp. 45–77. URL: http://www.jmis-web.org/articles/765.

[Pen+11] Yi Peng et al. "An empirical study of classification algorithm evaluation for financial risk prediction". In: *Appl. Soft Comput.* 11.2 (2011), pp. 2906–2915. DOI: 10.1016/j.asoc.2010.11.028. URL: https://doi.org/10.1016/j.asoc.2010.11.028.

[Pis03] Francisco Javier Martinez de Pisón Ascacibar. *Optimización mediante técnicas de mineria de datos del ciclo de recocido de una linea de galvanizado*. Universidad de La Rioja, 2003.

[Piv+13] Aleksander Pivk et al. "On approach for the implementation of data mining to business process optimisation in commercial companies". In: *Technological and economic development of economy* 19.2 (2013), pp. 237–256.

[PK03] Thomas Pyzdek and PA Keller. "The Six Sigma Handbook: A Complete Guide for Green Belts, Black Belts, and Managers at All Level". In: *New York [ua]: McGraw-Hill* (2003).

[PM15] Torsten Priebe and Stefan Markus. "Business information modeling: A methodology for data-intensive projects, data science and big data governance". In: *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015*. IEEE Computer Society, 2015, pp. 2056–2065. DOI: 10.1109/BigData.2015.7363987. URL: https://doi.org/10.1109/BigData.2015.7363987.

[Pou+16] Evangelos Pournaras et al. "Self-regulatory information sharing in participatory social sensing". In: *EPJ Data Sci.* 5.1 (2016), p. 14. DOI: 10.1140/epjds/s13688-016-0074-4. URL: https://doi.org/10.1140/epjds/s13688-016-0074-4.

[PR14] P.Bourque and R.E.Fairley. *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0*. IEEE Computer Society Press, 2014.

[Pre20] EP Press-release. *Parliament leads the way on first set of EU rules for Artificial Intelligence*. 2020. URL: https://www.europarl.europa.eu/news/en/press-room/20201016IPR89544 (visited on 10/20/2020).

[PTN18] Ken Peffers, Tuure Tuunanen, and Björn Niehaves. "Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research". In: *Eur. J. Inf. Syst.* 27.2 (2018), pp. 129–139. DOI: 10.1080/0960085X.2018.1458066. URL: https://doi.org/10.1080/0960085X.2018.1458066.

[Put+16] Deepak Puthal et al. "A Secure Big Data Stream Analytics Framework for Disaster Management on the Cloud". In: *18th IEEE International Conference on High Performance Computing and Communications; 14th IEEE International Conference on Smart City; 2nd IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2016, Sydney, Australia, December 12-14, 2016*. 2016, pp. 1218–1225. DOI: 10.1109/HPCC-SmartCity-DSS.2016.0170. URL: https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0170.

[Qin+15] Zengchang Qin et al. "Evolutionary collective behavior decomposition model for time series data mining". In: *Appl. Soft Comput.* 26 (2015), pp. 368–377. DOI: 10.1016/j.asoc.2014.09.036. URL: https://doi.org/10.1016/j.asoc.2014.09.036.

[QL16] Jianwei Qian and Rob Law. "Vincenzo Morabito: The future of digital business innovation: trends and practices". In: *J. Inf. Technol. Tour.* 16.4 (2016), pp. 459–461. DOI: 10.1007/s40558-016-0058-z. URL: https://doi.org/10.1007/s40558-016-0058-z.

[Raj15] Meera Rajan. "Credit scoring process using banking detailed data store". In: *International Journal of Applied Information Systems (IJAIS)* 8.6 (2015), pp. 13–20.

[RDW11] Fauziah Abdul Rahman, Mohammad Ishak Desa, and Antoni Wibowo. "A review of KDD-data mining framework and its application in logistics and transportation". In: *The 7th International Conference on Networked Computing and Advanced Information Management*. IEEE. 2011, pp. 175–180.

[Ren+17] Ricardo Rendall et al. "A unifying and integrated framework for feature oriented analysis of batch processes". In: *Industrial & Engineering Chemistry Research* 56.30 (2017), pp. 8590–8605.

[Res16] Marina Resta. "VaRSOM: A Tool to Monitor Markets' Stability Based on Value at Risk and Self-Organizing Maps". In: *Intell. Syst. Account. Finance Manag.* 23.1-2 (2016), pp. 47–64. DOI: 10.1002/isaf.1372. URL: https://doi.org/10.1002/isaf.1372.

[Reu+17] Thomas Reutterer et al. “A data mining framework for targeted category promotions”. In: *Journal of Business Economics* 87.3 (2017), pp. 337–358.

[Ric+20] John T. Richards et al. “A Methodology for Creating AI FactSheets”. In: *CoRR* abs/2006.13796 (2020). arXiv: 2006.13796. URL: https://arxiv.org/abs/2006.13796.

[RN04] Carol J Romanowski and Rakesh Nagi. “A data mining approach to forming generic bills of materials in support of variant design activities”. In: *Journal of Computing and Information Science in Engineering* 4.4 (2004), pp. 316–328.

[Rob02] Colin Robson. *Real world research: A resource for social scientists and practitioner-researchers*. Vol. 2. Blackwell Oxford, 2002.

[RS08] Marcel van Rooyen and Simeon J. Simoff. “A Strategic Analytics Methodology”. In: *ICSOFT 2008 - Proceedings of the Third International Conference on Software and Data Technologies, Volume ISDM/ABF, Porto, Portugal, July 5-8, 2008*. 2008, pp. 20–28.

[Run+12a] Per Runeson et al. *Case Study Research in Software Engineering - Guidelines and Examples*. Wiley, 2012. ISBN: 978-1-118-10435-4. URL: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118104358.html.

[Run+12b] Per Runeson et al. *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons, 2012.

[RV13] Cristóbal Romero and Sebastián Ventura. “Data mining in education”. In: *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 3.1 (2013), pp. 12–27. DOI: 10.1002/widm.1075. URL: https://doi.org/10.1002/widm.1075.

[Sal15] Johnny Saldana. *The coding manual for qualitative researchers*. Sage publications, 2015.

[SAS17] Inc. SAS Institute. *SAS® Enterprise Miner™ 14.3: Reference Help*. Cary, NC: SAS Institute Inc., 2017.

[Sch+20] Daniel Schiff et al. “What’s Next for AI Ethics, Policy, and Governance? A Global Overview”. In: *AIES ’20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*. Ed. by Annette N. Markham et al. ACM, 2020, pp. 153–158. DOI: 10.1145/3375627.3375804. URL: https://doi.org/10.1145/3375627.3375804.

[Seg+16] Laura Lynn Segarra et al. “A Framework for Boosting Revenue Incorporating Big Data”. In: *Journal of Innovation Management* 4.1 (2016), pp. 39–68.

[SG08] Simeon J. Simoff and John Galloway. “Visual Discovery of Network Patterns of Interaction between Attributes”. In: *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*. Springer, 2008,

pp. 172–195. DOI: 10.1007/978-3-540-71080-6_12. URL: https://doi.org/10.1007/978-3-540-71080-6%5C_12.

[Sha+10] Muhammad Shahbaz et al. “Data mining methodology in perspective of manufacturing databases”. In: *J Am Sci* (2010).

[Sil+14] Lasanthi N. C. De Silva et al. “Design science research based blended approach for usability driven requirements gathering and application development”. In: *IEEE 2nd International Workshop on Usability and Accessibility Focused Requirements Engineering, UsARE 2014, 25-25 August, 2014, Karlskrona, Sweden*. Ed. by Shah Rukh Humayoun et al. IEEE Computer Society, 2014, pp. 17–24. DOI: 10.1109/UsARE.2014.6890996. URL: https://doi.org/10.1109/UsARE.2014.6890996.

[Sin+14] Kamaldeep Singh et al. “Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests”. In: *Inf. Sci.* 278 (2014), pp. 488–497. DOI: 10.1016/j.ins.2014.03.066. URL: https://doi.org/10.1016/j.ins.2014.03.066.

[Sin+16] Surender Singh et al. “A cellular logic array based data mining framework for object detection in video surveillance system”. In: *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE. 2016, pp. 719–724.

[Sin19] Monetary Authority of Singapore. *Principles to Promote FEAT in the Use of AI and Data Analytics in Singapore’s Financial Sector*. Singapore: FEAT Committee, Monetary Authority of Singapore, 2019.

[Sin20] Monetary Authority of Singapore. *Model Artificial Intelligence Governance Framework*. Singapore: Singapore’s Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), 2020.

[SJ05] Moon Sun Shin and Kyeong Ja Jeong. “An Alert Data Mining Framework for Network-Based Intrusion Detection System”. In: *Information Security Applications, 6th International Workshop, WISA 2005, Jeju Island, Korea, August 22-24, 2005, Revised Selected Papers*. 2005, pp. 38–53. DOI: 10.1007/11604938_4. URL: https://doi.org/10.1007/11604938%5C_4.

[ŠL09] Vita Špečkauskienė and Arūnas Lukoševičius. “A data mining methodology with preprocessing steps”. In: *Information technology and control* 38.4 (2009).

[SLZ08] Zhenfeng Shao, Jun Liu, and Xianqiang Zhu. “Image Mining for Generating Ontology Databases of Geographical Entities”. In: *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (2008).

[SO08] Sumana Sharma and Kweku-Muata Osei-Bryson. "Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects". In: *41st Hawaii International International Conference on Systems Science (HICSS-41 2008), Proceedings, 7-10 January 2008, Waikoloa, Big Island, HI, USA*. 2008, p. 77. DOI: 10.1109/HICSS.2008.339. URL: https://doi.org/10.1109/HICSS.2008.339.

[SO09] Sumana Sharma and Kweku-Muata Osei-Bryson. "Framework for formal implementation of the business understanding phase of data mining projects". In: *Expert Syst. Appl.* 36.2 (2009), pp. 4114–4124. DOI: 10.1016/j.eswa.2008.03.021. URL: https://doi.org/10.1016/j.eswa.2008.03.021.

[Sol02] Jose Solarte. "A Proposed Data Mining Methodology and its Application to Industrial Engineering". PhD thesis. University of Tennessee, 2002.

[Som05] Ian Sommerville. "Integrated Requirements Engineering: A Tutorial". In: *IEEE Softw.* 22.1 (2005), pp. 16–23. DOI: 10.1109/MS.2005.13. URL: https://doi.org/10.1109/MS.2005.13.

[SP17] Swapnil Shrivastava and Supriya N. Pal. "A Big Data Analytics Framework for Enterprise Service Ecosystems in an e-Governance Scenario". In: *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2017, New Delhi, India, March 07 - 09, 2017*. 2017, pp. 5–11. DOI: 10.1145/3047273.3047274. URL: https://doi.org/10.1145/3047273.3047274.

[Spa18] S. Spalek. *Data analytics in project management*. CRC Press, 2018.

[SR13] P Sridevi and N Reddy. "Informative knowledge discovery using multiple data sources, multiple features and multiple data mining techniques". In: *IOSR Journal of Engineering* 31 (2013), pp. 20–25.

[SS16] Jeffrey S. Saltz and Ivan Shamshurin. "Big data team process methodologies: A literature review and the identification of key factors for a project's success". In: *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*. 2016, pp. 2872–2879. DOI: 10.1109/BigData.2016.7840936. URL: https://doi.org/10.1109/BigData.2016.7840936.

[Str+15] Martin Strohbach et al. "Towards a big data analytics framework for IoT and smart city applications". In: *Modeling and processing for next-generation big-data technologies*. Springer, 2015, pp. 257–282.

[Stu+20] Stefan Studer et al. "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology". In: *CoRR*

abs/2003.05155 (2020). arXiv: 2003.05155. URL: https://arxiv.org/abs/2003.05155.

[Sun+15] Jianshan Sun et al. "Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework". In: *International Journal of Production Economics* 165 (2015), pp. 185–193.

[SVL03] Sachin Singh, Pravin Vajirkar, and Yugyung Lee. "Context-Based Data Mining Using Ontologies". In: *Conceptual Modeling-ER 2003, 22nd International Conference on Conceptual Modeling, Chicago, IL, USA, October 13-16, 2003, Proceedings*. 2003, pp. 405–418. DOI: 10.1007/978-3-540-39648-2_32. URL: https://doi.org/10.1007/978-3-540-39648-2%5C_32.

[SW20] Keng Siau and Weiyu Wang. "Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI". In: *J. Database Manag.* 31.2 (2020), pp. 74–87. DOI: 10.4018/JDM.2020040105. URL: https://doi.org/10.4018/JDM.2020040105.

[SWB00a] Kate A Smith, Robert J Willis, and Malcolm Brooks. "An analysis of customer retention and insurance claim patterns using data mining: A case study". In: *Journal of the operational research society* 51.5 (2000), pp. 532–541.

[SWB00b] Kate A. Smith, Robert J. Willis, and Malcolm Brooks. "An analysis of customer retention and insurance claim patterns using data mining: a case study". In: *J. Oper. Res. Soc.* 51.5 (2000), pp. 532–541. DOI: 10.1057/palgrave.jors.2600941. URL: https://doi.org/10.1057/palgrave.jors.2600941.

[TC11] Konstantinos K Tsiptsis and Antonios Chorianopoulos. *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons, 2011.

[Tor+17] Pedro Torres et al. "Data analytics for forecasting cell congestion on LTE networks". In: *Network Traffic Measurement and Analysis Conference, TMA 2017, Dublin, Ireland, June 21-23, 2017*. 2017, pp. 1–6. DOI: 10.23919/TMA.2017.8002917. URL: https://doi.org/10.23919/TMA.2017.8002917.

[Tsa+15] Chun-Wei Tsai et al. "Big data analytics: a survey". In: *J. Big Data* 2 (2015), p. 21. DOI: 10.1186/s40537-015-0030-3. URL: https://doi.org/10.1186/s40537-015-0030-3.

[TVP17] Rita Tavares, Rui Vieira, and Luis Pedro. "A preliminary proposal of a conceptual educational data mining framework for science education: Scientific competences development and self-regulated learning". In: *2017 International Symposium on Computers in Education (SIIE)*. IEEE. 2017, pp. 1–6.

[Vak+19] Ville Vakkuri et al. “Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study”. In: *CoRR* abs/1906.07946 (2019). arXiv: 1906.07946. URL: http://arxiv.org/abs/1906.07946.

[Vak+20] Ville Vakkuri et al. “"This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments”. In: *Agile Processes in Software Engineering and Extreme Programming - 21st International Conference on Agile Software Development, XP 2020, Copenhagen, Denmark, June 8-12, 2020, Proceedings*. Ed. by Viktoria Stray et al. Vol. 383. Lecture Notes in Business Information Processing. Springer, 2020, pp. 195–210. DOI: 10.1007/978-3-030-49392-9_13. URL: https://doi.org/10.1007/978-3-030-49392-9%5C_13.

[Van+11] RJB Vanwersch et al. “Methodological support for business process redesign in health care: a literature review protocol”. In: *International Journal of Care Pathways* 15.4 (2011), pp. 119–126.

[Ven+03] Viswanath Venkatesh et al. “User Acceptance of Information Technology: Toward a Unified View”. In: *MIS Quarterly* 27.3 (2003), pp. 425–478. URL: http://misq.org/user-acceptance-of-information-technology-toward-a-unified-view.html.

[Ven06] John Venable. “The role of theory and theorising in design science research”. In: *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006)*. Citeseer. 2006, pp. 1–18.

[VKA19a] Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. “AI Ethics in Industry: A Research Framework”. In: *Proceedings of the Third Seminar on Technology Ethics, Tethics 2019, Turku, Finland, October 23-24, 2019*. Ed. by Minna M. Rantanen and Jani Koskinen. Vol. 2505. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 49–60. URL: http://ceur-ws.org/Vol-2505/paper06.pdf.

[VKA19b] Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. “Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study”. In: *Product-Focused Software Process Improvement - 20th International Conference, PROFES 2019, Barcelona, Spain, November 27-29, 2019, Proceedings*. Ed. by Xavier Franch, Tomi Mannisto, and Silverio Martinez Fernandez. Vol. 11915. Lecture Notes in Computer Science. Springer, 2019, pp. 331–338. DOI: 10.1007/978-3-030-35333-9_24. URL: https://doi.org/10.1007/978-3-030-35333-9%5C_24.

[VKA20] Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. “EC-COLA - a Method for Implementing Ethically Aligned AI Systems”.

In: *46th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2020, Portoroz, Slovenia, August 26-28, 2020*. IEEE, 2020, pp. 195–204. DOI: 10.1109/SEAA51224.2020.00043. URL: https://doi.org/10.1109/SEAA51224.2020.00043.

[VPB12] John R. Venable, Jan Pries-Heje, and Richard L. Baskerville. "A Comprehensive Framework for Evaluation in Design Science Research". In: *Design Science Research in Information Systems. Advances in Theory and Practice, 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings*. Ed. by Ken Peffers, Marcus A. Rothenberger, and William L. Kuechler Jr. Vol. 7286. Lecture Notes in Computer Science. Springer, 2012, pp. 423–438. DOI: 10.1007/978-3-642-29863-9_31. URL: https://doi.org/10.1007/978-3-642-29863-9%5C_31.

[VPB16] John R. Venable, Jan Pries-Heje, and Richard L. Baskerville. "FEDS: a Framework for Evaluation in Design Science Research Open". In: *Eur. J. Inf. Syst.* 25.1 (2016), pp. 77–89. DOI: 10.1057/ejis.2014.36. URL: https://doi.org/10.1057/ejis.2014.36.

[VPB17] John R. Venable, Jan Pries-Heje, and Richard L. Baskerville. "Choosing a Design Science Research Methodology". In: *Australasian Conference on Information Systems 2017, Hobart, Australia*. 2017.

[Wan15] Li-C. Wang. "Data mining in functional test content optimization". In: *The 20th Asia and South Pacific Design Automation Conference, ASP-DAC 2015, Chiba, Japan, January 19-22, 2015*. 2015, pp. 308–315. DOI: 10.1109/ASPDAC.2015.7059023. URL: https://doi.org/10.1109/ASPDAC.2015.7059023.

[Wan17] Li-C. Wang. "Experience of Data Analytics in EDA and Test - Principles, Promises, and Challenges". In: *IEEE Trans. on CAD of Integrated Circuits and Systems* 36.6 (2017), pp. 885–898. DOI: 10.1109/TCAD.2016.2621883. URL: https://doi.org/10.1109/TCAD.2016.2621883.

[WB13] Karl Wiegers and Joy Beatty. *Software requirements*. Pearson Education, 2013.

[WH00] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, London UK*. Springer–Verlag, 2000, pp. 29–39.

[WK16] Pornwatthana Wongchinsri and Werasak Kuratach. "A survey-data mining frameworks in credit card processing". In: *2016 13th International Conference on Electrical Engineering/Electronics, Computer,*

Telecommunications and Information Technology (ECTI-CON). IEEE. 2016, pp. 1–6.

[WLM11] Dumidu Wijayasekara, Ondrej Linda, and Milos Manic. "CAVE-SOM: Immersive visual data mining using 3D Self-Organizing Maps". In: *The 2011 International Joint Conference on Neural Networks, IJCNN 2011, San Jose, California, USA, July 31 - August 5, 2011*. 2011, pp. 2471–2478. DOI: 10.1109/IJCNN.2011.6033540. URL: https://doi.org/10.1109/IJCNN.2011.6033540.

[Woh+12] Claes Wohlin et al. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[WWX18] Yibo Wang, Mingming Wang, and Wei Xu. "A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation: A Big Data Analytics Framework". In: *Wireless Communications and Mobile Computing* 2018 (2018). DOI: 10.1155/2018/8263704. URL: https://doi.org/10.1155/2018/8263704.

[Xia09a] Liu Xiang. "Context-Aware Data Mining Methodology for Supply Chain Finance Cooperative Systems". In: *Fifth International Conference on Autonomic and Autonomous Systems, ICAS 2009, Valencia, Spain, 20-25 April 2009*. 2009, pp. 301–306. DOI: 10.1109/ICAS.2009.48. URL: https://doi.org/10.1109/ICAS.2009.48.

[Xia09b] Liu Xiang. "Context-aware data mining methodology for supply chain finance cooperative systems". In: *2009 Fifth International Conference on Autonomic and Autonomous Systems*. IEEE. 2009, pp. 301–306.

[Xia09c] Liu Xiang. "Integrating Context-Aware and Fuzzy Rule to Data Mining Model for Supply Chain Finance Cooperative Systems". In: *The Fourth International Conference on Software Engineering Advances, ICSEA 2009, 20-25 September 2009, Porto, Portugal*. 2009, pp. 471–476. DOI: 10.1109/ICSEA.2009.75. URL: https://doi.org/10.1109/ICSEA.2009.75.

[XQ08] Shuting Xu and Mable Qiu. "A privacy preserved data mining framework for customer relationship management". In: *Journal of Relationship Marketing* 7.3 (2008), pp. 309–322.

[Yan+16] Yatao Yang et al. "A Novel Hybrid Data Mining Framework for Credit Evaluation". In: *Collaborate Computing: Networking, Applications and Worksharing - 12th International Conference, CollaborateCom 2016, Beijing, China, November 10-11, 2016, Proceedings*. 2016, pp. 16–26. DOI: 10.1007/978-3-319-59288-6_2. URL: https://doi.org/10.1007/978-3-319-59288-6%5C_2.

[Yan09] Qiang Yang. "Post-processing data mining models for actionability". In: *Data mining for business applications*. Springer, 2009, pp. 11–30.

[YFH13] Zhun Yu, Benjamin CM Fung, and Fariborz Haghighat. “Extracting knowledge from building-related data—A data mining framework”. In: *Building Simulation*. Vol. 6(2). Springer. 2013, pp. 207–222.

[YH14] Bingchuan Yuan and John Herbert. “A Cloud-Based Mobile Data Analytics Framework: Case Study of Activity Recognition Using Smartphone”. In: *2nd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2014, Oxford, United Kingdom, April 8-11, 2014*. 2014, pp. 220–227. DOI: 10.1109/MobileCloud.2014.29. URL: https://doi.org/10.1109/MobileCloud.2014.29.

[YHE14] Bingchuan Yuan, John Herbert, and Yalda Emamian. “Smartphone-based Activity Recognition using Hybrid Classifier - Utilizing Cloud Infrastructure for Data Analysis”. In: *PECCS 2014 - Proceedings of the 4th International Conference on Pervasive and Embedded Computing and Communication Systems, Lisbon, Portugal, 7-9 January, 2014*. 2014, pp. 14–23. DOI: 10.5220/0004723900140023. URL: https://doi.org/10.5220/0004723900140023.

[Yin+14] Yue Ying et al. “Domain driven data mining for customer demand discovery”. In: *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE. 2014, pp. 1–6.

[Yin17] Robert K Yin. *Case study research and applications: Design and methods*. Sage publications, 2017.

[YS10] Lai Yang and Zhongzhi Shi. “An Efficient Data Mining Framework on Hadoop using Java Persistence API”. In: *10th IEEE International Conference on Computer and Information Technology, CIT 2010, Bradford, West Yorkshire, UK, June 29-July 1, 2010*. 2010, pp. 203–209. DOI: 10.1109/CIT.2010.71. URL: https://doi.org/10.1109/CIT.2010.71.

[YTX16] Wenquan Yi, Fei Teng, and Jianfeng Xu. “Noval stream data mining framework under the background of big data”. In: *Cybernetics and Information Technologies* 16.5 (2016), pp. 69–77.

[Yua+18] Hui Yuan et al. “Mining Individuals’ Behavior Patterns from Social Media for Enhancing Online Credit Scoring”. In: *22nd Pacific Asia Conference on Information Systems, PACIS 2018, Yokohama, Japan, June 26-30, 2018*. Ed. by Masaaki Hirano et al. 2018, p. 163. URL: https://aisel.aisnet.org/pacis2018/163.

[YWY14] Zhang Yun, Li Weihua, and Chen Yang. “Applying balanced scordcard strategic performance management to CRISP-DM”. In: *2014 International Conference on Information Science, Electronics and Electrical Engineering*. Vol. 3. IEEE. 2014, pp. 2009–2014.

[Zal+11] Marvin Zaluski et al. “Developing data mining-based prognostic models for cf-18 aircraft”. In: *Journal of Engineering for Gas Turbines and Power* 133.10 (2011), p. 101601.

[ZAS13] Mohamed M Zaghloul, Amr Ali-Eldin, and Mofreh Salem. “Towards a Self-service Data Analytics Framework”. In: *International Journal of Computer Applications* 80.9 (2013).

[Zha+05] Kaidi Zhao et al. “Opportunity map: a visualization framework for fast identification of actionable knowledge”. In: *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*. 2005, pp. 60–67. DOI: 10.1145/1099554.1099568. URL: https://doi.org/10.1145/1099554.1099568.

[Zha09] Zhibing Zhang. “An Efficient Neuro-Fuzzy-Genetic Data Mining Framework Based on Computational Intelligence”. In: *9th International Conference on Hybrid Intelligent Systems (HIS 2009), August 12-14, 2009, Shenyang, China*. Ed. by Ge Yu et al. IEEE Computer Society, 2009, pp. 178–183. DOI: 10.1109/HIS.2009.148. URL: https://doi.org/10.1109/HIS.2009.148.

[Zho+17] Ray Y Zhong et al. “Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors”. In: *International journal of production research* 55.9 (2017), pp. 2610–2621.

[ZLL14] Wenping Zhang, Raymond Y. K. Lau, and Chunping Li. “Adaptive Big Data Analytics for Deceptive Review Detection in Online Social Media”. In: *Proceedings of the International Conference on Information Systems - Building a Better World through Information Systems, ICIS 2014, Auckland, New Zealand, December 14-17, 2014*. 2014. URL: http://aisel.aisnet.org/icis2014/proceedings/DecisionAnalytics/5.

[ZS05] Mohammed Zaki and Tarek S. Sobh. “NCDS: data mining for discovering interesting network characteristics”. In: *Information & Software Technology* 47.3 (2005), pp. 189–198. DOI: 10.1016/j.infsof.2004.08.002. URL: https://doi.org/10.1016/j.infsof.2004.08.002.

[ZVC06] X. Zhang, D. Vogel, and Z. Chen. “Is Reward Always Effective as Incentive in Electronic Knowledge Repositories? A Game-Theoretical Perspective”. In: *Proceedings of the 12th Americas Conference on Information Systems*. Acapulco, Mexico, 2006.

Appendix A. SLR I AND SLR II STUDIES

A.1. SLR I 'Extension' and 'Integration' Studies

Table A1. 'Extension' paradigm data mining methodologies application studies for period 1997-2018

Main Adaptation Purpose	Publications
1) To implement fully scaled, integrated data mining solution	[Sun+15], [Hu+10], [Wan15], [Du+17], [Çin+11], [Dor08], [Güd+14], [SG08], [DPP11], [XQ08], [SJ05], [CBK16], [Zha09], [DD07], [Liu+18], [SLZ08]
2) To implement complex systems and integrated business applications with data mining model/solution as component or tool	[Mob07], [Sin+14], [Ala+11], [KKR13], [HSC04], [LCR17], [KMB13], [Ort+15], [LZX18], [ARA11], [Cap+17], [Kab16], [Kia+09], [KRG01], [BM98], [Sha+10], [Lee+01], [Lee+00], [Bar+01], [LWW18]
3) To implement data mining as part of integrated/combined specialized infrastructure, data environments and types (e.g. IoT, cloud, mobile networks)	[Str+15], [Mah+13], [NJ03], [GPK13], [WLM11], [YH14], [BG16], [CPT16], [Bil+14], [Ren+17], [ZLL14], [YHE14], [Hua+02], [Sin+16], [SP17], [Lem16], [Gan+96], [Tor+17], [Zho+17], [Put+16], [Kum+16]
4) To incorporate context-awareness aspects	[SVL03]

Table A2. 'Integration' paradigm data mining methodologies application studies for period 1997-2018

Main Adaptation Purpose	Publications
1) To integrate/combined with various ontologies existing in organization	[SO08], [SO09], [BC08], [Par+17], [Yin+14] [CC03]
2) To introduce context-awareness and incorporate domain knowledge	[CSZ05], [CZ08], [Xia09a], [Xia09b], [Pou+16], [CZ07], [Cao10]
3) To integrate/combine with other research/industry domains frameworks, process methodologies, and concepts	[Mar+07], [Zha+05], [Fra08], [HF09], [TVP17], [MH11], [AF17], [MMS09], [MMF10], [Sol02], [Mar+09], [CKH16], [AP15], [Ang14]
4) To integrate/combine with other organizational governance frameworks, process methodologies, concepts	[MB17], [Deb07], [CKA13], [RDW11], [YWY14], [RS08], [Küh+18], [Seg+16], [LJ17]
5) To accommodate or leverage upon newly available Big Data technologies, tools and methods	[Lu+17], [OEB17], [BKC11], [DGG09], [KM06], [ZAS13], [Nie+16]

A.2. SLR II 'Modification', 'Extension', and 'Integration' Studies

Table A3. Data Mining Methodologies in Banking - Modification' scenario example texts mapping

Business Problems	Publications
1) CRM/Customer Service	[SWB00], [KM15]
2) Data-driven decision-making	[Kad13]
3) Credit Risk, Market Risk	[MBA17], [Raj15]
4) AML	[Luo14], [Res16]

Table A4. Data Mining Methodologies in Banking - 'Extension' studies mapping

Adaptation Goals	Publications
1) To implement fully scaled, integrated data mining solutions	[Pen+11], [Cla18], [LMK10], [SR13], [Yan09], [Yua+18]
2) To implement complex systems, integrated business applications with data mining model/solution as component or tool	[BD18], [Ang+18], [DTA12]

Table A5. Data Mining Methodologies in Banking - 'Integration' studies mapping

Adaptation Goals	Publications
1) To introduce discrimination-awareness in data mining	[BP14]
2) To integrate/combine with other organizational frameworks	[Deb07]
3) To integrate/combine with other well-known frameworks, process methodologies and concepts	[Piv+13], [LLV10], [PM15], [BG11], [CZ07], [Cao10], [LTO17], [LJ17], [KV08], [Qin+15]

Appendix B. FIN-DM COMPONENTS

B.1. FIN-DM Key Enablers and Checklists

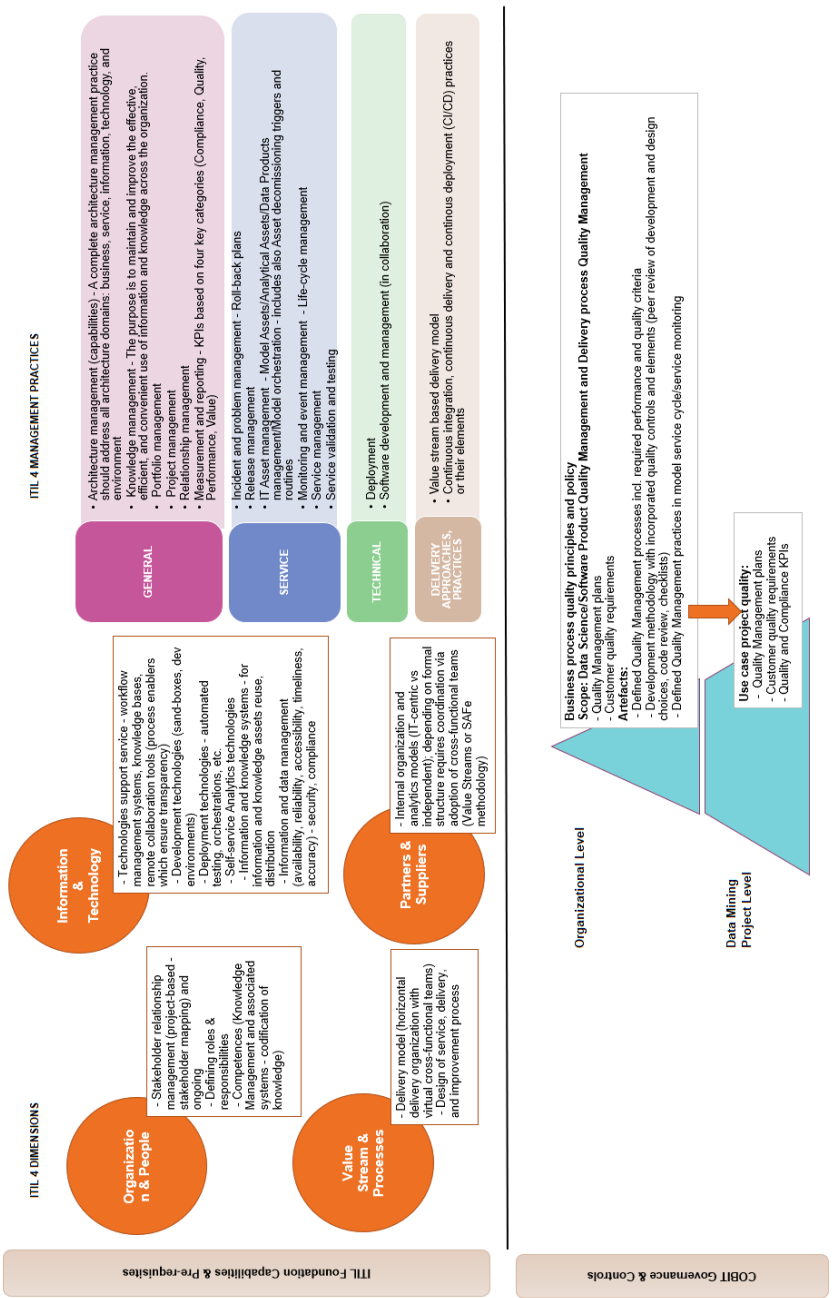


Figure B1. FIN-DM Key Enablers

REQUIREMENTS CHECKLIST				
PHASE	CHECKLIST NO.	DOMAIN (if applicable)		QUESTIONS/CHECKLIST POINTS
Requirements phase	**1**	**Business Requirements**	**1**	Business objectives, success criteria
			2	Translation into data mining objectives and success criteria
		Data	**3**	Data -List of datapoints, dataset description
			4	Data sourcing needs (systems) and timeline
		Tools	**5**	Tools (environments, software framework, programming language)
		Modelling	**6**	Modelling problem/scope, methods, techniques, reusability of existing modelling assets
			7	Analytical and model outputs requirements (related to **Actionability and Universality Requirements**):
				(1) list of analytical outputs, model(s)
				(2) model(s) interpretability and explainability requirements
				(3) analytical by-product(s):
				- descriptives, features/coefficients, results, vizualizations, interpretations, insights, etc.
		Actionability	**8**	Business usage requirements (related to **Actionability and Modelling Requirements**):
				a) define preliminary business usage scenario(s)
				b) ellicit how the solution will be embedded/integrated into business process
				c) ellicit and define required business process changes; draft their potential implementation plan
		Evaluation & Business Validation	**9**	Preliminary evaluation approach:
				a) data science evaluation (methods and metrics)
				b) business evaluation (methods and criteria)
			10	Key criteria set for selection of the best model/solution, covering business aspects (eg. Benefit, ease of use, interpretability) and technical aspects (eg. peformance, ease of implementation, infrastructure availability, etc.)
		Deployment	**11**	Preliminary consumption (based on **Actionability Requirements**) and associated deployment scenarios:
				d) performance requirements
				e) feedback loop
				f) life-cycle management requirements
				g) high-level draft of Target Architecture (with system lists)
				h) dataflow and workflow
			12	Risks, Dependencies mapped to ***Assess Situation, Project Plan***

Figure B2. FIN-DM Checklist 1

COMPLIANCE AND AI ETHICS CHECKLIST				
PHASE	**CHECKLIST NO.**	**DOMAIN (if applicable)**		**QUESTIONS/CHECKLIST POINTS**
Compliance & AI Ethics phase	**2**	**AI-input Data**	**1**	
		quality		Specify potential and existing, known data quality issues
		inherent data biases (G)		Specify potential and existing inherent data biases, most common being: (1) selection bias, (2) measurement bias
		AI Decision-making process	**2**	
		explainability (G)		Specify potential explainability scenarios and issues, methods
		transparency		List and document key AI ethics related design decisions and mitigations
		fairness (G)		Specify potential and known bias(es) caused by applying particular technique/modelling approach
				Specify fairness requirements/characteristics:
				(a) fairness conditions (relate to direct/indirect discrimination) and their testing
				(b) requirements for positive discrimination (if such exists), elimination of protected characteristics (eg. gender) or testing for them (if applicable)
				(c) potential discrimination scenarios and their testing (incl. counterfactual fairness testing)
				If applicable, calibrate fairness requirements with risk appetite/acceptable risk level of organization
		Human-centric AI solution	**3**	
		safety		Specify robustness requirements and associated testing (eg. Adversarial testing)
		neutral or positive impact on well-being		Specify monitoring, tuning (to control for concept drift) and re-training

Legend: G - term defined in Glossary

Figure B3. FIN-DM Checklist 2

MODEL LIFE-CYCLE MANAGEMENT PLAN				
PHASE	CHECKLIST NO.	DOMAIN (if applicable)		QUESTIONS/CHECKLIST POINTS
Modelling **Deployment & Implementation** **Post-Deployment**	**3**	**Model Governance**	**1**	**Specify business ownership, algorithm ownership, software solution ownership**
			2	**Define roles, responsibilities and processes**
		Model Maintenance, Service and Release Management	**3**	**Define and design monitoring process and parameters** based on 4 domains: (1) data (pipelines, data quality and reconciliations), (2) model parameters (diagnostics), (3) dependencies, (4) environment/system components
			4	**Define pontential incidents and continuity process (incl. roll-back plan)**
			5	**Specify change (new release) management incl.: (1) triggers and conditions for model rebuilds, (2) deployments/implementatios** (based on eg. CI/CD or other frameworks adopted by organization) (G)
			6	**Define model service cycle** incl. tunings, re-trainings both regular and triggered
			7	**Define final feedback loop process and associated activities**
		Model Review and Validation	**8**	**Define internal (if required external, incl. regulatory) review, verification, validation process**
		Model decomissioning	**9**	**Specify model decomissioning process incl.: (1) potential triggers/conditions for decomissioning, (2) decision-making, (3) dependencies/risks, (4) roles & responsibilities, etc.**

Figure B4. FIN-DM Checklist 3

DATA MINING, DATA SCIENCE CHECKLIST				
RELEVANT PHASE	CHECKLIST NO.	FOCUS		KEY CHECKLIST CONTROL POINTS
Data Preperation	**4**	**Data (pre)-processing**	**1**	Sampling correctness (incl. consistency across splits) and limitations
Modelling			**2**	Splits correctness, representation, concept drift
Evaluation			**3**	Consistent usage of the same data processing approach across splits
Deployment/Implementation			**4**	Data inputs, data conversions and transformations correctness
Post-Deployment/Monitoring			**5**	Data cleansing and enrichments correctness
			6	Treatment of outliers and seasonality, distiribution (eg. Skewness), anomalies
			7	Normalization and scaling - correctness and application
		Features engineering	**8**	Information leak
		Modelling	**9**	Rare target event predictability problems
			10	Overfitting
			11	Causation vs. correlation
		Modelling outcome correctness and validity	**12**	Conformance/correspondence between loss function, evaluation metric and business metric
		Code preparation for deployment	**13**	Code: (1) performing intended function, (2) conforming to agreed coding conventions, (3) modularity, (4) usage of third-party utilities and return errors, API parameters definitions correctness (if applicable) (5) latency (if applicable), (6) logging
		Reproducibility and robustness in production	**14**	Data capture, storage - source data input/output, data referential integrity, availability
			15	Correctness of source data conversions, transformations, enrichments

Figure B5. FIN-DM Checklist 4

B.2. FIN-DM Application Guidance

This is Application Guidance to FIN-DM (Financial service data mining process model) and should be used in conjunctions with FIN-DM components documentation. Application Guidance describes to potential FIN-DM users its components and guides users in applying the model.

1. FIN-DM Background

FIN-DM is an adaptation and extension of CRISP-DM (Cross-Industry Standard Process for Data Mining); it retains key CRISP-DM terminology and elements structure.

CRISP-DM is a hierarchical process model with four levels of abstraction (***general to specific***) consisting of ***phases, generic tasks, specialized tasks,*** and ***process instances*** respectively. At the top level, process is structured into six ***phases***, each phase consisting of several ***second-level generic tasks*** with respective ***outputs*** (reproduced in Figure 1 below). These two abstraction levels constitute CRISP-DM Reference Model. There are also ***third-level specialized tasks***, which are particular to data mining problem or situation or project specific as well as tool specific. They are further complemented by ***fourth-level process instance*** with account of actual activities within concrete data mining projects. CRISP-DM itself focuses on ***generic phases and tasks level (first and second)***, while detailing ***specialized level tasks and instances (third and fourth)*** are left to Reference Model users [1]. Likewise, FIN-DM covers ***phases and generic tasks*** based on CRISP-DM definitions while its application and scoping at specialized level is left to users.

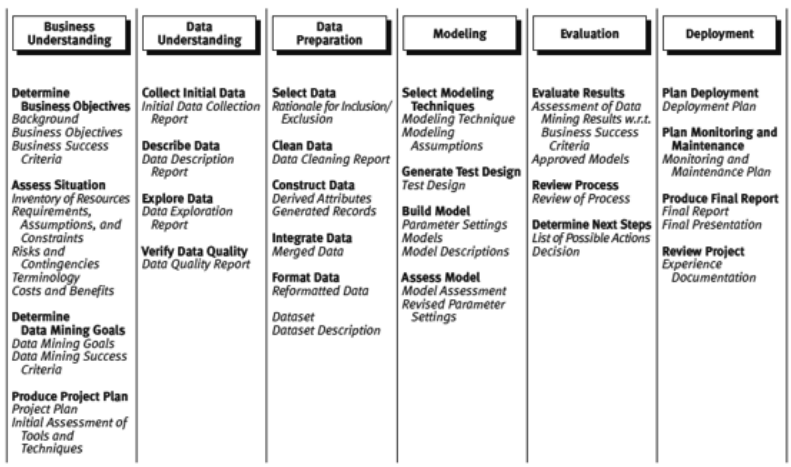


Figure 1 CRISP-DM Hierarchical View with second-level generic tasks and outputs (as in [1])

FIN-DM extends CRISP-DM at the three levels with increasing abstraction - starting from adding ***second-level tasks***, and then ***phases***. Lastly, ***frameworks*** or elements thereof originated from other domains are also introduced, they are positioned as specialized **domain extensions** relevant and applicable to whole FIN-DM life cycle, and thus, placed on the same hierarchical level as process itself.

The purpose of these extensions is to cover number of 'gaps' discovered in original CRISP-DM and listed in the *Table 1* below.

'Gap' Name	Definition
G1 - Requirements management and elicitation	Lack of tasks for validation and modification of existing requirements, elicitation of new ones
G2 - Interdependencies	Lack of iterations between different phases of CRISP-DM
G3 - Universality	Lack of support for various analytical outcomes, unsupervised and specialized techniques, deployment formats
G4 – Regulatory Compliance	Lack of tasks to address regulatory compliance (in particular, GDPR)
G5 - Validation	Lack of support for piloting models in real-life settings
G6 - Actionability	Lack of support for piloting models in real-life settings
G7 – Process	Lack of data mining process controls, quality assurance mechanisms, Critical process enablers (data, code, tools, infrastructure, and organizational factors), required for the effective execution of data mining projects are not taken into consideration

Table 1 Consolidated catalogue of CRISP-DM 'gaps' (as in [2])

2. FIN-DM Components

FIN-DM consists of 5 components Grouped into representation Layers:

1) FIN-DM ***Conceptual View*** – Layer 1
2) FIN-DM ***Hierarchical View*** and ***Key Enablers list*** – Layer 2
3) Accompanying Checklists (3 in total) – Layer 3
4) Glossary – key terms used in the FIN-DM
5) Application Guidance (this document) – description of the FIN-DM and its goals, FIN-DM intended usage and user categories

As noted, FIN-DM is based on original CRISP-DM process model for data mining projects execution and retains its elements. The new FIN-DM elements (additions or modifications) are marked in colours and constructed according to CRISP-DM original taxonomy:

- As additional ***individual generic tasks***,
- As additional ***phases*** (with specified set of ***generic tasks*** and ***outputs***),
- Additional ***frameworks*** and their elements (background note in section *5* below).

FIN-DM ***phases, tasks,*** and ***other frameworks elements*** are to be evaluated in the context of the data mining projects. Only relevant elements are to be picked up while irrelevant can be freely omitted. Also, FIN-DM is accompanied by the four checklists which can be evaluated, and relevant items used (entirely or some parts). FIN-DM allows to iterate between all phases in any sequence. Further, users are not prescribed to start with Business Understanding, but encouraged to evaluate depending on the project which phase to start with, e.g. with Data Understanding or combining Data and Business Understanding.

3. FIN-DM User Groups

Potential users are divided into the three broad user groups depending on the overall role, responsibilities, primary activities on the data mining project and RACI[1] mapping (as presented in the ***Table 2*** below).

	User Group 1	**User Group 2**	**User Group 3**	**User Group 3**
Role	(Top) Management stakeholders	Project members with business domain knowledge	Project members with project management role	Project members with development role
Profile	Functional Managers (business and tech domain)	Business users, business domain experts	Project manager(s)	Technical project delivery team - data mining experts, data /business analysts, data scientists, data engineers, software developers, etc.
Primary Activity	Overall oversight	Busines input (domain knowledge, validation, etc.)	Project management	Technical Development and Deployment end-to-end
RACI designation	I, C, R	C, R	R, A	R, A

Table 2 FIN-DM User Groups (as in [5])

Three user categories are differentiated as follows:
(1) User Group 1 - management stakeholders,
(2) User Group 2 - domain business users and experts, and
(3) Users Group 3 - technical delivery team (data mining experts, data analysts, data scientists, data engineers, etc.).

FIN-DM representation is mapped to match the respective Users Group as follows:

- Layer 1 representation - intended for all Users Groups, contains model ***Conceptual View***,
- Layer 2 representation - intended for User Group 2 and 3, in additional to ***Conceptual View*** contains model ***Hierarchical Process View*** and detailed ***Enablers*** lists,
- Layer 3 representation - primarily intended for User Group 3, contains all supplementary checklists (4) in addition to ***Conceptual View*** and ***Hierarchical Process View***
- Application Guidance and Glossary are intended for User Group 3 as reference material. They will support the data mining project manager in navigating the framework. Occasionally, it could serve as support for data mining project development team (especially, data scientists) as regards common terminology

FIN-DM Layer 1 is applicable and relevant for all user groups. It provides 'helicopter' view and assists with hands-on understanding on key phases of any data mining project. Also, it provides concise view on key pre-requisites required for such project's execution. Especially, it would assist project managers in explaining data mining project(s) execution to the strategic leaders and functional managers. For the latter, it gives concise overview of the key pre-requisites/enabler required to manage and execute portfolio of data mining projects and initiatives and perform data mining at scale.

FIN-DM ***Hierarchical View*** (Layer 2) is intended for User Group 2 and 3. It equips data mining project participants with the detailed understanding of each data mining project phase, the sequence of activities within each phases and outputs. Moreover, such process view also would assists project managers to explain business and domain experts' key activities where their engagement will be required.

[1] RACI – Responsible, Accountable, Consulted, Informed [5]

Lastly, FIN-DM Layer 3 intended for User Group 3 provides hands-on support in concrete tasks execution based on checklists. It will be of use for project managers and team members providing useful input to execute the project based on structured approach, keep track of its progress, scope and deliverables (manage backlog of the tasks effectively).

Example: determine the most applicable FIN-DM representation User might migrate and/or be involved in more than one primary group and assume additional activities beyond primary. For instance, business users (User Group 2) might get involved into testing activities and be closely related to technical development activities at certain stages of data mining project, thus, being involved into User Group 3 too. However, they are likely not to be required to use FIN-DM most specific Layer 3 checklists in these activities. Therefore, in this example, Layer 2 representation will remain most suitable and adequate for User Group 3 given their primary role.

4. Note on Additional Frameworks in FIN-DM

FIN-DM combines elements of ITIL[2] and COBIT[3] frameworks and they are present at the highest abstraction level in the ***Conceptual View*** (Layer 1). Further, the ***key enablers*** list is derived based on relevant element of these frameworks and is presented in conjunction with ***Hierarchical View*** (Layer 2). Apart from definitions in Glossary, we provide background and context of their usage below.

ITIL and COBIT Background

ITIL framework covers the whole life cycle of IT services [7], and focuses on defining comprehensive set of best practice processes for IT service management and support [6]. The key components of ITIL 4 framework are the *ITIL Service Value System (SVS)* and the *four-dimensional model* (as presented in reference box to the right). In ITIL paradigm, IT service management functions best when organized as a system. Hence, ITIL SVS describes the inputs to this system, its elements, the outputs, and details how the various components and activities of the organization work together to facilitate value creation through IT-enabled services. To support a holistic approach to service management, ITIL defines *four dimensions* that collectively are critical to the effective and

ITIL SERVICE VALUE SYSTEM (SVS)

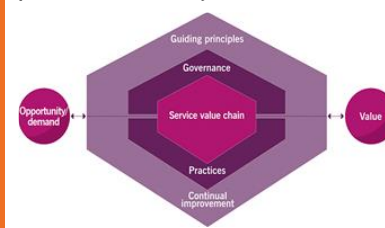
Guiding principles - recommendations that can guide an organization in all circumstances, regardless of changes in its goals, strategies, type of work, or management structure.

Governance - the means by which an organization is directed and controlled.

Service value chain - A set of interconnected activities (operating model) that an organization performs to deliver a valuable product or **service** to its consumers and to facilitate value realization.

Practices - Sets of organizational resources designed for performing work or accomplishing an objective. The ITIL SVS includes 14 general management practices, 17 service management practices, and three technical management practices.

Continual improvement - A recurring organizational activity performed at all levels to ensure that an organization's **performance** continually meets stakeholders' expectations.



THE FOUR-DIMENSIONAL MODEL

- Organizations and people
- information and technology
- partners and suppliers
- value streams and processes.

Source: *AXELOS (2020) ITIL ®: Foundation ITIL 4 Edition. TSO, London.*

[2] Information Technology Infrastructure Library, hereinafter, we refer to the latest version - ITIL 4
[3] Control Objectives for Information Technologies

efficient value delivery. They represent perspectives which are relevant to the whole SVS ([8]-[9]).

COBIT is a framework for the governance and management of enterprise information and technology, aimed at the whole enterprise. This is achieved by COBIT [10]:

COBIT 2019 CORE MODEL AND KEY COMPONENTS

Governance objectives – are grouped into EDM domain (Evaluate, Direct and Monitor). In this domain, the governing body evaluates strategic options, directs senior management on the chosen strategic options and monitors the achievement of the strategy.

Management objectives are grouped in four domains:

Align, Plan and Organize (APO) - addresses the overall organization, strategy and supporting activities for I&T.

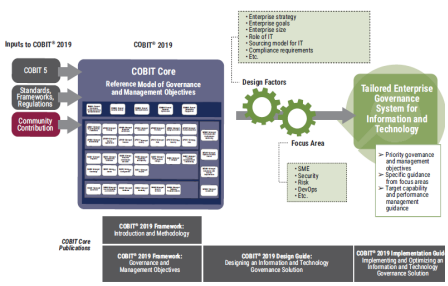
Build, Acquire and Implement (BAI) - treats the definition, acquisition and implementation of I&T solutions and their integration in business processes.

Deliver, Service and Support (DSS) - addresses the operational delivery and support of I&T services, including security.

Monitor, Evaluate and Assess (MEA) - addresses performance monitoring and conformance of I&T with internal performance targets, internal control objectives and external requirements.

Components of governance system – processes, organizational structures, principles, policies, and frameworks, information, culture, ethics and behavior, people, skills and competencies, services, infrastructures, and applications.

Design factors – enterprise strategy and goals, risk profile I&T-related issues, threat landscape, compliance requirements, role of IT, sourcing model for IT, IT Implementation Methods, Technology Adoption Strategy, enterprise size.



COBIT Overview

Source: ISACA, COBIT 2019: Introduction and Methodology. Schaumburg: ISACA, 2018.

(1) Defining the components to build and sustain a governance system,
(2) Defining the design factors that should be considered to build a best-fit governance system, and
(3) Grouping relevant components into governance and management objectives.

COBIT is based on a set of key governance principles translated into ***COBIT Core Model***. It consists of 40 high-level ***governance and control objectives*** grouped into five domains. To satisfy governance and management objectives, each enterprise needs to establish, tailor, and sustain a ***governance system*** built from several components proposed by COBIT. It is complemented by ***design factors*** which can influence the governance framework. Finally, ***goals cascade*** supports alignment and prioritization between enterprise goals and management goals. Also, COBIT includes guidance on performance management with focus on improving capabilities and attaining higher maturity levels.

COBIT is a comprehensive control and management framework aimed to ensure holistic IT governance and management throughout organization. COBIT does not include process steps and tasks. Due to such broad scope, COBIT is referred as 'integrator' establishing link between various IT practices and business requirements [6]. COBIT operates from viewpoint of entire enterprise while ITIL focuses entirely on IT and associated service management practices. ITIL can be adapted and used in conjunction with other practices and extensively with COBIT, and both practices are viewed as complementary [7].

Using Key Enablers List derived from ITIL and COBIT frameworks

Number of Foundational capabilities and prerequisites are selected from the ITIL framework as part of ***Key Enablers*** list. ITIL pre-requisites are complemented by COBIT elements related to quality management. The suggested ITIL-derived enablers and COBIT elements are intended to address CRISP-DM process 'gaps' (G7 in *Table 1*).

Based on ITIL typology, the following attributes across four ITIL dimensions are presented:

- Organization and People - stakeholders management and data mining competencies (via knowledge management and codification systems)
- Information and Technology - data management, model development, deployment and self-service technologies encompassing the whole data mining life cycle including sharing the results with users
- Partners and Suppliers - planning of resources and internal coordination
- Value Stream and Processes - delivery models (horizontal vs. virtual cross-functional teams) and the design of service, delivery, and improvement processes

The relevant dimensions and attributes are complemented by sets of selected ITIL Management practices covering general management (portfolio, project, relationships, etc.), service (primarily related to implemented data mining models service management), technical (deployment and software development practices), and delivery approaches (including delivery models, and continuous integration, delivery and deployment practices).

COBIT elements refer to quality management at overall organizational level (data mining process quality principles, policy, and quality management practices of data mining models), and individual data mining projects level, i.e. quality management plans for respective data mining projects based on stakeholders' quality requirements.

The ***Key Enablers*** is especially useful to be considered in the context of establishing industrialized data mining practices at scale and developing portfolio of data mining services and projects. The list is also beneficial to evaluate in the context of individual data mining projects. As noted earlier, only relevant elements are to be picked up while irrelevant can be freely omitted.

> *Example: considering Key Enablers in the context of data mining model to be integrated into customer-facing application* In case of such projects, availability of model results on continuous basis to customers becomes of critical importance. That implies that in case of model failure, fallback process needs to be followed. In this context, by reviewing *Key enablers, incidence management practices* are the most relevant and need to be picked-up. By adopting such practices from the rest of organization towards the given model or establishing such practices anew, organization will ensure required model service level consistently.

For the users' guidance and convenience, ITIL and COBIT elements are also mapped and inherently integrated where applicable into the FIN-DM hierarchical process via respective sub-components and tasks. For COBIT elements, these are data mining project quality management tasks reflected in the ***Risk Management and Quality Assurance sub-component*** of the ***Compliance phase***. For ITIL elements, these are data mining requirements reflected in the ***Requirements phase.***

5. Note on Project Management Aspects

FIN-DM primary focus is adapting data mining process towards needs of financial services industry. Therefore, some aspects are left for the users to specify more in the context of existing governance and management practices and frameworks adopted in their specific organizations. One of such aspects is project management where FIN-DM is less centred

beyond Requirements elicitation and management. Typically, project management practices are governed within organization and guided by respective paradigms (e.g. Agile, SAFE, etc.). Therefore, we propose for users to rely on them in conducting data mining projects. Such practices typically specify well the division of responsibilities between various groups of stakeholders in organization when projects are conducted.

At the same time, we include defining and sign-off of Roles and Responsibilities into the Business Understanding phase. Furthermore, we recommend relying on RACI framework (mentioned above) with the tentative categorization of project participants as in BABOK[4] Guide [13] as:

- **Responsible (R)** – person(s) who will be performing the work on the task,
- **Accountable (A)** – person(s) who is ultimately held accountable for successful completion of the task and is the decision maker,
- **Consulted (C)** - the stakeholder or stakeholder group who will be asked to provide an opinion or information about the task (often the subject matter experts (SMEs) are falling into these category),
- **Informed (I)** - a stakeholder or stakeholder group that is kept up to date on the task and notified of its outcome. In case of Informed the communication is one-direction (business analyst to stakeholder) and with Consulted the communication is two-way.

Roles and Responsibilities (as part of Checklist 1) can be detailed for each phase of the data mining projects, emphasizing type of involvement in key process phases of the respective specialists based on their core competences and functions (as in Example).

> *Example:* dedicated SMEs and End Users can be heavily involved in formulating, review and approval of initial data mining requirements. They also are required to be involved and perform validation and business testing of the final data mining model or outcome, as well as validate significant intermediate results of the project from business point of view. Data Scientists and Software/AI Engineers are responsible for technical development and implementation, but with different participation degree: in key FIN-DM phases of Data Understanding and Preparation, Modelling, Testing, Data Scientists lead and perform the work, while in Implementation phase, AI/ML Engineers are responsible to execute key software development and integration associated tasks.

Lastly, we also recommend for the users to consult widely accepted standards and best practice guides in project management field. The closest related, but not exhaustive list of such practice guides as points of initial reference could be PMBOK[5] body of knowledge [11], PRINCE 2[6] methodology [12], BABOK body of knowledge [13], and similar.

6. FIN-DM Application Principles and Recommendations

General principles

Adaptable and Extendable FIN-DM is highly adaptable – users are encouraged to contextualize process model to specifics of their concrete data mining project and omit any elements which are not applicable or relevant. As well, any modifications (including extensions) can be introduced by users too.

Unlimited iterations and accounting for interdependencies FIN-DM promotes and allows for any number of iterations across all phases and elements as deemed necessary and beneficial by data mining project management and project experts team. Further, users are

[4] BABOK – the globally recognized standard for the practice of business analysis
[5] PMBOK – project management body of knowledge generally recognized as best practice [11], widely accepted and used across the world
[6] PRINCE 2 – widely considered as the leading project management method [12]

free and encouraged when using ***Hierarchical Process Model***, to merge, closely integrate or parallelize phases at their discretion. Based on evaluation with users, we also provided indications of potential parallelization/merger of phases on the diagram. FIN-DM also supports discovering, and ongoing tracking and calibration of interdependencies throughout the whole data mining life cycle [2].

Compatible, Easy to integrate with other frameworks, practices and methods Given FIN-DM adaptiveness, flexibility and its concentration on data mining life-cycle, it is fully compatible and can be easily embedded, integrated with other popular product development and delivery frameworks, practices, methods and patterns used across different organizations. For example, it can be fit into Adaptive Software Development/XP programming with focus on software development, or alternatively Agile, Scaled Agile (SAFE), Scrum and its variations focusing on comprehensive flow of work management for complex systems, products and projects, etc.

Dynamic FIN-DM follows open architecture principles and is not prescriptive. Therefore, adding, changing, and modifying existing complements and elements in response to external disruptions, technological changes and emerging organizational needs can be implemented by users without impacting framework structure, other components, and content. FIN-DM is referenced and supports periodic realignment with other IT frameworks (ITIL, COBIT) which in turn are upgraded in due course and reflect latest organizational and technological developments in IT/technology domain.

Specific application recommendations

Practical adoption To practically adapt FIN-DM, we suggest considering two general application patterns – 'light-weight' mode and 'full' mode. This is especially relevant for two elements – Requirements phase and AI & Ethics subcomponent. Suggested application modes primarily correlate with size and complexity of company operations and its business model. Also, complexity of the data mining project, type of data mining problem and relevance of AI ethics and compliance concerns are key drivers.

For instance, if the company operates in limited geographical scale, is of small size in industry terms and is not delivering full universal banking services, the data mining project scope and data used will be naturally constrained. Therefore, many of Requirements can be simplified and/or omitted which also decreases number of the relevant tasks as well as necessity for many Requirements management and iterations activities.

In the similar way, if due to data mining use case context and data used, there are no triggers for AI ethics and compliance concerns, AI Ethics tasks and Checklist 2 can be approached in 'light-weight' mode, i.e. it is sufficient to use them as checkpoints in the beginning and end of the data mining projects.

Tools support Assigning and tracking FIN-DM phases, tasks, and checkpoints progress is best realized with support of respective technologies, such as collaborative issue tracking, project management, and workflow tools.

References

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. 2000. CRISP-DM 1.0 Step-by-step data mining guide, SPSS Inc.

[2] Plotnikova V., Dumas M., Nolte A., Milani F. 'Designing a Data Mining Process for the Financial Services Domain' (submitted).

[3] Wirth, R. and J. Hipp (2000). "CRISP-DM: Towards a standard process model for data mining." In: In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, London UK. Springer–Verlag, pp. 29–39.

[4] Wiegers, Karl. "More about software requirements: thorny issues and practical advice". Microsoft Press, 2005.

[5] Spalek, S. (2018). Data analytics in project management. CRC Press.

[6] Guldentops, E., Hardy, G., Heschl, J.G. and Taylor, S., 2010. Aligning Cobit, ITIL and ISO 17799 for Business Benefit: A Management Briefing from ITGI und OGC.

[7] Almeida, Rafael, Paloma Andrade Gonçalves, Inês Percheiro, Miguel Mira da Silva, and César Pardo. "Integrating COBIT 5 PAM and TIPA for ITIL Using an Ontology Matching System." International Journal of Human Capital and Information Technology Professionals (IJHCITP) 11, no. 3 (2020): 74-93.

[8] AXELOS (2020), ITIL® 4: Create, Deliver and Support. TSO, London.

[9] AXELOS (2020) ITIL ®: Foundation ITIL 4 Edition. TSO, London.

[10] ISACA, COBIT 2019: Introduction and Methodology. Schaumburg: ISACA, 2018.

[11] Project Management Institute (2017). A guide to the project management body of knowledge: PMBOK® GUIDE, Sixth edition, Newtown Square, Pennsylvania, US.

[12] AXELOS (2017), Managing Successful Projects with PRINCE2®. TSO, London.

[13] International Institute of Business Analysis (2015): BABOK®, version 3, A GUIDE TO THE BUSINESS ANALYSIS BODY OF KNOWLEDGE®, Toronto, Ontario, Canada.

ACKNOWLEDGEMENTS

I want to express immense gratitude to my supervisors, Prof. Marlon Dumas and Assoc. Prof. Fredrik P. Milani. Thanks for taking the risk to accept me as an industrial PhD student. Sincere thanks for your time, consistent support, and invaluable efforts in driving and directing my PhD journey towards successful completion. I am also very grateful to opponents and reviewers of the Thesis and my other research publications for valuable advice and feedback, which allowed to improve my work.

I am very thankful to Robert Kitt, who, back in the beginning of 2017, promoted and highlighted to me the idea and value of PhD studies. This thinking has been inspirational and instrumental in bringing me into the PhD path, and served as a catalyst to a new phase in my professional and personal development, so thank You, Robi!

I appreciate constant inspiration and support from my colleagues and team members at Swedbank. Our joint reflections and discussions provided invaluable input to develop and frame the ideas presented in this Thesis, too. It has been an absolute pleasure and honor to cooperate with all of You.

Last but not least, I owe a huge debt of gratitude to my family and closest friends. Without Your support and advice, the journey would be much more challenging to pursue, let alone complete. Sincere and special thanks to all of You!

SISUKOKKUVÕTE

FIN-DM: finantsteenuste andmekaeve protsessi mudel

Andmekaeve ja arendatud analüütika kasutamine otsustamise hõlbustamiseks on viimaste aastakümnete jooksul märgatavalt kasvanud. Kompleksete suurandmetete, andmekaeve ja arendatud analüütika projektide välja töötamine ja rakendamine vajab hästidefineeritud metodoloogiat ja protsesse. On välja käidud hulganisti juhendeid ja standardseid protsessimudeleid – CRISP-DM, KDD ja SEMMA[1] –, millede eesmärgiks on andmekaeve projektide läbiviimine ja juhtimine; ning nende kasutamine on märkimisväärselt kasvanud. Standardsete andmekaeve lähenemiste hulgas on kõige tähelepanuväärsem ja laiemat kasutust leidev CRISP-DM. See on tööstusharust sõltumatu, ja tihti kasutusele võtmisel kohandatud, eesmärgiga vastata sektorispetsiifilistele nõudmistele. CRISP-DM'i tööstusharuspetsiifilisi kohandusi on välja pakutud erinevates valdkondades, sealhulgas tervishoius, hariduses ning ka tööstustehnoloogia, tarkvaraarenduse, logistika jm. aladel. Kuid, nii palju kui meile teada on, puudub siiani CRISP-DMi kohandus finantsteenuste tööstusharus rakendamist leidvate andmekaeveprojektide suunamiseks ja struktureerimiseks – seda sektoris, millel on oma erialaspetsiifiliste nõuete kogumik. Käesoleva doktoritöö eesmärgiks on täita see tühimik, disainides, arendades ja evalveerides uue artefakti – finantsteenuste sektorispetsiifilise andmekaeve protsessi mudeli (FIN-DM).

Käesolevas uurimistöös oleme kasutusele võtnud Design Science Research Methodology (DSRM) kombineerituna lehter-lähenemisega (funnel approach). Töö saab alguse uurimusega, kuidas standardseid andmekaeve protsessi mudeleid kasutatakse erinevates tööstussektorites ning ka finantsteenuste sektoris – "makroperspektiivist", töötades läbi laia valiku akadeemilist ja erialakirjandust. Me oleme leidnud ja kirjeldanud hulga kohandamismustreid, mida saab rakendada standardsele andmekaeveprotsessile. Töö käigus oleme jõudnud tõdemuseni, et varasemad lähenemised ei pööra piisavalt tähelepanu kasutuselevõtuprobleemidele, mis mängivad andmekaevemudelite IT arhitektuuri ning organisatsioonide äriprotsessidesse integreeritud tarkvaratoodeteks arendamise juures tähtsat rolli. Kokkuvõtvalt nendime, et olemasolevate juhendite edasiarendused, millede otstarbeks oleks kombineerida andmeid, ning tehnoloogilisi ja organisatsioonilisi aspekte, võivad praegusi puudujääke vähendada. Seda teemat käsitledes tuleb nentida, et finantsteenuste sektoris on põhilisteks leitud kohandamisstsenaariumideks need, mis puudutavad tehnoloogiakeskseid (skaleeritavus), ärikeskseid (ellu rakendatavus) ja inimesekeseid (diskrimineerimisefektide kõrvaldamine) andmekaeve aspekte. Järgnevalt lähenesime probleemile "mikroperspektiivilt", viies läbi kaasusuuringu reaalselt eksisteerivas finantsteenuste organisatsioonis, uurimaks, kuidas standardiseeritud andmekaeve protsessid on praktikas ellu rakendatud. Tulemustena tõime

[1] CRISP-DM - Cross-industry standard process for data mining, KDD - Knowledge Discovery in Databases, SEMMA - Sample, Explore, Modify, Model, Asses

esile 18 leitud CRISM-DM'is esinevat puudujääki, koos nende järelmitega, ning praktikute poolt rakendatavad mehhanismid, selleks et puudujääkide mõjuga toime tulla.

Viimase sammuna kasutasime andmeid ja tulemusi "makro" ja "mikro" uuringutest, ning disainisime, arendasime ning evalveerisime kasutajatega the Financial Industry Process for Data Mining'u (FIN-DM). FIN-DM on kohandatud ja laiendatud CRISP-DM, toetamaks privaatsust võimaldavat andmekaevet, tegelemaks AI eetika riskidega, täitmaks mudeli riskijuhtimise nõudeid ja pannes kvaliteedi tagamise andmekaeve elutsüklis kesksele kohale. Raamistiku kasulikkus on saanud kinnitust tihedas koostöös andmekaeve ja IT praktikutega reaalselt eksisteerivas finantsteenuste organisatsioonis. Seeläbi toob FIN-DM välja tulemusi kahes aspektis: (1) toetades praktikuid, kes töötavad andmekaeve projektidega finantsteenuste sektoris; (2) andes lahendusi laiema IT-, andmekaeve- ja ärivaldkondade integratsiooni teemadevaldkonna raames. Veel tuleb öelda, et FIN-DM'ga on seotud mitmed eelised, aga ka puudused. Seda võib kasutada kui plaani või arengukava skaleerimisel ning tööstusliku andmekaeve funktsioonide loomisel; olles samas rohkem efektiivne organisatsioonides kus on juba loodud IT aluslahendused ning kasutusel standardsed IT tööstuse raamistikud (n. ITIL). Lisaks, tuleks kasuks CRISP-DM'i alane teadmine ning varasem töökogemus, loomaks võimaluse efektiivsemale organisatsioonisisesele kasutuselevõtule. Mõned välja toodud lahendused leitud puudujääkidele ei pruugi olla üksnes finantssektorispetsiifilised, vaid võivad olla rakendatavad teiste, sarnaste probleemidega tegelevate, sektorite kontekstis (n. telekom). Heaks näiteks sellistest üldistest lahendustest on *privaatsuse* ja *AI eetika* komponendid. Seega, on võimalik ka FIN-DM'i ja tema elementide laiem rakendamine – näiteks teistes sektorites, nagu telekom, ning organisatsioonides.

CURRICULUM VITAE

Personal data

Name, Surname:	Veronika Plotnikova
Date of Birth:	27 September 1979
Contact e-mail 1:	veronika.plotnikova@ut.ee
Contact e-mail 2:	veronika.plotnikova@hotmail.com

Education

2017–2021	University of Tartu (Estonia), PhD Computer Science
2001–2003	Stockholm School of Economics (SSE, Sweden), M.Sc. Finance *Alice och Knut Wallenberg scholarship for Master studies at SSE* *Staffan Burinstam Linder scholarship for studies in Sweden*
1998–2001	Stockholm School of Economics in Riga (SSE Riga, Latvia), B.Sc. Economics and Business Administration, graduated with the highest average on the course *PriceWaterhouse academic performance scholarship* *Rietumu Banka academic performance scholarship*

Certifications, Professional Memberships

2021	Certified SAFe 5 Agilist, SAFe® (Scaled Agile Framework)
2021	Certified Celonis Process Mining Expert, Celonis
2017–2021	University of Tartu/IIBA (International Institute of Business Analysis) – Certification in Digital Business Analysis
2020–now	ISACA Member, Latvian Chapter

Employment

2021–now	Head of AML Engineering Guild, Anti-Financial Crime BIO, Swedbank Group
2017–2021	Head of Data Science Guild 2, Head of Delivery and Service Division, Analytics & AI, Swedbank Group
2016–2017	Head of Data Analytics Department in Retail Segment, Swedbank Baltic Banking
2013–2016	Head of Analysis and Planning Department, Swedbank Latvia
2008–2013	Head of Controlling and Planning Division, DNB Latvia
2006–2008	Head of Analytics Unit in International Relations, Investment Banking and Capital Markets, Parex Bank
2003–2006	Analyst/Head of Middle Office Unit, Treasury and Capital Markets, SEB Latvia
2001–2003	positions in Assurance and Financial Advisory Services, KPMG (Latvian office)

Scientific work

Main fields of interest:

- data mining and advanced analytics life-cycle - methods, practices and delivery frameworks
- IS, IT and software development, delivery life-cycle and management - methods, practices and frameworks
- organizational adoption and practices of data mining, advanced analytics and IT development, delivery and management
- design science

ELULOOKIRJELDUS

Isikuandmed

Ees- ja perekonnanimi:	Veronika Plotnikova
Sünnikuupäev:	27. september 1979
E-post 1:	veronika.plotnikova@ut.ee
E-post 2:	veronika.plotnikova@hotmail.com

Haridus

2017–2021	Tartu Ülikool, informaatika doktorikraad
2001–2003	Stockholmi Kõrgem Majanduskool (SSE, Rootsi), rahanduse magistrikraad
1998–2001	Stockholmi Kõrgem Majanduskool Riias (SSE Riga, Läti), majanduse ja ärijuhtimise bakalaureusekraad, lõpetamisel kursuse kõrgeim keskmine hinne

Sertifikaadid, kutseorganisatsioonide liikmesus, keeleoskus

2021	SAFe 5 Agilist, SAFe® (Scaled Agile Framework) sertifikaat
2021	Celonis Process Mining Expert, Celonise sertifikaat
2017–2021	Tartu Ülikool, rahvusvahelise ärianalüüsi instituudi IIBA digitaalse ärianalüüsi sertifikaat
2020–now	Infosüsteemide Auditi ja Juhtimise Assotsiatsiooni (ISACA) liige, Läti osakond
Keeleoskus:	Inglise (ladus), Läti (ladus), Vene (emakeel)

Teenistuskäik

2021—praeguseni	rahapesu tõkestamise juht (AML Engineering Guild), finantskuritegude äriteabeametnik, Swedbank Group
2017–2021	andmeteaduse juht (Data Science Guild 2), andmeanalüütika ja tehisintellekti tarne- ja teenuseosakonna juht, Swedbank Group
2016–2017	jaemüügi valdkonna andmeanalüütika osakonna juht, Swedbanki Balti pangandus
2013–2016	analüüsi- ja planeerimisosakonna juht, Swedbank Läti
2008–2013	kontrolli- ja planeerimisosakonna juht, DNB Pank Läti
2006–2008	rahvusvaheliste suhete, investeerimispanganduse ja kapitaliturgude analüütikaüksuse juht, Parex Bank
2003–2006	raha- ja kapitaliturgude valdkonna keskkontori analüütik/juht, SEB Läti
2001–2003	ametikohad kindlustus- ja finantsnõustamisteenuste valdkonnas, KPMG (Läti haru)

Teadustegevus

Peamised uurimisvaldkonnad:

- andmekaeve ja arenenud analüütika elutsükkel ning nende meetodid, kasutusalad ja raamistikud;
- nfosüsteemid, infotehnoloogia ja tarkvaraarendus, tarne elutsükkel ja haldus ning nende meetodid, kasutusalad ja raamistikud;
- andmekaeve tavad ja organisatsioonis kohandamine, arenenud analüütika ja infotehnoloogia teenuste arendus, tarne ja haldus;
- disainiteadus.

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I Veronika Plotnikova, Marlon Dumas, and Fredrik Milani. "Data Mining Methodologies in the Banking Domain: A Systematic Literature Review". In:*Perspectives in Business Informatics Research -18th International Conference, BIR 2019, Katowice, Poland, September 23-25, 2019, Proceedings.* Ed. by Malgorzata Pankowska and Kurt Sandkuhl. Vol. 365. Lecture Notes in Business Information Processing. Springer, 2019, pp. 104–118.
Author contributions: Lead author. Conceived the idea, contributed to the study design, conducted data collection and analysis, and wrote the manuscript with support of the other authors.

II Veronika Plotnikova, Marlon Dumas, and Fredrik Milani. "Adaptations of data mining methodologies: a systematic literature review". In: *PeerJ Computer Science*, 6 (2020), e267.
Author contributions: Lead author. Conceived the idea, contributed to the study design, conducted data collection and analysis, and wrote the manuscript with support of the other authors.

III Veronika Plotnikova, Marlon Dumas, and Fredrik Milani. "Adapting the CRISP-DM Data Mining Process: A Case Study in the Financial Services Domain". In: *Research Challenges in Information Science - 15th International Conference, RCIS 2021, Limassol, Cyprus, May 11-14, 2021, Proceedings.* Ed. by Samira Si-Said Cherfi, Anna Perini, and Selmin Nurcan. Vol. 415. Lecture Notes in Business Information Processing. Springer, 2021, pp. 55–71.
Author contributions: Lead author. Conceived the idea, contributed to the study design, conducted data collection and analysis, and wrote the manuscript with support of the other authors.

Publications not included in the thesis

IV Mateush, Artem, Rajesh Sharma, Marlon Dumas, Veronika Plotnikova, Ivan Slobozhan, and Jaan Übi. "Building payment classification models from rules and crowdsourced labels: a case study." In: *International Conference on Advanced Information Systems Engineering*, pp. 85-97. Springer, Cham, 2018.

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.