



# The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks

Andre Massahiro Shimaoka  [ Instituto de Pesquisas Tecnológicas de São Paulo and Federal University of São Paulo | [andre.shimaoka@unifesp.br](mailto:andre.shimaoka@unifesp.br) ]

Renato Cordeiro Ferreira  [ Institute of Mathematics and Statistics, University of São Paulo (USP) | [renatocf@ime.usp.br](mailto:renatocf@ime.usp.br) ]

Alfredo Goldman  [ Institute of Mathematics and Statistics, University of São Paulo (USP) | [gold@ime.usp.br](mailto:gold@ime.usp.br) ]

✉ Instituto de Pesquisas Tecnológicas, Av. Prof. Almeida Prado, nº 532, Butantã, CEP 05508-901, São Paulo/SP, Brazil.

Received: 17 October 2023 • Accepted: 07 October 2024 • Published: 24 October 2024

**Abstract** The expansion of Data Science projects in organizations has been led by three factors: the growth in the amount of data generated, the evolution in storage capacity, and the increase in computational capabilities. However, most of these projects fail to deliver the expected value: 82% of the teams do not use any process model. Despite the popularity of Agile Methods, their adoption in Data Science projects is still scarce. Most of the existing research focuses on algorithms. There is a lack of studies on agility in Data Science. This Systematic Literature Review (SLR) was performed to identify and evaluate 16 studies that can answer how to adapt and apply CRISP-DM using different approaches — methods, frameworks, or process models. In addition, it shows how CRISP-DM has evolved over the last few decades, with derivations emerging from rigid processes to agile methods. This research then analyzes the 16 tailored models and examines the similarities and differences between CRISP-DM derivatives. As a result, it summarizes the CRISP-DM adaptation patterns identified, such as phase addition, phase modification, features and tools addition, and integration with other approaches. Consequently, this SLR showcases how CRISP-DM is a robust, flexible, and highly adaptable model that can be extended to different business domains. Finally, it proposes a theoretical guide to modify and customize CRISP-DM for Data Science projects.

**Keywords:** Data Science; Process Model, CRISP-DM, Agile

## 1 Introduction

The explosive growth of user-generated data alongside the evolution of storage capacity, data accessibility, and computational power are driving the popularization of Data Science within organizations [Ahmad *et al.*, 2022; Goyal *et al.*, 2020]. According to IDC [2020], global data will grow from 45 zettabytes in 2019 to 175 zettabytes by 2025, which represents a growth of approximately 289% over six years.

Data Science allows companies to reason over a high volume of data coming from different sources, enhancing their decision-making and data product creation [Ahmad *et al.*, 2022; Manirupa *et al.*, 2015]. It is widely adopted in many sectors such as finance, medicine, and marketing. Due to advances in technology and public acceptance, Data Science will continue to grow over the next few years [Ahmad *et al.*, 2022].

Since the introduction of the Agile Manifesto in the early 2000s, Agile methods are growing in popularity, mostly adopted for software development [Hoda *et al.*, 2018]. Agile practices have a large impact on the technology industry, improving communication, collaboration, quality, team productivity, client satisfaction, and successfully delivered products [Julian and Anslow, 2019].

### 1.1 Data Science

Data Science focuses on data product building, i.e., software products that extract information and generate knowledge from data. It applies algorithms to solve problems using data [Manirupa *et al.*, 2015; Provost and Fawcett, 2013]. The term Data Science has been more used than Data Mining for knowledge discovery and data-oriented problem-solving [Martínez-Plumed *et al.*, 2019].

CRISP-DM is an acronym for Cross Industry Standard Process for Data Mining. It was created in the 1990s by a group of organizations (Teradata, SPSS, Daimler-Chrysler, and OHRA) for data science projects [Saltz, 2020; Mariscal *et al.*, 2010]. A process model is a reference that structures projects and improves the understanding and communication between stakeholders. A Data Mining process requires several tools, techniques, and personnel, along with efficient management [Wirth and Hipp, 2000].

According to Martínez-Plumed *et al.* [2019], CRISP-DM incorporates principles and ideas of other processes, such as KDD and SEMMA. Given its robustness, it became the base for more recent proposals, such as Microsoft Team Data Science Process (TDSP) [Ahmed *et al.*, 2018; Costa and Aparicio, 2020].

## 1.2 Agile Methods and Data Science

According to Versionone [2020], organizations are increasingly adopting agile practices and techniques. The main reasons include i) to accelerate the delivery; ii) to improve management and prioritization; iii) to increase productivity; iv) to provide predictability of delivery; v) to improve quality; vi) to increase the alignment between Business and IT; vii) to reduce risk; and viii) to improve visibility.

In opposition to traditional methods, agile methods seek shorter development cycles, more interaction with stakeholders, incremental delivery, and flexibility for change [Matharu *et al.*, 2015]. The most common agile practices used by companies include unit tests, continuous integration, refactoring, automated tests, pair programming, and test-driven development [Versionone, 2020].

Although Data Science projects are becoming more popular, most research about it focuses on techniques (models and algorithms), whereas less focus is given to project management. 82% of teams do not use a process for Data Science. Therefore, Data Science projects fail or do not deliver the expected value [Saltz and Sutherland, 2019].

Data Science projects are aligned with agile principles. They seek constant feedback, quick response to changes, and insight from stakeholders. However, there are few references in the literature about Agile on Data Science projects [Larson and Chang, 2016]. Since Data Science is exploratory in nature, an agile approach can help manage expectations and achieve predefined goals [Riungu-Kalliosaari *et al.*, 2017].

Although CRISP-DM is an iterative approach, it is frequently used sequentially, contradicting agile practices: in CRISP-DM, there is no defined process saying when or how to perform an iteration [Saltz and Sutherland, 2019; Baijens *et al.*, 2020].

As a consequence, a project that follows only with CRISP-DM is unlikely to meet all the client's needs, such as incremental delivery, forecasting, and adaptability, among other benefits from agile methods. Moreover, CRISP-DM does not suggest software engineering practices by itself [Baijens *et al.*, 2020].

The goal of this study is to evaluate the existing literature that integrates CRISP-DM in different contexts. It is important to understand how CRISP-DM can be adopted together with other methodologies to show whether it can be used to enable agile data science projects. Therefore, this paper seeks to evaluate studies that can answer the following research question:

How is the process model CRISP-DM adapted or utilized with other methodologies in Data Science projects?

Considering the knowledge generated by this literature review, this research aims to map the adaptations of CRISP-DM with other methodologies.

The article is divided into seven sections. **Section 3** brings the theoretical foundation with an overview of CRISP-DM and other frameworks. **Section 4** presents the method of Systematic Literature Review (SLR) adopted in this work. **Section 5** provides the analysis and the results obtained. **Section 5.6** synthesizes and discusses the SLR by introducing a theoretical reference proposal. **Section 6** overviews the threats to validity and discusses the overall limitations of this research. Finally, **Section 7** presents final remarks, including, suggestions for future research, and practical applications in an enterprise environment.

**Section 5.6** synthesizes and discusses the SLR by introducing a theoretical reference proposal. **Section 6** overviews the threats to validity and discusses the overall limitations of this research. Finally, **Section 7** presents final remarks, including, suggestions for future research, and practical applications in an enterprise environment.

## 2 Related Work

Schröer *et al.* [2021] conducted a systematic literature review including papers published between 2017 and 2019. This research summarizes domains of application for CRISP-DM, highlighting health, education, and research. Moreover, this review analyzes how each of the six CRISP-DM phases was conducted in the studies, showcasing their different approaches. However, the study left some unaddressed points. It did not explicitly mention how modifications, additions, or removals to the CRISP-DM phases were executed, nor did it investigate integrations with other approaches, tools, and features. It also did not address the evolution of new versions of adapted CRISP-DM. Mixing the original and adapted versions complicated a clear analysis of the pros and cons of adopting an adapted CRISP-DM. Additionally, the study did not examine the use of the models in agile contexts.

The study conducted by Martínez-Plumed *et al.* [2019] aimed to assess whether CRISP-DM remains suitable for data science projects, considering significant changes and advancements between 2008 and 2018. The study was not a systematic review but examined the evolution of CRISP-DM. It found that CRISP-DM is aligned with many projects without requiring major modifications. However, it identified shortcomings in CRISP-DM, such as the lack of additional exploratory activities. The study compared CRISP-DM with new approaches individually but did not include comparisons among different adaptations of CRISP-DM or integrations with other frameworks, tools, and features.

## 3 Background

This section presents an overview of the main processes for data science, including CRISP-DM, KDD, SEMMA, Scrum, Kanban, and TDSP.

### 3.1 CRISP-DM

This section introduces the foundation for this research. It highlights the main concepts and phases of the CRISP-DM process model.

CRISP-DM is a technology-independent process model. It can be applied to any industry and it aims to turn data mining projects easier, faster, more repeatable, and more manageable. CRISP-DM defines the activities that should be done to develop a data mining project: it defines what should and should not be done [Mariscal *et al.*, 2010; Wirth and Hipp, 2000].

CRISP-DM describes a life cycle for data mining projects containing six phases. **Figure 1** shows that these phases can be sequential or cyclic, allowing for stopping and resuming between phases [Wirth and Hipp, 2000].

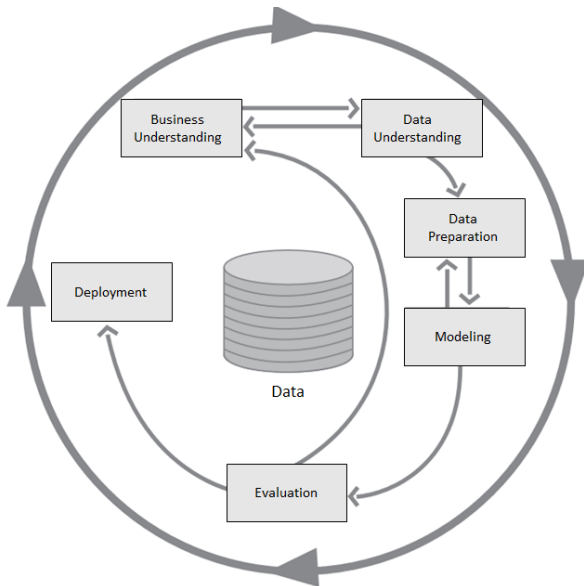


Figure 1. CRISP-DM Phases [Chapman *et al.*, 2000]

Chapman *et al.* [2000], Azevedo and Santos [2008], and Wirth and Hipp [2000] outline a summary of each CRISP-DM phase:

- **Business Understanding:** focuses on understanding the goals and requirements of the project from a business perspective. The data scientist converts requirements into a data mining problem definition, then elaborates a project plan.
- **Data Understanding:** focuses on the initial data gathering to gain familiarity with the data. The data scientist gains initial insights about the data and builds hypotheses.
- **Data Preparation:** focuses on building a data set that will be used later in the modeling phase. The data scientist executes preparatory tasks such as data selection, transformation, and cleaning. This phase can be conducted multiple times without a prescribed order.
- **Modeling:** focuses on comparing techniques and estimating parameters to solve the problem. During this phase, the data scientist might detect problems with the dataset, requiring them to return to previous phases for adjusting the data.
- **Evaluation:** focuses on the validation of the chosen model. The data scientist reviews the steps executed previously to ensure the model meets the business goals and any metrics defined to it.
- **Deployment:** focuses on the packaging of the model in a way the customer can use it. The data scientist may produce a solution varying from a simple report to implementing a repeatable data mining process for the entire organization (e.g., deploying a recommendation system for an e-commerce)

Although the steps are not strictly sequential, CRISP-DM assumes that they can be sequential, i.e., a step can not start before the previous one [Wirth and Hipp, 2000]. CRISP-DM allows looping back, but there is no defined process on how and when to do these iterations [Saltz and Sutherland, 2019]. **Table 1** summarizes the advantages and disadvantages of CRISP-DM for Data Science projects, including

their bibliographical references.

## 3.2 KDD

According to Fayyad *et al.* [1996], the Knowledge Discovery in Databases (KDD) process aims to uncover knowledge within data through data mining. It consists of five distinct phases:

- **Selection:** extracts data from various sources and chooses a dataset.
- **Preprocessing:** cleans and prepares the data for the next steps.
- **Transformation:** transforms data and execute feature selection.
- **Data Mining:** identifies patterns and trends in the data.
- **Interpretation/Evaluation:** assesses the identified patterns for their relevance and value, leading to the generation of knowledge.

CRISP-DM and KDD by Fayyad *et al.* [1996] show similarities, but KDD does not clearly define the phases of *Business Understanding* and *Deployment* [Azevedo and Santos, 2008]. While KDD combines evaluation and deployment within the *Interpretation/Evaluation* phase, CRISP-DM addresses these activities separately [Dåderman and Rosander, 2018]. According to Martínez-Plumed *et al.* [2019], CRISP-DM is a practical evolution of KDD, offering a more applicable framework for data science processes.

## 3.3 SEMMA

SEMMA is a structured data mining process created by SAS [2003] in the 1990s and used in the SAS Enterprise Miner tool. It consists of five phases:

- **Sample:** extracts data samples from a larger dataset.
- **Explore:** analyzes and understands the data statistically and graphically, identifying trends or anomalies.
- **Modify:** prepares the data for analysis by creating additional variables, transforming variables, and handling missing values.
- **Model:** applies data mining algorithms to find patterns and build predictive models.
- **Assess:** evaluates the effectiveness and accuracy of the created model.

CRISP-DM covers the entire data science project lifecycle, from problem understanding to deployment. In contrast, SEMMA focuses primarily on data management and modeling, with less emphasis on business problem comprehension [Palacios *et al.*, 2017; Azevedo and Santos, 2008]. Additionally, CRISP-DM is an open and non-proprietary framework, whereas SEMMA is developed by SAS and often associated with its tools [Palacios *et al.*, 2017].

## 3.4 Scrum

According to Schwaber and Sutherland [2020], Scrum is an agile and adaptive framework designed to address complex problems and promote continuous delivery. Scrum originated in 1995 but only gained widespread popularity in the 2000s. It defines specific roles, such as Scrum Master, developers, and Product Owner. The Scrum events are:

- **Sprint:** a time-boxed period aimed at creating an increment of the product.

Pros of CRISP-DM	Reference
Structured model focused on processes and the work to be done	Baijens <i>et al.</i> [2020] Mariscal <i>et al.</i> [2010] Wirth and Hipp [2000] Chapman <i>et al.</i> [2000]
Facilitates the communication in the project, providing a clear reference and common terminology	Wirth and Hipp [2000]
The dominant model in the market and the best-known in Data mining and Data Science	Martínez-Plumed <i>et al.</i> [2019] Saltz [2020] Mariscal <i>et al.</i> [2010]
Concerns about the requirements and business goals definition (not only the technical side)	Baijens <i>et al.</i> [2020] Chapman <i>et al.</i> [2000]
Incorporates principles and ideas of the majority of the model process and frameworks for Data Science	Martínez-Plumed <i>et al.</i> [2019] Mariscal <i>et al.</i> [2010]
Independent of tools and technologies, suitable to any industry	Wirth and Hipp [2000] Mariscal <i>et al.</i> [2010]
Cons of CRISP-DM	Reference
Does not focus on project management	Baijens <i>et al.</i> [2020] Mariscal <i>et al.</i> [2010] Wirth and Hipp [2000]
Is not predictable	Mariscal <i>et al.</i> [2010] Wirth and Hipp [2000]
Allows iteration, but it does not specify how and when to return to a previous step and perform the iteration	Baijens <i>et al.</i> [2020] Saltz and Sutherland [2019]
Commonly used in sequential and linear form	Saltz [2020] Wirth and Hipp [2000]
Does not follow agile principles and practices	Baijens <i>et al.</i> [2020] Saltz and Sutherland [2019]

**Table 1.** Pros and Cons of CRISP-DM Model.

- **Sprint Planning:** the work undertaken during the Sprint, as defined by the team.
- **Daily Scrum:** A daily meeting to discuss what has been done, what will be done, and any impediments.
- **Sprint Review:** the product increment is presented to stakeholders to gather feedback.
- **Sprint Retrospective:** the team discusses what went well and what could be improved in the next Sprint.

The artifacts include:

- **Product Backlog:** an ordered list of items and requirements as specified by the product team.
- **Sprint Backlog:** the items from the Product Backlog selected for the Sprint.
- **Increment:** the items completed during the Sprint that represent a potentially usable version of the product.

According to Baijens *et al.* [2020], Scrum and CRISP-DM serve different purposes and represent distinct approaches. Scrum is an agile method not originally designed for data science contexts. In data science, activities involving exploration and experimentation can complicate requirements definition and incremental delivery. These activities may also be challenging to align with Sprint events. In contrast, while CRISP-DM is not an agile methodology, it offers flexibility for adaptation.

### 3.5 Kanban

According to [Anderson, 2010], Kanban is a workflow management method aimed at improving efficiency and produc-

tivity in development processes. Kanban originated in the 1940s at Toyota, but began to gain prominence in the 2000s when it was applied to software development environments. This method enables teams to identify bottlenecks and continuously improve their processes based on:

- **Visualization of work:** creating a Kanban board to display all tasks in different columns representing the stages of the process.
- **Limiting work in progress (WIP):** establishing limits for the amount of work that can be in progress at each stage of the workflow.
- **Flow management:** monitoring metrics such as cycle time, throughput, and flow efficiency
- **Process Policies Explicit:** clearly defining and communicating the process rules and policies.
- **Feedback Loops:** using regular meetings and reviews to gather feedback
- **Collaborative improvement and experimental evolution:** fostering a culture of continuous improvement through collaborative experimentation and learning.

Kanban and CRISP-DM are distinct approaches but can be used together in data science. Kanban focuses on workflow management and team collaboration, while CRISP-DM provides a structured framework to guide the data science process [Saltz *et al.*, 2017].

### 3.6 TDSP

Team Data Science Process (TDSP) is a framework developed and owned by Microsoft [2024] for organizing and managing data science projects based on agile principles. TDSP can be integrated with Microsoft solutions, such as Azure Machine Learning. It defines the roles of Project Lead, Data Scientist, Data Engineer, and Solution Architect. The TDSP is structured into five main phases:

- **Business Understanding:** defines project objectives and identifies relevant data sources.
- **Data Acquisition and Understanding:** collects and transfers data to repositories, including data cleaning, transformation, and exploratory analysis.
- **Modeling:** applies feature engineering, model training, and using various algorithms to identify the best solution.
- **Deployment:** implements the model in a production environment and monitors its performance.
- **Customer Acceptance:** confirms the implementation of the product with the client.

Both methodologies, CRISP-DM and TDSP, have similar phases, such as *Business understanding*, *Data understanding*, *Data preparation*, and *Modeling*. However, TDSP is more iterative and adaptable to agile contexts. Additionally, TDSP includes a specific phase for *Customer acceptance*, while in CRISP-DM, this activity occurs during the *Evaluation* or *Deployment* phases. The defined roles and responsibilities in TDSP provide direction but may limit flexibility, particularly if phase adaptations or new roles are needed. As a proprietary method, the customization of TDSP is more restricted compared to the open and widely recognized CRISP-DM.

## 4 SLR - Planning

A Systematic Literature Review (SLR) was applied to map references and highlight relevant studies regarding CRISP-DM. This SLR identifies, analyzes, and compares methods, process models, and frameworks in the literature about CRISP-DM. Moreover, it catalogs the main adaptations and/or extensions concerning its original form.

A Systematic Literature Review is supported by a research protocol with the goal of identifying, selecting, interpreting, assessing, and summarizing the literature about a specific topic or research question [Kitchenham and Charters, 2007]. The purpose of this research protocol is to reduce researcher bias and allow reproducibility by defining strategies, criteria, and forms to be followed in the SLR [Nakagawa *et al.*, 2017].

According to Kitchenham and Charters [2007] and Nakagawa *et al.* [2017], a systematic review can be divided into three phases:

- **planning:** defines the systematic review goals and the research protocol,
- **conducting:** comprises the identification and selection of studies according to the search strategy and selection criteria, and
- **synthesis of the results:** summarizes the data describing and assessing the results.

This Systematic Literature Review seeks to identify relevant studies that help to answer the following research question:

How is the process model CRISP-DM adapted or utilized with other methodologies in Data Science projects?

This research question will be divided and organized using the PICO strategy from Petticrew and Roberts [2006]. The acronym PICO stands for Population, Intervention, Comparison, and Outcome:

- **population:** data science projects;
- **intervention:** concepts about process models based on CRISP-DM that were adapted or extended;
- **comparison:** comparison with the original CRISP-DM; and
- **outcome:** articles that bring the proposal for the use of CRISP-DM with other approaches in Data Science.

### 4.1 Data Source Selection Criteria

Data sources were chosen using the following criteria: i) it is available on the web; ii) it is a preferably relevant scientific base well-known in the Computer Science area; iii) it has search mechanisms for keywords; and iv) it is recognized by systematic literature review bibliographies or experts.

### 4.2 Database / Search Engine

Based on the selection criteria presented previously, three bibliographic sources, known in the Computer Science field, were chosen:

Sources	URL
ACM Digital Libray	https://dl.acm.org
IEEE Xplore	https://ieeexplore.ieee.org
Scopus	https://www.scopus.com

Table 2. Database / Search Engine.

### 4.3 Keywords

Keywords are terms extracted from the research questions that represent the goal of the SLR. They characterize the investigated theme and are used to elaborate a search string [Nakagawa *et al.*, 2017]. For the automated database search, the keyword CRISP-DM was used alongside the following terms: *Process Model, Reference Model, Framework, Methodology, Method, and Approach.*

Moreover, some terms were added to imply adaptations or integrations of CRISP-DM, such as *adapted, extended, modified, integrated, improved, adaptation, modification, extension, extendable, adaptable, integration, new approach, novel approach, novel architecture, new architecture, alternative approach, tailored, and improvement.*

Afterward, the “CRISP-DM” keyword was connected by the operator ”AND” with other synonyms and with terms that represent adaptations. At least one of these synonyms should be present in the search. Therefore, they are connected via “OR” operators.

The validation of the search string involved an iterative process of adjustments to ensure the identification of relevant studies in the systematic review. We conducted preliminary searches in Scopus with well-cited studies studies such as LTDM [Ahmed *et al.*, 2018] and DMME [Huber *et al.*, 2018], adjusting the terms and operators as needed to optimize coverage. **Figure 2** shows the search string used on each search source.

(CRISP-DM  
AND  
(adapted OR extended OR modified OR integrated OR improved OR adaptation OR modification OR extension OR extendable OR adaptable OR integration OR "new approach" OR "novel approach" OR "novel architecture" OR "new architecture" OR "alternative approach" OR tailored OR improvement)  
AND  
("Process Model" OR "Reference Model" OR "Life Cycle" OR process OR method OR Approach OR Framework OR Methodology))

Figure 2. Search String

### 4.4 Study Inclusion and Exclusion Criteria

The inclusion and exclusion criteria were defined to guide how to choose relevant studies for the SLR [Nakagawa *et al.*, 2017]. These criteria provide evidence over the research question and reduce the probability of research bias [Kitchenham and Charters, 2007]. Based on the research objective, the following inclusion (IC) and exclusion (EC) criteria were defined:

- IC.1** Full studies published in electronic format



- IC.2 Studies that adapt or extend the CRISP-DM process model.
- IC.3 Studies that present new methods or process models that were based on CRISP-DM.
- EC.1 Studies with no full-text access.
- EC.2 Duplicated studies in more than one source.
- EC.3 Studies not in Portuguese or English.
- EC.4 Studies that only focus on the algorithm technique or in data science models.
- EC.5 Studies that only mention or apply CRISP-DM in its original format.
- EC.6 Studies that do not contain a summary (abstract).
- EC.7 Review studies.

An article is included when it meets IC.1 and at least one other inclusion criterion. On the other hand, an article is excluded when it meets any exclusion criteria.

### 4.5 Quality Assessment Checklist

The quality of the SLR depends on the included studies. The Quality Assessment Checklist is a tool designed to evaluate the quality of these studies by applying stringent quality criteria and reducing bias. This helps to exclude low-quality studies that might compromise the synthesis of results [Yang *et al.*, 2021]. We developed a checklist comprising eight quality criteria:

- QA.1 Is the research objective or question clearly defined?
- QA.2 Did the study adequately justify the reasons for adapting the CRISP-DM model?
- QA.3 Did the study clearly describe how CRISP-DM was adapted or modified to meet specific needs, including any changes in processes, phases, or features?
- QA.4 Did the study provide practical or empirical evidence demonstrating the effectiveness or efficiency of the CRISP-DM model adaptation?
- QA.5 Did the study adequately ground the new model with bibliographic references?
- QA.6 Did the study clearly detail the research methodology used to create the new model?
- QA.7 Did the study discuss the limitations of the research appropriately?
- QA.8 Did the study explicitly and adequately discuss future research directions?

Each fully met criterion received 1 point, while partially met criteria received 0.5 points. The maximum possible score was 8 points. Studies that scored 4.5 points or above were approved, while those with a score below 4.5 were rejected.

### 4.6 Strategy for study selection

The data selection was divided into five steps. **Figure 3** presents graphically each of the selection steps.

Step 1 contains the automated search in each of the selected data sources. The search string was applied in the title and abstract of the studies. Step 2 shows the pre-assessment

that was made considering the exclusion and inclusion criteria after abstract reading. Step 3 considers the same inclusion and exclusion criteria, applied after a complete reading of the studies. Afterward, information from the studies was collected and the data extraction form was filled out as shown in **Figure 4**.

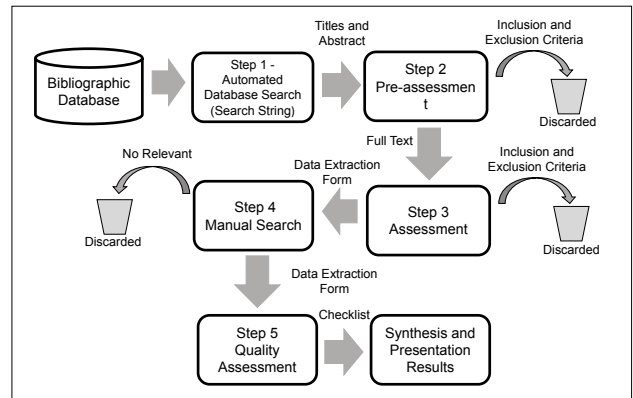


Figure 3. Studies Selection Process Steps

Step 4 is the snowballing process, a manual search was made observing the references about CRISP-DM cited in the articles. If the studies are consistent and relevant, the inclusion and exclusion criteria were applied, and the data extraction form was used. Subsequently, in step 5, the Quality Assessment Checklist is applied to evaluate the methodological quality of the studies included in the review.

### 4.7 Data Extraction and Synthesis Strategy

To minimize the researchers' bias, the data extraction form (showed in **Figure 4**) was used during data gathering. It captures information about model usage, justification, limitations, phases considered, and types of adaptation over in CRISP-DM. Using this standardized protocol, the interrelationships, differences, and similarities between the methods were shown.

Data Extraction Form		ID: <input type="text"/>
Title:	<input type="text"/>	
Authors:	<input type="text"/>	
Publication Year:	<input type="text"/>	Source: <input type="text"/> Language: <input type="text"/>
Name (Method or Model Process):	<input type="text"/>	
Scenario or Context of Use:	<input type="text"/>	
Justification for the use new Method:	<input type="text"/>	
Summary:	<input type="text"/>	
Phases (Steps) of the Model:	<input type="text"/>	
Adaptations in CRISP-DM:	<input type="checkbox"/> Change of existing phases in CRISP-DM <input type="checkbox"/> Creation of new Phases in CRISP-DM <input type="checkbox"/> Addition of features or tools complementary to CRISP-DM <input type="checkbox"/> Integration of CRISP-DM to other methods or frameworks	

Figure 4. Data Extraction Form Model

ID	Title	Author and Year	Model Name	Scenario or Context
13	Intelligent Big Data Analysis Architecture Based on Automatic Service Composition	Siriweera <i>et al.</i> [2015]	BDA Architecture	Not restricted to any specific domain, but suitable for Big Data projects
28	A methodology for prior management of temporal data quality in a data mining process	Diop <i>et al.</i> [2017]	DPM	For mining projects of temporal data
36	Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes	Schafer Schäfer <i>et al.</i> [2018]	QM-CRISP-DM	For data mining projects in the Manufacturing field
45	POST-DS: A Methodology to Boost Data Science	Costa and Aparicio [2020]	POST-DS	Not restricted to any specific scenario or context, but seeks to improve the organization and management of projects
46	Applying Scrum in Data Science Projects	Baijens <i>et al.</i> [2020]	SCRUM-DS	Not restricted to any specific domain, but suitable for agile projects
58	Towards a Process Model to Enable Domain Experts to Become Citizen Data Scientists for Industrial Applications	Merkelbach <i>et al.</i> [2022]	CRISP-DM tailored for domain experts	It is not restricted to any specific domain, but it allows domain experts to perform data science activities
65	Specializing CRISP-DM for evidence mining	Venter <i>et al.</i> [2007]	CRISP-EM	Used for discovery and mining of evidence in digital forensic investigation
132	CRISP-eSNeP: Towards a data-driven knowledge discovery process for electronic social networks	Asamoah and Sharda [2019]	CRISP-eSNeP	For social media platforms with a great volume of data (Big Data)
133	DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model	Huber <i>et al.</i> [2018]	DMME	For data mining projects in the Manufacturing Engineering field
143	A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects	Ahmed <i>et al.</i> [2018]	LTDM	Not restricted to any specific scenario or context, seeks the application of current concepts of Design Thinking and Lean Startup in Data Science
153	Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset	Ayele [2020]	CRISP-IM	Used for discovery and mining of new ideas (Innovation)
158	Crisp-dm/smes: A data analytics methodology for non-profit smes	Montalvo-Garcia <i>et al.</i> [2020]	CRISP-DM/SMEs	For Small and Mid-Size Enterprises (SMEs)
192	Data science as knowledge creation a framework for synergies between data analysts and domain professionals	van der Voort <i>et al.</i> [2021]	Knowledge Creation + CRISP-DM	It is not restricted to any specific context, but it integrates knowledge sharing in the Data Science
202	Designing a data mining process for the financial services domain	Plotnikova <i>et al.</i> [2022]	FIN-DM	For Financial Services Data Science projects to support regulatory compliance
203	Development of a Framework to Aid the Transition from Reactive to Proactive Maintenance Approaches to Enable Energy Reduction	Ahern <i>et al.</i> [2022]	IDAIC	For data science projects in the industrial data domain and the proactive maintenance of equipment
208	CRISP Data Mining Methodology Extension for Medical Domain	Niaksu [2015]	CRISP-MED-DM	For data mining projects in the medical and healthcare domain
209	CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories	Martínez-Plumed <i>et al.</i> [2019]	DST	Not restricted to any specific domain, but integrating additional exploratory activities

Table 3. Related Articles by the SLR

## 5 Results

The Systematic Literature Review was conducted in March 2023. **Figure 5** represents the SLR summary, showing the selected articles and the inclusion and exclusion criteria applied at each stage:

- **Step 1:** The automated database search identified 207 studies.
- **Step 2:** The pre-assessment involved reading the titles

and abstracts of the studies, and then applying the inclusion and exclusion criteria, resulting in the acceptance of 37 studies.

- **Step 3:** The assessment involved reading all 37 studies and reapplying the same inclusion and exclusion criteria, resulting in 15 studies accepted. To reduce the researchers' bias, the accepted studies were collected and noted in the data extraction form.
- **Step 4:** The snowballing process examined reference

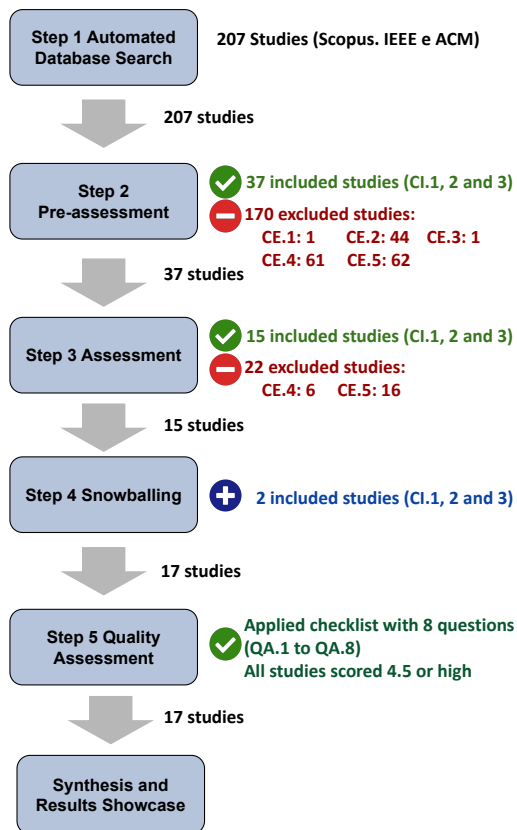


Figure 5. Systematic Review Evolution Summary

articles, then two additional papers were identified, resulting in a total of 17 selected studies.

- **Step 5:** All studies that underwent the quality assessment checklist achieved a score of 4.5 or higher.

The 17 new processes, their ID, name, and context are shown in **Table 3**. Details of each step can be found in the supplementary material<sup>1</sup>.

All the studies were analyzed with the information collected in the extraction form. Afterward, a comparison between the adaptations of CRISP-DM was made, observing the differences and similarities between the methods.

Most of the new models based on CRISP-DM were made to meet some specific scenario or context. However, some methods are generic. These models add elements of other methods such as agility, design thinking, lean startup, big data, and others.

Many models underwent updates in the CRISP-DM standard steps: Eight had their steps changed, and six had new steps included. When the steps were changed, the process life cycle was also adapted. **Table 4** shows a comparison between the models, highlighting their adaptations.

### 5.1 Phases modification

Despite the renaming, the six process models (CRISP-DM/SMEs, CRISP-EM, CRISP-eSNEP, CRISP-IM, CRISP-MED-DM, and LTDM) did not change the essence of the phases. They were only adapted to different contexts. Even when there were modifications, such as in the CRISP-eSNeP

ID	Model Name	Year	Modify	Add	Features or Tools	Integration
13	BDA Architecture	2015				X
28	DPM	2017		X		X
36	QM-CRISP-DM	2018			X	X
45	POST-DS	2020			X	
46	SCRUM-DS	2020			X	X
58	CRISP-DM for Domain Experts	2022	X	X		
65	CRISP-EM	2007	X			
132	CRISP-eSNeP	2019	X			
133	DMME	2018		X		
143	LDTM	2018	X		X	
153	CRISP-IM	2020	X			
158	CRISP-DM/SMEs	2020	X			
192	Knowledge Creation	2021		X		X
202	FIN-DM	2022		X		
203	IDAIC	2022	X	X		
208	CRISP-MED-DM	2015	X			
209	DST	2019				X
	<b>Total</b>		<b>8</b>	<b>6</b>	<b>4</b>	<b>6</b>

Table 4. Comparison between the 16 new methods based on CRISP-DM

method, the fundamentals from CRISP-DM remained unaltered. **Table 5** presents a comparison of the phases that were renamed, changed, grouped, or divided. They are further detailed in the following subsections. Any unmentioned phases are equivalent to the ones in CRISP-DM, following its naming conventions.

#### CRISP-DM/SMEs

Montalvo-Garcia *et al.* [2020] created CRISP-DM/SMEs for small and mid-sized companies. This model reduces CRISP-DM into five phases.

The *Business Understanding* phase was modified to a **Project Definition** phase, whereas *Data Understanding* and *Data Preparation* were grouped into **Data Management**.

- **Project Definition:** the project goals and success criteria are defined, observing the strategic plan of the Small and Mid-sized Enterprises (SMEs). Human resources and finances are estimated. Scope and risks are defined.
- **Data Management:** the data is collected from several sources. Afterward, it is integrated and formatted.

#### CRISP-EM

The CRISP-EM by Venter *et al.* [2007] is an adaptation of CRISP-DM for mining digital evidence in forensic analysis.

The *Business Understanding*, *Modeling*, *Evaluation*, and *Deployment* phases were modified to **Case Understanding**, **Evidence Modeling**, **Evaluation and Evidence Extraction**, and **Evidence Reporting**, respectively.

- **Case Understanding:** as in CRISP-DM, the business objectives and requirements are defined. However, this stage focuses specifically on investigating requirements.
- **Evidence Modeling:** the evidence models and techniques are selected for the events reconstruction.

<sup>1</sup><https://doi.org/10.5281/zenodo.12753088>



Process Model	Process Model Phase					
CRISP-DM	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
CRISP-DM/SMEs (158)	Project Definition	Data Management		Modeling	Evaluation	Deployment
CRISP-EM (65)	Case Understanding	Data Understanding	Data Preparation	Evidence Modeling	Evaluation and Evidence Extraction	Evidence Reporting
CRISP-eSNeP (132)	Integrated Business Knowledge	Big Data Platform Development	Data Acquisition and Storage	Model Development	Evaluation and Deployment	
			Data Cleaning and Formatting			
			Data and Graph Validation			
CRISP-IM (153)	Technology Need Assessment	Data Collection and Understanding	Data Preparation	Modeling for Idea Extraction	Evaluation and Idea Extraction	Reporting Innovative Ideas
CRISP-MED-DM (208)	Problem Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
LDTM (143)	Work Discovery	Analytical Approach	Data Preparation	Build MVP	Measure Value	Learn and Update
		Data Resources				

**Table 5.** Phases modification comparison between Original CRISP-DM and the 6 new methods

- **Evaluation and Evidence Extraction:** the results of evidence models are evaluated. In addition, the steps and activities performed are reviewed.
- **Evidence Reporting:** the knowledge acquired is presented in proper format and used as proof and evidence.

**CRISP-eSNeP**

Asamoah and Sharda [2019] remodeled CRISP-DM phases for Big Data and social network contexts. This model had the highest number of modifications.

The *Business Understanding*, *Data Understanding*, and *Modeling* phases were modified to **Integrated Business Knowledge**, **Big Data Platform Development**, and **Model Development**, respectively. The *Data Preparation* phase was expanded into **Big Data Platform Development**, **Data Acquisition / Storage**, and **Data Cleaning / Formatting**.

- **Integrated Business Knowledge:** the problems, business objectives, success criteria, and risks are detailed. Unlike CRISP-DM, there is a concern with the integration of the stakeholders from different departments involved.
- **Big Data Platform Development:** the Big Data infrastructure is implemented, allowing the processing of large amounts of data. Cloud platforms and tools that support analysis can be used.
- **Data Acquisition and Storage:** the different data types (such as texts, JSON, video, and audio) are collected from social network platforms.
- **Data Cleaning and Formatting:** the irrelevant data is eliminated. Furthermore, unstructured data may come in several formats (such as text, video, and sound). They are converted into a common format.
- **Data and Graph Validations:** the data accuracy and sample representativeness of the population chosen are validated.

- **Model Development:** the algorithms are selected to build descriptive and predictive models.
- **Evaluation and Deployment:** follow the same scope as CRISP-DM. However, this phase is a grouping of the two phases in CRISP-DM.

CRISP-esNeP conducts a case study on Twitter to analyze how influence affects information dissemination. They implemented and demonstrated each phase of CRISP-esNeP in a big data project

**CRISP-IM**

Ayele [2020] elaborated a process model that follows the five phases of CRISP-DM, specializing in idea mining.

The *Business Understanding*, *Data Understanding*, *Modeling*, *Evaluation*, and *Deployment* phases were modified to **Technology Need Assessment**, **Data Collection and Understanding**, **Modeling for Idea Extraction**, **Evaluation and Idea Extraction**, and **Reporting Innovative Ideas**.

- **Technology Need Assessment:** the business requirements are discovered. In addition, the trends and patterns for the innovation of the ideas are identified.
- **Data Collection and Understanding:** collection, cleaning, data reformatting, anomaly removal, and redundant data removal activities are performed.
- **Modeling for Idea Extraction:** the best models are selected through techniques identification such as text mining, social network analysis, statistical analysis, and bibliometrics.
- **Evaluation and Idea Extraction:** the ideas are extracted and the results are evaluated in relation to the objectives defined in the first phase.
- **Reporting Innovative Ideas:** the analysis results are communicated with a focus on the idea clarification. Furthermore, the lessons learned and innovative ideas are documented and published.

CRISP-DM	DMME (133)	DPM (28)	Knowledge Creation + CRISP-DM (192)	FIN-DM (202)
Business Understanding	Business Understanding	Business Understanding	Business Understanding	Business Understanding
-	Technical Understanding	Prior Temporal Data Understanding	Enlarging individual Knowledge	Requirements phase
-	Technical Realization	Prior Temporal Data Preparation	Sharing Tacit Knowledge	-
Data Understanding	Data Understanding	Data Understanding	Data Understanding	Data Understanding
Data Preparation	Data Preparation	Data Preparation	Data Preparation	Data Preparation
Modeling	Modeling	Modeling	Modeling	Modeling
Evaluation	Evaluation	Evaluation	Evaluation	Evaluation
-	Technical Implementation	-	Networking Knowledge	Compliance phase
Deployment	Deployment	Deployment	Deployment	Deployment
-	-	-	-	Pos-Deployment

**Table 6.** Phase addition comparison between Original CRISP-DM and the 4 methods

### CRISP-MED-DM

Niaksu [2015] created an adaptation of CRISP-DM including specific tasks in the medical domain. The *Business Understanding* phase was modified to **Problem Understanding**.

- **Problem Understanding:** it includes known tasks from CRISP-DM with the inclusion of specific tasks for the medical domain, such as the definition of clinical goals, the definition of objectives for the management of health care, and the evaluation of patient data privacy issues.

### LTDm

Ahmed *et al.* [2018] adapted CRISP-DM phases for the usage of Lean and Design Thinking tools.

The *Business Understanding*, *Modeling*, *Evaluation*, and *Deployment* phases were modified to **Work Discovery**, **Build MVP**, **Measure Value**, and **Learn and Update**, respectively. Besides that, the *Data Understanding* phase was divided into **Analytical Approach** and **Data Resources**.

- **Work Discovery:** the Design Thinking strategies are adopted to identify problems and propose solutions, also to define goals, project objectives, functional and non-functional requirements.
- **Analytical Approach:** the team starts to think about the possible techniques of statistics and machine learning.
- **Data Resources:** all the data resources associated with the problem domain are identified and collected.
- **Build MVP, Measure Value, and Learn and Update:** these stages are adaptations of the Modeling, Evaluation, and Deployment phases of CRISP-DM. The Lean Startup is applied through the construction of a MVP

(Minimum Viable Product), which is the simplest version of the solution. It is quickly implemented and tested by the users. Finally, the solution is incremented.

## 5.2 Addition of Phases

**Table 6** presents the new phases of the two models, comparing them with the original CRISP-DM phases. Four models included new steps focused on their business domain, but the purpose of the existing steps was kept unaltered.

### DMME

Huber *et al.* [2018] adapts the model process CRISP-DM, including three new stages.

- **Technical Understanding:** the business goals are transformed into measurable technical objectives, gathering documentation, and developing an experiment plan for measurement through sensors, interfaces, or software.
- **Technical Realization:** the technical tests are executed, the suitable data acquisition method is selected, and the experiment plan is conducted before the data processing.
- **Technical Implementation:** the infrastructure capable of executing models in real time is implemented for the deployment phase.

DMME facilitates building a model to detect parts being transported on a guide cart in an industry. The team uses the new phases of DMME to understand the machines and production processes, then installed the hardware and software on the machines.

### DPM

Diop *et al.* [2017] proposed a model based on CRISP-DM with the inclusion of two phases that manage temporal data. This model also integrates with Software Engineering activities such as the specification and development of software.

- **Prior Temporal Data Understanding:** the temporal data is identified and data requirements are defined. The conceptual schema of data is evaluated and temporal data is distinguished from general data.
- **Prior Temporal Data Preparation:** the logical data model and the temporal data list are validated and the data requirements are implemented. Next, mining software is implemented to be used in the data understanding and preparation phases.

DPM incorporates temporal data requirements (e.g., inmate incarceration history, incident records, and prisoner visit logs) during the software design phases of a prison administration software. The project utilized CRISP-DM to guide its data science activities.

**Knowledge Creation + CRISP-DM**

van der Voort *et al.* [2021] adds three new phases focused on knowledge creation in CRISP-DM.

- **Enlarging Individual Knowledge:** different actors have different sources of knowledge. This activity is a continuous process of learning and capacitation.
- **Sharing Tacit Knowledge:** the knowledge is shared with other individuals of the same organizational role. This stage creates knowledge for the Business Understanding phase of CRISP-DM.
- **Networking Knowledge:** the acquired knowledge must be disseminated throughout the organization so that it can be used by all stakeholders.

Knowledge Creation + CRISP-DM categorizes companies based on risk factor. In addition to the traditional CRISP-DM phases, the project incorporated two additional Knowledge Management stages. As a result, inspectors and data analysts actively collaborated in identifying risk factors and participated in workshops to share knowledge and expertise.

**FIN-DM**

Plotnikova *et al.* [2022] adds three new phases that focus on the financial services sector and meeting the compliance requirement.

- **Requirements:** the requirements are defined and managed, including business requirements, technological aspects, and the data mining itself data requirement.
- **Compliance:** the regulatory issue is covered, mitigating risks, and observing GDPR (General Data Protection Regulation) privacy issues.
- **Pos-Deployment:** the activities of monitoring and periodic quality reviews are performed.

**5.3 Addition and Modification of Phases**

Table 7 presents the methods that had phases added and altered at the same time. Besides having a higher complexity with these adaptations, the new process models did not change their purpose and objectives.

**IDAIC**

Ahern *et al.* [2022] renamed and adjusted the six original steps of CRISP-DM for the industrial domain. Besides

CRISP-DM	IDAIC	CRISP-DM for domain experts
Business Understanding	Domain Understanding	Preliminaries
Data Understanding	Data Contextualisation and Assessment	Domain and Data Understanding
Data Preparation	Data Preparation	Data Preparation
-	Operation Assessment	Design of Evaluation and Analytics Pipeline
-	Commissioning	Data Science Training
-	Domain Exploration	-
Modeling	Data Exploration and Algorithm Selection	Implementation of Evaluation and Analytics Pipeline
Evaluation	Results Exploration	Evaluation
Deployment	-	Deployment

Table 7. Phase addition and modification comparison between Original CRISP-DM and 2 models

that, the authors added five more phases to handle the maintenance of equipment.

- The phases **Domain Understanding, Data Contextualization and Assessment, Data Exploration and Algorithm Selection, and Results Exploration** are similar to CRISP-DM phases being renamed for the industrial context.
- The new three phases **Operation Assessment, Commissioning, and Domain Exploration** have complementary activities, such as the detection of equipment degradation and early identification of failures. With this, it is possible to proactively maintain mechanical equipment.

**CRISP-DM for domain expert**

Merkelbach *et al.* [2022] renames three phases of the original CRISP-DM and adds two new phases that allow domain specialists to perform data scientist activities.

- The phases **Preliminaries, Domain and Data Understanding and Design and Implementation of Evaluation and Analytics Pipeline** were altered considering activities for the new roles of trainer data scientist and developer domain specialist.
- In the new steps, the domain specialist gains the abilities and knowledge of data science necessary to implement models.

CRISP-DM for Domain Experts transitions Domain Experts into Data Scientists to optimize the assignment of storage locations for household appliances. In this regard, they engage in new activities, receive technical training, and implement data analysis pipelines under the guidance of data scientists, all of which complement CRISP-DM.

**5.4 Features and Tools Addition**

New features or tools were added in four process models: POST-DS, SCRUM-DS, LDTM, and QM-CRISP-DM.

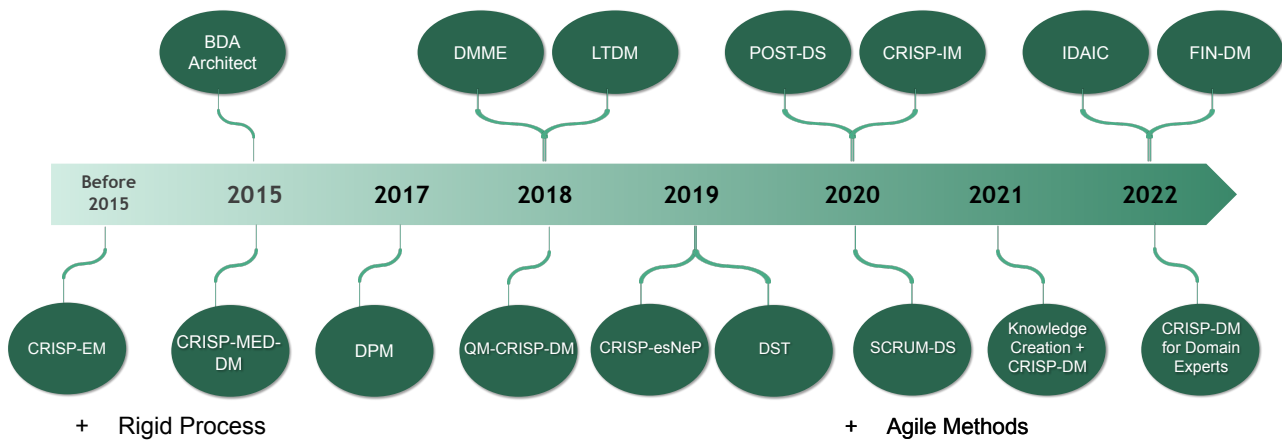


Figure 6. Evolution of CRISP-DM in the last decades

These methods are related to management improvements, such as project, requirement, and quality management. These additions do not change the original CRISP-DM purpose: they only complemented features not covered by CRISP-DM in its original form. The tools or features addition was performed as follows:

- **POST-DS**: adds organization and project management tools, such as RACI matrix and Gantt.
- **SCRUM-DS**: adds artifacts, events, and roles from the SCRUM framework: User Stories, Product Backlog, and Sprint Backlog.
- **LTDM**: adds concepts and features from Design Thinking for management and comprehension of requirements.
- **QM-CRISP-DM**: adds tools for Quality Management of DMAIC and Six Sigma.

### 5.5 Integration with other methods and frameworks

CRISP-DM allows the extension and integration with other models, processes, or frameworks. Therefore, four methods performed integrations with well-known frameworks in the market. These integrations did not change the CRISP-DM purpose, but rather only added features CRISP-DM did not offer in its original form. Here it follows a summary of CRISP-DM integrations with other methods:

- **SCRUM-DS**: integrates CRISP-DM with Scrum, with the use of its artifacts, events, and roles. It includes the events: Planning, Sprints, Daily, Reviews, and Retrospective. Moreover, it includes the roles: Product Owner, Developer, and Scrum Master.
- **QM-CRISP-DM**: integrates and overlaps CRISP-DM’s phases with DMAIC’s steps: Define, Measure, Analyze, Improve, and Control.
- **DPM**: integrates CRISP-DM’s phases with the software engineering process: specification, development, validation, and evolution of the software.
- **BDA Architecture**: integrates and automates CRISP-DM’s activities with Big Data infrastructure.

- **DST**: integrates CRISP-DM with exploratory activities such as exploring business goals, data sources, data value, and results.

Scrum-DS involved interviews with experts from three organizations to discuss how the events, artifacts, and roles of Scrum integrated with CRISP-DM. The demonstration and evaluation were based on these interviews.

QM-CRISP-DM applied CRISP-DM with quality management tools in each of its phases to develop an error prediction system for electronic production processes. For example, it used tools such as SIPOC (Supplier, Input, Process, Output, Customer) to define the scope and objectives in the *Business Understanding* phase.

BDA Architecture showcases the integration of Nested Automatic Service Composition (NASC) with CRISP-DM and the utilization of big data to automate the process of analyzing flight delay data from an airline.

DST presents seven practical examples of exploratory activities in different data science projects: 1) in Tourism Recommender, to understand tourist location and behavior data; 2) in Environment Simulator, to simulate environmental scenarios and test hypotheses; 3) in Insurance Refining, to refine policies in insurance models and evaluate risk data; 4) in Sales OLAP, to enhance data warehouse construction; 5) in Publication Repository, to analyze citation patterns and publication impact; 6) in Parking APP, to analyze parking usage data and enhance user experience; and 7) in Payment Geovisualization, to explore tourist spending data. These activities complement those of CRISP-DM.

Figure 6 represents how CRISP-DM has evolved over the last decades. The first models, such as CRISP-EM and CRISP-MED-DM, were rigid and used in sequential and linear forms. Since 2017, the derivative models are more flexible, used with agile methodologies, such as SCRUM-DS (with SCRUM elements) and LTDM (with Design Thinking and Lean Startup elements).

### 5.6 Discussion

This Systematic Literature Review showed that CRISP-DM is a process widely adopted in Data Science projects. It can

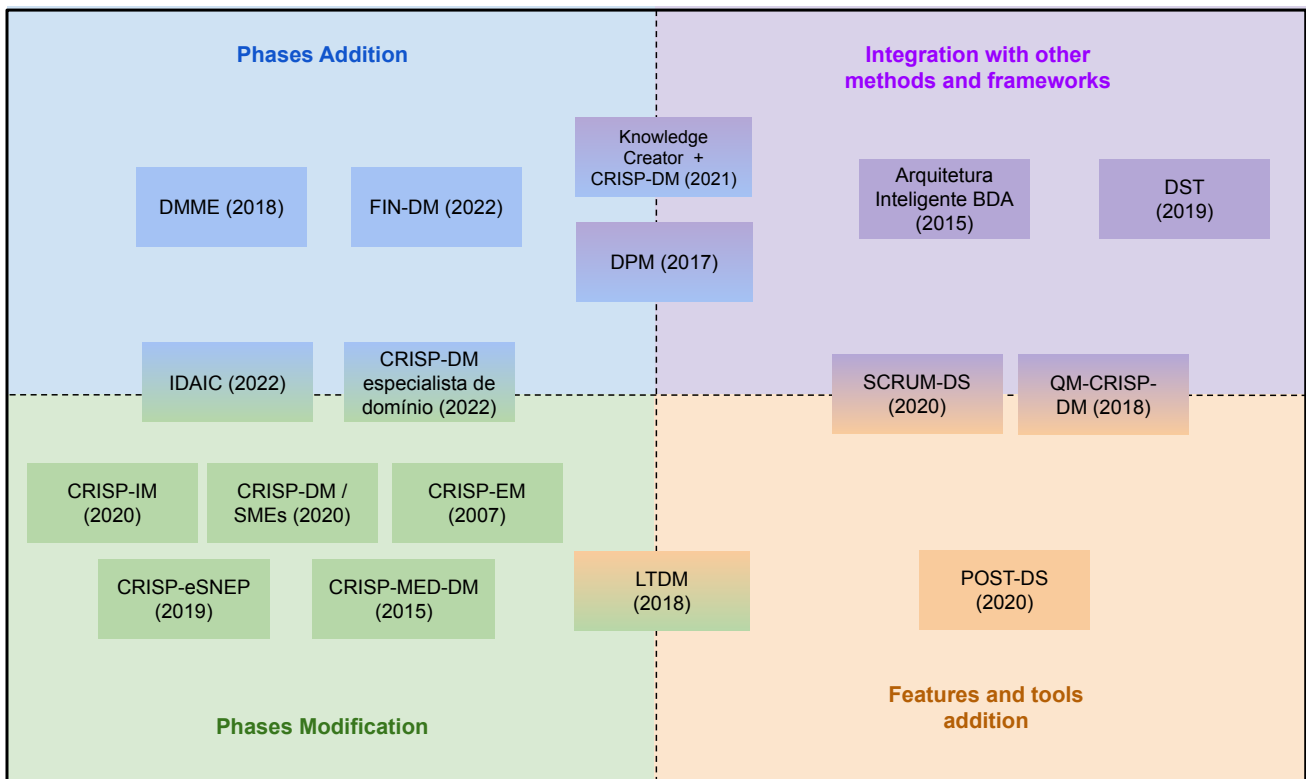


Figure 7. Mapping of the process models based on CRISP-DM that were adapted or extended

be adapted to various business domains. Allowed us to point out the similarities and differences of the aforementioned derivatives CRISP-DM.

CRISP-DM does not meet the current agile principles and fundamentals. There are few empirical studies about agility in Data Science within organizations [Ahmed *et al.*, 2018; Saltz and Suthrland, 2019]. Agility is an approach modestly explored in Data Science.

Nevertheless, CRISP-DM is adaptable and can be integrated with agile methods. For example, Scrum-DS, adds elements from Scrum into CRISP-DM by seeking agility in the form of working and management. However, it does not present features dedicated to software engineering.

On the other hand, DPM integrates itself with the traditional software engineering life cycle, but does not bring agile elements and practices. It was not possible to find any study that mixes Extreme Programming (XP) with CRISP-DM.

Based on the SLR, it is possible to answer the research question:

**How is the process model CRISP-DM adapted or utilized with other methodologies in Data Science projects?**

CRISP-DM is adaptable, and its phases can be used in their original form, applied to a specific context, or integrated with new approaches and practices. In addition, some studies use CRISP-DM with agility, showing that it can be compatible with the agile principles, while it can be integrated with the

software engineering discipline.

Based on the theoretical foundation and the systematic literature review, this research proposes a theoretical map for CRISP-DM adaptations (Figure 7). In the quadrant 1, there are the methods with new phases. In the quadrant 2, there are integrations with other methods or frameworks. In the quadrant 3, there are methods with phase changes. In the quadrant 4, there are methods that were complemented with resources and tools. There are methods that cover more than one quadrant, for example, the DPM which participates from quadrant 1 (Phases Addition) and quadrant 2 (Integration with other methods and frameworks). In Figure 7, these methods can be found between the two quadrants.

## 6 Threats to validity

### 6.1 Internal Threats

The main limitations of this systematic literature review are related to the inaccuracy of the data extraction and the study selection. During the data extraction, it was observed that articles have different levels of detail and do not have a pattern. This can cause insufficient information and deficiencies in the analysis of the studies. To reduce this issue, it was defined a data extraction form to standardize the format, level of detail, and to facilitate and comparison between studies.

Although a carefully crafted search string was employed, some studies may have been omitted because they did not contain specific terms within the defined scope. To address this, the search string underwent iterative validation and adjustment. However, due to inherent terminology variability,



some relevant studies might not have been included.

## 6.2 External Threats

To enable the impartiality of the systematic literature review, a research protocol was developed including the research objective and question, keywords, search string, and selection criteria. However, an external validity threat emerges due to the lack of standardization in keywords within software engineering and data science, which can introduce ambiguities. Therefore, even with a structured protocol, there is a risk that studies have been omitted. To mitigate this risk, the execution was documented with the search strategies applied in well-known bibliographic resources in the science computer field. Moreover, grey literature was avoided. Finally, a manual search step was performed observing the reference and citations of the works, to add new studies relevant to the research.

## 7 Conclusion

17 studies were found in the SLR. These studies highlighted new processes derived from CRISP-DM, which underwent adaptations for specific contexts and business domains. Therefore, they provided resources to answer the research question "How is the process model CRISP-DM adapted or utilized with other approaches in Data Science projects?".

The adaptations identified were phases modifications (Section 5.1), new phases addition (Section 5.2), phases addition and modifications (Section 5.3), new features and tools addition (Section 5.4), and integration with other methods (Section 5.5). These adaptations did not change the purpose of CRISP-DM.

It is also important to consider how other process models have influenced the historical and current of CRISP-DM. KDD, as a precursor, influenced the development of CRISP-DM in the 1990s. SEMMA emerged alongside CRISP-DM to support SAS tool, but had limitations in business understanding. During the 2000s, there was a growth in agile methods, such as Scrum, XP, and Kanban, as well as concepts like Design Thinking and MVP, which inspired adaptations of CRISP-DM, including Scrum-DS and LTDM. TDSP, inspired by CRISP-DM, introduced greater iterativity and adaptability to the agile context. These developments underscore how CRISP-DM remains a robust foundation, adaptable to the evolving needs of the data science field.

Data Science projects are compatible with agile principles and software engineering practices. However, the systematic literature review shows a lack of studies covering agility and software engineering in Data Science. The adoption of CRISP-DM with agile practices like Extreme Programming is seldom explored. Therefore, there is a need to increase the quantity and quality of studies about agility in Data Science.

The research goal to evaluate how the method CRISP-DM adapts to other approaches was accomplished. Besides, it was proposed a theoretical reference that makes it possible to explore these CRISP-DM adaptations and can be used to as-

sist in the comparison with other studies about process structure in Data Science.

In future studies, based on the theoretical reference presented, it is suggested to conduct empirical studies in a real corporate environment to measure the effectiveness and benefits of using CRISP-DM with agile practices and software engineering. Regarding the systematic literature review, its standardized protocol can be reproduced again for future studies.

## Declarations

### Acknowledgements

We thank Andrea Tiemi Abe Shimaoka for her assistance with the English language translation.

### Authors' Contributions

All authors contributed to the writing and reviewing of the article, as well as creating tables and figures to facilitate the understanding of the scientific study.

### Competing interests

The authors declare that they have no competing interests.

### Availability of data and materials

The datasets generated and/or analyzed during the current study are available at <https://doi.org/10.5281/zenodo.12753088>.

## References

- Ahern, M., O'Sullivan, D. T. J., and Bruton, K. (2022). Development of a framework to aid the transition from reactive to proactive maintenance approaches to enable energy reduction. *Applied Sciences*, 12(13). DOI: 10.3390/app12136704.
- Ahmad, N., Hamid, A., and Ahmed, V. (2022). Data science: Hype and reality. *Computer*, 55(2):95–101. DOI: 10.1109/MC.2021.3130365.
- Ahmed, B., Dannhauser, T., and Philip, N. (2018). A lean design thinking methodology (ldtm) for machine learning and modern data projects. In *2018 10th Computer Science and Electronic Engineering (CEECE)*, pages 11–14. DOI: 10.1109/CEECE.2018.8674234.
- Anderson, D. J. (2010). *Kanban: Successful Evolutionary Change for Your Technology Business*. Blue Hole Press, Seattle, WA. Available at: <https://personalpages.bradley.edu/~young/CS690S117old/Kanban.pdf>.
- Asamoah, D. A. and Sharda, R. (2019). Crisp-esnep: Towards a data-driven knowledge discovery process for electronic social networks. *Journal of Decision Systems*, 28(4):286–308. DOI: 10.1080/12460125.2019.1696614.



- Ayele, W. Y. (2020). Adapting crisp-dm for idea mining: A data mining process for generating ideas using a textual dataset. *International Journal of Advanced Computer Science and Applications*, 11(6). DOI: 10.14569/IJACSA.2020.0110603.
- Azevedo, A. and Santos, M. (2008). Kdd, semma and crisp-dm: A parallel overview. In *IADIS European Conf. Data Mining*, pages 182–185. Available at: <https://recipp.ipp.pt/handle/10400.22/136>.
- Baijens, J., Helms, R., and Iren, D. (2020). Applying scrum in data science projects. In *2020 IEEE 22nd Conference on Business Informatics (CBI)*, volume 1, pages 30–38. DOI: 10.1109/CBI49978.2020.00011.
- 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. Available at: <https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf>.
- Costa, C. J. and Aparicio, J. T. (2020). Post-ds: A methodology to boost data science. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. DOI: 10.23919/CISTI49556.2020.9140932.
- Diop, M., Camara, M. S., Fall, I., and Bah, A. (2017). A methodology for prior management of temporal data quality in a data mining process. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–8. DOI: 10.1109/ISACV.2017.8054906.
- Dåderman, A. and Rosander, S. (2018). Evaluating frameworks for implementing machine learning in signal processing: A comparative study of crisp-dm, semma and kdd. *DiVA Portal*. Available at: <https://www.diva-portal.org/smash/get/diva2:1250897/FULLTEXT01.pdf>.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37. DOI: 10.1609/aimag.v17i3.1230.
- Goyal, D., Goyal, R., Rekha, G., Malik, S., and Tyagi, A. (2020). Emerging trends and challenges in data science and big data analytics. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–8. DOI: 10.1109/ic-ETITE47903.2020.316.
- Hoda, R., Salleh, N., and Grundy, J. (2018). The rise and evolution of agile software development. *IEEE Software*, 35(5):58–63. DOI: 10.1109/MS.2018.290111318.
- Huber, S., Wiemer, H., Schneider, D., and Ihlenfeldt, S. (2018). Dmme: Data mining methodology for engineering applications – a holistic extension to the crisp-dm model. *Procedia CIRP*, 79:403–408. DOI: 10.1016/j.procir.2019.02.106.
- IDC (2020). The digitization of the world – from edge to core. Available at: <https://www.seagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf>. Last checked on Aug 20, 2024.
- Julian, Brendan and Noble, J. and Anslow, C. (2019). Agile practices in practice: Towards a theory of agile adoption and process evolution. In *Agile Processes in Software Engineering and Extreme Programming*, pages 3–18, Cham, Springer International Publishing. DOI: 10.1007/978-3-030-19034-7\_1.
- Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report. Available at: [https://legacyfileshare.elsevier.com/promis\\_misc/525444systematicreviewsguide.pdf](https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf).
- Larson, D. and Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5):700–710. DOI: 10.1016/j.ijinfomgt.2016.04.013.
- Manirupa, Cui, R., Campbell, D. R., Agrawal, G., and Ramnath, R. (2015). Towards methods for systematic research on big data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2072–2081. DOI: 10.1109/BigData.2015.7363989.
- Mariscal, G., Marban, O., and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2):137–166. DOI: 10.1017/S0269888910000032.
- Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., and Flach, P. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061. DOI: 10.1109/TKDE.2019.2962680.
- Matharu, G. S., Mishra, A., Singh, H., and Upadhyay, P. (2015). Empirical study of agile software development methodologies: A comparative analysis. *SIGSOFT Softw. Eng. Notes*, 40(1):1–6. DOI: 10.1145/2693208.2693233.
- Merkelbach, S., Von Enzberg, S., Kühn, A., and Dumitrescu, R. (2022). Towards a process model to enable domain experts to become citizen data scientists for industrial applications. In *2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS)*, pages 1–6. DOI: 10.1109/ICPS51978.2022.9816871.
- Microsoft (2024). Tdsp - processo de ciência de dados de equipe. Available at: <https://learn.microsoft.com/pt-br/azure/architecture/data-science-process/lifecycle>.
- Montalvo-Garcia, J., Quintero, J., and Manrique, B. (2020). *CRISP-DM/SMEs: A Data Analytics Methodology for Non-profit SMEs*, pages 449–457. Springer Singapore. DOI: 10.1007/978-981-15-0637-6\_38.
- Nakagawa, E., Scannavino, K., Fabbri, S., and Ferrari, F. (2017). *Revisão Sistemática da Literatura em Engenharia de Software: Teoria e Prática*. Elsevier Brasil. Book.
- Niaksu, O. (2015). Crisp data mining methodology extension for medical domain. *Baltic J. Modern Computing*, 3:92–109. Available at: [https://www.bjmc.lu.lv/fileadmin/user\\_upload/lu\\_portal/projekti/bjmc/Contents/3\\_2\\_2\\_Niaksu.pdf](https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/3_2_2_Niaksu.pdf).
- Palacios, H. J. G., Toledo, R. A. J., Pantoja, G. A. H., and Navarro, A. M. (2017). A comparative between crisp-dm and semma through the construction of a modis repos-

- itory for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems Journal*, 2(3):598–604. DOI: 10.25046/aj020376.
- Petticrew, M. and Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley. Book.
- Plotnikova, V., Dumas, M., Nolte, A., and Milani, F. (2022). Designing a data mining process for the financial services domain. *Journal of Business Analytics*, 0(0):1–27. DOI: 10.1080/2573234X.2022.2088412.
- Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59. DOI: 10.1089/big.2013.1508.
- Riungu-Kalliosaari, L., Kauppinen, M., and Männistö, T. (2017). What can be learnt from experienced data scientists? a case study. In *Product-Focused Software Process Improvement*, pages 55–70, Cham. Springer International Publishing. DOI: 110.1007/978-3-319-69926-4.
- Saltz (2020). Crisp-dm is still the most popular framework for executing data science projects. Available at: <https://www.datascience-pm.com/crisp-dm-still-most-popular>.
- Saltz, J. and Suthrland, A. (2019). Ski: An agile framework for data science. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3468–3476. DOI: 10.1109/BigData47090.2019.9005591.
- Saltz, J. S., Shamshurin, I., and Crowston, K. (2017). Comparing data science project management methodologies via a controlled experiment. In *Hawaii International Conference on System Sciences*. DOI: 10.24251/HICSS.2017.120.
- SAS (2003). *Data Mining Using SAS Enterprise Miner: A Case Study Approach*. SAS Publishing, 2nd edition. Available at: [https://support.sas.com/documentation/onlinedoc/miner/casestudy\\_59123.pdf](https://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf).
- Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534. DOI: 10.1016/j.procs.2021.01.199.
- Schwaber, K. and Sutherland, J. (2020). *The Scrum Guide*. Scrum.org. Available at: <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf>.
- Schäfer, F., Zeiselmaier, C., Becker, J., and Otten, H. (2018). Synthesizing crisp-dm and quality management: A data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 190–195. DOI: 10.1109/ITMC.2018.8691266.
- Siriweera, T., Paik, I., Kumara, B. T., and Koswatta, K. (2015). Intelligent big data analysis architecture based on automatic service composition. In *2015 IEEE International Congress on Big Data*, pages 276–280. DOI: 10.1109/BigDataCongress.2015.46.
- van der Voort, H., van Bulderen, S., Cunningham, S., and Janssen, M. (2021). Data science as knowledge creation a framework for synergies between data analysts and domain professionals. *Technological Forecasting and Social Change*, 173:121160. DOI: 10.1016/j.techfore.2021.121160.
- Venter, J., de Waal, A., and Willers, C. (2007). Specializing crisp-dm for evidence mining. In *Advances in Digital Forensics III*, pages 303–315, New York, NY. Springer New York. DOI: 10.1007/978-0-387-73742-3\_21.
- Versionone (2020). Crisp-dm is still the most popular framework for executing data science projects. Available at: <https://info.digital.ai/rs/981-LQX-968/images/S0A14.pdf>.
- Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. Available at: <https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>.
- Yang, L., Zhang, H., Shen, H., Huang, X., Zhou, X., Rong, G., and Shao, D. (2021). Quality assessment in systematic literature reviews: A software engineering perspective. *Information and Software Technology*, 130:106397. DOI: 10.1016/j.infsof.2020.106397.