INSTITUTO
UNIVERSITÁRIO
DE LISBOA

ETL for Data Science? A Case Study.

Nicole Furtado Oliveira

Master Degree in Computer Engineering

Supervisor:
PhD Luís Miguel Martins Nunes, Associate Professor,
ISCTE-IUL - Lisbon University Institute

Co-supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor,
ISCTE-IUL - Lisbon University Institute

November, 2021

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

ETL for Data Science? A Case Study.

Nicole Furtado Oliveira

Master Degree in Computer Engineering

Supervisor:
PhD Luís Miguel Martins Nunes, Associate Professor,
ISCTE-IUL - Lisbon University Institute

Co-supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor,
ISCTE-IUL - Lisbon University Institute

November, 2021

I dedicate this dissertation to my family,

who supported me

with unconditional love

throughout my life

Acknowledgement

A heartily thank you to Professor Luís Nunes, who has invited me to be part of the project that set the stage for this dissertation. He has been a teacher, a workmate, an advisor and a friend that has always kindly and consistently been there for me whenever I needed.

A thank you to Professor Elsa Cardoso, who had an important role in the theme specification of this study and firmly helped me to set up for success.

A very special thank you to kindhearted Susana Fernandes who patiently worked close to me in the project, brainstormed with me some ideas and inspired me to produce this dissertation the way it is.

A special thank you to Professor Raul Laureano, who took time to give me counseling and helped me build the trust I needed to overcome some obstacles in a moment of need.

A loving thank you to my parents that were the best as they could be to make all of who I am possible.

A loving thank you to my dearest friend Tiago Mileu who has given me support and strength in the last years of my studies.

A thank you to my godfather who made me feel proud of myself to have come this far.

A grateful thank you to my friend Hugo Henriques that guided me to come to ISCTE, my home for the last five years, and to pursue a career as a computer engineer.

Also, a thank you to all the people I have not mentioned but who played an important part in my life path that brought me here: family, friends, teachers, ISCTE colleagues and work colleagues.

Last but not least, a loving thank you to God to have put together all the necessary conditions to allow me to learn and grow, to be me and to be here.

This work was partly funded through national funds by FCT - Fundação para a Ciência e Tecnologia, I.P. under projects UIDB/EEA/50008/2020 (Instituto de Telecomunicações) and UIDB/04466/2020 (ISTAR).

Resumo

A *big data* tem impulsionado o desenvolvimento e a pesquisa da ciência de dados nos últimos anos. No entanto, há um problema - a maioria dos projetos de ciência de dados não chega à produção. Isto pode acontecer porque muitos deles não usam uma metodologia de ciência de dados de referência. Outro elemento agravador são os próprios dados, a sua qualidade e o seu processamento. O problema pode ser mitigado através da documentação de estudos de caso, pesquisas e desenvolvimento da área, nomeadamente o reaproveitamento de conhecimento de outros campos maduros que exploram questões semelhantes, como *data warehousing*. Para resolver o problema, esta dissertação realiza um estudo de caso sobre o projeto "IA-SI - Inteligência Artificial na Gestão de Incentivos", que visa melhorar a gestão dos fundos europeus de investimento através de *data mining*. As principais contribuições deste estudo, para a academia e para o desenvolvimento e sucesso do projeto são: (1) Um modelo de processo combinado dos modelos de processo de *data mining* mais usados e as suas tarefas, ampliado com os subsistemas de ETL e outras recomendadas práticas de *data warehousing* selecionadas. (2) Aplicação deste modelo de processo combinado ao projeto e toda a sua documentação. (3) Contribuição para a implementação do protótipo do projeto, relativamente a tarefas de compreensão e preparação de dados. Este estudo conclui que CRISP-DM ainda é uma referência, pois inclui todas as tarefas dos outros modelos de processos de *data mining* e descrições detalhadas e que a sua combinação com as melhores práticas de *data warehousing* é útil para o projeto IA-SI e potencialmente para outros projetos de *data mining*.

Palavras-chave: *data mining*; compreensão do negócio, compreensão de dados; preparação de dados; extração de dados; transformação de dados; limpeza de dados; normalização de dados; *pipeline*; ETL; *data warehousing*; modelos de processo; metodologias; CRISP-DM; KDD; Python; gestão de fundos de investimento; fundos de investimento para o setor privado; Fundos Estruturais e de Investimento Europeus; *machine learning*; inteligência artificial; ciência de dados; caso de estudo.

Abstract

Big data has driven data science development and research over the last years. However, there is a problem - most of the data science projects don't make it to production. This can happen because many data scientists don't use a reference data science methodology. Another aggravating element is data itself, its quality and processing. The problem can be mitigated through research, progress and case studies documentation about the topic, fostering knowledge dissemination and reuse. Namely, data mining can benefit from other mature fields' knowledge that explores similar matters, like data warehousing. To address the problem, this dissertation performs a case study about the project "IA-SI - Artificial Intelligence in Incentives Management", which aims to improve the management of European grant funds through data mining. The key contributions of this study, to the academia and to the project's development and success are: (1) A combined process model of the most used data mining process models and their tasks, extended with the ETL's subsystems and other selected data warehousing best practices. (2) Application of this combined process model to the project and all its documentation. (3) Contribution to the project's prototype implementation, regarding the data understanding and data preparation tasks. This study concludes that CRISP-DM is still a reference, as it includes all the other data mining process models' tasks and detailed descriptions, and that its combination with the data warehousing best practices is useful to the project IA-SI and potentially to other data mining projects.

Keywords: data mining; business understanding, data understanding; data preparation; data extraction; data transformation; data cleaning; data normalization; pipeline; ETL; data warehousing; process models; methodologies; CRISP-DM; KDD; Python; grant funds management; grant funds for the private sector; European Structural and Investment Funds; machine learning; artificial intelligence; data science; case study.

Index

Abbreviations and Acronyms

AI	Artificial Intelligence
AICEP	Agency for Investment and Foreign Trade of Portugal
	(PT - Agência para o Investimento e Comércio Externo de Portugal)
BI	Business Intelligence
DM	Data Mining
DS	Data Science
ETL	Extract-transform-load
FCT	Foundation for Science and Technology
	(PT - Fundação para a Ciência e a Tecnologia)
IAPMEI	Institute for the Support of Small and Medium Enterprises and Innovation
	(PT - Instituto de Apoio às Pequenas e Médias Empresas e à Inovação)
IES	Simplified Business Information
	(PT - Informação Empresarial Simplificada)
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
NIF	Tax Identification Number
	(PT - Número de Identificação Fiscal)
ODBMS	Operational Database Management Systems

Index of Figures

Index of Tables

CHAPTER 1

1 Introduction

This chapter presents the case study and its context, the goals and the motivation of this dissertation and the structure of the rest of the document.

1.1 The Case and it's Context

This dissertation performs a case study about the project "IA-SI - Artificial Intelligence in Incentives Management" (*IA-SI - Inteligência Artificial na Gestão de Incentivos*), from here onward simply called IA-SI. The project aims to improve the management of European grant funds and the control of the supported investments [1] through data mining (DM). Grant funds are financial help/ incentives destined to business investment [2]. To be able to obtain funding, businesses apply with a project describing the intended investments. Applications are carefully analysed by institutions that manage the grant funds and only some are selected/ accepted to receive funding.

The accepted projects, also called eligible, must present expenses that justify the granting of the incentive. Thus, one of the project's DM goals is to predict expenses that don't justify the granting of the incentive, that is, ineligible expenses. An ineligible expense can be, for example, an investment that can't be framed in the project's context.

The other DM goal is to predict project cancelation. A project can be cancelled, for example, if candidates don't comply with their application proposal or with the funding program terms.

The author of this dissertation has been integrated in the project's team for 15 months, alongside with other researchers and field experts. She is responsible for the data preparation for the prediction of ineligible expenses, most of the data understanding for both DM goals, part of the evaluation for both DM goals and related documentation. Therefore, and for scope management reasons, although more context is provided, the focus of this dissertation in on the data understanding and preparation for the DM problem of the prediction of ineligible expenses.

The project is under the Public Administration Digital Transformation Support System (SAMA - *Sistema de Apoio à Transformação Digital da Administração Pública*) and is financed with aproximately €300,000 by Compete grant funds [1], which is a managing authority for granting European funds [3], [4].

This project is innovative because:

(1) There is little application of DM for managing European grant funds in Europe [1]. [5]–[7] is a similar project (more about this in section 2.2).

(2) Although, currently, in Portugal, there are projects granted with funds that aim to modernize the Portuguese public administration using data science (DS) and artificial intelligence (AI) [1], [8] there is little known use of AI to the evaluation and monitoring of other types of investment projects.

European grant funds, sometimes called community funds [9], are important for economic growth and concord with the Portuguese constitution's objectives, therefore, they must have effective and efficient management and control, in order to enhance the benefits of their application. To this end, the project emerges as a joint initiative of IAPMEI (Institute for the Support of Small and Medium Enterprises and Innovation/ PT - *Instituto de Apoio às Pequenas e Médias Empresas e à Inovação*) [10] and AICEP (Agency for Investment and Foreign Trade of Portugal/ PT - *Agência para o Investimento e Comércio Externo de Portugal*) [9], having ISCTE – Lisbon's University Institute/ PT - *Instituto Universitário de Lisboa* in the technical-scientific role [1].

IAPMEI [10] and AICEP [9] are intermediate institutions [11] that manage the European grant funds destined to Portugal. They receive, yearly, hundreds of applications from companies presenting their projects. After the project is analysed, the decision is communicated to the company: it is either accepted or rejected. If accepted, the company submits over time several payment requests which include the expenses that justify granting the incentive. During the entire process, technicians from IAPMEI and AICEP carry out a careful analysis so that the funds are granted to companies with the most potential and that present proof of their correct application. Some projects can be cancelled if they don't comply with their application proposal or with the funding program terms, for example.

Because of the deluge of submitted applications, there have been delays in the communication of the project's eligibility or in the grants' payment to applicants [1].

The project IA-SI, in response to this and other challenges, aims to:

- "Improve the level of service and responsiveness to companies" [2];
- "Provide a less bureaucratic and more efficient service to companies, streamlining the relationship with companies by reducing the bureaucratic burden and reducing response times to payment requests, strengthening companies' trust in the State" [2];
- "Improve efficiency and reliability in project evaluation" [2];

Through the actions:

- Analysis of payment requests from beneficiary companies "using machine learning techniques" [2];
- Prevention and detection of situations where the grant funds are not used correctly [1];
- Identification of risk patterns in the data [1];
- Generation of risk scores to enable decision making.

And having as pillars:

1. “Increase in the degree of fulfilment of contracted objectives” [2];
2. Reduction of the administrative burden of projects that do not present a risky profile (the majority) and reduction of incidents that imply the devolution of granted funds [2];
3. “Increased reliability of execution and possibility of preventive actions to avoid these occurrences” [2].

Hereupon, the management of grant funds and the control of the investments is improved, encouraging a more ethical distribution of the grant funds. Thereat, this dissertation contributes to the United Nation’s “Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all” Sustainable Development Goal (goal 8) [12].

The Portuguese constitution asserts that it is up to the state to “promote the increase of social and economic well-being and the quality of life of people (…) within the framework of a sustainable development strategy” and to encourage and support “scientific creation and research, as well as technological innovation (...) to ensure (...) the strengthening of competitiveness and the articulation between scientific institutions and companies”. The industrial policy, included in the government's economic and social development plans, aims to “support small and medium-sized companies and, in general, initiatives and companies that generate employment and encourage exports or import substitution” and the “support for the international projection of Portuguese companies” [13, Sec. Article 81.º a), Article 73.º 4., Article 100.º d)].

To help meet these and other goals in line with the European strategy, namely the Sustainable Development Goals [12], Portugal is granted European Structural and Investment Funds [14]. These funds are managed in accordance with an agreement between Portugal and the European Commission, “which defines how the funds will be used during the current funding period” [15, Sec. How the funds are managed]. It is legislated that the management of these funds can be carried out by intermediary institutions [11], such as FCT (Foundation for Science and Technology/ PT - *Fundação para a Ciência e a Tecnologia*) [16], IAPMEI and AICEP.

These intermediate institutions receive applications from companies, in the case of IAPMEI and AICEP, and from researchers, in the case of FCT, whose purpose is to be granted part of these funds for their subsequent application. In this way, research and development [16] and the empowerment of companies are promoted, increasing competitiveness, business growth, internationalization, innovation and entrepreneurship [9], [10].

1.2 Goals

Given the IA-SI project's importance and interdisciplinarity, it presents itself as the ideal context for research, as this masters' dissertation. Also, research can contribute to the project's development and success. To this end, this study has the following goals:

1. Understand which is the most comprehensive DM process model or methodology to be the reference followed in the project.
2. Discover if and how can the data warehousing best practices be applied in a DM problem and extend the reference DM process model or methodology.
3. Document the application of the data warehousing best practices in the project, so that it can be shared to other researchers and practitioners.
4. Develop software prototype components related to data understanding and preparation tasks for the ineligibilities DM problem of the IA-SI project.
5. Contribute to the United Nation's "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all" sustainable development goal (goal 8) [12] through the support of the project's development, that aims to achieve a better and more ethical grant funds management distribution.

To attain these goals, this study proposes the following actions:

- Gather the most used DM process models and methodologies that appear in the literature.
- Study, compare and understand if they can be combined into a single DM process model.
- If they can, create a new unified reference DM process model or methodology; if not, choose the most appropriate one to be used in the project.
- Gather and study the data warehousing best practices that can be applied to a real DM problem.
- Understand if they can contribute to extend the previously obtained reference DM process model or methodology.
- If they can, extend the reference DM process model or methodology, apply it in the project and document this process.
- Program scripts in Python related to data understanding and preparation tasks for the ineligibilities DM problem of the IA-SI project.

1.2 Motivation

DS has developed a lot in the last years, partly because of the new challenges of big data, e.g.: how should large quantities of data that come from disparate sources be storaged, processed and analysed [17]. Big data arised with the continuous technological developments that made possible the collection of large quantities of data per multiple devices, like sensors and smartphones [17].

Along with this development, there has been a lot of research in the DS field [17], [18] and a high demand on DS related jobs, that is expected to continue its growth [19]. Despite all of the field's progress, surveys have shown that the majority of the DS projects are never put in production and that it is challenging for companies to keep up with the latest big data and AI trends [17].

The use of a DS methodology can help to achieve the prosperity of the DS project [17], [20], [21]. [20] did a survey yielding that 82% of the data scientists didn't follow a methodology but 85% of them thought that the creation of an improved DS methodology would be beneficial. For the ones that use a methodology, CRISP-DM (Cross Industry Standard Process for DM) has been the most common choice [17], [20], [21], however, [17] suggests that there is a need to review CRISP-DM to understand it's limitations and possible enhancements.

Data itself and its quality are paramount to the DS project success [17]. As such, all the tasks that involve its understanding and preparation deserve special attention. Especially, since data preparation tasks can take up to 90% of the DM project's time [22]. An overview of the most common data preparation processes can be found on appendix E.

Other important factor is the documentation of all the processes so that knowledge can be shared with other people in order to reproduce the undertaken experiences, to further explore the problem or apply the knowledge in other projects [17], [23]. [23] declare that DM "activities can naturally be captured as experience in the form of cases" and that DM "is a domain where few documented cases are available".

Also, [23] state that available DM methodologies "provide very little detailed knowledge for the novice miner on how to actually carry out a given step (…); what a non-specialist really needs are explanations, heuristics and recommendations on how to effectively carry out the particular steps of the methodology. (…) The use of a case-based reasoning approach (by imposing a less formal and structured knowledge representation approach) provides a first step towards fostering knowledge reuse of data mining activities". "Case based reasoning means using old experiences to understand and solve new problems" [24].

According to [25], "standardization of data mining process models should be an essential research line in present and future of data mining and knowledge discovery".

In [26], the author asserts that “serious efforts should be made to develop workable methodologies, techniques, and case studies to promote another round of booming research and development of data mining in real-world problem solving”.

Data warehousing is a mature field that has well established best practices [27]–[29] that can potentially be applied to DM problems, especially if there is a need to clean the source data, which is usually true. The two processes are mostly analogous. In data warehousing, data is collected from disparate sources, cleaned, conformed, transformed and loaded to the data warehouse. This process is called ETL (extract-transform-load). In a DM problem, data usually goes through the same process but, instead of being loaded to the data warehouse, it serves as input to train machine learning models.

Although there are papers that mention ETL in the context of a DM problem (e.g.: [30]–[34]), there, the term ETL is used almost as a synonym of data preparation. This study couldn’t find research where the ETL subprocesses defined in data warehousing practices are systematically applied in a DM problem, nor a process model or methodology that compares or combines data warehousing best practices with the most used DM process models and methodologies.

DM is fundamental to business intelligence (BI) [25] and so is data warehousing [29]. Over time, companies have been investing continuously in BI and it is expected that this area continues to flourish [25].

Kimball’s methodology fits in this research because it is well developed [29] and it has been widespread used with success in thousands of BI projects [28]. More specifically, the focus of this research is on Kimball’s ETL processes because:

(1) “The problems of cleansing data for a data warehouse and for data mining are very similar” [35].
(2) Pyle argues that “data preparation consumes 60 to 90% of the time needed to mine data – and contributes 75 to 90% to the mining project’s success” [22], similarly to ETL that is accounted by Kimball 70% of the effort and risk needed to implement and maintain a DW [36].
(3) Kimball’s methodology has the ETL process well documented and organized [29].

1.3 Document Structure

Chapter two describes the research methodology, related work and the most relevant literature that serves as background not only to undertake the project, but also to inform the reader. Chapter three compares, analyses and extends current DM process models and methodologies with the data warehousing best practices standardized by Kimball. Chapter four documents the project and the case study, where the results from chapter three are applied. Chapter five summarizes the contributions, conclusions and limitations of the dissertation and delineates future work that can be done.

CHAPTER 2

2 Literature Review

2.1 Research Methodology

To select relevant literature, several searches were carried out on Google Scholar that allows to "search all scholarly literature from one convenient place" [37] and on the content aggregator portal EDS - EBSCO Discovery Service, which allows, from a single platform, to conduct searches on thousands of content and metadata providers of approximately 70,000 books and 64,000 magazines [38].

The most important keywords used that yielded better results were:

- extract transform load etl machine learning artificial intelligence models
- extract transform load etl predict predictive analytics
- extract transform load etl data mining
- data mining process models methodologies
- data science process models methodologies
- knowledge discovery
- data science
- machine learning
- artificial intelligence
- etl methodologies
- etl
- data warehouse architectures
- data warehouse data mining
- data preprocessing
- data pre-processing
- data preparation
- european structural and investment funds
- grant funds management
- European Comission tool risk data mining manage structural investment funds

Some examples of content providers included on the research are:

- Ieee Computer Society

- Acm
- Springer
- Elsevier
- ResearchGate
- Institution of Engineering and Technology Journals
- International Journal of Computer Science
- International Journal on Intelligent Data Analysis
- Journal Of Big Data
- Emerald Publishing
- Medwell Journals
- American Statistical Association
- Blackwell Publishing
- Bloomberg, L.P.
- Cambridge University Press
- Oxford University Press
- Public Library of Science
- Canadian Science Publishing

A complementary Google search was undertaken, aimed at finding news or articles about the project IA-SI and some definitions. These were carefully selected only from reliable sources, such as government, press, or other reference websites. By doing this, on one hand, a more complete search was obtained and, on the other, the confidentiality of the project's sensitive information was ensured.

Also, some references were obtained from the references used in some of the read papers.

2.2 Related Work

Countries such as the United States of America, United Kingdom, Japan and Singapore are already successfully using artificial intelligence (AI) in the public sector [39].

"Through a system of neural networks, the UK Department of Work and Pensions was able to automate the processing of refunds and flag fraud situations automatically, ensuring that taxpayers who are effectively entitled to benefits receive them, preventing or significantly reducing fraudulent situations" [30], which shows that AI can help to manage and improve public services.

The use of AI in public administration systems has a direct impact on the lives of citizens due to the public services improvement such as health, social services and justice [39].

These projects aim to “improve services, treating data in such a way that they can be used to prevent problems rather than having to remedy them” and promote social well-being, quality of life and employment [8].

Through the discovery of patterns, large amounts of data available in public administration can be transformed into relevant information [8]. This information is useful for decision making and for improving the efficiency of operational processes, thus becoming valuable knowledge [40].

The IA-SI project is very similar to ARACHNE, an older project developed by the European Commission. ARACHNE is an operational tool that identifies the most risky grant fund beneficiary projects [7]. Its goals are [5], [7]:

- Support grant funds managing authorities.
- Detect and prevent fraud.
- Lower the error rate and fraud.
- Facilitate the continuous monitoring and management of projects.
- Prioritize focus on the most risky projects.

These goals are achieved by the actions [7]:

- Analyse project data (internal) and public external data.
- Provide risk indicators and alerts.
- Feedback loop with tool users to improve its quality.

To be able to use ARACHNE, the institutions have to provide to the European Commission project data (internal) [5]. This data is then complemented with external data by the European Commission and processed by ARACHNE, providing risk indicators [5].

The external data that is used includes the Orbis and the World Compliance databases [7]. Orbis has information about *circa* 100 million companies (general, financial, company related people data, etc.) and *circa* 100 million people (personal data, affinities between people, companies and roles, etc.) [5], [7]. World Compliance has data about politically exposed people (and associated people), sanction lists (like terrorists), enforcement lists (criminals) and adverse media list (companies or people that appeared in the news as linked to illicit activities) [5], [7].

Project data (internal) is provided by the managing authorities [5] and includes information about the partners and the suppliers, type of project, number of employees, project costs, project duration, etc [7].

Similarly to IA-SI, ARACHNE uses data mining to provide to the grant fund managing authorities alerts of fraud risks, conflict of interest and irregularities [5]. This information allows the entities to focus on the most risky projects, which leads to efficiency gains [5].

The feedback loop between managing authorities and the ARACHNE team is useful to signal differences between internal and external data and to manage and improve the overall ARACHNE tool quality [5].

ARACHNE also records operations, which, over time, allows the institutions to assess their performance [5].

The tool is free to be used by the institutions and training and technical support is provided by the European Commission [5].

ARACHNE contributes to improve the efficiency and effectiveness of grant funds management by calculating more than 100 risk indicators, grouped into seven risk categories such as procurement, contract management, eligibility, performance, concentration, other and reputational and fraud alerts [5].

These indicators are used to provide alerts that can help to allocate human resources to the projects that are riskier, helping to prevent, detect and fight against irregularities and fraud [5].

ARACHNE was used and tested in most of the European Member States' managing authorities, including in Portugal's [6], which is Compete [3], [4]. A managing authority is different of the intermediate institutions like IAPMEI and AICEP referenced earlier.

According to [41], managing authorities "are responsible for managing one or several Operational Programmes in accordance with the principle of sound financial management. The managing authority is also ultimately responsible for putting effective and proportionate anti-fraud measures in place taking into account the risk identified".

Operational programmes are "detailed plans in which the member states set out how money from the European Structural and Investment Funds will be spent during the programming period" [42].

The author of this dissertation couldn't find the reason why the use of ARACHNE was not implemented in the Portuguese intermediate institutions, but there is a study made by PWC that has information about ARACHNE's uptake.

According to it (see figure 2.1), only 33% of the managing authorities use ARACHNE.

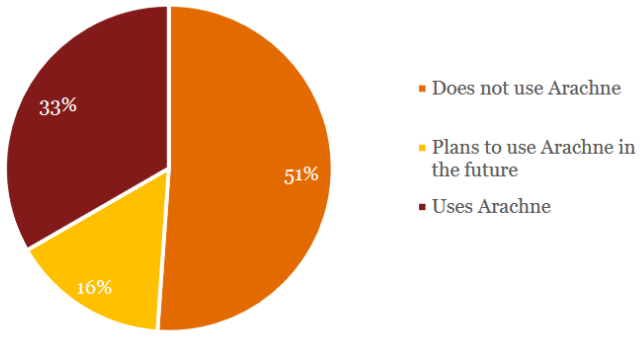

Figure 2.1 Use of ARACHNE across Operational Programmes [41].

Managing authorities shared with PWC some obstacles regarding the use of ARACHNE [41]:

- Data collection and accuracy issues: not all member states provided data and the part of the provided data has delays, which provides uncertainty in the results.
- High number of false positives, which weakens confidence in the results and reliability of the tool.
- Legislative barriers, in particular compliance with national data protection laws.
- The tool is not adjustable to the needs of each member state.
- Using the tool is an extra administrative burden.
- Some managing authorities didn't understand how the tool works and how to interpret the results, as they were not aware of the European commission's on-request available training.

Also, 56% of the managing authorities use other tools and databases that have similar functionalities to Arachne, for example to assess relationships between companies or detect potential fraud related to conflict of interest [41].

Still, PWC didn't find any cross-border, regional or European tools used by authorities that can replace ARACHNE's European coverage [41].

This study couldn't find other tools or projects as similar to IA-SI as ARACHNE. Nevertheless, it is worthwhile to mention that there are some tools/ projects in the area of corruption risk on public procurement [43]:

- The Subsidystories.eu project collects data about how each country member of the European Commission allocates its money from the European Structural and Investment Funds [44].
- The Monitoring European Tenders (MET) is a "risk assessment tool for public authorities to assess the degree of integrity of European public procurement procedures" [45].
- The Red Flags tool "provides an interactive tool that allows the monitoring of procurement processes and their implementation by citizens, journalists or even public officials and catch fraud risks at different stages of the procurement process" [46].
- The Organisation for Economic Cooperation and Development (OECD) public procurement toolbox is an "online resource that collects the best solutions for the prevention of corruption, which strengthen suitable management techniques in public procurement procedures in member countries of the OECD and others" [47].
- "The public procurement Due Diligence Tool published on the Business Anti-Corruption Portal is a tool developed for the assessment and prevention of corruption risks in public procurement" [47].

- "The corruption risk index (CRI) developed by the Corruption Research Centre Budapest (CRCB) is a composite index based on micro data that evaluates public procurement procedures using data of public procurement databases with the help of a complex indicator system" [47].

Regarding the most used data mining (DM) process models and methodologies, the authors of [25] and [48] were essential, covering the evolution up to 2019 between the two. This study couldn't find new or other relevant ones that weren't covered by those authors.

There are two main approaches, from which most of the others get inspiration: KDD (Piatetsky-Shapiro, 1991; Fayyad *et al.*, 1996 a, b) [49] and CRISP-DM (Chapman *et al.,* 2000) [50].

Approaches related to KDD include Cabela *et al.* (Cabena *et al.,* 1997) [48], Human-centred (Brachman & Anand, 1996; Gertosio & Dussauchoy, 2004) [51], Anand & Buchner (Anand & Buchner, 1998; Anand *et al.* 1998; Anand *et al.*, 1999) [52], Two Crows (Two Crows Corporation, 1998; Two Crows Corporation, 1999) [35] and SEMMA (SAS Institute, 2005) [53].

Approaches related to CRISP-DM include Cios *et al.* (Cios *et al.,* 2000; Cios & Kurgan, 2005) [54], RAMSYS (Moyle & Jorge, 2001; Blockeel & Moyle, 2002) [55], DMIE (Solarte, 2002) [56], Data Mining Engineering (Marbán *et al.*, 2007; Marbán *et al.*, 2008) [57], FMDS (IBM, 2015) [58], ASUM-DM (IBM, 2015) [59], [60], TDSP (Microsoft, 2016) [61] and CASP-DM (Martinez-Plumed *et al.*, 2017) [62].

Other approaches that are not related to KDD or CRISP-DM are: 6-σ (Harry & Schroeder, 1999; Pyzdek, 2003) [63], KDD Roadmap (Debuse *et al.*, 2001) [64], 5 A's (SPSS, 1999; de Pisón Ascacibar, 2003) [65] and Refined Data Mining Process (Márban *et al.*, 2010) [25].

There are also other approaches that are meant to be used only in exploratory projects (more about this on section 2.6). They are the D^3M methodology (Longbing Cao, 2010) [26] and the DST framework (Martinez-Plumet *et al.*, 2019) [48] - this one is also related to CRISP-DM.

More information about most used DM and process models and methodologies can be found in sections 2.3 and 2.6 and in appendix C. Appendix C.1 presents the ones related to KDD. Appendix C.2 presents the ones related to CRISP-DM. Appendix C.3 presents other approaches. Exclusively exploratory approaches can be found on appendix D.

2.3 Data Mining, Knowledge Discovery and Other Related Concepts

Historically, there are many terms for the discovery of patterns or models in the data: DM and knowledge extraction or discovery are some of them [49]. The terms knowledge discovery (KD), DM, data science (DS), machine learning, AI and even analytics seem to be used interchangeably nowadays.

The term AI was coined in the 1956 by John McCarthy [66] and initially meant human intelligence being exhibited by machines [67]. Nowadays, the concept has evolved meaning that machines imitate human intelligence [68]. In [69], John McCarthy defined it as ”the science and engineering of making intelligent machines, especially intelligent computer programs”.

The term machine learning was coined in 1959 by Arthur Samuel as the “field of study that gives computers the ability to learn without being explicitly programmed” [70]. It means that machines “exhibits the experiential “learning” associated with human intelligence, while also having the capacity to learn and improve its analyses through the use of computational algorithms” [67].

According to [71], “the science of the use of data” was first mentioned by Naur in 1966. The meaning of the DS concept has changed and evolved through the years and there is no consensual definition [71]. IBM defines DS as a “multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today’s organizations” [19]. It “encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions” [19]. It “combines the scientific method, math and statistics, specialized programming, advanced analytics, AI, and even storytelling to uncover and explain the business insights buried in data” [19].

IBM defines business analytics as “a set of automated data analysis practices, tools and services that help you understand both what is happening in your business and why, to improve decision-making and help you plan for the future” [72]. The term “business analytics” is associated with BI and they are often used together [72].

The term knowledge discovery in databases (KDD) was introduced by Gregory Piatetsky-Shapiro in 1989 [49]. It is the set of activities performed with the goal of obtaining high-level, valued knowledge from low-level data present in large datasets [49]. KDD is an interactive and iterative process [49] that emphasizes “that knowledge is the product of a discovery process guided by data, and it is a joint point of different research areas focused on data analysis and knowledge extraction from different points of view, such as databases, statistics, mathematics, logic or artificial intelligence” [25]. Through the years, many algorithms and tools were developed to aid the KDD process [25].

For the authors of [49], DM is a step in KDD. It´s methods and algorithms are used in KDD to discover useful, non-trivial patterns or models in the data [49]. Other authors describe DM as “a process of identifying interesting patterns in databases that can be used in decision making” or as “a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequently gain knowledge from a large database” and aims to “obtain useful, non-explicit information from data stored in large repositories” [73]. Other author sees DM as a new decision support analysis process, that is iterative and continuous to improve business [74]. Figure 2.2 shows and overview of the KDD process.

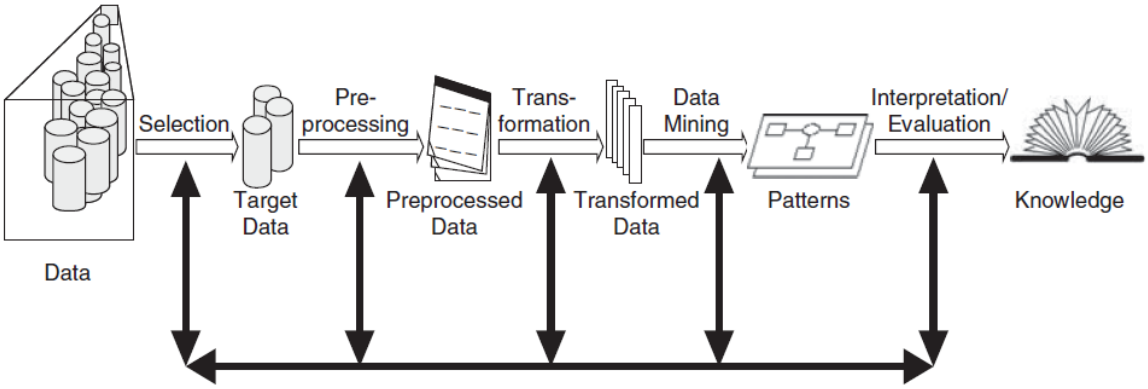

Figure 2.2 Overview of the KDD process [25].

The KDD process involves the following steps [25], [62]:

- Learning the application domain, required prior knowledge and its goals.
- Creating a target data set.
- Data cleaning and pre-processing (removing noise or outliers, deal with missing values, etc.).
- Data reduction and projection (find useful features to represent the data).
- Choosing the function of DM (summarization, classification, regression or clustering).
- Choosing the DM algorithm and the parameters.
- DM (it involves searching for patterns in the classification rules or trees, regression, clustering, sequence modelling, dependency, association rules and line analysis).
- Interpretation (it involves the selection, visualization, interpretation and translation of the discovered patterns into terms understandable by users).
- Using discovered knowledge, by incorporating it into a system to make decisions or simply document it.

KDD is strongly related to databases and data warehousing [49]; "the problems of cleansing data for a data warehouse and for DM are very similar" [35]. In the literature, there are three ways to combine DM and data warehouse (DW) technologies [75]:

1. Integration of DM tools on the DWs front-end GUI.
2. DW technology supporting the DM process by providing efficient database technology with fast responses to queries [76] and high-quality metadata that can be queried upon [77].
3. DM techniques supporting the DW design process.

In 1 and 2, the data has already been cleansed, consolidated and has maintenance procedures programmed for the purpose of building the DW and can be used for DM from the DW [35]. In 3, the data is prepared for DM and only then it is added to the DW.

2.4 Data Warehousing and Business Intelligence

There are two major reference authors that studied DWs: Kimball ([36], [29] and [78]) and Inmon ([79]). The term DW appeared around 1990 in Inmon's work [80].

To Kimball:

- "A data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making" [36]. "It is the necessary platform for business intelligence" [29].

To Inmon:

- "A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The data warehouse contains granular corporate data" [79].

Normally, data in the DW is presented in a dimensional structure (see figure 2.3), that is intuitive to business users and provides fast query performance [29]. Each of the data marts (further explained below) has a dimensional model that is composed of facts (measurements) and dimensions (context – who, what, when, where, why and how) [29]. More about dimensional modeling can be found in [81].

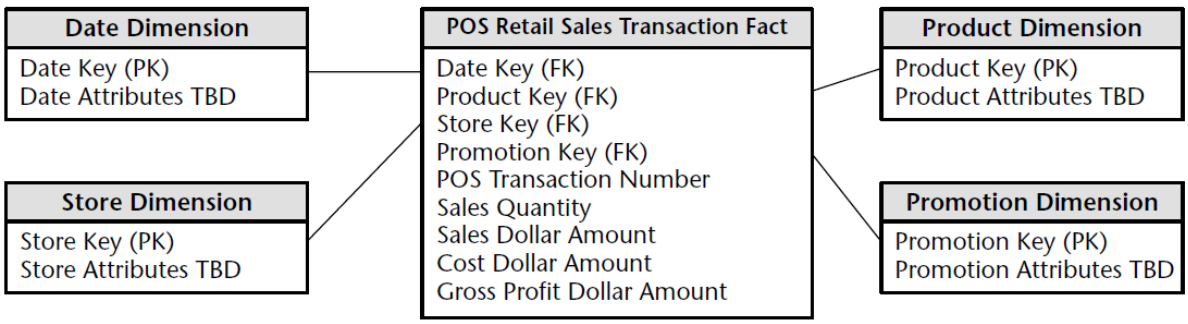

Figure 2.3 Example dimensional model for a POS retail sales business process [81].

The term BI emerged in the 1990s; it is defined as "a process, a product, and as a set of technologies, or a combination of these, which involves data, information, knowledge, decision making, related processes and technologies that support them" [82].

Data warehousing is the process of collecting and cleaning of transactional data [49], presenting it from a business perspective, so that it can be analysed to support strategic decisions and management of the organization [83]. DWs are the foundation of all decision support systems (DSS) processing [79].

DWs require historical data at different granularity levels: at a transaction level and summarized [80]; the transaction level is the most atomic level of the data and summarized data can be monthly sums of atomic data, for example. DWs provide an user-friendly way to users make queries with fast response times [80].

The systems that manage daily business transactions are called operational database management systems (ODBMS) [83].

Analytical queries made in ODBMS would be too heavy would make them slow [84], which would pose a problem because these systems are meant to be responsive [83]. Separating ODBMS from DSS improves performance in both systems [84].

Also, it is very hard to produce analysis or reports that need to get data from many different sources inside a company [79]. Those sources may provide inconsistent information [83], [84], what worsens the task even more. This is resolved by integrating them into a single format in the DW, enabling decision-making [84], credibility and productivity [79]. This integration can include data conversion, reformat, resequence, summary, etc. [79].

Data warehousing can be a long and expensive venture; some of its challenges are [79]:

- Different technologies on the ODBMS and on the DSS.
- Input files may need resequencing.
- Complex extraction.
- Need to perform operation keys restructure.
- Need to perform data reformat (e.g., YYYY/MM/DD to DD/MM/YYYY), data conversion (e.g., change maximum field size, change encoding, etc.) or data cleanse (domain checking, cross-record verification, simple formatting verification, etc.).
- Need to merge data from multiple sources.
- Different sources may have different keys that need to be reconciled.
- May have multiple outputs, with different levels of summarization.
- Default values must be supplied.
- The update with new records can slow down the ODBMS.
- Summarize data from different sources.
- Need to document data renaming.
- Need to unravel undocumented relationships in the source data.

- Parallel reads and parallel loads may need to be designed when the input is too large.
- The data warehousing process must conform to the corporate data model, i.e. to the information needs, systems and applications that are used in the company.
- The DW needs to reflect the history of the information (the time element must be added), while the ODBMS concerns the current state of information.
- The DW needs to provide useful information for decision-making, while the ODBMS needs to provide information to operate the business.
- Need to take in account the logistics of data transformations (i.e., where will the transformations happen) and transmission to the target machine that will host the DW.

Data warehousing can be carried out with in-house built programs or with a purchase of an existing tool; each options has its advantages and disadvantages [36]. If using in-house built programs, the company has to do all the needed ongoing maintenance [79]. The technology that automates data warehousing is called Extract-Transform-Load (ETL) tools or software: in it, the data is retrieved from different sources of data, transformed (conversions, reformats, corrections and integration, e.g.) and loaded into a final database [79].

There are many DW architectures available in the literature [83]. The aim of this study is not to explore this topic in detail, but it's noteworthy that Kimball and Inmon have different perspectives regarding it that can be found in their work and also on other articles like [80] and [83]. It is up to companies to choose the best architectures for their needs [83].

Inmon has a top-down approach where the DW is monolithic and the data is integrated and atomic therein [80]. After the DW is built, smaller departmental databases (or data marts) are created; they contain mostly derived data from the DW (summarized and denormalized, e.g.) and have a multidimensional structure [79]. Departments can be for example: accounting, marketing, engineering and manufacturing [79]. The DW and the data marts are a part of a Corporate Information Factory (CIF) that also comprises DSS applications, exploration warehouses, DM warehouses, alternate storage, etc. [79]. More about the CIF can be found in Imnon's work.

Kimball has a bottom-up approach where the DW is the sum of all data marts [80]. Data marts are organized by business process and their data is conformed between them (bus architecture); they are built one at a time and data in the following must conform to data in the preceding [36]. For this, data is copied from the operational source systems into a staging area, cleaned and integrated / conformed [36]. Different data marts can have different grains (atomic or summarized data), but the atomic form is saved in the staging area to be loaded to the data marts when need be [36].

One big difference between Inmon's and Kimball's data marts is that an Inmon's data mart serves one department, while a Kimball's data mart can serve multiple departments because they can share the interest in a specific business process [80]. Other difference is that Kimball's DW model is faster to deliver the first data mart and the price is more or less the same for the next data marts, while Inmon's DW model is slower to create the monolithic DW, the start-up price is high, but the price is lower for subsequent projects [80].

2.5 ETL and Data Warehousing

This study couldn't find the author of the term ETL (Extract-Transform-Load), but it has been used since 1970 by companies who need to integrate different sources of data into a single database [85]. "The Extract-Transform-Load (ETL) system is the foundation of the data warehouse" [36].

In 1985, the first extract program appeared; it's goal was to search and select data from a file or database based on some given criteria and transporting it to another file or database [79]. The extract program was very successful because, by getting the data out of the source, (1) its performance wasn't affected and (2) the user that retrieved the data had total control over it [79].

In [79], Inmon doesn't provide much detail as Kimball regarding ETL, so this study presents mostly Kimball's perspective regarding ETL. It is noteworthy that, at the end of [79], Inmon presents a development methodology for building a DW, but again he doesn't provide a comprehensive list and description of ETL tasks as Kimball does.

Figure 2.4 shows the sequence of high level tasks required for successful DW / BI projects [78] (more about this topic can be found in [29]).

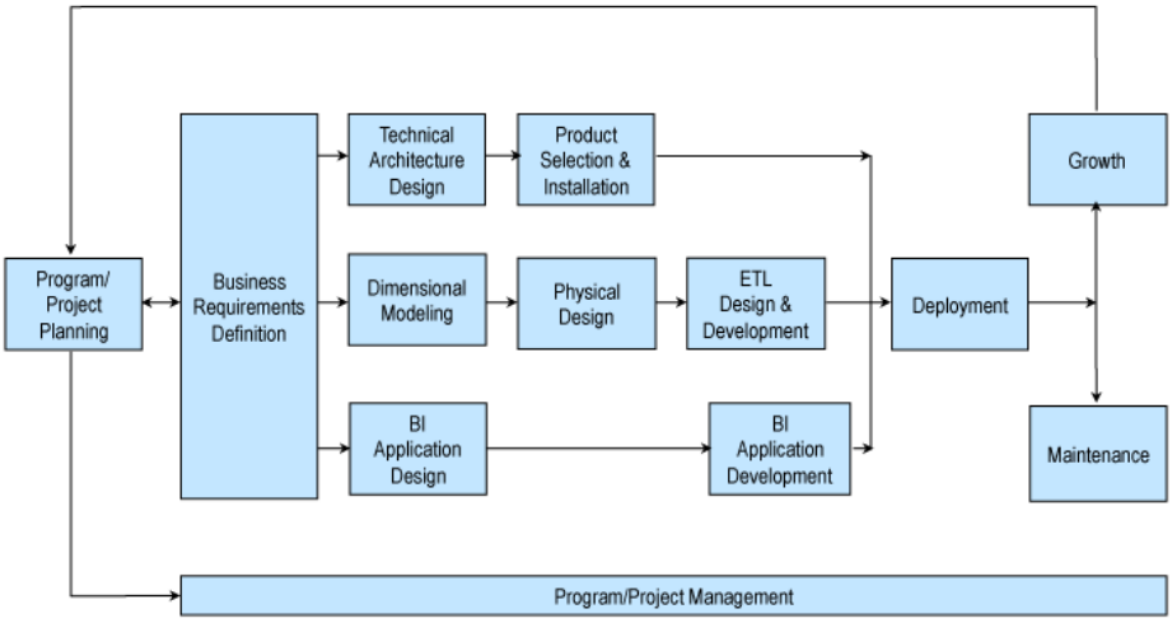

Figure 2.4 DW/BI project life cycle [78].

The DW design is constrained by "the business requirements, source data realities, budget, processing windows, and skill sets of the available staff" [78].

Being the core of a DW project, the ETL system [36]:

- Adjusts/conforms data from multiple sources to be used together.
- Corrects mistakes and missing data, adding value to it.
- Documents the lineage of data, providing traceability.
- Provides documented measures of confidence in data.
- Captures the flow of transactional data for safekeeping.
- Structures data to be usable by BI tools, as efficiently as possible.
- Enables data analysis and business decision making.

Architecture-wise, in Kimball's methodology, the DW has the back room and the front room (see figure 2.5) [36]. They are physically, logically, and administratively separate [36]. The back room is where all the data preparation happens and the front room presents that data to users [36].

Data preparation or data management or the back room comprises [36]:

- Extracting data from the original sources (legacy or transaction database systems).
- Quality assuring and cleaning data.
- Conforming the labels and measures in the data to achieve consistency across the original sources.
- Delivering data in a physical format that can be used by query tools, report writers, and dashboards (dimensional schemas).

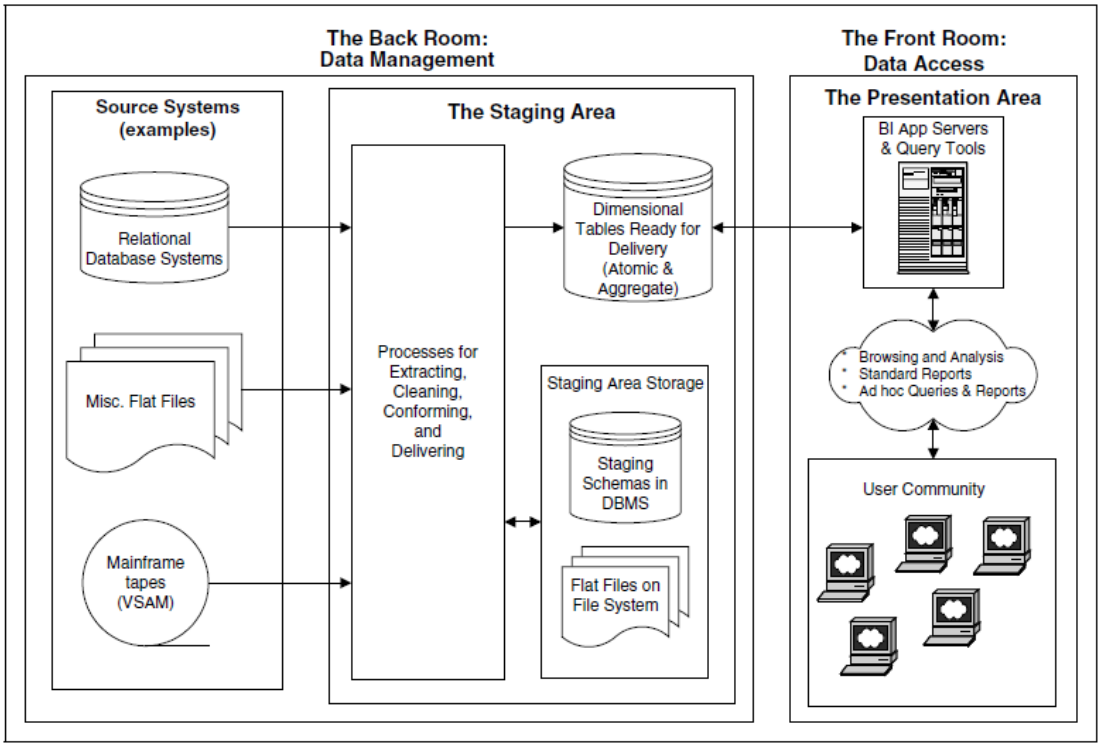

Figure 2.5 DW back and front rooms [36].

For Kimball, the build of an ETL system encompasses two simultaneous threads: the Planning & Design thread and the Data Flow thread [36].

In [36], ELT is generalized to the extract, clean, conform, and deliver steps in the data flow thread, as depicted in figure 2.6.

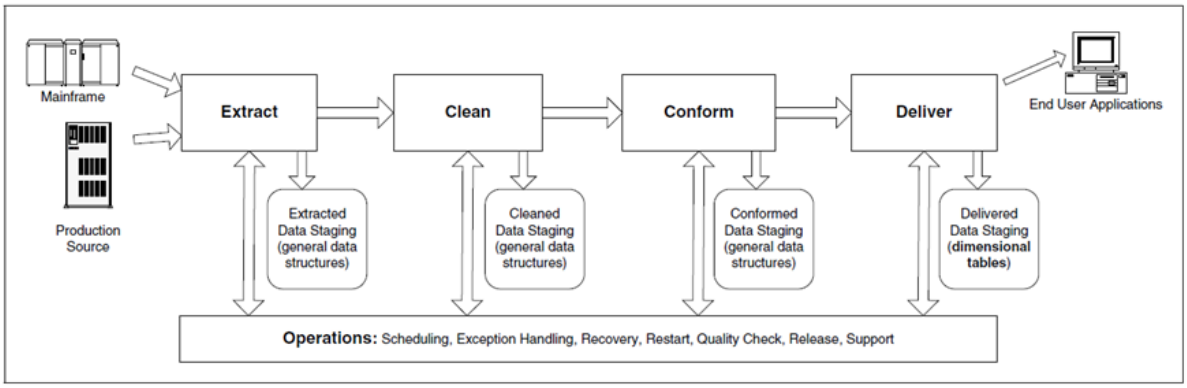

Figure 2.6 Data flow thread [36].

More detail about the data flow thread can be found in appendix A.1.

Simultaneous with the data flow thread, there is the planning & design thread [36], as represented in figure 2.7.

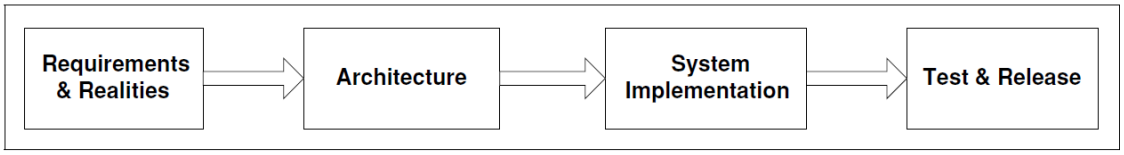

Figure 2.7 Planning & design thread [36].

More detail about the planning & design thread can be found in appendix A.2.

Kimball accounts ETL for 70% of the effort and risk needed to implement and maintain a DW [36], [29] and Inmon 80% [79], which is similar to the data preparation role in DM (60% to 90% of the time of a DM project [22]).

For ETL, data quality and data profiling are paramount:

- "The biggest risk to the timely completion of the ETL system comes from encountering unexpected data-quality problems. This risk can be mitigated with the data profiling techniques. (…) Quality requires a total commitment across every part of an organization. (…) However, the reality is that many organizations have not yet established formal data-quality environments" [36].
- "[Data profiling] employs analytic methods for looking at data for the purpose of developing a thorough understanding of the content, structure, and quality of the data. A good data profiling [system] can process very large amounts of data, and with the skills of the analyst, uncover all sorts of issues that need to be addressed", [36] stating Jack Olson's book *Data Quality: The Accuracy Dimension*. Data profiling guides the cleaning and conforming steps [36].

The ETL process can be adapted to individual needs: "sometimes it's ETL, sometimes it's ELT, or ELTL, or TETL" [29]. It is not recommended to have an unstructured approach to developing an ETL system, so the Kimball Group, based on thousands of successful DWs, has identified its 34 subsystems organized in 4 categories [78], presented in the figures 2.8 to 2.11 by the (*) numbers. These categories provide an organized description of ETL. This study focuses on the first and second category.

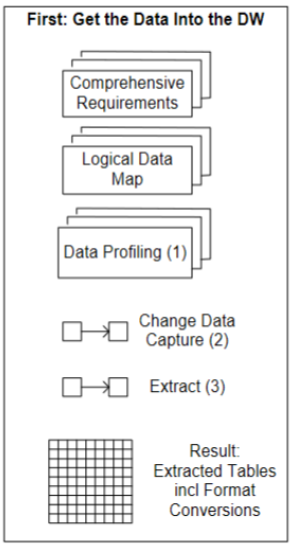

Figure 2.8 Get the data into the DW [78].

- Data Profiling (subsystem 1) – Explores a data source to determine its fit for inclusion as a source and the associated cleaning and conforming requirements [78].
- Change Data Capture (subsystem 2) – Isolates the changes that occurred in the source system to reduce the ETL processing burden [78].
- Extract (subsystem 3) – Extracts and moves source data into the DW environment for further processing [78].

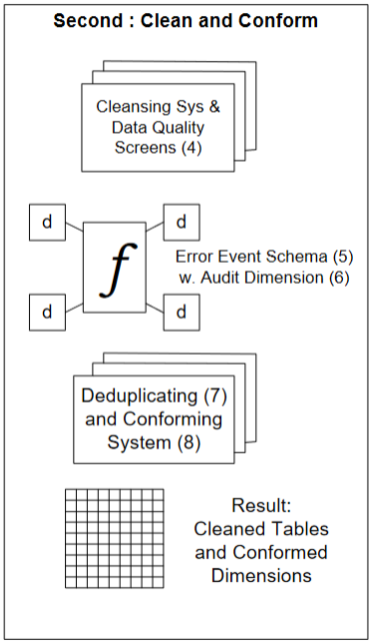

Figure 2.9 Clean and conform [78].

- Data Cleansing System and Data Quality Screens (subsystem 4) – Implement data quality processes to catch quality violations.
- Error Event Tracking (subsystem 5) – Captures and documents all error events that are vital inputs to data quality improvement [78].
- Audit Dimension Creation (subsystem 6) – Attaches metadata to each fact table as a dimension. This metadata is available to BI applications for visibility into data quality [78].
- Deduplication (subsystem 7) – Eliminates redundant members of core dimensions, such as customers or products. May require integration across multiple sources and application of survivorship rules to identify the most appropriate version of a duplicate row [78].
- Data Conformance (subsystem 8) – Enforces common dimension attributes across conformed master dimensions and common metrics across related fact tables [78].

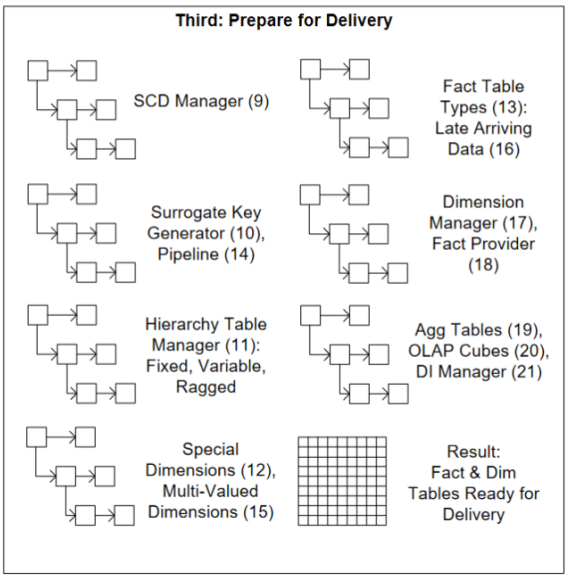

Figure 2.10 Prepare for delivery [78].

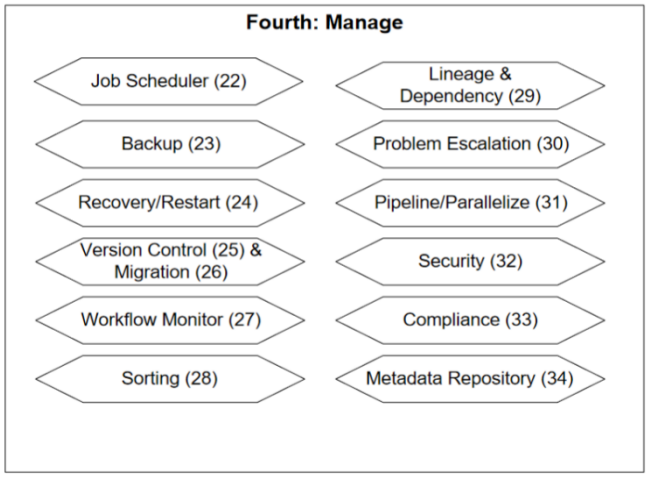

Figure 2.11 Manage [78].

From the third and fourth categories, only the Aggregate Builder (subsystem 19) is relevant for this study. The Aggregate Builder builds and maintains aggregates to be used seamlessly with aggregate navigation technologies for enhanced query performance [78].

2.6 Data Mining and Knowledge Discovery Process Models or Methodologies

Process models define what tasks to do; methodologies are process model instances that include all the information that process models have but also define how to carry out the tasks [25], [62].

The life cycle of a DM project "determines the order in which each activity is to be done. A life cycle model is the description of different ways of developing a project" [25]. A life cycle stipulates the criteria to validate a phase and the conditions to move to the next one, e.g.: provide intermediate deliverables [25].

DS can have a "top-down" approach (most common) where the business problem is first defined and then the data is analysed to find a solution or a "bottom-up" approach, also called exploratory [61], where the data is first analysed in order to extract business goals and then solutions are found [58].

Appendix C presents solutions that can fit both the “top-down” or “bottom-up” approaches if they are adapted to the project they are meant to be used in. For instance, some new steps may be created, others can be omitted, and the order of the steps can also change depending on the project, that is, the life cycle can be adapted. Exclusively bottom up or exploratory approaches can be found on appendix D.

There are many DM and KD process models and methodologies, but most of them are based in KDD or CRISP-DM (Cross Industry Standard Process for DM) [25].

The authors of [25] and the authors of [48] compiled the most used DM and KD process models and methodologies, covering up to 2019 between the two. This study couldn’t find new or relevant ones that weren’t covered by those authors.

More information about most used DM and KD process models or methodologies can be found in appendix C. Appendix C.1 presents the ones related to KDD. Appendix C.2 presents the ones related to CRISP-DM. Appendix C.3 presents other approaches.

A process model composes a set of inputs, outputs and well-defined activities and tasks related to a subject [25], [62]. “The goal of a process model is to make the process repeatable, manageable and measurable (to be able to get metrics)” [57]. A process is “a series of things that are done in order to achieve a particular result” [86], a process has a goal, it involves the path to achieve that goal. Defining a process can be so simple as to write it down [87].

[87], inspired in McCall’s software quality factors, enumerates the characteristics of a good process:

- Effectiveness: the process helps to achieve its intended result.
- Maintainability: the process should be well documented so that anyone who analyses it can understand it, make changes or improve it if necessary.
- Predictability: because the process is planned, it helps to predict how much time and people are necessary to execute it; also, if the process is executed consistently across different contexts, we can compare the different executions and learn from it.
- Repeatability: the process should be general enough to be replicable in different contexts or at least adapted; the exception to this are ad-hoc processes.
- Quality: the process should help to achieve the suited result or product with quality, fulfilling the stakeholder’s needs and conforming to its requirements.
- Improvement: the process is subject to continuous updates and improvements because of everchanging environments.
- Tracking: the process should be tracked and managed so that the intended results can be achieved with the desired quality; tracking it, powers the chance to improve it.

[87] stresses the importance of the documentation about the process (meta-process): besides helping in process conformance, it should lead to the quality of the result or product.

The process must be easy to follow through so that people adhere to it [87]. It should aim to not be seen as a bureaucratic task with little importance but to be a natural part of the day to day work [87]. Also "there must be some feedback from the user of the process as to how to improve the process, leading to the virtuous circle that is the Holy Grail of process engineers: Continuous Process Improvement or CPI" [87].

2.7 CRISP-DM

"CRISP-DM is the most widely used methodology for developing data mining projects" [25]. It is considered the *de facto* standard [25], [48]. "CRISP-DM comes up to resolve the problems that existed in data mining project developments" [25].

Rigorously, CRISP-DM is a mixing between a methodology and a process model, because it doesn't describe how all tasks must be done, so it's not a pure methodology [25]. Still, CRISP-DM provides much detail describing tasks that are valuable to inexperienced practitioners, especially in the user guide section [50].

"The CRISP-DM methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance" [50].

Each phase comprises several generic tasks, each generic task comprises several specialized tasks and each specialized task comprises several process instances, as showed in figure 2.12. Phases (see figure 2.13) and generic tasks compose the CRISP-DM process model which maps to a specific CRISP-DM process. That is, the CRISP-DM (more abstract) process model should be instantiated and adapted to the DM project specific context [50].

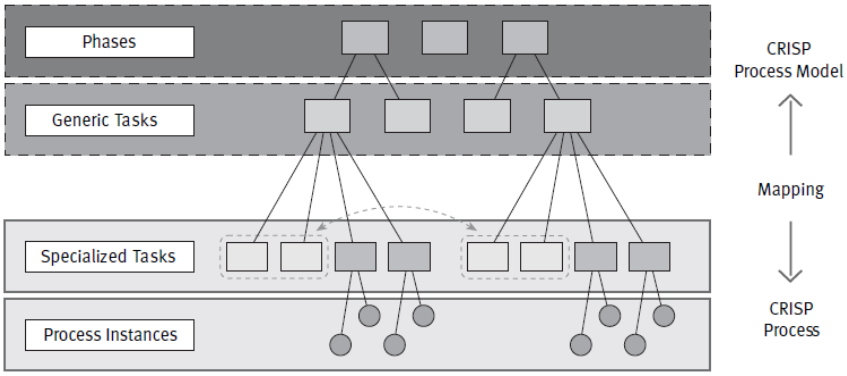

Figure 2.12 Four level breakdown of the CRISP-DM methodology [50].

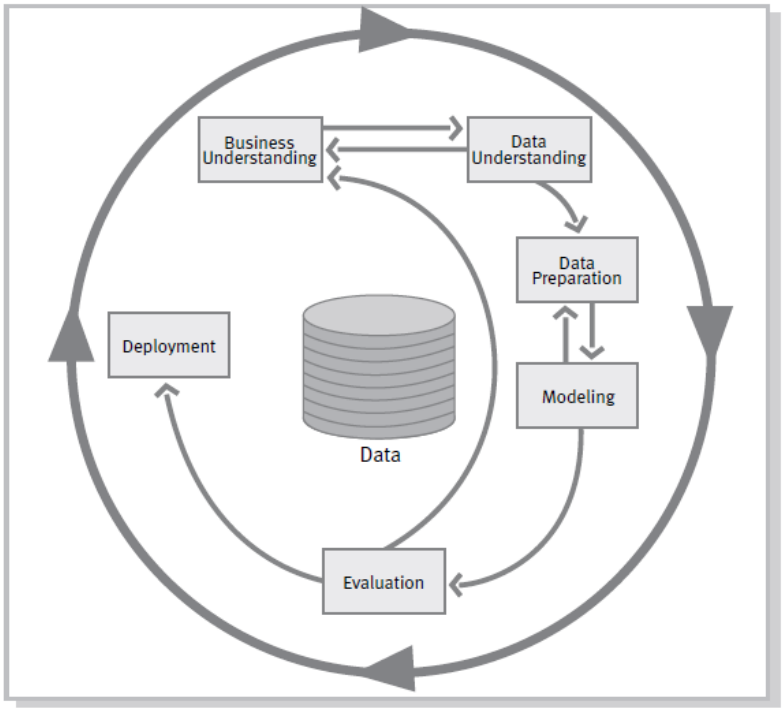

Figure 2.13 Phases of the CRISP-DM process model [50].

The CRISP-DM process model "provides an overview of the life cycle of a data mining project" and mirrors its iterative nature [50]. The tasks, depending of the DM project, have different relationships and orders of execution; this is also true regarding the phases [50]. "The arrows indicate the most important and frequent dependencies between phases" [50].

More information about each phase of CRISP-DM can be found in appendix B.1.

Summary tables for each phase with its associated generic tasks, notes about each task and the expected task outputs can be found in appendix B.2, adapted from [50].

[25] lists the main objectives and characteristics of CRISP-DM:

- Ensure quality of DM projects results.
- Reduce skills required for DM.
- Capture experience for reuse.
- General purpose (i.e., widely stable across varying applications).
- Robust (i.e., insensitive to changes in the environment).
- Tool and technique independent.
- Tool supportable.
- Industry-neutral.

Stating [25]: "Many changes have occurred in the business application of data mining since CRISP-DM 1.0 was published (…). Emerging issues and requirements include:

- The availability of new types of data (e.g., text, web and attitudinal data) along with new techniques for pre-processing, analysing and combining them with related case data.
- Integration and deployment of results with operational systems such as call centres and web sites.
- Far more demanding requirements for scalability and for deployment into real-time environments.
- The need to package analytical tasks for non-analytical end users and integrate these tasks in business workflows.
- The need to seamlessly integrate the deployment of results and closed-loop feedback with existing business processes.
- The need to mine large-scale databases in situ, rather than exporting an analytical data set.
- Organizations increasing reliance on teams, making it important to educate greater numbers of people on the processes and best practices associated with data mining and predictive analytics".

Being an ever-evolving area, new additions or updates to existing ones or new DM methodologies or process models are required [25]. Other methodologies like SEMMA or custom made ones are being adopted by professionals of the field [25].

In 2006, a CRISP-DM 2.0 version was being created [88], but, to this date, it is not published and it is unknown if or when it will be.

Because "[DM] projects not only involve examining huge volumes of data, but also managing and organizing big interdisciplinary human teams", [25] points out missing processes from CRISP-DM: project management processes, integral processes like quality management or change management that ensure project completeness and quality, and organizational processes, that help to achieve a more effective organization. This study doesn't agree with all of these claims because some processes are not directly related to a DM or KDD problem, but are complementary processes that are normally present in organizations.

In its conclusions, [25] suggests the study of other mature engineering fields, like software engineering (see [57]), to adapt its standards to DM to stimulate its growth and development, similarly to what is done in this study: adapting ETL to a DM problem.

2.8 Chapter Closure

This chapter documents the related work and the most important concepts to develop the project and inform the reader.

The next chapter (three) compares, analyses and extends current DM process models and methodologies with the data warehousing best practices standardized by Kimball.

CHAPTER 3

3 Analysis

From here onward, this study refers to data mining (DM) and knowledge discovery (KD) process models or methodologies simply as DM process models. The reasons for this are:

1. Simplification of language. Some authors even use the terms paradigm, framework or method to name their approach.
2. It is implied that DM leads to KD.
3. It is debatable if any of the presented methodologies are really a methodology. A methodology explains how to do the tasks. Since tasks in different DM problems are not always the same, the tasks presented in the methodologies may not apply.

3.1 Comparison and Unification of Data Mining and Knowledge Discovery Process Models or Methodologies

The first effort towards a hybrid process model that combines ETL and other data warehousing best practices with the most used DM process models is to unify the existing DM process models into one. To this end, this paper presents a comparative table - table 3-1 - where the tasks for different DM process models are therein aligned, according to its descriptions in the source papers. Hereupon, this study wants to understand if CRISP-DM encompasses all the information of other approaches, since it is the most used one in DM projects [17], [20], [21], [25], [48].

Although with some inspiration from [25]'s work, it is noteworthy that, by reading the DM process models' papers, this study didn't agree with many alignments in the tables proposed by [25].

On other note, the bottom-up or exploratory approaches are not included since they are not comparable with the others, because they don't present the tasks as a series of steps. Anyhow, the other approaches can be adapted to explorative DM projects. This study considers the exploration tasks as part of the steps that require them. For example, DST's Goal Exploration, according to CRISP-DM (see appendix B.2), can be done in the Business Understanding phase and DST's Data Value Exploration can be done in the Data Understanding phase.

Also, the Refined Data Mining Process' and Data Mining Engineering's Life Cycle Selection task (see appendices C.2 and C.3) is omitted because often it is not clear before the beginning of a DM project which tasks or in what order they will be addressed, that is, the order of tasks in a DM project, or even the undertaken tasks, can differ, depending on the results of the previous tasks. As many authors assert ([49], [74], [50], etc.), DM projects have an iterative nature; the process models are meant to be adapted to the individual needs of a given DM project, some tasks can be omitted, redone and new ones can be created.

Taking table 3-1 and the original papers into account, some approaches (refer to appendix C) leave the DM project open to perform Operating and Optimization Processes, Refinement, Feedback and Establish On-Going Support tasks. It is questionable if CRISP-DM includes these tasks. Still, in table 3-1, this study choses to align them with CRISP-DM's Deployment generic task because:

1. CRISP-DM includes a Plan Monitoring and Maintenance task in the Deployment generic task that is described as: "Determine when the data mining result or model should not be used any more. Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.), and what should happen if the model or result could no longer be used (update model, set up new data mining project, etc.)" [50]. Thus, CRISP-DM takes these tasks into account and prepares for their execution.
2. The undertake of these tasks can mean that the DM process must go back to the start; that is another iteration of CRISP-DM or other process model.
3. CRISP-DM generalizes better to the case that the DM project is outsourced; in this case, these tasks are unlikely performed by the same people, thus, the DM project would not include them.
4. CRISP-DM also includes a Determine Next Steps task in the Evaluation generic task to evaluate if the project is ready for the deployment or if another iteration is needed.

Therefore, it is understandable why CRISP-DM, even after 21 years, continues to be the *de facto* standard:

- Comprehensiveness: it includes all the other DM process models' tasks.
- Detailedness: it provides a lot of detail regarding tasks and outputs that belong to each generic task.
- Flexibility and versatility: it can be easily adapted to different DM projects.

For these reasons, in this study, CRISP-DM alone is compared and extended with ETL and other data warehousing best practices.

Table 3-1 DM process models and methodologies comparison table.

Name	Task																							
CRISP-DM (& RAMSYS & CASP-DM)	Business Understanding				Data Understanding			Data Preparation			Modeling			Evaluation			Deployment							
ASUM-DM	Conduct Readiness Assessment		Understand Business			Data Understanding			Prepare Data			Build Model			Evaluate Model			Conduct Analytical Knowledge Transfer		Deployment Processes		Operating and Optimization Processes		
FMDS	Business Understanding			Analytic Approach		Data Requirements	Data Collection	Data Understanding		Data Preparation			Modeling			Evaluation			Deployment				Feedback	
TDSP	Business Understanding						Data Acquisition and Understanding			Modeling									Deployment				Customer Acceptance	
KDD (only image)							Selection				Pre-Processing	Transformation	Data Mining			Interpretation/ Evaluation								
KDD (with description)	Understanding of the Application Domain						Creating a Target Data Set		Data Cleaning and Pre-Processing			Data Reduction and Projection	Choosing the DM Function	Choosing the DM Algorithm	Data Mining		Interpretation			Using Discovered Knowledge				
Refined Data Mining Process	Domain Knowledge Elicitation		Human Resource Identification		Problem Specification		Data Prospecting		Data Cleaning	Pre-Processing		Data Reduction and Projection	Choosing the DM Function	Choosing the DM Algorithm	Build Model	Improve Model	Evaluation	Interpretation	Deployment		Automate		Establish On-Going Support	
Data Mining Engineering	Project Management, Pre-Development and Requirements Processes						KDD Processes												Post-Development and Integral Processes					
6-sigma	Define						Measure						Analyse			Improve	Control							
5 A's	Asses						Access						Analyse						Act		Automate			
Human Centred	Task Discovery				Data Discovery				Data Cleaning				Model Development			Data Analysis	Output Generation							
SEMMA							Sample	Explore	Modify				Model			Assess								
Two Crows	Define Business Problem						Build DM Data Base	Explore Data	Prepare Data for Modeling				Build Model			Evaluate Model	Deploy Model and Results							
Anand & Buchner	Domain Knowledge Elicitation		Human Resource Identification		Methodology Identification	Problem Specification	Data Prospecting		Data Pre- Processing				Pattern Discovery			Knowledge Post- Processing								Refinement
Cabena *et al.*	Business Objectives Determination						Select				Pre-Process	Transformation	Mine			Analyse and Assimilate								
Cios *et al.*	Understanding the Problem Domain						Understanding the Data		Preparation of the Data				Data Mining			Evaluation of the Discovered Knowledge			Using the Discovered Knowledge					
KDD Roadmap	Resourcing			Problem Specification					Data Cleansing	Pre-Processing			Data Mining			Evaluation	Interpretation		Exploitation					
DMIE	Analyse the Organization			Structure the Work			Develop Data Model												Implement Model				Establish On-Going Support	

3.2 Comparison and Extension of CRISP-DM with ETL and other Data Warehousing Best Practices

Similarly to the DM process models and methodologies comparison table (table 3-1) this section presents comparative tables – tables 3-2 to 3-4 - where the CRISP-DM tasks and the ETL subsystems are therein aligned, according to its sources' descriptions. After, the CRISP-DM's methodology is extended with the data warehousing best practices and heuristics. These extensions are organized in subsections titled with the CRISP-DM's tasks (with the phases in parentheses) and the corresponding ETL subsystems (with the ETL systems in parentheses); except for the General ETL and other Data Warehousing Best Practices that can be applied to the whole DM process.

Some ETL subsystems are omitted because they are related to the DW dimensional nature or are out of this paper's scope. Some CRISP-DM tasks are omitted as well because they don't have a correspondent subsystem in ETL.

There are some differences regarding CRISP-DM tasks' order when comparing it to the ELT subsystems' order; perhaps because:

- CRISP-DM is presented as a process model, meant to generalize to different contexts, having "development" steps and deployment steps. Depending of the DM project, the deployment in itself is or is not undertaken in the context of the DM process [50].
- ELT is presented more as an architecture, that provides heuristics on how to implement an ETL system – it includes deployment in itself.

These differences are not a problem, because CRISP-DM tasks' order can be adapted depending on the DM project and the tasks' nature are similar in both worlds.

A necessary remark to make is that ETL's subsystem "data cleansing system and data quality screens" appears twice in the comparison, because ETL doesn't separate the data quality verifications of their implementation (data cleansing).

The conclusions of this comparison and extension are:

- In the specific parts of CRISP-DM where ETL is applicable, ETL's heuristics are more or less the same as CRISP-DM's; albeit ETL's data quality provides some additional interesting information.
- The biggest contributions to CRISP-DM are the general ETL and other data warehousing best practices that can be applied to a DM problem; its heuristics are not present in CRISP-DM.

ETL's relevant subsystems in its original order are:

- Data profiling

- Change data capture
- Extract
- Data cleansing system and data quality screens
- Error event tracking
- Audit dimension creation
- Deduplication
- Data conformance
- Aggregate builder

The following tables (tables 3-2 to 3-4) present the comparison between CRISP-DM tasks and ETL subsystems.

Table 3-2 CRISP-DM and ETL comparison table for CRISP-DM's Data Understanding.

<table>
<tr><td rowspan="2">CRISP-DM</td><td colspan="6">Data Understanding</td></tr>
<tr><td>Collect initial data</td><td>Describe data</td><td>Explore data</td><td colspan="3">Verify data quality</td></tr>
<tr><td rowspan="2">ETL</td><td>Extract</td><td colspan="3">Data profiling</td><td>Data cleansing system and data quality screens</td><td>Error event tracking</td><td>Audit dimension creation</td></tr>
<tr><td colspan="3">Get the Data Into the DW</td><td colspan="3">Clean and Conform</td></tr>
</table>

Table 3-3 CRISP-DM and ETL comparison table for CRISP-DM's Data Preparation.

<table>
<tr><td rowspan="2">CRISP-DM</td><td colspan="6">Data Preparation</td></tr>
<tr><td>Select data</td><td>Clean data</td><td colspan="3">Integrate data</td><td>Format data</td></tr>
<tr><td rowspan="2">ETL</td><td rowspan="2"></td><td>Data cleansing system and data quality screens</td><td>Deduplication</td><td>Data conformance</td><td>Aggregate builder</td><td rowspan="2"></td></tr>
<tr><td colspan="3">Clean and Conform</td><td>Prepare for Delivery</td></tr>
</table>

Table 3-4 CRISP-DM and ETL comparison table for CRISP-DM's Deployment.

<table>
<tr><td rowspan="2">CRISP-DM</td><td colspan="4">Deployment</td></tr>
<tr><td>Plan deployment</td><td>Plan monitoring and maintenance</td><td>Produce final report</td><td>Review project</td></tr>
<tr><td rowspan="2">ETL</td><td rowspan="2"></td><td>Change data capture</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>Get the Data Into the DW</td></tr>
</table>

3.2.1 General ETL and other Data Warehousing Best Practices

By analysing Kimball's work [36], this study gathered some best practices to build a successful ETL that mostly can be applied to a DM problem:

- Make ETL metadata driven. The ETL tool should read the transformations from metadata; they shouldn't be hard coded.
- Stage data for recoverability, backup or audit purposes, for instance staging after each step of ETL or whenever a big transformation takes place.
- Perform impact analysis. It involves evaluating what are the impacts if a table is changed by analysing its metadata. Once a table is created in the staging area, one must perform impact analysis before any changes are made to it.
- Create metadata, that is, document everything (refer to figure 3.1). Metadata is composed by table structure metadata (25%), data-cleaning results (25%), and process results (50%). Most of it should be done automatically by ETL tools. Some examples are: technical and business definitions, process metadata (like number of rows inserted, updated, deleted, and rejected; start time, end time, and duration for each process) and the logical data map (or data lineage).
- Use a logical data map. It is an ongoing document from the beginning to the end of ETL that serves as a blueprint for the transformation step (it appears in the first categorisation of the 34 subsystems). Is usually presented in a table or spreadsheet format and includes the following specific components:
    i. Target table name: The physical name of the table as it appears in the DW.
    ii. Target column name: The name of the column in the DW table.
    iii. Table type: Indicates if the table is a fact, dimension, or subdimension (outrigger) (the reader can find more about these subjects in [36]; this is not applicable to a DM problem).

iv. SCD (slowly changing dimension) type: Type-1, -2, or -3 slowly changing dimension (more about this in [36]; this is not applicable to a DM problem).

v. Source database: The name of the instance of the database where the source data resides. This component is usually the connect string required to connect to the database. It can also be the name of a file as it is an ongoing document it appears in the file system. In this case, the path of the file would also be included.

vi. Source table name: The name of the table where the source data originates. There will be many cases where more than one table is required. In those cases, simply list all tables required to populate the relative table in the target DW.

vii. Source column name: The column or columns necessary to populate the target. Simply list all of the columns required to load the target column. The associations of the source columns are documented in the transformation section.

viii. Transformation: The exact manipulation required of the source data so it corresponds to the expected format of the target. This component is usually notated in SQL or pseudo-code or even in natural language.

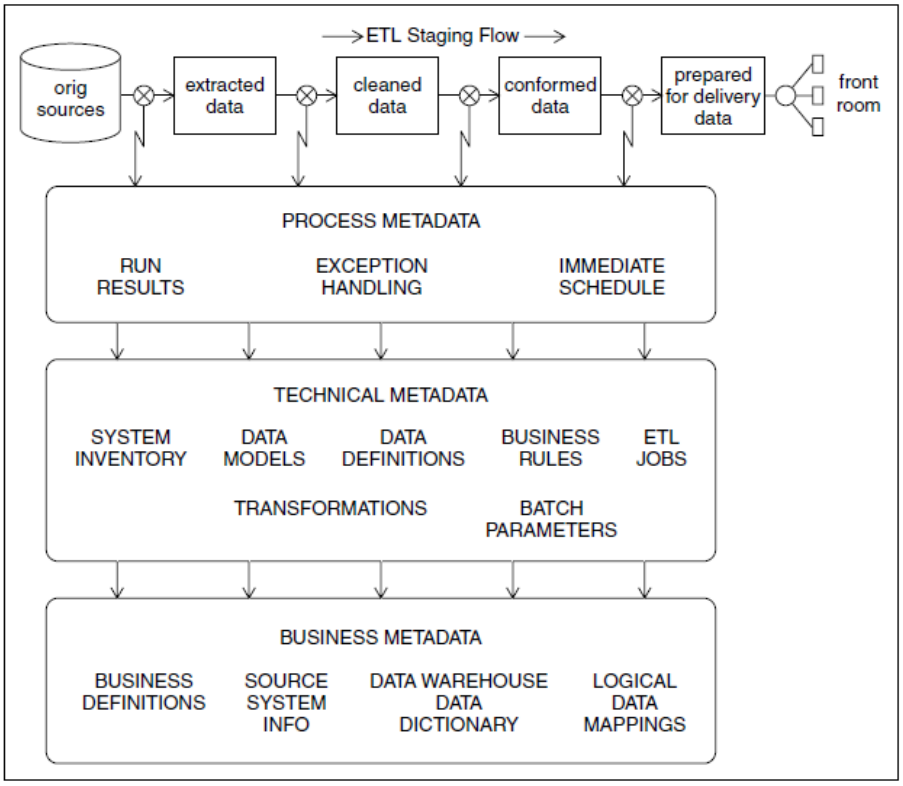

Figure 3.1 Metadata sources in the back room of the DW [36].

3.2.2 Collect Initial Data (Data Understanding) – Extract (Get the Data Into the DW)

Extract includes [36]:

- Reading source-data models: data should be sourced from the system-of-record, that is, the original source. There can be altered versions of the same source table that may not be suitable for ETL.
- Connecting to and accessing data.
- Sometimes performing low level cleaning like change encodings.
- Scheduling the source system, intercepting notifications and daemons.
- Capturing changed data: this should be planned in advance.
- Staging (writing) the extracted data to disk.

In CRISP-DM, the capturing changed data step of the ETL's extraction aligns with the Plan Monitoring and Maintenance task, more about it can be found in section 3.2.7.

[36] provides some hints to speed the extraction process (most only applies if using SQL):

1. Index the columns used in where clauses
2. Retrieve only the data you need
3. Use distinct and set operations like union, minus and intersect sparingly, because they are slow. Union all is faster, but it returns duplicates.
4. Use hint
5. Avoid not and <>
6. Avoid functions in your where clause

3.2.3 Describe and Explore Data (Data Understanding) - Data Profiling (Get the Data Into the DW)

In ETL, the data profiling metadata should include (more detail about this in Appendix B of Jack Olson's book, *Data Quality: The Accuracy Dimension*) [36]:

- Schema definitions.
- Business objects.
- Domains.
- Data sources.
- Table definitions.
- Synonyms.
- Data rules.

- Value rules.
- Issues that need to be addressed.

A data-profiling checklist should include [36]:

- Providing a history of record counts by day for tables to be extracted. (This is not applicable for most DM projects).
- Providing a history of totals of key business metrics by day. (This is not applicable for most DM projects).
- Identifying required columns.
- Identifying column sets that should be unique.
- Identifying columns permitted (and not permitted) to be null.
- Determining acceptable ranges of numeric fields.
- Determining acceptable ranges of lengths for character columns.
- Determining the set of explicitly valid values for all columns where this can be defined.
- Identifying frequently appearing invalid values in columns that do not have explicit valid value sets.
- Identifying column distribution reasonability.

Some of the items in the former list, actually belong to the Verify Data Quality category (section 3.2.4), but they were included in this category to keep ETL's organization for the reader's reference.

3.2.4 Verify Data Quality (Data Understanding) - Data Profiling, Data Cleansing System and Data Quality Screens, Error Event Tracking and Audit Dimension Creation (Get the Data Into the DW and Clean and Conform)

Data quality requires a strong commitment across every part of an organization [36].

The data quality process should be executed after data is extracted (before any kind of processing) and after processing it [36]. It includes analysing the quality of all the metadata or documentation [36]. A good quality verification system can process very large amounts of data [36]. Normally, screens / error checks run in parallel (see figure 3.4) [36].

In [29], metadata is described as "all the information that defines and describes the structures, operations, and contents of the" system; It can be technical (e.g. transformations), business (e.g. descriptions) or process metadata (e.g. logs).

For the data to be accurate, it has to be: correct, unambiguous, consistent and complete (record by record or in aggregated records) [36].

Having a fast ETL is conflicting with having a thorough data quality screening, as depicted in figure 3.2 [36].

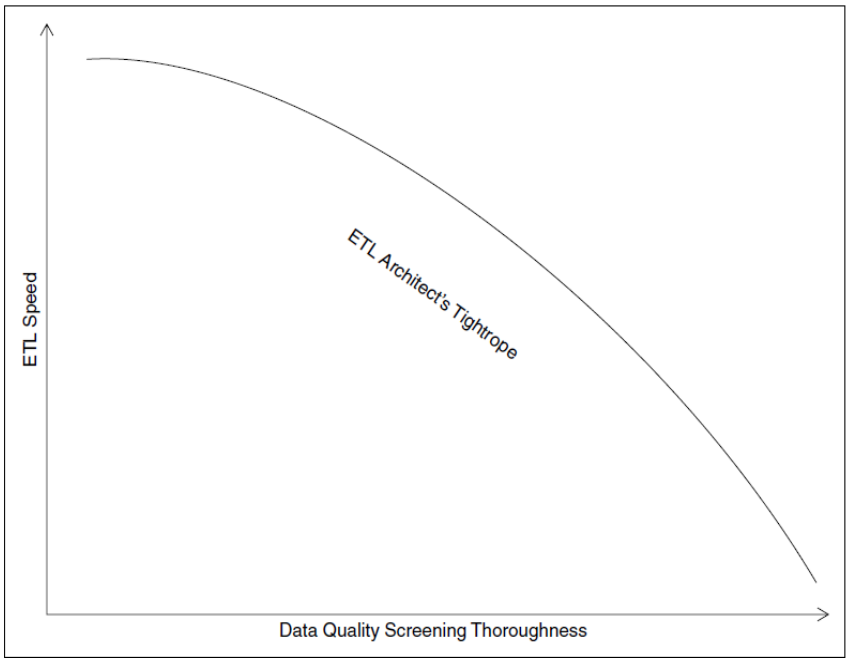

Figure 3.2 Speed vs completeness [36].

According to [36], there are four competing pressures or priorities in the data quality process, as depicted in figure 3.3:

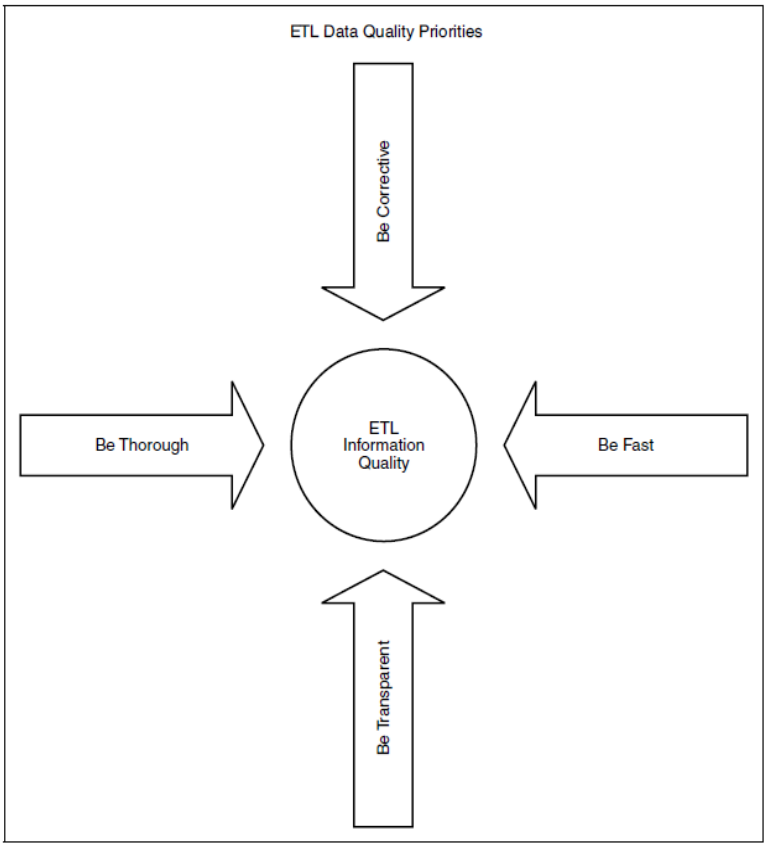

Figure 3.3 Data quality priorities [36].

It is not possible to address all of them at the same time, one must prioritize and balance them according to the business needs [36]. Too much transparency reduces the analytical strength and a too much correctness disguises operational deficiencies and slows organizational progress [36]. Derived or summarized attributes for data quality, opposed to raw data (transparency), can provide more analytical strength and errors in data help to detect points of improvement that lead to organization progress.

"Cleaning and conforming are the main steps where the ETL system adds value. (…) [These steps] generate potent metadata. Looking backward toward the original sources, this metadata is a diagnosis of what's wrong in the source systems. Data problems should be corrected at the source or as close possible to the source to mitigate potential costs and other problems. Ultimately, dirty data can be fixed only by changing the way these source systems collect data", which, in turn, can help re-engineering a business process [36].

The deliverables of data quality are the error event table and the audit dimension, structured in a dimensional model [36].

The error event table is essential for the data quality process; it captures all the records with errors from any table, discovered during the data quality check [36]. This table has information about the date, the quality check performed, the severity score of the error, the records that have errors, the error type (ex: incorrect, ambiguous, inconsistent or incomplete), the source system, the ETL stage in which the check is applied and the action to perform [36].

Regarding the action to perform during quality checks, a record can be [36]:

1. Passed with no errors.
2. Passed, flagging offending column values (most common).
3. Rejected.
4. The cause to stop the ETL job stream altogether.

The audit dimension, instead of focusing on the errors at the record level as the error event table, it captures information about the data quality at the table level, only for the final tables that are loaded to the from room, and in a summarized form [36]. This dimension includes important ETL-processing milestone timestamps and outcomes, significant errors and their frequency or occurrence, an overall data-quality score and a description of the fixes and changes that have been applied [36]. Therefore, the audit dimension contains aggregations of the error event table information, like score sums [36].

Data quality checks categories:

1. Column property enforcement; check if the columns extracted contain the expected values from the source:
    i. Null values in required columns.
    ii. Numeric values that fall outside of expected high and low ranges.

iii. Columns whose lengths are unexpectedly short or long.

iv. Columns that contain values outside of discrete valid value sets.

v. Adherence to a required pattern or member of a set of patterns.

vi. Hits against a list of known wrong values where the list of acceptable values is too long.

vii. Spell-checker rejects.

2. Structure enforcement; check:

   i. Primary and foreign keys and referential integrity.

   ii. Explicit and implicit hierarchies and relationships among groups of fields.

3. Data and value rules enforcement; check:

   i. If the business rules are being broken in the data.

   ii. For invalid values, like a client who is a woman but has the gender male.

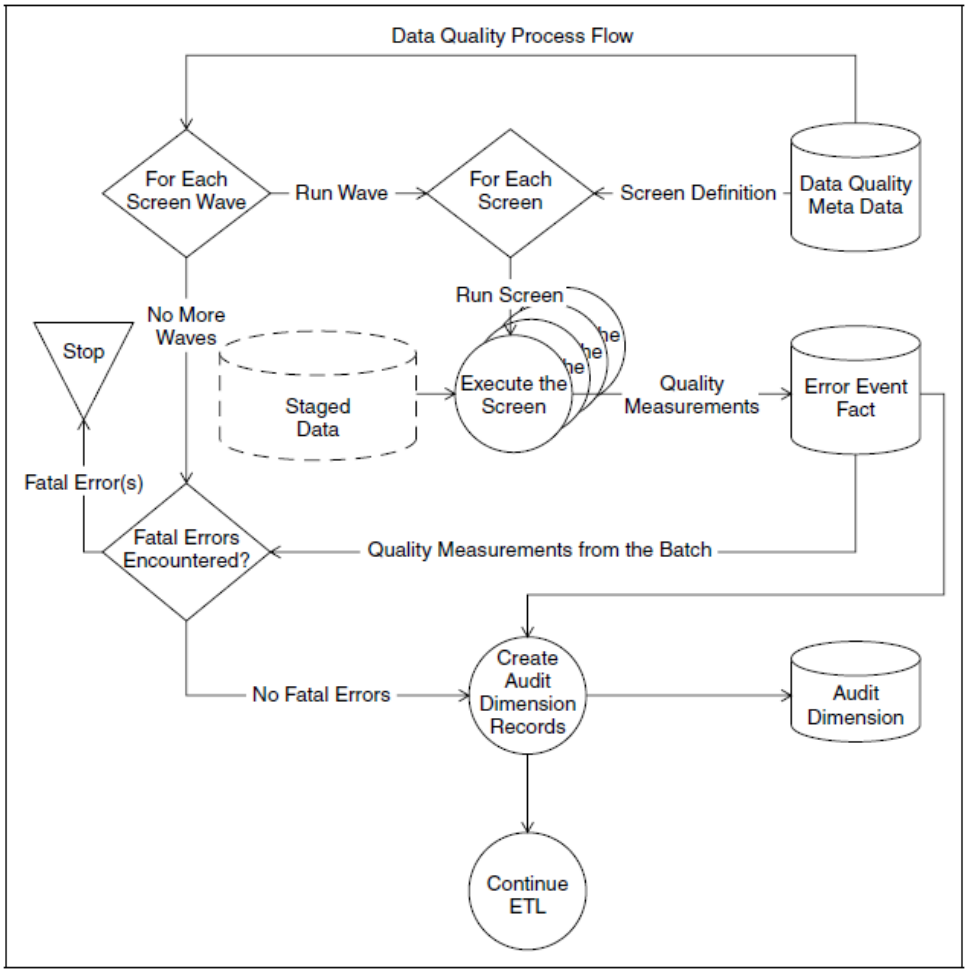

Figure 3.4 Data quality process flow [36].

3.2.5 Clean Data (Data Preparation) - Data Cleansing System and Data Quality Screens (Clean and Conform)

ETL's data cleanse is where the actions presented in the error event table and in the audit dimension take place.

Having missing values in data or nulls, means data is not complete or is ambiguous, because a missing can mean that the value is unknown or that an answer is not applicable on that record. [36] advises to replace nulls with a value that represents "missing value" in some circumstances, like when there is a need to perform joins on that specific column; this prevents data loss.

3.2.6 Integrate and Format Data (Data Preparation) - Deduplication, Data Conformance and Aggregate Builder (Clean and Conform and Prepare for Delivery)

The conform step can be summarized in:

- Standardizing.
- Structuring data as a series of dimensional schemas.
- Matching and deduplication (duplicates come from the integration of data from different sources).
- Surviving (combine duplicates that are not exact copies to create better quality records to be loaded to the from room).

In ETL, after the clean and conform steps, comes the deliver step. Its focus is on the dimension tables processing [78]. To create the dimension tables, this step includes all the transformations needed in the data, like data aggregation [78]. Finally, the transformed data is delivered to the from room [36]. In the case of a DM problem, at this point, data is ready for modeling and subsequent tasks. Because CRISP-DM's aggregations are done in the Integrate Data task, the Aggregate Builder subsystem is herein included and not in the Construct Data task, for example.

3.2.7 Plan Monitoring and Maintenance (Deployment) - Change Data Capture (Get the Data Into the DW)

As new transactional data is added to the ODBMS, there is a need to capture changed data and update the DW data: incremental load [36]. This can also happen in DM projects: new data can rend the

previous trained model obsolete if its characteristics are different than the ones of the data that was used to train the model. Hence, there can be a need to retrain the models with new data.

There are some incremental load options, some better than others. The best, according to [36], are (others in [36]):

1 Use audit columns on the source and target tables: they are usually updated automatically by database triggers with the date and time of when a record is inserted or updated. Then, the dates are compared to the last incremental load's date to select only the right records to be inserted to or updated in the DW. Can go wrong if the audit columns are updated normally with an application but some record is edited in the back-end for some reason without updating the audit columns. The last incremental load's date is the date time found in the source system audit columns at the time of each extract and it is saved for each source table.

2 Use a process of elimination. Keep a copy of the extracted data in the staging area. At the next extraction, the entire source tables are copied into the staging area and compared with the saved tables: only new records are copied to the DW. It's not very efficient, but it is the most reliable technique. It also detects deleted rows from the source. For this, initial and incremental loads with a previous_load and a current_load table can be used. In pseudo code it would be something like:

   - *current_load = new_data*
   - *data_to_load = current_load - previous_load*
   - *transform data_to_load*
   - *load_to_DW data_to_load*
   - *delete previous_load*
   - *rename current_load to previous_load*
   - *create_empty current_load*

   The minus operation can be slow if inside the back room/ ODBMS, but it is a good option if done with an external tool.

[79] points another option, not mentioned by [36]:

- Detect new inserts or updates in ODBMS, extract, transform and load them into the DW immediately. This can, however, create performance issues in ODBMS.

3.3 Chapter Closure

In this chapter, the most relevant DM process models and methodologies in the literature were compared to attempt to unify them into one. ETL and other data warehousing best practices were selected from Kimball's methodology and hereupon described. Also, since CRISP-DM includes all of the other processes tasks, CRISP-DM was compared to and extended with the selected ETL and other data warehousing best practices. These extensions were applied in the project and are documented, alongside with relevant project context, in the next chapter (four).

CHAPTER 4

4 Case Study

This chapter presents how ETL and other data warehousing best practices were applied in the project IA-SI, to foster knowledge reuse.

For scope management purposes, only the tasks of ETL in which the author was fully or partly responsible are included. These correspond mostly to the prediction of the ineligible expenses goal, excluding the text mining tasks that were also done in the project. CRISP-DM's business understanding phase is included to provide context but the modeling and evaluation phases are not included, since they are out of this dissertation scope because they don't a have a correspondent ETL subsystem.

Sections from 4.1 to 4.5 introduce project details, starting with figure 4.1 that shows the project's data mining process model. The rest of the chapter is organized according to the CRISP-DM tasks that have a correspondent ETL subsystem, except for the general best practices, similarly to section 3.2, and for the business understanding phase, as explained before.

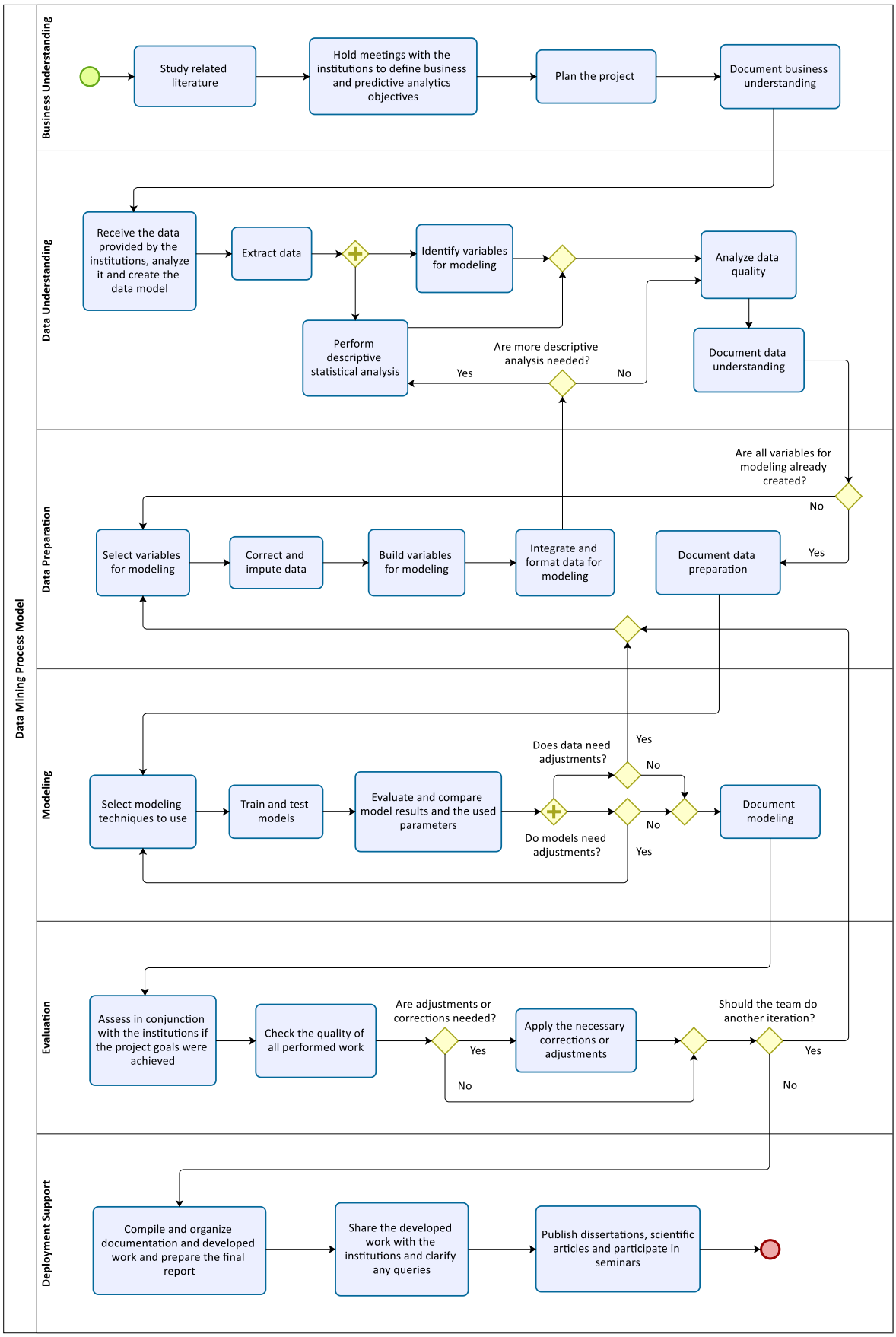

Figure 4.1 Case's data mining process model.

4.1 Actors

The actors that play a role in the project are:

- **Applicant**: a person or company that applies to the grant funds with a project application.
- **Supplier**: a person or company that sells something that the applicant's project needs.
- **Partner**: a person or company with which the applicant has some form of alliance.
- **Consultant**: a person or company that creates the applications or helps the applicant in doing so.
- **Team**: ISCTE people (technicians or experts) that work in the IA-SI project.
- **Team Technician**: ISCTE person that works mainly on the operational aspects of the IA-SI project.
- **Team Expert**: ISCTE person that works mainly on the administration and consulting aspects of the IA-SI project.
- **Institutions**: IAPMEI and AICEP; they manage the grant funds and applications for these and support the team in the execution of the IA-SI project.
- **Institution Technician**: IAPMEI or AICEP person that works mainly on the operational aspects of the management of grant funds and applications for these and supports the team for the execution of the IA-SI project.
- **Institution Manager**: IAPMEI or AICEP person that works mainly on the administration of the management of grant funds and applications for these and supports the team for the execution of the IA-SI project.

4.2 Data

Most of the data was provided by the institutions; some of it was created by the team. The data from both institutions is very similar. Whenever there is a need, this document makes the necessary distinctions. The data summary is presented in table 4-1.

Source Name	Source Description	Data Description	File Type	Sourced By
Candidaturas	Submited application forms	Applicant's personal and project information	XML	Institutions
FACI	Investment application analysis	Scores given to the applications	XML	Institutions
PPI	Interim payment request	Expenses characteristics	XML	Institutions
APPI	Interim payment request analysis	Expenses eligibility analysis	XML	Institutions
FACIE	Analysis of the closing of the investment application	Last expenses eligibility analysis	XML	Institutions
Anexo A	Part A of the simplified business information (IES)	Balance sheets and income statement	XML	Institutions
Anexo R	Part R of the IES	Economic activities code fields and their designation	XML	Institutions
BDCPME	Small and medium enterprises certification database	Dates of the constitution of companies and dates of their opening of activity	XLSX	Institutions
Anulações	Project cancellations	Projects that were cancelled and the reason	XLSX	Institutions
CAE	Grouped economic activities code fields	Grouped economic activities code fields and their descriptions	XLSX	Team
NUTS	Portuguese NUTS II regions	Portuguese councils and the correspondent NUTS II regions	XLSX	Team

Table 4-1 IA-SI project's data sources summary.

4.3 Data Mining Problem

As explained before, institutions receive and manage European grant funds. Figure 4.2 illustrates the grant funds cycle.

First, businesses apply with a project of investment to be able to receive those funds. The applications originate *candidaturas* data. The institutions analyze the applications and decide which should be supported/ which are eligible; this creates the FACI data. For the eligible projects, a contract is signed between the two parts. At this point, the applicant can receive money in advance or present expenses proof that justify the offered money and/ or justify granting more money. The presentation of expenses originates PPI data. The expenses are analyzed by the institutions, who investigate their eligibility and classify them as eligible or ineligible. This creates APPI data. As long as applicants present proof of expenses that are classified as eligible, and they don't exceed their budgeted grant fund, they continue to receive funds to undertake their projects. When the applicants' projects of investment are

done, they submit the last expenses proof and these are analyzed by the institutions, similarly to the PPI analysis. This originates FACIE data. At this point, sometimes, if there can be grant fund reimbursement if applicants received more money than they should. Finally, the contract between the two parts comes to an end.

An expense can be an item or a group of items. A group of expenses form a PPI. A group of expenses that were analyzed by the institutions form an APPI. The last group of expenses of a project that were analyzed by the institutions form a FACIE.

Given the high volume of presented expenses, some grant fund payments are delayed. To avoid this, and also to minimize grant fund reimbursements, the project IA-SI has the data mining (DM) goal/ target of predicting ineligible expenses.

Ineligible expenses are expenses that are not framed within the contract. It can be that the type of investment is out of the presented project's scope, or that the investment surpasses the budgeted grant fund, or mistakes in expense information or, the worst reason, false declarations.

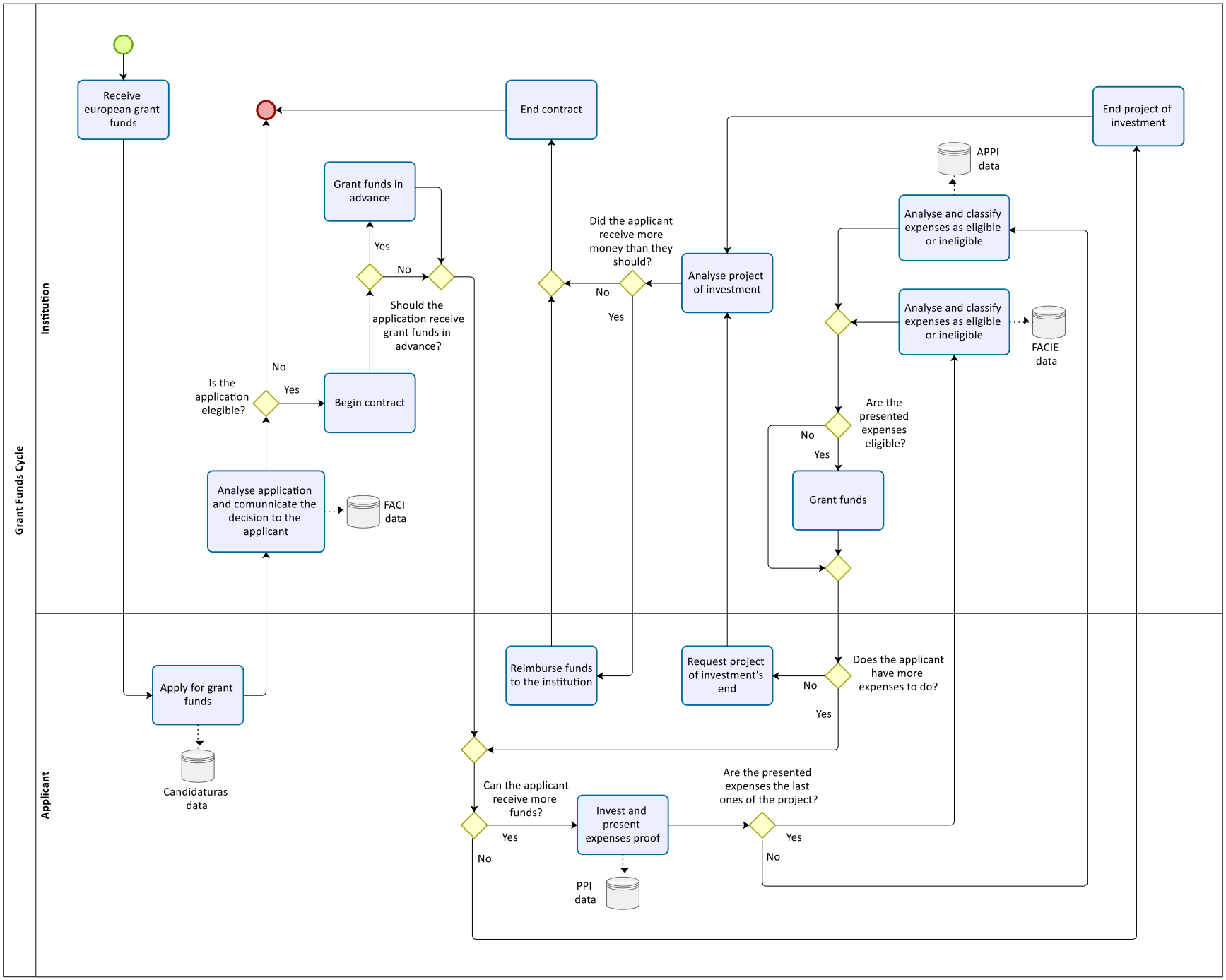

Figure 4.2 The grant funds cycle.

4.4 Data Mining Software Prototype

The team developed all the necessary components of a custom-made DM software prototype written in the Python programming language, as opposed to using a commercial tool. The reasons for this are:

- To have no licenses costs.
- To have freedom to personalize the solution completely.

The author of this dissertation was fully responsible for the development of the software prototype components of data understanding and preparation for the ineligibilities DM problem explored here, except for the programming of the extraction, part of the programming of the IES transformation and the text mining variables.

The author created various scripts (different files), each one with its purpose:

- **Main** script. The entry point to the program is the "main" script, that is the centre for all operations. This script makes calls to the other scripts to run their methods.
- **Data preparation** scripts. As data preparation has a lot of code, there is a script to prepare each data source, for example, a script for *candidaturas*, other for FACI, other for PPI, etc.
- **Data correction** script, so that all the corrections are in the same place.
- **Derivate transformations** script, that is transformations that use more than one data source.
- **Imputations** script.
- **Record and column selectio**n script; it deletes the lines and columns that are not going to be used in modelling.
- **Data dictionary** script.
- **Plots and statistics** script.
- **Parameters** script; it has all the parameters used in the rest of the program.
- **Utilities** script; it has utility methods that are called by other scripts.

The line of execution of the software prototype is explained next:

1. The program starts by creating a backup of previously created data (if it exists already) for debugging purposes.
2. Then, all necessary folders are created to save the data in all intermediate steps that are important for traceability and auditing purposes.
3. Afterwards, all transformations by type of data source are made.
4. After these transformations, all the data is joined/ merged into a single table.

5. Then corrections are made to the data. The reason why corrections were only made at this point is because the correction of some columns from one source requires transformed columns from another source.
6. Then, more variables are created that depend on various sources or that can only be created after the corrections have been applied.
7. Then, the known values are imputed (mostly the transformed columns that were null with the merge and monetary values where null values must be imputed "0").
8. At this point, the data can be optionally used to perform statistics (data dictionaries and plots, for example).
9. Finally, undesirable rows and columns for model training (nulls for example) are removed, binary variables (dummies) are created and the data is normalized, thus leaving the data ready for modelling.

4.5 Chosen Features

To be able to predict the target, features were chosen taking into consideration meetings with the institutions, data understanding, the experts experience and, of course, the DM problem. Table 4-2 shows the IA-SI project's summary of the chosen features.

Feature Type	Feature Count	Features Description
Applicant	58	Application scores, budgeted grant fund, years passed from the company activity start till the application, types of financing (self, owner's capital or other), main economic activity code, amount of sales to foreign countries before starting the project, type of project (new establishment, production increase, production diversification, production re-engineering, etc.), project sector (health, technology, environment, etc.), company and project locations, company size, is the Tax Identification Number – NIF – valid and the NIF type (foreign, individual, collective or other type).
Expense	42	Invoice details (total value, expense value, days between invoice date and payment date, etc.), target (eligible or

		ineligible) and type of expense (marketing, establishment build/ rebuild, hiring new human resources, studies, fairs and exhibitions, machines and equipment, presence in the web, computer software and equipment, etc.).
Financial Information from IES (for the applicant, supplier and consultant)	28	Total assets, assets net income, financial autonomy, turnover growth, net turnover result, financing value, liquidity, return on owner's capital, number of employees, is the financial information missing, etc.
Supplier	18	Is the supplier missing, the supplier participates in other projects with the same suppliers of a given project, number of times this supplier appears across all expenses, number of times this supplier appears in a given project, number of times this supplier appears in a given PPI, number of different projects that are supplied by this supplier, if the supplier only supplies this project, supplier expense value by project, supplier expense value by PPI, total supplier expense value, supplier expense value percentage by project, supplier expense value percentage by PPI, is the NIF valid and the NIF type (foreign, individual, collective or other type).
Consultant	8	Is the consultant missing, number of projects where this consultant appears, if the consultant's NIF was imputed (more about this below), is the NIF valid and the NIF type (foreign, individual, collective or other type).
Combined	4	If this supplier is also the project's consultant, this applicant acts as a consultant in how many projects, this applicant acts as a supplier in how many projects, if this supplier is also a project's partner.
Partner	3	Number of partners by project, does this project have at least one partner and does this project have at least one partner that partners with another project.
Total	**161**	**-**

Table 4-2 IA-SI project's summary of the chosen features.

4.6 Applied General ETL and other Data Warehousing Best Practices

Before exploring what was done in regard to each of CRISP-DM tasks/ ETL subsystems, an overview of the general ETL and other employed data warehousing best practices were:

- In the process of the creation of the features table, a copy was saved in different stages (as per "stage data for recoverability, backup or audit purposes" [36]):
    - Optional initial full backup of previous saved data before making any changes to the ETL program.
    - Backup selected columns for each of the data sources and for each of the used tables.
    - Backup transformed columns for each of the data sources and for each of the used tables.
    - Backup after merging all the transformed columns.
    - Backup after corrections are made.
    - Backup after derived transformations (transformations that combine two or more transformed fields) are made.
    - Backup after imputing missing known values (missing values that we know the value, more about this in section 4.9.1).
    - Backup after imputing missing unknown values (more about this in section 4.9.1).
    - Backup after dropping unwanted rows and columns (more about this in section 4.9).
    - Backup after creating dummie variables (variables that are transformed from categorical to binary, getting the value 0 or 1).
    - Backup after normalizing variables (more about this in section 4.9.2).
- The team created a logical data map (as per "use a logical data map" [36]), that also includes a summary of the quality conclusions (more about this below), with the following information:
    - Institution
    - Source file (table)
    - Source field/ column
    - Source description
    - Source type (quantitative, qualitative, date, etc.)
    - Ambiguities/ inconsistencies/ nonconformities (descriptions)
    - Quality criticality degree (low, medium, high)
    - Quality solutions (descriptions)
    - Destination file (table)
    - Feature type (applicant, supplier, etc.)

    - Destination type (quantitative, qualitative, date, etc.)
    - Destination field/ column
    - Destination description
    - Notes (like alternative fields or other additional information)
    - Selection notes (which rows were selected)
    - Cleaning notes (performed cleansing)
    - Transformation notes (performed transformations)
- The following logical data map columns were used to fill in some columns of the data dictionary (as per "make ETL metadata driven" [36]; more about the data dictionary below):
    - Institution
    - Source file (table)
    - Source field
    - Source description
    - Source type (quantitative, qualitative, date, etc.)
    - Destination file (table)
    - Feature type (applicant, supplier, etc.)
    - Destination type (quantitative, qualitative, date, etc.)
    - Destination field
    - Destination description

4.7 Business Understanding

In the business understanding phase, the author researched about the context (for example information about grant funding), similar case studies (for example documented cases of DM use in the public administration or in the management of grant funds) and the data mining process models and methodologies and the data warehousing best practices that can be applied to a DM process. The goal was to get some background information and to understand what could be learned to be able to set a plan for the DM project.

The team had meetings with the institutions that were very important to understand all the business concepts, like how does the grant funds cycle work and how is the data created, it's meaning and what is an ineligible expense, for example. After some meetings, the team, jointly with the institutions, were able to define the DM project goals: predict expenses ineligibility and predict cancelled projects.

The next task was to design a plan of the project. This was possible due to the experience of the team experts, the undertaken research and the information gathered from the meetings with the institutions.

The last task of the data understanding step was its documentation which produced the following outputs:

- Context and literature review (relevant and public parts previously presented in chapters two and three).
- Grant funds cycle (previously presented).
- DM goals (previously presented).
- Project plan (summarized in figure 4.3 in a Gantt chart, however it is possible that the project is extended until March 2022).

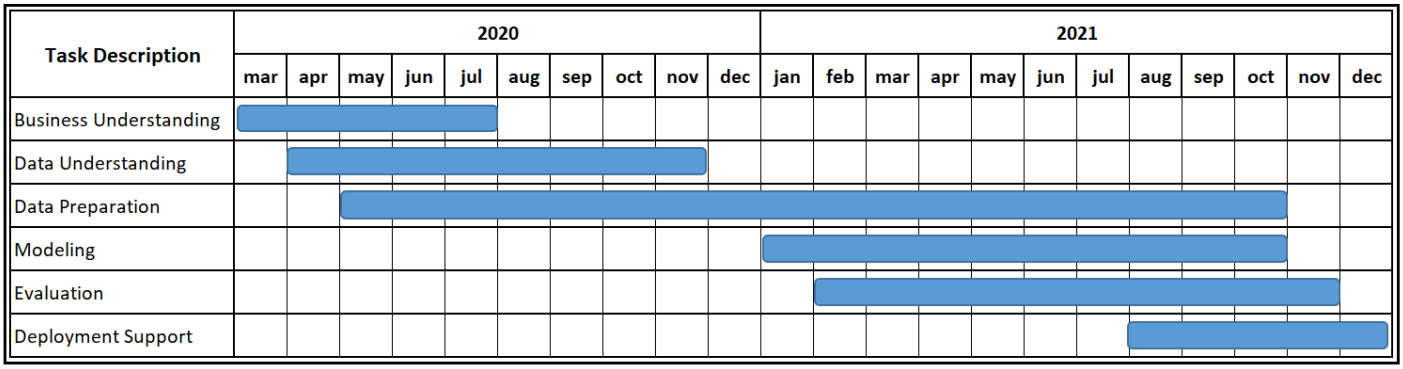

Figure 4.3 Summarized project plan.

4.8 Data Understanding

4.8.1 Collect initial data - Extract (Get the Data Into the DW)

Because data was supplied in the XML format, the team had to extract it to the CSV format (as per "extract includes […] performing low level cleaning like change encodings" [36]). To this end, the XML files' file structure was carefully analysed to infer entities and the relationships between them. From this, the team designed a relational data model to conceptually organize the information into tables/ files and proceeded with the programming. The author helped in these tasks, but was not the main responsible.

The XML files are composed of five data groups produced by the institutions - *Candidaturas*, FACI, PPI, APPI and FACIE – and IES data – *Anexo A* and *Anexo R*.

Regarding the data produced by the institutions, *Candidaturas* data has a XML file per application, the other four types can have more than one. All files have an application number and all files except *Candidaturas* have a document number. This information is on the file name.

For each of the five data groups produced by the institutions, the extracted data is composed of a main table and related tables, that have 1:∞ relationships with the main table. This allows data from the main and the related tables to be joined/ merged. For example, the main *Candidaturas* table has information about the business like its location and size; examples of its related tables are the business partners and employers.

Having this into account, the defined table keys (PK stands for primary key and FK stands for foreign key) for the data produced by the institutions were the following:

Candidaturas:

- Main table: application number (PK, FK)
- Related tables: application number (PK, FK), record number (PK)

FACI, PPI, APPI, FACIE:

- Main table: application number (PK, FK), document number (PK, FK)
- Related tables: application number (PK, FK), document number (PK, FK), record number (PK)

Still regarding the data produced by the institutions, the tag levels indicate how the XML information is transposed to a relational model. Each level 1 tag originates a column prefix, each level 2 tags that don't have more tags inside originate a column suffix. Level 2 tags that have more tags inside are all called Reg, so they are not used in column naming. Each level 3 tags originate a column suffix and imply 1:∞ relationships between the main and the related tables, respectively. For example, imagining a FACI file of application number 123, with the document number 1 that has the following structure (l stands for level and t stands for tag):

```
<l1_t1>
        <l2_t4> data 1 </l2_t4>
        <l2_t5> data 2 </l2_t5>
</l1_t1>
<l1_t2>
        <l2_t6> data 3 </l2_t6>
</l1_t2>
<l1_t3>
        <Reg>
```

```
            <l3_t7> data 4 </l3_t7>
            <l3_t8> data 5 </l3_t8>
            <l3_t9> data 6 </l3_t9>
        </Reg>
        <Reg>
            <l3_t7> data 7 </l3_t7>
            <l3_t8> data 8 </l3_t8>
            <l3_t9> data 9 </l3_t9>
        </Reg>
</l1_t3>
```

This structure originates a record in the FACI's main table, like depicted on table 4-3:

Table 4-3 FACI main table example.

Application number	**Document number**	**l1_t1/l2_t4**	**l1_t1/l2_t5**	**l1_t2/l2_t4**
123	1	data 1	data 2	data 3

And originates two records in the related FACI table called **l1_t3**, depicted on table 4-4:

Table 4-4 FACI related table example.

Application number	**Document number**	**Record number**	**l2_t6**	**l2_t7**	**l2_t8**
123	1	1	data 4	data 5	data 6
123	1	2	data 7	data 8	data 9

The extraction of the IES XML data was done independently by another team member. In this data type, there is a XML file for each NIF and the IES delivery year. As such, the primary keys are the NIF and the IES delivery year, that are also in the file name. The IES XML structure was similar, but less intricate than the other ones, as there are only two tag levels. In the end, all the IES information can be organized in a single table. Only the last IES of a company was used. For example, for the NIF 123 of the year 2020, the pattern:

<l1_t1>

<l2_t4> data 1 </l2_t4>

<l2_t5> data 2 </l2_t5>

</l1_t1>

<l1_t2>

<l2_t6> data 3 </l2_t6>

</l1_t2>

Originates the following IES data (table 4-5):

Table 4-5 IES data.

NIF	Year	l1_t1/l2_t4	l1_t1/l2_t5	l1_t2/l2_t4
123	2020	data 1	data 2	data 3

After the CSV files were saved (as per "staging (writing) the extracted data to disk" [36]) the author of this dissertation built a MySQL relational database using Python and XAMPP. The idea was to have a server with the database, so that all team members could access and prepare variables in SQL language. However, after some data understanding and preparation using MySQL Workspace, this idea was abandoned and only the CSV files were used directly using Python libraries (more about this in section 4.9). This happened because data preparation was too complex to be done in SQL; also the author (or other team members) had limited SQL experience and SQL is much harder to debug.

Having this into account, the author also created a Python script to convert the CSV files to XLSX (Excel). In this fashion, the team was able to easily explore the data with Excel, Modeler, through Python programming, or other means, depending of the team member's tools knowledge. The Excel files helped to explore data and identify relevant data to build the features table (as per "retrieve only the data you need" [36]).

4.8.2 Describe and Explore Data - Data Profiling (Get the Data Into the DW)

Regarding data description and exploration, the author of this dissertation created a data dictionary with Panda's describe, sort_values, value_counts, map and apply, the data model using MySQL Workbench and several visualizations and statistics using Excel (pivot tables and Power Query), seaborn's displot, matplotlib's plot and imbalanced learn's RandomUnderSampler (to balance the datasets).

The data model was created with the MySQL Workbench software through reverse engineering, that is, having all the relational tables, build by the author, with their primary and foreign keys, the data model diagrams were generated automatically. They are quite complex, lengthy, include project's private information and not the main focus of this document. Therefore, only the composition of the institutions produced and extracted XML data is hereupon presented:

- IAPMEI produced data:
    - 168 tables/ files
    - 3906 fields
- AICEP produced data:
    - 135 tables/ files
    - 3362 fields

The built data dictionary has the following columns:

- Source
- Field
- Description
- Type (quantitative, qualitative, date, etc.)
- Institution
- Feature type (applicant, supplier, etc.)
- Average
- Standard deviation
- Minimum value
- 2% quartile
- 5% quartile
- 25% quartile
- Median
- 75% quartile
- 95% quartile
- 98% quartile
- Maximum value
- Number of not nulls
- Percentage of not nulls
- Number of nulls
- Percentage of nulls
- Number of different values

- Data types (int, float, etc.)
- Most frequent values top 20 (top 20 most frequent values in the format {value: frequency of value})
- Less frequent values top 20 (top 20 less frequent values in the format {value: frequency of value})
- Number of the most frequent value (the most frequent value in the format {value: frequency of value})
- Number of the less frequent value (the less frequent value in the format {value: frequency of value})
- Percentage of the most frequent value
- Percentage of the less frequent value

Visualization of data through distribution plots was also created by the author of this study, using Python, which also contributed to a deeper data understanding. In the probability density plots, data was under sampled to balance the target classes and values were rescaled using the logarithm of base 10 for a better visualization. Probability density plots were used for quantitative variables and stacked bars plots were used for qualitative variables. Some plots examples are ilustrated in figures from 4.4 to 4.9:

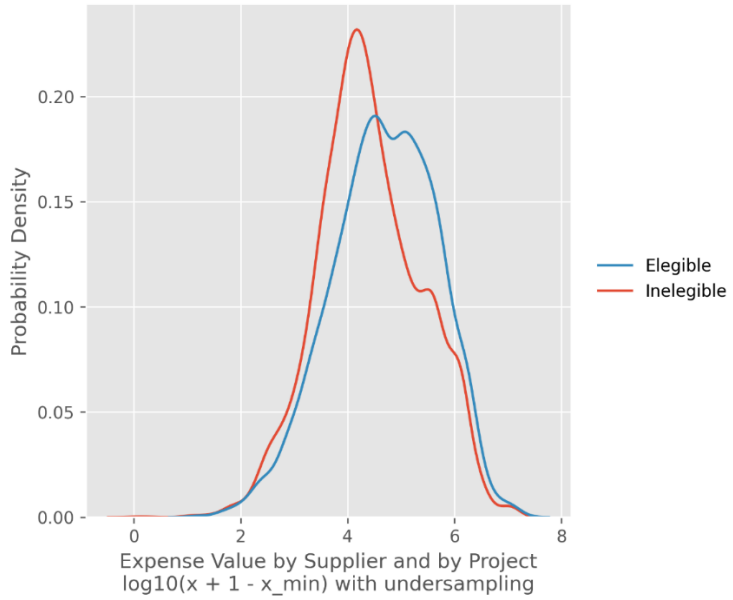

Figure 4.4 Target distribution plot of the expense value by supplier and by project (IAPMEI).

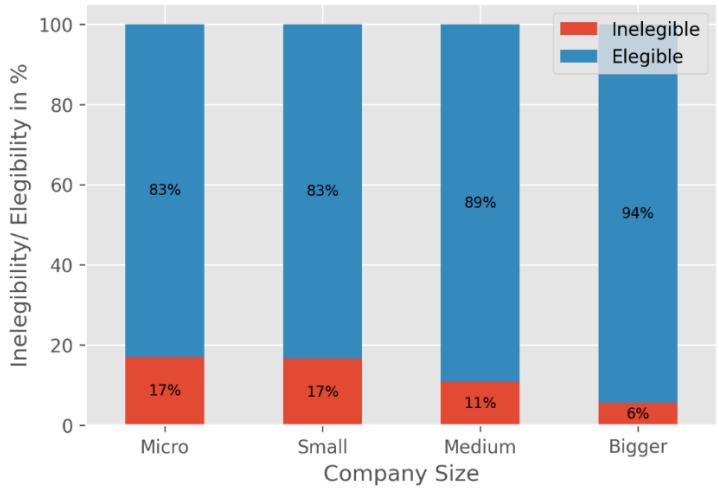

Figure 4.5 Target distribution plot of the company size (IAPMEI).

By analysing figures 4.4 and 4.5, it can be concluded that smaller expenses and companies have a bit more ineligibility than bigger ones.

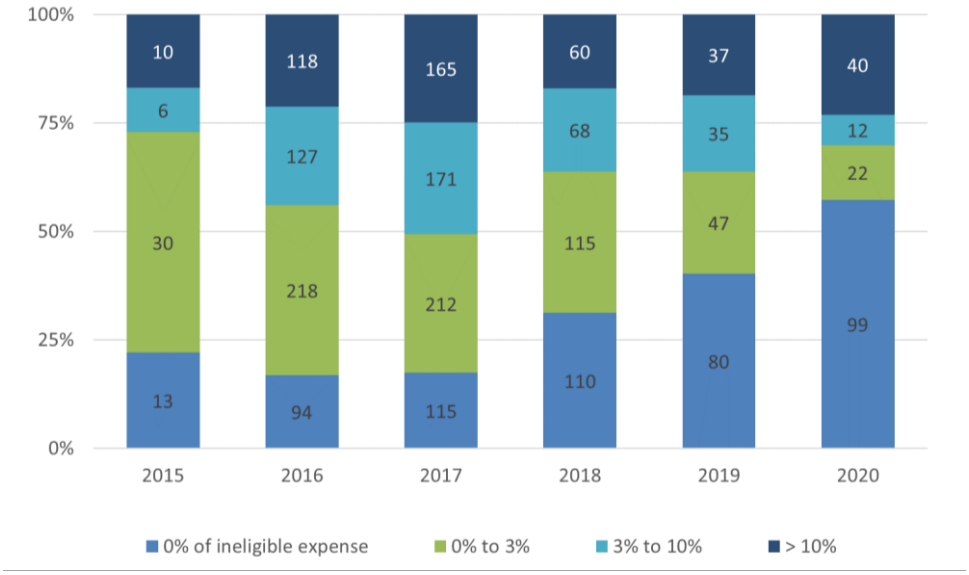

Figure 4.6 Proportion of application's ineligible expenses in eligible expenses per application year (IAPMEI).

Besides the former, figures from 4.7 to 4.9 illustrate general statistics were also produced (again by the author):

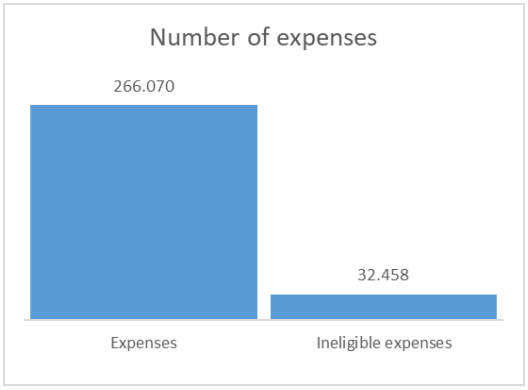

Figure 4.8 Number of expenses (AICEP).

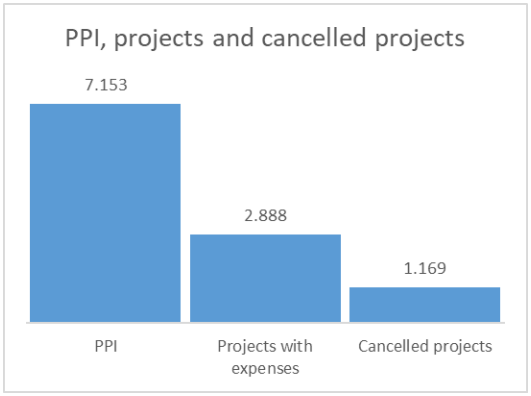

Figure 4.7 PPI, projects and cancelled projects (AICEP).

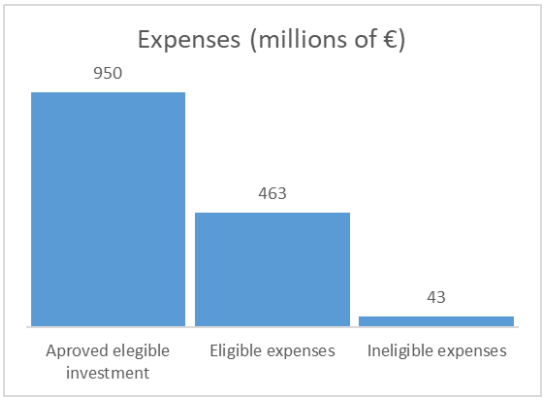

Figure 4.9 Expenses, millions of € (AICEP).

Other produced statistics that won't be presented here due to data confidentiality or scope management reasons are:

- Number and percentage of projects, percentage of canceled projects, total and percentage of expenses in millions of € and percentage of ineligible expense per:
  - Region, sector of activity and company size;
- Number and type of canceled projects per its application year;
- Number of projects per end delay years relative to the application end year (to measure how many projects had one or more years of delay relative to their initial plan and how many met the schedule).

The data exploration and data profiling activities were crucial to identify relevant variables for modeling and provided some hints about the data quality.

4.8.3 Verify Data Quality - Data Profiling, Data Cleansing System and Data Quality Screens, Error Event Tracking and Audit Dimension Creation (Get the Data Into the DW and Clean and Conform)

As CRISP-DM and ETL advise, the team performed the data quality verification both before transformations (after the extraction) and also after transformations. Also, the team confirmed that the quality verification process is a longstanding one that, sometimes, delayed dependent project tasks (as Kimball advocates in [36]).

Most of the data quality verifications were performed by another member of the team, not the author of this dissertation. To verify the data quality, she used the data dictionaries, created by the author, and Modeler's data audit node, which provided the following extra information that was not in the data dictionary:

- Skewness
- Number of outliers (between 3 and 5 standard deviations above or below the average)
- Number of extreme outliers (more than 5 standard deviations above or below the average)
- Number of empty strings
- Number of white spaces

This information could have been also included in the data dictionary, but was kept separate to keep it simpler.

Through the analysis of these fields and the data dictionary, as mentioned earlier, a description of the ambiguities/ inconsistencies/ nonconformities, the quality criticality degree (low, medium, high) and a description of the quality solutions was created and documented in the logical data map.

Besides these quality checks, the author verified the "primary and foreign keys and referential integrity" [36] and, whenever two files were merged/ joined, the result's number of rows and columns was verified to insure its correctness.

As [36] asserts, found errors during quality verifications can help to improve systems and reduce costs. The team found the following errors that can lead to the institutions source systems' improvement:

- There are null fields that should be mandatory:
    - Many monetary values are null in the data, but they should be "0". The system could fill them with a "0" placeholder before people introduce values.
    - For the consultants' and suppliers' NIF, there could be a "no consultant" and a "foreign supplier" option, respectively, so that the field is never filled in blank (as [36] advises).

- There are fields that have errors that can be avoided with automatic checks:
    - NIF is not validated. There is, for each country, an algorithm to verify if the introduced NIF is correct that can be used.
    - Dates are not validated. This is mostly done with business or other rules, e.g., an expense date has to come before the expense payment date and after the company is formed.
    - Fields don't have all the same format. E.g.: some dates are in dd-mm-yy format, others in yy/mm/dd format; some monetary values have thousands separators, others don't; some monetary values have a "," decimal separator, others a "."; some monetary values have unwanted spaces.
    - Some values are out of the correct domain. E.g.: negative values in expenses and number of workers that is not concordant with the company's size.

4.9 Data Preparation

Data Preparation was performed by the author using Python. The first step of the data preparation was an initial variables selection. This was done across all of the relevant disparate tables. Then the data had to be merged into a single table – the features table. After, this data was transformed, cleaned and imputed. Transformations were done using Python libraries such as Pandas' groupby, unstack, transform, fillna, merge and apply, NumPy's where and scikit-learn.

Then, there was a need to delete some affected records or columns, using Pandas' dropna. There were some variables that were used to build the final modeling variables that were deleted, as they were not intended to be used in the modeling task. Also, all the columns that had more than 5% of missing values were deleted, using the code df.dropna(axis=1, thresh=(len(df)-(len(df)*0.05)))). Finally, all records with at least a missing value were also deleted.

After, the dummie variables were created using Pandas' get_dummies. The reason for this being done after deleting records and columns was because some categories could disappear with the deletions. Therefore, a dummie variable with nulls in all values would be useless. Finally, the categoric variables that were used to build the dummies were also deleted.

The data preparation task was very arduous and lengthy not only because of the number of different variables to be created, but also because of the complexity of some transformations. Also, having started to perform some transformations using SQL delayed the process because it was much harder to do transformations using SQL than Python, partly due to the team's SQL inexperience, partly because SQL is much harder to debug.

4.9.1 Clean Data - Data Cleansing System and Data Quality Screens (Clean and Conform)

The data was cleaned according to the issues found in the data quality verifications and according to the performed data understanding, using Pandas' replace, fillna and apply and Numpy's where.

Because of the existing missing values, it is worthwhile to summarize the undertaken imputations:

- Null partners were imputed in some cases with the value "1" using the company's nature (if the company is a single person one).
- Null consultants were imputed using the expense and supplier nature (if expenses were consulting expenses and if the supplier was a consultant).
- Some null monetary values were imputed with their known value: "0".
- Some nulls were imputed with a "null" string in order to build a dummie variable to be able to flag that value as a missing.
- The true missing IES values were imputed with "0". Another version of the true missing IES values imputations was also created with imputations a mean below the minimum value across records to test if the predictive results could improve. The records with missing IES values that shouldn't be missing weren't used to train or test the DM models.

4.9.2 Integrate and Format Data - Deduplication, Data Conformance and Aggregate Builder (Clean and Conform and Prepare for Delivery)

The required performed integration and conforming was:

- The data from disparate sources had to be merged into a single table – the features table.
- Dates and monetary values were standardized to the same format.
- APPI and FACIE expenses were integrated into a single table in which the relative origin was flagged.

- Records were aggregated to build some features, e.g.: sums by project, by PPI and by supplier.

Finally, being the last step before modeling, all monetary values were normalized using sklearn's preprocessing normalize method. This transforms all values between 0 and 1. This was done to decrease these features' scales so that they could have similar scales of other features, as most are between 0 and 1. This is done because, when using different scales, features with bigger values can have a bigger influence in the results, even though the predictive importance is not superior comparing to other features.

4.10 Deployment

The team is only supporting deployment; is not going to be fully performed by the team.

4.10.1 Plan Monitoring and Maintenance - Change Data Capture (Get the Data Into the DW)

In the software prototype components that the author programmed, historical data was used to build features that are used to train the machine learning models. In the final deployed version, when new expenses (PPI) arrive in the system, in order to predict their ineligibility and to build some of the aggregated features (like supplier expense value by project and total supplier expense value, for example), they have to be updated using the old data. Also, as time passes, it is possible that the data characteristics change; in this case, the DM models have to be updated. For these two purposes, ETL's change data capture options (see section 3.2.7) are available and are useful to be analyzed and chosen for deployment by the institutions, with the team's support.

4.12 Chapter Closure

In this chapter, the case study was presented, organized by the CRISP-DM phases and tasks and their correspondent ETL subsystems. The next and final chapter (five) presents the conclusions, contributions and limitations of this study and delineates future work that can be done.

CHAPTER 5

5 Conclusion

This study researched and compared the most used DM process models and concludes that CRISP-DM includes all the other DM process models' tasks. CRISP-DM provides much detail that is valuable to inexperienced practitioners, especially in the user guide section [71]. As such, this dissertation agrees with [25], [48]: CRISP-DM is the *de facto* DM process model, even after 21 years of its publication.

Because CRISP-DM includes all the other DM process models' tasks, this study compared and extended CRISP-DM with concepts derived from the data warehousing best practices. These extensions bring additional information to CRISP-DM, especially regarding data quality. The biggest contribution of the data warehousing best practices to be applied in a DM process are the general heuristics that can't be found in the CRISP-DM descriptions.

Another contribution of this study was the application of the extended CRISP-DM with data warehousing best practices to the IA-SI project its documentation to share to foster knowledge reuse.

CRISP-DM was useful to the project because it served as a good basis so that the non-expert team members could learn how a DM project should be done, since:

- It includes all the other DM process models' tasks; it is comprehensive.
- It provides a lot of detail regarding tasks and outputs that belong to each generic task; it is detailed.
- It can be easily adapted and generalized to different DM projects; it is flexible and versatile.

The data warehousing best practices extensions to CRISP-DM were useful to the project because:

- Frequent data backups allowed an easier debugging and contributed to the data mining custom made program's modularization.
- The logical data map provided a structured way to document all the extensive and complex transformations and improved data traceability and, therefore, communication between team members and also with the institutions.
- By programmatically copying some information from the logical data map to the data dictionary, harmonization was ensured between the two documents.
- It offed organized change data capture options that can be chosen by the institutions with the team's advice and support.
- If provides extra information that is not in CRISP-DM's descriptions that is useful to use in a DM project.

Even if data used in a DM problem comes from a data warehouse (DW), most of the data warehousing best practices still apply. For example, the data still needs to be transformed in some way.

Herewith, this dissertation concludes that the data warehousing best practices can and should be applied to a DM problem. They extend the information presented in CRISP-DM and provide a positive contribution when applied to a DM project.

To sum up, table 5-1 presents the contributions of this dissertation:

Table 5-1 Goals, contributions and conclusions summary.

Goal	Contribution	Conclusion
1. Understand which is the most comprehensive DM process model or methodology to be the reference followed in the project.	An analysis, comparison and mapping between the most common DM process models and their tasks.	This showed that CRISP-DM includes all of the other DM process models' tasks, so there is no need to create a new DM process model that is a combination of the others. CRISP-DM provides much detail that is valuable to inexperienced practitioners, especially in the user guide section [71]. CRISP-DM, being comprehensive, detailed, flexible and versatile, served as the base reference to be followed in the project.
2. Discover if and how can the data warehousing best practices be applied in a DM problem and extend the reference DM process model or methodology.	A selection of the data warehousing best practices that can be applied to a DM problem.	There are many data warehousing best practices that can be applied to a DM problem.
	An analysis, comparison and mapping between CRISP-DM's tasks and the ETL's subsystems and other data warehousing best practices.	ETL and other data warehousing best practices contribute to extend CRISP-DM.

3. Document the application of the data warehousing best practices in the project, so that it can be shared to other researchers and practitioners.	Application and documentation of the extended CRISP-DM to the project.	The data warehousing best practices were useful to the project, as explained above. Even if data used in a DM problem comes from a data warehouse (DW), most of the data warehousing best practices still apply. For example, the data still needs to be transformed in some way.
4. Develop software prototype components related to data understanding and preparation tasks for the ineligibilities DM problem of the IA-SI project. 5. Contribute to the United Nation's "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all" sustainable development goal (goal 8) [12] through the support of the project's development.	Contribution to the prototype implementation on tasks that encompassed data extraction, validation, transformation, data set creation and documentation of the full process, focusing on the prediction of ineligible expenses DM goal. This was done mostly through Python programming, since the team didn't use any commercial tool. Excel, Modeler and MySQL Workbench were also used.	Data preparation is a meticulous and slow process just as Pyle advocates, taking up 60 to 90% of the time needed in DM [22]. Using SQL to perform some transformations also slowed the process because it was much harder using SQL than Python, partly due to the team's SQL inexperience, partly because SQL is much harder to debug.

A limitation of this dissertation is that the real time possibility of ETL hasn't been explored. The IA-SI project doesn't require real time updates in its system, but other DM projects could benefit from that analysis. Other limitation is the fact that the ETL and other data warehousing best practices were only applied and documented for one project. There can be more insights to share with others if they are applied to more projects.

Thus, as future work:

- Other researchers can apply the selected data warehousing best practices to other projects and document their insights.
- Other researchers can explore how can ETL's real time updates be applied in a DM project and the benefits therein.
- Other experienced DS practitioners can try to extend CRISP-DM further with their own heuristics or other heuristics taken from other area, similarly to what this study attempted (data science area *versus* data warehousing area).
- More DM case studies that test all the new heuristics can be documented and shared with others.
- All the new extensions can possibly be organized and compiled to create a "new" CRISP-DM.

Statement of Independent Work

I hereby confirm that this dissertation was written independently by myself without the use of any sources beyond those cited, and all passages and ideas taken from other sources are cited accordingly.

References

[1] IAPMEI, “Estado aposta na inteligência artificial para acelerar fundos,” 2019. [Online]. Available: https://www.iapmei.pt/getattachment/a2bd09ec-996e-4098-b1ad-28e5dc9dd14a/Expresso09112019.pdf.aspx?lang=pt-PT. [Accessed: 11-Jan-2021].

[2] Portugal 2020, “Financiamento para usar Inteligência Artificial na gestão de Sistemas de Incentivos | Portugal 2020,” 2019. [Online]. Available: https://www.portugal2020.pt/content/financiamento-para-usar-inteligencia-artificial-na-gestao-de-sistemas-de-incentivos. [Accessed: 11-Jan-2021].

[3] Compete 2020, “Sobre nós,” 2017. [Online]. Available: https://www.compete2020.gov.pt/sobre-nos/Missao. [Accessed: 11-Jan-2021].

[4] E. Commission, “Managing Authorities.” [Online]. Available: https://ec.europa.eu/regional_policy/en/atlas/managing-authorities//. [Accessed: 05-Aug-2021].

[5] E. Commission, “ARACHNE risk scoring tool.” [Online]. Available: https://ec.europa.eu/social/main.jsp?catId=325&intPageId=3587&langId=en. [Accessed: 05-Aug-2021].

[6] E. Commission, “ARACHNE - Presentation on Arachne usage by Interreg programmes.” [Online]. Available: https://www.interact-eu.net/library?title=arachne&field_fields_of_expertise_tid=All&field_networks_tid=All. [Accessed: 05-Aug-2021].

[7] E. Commission, “ARACHNE PROJECT 2014.” [Online]. Available: https://ec.europa.eu/regional_policy/archive/conferences/anti_corruption/2014_10/doc/wg2_arachne.pdf. [Accessed: 05-Aug-2021].

[8] Incode 2030, “Governo promove boas práticas de utilização de inteligência artificial na Administração Pública | Portugal INCoDe.2030.” [Online]. Available: https://www.incode2030.gov.pt/newsletter/01/governo-promove-boas-praticas-de-utilizacao-de-inteligencia-artificial-na-administracao-publica. [Accessed: 11-Jan-2021].

[9] AICEP, “Projetos com apoio comunitário.” [Online]. Available: http://portugalglobal.pt/PT/sobre-nos/projetos-apoio-comunitario/Paginas/pac.aspx. [Accessed: 11-Jan-2021].

[10] IAPMEI, “IAPMEI - Documentos - Incentivos.” [Online]. Available: https://www.iapmei.pt/PRODUTOS-E-SERVICOS/Incentivos-Financiamento/Documentos-Incentivos.aspx. [Accessed: 11-Jan-2021].

[11] Diário da República, “Diário da República Eletrónico,” 2014. [Online]. Available: https://dre.pt/legislacao-consolidada/-/lc/115287303/201904182056/diploma/3?rp=indice. [Accessed: 11-Jan-2021].

[12] United Nations, “THE 17 GOALS.” [Online]. Available: https://sdgs.un.org/goals. [Accessed: 31-Aug-2021].

[13] Assembleia da República, “Constituição da república portuguesa,” *Assembleia da República Portuguesa*, 2005. [Online]. Available: https://www.parlamento.pt/Legislacao/Paginas/ConstituicaoRepublicaPortuguesa.aspx. [Accessed: 11-Jan-2021].

[14] Portugal 2020, “O que é o Portugal 2020 | Portugal 2020.” [Online]. Available: https://www.portugal2020.pt/content/o-que-e-o-portugal-2020. [Accessed: 11-Jan-2021].

[15] European Commission, “European Structural and Investment Funds | European Commission.” .

[16] FCT, “FCT — Projectos — Concursos.” [Online]. Available: https://www.fct.pt/apoios/projectos/concursos/2017/index.phtml.pt. [Accessed: 11-Jan-2021].

[17] I. Martinez, E. Viles, and I. G. Olaizola, “Data Science Methodologies: Current Challenges and Future Approaches,” *Big Data Res.*, vol. 24, p. 100183, 2021, doi: 10.1016/j.bdr.2020.100183.

[18] F. Ruiz-lopez, J. Perez-ortega, J. Ortiz-hernandez, and Y. Hernandez-perez, “Systematic Review of Methodologies in Data Science,” 2021.

[19] IBM, “Data Science,” 2020. [Online]. Available: https://www.ibm.com/cloud/learn/data-science-introduction. [Accessed: 31-Aug-2021].

[20] J. Saltz, N. Hotz, D. Wild, and K. Stirling, “Exploring project management methodologies used within data science teams,” *Am. Conf. Inf. Syst. 2018 Digit. Disruption, AMCIS 2018*, pp. 1–5, 2018.

[21] W. Y. Ayele, “Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020, doi: 10.14569/IJACSA.2020.0110603.

[22] P. M. Gonçalves and R. S. M. Barros, “Automating data preprocessing with DMPML and KDDML,” *Proc. - 2011 10th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2011*, pp. 97–103, 2011, doi: 10.1109/ICIS.2011.23.

[23] M. Charest, S. Delisle, O. Cervantes, and Y. Shen, “Intelligent data mining assistance via CBR and ontologies,” *Proc. - Int. Work. Database Expert Syst. Appl. DEXA*, pp. 593–597, 2006.

[24] J. L. Kolodner, “An Introduction to Case-Based Reasoning,” doi: 10.1136/bmj.4.5576.398.

[25] G. Mariscal, Ó. Marbán, and C. Fernández, “A survey of data mining and knowledge discovery

process models and methodologies," *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, 2010, doi: 10.1017/S0269888910000032.

[26] L. Cao, "Domain-driven data mining: Challenges and prospects," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 755–769, 2010, doi: 10.1109/TKDE.2010.32.

[27] "Kimball Techniques." [Online]. Available: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/. [Accessed: 31-Aug-2021].

[28] "Kimball DW/BI Lifecycle Methodology." [Online]. Available: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dw-bi-lifecycle-method/. [Accessed: 31-Aug-2021].

[29] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, "The Data Warehouse Lifecycle Toolkit, 2nd Edition: Practical Techniques for Building Data Warehouse and Business Intelligence Systems," 2008.

[30] A. C. Onal, O. Berat Sezer, M. Ozbayoglu, and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 2037–2046, 2017, doi: 10.1109/BigData.2017.8258150.

[31] O. Gervasi *et al.*, *Computational Science and Its Applications – ICCSA 2020*. 2020.

[32] M. Unger, "Data acquisition and the implications of machine learning in the development of a Clinical Decision Support system," *Proc. - 2021 IEEE/ACM 1st Work. AI Eng. - Softw. Eng. AI, WAIN 2021*, pp. 101–104, 2021, doi: 10.1109/WAIN52551.2021.00022.

[33] S. Batasova, M. Efimova, I. Kholod, and A. Semenchenko, "Preparation of distributed heterogeneous data for data mining," *Proc. Int. Conf. Soft Comput. Meas. SCM 2015*, pp. 205–207, 2015, doi: 10.1109/SCM.2015.7190457.

[34] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Inf.*, vol. 11, no. 4, 2020, doi: 10.3390/info11040193.

[35] Two Crows, *Introduction to Data Mining and Knowledge Discovery*. 1999.

[36] R. Kimball and J. Caserta, *The Data Warehouse ETL Toolkit*. 2004.

[37] "About Google Scholar." [Online]. Available: https://scholar.google.com/intl/en/scholar/about.html. [Accessed: 28-Apr-2021].

[38] ISCTE, "Bases de dados e localizadores de recursos - Iscte – Instituto Universitário de Lisboa." [Online]. Available: https://www.iscte-iul.pt/conteudos/estudantes/biblioteca/recursos/532/bases-de-dados-localizadores-de-recursos. [Accessed: 11-Jan-2021].

[39] J. Peixoto, "Qual o impacto da Inteligência Artificial nos serviços públicos?" [Online].

Available: http://www.idcdx.pt/diretorio/qual-o-impacto-da-inteligencia-artificial-nos-servicos-publicos/. [Accessed: 11-Jan-2021].

[40] Incode 2030, “O desafio da inteligência artificial na Administração Pública | Portugal INCoDe.2030.” [Online]. Available: https://www.incode2030.gov.pt/newsletter/01/o-desafio-da-inteligencia-artificial-na-administracao-publica. [Accessed: 11-Jan-2021].

[41] PWC, “DG REGIO – Preventing fraud and corruption in the European Structural and Investment Funds – taking stock of practices in the EU Member States.” [Online]. Available: https://ec.europa.eu/regional_policy/sources/docgener/studies/pdf/implem_article125_fraud_en.pdf. [Accessed: 01-Sep-2021].

[42] European Commission, “Operational programme.” [Online]. Available: https://ec.europa.eu/regional_policy/en/policy/what/glossary/o/operational-programme. [Accessed: 01-Sep-2021].

[43] C. Bratsas, E. Chondrokostas, K. Koupidis, and I. Antoniou, “The use of national strategic reference framework data in knowledge graphs and data mining to identify red flags,” *Data*, vol. 6, no. 1, pp. 1–20, 2021, doi: 10.3390/data6010002.

[44] “Subsidystories.eu.” [Online]. Available: http://subsidystories.eu/. [Accessed: 01-Sep-2021].

[45] “Monitoring European Tenders.” [Online]. Available: http://digiwhist.eu/. [Accessed: 01-Sep-2021].

[46] “Red Flags.” [Online]. Available: https://www.redflags.eu/. [Accessed: 01-Sep-2021].

[47] “Red Flags project.” [Online]. Available: https://www.redflags.eu/files/redflags-summary-en.pdf. [Accessed: 01-Sep-2021].

[48] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2019, doi: 10.1109/tkde.2019.2962680.

[49] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework,” vol. 9, no. 6, pp. 851–860, 1996.

[50] P. C. Ncr *et al.*, “CRISP-DM,” *SPSS inc*, vol. 78, pp. 1–78, 2000.

[51] C. Gertosio and A. Dussauchoy, “Knowledge discovery from industrial databases,” *J. Intell. Manuf.*, vol. 15, no. 1, pp. 29–37, 2004, doi: 10.1023/B:JIMS.0000010073.54241.e7.

[52] A. G. Buchner, M. D. Mulvenna, S. S. Anand, and J. G. Hughes, “An Internet-Enabled Knowledge Discovery Process,” *Proc. 9th Int. Database Conf. Hong Kong*, vol. 1999, no. February 2014, pp. 13–27, 1999.

[53] SAS, “Introduction to SEMMA,” 2017. [Online]. Available: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm. [Accessed: 31-Aug-2021].

[54] K. J. Cios and L. A. Kurgan, *Six-Step Knowledge Discovery and Data Mining Process*. 2005.

[55] S. Moyle and A. Jorge, "RAMSYS - A methodology for supporting rapid remote collaborative data mining projects," *Proc. ECML/PKDD'01 Work. Integr. Asp. Data Mining, Decis. Support Meta-Learning*, no. 1, pp. 20–31, 2001.

[56] J. Solarte, "A Proposed Data Mining Methodology and its Application to Industrial Engineering," 2002.

[57] Ó. Marbán, G. Mariscal, and J. Segovia, "A Data Mining & Knowledge Discovery Process Model," *IntechOpen*, no. February, p. 436, 2009.

[58] John B. Rollins, "Foundational Methodology for Data Science," *IBM Anal.*, pp. 1–4, 2015.

[59] IBM Corporation, "Interactive Analytics Solutions Unified Method (ASUM)," 2015. [Online]. Available: http://i2t.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html_wbs.html?proc=_0eKIHlt6EeW_y7k3h2HTng&path=_0eKIHlt6EeW_y7k3h2HTng. [Accessed: 31-Aug-2021].

[60] IBM Corporation, "ASUM: Analytics Solutions Unified Method," *Anal. Serv. Datasheet*, 2016.

[61] Microsoft, "What is the Team Data Science Process?," 2020. [Online]. Available: https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview. [Accessed: 31-Aug-2021].

[62] F. Martínez-Plumed *et al.*, "CASP-DM: Context Aware Standard Process for Data Mining," pp. 1–38, 2017.

[63] T. Pyzdek, *The Six Sigma Project Planner*. 2003.

[64] J. C. W. Debuse, B. De Iglesia, C. M. Howard, and V. J. Rayward-Smith, "A methodology for knowledge discovery : a KDD roadmap," *SYS Tech. Rep. SYS-C99-01*, no. 2552, pp. 1–20, 1999.

[65] F. Ascacibar, *Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado*. 2003.

[66] S. L. Andresen, "John McCarthy: Father of AI," *IEEE Intell. Syst.*, vol. 17, no. 5, pp. 84–85, 2002, doi: 10.1109/MIS.2002.1039837.

[67] J. M. Helm *et al.*, "Machine Learning and Artificial Intelligence : Definitions , Applications , and Future Directions," pp. 69–76, 2020.

[68] J. N. Kok, E. J. W. Boers, W. A. Kosters, P. Van Der Putten, and M. Poel, "ARTIFICIAL INTELLIGENCE: DEFINITION, TRENDS, TECHNIQUES, AND CASES."

[69] J. Mccarthy, "What is Artificial Intelligence?," pp. 1–14, 2004.

[70] Stanford, "Arthur Samuel: Pioneer in Machine Learning." [Online]. Available: http://infolab.stanford.edu/pub/voy/museum/samuel.html. [Accessed: 31-Aug-2021].

[71] U. Fayyad and H. Hamutcu, "Analytics and Data Science Standardization and Assessment

Framework," *Harvard Data Sci. Rev.*, no. 2, pp. 1–33, 2020, doi: 10.1162/99608f92.1a99e67a.

[72] IBM, "Business analytics." [Online]. Available: https://www.ibm.com/analytics/business-analytics. [Accessed: 31-Aug-2021].

[73] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, 2011, doi: 10.1016/j.dss.2010.08.006.

[74] C. Kleissner, "Data mining for the enterprise," *Proc. Hawaii Int. Conf. Syst. Sci.*, vol. 7, no. c, pp. 295–304, 1998, doi: 10.1109/hicss.1998.649224.

[75] C. Sapia, G. Hofling, M. Muller, C. Hausdorf, H. Stoyan, and U. Grimmer, "On supporting the data warehouse design by data mining techniques," *Venue GI-Workshop Data Min. Data Warehous.*, pp. 1–8, 1999.

[76] S. Choenni and A. Siebes, "Query optimization to support data mining," *Int. Conf. Database Expert Syst. Appl. - DEXA*, no. October 1997, pp. 658–663, 1997, doi: 10.1109/dexa.1997.617408.

[77] M. A. Jeusfeld, C. Quix, and M. Jarke, "Design and analysis of quality information for data warehouses," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1507, no. 1, pp. 349–362, 1998, doi: 10.1007/978-3-540-49524-6_28.

[78] "ETL Architecture's 34 Subsystems - Kimball Group." [Online]. Available: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/etl-architecture-34-subsystems/. [Accessed: 09-May-2021].

[79] W. H. Inmon, *Building the Data Warehouse*. 2002.

[80] M. Breslin, "Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon models," pp. 6–20, 2004.

[81] R. Kimball, *The Data Warehouse Toolkit (Second Edition)*, vol. 45, no. 2. 2003.

[82] A. Shollo and K. Kautz, "Towards an understanding of business intelligence," *ACIS 2010 Proc. - 21st Australas. Conf. Inf. Syst.*, 2010.

[83] Q. Yang, M. Ge, and M. Helfert, "Analysis of data warehouse architectures: Modeling and classification," *ICEIS 2019 - Proc. 21st Int. Conf. Enterp. Inf. Syst.*, vol. 2, pp. 604–611, 2019, doi: 10.5220/0007728006040611.

[84] V. Filatov, V. Semenets, and O. Zolotukhin, "DATA MINING IN RELATIONAL SYSTEMS," vol. 3, no. 13, pp. 65–76, 2020.

[85] SAS, "ETL, What it is and why it matters." [Online]. Available: https://www.sas.com/en_us/insights/data-management/what-is-etl.html. [Accessed: 31-Aug-2021].

[86] “Oxford Advanced Learner’s Dictionary: process.” [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/english/process1_1?q=process. [Accessed: 26-May-2021].

[87] S. Tyrrell, “The Many Dimensions of the Software Process.” [Online]. Available: https://dl.acm.org/doi/fullHtml/10.1145/333424.333435. [Accessed: 26-May-2021].

[88] C. Shearer, “First CRISP-DM 2.0 Workshop Held,” 2006. [Online]. Available: https://www.kdnuggets.com/news/2006/n19/4i.html. [Accessed: 31-Aug-2021].

[89] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. 2015.

Appendices

Appendix A - ETL Data Flow and Planning & Design Thread Details

A.1 Data Flow Thread

Extract includes:

- Reading source-data models.
- Connecting to and accessing data.
- Sometimes performing low level cleaning like change encodings.
- Scheduling the source system, intercepting notifications and daemons.
- Capturing changed data.
- Staging (writing) the extracted data to disk.

Clean involves:

- Enforcing column properties.
- Enforcing structure.
- Enforcing data and value rules.
- Enforcing complex business rules.
- Building a metadata foundation to describe data quality.
- Staging (writing) the cleaned data to disk.

Conform includes:

- Standardizing.
- Conforming business labels (in dimensions).
- Conforming business metrics and performance indicators (in fact tables).
- Deduplicating.
- Householding.
- Internationalizing.
- Staging (writing) the conformed data to disk.

Deliver involves:

- Loading flat and snowflaked dimensions.
- Generating time dimensions.
- Loading degenerate dimensions.
- Loading subdimensions.

- Loading types 1, 2, and 3 slowly changing dimensions.
- Conforming dimensions and conforming facts.
- Handling late-arriving dimensions and late-arriving facts.
- Loading multi-valued dimensions.
- Loading ragged hierarchy dimensions.
- Loading text facts in dimensions.
- Running the surrogate key pipeline for fact tables.
- Loading three fundamental fact table grains.
- Loading and updating aggregations.
- Staging the delivered data to disk.

The previous steps of the data flow are overseen and accompanied by the operations step, which includes:

- Scheduling.
- Job execution.
- Exception handling.
- Recovery and restart.
- Quality checking.
- Release.
- Support.

A.2 Planning & Design Thread

Requirements and realities include:

- Business needs.
- Data profiling and other data-source realities.
- Compliance requirements.
- Security requirements.
- Data integration.
- Data latency.
- Archiving and lineage.
- End user delivery interfaces.
- Available development skills.
- Available management skills.

- Legacy licenses.

Architecture involves deciding about:

- Hand-coded versus ETL vendor tool.
- Batch versus streaming data flow.
- Horizontal versus vertical task dependency.
- Scheduler automation.
- Exception handling.
- Quality handling.
- Recovery and restart.
- Metadata.
- Security.

System implementation includes:

- Hardware.
- Software.
- Coding practices.
- Documentation practices.
- Specific quality checks.

Test and release involve the design of the:

- Development systems.
- Test systems.
- Production systems.
- Handoff procedures.
- Update propagation approach.
- System snapshotting and rollback procedures.
- Performance tuning.

Appendix B - CRISP-DM Details

B.1 CRISP-DM Phases' Descriptions

[50] defines each phase:

Business understanding: "This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives".

Data understanding: "The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information".

Data preparation: "The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

Modeling: "In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary".

Evaluation: "At this stage in the project, you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached".

Deployment: “Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying “live” models within an organization’s decision making processes— for example, real-time personalization of Web pages or repeated scoring of marketing databases. 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 across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models”.

B.2 CRISP-DM Summary Tables (Tasks, Notes and Outputs)

Table 1 Business Understanding tasks and outputs, adapted from [50].

Business Understanding		
Generic Task	**Task Notes**	**Output**
Determine business objectives	What are the customer goals?	Background (information about the business)
		Business objectives
		Business success criteria
Assess situation	Assess resources, constraints, assumptions and other factors.	Inventory of resources
		Requirements, assumptions, and constraints
		Risks and contingencies
		Terminology
		Costs and benefits
Determine DM goals	What do we want to predict?	DM goals
		DM success criteria
Produce project plan	Include the initial selection of tools and techniques.	Project plan
		Initial assessment of tools and techniques

Table 2 Data Understanding tasks and outputs, adapted from [50].

Data Understanding		
Generic Task	**Task Notes**	**Output**
Collect initial data	Load the data and integrate it if applicable.	Initial data collection report
Describe data	Describe its general properties.	Data description report
Explore data	Use querying, visualization, and reporting techniques.	Data exploration report
Verify data quality	Completeness, correctness, etc.	Data quality report

Table 3 Data Preparation tasks and outputs, adapted from [50].

Data Preparation		
Generic Task	**Task Notes**	**Output**
Select data	Columns and rows used in the analysis.	Rationale for inclusion/exclusion
Clean data	Insert defaults, estimate missing data, clean noise, etc.	Data cleaning report
Construct data	Create derived or transformed attributes and generate new records.	Derived attributes
		Generated records
Integrate data	Combine data coming from multiple sources.	Merged data
Format data	Syntactic modifications for modeling, like randomize order of records or remove all punctuation.	Reformatted data
-	-	Dataset
		Dataset description

Table 4 Modeling tasks and outputs, adapted from [50].

Modeling		
Generic Task	**Task Notes**	**Output**
Select modeling technique	Like neural network generation with back propagation.	Modeling technique
		Modeling assumptions (no missing values allowed, only uniform distributions allowed, etc.)
Generate test design	Plan the train, test, and evaluation of the models.	Test design (number of iterations and folds, etc.)
Build model	List the parameters chosen and train the model or models.	Parameter settings
		Models
		Model descriptions
Assess model	Accuracy, comparison between models, used parameters, etc.	Model assessment
		Revised parameter settings

Table 5 Evaluation tasks and outputs, adapted from [50].

Evaluation		
Generic Task	**Task Notes**	**Output**
Evaluate results	Is the model suited for the business needs?	Assessment of DM results with respect to business success criteria
		Approved models
Review process	Check all the process for correctness, completeness, etc.	Review of process
Determine next steps	Deploy or iterate again? Create new DM projects?	List of possible actions
		Decision

Table 6 Deployment tasks and outputs, adapted from [50].

Deployment		
Generic Task	**Task Notes**	**Output**
Plan deployment	DM necessary steps and how to perform them.	Deployment plan
Plan monitoring and maintenance	To insure the correct usage of DM results.	Monitoring and maintenance plan
Produce final report	Summary of the project, its experiences and results.	Final report
		Final presentation
Review project	Points to improve and points of success, acquired knowledge and generalizations that can be used in other projects.	Experience documentation

Appendix C - Data Mining and Knowledge Discovery Process Models or Methodologies

C.1 Related to KDD

This section presents the most relevant DM and KD process models or methodologies that are related to KDD, as depicted in figure 1.

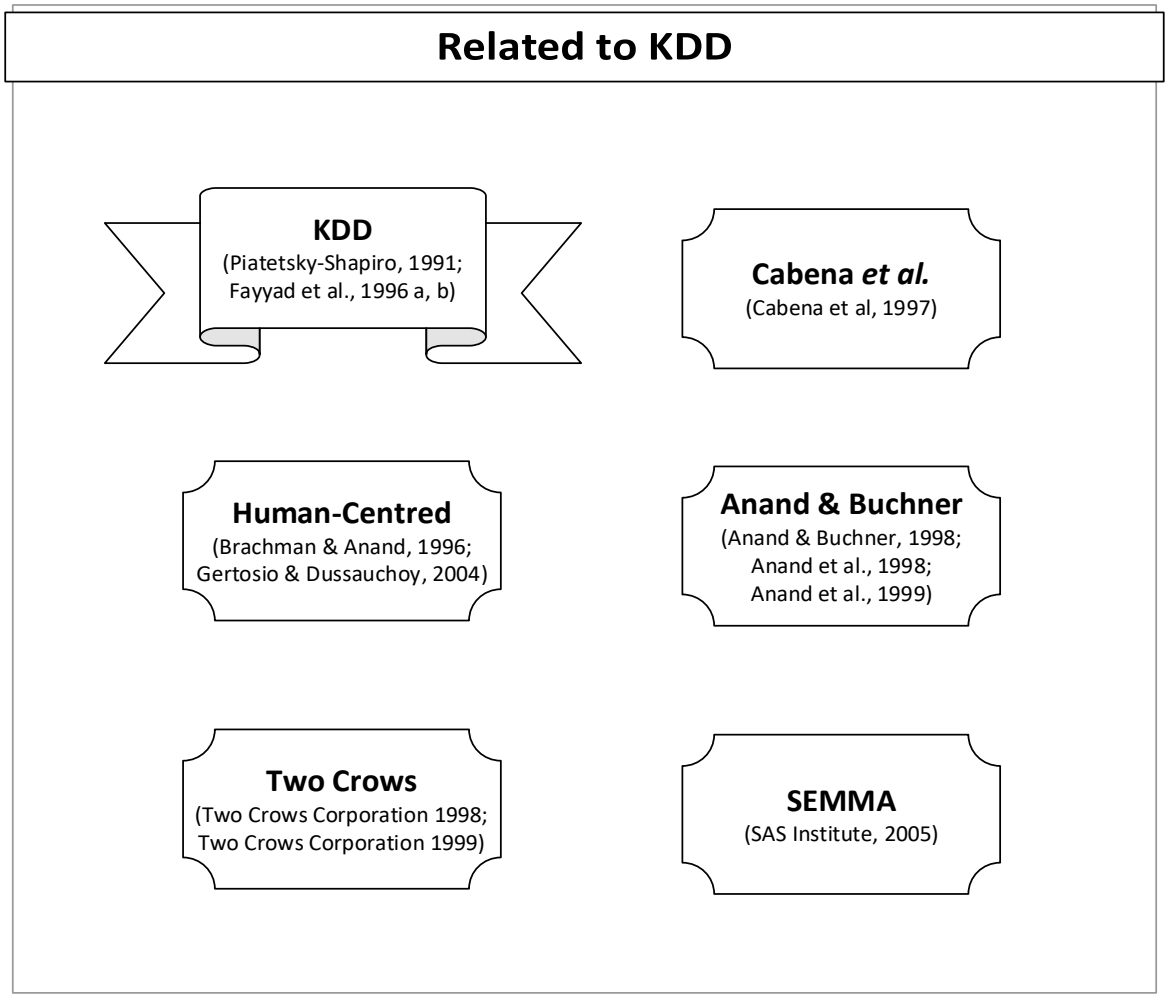

Figure 1 DM and KD process models or methodologies related to KDD and the papers' authors and years.

Although the figure 2 doesn't include it, Cabena *et al*. DM process starts with business objectives determination, ending with assimilation of knowledge [25]. It is used in the marketing and sales domain [48].

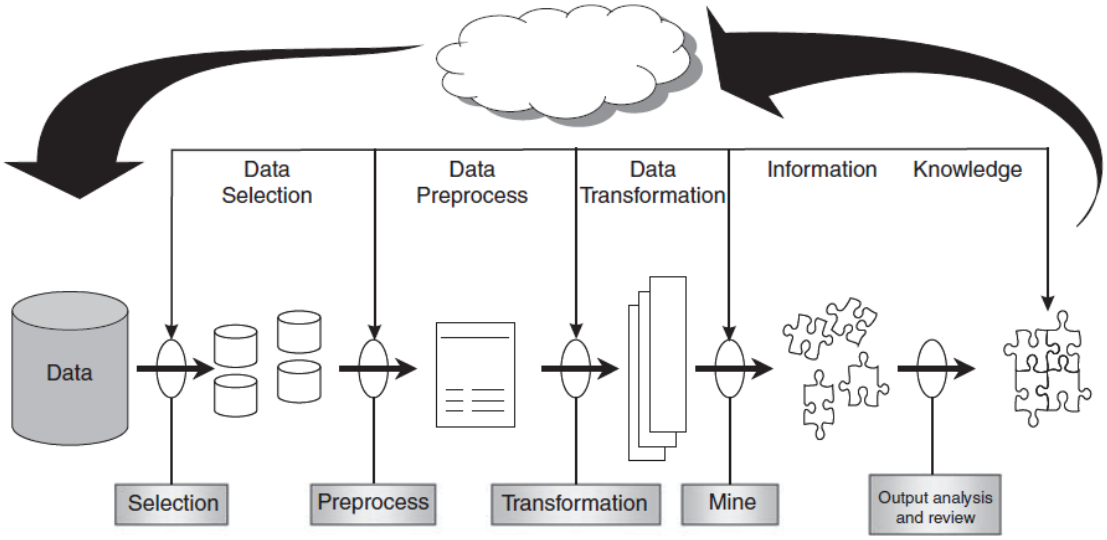

Figure 2 DM process according to Cabena *et al.* [25].

The human-centred process, in figure 3, [51] is a practical view of KDD [25]. Being very similar to KDD, it focuses more on the data miner and his decisions instead of focusing more on the data [25].

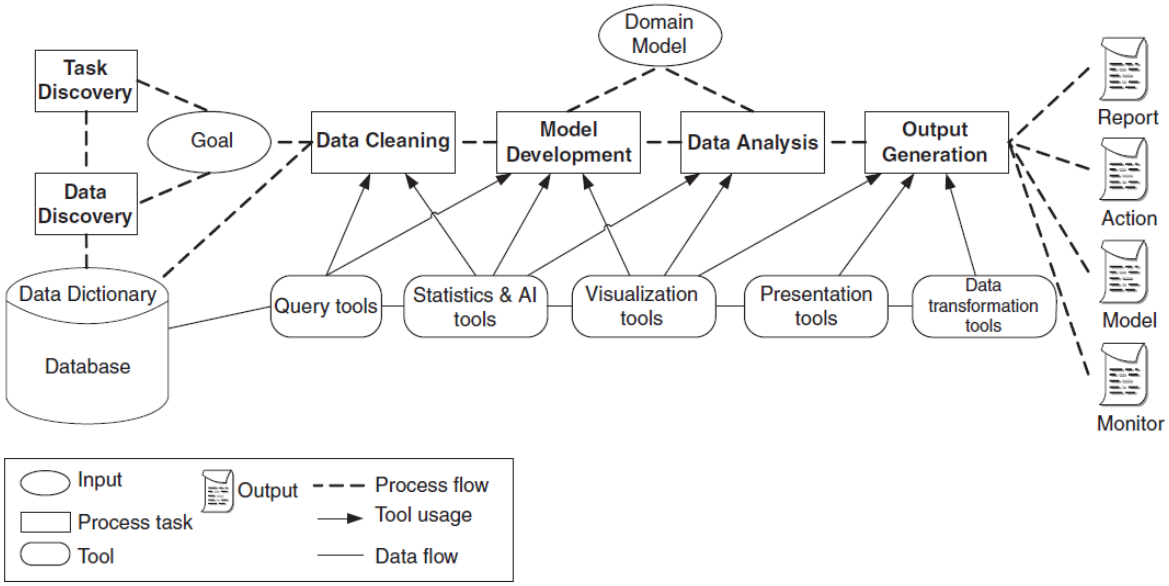

Figure 3 Human-centred process [25].

Anand and Buchner's process model, in figure 4, [52] is more oriented to web mining projects [25].

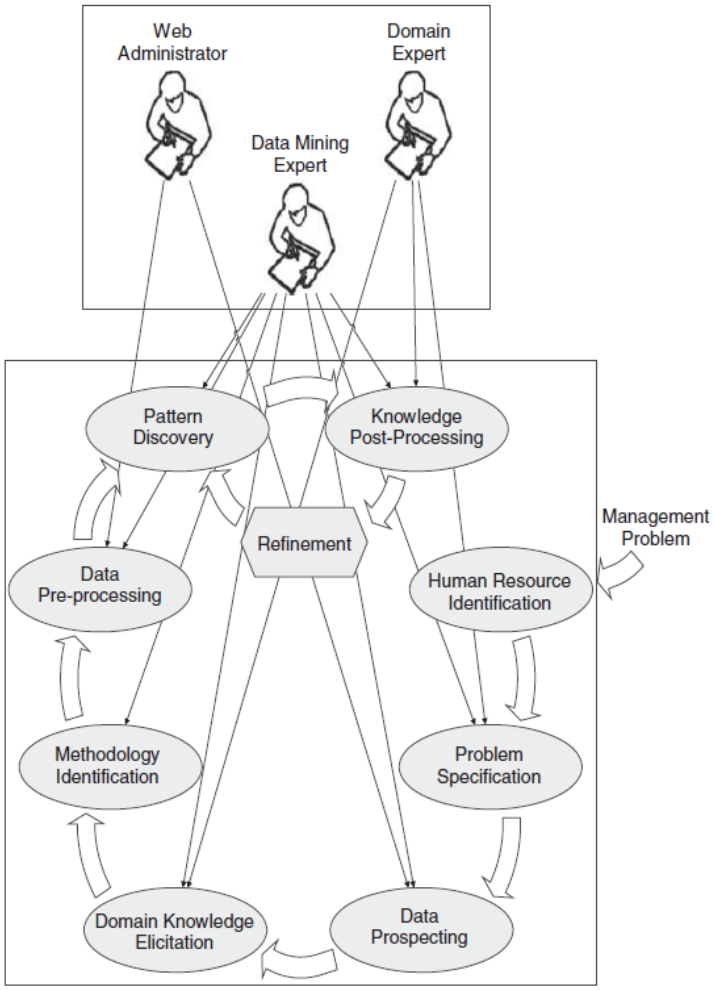

Figure 4 Anand and Buchner process model [25].

Figure 5 shows the Two Crows DM process model.

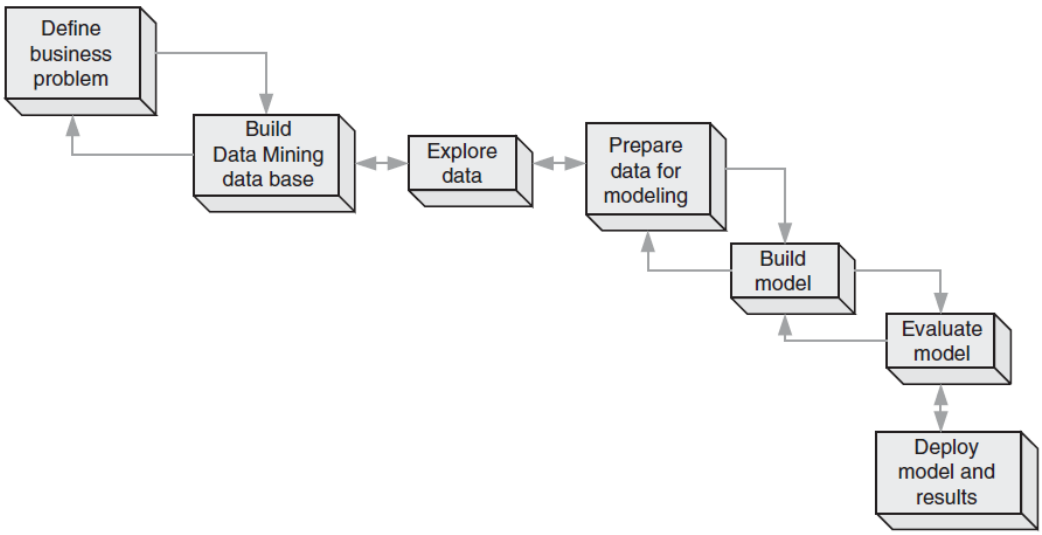

Figure 5 Two Crows DM process model [25].

SEMMA – sample, explore, modify, model, assess, in figure 6, [53] is the methodology that Enterprise Miner (a tool developed by SAS, a leader company in BI) is based on [25]. SEMMA, compared with KDD, doesn't include the first step, learning the application domain, and using the discovered knowledge [25].

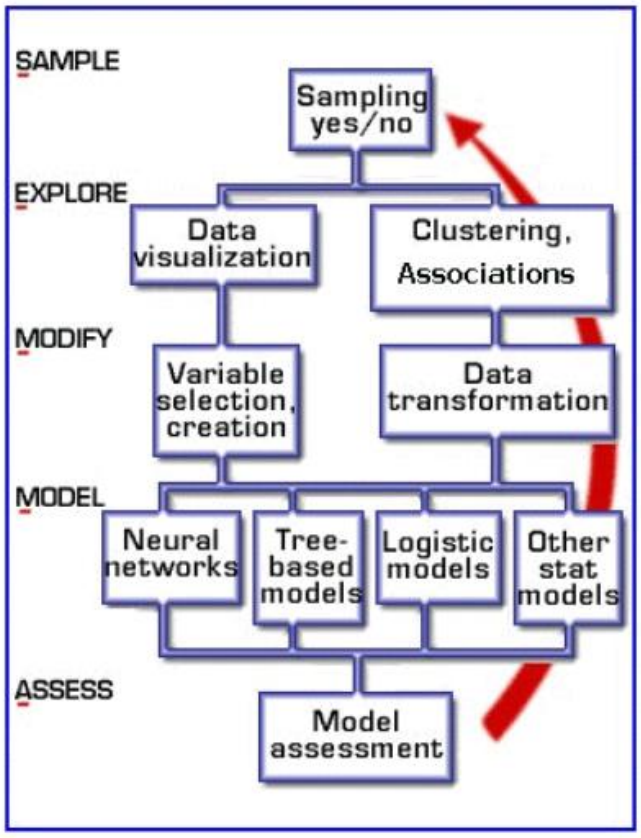

Figure 6 SEMMA methodology steps [53].

C.2 Related to CRISP-DM

This section presents the most relevant DM and KD process models or methodologies that are related to CRISP-DM, as portrayed in figure 7.

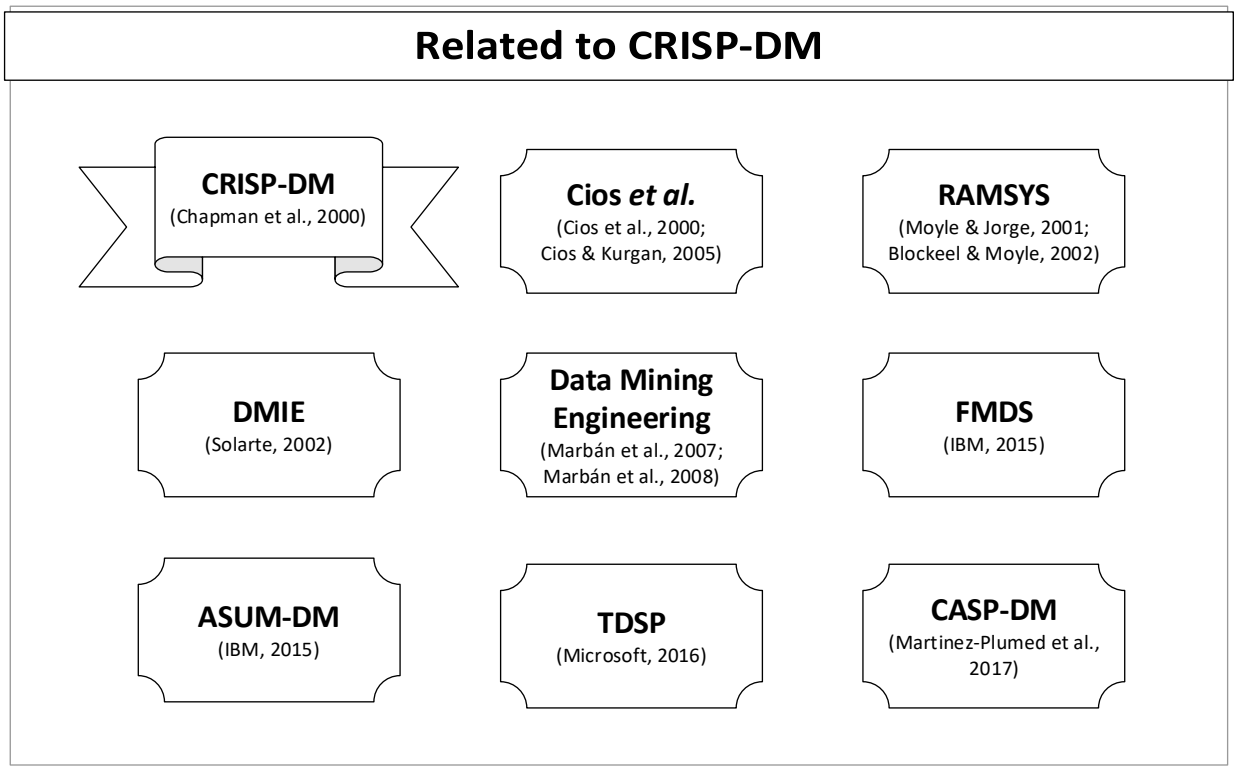

Figure 7 DM and KD process models or methodologies related to CRISP-DM and the papers' authors and years.

The Cios *et al*. process model, in figure 8, [54] is an adaptation of CRISP-DM to the needs of the academic research community [25]. It affirms that the knowledge discovered for a particular domain may be applied in other domains [25].

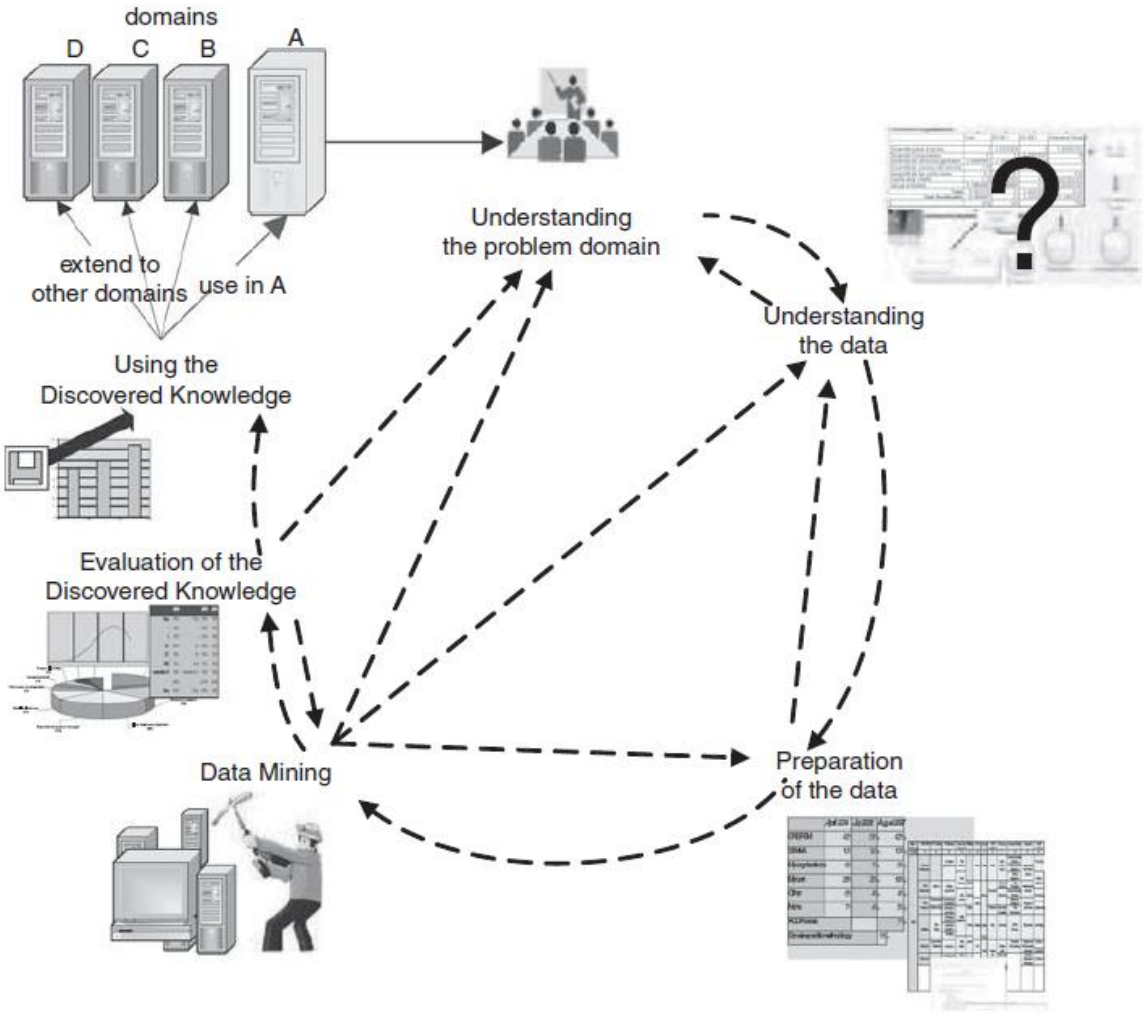

Figure 8 Cios *et al*. process model [25].

The Rapid Collaborative Data Mining System (RAMSYS), in figure 9, [55] methodology is meant to be used by groups who are working remotely, but work together on the same DM problem in a collaborative way [25]. Its principles are constant and updated knowledge sharing and communication, light management, enabling of member rotation and information security [25]. The different groups access the information vault to retrieve and share information [25]. RAMSYS introduces a new task to CRISP-DM, Model Submission, where the current best models from each of the remote teams are selected, evaluated and delivered [25].

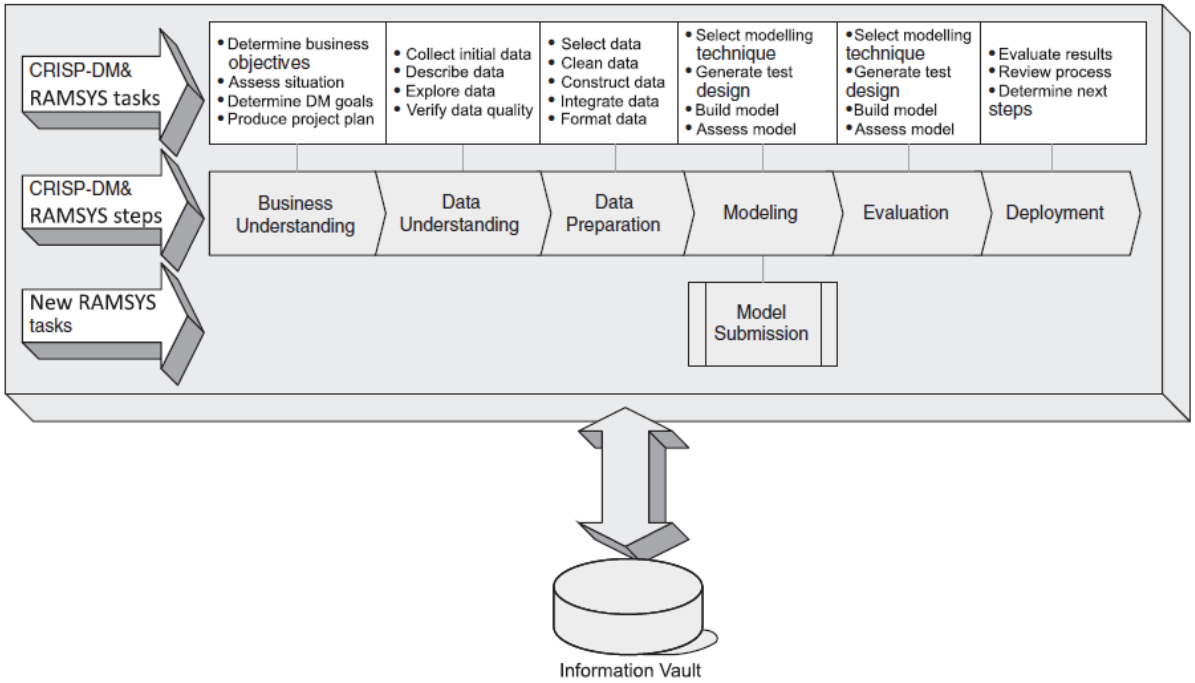

Figure 9 Rapid Collaborative Data Mining System (RAMSYS) methodology [25].

The DM for industrial engineering (DMIE) methodology, in figure 10, [56] is oriented to the industrial engineering domain [25]. It adds a last phase that focuses on support and maintenance, involving data backups, data maintenance, DM model updates and software updates when needed [25].

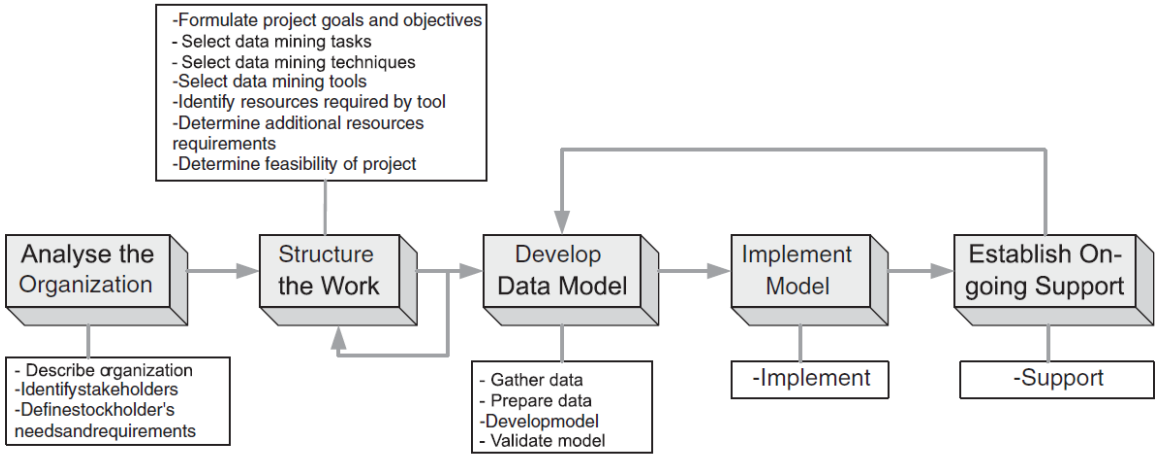

Figure 10 Data Mining Process for Industrial Engineering (DMIE) [25].

The process model for data mining engineering, in figure 11, [57] took inspiration in the software engineering field, that has over 40 years of experience [25]. It adds to CRISP-DM tasks and activities required in an engineering process [25]. Besides CRISP-DM, it is based on KDD and on two software engineering standard process models: IEEE 1074 and ISO 12207 [25].

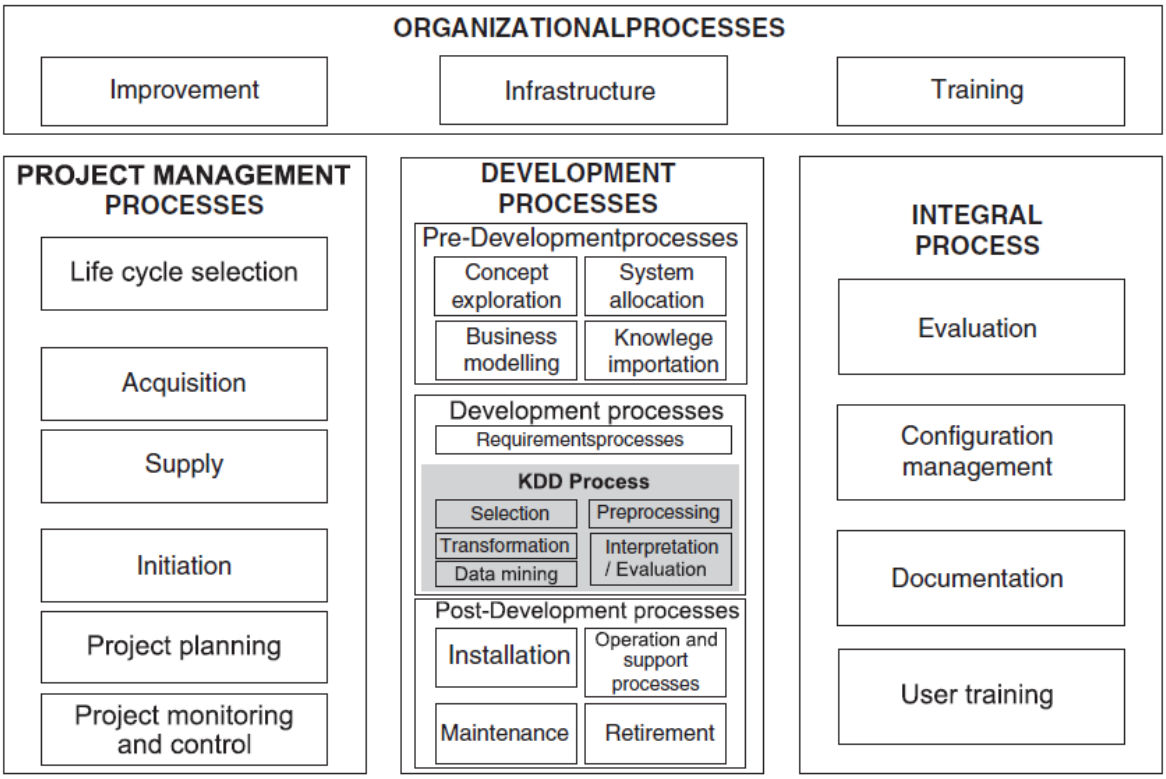

Figure 11 Process Model for Data Mining Engineering [25].

The Foundational Methodology for Data Science (FMDS), in figure 12, is an iterative process model designed by IBM that emphasises a number of the new practices such as the use of very large data volumes, the incorporation of text analytics into predictive modelling and the automation of some of the processes [58].

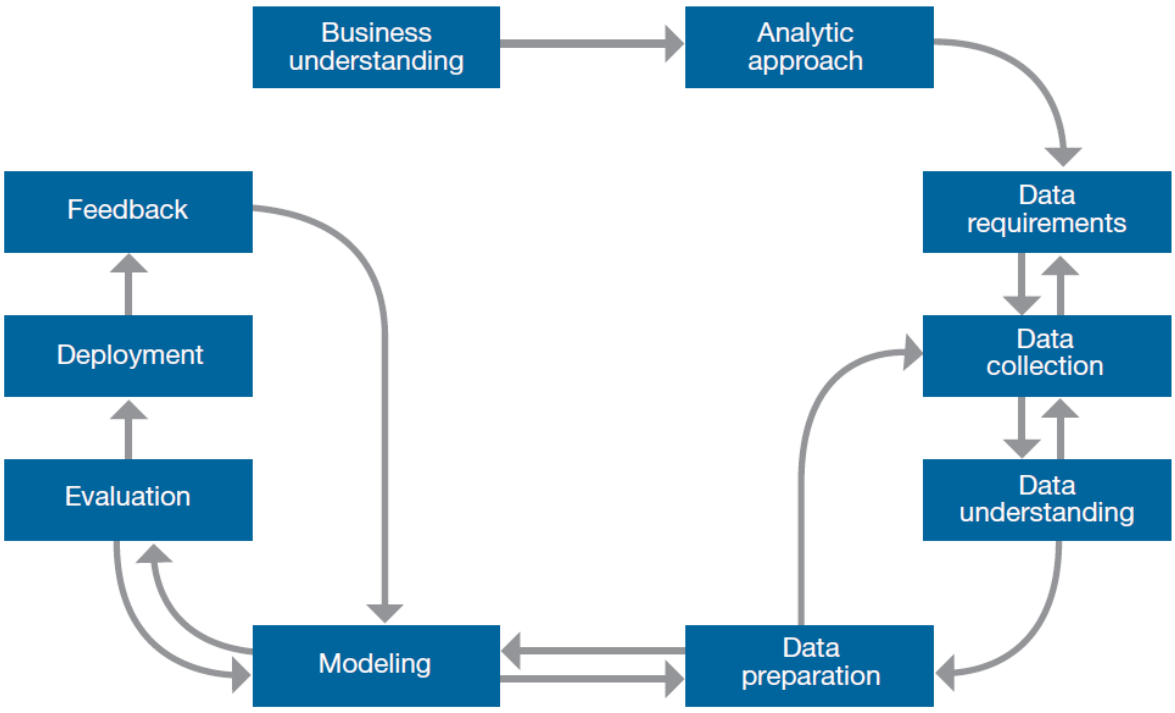

Figure 12 Foundational Methodology for Data Science (FMDS) [58].

Analytics Solutions Unified Method for Data Mining (ASUM-DM), in figure 13, is a methodology which refines and extends CRISP-DM with project management, infrastructure and operations, deployment and operating and optimization processes [59]. It provides templates and guidelines and was designed by IBM to enable a successful implementation of their Analytics solutions by their clients [60].

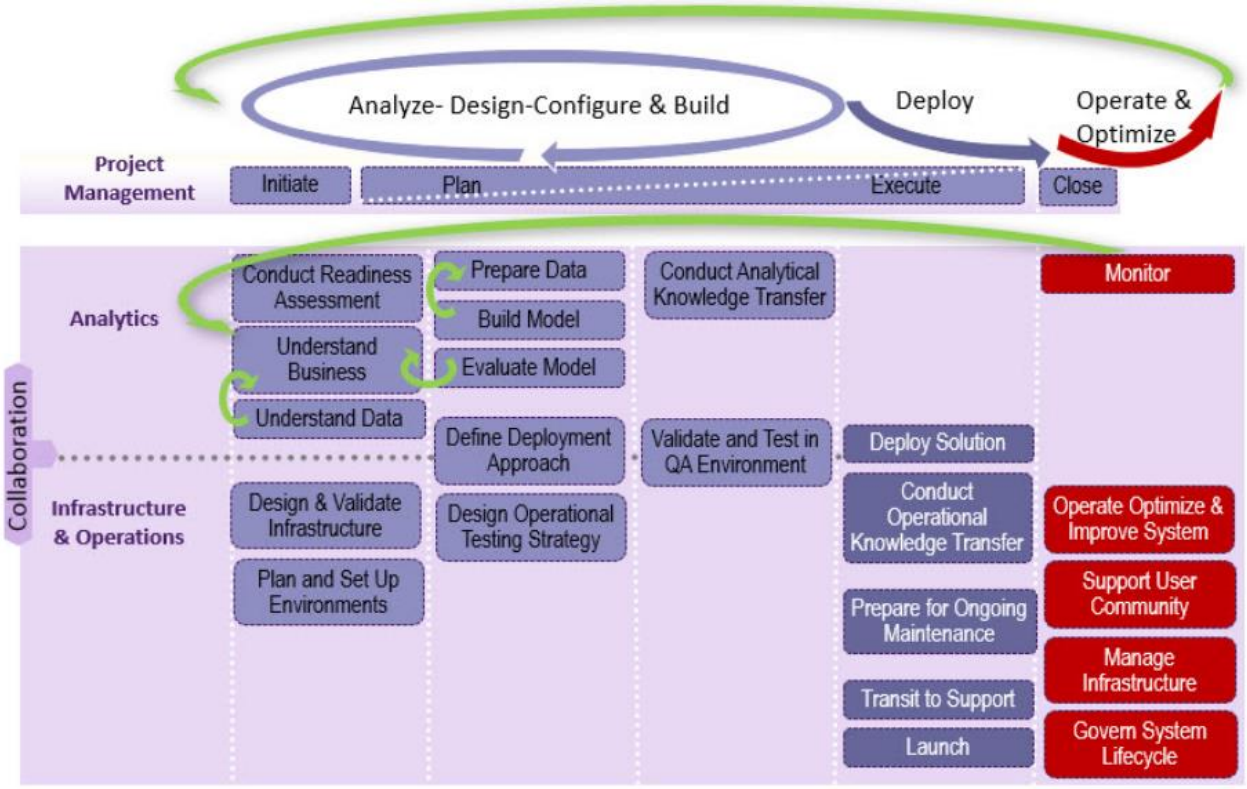

Figure 13 Analytics Solutions Unified Method for Data Mining (ASUM-DM) [59].

The Team Data Science Process (TDSP), in figure 14, was designed by Microsoft and it "is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. TDSP helps improve team collaboration and learning by suggesting how team roles work best together" [61].

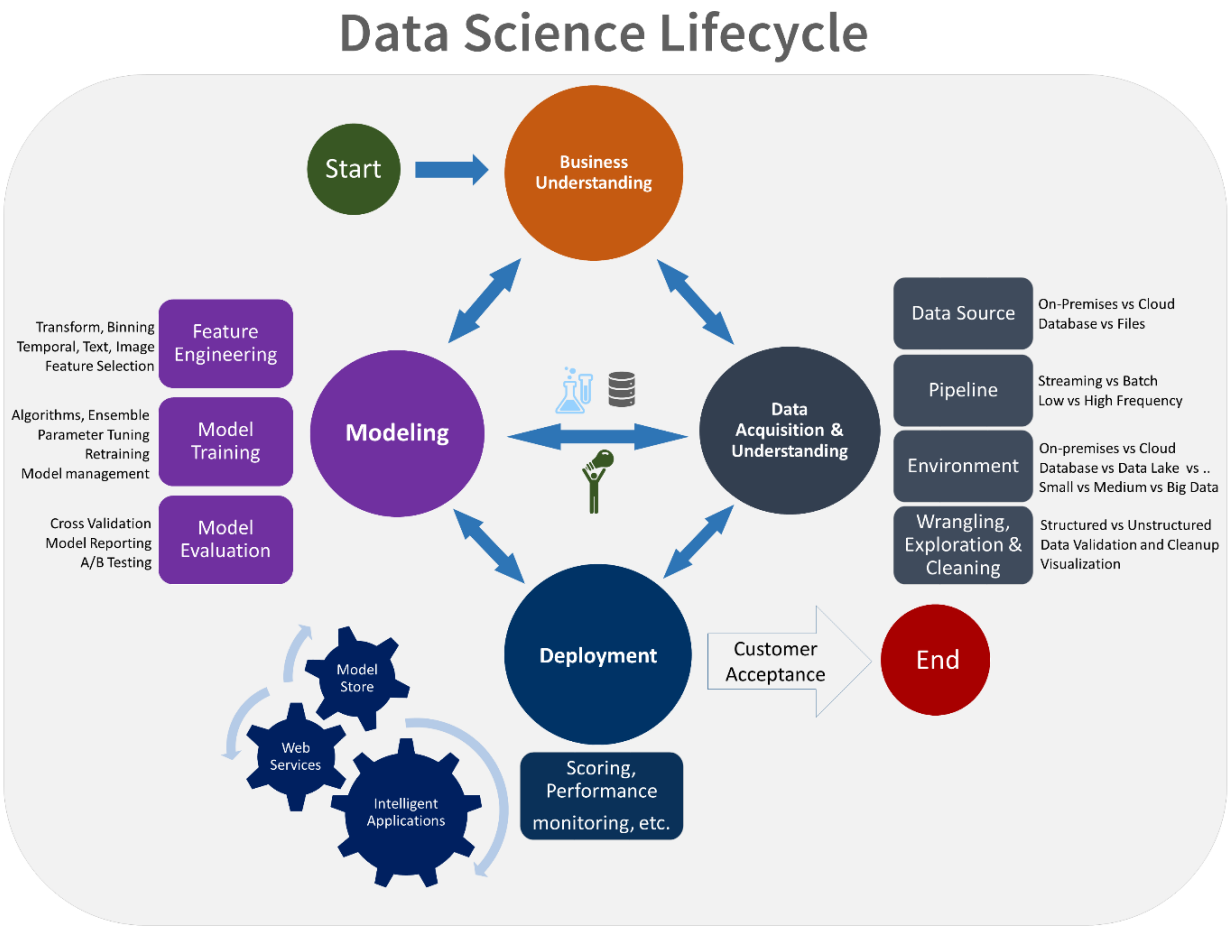

Figure 14 Team Data Science Process (TDSP) lifecycle [61].

Context-Aware Standard Process for DM (CASP-DM) proposes new activities and outputs and enhance some CRISP-DM outputs, focusing on context awareness and change and allowing model reuse [62]. The new proposed activities are for CRISP-DM's modeling phase and are [62]:

- Reframe Setting: If the model is versatile, this activity allows the model to be used in other contexts.
- Revise Model: If the model is not versatile but it is revisable, this activity allows the model to be used in other contexts.

More information about the new or enhanced CRISP-DM outputs can be found in [62].

C.3 Other approaches

This section presents the most relevant DM and KD process models or methodologies that are not related to KDD or CRISP-DM, as depicted in figure 15.

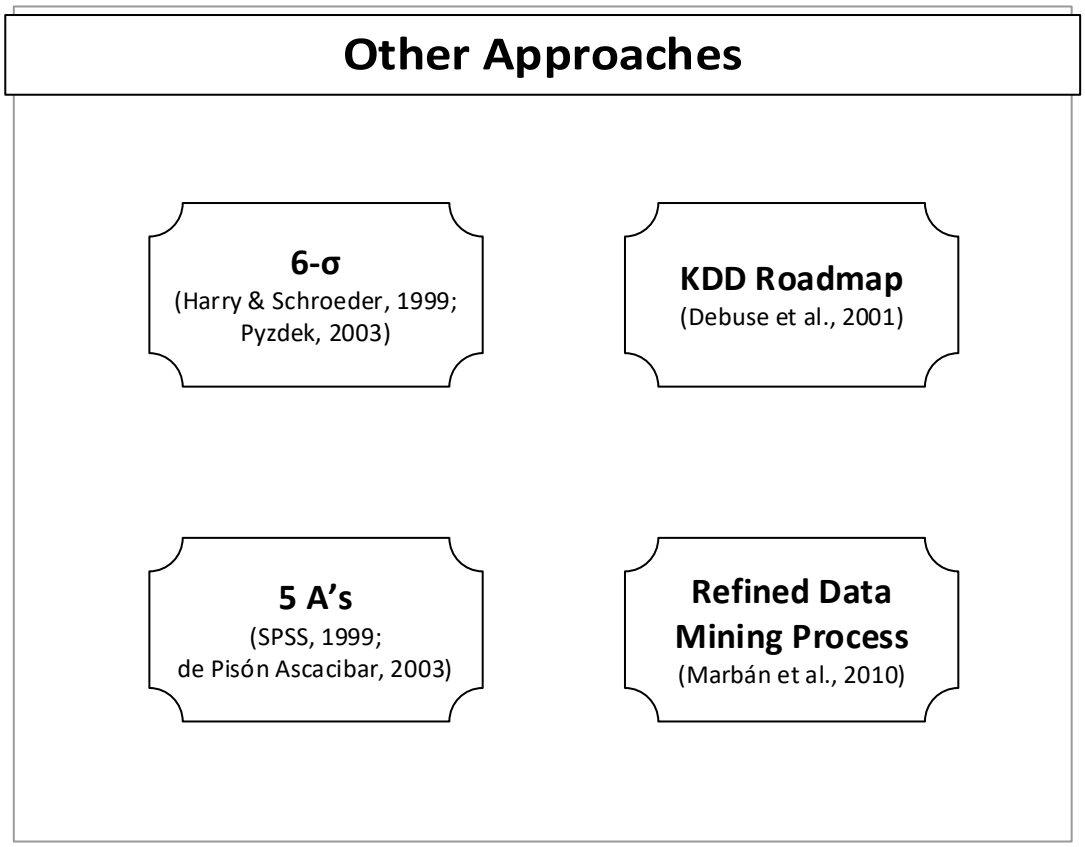

Figure 15 DM and KD process models or methodologies not related to KDD or CRISP-DM and the papers' authors and years.

The 6-σ [63], in figure 16, created in 1999 by Motorola, "is a paradigm for quality and excellence in management. In other words, it defines how to improve quality and customer's satisfaction and, at the same time, reduce production costs. (…) The 6-s method has also been applied in DM projects" [25].

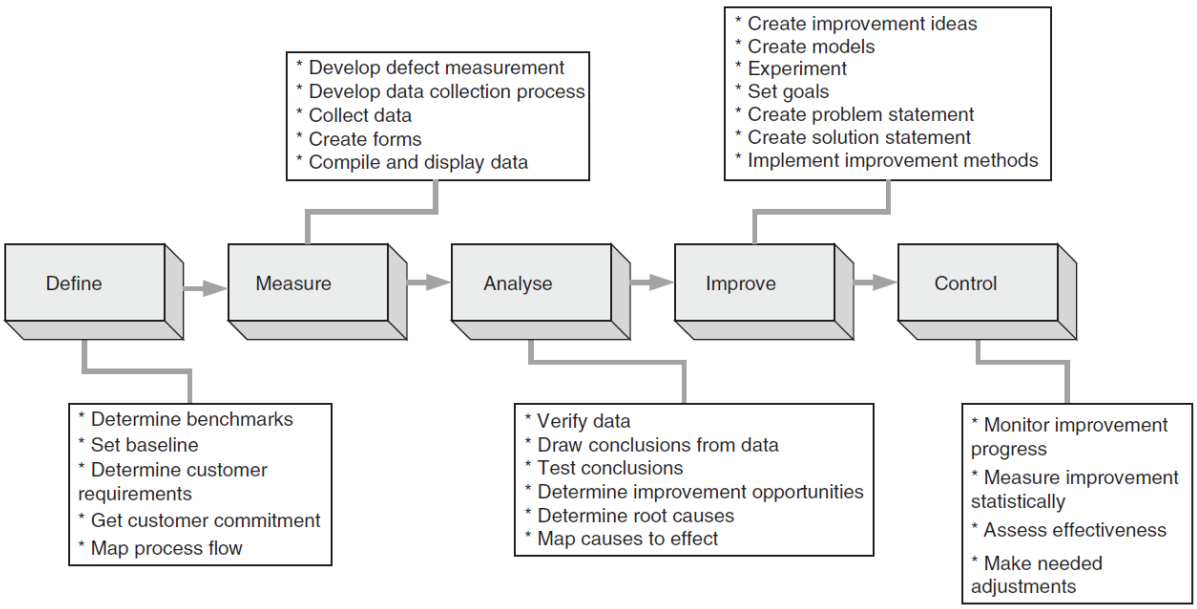

Figure 16 6-σ paradigm [25].

KDD Roadmap, in figure 17, [64] is an iterative methodology that contributed with the resourcing task [25]; it consists in getting all resources needed for the project, e.g. algorithms and data [64].

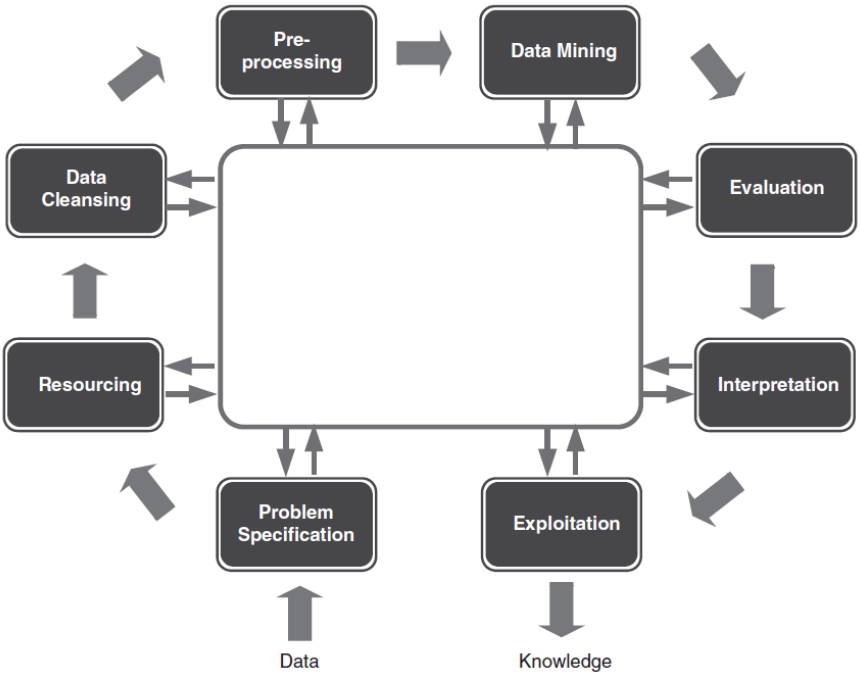

Figure 17 KDD Roadmap [25].

The five A's methodology, in figure 18, [65] doesn't describe how to do DM, but presents the automate step that highlights the possibility of non-expert users test the trained models with new data [25].

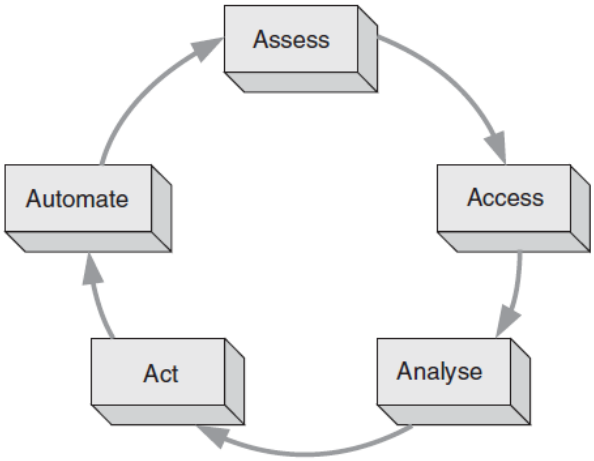

Figure 18 5 A' methodology phases [25].

[25] proposes a theoretical, iterative Refined Data Mining Process, based on its comparation of previous and relevant DM and KD process models or methodologies (see tables from 7 to 9). It is composed of 3 processes - analysis, development and maintenance - and 17 subprocesses, presented below along with adapted descriptions. It is an effort to combine all of the tasks and phases of other DM methodologies and process models [25]. [25] asserts that, by including more specific phases than other approaches, the Refined Data Mining Process will be easier to follow and facilitates the identification of dependencies between the subprocesses.

Table 7 Refined Data Mining Analysis Process, adapted from [25].

Analysis	
Subprocess	**Description**
Life cycle selection	Acquisition, supply and life cycle selection.
Domain knowledge elicitation	Understand data and the problem.
Human resource identification	Identify needed human resources.
Problem specification	Establish concrete DM objectives; select tasks, techniques and tools.
Data prospecting	Collect data.
Data cleaning	Search for and remove errors, sample data, deal with outliers, missing and unreliable values and possibly balance data.

Table 8 Refined Data Mining Development Process, adapted from [25].

Development	
Subprocess	**Description**
Preprocessing	Preprocess data.
Data reduction and projection	Find useful features to represent the data, use dimensionality reduction or transformation methods to reduce the number of variables under consideration or to find invariant representations for the data.

Choosing the DM function	Decide the purpose of the model derived by the DM algorithm and decide DM techniques to apply (e.g., summarization, classification, regression and clustering).
Choosing the DM algorithm	Decide which models and parameters may be appropriate.
Build model	Train DM models.
Improve model	Review the model and try to improve it.
Evaluation	Test new data.
Interpretation	Interpret discovered knowledge.
Deployment	Apply the discovered knowledge / make use of the created models.
Automate	Facilitate to non-experts the training of the models and using the discovered knowledge.

Table 9 Refined Data Mining Maintenance Process, adapted from [25].

Maintenance	
Subprocess	**Description**
Establish on-going support	Retrain models because of new data, update software, etc.

Appendix D - Data Mining and Knowledge Discovery Methodologies: Exclusively Bottom Up or Exploratory Approaches

This section briefly presents some methodologies that are said by their authors to be flexible to be used in projects that have an exploratory nature, that is, projects that don't have a set goal from the start (see figure 19). These approaches are not so relevant as the previous ones for the purpose of this paper, still they are included for the reader's reference, since the authors provide some interesting contributions. Namely, [26] lists many issues of DM studies that were important for this paper.

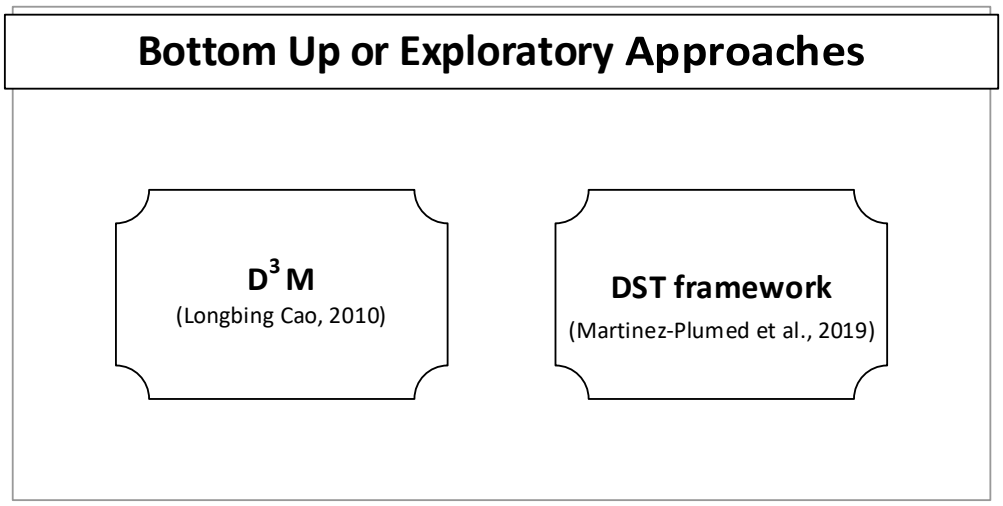

Figure 19 DM and KD methodologies for bottom up or exploratory approaches and the papers' authors and years.

The domain-driven DM (D^3 M) methodology promotes the paradigm shift from "data-centred knowledge discovery" to "domain-driven, actionable knowledge delivery" (see figure 20) [26]. The "actionability" of delivered knowledge "measures the ability of a pattern to prompt a user to take concrete actions to his/her advantage in the real world" [26].

"Targeting real-world problem solving, knowledge discovery is further expected to migrate into actionable knowledge discovery and delivery (AKD). AKD aims to deliver knowledge that is business friendly, and which can be taken over by business people for seamless decision making. (…) AKD is critical in promoting and releasing the productivity of KDD for smart information extraction, business operations, and decision making. (…) D^3 M emphasizes the development of methodologies, techniques, and tools for actionable knowledge discovery and delivery by incorporating relevantly ubiquitous intelligence surrounding data-mining-based problem solving. Ubiquitous intelligence consists of in-depth data intelligence, human intelligence, domain intelligence, network intelligence, and organizational/social intelligence. It is essential to synthesize such ubiquitous intelligence in actionable knowledge discovery and delivery." [26].

"The main task of D^3 M is to develop AKD-oriented problem-solving systems. AKD-oriented D^3 M, on top of the data-driven framework, aims to complement the shortcomings of traditional data mining, through developing proper methodologies and techniques to incorporate domain knowledge, user needs, the human role and interaction, as well as actionability measures into KDD process and systems. It is data and domain intelligence working together to disclose a hidden story to business, and to satisfy real user needs. End users hold the final decision in evaluating the findings and business deliverables" [26].

Aspects	Data-Driven	Domain-Driven
Rationale	Data tells the story	Data and ubiquitous intelligence disclose problem-solving solutions
Objective	Innovative and effective algorithms	Effective problem-solving
Data	Abstract, synthetic and refined data	Real-life data and surrounding information
Process	One-off	Multiple-step, iterative and interactive on demand
Mechanism	Automated	Human-centered or human-mining-cooperated
Infrastructure	Closed pattern mining systems	Closed-loop problem-solving systems in open environment
Usability	Predefined models and processes	Ad-hoc, dynamic and customizable models and processes
Deliverable	Patterns	Business-friendly decision-support actions
Deployment	Solid validation	Well-founded artwork in problem-solving
Evaluation	Technical metrics	Tradeoff between technical significance and business expectation

Figure 20 Data-driven DM and domain-driven DM comparison [26].

The Data Science Trajectories (DST) framework, in figure 21, expands CRISP-DM (which comprises the goal-oriented activities) by including exploratory activities and data management activities [48]. It doesn't include arrows between the activities to emphasize that the order depends on the project and some activities are not performed in all projects: each project has its trajectory [48]. [48]'s authors argue that DM is goal-oriented and concentrated on the process while DS is data-oriented and exploratory.

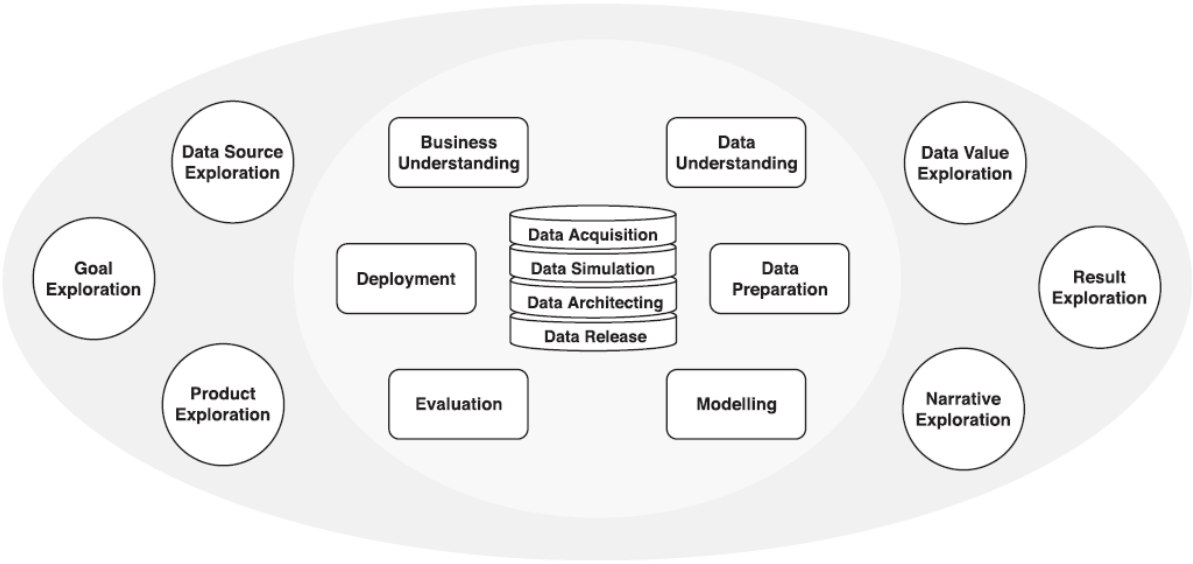

Figure 21 Data Science Trajectories (DST) framework [48].

Appendix E - Data Preparation

This section presents a general overview of the most common data preparation processes; as it is not the main purpose of this study to exhaustively explore this topic (more in [89]).

Data preparation's importance in DM is paramount. Pyle argues that "data preparation consumes 60 to 90% of the time needed to mine data – and contributes 75 to 90% to the mining project's success" [17].

Data must be prepared before serving as input to an algorithm [89, p. 11]. Otherwise, results might not be considered accurate knowledge or there can be errors during runtime that prevent obtaining results [89, p. 11].

[89, p. 11] considers the following processes in data preparation (see figure 22):

- Data Cleaning: e.g., correct bad data, remove incorrect data out of the data set and reduce the unnecessary detail of data.
- Data Transformation: e.g., create dummies [89, p. 54] and record summarization.
- Data Integration: e.g., identification and unification of variables and domains, the analysis of attribute correlation, the duplication of tuples and the detection of conflicts in data values of different sources.
- Data Normalization: normalizing the data attempts to give all attributes equal weight and it is particularly useful in statistical learning methods.
- Missing Data Imputation: with a reasonable estimate of a suitable data value.
- Noise Identification: detect random errors or variances in a measured variable.

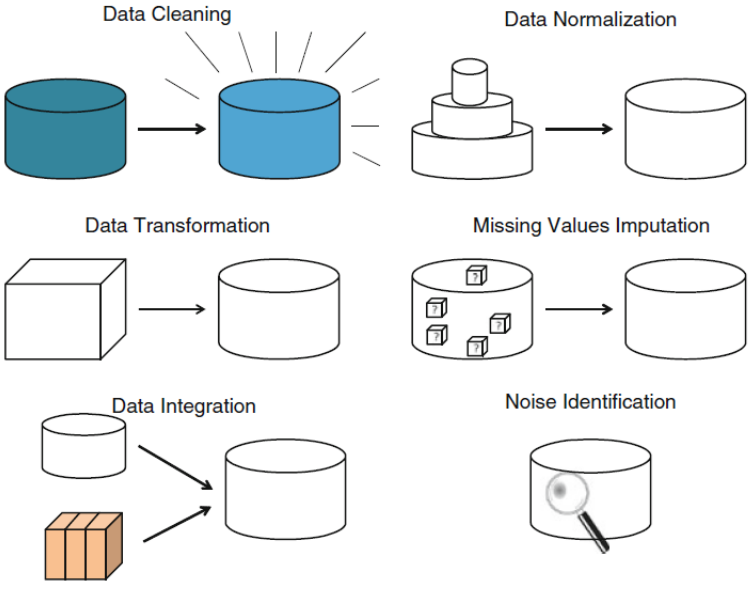

Figure 22 Forms of data preparation [89, p. 12].

Another set of techniques is data reduction (see figure 23). It consists in downsizing the amount of data used while maintaining its essential characteristics [89, p. 13]. It's not a mandatory step, except when the data size exceeds the limit accepted by certain algorithm [89, p. 13]. Forms of data reduction [89, p. 15]:

- Feature Selection: removing irrelevant or redundant features (or dimensions).
- Instance Selection: choosing a subset of the total available data.
- Discretization: transform numerical attributes into discrete attributes with a finite number of intervals, obtaining a non-overlapping partition of a continuous domain; an association between each interval with a numerical discrete value is then established.

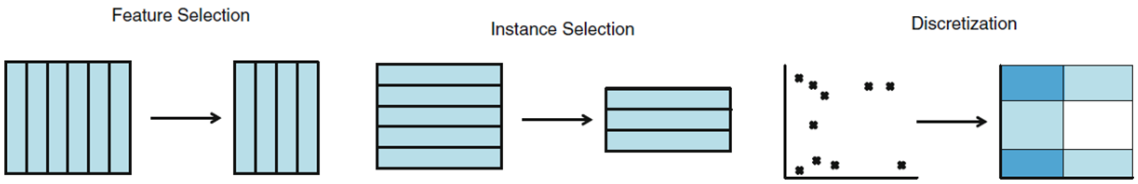

Figure 23 Forms of data reduction [89, p. 14].

Data can be reduced, but it can also be expanded with [89, p. 16]:

- Feature Extraction: create new attributes from existing ones, with a mapping, for example.
- Instance Generation: create new instances from existing ones, using neighbour records, for example [89, p. 221].