applied sciences

MDPI

Article

Development of a Framework to Aid the Transition from Reactive to Proactive Maintenance Approaches to Enable Energy Reduction

Michael Ahern [1,2], Dominic T. J. O'Sullivan [1,2] and Ken Bruton [1,2,*]

[1] Intelligent Efficiency Research Group (IERG), Department of Civil and Environmental Engineering, University College Cork, T12 CY82 Cork, Ireland; 119227513@umail.ucc.ie (M.A.); dominic.osullivan@ucc.ie (D.T.J.O.)
[2] MaREI Centre, Environmental Research Institute, University College Cork, T12 CY82 Cork, Ireland
* Correspondence: ken.bruton@ucc.ie

Abstract: The disparity between public datasets and real industrial datasets is limiting the practical application of advanced data analysis. Therefore, industry is stuck in a reactive mode regarding their maintenance strategy and cannot transition to cost-effective and energy-efficient proactive maintenance approaches. In this paper, an integration-type adaptation of the CRISP-DM data mining process model is proposed to combine domain expertise with data science techniques to address the pervasive data issues in industrial datasets. The development of the Industrial Data Analysis Improvement Cycle (IDAIC) framework led to the novel repurposing of knowledge-based fault detection and diagnosis (FDD) techniques for data quality assessment. Through interdisciplinary collaboration, the proposed framework facilitates a transition from reactive to proactive problem solving by firstly resolving known faults and data issues using domain expertise, and secondly exploring unknown or novel faults using data analysis.

Keywords: data analytics; data mining; fault detection and diagnostics; industrial AI; data quality; building AFDD

Citation: Ahern, M.; O'Sullivan, D.T.J.; Bruton, K. Development of a Framework to Aid the Transition from Reactive to Proactive Maintenance Approaches to Enable Energy Reduction. *Appl. Sci.* **2022**, *12*, 6704. https://doi.org/10.3390/app12136704

Academic Editor: Jason K. Levy

Received: 8 June 2022
Accepted: 27 June 2022
Published: 1 July 2022

1. Introduction

The industrial sector must reduce its energy demand as it is responsible for almost 40% of global final energy consumption [1], yet considerable increases are anticipated. The US Energy Information Administration (EIA) project that the industrial sector's energy use will grow nearly twice as fast as any other end-use sector between 2021 and 2050 [2]. The industrial sector therefore faces a significant challenge to attain ambitious sustainability targets, such as doubling the global rate of energy efficiency improvement [3]. In a competitive global market, ensuring sustainability targets are met, while simultaneously driving growth, is a key significant challenge. In Europe, the European Union (EU) outlined the need to protect, conserve and enhance the EU's natural capital, transition to climate neutrality, while also reducing waste in the European Green Deal [4]. The EU recognized the severity of the sustainability challenges and the need for a strategic change, hence the "Twin Transition" was proposed [5]. To ensure Europe achieves sustainability goals to become greener, but also increase competitive advantage, the EU proposed digitalization as a key enabler for both goals. Information and Communications Technology can enable a 20% reduction of global CO_2 emissions by 2030, which could provide the necessary environmental protection while avoiding a trade-off with economic prosperity [6]. The International Renewable Energy Agency (IRENA) estimates that if manufacturing companies adopted the best available technologies, then the consumption globally would reduce by about a quarter [7].

The industrial sector is comprised of many different subsectors such as manufacturing, pharmaceuticals, food, construction and oil and gas. While the constituent assets and processes that account for the energy consumption in these subsectors will differ, indoor facilities have a common, energy intensive requirement to provide ventilated working environments. Heating, Ventilation and Air Conditioning (HVAC) systems, which provide fresh air as well as heating and cooling needs, consume about 50% of building energy consumption on average [8]. In an industrial facility, this figure reduces to about 40% on average [9], while 34% of hi-tech facilities managers say HVAC is their site's biggest energy cost [10]. There is high potential for energy savings in HVAC systems as they are self-correcting in nature, which facilitates the occurrence of unnoticed faults that may account for up to 20% of the energy consumed by these systems [11]. Monetarily, Mills [12] estimated that common faults such as duct leakage, dampers not working properly and valve leakage cost 2.9, 0.5 and 0.1 $billion/year, respectively, in commercial buildings, based on the findings by Roth et al. [11]. HVAC systems therefore need to be well maintained, but maintenance costs, time for repair, and replacement of components that have not reached the end of their useful life must be balanced with the potential energy savings. As noted by O'Donovan et al. [13], "equipment maintenance can exceed 30% of total operating costs, or between 60 and 75% of equipment lifecycle cost [14]". Therefore, selecting an appropriate maintenance strategy is critical to be cost effective. Traditional maintenance practices are reactive in nature with a "fail and fix" philosophy [15]. These approaches are known as reactive or corrective maintenance whereby the main strength is reduced maintenance costs [16]. However, the increased time interval between maintenance activities could lead to hidden costs such as suboptimal operation, while also increasing the likelihood of costly downtime [16]. An alternative approach would be an evolution from reactive maintenance to proactive "predict-and-prevent" maintenance [15]. Proactive maintenance would encompass strategies such as condition-based maintenance, whereby measurements are monitored to determine if an issue is about to occur or already has occurred [13], as opposed to predictive maintenance, whereby the optimal time for maintenance is determined by estimating the remaining useful life of machine components [13]. These smarter approaches have the capability to not only increase the effectiveness of maintenance activities, but also ensure the system remains operating in an energy efficient manner. An estimated 10–40% of HVAC energy use could be saved by an appropriate maintenance strategy [17], although maintenance costs must also be incorporated. Therefore, proactive approaches require information about the operation of the machines in the form of data to enable more informed decision making regarding maintenance activities. Therefore, a successful pervasive digital transition is a key to enabling highly scalable proactive maintenance that will deliver the savings needed to achieve policy ambitions in terms of energy efficiency.

2. Background and Related Research

According to a survey of hi-tech manufacturers in the UK, 40% are stuck in reactive mode for their HVAC maintenance strategy [10]. It appears this is caused by a lack of information on the operation of these systems, as 62% of facility managers in this survey admit they are deficient in relation to the collection and analysis of their HVAC data [10]. However, in the literature, a myriad of data-driven approaches have been developed and there is an increasing trend [18]. The disparity between academia and practical applications appears to be caused by a difference in the data. According to Zhao et al. [18], most studies in the literature utilize experimental or simulated data from laboratory tests such as ASHRAE RP-1312 [19] or those made available by the Lawrence Berkeley National Laboratory [20]. While these datasets address a severe lack of publicly available benchmark datasets, they are not representative of real-world datasets. In 2022, Huang et al. [21] found that the performance of FDD strategies developed on simulated data and directly applied to real building data is less that satisfactory. In practical applications, the granularity of the data is less sufficient, the metadata is not well described, and there is a lack of well-labelled faults with multiple severity ratings. Therefore, many of the studies in the literature do

not address the challenges that are faced in practical applications of data-driven analysis. To summarize the data issues identified in the literature, Table 1 categorizes the main data-related challenges outlined by systematic reviews of data-driven applications in both generic industrial and HVAC-specific studies. The table is by no means exhaustive but does provide an indication that many challenges are pervasive in both fields.

Table 1. Classification and comparison of the issues identified in systematic reviews in the area of advanced data analysis in the industrial domain.

Classification	Challenge	General Industrial				HVAC-Specific	
		Dogan and Birant [22]	**Bertolini et al. [23]**	**Wuest et al. [24]**	**Dalzochio et al. [25]**	**Mirnaghi and Haghighat [26]**	**Zhao et al. [18]**
Broken	Data Availability	X	X	X	X	X	
Bad Quality	Noisy Data	X				X	
Background	Algorithm Selection	X	X	X	X		
	Evaluation of Results	X	X	X	X		
	Distinguishing between sensor and component faults						X

To classify the challenges outlined in Table 1, the data issue classification proposed in the Industrial AI paradigm is adopted [27]. Professor Jay Lee [27] coined the term "Industrial AI" to describe the application of artificial intelligence in the industrial domain. Data issues are classified as broken, bad quality and background, and are collectively known as the "3B" [27]. Firstly, data may be considered broken if it does not fulfil the requirements of the analysis. Professor Lee [27] outlines that "data can't just be numerous—it must be comprehensive". As in Table 1, data availability is a common issue in the literature, whereby the data is described as missing, incomplete or insufficient. Data may be unavailable for reasons such as lack of sensors, faults in communication protocols or damaged sensor networks. In the case where the necessary information is available, it must accurately represent the physical phenomena it measures. Bad quality data does not reach the standard required to apply data-driven techniques and is often considered as noisy. Noisy data is composed of issues such as incorrect data, improper data, duplicated data and inconsistent data, which is often caused by sensor faults. The different fault types are visualized in Figure 1 and are commonly classified as drift fault (a), bias fault (b), precision degradation fault (c), spike fault (d) and stuck fault (e) [28]. The presence of a sensor fault may lead a data-driven algorithm to incorrectly classify the anomaly as a component fault, which would lead to a lack of trust in these methods by industrial practitioners. Distinguishing between sensor faults and component faults is a key challenge in HVAC applications [18], which requires an understanding of the operation of the equipment to overcome. This is a key component of the background issue that relates to the need for understanding the context of the data. In practical applications, the data is often inadequately labelled, which means the interpretation of domain experts is needed to clarify the meaning of each measurement [22]. Uncertainty with regards to the meaning behind the data leads to the prevalent challenges of selecting an appropriate algorithm and evaluating its results. If the meaning behind the data is not well understood, the value of data-driven analysis diminishes as the results cannot be adequately interpreted.

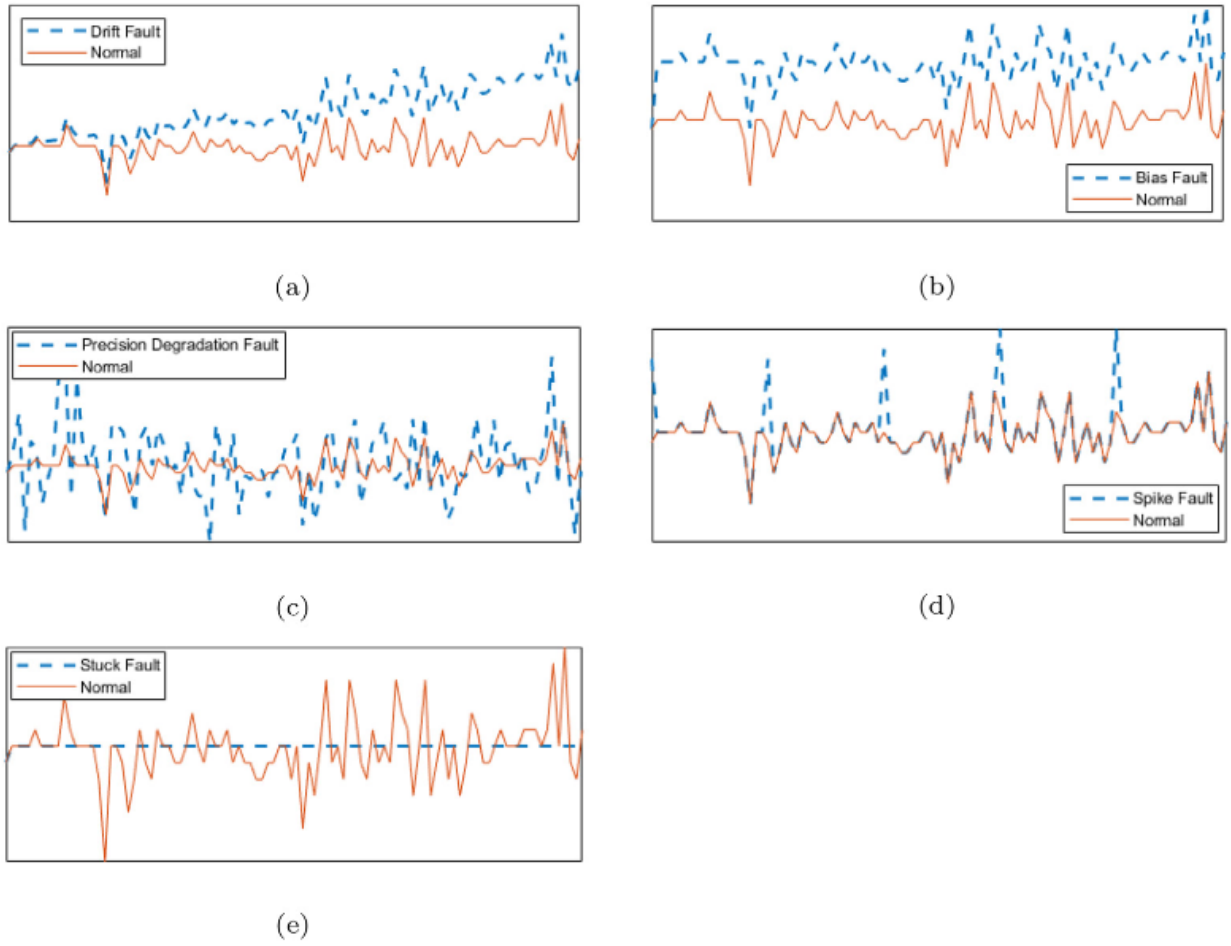


Figure 1. Illustrations of (**a**) drift, (**b**) bias, (**c**) precision degradation, (**d**) spike, and (**e**) stuck faults. Reprinted from Jan et al. [28], with permission from Elsevier, 2020.

The "3B" data issue taxonomy provides a classification for the plethora of data-related challenges in the industrial setting that extend beyond the challenges outlined in Table 1. We interpret the nature of the data issues as in Figure 2, such that data challenges may be classified in a number of regions in the Venn diagram. For example, the challenge of distinguishing sensor faults from component faults is a key challenge in HVAC applications [18], which may be caused by faulty sensors (Bad quality) or the physical failure of mechanical components (Background). Therefore, rather than prescriptively resolving a myriad of specific data challenges, a means of managing the challenges that lie within the Venn diagram is a potentially more sustainable solution. The inadequacy of real-world industrial data appears to be a main barrier to the implementation of data-driven predictive maintenance systems; therefore, managing the 3B's is a requirement for successful digital transition. In summary, we are of the opinion that energy reduction-related decision making in the HVAC domain would be aided if knowledge-based and data-driven methods were integrated. There are two main research gaps that are impeding the realization of the benefits of an integrated approach. Firstly, the industry is stuck in reactive mode regarding its maintenance strategy and there is a lack of direction in terms of the transition to proactive methods. Secondly, the pervasive nature of real-world data issues composes a significant barrier to practical application of data-driven approaches and there is a lack of a means to manage the "3B" data issues in the HVAC domain. The evolution of data-driven analysis is reviewed in the following section, with particular attention given to the value data-driven analysis provides in the industrial setting and the gaps that need to be filled to ensure this value is realized.

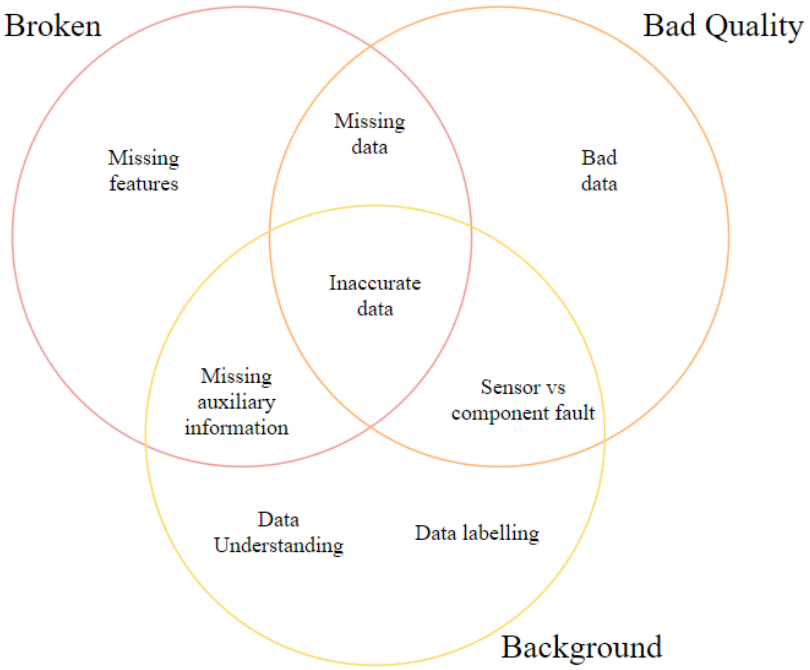
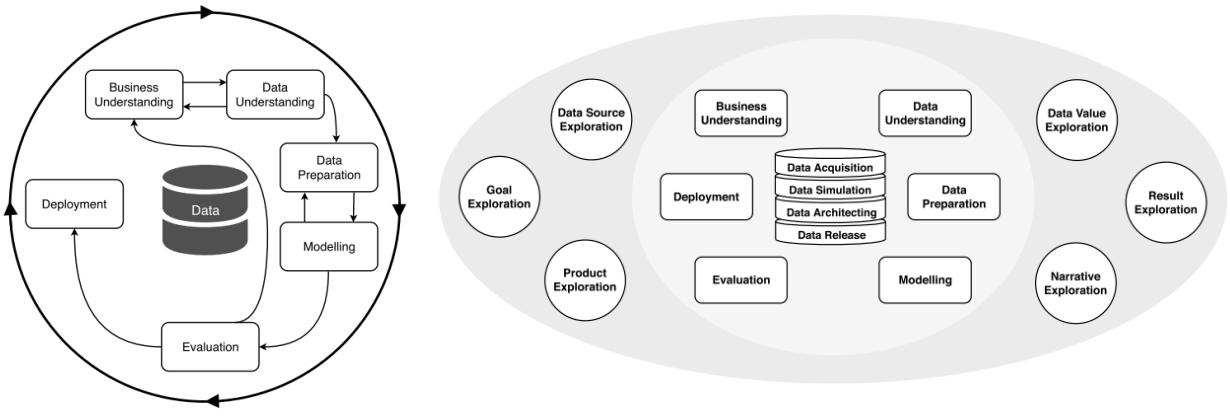


Figure 2. Data Issue Map.

3. Literature Review

3.1. State of the Art Review of Data Mining

In 1996, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was conceived, providing a "blueprint for conducting a data mining project" [29]. Data mining may be described as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [30]. CRISP-DM provides a structured six phase approach with highly flexible transitions between the phases, denoted by the multiple arrows and cyclical nature outlined in Figure 3a. CRISP-DM adequately satisfied the needs of data analysts and became established as the de facto standard for industry [29]. In 2000, Shearer envisaged that extensions and improvements were to be expected [29], and this evolution was analyzed in 2019 by Martínez-Plumed et al. [31].



(**a**) (**b**)

Figure 3. Data analysis models: (**a**) The CRISP-DM process model of data mining [32]; (**b**) DST Map [31].

The evolutions satisfy specific needs while retaining the principles and ideas of CRISP-DM or Knowledge Discovery in Databases (KDD). In 1996, Fayyad et al. [32] were of the opinion that "KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process". However, in more recent times, data mining is often used as a synonym for KDD [31], and we select data mining as the preferred term in this paper. Figure 4 outlines some of the most influential methodologies in the development of CRISP-DM, along with some adaptations that it has inspired. Figure 4 is an adaptation of the evolution illustrated by Mariscal et al. [33] which was released in 2010. Mariscal et al. [33] name KDD and CRISP-DM as the two main approaches that are referred to as the "canonical" methodologies by Martínez-Plumed et al. [31], depicted in grey in Figure 4. The most recent adaptations appear to be CRISP-DM related, addressing a variety of specific needs, such as the project management need in IBM (ASUM-DM [34]), the need for collaboration between geographically diverse groups (RAMSYS [35]) and the need for context change (CASP-DM [36]). More recently in 2020, Plotnikova et al. [37] performed a systematic literature review of data mining adaptations, categorizing the derivative work as modification, extension and integration-type adaptations. Modifications aim at solving a particular problem in a given case study, while extensions are aimed at altering data mining methodologies to account for specialized environments and incorporate context-awareness [37]. In contrast, integration-type adaptations are aimed at combining data mining methods with other domain frameworks, methodologies and concepts [37]. Prior to 2008, most data mining methods (CRISP-DM) were applied "as is", whereas the use of adapted data mining methodologies slowly began to take precedence in recent years [37]. Therefore, a key finding of the literature review is that data mining methodologies now need to be framed as part of "a broader ecosystem of methodologies", rather than the traditional, isolated implementation [37].

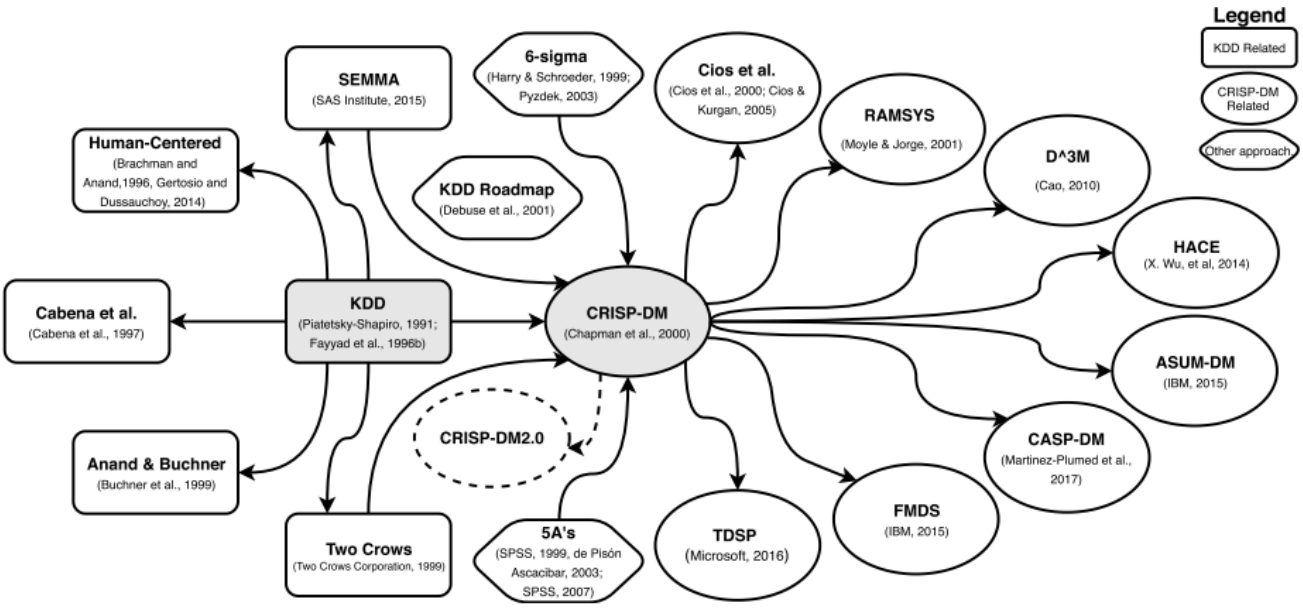


Figure 4. Evolution of Data Mining and Data Science models and methodologies [31].

A change in the data mining landscape is also observed by Martínez-Plumed et al. [31], noting that the term "data science" is preferred to the use of data mining. Martínez-Plumed et al. [31] make the argument that "the key difference (we perceive) between Data Mining twenty years ago and data science today is that the former is goal-oriented and concentrates on the process, while the latter is data-oriented and exploratory". The term exploratory data analysis (EDA), a closely related term to data mining, was first introduced in 1977 by John Tuckey [38] to describe the detective-like analysis of data—that is to say, to not take the data at face value or as absolute truth. Therefore, it is important to understand the data

and the problem that needs to be solved to determine if the analysis is goal-oriented or data-oriented.

3.2. Data Mining and Data Science for HVAC Maintenance

In the context of data-driven maintenance, both approaches are useful. For known faults in mechanical equipment, whereby the failure modes are known, a goal-oriented CRISP-DM approach is necessary. By contrast, for unknown faults, those that are beyond the current knowledge of domain experts, a data-oriented and exploratory approach is necessary. This is a key concept in the Industrial AI paradigm [27].

The majority of manufacturing problems can be classified as either visible problems or invisible problems [39], as in Figure 5. Invisible problems encompass "machine degradation, lack of lubrication, loss of accuracy, wear and tear of parts, and resource waste" [39], while visible problems are often the result of invisible factors such as "component failure, equipment downtime, machine break-downs, decrease in product quality" [39]. These problems need to be both solved and avoided, requiring different approaches. The visible problems are familiar to domain experts and goal-oriented CRISP-DM solutions are suitable, especially in the Visible and Avoid quadrant [39]. The invisible problem quadrants will require an exploratory approach as the problem is less defined. Based upon these differing trajectories, Martínez-Plumed et al. [31] introduced the Data Science Trajectories (DST) map. The outer circle, in Figure 3b, contains the exploratory activities and the inner circle contains CRISP-DM or goal-oriented activities. For industry to transition from reactive to proactive approaches, goal-oriented CRISP-DM visible problem solving and avoidance will need to evolve to data-oriented exploratory invisible problem solving and avoidance. That is to say, the visible problems must be adequately managed before the value of data-driven analysis may be realized through invisible problem solving. While originating from the two distinct disciplines of engineering and computer science, an integrated approach would fill a knowledge gap, enabling a seamless transition from reactive maintenance to proactive maintenance practices in the era of digitalization. Many integration-type adaptation studies have shown that CRISP-DM may be successfully combined with domain methodologies [37]; however, evolving from visible problem solving to invisible problem solving requires further study. As outlined in Table 1, the 3B's are halting this progression. Therefore, a means of identifying and resolving these issues is a fundamental component of enabling data-driven proactive maintenance. In the following section, the state of the art in visible and invisible problem solving in generic industrial and HVAC-specific applications is discussed to ascertain the gaps that data issues create to enable the development of an integration-type solution.

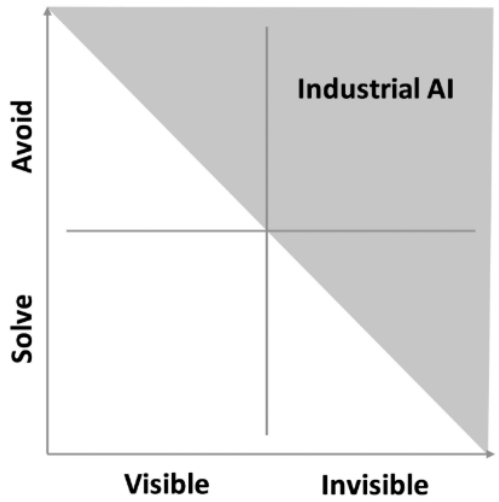


Figure 5. The four quadrants of industrial AI opportunity space [39].

3.3. Visible and Invisible Problem Solving

Visible problems are well understood by domain experts as they both solve and avoid these problems on a daily basis. While data may inform the domain experts decision making, auxiliary information such as past experience and comfort complaints provide the basis for many of the decisions made in practical applications. Data analysts do not possess this valuable information, and given the complex nature of manufacturing [22], a knowledge gap forms that creates the background issue. In HVAC applications, visible problem solving using fault detection and diagnosis (FDD) is a mature area of research. While differences exist in the literature on the classification of HVAC FDD methods, the categorization proposed by Mirnaghi and Haghighat [26] is appropriate for this study, such that methods may be knowledge-based, model-based or data-driven. A model-based method aims at estimating parameters and signals by applying a dynamic process, while a data-driven method "does not require any intervention of human knowledge or physical models, and it only needs real system operational data" [26]. Knowledge-based methods, on the other hand, are defined as the qualitative part of model-based and data-driven methods. In 2008, Fan et al. [40] noted that most knowledge-based methods are rule-based expert systems, which have been developed to elicit the tacit knowledge of experts [41]. Known (visible) problems are detected and diagnosed using a series of "IF-THEN" rules. For air handling units (AHUs), a fundamental component of HVAC systems, House et al. [42] developed the AHU Performance Assessment Ruleset (APAR) that utilized the understanding of the conservation of energy and mass laws. It appears that for HVAC applications, knowledge-based methods may adequately solve the visible problems. However, these methods are limited to the experience of the domain expert. Therefore, they are not capable of detecting faults that are not flagged in historical data or those faults which are beyond the engineer's experience [26]. To improve the problem-solving performance, data-driven techniques are required to detect these unknown faults. However, industry cannot make this transition as current practice (knowledge-based approaches) does not require a comprehensive dataset, which facilitates the propagation of digital transition ceasing data issues. Therefore, the gap that appears to emerge is that current visible problem-solving techniques do not sufficiently improve the quality of the dataset to enable data-driven techniques for invisible problem solving.

Invisible problems often lead to the visible problems [39]; therefore, early identification of these problems enable proactive decision making to avoid inefficient operation. By their nature, invisible problems are not easily identified and advanced techniques, beyond human comprehension, are required to reveal them [27]. One of the key enabling technologies driving the transition to more sustainable practices [43], as identified in the 2021–2027 Horizon Europe funding programme, is artificial intelligence (AI).

AI is a branch of computer science that aims at mimicking the brain's ability to learn from experiences. A subset of AI is machine learning (ML) where the main definition, from 1959, but which is still valid today [24], is allowing computers to solve problems without being specifically programmed to do so [44]. AI and ML give machines the ability to learn and decipher the solution to various problems that would normally rely on human intelligence. The benefit of AI techniques is that it can be far easier to train a system to identify abnormal behavior by presenting it with examples of desired input-output behavior, rather than manually model or program for every possible input [45]. ML has been successfully applied to perform tasks such as computer vision, speech recognition, natural language processing and robot control [45] in sectors such as marketing, finance, telecommunications and network analysis [46]. This technology could enable manufacturers to get greater insight into the performance of their assets by clearly identifying trends and characteristics to solve invisible problems that have not yet been fully recognized [27]. The insight and evidence contained in the data also enables predictive analysis, which is a key benefit of AI [27]. Through exploration of complex relationships in the data, new knowledge is generated that may be iteratively improved in a sustainable and continuous manner to increase the efficiency of industrial systems [27].

There are three basic paradigms in ML as in Figure 6: supervised, unsupervised and reinforcement learning [45]. Supervised learning, the most widely used method [45], learns a function that maps an input to an output based on labelled training data. The most common supervised learning algorithms include neural networks, support vector machines, decision trees, logistic regression and naïve bayes classifiers [23]. The output variables may be either categorical or continuous and are known as classification tasks and regression tasks, respectively. Examples in the industrial context outlined by Bertolini et al. [23] include process failure detection (classification) and physical property prediction such as thickness or surface roughness from parts processed by a numerical control machine (regression). Unsupervised learning, on the other hand, does not receive any output information. Its goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. [47]. Ghahramani. [47] suggests that unsupervised learning can be thought of as finding patterns in data that would otherwise be considered as unstructured noise. Classic examples include clustering and dimensionality reduction. Lastly, reinforcement learning differs by interacting with its environment and producing actions. These actions change the state of the environment that causes the algorithm to receive rewards or punishments. According to Ghahramani. [47], the goal of the algorithm is to maximize the rewards and minimize the punishments it receives over its lifetime. It is closely related to control theory in engineering.

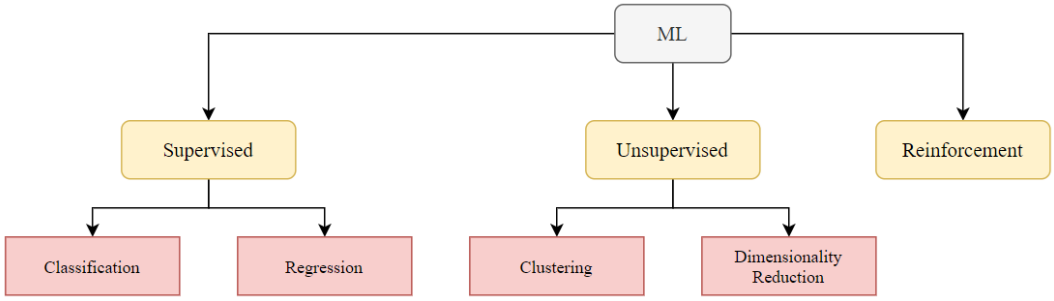


Figure 6. Categorization of ML paradigms.

While ML is one of the key enablers to evolve a traditional manufacturing system up to the Industry 4.0 level [48], ML is not applicable to every industrial problem. Lee et al. [15] recommends that systems with differing complexity and uncertainty levels should have different maintenance strategies, and by extension, are not in need of ML solutions. In the case of HVAC systems on industrial sites, there is low availability of facilities personnel to tend to these assets, often outnumbered by 20 AHUs to one technician [41]. Therefore, costly external consultancy is required to ascertain the cause of complex issues [49] and prioritize the list of issues that need to be resolved. For this reason, the ideal solution is a proactive, data-driven maintenance strategy. Data-driven methods perform well in large-scale and complex systems, but they require high quality and sufficient training data [26], which has not been a priority in the past.

Transitioning to data-driven, proactive and energy efficient practices faces many challenges. In practical applications, these challenges are not being adequately addressed as many hi-tech facility managers admit to being insufficient in terms of their collection and analysis of HVAC data [10]. Visible problem solving is the key focus in industry, but the frequency of their occurrence may be reduced through invisible problem solving. The collaboration of traditionally different disciplines (engineering and computer science) "is necessary to drive progress" [24], proven by the increasing number of studies analyzing the integration of domain specific frameworks, methodologies, and concepts with data mining methodologies [37]. This is a difficult task, however, as the historical lack of emphasis on data analytics has created "a skills gap when it comes to a data centered-mindset amongst the manufacturing workers" [50]. For this reason, we propose that a novel framework is

developed based on an integration-type adaptation of CRISP-DM. The lack of direction in terms of the transition to proactive maintenance would be addressed by solving and avoiding visible problems using knowledge-based approaches and solving and avoiding invisible problems using data-driven approaches. Secondly, further investigation of the data issues is required to manage the "3B" data issues. We propose that CRISP-DM is extended to put more emphasis on the assessment of the data to aid the detection of data issues, such as those in Figure 2. In the following section, the necessary components of CRISP-DM are supplemented with the activities to enable invisible problem solving in Table 2, to develop a unified framework to transition to proactive maintenance practices.

Table 2. Data issue threat to visible and invisible problem solving.

Issue	Visible	Invisible	Path from Visible to Invisible
Broken	Low	High	• Determine the sufficiency of the current dataset • Determine the value more data provides
Bad Quality	Medium	High	• Determine the sufficiency of the data quality • Determine the value better data quality provides
Background	High	Medium	• Determine the operating status of the system • Determine the sufficiency of the current problem understanding • Determine the value exploratory-based analysis may provide

4. IDAIC Framework

A process to develop a framework that maintains the necessary phases of CRISP-DM while also making the necessary adaptations to integrate domain knowledge and improve the AI/ML training dataset is needed. The process followed by Corrales et al. [51] and Almutiry et al. [52] is adopted as data quality frameworks were successfully developed from this approach. The tasks to develop the framework are gathering, filtering and mapping, and clustering. In the gathering section, the necessary activities to be included in the framework are collected. In the following task, filtering and mapping, the activities gathered are grouped where necessary to distil these activities into phases. In the last section, the phases are further clustered to achieve the high-level goals of the framework.

The threat that each data issue has to visible and invisible problem solving is summarized in Table 2. Broken datasets are considered a low threat to visible problem solving as the data requirements are low and partial application of knowledge-based techniques are possible. However, broken datasets are a high threat to invisible problem solving as data-driven approaches require comprehensive datasets. In the case of bad quality data, the threat to visible problem solving is considered as medium as error thresholds may account for bias or precision degradation faults. However, the presence of drift, spike and stuck faults may severely reduce the visible problem-solving performance. The need for high quality data in data-driven approaches results in a high threat level for invisible problem solving. Lastly, the need for domain understanding to solve visible problems is high and therefore the background issue is a significant threat. For invisible problem solving, an understanding of the domain is necessary but not as in-depth as visible problem solving as anomalies in the data are used to detect faults rather than expert interpretation of the data.

To enable a transition to proactive analysis, the issues in need of the most attention are broken data and bad quality data, as the background issue may be adequately addressed through the integration of comprehensive visible problem solving. The low data requirements of visible problem solving means that the breadth of the data analyzed in this phase is lower than the invisible problem solving stage. For this reason, the data analysis in the visible problem phase will need to be extended beyond the key features needed to detect faults, to analyze other features that may be useful for the invisible problem stage. That is to say, to address the broken data issue, features that are not currently analyzed in the visible problem space will need to be analyzed using domain knowledge. Similarly, the data quality requirements are not as high in the visible problem space as in the invisible

problem space, but more rigorous analysis of data quality is needed to ensure that the data meets the invisible problem-solving requirements.

The principles and phases of CRISP-DM appear to be universally accepted as necessary components of data analysis. However, data in the industrial domain poses some extra difficulties, as mapped in Figure 2, which require more guidance for practical implementation. The trend that emerges is that while data science methodologies as such CRISP-DM provide sufficient structure to perform data analysis, they do not provide sufficient domain understanding to tackle visible problems in the industrial setting. The main reason CRISP-DM needs to be adapted is to address this lack of domain understanding, by integrating domain specific methodologies, such as those developed in the area of FDD. Secondly, the lack of necessary data and insufficient data quality is limiting the value of data-driven analysis in the real-world industrial setting. Hence, the training data for ML application is not sufficient in terms of quantity or quality to realize the benefits of data-driven approaches such as predictive analysis. Therefore, more emphasis on these issues is required to obtain an AI/ML ready dataset than "as is" CRISP-DM implementation provides.

In Figure 7, the CRISP-DM phases in Figure 3a and the activities to transition from visible to invisible problem solving in Table 2 are gathered. Based on the comparative study of popular data mining processes (KDD, SEMMA and CRISP-DM) by Azevedo and Santos [53], a description of each CRISP-DM phase is also included in the gathering section. The data understanding phase is therefore repeated to describe the activity of acquiring the relevant data (2) and the activity of searching for unanticipated trends and anomalies (3). In the following section, similar activities are grouped. For example, the activities of determining data quality (C), determining the value of additional data (B) and determining the value of better quality data (D) may all be achieved in a data assessment phase. While the data understanding phase of CRISP-DM contains a "verify data quality" task [54], it does not provide sufficient detail to achieve tasks B, C and D. The data understanding phase (2+3) is therefore broken into two phases: data contextualization (2+A) and data assessment (3+B+C+D). Data contextualization integrates the activity of acquiring the data (2) and determining if the dataset is sufficient (A). Lastly, in alignment with the DST map proposed by Martinez-Plumed et al. [31], the phases are clustered to achieve the goals of visible (goal-oriented) and invisible (data-oriented) problem solving.

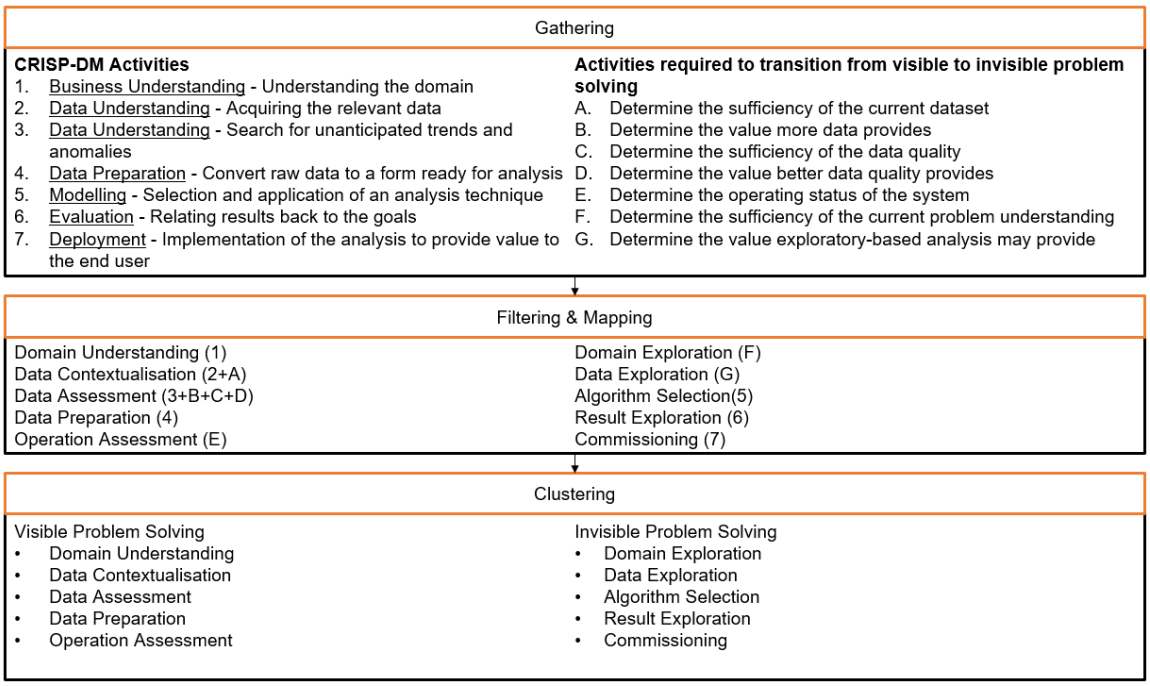


Figure 7. Framework Development.

The framework is illustrated in Figure 8, retaining a similar structure to that of CRISP-DM and the DST map. The framework is named the Industrial Data Analysis Improvement Cycle (IDAIC), as the purpose of the framework is to improve the analysis of industrial data so that industry may transition from reactive to proactive decision making. The main contributions made to CRISP-DM is summarized as follows:

- Extension of Data Understanding phase to Data Contextualization and Data Assessment. These phases aim to provide greater guidance on interpreting the meaning and sufficiency of the dataset and measuring the value of the data, respectively.
- Addition of an Operation Assessment phase to integrate domain knowledge for the identification of known faults that would assess the health status of the asset.
- Addition of a Domain Exploration phase to elicit the tacit knowledge of experts for the guided exploration of areas of improvement through invisible problem solving.
- Addition of a Data Exploration phase to determine if the invisible problem-solving aspirations may be realized with the available dataset.
- Addition of a commissioning phase to bridge the gap between domain knowledge-reliant goal-based visible problem solving and data-reliant exploratory-based invisible problem solving. The commissioning phase also incorporates the task of implementing the analysis for value creation that has previously been satisfied by the Deployment phase in CRISP-DM.

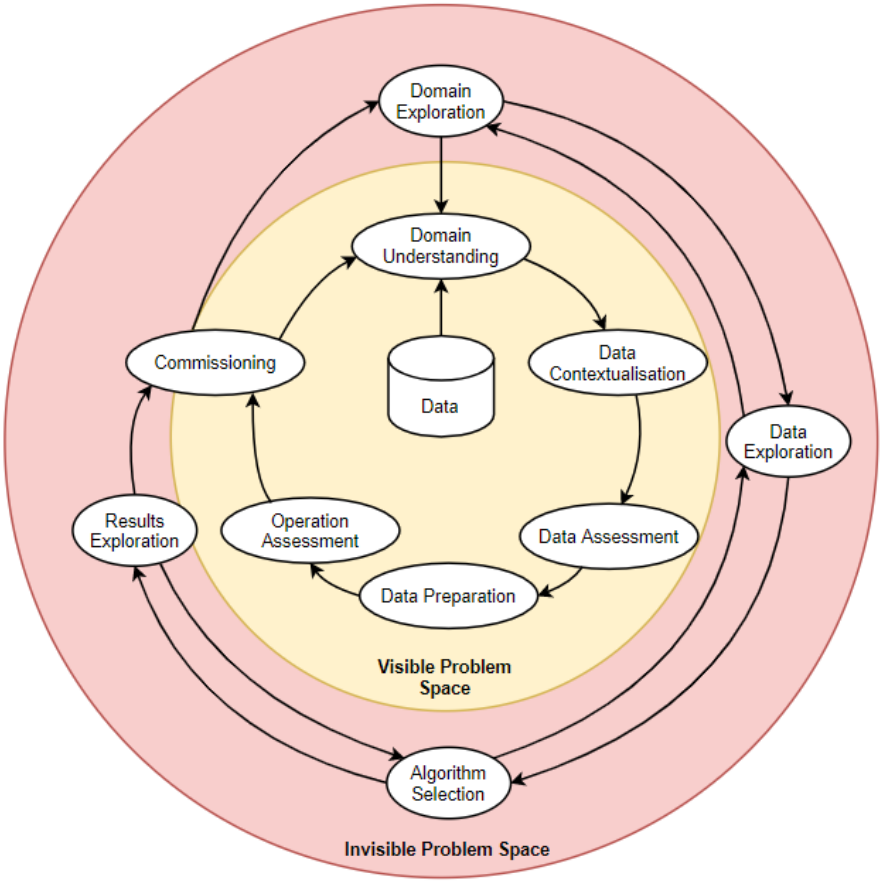


Figure 8. Industrial Data Analysis Improvement Cycle.

The Business Understanding, Modelling and Evaluation phases are renamed to Domain Understanding, Algorithm Selection and Result Exploration. Novel contributions to these phases are not proposed in this paper. The following sections describe the phases in the IDAIC framework, along with possible methodologies to satisfy specific phases.

4.1. IDAIC Phases

4.1.1. Domain Understanding

The first phase in the IDAIC framework aims at collecting the necessary information to obtain a comprehensive understanding of the asset or system under study. In accordance with the Industrial AI paradigm [27], we categorize the required inputs as people, systems and things. The key people, or the roles that must be fulfilled, are those of a domain expert and a data expert. The understanding of the domain and the understanding of data analytics are two key skills needed for successful analysis of industrial data. These people must then integrate with the systems (such as data from heterogeneous sources), and things (such as documentation and equipment specifications). This phase is similar to Business Understanding in CRISP-DM; however, given the engineering context, Domain Understanding is considered to be more aptly named.

4.1.2. Data Contextualization

The problem understanding that is elicited from the integration of the relevant people, systems and things is then used to contextualize the problem. The relevant data is not only acquired, but also evaluated at a high level to ascertain if the data is sufficient. Tasks that would be beneficial to determine the comprehensiveness of the data may include measuring the granularity of the data, identifying labels in the data, evaluating the architecture of the data and determining the real time capabilities of the data. Each of these tasks would aid in the evaluation of the data to determine if the dataset is fit for purpose, or in another word, a data gap analysis. This phase is similar to data understanding in CRISP-DM; however, more emphasis is given to determining the problems that the dataset can and cannot help to solve. This is achieved through an integration of highly interpretable expert rules that enables the quantification of the problems or faults that may be solved with the available data. Therefore, the output of this phase is an understanding of the dataset with emphasis on the identification of the visible problems that may be solved.

4.1.3. Data Assessment

Once the high-level capabilities of the dataset are understood, the data may be assessed in further detail to evaluate the quality of the data. The data assessment phase is focused on managing the bad quality issue but may also provide insight into the completeness of the data (broken) and the operation of the asset (background). The area of measuring data uncertainty, or data quality assessment, is not a mature area of research in industrial applications. In 2015, Cai and Zhu [55] proclaimed that a unified and approved data quality standard had not been agreed for big data and that research in the area had just begun. However, many data quality assessment methodologies have been developed in other domains, which have been systematically reviewed by Batini et al. [56] in 2009. The authors found accuracy (Acc), completeness (Comp), and consistency (Cons) to be the dimensions of highest consistency in the literature [56]. Therefore, in the data assessment phase we propose that the assessment of these dimensions is the main focus, and the associated metrics to assess these dimensions are heavily influenced by domain knowledge. In Figure 9, a decision tree is proposed to manage the 3B's based on the outcome of the data assessment. If the required data is missing, then a digitalization plan is required to obtain the data, such as the approach proposed by Clancy et al. [57]. If the data is inaccurate, then the sensors need to be manually checked and may require recalibration or replacement. Finally, if the consistency of the data is low, or the data is not in accordance with the domain expert's experience, sensors may also be in need or recalibration or replacement. The difference from the accuracy dimension is that inconsistent data may be caused by a component or control fault; therefore, an operation assessment is needed to determine the source of the anomaly. However, the data must firstly be appropriately prepared.

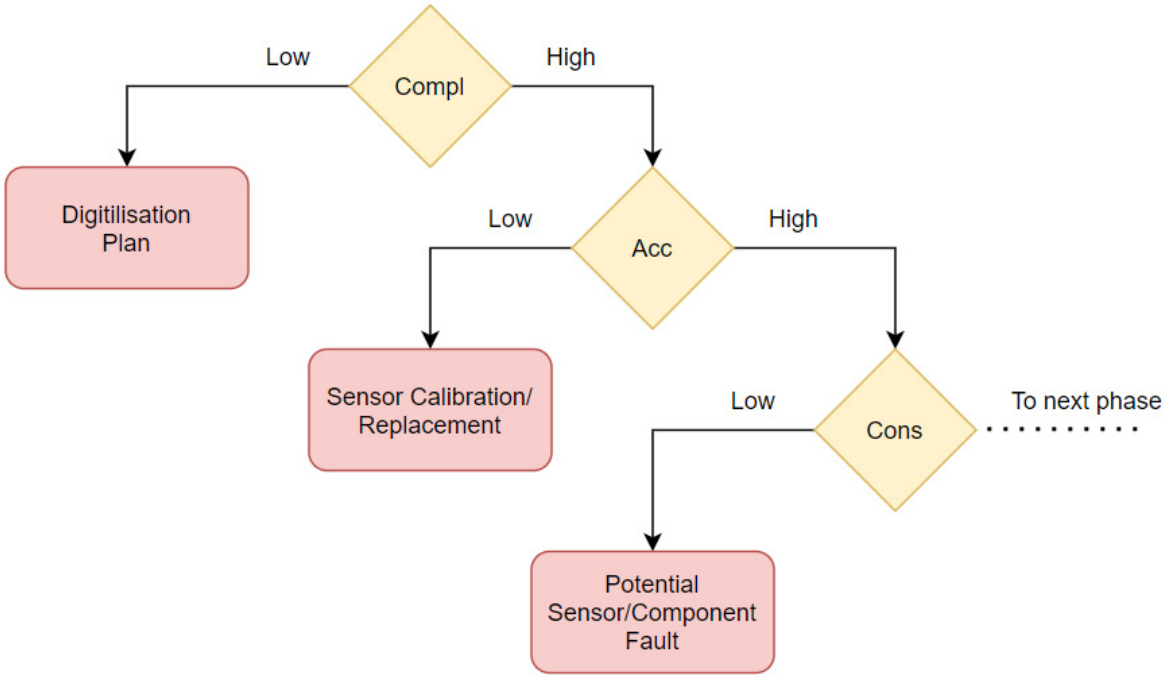


Figure 9. Data Assessment Decision Tree.

4.1.4. Data Preparation

Data preparation is commonly implied as the most time-consuming task in a data analytics project. The data preparation phase in the IDAIC framework is inspired by the CRISP-DM data preparation phase. However, tasks such as data cleaning are more informed from the domain-integrated data assessment phase, facilitating faster preparation of the data given the known issues with the dataset. The data contextualization phase is also helpful to provide clarity in this phase for tasks such as changing the naming convention or aligning heterogeneous data sources based on the granularity of time-based data.

4.1.5. Operation Assessment

Assessing the operation, or solving visible faults, is aligned to the area of FDD. The visible problems must be solved before the value of data-driven analytics may be realized. The strengths and weaknesses of each FDD classification outlined by Mirnaghi and Haghighat [26] is shown in Table 3. Model-based methods are not appropriate for this phase as significant effort is required to develop models that are highly susceptible to error. Data-driven methods are the most promising techniques for highly scalable solutions, but in the visible problem-solving space, the quantity and quality of the data is not suitable. Therefore, knowledge-based approaches appear to be the best due to their low data requirements. While knowledge-based approaches are limited to the domain expertise available, these methods sufficiently address visible problems in a highly interpretable manner. Knowledge-based methods integrate domain expertise into the data analysis pipeline and ensure the system is operating as expected so that invisible problem solving is enabled. The operation phase will also aid in the detection of the most problematic components in the system, which may be used to inform the domain exploration phase in the invisible problem space.

Table 3. Comparison of strengths and weaknesses of HVAC FDD methods. Reprinted from Mirnaghi and Haghighat [26], with permission from Elsevier, 2020.

Method	Strength	Weakness
Knowledge-based FDD	• The detailed mathematical model is not needed available • Suitable for the system with a small number of inputs, outputs, and states	• It highly relies on domain expertise while a wide range of faulty and failure cases are beyond engineers' experiences • Not capable of detecting novel failures which are not flagged in the historical data • Cannot work well for complex and large-scale HVAC systems
Model-based FDD	• Works properly when a good HVAC physical model is available	• It is not proper for large-scale and complex HVAC systems • The performance would be limited because of modeling and linearization error
Data-driven based FDD	• Less model development time and cost • No dependency on the model • Easy to retrain • Efficient use of system data • High accuracy • Good performance in works with large-scale and complex systems • Minimizing cognitive errors which are beyond the engineers' knowledge	• Need to collect the high quality and sufficient training data for supervised models • High dependency on the quality and the quantity of the collected data

4.1.6. Commissioning

The issues identified in the data assessment and operation assessment must then be actioned upon. In HVAC applications, the manual activity of ensuring systems are operating in accordance with design specifications is known as commissioning. In a review of FDD methods for AHUs, Bruton et al. [49] outlined the benefits of FDD aided commissioning in relation to the four ideal types of commissioning presented in the International Energy Agency's Annex 40 [58]. In the first iteration of the IDAIC framework, the phase will either be retro-commissioning or re-commissioning, based upon the occurrence of an initial commissioning phase. In subsequent iterations of the framework, the activity is known as on-going or continuous commissioning as the occurrence of faults are quickly identified and appropriate maintenance is carried out to maintain and improve the system performance. The integration of the commissioning phase is a key adaptation to CRISP-DM as it enables the transition from the visible to invisible problem solving, or reactive to proactive maintenance activities.

4.2. Invisible Problem Space

4.2.1. Domain Exploration

Once the system is re-commissioned or retro-commissioned, the invisible problem-solving activity is enabled. The outer ring in Figure 8 aims to explore the data to detect early signs of degradation that would inform decision makers on the remaining useful life of components to plan maintenance activities accordingly. The first phase is to determine if the current problem understanding is sufficient. To do this, exploration of the domain is undertaken to establish if the system is operating in a stable and efficient manner. While the operation assessment aims at detecting and diagnosing visible problems, the domain exploration aims at assessing the efficiency of the system and propositioning the areas where data-driven analysis may provide value. For example, the operation assessment may identify a passing valve that is then replaced in the commissioning phase. However, this valve may have reached the end of its useful life prematurely, resulting in costly replacement and insufficient knowledge as to the cause of the rapid degradation of this valve. In this scenario, data-driven analysis may provide insight into the degradation pattern of the new valve, creating valuable information for the facilities management team that enables smarter, proactive decision making. In collaboration with the domain expert, the features of interest may be targeted for analysis.

4.2.2. Data Exploration

In the following phase, the proposed concepts, or areas for further exploration, are assessed using the data. The viability of a data-driven approach is justified based on the ability of the dataset to address the invisible problem. Common exploratory data analysis and ML techniques are tested in this phase to uncover hidden relationships or unexpected patterns or anomalies in the data that may substantiate the signs of an invisible problem.

4.2.3. Algorithm Selection

The most appropriate algorithm or technique is then selected in the penultimate phase. At the high level, ML algorithms as may be classified as supervised, unsupervised or reinforcement as in Figure 6. Based on the findings in the data exploration phase, an appropriate algorithm is selected that accounts for domain specific auxiliary information that may not be captured in the data.

4.2.4. Results Exploration

Lastly, the meaning of the results is explored. The output of the data-driven analysis is critically appraised before an action item is raised for the commissioning phase. In this manner, the visible and invisible problem-solving tasks are integrated in a single framework for continuous problem-solving improvement.

4.3. Conclusions on the IDAIC Framework

The outer ring in Figure 8, or invisible problem solving, is exploratory in nature; therefore, backward arrows are included to denote the back-and-forth nature of this activity. While this is also the case for visible problem solving, the latter phases of the IDAIC framework are less prescriptive and more transitions between phases are necessary. Furthermore, not every possible transition is shown in Figure 8. Similarly to the DST map proposed by Martinez-Plumed et al. [31], not every phase must be completed for successful industrial data analysis. The IDAIC framework merely provides an integrated solution for early adopters of digitalization to manage the 3B data issues throughout a data-driven industrial analysis project. It is not designed to eradicate every data challenge faced in industrial data analysis through rigid implementation. For example, in systems that are well maintained, visible problems may not occur rendering the operation assessment phase redundant.

The contributions of the novel CRISP-DM integration-type adaptation proposed in this paper are summarized in Table 4. The main barriers to data-driven analysis in the industrial sector are the occurrence of broken and bad quality datasets, along with the lack of understanding of the domain (background). The necessary activities to manage these issues has been outlined in Table 2 and an example of specific tasks in the framework that contribute to the fulfilment of the data analysis enabling activities is outlined in Table 4. While we interpret the data challenges landscape as an overlapping Venn diagram as in Figure 2, an example of managing each issue is outlined in Table 4. Firstly, the completeness of the dataset is assessed by performing a data gap analysis in the data contextualization phase. In collaboration with the domain expert, the possible faults and data required to identify their occurrence are listed. A high-level quantitative measure of the completeness of the dataset may then be obtained to understand the value of the current dataset and the opportunity that exists if the system is further digitalized. Secondly, the quality of the dataset is evaluated by measuring the deviation from expected values as defined through domain integration. In the semantic accuracy analysis task, traditional expert rule sets have been repurposed to measure the quality of the dataset rather than identify faults in the system. Thirdly, knowledge-based FDD has been integrated into the data analysis to combine the benefits of data-driven and knowledge-based FDD. The high interpretability of these techniques facilitates better understanding of the operation of the system, reducing the knowledge gap between engineering and data science type approaches.

Table 4. Novel tasks in the IDAIC framework to address the research objectives.

IDAIC Phase	Primary Data Issue	Novel Task/Integration	Proposed Methodology/Solution
Data Contextualization	Broken	Data gap analysis	Number of problems that can/cannot be solved with the data available using an expert rule set
Data Assessment	Bad quality	Semantic Accuracy Analysis	Expert rule-based analysis of data quality
Operation Assessment	Background	Knowledge-based FDD	Expert rule implementation

5. Discussion

A novel framework is proposed to integrate traditional engineering or knowledge-based approaches for visible problem solving, with techniques that originate from the field of data analysis for invisible problem solving, for the purpose of advancing industrial data analysis and reducing energy consumption. The Industrial Data Analysis Improvement Cycle (IDAIC), Figure 8 is a framework that simultaneously aims to provide short term value by reactively analyzing data using knowledge-based approaches, while also facilitating the transition to proactive analysis by accessing data quality and monitoring its improvement over time. The conventional data analysis approach must be supplemented with strong domain knowhow and therefore the CRISP-DM methodology is extended through the novel integration of data assessment techniques, expert systems and exploratory data analysis. The IDAIC framework begins with understanding-centric phases, whereby the domain and the data are assessed to determine the value that the problem-solving activities may realize. Data assessment is then performed to determine the quality of the dataset or the fitness of the dataset for AI & ML application. Data points that are deemed to be of concern are marked for further analysis in the operation assessment phase, with appropriate pre-processing activities performed in advance. The operation of the asset is assessed using knowledge-based FDD techniques, such as expert rules, to identify visible problems in the system. While problems that are identified may be caused by those bad quality data points marked from the data assessment phase, it may incentivize industrial practitioners to recalibrate or replace faulty sensors when highly interpretable expert rules cannot be applied. While the IDAIC framework aims to enable industry to perform advanced analysis, the overall goal is energy consumption reduction; therefore, the analysis must inform the physical changes that are made to the system. Commissioning may be defined as "a quality-oriented process for achieving, verifying and documenting whether the performance of a building's systems and assemblies meet defined objectives and criteria" [58]. While sensor faults are resolved in commissioning activities, the data quality demands of AI and ML solutions require vigorous data assessment and maintenance of the sensor. Therefore, the commissioning phase in the IDAIC framework not only performs the necessary maintenance activities to achieve energy efficient operation, but also aims at performing the necessary maintenance activities to collect the highest quality of data possible.

At this point in the IDAIC framework, domain knowledge has been sufficiently integrated to solve visible problems and the asset or system is operating in a faultless, energy efficient state with good quality data continuously being collected. Industry may then switch from a reactive maintenance approach to a proactive maintenance approach with data-driven techniques. However, while domain expertise informed the inner goal-oriented circle, the outer circle is data-oriented and exploratory in nature. Similar to the DST map in Figure 3b proposed by Martinez-Plumed et al. [31], the invisible problem solving and avoidance begins with goal, or domain, exploration. Strong collaboration is required between data analysts and domain experts in this phase to elicit potential areas of interest to apply data-driven algorithms. For example, the cause of a recurring fault or failure may be unknown and an unsupervised ML algorithm may provide insight through pattern recognition or clustering techniques. Similarly, ML may be used to anticipate a particular fault that is deemed critical by the domain expert, providing an estimate of the remaining useful life in a predictive maintenance strategy. Once a goal has been selected, the data exploration takes place to uncover if all relevant information is available and

of sufficient quality to perform the analysis. If this is not the case, a return to the goal exploration phase is needed to either update the goal based on the information available or else return to the gathering phase to obtain this information. Data exploration is a necessary intermediary phase to determine if the aspirations of the defined goal are attainable. If the data-oriented goal is achievable, an appropriate algorithm must then be selected. The challenge of algorithm selection and result evaluation is commonly stated in the literature as shown in Table 1. Through solving the visible problems and determining if the data-oriented goal is achievable in the domain exploration phase, we the authors envisage that much of the uncertainty in data-driven analysis will be removed, enabling a clearer decision to be made on algorithm selection. The algorithms results are then evaluated in the final phase of the IDAIC framework, whereby the results are used to inform decision makers to make energy reduction decisions.

To test the effectiveness of the proposed framework, we plan to apply the IDAIC framework to an air handling unit (AHU) in a large medical devices manufacturing facility. While full implementation will be documented once complete, a preliminary finding is introduced in this paper to illustrate the benefit of practical application of the IDAIC framework in the real world. During the COVID-19 pandemic, our industrial partners made the decision to operate a recirculating air AHU as a full fresh air AHU to maximize the fresh air supplied to the production floor. This resulted in a novel mode of operation for a recirculating air AHU, which reduced the effectiveness of knowledge-based approaches. Early implementation of the IDAIC framework resulted in the identification of spike faults and bias faults in the data assessment phase which would also reduce the effectiveness of data-driven approaches. Through collaboration with onsite facilities personnel, a novel knowledge-based approach was developed to detect faults during this mode of operation in the operation assessment phase. While further detail regarding the development of this fault detection approach is left for future work, early indications suggest that the IDAIC framework may facilitate a collaboration with onsite engineers to enable the improvement of data analysis, despite the impediments presented in the real world.

6. Conclusions and Future Directions

A conceptual framework has been proposed in this paper to aid the transition from reactive to proactive maintenance approaches to enable energy reduction. An integration-type adaptation of CRISP-DM is developed to incorporate domain expertise into the analysis. The novel integration of engineering knowledge and a data science process model has led to a structured cycle whereby knowledge-based approaches solve and avoid visible problems, and data-driven approaches solve and avoid invisible problems. The IDAIC framework therefore addresses the lack of direction to transition from reactive to proactive maintenance approaches. The second novel contribution of this paper is the extension of CRISP-DM for the purpose of managing the "3B" data issues. The contributions of greatest importance are firstly, the data contextualization phase whereby domain expertise is incorporated to understand if the features available are capable of solving visible problems. Secondly, the data assessment phase incorporates knowledge-based FDD for the novel purpose of assessing data quality rather than problem identification. Thirdly, knowledge-based FDD is applied to retrospectively analyze the asset or system to isolate visible problems that must be solved before the value of data-driven analysis may be realized on invisible problem solving.

While the proposed IDAIC framework has been discussed at a high level in this paper, practical implementation must follow in future work. As noted by Wirth and Hipp [54], "if the early adopters fail with their data mining projects, they will not blame their own incompetence in using data mining properly but assert that data mining does not work". As discussed in Section 5, we plan to apply the IDAIC framework to an air handling unit in a large medical devices manufacturing facility to test the applicability of the framework in the real world.

Author Contributions: Conceptualization, M.A., D.T.J.O. and K.B.; methodology, M.A. and K.B.; investigation, M.A., D.T.J.O. and K.B.; resources, D.T.J.O. and K.B.; writing—original draft preparation, M.A.; writing—review and editing, M.A., D.T.J.O. and K.B.; visualization, M.A., D.T.J.O. and K.B.; supervision, D.T.J.O. and K.B.; project administration, D.T.J.O. and K.B.; funding acquisition, D.T.J.O. and K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has emanated from research that was part-funded by Science Foundation Ireland (SFI) through MaREI, the SFI Research Centre for Energy, Climate, and Marine [Grant No: 12/RC/2302_P2], with supporting funding obtained from DePuy Synthes. For the purpose of OpenAccess, the author has applied a CCBY public copyright license to any Author Accepted Manuscript version arising from this submission.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Energy Outlook 2021-Analysis-IEA. Available online: https://www.iea.org/reports/world-energy-outlook-2021 (accessed on 1 June 2022).
2. U.S. EIA. *Annual Energy Outlook 2022*; 2022. Available online: https://www.eia.gov/outlooks/aeo/pdf/AEO2022_Narrative.pdf/ (accessed on 1 June 2022).
3. UN General Assembly. Transforming Our World: The 2030 Agenda for Sustainable Development; Report No. A/RES/70/1. 2015. Available online: https://sustainabledevelopment.un.org/post2015/transformingourworld/publication (accessed on 7 June 2022).
4. European Commission. *Annex to the Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions—The European Green Deal. Roadmap—Key Actions*; European Commission: Brussels, Belgium, 2019.
5. European Commission. Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions—A New Industrial Strategy for Europe. 2020. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0102 (accessed on 7 June 2022).
6. Global e-Sustainability Initiative. *#SMARTer2030—ICT Solutions for 21st Century Challenges*; Global e-Sustainability Initiative: Brussels, Belgium, 2015.
7. IRENA. *Global Energy Transformation: A Roadmap to 2050*; International Renewable Energy Agency: Abu Dhabi, United Arab Emirates, 2018.
8. Pérez-Lombard, L.; Ortiz, J.; Pout, C. A review on buildings energy consumption information. *Energy Build.* **2008**, *40*, 394–398. [CrossRef]
9. Bruton, K.; Raftery, P.; O'Donovan, P.; Aughney, N.; Keane, M.; Sullivan, D.O. Development and alpha testing of a cloud based automated fault detection and diagnosis tool for Air Handling Units. *Autom. Constr.* **2014**, *39*, 70–83. [CrossRef]
10. CIM. The Energy Blind Spots. 2022. Available online: https://cim.io/documents/energy-blind-spots/ (accessed on 1 June 2022).
11. Roth, K.W.; Westphalen, D.; Feng, M.Y.; Llana, P.; Quartararo, L. *Energy Impact of Commercial Building Controls and Performance Diagnostics: Market Characterization, Energy Impact of Building Faults and Energy Savings Potential*; Office of Building Technology, Department of Energy: Washington, DC, USA, 2005.
12. Mills, E. Building commissioning: A golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. *Energy Effic.* **2011**, *4*, 145–173. [CrossRef]
13. O'Donovan, P.; Leahy, K.; Bruton, K.; O'Sullivan, D.T. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *J. Big Data* **2015**, *2*, 1–26. [CrossRef]
14. Dhillon, B.S. *Maintainability, Maintenance, and Reliability for Engineers*; CRC Press: Boca Raton, FL, USA, 2006.
15. Lee, J.; Wu, F.; Zhao, W.; Ghaffari, M.; Liao, L.; Siegel, D. Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mech. Syst. Signal Process.* **2014**, *42*, 314–334. [CrossRef]
16. Bumblauskas, D.; Gemmill, D.; Igou, A.; Anzengruber, J. Smart Maintenance Decision Support Systems (SMDSS) based on corporate big data analytics. *Expert Syst. Appl.* **2017**, *90*, 303–317. [CrossRef]
17. Schein, J.; Bushby, S.T.; Castro, N.S.; House, J.M. A rule-based fault detection method for air handling units. *Energy Build.* **2006**, *38*, 1485–1492. [CrossRef]
18. Zhao, Y.; Li, T.; Zhang, X.; Zhang, C. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renew. Sustain. Energy Rev.* **2019**, *109*, 85–101. [CrossRef]

19. RP-1312—Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units—ASHRAE Store. Available online: https://www.techstreet.com/standards/rp-1312-tools-for-evaluating-fault-detection-and-diagnostic-methods-for-air-handling-units?product_id=1833299 (accessed on 25 March 2021).
20. Lin, G.; Kramer, H.; Granderson, J. Building fault detection and diagnostics: Achieved savings, and methods to evaluate algorithm performance. *Build. Environ.* **2020**, *168*, 106505. [CrossRef]
21. Huang, J.; Wen, J.; Yoon, H.; Pradhan, O.; Wu, T.; O'Neill, Z.; Candan, K.S. Real vs. simulated: Questions on the capability of simulated datasets on building fault detection for energy efficiency from a data-driven perspective. *Energy Build.* **2022**, *259*, 111872. [CrossRef]
22. Dogan, A.; Birant, D. Machine learning and data mining in manufacturing. *Expert Syst. Appl.* **2021**, *166*, 114060. [CrossRef]
23. Bertolini, M.; Mezzogori, D.; Neroni, M.; Zammori, F. Machine Learning for industrial applications: A comprehensive literature review. *Expert Syst. Appl.* **2021**, *175*, 114820. [CrossRef]
24. Wuest, T.; Weimer, D.; Irgens, C.; Thoben, K.D. Machine learning in manufacturing: Advantages, challenges, and applications. *Prod. Manuf. Res.* **2016**, *3277*, 1–23. [CrossRef]
25. Dalzochio, J.; Kunst, R.; Pignaton, E.; Binotto, A.; Sanyal, S.; Favilla, J.; Barbosa, J. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Comput. Ind.* **2020**, *123*, 103298. [CrossRef]
26. Mirnaghi, M.S.; Haghighat, F. Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review. *Energy Build.* **2020**, *229*, 110492. [CrossRef]
27. Lee, J. *Industrial AI: Applications with Sustainable Performance*; Springer: Berlin/Heidelberg, Germany, 2020.
28. Jan, S.U.; Lee, Y.D.; Koo, I.S. A distributed sensor-fault detection and diagnosis framework using machine learning. *Inf. Sci.* **2021**, *547*, 777–796. [CrossRef]
29. Shearer, C. The CRISP-DM model: The New Blueprint for Data Mining. *J. Data Warehous.* **2000**, *5*, 13–22.
30. Hand, D.; Mannila, H.; Smyth, P. Principles of data mining. *Drug Saf.* **2001**, *30*, 621–622. [CrossRef]
31. Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramírez-Quintana, M.J.; Flach, P. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Trans. Knowl. Data Eng.* **2019**, *4347*, 1. [CrossRef]
32. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *Adv. Knowl. Discov. Data Min.* **1996**, *17*, 37–54.
33. Mariscal, G.; Marbán, Ó.; Fernández, C. A survey of data mining and knowledge discovery process models and methodologies. *Knowl. Eng. Rev.* **2010**, *25*, 137–166. [CrossRef]
34. ASUM-DM Teaser. Available online: http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html (accessed on 28 March 2022).
35. Moyle, S.; Jorge, A. RAMSYS-A methodology for supporting rapid remote collaborative data mining projects. In Proceedings of the ECML/PKDD01 Workshop: Integrating Aspects of Data Mining, Decision Support and Meta-Learning (IDDM-2001), Freiburg, Germany, 3–7 September 2001; pp. 20–31.
36. Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Flach, P.; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramírez-Quintana, M.J. Context Aware Standard Process for Data Mining. *arXiv* **2017**, arXiv:1709.09003.
37. Plotnikova, V.; Dumas, M.; Milani, F. Adaptations of data mining methodologies: A systematic literature review. *PeerJ Comput. Sci.* **2020**, *6*, 1–43. [CrossRef]
38. Tukey, J.W. *Exploratory Data Analysis*. 1997, pp. 5–24. Available online: http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf (accessed on 7 June 2022).
39. Lee, J.; Singh, J.; Azamfar, M.; Sun, K. Industrial AI: A Systematic Framework for AI in Industrial Applications. *Zhongguo Jixie Gongcheng/China Mech. Eng.* **2020**, *31*, 37–48.
40. Fan, C.M.; Lu, Y.P. A bayesian framework to integrate knowledge-based and data-driven inference tools for reliable yield diagnoses. In Proceedings of the 2008 Winter Simulation Conference, Miami, FL, USA, 7–10 December 2008; pp. 2323–2329.
41. Bruton, K.; Coakley, D.; Raftery, P.; Cusack, D.O.; Keane, M.M.; O'Sullivan, D.T.J. Comparative analysis of the AHU InFO fault detection and diagnostic expert tool for AHUs with APAR. *Energy Effic.* **2015**, *8*, 299–322. [CrossRef]
42. House, J.M.; Vaezi-Nejad, H.; Whitcomb, J.M. An expert rule set for fault detection in air-handling units/discussion. *Ashrae Trans.* **2001**, *107*, 858.
43. Ortega-Gras, J.J.; Bueno-Delgado, M.V.; Cañavate-Cruzado, G.; Garrido-Lova, J. Twin transition through the implementation of industry 4.0 technologies: Desk-research analysis and practical use cases in Europe. *Sustainability* **2021**, *13*, 13601. [CrossRef]
44. Samuel, A.L. Some Studies in Machine Learning. *IBM J. Res. Dev.* **1959**, *3*, 210–229. [CrossRef]
45. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 6245. [CrossRef]
46. Wang, H.; Ma, C.; Zhou, L. A brief review of machine learning and its application. In Proceedings of the 2009 International Conference on Information Engineering and Computer Science ICIECS 2009, Wuhan, China, 19–20 December 2009; pp. 9–12.
47. Ghahramani, Z. Unsupervised Learning. *Adv. Lect. Mach. Learn.* **2004**, *55*, 80–112.
48. Xu, L.D.; Xu, E.L.; Li, L. Industry 4.0: State of the art and future trends. *Int. J. Prod. Res.* **2018**, *56*, 2941–2962. [CrossRef]
49. Bruton, K.; Raftery, P.; Kennedy, B. Review of automated fault detection and diagnostic tools in air handling units. *Energy Effic.* **2014**, *7*, 335–351. [CrossRef]
50. Lenz, J.; Wuest, T.; Westkämper, E. Holistic approach to machine tool data analytics. *J. Manuf. Syst.* **2018**, *48*, 180–191. [CrossRef]

51. Corrales, D.C.; Ledezma, A.; Corrales, J.C. A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal. *J. Comput.* **2015**, *10*, 396–405. [CrossRef]
52. Almutiry, O.; Wills, G.; Alwabel, A.; Crowder, R.; Walters, R. Toward a framework for data quality in cloud-based health information system. In Proceedings of the Conference on Information Society (i-Society 2013), Toronto, ON, Canada, 24–26 June 2013; pp. 153–157.
53. Azevedo, A.; Santos, M.F. KDD, SEMMA and CRISP-DM: A parallel overview. *IADS-DM* **2008**, 182–185. Available online: https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview (accessed on 7 June 2022).
54. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; pp. 29–39.
55. Cai, L.; Zhu, Y. The challenges of data quality and data quality assessment in the big data era. *Data Sci. J.* **2015**, *14*, 1–10. [CrossRef]
56. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **2009**, *41*, 1–52. [CrossRef]
57. Clancy, R.; O'Sullivan, D.; Bruton, K. Data-driven quality improvement approach to reducing waste in manufacturing. *TQM J.* **2021**. ahead-of-print. [CrossRef]
58. Visier, J.C.; Buswell, R.A. *Commissioning Tools for Improved Building Energy Performance*. 2010. Available online: http://www.iea-ebc.org/Data/publications/EBC_Annex_40_Commissioning_Tools_for_Improved_Energy_Performance.pdf (accessed on 7 June 2022).