



TRANSLATIONAL ARTICLE

Embedding data science innovations in organizations: a new workflow approach

Keyao Li¹ , Mark A. Griffin¹, Tamryn Barker², Zane Prickett², Melinda R. Hodkiewicz³ , Jess Kozman⁴ and Peta Chirgwin⁵

¹Future of Work Institute, Curtin University, Perth, WA, Australia

²CORE Skills, Perth, WA, Australia

³School of Engineering, University of Western Australia, Perth, WA, Australia

⁴Katalyst Data Management, Perth, WA, Australia

⁵Chameleon Mettle Group, Perth, WA, Australia

Corresponding author: Keyao Li; Email: Keyao.li@curtin.edu.au

Received: 02 February 2023; **Revised:** 04 September 2023; **Accepted:** 09 October 2023

Keywords: data roles; data science innovation; multi-disciplinary team; process workflow; skills

Abstract

There have been consistent calls for more research on managing teams and embedding processes in data science innovations. Widely used frameworks (e.g., the cross-industry standard process for data mining) provide a standardized approach to data science but are limited in features such as role clarity, skills, and cross-team collaboration that are essential for developing organizational capabilities in data science. In this study, we introduce a data workflow method (DWM) as a new approach to break organizational silos and create a multi-disciplinary team to develop, implement and embed data science. Different from current data science process workflows, the DWM is managed at the system level that shapes business operating model for continuous improvement, rather than as a function of a particular project, one single business unit, or isolated individuals. To further operationalize the DWM approach, we investigated an embedded data workflow at a mining operation that has been using geological data in a machine-learning model to stabilize daily mill production for the last 2 years. Based on the findings in this study, we propose that the DWM approach derives its capability from three aspects: (a) a systemic data workflow; (b) multi-disciplinary networks of collaboration and responsibility; and (c) clearly identified data roles and the associated skills and expertise. This study suggests a whole-of-organization approach and pathway to develop data science capability.

Impact Statement

This study could contribute to the community of data science in the following aspects: (a) introducing a new data workflow method (DWM) to address the limitations of current process workflows, enhancing both social and individual capital that is necessary for successful embedding; (b) more practically, with the support of interview findings, this study points to a whole-of-organization pathway to foster data science capability for the digital future in industrial operations; and (c) presenting an in-depth analysis to enhance the understanding of the DWM and its application in operational contexts. This insight showcases how data science could be adopted not only as technical solutions for a single project, but also to improve organization data strategy, and in turn overall business operations.

1. Introduction

Emerging technologies are generating vast amounts of data and, as the maturity of cloud resources is removing many restrictions of data storage and sharing, the data science industry is growing rapidly across different sectors (Cheng et al., 2015; Jayapandian and Rahman, 2017). The mining industry, which is the focus of our case study below, is no exception to the pervasive data revolution (Gruenhagen and Parker, 2020; Qi, 2020). However, many data science initiatives suffer from temporary, poorly defined, and ad hoc processes; with inefficient team coordination and information sharing (Bhardwaj et al., 2014; Saltz and Krasteva, 2022). To illustrate, Fortune 1000 companies reported struggling to make progress in their data and analytics investments in 2021, with only 29.2% achieving intended business outcomes, and just 30% having a well-articulated data strategy (Bean, 2021). A 2021 survey of executives revealed that companies were having difficulties executing their data-driven transformation, with only 24.4% embraced a data culture (NewVantage Partners, 2021). Despite more (65%) companies established a chief data officer role in 2021 comparing to only 12.0% in 2012, clarity on responsibilities and reporting relationships remained in flux with only 33.3% reporting that the chief data officer role was sufficiently defined.

There have been consistent calls for more research on managing teams and embedding data science processes within organizations (Martinez et al., 2021). However, data science mostly has been studied as time-bound technical projects with team process and collaboration often considered only within the data analytics team (Wang et al., 2019; Saltz and Krasteva, 2022). Widely used frameworks (e.g., the cross-industry standard process for data mining) provide a standardized approach to data science but are focusing on technical features of implementation. As we outline below, existing frameworks are limited in features such as role clarity, skills, and cross-team collaboration that are essential for developing organizational capabilities in data science. In this study, we introduce a data workflow method (DWM) as a new approach to break organizational silos and create a multi-disciplinary team to develop, implement and embed data science into an ongoing business process.

This article is developed as follows. First, we review current data science process workflows and identify a key limitation in the lack of attention to human and social capital in organizations. Based on the review, we propose a new DWM to embrace clearly defined roles, responsibilities, and a collaborative team process that go beyond organizational boundaries to integrate data science. We then introduce the operationalization of the DWM with a case study in the mining industry. Discussions of the findings highlight a whole-of-organization development pathway of organizational capability to promote data science.

1.1. Current data science process workflows

Saltz and Krasteva (2022) conducted a systematic review to explore the current approaches for executing data science projects. They found that 40% of the reviewed studies deployed a workflow approach to organize with pre-defined phases, steps, activities, and tasks. A data workflow comprises the transformational steps through which data is generated, captured, interpreted, and applied. This sequence of steps that constitute a data workflow is critically important because each step might cut across organizational boundaries within which information is traditionally managed. Indeed, a major disruptive benefit of data science innovation is the potential to connect previously disparate silos of data (Abedjan et al., 2019). Therefore, it is important to view a data workflow as a core organizational process that is managed at the system level rather than as a function of a particular project, a single business unit, or individuals.

Workflow design for data science has been discussed in recent frameworks. Knowledge discovery in databases (KDD) is a classic data science process workflow with a goal to derive useful knowledge from large data repositories (Brachman and Anand, 1996; Fayyad et al., 1996). KDD process has a strong emphasis on identifying customers' needs and understanding the application domain. However, KDD was developed in an earlier period without many of the complexities of modern data science projects, which involve multiple stakeholders and associated tasks. The cross-industry standard process for data mining (CRISP-DM) (Chapman et al., 2000) is another widely used data science workflow. CRISP-DM

has been considered a standard for most of the analytics, data mining, and data science processes (Mariscal et al., 2010; Martinez-Plumed et al., 2019). CRISP-DM provides a structured phased process, which encourages the documentation and retention of knowledge along the process. However, CRISP-DM focuses on project-based technical processes and does not provide guidance for integration with ongoing business processes. Limited integration with business processes is a key reason for failures in data science implementation (Martinez et al., 2021). Moreover, CRISP-DM includes few details for iterative development of the phases, resulting in insufficient guidance for continuous improvement of the key activities. Most importantly, neither KDD nor CRISP-DM address challenges inherent in the operation of multi-disciplinary teams and multiple roles in the data science life cycle that are more common in modern data science projects. Lack of attention to team processes and collaboration among different work groups limits the helpfulness of these workflow models in guiding efficient and productive data science innovations (Saltz, 2021).

Team Data Science Process (TDSP) launched by Microsoft in 2016 made a breakthrough in supporting collaborative teamwork. TDSP defined six team roles in a data science project lifecycle: solution architect, project manager, data engineer, data scientist, application developer, and project lead (Microsoft, 2020). It further introduced task allocation and team interactions among these roles in the data science unit. Nevertheless, TDSP considers data science to be a project-based process, completed by a data science unit alone. It did not encompass the collaboration between a data science unit and other business units such as frontline workers, subject matter experts, and senior management in the organization. Thus, TDSP omits the collective effort required across an organization in contributing to the completion and embedding of data science solutions. Such organizational boundaries can impose significant barriers that prevent a data science workflow becoming a systemic process.

More recently, new workflows have been developed as extensions to or specializations of the above three workflows, with relevant adjustments for specific use cases. However, similar limitations in team process and continuous improvement remained. For example, Costa and Aparicio (2020) integrated the importance of scheduling, roles, and tools into the original CRISP-DM steps. They proposed a POST-DS (Process Organization and Scheduling electing Tools for Data Science) with a responsibility matrix, using techniques such as RACI (Responsible, Accountable, Consulted, Informed) to identify the possible roles involved and what tools can be used in each activity. However, due to the inadequate details on clarity of those roles, POST-DS is also limited in supporting teams and networks building. On the other hand, Zhang et al. (2020) recognized the understudied perspective of collaborative practices and extended data science team by incorporating roles of nontechnical team members in their data workflow. Their study revealed role activities during the stages of a data science project and identified communication pathways of the collaborative relationships among the roles. However, it was unclear how the roles and phased stages were initially developed in their data workflow. And above all, the new six-stage data workflow still structured data science as temporary projects, with insufficient focus on iterative development and integration with ongoing business process. Table 1 presents the features of the above-mentioned data science process workflows.

1.2. The missing components: human and social capital

1.2.1. Human capital: knowledge, skills, and abilities of different roles

The complexity of data science demands effective teamwork, in which team members play different roles and combine their skills, knowledge, and abilities to collective outcomes. Unfortunately, the above data science process workflows do not adequately integrate the essential role of people (Martinez et al., 2021). Data science initiatives are built upon the creativity and innovativeness of people, so this support is critical for successful innovation (Chatterjee et al., 2022).

The concept of human capital enables a more complete and systematic view of the way people contribute to data science innovation. Human capital describes the organizational benefit that is derived from the knowledge, skills, abilities, and experience of the people in an organization (Schultz, 1961;

Table 1. Data science process workflows

Workflow	Phases	Roles	Continuous feedback loop	Limitations/Improvements in the human aspect	Limitations/Improvements in the social aspect	Use case	References
Knowledge discovery in database (KDD)	<ul style="list-style-type: none"> • Selection • Pre-processing • Transformation • Data mining • Interpretation/Evaluation 	Lack of role definition	No	KDD does not address complexities of modern data science projects with various roles and skills requirements	<ul style="list-style-type: none"> • Lack of team definition • Little guidance on connecting the team members, who are likely from different business units 	Finding valid and useful patterns in data	Fayyad et al., 1996
Cross-industry standard process for data mining (CRISP-DM)	<ul style="list-style-type: none"> • Business understanding • Data understanding • Data preparation • Modeling • Evaluation • Deployment 	Lack of role definition	Yes	CRISP-DM does not address various roles and skills requirements	<ul style="list-style-type: none"> • Lack of team definition • Little guidance on connecting the team members, who are likely from different business units 	Providing structured phases to data mining projects	Chapman et al., 2000; Saltz and Krasteva, 2022
The team data science process (TDSP)	<ul style="list-style-type: none"> • Business understanding • Data acquisition and understanding • Modeling • Deployment 	<ul style="list-style-type: none"> • Solution architect • Project manager • Data engineer 	Yes	<ul style="list-style-type: none"> • Little guidance on the skills development, assessment, and training of the roles 	<ul style="list-style-type: none"> • Inadequate attention has been paid on the collaboration among different 	Building predictive analytics solutions and intelligent applications	Microsoft, 2020

Continued

Table 1. Continued

Workflow	Phases	Roles	Continuous feedback loop	Limitations/Improvements in the human aspect	Limitations/Improvements in the social aspect	Use case	References
	<ul style="list-style-type: none"> • Customer acceptance 	<ul style="list-style-type: none"> • Data scientist • Application developer • Project lead 			<ul style="list-style-type: none"> • business units in the organization • Neither on the development of a data culture 	within the data science group	
The POST-DS (Process organization and scheduling Electing tools for data science) POST-DS	<ul style="list-style-type: none"> • Business understanding • Data understanding • Data preparation • Modeling • Evaluation • Deployment 	Roles to be identified based on specific projects, example roles: <ul style="list-style-type: none"> • Business Analyst • Data Engineer • Data Scientist • Web Designer 	No	<ul style="list-style-type: none"> • Little guidance on the skills development, assessment, and training of the roles 	<ul style="list-style-type: none"> • Lack of clarity on team collaboration process • Neither on the development of a data culture 	Identifying processes, organization, scheduling, and tools to align with the overall project management	Costa and Aparicio, 2020
A 6-stage data science workflow	<ul style="list-style-type: none"> • Understand problem and create plan • Access and clean data 	<ul style="list-style-type: none"> • Engineer/Analyst/Programmer • Communicator 	No	<ul style="list-style-type: none"> • Little guidance on the skills development, assessment, and 	<ul style="list-style-type: none"> • Introduce roles and tasks in teams, including not only technical 	Understanding and supporting the collaborative aspect of data	Zhang et al., 2020

Continued

Table 1. Continued

Workflow	Phases	Roles	Continuous feedback loop	Limitations/Improvements in the human aspect	Limitations/Improvements in the social aspect	Use case	References
	<ul style="list-style-type: none"> • Select and engineer features • Train and apply models • Evaluate model outcomes • Communicate with clients or stakeholders) 	<ul style="list-style-type: none"> • Researcher/Scientist • Manager/Executive • Domain Expert 		training of the roles	<ul style="list-style-type: none"> members, but also nontechnical members • However, not sufficient on developing a data culture and on workflow embedding 	science teamwork	
DWM	<ul style="list-style-type: none"> • Identify data opportunities • Seek data solutions • Deploy solutions • Embed solutions 	<ul style="list-style-type: none"> • Data lead • Data creator • Data custodian • Data composer • Data consumer • Data enabler 	Yes	<ul style="list-style-type: none"> • Clearly defined role clarity linked to role-based skills requirements 	<ul style="list-style-type: none"> • A multidisciplinary team with clearly identified networks • Collaborating team process contributes to an embedded data culture 	Embedding data science as a core organizational process that shapes business operating model	This study

Becker, 2009). The importance of human capital in facilitating technology innovation, more specifically, the quality of human capital, and the ability to develop and leverage were recognized as instrumental, rather than the mere possession of human capital by an organization (Danquah and Amankwah-Amoah, 2017). Extensive evidence suggests that high-skilled human capital is vital for innovation at the organization level, affecting an organization's propensity to collaborate, innovate and maintain its growth momentum (McGuirk et al., 2015; Timothy, 2022). Based on human-capital theory and empirical insights, we expect that higher level of skilled human capital is more likely to enable organizations to adapt and embed new data science applications.

A lack of skills across different roles has been recognized as a cause of failure in data science processes (Martinez et al., 2021). Very often, the skills in data science refer to skills in using analytics such as predictive analysis, data modeling, and visualization to predict the future (Chen et al., 2012). However, skills required for a data science team are multi-dimensional, including not only technical qualifications and expertise, but also a range of experience and abilities, such as coping with stress, problem-solving, collaborating and cooperating, adaptive to changes, resilience, openness, and risk awareness (Kautz et al., 2014; Halwani et al., 2021). Some studies have grouped these skills under an umbrella term "soft skill" in data science (Ismail and Abidin, 2016; da Silveira et al., 2020), even though this concept does not distinguish the range of interpersonal and nontechnical skills that are required in complex team projects. It is also unclear how broad concepts such as "soft skills" help to identify that path through which individuals contribute to an efficient data culture and a collaborative team environment. Without defining specific knowledge, skills, and abilities (KSAs) for different roles in the team, it is difficult to specify how the team could act to integrate data workflows into business operation models. Therefore, there is an increasing need to embrace role clarity and the associated KSAs in the current data science workflow approach.

1.2.2. Social capital: a data culture and collaborative team process

Social capital describes an organization's potential that is derived from the ongoing network of communication and information sharing that occurs among people and teams in an organization (Nahapiet and Ghoshal, 1998). Social capital creates an enabling context for accessing and contributing the collective knowledge and resources (Setini et al., 2020). It stimulates knowledge sharing, organizational performance, and innovation (Cofré-Bravo et al., 2019). Social capital promotes coordination and cooperation among people by boosting their willingness to engage in intensive interactions to achieve higher goals of the team and organization (Leana and Van Buren, 1999; Metz et al., 2022).

Organizational culture is defined as "the shared perceptions of organizational work practices" and it forms "the glue that holds the organization together and stimulates employees to commit to the organization and perform" (Van den Berg and Wilderom, 2004, p. 572). As a particular type of organizational culture, data culture embodies shared beliefs and behaviors with cohesive collaboration across the organization for improved business performance through data-driven insights. A positive data culture has been proposed to be key element of the social capital required for successful data science innovations (Kesari, 2021). When a data culture is embedded in an organization, it facilitates organizational data assets to be trusted and shared by multiple work groups (Waller, 2020). A positive data culture motivates employees to solve business challenges through leveraging data analytics results, in turn, enhance operational performance (Kwon et al., 2014; Arif et al., 2019). Despite its importance, building a positive data culture is rarely mentioned in the current data science process workflows.

A healthy data culture can be strengthened through collaborative team processes (Baker et al., 2006; Boyd and Crawford, 2012). Teams are the main group structure capturing the variety of interpersonal processes that occur through the work activities of individuals (Hackman, 1990; Hoegl et al., 2004). The complexity of data science applications normally requires seamless cooperation from multiple teams from different professional backgrounds (Martinez et al., 2021). Therefore, the notion of a positive data culture highlights the need to foster partnership and communication around data insights across multi-

disciplinary teams and to ensure ongoing commitment to leveraging data for informed decision making and solutions.

1.3. The DWM

The development of the DWM was inspired by the capability framework by Griffin et al. (2014), which was originally proposed to assess “fitness to operate” in the offshore oil and gas industry. In their framework, the microfoundations of capability comprise three forms of capital: organizational, social, and human capital. In combination, the three forms of capital enable an organization to successfully perform, adapt, and survive. However, we argue that previous data science frameworks address only aspects of organizational capital, with limitations on iterative development and embedding. Further, these frameworks pay almost no attention to the social and individual capital that are necessary for successful integration into overall business operations (as seen in Table 1). This study proposed the DWM intended to integrate the three enabling capitals that can support organizations to develop, implement, and embed data science.

The DWM defines a multi-disciplinary team approach in a four-stage iterative cycle to data science: identifying data opportunities, seeking data solutions, deploying solutions, and embedding solutions. To ensure the effectiveness of the DWM stages, clear role clarity was advocated with six different data roles carrying specific responsibilities. These include a data lead, who strives to seek opportunities and benefits of leveraging data for business outcomes; a data creator, who collects data from various sources across the organization; a data custodian, who manages and cleans the dataset for the workflow; a data composer, who analyses data to generate actionable insights; a data consumer, who utilizes data to perform daily tasks; and a data enabler, who promotes and oversees the data workflow within the team. Within this multi-disciplinary team arrangement, team members are required to engage with and navigate the heightened level of diversity and ever-evolving dynamics of talents. The success of the DWM relies on the skilled team members effectively performing their data roles. Thus, it is paramount to understand the diverse skill requirements linked to the data roles and how to acquire them through talent recruitment and development.

The DWM promotes a healthy data culture and collaborative team process by connecting organization members not only in the data science unit, but also in other business units. Different from current data science process workflows that consider data science as one-off projects, the DWM regards data science as an ongoing business to be integrated into the core value of the organization. It explains the systemic focus of the DWM to approach data science as a core organizational process, rather than as a function of a particular project, isolated individuals, or a single business unit. It is managed at the system level that involves multiple interconnected business units with coordinating efforts for the unified goal of continuous improvement of business operating model.

Notably, the data roles in the DWM reflect operational needs across a data workflow, and these roles might not align with the official position titles of the team members in the organization. It is possible that a data role is played by multiple members in the team, and in some cases such as in a smaller team, different data roles are handled by one team member. To better implement the DWM approach, we assessed the specific skills required for different data roles. This was done by examining an integrated data workflow within a mining operation (Figure 1).

2. Method

2.1. The study context

This article reported an in-depth case study of a machine learning data workflow in the Australian mining industry. The mining organization has a strong innovation focus of using data and integrated systems to identify and deliver new solutions. They have a history of employing communication-based automatic systems to modify traditional ways of working, and to integrate business systems, processes, and technology to maximize efficiency. The selected data workflow involved implementing the controls



Figure 1. Data workflow method.

and support process of a new machine learning model within operations of the organization. Using collected geological data, this machine learning model was used to predict plant production for each drill hole, generating insights for both long-term and short-term scheduling and execution decisions to stabilize yield in mining sites.

This case is selected because it is an exemplar of data science process being developed and implemented by a multi-disciplinary team formed by members across different business units in the organization, including the geology team, data scientist team, scheduling team, execution team, and the leadership team. Over a 2-year period, the data science process moved from a proof-of-concept stage to an embedded process workflow within the organization's operational systems. Therefore, the selected case provides an opportunity to explore team collaborations and individual expertise throughout the data science lifecycle. The development timeline of the data workflow is shown in Figure 2.

2.2. Data collection

We employed purposeful sampling (Patton, 2002) by deliberately approaching the participants involved in the selected machine learning data workflow who could provide first-hand experience in the process.

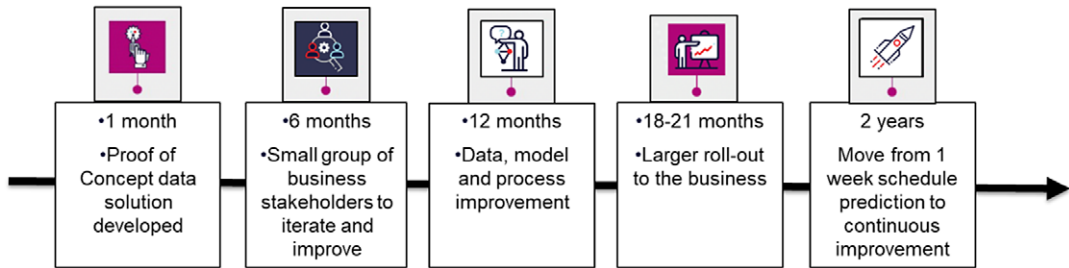


Figure 2. The development timeline of the machine learning workflow.

Table 2. Profile of the stakeholders in the case study

Pseudo	Professional background	Position title	Working experience (Year)
Adam	Geologist	Manager Geology and Scheduling	>20
Jane	Geologist	Geologist	6–10
Mary	Accountant	Manager	11–15
John	Data Scientist	Data Scientist	11–15
Betty	Data Scientist	Data Scientist	<5
Peter	Engineer	General Manager	>20
Eric	Metallurgist	Officer Crusher Delivery	11–15
David	Control Systems Engineer	Specialist Demand System	11–15
Joseph	Mining Engineer	Scheduler (Weekly)	6–10
Robert	Mining Engineer	Superintendent Scheduling and Execution	>20
Justin	Drill and Blast	Superintendent Mobile Plant Control	11–15
Charles	Dispatch	Supervisor Control	6–10
Daniel	Fixed and Mobile Plant Control	Scheduler Mine Execution	11–15
Steven	Dispatch	Supervisor Control	6–10
Donald	Surveyor, Mine Engineer	Scheduler Mine	6–10
Thomas	Geologist	Mine Geologist	6–10
Richard	Geologist	Mine Geologist	16–20
Paul	Geologist	Senior Geologist	11–15

The participants came from diverse professional background, including geologist, mining engineer, accountant, and so on. The diversity of the workflow participants represents a compelling exemplar of a multi-disciplinary team composition, including team members from not only the data science unit (only two of them), but also other business units. This aligns with the case study’s objective, which is to interview individuals not only in formal data science roles but also those outside of them to understand the connections among various roles. Table 2 shows profiles of the interviewees involved. Nearly 60% of them have more than 10 years’ of working experience.

For the purpose of this study, during the initial stages, we asked about interviewees’ general descriptions on the data workflow, their roles and responsibilities therein. Based on their responses, we probed more into whom they had been working with, the supports and barriers they encountered. Example questions include, “Please describe your role and responsibilities in this machine learning workflow,” “Who do you normally collaborate with throughout this workflow?” During later stages, interview

questions were focused more on the emerging concepts and themes from the thematic analysis, with more attention being paid to how the collaboration mechanisms were developed among different data roles and what kinds of skill and expertise were necessary to support their completion of the process. Within this general structure, we further asked for detailed illustrations through tangible examples and their observations and perceptions. We ensured a sufficient level of saturation by confirming that no significantly new information emerged from the last interviews. All the interviewees gave their consent for voluntary participation and being audio recorded. The university's ethics committee approved the research. The audio recordings were transcribed verbatim and anonymized during the transcription process. The average time for each interview is 60 minutes. In total, 18 stakeholders participated in the interviews and provided their observations in this data workflow.

2.3. Analytic approach to the case study

Transcribed interviews were organized with the software package NVivo 12. The process employed a thematic analysis approach facilitated by NVivo, wherein the authors delved into the data, identifying, analyzing, and integrating themes that emerged from the rich insights and perceptions of the participants collected during the interviews (Pratt et al., 2006). The use of NVivo12 was not focused on automated modeling, but rather on enhancing the efficiency and rigor of the manual thematic analysis process. When analyzing the interviews, data were organized continuously with the help of a work-in-progress case summary to map the process over time, as well as the functions and activities of the stakeholders involved (Patton, 2002). This assisted with compiling a storyline of the case indicating how events and dynamics unfolded.

Throughout the iterative analysis, we employed the literature on data science innovation and workflow approach to make sense of the interview data. For example, Barton and Court (2012) argued that the data science process should begin with identifying business problems and opportunities, and how data analytics could potentially enhance firm performance. Akter et al. (2016) further suggested the importance of identifying and managing risks in forming the data analytics capability to achieve successful applications. Therefore, "Understand business priorities and potentials," "Understand the potential of data analytics," and "Understand the uncertainty and risks" were used to interpret the dimensions of the interviewees' tasks at the initial stages and were further aggregated into the first phase of the data workflow: "Identify opportunities." Having sufficient domain knowledge to understand operation problems (Kim et al., 2012; Gokalp et al., 2021), data characteristics (Singh and Singh, 2019), and technical knowledge and expertise for data analytics (Gupta and George, 2016; Mikalef et al., 2019) were highlighted as critical in building data analytics solutions. Hence, these were used to convey the underpinning meaning and represent the key components at the second phase of the data workflow as "Seek solutions." Similarly, the development of higher-order themes and aggregated phases in the data workflow were all supported and consistent with previous theoretical foundations. Relevant underlying theories that supported the formation of the data structure are presented in Table 3. To ensure triangulation, the findings discussed in this study were mentioned by multiple users from diverse business units, thus representing different perspectives (Myers and Newman, 2007). As a result, throughout the machine learning process workflow studied, four different phases and six data roles stood out. Table 3 depicts the data structure with first-order, second-order and aggregated themes emerged from this study.

3. Results and insights

An overview of the data workflow in Figure 3 shows that the workflow process extends across multiple teams, roles, and activities. For example, the geology team collects and organizes the geological data from different data sources; the data scientist team uses machine learning techniques to predict the yield of ore; scheduling and execution teams work on both long-term and short-term planning based on the prediction results. To further understand the responsibilities and capabilities throughout the data workflow, we first

Table 3. Data structure of the stages in the data workflow

Task description (First-order themes)	Task dimension (Second-order themes)	Stages of workflow (Aggregated concept)	Theoretical foundations
Predicting the yield outcome is key to bring business benefits	Understand business priorities and potentials	Identify opportunities	Barton and Court, 2012; Akter et al., 2016; Gokalp et al., 2021
Machine learning and statistical algorithms could help predict the yield outcome	Understand the potential of data analytics		
Understand the possibilities of things going wrong and the limitations of the predictive approach, so as to make corresponding contingency plans	Understand the uncertainty and risks		
Have sufficient domain knowledge to understand the practical meanings and impact of model input and output	Understand the operation problem	Seek solutions	Kim et al., 2012; Gupta and George, 2016; Singh and Singh, 2019; Ngo et al., 2020; Gokalp et al., 2021
Ensure data collected are accurate and representative	Have fit-for-use data		
Build predictive model to inform operation decisions	Build prediction model		
Understand the why, how and so what of the process	Understand the entire data pipeline	Deploy solutions	Wamba et al., 2015; Akter et al., 2016; Munappy et al., 2020
Trust and recognize the potential of using data analytics	Trust the model and the process		
Provide real-world observations to build back into the model	Review and provide feedback		
Promote the positivity/acceptance of the model across business	Perpetuate a positive data culture	Embed solutions	Gupta and George, 2016; Ngo et al., 2020; Shamim et al., 2020
Communicate the long-term benefits of using data analytics	Communicate long-term business value		
Support people to adapt to changes and new ways of working and thinking	Facilitate the transition to new ways of working		

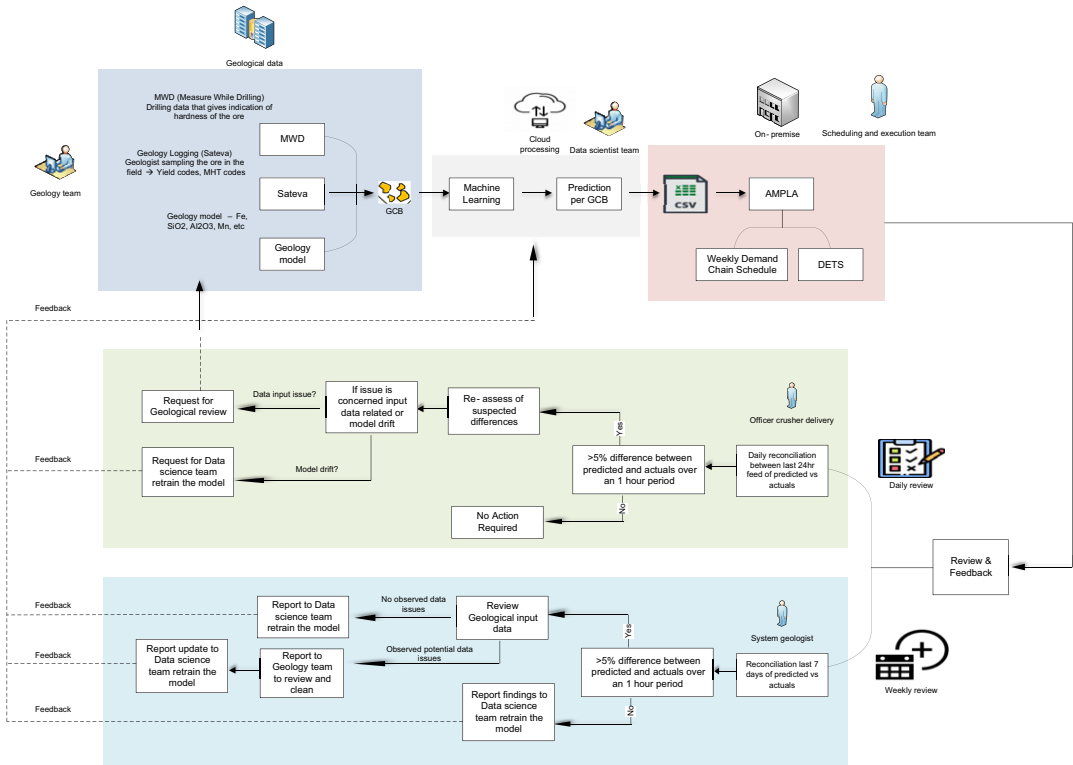


Figure 3. The process workflow of implementing a predictive machine learning model in an Australian mining company.

identify key stages based on the descriptions of the participants. Then, we analyze the roles, functions, and partnerships across different stages.

3.1. Four stages of the DWM

Based on the interview transcriptions, data analysis began by identifying descriptions of key tasks and activities conducted by workflow participants as initial coding themes. Following Gioia et al.'s (2013) approach, we developed first-order and second-order codes to better understand the key functions in different stages of the workflow. First-order themes were generated based on respondents' descriptions of their key activities. The second-order interpretations represent the essence of work in the frame of work dimensions at a more abstract level. Finally, four distinct stages of the data workflow emerged as can be seen in Table 3.

The first stage is to recognize opportunities of using data analytics to improve operational excellence and achieve business advantages. More specifically, the first stage recognized the underlying potential of using geological data to predict the yield of mine, which in turn informed weekly demand chain schedules and optimized the allocation of resources in execution and mobile control. Identifying the opportunities of using data analytics requires a deep understanding of both business priorities and how data analytics work. Rich experience in mining operations could bring forth the sensitivity to business needs. Adequate knowledge of data analytics techniques, such as how machine learning could be used to explore data patterns among massive data loads and in turn generate predictions, could elucidate the prospect of solving business challenges with data-driven science. In addition, it is also important for stakeholders in this stage to understand the risks that things might go wrong and the limitations of prediction models, so that corresponding contingency plans could be made. Therefore, identifying data opportunities is the

starting step of the data workflow, and boundary spanners with both domain expertise and technical understanding have been found critical. As mentioned by the general manager Peter during the interview:

Our process plant for all of its complexity, at the heart of it, it separates material into sizes, and anything above one millimetre in size, goes straight through to our shipped product. Anything less than one millimetre in size, goes through a bunch of processes to further segregate iron and not iron. So, the whole productivity of our business is based on managing the minus one-millimetre amount. There is a huge potential in using a machine learning algorithm to find the correlations among key variables to generate the most repeatable, predictable product outcome.

The second stage is to seek data solutions. More specifically, building machine learning models to analyze existing geological data and generating prediction results that went into the plant. Data scientists John and Betty are the key participants in this stage. In this case, the data scientist team built a machine learning model to identify the correlations between key variables of the geological data and the yield outcome, which could be further used to explain the composition of ore unit. Through frequent communications between the data scientist team and the general manager (Peter), the data scientist team provided feedback on the feasibility of the intended outcome of a more accurate yield prediction based on analyzing geological observations. This resulted in an enhanced understanding and determination of Peter to support this data initiative, and keep abreast of newest trends in data analytics and business opportunities. As John mentioned in the interview:

It is very important to understand composition of the ore unit that goes into the plant, identify if there's any relationships or any trends that we can pick up on, and inputs into the plant, and then to see if we can predict what comes out... We take machine learning to build a predictive model and operationalize it and plug it into systems and ensure the model isn't drifting and could provide stable and reliable prediction results.

As the data scientist team emphasized in the interview, the quality of the input data was critical in generating meaningful and accurate prediction results. Therefore, the geology team played a significant role in producing data solutions, as they were the ones who ensured the accuracy and representativeness of the geological datasets collected from multiple sources. As commented by a geologist:

The team (mine geologists) goes out and collects all the samples from all the blast cones they log, and then the lab sends the assays back in, and we have them in a system that are available for the data science team to look into. These data could help feed into the machine learning capability of the data science team and help to train the machine on what to look for.

The third stage in the studied data workflow is to deploy the data analytics results from the machine learning model to support operational decisions. As many participants mentioned during the interviews, having a complete understanding of the rationale, the development, and potential implications of the process would help them appreciate and implement the model prediction results for application:

Being an engineer, I need to understand the why. It is important that I understand the data that I'm using, where were they from and the purpose and the outcome of the analyses. If I cannot understand how it (the model) is created and what goes into creating it, and how frequently it needs to be updated, I can't ensure I can get it right when implementing it.

The prediction results could benefit multiple operational teams in this case, including scheduling team, execution team, and mobile plan control team. As Donald (mine scheduler) commented that:

I am responsible for scheduling the feed using data, but specifically, that figure comes into play in terms of the material handling of what we scheduled to feed. So that number is a direct representation of what we expect to see there...instead of using a generic assumed recovery percentage,

using this machine learning data, we're going down the path of trying to predict on a daily basis what kind of recovery we should expect, rather than just an average for a week...

Review and feedback were important activities in the deployment stage. Both daily and weekly reviews were conducted to ensure the studied data workflow was an iterative and continuous improvement process. Through daily and weekly reviews, officer crusher delivery (Eric) and systems geologist (Thomas) made reconciliation between the predicted results from the model and the actual situations as observed onsite. If larger than 5% differences were identified between the predicted and actuals, further reviews and examinations were conducted to either request for geological review on the accuracy of data input, or request for a data scientist team to retrain the model. By receiving more feedback on the performance and accuracy of the prediction results, the data scientist team received more information and resources to retrain the machine learning model, thus increased its predictive power. Therefore, review and feedback processes were instrumental as they not only improved the prediction model, but also enhanced the mutual understanding between the data scientist team and operational teams, as a result trust within the team was cultivated. As mentioned by Eric (officer crusher delivery) on the importance of review in deployment:

We're a part of that feedback loop and trying to feed that information back to the data scientist team to update the data, try and build back into the model to improve the accuracy of its prediction, that's continuous improvement, it's also in our best interest to try and tighten that as much as possible...I feel like that goes down to this whole trust element, it is about the moral support. When some models predictions went wrong, we feed that back to the technical team and trust it will improve the model and predictions.

The last stage is to embed the data workflow into an organizational work system for streamlined data usage in the future. Despite the benefits of this machine learning prediction model, it needed great effort and determination to embrace the changes it brought to the existing work process. Dealing with doubt, frustrations, and distrust when model predictions went wrong can be challenging, especially when the model was not ideally accurate at the beginning. Therefore, the embedding of the new machine learning model required the support from the management team in this data workflow, who facilitated the implementation and assisted their team members in trusting the process and supporting the improvement of data analytics models. As a result, the iterative data workflow was integrated into organization data strategy, and the established communication networks in the workflow broke down the barriers of organizational boundaries, and thus contributed to the transparency and overall efficiency of data management. To achieve these, managerial support played an indispensable role in enabling the acceptance of the machine learning model within business. As mentioned by Adam (Manager Geology and Scheduling):

My job is to ensure that we track live data for all our material flows from what we did to what we processed, what we crushed, and so forth...I'm to ensure that all those processes and systems, they work together harmoniously, so our team members can actually get the information needed to do the job... and once we had these new models for prediction, how do we then force it down into our execution systems so that people who are making those task based decisions have got relatively up to date data...and ultimately all the decision makers have that information available to them through whichever system is most appropriate for them.

3.2. Data roles and associated KSAs in the DWM

Six different data roles were manifested in the data workflow: data creator, data custodian, data composer, data consumer, data enabler, and data lead (see Table 4). In this section, the responsibilities, capabilities, and behaviors that support each data role were discussed with examples.

Data creator is a role to collect data for the entire data workflow. In this case of leveraging geological data, the geology team played the key role in collecting sample data from the blast cones and the drills. As they have subject matter knowledge in the geological characteristics of the mine, therefore, they are the domain experts to provide and demonstrate the meaning and context of the backgrounds behind the data for other stakeholders involved in the data workflow. As mentioned earlier, accurate and good-quality data input are the pre-requisites for achieving plausible data solutions. Therefore, it is essential that data creators ensure the accuracy, quality, and representativeness of the data by collecting from the right data sources with the appropriate collection methods. Once the data creators (geology team in this case) entered data accurately into the system, the process of leveraging data could progress. As Paul commented on the role of data creator in the interview:

The team (mine geologists) goes out and collects all the samples from all the blast cones they log, and then the lab sends the assays back in, and we have them in a system that are available for the data science team to look into. These data could help feed into the machine learning capability of the data science team and help to train the machine on what to look for... we're responsible for the quality of that data.

Data custodian is another data role that works closely with the input data, and they mainly maintain and manage the dataset to ensure the input data are cleansed and could represent the accurate meaning. Thus, they share similar responsibilities as the data creators, and they also provide the subject matter knowledge to explain the underlying geological meaning of the data to the rest of the stakeholders in the workflow. Therefore, data creator and custodian work closely together in the geology team and sometimes one could play both roles when team is small. Nevertheless, one specific aspect of the data custodian role was highlighted in the interview. Based on their holistic understanding of the underlying meaning of the geological dataset, they were required to interpret and make corrections to reconcile the accuracy of the model predictions when differences exist between the predicted model results and actuals onsite. As mentioned by Jane:

We (data custodian) generally have to interpret that data and ensure and reconcile the accuracy of the predictions. The onsite crusher delivery person technically reports through me on a daily basis or even quicker than a daily basis- that data is interrogated on its performance and to be adjusted as necessary. We then make a real-world reconciliation between how that data performed and what the actual outcomes are. For example, if we see something in the real world at the dig face of the excavator unit where it doesn't match that data, then we make corrections.

Data composer is the key player in developing data solutions. John and Betty from the data scientist team played the role of data composer in this case study. They used the input data from the geology team to develop machine-learning models for prediction. Thus, data composers ensure the model functions well and then operationalize and plug it into systems. To generate accurate prediction results that could inform operations, it is required of data composers to understand the business priorities and include necessary boundary conditions into the model with practical considerations. Having a problem mindset is therefore critical as it is expected of the data composers to come up with actionable insights to support decision-making, rather than fancy algorithms that are not realistic. In addition, to gain trust and promote the usage of the model output, it is important that data composers could demonstrate the efficiency and rationale behind the model to a nontechnical audience. This will require data composers to understand the professional background, interests, and even personality of the stakeholders who are the end users of the predictive results. As John mentioned during the interview:

My role (data composer) also involves ensuring the model outputs are understood and applied for operations. Like a salesperson, we make sure the model is not only running with relatively accurate results, but also is believed and embedded. This requires us to demonstrate the rationale to the

Table 4. Data roles, knowledge, skills, and abilities (KSAs) and demonstrated behaviors

	KSAs	Behavior requirement	Positions in the data workflow
<i>Data creator</i> Collect data from different data sources	<ol style="list-style-type: none"> 1. Know the data sources that can support this data workflow 2. Know how to enter the data accurately in the system for this data workflow 3. Know how to check the quality of the data in this data workflow 	<ul style="list-style-type: none"> • Know where to collect data • Be able to explain the data source • Manage the data inputs to be accurate and representative • Conduct weekly and monthly data validations • Conduct both people-driven and automated checks 	Mine Geologist; Senior Geologist
<i>Data custodian</i> Own, clean, and maintain good quality dataset	<ol style="list-style-type: none"> 1. Know how to check the quality of the data in this data workflow 2. Know how to clean the data for this data workflow 3. Understand the potential risks of managing the data for this data workflow 	<ul style="list-style-type: none"> • Conduct weekly and monthly data validations • Conduct both people-driven and automated checks • Check the discrepancies between actual and predicted results • Identify and remove outliers that might bias the model • Explore alternatives and make adjustments • Understand the possibilities that data could be erroneous • Be able to predict the potential risks • Instigate particular controls and responses • Have contingency plans 	Mine Geologist; Senior Geologist

Continued

Table 4. Continued

	KSAs	Behavior requirement	Positions in the data workflow
<i>Data composer</i> Analyze data to generate actionable insights	<ol style="list-style-type: none"> 1. Build models to generate predictions for this data workflow 2. Have a problem-solving mindset for this data workflow 3. Demonstrate the efficiency of the data analysis to a nontechnical audience in this data workflow 	<ul style="list-style-type: none"> • Build the minimal viable product (e.g., a simple validated insight, a functioning tool) • Be able to improve the model with additional information and feedback from the users • Figure out the problem, focus on the big picture, and explore different ways of problem-solving • Promote the simplicity in models while ensuring quality through benchmarking • Look at data continually with a more holistic approach and ensure it performs and serves the longer-range plan • Understand what is important to the audience (e.g., Active listening) • Understand people with different professions. Backgrounds might have different perspectives • Be able to explain data solutions with short and simple presentations • Communicate the limitations of the data solutions 	Data scientist

Continued

Table 4. *Continued*

	KSAs	Behavior requirement	Positions in the data workflow
<i>Data consumer</i> Utilize data to perform their daily tasks and make decisions	1. Embrace emerging technologies and techniques in this data workflow	<ul style="list-style-type: none"> • Open to new methods and techniques 	Officer crusher delivery; Scheduler (Weekly); Supervisor Control; Scheduler Mine Execution; Supervisor Control; Scheduler Mine
	2. Use the insights from this data workflow for new ways of working	<ul style="list-style-type: none"> • Use prediction results to inform new plans • Use the insights to point to future improvement directions 	
	3. Provide constructive feedback on data practices in this data workflow	<ul style="list-style-type: none"> • Communicate real-world situations and observations to the geology team and the data scientist team 	
<i>Data enabler</i> Promote the use of data for decision-making within the team	1. Enable others to use the insights from this data workflow	<ul style="list-style-type: none"> • Ensure the accessibility of up-to-date data to all the stakeholders • Communicate data vision within business 	Manager Geology and Scheduling; Manager; Specialist Demand System; Superintendent Scheduling and Execution; Superintendent Mobile Plant Control
	2. Promote collaboration on data usage across different teams in this data workflow	<ul style="list-style-type: none"> • Ensure people are engaged • Provide training and upskilling to the champions in the team • Encourage boundary-spanning positions • Organize regular meetings with different departments 	
	3. Build and sustain trust in this data workflow across the business	<ul style="list-style-type: none"> • Improve end users' understanding on the model • Explain the “why,” the process and emphasizes the potential 	

Continued

Table 4. Continued

KSAs	Behavior requirement	Positions in the data workflow
<p><i>Data lead</i> Actively seek potentials, align and adjust the data workflow with business KPIs</p>	<ol style="list-style-type: none"> 1. Focus on keeping this data workflow aligned to business KPIs 2. Have a positive influence on the organizational data culture in this data workflow 3. Promote continuous improvement of this data workflow 	<ul style="list-style-type: none"> • Reiterate and communicate the value in the long-term, rather than focusing on the shortcomings • Adjust the data workflow to get the intended output (business KPIs) • Live a working approach in using and reporting data to support decision-making • Get people involved, aligned, and motivated • Keep a positive attitude to the data workflow and avoid negativity • Commit ongoing resources to a fully embedded and systematic process • Provide feedback based on frontline inspections to improve the model <p>General Manager</p>

operations team and educate them with a deeper understanding of the model, thus being supportive in developing and improving the prediction model.

Data consumer is the role that applies the predictive modeling results for operations. In this case study, data consumers include mine scheduling team, execution team, and dispatch team. It is important that data consumers are open to embrace the emerging technologies and techniques in improving productivity. An open mindset and the enthusiasm in exploring innovative ways of production could remove the reluctance to changes and enable them to try out the predictive results from the machine learning models. When the predictive results did not align with actuals as observed onsite, it was vital that data consumers provided timely feedback to improve the machine learning model. Thus, data consumers were imperative in supporting the data workflow and the data composers with onsite observations to retrain the machine learning model with increased efficiency. Eric from the mine execution team, said well:

We (data consumer) are a part of a feedback loop. When we see the machine learning data is not quite what we would expect to see based on past experience or infield observations, we try to feed that information back to John and Betty to try and update the data (model). It's in our best interest to try and tighten that as much as possible. We do have that engagement piece there to try and get it as close as possible.

Data enablers supported and promoted the data workflow by exercising their leadership impact. They ensured the usability of data solutions by enabling the accessibility of up-to-date data to different stakeholders in the process. They played a champion role in building and sustaining trust of the predictive modeling results through communications of business value and outcomes. Their positive attitude to the imperfection of current predictive power and recognition of the potential improvement were essential in facilitating the acceptance of the new data workflow. Data enablers advocate the collaboration on data usage across different stakeholders to ensure people are engaged and supported in the new ways of working. As Justin commented during the interview:

There has been a little bit of a shift in my responsibilities, my role now is more to educate the supervisors within the mobile plant team, on how to respond and what the triggers are and the actions they can take and who they can escalate to. This involves more education and making them feel more comfortable about the new process and new standard. Whenever a business introduces a new measure, everyone starts to freak out, therefore the biggest part of my job is people management, to support the 46 members in my team.

Data lead is the key role in this case study, played by general manager Peter, who initiated the data workflow through proactive explorations of the possibilities to stabilize yield outcome. Data lead in this study had a clear understanding of business priorities and he aligned the data workflow with clearly defined business key performance indicators. The data lead guided the whole data pipeline and delivered a positive leadership influence on the organizational data culture. To ensure data and analytics remaining at the center of the organization's digital strategy, he committed and allocated organizational resources to make sure the data workflow could be embedded into the organization. By supporting this and more data workflows, data lead introduced this standardized and customized process of generating value through data into an ongoing business model and ultimately improved the maturity and capability of the organization in data usage. As Peter mentioned in the interview:

My role began from connecting a couple of dots and seeing the possibility. And then engaging with John and his team on what and how we might grab existing data and connect it together and explain how that may enable us to predict the outcome in a way that we haven't been able to do before. So, there was a level of novel interpretation of existing data to help us improve the yield outcome, productivity and deliver return of investment.

4. Discussion

Stories of failed implementation and the growing attention to data science as a business-critical capability highlight the need to consider more systemic and integrated approaches. The proposed DWM approach guides organizations to recognize, integrate, and adapt data science innovations through (a) a systemic workflow that contributes to the continuous improvement of business operation model; (b) multi-disciplinary networks of collaboration and responsibility; and (c) clearly identified roles and the associated skills and expertise. We discuss the development of the above as follows.

4.1. Building a systemic data workflow

For data science applications, innovation capability is accumulated when an effective data workflow process is embedded in routines, systems, and processes through which business operations are conducted (Saltz and Krasteva, 2022). Thus, it is important to have iterative phases of work through which data is generated, captured, interpreted, and applied. Our case study explored a data workflow in a mining operation that was implementing machine learning with geological data to inform production decisions. Insights from this well-organized and integrated operation-specific data workflow informs four stages: identifying data opportunities, data solutions, deploying solutions, and embedding solutions. It is important to apply the workflow as an iterative lifecycle that is managed at the system level to optimize data usage for continuous improvement, rather than as a function of a particular project, one single business unit, or certain individuals. In this selected case, some features of the mining industry can militate against a rapid adoption of technological data science innovations. Authors have noted that reluctance can be based on features such as conservative culture, high capital intensity, volatile market dynamics, high levels of uncertainty, and physical hazardous mining environments (Bartos, 2007). By highlighting the phase of embedding solutions in the DWM, the study emphasizes the importance of integrating the workflow into organizational digital strategy, and ultimately improving the business operation model to tackle the industry challenge of data proficiency. This systemic approach is what differentiates the studied data workflow (Figure 3) and other data science process frameworks (such as POST-DS and the 6-stage data science workflow in Table 1). This approach overcomes organizational boundaries and considers how different business units work together, ensuring optimal efficiency and alignment with the unified goal of continuous improvement for the whole organization.

4.2. Developing a multi-disciplinary team across the organization

Social capital is embodied in the collaborative team process to promote data science across the organization. In this case study of DWM, six different data roles and their partnership were identified in a mining-specific machine learning data workflow. The links identified in this case study have practical implications for coordination and management of data science applications in operations. For example, in the stage of seeking data solutions, the alliance between the geology team and the data scientist team was found indispensable. Geology team, as the data creator and custodian in this case, are the domain subject experts in analyzing geological data. They provide essential background information, such as the sources of geological data and assessment of data quality, to the data composer: the data scientist team. Domain knowledge was instrumental in improving data scientists' understanding toward the practical meaning of model inputs, and thus the limitations and implications of the model solutions. Similarly, during the deployment stage, the close collaboration between data consumers and data composers enabled informative field observations to be incorporated in refining the machine learning model, leading to improved accuracy of the prediction results for operational decision-making. Meanwhile, by being involved in the model development process, data consumers obtained more knowledge on the data science applications. This increased data consumer's understanding of the technical processes as well as close cooperation with the team, resulting in their enhanced trust and tolerance in embracing trial and error throughout the stages. Over time, active communications on the value generated by data and analytics could build a strong data culture in the organization. A positive data culture motivates and encourages team members during

setbacks and challenges through new data applications. This study highlights the importance and provides insights into building collaborative partnerships, communication networks, and a positive data culture in data workflows. As the data enabler Justin mentioned during the interview:

If everyone believes in that (the process), and it's important to show the willingness to make decisions based on it, that standing by this, then I think this positive culture cultivated could go a long way...Having this positive data culture really moves the direction in leaps and bounds when you're constantly trying to get people who are resistant to change or getting past those barricades.

4.3. Identifying data roles with appropriate KSAs

This case study of DWM also sheds light on the significance of high-skilled human capital in data science process. This study argues that to improve overall data capability, it is essential to have multiple data roles with appropriate skills, knowledge, and expertise. It is suggested that a role-based approach to talent management might be appropriate for improving overall effectiveness. The identification of six different data roles indicated that high level of innovation proficiency came from individuals fulfilling their specific roles and responsibilities. Therefore, role-based work design practices, such as increasing role clarity, providing more decision-making authority, granting more feedback, and reducing role conflict, could support employees to become more engaged and productive in the data initiatives (Parker et al., 2006; Griffin et al., 2007). Moreover, by identifying the key responsibilities and KSAs of the data roles, this study points to capability building, training, and upskilling for future workforce in mining. It is important to note that, the key KSAs requirements for different data roles are different. For example, as seen in Table 4, for data composer, the most important KSAs is to build accurate predictive models to solve operation problems. However, for data enabler, it is of the utmost importance to encourage the understanding and usage of the new data science solutions through communicating the benefits of data analytics across the business. The difference in KSAs requirements is reasonable when different professional backgrounds of the data roles are considered. This also reflects that the DWM underscores the participation of different business units and the collective efforts from the entire organization. Thus, it is important to develop separate capability development and training plans for different business units and departments. In this study, we identified the key capabilities for each data role to provide insights on the development of upskilling materials for digital talent in the mining sector and beyond. In addition, the corresponding behaviors for the key capabilities (in Table 4) can inform performance measurement and skill assessment when preparing digital workforce in mining operations. Furthermore, the KSAs and demonstrated behaviors outlined in Table 4 offer valuable insights into talent recruitment and selection. Management can use this as a guideline framework to form a multi-disciplinary team composed of high-skilled members with pertinent transferable skills for different data roles.

4.4. Limitations

Despite the contributions drawn from this case study, there are certain limitations that should be acknowledged. These limitations also point to issues of data-driven innovation within the mining context that need to be addressed by future research. A key objective of this study was to manifest a functioning data workflow in the mining operation context and how this could contribute to data value realization. This objective was achieved by investigating a data science process workflow in an Australian mining organization with a strong innovation focus and data culture. Thus, all the six different data roles in the DWM emerged with different skills, expertise, and abilities from the selected case. This represents a compelling exemplar of a multi-disciplinary team with cohesive collaborating efforts to achieve organization goals. Notably, the data roles in the DWM reflect operational needs across a data workflow, and these roles might not align with the official position titles of the team members in the organization. It is possible that a data role is played by multiple members in the team, and in some cases such as in a smaller team, different data roles are handled by one team member. In those cases, senior management support is

critical in promoting and communicating the value of data and in establishing a data culture. Senior management commitment serves as the vital link connecting employees with their organization, fostering a strong psychological attachment to the goals of the organization. When the employees are committed to the goals of the organization, they are more likely to participate in the new initiatives and applications of data-driven approaches aimed at improving business outcomes. In addition, the results of KSAs and their behavior requirements in [Table 4](#) are manifestations based on the investigation of the workflow and the participants from this case study, it should not be treated as an exhaustive list of skill requirements for the diverse data science applications.

Previous scholars have justified that a single case is acceptable if well-examined, since quality and depth of analysis are more important than the number of cases in one study (Sarker, 2021). However, comparative studies would be helpful to reveal to what extent the identified data workflow and data roles are generalizable across backgrounds and even countries and regions. Besides, the qualitative design used in this study had its limitation in quantifying the relevant effects. Thus, future quantitative approaches would be helpful to better understand the impact of well-defined data workflows on data usage in operations.

The concept of data science innovation encompasses multiple features of new technology, business processes, and management practice. The organizational capability to implement data science has received growing attention, yet its development pathway remains unclear. The findings of this study point to future research directions on unpacking data roles, communication networks, and workflows in data initiatives, and how these could collectively contribute to organizational capabilities to recognize and harness data usage and analytics.

5. Conclusions

Potential benefits notwithstanding, data science applications are often treated as a function of a particular project, one single business unit, or isolated individuals, rather than being integrated into the organization's core value and ongoing operations. We introduce a new DWM as a multi-disciplinary team approach in a four-stage iterative cycle to data science, with six different data roles carrying specific responsibilities. The DWM promotes a healthy data culture and involves multiple interconnected business units with coordinating efforts for the continuous improvement of the business operating model. The effectiveness of the DWM relies on its skilled team members with necessary knowledge, abilities, and expertise to perform their data roles. Thus, it is paramount to understand the different skill requirements linked to the data roles and, in turn, how these skills can be obtained through talent recruitment and development. This study points to a whole-of-organization strategy to build data capability that enables the digital future in mining operations and beyond.

Author contribution. Conceptualization and design: All authors; Data curation: T.B., K.L., Z.P.; Formal analysis: K.L., M.A.G.; Project administration: K.L., T.B., Z.P.; Supervision: M.A.G., M.R.H.; Writing—original draft: K.L., M.A.G.; Writing—review and editing: All authors.

Competing interest. The authors have no conflicts of interest to declare that are relevant to the content of this article.

Data availability statement. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding statement. K.L. is supported by the Australian Research Council through the Centre for Transforming Maintenance through Data Science (grant number IC180100030), funded by the Australian Government.

References

Abedjan Z, Boujemaa N, Campbell S, Casla P, Chatterjea S, Consoli S, Costa-Soria C, Czech P, Despenic M, Garattini C, Hamelinck D, Heinrich A, Kraaij W, Kustra J, Lojo A, Sanchez MM, Mayer MA, Melideo M, Menasalvas E, Aarestrup FM, Artígot EN, Petković M, Recupero DR, Gonzalez AR, Kerremans GR, Roller R, Romao M, Ruping S, Sasaki F, Spek

- W, Stojanovic N, Thoms J, Vasiljevs A, Verachtert W and Wuyts R** (2019) Data science in healthcare: Benefits, challenges and opportunities. In *Data Science for Healthcare*. Cham: Springer, pp. 3–38.
- Akter S, Wamba SF, Gunasekaran A, Dubey R and Childe SJ** (2016) How to improve firm performance using big data analytics capability and business strategy alignment?. *International Journal of Production Economics* 182, 113–131.
- Arif S, Zainudin HK and Hamid A** (2019) Influence of leadership, organizational culture, work motivation, and job satisfaction of performance principles of senior high school in Medan City. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)* 2, 239–254.
- Baker DP, Day R and Salas E** (2006) Teamwork as an essential component of high-reliability organizations. *Health Services Research* 41(4), 1576–1598.
- Barton D and Court D** (2012) Making advanced analytics work for you. *Harvard Business Review* 90, 78.
- Bartos PJ** (2007) Is mining a high-tech industry?: Investigations into innovation and productivity advance. *Resources Policy* 32(4), 149–158.
- Bean R** (2021) Why is it so hard to become a data-driven company. *Harvard Business Review Digital Articles*.
- Becker GS** (2009) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago, IL: University of Chicago Press.
- Bhardwaj A, Bhattacharjee S, Chavan A, Deshpande A, Elmore AJ, Madden S and Parameswaran AG** (2014) Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798*.
- Boyd D and Crawford K** (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5), 662–679.
- Brachman RJ and Anand T** (1996) The process of knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*. Menlo Park: The MIT Press, pp. 37–57.
- Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz TP, Shearer C and Wirth R** (2000) *CRISPDM 1.0: Step-by-Step Data Mining Guide*. Chicago: SPSS, Inc.
- Chatterjee S, Chaudhuri R and Vrontis D** (2022) Bright and dark side of knowledge management practices in firms using information systems: Examining different moderating impacts. *VINE Journal of Information and Knowledge Management Systems* 53, 880–900.
- Chen H, Chiang RH and Storey VC** (2012) Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 36, 1165–1188.
- Cheng H, Rong C, Hwang K, Wang W and Li Y** (2015) Secure big data storage and sharing scheme for cloud tenants. *China Communications* 12(6), 106–115.
- Cofré-Bravo G, Klerkx L and Engler A** (2019) Combinations of bonding, bridging, and linking social capital for farm innovation: How farmers configure different support networks. *Journal of Rural Studies* 69, 53–64.
- Costa CJ and Aparicio JT** (2020) POST-DS: A methodology to boost data science. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. Seville: IEEE, pp. 1–6.
- da Silveira CC, Marcolin CB, da Silva M and Domingos JC** (2020) What is a data scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias* 8(1), 25–39.
- Danquah M and Amankwah-Amoah J** (2017) Assessing the relationships between human capital, innovation and technology adoption: Evidence from sub-Saharan Africa. *Technological Forecasting and Social Change* 122, 24–33.
- Fayyad U, Piatetsky-Shapiro G and Smyth P** (1996) From data mining to knowledge discovery in databases. *AI Magazine* 17(3), 37.
- Gioia DA, Corley KG and Hamilton AL** (2013) Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods* 16(1), 15–31.
- Gökalp MO, Gökalp E, Kayabay K, Koçyiğit A and Eren PE** (2021) Data-driven manufacturing: An assessment model for data science maturity. *Journal of Manufacturing Systems* 60, 527–546.
- Griffin MA, Neal A and Parker SK** (2007) A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal* 50(2), 327–347.
- Griffin MA, Hodkiewicz MR, Dunster J, Kanse L, Parkes KR, Finnerty D, Cordery JL and Unsworth KL** (2014) A conceptual framework and practical guide for assessing fitness-to-operate in the offshore oil and gas industry. *Accident Analysis & Prevention* 68, 156–171.
- Gruenhagen JH and Parker R** (2020) Factors driving or impeding the diffusion and adoption of innovation in mining: A systematic review of the literature. *Resources Policy* 65, 101540.
- Gupta M and George JF** (2016) Toward the development of a big data analytics capability. *Information & Management* 53(8), 1049–1064.
- Hackman JR** (1990) *Groups that Work and Those that Don't* (No. E10 H123). San Francisco, CA: Jossey-Bass.
- Halwani MA, Amiriaee SY, Evangelopoulos N and Prybutok V** (2021) Job qualifications study for data science and big data professions. *Information Technology & People* 35, 510–525.
- Hoegl M, Weinkauff K and Gemuenden HG** (eds) (2004) Interteam coordination, project commitment, and teamwork in multiteam R&D projects: A longitudinal study. *Organization Science* 15(1), 38–55.
- Ismail NA and Abidin WZ** (2016) Data scientist skills.
- Jayapandian N and Rahman AMZ** (2017) Secure and efficient online data storage and sharing over cloud environment using probabilistic with homomorphic encryption. *Cluster Computing* 20(2), 1561–1573.

- Kautz T, Heckman JJ, Diris R, Ter Weel B and Borghans L** (2014) Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success.
- Kesari G** (2021) 4 key signals that indicate a data culture is thriving in your organization. *Forbes*.
- Kim G, Shin B and Kwon O** (2012) Investigating the value of sociomaterialism in conceptualizing IT capability of a firm. *Journal of Management Information Systems* 29(3), 327–362.
- Kwon O, Lee N and Shin B** (2014) Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management* 34(3), 387–394.
- Leana CR and Van Buren HJ** (1999) Organizational social capital and employment practices. *Academy of Management Review* 24(3), 538–555.
- Mariscal G, Marban O and Fernandez C** (2010) A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review* 25(2), 137–166.
- Martínez I, Viles E and Olaizola IG** (2021) Data science methodologies: Current challenges and future approaches. *Big Data Research* 24, 100183.
- Martínez-Plumed F, Contreras-Ochando L, Ferri C, Hernández-Orallo J, Kull M, Lachiche N, Ramírez-Quintana MJ and Flach P** (2019) CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering* 33(8), 3048–3061.
- McGuirk H, Lenihan H and Hart M** (2015) Measuring the impact of innovative human capital on small firms' propensity to innovate. *Research Policy* 44(4), 965–976.
- Metz I, Stamper CL and Ng E** (2022) Feeling included and excluded in organizations: The role of human and social capital. *Journal of Business Research* 142, 122–137.
- Microsoft Corporation** (2020) Team data science process documentation [Online]. Available at <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process> (accessed 28 April 2020).
- Mikalaf P, Boura M, Lekakos G and Krogstie J** (2019) Big data analytics capabilities and innovation: The mediating role of dynamic capabilities and moderating effect of the environment. *British Journal of Management* 30(2), 272–298.
- Myers MD and Newman M** (2007) The qualitative interview in IS research: Examining the craft. *Information and Organization* 17(1), 2–26.
- Nahapiet J and Ghoshal S** (1998) Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review* 23(2), 242–266.
- NewVantage Partners** (2021) The journey to becoming data-driven: A progress report on the state of corporate data initiatives.
- Ngo J, Hwang BG and Zhang C** (2020) Factor-based big data and predictive analytics capability assessment tool for the construction industry. *Automation in Construction* 110, 103042.
- Parker SK, Williams HM and Turner N** (2006) Modeling the antecedents of proactive behavior at work. *Journal of Applied Psychology* 91(3), 636.
- Patton** (2002) *Qualitative Research and Evaluation Methods*, 3rd Edn. Thousand Oaks, CA: Sage.
- Pratt MG, Rockmann KW and Kaufmann JB** (2006) Constructing professional identity: The role of work and identity learning cycles in the customization of identity among medical residents. *Academy of Management Journal* 49(2), 235–262.
- Qi CC** (2020) Big data management in the mining industry. *International Journal of Minerals, Metallurgy and Materials* 27(2), 131–139.
- Saltz JS** (2021) Crisp-DM for data science: Strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data)*. Orlando, FL: IEEE, pp. 2337–2344.
- Saltz JS and Krasteva I** (2022) Current approaches for executing big data science projects—A systematic literature review. *Peer J Computer Science* 8, e862.
- Schultz TW** (1961) Investment in human capital. *The American Economic Review* 51(1), 1–17.
- Setini M, Yasa NNK, Supartha IWG, Giantari IGAK and Rajiani I** (2020) The passway of women entrepreneurship: Starting from social capital with open innovation, through to knowledge sharing and innovative performance. *Journal of Open Innovation: Technology, Market, and Complexity* 6(2), 25.
- Singh NP and Singh S** (2019) Building supply chain risk resilience: Role of big data analytics in supply chain disruption mitigation. *Benchmarking: An International Journal* 26, 2318–2342.
- Shamim S, Zeng J, Khan Z and Zia NU** (2020) Big data analytics capability and decision making performance in emerging market firms: The role of contractual and relational governance mechanisms. *Technological Forecasting and Social Change* 161, 120315.
- Timothy VL** (2022) The effect of top managers' human capital on SME productivity: The mediating role of innovation. *Heliyon* 8(4), e09330.
- Van den Berg PT and Wilderom CP** (2004) Defining, measuring, and comparing organisational cultures. *Applied Psychology* 53(4), 570–582.
- Waller D** (2020) 10 steps to creating a data-driven culture. *Harvard Business Review*.
- Wamba SF, Akter S, Edwards A, Chopin G and Gnanzou D** (2015) How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics* 165, 234–246.

- Wang AY, Mittal A, Brooks C and Oney S** (2019) How data scientists use computational notebooks for real-time collaboration. *Proceedings of the ACM on Human–Computer Interaction* 3, 1–30.
- Zhang AX, Muller M and Wang D** (2020) How do data science workers collaborate? Roles, workflows, and tools. *Proceedings of the ACM on Human–Computer Interaction* 4, 1–23.

Cite this article: Li K, Griffin MA, Barker T, Prickett Z, Hodkiewicz MR, Kozman J and Chirgwin P (2023). Embedding data science innovations in organizations: a new workflow approach. *Data-Centric Engineering*, 4, e26. doi:10.1017/dce.2023.22