# The Influence of the Open-Endedness of Data on the Data Scientists' Work Practice and Occupational Identity

**by**

**Febriana Wisnuwardani**

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Business and Management

University of Warwick, Warwick Business School

December 2022

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisors: Eivor Oborn, Panos Constantinides, and Manos Gkeredakis, for their endless support, constant encouragement, and constructive suggestions during the completion of my PhD thesis. Additionally, this endeavour would be impossible without the generous support from Warwick Business School, which sponsored my research.

I also sincerely thank all staff and colleagues at Warwick Business School, especially in the ISM (Information Systems Management) research group, for the support and invaluable learning experience. My PhD journey would be only enjoyable with all my friends at the University of Warwick. Despite different time zones and locations, I also thank my interviewees for supporting my research.

Lastly, I would like to mention my parents for their moral support every step of the way. I also want to thank my partner, Kemal Maulana Kurniawan, for the emotional support, appreciation, and encouragement that keep my motivation high during the process. My achievements will not be possible without them.

## DECLARATION

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Febriana Wisnuwardani December 12, 2022

## ABSTRACT

Data scientists have emerged as the primary knowledge workers in the age of big data and AI. More research needs to focus on the actual work of data scientists in interacting with data. Data scientists are highly dependent on data, and data plays a significant role in shaping what data scientists do and who they are. The openness of data interpretation challenges data scientists to extract insights for their business clients. Therefore, this research focuses on studying the influence of the open-endedness of data on the data scientists' work practices and occupational identity. This research aims to explain (1) how data scientists navigate the open-endedness of data to extract valuable insights and (2) how the open-endedness of data shapes their occupational identity. By conducting semi-structured interviews and participant observation, this research gains two key findings. First, data scientists navigate the open-endedness of data by performing a validation process that consists of three phases: validating problems, validating data, and validating algorithms. In doing the validation, data scientists engage in the act of making judgements. Second, because of the need to constantly make judgements, there is a contradiction between the identity that data scientists enact and espouse. Data scientists espouse objectivity while enacting their subjective judgments. This research contributes to the literature about data scientists, particularly their work practice and occupational identity.



Keywords: data scientists, work practice, occupational identity, data.

## ABBREVIATIONS

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **AUC** | Area Under Curve |
| **BA** | Business Analyst |
| **BDS** | Business Data Scientists |
| **BI** | Business Intelligence |
| **CICT** | Corporate Information and Communication Technology |
| **CRISP-DM** | Cross-Industry Standard Process for Data Mining |
| **DA** | Data Analyst |
| **DE** | Data Engineer |
| **DS** | Data Scientist |
| **EDS** | Engineering Data Scientists |
| **HDS** | Healthcare Data Scientist |
| **IS** | Information Systems |
| **IT** | Information Technology |
| **ML** | Machine Learning |
| **MSE** | Mean Squared Error |
| **PDS** | Physics Data Scientists |
| **SE** | Software engineer |
| **SQL** | Structured Query Language |
| **SVP** | Senior Vice President |

# CHAPTER 1

# INTRODUCTION

## 1.1 The research motivation

The proliferation of big data utilisation among organisations indicates the increasing interest in using data to support various business activities. Organisations invest more in gathering large quantities of data, analysing data, and gaining insights or values from data to create better decisions. In order to harness the benefits of big data, many organisations hire people with the skills to analyse and extract values from big data. This trend leads to the emergence of a new type of occupation in the analytics and data science sphere, namely the data scientist. Davenport and Patil (2012) claim that the data scientist profession is the "*sexiest job in the 21st century.*" It has made the demand for data scientists explode - data scientists' employment is expected to increase by 15 per cent by 2029 (U.S. Bureau of Labor Statistics, 2021). In order to extract insights and knowledge from data, data scientists are expected to master a combination of skills from different disciplines, such as computer science, mathematics, statistics, and business (Davenport and Patil, 2012). They must apply computational and mathematical methods to draw inferences from data to create value and improve business outcomes (Hamutcu and Fayyad, 2020).

Despite the growing interest in this profession, little research has studied the work of data scientists as the professionals who extract meanings from data

and create, develop, modify and put the algorithms into use. Most data science studies have focused on implementing data science in various contexts (Waller and Fawcett, 2013; Engin and Treleaven, 2018; Spruit and Lytras, 2018), but only a few studies focus on how data science is enacted in practice. Seaver (2017) calls to shift the studies about data science to focus more on the reality of extracting valuable insights or meaning from data. The process of doing data science offers much vagueness that challenges data scientists to perform their work. Several popular industry frameworks illustrate what data scientists do to analyse patterns in data and extract values from data, for instance, the CRISP-DM (*Cross-Industry Standard Process for Data Mining*) framework. However, borrowing the term by van der Aalst (2014, pp. 11–12), there are a lot of "*unknown unknowns*" or "*things we do not know we do not know*" in the process. In other words, those frameworks cannot explain the actual work practices of data scientists to overcome the challenges of doing data science.

Data scientists are challenged by the vagueness of data in their work practice in extracting insights from data (Saltz and Grady, 2017). Data interpretation is open-ended (Monteiro and Parmiggiani, 2019), which means the work of data scientists is constantly open to changes. Several studies have started to pay more attention to this and offer insights to study vagueness in working with data. For example, some scholars have shown that data is a product of the construction of human interpretation (Lupton, 2015; Feinberg, 2017; Veel, 2018; Alaimo and Kallinikos, 2020). Humans' interpretation of the data can always be changed, which brings vagueness to the process of extracting

insights from data. The meaning of data is contextual so that the meaning can be obsolete, irrelevant, and inconsistent. In this kind of condition, data scientists need to make judgements to overcome the vagueness (Tanweer, Fiore-Gartland and Aragon, 2016; Pink *et al.*, 2018). The vagueness and uncertainty of data make the work practices of data scientists messy and hard to scrutinise (Seaver, 2017). Studying the open-endedness of data by linking it to studies about work and occupational identity can become a helpful lens in demonstrating what data scientists do to navigate the vagueness of data.

Data scientists also face another challenge regarding how they define their occupational identity. Many companies see data scientists as technical people who can reap knowledge from data to help companies make better rational decisions (Cote, 2021) however, their expected work and roles are not well-defined and standardised (Bowne-Anderson, 2018). Companies employ data scientists and create job descriptions based on their understanding of the field. Some companies think that data scientists can solve all business problems with data. Some think data scientists' tasks include all actions involved in deriving insights from data which overlaps the work of data analysts, data engineers, and BI analysts. At the same time, others think that data scientists only build machine learning and AI models. Data scientists are likewise learning their craft independently, frequently feeling that their work does not meet their expectations (Grootendorst, 2021). There is a misalignment of expectations about data scientists' jobs between data scientists and companies (Day, 2021). Data scientists face difficulties in creating and maintaining an occupational jurisdiction which makes them have unclear ideas

about the authority over their tasks (Bechky, 2003). Data scientists need to work around the legitimacy of their work. From the practitioners' point of view, hiring data scientists is not enough to successfully gain benefits from applying data science. Organisations need to understand data scientists as their talents to attract them and leverage their potential strategically (Davenport and Patil, 2012). Data scientists' occupational identity challenge hinders organisations from leveraging data scientists' full potential.

This study focuses on how the open-endedness of data can help enrich the understanding of data scientists' work and occupational identity. IS scholars are paying more attention to understanding data occupations that play an important role in doing knowledge work regarding AI adoption (Lebovitz, Lifshitz-Assaf and Levina, 2022; Waardenburg, Huysman and Sergeeva, 2022). Studying the effect of big data and AI on occupations is becoming more critical because more occupations will emerge in response to, and be affected by, those technologies. Understanding data scientists' work and occupational identity can be the step to leveraging the potential of data workers. As Ashcraft (2007) explained, understanding occupational identity is important to observe how the members of an occupation become successful in their careers. Organisations can gain insights to understand and provide better management for data scientists to motivate and retain them.

### 1.2 The research objectives and questions

Although there is a nascent field of studies focusing on data scientists' occupational identity (Avnoon 2021; Gehl 2015; Vaast and Pinsonneault 2021), there is still little attention to how data scientists navigate the openness

of data interpretation. The objective of my research is to fill this gap in understanding how the open-endedness of data influences the work practices and occupational identity of data scientists. This is an important research gap because data scientists are highly dependent on data while data plays a significant role in shaping how their work is organised (Gehl, 2015). To achieve this objective, first, I aim to unpack data scientists' work practice by examining how they interact with data, which are inherently open-ended. Then, I examine how data scientists perceive their occupational identity based on what they do in handling the open-endedness of data. Finally, my research gives a new perspective on understanding data scientists' occupations as the primary knowledge workers in the big data and AI age. To conduct this study, I develop two research questions:

1. "*How do data scientists navigate the open-endedness of data to extract valuable insights?*
2. *"How does the open-endedness of data shape data scientists' occupational identity?*

The first research question examines data scientists' work practices in overcoming the openness of data interpretation. Many people and organisations believe that data scientists are able to extract insights from data because of their technical skills. However, in this study, I capture and understand the actual practices that data scientists develop in responding to the nature of data. Then, the second research question focuses on examining how dealing with the open-endedness of data shapes data scientists' occupational identity. Because data scientists are still emerging (Kimmons and Veletsianos, 2014), data scientists are still in the phase of shaping and

negotiating their identity. Therefore, studying occupational identity is vital for data scientists' context.

This research aims to capture data scientists' experiences of working with data. Therefore, this research conducted semi-structured interviews with data scientists to ask them about what they do, the challenges of turning data into valuable insights, their expectations, and their thoughts about their identity. The interviews with data scientists are also complemented by interviews with other data occupations that work alongside data scientists for justification. This research also gathers data from an observation of various data science communities. To achieve the research objective, I adopt the interpretive approach to analyse the data to gain a richer understanding of the data scientists' perspective.

### 1.3 Thesis structure

This thesis has seven chapters. This introduction chapter covers the motivation and the rationale of my research questions. In this section, I explain the overview of the rest of the chapters that follow:

***Chapter 2 - Literature Review***

This chapter provides a review of the relevant literature of my study. I review what previous literature has studied about the data scientist occupation. Then, I explain the lens to study occupation by Anteby, Chan, and DiBenigno (2016) that I adopt as the key theoretical lens that I use to help me study data scientists' work and occupational identity. I also review what literature has said about doing data science and data scientists' occupational identity.

***Chapter 3 - Methodology***

In Chapter 3, I explain the research paradigm that I adopt to conduct my research. Then, I specify the data collection methods and explain them in detail. The explanation of the data collection methods includes the type of data that I collected, the sources, and my reflection on what I did in each method. Next, I explain the process of how I did the data analysis and the methods and approaches that I used to do the analysis. Lastly, I present the data structures that I produce from the data analysis. There are two data structures that are organised to answer both of my research questions.

***Chapter 4 - Data Scientists' Validation: A Process to Navigate the Open-Endedness of Data***

This chapter is the first analysis chapter that aims to answer the first research question. In this chapter, I focus on analysing what data scientists do to navigate data. I illustrate my findings in the form of a process model that shows the data scientists' validation process that I identify in the data. This chapter is written based on the first data structure that I presented in Chapter 3. Therefore, I present the analysis by following three themes that I develop in the analysis, namely: 1) validating process, 2) validating data, and 3) validating algorithms.

***Chapter 5 - Data Scientists' Espoused and Enacted Occupational Identity***

Chapter 5 is the second chapter of my analysis. This chapter focuses on answering the second research question about how data shapes data scientists' occupational identity. I develop three themes to explain it from my empirical data. I organise this chapter based on three themes in the second

data structure that I produced in the data analysis. The three themes are labelled as follows: 1) data scientists' espoused identity, 2) data scientists' enacted identity, and 3) managing the inherent identity tensions. I create a model that shows the paradox of the data scientists espoused and enacted identity.

***Chapter 6 - Discussion***

This chapter provides a discussion of how my findings answer the research questions by tying them with the literature. The discussion chapter allows me to develop the theorisation of my findings. In this chapter, I also explain the contributions of my research and the implications of my findings on broader topics. I organise this chapter into three main sections. First, I focus on the discussion of data scientists' validation process by abstracting this process from the literature. Second, I discuss how the tensions between data scientists espoused and enacted identity. Third, I provide a discussion about the relationship between data scientists' validation process and the emergence of their identity tension.

***Chapter 7 - Conclusion***

This chapter is the final chapter in this thesis manuscript. I provide an overview of the research that covers what I have done in the previous chapters. Then, I discuss the theoretical and practical contributions that my research offers. Lastly, I identify the limitations of my research and the recommendations for future research.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

In this chapter, I provide a critical review of the literature. The high-level focus of my research is on the influence of data on the data scientists' work practices and occupational identity. To understand this, I review studies relevant to data scientists, the process of doing data science, occupational identity, and the concepts of data. I also provide the theoretical foundations of my research. I explain which lenses I use to frame my research. In relation to data, where there are currently a number of literature streams and conversations, I specifically focus on the nature of data as being open to interpretation. This conceptualisation of data is an underpinning assumption of my research.

In the next section, I start by reviewing the literature about data scientists as an emerging occupation, particularly as data professionals. I examine how previous studies define who data scientists are and what data scientists do. This topic forms the background of the following sections in the literature review.

## 2.2 The emergence of data scientists' occupation

Data scientists have emerged as one of the most in-demand professions in the age of big data and AI. In this section, I explain what the literature has said about the emergence of data scientists as an occupation and what makes them considered important by organisations.

### *2.2.1 Data scientist as an emerging occupation*

The emergence of occupation is influenced by the forces or systems surrounding it (Abbott, 1988). According to Evans (1987, p. 627), occupation is "the active or '*doing*' process of a person engaged in goal-directed, intrinsically gratifying, and culturally appropriate activity." To study an emerging occupation, one cannot separate the occupation from the various factors that influence it. Examples of the elements or forces that influence the emergence of occupation are the surrounding economic and technological developments. Many occupations have been established to fill the workforce demand to harness the benefits of implementing technology. For example, the widespread use of television created a demand for television repairs (Barley, 1996). Because of this, the occupation of "*television repairmen*" emerged to answer the needs. Barley (1996) argues that the emergence of many occupations was tied to the commercialisation of new technology.

With the significant increase in the use of big data and algorithms, a similar situation as Barley's (1996) example is occurring today. There is increasing availability and access to large volumes of structured or unstructured data generated by systems, people, sensors, or digital traces from human activity using digital technology (Saltz and Grady, 2017). It coincides with the creation of new jobs, tasks, activities, and industries (Acemoglu and Restrepo, 2019), creating new and emerging occupations. The application of data science in organisations calls for new professionals that are expected to have the ability to find patterns and learn from increasingly large data (The Royal Society, 2019, p. 17). Many occupations have emerged to answer the demand for handling big data and enabling data-driven approaches to work (Saltz and

Grady, 2017). The data scientist emerged as one of the most well-known occupations that can harness the benefits of big data for diverse purposes.

Historically, various disciplines focus on finding valuable patterns in data - for example, data mining, knowledge extraction, information discovery, information harvesting, data archaeology, data pattern processing, and knowledge discovery (Fayyad, Piatetsky-Shapiro and Smyth, 1996). The disciplines were performed by various practitioners such as scientists, statisticians, data analysts, librarians, computer scientists, and others (Fayyad, Piatetsky-Shapiro and Smyth, 1996; The Royal Society, 2019). The term 'data science' first emerged in the 1960s to call for a new field of science which focuses on learning from data (Donoho, 2017). The term "data scientist" was first coined by DJ Patil – also known as the first data scientist in the United States – in 2011 (Chibber, 2018). Patil was the Head of Data Products on LinkedIn when the HR team asked for his help to clean up the organisational chart. Patil found that there were too many job titles using the term "data". With the suggestion of his friend on Facebook, Patil decided to create a new term to call this kind of professional "data scientist" (Chibber, 2018).

Data scientists emerge as professionals who provide valuable insights from data. According to the Data Science Association's (2020) website, a data scientist means "*a professional who uses scientific methods to liberate and create meaning from raw data*"[1]. Data scientists encompass different disciplines such as statistics, data mining, machine learning, and computer

[1] Source: https://www.datascienceassn.org/code-of-conduct.html

science (van der Aalst, 2014; George *et al.*, 2016). Provost and Fawcett (2013) emphasise that data scientists' skills need to be complemented by domain knowledge from various fields (e.g., environmental science, healthcare, law, and others) depending on the context of the problems. In this way, data scientists can support and guide organisations to answer questions with data (Agarwal and Dhar, 2014).

### *2.2.2 The importance of data scientists in organisations*

Data scientists have become attractive in the job market. Many organisations increasingly compete to hire high-quality data scientists. The growing demand for data scientists makes this profession one of the professions with the highest salary (Rzeznikiewicz, 2022). The prospective incentives also attract many graduates and professionals to become data scientists. Therefore, it is important to unpack what literature has said about what makes data scientists considered important for many modern organisations. The rise of data scientist professions is intertwined with the increasing importance and attention on data-driven decision making (Provost and Fawcett, 2013).

In the post-industrial age, most organisations desire to be more rational in making decisions. Organisations generally strive to make rational decisions to optimise efficiency, predictability, quantification, and control (Walker *et al.*, 2008). Efficiency is important to allow organisations to handle a large number of tasks with minimum costs. Predictability decreases uncertainties that enable organisations to have high confidence in knowing what and when something will happen. Quantification allows organisations to perform evaluations of the

organisations' achievements to strive for excellence. Lastly, control enables automatic functioning that reduces human judgements and replaces them with rules and structures. Nevertheless, the overarching goal of organisations' rationality is to make decisions that minimise costs and maximise profits (Simon, 1990).

With this rationality, most modern organisations optimise their decisions using data to achieve maximum utility. Data has been useful in helping organisations in making rational decisions. In the age of big data, with the advancement of computational powers and breakthroughs in mathematical methods, data offers more significant promises in helping organisations make better rational decisions to gain more economic value (Regalado, 2014). To achieve this, organisations need to restructure their organisations and hire people who can help them tap into the benefits of big data (Perrons and Jensen, 2015).

Data scientists are deemed important in helping organisations to make rational decisions to handle challenges that become more complex and novel in the age of big data. Rationality is heavily connected with decision-making, which is the scope that data scientists are expected to optimise. Organisations expect data scientists to build algorithms as agents that could make rational choices (Parkes and Wellman, 2015). Algorithms in advanced computational artefacts such as AI offer an ideal conception of rationality. AI algorithms could learn based on data to conceive perception and action based on defined goals and conditions (Parkes and Wellman, 2015). By using those algorithms, organisations hope to improve their work's efficiency, effectiveness, and

objectivity (Waardenburg, Sergeeva and Huysman, 2018). Thus, hiring data scientists is vital to help organisations - particularly ones who want to be the leaders of big data analytics - gain, improve, and maintain their competitive advantage.

### 2.3 The lenses to study data scientists' occupation

Many theoretical lenses can be used to study data scientists. Ones could adopt theoretical lenses that have been used to study other occupations. Researchers have studied occupations in various fields, such as economics, sociology, and business studies (Abbott, 1993). They studied other occupations by focusing on various things, for example, practices, strategies, and interaction between different occupations (Bechky, 2003; Greenwood, 2005), organisational dynamics, especially regarding power and identity (O'Mahoney and Sturdy, 2016), the influence of the economic system surrounding the occupation (Dedoulis and Caramanis, 2007; Stuart, 2013).

According to my research objectives (Section 1.2), my study aims to explain how the open-endedness of data influences the work practices and occupational identity of data scientists. Work practices and occupational identity are intertwined because the work practices that people do shape and are shaped by how people convey their identity (Anteby, 2010). With this understanding, in this dissertation, I draw on previous research to focus on the 'doing' and the 'being' (Ennals *et al.*, 2016) of data scientists. The 'doing' lens concentrates on understanding how occupational members carry out and engage in the practices that impact individual, occupational, and organisational results (Anteby, Chan and DiBenigno, 2016). The 'doing' lens

helps my research focus on data scientists' work practices. The ‘being’ lens focuses on understanding how occupations define their sense of identity according to their occupational roles (Ennals *et al.*, 2016). The ‘being’ lens helps to focus on studying what it means to be a data scientist for their occupational identity.

However, one of the challenges in studying emerging occupations is that the occupation does not yet have a strong occupational identity (Khan, 2022). As explained in section 1.1, data scientists still have an identity challenge because their jobs boundary are unclear and, sometimes, redundant with other occupations. The term “data scientists” frequently refers to various groups of occupations whose tasks are linked to data science (Wang *et al.*, 2019). It challenges my study in defining the scope of my research participants. Because of the diverse way of defining data scientists, following Wang *et al.* (2019), my study exclusively looks at self-described "*data scientists*.” I apply this term to study the research participants who actively practise data science and describe themselves as “*data scientists*”.

## 2.4 The work practices of data scientists

By concentrating on the doing, this section aims to review prior studies of data scientists’ work. As previously stated, the doing lens offers a lens for examining the activities carried out by data scientists that impact organisational outcomes. Studying data scientists' practices entails learning how they carry out their work. This part focuses on comprehending what has been written in the literature about conducting data science. To review the literature in a more structured way, I divided this section into several

subsections. First, I focus on how data science has been practically applied in the industry. Second, I focus on some challenges in understanding data scientists' work practices. Third, I touch on how data indicates a profound role in shaping data scientists' work practices.

### *2.4.1 The application of data science in practice*

Data science has been applied to solve various problems in different industries for different purposes. The application of data science is prevalent in the retail industry. It is used to analyse data regarding purchases, products, and consumers. For example, the use of data science to build repeat purchase recommendations that are personalised to customers, as found in Amazon.com (Bhagat *et al.*, 2018). Besides retail, data scientists are applied in other sectors. For example, data science is applied in the healthcare setting to create a patient-centric healthcare system to provide personalised services to patients (Spruit and Lytras, 2018). Data science is also helpful in improving production and manufacturing sectors by enhancing supply chain management (Waller and Fawcett, 2013). Data science application offers predictive analytics on large volumes of data to improve, i.e., forecasting and inventory management. Data science can also be applied in the transportation industry. Abo and Voisin (2014) show how data science is implemented to improve railway safety-related systems. Those examples illustrate that data science provides techniques and tools that can be applied in different contexts.

Although data science has been applied for various purposes, the way data science is performed varies. Several works of literature conceptualise data

science approaches to understand the work of data scientists. For example, Sambasivan et al. (2021) conceptualise data science work into two types: data work and model work. Data work is defined as the upstream work in the data science pipeline where practitioners collect and label data and then clean and analyse data. Model work comprises the downstream work where practitioners develop the models through model selection, model training, model evaluation, and model deployment.

Muller et al. (2019) conceptualise data scientists' five approaches to working with data: data discovery, data capture, data design, data curation, and data creation. Data discovery construes data scientists' passive stance towards receiving data as 'given' naturally by the environment. Data capture refers to the more active role of data scientists in choosing data and determining how to get the data. Data curation describes an active approach to selecting aspects of data that will be useful for a particular usage. Data design is the approach in which data is designed or produced to make it analysable. Lastly, data creation is the approach that emphasises the significance of human intervention in shaping and validating data and creating the 'ground truth'.

On the practitioners side, there are frameworks that illustrates the process of doing data science based on how various practitioners across industries do data science. There are a lot of data science process frameworks that are developed according to different needs. Table 1 shows the steps of the data science process from the previously mentioned frameworks.

*Table 1 Data science process from various frameworks*

| **Data Science Process Framework** | **Steps** |
|---|---|
| CRISP-DM | Business understanding, data understanding, data preparation, modelling, evaluation, and deployment |
| SEMMA | Sample, explore, modify, model, and assess. |
| KDD | Data selection, data preprocessing, transformation, data mining, and interpretation/evaluation. |
| OSEMN | Obtain data, scrub data, explore data, model, and interpret. |
| The Data Science Process | Ask an interesting question, get the data, explore the data, model the data, and communicate and visualise the results. |
| Lifecycle of an ML Project | Planning & project setup (define project goals, choose metrics, evaluate baselines, set up codebase), data collection & labelling (strategy, ingest, labelling), training & debugging (choose simplest, implement model, debug model, look at training/validation/test, prioritise improvement), and deploying & testing (pilot in production, testing, deployment, monitoring) |
| CD4ML | Model building, model evaluation and experimentation, model production model, testing, deployment, and monitoring and observability |
| NBDRA | Data collection, data preparation/curation, analytics, visualisation, and access. |
| EDSF | Design, modelling, execution, monitoring, and optimization. |
| Azure TDSL | Business understanding, data acquisition & understanding, deployment, and modelling |
| IBM Fundamental Methodology for Data Science | Business understanding, analytic approach, data requirements, data collection, data understanding, data preparation, modelling, evaluation, deployment, and feedback. |

In relation to knowledge work, data science bring more scientific methods to generate knowledge for organisations. Data science comprises a more structured and rational way for organisations to discover and generate knowledge. By building mathematical modelling on business problems, data scientists could generate knowledge through a scientific approach fit with the

functionalist view of knowledge - that see knowledge as a representation of "truth" (Alvesson, 2001) - which enables organisations to rely on scientific methods to generate knowledge to achieve economic goals (Kleinman and Vallas, 2001).

Despite giving more structure in extracting insights from data, doing data science has many challenges to be implemented. Many practitioners applying data science still find difficulties in performing a successful data science project because of the need to understand how to align data with their objectives (Joshi *et al.*, 2021). Aaltonen and Penttinen (2021, p. 5924) argue: "*While computer and data science can tell us how a structure can be imposed on or extracted from data, they do not explain how or why a particular way of structuring data renders them useful in an industry or organisational setting*." Unpacking what and how data scientists perform data science to align with the organisations' objectives – e.g., creating a structure on data, choosing their problems, and developing algorithms – is important to understand the data scientists' reasoning and judgements in doing their work.

### *2.4.2 The challenges in understanding data scientists' work practices*

Understanding data scientists' actual work practices still become a challenge for many scholars. Although data science is done following structured steps, there are many elements of data scientists' work that are hard to scrutinise. The opacity of algorithms exacerbates the opacity of understanding data scientists' work practices. The algorithms' opacity comes from the difficulties in understanding how algorithms learn and what they know (Dourish, 2016). In general, people know that algorithms make choices based on the

commonalities or patterns from a huge amount of training data (Dourish, 2016). However, the way algorithms arrive at the choices is still hard to explain.

Research has highlighted that data science practice involves negotiation to build trust in applying the algorithms (Passi and Jackson, 2018). Often, data science practitioners create several practices to perform a successful negotiation. For instance, data scientists renegotiate the perceived success or failure of the algorithms, leverage their intuition to justify the process and the results, and underline the importance of algorithms' results to negotiate the need to understand the 'inspectable' process of how the algorithms work (Passi and Jackson, 2018). One of the ways to exercise a successful negotiation is by gaining 'deliberative accountability', which, in the data scientists' context, means involving related actors in the organisations in assessing the algorithms' credibility to gain the algorithms' trustworthiness (Passi and Jackson, 2018).

There is a tendency to 'blackbox' some processes in algorithms' development that are hard to scrutinise. The decisions to blackbox or explain some parts of algorithms building are in the developers' hands (Innerarity, 2021). The need to explain how algorithms work depends on when knowledge becomes important and for whom. Blackboxing allows developers to concentrate on designing the properties and functions of the algorithms to reach the defined goals (Innerarity, 2021). The attention on algorithms' development becomes mostly drawn to the results. Evaluating the trustworthiness of the algorithms is based on how the algorithms could produce, most of the time, reliable results

(Durán and Jongsma, 2021). The results are measured according to several sets of metrics, for example, algorithms' accuracy.

As a consequence, there is a hiddenness in data scientists' work practices. Especially in how data scientists justify algorithms to achieve or produce specific end results (Saltz, Shamshurin and Connors, 2017). Algorithms are built based on the design choices of humans that build them, but the choices often are not easily understood by other people (Waardenburg, Sergeeva and Huysman, 2018). The complex technical calculation in developing algorithms hinders most people from scrutinising and understanding the justifications and reasoning behind algorithms building. The justifications for the choices remain hidden or are left to be understood only by a few highly specialised professionals (Dourish, 2016), such as data scientists. Therefore, there is a need to study data scientists' work practices more thoroughly to understand what they really do in extracting insights from data.

### *2.4.3 The indication of how data affects the data scientists' doing*

Data is central to data scientists' work practices. Therefore, it is necessary to connect data science with the characteristics of data and how they might affect each other to understand better data scientists' doing. Previous studies indicate that the open-endedness of data affects how people do data science. Several examples show how practitioners handle the challenges of working with data. Pink *et al.* (2018) studied data as materials and the process of data construction. The study explains that: *"data is not always accurate, complete, or fully aggregated representations of what individuals or social groups have done*" (Pink *et al.*, 2018, p. 10). In the process of data construction, data is

continuously damaged and repaired through creative improvisations. Furthermore, Pink *et al.* (2018) also emphasise that even if we consider data to be always damaged, it is only broken in a contextual way which means that data is broken if it does not function according to how it is intended to be used.

Another study by Tanweer, Fiore-Gartland and Aragon (2016) shows how data shapes data repair as a part of data scientists' routine. The study focuses on data breakdown and repair activity in the process of doing data science. They suggest that there are a lot of moments of data breakdown, which is when the data science project is stopped due to material limitations of data related to - for example - relevance, consistency, density, and redundancy of data. Those data breakdowns call for the repair that becomes routine and an opportunity for data scientists to generate new imaginations and configurations of their data sets' materiality.

The open-endedness of data interpretation offers vagueness and uncertainty to practitioners who want to utilise data. Mikalsen and Monteiro (2021) show that vagueness and uncertainty make practitioners develop work practices to make interpretations of data, namely accumulation, reframing, and prospecting. Those studies illustrate how the openness of data interpretation brings opportunities and challenges that shape how people work with data. They also provide indications for my study to examine how the open-endedness of data influences data scientists' work practices and how they define their identity according to what they do with data.

## 2.5 The occupational identity of data scientists

Another lens that I adopt in studying data scientists is the “being” lens. The being lens focuses on examining how and what data scientists define their occupational identity. In the following sections, I review previous studies about data scientists’ occupational identity and the ambiguities in data scientists’ occupational identity. Connecting occupational identity to work practices, I also review the studies of how occupational members manage ambiguities to claim the authority to perform their work. Then, I review what previous studies indicate about how the nature of data could help study data scientists’ occupational identity better.

### *2.5.1 Studies of data scientists’ occupational identity*

To study data scientists’ occupational identity, it is important to briefly review what occupational identity is. According to Skorikov and Vondracek (2011, p. 694), occupational identity refers to a person’s conscious awareness that represents the interests, goals, abilities, values and the complex and evolving structure of meanings that link one’s motivation and competencies with the career roles. Occupations influence and shape individuals’ identity according to what the persons do at work (Bechky, 2011) because it plays as the mechanism that allows people to form and express their identities.

With the increasing popularity of the data scientist profession, there is a growing interest among scholars in understanding this occupation and especially learning about their identity. From the sociological perspective, Avnoon (2021) examines data scientists’ identity and identifies tensions regarding their skills. Data scientists are expected to have a wide range of

skills to apply data science to solve problems in various contexts. The study shows that the tensions regarding skills are the occasion for data scientists to construct their identity. For example, data scientists bridge the gap between the scientists' and engineers' identity by creating a new omnivorous identity (Avnoon, 2021). The omnivorousness allows data scientists to build their sense of identity despite being expected to master multiple theories, acquire various domain knowledge, and have technical and social skills. The study also shows that data scientists develop the skill to self-learn to adapt to the rapid pace of innovation and technological development. Technology contributes to influencing data scientists' identity.

From IS perspective, Vaast and Pinsonneault (2021) bring insights into how digital technologies create tensions in data scientists' occupational identity. First, data scientists constantly face the tension between similarity and distinctiveness compared to other professions that work in similar fields and use similar technologies. There is an ambiguity in terms of their occupational boundary. The answers to "*who data scientists are?*" are not standardised. This issue is aligned with other arguments, which state that data scientists working in different companies have different definitions of who they are (Donoho, 2017). The openness of data scientists' task boundaries makes the practitioners adapt by narrowing the task boundary to specialisations. Many organisations have built strategies to gain data scientists' skills (Waller and Fawcett, 2013). Organisations form a data science team by transforming workers with backgrounds in software engineering or business analytics to become data scientists (Wang *et al.*, 2019). Because of this, there is an

unclear boundary that differentiates data scientists' occupations from other professionals (Saltz and Grady, 2017).

This phenomenon is found across the globe. In the US, The Bureau of Labor Statistics also does not come with a clear definition of the jobs and skills of data science-related occupations (Burning Glass Technologies, BHEF and IBM, 2017, p. 5). EDISON project, a European Union (EU)-funded effort to increase the number of qualified and competent data scientists across Europe, defines data scientists as professionals who "*find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of datasets and create visualisations to aid in understanding data, build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data*" (Saltz and Grady, 2017, p. 2357). Other sources define data scientists as professionals who analyse data to produce valuable insights (Matveeva, 2019) and tackle big data problems by extracting actionable knowledge from data (Song and Zhu, 2016). Data scientists are defined in a general manner which makes their work become easily redundant with other occupations.

The second point from Vaast and Pinsonneault (2021) is that data scientists are also constantly facing the tension between persistence and obsolescence due to digital technologies that are continually changing. Older studies have shown that technology change and occupational identity can shape each other. A longitudinal study by Nelson and Irwin (2014) shows how librarians redefine their occupational identity when responding to the development of

internet search technology over time. The study builds a model that illustrates the interaction between librarians' occupational identity and internet search technology. Their interaction over time creates a specific pattern of identity change and discursive responses while the internet technology changes. However, in the data scientists' context, data scientists' work relies heavily on digital technologies, which do not bring their identity into stabilisation (Vaast and Pinsonneault, 2021). Data scientists' identity continually changes in cyclic processes (Vaast and Pinsonneault, 2021). There is continued growth in their occupation, which continually shapes and reshapes their occupational identity.

Another study demonstrates how data scientists' occupation is growing by identifying many new occupational titles of data scientists that are rising. Data scientists tend to become more specialised. For example, according to the context of the projects, there are several types of data scientists, such as Healthcare Data Scientists (HDS) who work on medical-related projects, Physics Data Scientists (PDS) who focus on data science applications in physics, and Business Data Scientists (BDS) that support companies to solve business problems (Ramzan *et al.*, 2021). Other classifications of data scientists are divided according to the task specialisation. For example, Engineering Data Scientists (EDS) focus on maintaining the data quality and the data systems, Machine Learning Data Scientists focus on building machine learning models, Artificial Intelligence Data Scientists develop models to build AI systems, and generalist data scientists perform the whole work of extracting insights from data from end-to-end (Hamutcu and Fayyad, 2020; Ramzan *et al.*, 2021).

Studies about data scientists' occupational identity demonstrate the ambiguities in data scientists' occupation. The various occupational titles and tasks related to data scientists show that this emerging occupation is still finding its way to fit into organisational and market requirements. Because of the ambiguity of their occupational identity, data scientists continually face uncertainties in their work, which affect how data scientists gain authority in the work. Connecting occupational identity to work, there is a need to explore how data scientists manage the ambiguity to gain and maintain the authority in performing their work.

### *2.5.2 The need to understand how data scientists manage the ambiguous occupational identity to gain authority.*

The challenge in interpreting data might bring ambiguities to data scientists' work. According to previous studies, ambiguous identity makes it hard for professionals to secure authority in performing their work. Occupational members could gain the authority to perform the work based on how occupational members have the understanding of themselves as professionals (Brown *et al.*, 2010). By making a clear jurisdiction, occupational members could legitimate their work (Fayard, Stigliani and Bechky, 2017). Since their identity is not well-defined and constantly changing, data scientists are challenged with the ambiguity of defining their work jurisdiction, which questions the scope of their authority. The ambiguous identity makes occupational members difficult to demonstrate their competence and define their work evaluation (Alvesson, 2001). However, little is known how data

scientists manage the ambiguity to demonstrate their competence in working with data.

Despite the ambiguity in interpreting data, data scientists still receive a high level of demand to extract insights from data and trust in their recommendations. There is a need to study how data scientists manage the ambiguity in their occupational identity and gain authority in their work. In general, knowledge workers are vulnerable to ambiguous occupational identity, and they come up with various ways to manage it (Alvesson, 2001). With the advantage of the pervasiveness of big data, data scientists might have another way of managing their identity ambiguity. The ambiguity might also allow data scientists to shape and negotiate their identity in a particular way as they want it to be perceived, which legitimises their work practices (Brown *et al.*, 2010). Studying this issue brings insights on how data scientists' gain a sense of control over their work.

Previous studies provide examples of the way occupational members gain authority in their work. For example, a study shows how different occupational communities establish jurisdiction by representing their authority through workplace artefacts (Bechky, 2003). The workplace artefacts mediate interactions between the occupational communities and establish the authority over the tasks and work processes they claim. Another study shows how R&D professionals gain and maintain authority in responding to open innovation (Lifshitz-Assaf, 2018). They respond differently based on whether they want to refocus their work. When they want to refocus their work, they dismantle

their knowledge-work boundaries fully or partially to enable the exchange of external and internal knowledge. However, when the professionals do not want to refocus work and want to protect their boundaries, they fence their boundary of knowledge work.

The examples show a similar strategy for gaining and maintaining authority through control over specialisation. Ambiguous context makes professionals seek to gain a sense of control by claiming their specialist status (Alvesson, 2001). They tend to create boundaries to make certain parts of their tasks or knowledge hidden. Professionals must make a major effort to define a specialised field of activity in order to overcome the difficulties presented by ambiguity, which enable them to convince others that they have authority on issues related to their claimed field of expertise (Mallett and Wapshott, 2012). By controlling the boundary of their tasks, their work practices become legitimised, and they would be deemed experts in the area that only they know and can do.

The perspectives provided by those studies could be adopted to examine the shaping of data scientists' identity. As the key knowledge workers in the age of big data and AI, data scientists are deemed experts in working with data. Therefore, examining data scientists' occupational identity by considering the nature of data may bring another perspective to the conversation. The study can examine which of the practice areas are claimed under data scientists' control.

### *2.5.3 The indication of how data affects data scientists' "being"*

Data scientists' occupation emerges because of the rise of big data; therefore, data plays an important role in bringing both challenges and opportunities in shaping data scientists' tasks and roles. Paying attention to data scientists' interaction with data might bring a better lens to understanding data scientists' occupational identity. Researchers can examine what data scientists think about their identity, for example, based on how they overcome the vagueness and ambiguity of interpreting data and how they bring insights from data to their business clients. Putting aside how data scientists interact with data might hinder researchers from understanding the actual work of data scientists and how their identity is shaped through their interaction with the technology they need to extract insights from data.

Data is increasingly cheap and ubiquitous because of the ease of producing and sharing data through digital technology (Gehl, 2015). Recent studies have shown how the pervasive use of data to build AI and the use of AI influences identity. For example, a study examines AI's influence on auditors' occupational identity (Goto, 2021). The paper explores whether and with what roles, the interpretation of technology and institutional logic coexist and interact in the shifting identity. Goto (2021) shows that the use of AI allows for the creation of new identities of auditors in order to achieve three goals: enhancing managerialism, enhancing professionalism, and sustaining their professional legitimacy.

Another study focuses on the implication of AI to loan consultants' professional identity (Strich, Mayer and Fiedler, 2021). The study reveals that the

implementation of AI to substitute the loan consultants' decision-making limits the opportunity of the loan consultants to influence, adapt, or overrule the decision-making processes. The lack of interaction with the AI systems makes the loan consultants create another identity to redefine their roles. Endacott and Leonardi's (2021) study examines the influence of AI-based scheduling applications on users. This study shows that the applications influence users' personal and professional identities because the users use them to assist their professional activities. By studying the user's identity, the study offers insights into the factors that motivate users to contribute to improving the AI-based application.

Literature also indicates the need to study how data and data scientists influence each other. For example, by examining how big data and data scientists interact and are managed by organisations, Gehl (2015) argues that big data can be used to control data scientists. This is because organisations hire data scientists to handle the mess of big data and create the value and knowledge that organisations desire (Gehl, 2015). Data increasingly defines the work and performance of data scientists. However, only a few studies include data scientists' perspectives on how data influences them.

Data brings many questions in defining data scientists' occupational identity, for example, in terms of occupational boundaries (Black, Carlile and Repenning, 2004), the required skills (Acemoglu and Restrepo, 2019), job obsolescence (Crosby, 2002), and their autonomy in generating knowledge (Gorman and Sandefur, 2011). There are many opportunities to extend the

conceptual understanding of data scientists' occupational identity by studying data scientists' interaction with data. In short, there are a lot of ways to study data scientists' occupational identity, and there is a need to explore the effect of data on their identity to enrich our understanding of data scientists as the primary knowledge worker in the age of big data.

## 2.6 The concept of the open-endedness of data

In this section, I aim to conceptualise the nature of data that is always open to interpretation. This is the key assumption that I adopt in this research. The data that I focus on is data in digital forms. Data scientists mainly work with digital data to compute them using programming and statistical software. I review the open-endedness of digital data to understand data scientists' doing and being better. I begin by reviewing the common conceptualisation of data and how several scholars endorse the concept that sees data as an object that is constructed through human interpretation. Then, I continue by clarifying how data interpretation is open-ended.

### *2.6.1 The conceptualisation of data*

Conceptualising data cannot be separated from the data, information, and knowledge hierarchy. Data is commonly defined as a collection of simple facts that may be organised to generate information (Tuomi, 1999). Information is transformed into knowledge when it is understood, contextualised, or given meaning. To be utilised, data is organised and stored in a relational database. Based on the formal structure of relational databases, data is classified into three types: structured data, unstructured data, and semi-structured data

(Rusu *et al.*, 2013). Structured data is the type of data that conforms with the data models of the relational database. Semi-structured data is a type of structured data that does not follow the formal structure of data models associated with relational databases or other types of data tables but still includes tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Unstructured data (or unstructured information) is data that does not fit into relational tables or does not have a specified data schema. Unstructured data is often text-heavy, but it can also include data like dates, statistics, and facts.

However, the definition of data varies among IS scholars. Zins (2007) reviewed 130 definitions of data from IS literature and concluded that data definitions could be conceptualised through various approaches. Dourish and Gómez Cruz (2018) conceptualise data through the approach of datafication. They define data originally as "jottings, artefacts, feelings, and experiences" that were going through a datafication process to become data. According to Cukier and Mayer-Schönberger (2014), datafication transforms social actions into quantified data. Dourish and Gómez Cruz (2018) then highlight that the transformation process involves interpretative and imaginative work, meaning that data is inevitably subjective and possess a certain purposive role.

Other scholars also highlight the subjectivity in viewing something as data. Gherardi and Benozzo (2021, p. 3) explain that: "*data are not 'there' or 'here', waiting to be collected, observed, analysed, and interpreted […] data are interesting not (only) for their meaning but because they do something to us (they seduce, attract, disgust, affect*)." Going back to the etymology, the

conceptualisation of data is different from to “facts” and “evidence” (Gitelman, 2013). “Data” is the plural form of the Latin word “datum”, which means “dare” or “to give”, while “fact” came from the Latin word “*facere*”, which means “done, occurred, or exists”. Different from the meaning of data, “evidence” came from the Latin word “*videre*” which means “to see”. The etymology distinguishes the three words: facts are ontological, evidence is epistemological, and data is rhetorical (Gitelman, 2013, p. 18). According to its etymology, data is used to bring or give an argument. Therefore, data is never an “*isolated piece of simple facts*” (Tuomi, 1999, p. 7).

### *2.6.2 The views of data as non-objective and contentious*

Many social science research emphasises that data is non-objective and contentious. Data is not merely a simple fact. It is ingrained with the value choices of those who construct and manipulate them. Based on an ethnographic study, research shows that data is made from and exchanged through stories (Neff et al., 2017). The meaning of data is shaped through stories that help make sense of the desired outcome of using data. Data is constructed based on interpretation, which relies on the social context and emerges through conversation, negotiation, and action (Neff *et al.*, 2017).

Another research also contested the objectivity of data construction, particularly the classification of data as structured and unstructured (Aaltonen and Penttinen, 2021). They examine the contrast between structured and unstructured data as carriers of facts, drawing on the theory of digital objects. They argue that data do not have a structure but are created by the structure

that grants data the ability to reflect contextual truths. Therefore, they believe that data structure should be considered as a matter related to a contextual goal.

The use of data in data science also reflects that the practice of doing data is not objective. Algorithms are built based on the choices of the developers. The choices reflect and are affected by the developers' value, knowledge, and expertise (Lebovitz et al., 2021). Consequently, evaluating algorithms' performance becomes challenging because assessing value, knowledge, and expertise is far from objective (Lebovitz et al., 2021). In using the algorithms, algorithmic predictions need to be interpreted by 'credible' professionals with specific knowledge to understand and translate algorithms' predictions to wider audiences – known as the 'algorithm brokers' (Waardenburg, Huysman, and Sergeeva, 2022). However, due to the algorithms' opacity, the knowledge brokers use the judgements to overcome their knowledge boundary in understanding the algorithms' reasoning.

Based on the previous examples, many social science scholars acknowledge that data and – thus, the associated practices in data science – are not objective. The process of data construction is interpretive. The interpretation of data is never fixed because data always moves from one condition to another; data is open to changes according to the intervention. There is no specific end to the interpretation of data (Ribes, 2017).

### *2.6.3 The open-endedness of data*

Scholars have started to pay more attention to conceptualising the nature of data, which is always changing and how the implications that it brings. One of the conceptualisations of the nature of data is that data is "open-ended" (Monteiro and Parmiggiani, 2019). Data can constantly be expanded, deleted, amended, and modified (Aaltonen, Alaimo and Kallinikos, 2021). Therefore, there are no specific ends to data. Digital data can be transcoded; therefore, they are manipulable and mutable (Bates, Lin and Goodale, 2016). The mutability of data as digital objects makes data easily move between and are shaped by different groups of practitioners, organisations, and projects.

Regarding its materiality, some scholars argue that data are not conceptualised only as modular artefacts that can be combined following the logic of modularity; data are defined as cognitive elements (Alaimo and Kallinikos, 2020). This conceptualisation of data emerges by defining the nature of data as marks and signs of tokens that are constructed and designed by humans' interpretation to describe, index, and represent reality (Alaimo and Kallinikos, 2020). Data is a product of cognitive work; therefore, as cognitive elements, the open-endedness of data is not solely defined based on how data can be constantly modified materially. It is defined based on the openness of the interpretation or meaning of data. The reality that data represents depends on the knowledge of the people who make the interpretation and how they view the world.

In a knowledge generation context, the open-endedness of data makes room for sensemaking and knowledge creation (Aaltonen, Alaimo and Kallinikos,

2021). What people can produce from data is not merely artefacts, but insights and knowledge for learning, inferring or predicting (Alaimo and Kallinikos, 2020). Because data is open to interpretation, it can be used as a medium to carry out various purposes and claims. The knowledge created from data is influenced by humans' intentions and humans' ability to make meaning (Veel, 2018). Humans embed their bias, interpretation, and subjectivity the moment a data set is constructed (boyd and Crawford, 2012). The same data set may be labelled differently according to the purpose of the actors who are using it. The massive opportunity to harness the open-endedness of data for knowledge creation comes with consequences.

Monteiro and Parmiggiani (2019) indicated that the open-ended quality of data offers a medium for ideological and rhetorical strategies; thus, data can be political. For example, according to Raji (2020) in April 2020, the US government claimed a successful outcome of Covid handling by comparing the final number of projected death count – 100,000 people – with the original projections of 2.2 million, a condition without intervention. The government used the original projection data as a mark to index the final projection, arguing that their handling was successful. The example shows that with the open-endedness of data interpretation, different people with different intentions may utilise data differently; thus, the meaning of data is subjective. Ribes (2017) argues that, in most cases, data annotation, labelling, and sensemaking are means to be open-ended; therefore, when there are specific ends that claim data as "*hard facts*", it is an indication to consider the intention behind the data.

## 2.7 The scope of the thesis

The previous sections have covered what the literature has explained about the data scientist profession. Studying data scientists' occupations can be done using various lenses. Therefore, in this section, I am clarifying my research scope based on the theoretical lens I adopt to study the data scientist profession. Adopting the lens to study occupations based on previous studies, I focus on studying data scientists' "doing" and "being". In other words, my research aims to examine data scientists' work practices and occupational identity. I identified the gaps in the study about data scientists' work practices and occupational identity from the literature review. Very few studies considered those aspects of the data scientist profession in relation to how they interact with data. The literature has shown that the interpretation of data is limitless and open-ended, and data scientists face both challenges and opportunities in developing work practices and shaping their occupational identity. In contributing to filling in the literature gaps, my research is scoped into studying the influence of the open-endedness of data on the data scientists' work practice and the shaping of their occupational identity. The conceptualisation that data interpretation is open-ended becomes the assumption of this research to study how data scientists handle data open-endedness in their work.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

This chapter explains the methodological approach to my research. I will provide an explanation of the research paradigm, data collection, and data analysis of my study. The methodology of this research is designed to answer the research questions that have been formulated in the previous chapters. The research questions determine the suitable approach to conduct the research. I divided this chapter into three subchapters. First, I will explain the research design, which includes the research paradigm that underpins my research's epistemological position, the research approach, and the research method. Second, I will explain the data collection methods used in collecting the empirical data. I will include a deep understanding of my positionality in the field and empirical context. Third, I will explain the data analysis. This section includes an explanation of how the empirical data was analysed and the data structures that were built during the data analysis.

## 3.2 Research design

This research is designed according to the research questions. This study focuses on studying data scientists' work and occupational identity in light of the open-endedness of data. I adopt the research paradigm suitable for answering my research questions. Then, I adopt the research method to guide the data collection and data analysis stages specifically. The further explanation of my research design is as follows.

### *3.2.1 Research paradigm*

How a researcher conducts the research depends on the research paradigm that the researcher adopts. The research paradigm represents the researchers' perspective and framework for understanding the world (Denzin and Lincoln, 2018). My research studies a phenomenon that has not been studied extensively. Studying data scientists' work and identity is relatively new, especially among IS scholars. Therefore, I adopt the research paradigm that enables me to construct and understand how and why the phenomenon emerges. In this section, I will explain my research paradigm by dividing it into two subsections. First, I will explain the research epistemology to explain my philosophical position in studying the phenomenon being studied. Second, I will explain the research approach that I adopt to guide me in conducting the study.

#### *3.2.1.1 The research epistemology: Interpretivism*

For researchers in social science, choosing the epistemological position before embarking on the research journey is important. The epistemological stance of the researcher determines how the study will arrive at the knowledge production. According to Bryman (2012), there are two main streams of research epistemology in business and management studies, namely positivism and interpretivism. The two epistemological stances contrast with each other. Positivism separates the position between the researchers and the object of research (Blaikie, 2003). On the other hand, interpretivism acknowledges that the researchers' position cannot be separated from what is being researched (Creswell, 2013; Denzin and Lincoln, 2018). Interpretivism

believes that the meaning behind a phenomenon is not out there to be found but is constructed (Myers, 2017). Based on the aim of this research, I chose interpretivism as the epistemological position of this research. To understand how the open-endedness of data influences data scientists' work practice and occupational identity, I need to ask questions to data scientists and observe their activities. I need to interpret my conversation with the practitioners and the observations. Taking a positivist stance could not enable me to address the research questions of this research.

According to Bryman (2012), the goal of interpretivism is to comprehend how people behave. Interpretivism is more interested in an empathic understanding of human activity. Through an interpretive lens, the phenomena are studied as situated in a natural context in an effort to make sense of or interpret phenomena in terms of the meanings people assign to them (Denzin and Lincoln, 2018). When social scientists take an interpretative perspective, they don't just describe how people in a social group see the world. The social scientist's goal is to integrate the interpretations into a social scientific framework (Bryman, 2012). Corbin and Strauss (2015) emphasise the need for the researchers' sensitivity in taking an interpretive stance which means that researchers must keep in mind the experiences and backgrounds that provide them with the ability to understand the empirical data. The researchers impose their viewpoints on the data, but they need to divorce themselves from their own assumptions. In other words, the researchers need to be reflexive and aware of their bias.

Interpretivism is also commonly adopted in the IS field. Since the beginning of the 1990s, interpretivism has been one of the major streams of epistemological stance among IS scholars (Walsham, 1995). Some IS bodies of work that adopt interpretivism are: systems design, organisational intervention and management of IS, the social implication of IS, and Artificial Intelligence (Walsham, 1995, p. 4). My research studies the sociological aspect of data science by focusing on the data scientists' profession. To explore the social problem of IS, I need to build a comprehensive picture of the situation, for example, based on an in-depth perspective from the informants (Creswell, 2013). Therefore, I follow the interpretivism stream in situating my research in the IS studies' epistemological stance.

#### *3.2.1.2 The research approach: An inductive and qualitative study*

According to Bryman (2012), there are two main research inquiries to guide a study, namely deductive and inductive inquiry. The deductive inquiry is suitable for researchers who deduce hypotheses from existing theories and then confirm the hypotheses to revise the theory. On the other hand, inductive inquiry is suitable for researchers who build an abstraction from their observation (Bryman, 2012). Inductive inquiry moves the research from specific empirical phenomena to building concepts and theories (Locke, 2007). Researchers who make inductive inquiries gather the data first and then develop theories according to the findings. Based on the research questions, the latter predominantly fits my research. I do not aim to test or falsify existing theories. I aim to answer new questions about a phenomenon of interest instead; therefore, an inductive inquiry is more suitable for my research

(Locke, 2007). Inductive inquiry allows me to answer the “how” questions of the phenomenon of interest and make a theoretical explanation out of it (Woiceshyn and Daellenbach, 2018).

In adopting the inductive inquiry, I conduct qualitative research. Qualitative research in social science is a type of research that places researchers in the world to explore social problems (Creswell, 2013; Denzin and Lincoln, 2018). Qualitative researchers use field notes, interviews, conversations, photographs, and memos as the representation of the phenomena and study them in their natural setting. Understanding how things happen can be done better with qualitative study (Corbin and Strauss, 2015). I aim to study how the open-endedness of data influences data scientists’ work and identity. This kind of study is better to be approached by gathering qualitative data, for example, by asking data scientists’ opinions about who they are and what they do. Moreover, previous studies about work and identity were mostly done by studying individuals in their natural settings (Alvesson, 1994; Ashcraft, 2007). Several IS scholars who study work (Orlikowski, 2000; Bailey, Leonardi and Barley, 2011; Leonardi, 2011) and identity (Vaast and Pinsonneault, 2021) also conducted the research in the participants’ natural settings. Following the previous studies, I conduct my study by analysing and interpreting data scientists’ opinions and conversations by asking data scientists and observing their activities in their community.

### *3.2.2 The research method*

After choosing the approach of inquiry, I specified the method to guide my data collection and analysis. To answer my research questions, I adopted a

grounded theory approach for data collection and data analysis. Historically, grounded theory was developed by Glaser and Strauss (1967) for the purpose of building theory, an abstract analytical schema of a phenomenon, from empirical data (Creswell, 2013; Corbin and Strauss, 2015). Ideally, grounded theory researchers do not enter the empirical field with preconceptions or pre-formulated hypotheses (Urquhart, Lehmann and Myers, 2010). When I began to conduct the study, I did not have any theoretical hypotheses, and I did not aim to verify or falsify any theories. I wanted to enhance the theory on how big data and AI affected occupations. However, I already had preconceived theoretical ideas about how AI and big data could disrupt occupations, so my research is not a pure form of grounded theory research. Nevertheless, my study is still relevant to the grounded theory approach because my research aims to enhance theories from a new phenomenon.

According to Charmaz (2014), grounded theory is a research method that consists of systematic yet adaptable methods for gathering and analysing qualitative data for the creation of theories directly from the data. By adopting a grounded theory approach, my research starts with gathering data relevant to my initial research questions by interviewing professionals who work in the field of data analysis and AI. Then, I perform the coding and memoing and constant comparative analysis. I used comparison approaches and iterative tactics of switching back and forth between data and analysis. The data collection and analysis were done simultaneously. I followed these steps iteratively until I developed the findings into an abstraction. Adopting a grounded theory approach, particularly in data collection and analysis, gives

me the guidelines to refine my research focus according to what I found in the empirical data. The details of the data analysis step that I did in this research are explained in Section 3.5.

### 3.3 The reflection of the researcher's position

In conducting qualitative research, my background as a researcher influences my data collection. There is a need to reflect on my background and my position in the research.

My educational background ranges from industrial engineering to information systems management. To some extent, my educational background helps me understand the technical concepts or terms that I gained from the interview and observation, for example, the software that data scientists commonly use.

My professional experiences also give me insights and perspective into studying data scientists' work. I have never worked as a data scientist before, but I worked as a digital marketing staff in an IT start-up company. I worked closely with the data team consisting of data scientists, analysts, and engineers. This experience gives me practical knowledge about working with data scientists and what data scientists do. So, I could imagine the participants' stories about their work by reflecting on my experience. I also worked as a management consultant handling IS strategies. This experience gives me the perspective from the management side about the increasing demand for data scientists for incumbent organisations who undergo a digital transformation and aim to harness benefits from data.

My background puts me in a position to see data scientists' work from a business and management point of view. The way I interpret the empirical data is biased toward managerial interests in understanding data scientists' work and identity.

## 3.4 Data collection

The data collection was conducted by doing semi-structured interviews, an online participant observation, and gathering additional secondary data from various sources. In the following sections, I will explain and reflect on each data collection technique I used in this research. I will also explain the data source and the justifications that I made to gain relevant data.

### *3.4.1 Semi-structured interviews*

I conducted 49 semi-structured interviews with data professionals across different industry fields. There are several interviews that are repeated with the same person. The interview consists of 23 interviews with 17 data scientists. The rest of the interviews are conducted with other data professionals that work alongside data scientists, for example, data analysts, business intelligence analysts, business analysts, AI consultants, and product managers. All interview participants are Indonesian and work in business organisations, except one participant who represents a data science community. Table 2 provides the details of the interview, including the occupational title, participants' identification, participant's industry field interview setting, and total interviews per occupation.

*Table 2 Research interview details*

| Occupational Title | ID | Industry Field | Company size | Years of Working* | Interview Setting | Number of Interview | Total Interviews per Occupation |
|---|---|---|---|---|---|---|---|
| Data Scientist | DS1 | IT and Transportation | Medium (1,001-5,000 employees) | 2 | Online | 1 | |
| | DS2 | IT and Recruitment Service | Small (51-200 employees) | 3 | Face-to-face and online | 3 | |
| | DS3 | IT and Recruitment Service | Small (51-200 employees) | 1 | Face-to-face | 1 | |
| | DS4 | IT and Hospitality | Small (201-500 employees) | 4 | Face-to-face | 1 | |
| | DS5 | IT and Transportation | Medium (1,001-5,000 employees) | 3 | Face-to-face | 1 | |
| | DS6 | IT and Legal Service | Small (51-200 employees) | 2 | Online | 2 | |
| | DS7 | News and Media | Large (10,001+ employees) | 4 | Online | 1 | |
| | DS8 | IT and Transportation | Medium (1,001-5,000 employees) | 1 | Online | 1 | |
| | DS9 | E-commerce | Medium (1,001-5,000 employees) | 4 | Online | 1 | |
| | DS10 | News and Media | Large (10,001+ employees) | 3 | Online | 1 | |
| | DS11 | IT Services and IT Consulting | Small (11-50 employees) | 3 | Online | 2 | |
| | DS12 | IT Services and IT Consulting | Large (10,000+) | 4 | Online | 1 | |
| | DS13 | Finance and Banking | Large (10,000+) | 2 | Online | 2 | |
| | DS14 | E-commerce | Medium (1,001-5,000 employees) | 2 | Online | 2 | |
| | DS15 | E-commerce | Small (51-200 employees) | 2 | Online | 1 | |
| | DS16 | E-commerce | Medium (1,001-5,000 employees) | 3 | Online | 1 | 23 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | DS17 | E-commerce | Medium (1,001-5,000 employees) | 3 | Face-to-face | 1 | |
| Data Analyst | DA1 | IT and Hospitality | Medium (1,001-5,000 employees) | 4 | Face-to-face | 1 | |
| | DA2 | E-commerce | Medium (1,001-5,000 employees) | 4 | Face-to-face | 1 | |
| | DA3 | IT Services and IT Consulting | Small (51-200 employees) | 3 | Online | 1 | |
| | DA4 | IT and Transportation | Medium (1,001-5,000 employees) | 3 | Online | 1 | |
| | DA5 | IT and Transportation | Medium (1,001-5,000 employees) | 5 | Face-to-face | 1 | 5 |
| Business Intelligence Analyst | BI1 | IT and Transportation | Medium (1,001-5,000 employees) | 3 | Face-to-face | 1 | 1 |
| Business Analyst | BA1 | IT and Transportation | Medium (1,001-5,000 employees) | 3 | Face-to-face | 1 | 1 |
| AI Consultant | AIC1 | IT Services and IT Consulting | Small (11-50 employees) | 1 | Face-to-face | 1 | 1 |
| AI Research Scientist | ARS1 | IT Services and IT Consulting | Small (51-200 employees) | 1 | Online | 1 | 1 |
| Software Engineer | SE1 | E-commerce | Medium (1,001-5,000 employees) | 4 | Face-to-face | 1 | 1 |
| Product Manager | PM1 | IT and Transportation | Medium (1,001-5,000 employees) | 4 | Online | 2 | |
| | PM2 | IT and Transportation | Medium (1,001-5,000 employees) | 4 | Face-to-face | 1 | 3 |
| SVP of CICT | CICT 1 | Oil and Gas | Large (10,000+) | 20+ | Online | 1 | 1 |
| Digital Marketing | DM1 | Oil and Gas | Large (10,000+) | 10+ | Face-to-face | 1 | |
| | DM2 | Oil and Gas | Large (10,000+) | 10+ | Face-to-face | 2 | |
| | DM3 | Oil and Gas | Large (10,000+) | 10+ | Face-to-face | 2 | 5 |
| IT Manager | ITM1 | Oil and Gas | Large (10,000+) | 10+ | Face-to-face | 1 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| IT Asst. Manager | ITAM 1 | Oil and Gas | Large (10,000+) | 10+ | Face-to-face | 1 | 2 |
| | ITAM 2 | Oil and Gas | Large (10,000+) | 10+ | Face-to-face | 1 | |
| Business Growth Manager | BGM 1 | IT and Transportation | Medium (1,001-5,000 employees) | 3 | Face-to-face | 1 | 1 |
| Data Science Community Manager | DSC M1 | IT Community | Medium (1,001-5,000 members) | 3 | Online | 1 | 1 |
| Human Resource Manager | HR1 | E-commerce | Medium (1,001-5,000 employees) | 2 | Online | 1 | 2 |
| | HR2 | E-commerce | Medium (1,001-5,000 employees) | 4 | Online | 1 | |
| | | | | | | 49 | |

*Years of working are calculated at the time of the interview.

I conducted interviews with professionals who work alongside data scientists to justify data scientists' views and opinions from external perspectives. Some of the interviews were done face-to-face by visiting the participants' office building in Indonesia. Some participants gave me access to enter the office building and conducted the interview in their meeting rooms. Therefore, the face-to-face interviews give me the nuance of the participants' workplace. However, due to Covid-19 restrictions, most of the interviews were done virtually using video conference platforms (i.e., Microsoft Teams). The interviews lasted from 30 minutes to 2 hours. I also send the "*Participant Information Leaflet*" and the consent form before doing the interviews, so the participants can read the information regarding the research ethics, e.g., the purpose of the research, how the data is recorded, how the data is stored, and how the interviews will be pseudonymised. The example of the participant information leaflet and the consent form is attached in Appendix B and

Appendix C, respectively. The interviews are audio-recorded using my recorder device based on the participants' consent. Most interviews are done in Indonesian, with one in English. I transcribed and translated each interview manually.

The participant selection was made through the snowballing approach. I began the interviews with the respondents that I knew, and then I asked them to introduce me to other potential respondents. The interviews started with several different professionals, such as data scientists, data analysts, an AI consultant, and an SVP of CICT. In the early phase of the interviews, I asked several general and open-ended questions about their perspectives on big data utilisations, the tools and techniques that they use (e.g., data analytics, machine learning, and artificial intelligence), and the example of the application on their organisations. Then, the questions are generally followed by questions about the challenges and opportunities that they see from utilising big data. Some examples of the starting questions that I asked are shown in Figure 1; further examples are shown in Appendix A.

**General questions:**

1. What is your occupational title?
2. Can you explain your day-to-day responsibilities?
3. Who do you usually work with?
4. Have you ever had a project or work using data? If so, can you explain the project?

*Figure 1 The example of the general questions*

After asking the general questions, I adapted the thematic interview questions depending on who the respondent was. For example, if the respondent is a data scientist, I will ask further questions about their experience and views of the data scientist's profession. For example, the followed-up questions for data scientists are shown in Figure 2. If the respondent is not a data scientist, then I will also ask them to explain an example of a project that they have worked on with a data scientist

**Follow-up questions:**

1. In your opinion, what are the skills that a data scientist should have?
2. What is the difference between data scientists and other data professionals (e.g., data analysts and data engineers)?
3. Could you tell me about your journey in becoming a data scientist?
4. Could you explain how you learned to become a data scientist and the learning resources that you used?
5. What are the challenges that you encounter when working as a data scientist?
6. What is your opinion about data science implementation in Indonesia?

*Figure 2 The example of the follow-up questions*

From those interviews, I found an interesting phenomenon about data scientists' occupations. There is a growing interest in this occupation among organisations and professionals. The SVP of CICT and other occupations in the managerial role see data scientists as strategic occupations that organisations need to hire to harness the benefits of big data. Then, some technical professionals that I interview see data scientists as promising occupations in the age of big data. Some data scientists also think that their skills will become more important for strategic applications. However, some

data scientists think that there are some ambiguities and confusion about who data scientists are and what data scientists do. Data scientists often found mismatches between their expectations about their work and the organisation's expectations of them. The impression that I got from the interview is that many data scientists were still trying to figure out their identity.

The confusion and ambiguity about data scientists' identity led me to narrow down my interview to investigate data scientists' occupational identity. The decision to narrow down the research focus to occupational identity is made through theoretical sampling. I analysed the pattern of codes in the interview data and explicated them with potential theoretical lenses. The interview is done simultaneously with data analysis. By using the semi-structured interview, I could modify the questions according to the justification of the research focus. As shown in Table 2, I interviewed several respondents more than once if I needed to ask follow-up questions to clarify and gain further information for the data analysis. Following Alvesson, Ashcraft and Thomas (2008), I use semi-structured interviews to understand the identity construction to draw attention to the relationship between "what one does" and "who one is." The semi-structured interviews can help to gain the data scientists' narrative about their doing and their being.

### *3.4.2 Online participant observation*

Interviews are useful in studying identity to get explicit quotes describing the perspective of data scientists' identity. Through interviews, I can ask data scientists to explain what they think about themselves. However, interviews have some limitations in studying identity; the participants' answers about

identity during interviews may not comprehensively explain who they are. Therefore, I also complemented the interviews with empirical data from online observation in two data science communities in Indonesia. I pseudonymised the name of both communities as *Data Professionals Community* (DPC) and *Artificial Intelligence Community* (AIC). I observed data scientists' perspectives and how they speak about their work and their profession in their activities.

During the observation, I attended and participated in some community activities. The communities usually had face-to-face gatherings for seminars or sharing sessions. However, due to Covid-19 restrictions, all the activities were done online via video conference platforms. Therefore, I can only do an online participant observation of the communities. The communities held several public webinars to talk about data science implementation and development in Indonesia. I attended and took notes on nine webinars that were conducted by the communities. In addition, I also attended and participated in data science webinars held by a technology company in Jakarta, Indonesia. The details of the number of webinars attended, and the organisers are shown in Table 3.

*Table 2 Number of webinars attended and the organisers.*

| **Organisers** | **Number of Webinars Attended** | **Observation Hours in Webinars** |
|---|---|---|
| Data Professionals Community (DPC) | 6 | 10 |
| Artificial Intelligence Community (AIC) | 3 | 4.5 |
| A technology company | 1 | 5 |
| **Total** | **10 Webinars** | **19.5 hours** |

The webinars that I attended and observed comprise various topics. The examples of the title of the webinars that I observed are “Data Team Work-Life during Quarantine”, “The Identification of Human’s Gender-based on Caninus Dental Panoramic using Backpropagation Algorithm”, “Cloud, Data Privacy & Security, Open Data”, and “Building Data Team for Digital Transformation Journey”. The topics range from sharing the technical implementation of data science in practice and research and insights about working as data professionals in organisations. The webinars gave me insights and nuances about what topics data scientists and other data professionals are interested in. Also, by looking at the speakers, I gained information about the prominent data professionals in Indonesia. I could also see the people that I could contact to do interviews. Some of my interview respondents are the people in the communities that I saw in the webinars.

During the webinars, I took notes on the seminar's key points. I also took notes of the number of people attending. However, the main thing that I paid attention to is the discussion between the webinar speakers and the audience during the question-and-answer session. Sometimes I also participated in the discussion by asking questions relevant to the seminar topics. By asking questions, I gained experience in participating in community activities. I took notes of the discussion to see what topic the audience thought was important. For example, I found an interesting insight from a webinar observation. The webinar is about “Data Team Work-Life during Quarantine”. The webinar is about how data professionals manage their work at home, coordinate with colleagues remotely, use collaboration tools, and also balance work and life. There were 107 participants that joined the webinar, and there were 53

questions on the *sli.do*. However, most of the questions were not relevant to the topic. There were only two questions that were relevant to the topic, which asked what kind of innovation the data team could help during the pandemic and confirmed whether working from home is busier for the data team. The rest of the questions are about how the speakers decided to work as data professionals, how to start learning data science, the differences between data scientists, data analysts, business analysts, business intelligence analysts, and market analysts, and other similar questions around getting to know data occupations better. The questions reflect the interest of people who join this community despite the topic being brought up in the webinar. Many of the audiences are interested to know more about the occupations related to data, the differences, and what these professionals do.

In addition to webinars, the communities also post several podcasts and one radio talk. The communities regularly post podcasts to talk about relevant issues among data professionals, for example, the development of technical methods and the career issues and conditions in Indonesia. The podcasts discuss general topics related to the use of data. Some podcasts are also produced in alignment with the current issues in Indonesia, such as a podcast with the title “How Data Perform in Quick Count/ Survey” produced during the general election in the country. Also, two podcasts – “Learn to be a Data Analyst (from home)” and “Telemedicine: Healthcare from a Distance in the New Normal” – were published during the pandemic. Most of the episodes discuss technical or practical aspects of utilising data. However, there are some episodes that bring managerial issues, such as “How to Grow the Data Science Community in the East Java region” and “How to Deal with the Data

Scientist Recruitment Process". I also listened to and took notes on one radio talk show participated by one of the communities discussing "Why do data practitioners in Indonesia have high salary standards?"

*Table 3 Period of observations in each community*

| **Community** | **Platform** | **Period** |
|---|---|---|
| Data Professionals Community (DPC) | Group chat | January 1st, 2020 - June 11th 2020 |
| Artificial Intelligence Community (AIC) | Google group | February 25th, 2020 - April 20th, 2020 |

At the beginning of the data collection, I gained access to the group chat and the google groups of the communities. I did the observation and note-taking for certain periods, as shown in Table 4. I did the observation for six months in the DPC group chat. I ended the observation when I thought I did not find any new topics in the conversation. However, for AIC, the google groups are not used for conversations. It is mainly used to broadcast events. Therefore, I decided to end the observation two months after I joined because there were few conversations that I could observe.

For DPC, I could see the conversations and all the content that is shared by the members in the group chat. The interaction in the group chat is continuous and active. The group chat is the main platform for all members to communicate. It is an active group chat with 1,579 registered members. Group chat is mainly used as a platform for knowledge sharing. One of the ways of knowledge sharing is through questions and answers. The members give relatively fast responses to answer the questions, even though there are also

some questions that are left unanswered. Some questions also lead to discussions that allow members to continue the conversations privately. In that way, members can talk to each other personally despite the geographical differences. There are various topics for the questions and answers. Members are asking for recommendations and experiences. However, the topics are not categorised because all conversations happen within a single chat room.

Besides questions and answers, the members also voluntarily share any information that might be relevant to other members. Members often voluntarily share job vacancies that are not only aimed at data scientists but also at data engineers, data analysts, software engineers, business analysts, and other occupations. Besides job vacancies, members are also keen to share learning sources such as seminars, summer school programs, boot camps, workshops, programming language libraries, article links, other learning channels, and others. Some posts were posted without any context or messages. Some posts also had responses, but others did not. But members keep sharing these kinds of posts even though no one has responded to their posts.

Doing an online participant observation of data science communities is beneficial in giving me general conversations among data professionals. By participating in their communities, I gained experiences in their activities in sharing and exchanging knowledge about data science development and issues. The notes that I took from the observations can complement the interview data. For example, I can validate the occupational ambiguity issues that data scientists face by comparing the interview data and the topic that the

communities discuss. They speak similarly about the unclear definition and the boundaries about who data scientists are and what data scientists do. However, observing communities also have limitations. I could not see how data scientists perform their work in practice. In the future, the data collection can be improved by doing a participant observation study to shadow data scientists when they perform their work.

### *3.4.3 Additional documents*

*Table 4 Additional documents*

| **Documents** | **Sources** | **Quantity** |
|---|---|---|
| Brochure of data science course | A data science learning institute | 1 |
| Copy of webinar slides | DPC and AIC | 5 |
| Copy of data professionals research presentation | DPC | 1 |

To complement the online participant observation, I also analysed several additional documents (Table 5). I collected a brochure of a data science learning course from an institute, a copy of the webinar slides, and a copy of a research presentation about data talent in Indonesia conducted by one of the data science communities. The data science learning course brochures gave me information about the expected skills to become a data scientist and the learning sources. The copy of the webinar slides provides documentation of the webinar materials that complement my notes during observation in the webinar. Lastly, the copy of the research presentation about data talent in Indonesia provides additional data about the condition of data professionals in Indonesia and how the communities tried to map and understand the issues.

## 3.5 Data analysis

There are several techniques I did to analyse the interview and observation data using the grounded theory approach. First, I did the transcription and translation of the interview data manually by myself. Then, I did the coding iteratively on the data. Simultaneously with the coding, I did memo writing of the interview transcriptions and the notes from the observations. Then, I did a thematic analysis based on the coding and the memo that I had created. The thematic analysis was done to identify the emerging themes from the data. While doing the thematic analysis, I created the data structure. The data structure was also created and revised iteratively according to the coding, memo, and thematic analysis changes. In the following section, I will explain the details and process of doing each technique in the data analysis.

### *3.5.1 Transcription and translation*

As explained before, I collected the data using semi-structured interviews and taking notes during the observation. The data I collected during the interviews is in the form of audio recordings and the text of notes that I wrote during the interviews. The observation produces the written text of the observation notes. To analyse the data, I need to have the data in written form to avoid missing any important part of the data and to compare and contrast the data conveniently. I have the data in written form, except for the audio recording. Therefore, I need to transcribe the audio recordings to transform them into a written format. I manually transcribe each interview. After I had done the interview, I spent several hours transcribing the audio recording. Transcribing the audio recording directly after the interviews gave me fresh memories and

insights into what happened in the interview. Therefore, I could take notes of the hunches that emerged when I did the interviews and the transcribing.

I transcribed the interviews by myself without any software to help. Although it took longer time and required more effort, there are several benefits that I gained during the manual transcribing process. I could understand the interviews better and memorise several parts of the data. Therefore, I could recall some parts of the data more easily in the analysis phase. I could make better connections between the data and interpret the data easier. Most of the interviews were done in the Indonesian language. Due to the limitation of translation software that can perform accurate translations in my language, I translated the interview data by myself. I left the transcription in Indonesian, but I translated several important quotes that represent the coding. The translated quotes are the ones that I used to discuss and interpret with my supervisors. Nevertheless, the long hours that I spent on transcription and translation helped me to do the other techniques in the data analysis process.

### *3.5.2 Iterative substantive coding*

After I had transcribed the interview data and the observation notes, I performed substantive coding to find the pattern in the data. Substantive coding means I code to conceptualise the empirical data by finding the emerging core categories. I did the coding using software such as NVivo and Microsoft Excel spreadsheets. The coding is done iteratively and simultaneously with memo writing. There were several steps of coding that I performed during my data analysis processes. By following a grounded theory approach, I adopted the sequence of the cycle of coding that is usually done

in a grounded theory study (Saldaña, 2016). In the first cycle, I did an "*initial coding*" – which refers to "*open coding*" in earlier publications (Corbin and Strauss, 2015) – to produce codes based on the original terms from the data. An initial coding helps me to understand the data from the respondents' original perspectives and language. Second, I continued the coding cycle by doing "*focused coding*" – or "*selective coding*" in earlier publications (Saldaña, 2016). In this phase, I categorise the codes from the first cycle into categories. So, I analysed the text data based on the categories and, if necessary, collected more data based on the selected categories.

The first cycle of coding, the "*initial coding*", is done by doing line-by-line coding on the interview transcripts and the observation notes. I began the initial coding when I started the data collection. After I have the interview transcript ready, I import the transcript to NVivo to begin the initial coding. I tried to be open in coding the data at this stage to get familiar with the data first. I wanted to understand what terms data scientists use to explain certain things. I compare the data with other data to produce the codes. In the beginning, I produced 218 codes. The example of initial codes that I got in the early phase is "*choosing algorithms from the open-sourced libraries*", "*considering whether the problems are relevant*", "*comparing data scientists with data analysts*", "*understanding the data*", and so on. The codes that I got from the first attempt at coding were also changed several times. As I collected more data and more information and understanding, I refined the codes accordingly. After getting familiar with the data through initial coding, I found the sequence of processes in the data – indicated by the use of gerunds (*-ing*) in the codes. So, I employed process coding by ordering the categories based

on the order of what activities data scientists do to complete a project and adding more codes that are related to the process codes. After that, I took my perspective back to my research question. I reflected on the codes and tried to connect the codes with the questions. From that reflection, I could approach my first research question about data scientists using process coding. I refined the research questions according to the initial findings that I found from the empirical data. I also took the same approach to my second research question.

The second step of coding is "*focused coding*". In this stage, I performed a higher level of coding by categorising the codes from the initial coding. I compare the data with the codes and the codes with the other codes to find the similarity and distinctiveness to create the categorisations. I tried to find codes that connect with each other and can be classified into one category. An example of the categorisation of codes is shown in Table 5.

Table 5 Example of categorisation of codes

| Quotes | The codes of data scientists' work | The categories of data scientists' work |
|---|---|---|
| "At the early stage of the project, in the team, I'll spend myself having a chat with the security researchers. The one who really knows what's going on. So, my job is basically to create a detector to block suspicious behaviour. Because I didn't have the domain knowledge, usually in the early days, I talked to the security researchers, the ones who know the types of attack that I've been working on." | Understanding problem context. | Scrutinising problem existence |
| "Usually, I spend around two weeks to define the problems by discussing with anyone who is involved in the project (e.g., the business clients, data analysts, or data engineers). We need to choose the right algorithms, so we need to clarify the problems. The business clients give us access to their data to help us understand the problems, their business, and the data they have." | | |
| "After understanding the problems, data scientists must decide whether the problems are worth solving. We need to break down the problems as clearly as possible to ensure that the problems exist and if it is beneficial for the company to solve the problems. " | Examining problem existence. | |
| "Data scientists must validate the existence of the problem. We need to make hypotheses of the problems and test them. Then, we should determine the next actions to solve the problems. " | | |

Table 5 shows that I classified two codes into one category. The codes are "*understanding the problem context*" and "*examining the problem's existence*." Before I can categorise the two codes into one category, I need to compare and contrast the codes with each other. After the comparison, I interpreted that the codes are relevant because they talk about "*the problems*". The first code

is about the data scientists' task to understand the problem context, and the second code is about the task of examining the existing problem that data scientists want to solve. Therefore, I categorised the codes as "*scrutinising the problem existence*" because the data scientists understand and examine the problem as a part of tasks to check whether the problem exists. The two codes can be categorised into one category because they are related and can be merged into one category of task. I did this focused coding on the other codes. However, not every code can be categorised. Therefore, focused coding is also a phase of selection. Some codes cannot be categorised into any categories. So, I left and kept them uncategorized. Then, there are several categories that are not relevant to my research focus. So, I did not include them in my focused coding. The stage of doing focused coding requires me, as the researcher, to create a decision about the direction in which the research focuses.

Nonetheless, this process of substantive coding is iterative. I did the initial coding several times. In the beginning, I was very open to any codes to understand the data. Then, as I gained more data and understanding, I revised the codes and added more codes to give more clarity to make sense of the data. I started the focused coding only after I felt more familiar with the data. The focused coding was also done repetitively because I changed the focus of my research several times. I collected more data to clarify the focus of my research based on the emerging categories. The codes and the research questions are also refined simultaneously. The process of coding requires a lot of interpretative and reflective practice to make sense of the data and generate the codes. To make the decision about which direction I will focus

on, I also need to reflect on my current data and analysis. Therefore, I also wrote analytic memos to complement my coding process.

### *3.5.3 Analytic memo writing*

According to Birks, Chapman and Francis (2008), memo writing is useful for documenting and structuring the insights and hunches that emerge from interpreting the data. In qualitative research, the researchers are the instrument of research that interpret and extract meaning from the data (Saldaña, 2016). Therefore, memoing is an effective technique for performing reflexive practice in interpreting data. As an instrument of research, I cannot avoid being biased in interpreting the data. By memoing, I could acknowledge and highlight the inevitable bias to reflect on the subjective influences of my own interpretation on the data collection and interpretation (Birks, Chapman and Francis, 2008). My background as a researcher with a business and management perspective inevitably influences how I interpret and understand data scientists' points of view. Memo writing helps me to clarify my point of view. Thus, it helps me reflect and identify my biases and knowledge limitations in understanding and interpreting the data. For example, when a business growth manager talks about the "lack of understanding" from the data team towards the business teams' problems, I can imagine the frustration that the business teams feel by reflecting on my experience. However, this could lead to my interpretation bias toward the business teams' story. I took notes in the memo to acknowledge my bias and how I could handle it. To enhance my limited knowledge, I asked about the same topic to data scientists. They gave me another perspective on the challenges that the data team encounters in understanding business problems. By taking notes in the memo, I could

identify how my bias could influence my interpretation and build a conversation from various perspectives to get a clearer view of the topic.

Memoing not only helps me to perform reflexive practice but also helps me perform communication with myself and other people (Corbin and Strauss, 2015). Memoing is useful to enhance engagement with the data. While reading the data, there were insights and ideas emerged in my mind. I documented them in the memo to keep them and reflect on them (Orona, 1990). I could build communication with myself to think out loud in written text. I can reflect on the ideas that I had written to test the rationality of my ideas and choose the direction of my analysis. As Birks, Chapman and Francis (2008) said, writing memos helps in articulating my assumptions and perspectives. Memoing helps me to document a richer analysis of the data to complement the other techniques. Besides helping me to communicate with myself, memoing also helps me to distribute my ideas to my supervisors. The memo can be material to be shared and discussed with other people. The ideas that I wrote in the memo are more structured and clearer, so they could be more understandable to be refined by my supervisors.

In my data analysis process, I wrote analytical memos in the form of informal writing and the description of insights from the data. An analytical memo is written to construct analytic thought and create an abstraction of the data to develop the theory (Lee *et al.*, 2019). Therefore, I used the analytical memo to help me perform the theoretical sampling to collect more data based on the emerging themes or concepts from the empirical data (Corbin and Strauss, 2015). For example, by writing analytical memos, I could see that taking

“occupational identity” is a relevant theory to analyse my data. The analytical memo helps me to conceptualise the findings in a more abstract form. Then, I tried to collect more data about occupational identity to explore and deepen insights into the themes in the data. Memo writing is not only one step in the data analysis (Charmaz, 2014). I wrote analytical memos throughout the data analysis iteratively while doing other data analysis techniques, such as coding. Analytic memo writing helps me to generate the codes and take notes of my thought process in generating the codes. So I can keep the documentation of my thought process when analysing the data.

### *3.5.4 Thematic analysis*

To make sense of the data more abstractly, I performed a thematic analysis to generate themes from the empirical data. According to Boyatzis (1998), thematic analysis is a way to make sense of qualitative information systematically.. Thematic analysis is also a way to present the meaningful conceptualisation of qualitative data (Boyatzis, 1998). The themes generated in this stage are the theorisation and integration of the codes and categories identified in the previous stages of analysis (Charmaz, 2014; Corbin and Strauss, 2015). I looked more thoroughly at the empirical data to see the pattern and then explicated them with possible relevant theories. The themes should be an abstraction of the categories. I used the description of the connection between the codes that I wrote in the analytic memos to guide the abstraction. The process of producing the themes can also change the codes and categories in the previous stage of analysis.

The themes can be generated by an empirical data-driven approach, a theory-driven approach, by comparing empirical data and theories, or by doing all together (Boyatzis, 1998). At the beginning of my research, I clearly stated my theoretical assumption of the nature of "*data*" as being open-ended. By using this theoretical assumption, I tried to see the pattern of what data scientists do to harness the insights from data. I keep this theoretical assumption in my mind when I code the data. I tried to connect the substantive codes with the theoretical assumptions. After I had produced the substantive categories, I tried to think about how the open-endedness of data is related to what data scientists do.

By iteratively comparing and relating the substantive codes and categories with my theoretical assumption, I produce several themes from the empirical data. I generated the label of each theme. Following the guideline from (Saldaña, 2016), the label of the themes must represent the codes and categories in a more subtle and tacit form. They must be clear, concise, and short enough to describe the meaning of the phenomenon being studied (Boyatzis, 1998). Then, the themes that I have generated are structured by adopting the Gioia method in structuring the data (Gioia, Corley and Hamilton, 2013). I present the data by showing the codes, the categories, and the themes in order to show which codes constitute the generation of each category. The codes are presented in the first order of analysis. After that, the categories are presented in the second order of analysis. Lastly, the themes are presented in the last order as the abstraction of the whole code and categories. I will show and explain the data structure of this research in the next subchapter.

## 3.6 The data structure

In this section, I will present and explain the data structure I produced from the data analysis. I adopted the Gioia method in presenting the data structure (Gioia, Corley and Hamilton, 2013). In this research, I have two research questions. Therefore, I have done two separate analyses and created two data structures for each research question. I divide this section into two: first, I will show the data structure for the data scientists' doing, and second, I will present the data structure for the data scientists' being. The data structures will show the codes, categories, and themes for each analysis in detail.

### *3.6.1 The data structure of data scientists' "doing"*

The first data analysis I did focused on what data scientists do. I aim to understand "the doing" of data scientists to understand the step-by-step practices that they do to extract insights from data, especially in the context of data being open-ended. I asked the data scientists to explain what they do in their jobs. I collected numerous explanations about what data scientists do from the interviews and some additional data from the observation. I read and analysed the narrative text of the interview data, performed iterative coding, and took reflection notes in the analytic memo. By doing all of them repetitively and simultaneously, I produced the codes, categories, and themes about data scientists' doing. The process of creating the data structure for data scientists' doing data structure was not created in the first attempt. It consists of several iterative steps.

First, I made myself familiar with the data to understand what data scientists do. I coded each data scientist's action from the interview and observation

data. From the initial coding, I recognised a pattern of process. The empirical data shows a sequence of actions that data scientists do with data to complete their projects. This sequence of activities can be studied using the process study (Corbin and Strauss, 2015). Some data scientists explain how they extract insights from data chronologically in the process. Then, I also recognised the same pattern in the other empirical data. During the initial codes, I reflected on this process of the theoretical concept of data. I found a stream of scholars that conceptualise the transformation of data. So, I decided to adopt a process perspective to understand what data scientists do in chronological order and what they do with data during the process. Therefore, during the initial coding, I employ process coding to study data scientists' narratives about what they do in chronological order. I organised the actions that data scientists do to complete a project, starting with what they do first until the end.

After I had the codes and categories in order, I continued the process to the abstraction stage. I went back and forth between the codes and the memo to reflect on the context of the "open-endedness of data." I used those reflections as references to label the themes. The process of labelling was not a simple task. I revised the label numerous times to choose the terms that represented the meaning that I wanted to present from the data. After several iterative processes, I formulated the labels for each theme. Based on my reflective analysis, there are three themes about what data scientists do that emerge from the empirical data. The themes are: 1) validating problems, 2) validating data, and 3) validating algorithms. Those themes emerged from the

theorisation of codes and categories that I have organised in order. The data structure of my first analysis is shown in Figure 3.

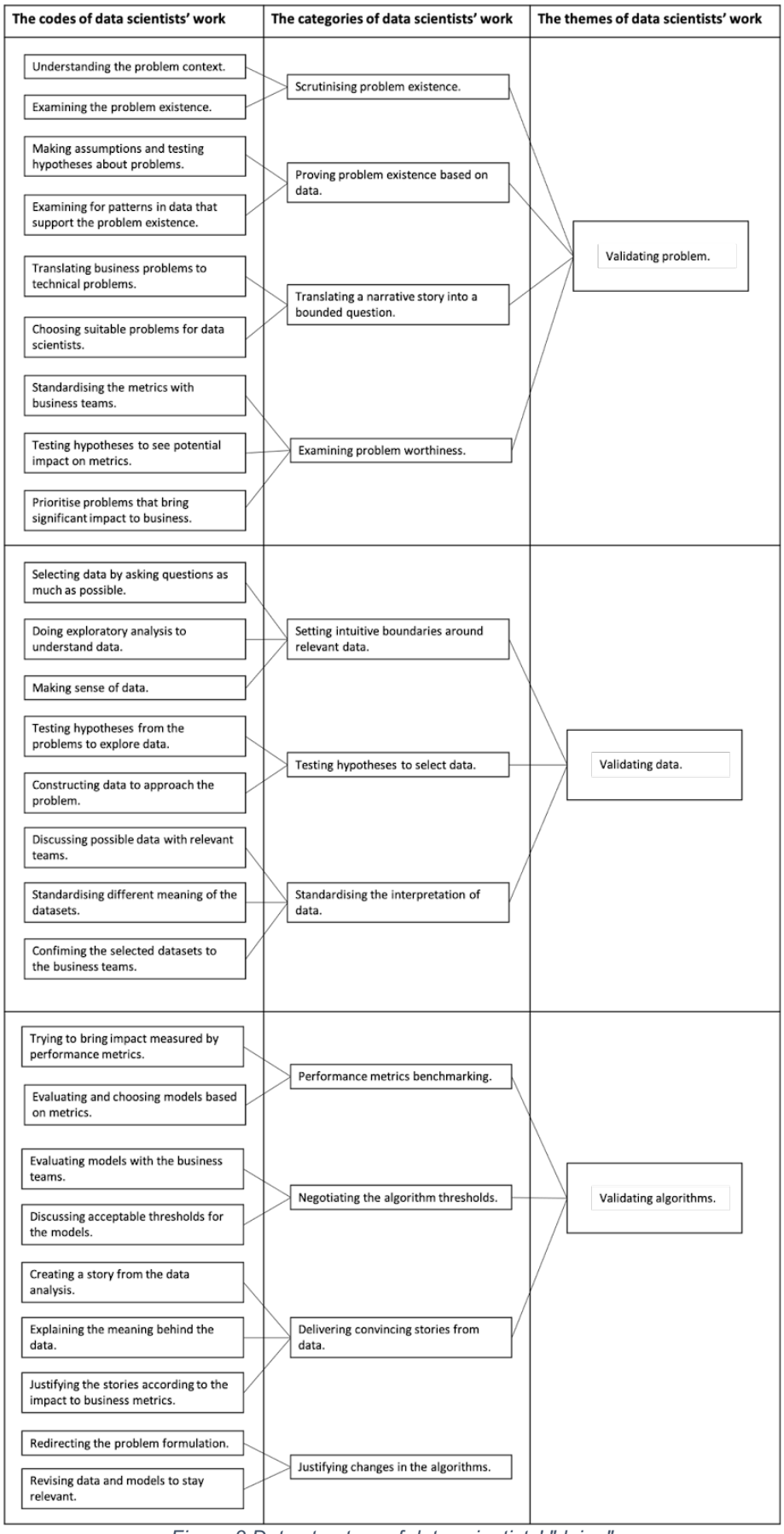


*Figure 3 Data structure of data scientists' "doing"*

### *3.6.2 The data structure of data scientists' "being"*

The second analysis that I performed focuses on data scientists' occupational identity. The analysis is done to answer the second research question about how the open-endedness of data influences data scientists' occupational identity. My interview data also comprises data scientists' opinions about who they are. I used the same empirical data sources – the interview and the observations. I also performed the analysis with a similar approach to what I did in analysing data scientists' doing. However, I took a different lens in analysing the empirical data to study data scientists' occupational identity. Because I took a different lens, the way I interpreted and organised the data is different compared to what I did in the first analysis.

In this analysis, I aim to examine the perspective of data scientists regarding their identity. I do not focus on studying how their identity is shaped but only on how data scientists portray themselves. Therefore, I did not do process coding in the initial coding. I performed the initial coding by identifying the statements in the interview that indicates how data scientists describe who they are. Besides the interview, I also went through the observation notes to identify any indication of what data scientists say about their identity and how they tell other people about it. In the early phase of initial coding, I code the statements using exactly the same words and terms as what data scientists say about themselves. Then, in the next iteration, I also add some expressions that are produced by my interpretation.

Through coding and memoing, I found a pattern of identity themes in the empirical data. The categories define data scientists' identity based on who

they think they are supposed to be and who they are based on what they enact. I theoretically labelled the themes as “espoused identity” for the former, and “enacted identity” for the latter. By analysing those two themes, I found inherent tensions in data scientists’ identities – which I will explain further in Chapter 4. Therefore, besides the two themes of identities, I also found another theme about how data scientists manage the inherent tensions. I label this theme as “managing the inherent identity tension.” The code, categories, and themes that emerge in this analysis are shown in Figure 4.

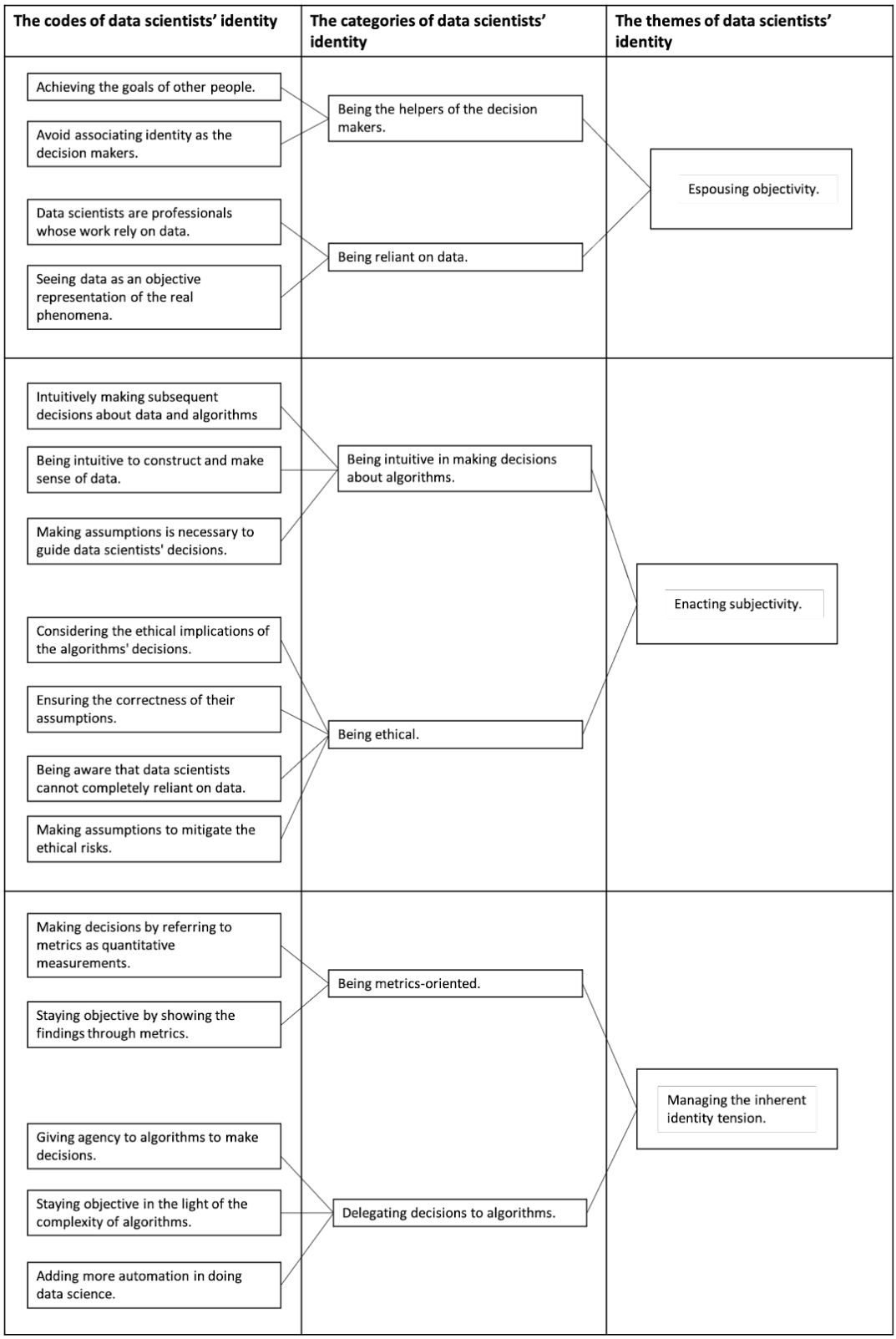


*Figure 4 Data structure of data scientists' "being"*

## 3.7 Summary of the methodology

In this chapter, I have explained the methodological aspect of my study. I have explained the research paradigm by making statements about my research approach and epistemological position. The choice of the research paradigm is taken based on my research aims. I chose inductive and qualitative inquiry to enable me to study a topic that is not yet well-explained by existing theories. Then, I chose the grounded theory method to provide me with a rigorous guideline in analysing qualitative data and developing theoretical concepts that explain data scientists' work and identity. I also chose to join the interpretive epistemological stream in the IS field to enable me to study the social phenomena of data science in a more comprehensive way.

My choice of research paradigm determines how I collected and analysed the data. I gathered the qualitative data by doing semi-structured interviews with data scientists and other professionals that work alongside data scientists. I also took notes of the observation to data science communities. Then, I analysed the data by following the grounded theory steps. I performed the coding, memo-writing, constant comparative analysis, and theoretical sampling iteratively. Those steps enable me to perform a thematic analysis to develop the themes from the empirical data. Reflecting on my experience in doing the data collecting and analysis, the grounded theory method requires me as the researcher to be adaptable to changes that happen during the process. There were many times I refined the research questions after I performed the coding, memo writing, and theoretical sampling. Nonetheless, the method enables me to gain the findings truthfully from the data.

I created the structure of the findings from the empirical data following the Gioia method. The further details of the data analysis for each data structure are explained in the next two chapters – Chapter 4 and Chapter 5.

# CHAPTER 4

# DATA SCIENTISTS' VALIDATION: A PROCESS TO NAVIGATE THE OPEN-ENDEDNESS OF DATA

## 4.1 Introduction

This chapter discusses the first section of the analysis that focuses on the process of navigating the open-endedness of data in the data science process. The aim of this chapter is to unpack the kinds of work that data scientists perform to extract valuable insights from data. Based on the empirical data, this analysis identified what data scientists do based on what they tell. It allowed the analysis to capture data scientists' views and reasonings in overcoming the problems in their work.

As shown in the methodology chapter, the analysis is done by tracing the bundle of tasks that data scientists tell in completing a project. Data scientists perform bundles of tasks that include qualitative and interpretive tasks (e.g., framing, brainstorming, negotiating, and storytelling), apart from quantitative and calculative tasks (e.g., hypotheses testing, data wrangling, and modelling). Data scientists perform the bundles of tasks interchangeably to perform validations that navigate the open-endedness of data. Dividing the process into phases, the analytical themes show that data scientists are doing validation in three phases of the ongoing process: (1) validating problem statements, (2) validating data, and (3) validating algorithms, as shown in Figure 5.

The process begins when data scientists receive or find a problem. Then, data scientists perform bundles of tasks to validate the problem, the data, and the algorithms until they accept the algorithms to be deployed to the systems. In the following sections, I will explain the details of the tasks that data scientists do throughout the validation process to navigate the openness of data interpretation.

### 4.2 Validating problem

The openness of how data can be interpreted suggests innumerable ways of making sense of and utilising data t for various purposes. Because there are so many options for the direction of interpretation, data scientists need to specify the purposes that define the intended outcomes. At the beginning of a data science project, data scientists choose and frame specific problem statements. The problem statements will guide the project and create the project boundaries in harnessing data towards the outcomes. The process of framing the problem statements requires data scientists to sit together with the business teams to understand and discuss the problems. Data scientists need to judge which problems are worth solving. In addition, data scientists perform calculative work to augment the problem framing with evidence to validate their problems. In the following sections, I will explain the steps that data scientists take in validating the problems.

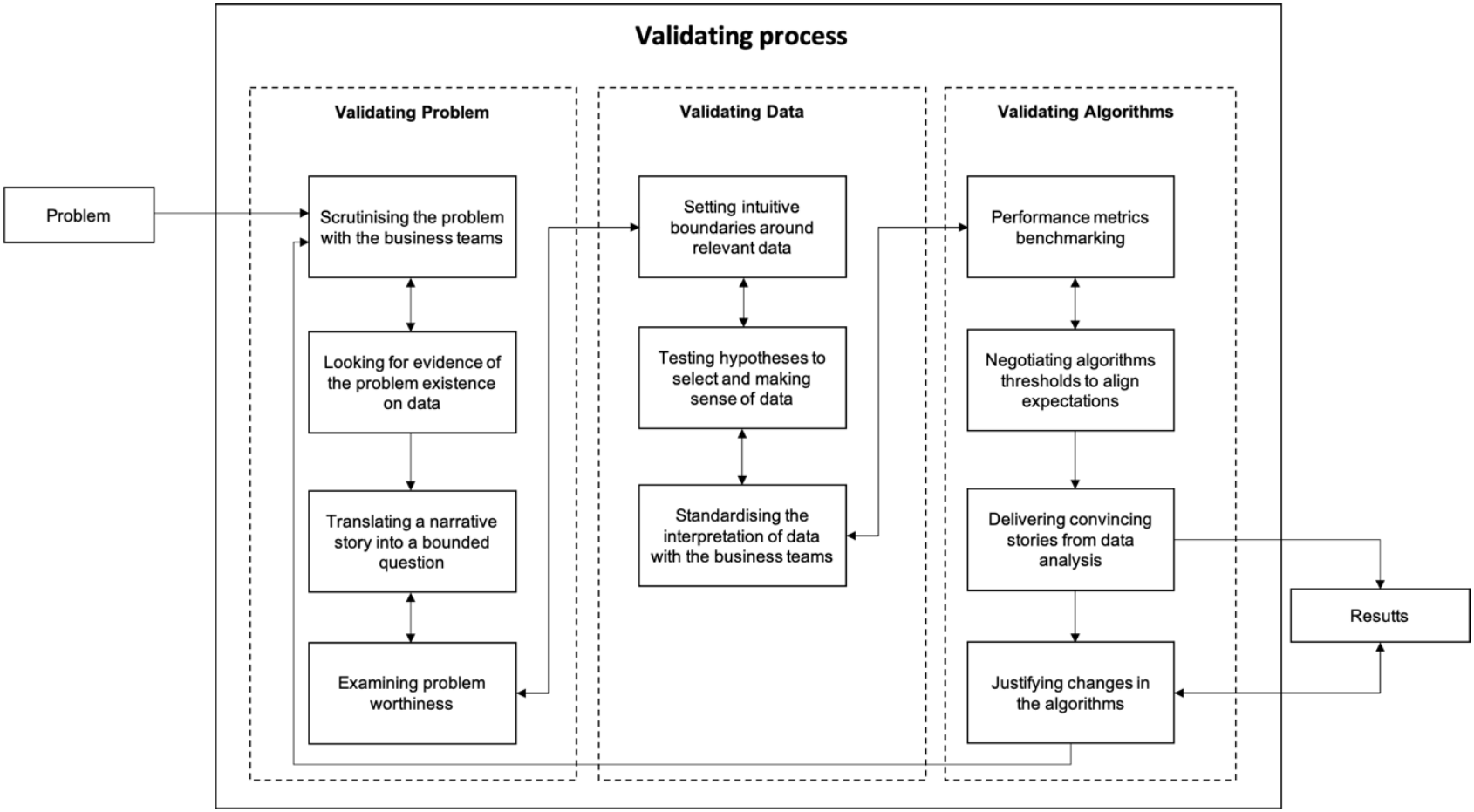


*Figure 5 The data scientists' validation process*

### *4.2.1 Scrutinising problems through communication with business teams*

At the beginning of the project, commonly, data scientists receive problems from the business teams[2]. The business teams come up with problems that they would like to solve using data. The first step that data scientists take after receiving the problems from the business teams is to scrutinise the problem. To do this, data scientists try to get a deep understanding of the problem context by asking questions to the business teams extensively. While data scientists are trained to analyse data, in doing so, they need to understand relevant domain knowledge situating the problems. Data scientists might have limited knowledge to understand the problem context. Therefore, asking the relevant business teams to understand the domain knowledge is important to scrutinise the problem. One data scientist who works to support an IT security department gave an example of how they try to understand the problems:

> *"At the early stage of the project, in the team, I'll take the time to chat with the IT security researcher. The one who really knows what's going on. To give context, my job is basically to create a detector to block suspicious behaviour. Because I don't have the domain knowledge, usually in the early days I talked to the IT security researchers, the ones who know the types of (cyber) attack that I've been working to prevent." – DS12.*

DS12 was asked to create a model to automatically block suspicious behaviour. As a data scientist with 4 years of experience working in various industries other than IT security, DS12 has limited domain knowledge in IT security. So, DS12 needed to spend several times asking the IT security researcher about the range of breaching behaviours. The limited domain

[2] The "business teams" is the term that data scientists often refer to other teams within the organisations whose work oriented on the core business activities, for example, the sales team, marketing team, operations team, and others.

knowledge creates a level of dependence in dealing with data as data is relational to the context. The relationality of data entangles data with people, work practices, and the environment. Therefore, for data scientists, engaging in close communication with the people or team that have the relevant knowledge about the problem is important to be able to solve the problems with data effectively.

By breaking down the problems and asking questions to relevant teams as much as possible, data scientists not only understand the problem to solve it but also scrutinise whether the problem exists or is worth solving. The problems come to them through the business teams' narrative stories which means that the stories of the problem are constructed based on the business teams' understanding and assumptions. One data scientist explained:

> "*My data team leader said to me that the problems that the business teams identified might not be the real problems" – DS14.*

Data scientists think that they need to examine whether the problems that they receive are real problems. What they mean by "real problem" is the root cause of the problems, not only the symptoms that the business teams identify.

Solving problems requires resources and costs; therefore, data scientists need to investigate the existence of the problem. Data scientists scrutinise the problem to find the real problem and ensure that it can be solved using their skills. As explained by a data scientist:

> "*Business teams come to us with a problem. But, as data scientists, we need to clarify the problems; are they suitable for us? […] We need to validate the problems; does the problem really exist? We break down the problem as clearly as possible to make sure that the problem exists or happens. […]" – DS15.*

In this way, they validate the kind of problems that are suitable for data scientists. By scrutinising the problems identified by the business teams, data scientists gain an understanding of the problems choosing the focus and creating a boundary around the problem. Data scientists can choose what must be included in and excluded from the problem scope. However, scrutinising the problems is not enough to describe the problem in plain language subjectively. Scrutinising the problem should be supported by investigating the data to frame the problem with hard data objectively which leads data scientists to the next step.

### *4.2.2 Looking for evidence of the problem's existence in data*

Scrutinising the problem is a crucial stage because it guides the actions and decisions that data scientists perform in the next stages. As explained in the previous section, scrutinising the problem involves judgements about the existence of the problems and which problems they should solve. Therefore, data scientists need to confirm their hunches with evidence by performing calculative work with data. Data scientists confirm their understanding and hunches by creating and testing their hypotheses about the problem. They investigate data to see the pattern that supports their understanding of the problems.

> "*After we received the problems from the business teams, we checked data to see the patterns in the data.*" – *DS2.*

The output of the hypothesis testing can be used as evidence to validate the problem's existence. Numerical data can support their judgement about the problems or otherwise falsify their judgement which leads them back to inform the business teams to scrutinise the problem's existence.

### *4.2.3 Translating a narrative story into a bounded question*

The problems that data scientists receive from the business teams' stories are often too broad and poorly defined. Therefore, after scrutinising the problems, data scientists articulate the boundaries around the scope of the problem by formulating a focused question. A data scientist explains:

> "*As data scientists, we have to be able to solve the problem. So, the most important thing is the problem statement. We only need one or two questions. Very often, business teams tell us the problem in a long story. Then I said 'I only need the question. What is the question?' If data scientists cannot help them to formulate the questions, then what are data scientists for?*" *– DS5.*

Data scientists need a clear question as a problem statement to guide them. Together with the business teams, data scientists turn the business teams' stories into specific questions. They expect themselves to have the skills to help the business teams formulate the question.

However, in this step, it is important to highlight that data scientists, together with the business teams, impose their judgement in formulating the question. Formulating the questions is not an exact science but rather depends on the context, the perspective and the purposes of those who ask the questions. Problems can be defined in many ways according to the purpose that an individual or an organisation wants to achieve. As boyd and Crawford (2012, p. 674) argue: "*who is asking the questions determines which questions are asked*". Data scientists formulate the questions in a way that the questions can be solved using the data science approach. A data scientist said:

> "*In my case, I handle problems related to marketing. For example, evaluating the impact of the marketing campaign on sales. But we can evaluate the impact in many ways. We can have a different definition of impact, maybe we can see it based on the traffic or something else. […] Very often, the business teams give us a very general problem,*

> *only asking for impact. So, in the end, we have to be able to translate this kind of business problem to something that can be solved by statistics or machine learning, for instance." – DS4*.

The quote above states that data scientists "translate" business problems into mathematical problems. In "translating", data scientists define the impact in certain ways and refer the problem formulation to the impact they want to bring. Based on the impact that they have defined, data scientists exclude other definitions of the impact that cannot be addressed using mathematical methods.

More than translating, data scientists choose the questions that can be solved using mathematical methods to achieve the impact that data scientists had defined. For example, some data scientists describe the suitable questions for them are ones that need complex statistical methods or modelling, specifically predictive and prescriptive analyses:

> "*When companies need more complex analysis […] for example when the problems need to be solved using machine learning or prescriptive analysis, optimisation, or advanced statistical modelling […], they ask data scientists*." – *DS4.*

DS4 was a data scientist who worked within an IT and Hospitality company with other data professionals, such as BI analysts and data analysts. Therefore DS4 could categorise the analytic problems specifically for data scientists. The organisational context in defining data scientists' problem scopes frames the problem categorisation. It also frames the data scientists' relational connection on specific domains – e.g., prescriptive and predictive analytics. As an example, the following quote from a Business Growth Manager represents the organisation's expectation of the problem to be solved by data scientists:

> *"We need data scientists to create machines that can learn the customers' behaviours, so we can predict better prices for our products […]" - BGM1.*

While DS4 scopes data scientists' problems on prescriptive analysis, BGM1's quote shows that data scientists are expected to solve predictive analysis problems. Therefore, the organisations' framing influences data scientists' doing by making variations of data scientists' problem scoping in translating the narrative stories into questions.

Some data scientists also emphasise that the questions that are worth pursuing through data science are those that bring significant and long-term impact, as one of the data scientists explained:

> "*Usually, data scientists solve problems which bring long-term and huge impact. Our models are usually implemented in the product. Creating models to be implemented in products requires a long process and a huge investment […] Data science is quite complex in terms of modelling. Therefore, when we create a solution from data science, we expect a huge impact*." *– DS5.*

DS5 worked in a similar company with DS4. The quote shows that data scientists expect a high return to business from the outcome of their work; thus they make decisions on which problems become considered important enough to solve. Some of them think that their work needs a lot of investment. They only want to focus on important problems that bring a huge impact on the business. Therefore, ensuring that solving questions is worthy for the business is important for some data scientists.

### *4.2.4 Examining problem worthiness*

The formulation of the questions is then complemented by calculative work to examine the worthiness of solving the defined questions. In so doing, data

scientists create hypotheses and test the potential economic value of solving the problem. Data scientists and the business teams define the performance metrics that will be used to measure the outcomes of the analysis or the model. In this stage, data scientists predict the potential value of pursuing the problem, given the performance metrics. If data scientists are convinced that the problems are worthy, they validate the problems and design the actionable steps. One data scientist explained:

> *"If the company stakeholders trust data – or are data-driven – they won't do something carelessly without looking at data. Usually, in the problem formulation, data scientists provide strong cases to pursue or to not pursue a particular project." – DS15.*

The calculative work in examining the problem's worthiness shows that data scientists measure the worthiness of solving a particular problem through numbers. Data scientists could make a strong argument about the problem's worthiness by presenting the predicted economic value of solving the problem. Data scientists need to see the numbers to be convinced that they will solve important problems that bring measurable impact to organisations.

### 4.3 Validating data

After framing the problem statements, data scientists need to select data. They translate their understanding of the problems to select relevant data. Data scientists apply the understanding of the domain knowledge that they gained from scoping the problem. Because of the open-endedness of data, data scientists need to set an intuitive boundary in selecting relevant data. Then, data scientists need to provide evidence by testing hypotheses in selecting the data. Finally, to validate the selected data, data scientists need to standardise their interpretation of the data with the business teams.

### *4.3.1 Setting intuitive boundaries around relevant data*

Data scientists need to select relevant datasets according to the problems they want to solve. There are infinite options of datasets that may be selected. Datasets selection requires a deep understanding of the domain knowledge related to the problems. Data scientists might not have in-depth knowledge about the problem context. Therefore, in the early stage of data selection, data scientists brainstorm in discussing suitable data with fellow data scientists in one team or with the relevant teams that are knowledgeable of the problem. One data scientist said about this brainstorming stage:

> "*Usually, I discuss the possible data with other data scientists in my team. If possible, we also brainstorm with the business teams. Sometimes, we conduct a regular meeting with the business teams to brainstorm and report the progress*." *– DS11.*

Selecting datasets requires work that involves deliberation and brainstorming with other parties, ideally the teams who own the problems.

In brainstorming the relevant datasets, data scientists start by asking as many questions as possible. They brainstorm hypothetical questions about what variables or datasets might be related to the problems that they are solving. One data scientist offered an example:

> *"To explore and select the variables we want to include in the model, we started by creating a list of questions that we want to know from the datasets. For example, I handled a project about dairy products. We wanted to know the relationship between the cows' age and the raw material price (e.g., the milk). We created a list of hypothetical questions, for instance, we hypothesised that the cows' age maturity affects the size of milk production. We wrote other hypothetical questions that emerge in our mind." – DS11.*

Data scientists started to create questions based on hunches they have about the problems. The hypothetical questions emerge intuitively based on their

common senses. That is the reason why brainstorming is important to generate hypothetical questions because brainstorming can generate ideas despite their limited understanding of the problem context.

Nevertheless, data scientists require some level of understanding in certain domain knowledge. Without this knowledge, exploring possible data can be exhaustive.

> *"We start by asking many questions. We try to explore possible factors that might influence the variable that we are looking at […] Well, we can't explore all possible data because there are just too many. So, it is important to have some knowledge related to the problem." – DS11.*

Data scientists also need to do their own research to gain relevant domain knowledge. Having a certain level of knowledge can help data scientists to generate questions efficiently. Learning new domain knowledge independently is an important skill for data scientists to be able to sharpen their intuition. As said by a data scientist in a webinar:

> *"This is an important skill for data scientists. Using our intuition, we select possible data and then test them. Choose the ones that are relevant" – a data scientist (from the observation to a technology company's webinar, 2020).*

Working with intuition is essential in the first step of selecting data. Very often, data scientists do not know what to do, so exploration is important. Their intuition is the one that is guiding their exploration in finding relevant data to solve their problems.

### *4.3.2 Testing hypotheses to select data and make sense of data*

The previous step involves a high level of intuitive work for data scientists. Therefore, data scientists need to complement it with calculation. Data scientists need to test the hypotheses to select the datasets. In testing the

hypothesis, they look at the pattern and relationship between the variables to delineate datasets. One data scientist gave an example:

> *"We examine the correlation between variables to see the relationship of each variable. If we find multicollinearity – which means that there is strong correlation between one variable with other variables – then, we exclude that variable." – DS11.*

Data scientists test their hypotheses and make sense of the data by examining the pattern. When a selected data is excluded, data scientists might need to select another data.

After testing hypotheses, data scientists might go back to the brainstorming process to create other hypotheses. The process of doing intuitive work in generating hypothetical questions with hypothesis testing is iterative until they decide on the selected data. One data scientist gave an example of how they selected data. This data scientist had a project in corporate banking, and they wanted to predict which companies are highly likely to default on debt:

> *"I tried to understand the factors that influence companies' debt default. I started with what I knew. I tried looking at the CEOs' credit scores, their personal finance, and also the company's financial data. I tried to explore them, add other data, and redirect my analysis […] What I know guides what I can explore. So, I also keep collecting data as I go […]" – DS14.*

DS14 tested their hypotheses to select data, but in the process, DS14 might find other hypothetical questions. As they gained more understanding, DS14 created other hypotheses, collected more data and redirected the analysis. The process of selecting data is iterative and guided by hypotheses. To validate the datasets, data scientists need to harmonise their interpretation with the relevant teams who own the problems.

### *4.3.3 Standardising the interpretation of data with the business teams*

Data scientists confirm the selected data from the hypotheses testing to the business teams. Data is open to different interpretations in many ways depending on the context (boyd and Crawford, 2012). The interpretation, and by extension, the valuable insights gained, is mediated by data scientists. Therefore, data scientists need to confirm whether the business teams have the same interpretation and agree with the selected data. One data scientist explained:

> *"After I select the data, I will communicate back and forth to the business teams (e.g. the IT security team) to make sure that I collect the right data" - DS12.*

Data scientists are solving the business teams' problems, so they need the business teams' opinions to verify the data. Nonetheless, the business teams are the ones who own the problems. Data scientists are helping them to solve their problems, so data scientists think that they will need the business teams' opinion in selecting the data to ensure that data scientists' judgement aligns with the business teams' understanding and purposes.

Checking with the business teams is important for data scientists. On several occasions, data scientists found that different business teams have different interpretations of data. The following quote by DS2 illustrates the example (DS2 worked as a data scientist in a small size IT and recruitment service company):

> *"In my company, we have the term 'active user'. Different teams have different perceptions of this term. The marketing team sees an 'active user' as someone that has signed up and created an account on our website. While the operations team thinks an 'active user' is a person who has signed up and also made several transactions for at least one month. They have different perceptions because they have different needs. This is dangerous for the data team. When they ask us to analyse using 'active user' data, we may use the wrong data. So, we*

> *tried to standardise the naming of data. For example, we use 'registered user' to define the former and use 'active user' for the latter. We tried to include the business teams to standardise the naming of data together." - DS2.*

The example by DS2 shows different interpretations of data. The marketing team sees 'active users' in that way because their goal is to convert people to sign up on the website. The success of the marketing team is defined by the conversion rate. While the operations team has a different goal. They need to increase transactions as much as possible. So, they need data about people who have made transactions on the website. DS2 who worked in a small size company had a significant authority in organising the data. DS2 then take a position to standardise the differences in order to accommodate different needs. Data is constructed to serve particular purposes. By discussing with the relevant teams, DS2 standardised the meaning of data so that every team will have a mutually intelligible interpretation of the data.

The confirmation of the selected data provides a chance for data scientists to standardise different interpretations of the data. And, if the selected data is accepted by the business teams, then the data is validated by the business teams as the ones that have deeper domain knowledge of the problem. The combination of the brainstorming and confirmation of the business teams and the hypotheses testing produces a validation of the datasets. Brainstorming enables data scientists to create hypotheses and selects the data by testing the hypotheses. Then, the confirmation to the business teams brings the perspective from the business teams about the selected data. Data scientists might justify the data selection and standardise different interpretations of the data. This confirmation provides legitimation to the selected numerical data

which further validates data. Data scientists might use the datasets and move to the next steps.

## 4.4 Validating algorithms

After selecting the data, data scientists feed the data into the algorithms. Data scientists develop, modify, or manipulate algorithms to solve problems. In validating the algorithms, data scientists then choose the model according to the performance metrics. However, data scientists encounter several challenges in this stage, for example, dealing with the trade-offs between measuring the algorithms based on statistical metrics against the business metrics. Besides, data scientists also face a challenge in communicating the algorithms to the business teams. In this section, I will explain how data scientists overcome those challenges in validating the algorithms.

### *4.4.1 Performance metrics benchmarking*

There are many ways data scientists can develop their algorithms. In certain situations, data scientists develop algorithms from scratch based on requests from the business teams. However, commonly, data scientists search for existing algorithms that are written by other people on forums or websites (e.g., GitHub, GitLab, and StackOverflow), and then modify the algorithms according to their purpose. One of the data scientists working in a small size start-up company explained:

> *"When I received the problem from the Head of Data, I searched for the algorithms in forums or websites, such as GitLab and GitHub. Then I improved and modified the algorithms. I think that is a common practice in many companies. Usually, I tried the algorithms one by one by using the data then I compared the accuracy. I chose the algorithm that gives the highest accuracy." – DS3.*

The quote represents the common practice of data scientists in business organisations who find algorithms from forums or websites. When they are choosing the algorithms, they do not choose by breaking down the algorithms to understand the detailed steps. Data scientists undertake performance benchmarking by testing the algorithms by inputting the data and evaluating the results based on the statistical performance metrics, for example, accuracy, logarithmic loss, confusion matrix, the area under the curve, F1 score, mean absolute error and mean squared error (MSE).

Following the statistical rule, a good algorithm is one that gives a high value to each of the performance metrics. However, sometimes increasing the performance metrics conflicts with the economic value that the business teams' favour, as explained by one data scientist:

> "*We have several metrics to evaluate algorithms, for example, accuracy, recall, precision, etc. If we are not making judgements based on business values and impact, we will absolutely decide to improve the accuracy, precision, and recall as high as we can. But, in the implementation, we can't do that. Still, the metrics should be high, but these metrics do not matter that much. For example, when we want to improve the accuracy by 5-10%, we have to consider whether or not the effort to improve will have a significant impact (on the business). If not, then the improvement is unnecessary*." – *DS14.*

DS14 was very aware of the company's limited resources. Increasing the performance metrics needs time and cost. DS14 thought they cannot choose the algorithms based on merely statistical performance metrics. There might be conflicting concerns between data scientists and business teams in terms of the threshold of accepting the models. For example, data scientists want the algorithms to have a high value of accuracy, while the business teams want the algorithms to be developed quickly, and cheaply, and create significant economic value.

### *4.4.2 Negotiating algorithms thresholds to align expectations*

To overcome the challenge in performance metrics benchmarking, data scientists negotiate acceptable thresholds with the business teams. Naturally, data scientists evaluate models based on statistical metrics. But they are aware that they are serving business needs. A data scientist said:

> *"Our end goal is bringing impact to the business. If we don't think about the impact, we might only think about improving the metrics – accuracy, recall, precisions, etc. Improving the accuracy matters, but the most important is the impact on business. For example, if we improve the accuracy by 5% or 10%, what is the impact on the business? Does it bring a significant impact? If not, then, it's not worth modifying the algorithms to increase the accuracy."– DS14.*

Data scientists think they need to accommodate business goals. But improving the performance of the algorithms based on the statistical metrics is still important. Data scientists need to consider both measurements in defining the algorithm's threshold. Nonetheless, both measurements are something negotiable. Therefore, in this stage, data scientists negotiate the thresholds with the business teams to accept the models.

Data scientists need to communicate and negotiate the business team's expectations about the metrics because they think that the decisions are to be made by the organisation. One data scientist explained:

> *"Our decisions depend on the company. As a data scientist, of course, I want to prioritise increasing accuracy. But if the company wants to obtain the results quickly, then I don't have much time to increase accuracy. I need to sacrifice the algorithm's quality a little bit. At least we have a minimum threshold for accuracy. So, we negotiate the threshold with the business teams to complete the project quickly while maintaining the quality." – DS2.*

The quote shows the example of the moment when a data scientist wants to prioritise accuracy, but the business teams want the models to be ready

quickly. DS2 negotiated the minimum threshold of accuracy to maintain the model quality given the limited time from the business teams. Data scientists try to reach a win-win solution, so both parties (i.e., the data scientists and the business teams) are satisfied.

Negotiating the algorithms' acceptable threshold is a crucial step in aligning the data scientists and business teams' expectations and achieving their shared goal. At the end of the day, data scientists' goal is to help business teams to solve problems and achieve their business goals. One data scientist explains:

> *"The key metric of the company is to increase revenue. The business teams often come to us and say, "we need an algorithm with x% accuracy to get a 50% increase of the business value." Then, it becomes the data scientists' goal. Our goals should be aligned with the business team's goals because we are supporting them." – DS13.*

Data scientists are hired by organisations to help them reach their organisational goals. Therefore, data scientists need to always communicate and align both parties' expectations to develop the algorithms in a way that is accepted by both parties.

### *4.4.3 Delivering convincing stories from data analysis*

To validate the algorithms, data scientists must ensure the algorithms can solve the business teams' problems as expected. Therefore, data scientists must deliver convincing results from their analysis. Data scientists communicate this by narrating stories. Just like telling stories, there is an art through which data scientists convey meaning through storytelling.

> *"We do data storytelling. When we do it, first, we choose simple terms and language. Second, we create narratives. We decide the points that we want to deliver and the storyline. What did we want to deliver at the beginning, and how was the journey? It is like writing an article*

*in a newsletter, but what we show is the data. We have to create the story from the start until we get the results." – DS14.*

Data scientists consider the language preferences of their audiences - for example, whether the audience understands data visualisation. If they think that the audience has limited knowledge to read data visualisation, then they will deliver the result using simple and easy-to-read language and charts. Using appropriate terminology, the meaning from the data fits into a story plot structured around the originally framed problem. The outcomes of the algorithmic modelling suggest scripts for the story characters, namely the business stakeholders.

In delivering convincing results to business teams, communication is often still a challenge for many data scientists. The challenge lies in having a different perspective on the data and the business problems. Data scientists and business teams have different knowledge backgrounds. What is important for data scientists might not be important for the business teams. In order to deliver the analysis effectively, data scientists must learn how to communicate in a way the business teams can easily understand and see the worthiness of deploying the results. As one of the respondents said:

*"When we are talking about algorithms with other data scientist fellows, we talk about the accuracy, sensitivity, precision, and metrics of the algorithms. But the business teams don't care about those things. What they want to hear is what the algorithms can do to increase business revenue or cut costs. For example, even though the accuracy of the algorithm that we build is 3-5% below the target, the business teams can accept it as long as the cost is cheaper. Now, we understand their expectations. The thing is to translate our technical analyses into business language. How our findings can improve business. That is what they want to hear." – DS4.*

To communicate clearly, data scientists must learn to translate their findings into the impact on the business. Data scientists must be aware of the things that the audience (i.e., the business teams) need and would like to know. Data scientists could translate the findings from the analysis to show how they could increase revenue, efficiency, and other business metrics that are deemed important for the business teams.

### *4.4.4 Justifying changes in the algorithms*

After the algorithms are accepted by the business teams, then data scientists must maintain the relevance of the algorithms to the ever-changing situation. The open-endedness of data means data is generative, extendable, and can be broken when the context of their interpretation has changed. Data is very fragile to the changes that happen in the problem context. In order to keep the results relevant, data scientists must justify any changes needed in the model. The Covid-19 pandemic is a great example to show when certain data can be no longer relevant. Data scientists work with historical data to provide insights into current problems and also predictions for the future. The pandemic has created changes in almost every business model, rendering historical data irrelevant to today's condition and future predictions. Data scientists could no longer work with historical data in the same way. One of the interviewees reflects on how the pandemic affects data.

> "*Covid 19 really affects our data analysis. Let me explain the context in our bank. All our customers are entrepreneurs and women who have micro to small businesses and live in rural areas. We analyse data by classifying which customers will be offered lending/financing in the future and which ones do not. The classification is done based on data on the ability to pay and entrepreneurial spirit. In the pandemic, the situation is different. The customers still have a high entrepreneurial spirit, but their businesses cannot perform as well as before the Covid 19. Their income affects their ability to pay. For*

> *example, on average, our customers used to have the ability to pay IDR 2 million each year, but now they only can pay IDR 500,000 each year. If this data is fed to our current algorithm, then most of our customers will be classified as bad customers. We need to make modifications by providing a lower amount of lending with longer tenure to maintain customers' ability to pay. We also modify the algorithm to match our new policy and data.*" *– DS13.*

Data scientists need to monitor and evaluate regularly to see whether or not the data is still relevant. Data scientists must know when and how data becomes irrelevant and outdated and make justifications on the algorithms accordingly. This shows that data scientists must always make judgments about the algorithms even though they have been implemented and deployed in the systems.

The open-endedness of data also affects the algorithms' relevance. When there are changes in problems and data, a resilient algorithm should be able to automatically adapt. Therefore, data scientists must monitor the relevance of the algorithms for a certain period:

> "*In the first 30 days after deployment, I monitor the algorithms every day. What if there is new data, would the algorithm still be relevant? If I see that the algorithm can adapt to new data and problems, then when I think it works well. I don't have to monitor it every day.*" *– DS13.*

By monitoring the algorithm's performance and justifying the changes in the algorithms, data scientists maintain the validity of the algorithm's outputs so the algorithms grow and stay relevant over time.

### 4.5 Summary of the analysis

In conclusion, to extract valuable insights from harnessing data, data scientists take ownership of the ambiguity of data by doing validation as a process. Data scientists perform bundles of tasks to validate the problem, the data, and the

algorithms continuously and iteratively. Data scientists' validations produce legitimacy to the outputs in each phase of a data science project. In the process, data scientists also need to deal with the differences between data scientists' and business teams' interpretations and expectations of the data. This tension shows that data scientists' validation is as open-ended as the data because what they have accepted as valid does not stay valid forever. Data scientists must keep making a judgement as a continuous and iterative process throughout the data science project. Data scientists need to deal with this tension by performing discursive work, such as brainstorming and negotiating, with the business teams. Doing data science is not only about performing robust calculative work. Often, data scientists require subtle approaches to harmonise the data scientists' and the business teams' interpretation of the data.

# CHAPTER 5

# DATA SCIENTISTS' ESPOUSED AND ENACTED OCCUPATIONAL IDENTITY

## 5.1 Introduction

In the previous chapter (Chapter 4), I analysed the doing of data scientists in navigating the process of extracting insights from data. To continue the analysis, in this chapter, I am focusing on "the being" of data scientists. This chapter is the second part of my analysis that focuses on answering the second research question about the influence of data on the data scientists' occupational identity. Based on the empirical evidence, I found a latent identity tension that arises in the practices that data scientists enact, which contradict the identity which they espouse. Though they did not overtly acknowledge the tension, the contradictions between data scientists' enacted identity – the identity based on what they have to do – against their espoused identity – the identity based on the projection of what they think they try to do – spark an identity tension that becomes evident through the disconnect between their doings and who they aim to be. Data scientists espouse objectivity by being reliant on data. However, because of the limitless interpretation of data, data scientists need to enact subjectivity in extracting insights from data. Thus, to enact their roles, they need to put aside the goal of objectivity and make ongoing subjective choices.

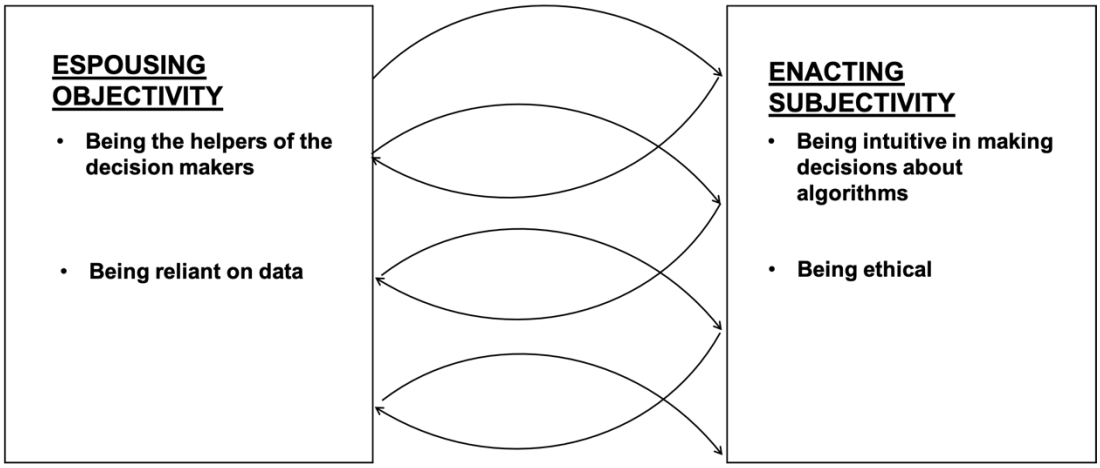


*Figure 6 The illustration of data scientists' identity tension*

In this chapter, I explore the two dimensions of data scientists' occupational identities and how they manage the tension between them. I create an illustration of data scientists' identity tension (Figure 6) which shows the ongoing tension between espousing objectivity and enacting subjectivity. I will explain the details of the identity tension in the following sections. First, I will discuss how data scientists espouse their identity as being objective. Then secondly, I will continue to discuss data scientists' enacted identity as being subjective. Lastly, I will explain the ways that data scientists manage latent tensions.

## 5.2 Espousing objectivity

Objectivity is something important for data scientists. As people who have the technical skills and knowledge to extract insights from data, data scientists espouse their identity as being objective. Data scientists try to use data as the basis of their work, especially data in numerical form. There are several ways to present this identity. In this section, I will explain the ways data scientists espouse being objective, such as by being the helpers of the decision makers

– rather than actually making decisions – and being fully reliant on data to direct their insights.

### *5.2.1 Being the helpers of the decision makers*

Data scientists have their ideal picture of who they are supposed to be. There are several pieces of evidence from the interviews that show how data scientists try to project their espoused identity. Considering the relational aspects of their work, data scientists tend to work for someone else – which can be other individuals, groups, or organisations. They analyse data to solve problems or achieve the goals of other people. One data scientist indicates this in the interview when defining who they are:

> *"In my opinion, data scientists are the people who help companies to increase business value - for example, efficiency and revenue - based on data." – DS13.*

The example shows that data scientists define themselves as people who help companies by working with data. This data scientist frames their identity as the helpers that contribute to the improvement of economic value. However, other data scientists point out that data scientists could contribute to other forms of value, depending on the objectives of the decision makers. One data scientist explained:

> *"My contribution depends on whom I am working with. If I am working at an e-commerce company, then the contribution is to help customers, sellers, or other people who use the platform. If I am working at a taxation office, then my impact is to help governments, citizens, or other organisations/parties that involve in the business process…In terms of value, I tried to classify value into three: resource, social, and experience…All of them depend on the organisations or people we are working with" – DS14*

According to my analysis, the most common phrase that data scientists use to define who they are is by saying that they are "the helpers of the decision-makers." The decision-makers can be anyone in the organisation who owns the problems or has the authority to execute the actions. A data scientist explains the examples of who the decision makers are:

> *"If you are working for the product team, then the decision maker is the product manager. If you are working for the strategic level, then the decision makers are the C-level."- DS16.*

Data scientists must know to whom they serve and what things are important for the decision-makers. Data scientists inform decisions, not only based on their mathematical skills nor their industry experience but also the context of the decision makers. It is necessary to ensure that their work is usable for the decision-makers as means to help them make decisions. Data scientists should align their work with the decision-makers. The decision makers vary across contexts which also shapes data scientists' identity. Data scientists position themselves as the people who work with data and use their work to help other people, particularly those who own key problems.

To be the helpers of the decision makers, many data scientists emphasise the disassociation of their identity with being the decision-makers, which they imply requires subjective judgments. They only want to give advice and leave the decisions to the decision-makers. One of the data scientists' statements indicates this when they are explaining who data scientists are:

> *"I think data scientists should not be the people who make decisions. We have to have an objective look at the data. When we are being the decision-makers, we have a stake in the decisions, right? Therefore, we can't have that objectivity in the data. So, I would say a data scientist is some kind of '*advisor*'. Our position is under the decision-makers. The decision-makers make decisions based on information*

> *and knowledge, don't they? So, data scientists are the people who supply the decision makers with the necessary knowledge in any required forms." - DS16.*

This statement shows that by defining their identity as the helpers of the decision-makers, data scientists espouse their objectivity, removed from the need to exercise soft judgments. The statement emphasises that objectivity is important for data scientists' professional reputation as people who extract insights from data. Data scientists thus protect their objectivity in performing their work.

As helpers for the decision-makers, data scientists see themselves as deriving meanings from data that other people cannot see. Therefore, they try to help people to interpret data the way they interpret it. They become a looking glass to help people apprehend the data scientist's objective interpretation of data. One data scientist vividly explained this:

> *"In my opinion, data scientists are the people who can see what is in the data that other people cannot see. You have all of these data in the world […] Data is generated daily. From this massive and messy data, data scientists should be able to generate knowledge or insights, and how those knowledge and insights help people to make decisions better." – DS16.*

The quote above shows the perspective of data scientists about themselves in relation to the open-endedness of data. "See things in data" means that data scientists believe they can understand and see the meaning behind the data. They acquire the ability to see the connection and pattern of data and thus extract the meaning from the pattern. Then, the claim continues by adding that data scientists try to see things that "…other people cannot see." This part of the claim shows data scientists' confidence in the uniqueness of their abilities

that other people do not have. With this unique ability, they can help other people to understand data the way they understand the data and navigate the process of extracting meaningful insights from data to improve the business.

Data scientists think that serving other people helps them to protect their objective look at data. They embrace their identity as the helpers of the decision-makers because it avoids them from compromising their neutrality in making decisions. For data scientists, being objective means trying to make decisions that do not only favour data scientists' biases. Like other occupations, biases are inevitable for data scientists when making decisions. One data scientist explains one example of the biases the data scientists generally have:

> *"Usually, data scientists tend to 'chase for shiny toys.' We tend to try to use sophisticated, complex, trendy, and state-of-the-art models. We want to chase a high accuracy. If we only care about the performance metrics, we will chase after those kinds of models." – DS6.*

One of data scientists' biases is the tendency to choose complex models to perform their analysis. Being able to use complex models increases their self-esteem to some extent. It challenges them and satisfies their curiosities. For DS6 who worked in a medium sized company whose performances are easily seen, using complex models can make DS6 appear to provide more valuable advice because they can create an impression that they perform more calculations behind the models. DS6 could gain a recognition of their mathematical skills. In addition, generally, developing an algorithm that produces a high value of performance metrics is of the greatest importance for data scientists. A high value of performance metrics indicates data scientists' performance.

But by helping other people, data scientists can prevent themselves from bias in making decisions since they focus on the business teams' concerns first. For data scientists, the business teams' concerns precede any other factors in their judgement. DS6 continued the explanation:

> *"There were sometimes when I developed a complex model. The deployment needs a more expensive server to run the model. While actually, without having a model that complex, I could already achieve the business metrics. So, the external metrics outside the data science team protect me from being biased in developing complex models." – DS6.*

If data scientists link about their biases to making decisions - for example, to chase complex models. When they are working with other people, what matters for data scientists is making a positive impact on the people they help.

### *5.2.2 Being reliant on data*

Data scientists espouse their objective identity by being reliant on data. The recommendations that data scientists' advise to the decision-makers are based on numerical data. They also produce recommendations using quantitative calculation methods, such as statistical or mathematical methods. The identity of being objectively reliant on data defines who they are. A data scientist indicates:

> *"I think because we work as data scientists, we are expected to see everything based on data. When we do something, we automatically think about: what do the data say? What data should we see? How to get the data? Our job trains us to always conform to data." – DS2.*

The statement above shows an example of other people's expectations of them. They think that other people see them as professionals who always think and do based on data. Data scientists are expected to always refer to data in performing their work. This expectation is manifested not only in their practice (as shown in Chapter 4) but also in their identity.

Working with data enables data scientists to be objective because they define data as an objective representation of the real world, a reflection of reality. A data scientist gave an explanation about this:

> *"What is important for us is to show data. We show people that this is what the data says […] We can't change the data because data reflects reality. If other people's data show different things, then maybe we have different sources of data." – DS2.*

Data is understood to show what happens in the world as it is. Therefore, being reliant on data is important because it enables them to project objectivity. By referring to data, data scientists can create a projection of identity around making their recommendation based completely on what the data says, and thus a mirroring of reality.

The identity of being reliant on data is shown in several quotes throughout the interviews. When data scientists are asked to define who they are, commonly, they explain that they are people who do everything based on data, with which they can perform complex calculations, often beyond the comprehension of others. A data scientist gave an example of how they describe who a data scientist is:

> *"If I can describe who a data scientist is, I guess it is a person who analyses the data, creates something valuable from the data that is not obvious, and derives some values from it using machine-learning algorithms, statistics - any method basically." – DS12.*

This example shows an image that data scientists are supposed to be able to work with data and create something valuable from data. They draw on data in order to produce rationality for the organisation that is unable to obtain objective insight on its own.

For data scientists, having phenomena transformed into numbers is very important. A data scientist explains the importance of transforming phenomena into numbers, rendering these phenomena into rational form:

> *"Data scientists work with numbers, so we try to collect numerical data. There are several conditions where we need to work with non-numerical data, for example, text. But, in the process, we transform texts into numbers. We can also create categorical data and transform them into numbers […] when we need to capture customer-experience data, we create ratings in numbers. I think almost everything can be transformed into numerical data. We just need to be more creative to collect and transform them." – DS2.*

Transforming phenomena into numbers is important for performing calculations. Data scientists need data in numerical form to fit them into the mathematical and computational approach. In analogy, numerical data is data scientists' raw material that will be processed using mathematics, statistics, and computational techniques to produce their end products (e.g., insights or systems). To do this, data scientists argue that they need to be creative in transforming the data. In their view of reality everything can be transformed into numerical data. With this belief, data scientists hink that they need to be particularly skilled and even creative in translating phenomena into numbers. This example shows that by defining their identity as being reliant on data, they project the ideal skills that data scientists should have.

Data scientists also use data as evidence to support their arguments. For example, data scientists rely on numerical data to define a classification of a phenomenon. Data scientists can build algorithms that can classify something as true or false and assign them into categories based on calculating numerical data. The algorithms can classify people, animals, things, or conditions by presenting them as numbers. The approach accomplished by

quantifying phenomena into numerical data and then calculating it using mathematical methods allows data scientists to be objective in their attempt to apprehend the truth. One example of this is credit scoring. A data scientist who worked in a bank gave an example:

> *"We developed a credit scoring algorithm to classify our customers. We classify them into two categories based on their score: low-risk lenders and high risks lenders. We define several variables to calculate the score." – DS13.*

The example shows that the data scientist chose variables and assigned representations using numerical data. Then, the numerical data is calculated to define the score of each customer. Data scientists define a specific number as a threshold to classify or categorise the customers. By calculating numerical data, data scientists can espouse their identity as being objective in classifying customers. They can define the value of each customer through the scoring and defining which customers the bank should retain and which ones it should not. The scores become evidence to indicate what the bank can accept as the truth.

### 5.3 Enacting subjectivity

While data scientists espouse their identity of being objective by making data a representation of the real world, the process of extracting insights from data shows that data scientists enact practices that involve subjectivity and intuition. This presents a latent tension between their espoused identity of being objective and their enacted identity as being subjective. Data scientists want to project their identity as being objective, but paradoxically, they can only reach that by being subjective in practice.

### *5.3.1 Being intuitive in making decisions about algorithms*

As explained in section 5.2.1., data scientists disassociate their identity from being the decision-makers. However, in practice, they need to make decisions to choose the problems, select the data or variables, and decide the thresholds to evaluate and accept in the algorithms. Data scientists cannot avoid being decision-makers. Their decisions about the algorithms affect the decision-makers' decisions. This tension is shown in one of the data scientists' statements in the interview:

> *"In my opinion, the decision-making process should not be a part of what data scientists do […] Even though, actually, I indirectly poise the decisions to a particular direction." – DS16.*

The quote shows the contradiction between the ideal pictures of what data scientists should do and what they really do. This person acknowledges that data scientists are driving the decisions in a particular direction. In order to be the helpers of the decision makers, data scientists must make judgements about the algorithm development. Moreover, they need to be intuitive in making decisions.

Contrary to data scientists' attempts to disassociate their identity from being the decision makers, data scientists enact intuitive work to make subsequent decisions in order to validate data and algorithms. For example, to develop algorithms that achieve economic value, data scientists choose which data and algorithms will help them achieve the value. A data scientist gives an example of how they choose the data by using intuition:

> "*For example, there was a time when a product team came to us with a question, "how do we increase customer conversion rate?" The product teams want to increase the number of people who visit our store and then make transactions. They came up with the hypothesis*

> *that if we increase the number of store visitors, then the conversion rate will increase. They think that the other parameters - apart from the number of store visitors - are constant. I think that's definitely incorrect. Their assumption is incorrect, and their analysis will go in the wrong direction. So, I explained to them what data they should look at. First, they should create a segmentation of the store visitors - for instance, which ones take a basket when they enter the store and which ones do not. People who do not carry a basket may not have any intention of purchasing anything. Maybe they just visit the store while waiting for a taxi. We want to increase the number of people who take a basket. Then, we can analyse the types of people who carry the basket. For example, we can see in that pattern that the majority of those people are women within a particular range of age. Most of them do not bring kids when they enter the store. And they usually come after 3 pm. This is the segment of customers that we should focus on. We should increase the number of visits of this segmentation of people, not everybody." – DS17.*

This example shows how data scientists direct their analysis by explaining what data or variables the business teams should focus on. They are revising the business teams' assumptions, which they consider as incorrect. Yet, data scientists subjectively perform an intuitive analysis into selecting the right data. Their intuition is developed based on their knowledge of the domain. In the example, the data scientist argued that they could not treat other variables apart from the number of visitors as constant. This argument can emerge because the data scientist understands the problem's context. They combine their intuition with analytical thinking to reveal which data is the most relevant and effective to answer the questions, as different data will point to different solutions. The intuition drives their analysis to bring out the pattern of the type of store visitors that are most likely to make a purchase. Then, they argue that the analysis should focus on this type of store visitor. The combination of domain knowledge and computation produces a convincing argument that can direct the analysis.

In practice, intuition is very important in doing data science. As the people who develop and build algorithms, data scientists need to use their intuitive judgement to make decisions. In a data science podcast, a Human Resource Specialist describes that having intuitive thinking is a required skill to be a data scientist:

> *"Data scientists need to solve problems with data, so technical skills are important. But the most important [thing] is having an intuition about the problem." – an HR Specialist (DCP community podcast).*

Because data is open to interpretation, data scientists need to be intuitive to construct and make sense of data. From all possible data, data scientists choose which data are relevant to the problems. In choosing the data, they need to be intuitive in gathering the options of relevant data. They also need to perform discursive work to reach agreements about their interpretation of the data. For example, section 4.2.1 indicates how data scientists are being intuitive to scrutinise the narrative problems presented by the business teams to identify and frame the real problems.

Making assumptions based on intuition is necessary to guide data scientists in being reliant on data - for example, to transform phenomena into numbers. Data scientists' knowledge about a particular domain is often limited. In these cases, their assumptions can fill in the knowledge gap to guide their interpretation. To understand how data scientists embed their assumptions in transforming phenomena into numbers, I will draw on an example given by a data team leader in a medium sized company when they label a training dataset. The team handled several different product segments and they wanted to create an algorithm to classify the target customers for "moms and babies" products. Therefore, their main target customers are mothers who

have babies. To develop this classification algorithm, they need to create a training dataset that contains labelled data of “moms and baby” customers:

> *“When we want to create a persona for moms that have babies, we need to know what kind of people fit this persona. We need to create a training dataset by labelling which people fit the persona and which ones do not. For example, we label “1” to the former and “0” to the latter. The model will learn to classify the people from the training dataset. The problem was we didn’t know how to label the training dataset […] What we did was create assumptions about the profile and behaviour of moms with babies. Making assumptions is difficult. We need some experience or knowledge to label the data. So, we tried to create assumptions as rigid as possible about what moms with babies usually do. If a person does a particular thing, then the person is most likely a mom with babies. On some occasions, we verified our assumptions by calling and asking a sample of people.” – DS17.*

This example shows that labelling data to numbers (e.g., 1 and 0) requires data scientists to create assumptions. Making assumptions involves subjectivity to define which people fall into the “moms and baby” category. To create the right assumptions, data scientists need to have some level of expertise or knowledge in a particular domain. Their argument about this requirement is shown in their statement which says “we need some experience or knowledge to label the data.” To label data by themselves, they rely on their knowledge or the relevant teams’ expertise in creating the assumptions. Data scientists can perform several kinds of verifications to ensure their assumptions are most likely correct. Some of them are aware that making the right assumptions is important to perform a correct classification. The example above also shows that in order to have data that can be used as the objective representation of reality, data scientists must, first, construct the data by embedding their subjective assumptions in the process. It shows that to be able to rely objectively on data, data scientists must enact subjectivity in constructing the data.

### *5.3.2 Being ethical*

In order to make objective decisions, data scientists also need to involve their subjectivity in considering the ethical implications of the algorithm's decisions. To be ethical, data scientists cannot disassociate their identity from being the decision-makers. They need to decide not only what is right or wrong but also what is good and bad. Moreover, to make such decisions, they again need to involve their intuition and judgement. They need to ensure that their interpretation of data does not bring negative consequences to the affected parties. However, the view of the importance of ethics is also contextual. For certain projects or problems or industrial contexts, some data scientists think that mitigating the risks of bringing unethical consequences is of great importance. But, their view might be different when working on different problems. A data scientist explained:

> "*For scientists, academia, and researchers, to be able to explain an algorithm in-depth is important. But for us, data scientists who use algorithms to serve decision making in businesses, we only need to explain the algorithms at a certain degree. The decision makers – the business stakeholders – do not need to know how the algorithms work in-depth. Most of the time the explanation is sufficient at the level of: what are the variables behind the results? […]But in several cases, data scientists need to explain how algorithms work in a deeper level. For example, there are several sensitive cases in healthcare, finance, and law […] we need to ensure to breakdown how the algorithms work to satisfy the decision-makers concerns about ethics."* – DS17

As the people who understand data and interpret them for the decision-makers, data scientists think they are responsible for addressing the risks. To help the decision-makers mitigate the ethical implications, data scientists need to make decisions that involve assumptions and judgement, and, in fact, in doing so, they cannot fully rely on data. To mitigate the ethical risks, data

scientists need to perform subjective work. Their assumptions become embedded in the algorithm's rules, which is contrary to their espoused identity of being objective.

One data scientist who works in an IT and recruitment service company explains an example of the bias in algorithms and where it comes from:

> "*I think bias is inevitable. It is embedded in the logic of the developers [of the algorithms]. For example, I developed an algorithm to score the candidates. We [the data scientists] selected variables to define the candidate's score. We chose skills and position levels. To calculate the scoring, we chose to weigh the position level more. So, the higher the candidates' position level, the greater their score. However, I realised that every company might have different situations or factors that affect promotions. Maybe there are people who have excellent skills but have no opportunity to promote to a higher position level. Maybe there are people who do not have a clear career path, so they are stuck at the same level. So, I think we should have considered other variables to be weighed more - for instance, years of experience, how long the candidates have the skills or the frequency of applying particular skills. We were biased toward people who have higher position levels. So, yes, I guess bias depends on who develops the algorithms*." *– DS2.*

Some data scientists are aware that their algorithms contain biases that come from themselves (i.e., their assumptions). Their data selection manifests the biases of the data scientists. In DS2's industrial context, ethics are very important to match potential candidates with the right job, especially in scoring the candidates. The example shows that DS2 was aware that by weighing the candidates' position level more than other variables, the algorithm may put some people at a disadvantage. There are a lot of factors that can affect a person's promotional journey. There may be a candidate that is not in a high position, but this does not mean that the person is not qualified or talented. Failing to recommend a highly qualified candidate to the recruiters is not only

a disadvantage for both the candidates and recruiters as the key platform users but also for the company for its reputation and the customers' trust.

Being aware of their biases, data scientists need to consider a lot of aspects to ensure the correctness of their assumptions. The data scientist continued the explanation:

> "*Our previous assumptions for the scoring are only from the supply side [the candidates]. We also need to align it with the demand side [the recruiters]; which variables the recruiters think are more important. But in the end, we [data scientists] are the ones that decide the weighing*." - *DS2.*

In this example, the algorithms will be used to help the recruiters find the best candidates for them. So, the recruiters' preferences matter. When the recruiters are happy with their preferred candidates, then the algorithms fit the recruiters. However, the above quote also shows an interesting statement. The data scientist acknowledges that even though their assumptions consider other people's perspectives, the data scientists themselves are still the ones who choose the assumptions. The platform users or customers' biases will only influence data scientists' decisions. They do not define the decisions. In the end, data scientists acknowledge that they are the decision-makers who decide what will be included and not included in the algorithms.

Another contradiction between data scientists espoused and enacted identity is in their view about being reliant on data. With the fact that data scientists enact subjectivity in being ethical, data scientists are also aware that they cannot be fully reliant on data. Being completely reliant on data can make them fall into unethical conduct. One data scientist – who is working at an IT service

and transportation company– explains an example of their concerns about the ethical consequences because of data:

> *"One of our teams' concerns is the cost of misclassification. We should always be careful of what might happen if the algorithms make a wrong classification when applied in the real world. In my case, I created an algorithm to suggest a 'quick reply' on the drivers-customers' chat room. I wanted to add gender-based salutations in the quick reply - for example, to reply 'Okay, Sir' or 'Okay, Madam.' The [data science] problem is that we need to predict the gender according to names because we can't access the personal data of the drivers and customers. There is a huge risk of misclassification because most drivers are male, maybe 95% of them. Our classification might be skewed to males. We need to think about what will happen if the algorithm suggests an incorrect salutation." – DS6.*

Classifying gender is not a simple task, especially when data scientists do not have the data or cannot have access to the data about the preferred gender. The data scientist in the example took 'names' data to predict the gender and classify them. This is where the potential unethical consequences emerge. By making the prediction rely only on names, the algorithms will get a higher risk of misclassification. Because the majority of the drivers are male, the algorithms will be more likely to predict the drivers' gender as male. The cost of misclassification this example mentions will affect not only the driver or customers but also the company. Incorrect salutations might give negative experiences to drivers or customers, which may affect the company's image - for example, in terms of inclusivity. Therefore, data scientists must carefully make assumptions to predict the ethical risks if they use certain data.

Data scientists can mitigate against the ethical consequences that may happen. They need to use their intuitive thinking to find the best way to mitigate the risks. A data scientist explains an example of how he mitigated the risk of misclassification:

> *"We mitigate the risk in various ways. First, we determine an acceptable threshold to ensure that the probability of our prediction is accurate. In that case, we set a threshold of 99% accuracy. If the accuracy of the prediction is above 99%, then the algorithm suggests a gender-based salutation; otherwise, the salutation is gender-neutral. Secondly, we use the second line of defence by double-checking our prediction using National Identification data. However, some gender data in the National Identification is inaccurate, so we can't rely on that data 100%."- DS6.*

The data scientists use two layers of mitigation. The first layer sets a threshold for validating the algorithm, and a second layer is a heuristic approach by benchmarking with other data. In this case, the threshold was defined through negotiation and discussion with the product team and legal team. The threshold was chosen through comparison. The data scientists tried to run the algorithms using several thresholds and chose the ones that made them most confident. Data scientists subjectively think about the possible consequences for the company - for example, the misclassification could make the customers or drivers share their unpleasant experiences on social media. That is one of the reasons they agreed to use a very strict threshold (e.g., 99% accuracy) to avoid the risk. In addition, the data scientists performed a second layer of mitigation by benchmarking the algorithms using gender data in the National Identification card. They used the National Identification card data only as the benchmark data because they think they cannot rely on gender data in that card either. The gender data in National Identification might not represent the subject's real gender. Nonetheless, they made a heuristic approach to mitigate the potential risks of only relying on 'names' data. This example shows a contradiction between data scientists' espoused identity as being reliant on data and the way they mitigate ethical risks in practice. Data might not be an

accurate representation of reality, so, to be ethical, data scientists aim to exercise judgement to assess the ethical implications and mitigate them.

### 5.4 Managing the inherent identity tension

The tension between being objective and subjective shows that data scientists can only project their identity as being objective by enacting subjectivity. Data is supposed to eliminate the problems of subjectivity, but it does not yet solve this. In contrast, the nature of data that is open to interpretation requires data scientists to be subjective when extracting useful insights. Objectively making decisions based on data can be achieved by performing subjective work - for example, intuitively choosing the problems, data, and algorithms. According to the empirical data, data scientists are not actually aware of this inherent tension and the tension is not something that data scientists struggle to solve. The tension between their espoused and enacted identity is not explicitly stated in the empirical data; rather, it is identified theoretically through my analysis.

My analysis also shows that data scientists naturally develop some strategies to decouple identity tensions. Their strategies enable them to maintain their fragmented identities in a harmonious way. Their strategies enable them to be the helpers of the decision-makers while also making decisions. They could also be reliant on the data to see it as the representation of the truth while also acknowledging and mitigating the ethical consequences of the data selection. Those strategies are categorised into two: being metrics-oriented and delegating decision-making to algorithms.

### *5.4.1 Being metrics-oriented*

The contradictions between what data scientists espouse – being objective – and what they enact – being subjective – show an inherent tension in data scientists' identity. Data scientists need a way to handle the contradictions. My analysis found that data scientists naturally handle the tensions by orienting their subjective judgement and interpretation toward quantitative indicators of measurement, which data scientists often call 'metrics'. Because of the limitless possibility to interpret data, metrics play an important role in measuring the value of data scientists' work quantitatively. For example, data scientists think that they should extract value from data. As explained by one of the data scientists:

> *"To sum it up briefly, I think a good data scientist is one that can extract value in many ways. What measures the quality of a data scientist is the value that they extract from data." – DS16.*

Creating value is important for data scientists. However, the value itself can be defined subjectively, based on who interprets it. Data scientists seek a quantitative measurement to define the value. Therefore, the metrics – which are defined by the decision makers or by the data scientists themselves - act as the measurement of value. The metrics represent not only the value that data scientists create but also the data representativeness of the phenomena of interest. In this section, I will describe how being metrics-oriented can solve data scientists' inherent identity tensions.

First, I will draw on the contradiction between how data scientists disassociate their identity from being the decision-makers against what they enact. As described in Subchapter 5.3., although data scientists want to be independent of being the decision makers, they acknowledge that they point the decision

makers' decisions in certain directions. Data scientists cannot be fully independent of decision-making. However, by referring to metrics, data scientists can protect their objectivity in directing decisions. One data scientist explained how metrics protect their objectivity:

> *"One of the things that protects my objectivity is the business metrics. If the data shows that my work can help a business reach their targeted metrics, then my duty is fulfilled." – DS6.*

Business metrics become the goals that data scientists use to legitimise their work. Every judgement that data scientists perform in their decision-making must be made to help organisations achieve the targeted business metrics. Moreover, by having the business metrics in numerical form, they can show the value of their work objectively. For example, as explained in Chapter 4, data scientists try to make decisions based on the potential economic value that organisations can gain. They perform hypothesis testing, using data to see the potential value and then making decisions based on the results. The calculations that they perform in doing hypotheses-testing protect their objectivity. The numerical data represents the economic value of data scientists' work and becomes hard evidence that defines data scientists' success.

By showing the metrics, data scientists can stay objective in directing decisions. For some data scientists, the numerical value of metrics is seen as a fact. This perspective enables data scientists to create convincing arguments by referring to the metrics. A data scientist said in the interview:

> *"Data is clear evidence for us. For example, we want to improve customer satisfaction. With data, we can measure satisfaction in the form of numbers. Then, we set the targeted number as a goal. How much satisfaction do we want to increase in percentage? It becomes*

*something that we can measure. If we can reach the targeted percentage, then we can say that we are successful." – DS2.*

Metrics act as quantitative measurements to produce numerical data that, later, data scientists treat as evidence of what happens in the real world. The data becomes the hard fact that data scientists use to make decisions - for example, to validate their work and make those decisions. As shown in Chapter 4, data scientists measure the performance of algorithms to validate the algorithms that they will deploy. They make this decision based on the performance data. Data scientists adopt a set of statistical metrics to define the performance of their algorithms, such as classification accuracy (commonly called 'accuracy'), logarithmic loss, confusion matrix, area under the curve, F1 score, mean absolute error, and MSE (Table 7). Metrics standardise the definition of value that is expected by data scientists and decision-makers. By showing those metrics, data scientists can stay objective in giving advice and recommendations.

*Table 6 The definition of statistical metrics to evaluate algorithms*

| **Statistical Metrics to Evaluate Algorithms** | **Short Definition** |
|---|---|
| Classification Accuracy | The ratio of the number of correct predictions to the total number of input samples. |
| Logarithmic Loss | The measurement of error by penalising false classification. |
| Confusion Matrix | The output matrix that shows the performance of the model by classifying cases that fall into true positives, true negatives, false positives, and false negatives. |
| Area Under Curve (AUC) | The probability that a classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. |
| F1 Score | The harmonic mean between precision and recall that tells how precise and how robust a classifier is. |
| Mean Absolute Error | The average of the difference between the original values and the predicted values. |
| Mean Squared Error (MSE) | The average of the square of the difference between the original values and predicted values. |

The second contradiction that data scientists handle with metrics is their espoused identity of being reliant on data against what they must do in practice in selecting the data. In section 5.3. I found that to be reliant on data, data scientists enact subjectivity in selecting data and weighing the data. Data scientists make assumptions to perform those tasks. Enacting subjectivity is inevitable to get the work done, even though it does not reflect their view of seeing data as an objective representation of reality. However, metrics can help data scientists to sustain this view. An example by a data scientist illustrates this:

> "*In deciding the variables and the weighing, we went back to the metrics (e.g., the accuracy). Usually, the business teams will ask us to revise the algorithm when they are unsatisfied with the metrics. So, we modified the algorithm and tried it with a dataset of hired candidates. For example, the accuracy of the old scoring algorithms is 70%, while the new algorithm brings 90% accuracy. It means that with the new scoring algorithms, the candidates that have a high score are more likely to be hired. Then, we decided to apply the new algorithm*."- *DS2.*

The example shows that data scientists, together with the decision-makers, choose data and thresholds that produce the best metrics score – e.g., accuracy. The metrics help the negotiation process that data scientists and the decision-makers undertake. For example, the C-Level management might need performance metrics as proof of the company's data capability to attract investors. Therefore, data scientists must align the thresholds with the C-Level's desired metrics. The decisions will depend on the metrics that the decision-makers think of as important.

Yet. the quantitative metrics that data scientists achieve from their work hide the subjective nature and positions of the findings as objective. A data scientist

gives an example of how data scientists' work is accepted based on quantitative metrics:

> *"There are several ways to evaluate the output from data scientists' work. For example, the accuracy of the algorithms […] and the business metrics. If my algorithms achieve the targeted accuracy or help the business teams reach their targeted metrics, then most likely, my algorithms will be used." – DS6.*

The quote shows an example of how the end results of algorithms are validated and accepted when they reach the targeted metrics. Even though the process of extracting insights from data is done through intuitive work, the findings are accepted objectively by organisations because they are supported by quantitative metrics. The metrics act as an objective measurement of the success of their work and put distance between themselves and their findings. Through this, data scientists can maintain their position to objectively guide the decision-makers by giving them the confidence to speak and make decisions based on data.

### *5.4.2 Delegating decisions to algorithms*

Besides being metrics-oriented, data scientists can maintain the discrepancy of their identities harmoniously by delegating decision-making to algorithms. The decisions that the algorithms make are computed automatically based on numerical data. Data scientists can maintain their objectivity by being independent of the decisions that the algorithms make. With more advanced machine learning and artificial intelligence methods, data scientists can train computers to perform cognitive tasks and, thus, give computers agency to make decisions. In this section, I will explain how delegating decisions to algorithms helps data scientists stay objective, despite having to enact subjectivity in practice.

First, I will focus on the tension in disassociating data scientists' identity as decision-makers against the inevitable enactment of subjectivity. In developing the algorithms, data scientists inevitably embed their biases (as described in section 5.3.2.). For example, data scientists need to decide the data, the algorithms, and the weight of variables. Despite this subjective enactment, data scientists can stay objective by imputing computer agency and letting them make decisions. For example, data scientists can let machines decide the classification of data based on the given variables. In the interview, a data analyst in an IT service and transportation company gave an example. This data analyst had experience working in a team with data scientists. He gave an example of a project where they developed an algorithm to automatically identify fake orders based on the receipts that the drivers uploaded to the application:

> *"We developed a '*receipt detection' *model. This model is used to identify fake orders based on the receipts that the drivers uploaded to the app. We trained the model to identify the images of valid receipts. We created a scoring model to classify which images are valid receipts and which ones are not. For example, if the images of the receipts are blurry, too dark, or have random objects, then the algorithm will classify them as invalid receipts. […] If the algorithms detect that a driver uploaded several invalid receipts in a day, they will classify the drivers as fraudsters who submitted fake orders and automatically suspend the driver's account."- DA5.*

The example shows the implementation of machine learning to identify fraud automatically. The computers are trained by feeding them with labelled data – in this case, labelled images. The algorithms are also given the authority to classify the drivers as fraudsters if the algorithms detect a certain number of invalid receipts from the drivers. The invalid receipts are evidence that indicates a fraudulent practice. The logic behind the algorithms creates a

condition wherein if there are no images of valid receipts, then the algorithms cannot treat the order as valid. So, the algorithms classify the orders as fake. Letting the algorithms detect the frauds and classify the drivers enabled data scientists, as the developers, to be objective. Classifying drivers as fraudsters can bring an emotional impact on the related parties. However, by doing it automatically by machines, data scientists could ensure objectivity because the classification is done with minimum human interventions.

Seeing the algorithm as a black box helps data scientists maintain their objectivity. The limitation to understanding the explainability of the machine to produce decisions protects data scientists. However, on some occasions, data scientists need to try and explain how the algorithm works. A data scientist gave an example:

> *"I understand that it will feel weird if, for example, an algorithm predicts that I will not be able to pay the debt without knowing how the algorithm comes to such a conclusion […] In some cases, explainability is important, so what I did is to try and explain the algorithms as much as possible. For example, I explain what variables are fed to the algorithms. Now, there are also other algorithms that can be used to explain how an algorithm works. Sometimes I use them if needed. But in the end, the level of explainability that I tried to explain depends on what the clients ask." – DS14.*

In the case when data scientists need to explain how the algorithms work, data scientists could also use another algorithm to perform this task. By delegating this task to another algorithm, data scientists also decrease the human intervention in the process of unpacking the algorithm's black box. Thus, data scientists can stay objective because they do not provide much intervention. Moreover, the quote above describes that the level of explainability data scientists need to provide depends on the clients. It shows that, sometimes,

the explainability of the algorithms is not necessary for completing their job. Thus, data scientists can appear objective in light of the complexity of the algorithms.

Another inherent tension is the contradiction between being reliant on data and the inevitable subjectivity to judge the ethical risks of data selection and weighing. Algorithms can make decisions according to the threshold numbers that are defined by data scientists. For example, data scientists can define thresholds to assign the data into different categories. A data scientist in a ride-hailing company gives an example of how they use quantified thresholds to categorise drivers' genders by predicting the genders based on names:

> *"We gathered the data of drivers' names, and then we developed a predictive model. We set a threshold of 99% accuracy. For example, if our algorithm predicted that Driver A is 99.99% male, we will accept the prediction that Driver A is male because it is higher than the threshold. If it falls below the threshold, we will treat Driver A as gender-neutral." – DS6.*

By categorising data using the threshold number, data scientists use quantified data to decide which data belongs to which category. The threshold can define which unit of data is categorised as black or white, right or wrong, and class A, B, C or D. The example shows that if the prediction is not accepted as true, then they will classify the driver's gender as "gender-neutral". The decision to set '99% accuracy' as a threshold was made by considering the ethical risks of misclassifying the gender. As described in section 5.3.2., data scientists cannot rely completely on data because the data (e.g., gender data) might not represent reality. Therefore, data scientists need to use their subjective judgement to mitigate the risks.

However, it does not reflect their view of data as something they can rely on to be objective. So, there are other ways to maintain this view - for instance, by adding more automation in the process of data labelling. Automating some data science processes is more achievable for data scientists working in large multinational companies, which are more likely to have access to more variety of data across nations and larger volumes of data. For example, they could gain more 'reliable' data to make computers learn to classify data automatically. A data scientist who worked in a large multinational company explained how it works:

> *"Using the neural network, we can train the algorithms to learn to classify data. The more data that we feed to the algorithms, the smarter they are." – a data scientist (from the observation of a technology company's webinar, 2020).*

The algorithms can give better performance if they are fed with more data. This data scientist could leverage their access to data to automate the process of data feeding.

Using machine learning, data scientists could train algorithms to identify images and automatically classify them. A member of the AIC community explains this in one of their webinars:

> *"We could use many features to identify the image of an object. Therefore, we could build an automatic way to determine the critical features to identify - for example, cats or other objects. We need a way to determine the features scalably." - a data scientist (from the observation to the AIC community webinar, 2020).*

This example shows how data scientists train machines to automatically identify objects and classify them into defined categories - for example, images of cats. The example shows evidence that data scientists could reduce human

intervention in the process. The greater the portion of tasks being conducted by algorithms; the more objective data scientists could espouse to be.

The examples in this section depict how access to more variety of data and a larger volume of data could allow data scientists to delegate more tasks to algorithms to understand and define what truly happens in the world. For example, using numbers, data scientists let the algorithms define a value of a thing or a person, decide what or who is right or wrong, and classify them. Data scientists could project more objectivity by appearing to have less intervention in algorithms' decisions.

### 5.5 Summary of the analysis

The open-endedness of data influences the way data scientists define their identity. Data scientists espouse their identity as being objective by means of being the helpers of the decision makers and being reliant on data. They can project their objectivity by being independent of the interests of the decision-makers. They also always give recommendations and advice based on what the data says. However, the practice they enact shows a contradictive identity. Data scientists are being subjective in performing their work. Enacting subjectivity behind the scenes is necessary in order to realise their espoused identity of being objective. They should be intuitive in making decisions about algorithms and be ethical in mitigating the consequences of the algorithms that they build. In managing this tension, data scientists naturally develop strategies that keep the tension invisible and maintain the fragmented identities harmoniously. They become metrics-oriented to measure the value of their work numerically and provide pieces of evidence of the validity of their

advice. Data scientists also give agency to algorithms to make decisions which protect their objectivity.

# CHAPTER 6

# DISCUSSION

## 6.1 Introduction

This chapter discusses the findings from my analyses in Chapter 4 and Chapter 5. I discuss the findings by tying them together with the literature. I examine how the openness of data interpretation influences 'the doing' and 'the being' of data scientists. Using this lens, I have brought up two findings in the previous chapters: 1) the validating process data scientists do to navigate the open-endedness of data and 2) the tensions between data scientists espoused and enacted identity. In this chapter, I discuss the two findings to answer the research questions more thoroughly and discuss how they relate. In the first section, I discuss how the open-endedness of data influences the doing of data scientists. I have specifically clarified the conceptualisation of data as being open to interpretation (Monteiro and Parmiggiani, 2019; Alaimo and Kallinikos, 2020). My research adopts this concept of data and studies it in the process of doing data science. According to my empirical analysis, data scientists develop practices to address the challenges and seize opportunities to extract insights from data. I theorise these insights by drawing on the relevant literature.

In the second section, I focus on the being of data scientists. Particularly the influence of the open-endedness of data on the data scientists' occupational identity. My research contributes to previous studies (Vaast and Pinsonneault, 2021) by showing that data creates an occupational identity tension for data

scientists. I discovered that the open-endedness of data brings challenges and opportunities for data scientists as an emerging occupation to shape what they do and who they are. The occupational perspective to study IS has hardly been adopted within the IS field. Therefore, my research contributes to the understanding of data science phenomena.

Next, in the third section, I explain the relationship between what data scientists do and who data scientists are. By discussing the two findings at a higher level of analysis, I explain how the validating process that data scientists do shape and reflect their espoused and enacted identity. This level of analysis produces the theoretical model that illustrates my theorisation.

## 6.2 The doing: Data scientists perform an open-ended validating process.

Prior studies indicate that the nature of data influences the practice of doing data science (Gray, Gerlitz and Bounegru, 2018; Muller *et al.*, 2019; Seidelin, Dittrich and Grönvall, 2020). Studies have also discussed how algorithm users develop practices to deal with the vagueness of data and algorithms (Monteiro and Parmiggiani, 2019; Mikalsen and Monteiro, 2021). They perform ongoing negotiations to build trust in their algorithms (Passi and Jackson, 2018). My study adds to this conversation by clarifying the process of how data scientists deal with the opacity in building and applying algorithms based on their perspective – how they tell what they do. My findings in Chapter 4 enhance this conversation by showing that the open-endedness of data puts data scientists in a position that requires them to engage in performing ongoing "validation". Data scientists need to make decisions and navigate the process

of data production, processing, distribution, and utilisations to make data, and consequently, their work, accepted as valid.

As explained in Chapter 2, the construction of data is interpretive thus data is non-objective and contentious (Aaltonen and Pentinnen, 2021; Neff et al., 2017; Lebovitz, Lifshitz-Assaf and Levina, 2021; Waardenburg, Huysman, and Sergeeva, 2022). Data undergoes a transformational journey to be used as a 'reference for the truth' (Bates, Lin and Goodale, 2016). By unpacking data scientists' validation process, my research clarifies how data can become the reference of truth from the data scientists' perspective as the experts that underpin algorithmic technology. It also contributes to clarifying the hiddenness of data scientists' work process in how they deal with the opacity in algorithm building. It becomes important to understand this process because there has been an increasing reliance on algorithmic technology for critical decision-making (Duan, Edwards and Dwivedi, 2019; Yang, Steinfeld and Zimmerman, 2019; Strich, Mayer and Fiedler, 2021). In this section, I conceptualise that data scientists' validation process is as open-ended as data and requires their judgement, shaping how people see data and subsequently act based on it.

I conceptualise data scientists' validation as an ongoing act of making judgements. Going back to its literal meaning, according to Oxford Dictionary (2021), validation is *"the act of proving that something is true or correct"*; "*the act of stating officially that something is useful and of an acceptable standard*". Data scientists judge and approve all aspects of the data science process to

produce useful and acceptable results according to the defined purposes. Data scientists perform ongoing validation to navigate the open-endedness of data and gain ongoing insights from data. The open-endedness of data brings vagueness to the overall data science process (O'Neil and Schutt, 2014). Therefore, data scientists must not only perform validation at one point in time (i.e., after selecting the data to ensure its correctness). Instead, they need to continuously perform validation to the problems, data, and algorithms to clarify the vagueness across a data science process continuum. The validating process is how data scientists - as the algorithm developers - deal with the opacity of working with data.

Data interpretation is contextual, so the meaning of data never stays the same and is not absolute (Pink et al., 2018). Data is constructed, designed (Feinberg, 2017; Seidelin, Dittrich and Grönvall, 2018), and validated through ongoing validation. Nevertheless, data that has been validated might not remain valid permanently. As explained in Chapter 2, we know data is the product of cognitive work. Assessing validity of data relevant to the problem is also part of cognitive work that constructs the meaning of data. As shown in the empirical findings of Chapter 4, in some contexts, data scientists need to gather other data in the process because the initially selected data is no longer relevant to the current situation. Data scientists need to modify the algorithm's thresholds to treat data differently according to the changing problem. For example, my empirical finding shows that when Covid-19 occurred, the algorithms became less relevant to solving problems. The Covid-19 pandemic changes the context of the problems. The meaning of data can become

irrelevant, and so do the algorithms. Data scientists must modify the algorithms and change the data labelling to fit the shifting context.

The contextuality of data necessarily makes data scientists' validation open-ended. The essence of doing data science is to try to represent the world through data and bring the outputs (e.g., insights, predictions, recommendations, classifications) back to the world to let users interact with them (O'Neil and Schutt, 2014). The interaction between the users and the outputs from data science creates new data that could bring new insights. The constantly changing context and dynamically growing data being produced make data scientists need to modify their judgements. To construct data as the representation of the truth (Lebovitz, Levina and Lifshitz-Assaf, 2021), data scientists need to accommodate the endless potential of data interpretation in the changing world. In this way, data scientists create and maintain the coherence of the new data required to meet evolving and dynamic real-world problems.

The contextuality of data (e.g., the organisational context, the industrial context, and the methodological context) also creates various ways for data scientists to perform the validation. For instance, as shown in the analysis, data scientists who work in companies with other various data occupations (e.g., data analysts, data engineers, and BI analysts) might classify and frame data science problems differently from data scientists in companies with limited data occupations. The former might have more specialised problems than the latter. The various ways of doing validation based on the contextuality

of data show that data scientists' validation is performed differently depending on the data scientists' judgement on the given context.

From the knowledge generation point of view, making judgements through the validation process is vital for creating new knowledge from data for organisations. Yet, the validation process shows that there are opacities in the process of doing data science that are not easily understood by data scientists, as well. In the knowledge generation process, it is common for knowledge workers to face uncertain situations, which makes it necessary for them to make judgments or inferences using their intuition (Tsoukas, 1996). In the context of data scientists, data offers many uncertain situations. Even though data science offers a scientific approach to making causal inferences for new knowledge, subjective judgements are always needed to complement the process. Their judgements make up the hidden elements of their knowledge work that are hard to scrutinise. Subjective judgements are necessary for the process of knowledge generation to deal with uncertainty (Nonaka and Toyama, 2005). The contextual and dynamic interpretations of data constantly create uncertainty, leaving scope for data scientists to make judgements.

Data scientists' judgements are used to create a logical consistency through stories that validate congruence between the problem, data, and algorithms. For example, at the beginning of the project, data scientists define the problem statements they can, should or will solve. These formulated problems then determine what kind of data will be needed, in what forms, and how the data moves (i.e., how to collect and store the data). By validating problems, data

scientists also shape their, and those of others, purposes. When data is seen as correct, it is because it is being seen from a mindset framed according to the problem being validated. To maintain a consistent story, data scientists constantly *fact-check* to see the fitness of data with the story. Data scientists need an understanding of the problem to review and guide when the process should go back to repurposing, re-selecting the data, or modifying the algorithms.

In order to ensure that they work is accepted as valid, data scientists need to gain the 'deliberative accountability' (Passi and Jackson, 2018). Data scientists seek to ensure that the other related actors in the organisation agree that they the results from their algorithms bring significant impact to organisation's goals. Data scientists perform the validation process dialogically by involving related teams who have knowledge about the problems. Data scientists validate knowledge through dialogue with the business teams. The dialogues with the related actors or teams shape the agreed perspective and understanding of data – which imbue meanings to data – and situate the validity of the algorithms according to the defined organisational goals.

In relation to organisations' common rationality, data scientists' validation process contributes to unpack how data scientists make their work practices important while managing to keep any seeming 'irrationality' hidden. Organisations' goals approaching rationality are mainly economic goals (Simon, 1990). Therefore, in evaluating the performance of algorithms, data scientists perform benchmarking to choose the algorithm that brings the

highest score on metrics they have agreed upon with the clients. The common practice in evaluating algorithms' performance based on quantitative measurement is related to modern organisations' rationality. The numerical value of the metrics can become an indicator of rational progress (Koch *et al.*, 2021) since making decisions based on quantitative measurements has been understood by organisations as a way of making rational decisions (Walker *et al.*, 2008). Even though data scientists engage in subjective judgements in building algorithms, the quantitative metrics makes their subjective judgements unapparent. The metrics also measure their contribution to organisations, thus making the significance of their work evident.

Making judgements in the validation process can come with consequential ethical concerns. Because the validation phases are interrelated, when the process's end-product is validated, it validates the whole process, including all the hidden biases and interests ingrained within data scientists' judgement. For example, in terms of the benchmarking practice in choosing algorithms, many scholars in critical data studies criticise it as inadequate to be used as the only practice to evaluate algorithm performance (Raji *et al.*, 2021). Algorithms benchmarking naturalise datasets and models (Miceli, Schuessler and Yang, 2020), thus giving them the representational capacity, which renders subjectivity and biases invisible. People will then more readily accept that data and algorithms represent the truth. Unintended consequences might appear if data scientists prioritise in making decisions merely to optimise the metrics. When data scientists are unaware of the ethical consequences of their

biases, there are groups of people who might experience unfairness from the deployment of the algorithm (Buolamwini and Gebru, 2018).

Understanding data scientists' validation as a process gives us an enhanced perspective to see that, because validation is continuous and iterative, the biases embedded by data scientists' judgements are leveraged in the whole process. Data construction is biased and subjective (boyd and Crawford, 2012), but in the end, data is often accepted and used as an objective representation of truth because it has been validated. The logical storylines that data scientists deliver can validate the biases embedded in the process (Paullada *et al.*, 2021, p. 3). Consequently, data scientists need to pay attention to the biases hidden by the validation process.

### 6.3 The being: Data scientists' two-fold identity as a way to gain authority.

Chapter 2 shows that many studies show that data scientists face ambiguities in defining their identity. The ambiguities create various identity tensions (Avnoon, 2021; Vaast and Pinsonneault, 2021). In Chapter 2, I also made a point that the way data scientists manage to gain authority to work despite having ambiguous identity remains unclear. In this section, I explain how considering the influence of data - in particular, the fact that data lend themselves to multiple interpretations - brings another perspective about the ambiguity of their identity and how they respond to the ambiguity to gain a sense of control over their work.

Previous studies about data scientists' occupational identity have shown that data scientists face several identity tensions. Vaast and Pinsonneault (2021) have explained how digital technology influences data scientists' identity tensions. The ever-changing digital technology creates an ambiguity in the persistence of data scientists' occupation. Data scientists are threatened by the development of digital technology that requires newer skills. The ambiguity creates an identity tension between 'obsolescence against persistence' (Vaast and Pinsonneault, 2021). By focusing on digital data that plays a central role to data scientists' work, my research brings another perspective to the study of data scientists' identity. My study finds the open-endedness of data brings the ambiguity to data scientists' work. The multiple interpretation of data makes data scientists need to navigate the interpretation in order to achieve the desired goals. Because of this ambiguity, data scientists gain an identity tension. Data scientists espouse their identity as being objective, while enacting their identity as being subjective, in constructing data and algorithms. The interpretation of data is the product of humans' sensemaking that involves subjective judgements. The dynamic interpretation of data requires data scientists to rely on their knowledge and intuition to interpret data creatively.

To espouse their identity, data scientists rely on the common perceptions that view data, especially quantitative data, as objective and reliable (Kemper and Kolkman, 2019). Identity is shaped through self-narratives (Christiansen, 1999). Data scientists use this common perception of data to create consistent narratives, promoting an objective identity. Data scientists portray themselves as being reliant on data and acting based on it; they create recommendations

based on data. By being reliant on data, data scientists could mirror the narrative of data objectivity to their identity. Even though data is constructed by humans' interpretation, once data is validated, the meaning embedded is seen as facts. Therefore, validated data feigns the objectivity of data scientists, making data scientists' objectivity appears unchallenged.

My findings contribute to the growing conversation about data open-endedness. As explained in Chapter 2, apart from having digital material characteristics (e.g. mutable, editable, recombinable), digital data has a distinct material nature as cognitive elements (Alaimo and Kallinikos, 2020). Data can be seen as signs or tokens representing facts or reality and containing ideologies (Alaimo and Kallinikos, 2020). Data is constructed and produced by humans' interpretation and sensemaking. As cognitive elements, data shape data scientists' occupational identity differently from other digital material. Data has no specific ends in terms of interpretation because its meaning or interpretation is dynamic to the ever-changing context. This nature of data brings opportunity for data scientists to construct meaning of data while also challenging data scientists to make judgements in making sense of and validating data constantly. The material nature of data as products of cognitive work requires data scientists to be intuitive in responding to unfamiliar problems or contexts. The fluidity of data as cognitive elements challenges the common perspective that sees data as objective. It makes enacting subjective work inevitable for data scientists to make subsequent decisions to make inferences from data to gain knowledge and solve their problems.

Although, in general, data scientists face a tension between objectivity and subjectivity, the way data scientists define who they are and what they do is contextual. For instance, the analysis shows that the decision-makers' context matters in defining data scientists' identity. For example, in business organisations, data scientists define themselves as people who help decision-makers improve economic value. However, the forms of value vary depending on the context. Another example is ethics. In certain contexts, ethics are seen as more important. Acknowledging how data scientists define who they are and what they do is important to understand that there are many ways of explaining data scientists' subjectivity and how they project their objectivity.

It is important to highlight that human interpretation of data is not a contaminant to data; rather, it is necessary to make sense of data (Neff *et al.*, 2017). The material nature of data shows that data is a product of cognitive interpretation. Therefore, data scientists need to make a judgement and use their intuition to make sense of data to make data meaningful. Especially in the era where data is produced in a much larger volume, the challenge for organisations is also finding meaningful knowledge (Manesh et al., 2020). Data scientists' subjective interpretation is necessary to extract meaningful knowledge for organisations. Their subjective judgements are also important in mitigating the ethical consequences that may follow from deploying the outcomes of their work back to the world. However, the subjectivity that data scientists enact is kept hidden. They must fit with the organisations' expectations of how organisations could gain knowledge.

My study adds to the study about organisational rationality by showing how it has a slight effect in shaping data scientists' identity. I build on this point by emphasising that the commonly adopted scientific management in organisations' rationality in decision-making (Nonaka and Toyama, 2005), creates scope for data science to help them make decisions with larger and more varied data using computational and statistical methods. When organisation rationality is limited, for example, in solving new or complex problems, organisations tend to rely on rules (Agrawal, Gans and Goldfarb, 2019). Data scientists offer the improvement of organisational rationality to solve problems by embedding rules in machines. This view indicates that objectivity is more valued in organisations. Organisations avoid dealing with subjectivity and aim to achieve objectivity using science to connect facts (Nonaka and Toyama, 2005). The algorithms that work based on rules help organisations think they achieve objectivity. With this expectation, data scientists must conform to the rationality accepted by organisations.

With this rationality, organisations are comfortable when they can quantitatively measure their performance and productivity, especially regarding financial results (Maister, 2000). What organisations want to see is the quantified value of the algorithms. Data scientists satisfy the organisations' desires by presenting the quantified metrics of the algorithms' performance (e.g., the accuracy, precisions, etc.) and how the algorithms can contribute to the economic values. The subjective judgements that are made in the process are taken-for-granted and remain hidden because they are not acknowledged to the organisations. Data scientists must show that they are being objective

to present a face of rationality, so the organisations have the confidence that they are being objective in making decisions based on quantifiable measurements.

The findings about data scientists' espoused and enacted identity adds to the studies about authority. The unclarity of ones' occupational identity makes the authority to perform work hard to gain (Brown *et al.*, 2010). Studies have indicated that the discrepancy between the espoused image and enacted practices is quite prevalent among organisations as the result of the interaction of micro and macro factors (Bromley and Powell, 2012; Leroy *et al.*, 2022). The discrepancy between them acts as a mechanism to secure legitimacy (Bromley and Powell, 2012). My study builds on this point at the individual level to argue that the shaping of data scientists' twofold identity is a response to manage the ambiguity of data open-endedness. Data scientists align with the organisations' rationality to secure the authority in their work. The discrepancy between data scientists' espoused and enacted identity makes certain parts of their tasks 'de-coupled' - the parts in which they are making subjective judgements in their choices. They make judgements, but it is hard to trace the connection between these judgements and the 'objective' algorithms they build. Thus, they could claim authority by maintaining this disconnect.

Consequently, data scientists could gain authority as the 'rulers' whose rules are embedded in the algorithms. In Chapter 5, I have shown that data scientists tend to distance themselves from being the decision-makers when describing their identity. Just like other professions, data scientists have their

own scope in making decisions, therefore it is important to unpack what decision-maker means according to data scientists. My findings show data scientists tend to describe decision-makers as the people who make the ultimate decisions of the organisational strategies and actions. However, consciously or unconsciously, the influence of data scientists' decisions become more prevalent and powerful in driving the organisations' decision-makers because it is embedded as rules that are deemed rigorous and objective.

The discussion about how the open-endedness of data shapes data scientists' occupational identity provides an insight about their authority in creating influential rules. Understanding how their authority is put into practice can be done by connecting data scientists' identity to their validation process, which is discussed in the following section.

### 6.4 The rulers in the shadows: Data scientists' twofold identity enables data scientists to make decisions in the validation process.

In the previous sections, I have discussed two main findings of my research. First, data open-endedness influences data scientists' doing by making data scientists perform an open-ended validation process. Second, in terms of being, data scientists keep their subjectivity hidden and present themselves as objective instead. This section will discuss how the two main findings are related. Anteby (2010) highlights that work practice is intertwined with occupational identity. Work practices can shape occupational identity, and identity can be enacted or realised through work practices (Ashcraft, 2007).

By drawing on those arguments, this study shows that what data scientists do and who data scientists are, are related to each other and are both affected by the open-endedness of data.

I created a figure of the model that illustrates how data scientists' doing and being are connected with each other (Figure 7). Because my research analysis is mainly based on what data scientists say, my analysis could identify that data scientists mostly think they need to make decisions in developing algorithms. In contrast, they think, ideally, they need to distance their identity from being the decision-makers to protect their objectivity. Therefore their validation process is tied to their identity tension.

I have illustrated the data scientists' identity tension in a model, as shown in Figure 6 in Chapter 5. Data scientists' identity tension creates a spiralling relationship with each other because data scientists can only espouse objectivity by enacting subjectivity. By adding the validating process from Figure 5 in Chapter 4, Figure 7 shows that identity tension happens along the validation process. The spiralling relationship between the espoused and enacted identity occurs continuously during data scientists' activities to validate the problems, data, and algorithms. The model shows that each phase in the validating process is iterative and goes back and forth. When this iterative process is mapped to the model of data scientists espoused and enacted identity, the model in Figure 7 reveals how the two-fold identity is continually produced throughout the validating process.

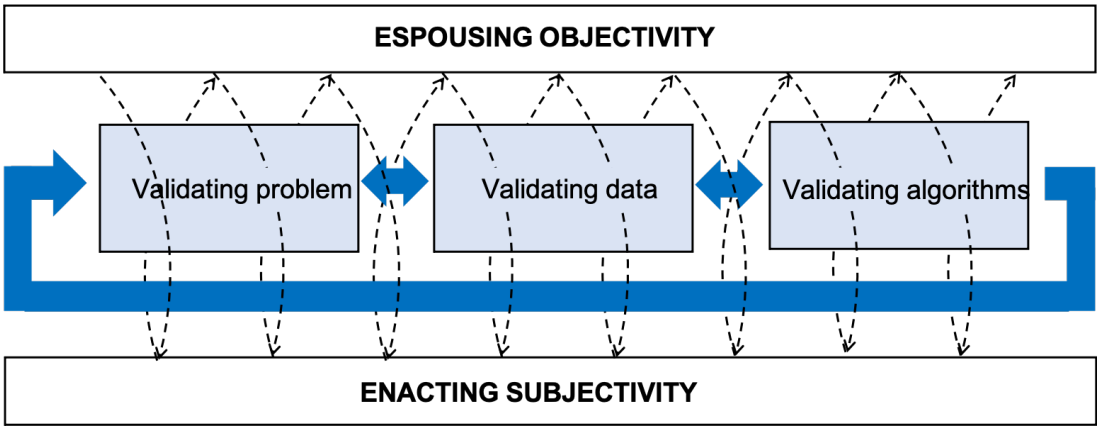


*Figure 7 Data scientists' two-fold identity is intertwined with the validation process.*

Data scientists' identity becomes justified during the validation process. The open-endedness of data brings ambiguity to data scientists' work practices and occupational identity. Therefore, just like how they want data to be accepted as valid, data scientists need to get the work done while justifying their rational, objective identity to continue rendering their work acceptable or valid. Espousing objectivity is necessary for data scientists to fulfil the organisation's expectations and achieve their professional goals as helpers of decision-makers. However, despite wanting to espouse their identity as objective through their judgements, data scientists will always fall into enacting subjectivity. They will appear to be subjective over time because their judgements are never static. It is contextually bound and the context constantly changes. Consequently, data scientists' twofold identity is formed through a tension of remaining objective toward the data and being subjective in their interpretation.

The necessity to appear objective shrouds the enacted subjectivity, and I suggest this can be linked to shadow work (Beane, 2019; Bharatan, Swan and Oborn, 2022). Literature indicates that technology implementation creates grey areas where practitioners need to enact certain practices in shadows to be able to escape notice and thus be freer to make mistakes when experimenting with technology to achieve their goals (Beane, 2019). In the data scientists' context, the subjective judgements are left in a shadow, where subjectivity is enacted hidden, often shrouded in the complex and vast data produced and mathematical precision which attracts attention, even admiration. Much of data scientists' subjective work remains opaque, particularly to non-data scientists like business owners. The production of facts as the result of data scientists' work relies on this shadow so that the latent tension of ongoing subjectivity remains hidden. Organisations hope to become increasingly rational through the work of data scientists that produces the narratives of science, logics, and facts to approach the "truth". The hiddenness of their subjectivity enables data scientists to gain a reputation for being objective, as expected by organisations. Data scientists make their work useful to organisations' decision-making; as a result, they claim the legitimacy to influence organisations' decisions.

As such, I suggest data scientists become the *rulers in the shadows*; they could make decisions and impose their subjective judgements in the shadow of the conception that sees data and algorithms as objective. As explained in Section 6.3, data scientists gain authority as the rulers whose decisions are influential and embedded in the algorithms. But their authority as the rulers

remains unapparent due to the hiddenness of their subjective judgements (As explained in section 6.2). They could have soft power in driving the decision-makers' decisions by subtly telling them what to do based on what the data says. The dominant perspective that sees data as an objective representation of reality enables data scientists to have this decision-making power without challenging their objectivity. As shown in empirical findings in Chapter 5, data scientists delegate power to the algorithms they build to make decisions. Data scientists tend to create more automation for algorithms to reduce human intervention, thus giving more "objectivity" to algorithms to appear accountable for making fair decisions.

Data, which is central to data scientists' identity, gives a strong entitlement (Amo *et al.*, 2022) to data scientists to make decisions. Data scientists' decisions on data are influential in guiding organisations to extract meaningful knowledge. Even though data scientists distance themselves from being seen as people with the power to make decisions, in the end, data scientists' judgements influence the decision-makers' approval. This hidden decision-making authority enables data scientists to validate what is worth knowing from the openness of how data can be interpreted, for example, what problems are worth solving and what data should be selected. Data scientists also can decide what actions they and the organisations need to take to acquire their desired knowledge, for example, brainstorming the possible relevant data and creating a coherent data story.

Data scientists' decision-making authority impacts the wider society. I argue that data scientists emerge as a new power elite among knowledge workers. Knowledge workers are generally expected to use their knowledge to navigate vagueness in solving problems. Other knowledge workers, e.g., management consultants, also exercise decision-making power using their knowledge (Gill, 2015). However, data scientists have a bigger influence because of the conception of data as the objective representation of reality. They shape what people believe as reality through the data and algorithms they construct. Their judgements are ingrained in the algorithms, which are pervasively used to shape the ideas and behaviour of society (Faraj, Pachidi and Sayegh, 2018). Conforming to the belief that data is objective empowers data scientists to wield extensive influence in society. Identity shaping enables data scientists' decision-making power to grow and establish the organisation and society's trust. Data scientists are given the authority to manipulate and modify data, creating a new power in the digital age of capitalism in shaping how people feel, think, and create meaning through algorithms that help companies reach their economic goals (Zuboff, 2019).

The discussion indicates the implication of my study to study the hidden power exercises in doing data science. There is a need to pay more attention to the exercise of power in data science (Miceli, Schuessler and Yang, 2020; Denton *et al.*, 2021; Ehsan *et al.*, 2021). There has been a conversation about the relationship between identity and power. Identity can show how individuals exercise power in organisations (Foldy, 2002; Alvesson, Ashcraft and Thomas, 2008). My study contributes to this literature by explaining how

studying occupational identity could explain how data scientists gain their hidden decision-making power. Moreover, my study also connects this topic to the open-endedness of data that enables data scientists to gain decision-making power.

The emergence of data scientists' hidden decision-making powers through their identity raises questions about other data occupations. Data might influence the identity of other professionals who work similarly to data scientists. Previous studies have examined the occupational identities of various knowledge workers, such as advertising agents (Alvesson, 1994), management consultants (Gill, 2015), and accountants (Goto, 2021). However, they have not yet touched on data occupations, which are the primary knowledge workers in the age of AI. The insights from my research can also be the stepping-stone to extending the study of data scientists to understand data occupations better.

## 6.5 Summary of the discussion

This chapter discusses the empirical findings I have provided in Chapter 4 and Chapter 5 by explaining how my study answers the research questions. The discussion is done by explicating the findings with theories to identify how my research contributes to the existing theories.

First, my research theorises that data scientists perform an open-ended validation process. Open-ended validation implies that the act of making judgements throughout the validation process embeds the biases in the algorithms that are validated. The biases embedded in the algorithms might

bring unintended ethical consequences when the algorithms are deployed in the real world - which warrants attention for data scientists to mitigate the consequences.

Second, I theorise that data scientists acquire a two-fold identity which is shown by how they espouse objectivity and enact subjectivity. The discussion of this theorisation identifies the influence of the nature of data as cognitive elements that require data scientists to enact subjectivity in interpreting data and the organisations' rationality that naturally makes data scientists' subjectivity hidden. Also, how data scientists gain the authority as the 'rulers'.

Then, I also discuss the relationship between data scientists' validation process and their identity tension by creating a model to illustrate that their identity tensions constantly emerge because their identity is attached to the activities in the validation process. The discussion of this theorisation identifies an indication of the hidden decision-making power exercise based on their identity. Data scientists enact subjectivity in making decisions about the interpretation of data, then their espoused objectivity hides their decision-making power through numbers and automation. Data scientists could be the rulers in the shadows, enabling them to make decisions in validating problems, data, and algorithms without challenging their objectivity. Data scientists' two-fold identity is not only a necessary consequence of data scientists' practice but also the enabler for them in gaining the authority to make decisions in the validation practices.

My theorisation then implies future research by opening up the conversation about data scientists' profession to a broader topic. My findings offer insights into the implications of data scientists' doing and being for other issues, which is further discussed in the next chapter.

# CHAPTER 7

# CONCLUSION

## 7.1 Introduction

In this chapter, I will summarise the contribution of my research. First, I will start by summarising the overview of what I have done in conducting the research. Then, I will explain my research contribution according to the theoretical and the practical contribution. Next, I will also point out the limitations of my research. I will reflect on how my research is designed and how it is conducted. Lastly, I will explain the directions for future research to expand and improve the study of the data scientist profession.

## 7.2 Summary of the research

As the main worker in the big data and AI age, data scientists encounter vagueness and ambiguity regarding their profession (Saltz and Grady, 2017). There is no single standardised definition of what data scientists do and who data scientists are. Organisations hire data scientists and develop the job descriptions according to their understanding - i.e., all activities related to extracting insights from data. Data scientists are also figuring out their profession - they often feel their job does not match their expectations.

The vagueness and ambiguity of the data scientists' profession call for research that examines the work and occupational identity of data scientists. Several studies have started to examine data scientists' occupational identity (Avnoon, 2021; Vaast and Pinsonneault, 2021). The increasing influence of

data scientists and algorithms they develop indicate a stronger imperative to understand how they work, define their identity, and how they gain authority. Drawing on the assumption that data is open-ended (Monteiro and Parmiggiani, 2019), this study examines data scientists' work and occupational identity by focusing on two things: first, how data scientists navigate the open-endedness of data to extract valuable insights, and second, how the open-endedness of data shapes data scientists' occupational identity.

I adopted two lenses to study data scientists' work and identity. The first lens is about "doing", which helps me to theorise what data scientists do to navigate data. Second, the lens about "being" helps understand the data scientists' identity. Previous studies about doing data science indicate that the nature of data creates ambiguities in doing data science (Tanweer, Fiore-Gartland and Aragon, 2016; Passi and Jackson, 2017; Pink *et al.*, 2018). Therefore, using the doing lens, my study adds to the conversation to examine how data scientists develop practices to overcome ambiguities. Then, my study also connects what data scientists do with their occupational identity. I draw on the current study about data scientists' identities (Avnoon, 2021; Vaast and Pinsonneault, 2021) to examine how working with data influences their identities.

To conduct the study, I used a qualitative research approach by conducting semi-structured interviews with data scientists and several other occupations that work alongside them. I also did an online participant observation of some data science communities. I focus on learning how they explain what data

scientists do and how they perceive who they are. I analysed the data using a grounded theory approach (Charmaz, 2014). I analysed the data by doing two layers of coding - initial coding and focused coding (Saldaña, 2016), writing analytic memos (Birks, Chapman and Francis, 2008), and performing a thematic analysis (Boyatzis, 1998) iteratively. I analysed the empirical data by making a constant comparison with the theories. Therefore, while analysing the data, I also did a theoretical sampling by gathering other empirical data to confirm the sensitivity of the theories to the empirical data. Lastly, by adopting the Gioia method for data structuring (Gioia, Corley and Hamilton, 2013), I produced two data structures that represent my findings of data scientists' doing and data scientists' being.

In the first chapter of my analysis (Chapter 4), I focus on examining what data scientists' do to navigate the open-endedness of data. I have built a process model to show that data scientists' engage with a validation process throughout a data science project. The validation process consists of three phases: 1) validating problems, 2) validating data, and 3) validating algorithms. I label the process as the "validation" process because data scientists need to make a judgement to decide what is accepted and not accepted in the process. They decide what problems to solve, the data selected, how it is interpreted, and what algorithms fit. In doing the validation, data scientists need to intuitively use their judgement to make decisions because the open-endedness of data offers ambiguities and uncertainties. Then, to make their judgement valid, data scientists complement it with the metrics they treat as the numerical evidence that measures their success. Data scientists need to engage with the validation process iteratively and continuously because the

interpretation of data is never fixed. Data scientists need to constantly validate to ensure that the problems, data, and algorithms are still valid to be deployed in a constantly changing world.

Then, in Chapter 5, I analysed the data scientists' being. I found that data scientists experience an identity tension between the identity they try to present to other people - their espoused identity - and the identity shown through how they act - their enacted identity. Data scientists want to show that they are objective beings. They are being objective by distancing their identity from being the decision-makers by saying that they are the helpers of the decision-makers. They also appear to act and argue by relying on data. However, my findings show that because data is open to interpretation, data scientists perform intuitive and subjective work, thus enacting their identity as subjective beings. They can also achieve the goals of being objective by being subjective in performing their work. The open-endedness of data makes data scientists gain a two-fold identity which consists of their espoused and enacted identity. I clarify how the empirical findings answer both of my research questions by conceptualising data scientists' doing and being regarding how data scientists navigate the data open-endedness.

Regarding the first research question, my research conceptualises that data scientists navigate data by constantly engaging with the validating process recursively. I conceptualise data scientists' validation as an act of making judgements. The open-endedness of data makes the validating process as open-ended as data, thus making data scientists constantly make judgements.

The validation process is needed to make logical narratives that represent rationality as organisations desire.

Then, in terms of the data scientists' being. Open-endedness of data shapes data scientists' two-fold identity, namely the espoused identity and enacted identity. In performing their work, data scientists experience an ongoing identity tension between espousing objectivity and enacting subjectivity because data is always open to interpretation. To fit with organisational rationality, data scientists' subjectivity remains hidden. It allows data scientists to gain authority in performing their work.

I also conceptualise the relationship between data scientists' validation process and identity tension. Data scientists' enacted identity as being subjective is attached to the activities of the validation process, and so is their espoused identity. The hiddenness of their subjective judgments and the conception that sees data as an objective representation of reality makes data scientists become the 'rulers in the shadows.' My research makes several contributions theoretically and practically. I will explain the contribution of my research in further detail in the next section.

### 7.3 Contributions

In this section, I will explain my research contributions to the theories and practices. I will divide this section into two subsections. First, I will describe what my research contributes to the theory. From a theoretical perspective, I hope to contribute to the IS field by enhancing the understanding of the data scientist's profession as the main knowledge workers in the age of big data

and AI. Second, I will explain the practical contribution of my research. I expect my research to bring insights to the practitioners who work as data scientists - or other people who do data science - and the organisations who hire data scientists and utilise their work to solve real-world problems.

### *7.3.1 Theorising data scientists' doing and being*

I study data science, which is a topical research theme with a growing interest in the IS field, using the occupational point of view. My study responds to the call (Provost and Fawcett, 2013) that emphasises the need to study data science by connecting it to other essential concepts to make data science work to serve businesses. Current studies about data science have been focusing on how data science can be implemented to solve problems in various contexts (Han, Liu and Zhang, 2016; Spruit and Lytras, 2018; Tung, 2019). Studying data science using an occupational lens is relatively new among IS scholars.

Therefore, my study adds to the growing conversation about data scientists' occupational identity in the IS field (Vaast and Pinsonneault, 2021). It aims to enhance the understanding of the social aspect of doing data science. My study brings another perspective in connecting the concept of "data open-endedness" with the concepts of work and occupational identity. By connecting those concepts from different knowledge domains, I can provide an alternative theoretical perspective to studying data scientists' profession from the lens of the theories of "data". My research enhances the theories about identity in the IS field by considering the study of work as a response to the call by previous studies (Barley and Kunda, 2001; Ashcraft, 2007).

Based on the discussion (Chapter 6), there are three main contributions from my study. The first contribution contributes to answering the first research question which is about the influence of data open-endedness on the doing of data scientists. My study has shown that, in dealing with the open-endedness of data, data scientists are engaging with continuous and open-ended validation in which they involve making subjective judgements. Data scientists are working in an environment that is constantly changing and fluid. Working with data makes data scientists need to deal with vagueness and ambiguities constantly. Thus, they constantly justify the data's relevance to the problems they are solving. My study highlights the increasing insight into how subjectivity remains hidden because the focus of doing data science is on the data and what the complex algorithms produce. The outputs and the technologies grab the attention given their novelty and immensity. Contributing to the studies that increasingly put attention on the subjective nature of doing data science, this research could provide a new or different lens to study transparency in producing algorithms.

The second contribution answers the second research question which asks about how data open-endedness shapes data scientists' being. This study has focused on the ways that the openness of data makes data scientists gain a twofold occupational identity. Data scientists enact subjectivity while espousing objectivity consistently along with the journey of data that is constantly being shaped and reshaped to stay relevant to the problem. Their twofold occupational identity is imperative in managing the ambiguities of data interpretation while also fitting with the organisational rationality. By having

their twofold occupational identity, data scientists could gain authority in making rules - that are embedded in the algorithms - which are deemed objective, despite involving subjective judgements in making the design choices.

Third, my research also contributes in explaining the relationship between data scientists' doing and being. This research has provided further evidence that occupational identity is co-constituted through the enactment of work practices. The discussion shows that data scientists' identity tensions constantly emerge as both the consequence and enabler of data scientists' validation process. With the hiddenness of data scientists' subjective work and the authority gained to create the rule of algorithms, data scientists become the 'rulers in the shadows'.

To clarify my research contribution and the future research directions, I created a table (Table 8) that summarises the theoretical concepts I contribute, what prior literature has explained about it, my research contribution, and the future research directions. To explore further, in this section, I will discuss the implications of the theorisation of my findings and how they can open up new questions for future research directions on several aspects as examples. I develop the connection to those aspects based on the insights from my findings and the current conversation among researchers that I consider relevant. Those aspects are not exhaustive, but they provide examples of how my study opens the link and connection to other concepts for future research about data scientists' professions.

*Table 8 Summary of theoretical research contribution and future research directions*

| Concept | Prior Literature | My Research Contribution | Future Research Directions |
|---|---|---|---|
| Data scientists' validation as a process. | Social science studies have a growing conversation about the non-objectivity of data science practices. Studies show that data science practitioners and users need to develop practices as workarounds in applying data science (Passi and Jackson, 2018; Aaltonen and Penttinen, 2021; Neff et al., 2017; Lebovitz et al., 2021; Waardenburg et al., 2022). | Adding to the conversation, my research draws on what data scientists say about their work. I identified that data scientists engage in doing validation process throughout the data science project, which includes: 1) validating problems; 2) validating data; 3) validating algorithms. The validation process is as open-ended as data as a process because it is punctuated with the data journey. Data scientists need to recursively perform validation to produce and maintain a valid result of their work. | ● Studying the validating process from the team or organisational unit of analysis.<br>● How do data scientists and organisations exercise power in doing data science?<br>● How is the trust from the algorithm users built towards data scientists' validating process?<br>● Can data be the representation of the truth?<br>● Studying data scientists doing and being using longitudinal approach and case studies.<br>● Considering a wide variety of data scientists. |
| Data scientists' twofold occupational identity | Studies about data scientists' identity have focused on the influence of the characteristics of digital technology on occupational identity (Vaast and Pinsonneault, 2021; Avnoon, 2021). | My research takes another lens to study the influence of data on identity. By seeing data as a process, my research found another data scientists' inherent identity tension: espousing objectivity versus enacting subjectivity. Data scientists need to constantly espouse objectivity while enacting subjectivity. Data scientists' identity is tied to data that is constantly being shaped and reshaped. | |
| Data scientists' twofold identity as the enabler of the validation process. | Work practices can shape occupational identity and identity can be enacted or realised through work practices (Ashcraft, 2007; Anteby, 2010). | Data scientists' identity tension emerges because data scientists continuously engage with the act of making judgements in the validating process that requires data scientists to espouse objectivity while enacting subjectivity constantly. | |

### *7.3.2 Practical contributions to data science*

My research also has practical implications. The data scientist profession is one of the most in-demand professions. However, this high demand for data scientists does not come with a clear understanding of who data scientists are and what they do. Therefore, my research aims to clarify both points to help practitioners understand the data scientist profession and be aware of what it means to do data science for data scientists, other professionals who do data science, and organisations.

From my findings, practitioners could take some insights into ethics in doing data science. Data science is not a purely objective practice. It involves subjective judgements in response to the vagueness and uncertainty of working with data. When data scientists engage in validation, they embed their biases in making decisions about the problems, data, and algorithms. The end products of data science reflect the judgement of the human that builds them. However, the subjectivity embedded in algorithms and where it comes from remains hidden. It creates a challenge to see who are responsible for the ethical challenges or the impacts of algorithms. Practitioners could consider acknowledging the subjectivity in the choices in designing the algorithms to make the process of doing data science more transparent.

Data scientists and organisations could also be aware of the authority they hold in shaping how people see the world and behave based on it. Because data is open-ended, data scientists can decide what kind of data they need to collect and how to use it, including biases in making the decisions. Data scientists and organisations come with a responsibility to use the power for

good purposes and to be ethical. With the authority to navigate the openness of data interpretation, it is up to data scientists and organisations about what kind of subjectivity the algorithms reflect on. Data scientists and organisational leaders could pay more attention to the accountability processes of building algorithms. For example, they could question more about why they use and collect certain data and the potential implications of their choices.

My research can also be a hint for organisations to build the strategy of incorporating data scientists into organisations in a better way. Understanding the issue related to occupational identity is important for organisations to develop a human resource strategy to improve employee well-being and work satisfaction (Ennals *et al.*, 2016). Understanding what data scientists do and who they think they are may give organisations a clue to allocate the optimum tasks and job placement to leverage the data scientists' potential. For example, by knowing the data scientists' identity tension, organisations can understand data scientists' expectations about who people think they are. It can be the basis for designing tasks suitable for data scientists and create a more precise occupational boundary for data scientists.

### 7.4 Research limitations

Reflecting on how I conducted this study, there are several limitations that I identified in this study. I found many of the limitations related to my research methods and the scoping of my research.

The first limitation is related to the data collection. As explained in Chapter 3, I collected the empirical data by doing semi-structured interviews and online participant observation in data science communities. Doing semi-structured

interviews gave me access to the narratives from data scientists about what they perceive about what they do and who they are. By doing the online participant observations of the communities, I could also gain access to the conversations, discussions, and activities among data scientists and other data workers. However, those are not the only data collection methods that I planned. I was planning to do a participant observation to observe how a data scientist team completes a project. Participant observation is needed to understand better what data scientists do. I could not perform it due to the Covid-19 restrictions. Many organisations were implementing the work-from-home schema, so it was hard to observe individuals in the teamwork because they do that in their home offices. Conducting a study in data science communities was my best option because they mainly do their activities online (i.e., in their group chats and webinars). Another limitation of this research is the unit of analysis. The access that I got was only to individual data scientists. My research did not conduct a study in a specific organisation. Therefore, I could not get the perspective from the organisational unit of analysis. This unit of analysis can provide a more strategic perspective that can enrich our understanding of the organising data scientist profession to achieve organisational purposes.

The second point of limitation is the scope. My study only focused on data scientists working in business organisations - or as (Ramzan *et al.*, 2021) call them, the business data scientists. There is a wide variety of data scientists who can be differentiated based on the contexts of where they work and the skill specialisation they possess (as explained in Chapter 2). Because my research focuses on business data scientists, I could gain insights into how

economic value (such as profit and cost efficiency) becomes a shared goal for data scientists. There could be other insights into the research being done to data scientists who work in other contexts, for example, government or policymakers, non-profit organisations, and healthcare. The research could also gather insights from various data scientists, for example, specialising in computer vision, natural language processing, and prescriptive analytics. I could not perform it in this research because extending the scope to study a wide variety of data scientists might need more time and effort that are limited for a PhD project.

### 7.5 Future research

Understanding data scientists' nature of work raises implications for other aspects of doing data science that researchers might need to pay attention to. My study covers a new area of research within the IS field. The topic of data scientists' work practice and occupational identity has not yet been well-theorised in the IS field. In general, my research brings a new perspective on the social aspect of data science regarding the professionals that do data science. The theorisation raises further questions and opens the conversation about data scientists to a broader perspective. By understanding what data scientists do and who they are, researchers can connect to various concepts to study different dimensions and perspectives in studying data scientists to develop new research questions. As shown in Table 8, the findings of this research can be expanded to create new ontological, epistemological, and methodological questions.

Future studies could study data scientists' doing and being from the organisational unit of analysis. A higher level of the unit of analysis allows researchers to study the other factors that influence data scientists' decisions in doing validation and the shaping of their identity. Also, there may be distinctions in the validating process between data scientists who engage in project work individually or as a team outside of organisational structures and those who are integrated into those organisations as members of existing functional units. The difference between the validation process performed by data scientists within and across organisations could be studied in more detail in future studies.

Future studies could take an institutional view in understanding how power between data scientists and the organisations are exercised. For example, studies could take the institutional view and use the 'decoupling' theory (Bromley and Powell, 2012) to understand how data scientists and organisations manage the discrepancies between the means and ends of doing data science. This lens of study could help researchers in identifying the environmental factors that affect the discrepancies and how data-driven organisations gain and maintain their authority.

The topic about algorithms users' trust is also worth exploring. The hiddenness of data scientists' work might indicate an issue in building clients' trust. Data scientists must appear professional in giving their recommendations. Therefore, they try to espouse themselves as being objective to gain users' trust because working with data makes them do subjective work. Current

studies have examined trust from the users' side. For example, some studies discuss how algorithms users develop their own way of seeing and utilising algorithms (Waardenburg, Sergeeva and Huysman, 2018; Mikalsen and Monteiro, 2021). It is important to study the topic of trust from the data scientists' side as the developer of the algorithms.

Future research can also ask an ontological question about data as the representation of the truth, i.e. "With the bias embedded in the process, can data be the representation of truth?" Studies could draw on previous studies that examine how the "ground truth" in algorithms are defined (Lebovitz, Levina and Lifshitz-Assaf, 2021).

Future research can also address my research limitation by doing a longitudinal participant observation by using other data collection and analysis methods. For example, future research can do a case study or multiple case studies in organisations. Doing case studies can help researchers to understand the process of creating algorithms. Researchers could also examine the interaction between data scientists and other roles in teams or organisations in building the algorithms. The case studies research can be used to examine various concepts in doing data science, for example, data-driven decision making (Provost and Fawcett, 2013), AI-assisted learning (Lai, 2021), and how the algorithm's product is deployed to the real world. Doing a case study can also help to extend the research to touch on other data workers, such as data analysts and data engineers, that often work together with data scientists and have interchangeable tasks and roles within organisations.

Another recommendation is to research a wide variety of data scientists. As I explained in the previous section, my research limitation is the scope of only studying business data scientists. Future research can explore a broader type of data scientists and include them in the scope. For example, the genomics data in drug development and citizen's personal data for policy making. The wide variety of data scientists can give the researchers other perspectives about the value (Fayard, Stigliani and Bechky, 2017) that other types of data scientists think are important for their professional goals. Different data contexts may create different tensions both for the validation process (i.e., their doing) and data scientists' occupational identity (i.e., their being). Also, researchers can explore the similarities or differences between different types of data scientists regarding the control to make decisions in building algorithms, their effort and approaches to being ethical, and their ways to gain the required skills and accountability.

# APPENDICES

## APPENDIX A: An Example of The Interview Guideline

This is an open-ended interview. The aim of this interview is to understand what data scientists do, what the tools and technologies that are related to their work, and how they make use of the learning resources in their work. The respondent is asked to answer the questions by reflecting on a project they have done. The general questions are:

Part I. Day-to-day responsibility

1. What do you do as a data scientist? What is your day-to-day job/responsibilities?
2. What is the term 'data scientist' based on your understanding?
3. What are the tools and technologies you use daily?

Part II. Reflection on a project (in this part, the respondent is asked to reflect on a project)

1. What kind of project was it? How long did you finish the project?
2. What was your role as a data scientist in the project?
3. Who did you work with and talk to in the project?
4. How did you coordinate with each of them?
5. What kind of tools and technologies did you use to complete the project? How did you use each of them?
6. How did you search for the data?
7. Did you use any learning resources to help you complete the project? What were they?

Part III. Occupational Identity and Boundary

1. Do you ever feel your job is redundant with the job of other occupations, such as data analyst, data engineer, ML engineer, etc.?
2. What makes the job of data scientist different from the other similar occupations (data analyst and data engineer)?
3. Do you feel the need to have a clear occupational boundary between you, as a data scientist, and other colleagues?

Part IV. Learning

1. Do you feel the need to constantly learn? And, why? (Update skills, sharpen tools, etc.)

Part V. Community

1. Do you join any professional community to support your work? And, why?

## APPENDIX B: Participant Information Leaflet

**Project information sheet:** Understanding How Data Scientists' Work Practices Shape Their Occupational Identity

**Researcher:**

**Supervisor:**

**Date:**

You are invited to act as a research participant for the above project. Your participation in this project is entirely voluntary. You may withdraw from participating in this project at any time, with no negative consequence to yourself or the organisation for which you work.

This research aims to understand how data scientists' work practices, as they work with data and algorithms, shape their occupational identity.

The project involves interviewing participants virtually through video conference.

Your involvement in this project will help the researcher to study the phenomena and complete the PhD project.

Participation in this project will involve being interviewed by the researcher named above on the theme of what it means for individuals to be data scientists and how data scientists shape their occupational identity.

It is not expected that you will experience any risks through participating in this project. Data will be anonymised from the start, with no names or specific positions recorded as part of the interview material. Your consent form will be stored in a locked office at the University of Warwick, and transcripts of interview data will be anonymised before being printed and stored in the same place. The transcripts will also be stored electronically on the lead researcher's password-locked laptop. All material may be destroyed after ten years from the completion of the research. The material from this research may be published. You can request a copy of the publication from the researcher named above.

Should you have any further questions about this research, please contact Febriana Wisnuwardani (*email address*).

You may also contact the University of Warwick Research and Impact Services, University House, University of Warwick, Coventry, CV4 8UW, UK. 02476575732 should you have the wish to make a complaint about the conduct of the researcher.

## APPENDIX C: Consent Form

**CONSENT FORM**

Title of Project: Understanding How Data Scientists' Work Practices Shape Their Occupational Identity

Name of Researcher:

Name of Lead Supervisor:

Date:

**Please initial box**

1. I confirm I have read and understand the information sheet dated (*insert date*) for the above study. I have had the opportunity to consider the information, ask questions of a member of the research team and have had these answered satisfactorily. ☐

2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. ☐

*3.* I understand that my information will be held and processed for the following purposes: to be analysed by the researcher for the purpose of completing the PhD research. ☐

4. I agree to take part in the above-named study, and I am willing to be interviewed and have my interview audio recorded. ☐

| _____ | _____ | _____ |
|---|---|---|
| Name of participant | Date | Signature |

| _____ | _____ | _____ |
|---|---|---|
| Name of Researcher | Date | Signature |

## REFERENCES


van der Aalst, W.M.P. (2014) 'Data Scientist: The Engineer of the Future', in K. Mertins et al. (eds) *Enterprise Interoperability VI*. Cham: Springer International Publishing, pp. 13–26. Available at: https://doi.org/10.1007/978-3-319-04948-9_2.

Aaltonen, A., Alaimo, C. and Kallinikos, J. (2021) 'The Making of Data Commodities: Data Analytics as an Embedded Process', *Journal of Management Information Systems*, 38(2), pp. 401–429. Available at: https://doi.org/10.1080/07421222.2021.1912928.

Aaltonen, A. and Penttinen, E. (2021) 'What Makes Data Possible? A Sociotechnical View on Structured Data Innovations', in. *Hawaii International Conference on System Sciences*. Available at: https://doi.org/10.24251/HICSS.2021.716.

Abbott, A. (1988) *The System of Professions*. The University of Chicago.

Abbott, A. (1993) 'The Sociology of Work and Occupations', *Annual Review of Sociology*, 19, pp. 187–209.

Abo, R. and Voisin, L. (2014) 'Formal Implementation of Data Validation for Railway Safety-Related Systems with OVADO', in S. Counsell and M. Núñez (eds) *Software Engineering and Formal Methods*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 221–236. Available at: https://doi.org/10.1007/978-3-319-05032-4_17.

Acemoglu, D. and Restrepo, P. (2019) 'Automation and New Tasks: How Technology Displaces and Reinstates Labor', *Journal of Economic Perspectives*, 33(2), pp. 3–30. Available at: https://doi.org/10.1257/jep.33.2.3.

Agarwal, R. and Dhar, V. (2014) '**Editorial** —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research', *Information Systems Research*, 25(3), pp. 443–448. Available at: https://doi.org/10.1287/isre.2014.0546.

Agrawal, A., Gans, J.S. and Goldfarb, A. (2019) 'Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction', *Journal of Economic Perspectives*, 33(2), pp. 31–50. Available at: https://doi.org/10.1257/jep.33.2.31.

Alaimo, C. and Kallinikos, J. (2020) 'Managing by Data: Algorithmic Categories and Organizing', *Organization Studies*, p. 017084062093406. Available at: https://doi.org/10.1177/0170840620934062.

Alvesson, M. (1994) 'Talking in Organizations: Managing Identity and Impressions in an Advertising Agency', *Organization Studies*, 15(4), pp. 535–563. Available at: https://doi.org/10.1177/017084069401500403.

Alvesson, M. (2001) 'Knowledge Work: Ambiguity, Image and Identity', *Human Relations*, 54(7), pp. 863–886. Available at: https://doi.org/10.1177/0018726701547004.

Alvesson, M., Ashcraft, K.L. and Thomas, R. (2008) 'Identity Matters: Reflections on the Construction of Identity Scholarship in Organization Studies', *Organization*, 15(1), pp. 5–28. Available at: https://doi.org/10.1177/1350508407084426.

Amo, L.C. *et al.* (2022) 'Technological Entitlement: It's My Technology and I'll (ab)use It How I Want To', *MIS Quarterly*, 46(3), pp. 1395–1420. Available at: https://doi.org/10.25300/MISQ/2022/15499.

Anteby, M. (2010) 'Markets, Morals, and Practices of Trade: Jurisdictional Disputes in the U.S. Commerce in Cadavers', *Administrative Science Quarterly*, 55(4), pp. 606–638. Available at: https://doi.org/10.2189/asqu.2010.55.4.606.

Anteby, M., Chan, C.K. and DiBenigno, J. (2016) 'Three Lenses on Occupations and Professions in Organizations: *Becoming, Doing, and Relating*', *Academy of Management Annals*, 10(1), pp. 183–244. Available at: https://doi.org/10.5465/19416520.2016.1120962.

Ashcraft, K.L. (2007) 'Appreciating the "work" of discourse: occupational identity and difference as organizing mechanisms in the case of commercial airline pilots', *Discourse & Communication*, 1(1), pp. 9–36. Available at: https://doi.org/10.1177/1750481307071982.

Ashcraft, K.L. (2017) 'The Glass Slipper: "Incorporating" Occupational Identity in Management Studies', *Academy of Management Review* [Preprint]. Available at: https://doi.org/10.5465/amr.2010.0219.

Avnoon, N. (2021) 'Data Scientists' Identity Work: Omnivorous Symbolic Boundaries in Skills Acquisition', *Work, Employment and Society*, 35(2), pp. 332–349. Available at: https://doi.org/10.1177/0950017020977306.

Bailey, D.E., Leonardi, P.M. and Barley, S.R. (2011) 'The Lure of the Virtual', *Organization Science* [Preprint]. Available at: https://doi.org/10.1287/orsc.1110.0703.

Bain, A. (2005) 'Constructing an artistic identity', *Work, employment and society*, 19(1), p. 22. Available at: https://doi.org/10.1177/0950017005051280].

Barley, S.R. (1986) 'Technology as an Occasion for Structuring: Evidence from Observations of CT Scanners and the Social Order of Radiology Departments', *Administrative Science Quarterly*, 31(1), pp. 78–108. Available at: https://doi.org/10.2307/2392767.

Barley, S.R. (1996) 'Technicians in the Workplace: Ethnographic Evidence for Bringing Work into Organizational Studies', *Administrative Science Quarterly*, 41(3), p. 404. Available at: https://doi.org/10.2307/2393937.

Barley, S.R. and Kunda, G. (2001) 'Bringing Work Back In', *Organization Science*, 12(1), pp. 76–95. Available at: https://doi.org/10.1287/orsc.12.1.76.10122.

Bates, J., Lin, Y.-W. and Goodale, P. (2016) 'Data journeys: Capturing the socio-material constitution of data objects and flows', *Big Data & Society*, 3(2), p. 205395171665450. Available at: https://doi.org/10.1177/2053951716654502.

Beane, M. (2019) 'Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail', *Administrative Science Quarterly*, 64(1), pp. 87–123. Available at: https://doi.org/10.1177/0001839217751692.

Bechky, B.A. (2003) 'Object Lessons: Workplace Artifacts as Representations of Occupational Jurisdiction', *American Journal of Sociology*, 109(3), pp. 720–752. Available at: https://doi.org/10.1086/379527.

Bechky, B.A. (2011) 'Making Organizational Theory Work: Institutions, Occupations, and Negotiated Orders', *Organization Science*, 22(5), pp. 1157–1167. Available at: https://doi.org/10.1287/orsc.1100.0603.

Bechky, B.A. (2020) 'Evaluative Spillovers from Technological Change: The Effects of "DNA Envy" on Occupational Practices in Forensic Science', *Administrative Science Quarterly*, 65(3), pp. 606–643. Available at: https://doi.org/10.1177/0001839219855329.

Bhagat, R. *et al.* (2018) 'Buy It Again: Modeling Repeat Purchase Recommendations', in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery (KDD '18), pp. 62–70. Available at: https://doi.org/10.1145/3219819.3219891.

Birks, M., Chapman, Y. and Francis, K. (2008) 'Memoing in qualitative research: Probing data and processes', *Journal of Research in Nursing*, 13(1). Available at: https://journals.sagepub.com/doi/abs/10.1177/1744987107081254 (Accessed: 14 October 2022).

Bharatan, I., Swan, J. and Oborn, E. (2022) 'Navigating turbulent waters: Crafting learning trajectories in a changing work context', *Human Relations*, 75(6), pp. 1084–1112. Available at: https://doi.org/10.1177/00187267211010366.

Black, L.J., Carlile, P.R. and Repenning, N.P. (2004) 'A Dynamic Theory of Expertise and Occupational Boundaries in New Technology Implementation: Building on Barley's Study of CT Scanning', *Administrative Science Quarterly*, 49(4). Available at:

https://journals.sagepub.com/doi/abs/10.2307/4131491 (Accessed: 13 October 2022).

Blaikie, N.W.H. (2003) *Analyzing quantitative data - from description to explanation*. California: SAGE Publications, Inc.

Bonter, D.N. and Cooper, C.B. (2012) 'Data validation in citizen science: a case study from Project FeederWatch', *Frontiers in Ecology and the Environment* [Preprint]. Available at: https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/110273 (Accessed: 14 October 2022).

Bowne-Anderson, H. (2018) 'What Data Scientists Really Do, According to 35 Data Scientists', *Harvard Business Review*, 15 August. Available at: https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists (Accessed: 11 October 2022).

Boyatzis, R.E. (1998) *Transforming qualitative information: thematic analysis and code development*. Thousand Oaks, CA;London; Sage Publications (Book, Whole). Available at: https://go.exlibris.link/nrzF1R6p (Accessed: 14 October 2022).

boyd,  danah and Crawford, K. (2012) 'CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication & Society*, 15(5), pp. 662–679. Available at: https://doi.org/10.1080/1369118X.2012.678878.

Breck, E. *et al.* (2019) 'Data Validation for Machine Learning', *Proceedings of the 2nd SysML Conference*, p. 14.

Bromley, P. and Powell, W.W. (2012) 'From Smoke and Mirrors to Walking the Talk: Decoupling in the Contemporary World', *Academy of Management Annals*, 6(1), pp. 483–530. Available at: https://doi.org/10.5465/19416520.2012.684462.

Brouillette, M. (2019) 'At the forefront of medical research: our cover authors share their hopes for their fields', *Nature Medicine*, 25(12), pp. 1800–1803. Available at: https://doi.org/10.1038/s41591-019-0692-z.

Bryman, A. (2012) *Social research methods*. Oxford: Oxford University Press.

Bucher, R. and Strauss, A. (1961) 'Professions in Process', *American Journal of Sociology*, 66(4), pp. 325–334.

Buolamwini, J. and Gebru, T. (2018) 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Proceedings of Machine Learning Research*, 81, pp. 1–15.

Burning Glass Technologies, BHEF and IBM (2017) *The Quant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market*. Burning Glass Technologies.

Brown, A.D. *et al.* (2010) '"Invisible walls" and "silent hierarchies": A case study of power relations in an architecture firm', *Human Relations*, 63(4), pp. 525–549. Available at: https://doi.org/10.1177/0018726709339862.

Carollo, L. and Guerci, M. (2018) '"Activists in a Suit": Paradoxes and Metaphors in Sustainability Managers' Identity Work', *Journal of Business Ethics*, 148(2), pp. 249–268. Available at: https://doi.org/10.1007/s10551-017-3582-7.

Charmaz, K.C. (2014) *Constructing Grounded Theory*. 2nd edn. SAGE Publishing.1 (Book, Whole).

Chibber, K. (2018) *The origins of the job title "data scientist"*, *Quartz*. Available at: https://qz.com/work/1435689/the-origins-of-the-job-title-data-scientist/ (Accessed: 11 October 2022).

Christiansen, C.H. (1999) 'Defining Lives: Occupation as Identity: An Essay on Competence, Coherence, and the Creation of Meaning', *American Journal of Occupational Therapy*, 53(6), pp. 547–558. Available at: https://doi.org/10.5014/ajot.53.6.547.

Corbin, J.M. and Strauss, A.L. (2015) *Basics of qualitative research: techniques and procedures for developing grounded theory*. Fourth. Los Angeles: SAGE (Book, Whole). Available at: https://go.exlibris.link/BdztdW0v (Accessed: 13 October 2022).

Cote, C. (2021) *What Is Data Science? 5 Applications in Business*, *Business Insights Blog*. Available at: https://online.hbs.edu/blog/post/what-is-data-science (Accessed: 11 October 2022).

Creswell, J.W. (2013) *Qualitative inquiry and research design - Library Search*. 3rd edn. Los Angeles: SAGE Publications, Inc.

Crosby, O. (2002) 'New and emerging occupations', *Occupational Outlook Quarterly*, 46(3), pp. 16–25.

Cukier, K. and Mayer-Schönberger, V. (2014) 'The Rise of Big Data: How It's Changing the Way We Think about the World', in M. Pitici (ed.) *The Best Writing on Mathematics 2014*. Princeton University Press, pp. 20–32. Available at: https://doi.org/10.1515/9781400865307-003.

Data Science Association (2020) *About the Data Science Association | Data Science Association*. Available at: https://www.datascienceassn.org/ (Accessed: 11 October 2022).

Davenport, T.H. and Patil, D. (2012) *Data Scientist: The Sexiest Job of the 21st Century*. Available at: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (Accessed: 11 October 2022).

Day, R. (2021) *Top 3 Assumptions and Expectations that are Tiring in Data Science*, *Medium*. Available at: https://towardsdatascience.com/top-3-

assumptions-and-expectations-that-are-tiring-in-data-science-999e0455b480 (Accessed: 11 October 2022).

Dedoulis, E. and Caramanis, C. (2007) 'Imperialism of influence and the state–profession relationship: The formation of the Greek auditing profession in the post-WWII era', *Critical Perspectives on Accounting*, 18(4), pp. 393–412. Available at: https://doi.org/10.1016/j.cpa.2006.01.010.

Denton, E. *et al.* (2021) 'Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation'. arXiv. Available at: http://arxiv.org/abs/2112.04554 (Accessed: 13 August 2022).

Denzin, N.K. and Lincoln, Y.S. (2018) *The SAGE handbook of qualitative research*. Los Angeles: SAGE.

Dhar, V. (2013) 'Data science and prediction', *Communications of the ACM*, 56(12), pp. 64–73. Available at: https://doi.org/10.1145/2500499.

D'Mello, M. and Sahay, S. (2007) '"I am kind of a nomad where I have to go places and places"… Understanding mobility, place and identity in global software work from India', *Information and Organization*, 17(3), pp. 162–192. Available at: https://doi.org/10.1016/j.infoandorg.2007.04.001.

Donoho, D. (2017) '50 Years of Data Science', *Journal of Computational and Graphical Statistics*, 26(4), pp. 745–766. Available at: https://doi.org/10.1080/10618600.2017.1384734.

Dourish, P. and Gómez Cruz, E. (2018) 'Datafication and data fiction: Narrating data and narrating with data', *Big Data & Society*, 5(2), p. 205395171878408. Available at: https://doi.org/10.1177/2053951718784083.

Duan, Y., Edwards, J.S. and Dwivedi, Y.K. (2019) 'Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda', *International Journal of Information Management*, 48, pp. 63–71.

Durán, J.M. and Jongsma, K.R. (2021) 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI', *Journal of Medical Ethics*, p. medethics-2020-106820. Available at: https://doi.org/10.1136/medethics-2020-106820.

Ehsan, U. *et al.* (2021) 'The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations'. arXiv. Available at: http://arxiv.org/abs/2107.13509 (Accessed: 13 August 2022).

Endacott, C. and Leonardi, P. (2021) 'Identity-Based Motivations for Providing the Unpaid Labor That Makes AI Technologies Work', *Academy of Management Proceedings*, 2021(1), p. 12195. Available at: https://doi.org/10.5465/AMBPP.2021.12195abstract.

Engin, Z. and Treleaven, P. (2018) 'Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies', *The British Computer Society* [Preprint]. Available at:

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8852885 (Accessed: 11 October 2022).

Ennals, P. *et al.* (2016) 'Shifting occupational identity: doing, being, becoming and belonging in the academy', *Higher Education Research & Development*, 35(3), pp. 433–446. Available at: https://doi.org/10.1080/07294360.2015.1107884.

Faraj, S., Pachidi, S. and Sayegh, K. (2018) 'Working and organizing in the age of the learning algorithm', *Information and Organization*, 28(1), pp. 62–70. Available at: https://doi.org/10.1016/j.infoandorg.2018.02.005.

Fayard, A.-L., Stigliani, I. and Bechky, B.A. (2017) 'How Nascent Occupations Construct a Mandate: The Case of Service Designers' Ethos', *Administrative Science Quarterly*, 62(2), pp. 270–303. Available at: https://doi.org/10.1177/0001839216665805.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From Data Mining to Knowledge Discovery in Databases', *American Association for Artificial Intelligence*, p. 18.

Feinberg, M. (2017) 'A Design Perspective on Data', in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. *CHI '17: CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, pp. 2952–2963. Available at: https://doi.org/10.1145/3025453.3025837.

Foldy, E.G. (2002) *Being all that you can be: Identity and power in organizations*. Ph.D. Boston College. Available at: https://www.proquest.com/docview/304789756/abstract/A4DA0086A5F14F92PQ/1 (Accessed: 14 October 2022).

Gee, J.P. (2000) 'Chapter 3 : Identity as an Analytic Lens for Research in Education', *Review of Research in Education*, 25(1), pp. 99–125. Available at: https://doi.org/10.3102/0091732X025001099.

Gehl, R.W. (2015) 'Sharing, knowledge management and big data: A partial genealogy of the data scientist', *European Journal of Cultural Studies*, 18(4–5), pp. 413–428. Available at: https://doi.org/10.1177/1367549415577385.

George, G. *et al.* (2016) 'Big Data and Data Science Methods for Management Research', *Academy of Management Journal*, 59(5), pp. 1493–1507. Available at: https://doi.org/10.5465/amj.2016.4005.

Gherardi, S. and Benozzo, A. (2021) 'Shadow organising as dwelling in the space of the "not-yet"', *Studies in Continuing Education*, p. 16. Available at: https://doi.org/10.1080/0158037X.2021.1900097.

Gill, M.J. (2015) 'Elite identity and status anxiety: An interpretative phenomenological analysis of management consultants', *Organization*, 22(3), pp. 306–325. Available at: https://doi.org/10.1177/1350508413514287.

Gioia, D.A., Corley, K.G. and Hamilton, A.L. (2013) 'Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology', *Organizational Research Methods*, 16(1). Available at: https://journals.sagepub.com/doi/abs/10.1177/1094428112452151 (Accessed: 14 October 2022).

Gitelman, L. (ed.) (2013) *Raw Data Is an Oxymoron*. Cambridge, Massachusetts: The MIT Press.

Glaser, B.G. and Strauss, A.L. (1967) *The discovery of grounded theory: strategies for qualitative research*. London, [England] ; New York, [New York]: Routledge (Book, Whole). Available at: https://go.exlibris.link/Sj022X6b (Accessed: 14 October 2022).

Gorman, E.H. and Sandefur, R.L. (2011) '"Golden Age," Quiescence, and Revival: How the Sociology of Professions Became the Study of Knowledge-Based Work', *Work and Occupations*, 38(3), pp. 275–302. Available at: https://doi.org/10.1177/0730888411417565.

Goto, M. (2021) 'Collective professional role identity in the age of artificial intelligence', *Journal of Professions and Organization*, p. joab003. Available at: https://doi.org/10.1093/jpo/joab003.

Gray, J., Gerlitz, C. and Bounegru, L. (2018) 'Data infrastructure literacy', *Big Data & Society*, 5(2), p. 205395171878631. Available at: https://doi.org/10.1177/2053951718786316.

Greenwood, R., Suddaby, R. and Hinings, C.R. (2002) 'THEORIZING CHANGE: THE ROLE OF PROFESSIONAL ASSOCIATIONS IN THE TRANSFORMATION OF INSTITUTIONALIZED FIELDS', p. 24.

Grootendorst, M. (2021) *The Truth about Working as a Data Scientist*, *Medium*. Available at: https://towardsdatascience.com/the-truth-about-working-as-a-data-scientist-99ed40a600d2 (Accessed: 11 October 2022).

Hamutcu, H. and Fayyad, U. (2020) 'Analytics and Data Science Standardization and Assessment Framework', *Harvard Data Science Review* [Preprint]. Available at: https://doi.org/10.1162/99608f92.1a99e67a.

Han, B., Liu, C.L. and Zhang, W.J. (2016) 'A Method to Measure The Resilience of Algorithm for Operation Management', *International Federation of Automatic Control*, p. 6.

Himes, D. (1996) 'Emerging Occupations', *Monthly Labor Review*, 119(3), p. 31.

Iedema, R. and Scheeres, H. (2003) 'From Doing Work to Talking Work: Renegotiating Knowing, Doing, and Identity', *Applied Linguistics*, 24(3), pp. 316–337.

Innerarity, D. (2021) ‘Making the black box society transparent’, *AI & SOCIETY*, 36(3), pp. 975–981. Available at: https://doi.org/10.1007/s00146-020-01130-8.

Joshi, M.P. *et al.* (2021) ‘Why So Many Data Science Projects Fail to Deliver’, *MIT Sloan Management Review*, 62(3). Available at: https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/ (Accessed: 8 December 2022).

Kane, A.A. and Levina, N. (2017) ‘“Am I Still One of Them?”: Bicultural Immigrant Managers Navigating Social Identity Threats When Spanning Global Boundaries: Navigating Identity Threats in Spanning Boundaries’, *Journal of Management Studies*, 54(4), pp. 540–577. Available at: https://doi.org/10.1111/joms.12259.

Kemper, J. and Kolkman, D. (2019) ‘Transparent to whom? No algorithmic accountability without a critical audience’, *Information, Communication & Society*, 22(14), pp. 2081–2096. Available at: https://doi.org/10.1080/1369118X.2018.1477967.

Khan, M.A.H. (2022) ‘Identity Construction in Emerging Occupations: How Muslim Chaplains Bricolage Occupational Identity’, *Academy of Management Proceedings* [Preprint]. Available at: https://doi.org/10.5465/AMBPP.2022.17475abstract.

Khatiwada, S. and Veloso, M.K. (2019) ‘New Technology and Emerging Occupations: Evidence from Asia’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3590128.

Kimmons, R. and Veletsianos, G. (2014) ‘The fragmented educator 2.0: Social networking sites, acceptable identity fragments, and the identity constellation’, *Computers & Education*, 72, pp. 292–301. Available at: https://doi.org/10.1016/j.compedu.2013.12.001.

Kleinman, D.L. and Vallas, S.P. (2001) ‘Science, Capitalism, and the Rise of the “Knowledge Worker”: The Changing Structure of Knowledge Production in the United States’, *Theory and Society*, 30(4), pp. 451–492.

Koch, B. *et al.* (2021) ‘Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research’, *arXiv:2112.01716 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/2112.01716 (Accessed: 31 December 2021).

Lai, C.-L. (2021) ‘Exploring University Students’Preferences for AI-Assisted Learning Environment: A Drawing Analysis with Activity Theory Framework’, *Educational Technology & Society*, 24(4), pp. 1–15.

Lebovitz, S., Levina, N. and Lifshitz-Assaf, H. (2021) ‘Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts’ Know-What’, *MIS Quarterly*, 45(3), pp. 1501–1526. Available at: https://doi.org/10.25300/MISQ/2021/16564.

Lebovitz, S., Lifshitz-Assaf, H. and Levina, N. (2022) 'To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis', *Organization Science*, 33(1), pp. 126–148. Available at: https://doi.org/10.1287/orsc.2021.1549.

Lee, D. *et al.* (2019) 'Enhancing Research Quality through Analytical Memo Writing in a Mixed Methods Grounded Theory Study Implemented by a Multi-Institution Research Team', in *2019 IEEE Frontiers in Education Conference (FIE)*. *2019 IEEE Frontiers in Education Conference (FIE)*, Covington, KY, USA: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/FIE43999.2019.9028469.

Leroy, H.L. *et al.* (2022) 'Walking Our Evidence-Based Talk: The Case of Leadership Development in Business Schools', *Journal of Leadership & Organizational Studies*, 29(1), pp. 5–32. Available at: https://doi.org/10.1177/15480518211062563.

Leonardi, P.M. (2011) 'When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies on JSTOR', *MIS Quarterly*, 35(1), pp. 147–167.

Levina, N. (2021) 'ALL INFORMATION SYSTEMS THEORY IS GROUNDED THEORY', *MIS Quarterly*, 45(1), pp. 489–494.

Lifshitz-Assaf, H. (2018) 'Dismantling Knowledge Boundaries at NASA: The Critical Role of Professional Identity in Open Innovation', *Administrative Science Quarterly*, 63(4), pp. 746–782. Available at: https://doi.org/10.1177/0001839217747876.

Locke, E.A. (2007) 'The Case for Inductive Theory Building†', *Journal of Management*, 33(6), pp. 867–890. Available at: https://doi.org/10.1177/0149206307307636.

Lupton, D. (2015) *Digital sociology*. Abingdon, Oxon: Routledge, Taylor & Francis Group.

MacKenzie, R., Marks, A. and Morgan, K. (2017) 'Technology, Affordances and Occupational Identity Amongst Older Telecommunications Engineers: From Living Machines to Black-Boxes', *Sociology*, 51(4), pp. 732–748. Available at: https://doi.org/10.1177/0038038515616352.

Mallett, O. and Wapshott, R. (2012) 'Mediating ambiguity: Narrative identity and knowledge workers', *Scandinavian Journal of Management*, 28(1), pp. 16–26. Available at: https://doi.org/10.1016/j.scaman.2011.12.001.

Matveeva, S. (2019) 'What Data Scientists Do And How To Work With Them', *Forbes*, p. 5.

Miceli, M., Schuessler, M. and Yang, T. (2020) 'Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision', *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), pp. 1–25. Available at: https://doi.org/10.1145/3415186.

Mikalsen, M. and Monteiro, E. (2021) 'Acting with Inherently Uncertain Data: Practices of Data-Centric Knowing', *Journal of the Association for Information Systems*, 22(6), pp. 1715–1735. Available at: https://doi.org/10.17705/1jais.00722.

Monteiro, E. and Parmiggiani, E. (2019) 'Synthetic Knowing: The Politics of the Internet of Things', *MIS Quarterly*, 43(1), pp. 167–184. Available at: https://doi.org/10.25300/MISQ/2019/13799.

Muller, M. *et al.* (2019) 'How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. *CHI '19: CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, pp. 1–15. Available at: https://doi.org/10.1145/3290605.3300356.

Myers, M.D. (2017) *Coming of age : Interpretive research in information systems*. 1st edn, *The Routledge Companion to Management Information Systems*. 1st edn. Routledge, pp. 83–93. Available at: https://doi.org/10.4324/9781315619361-7.

Neff, G. *et al.* (2017) 'Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science', *Big Data*, 5(2), pp. 85–97. Available at: https://doi.org/10.1089/big.2016.0050.

Nelson, A.J. and Irwin, J. (2014) '"Defining What We Do—All Over Again": Occupational Identity, Technological Change, and the Librarian/Internet-Search Relationship', *Academy of Management Journal*, 57(3), pp. 892–928. Available at: https://doi.org/10.5465/amj.2012.0201.

O'Mahoney, J. and Sturdy, A. (2016) 'Power and the diffusion of management ideas: The case of McKinsey & Co', *Management Learning*, 47(3), pp. 247–265. Available at: https://doi.org/10.1177/1350507615591756.

O'Neil, C. and Schutt, R. (2014) *Doing Data Science: Straight Talk from the Frontline*. United States of America: O'Reilly. Available at: https://books.google.com/books/about/Doing_Data_Science.html?id=ycNKAQAAQBAJ (Accessed: 12 October 2022).

Orlikowski, W.J. (2000) 'Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations', *Organization Science* [Preprint]. Available at: https://doi.org/10.1287/orsc.11.4.404.14600.

Orona, C.J. (1990) 'Temporality and identity loss due to Alzheimer's disease', *Social Science & Medicine*, 30(11), pp. 1247–1256. Available at: https://doi.org/10.1016/0277-9536(90)90265-T.

Oxford Dictionary (2021) 'Oxford Learner's Dictionaries'. Available at: https://www.oxfordlearnersdictionaries.com/ (Accessed: 9 May 2021).

Pachidi, S. *et al.* (2021) 'Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing', *Organization Science*, 32(1), pp. 18–41. Available at: https://doi.org/10.1287/orsc.2020.1377.

Parkes, D.C. and Wellman, M.P. (2015) 'Economic reasoning and artificial intelligence', *Science*, 349(6245), pp. 267–272. Available at: https://doi.org/10.1126/science.aaa8403.

Passi, S. and Jackson, S.J. (2017) 'Data Vision: Learning to See Through Algorithmic Abstraction', *Data Visualization*, p. 12.

Paullada, A. *et al.* (2021) 'Data and its (dis)contents: A survey of dataset development and use in machine learning research', *Patterns*, 2(11), p. 100336. Available at: https://doi.org/10.1016/j.patter.2021.100336.

Perrons, R.K. and Jensen, J.W. (2015) 'Data as an asset: What the oil and gas sector can learn from other industries about "Big Data"', *Energy Policy*, 81, pp. 117–121. Available at: https://doi.org/10.1016/j.enpol.2015.02.020.

Pink, S. *et al.* (2018) 'Broken data: Conceptualising data in an emerging world', *Big Data & Society*, 5(1), p. 205395171775322. Available at: https://doi.org/10.1177/2053951717753228.

Price, W.N. (2018) 'Big data and black-box medical algorithms', *Science Translational Medicine*, 10(471), p. eaao5333. Available at: https://doi.org/10.1126/scitranslmed.aao5333.

Provost, F. and Fawcett, T. (2013) 'Data Science and its Relationship to Big Data and Data-Driven Decision Making', *Big Data*, 1(1), pp. 51–59. Available at: https://doi.org/10.1089/big.2013.1508.

Raji, I.D. (2020) 'The Discomfort of Death Counts: Mourning through the Distorted Lens of Reported COVID-19 Death Data', *Patterns*, 1(4), p. 100066. Available at: https://doi.org/10.1016/j.patter.2020.100066.

Raji, I.D. *et al.* (2021) 'AI and the Everything in the Whole Wide World Benchmark', *arXiv:2111.15366 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2111.15366 (Accessed: 24 December 2021).

Ramzan, M.J. *et al.* (2021) 'A Conceptual Model to Support the Transmuters in Acquiring the Desired Knowledge of a Data Scientist', *IEEE Access*, 9, pp. 115335–115347. Available at: https://doi.org/10.1109/ACCESS.2021.3105038.

Regalado, A. (2014) *The Power to Decide*, *MIT Technology Review*. Available at: https://www.technologyreview.com/2014/01/22/174493/the-power-to-decide/ (Accessed: 8 December 2022).

Ribes, D. (2017) 'Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures', in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. *CSCW '17: Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, pp. 1514–1526. Available at: https://doi.org/10.1145/2998181.2998344.

Rusu, O. *et al.* (2013) 'Converting unstructured and semi-structured data into knowledge', in. *11th RoEduNet International Conference*, pp. 1–4. Available at: https://ieeexplore.ieee.org/abstract/document/6511736/ (Accessed: 12 October 2022).

Rzeznikiewicz, D. (2022) *How to become a data scientist: A guide to the education, skills, and necessary experience*, *Fortune*. Available at: https://fortune.com/education/business/articles/2022/01/25/how-to-become-a-data-scientist-a-guide-to-the-education-skills-and-necessary-experience/ (Accessed: 11 October 2022).

Saldaña, J. (2016) *The coding manual for qualitative researchers*. Third. Los Angeles: SAGE (Book, Whole).

Saltz, J., Shamshurin, I. and Connors, C. (2017) 'Predicting data science sociotechnical execution challenges by categorizing data science projects', *Journal of the Association for Information Science and Technology*, 68(12), pp. 2720–2728. Available at: https://doi.org/10.1002/asi.23873.

Saltz, J.S. (2021) 'CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps', in *2021 IEEE International Conference on Big Data (Big Data)*. *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2337–2344. Available at: https://doi.org/10.1109/BigData52589.2021.9671634.

Saltz, J.S. and Grady, N.W. (2017) 'The ambiguity of data science team roles and the need for a data science workforce framework', in *2017 IEEE International Conference on Big Data (Big Data)*. *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA: IEEE, pp. 2355–2361. Available at: https://doi.org/10.1109/BigData.2017.8258190.

Seaver, N. (2017) 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems', *Big Data & Society*, 4(2), p. 205395171773810. Available at: https://doi.org/10.1177/2053951717738104.

Seidelin, C., Dittrich, Y. and Grönvall, E. (2018) 'Data Work in a Knowledge-Broker Organisation: How Cross-Organisational Data Maintenance shapes Human Data Interactions', in. *Proceedings of the 32nd International BCS Human Computer Interaction Conference*. Available at: https://doi.org/10.14236/ewic/HCI2018.14.

Seidelin, C., Dittrich, Y. and Grönvall, E. (2020) 'Foregrounding data in co-design – An exploration of how data may become an object of design', *International Journal of Human-Computer Studies*, 143, p. 102505. Available at: https://doi.org/10.1016/j.ijhcs.2020.102505.

Simon, H.A. (1990) 'Bounded Rationality', in J. Eatwell, M. Milgate, and P. Newman (eds) *Utility and Probability*. London: Palgrave Macmillan UK (The New Palgrave), pp. 15–18. Available at: https://doi.org/10.1007/978-1-349-20568-4_5.

Skorikov, V.B. and Vondracek, F.W. (2011) ‘Occupational Identity’, in S.J. Schwartz, K. Luyckx, and V.L. Vignoles (eds) *Handbook of Identity Theory and Research*. New York, NY: Springer New York, pp. 693–714. Available at: https://doi.org/10.1007/978-1-4419-7988-9_29.

Slota, S.C. *et al.* (2020) ‘Prospecting (in) the data sciences’, *Big Data & Society*, 7(1), p. 205395172090684. Available at: https://doi.org/10.1177/2053951720906849.

Song, I.-Y. and Zhu, Y. (2016) ‘Big data and data science: what should we teach?’, *Expert Systems*, 33(4), pp. 364–373. Available at: https://doi.org/10.1111/exsy.12130.

Spruit, M. and Lytras, M. (2018) ‘Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients’, *Telematics and Informatics*, 35(4), pp. 643–653. Available at: https://doi.org/10.1016/j.tele.2018.04.002.

Strich, F., Mayer, A.-S. and Fiedler, M. (2021) ‘What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees’ Professional Role Identity’, *Journal of the Association for Information Systems*, p. 21.

Stuart, P.H. (2013) ‘Social Work Profession: History’, *Encyclopedia of Social Work* [Preprint]. Available at: https://doi.org/10.1093/acrefore/9780199975839.013.623.

Suddaby, R. and Greenwood, R. (2005) ‘Rhetorical Strategies of Legitimacy’, *Administrative Science Quarterly*, 50(1), pp. 35–67. Available at: https://doi.org/10.2189/asqu.2005.50.1.35.

Tanweer, A., Fiore-Gartland, B. and Aragon, C. (2016) ‘Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process’, *Information, Communication & Society*, 19(6), pp. 736–752. Available at: https://doi.org/10.1080/1369118X.2016.1153125.

Thakkar, D., Kumar, N. and Sambasivan, N. (2020) ‘Towards an AI-powered Future that Works for Vocational Workers’, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. *CHI ’20: CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, pp. 1–13. Available at: https://doi.org/10.1145/3313831.3376674.

The Royal Society (2019) *Dynamics of data science skills | Royal Society*. Available at: https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/ (Accessed: 11 October 2022).

Tung, K. (2019) ‘AI, the internet of legal things, and lawyers’, *Journal of Management Analytics*, 6(4), pp. 390–403. Available at: https://doi.org/10.1080/23270012.2019.1671242.

Tuomi, I. (1999) ‘Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational

Memory', *Journal of Management Information Systems*, 16(3), pp. 103–117. Available at: https://doi.org/10.1080/07421222.1999.11518258.

Urquhart, C., Lehmann, H. and Myers, M.D. (2010) 'Putting the "theory" back into grounded theory: guidelines for grounded theory studies in information systems', *Information Systems Journal*, 20(4), pp. 357–381. Available at: https://doi.org/10.1111/j.1365-2575.2009.00328.x.

U.S. Bureau of Labor Statistics (2021) *Computer and Information Research Scientists : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*. Available at: https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm (Accessed: 11 October 2022).

Vaast, E. and Pinsonneault, A. (2021) 'When Digital Technologies Enable and Threaten Occupational Identity: The Delicate Balancing Act of Data Scientists', *MIS Quarterly*, 45(3), pp. 1087–1112. Available at: https://doi.org/10.25300/MISQ/2021/16024.

Veel, K. (2018) 'Make data sing: The automation of storytelling', *Big Data & Society*, 5(1), p. 205395171875668. Available at: https://doi.org/10.1177/2053951718756686.

Waardenburg, L., Huysman, M. and Sergeeva, A.V. (2022) 'In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms', *Organization Science*, 33(1), pp. 59–82. Available at: https://doi.org/10.1287/orsc.2021.1544.

Waardenburg, L., Sergeeva, A. and Huysman, M. (2018) 'Hotspots and Blind Spots: A Case of Predictive Policing in Practice', in U. Schultze et al. (eds) *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology*. Cham: Springer International Publishing (IFIP Advances in Information and Communication Technology), pp. 96–109. Available at: https://doi.org/10.1007/978-3-030-04091-8_8.

Walker, G.H. *et al.* (2008) 'A review of sociotechnical systems theory: a classic concept for new command and control paradigms', *Theoretical Issues in Ergonomics Science*, 9(6), pp. 479–499. Available at: https://doi.org/10.1080/14639220701635470.

Waller, M.A. and Fawcett, S.E. (2013) 'Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management', *Journal of Business Logistics*, 34(2), pp. 77–84. Available at: https://doi.org/10.1111/jbl.12010.

Walsham, G. (1995) 'The Emergence of Interpretivism in IS Research', *Information Systems Research*, 6(4), pp. 376–394. Available at: https://doi.org/10.1287/isre.6.4.376.

Wang, D. *et al.* (2019) 'Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI', *Proceedings of the ACM on*

*Human-Computer Interaction*, 3(CSCW), pp. 1–24. Available at: https://doi.org/10.1145/3359313.

Whitley, E.A., Gal, U. and Kjaergaard, A. (2014) 'Who do you think you are? A review of the complex interplay between information systems, identification and identity', *European Journal of Information Systems*, 23(1), pp. 17–35. Available at: https://doi.org/10.1057/ejis.2013.34.

Woiceshyn, J. and Daellenbach, U. (2018) 'Evaluating inductive vs deductive research in management studies: Implications for authors, editors, and reviewers', *Qualitative Research in Organizations and Management: An International Journal*, 13(2), pp. 183–195. Available at: https://doi.org/10.1108/QROM-06-2017-1538.

Yang, Q., Steinfeld, A. and Zimmerman, J. (2019) 'Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11. Available at: https://doi.org/10.1145/3290605.3300468.

Yoo, Y., Henfridsson, O. and Lyytinen, K. (2010) 'Research Commentary—The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research', *Information Systems Research*, 21(4), pp. 724–735. Available at: https://doi.org/10.1287/isre.1100.0322.

Zins, C. (2007) 'Conceptual approaches for defining data, information, and knowledge', *Journal of the American Society for Information Science and Technology*, 58(4), pp. 479–493. Available at: https://doi.org/10.1002/asi.20508.

Zuboff, S. (2019) *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. London: Profile Books (Book, Whole). Available at: https://go.exlibris.link/bvq3T1Sx (Accessed: 15 October 2022).