Beyond the Dataset: Understanding Sociotechnical Aspects of the Knowledge Discovery Process Among Modern Data Professionals

by

Anson Ho

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Masters of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2017

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Data professionals are among the most sought-out professionals in today's industry. Although the skillsets and training can vary among these professionals, there is some consensus that a combination of technical and analytical skills is necessary. In fact, a growing number of dedicated undergraduate, graduate, and certificate programs are now offering such core skills to train modern data professionals. Despite the rapid growth of the data profession, we have few insights into what it is like to be a data professional on-the-job beyond having specific technical and analytical skills. We used the Knowledge Discovery Process (KDP) as a framework to understand the sociotechnical and collaborative challenges that data professionals face. We carried out 20 semi-structured interviews with data professionals across seven different domains. Our results indicate that KDP in practice is highly social, collaborative, and dependent on domain knowledge. To address the sociotechnical gap, the need for a translator within the KDP has emerged. The main contribution of this thesis is in providing empirical insights into the work of data professionals, highlighting the sociotechnical challenges that they face on the job. Also, we propose a new analytic approach to combine thematic analysis and cognitive work analysis (CWA) on the same dataset. Implications of this research will improve the productivity of data professionals and will have implications for designing future tools and training materials for the next generation of data professionals.

Acknowledgements

The support and guidance provided by my thesis co-supervisors, Dr. Catherine Burns from the University of Waterloo and Dr. Parmit Chilana from Simon Fraser University were vital for the development of this research and my growth as a researcher. This research was made possible through the Natural Sciences and Engineering Research Council of Canada (NSERC).

I would like to thank Dr. Shi Cao and Dr. James Wallace for being my thesis readers. I am grateful for their feedback and comments on this thesis. Thank you for the support and comments to help improve my research and thesis.

Special thanks go out to Ethan Li, Jennifer Fong, Johnson Kan, Rachel Cao, and Nathaniel Hudson for their input and feedback to this thesis. Special thanks to the past and present members of the Advanced Interface Design Lab, in particular, Plinio Morita, Behzad Aghaei, Yeti Li, Wayne Giang, Leila Rezai, Elizabeth Kittel, and Laton Vermette. Also, I would like the thank the Human Factors and the Engineering HCI communities at the University of Waterloo.

I must express my gratitude to my parents and Kristen Chai-Chong for providing me with the support and encouragement throughout my years of study and writing this thesis.

Finally, I must thank you, the reader, for reading and joining me on the adventure to learn about modern data professionals' challenges and coping strategies throughout their workflow. I hope you find my thesis insightful and this thesis informs your future research.

Dedication

I would like to dedicate the thesis to you, the reader.

Table of Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Definition
AH	Abstraction Hierarchy
ConTA	Contextual Task Analysis
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CSCW	Computer-Supported Collaborative Work
CWA	Cognitive work analysis
DL	Decision Ladder
HCI	Human-Computer Interaction
HF	Human Factors
KDD	Knowledge Discovery in Databases
KDP	Knowledge Discovery Process
WDA	Work domain analysis

Chapter 1
Introduction

Data analysts, data scientists, and data engineers (henceforth, data professionals) are among the most sought-out professionals in today's industry [21]. Although the skill sets and training can vary among data professionals [15,27], there is some consensus that a combination of technical and analytical skills is necessary (e.g., machine learning, operations research, programming, statistics, and business knowledge) [36]. In fact, a growing number of dedicated undergraduate, graduate, and certificate programs are now offering such core skills to train modern data professionals.

Despite the rapid growth of the data profession, we have few insights into what it is like to be a data professional on-the-job, beyond having specific technical and analytical skills [70]. Models such as the knowledge discovery process (KDP) propose that to derive insights, professionals have to understand the problem at hand, acquire the necessary data, validate the data, and visualize and communicate results. We believe that many of the current challenges for data professionals are not strictly technical. Each of these stages can potentially present unique sociotechnical and collaboration challenges. For example, Ackerman's lens of the sociotechnical gap [1] points out that a potential divide can occur when there is difference between what people want compared to what is technically possible within the complex interactions that take place between individuals, groups, and technical systems. Although Human Factors (HF), Human-Computer Interaction (HCI) and Computer-Supported Collaborative Work (CSCW) have a long history of focus on data visualization and communication [41], the other phases of the KDP have surprisingly received less attention, even though they are critical to the modern data professional.

In this thesis, we investigate the day-to-day work of data professionals from a sociotechnical perspective [1,69]. We consider two key research questions:

1. What are the day-to-day sociotechnical challenges that data professionals face across the different stages of the KDP?
2. How do data professionals currently cope with these challenges?

We carried out 20 semi-structured interviews with data professionals across seven different domains, including healthcare, geography, corporate, consulting, education, finance, and technology. The interviews focused on eliciting and understanding the collaborations that exist throughout the

discovery process. Each of the seven domains represents sectors in which data professionals are most sought after and present unique data analysis challenges.

For our analysis, we used two different lenses to explore the interview data. First, we conducted a thematic analysis [9] to study the data from a bottom-up approach. Next, we looked at the data through cognitive work analysis (CWA), a work-centered theoretical framework used to analyze the cognitive work that occurs within a system. The CWA complemented the thematic analysis by providing a top-down lens, allowing us to further understand the environmental constraints. The combination of the two analytical techniques allowed for multiple perspectives to become salient in our analysis.

Our key results show that although data professionals have to constantly adapt their technical and statistical skills to solve a new problem, they also spend a significant amount of time to talk to other team members, clients, or domain experts. These conversations occur not only in communicating results, but throughout the process of knowledge discovery, suggesting that the KDP in practice is highly social, collaborative, and dependent on domain knowledge. In addition to being experts in data manipulation and analysis, data professionals indicated that they must be well-versed in conversing with different players from the problem inception stage to the presentation of the results.

The main contribution of this thesis is in providing empirical insights into the work of data professionals, highlighting sociotechnical challenges that they face on the job. One implication of our results is that there is need to design data acquisition, analysis, and communication tools that consider the social and domain-specific aspects of the KDP. There is also need to further investigate how future data professionals can be trained so they can better adapt to not only the technical demands of the job, but also cope with sociotechnical challenges.

1.1 Thesis organization

The thesis is organized into the following chapters:

Table 1: Thesis organization

Chapter	Description of the Chapter
1. Introduction	Introduction contains the motivation, main research question, and the organization of the thesis
2. Literature review	Literature review contains a detailed survey of relevant literature, including the KDP, common ground, and studies about modern data professionals
3. Method	Method describes the 20 semi-structured interviews that were conducted including the sampling technique and interview guide
4. Thematic analysis	Thematic analysis outlines the first analytical approach to our collected data. This chapter also presents the results of the analysis.
5. Cognitive work analysis	CWA outlines the second structured method of analyzing the data. This chapter presents two models.
6. Discussion	In this chapter, we describe the implications for design future tools and training future data professionals. We will also outline the limitations of the current study and the learnings from using two different analytical methods.
7. Conclusion	The conclusion summarizes the main findings of the thesis and describes future research opportunities.

1.2 Contribution

This thesis makes the following main contributions:

1. Establishes an empirical understanding of the human aspect of the KDP, highlighting the sociotechnical gap in each of the phases.
2. Illustrates the emerging need of the translator within the KDP
3. Understands the role of domain knowledge within the KDP
4. Applies CWA to describe the data profession as a complex sociotechnical system
5. Compares two analytical techniques: thematic analysis and CWA on the same dataset by applying both a constructivist and ecological approach
6. Highlights the design opportunities that exist in building the next generation of tools for data professionals to consider the social and domain-specific aspects of the KDP
7. Illustrates the implications for training the next generation of data professionals so that they can cope with the sociotechnical challenges

Insights from this thesis will be beneficial for 1) designers inventing new tools for data professionals; 2) data professionals trying to improve their workflows; 3) researchers seeking to reduce and understand barriers within the KDP; 4) educators designing new training programs for the next generation of data professionals.

Chapter 2
Related work

Data science has become a popular topic in recent years, as companies and industries have recognized the value of data to improve business processes. Data has already been used to revolutionize medical practice, modernize public policy, and inform business decisions [49]. The role of the data scientist has even been labeled as "the sexiest job of the 21st century" [21]. Despite the enthusiasm around data science, research focused on understanding the role of data professionals and their day-to-day challenges is only beginning to emerge.

The demand for data science and data technology within the industry is growing faster than the supply of talent. McKinsey stated: "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions." [40]. In another perspective, the demand for data professionals will exceed the supply by 50- 60% [40].

This chapter reviews the related literature and discussions relevant to our thesis topic and highlights how this thesis provides a novel contribution to the field of HCI and HF. We will first discuss related research on data professionals and their proposed workflows, such as the KDP. Next, we will describe the sociotechnical lens which we use to understand the challenges that exist within data professionals' workflows. Specifically, we will describe Ackerman's lens of the sociotechnical gap used to understand the sociotechnical challenges that data professionals face. Finally, we will review the combined used of thematic analysis and CWA.

2.1 Study of Data Professionals

Data professionals are high-ranking professionals with the skill and curiosity to observe patterns in the world of big data. The job title has only been around for a few years, coined in 2008 by D.J Patil and Jeff Hammerbacher [21]. The increased demand for data professionals within the industry reflects the increased need for companies to understand information in varieties and volumes never encountered before.

The academic community is only beginning to understand the work practices of modern data professionals. We have few insights into the data professional's challenges. In this section, we will

discuss key findings on 1) skill sets and the types of a modern data professionals, 2) how data professionals work in organizations, and 3) impact of domain knowledge on data science.

2.2 Data Professional Skills

Harris et al. [36], surveyed 250+ data professionals to understand their skills and experiences. Harris et al., explore a more precise vocabulary for describing the work, based on how data scientists work and describe themselves and their skills. They describe four different types of data scientists: *Data Businesspeople*, *Data Creatives*, *Data Developers* and *Data Researchers*. Harris et al. focused on understanding the various kinds of technical skills needed for each type of data scientist such as Bayesian statistics, product development, and visualization. Their study highlighted the need for "T-shaped" data professionals, who have a breadth of skills with depth in a single skill area. Harris et al. focus on evaluating only the technical expertise required to be a data professional. The study did not describe the types of work and changes that data professional's face on the job.

A small number of studies have systematically focused on how data professionals are embedded in a company. Fisher et al. [30] interviewed sixteen data analysts at Microsoft to identify pain points regarding specific data analysis tools. Fisher et al. uncovered issues with tools for specific cloud-based scenarios such as data integration, cost estimation problems from cloud computing, shaping data in a cloud computing platform and the need for fast iterations on analysis results. Even though Fisher et al. studied data professionals within an organization, they did not describe the roles the data professionals play within a team, nor do they describe the sociotechnical challenges that they may face. Our findings are complementary to Fisher et al. as they add a comprehensive perspective of the data professional's technical and non-technical challenges.

2.3 The role of domain knowledge

Data science is a combination of processes and tools made relevant by domain knowledge. Without domain knowledge, the output of data science is not meaningful. The role of the domain expert and domain knowledge with the modern data professional has been given little attention so far. Researchers have admitted the need for domain knowledge to lead the KDP [45,70]. Others have suggested the use of domain knowledge as a method to help with constraining the KDP and to avoid over-fitting of data algorithms [22]. Yoon et al. [72] describe domain knowledge as a tool to describe relationships among data attributes. Most of the existing studies concentrated on the use of domain knowledge within the data analysis phase. Only Kopanas et al. [45] has explored the role of domain

knowledge through the various stages of a large scale data mining project. Our results complement those of Kopanas et al. by describing how data professionals obtain domain knowledge in non-data mining projects. In this thesis, we attempt to explore the role of domain knowledge within the different phases of the KDP.

2.4 Knowledge Discovery Process

Most studies in HCI and CSCW to date have focused on specific aspects of the data professional workflow like data analysis and visualization. We found only a small number of studies that systematically explore the end-to-end workflow of data professionals. Even though the workflow models focused on heavily on technical and data mining aspects of the workflow process, the models provide a foundation to understand the workflow of modern data professionals.

Most discussions of data science workflows are derived from the data mining literature. Simply understanding the algorithms used for data analysis is not sufficient for a successful KDP project. In the data mining literature, the standardized process model for extracting useful knowledge from data is known as the KDP. The KDP process describes the nontrivial process of identifying, novel, potentially useful, and ultimately understandable patterns in data. The model can help organizations better plan and execute a project.

Since the 1990s, several different KDP models have been developed by academia (Fayyad et al. [27], Anand & Buchner [3]), industry (Cabena et al. [13]) and both (Cios et al. [18]). The main differences between each model are in the number and scope of each step of the model [19]. KDP models range from the tasks to understand the domain, data preparation and analysis, to evaluation and application of generated knowledge. It is important to note that KDP is iterative including many feedback loops and repetition [19]. A common feature of all models are the inputs and outputs. Typical inputs into the KDP include data of different formats such as numbers, videos, and images, while the output generated is new knowledge in the form of rules, patterns, models, and trends.

In this section, we will describe the two workflow models from the data mining literature. We will first describe the original academic KDP model as developed by Fayyad et al. [27]. The original model built the foundation for future KDP. Next, we will describe the knowledge discovery model that will be used as a framework of this thesis. The model helps us understand and frame the “typical” data science workflows.

2.5 Academic Knowledge Discovery Process model

The first KDP model is Knowledge Discovery in Databases (KDD), one of the most popular academic models. KDD is a nine-step model by Fayyad et al. [26–28]. The model was used to guide users of data mining tools through knowledge discovery. The main emphasis is to help provide a sequence of steps to execute knowledge discovery in any domain.

The KDD comes from the lens of data miners and academia, focused heavily on data model and data mining tools. The nine steps outlined by Fayyad et al. are:

1. **Developing an understanding of the application domain and relevant prior knowledge**: Learning and understanding the relevant knowledge and identify the goal of the KDP.
2. **Create a target data set**: Selecting a data set to perform the discovery task. This step often includes querying existing data to select the desired subset.
3. **Data cleaning and preprocessing**: Operations are performed to remove noise, dealing with noise and missing data.
4. **Data reduction and projection**: With dimensionality reduction or transformation methods, the number of variables in the data set can be reduced.
5. **Choosing the appropriate data mining task**: Matching the goal of the KDP to a data mining method such as classification, regression, and decision trees.
6. **Exploration analysis, model and thesis selection**: In this step, the data mining algorithm(s) and methods are selected
7. **Data mining**: This step generates patterns into a particular representation such as classification rules, regression models, and decision trees.
8. **Interpreting mined patterns:** This step involves visualization of extracted patterns and models.
9. **Consolidating discovered knowledge**: The final step of the process is to incorporate the knowledge discovered into another system for future action, or report the knowledge to the stakeholders.

The KDD is an iterative process; however, the authors of the model do not describe the relationship between any two steps. The model provides a detailed technical description with respect to data

mining and data analytics, but lacks descriptions of the humanistic aspects and business aspects of the process. Also, KDD is process heavy, focused on the data mining task in academia. Such model may not be appropriate for describing the modern data professional's workflow.

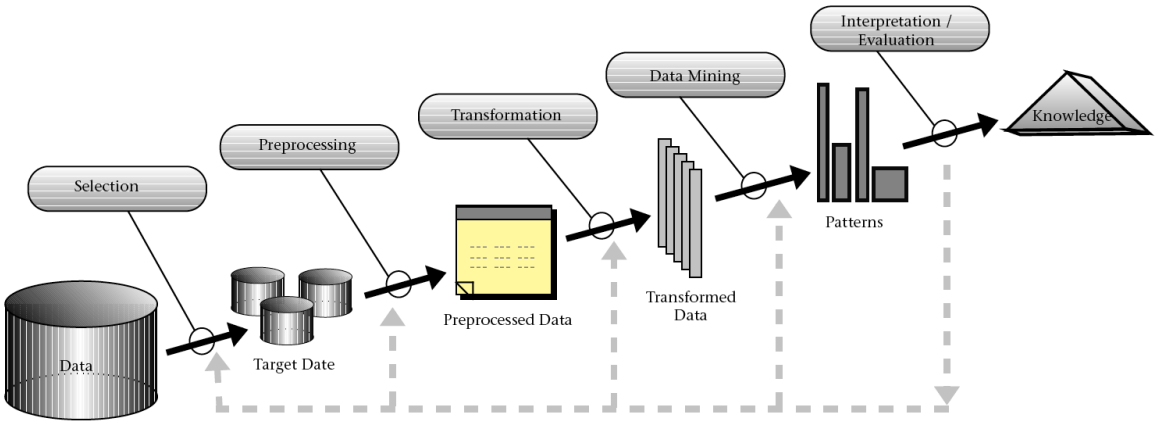

Figure 1: Overview of steps in Knowledge Discovery in Databases (KDD) [28]

2.6 Hybrid Knowledge Discovery Process model

In our study, we attempt to understand the interactions between different roles throughout the hybrid KDP. The hybrid model combines both the KDD as described in the previous chapter and the Cross-Industry Standard Process for Data Mining (CRISP-DM) [66]. The development of academic and industrial models has led to the development of hybrid models which combine aspects of both. The main difference of these models, provided a more general research oriented description of each step as well as introduce a data analysis step instead of modelling. Moving forward the hybrid KDP will just be called the KDP. The discovery process concerns the entire knowledge extraction process, including how data is stored, accessed, how to efficiently use algorithms to analyze data sets, how to interpret and visualize the results and how to model and support the interaction between human and the machine. We present an adapted version of Cios's hybrid model in Figure 2. Specifically, in the model we reduced the importance of data mining.

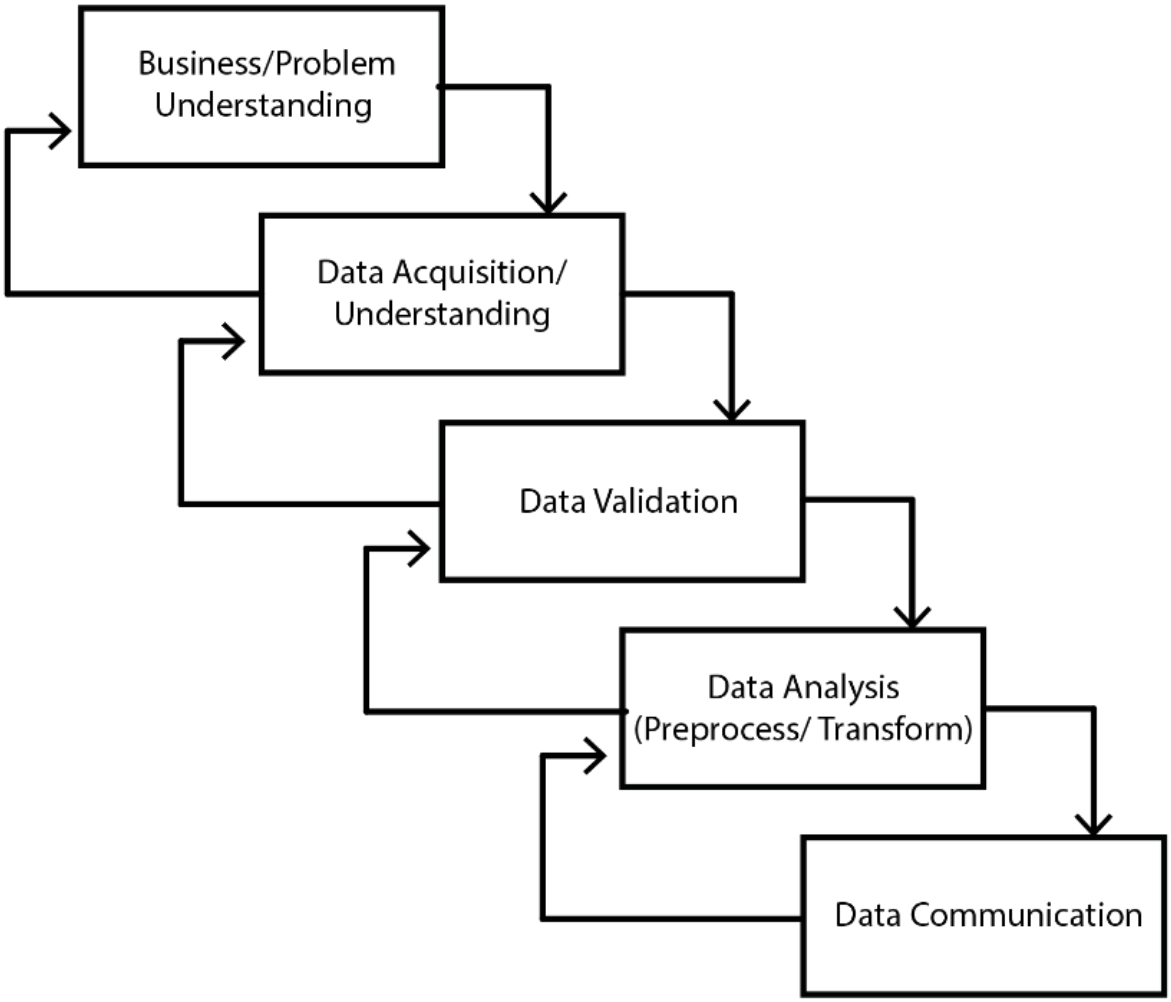

Figure 2: Stages of the Knowledge Discovery Process

2.6.1.1 Business/ Problem Understanding

Understanding the problem and business is an initial step that involves working closely with stakeholders to define a problem and determine problem goals, terminology, questions and identify key stakeholders. Project goals and questions are translated into data/ technical goals.

2.6.1.2 Data acquisition and understanding

Data acquisition and understanding is the second step. This step often includes data collection and familiarization of the data and deciding what tools and methods should be used to solve the problem. Very often there is a feedback loop into this phase, as there is a need for additional domain knowledge to understand the data. After collecting the data, data professionals must understand the data that they have collected: what do the values mean? This phase can be broken down into four main steps: 1) collection of data, 2) description of data, 3) exploration of data and 4) verification of data quality [66].

2.6.1.3 Data validation

Data verification can be broken into its own step. Data are checked for completeness, redundancy and missing values. One of the most important aspects of this is to verify the usefulness of the data and data patterns with respect to the initial problem.

2.6.1.4 Data analysis

Finally, data analysis occurs with the acquired data. In this step, data professionals are required to find patterns within the dataset. Evaluation includes understanding patterns, validating the patterns and interpreting the patterns. Each data professional may analyze their data differently. In some cases, it might be data mining or data modeling. After a pattern is established, the results must be checked to see whether the insights are novel and interesting to the original goal.

2.6.1.5 Data visualization and knowledge communication

Once results are interpreted, the results are communicated back to the stakeholders. This is a critical step in the KDP, as this phase is how data professionals are able to provide value to their organizations. In this step, data patterns must be turned into actionable insights. Many companies struggle to make sense of their data and create value with their data insights [23]. Forrester reports 74% of firms want to make data-driven decisions, but only 29% are successful at connecting analytics into action.

2.6.2 Challenges within the Knowledge Discovery Process

Challenges within the KDP have mainly been presented by studying analysts working on specific problems, such as within intelligence analysis [16,25,42,43]. Researchers have mainly focused on understanding the process and technical challenges within the process, as well as understanding

common issues with different data analytic tools [29,46,65]. Although there is overlap in the high-level analytic process of intelligence analyst and the modern data professional, intelligence analysts often work with different dataset. Modern data professionals often work with large data sets than documents and emails. Intelligence analysts have different goals, and consequently perform different task throughout the process.

Other researchers have solely focused on understanding the technical challenges and tasks that are needed within the data analysis and data visualization phases [2,39,46,63,65]. Researchers like Amar et al. [2] focus on describing the specific task that data professionals need to complete in data visualization. In contrast, we focus on describing the challenges rather than the specific tasks that data professionals perform throughout the process. Russell et al. [63] focus on characterizing high-level sense-making activities that are required when analysing data. Our study extended the study by describing how data professionals leverage these sense-making activities. Kwon and Fisher [46] discuss challenges novice experience when using visualization tools. In our study, we focus on understanding the sociotechnical challenges rather than understanding the technical difficulties. Even though many researchers have solely focused on understanding the challenges and tasks that exist within data analysis and data visualization, few have outlined the sociotechnical challenges that may exist within the discovery process.

Other researchers have focused on the importance of capturing data provenance—the documentation of inputs, systems, and processes that influence data—throughout the KDP [14,31,34]. Systems proposed include automatically logging data interactions and manual annotations. Our results complement these results, but discussing tracking data provenance as a method of building the trust that data professionals have with their data.

Kandel et al. [41] provide some insight into the sociotechnical challenges of enterprise analysts by presenting analysis of 35 interviews with analysts in healthcare, retail, and finance. Kandel et al. characterize the process of data analysis and touch upon how the organizational structure may impact an analyst, particularly in the context of adopting visual analytic tools. In contrast, we focus on understanding the impact of the organizational structure throughout the KDP as well as describing the challenges from a sociotechnical perspective. We also describe the environmental constraints that may impact a data professional and how they overcome such constraints.

2.7 Sociotechnical system

Throughout this thesis, we apply a unique lens to the data profession, viewing the domain a sociotechnical system. A sociotechnical system contains both social (human-related) and technological (non-human) aspects that interact together to pursue a common goal [61]. In other words, humans within a system must interact with each other through technology. This introduces a social dimension in a technical system. A sociotechnical system has the following four features [69]:

1. Work is in a physical environment but also in a social environment
2. Work includes the communication of data, information, and knowledge
3. Work is often performed collaboratively and cooperatively
4. Social interaction adapts and self-evolves through work.

2.7.1 Sociotechnical gap

Within a sociotechnical system, the social dimension influences the technical dimension. This interaction between a user within a system creates different social needs that the technical dimensions must need. If the social needs are not met, there is a discrepancy between the social and technical dimensions. This is also known as the sociotechnical gap. Ackerman defines the sociotechnical gap as "the divide between what we know we must support socially and what we can support technically" [1]. The sociotechnical gap has been used to understand challenges in a variety of different domains such as software development [68], social media [50], decision support systems [62], and location-aware computing [32]. In our study, we explore the social technical gap within the data profession.

2.7.2 Sociotechnical gap and Computer-Supported Collaborative Work

The sociotechnical gap is the fundamental problem within CSCW. Ackerman [1] argues that understanding and reducing the sociotechnical gap is the reason for CSCW's existence. CSCW is uniquely positioned to address this gap, as CSCW exists at the intersection of technology and social settings. Throughout this thesis, we aim to identify the sociotechnical challenges that exist throughout the KDP, which could have arisen from the gap. By understanding the sociotechnical gap, the CSCW community can address the gap through technology. By understanding and addressing the sociotechnical gap, we might be able to reduce the sociotechnical gap through improved technology and improved data professional training.

Throughout the thesis, we focus on understanding the sociotechnical gap within the KDP from the organizational perspective. Unlike previous researchers [2,39,46,63,65], we analyze KDP from a system-wide process, rather than only at the individual or small group level. Gurdin's conceptual framework of CSCW helps understand how to better design for working together [35]. It is important study *the systems developed to support organizational goals as they act through individuals, groups, and projects* [35].

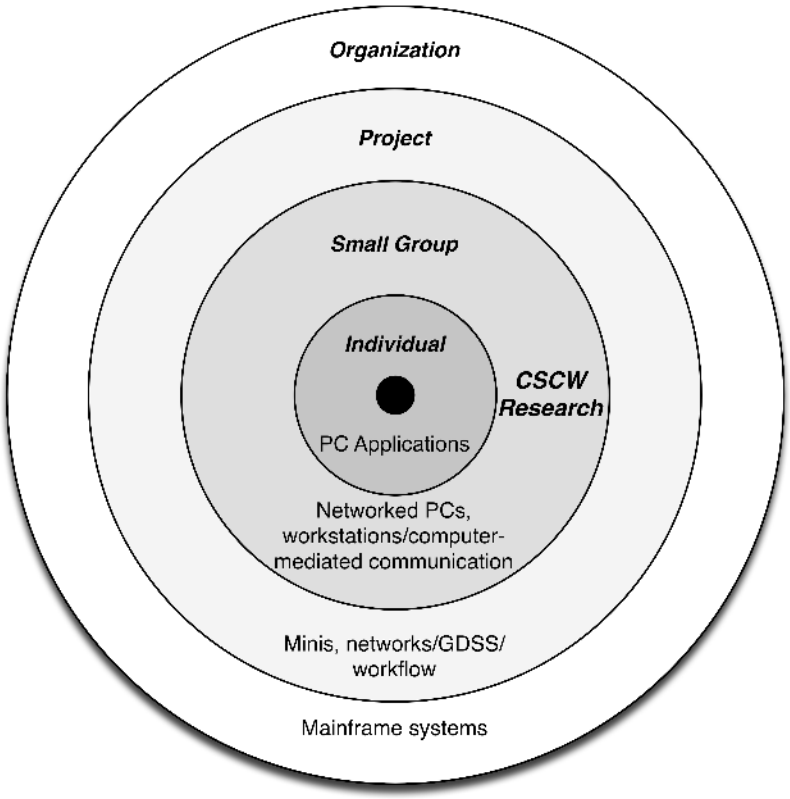

Figure 3: Recreation of Grudin's model for design space of CSCW [35]. Diagram from Lee and Paine [47]

2.8 Combining the use of thematic analysis and cognitive work analysis

In this thesis, we combined the use of two different analytical techniques: thematic analysis and CWA. Thematic analysis is a widely used qualitative analytic method in social sciences to identify, analyze and report themes within a dataset to answer a research question. In contrast, CWA is a structured framework used to analyze complex sociotechnical systems. Few researchers have combined these two analytic techniques together. In this section, we will describe how other researchers have used a hybrid of these two methods.

When researchers leverage CWA, they often do so to understand the environmental constraints and use thematic analysis as a method of identifying and summarize the environmental constraints. Effken et al., [24] used CWA as a method to fit decision support tools to nurse managers' workflow and the constraints were identified using thematic analysis. CWA was a method to constraint and frame the thematic analysis. Naweed [52] leveraged CWA and thematic analysis to investigate the skills of modern and traditional train drivers. Their approach was to first explore the domain, its task and strategies, and used thematic analysis as a tool to repack the findings into a conceptual model to be easily explained.

In contrast, we use thematic analysis as a method to first freely explore the data, before leveraging CWA to understand the specific domain constraints that impact data professionals. In doing so, we can approach the data from both an ecological (top-down) and cognitivist (bottom-up) perspective.

2.9 Chapter Summary

This chapter reviews relevant background literature about data professionals and their proposed workflows. We describe the KDP as a framework in which this thesis will use to describe sociotechnical challenges in the data profession. This chapter also reviews the Ackerman's lens of the sociotechnical gap used to understand the sociotechnical challenges that data professionals face. The next chapter will elaborate on the study methodology.

Chapter 3
Approach

This chapter describes the process by which the interviews were conducted to investigate the sociotechnical challenges that data professionals face. The sampling strategy and recruitment are first described followed by a description of the interviewees, and the interview procedure.

3.1 Sampling and recruitment

With this research study, we aimed to reflect the diversity within a given population, rather than create statistical generalizability. A purposive sampling strategy was used to generate insights by "selecting information-rich cases strategically and purposefully" [54]. With purposeful sampling, we deliberately seek to include outliers to understand more about the topic, rather than to attempt to generalize from our sample to the general population. This allows for deviant scenarios to be illuminated through the findings. Participants were selected based on three criteria:

1. Self-identification as a "data professional"
2. Stored, manipulated, or obtained insights from data
3. Collaborate with multiple teams or team members.

Also, a maximum variation sampling strategy, a subset of purposive sampling strategy, was used to select our participants [53]. The goal of maximum variation sampling aims to sample for heterogeneity and select a small number of cases that maximize the diversity relevant to the research question. The participants were recruited for diverse industries, experience, job functions, and educational background from North America and Europe.

The sample size was guided based on the concept of theoretical saturation [33]. "As a study goes on, more data does not necessarily lead to more information" [48]. Theoretical data saturation is reached when sampling more data will not result in new information related to the research question [64]. In other words, when the interviews illustrate similar instances repeatedly, themes and theories can emerge.

A total of twenty participants were recruited using two specific methods:

1. Twelve participants were directly contacted through email or LinkedIn and selected through the maximum variation sampling technique. The participants had job functions which fit the data professional description such as data engineer, data scientist, and data analyst.
2. Finally, using a quasi-snowball sampling technique [67], the final six participants were recruited based on the recommendation of other participants. Often, the recommendations were colleagues of the participants who helped complete the story. For example, a data analyst would recommend a data engineer to be part of the study. We filtered the recommendations from the participants based on the maximum variation sampling procedure.

All procedures obtained clearance from the University of Waterloo Office of Research Ethics (ORE) under the project 21312, titled "Understanding and Supporting Data Professionals In Complex Domains."

3.2 Interview procedure

The interviews were conducted between March and April 2016. When possible, interviews were carried out in person at the participants' workplace in a meeting room. All audio was recorded using QuickTime. Interviews P4, P14, and P17 were performed over Skype due to distance limitations. The interviews lasted between 45 and 90 minutes and were audio recorded and later, transcribed using a transcription service.

3.3 Interview questions

A semi-structured interview protocol, combining a predetermined set of open questions with freeform questions, was used to address the research questions. A structured list of eighteen questions was used. The researcher probed into interesting responses through an unstructured conversation. The interview had three main sections: background information, a walkthrough of two projects, and understanding the effects of domain knowledge on the KDP.

During the interview, we first asked interviewees to describe their educational background, age, experience, their current role, their team, and their typical day.

We next asked participants to describe in detail a specific project in which they had to ask and answer a question using data. We focused on eliciting and understanding the collaborations that existed throughout the project. Next, we introduced the KDP to the interviewees. We presented the

process using a diagram and a verbal one sentence description of each phase. The interviewees were asked to comment on the process and walk through a second project in detail using the KDP as a framework. The interviewees were asked about the difficulties they faced in each phase of the KDP. We probed into the strategies the interviewees used to cope with the challenges in each phase.

The rest of the interview focused on understanding the effects of a domain on the KDP as well as different methods of communication. Time was left at the end for an open discussion to allow interviewees to share anything else that they think would help us understand their work and their process. Finally, interviewees were asked to share names of colleagues that they think would be potential participants.

The interview data were later analyzed using two methods, thematic analysis and CWA. The results from both analyses are presented in the following two chapters.

3.4 Participants

The study consisted of twenty semi-structured interviews from 14 companies, spanning over 45 to 90 minutes. Table 2 below outlines all 20 participants. 14 were male, and 6 were female from the ages from 22 to 55. The participants had a diverse set of job titles such as data engineer, data scientist, data analyst, consultant, and quality improvement specialist. They varied in seniority from 1 to 20 years and worked in organizations of different sizes, from start-ups to multi-national companies. Their educational backgrounds ranged from some high school to a PhD degree (with most participants having some form of higher education). Most of our participants had formal training in finance, engineering, and statistics. Our participants worked from companies distributed across Waterloo (ON), Toronto (ON), Seattle (WA), and Copenhagen (DK). A diverse set of participants ensured that a diverse set of insights could be extracted. For example, we explicitly included data professionals that worked in non-traditional data sets such as map and gaming.

Table 2: Participant background

Participant	Job Title	Education Background	Experience	Industry
P1	Corporate Strategy	PhD	8	Corporate
P2	Operational Research	PhD	15	Healthcare
P3	R&D User Researcher	Bachelors	4	Technology
P4	Manager, Maps/Data/GIS	PhD	10	Geography
P5	Data Analytics and Reporting	Bachelors	7	Corporate
P6	Data Scientist	Bachelors	1	Technology
P7	Data Engineer	High school	6	Consulting
P8	CEO	Bachelors	5	Finance
P9	Business Intelligence Analyst	Bachelors	8	Finance/ Technology
P10	PhD Student	Masters	4	Healthcare
P11	Planning –Evaluate	Bachelors	10	Corporate
P12	Institutional Analyst	Bachelors	18	Corporate
P13	CEO	Bachelors	2	Technology
P14	Game Consultant	PhD	20	Gaming
P15	Consultant	PhD	10	Healthcare
P16	Data Engineering	Bachelors	6	Finance/ Technology

Participant	Job Title	Education Background	Experience	Industry
P17	Data Scientist	PhD	16	Technology
P18	Business Development	Bachelors	1	Healthcare
P19	Data Scientist	Bachelors	10	Technology
P20	Data Engineer	PhD Candidate	6	Healthcare

3.5 Chapter summary

Twenty data professionals from fourteen companies participated in 45 to 90 minutes semi-structured interviews. The interview probed into the collaboration and communication challenges and coping methods that the data professionals used. The analysis and findings are presented in the following two chapters, thematic analysis, and CWA.

Chapter 4
Thematic analysis

The purpose of the chapter is to describe the first lens of our analysis using thematic analysis. First, we will describe the method and then, walk through the findings using the KDP as an organizational structure. Our results illustrate the complexity of the KDP and highlighting the sociotechnical challenges that a data professional may face.

4.1 Method

Thematic analysis is a widely used qualitative analytic method used in social, behavioral, and applied sciences. The goal of thematic analysis is to identify, analyze, and report patterns (themes) across a dataset to answer a research question [9]. Patterns are often identified through a rigorous process of data familiarizing, data coding, and theme development. Based on Braun et al. [9], thematic analysis consists of six phases:

Table 3: Six phases of thematic analysis

Phase	Description of the process
1. Familiarization with the data	Transcribing data (if necessary), reading and re-reading the data, noting down initial ideas.
2. Generating initial codes	Coding interesting features of the data in a systematic fashion across the entire dataset, collating data relevant to each code.
3. Searching for themes	Collating codes into potential themes, gathering all data relevant to each potential theme.
4. Reviewing themes	Checking in the themes work about the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic "map" of the analysis.
5. Defining and naming themes	Ongoing analysis to refine the specifics of each theme and the overall story the analysis tells; generating clear definitions and names for each theme.

Phase	Description of the process
6. Producing the report	The final opportunity for analysis. Selection of vivid, compelling extract examples, final analysis of selected extracts, relating back to the analysis to the research question and literature, producing a scholarly report of the analysis.

Even though the phases are listed sequentially and build on previous steps; thematic analysis is typically a recursive process.

4.1.1 Thematic analysis approaches

There are many different approaches to generate the codes and themes for thematic analysis.

- **Inductive "bottom-up"**: Coding and themes are directed by the content of the data
- **Theoretical "top-down"**: Coding and themes are directed by existing concepts or ideas
- **Semantic**: Coding and themes reflect the explicit content of the data
- **Latent**: Coding and themes reflect the concepts and assumptions behind the data
- **Realist**: Focuses on reporting an assumed reality as described in the data
- **Constructionist**: Focuses on reporting on how a certain reality is created by the data

It is common that more inductive, semantic, and realist approaches are clustered together, whereas deductive, latent and constructionist approaches are clustered together. The specific approach that a researcher takes is not as important as the analysis being theoretically consistent and coherent.

The two primary methods that themes can be identified are: inductive and theoretical. An inductive approach refers to themes are strongly linked to the data and is similar to grounded theory. The themes emerge may have little to no relationship to the specific questions asked of the participants. An inductive approach does not consider any pre-existing framework or researcher's preconception. On the other hand, a theoretical analysis is driven by a theory and a researcher's research questions.

4.1.2 What is a theme?

A theme represents an important aspect of the data in relation to the research question and "represents some level of patterned response or meaning within the data set" [9]. It is important to note that more

instances of a theme do not necessarily mean the theme is more crucial than another [9]. In addition, there is no cutoff on how prevalent a theme is. For example, it is not the case that a theme is only a theme if it is present within 50% of one's data and not a theme if it only appeared 40% of the time. It is up to the researcher to determine what a theme is.

4.1.3 Multiple coders

A range of technical fixes including purposive sampling, grounded theory, multiple coding, inter-reliability score has been used in the past to confer rigor [5]. Many researchers such as Barbour [5] and Barry [6] believe that coding data with multiple researchers does not result in better coding or more accurate results. Multiple coders only result in different coding. The need for multiple coders and inter-rater reliability assumes that there is an accurate reality within the data that can be captured through the thematic analysis. Instead, thematic analysis is flexible and organic, with the themes evolving through the coding process, with no one accurate method to code the data. Inter-rater reliability scores can only illustrate that two researchers have been trained to code data identically, rather than their codes and analyses are "accurate."

For example, Armstrong et al. had six experienced researchers who independently coded one focus group transcript and it illustrated substantial deviations between how each researcher coded the data [4]. Some researchers argue that this is not surprising given the complexity of qualitative data and the range of backgrounds of the researchers [5].

The degree of concordance is not important; the value is within the disagreements and discussion used to refine the coding. The great benefit of having multiple coders is the ability to explore alternative interpretations. In other words, multiple coders allow for one to act as the "devil's advocate," and provide different perspectives. This specific exercise is used to encourage thoroughness within the analysis. Whether the analysis is carried out by a sole researcher or by a team, is irrelevant. Instead, ensuring the analysis follows a systematic process and is transparent within the concluding report is more important.

4.1.4 Quantifying qualitative data

Within a qualitative study, we do not quantify our results. Pratt [56] describe five dangers of quantifying qualitative data:

1. It may trigger a quantitative/ deductive mindset among reviewers
2. It may be misleading (small changes in response corresponds to large changes in percentage counts)
3. It may overlook "taken for granted meanings"
4. It may do "violence to experience," inadequately representing the voices of the individuals studied
5. It may simply create the "worst of all worlds": not enough of a sample for a statistically significant test, and too anemic a representation to adequately represent rich data

In addition, other researchers like Pyett [57] argue that "counting responses is missing the point of qualitative research, as frequency does not determine value". Quantifying responses from semi-structured interviews may not practical. During semi-structured interviews, the same questions are not always asked to obtain the same insightful stories. Even if the same questions are asked, they are often not asked in the same verbiage or order. Semi-structured interviews are analyzed through thematic analysis to answer *what, how,* and *why* questions. In comparison, research methods such as structured interviews and surveys are more appropriate to answer *how frequent* and *how prevalent* questions.

Measuring prevalence is not crucial to thematic analysis. Part of the flexibility of thematic analysis is it allows the researcher to determine the themes and prevalence in a variety of ways. What is most important is that the results are presented in a consistent method. Within this thesis, we adopt conventions for representing prevalence using descriptions such as "the majority of participants" [9] and "a number of participants" [10].

4.2 Data analysis

The interview data collected using the method as described in Chapter 3 – Approach, served as the foundation of the thematic analysis and inductive analysis [20]. To facilitate the in-depth analysis, the interviews were all transcribed using a transcription service. Each transcript was reviewed and

corrected for inaudible sections. By reading and re-reading the transcriptions, we became more familiar with the data and jotted down initial thoughts.

After the transcripts, had been corrected and reviewed, the transcripts were imported into *Atlas.ti*, a robust qualitative data analysis software for large bodies of textual, audio, and video data. *Atlas.ti* is a tool used by many institutions and corporations such as Harvard University, Google, and Microsoft. *Atlas.ti* allows researchers to explore transcripts in a systematic manner enabling open coding to be conducted. The software enables researchers to evaluate quotes side-by-side and established relationships between codes to create themes.

The thematic analysis began with a line-by-line coding of the transcripts, also known as microanalysis [20]. In the beginning, the goal of microanalysis of the process was to discover categories and to uncover relationship among different concepts. After the categories were established, the analysis focused on validating the themes and verifying relationships [20]. The coding process focused on first understanding the sociotechnical challenges and coping strategies. The phases of the KDP were used as initial codes [55]. Following the process as outlined by Hsieh & Shannon [38], any quotes that were not categorized that did not belong to the initial coding scheme were given a new code. The new codes were used to keep track of new concepts and themes that emerged from the transcript.

It is important to highlight that microanalysis does not mean that every single word was coded. Rather, we scanned the transcript for relevant material and quotes. When a section of the transcript was identified as being potentially useful, a line-by-line procedure was used to assign and code the relevant sections.

After the microanalysis was conducted, the codes were reviewed and grouped into potential themes. The themes were checked in relation to the codes. Next, using the generated themes as codes, a second pass of the data was conducted to validate the themes, verify relationships and strengthen each theme. The themes were checked against the related literature to evaluate the novelty of the findings. Each theme was also discussed within the research team as themes emerged. Discussions within the research teams were used to evaluate the strength and gaps within the themes.

As thematic analysis is an iterative process, it was carried out in parallel with the interview process. By conducting the analysis throughout the interviews, a theoretical data saturation level could be established.

4.3 Findings

We now report our main findings of illustrating the sociotechnical challenges experienced by data professionals in the context of the different phases of the KDP. Our results validate and extend the KDP as a framework to describe data professionals' workflow. Specifically, we will discuss the challenges faced by data professionals and the coping strategies used in each stage of the knowledge discovery process.

We have organized our primary results on the sociotechnical challenges that data professionals face around the critical phases of the KDP described above: 1) business/problem understanding; 2) data acquisition; 3) data validation and analysis; 4) data visualization and communication. We will illustrate how the underlying organization and domain create friction throughout a data professional's workflow. Finally, we will describe the emerging need of the translator throughout the process. Recognizing and identifying these challenges is the initial step in addressing the challenges. It is important to note that the prevalence and severity of each challenge are unique per data professional based on different factors such as the domain and the organization.

4.3.1 Problem understanding

Problem understanding is the initial step in the discovery process. To fully understand the problem, data professionals are required to understand the domain, business, and problem space. Two main challenges exist within this phase: 1) Data professionals are often presented with lofty goals, 2) Data professionals have difficulty obtaining knowledge about the problem space.

Within the problem understanding phase, data professionals need to understand the problem before moving to the other phases of the KDP. A data professional consultant described understanding the problem and business needs as a method of informing the solution. He explained that the technical aspects of the discovery process follow a similar basic framework in that he must understand the business requirements and apply them to the technological framework:

> *The majority of my time personally is spent in probably split between these [problem and business understanding] and making sure that the customers and people that we're working with understand what they're building and why… it was still the same basic framework and the same basic architecture and just transposed into a different set of requirements[P7].*

However, achieving understanding of what stakeholders needed was difficult. Organizations and stakeholders did not have a clear question that can be addressed using data. Questions to which stakeholders wanted the answer may require metrics that were not collected or metrics that could not be measured. The lack of clarity resulted in ill-defined questions or lofty goals, as described by a participant working in higher education:

> *[Stakeholders] have kind of grand ideas of what they want. But trying to narrow them down to say "ok you need one number or trend line to measure what you're trying to show" [P5]*

In fact, stakeholders or clients often did not appear to know what they actually wanted with data:

> *...the product team may have their request and sometimes they don't even know what they exactly want. It's that I want this, but it's like Henry Ford. When everyone was asking for a faster horse but Henry Ford knows what they actually want is a car. Sometimes they will submit their requests, we have a backlog system, we can't trash those requests, but sometimes you have to be careful with those requests. [P17]*

Next, we highlight different methods in which data professionals attempted to define the problem space in collaboration with stakeholders. The first method was to understand the problem domain. Regardless of the complexity of the domain, data professionals needed to understand the domain and the goals of stakeholders to refine the problem. In other words, by understanding the problem domain, data professionals could understand the context and goal of the stakeholders. Understanding the problem domain was difficult as it often required learning from domain experts, whom have limited amounts of time. Domain experts differed in each problem space, but ranged from product owners, doctors, accountants, to developers. A data professional within the finance industry highlighted the need of using domain knowledge as a method to refine the ill-defined problem:

> *The first thing we will walk through what is your motivation. Why do you want to solve this problem? Then we will ask them what kind of decisions are you trying to make based on the results and what kind of actions are you planning to take based on the results? How much business impact can this give us? Then the next question once we figure out that is what data do we have today? What do we need in order to answer those questions? [P17]*

Across most of our interviews, we saw data professionals seeking answers from domain experts in different faces of the KDP. Questions were often used to understand the functional purpose of an organization or problem space. For example, a data professional within the technology sector consulted a business owner to refine the problem space:

> *First I think you yourself have to understand what that business does, how that business functions. For instance, what is the value proposition of that group? What kind of value are they trying to drive for the business? How are they making money? What expenses are they looking at? And then what the bottom line, for me just coming from an accountant background that's one big thing I need to understand, how they generate revenue and then second, what other maybe not so tangible success factors are they looking at right and how do they add to the company and then from there you can say, "How can I add to help them? How can I drive value to help them accomplish their goals? [P9]*

Helping stakeholders understand what can be measured was the second strategy data professionals used to redefine a lofty goal. One data professional in higher education explained the need to understand *what* can be measured when presented with a problem. For example, stakeholders may want to understand *why* students succeed in certain courses or programs. However, a data professional must understand *how* success can be operationalized and measured in this context:

> *The first question is how do we measure something like this. Can we even measure what you're trying to look for? If the question is what makes students most successful? How do we measure success? [P5]*

A data professional from the games industry also described a similar strategy to help define lofty goals. Stakeholders often want to understand *why* something is happening, yet data can only track *what* people are doing:

> *An analyst can go to a designer asking them what types of behaviors are in you interested in having tracked. The designer will often ask them questions, give them behaviors that we cannot track. For example, what is the motivation of the player? Which for a designer is incredibly interesting? They want to know how people respond to their art and craft. If you then go to a back-end engineer saying, "Track the motivations of the players." They're like, "How exactly do you want to do that? Because, all I can track is what people do not why." [P14]*

An extra challenge with lofty goals was that data professionals experience frequent project requirements changes. The project requirements frequently change as problem understanding improves for both stakeholders and data professionals:

> *[Understanding what the business wants is] the biggest piece right and a lot of the times businesses may think they want one thing but then as you dig deeper you realize that that want something else [P9]*

4.3.1.1 Problem understanding is "social"

It is important to highlight that problem understanding and definition is not done in silos. Data professionals often sit together with the experts to determine details of a research question. The frequency of these interactions varies from constant side-by-side to monthly communication:

> *I have one on one's with the director of that business group on a weekly basis and that really helps me to understand what the group is trying to do, what is it that they're really trying to achieve and that really helps me do a lot and can make better decisions when I do my analysis. Bringing data points that they may not have asked for but I think is actually useful to them [P9]*

As the team has a variety of different skills and experiences, the problem of achieving common ground can be seen. One specific challenge with asking questions is knowing how to frame and ask the questions. A novice data professional working within the medical field illustrates this issue:

> *I never know how to frame it into the right question, so I just ask a question and then I ask another question afterward, if I'm not getting the answers I'm looking for. Sometimes I might be completely down the wrong track and you don't' realize it until you notice that there's a hole somewhere inside your data analysis. It takes a lot of sort of thinking about the problem. And I have a lot of time to think about it. So you find a lot of holes [P10].*

Overall, we found that problem understanding was a collaborative process, with data professionals challenged with lofty goals from stakeholders, requiring significant effort and time to understand the underlying problem.

4.3.2 Data acquisition and understanding

Once data professionals understand the problem, as per the KDP, they need to acquire the appropriate data. Although several streams of research are focused on the technical challenges of data acquisition such as data storage, data cleaning, and data management [2,39,46,63,65], we focused on our analysis on uncovering the sociotechnical roadblocks that prevent a data professional from acquiring and understanding data. Within this phase, participants described two main challenges: 1) Difficulty knowing what data is available and its source, 2) Trusting and understanding the data extracted.

It is important to point out that data acquisition and understanding can only occur if a data professional understands the domain.

4.3.2.1 What data is available?

Once a problem has been clearly defined, data professionals must understand to what data is available and what they can actually access. Only once data professionals understand what data is available and accessible, can they create a plan to extract data from the databases, validate the data and analyze it. This step of understanding the data often includes understanding the business processes that lead to the creation of data within a system. By understanding the origin of a data point, data professionals can validate and comprehend the meaning behind each data point. A data professional from an educational institution explained:

> *There are business processes that lead to that data in the system. So we try to understand how does this come to be, what is this number mean, why is there, and how does it represent the world. Once we answer that, we give them a sample set to say here's a preliminary look at what the data looks. [P5]*

While business processes define what data can be available, environmental constraints may have an effect on what data can be stored. An interviewee explained how environmental constraints of the oil and gas sector can affect the data that can be collected:

> *In oil and gas has a situation where they've got oil rigs and those oil rigs get disconnected from the internet all the time because of weather or solar flares. That means that their data acquisition policies and processes have to be different from the fact that they lose connection to the rest of the world for 2, 3 days at a time. That changes the technology implementation such that you have to have a data acquisition process on the rig itself and then a process to pull that back to a centralized location on a schedule [P7]*

In comparison, the same interviewee described how legal and regulations restrict the amount of data to which a data professional could access:

> *[In finance], their regulations affect what data we can use. For example, financial information in Brazil for example is not allowed to leave the country at all, flat out. You cannot have any personal information about any financial transaction leave that country [P7].*

Third-party data sources can also affect the type of data that is available for analysis. Third-party data sources included any data sets that is not collected by the organization itself. A data professional working with third-party data sources explained the struggles with such data, as he was unable to trust the origin of each data point:

> *It started off with getting raw data, ensuring that it's correct. We had raw data coming from a third-party tool, so we get the raw JSON coming in on a scheduled basis. There were tons of issues in terms of the API, that they were not sending the right data. Turns out there were rows that were deleted that they didn't tell us were deleted, so we were just assuming that they existed… When they were sending us the data, they weren't sending us the history. They were just sending us this one attribute at this point in time. [P16]*

The lack of trust within extracted data extend beyond third-party data. This trust could be established by keeping track of data provenance (i.e. the origin of the data and whether it is still updated):

> *What's really interesting about this type of analysis, especially when it's passing hands or even within an organization where the project may span multiple years, typically the person who's coded the database is not the person who's looking at the data afterward. [Name] had to spend a lot of time creating a legend trying to clearly identify what the database variables were, how he's coded them either by using ranges or whatever it may be, like time periods to silo out the data. He created a big legend for that which took most of the time I would say really because cleaning the data is the big pain. It was definitely not instinctual to just get into the data and start to look at it. We definitely had to clean it and make sense of it first [P20].*

4.3.2.2 What does the data mean?

The challenges that data professionals described were not limited to extracting data—in fact, a common problem that data professionals faced was the capacity to understand the extracted data as well as the source. Data professionals described how they were required to understand how data extracted is a representation of the world. A participant working in higher education, explains that even though he could extract the data and see the data, he did not understand what the data described:

> *When I first started, I had a real problem looking… I can see all the data that's coming out, but not understand what I should be pulling. You may have several measures that are named slightly differently… An example of that would be the count of students can be slightly different depending on what you're looking for... If you're picking the wrong one you're going to get vastly different results…Based on the question you may have you'll know what you should be picking. Initially, I had an issue because I didn't know where to get that context [P19]*

All our participants regardless of the domain described similar scenarios of not being able to comprehend extracted data. A data professional in finance recalled misunderstanding the origin and the meaning of two data points labeled in a similar manner:

> *[When] I first came here there was a lot of data points, and honestly there would be like current partnership type and then just a partnership type and there's tables that have the same name except four letter difference [P9].*

The problem appeared to compound as a misunderstanding of terminology often occurred within teams. One of the participants' narrative highlights this frequent pain point of understanding data:

> *It's more a misunderstanding of what terms mean. Team A may be defining a merchant as a billing relationship, but somebody in Team B may be looking at it as "one shop is a merchant." It depends on how the terms are defined, and then that will give you a very different understanding. When you're calculating some value, especially money-related, anything finance related, it has to be very clear what that value is or how you're calculating it and at what point that data is valid and when it is not valid. [P16]*

Once data professionals understand the data, problems often emerge. The root causes for discrepancies and anomalies within the data must be understood and addressed before the data analysis can be conducted. Without addressing and understanding the data discrepancies, data

professionals may observe misleading patterns. However, one data professional describes the importance of choosing the appropriate method of dealing with the problem. The method of addressing issues between colleagues and external partners differed:

> *There is a difference between saying, "Your data is wrong and I don't understand your data. Somebody [in the] presentation, said, "Saw some data, didn't like the number, you've got a mistake in a number, that number's wrong." What happened after that, in the hallway, was a confrontation between the 2 parties... There was heated discussion and raising of voices and defensiveness. It wasn't very collegial. [The] whole thing could have been avoided, if proper tact had been applied. Even though we might think that we understand our data 100% we should assume that something is missing and I am still wrong. [P11]*

4.3.3 Data validation and analysis

Our interviewees described that there were two kinds of validation activities that they normally participated in: 1) validating analysis results using common sense and basic statistical skills, and 2) validating results using the domain and context specific knowledge. While mastery of statistical concepts was important among our interviewees, it was often not enough to adequately validate the data thoroughly:

> *Even when you know the data type, and expected values, and their relationship, we still get questions [P11]*

All participants mentioned using common sense data validations techniques have been used to conduct sanity checks on the data set. For example: "*Is there are a date where a gender male or female value should be?*":

> *"You told me that age has expected value between 20 and 99. Why am I seeing A in the age column?" The data type has to be numerical, because it's a percentage of something. I see a word in there. You know something is wrong. [P11]*

However, common sense cannot be used to validate every scenario. Without the use of domain knowledge to validate the data, problems could never be detected and ultimately lead to a misleading result and decision:

> *For example, we know students couldn't take 50 classes in a term. So we see outliers or see something. That's a really simple example. But in other cases where it's not so easy, we may never know. Until a power user or someone says "chemistry students can't take ECON 101 how do you have ECON 101 there? [P5]*

The use of domain knowledge to validate data was difficult as data professionals often did not have the necessary domain knowledge. For example, a data professional working in healthcare was trying to analyze blood pressure data from an application to make recommendations. She was unable to validate this data as she lacked the domain knowledge. To do so, she needed to collaborate with clinicians:

> *Sometimes I would have to ask them [clinicians] is this a valid answer? Does this average blood pressure look right? When it comes to analyzing the data set ... does this relative risk look okay, or is this way out of bounds of what you would have expected? [P15].*

When analyzing the results, data professionals often notice unusual trends within the results. On investigation, issues may arise because of an error in the code or an error with the initial data. This often requires a data professional to re-analyze and review the data:

> *It would be something along the lines of you're evaluating the result of your experiment, your model or something, and you're looking at new data and you see something that doesn't fit into your preconceived notions of how the system could work and then that raises a red flag in your brain. You go, and you look at the data more closely and you realize there is some data that you missed in your validation phase [P5].*

4.3.4 Data visualization and knowledge communication

The final step of the KDP is data validation and communication. As discussed previously, HCI has a long history of developing effective visualization and communications. Our results illustrate the phase as highly social with the need of domain knowledge to provide context to the data patterns. This phase is also the most critical, as without proper communication of actionable insights, data professionals are not able to provide value to stakeholders and their organizations. This is essential as

74% of firms want to be data driven, but only 29% can act on the insights [23]. Our results illustrate that data visualization does not only serve as a vehicle for communication knowledge, but also as an intermediate step to understand the data.

The importance of data visualization and communication can be reflected in the time spent on this phase. A data professional from a large technology firm describes:

> *If I have to divide my time then I will spend 30% of the time to acquire the data, to validate the data. Then 30% of the time to do the real work, to do the data analysis, monitoring and etc, then 40% of the time to prepare for the presentation [P17].*

The challenges within knowledge communication are the ability to tell a story and to create actionable insights which impact the business. There were two aspects of data visualization that interviewees described: 1) using visualization techniques to enhance their understanding of the data and 2) presenting their findings to clients and other team members. One data scientist at a large technology company explained how visualization could also be used as an intermediate step in the analysis process to get feedback, and not solely for communication:

> *I will grab someone, another data scientist from my team and just go through the slides with him or her before the meeting and I'm just going to monitor how he or she is going to react. Then he or she needs to actually understand the graphs. If so that means perhaps I don't need to make some changes to my slide X [P17].*

One of the most challenging aspects of data visualization was communicating the results to an audience in an unfamiliar domain. For example, a data professional working in producing video games different data visualizations are required for different stakeholders:

> *Let's say that I am showing a spreadsheet built in Excel that showcases a variety of different indications about the behaviors of players in a game. You can then show such a spreadsheet, filled with numbers, to a level designer, who's used to thinking of terms of height and textures and lighting. They will look at the numbers and they won't understand anything. If I show them a heat map, if I show them, here is your game level and here is where people die, here is where people pick up weapons, here's where people encounter and talk to NPCs. That is more in their language, right? This is how they are used to thinking about the behavior of the customers, right? [P14]*

Data professionals not only need to figure out *how* to communicate the knowledge, but also *what* to communicate. There are often privacy issues that cannot be shared or reported. A data professional working in healthcare describes this problem:

> *It'd be very dangerous to just leave them into a space where they can see anything especially when it's dealing with health data. People can see very personalized health information and we don't want that [P13]*

Privacy concerns or personally identifiable data occur across different sectors other than health. Similarly, a data professional working in higher education explained that there are some data points which cannot be shared:

> *Usually, the conversation, manager level wise, has to happen when you're talking about permissions. "Are we allowed to report this data," or "This is the level that we're being ... The U15 is asking us for this level of data that includes student information, they want record level student information." We have to go through the Registrar's Office, and we also go through the ethics, same thing that you've done. [P12]*

When communicating data to stakeholders, novice data professionals struggle to focus on the communicating key findings and not on the details:

> *It took me a long time to get out of specifics and into the general. When you are in school you are taught to look at the specifics… When you're in a work environment they don't care about the specifics unless something goes wrong, only if something goes wrong you'll want to go in the specifics. [P1]*

The goal of data communication or knowledge dissemination is critical to provide impact within an organization. This phase is how data professionals can provide value. One important aspect of data communication is the ability for a data professional to defend their methods and choices to gain credibility and trust from the team:

> *They may ask you, they may argue. Have you thought about other methods? Have you thought about different algorithms?... It's like defending a thesis. [P17]*

4.3.5 Summary of challenges in the Knowledge Discovery Process

Our results indicate that the KDP as highly social, collaborative and dependent on domain knowledge. By understanding the sociotechnical aspects of the KDP, we became aware of many challenges data professionals faced in every stage of the workflow. We highlighted five key themes related to these challenges:

1. Dealing with lofty goals, resulting in frequently changing project requirements that often affect the purpose and scope of the project.
2. Understanding the problem domain
3. Trusting and understanding acquired data
4. Unable to fully validate data and results
5. Understanding *how* and *what* to communicate

A summary of the challenges can be viewed in Figure 4.

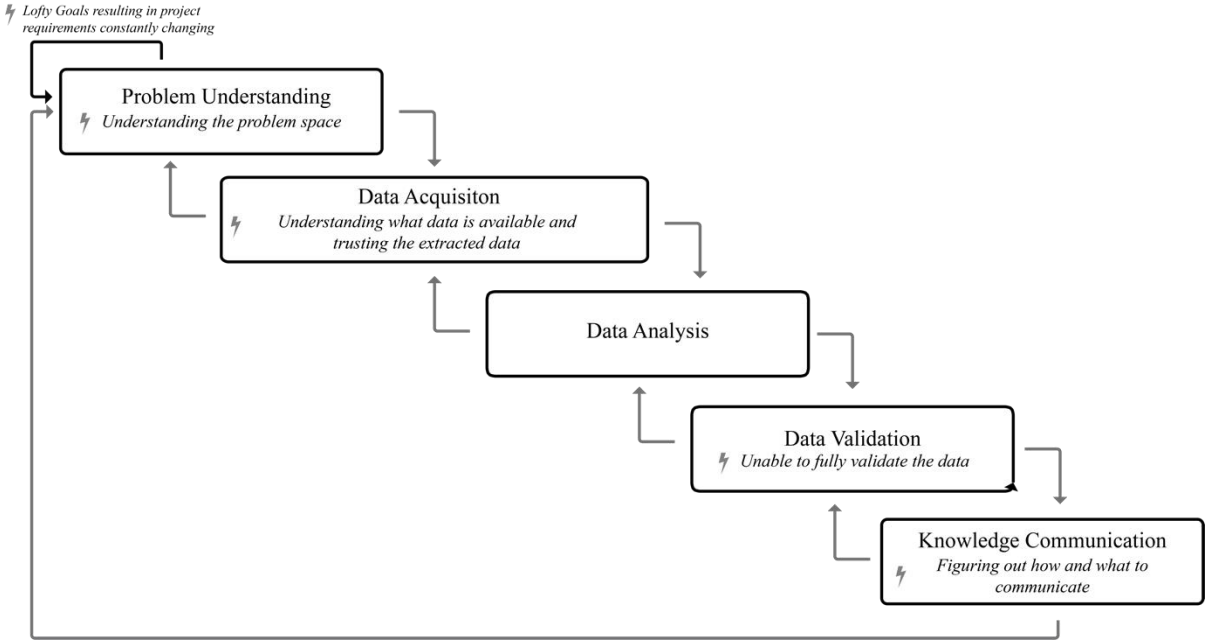

Figure 4: Outline of the challenges within the Knowledge Discovery Process

4.4 Data made relevant by domain knowledge

The underlying need for domain knowledge is apparent throughout the KDP. Domain knowledge can be used to make sense and refine a goal, validate and identify data anomalies, as well as convert data patterns into actionable knowledge. Given the importance of domain knowledge, we observed that data professionals exhibit multiple efforts to obtain domain knowledge. One of the most common initiatives our participants described was using web searches to gain a basic understanding of the terminology need to ask the correct questions and achieve common ground with different stakeholders:

> *If you're walking into something blind, then Google search is better than nothing… There's a discovery process in every single one of these meetings. Generally, what happens is I'll walk in and I'll say I talked to the sales team, I talked to the account team, I talked to X, Y, and Z. This is what they told me, this is what I understand, help me fill in the gaps [P7].*

Even though data professionals took the initiative to understand the domain, it was insufficient for understanding the nuances and organization specific details. Data professionals relied heavily on domain experts to learn and clarify details. The strategies in which data professionals attempted to understand a domain differ according to the domain. A consultant described some strategies he used to build his domain expertise before meeting a client:

> *It depends on the domain. In a lot of times it's basically just talking to the customers, the businesses have problems. They know they have problems, they know what the problems are, but they don't necessarily know have a methodology for addressing those problems… In the case of oil and gas, you tend to learn a lot more from talking to people in the industry than you do from Google searches. If you're walking into something blind, then Google search is better than nothing. [P7]*

Other interviewees went as far as integrating themselves into the domain to understand exactly how domain experts worked and functioned. A data professional working in healthcare transportation described the need of being in the field to understand the work processes. This type of work required additional effort beyond their expected duties and day job:

> *In the beginning, I spent a lot of time understanding the field… So understanding how the processes work. So I'm trying to analyze medical transfers, right? So I know how the transfers are supposed to be done. I've watched dispatchers do multiple transfers. I've talked to the doctors on how they use this data to change their transfers. So I've done a lot of observation studies to understand the process…you have to gain enough domain expertise that you can at least ask the right questions a lot of the time. If I had a specific question I had, I would go and maybe I would conduct another study to understand that question element [P10].*

However, there was a saturation point in which understanding more about a domain may not help a data professional understand more about a problem and data set. A data professional consultant explained this further:

> *At a certain point, the problem understanding [and] domain understanding is useful for relevance and useful for understanding a set of requirements but after that it doesn't necessarily impact what you're doing with the data. [P7]*

The impact of domain knowledge in each of the phase is summarized in Table 4.

Table 4: Overview of the use of domain knowledge in the Knowledge Discovery Process

Phase	Impact	Role of domain knowledge
Problem Understanding	High	Domain knowledge is used to refine ill-defined problems and the provide context to the problem space.
Data acquisition and understanding	Medium	Business processes can be used to help data professionals understand the origin and relationships between data points.
Data validation	High	Domain knowledge is used to validate the data set and any data patterns.
Data Analysis	Low	
Data visualization and knowledge communication	High	Domain knowledge is needed to convert data patterns into actionable insights that be used by an organization

4.5 Organizational structure

In this section, we outline challenges that the organization may impose on the data professionals. We highlight communication issues and technological constraints that can occur based on the organization process and structure.

Many participants described how organizational hierarchy sometimes resulted in a game of "broken telephone" where different pieces of information were lost when moving up and down an organizational hierarchy. In this scenario, a game of "broken telephone" occurs between a data analyst and a manager between two stages of the KDP process: problem understanding and data acquisition. One of the issues highlighted by the participants, was the need to understand the problem "up" the reporting chain and conveying the same message "down" the reporting chain. This often results in the message being "lost in translation":

There might be a question that came from up top and by the time it gets four levels down, it's like broken telephone, something changes along the way. [Data analyst] are trying to pull something but because they haven't been involved in the full context of all the discussions and we [managers] don't necessarily want them to be there because it could be months of meeting before we arrive at something. But because they don't know all the discussion, the barriers and all the things we kind of had to break through to arrive at that question. To them providing a data point is easy but may not tell the whole thing. That might be the wrong data point or it could be one of many issues... I think the biggest challenge is working up to the hierarchy and then working back down and think getting lost in translation. [P5].

Similarly, this issue can be seen when organizations separate each phase of the KDP into specific roles. A data professional who has 5 years of experience of working in both a financial and technological domain, compares two different organizational structures. In finance, the data professional worked in silos, resulting in an assembly line behavior and messages being "lost in translation". The interviewee described the organization structure as:

[In my previous financial organization] It's almost like assembly line, and when you do assembly line people just get really good at what they do and therefore it's like, "I'm the BA I can do really great requirements, I'll throw it to you and you can translate it really well to the technology team". Then it's almost like a segmentation of duties... We use to play broken telephone and they don't have a good understanding of what the business looks for, like the business wants a Honda and they're giving them a Ferrari right, like things like that I've seen many time where a lot of projects even failed because it was not what the business was looking for. Then there's misinterpretation of business requirements. [P9]

In contrast, the interviewee described the technological organization as a collaborative organization, where she could see the value of her work and understand the context of each problem better. Most of our data professionals that worked in a collaborative environment described similar experiences:

> *Here you have to understand everything so I think it is harder to maybe hire for somebody who really has that full stack understanding from understanding the problem to do all that and getting the data piece in as well and on top of that it takes more time in terms of, you have to do all this, as opposed to, "Here you're only doing up to here and then you're done and you're going onto your next project", whereas here you're doing the whole thing so it would take longer to really get that whole project done. When I was in [my previous organization], I would probably work on five projects at a time. Here I probably working on two big projects right now but the thing is I think my understanding of it is so much better, so I feel like the results of it will be that much better. I feel like I can see the value of my work here. [P9]*

On top of the effect of the organization structure on a data professional's role, the organization may constrain the type of technologies available to a data professional. Organizations have limited resources to purchase software licenses, which may restrict and redirect the work that data professional can complete. A data professional in the education industry illustrated that the organization only had Microsoft-based tools and was not allowed to use any open-source software like R. This organization constraint limits the type of work that data professionals could conduct:

> *[In this organization], I use Microsoft-based solutions; Microsoft Office, SharePoint, Tableau. That's our set of tools... Based on where I came from and what tools I used to have, Sequel Server, the Microsoft Stack, all these web-based solutions; I really had everything at my fingertips. Now I've come here and I've been told, "We use Excel and Database Access."... I don't know if you've watched or know of MacGyver where he's given very little tools and has to build something. [P19]*

4.6 Data professionals as translators

Establishing common ground between stakeholders and data professionals is a challenge throughout the KDP. To compensate, the need for a "translator" has emerged within the KDP to assist with establishing common ground. Evidence of the need for translators exists within our interviews. The KDP is comprised of multiple stakeholders with different backgrounds and experiences.

In this section, we describe the two different types of translators we identified in participant interviews. These types are not intended to provide an exhaustive taxonomy of the different translators that exist. Rather, the intent is to provide characteristics of the kinds of translators that arose from our interviews. The two different types of translators that our participants described are 1) human as a translator and 2) artifacts as a translator. In this section, we will outline the scenarios in which each type of translator is useful. In the discussion chapter, we will further reflect on the implications of the role of the translator on education, and the KDP.

4.6.1 Human as a translator

The first type of translator is a human throughout the KDP. Within our interview, the data professional assumed the role of a translator. The role of the translator is to facilitate a two-way conversation between the data and stakeholder. The translator is needed most to convert business requirements into data requirements and well as convert data insights into actionable business insights. In a sense, a translator converts concepts in one language to equivalent concepts in a target language. In our interview, this unique role is often taken up by the data professional. The translator is needed most in the problem definition and problem understanding and the knowledge communication stages of the discovery process.

The different languages that various stakeholders speak throughout the KDP is described by a participant in the games industry who described his experience conversing with a designer:

> *A very common problem is that you have different languages, different types of training between the different types of stakeholders. That can put a barrier for communication because an analyst can go to a designer asking them what types of behaviors are you interested in having tracked. The designer will often ask them questions, give them behaviors that we cannot track. [P14]*

Most our participants exhibited the need to translate data, concepts, and requirements throughout their job. One data professional working in higher education expressed the need to translate all the business requirements into data requirements:

> *Everything was business and they didn't have anyone quantitative, so numbers. So they were all talking business. And mentally I had to convert it and actually do it myself. [P5]*

Another data professional in technology went as far as describing being a translator was his primarily job function:

> *The thing is, my job is primarily to facilitate taking the idea from the client who's usually not very tech-savvy and converting it to what tech-savvy people know. When you try to bring the tech-savvy language into the client space, it does one of 2 things. Either it confuses them really badly and that's never a good thing because if they get confused, then they start getting emotional about. Because they'll say, "Oh, I should understand this." They start freaking out about things even though it's not a problem. [P13]*

A translator is most needed within the problem understanding phase to convert business requirements into data requirements. These translations are not straightforward, as translators must be able to stitch together the appropriate data. For example, a data professional described the need to understand the technology as a method to comprehend what business requirements were feasible:

> *They [stakeholders] start putting new things or making suggestions about what we should do and it totally screws things up. Because if you have to include things that they don't understand, then it ends up breaking the [technological] architecture pretty badly. [P13]*

4.6.2 Artifacts as translators

The second type of translator is an artifact often created by data professionals. These artifacts were often used for documentation to translate and describe the attributes of data. One interviewee in education, who worked with multiple clients created a "data dictionary." A "data dictionary" is a created artifact used to describe in detail the data set, assumptions, and results in the discovery process. Artifacts are powerful in outlining the details and technical aspects of the discovery process. These artifacts are most important within the data analysis and verification stages of the discovery process to help data professionals understand how a dataset may represent the world. A data professional described a data dictionary as:

> *[Data dictionary is] a template that we've created that has the title, description, what their research question was that led to that data, and have the technical conclusion and any notes that we did [P15].*

Our interviewees created different variations of a data dictionary. Each artifact created was used to communicate to stakeholders about the specific data attributes and details. A data professional and

accountant described spending a full year to convert accounting logic and terminology into technical terms in which a database could understand:

> *It was a pain in the XXX to do once, now this is quantified. The rules are set, but what we had to do is translate the world of accounting into a structure so that we could organize the database the way we want it, so that I could pull the data out in any way I want it [P8].*

The use of spreadsheets as a translation tool was a common theme between our data professionals. Another data professional at a technology company mentioned the use of Microsoft Excel as a method to convert technical database columns to business terms:

> *The finance guys organize it like this in the spreadsheet and explain it to the technical guys. Then the technical guys go, of course, that makes sense and they go do it…Mapping those relationships, the other spreadsheet ... All of that had to be organized first because it's not arbitrary and the rules are not ours. They're accounting rules. So we didn't make them, we had to translate them into the system. That's what I meant by it was a pain in the ass, because that wasn't fun to do. [P7].*

One of the drawbacks with artifacts is that they often become stale and out of date. However, the type of organization affects the usefulness of documentation as described by a data professional who worked in an e-commerce company and a finance company:

> *Here [at an e-commerce tech company] we're growing so fast, we can document something and then in six months' time it's out of date, whereas with banking because it's been going on for so many years, once you document it the changes aren't significant, so that even if you were to take a document from five years ago there might be changes but it's still usable [P9].*

Another issue with documentation, as described by one interviewee, was that documentation was scattered in across multiple places:

> *Initially, I had an issue because I didn't know where to get that context. Sometimes people had it within some Word document that they have sitting somewhere. Sometimes it's in an e-mail; someone has a really great e-mail description of it [P19].*

Through the examples, we can see artifacts are often used to be translators to communicate concrete information about the data.

4.6.3 Evidence of the translator throughout our participants

The need and the role of the translator are displayed throughout our interviews. In this section, we will describe the type of translations that were conducted per participants and any artifacts they created.

Table 5: Participant description

Participant	Job Title	Type of Translations
P1	Corporate Strategy	• P1 acted as the middleman between the business and the data engineers • Translating business requirements into data needs with a focus on converting data insights into actionable insights
P2	Operational Research	• Embedded himself into the medical field to "learn their language" and become the translator between the hospital and the research team • Leverages Microsoft Excel as a means to transform medical terminology into mathematical terminology
P3	R&D User Researcher	• Required to convert his team's design needs into a machine learning team • Translates results from the machine learning team into design requirements and decisions
P4	Manager, Maps/Data/GIS	• Understands the need of his clients to suggest the appropriate data sets and resources
P5	Data Analytics and Reporting	• "I'm more of the middle man between business and technical." • Creates a data dictionary to communicate research question, data set and data insights

Participant	Job Title	Type of Translations
P6	Data Scientist	• N/A
P7	Data Engineer	• Learns about a domain through internet searches before meeting clients to understand the business requirements and suggest a technological solution
P8	CEO	• Created a document converting data columns into financial terms
P9	Business Intelligence Analyst	• Converts business requirements into technological requirements – "I would say probably the middle man between technology [and the business], I guess I have a better understanding of the business but can also do some of the tech work, but I'm not so deep into it, so I don't really have that much stats that I understand, I understand the top stack [P9]"
P10	PhD Student	• Translates data results into actionable design decisions • P10 embedded himself into the domain to understand the domain better
P11	Planning – Evaluate	• Converts data results and insights into a report that all stakeholders will understand
P12	Institutional Analyst	• N/A
P13	CEO	• Describes his primary job function is a translator
P14	Game Consultant	• Describes the importance of being able to communicate and translate data insights to designers

Participant	Job Title	Type of Translations
P15	Consultant	• Understands the need of her clients to suggest the appropriate data sets and resources
P16	Data Engineering	• As an engineer, his focus is to translate data sets into business knowledge
P17	Data Scientist	• Converts data insights into business insights that can be understood by the stakeholders
P18	Business Development	• "I would be translating that analysis into terms our stakeholders would understand. The analysis, they perform the analysis but then I have always had roles where we are taking the analysis and creating, writing the story that explains with how"
P19	Data Scientist	• Prepares a data dictionary to communicate research question, data set and data insights
P20	Data Engineer	• Translates data insights into actionable insights that stakeholders can act upon

4.7 Chapter summary

In this chapter, we analyze our data using thematic analysis. We highlight the sociotechnical challenges within each of the stages of the KDP. Our results indicate that the process as a highly social, collaborative and dependent on domain knowledge. We highlighted five key themes related to these challenges:

1. Dealing with lofty goals, resulting in frequently changing project requirements that often affect the purpose and scope of the project.
2. Understanding the problem domain
3. Trusting and understanding acquired data
4. Unable to fully validate data and results
5. Understanding *how* and *what* to communicate

Next, we highlight the role of the organization and domain knowledge within discovery process. Finally, we reflect upon the two types of translators that emerged from our interviews 1) humans as a translator and 2) artifacts as a translator.

Chapter 5
Cognitive Work Analysis

The chapter introduces CWA as a second lens to analyze the data professional's sociotechnical challenges within their workflow. To do so, we follow a systematic approach to evaluate the impact of the work domain and organization on a data professional's workflow. First, we will describe CWA, a framework consisting of 5 phases:

1. Work domain analysis (WDA)
2. Control task analysis (ConTA)
3. Strategies analysis
4. Social organizational and cooperation analysis
5. Worker competencies analysis.

Next, we will conduct the first two phases of CWA to uncover the constraints and task within the KDP. The goal of CWA is to discover the constraints that shape the workflow and behavior of data professionals.

5.1 Method

CWA is a framework developed by Kim Vicente to design "safe, productive and healthy computer-based work" [69]. CWA focuses on analyzing complex sociotechnical systems or "systems containing social, psychology and technical elements" [69]. CWA is based on an *ecological* approach, which suggests that human behavior cannot be predicted unless one understands the situation or context or environment the *actor* was in. From this perspective, it is most important to identify the constraints that the environment imposes on user's actions [69]. This is contrary to the dominant *cognitivist* perspective in psychology, and HCI, which suggest that the human cognitive system constrains the environment [71]. From this perspective, it is most important to identify a *user's* mental model of the work domain and then design solutions for the work domain. In other words, CWA considers people as actors involved in their work-related actions, as supposed to users of the system.

CWA focuses on categorizing constraints and describing how these constraints shape the work domain. CWA consists of 5 phases. Table 6 describe each phase of the analysis and the question it seeks to answer.

Table 6: Description of the five phases of CWA

Phase	Purpose
Work Domain Analysis	Identify a fundamental set of constraints on the actions of any actor
Control Task Analysis	Identify the constraints on *what* needs to be done
Strategies Analysis	Identify *how* it can be done
Social Organization and Cooperation Analysis	Identify how social and technical factors build on and inherits the constraints identified in the first three phases
Worker Competencies Analysis	Identify how the constraints in the first four phases influence the competencies of the actors needed to function effectively in the domain

There are three different types of work analysis approaches. *Normative approaches* describe how a system should behave. *Descriptive approaches* describe how a system behaves in practice. *Formative approaches* specify the requirements that must be satisfied so that a system could behave in a desired way [69].

In the following sections, the first two phases of a WDA and ConTA will be described and conducted. A formative approach was taken for throughout our analysis.

5.1.1 Phase 1: Work Domain Analysis

WDA is the first stage of CWA, and it is used to examine the fundamental constraints that a user wants to control or have information about. A work domain is "the system being controlled, independent of any particular worker, automation, event, task or interface" [69].

Introduced by Rasmussen et al. [60], the abstraction hierarchy (AH) is the primary tool used to analyze work domain constraints. The AH has five levels: 1) functional purpose, 2) abstract function, 3) generalized function, 4) physical function, and 5) physical form. Each level represents a different abstraction of the system. Using *how* and *why* relationships the different levels of the AH can be linked together. An AH illustrates the system constraints needed to achieve a work domain's purpose, independent of any specific actions or people.

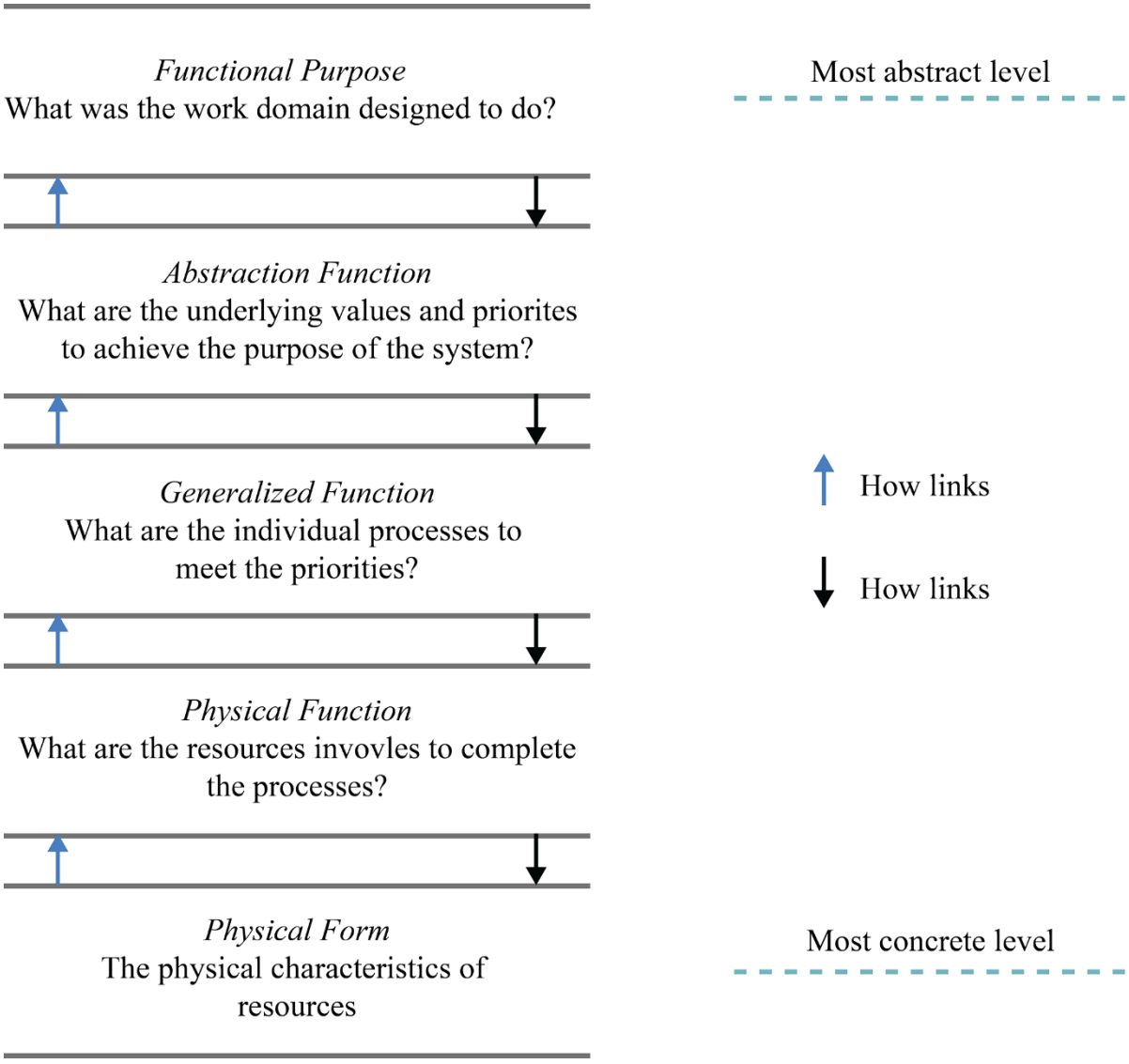

Figure 5: The five levels of the abstraction hierarchy and their connections (Adapted from Burns and Hajdukiewicz [12])

The top level of the AH focuses on defining the purpose and goals of the system whereas the lowest level focuses on the physical components of the system. Each level in the AH is a unique perspective of the work domain.

Various types of data collection methods can be used when performing WDA. In the past, field observations [59], document analysis [7], and interviews with subject matter experts [51] have been used for WDA.

5.1.2 Phase 2: Control Task Analysis

The second phase of CWA is ConTA, which explores the *actions* involved to achieve the domain's functional purpose. Unlike in Phase 1 of CWA, the focus of the analysis changes from the thing being controlled (the work domain) to the requirements associated with effective control (control task). The output of this phase describes in detail what tasks must be performed within a domain. However, ConTA does not describe *how* the task should be performed nor *who* should perform the task. These aspects are the focus of strategies analysis, social organization and co-operation analysis, and worker competencies analysis. The basic relationship between the control task and the work domain can be seen in Figure 6.

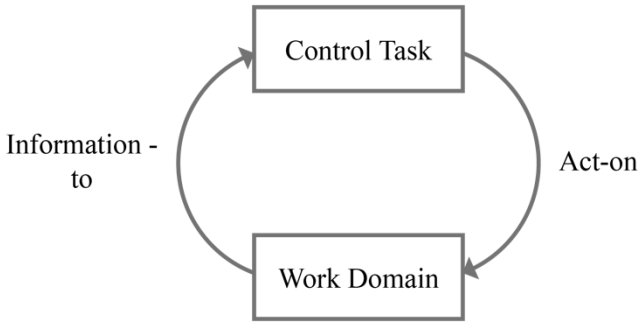

Figure 6: Basic relationship between the control task and work domain (Adapted from Vicente [69])

A decision ladder (DL) is used to support to ConTA. A DL is a tool for describing *what* task must be done to achieve the functional purpose. DLs are comprised of *information-processing activities* and *states of knowledge.* Information-processing activities are cognitive or computational activities in which an actor must engage to complete a task (depicted as boxes). States of knowledge are the products of the activities (depicted as circles). In this model, multiple steps are involved in decision-making, transforming one state of knowledge into another. While actors may follow a linear sequence between different stages of the DL, expert actors are likely to take shortcuts known as *shunts* and *leaps* [60]. Shunts connect activities to their output states of knowledge, and leaps present a link between two states of knowledge through association.

It important to highlight that the constructs identified on the DL can be associated with any actors, human or machine automation. This is possible as in this analysis, *what* needs to be done and *who* should do it is decoupled.

Similar to WDA, research suggests empirical research methods for performing ConTA. In the past, field research, task analysis methods, and surveys [59] were used to inform the ConTA.

5.2 Analysis

CWA is a framework for analyzing complex sociotechnical systems. A sociotechnical system is one that relies heavily on its overall function on social processes of communication. Vicente [69], describe different types of complexity in a sociotechnical system as 1) large problem space, 2) social, 3) heterogeneous perspectives, 4) distributed, 5) dynamic, 6) potentially high hazards, 7) many coupled subsystems, 8) automated, 9) uncertain data, 10) mediated interactions via computers, and 11) disturbances.

To begin the analysis, we examine the nature of data professionals and their organizations by using the two of the most apparent complex system characteristics within the data profession. The analysis helps outline the fit and the use of CWA within this domain.

Social: KDP is highly social process, requiring multiple stakeholders with different perspectives to work together as a team. As seen in Chapter 3, data professionals are required to communicate with various stakeholders throughout the process from understanding the problem and the domain to communicating the results.

Uncertainty: Data professionals encounter uncertainty within data daily. Boukhelifa et al. describe five different types of uncertainty: 1) errors, 2) imprecise or inaccurate data, 3) inconsistency, 4) missing or unknown data, and 5) vagueness, ambiguity, and fuzziness [8]. The 5 types of uncertainty can be reflected within our participants. For example, uncertainty may be introduced through errors in capturing data where the data may not reflect the ground truth or through imprecise measurements due to sensor error.

CWA is an appropriate analytical method due to the complexity of the work domain of data professionals. Specifically, we took a formative approach to WDA, seeking to describe the constraints of the work at a perspective of abstraction that was independent of the work of any one data analyst in any one context. This is a challenging approach, but allows for a more insightful look at the cognitive work which was occurring in the domain. In particular, WDA investigates the constraints that drive the sense-making function in which data professionals engage.

5.3 Phase 1: Work Domain Analysis

In the first analytical phase, we conducted two WDAs using the interview data collected as described in Chapter 3. Separated AHs of the data professional and the problem space were created. The goal of the WDA is to understand:

1. What was the goal of the data professional and how is it achieved?
2. What is the relationship between the constraints and the achievement of higher-order purposes?
3. Observe how the problem being analyzed impacted and constrained the work of the data professional.

The AH is used as the output of phase one of the analysis as well as structured analytical activities. To start, we identified the Functional Purpose (Goal) and the Physical Functions of the work domain. Next, we progressed inwards from the top and bottom of the AH by analyzing the *why* and *how* relationships to populate the Abstract and Generalized Functions. This specific approach to the WDA was informed by existing WDA examples [69]. The accuracy and completeness of the WDA were evaluated by probing *why* and *how* questions while traversing through the AH.

5.3.1 Data professional Work Domain Analysis

In the first WDA, we focus on understanding the work domain of data professionals. This sets up a system boundary [11] for the WDA. Figure 7 illustrates the abstraction hierarchy.

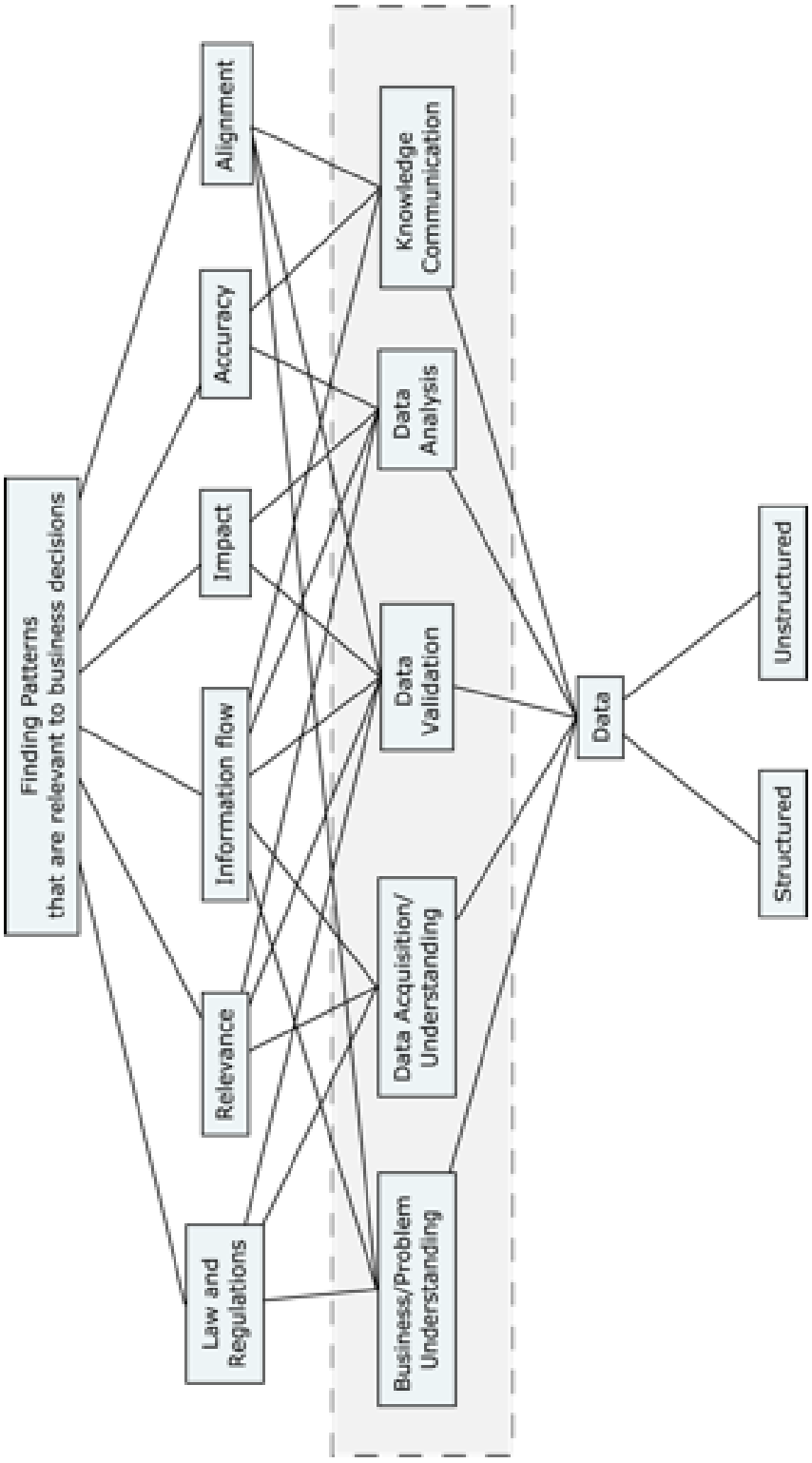

Figure 7: Data Professional abstraction hierarchy

5.3.1.1 Functional purpose

The functional purpose illustrates the objective of the data professional's task. The purpose of the KDP is to find meaningful patterns in a resource-efficient manner. (Figure 7, Level 1). The purpose can be separated into two different segments: meaningful and efficient. The goal of KDP is to extract patterns from data that can be acted upon. This first aspect of the functional purpose is critical, as only 29% of organizations are successful at connecting analytics into action [23]. Meaningful patterns are necessary for organizations to make data-driven decisions. The second aspect of the functional purpose is efficiency. Data professionals have limited time and monetary resources. If insights are not timely, the patterns may become too stale for stakeholders to act on. For example, if a business decision needs to be made by the end of week, a data professional cannot provide insights the following week.

5.3.1.2 Abstract function

The abstract function defines the principles, priorities, and values constraining how the functional purpose can be achieved. We identified six main abstract functions:

- **Alignment**: When a pattern is tied closely towards business goals and strategy, the more likely the pattern will lead to action. For example, patterns that impact businesses' key performance indicators (KPIs) such as engagement will inherently create a sense of urgency.
- **Accuracy**: As a priority, patterns that are found should be an accurate representation of the world.
- **Information Flow**: The principle of information flow governs how data is transferred, manipulated, and analyzed.
- **Relevance**: As a priority, patterns that are discovered must be relevant to the research question and goal. The same knowledge may be relevant for a set of stakeholders and irrelevant to others.
- **Impact**: Patterns that are discovered are only as valuable as the impact that the insight can have on an organization. This is a principle that data professionals must follow in order to achieve the functional purpose.

- **Law, regulations, and policies**: The laws, regulations, and policies of the domain affect the type of data to which a data professional may have access, limiting the types of questions and analysis that can be conducted.

Each of the abstract functions feeds into the generalized function to ensure that insights are actionable.

5.3.1.3 Generalized function

At the generalized function level, the five step KDP as illustrated in Figure 2 describes the main processes within the domain (Figure 7, Level 3). The specific functions are: 1) problem understanding, 2) data acquisition and understanding, 3) data validation, 4) data analysis, and 5) data visualization and communication. Using the CWA as a visualization tool, the constraints and potential sociotechnical challenges that exist within each phase can be highlighted. For example, we can observe the impact in which the abstract function affects and constrains each phase of the KDP.

5.3.1.4 Physical function

The physical function shows the physical components within the domain. Data is the main physical function (Figure 7, Level 4).

5.3.1.5 Physical form

At bottom level of the AH is the physical form, illustrating the operational conditions or attributes. There are two main categories of data:

1. **Structured**: Data that has a high degree of relationship, such as the use of relational databases. Structured data has specific characteristics such that each row has similar attributes or columns that make it easily searchable.
2. **Unstructured**: Data that has no identifiable or pre-defined structure and has no value until identified and stored in an organized fashion. Unstructured data accounts for 80% of information that is used to make business decisions [37]. An example of unstructured data is a customer review or call center conversations.

Each data set consists of multiple data points, each with different attributes such as origin, source, insert time, update time, and history.

5.3.2 Problem space Work Domain Analysis

Each data professional works in a unique problem space and domain, and are presented with their own set of unique constraints and processes. The interconnected relationship between the problem space and the data professional should be explored. The model represents the problem space within which the data professionals are working. The objective of the problem space model was not to model a specific problem space, but to take a formative approach and examine the sense-making problem that data analysts face. With this approach, the work of data analysts can be understood without reference to the descriptive contexts in which they work. This follows a similar approach that Burns et al. [11] used to model the problem-solving task that frigate commanders needed to conduct in naval command and control scenarios. Commanders must evaluate based on specific criteria whether another ship is friendly or enemy. This sense-making task was described in their WDA formatively using an AH of sense-making functions that are used by the frigate commanders to understand the nature of the contacts within their control space. A similar approach has been taken to formatively look at the sense-making task of data analysts by modeling the functions that data analysts must satisfy to understand their data problem space appropriately.

To create the second model, the system boundary of the problem space was established. The system boundary was set to focus on the constraints and functions that the data analyst must learn about and satisfy to have a strong understanding of the problem. This boundary could be small or large, depending on the scope of the problem space in which the data analyst must work. The AH is a model of sense-making queries and functions of the things the data analyst must learn about and understand to be able to understand the data problem properly. By completing this AH, the analyst solves the data problem they are solving. Being formative, this model extends across different contexts.

In this chapter, we will provide examples of a hospital CWA based on the work of Chow [17], and interview data we collected as described in Chapter 3. Each level of the AH focuses on uncovering the questions that are needed to be answered to help a data professional understand the problem space and go through the KDP. The problem space WDA can be seen in Table 7.

Table 7: Problem space work domain analysis

Phase	Description
Functional Purpose What is the purpose of the organization?	The fundamental purpose of the organization or team becomes the functional purpose. In the majority of businesses, the purpose will be to increase revenue. From Chow [17], in a hospital setting, the functional purpose maybe to improve patient safety and reduce readmission rates
Abstract Function What are the values and priorities of the organization?	In a hospital, the priorities may be quality of care, time to treatment, and number of patients treated.
Generalized Function What are the standardized procedures of the business to achieve the priorities?	In a hospital, the functions may be to admit, treat, and release patients.
Physical Function What are the resources involved to complete the processes?	In a hospital setting, the availability and state of the facilities and personnel are a constraint within the system.
Physical Form What are the attributes of the resources involved in the process?	In a hospital setting, the location and hours of the facilities, ER, and physicians may be the attributes of the resources.

The generic model illustrates the problem space data professionals are trying to understand, using the data available to them and the KDP. In the following section, we will describe how the model can be used by data professionals. The analogy is similar to how in Burns et al [11], frigate commanders used an undefined model that the commanders "fill out" as they evaluate whether another ship is friendly. In parallel, data professionals "fill out" the model as they learn about the domain of analysis.

5.3.3 CWA and the Knowledge Discovery Process

The CWA presents a unique systems-based approach on the KDP. In this section, we will describe the relationship between the two created models (Figure 7 and Table 7) and the discovery process. Traditionally, literature on the KDP has focused on the activities that are executed at each of the phases. The results of the CWA provide insights to the following questions about the KDP and data professionals:

- What is the purpose of the KDP?
- What information is required for data professionals to achieve their goal?
- What environmental factors impact the KDP?

Through the WDA (Figure 7), we can clearly model and outline the constraints that may impede a data professional at a specific phase of the discovery process. In Figure 7, Level 2, we outline the priorities, value and principle that influence each phase of the data professional's workflow. For instance, the importance of discovering meaningful patterns that drive data-driven decisions through the principles of alignment, accuracy, and impact. One data professional described the impact of these principles are used to help refine and determining a problem:

> *The first thing we will walk through what is your motivation. Why do you want to solve this problem? Then we will ask them what kind of decisions are you trying to make based on the results and what kind of actions are you planning to take based on the results? How much business impact can this give us? Then the next question once we figure out that is what data do we have today? What do we need in order to answer those questions? [P17]*

However, when structure of the organization prevents data professionals from establishing these principles and values, data professionals are not only unable to properly conduct data analysis and

answer the most impactful questions, but are instead demotivated. One data professional in a hierarchical organization described:

> *You always want to be connected to what the outcome and the direction. If you're so far disconnected from kind of the strategic part, you never see the result of what your work is. You need to know what your work is impacting. So aside from the data analysis and getting things right and the context, it's maybe not being as motivated because you're not close [P5]*

It is important to note that the model complements our findings from our previous chapter, by highlighting the sociotechnical constraints that influence each phase of the data professional's workflow.

5.3.4 The relationship between the data professional and the problem space

Through the two CWA models, the data professional WDA (Figure 7) and the problem space WDA (Table 7), we can recognize the relationship and the constraints between the models. The goal of the data professional is to understand and find patterns within the organization using data. As discussed in the previous chapter, data professionals need to domain knowledge to generate insightful conclusions. The problem space AH (Table 7) is an opened and undefined model that data professionals can "fill-out" as they learn about the problem space, also known as domain knowledge.

The domain knowledge that data professionals require can be broken down into 5 different types of domain constraints: functional purpose, abstract function, generalized function, physical function, and physical form. A data professional working in finance described the need to understand different constraints of the domain and organization during the problem understanding and data acquisition phases of the KDP. By understanding the constraints, the data professionals "fill-out" the AH of the problem space.

Fundamentally, it is important for data professionals to understand the purpose of the domain:

> *First, I think you yourself have to understand what that business does [P9]*

After understanding the purpose, data professionals described the need to understand the values and priorities of the business—the abstract function of the domain:

> *What kind of value are they trying to drive for the business? How are they making money? What expenses are they looking at? [P9]*

Next, data professionals explained the need to understand the business processes:

> *Next you have to understand, how that business functions? what is the value proposition of that group? [P9]*

The need to understand business processes could assist with understanding the original source of the data:

> *So we try to find that information and try to understand the structure, the technical side of it. Because there are business processes that lead to that data in the system. So we try to understand how does this come to be, what is this number mean, why is there, and how does it connect to a student. Those sort of really technical questions [from] a database perspective in which we get requests. And then once we answer that, we give them a sample set to say here's a preliminary look at what the data looks. [P5]*

Finally, data professionals must understand the general function and form. Documentation is often used to communicate the general function and form. For instance, a company spent a year creating a "data dictionary" converting accounting terms into technical terms. By understanding the general function and form, data professionals could interpret the data to view the world through the lens of data:

> *We took a year to map out the relationships on this spreadsheet. We converted the technical terms in this Column A into accounting terms in Column C [P18]*

Using CWA, we observe the types of domain knowledge that data professionals need to obtain and the reasons why. The types of domain knowledge that data professionals need to obtain can be categorized into the five levels of the AH. The need for data professionals to understand each of the five different types of domain constraints results in the need for translators. Higher levels of abstraction are often understood by data professionals through stakeholders and conversations, whereas lower levels of abstractions are often conveyed using documentation. As seen in the previous chapter, documentation focuses on conveying the nuances within data.

By understanding the types of domain knowledge that is necessary, we are able to uncover what the "saturation point" in which understanding more about a domain may not help a data professional understand more about a problem and a data set. A data professional consultant explained:

At a certain point, the problem understanding [and] domain understanding is useful for relevance and useful for understanding a set of requirements but after that it doesn't necessarily impact what you're doing with the data. [P7]

By understanding the types of domain knowledge that is required, it may help explain why data analysts spend a significant amount of time learning this information. With the amount of investment needed to learn about a domain, most of our participants agreed that once embedded into a domain, it is unlikely that they would switch domains.

5.4 Phase 2: Control Task Analysis

A ConTA was conducted next to understand *what* specific task needs to be conducted to achieve the functional purpose of finding patterns. The goals of the ConTA are 1) map out how the KDP can be described as a DL, and 2) explore bottlenecks and design opportunities within the DL.

The ConTA was informed by publicly accessible knowledge and case studies of the KDP. Interview data collected, as described in Chapter 3, was the second source used to inform the ConTA. During the interview, each participant was asked to describe two projects. The projects described by the participants were used to help develop the ConTA. Finally, a literature review was performed to evaluate and inform the ConTA.

In Figure 8, we present a DL from the initiation of human activity to the execution of a task as described by Rasmussen [58]. The DL is divided into five regions, representing the stages of the KDP. The following subsections will describe each stage in detail. It is important to note that each stage can further be broken down into individual DLs.

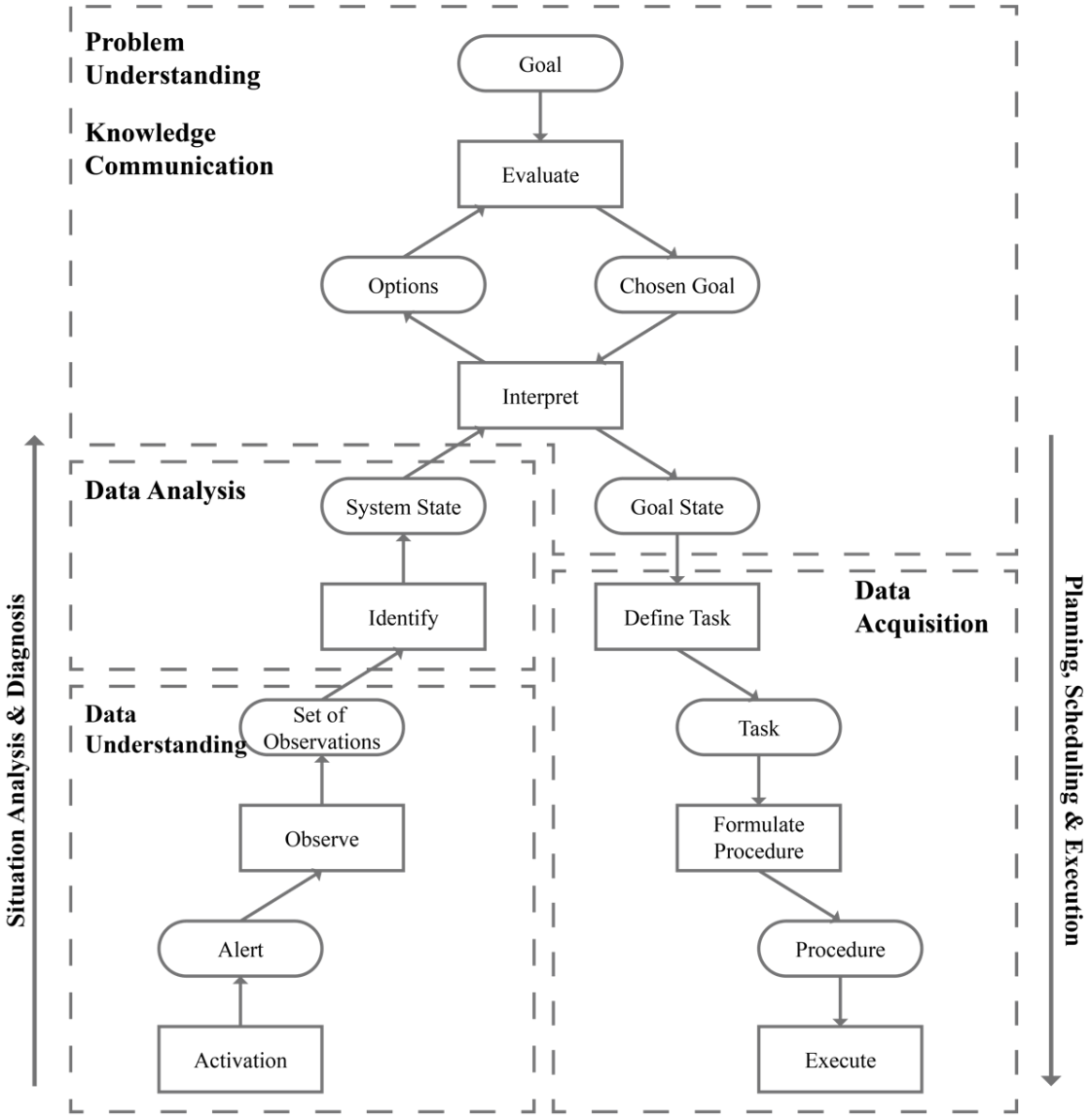

Figure 8: Stages of Knowledge Discovery Process represented on a decision ladder

5.4.1 Goal

The goal of the discovery process is often derived from the problem understanding phase. The goal is collaboratively derived from the data professional and stakeholders.

5.4.2 Interpret, evaluate, and (re-)interpret

Data professionals are required to interpret the problem statement in order narrow down the feasibility and scope of the problem. The human decision-making process is a "complex mental process that requires a high-level of abstraction of the domain knowledge" [58]. This often takes multiple iterations and dialog between the data professional and stakeholder. In this phase, experienced data professionals can effectively communicate and ask targeted questions to evaluate and select an appropriate goal. Experienced data professionals are able to also have a clear view of the available and feasible options. In contrast, if the goal is unique or there is a novice data professional, the data professional must spend longer in this phase to evaluate the possible options and solutions. The output of this phase is the *Goal State*.

5.4.3 Define task, formulate procedure, and execute

The right-hand side of the DL describes the actions [58]. Expert data professionals within a specific domain will be able to define the task and formulate an action plan quickly by understanding the types of data that is required to perform the KDP.

5.4.4 Observe

Once the data is acquired, data professionals are required to understand the acquired data by getting to know the data and validate the data. Often preliminary analysis is conducted to allow for a data professional to understand the data. Novice data professionals that may not understand the terminology may spend longer in this phase to get acquainted with the data through the assistance of documentation or stakeholders.

5.4.5 Identify

Data professional review observed information to discover about the underlying system state [58]. In doing so, data professionals may model the data or apply machine learning techniques on the data in order make sense and establish an understanding of the system state. The output of this model may include parameter and a rationale for their choices.

5.4.6 Interpret, evaluate, and (re-)interpret

In completing the DL, data professionals must interpret the system state and present the options and recommendations to stakeholders. To present the results in a succulent manner, data professionals must translate the findings to the stakeholders. At this point, the data patterns may reveal more questions that need to be answered.

5.4.7 Results of DL analysis

The ConTA demonstrates how the KDP can be mapped onto a DL. This is likely not a surprising mapping, given that a DL is an information processing model and the KDP is also an information processing model.

In many cases, there may be more intricate processing steps that could be explored within each of the steps in the KDP. The interviews held for this work did not go deeply enough to explore these issues. It would be interesting for future work to open these sections of the KDP and look for finer processing steps with the DL. This exercise could be quite helpful in determining where software support, intelligence, or decision-making support could be added to help data professionals.

5.5 Chapter summary

In this chapter, we analyzed our data through an ecological approach by observing how the work domain can impose certain constraints and demands on data professionals. The constraints and demands inherently create challenges for data professionals with which they need to cope. We conducted the first two phases of CWA: WDA and ConTA. The two types of analyses uncovered why domain knowledge is needed and what types of domain knowledge is needed in the KDP. Data professionals need to understand all aspects of the work domain of the organization before identifying the patterns needed to influence the organization.

Chapter 6
Discussion

This thesis investigated the challenges and coping strategies of modern data professionals. The results from our exploratory study illustrate the complexity of the KDP and the importance of sociotechnical skills and domain knowledge throughout the workflow of the modern data professional. There are several implications of these findings for improving the current practice of data professionals. We revolve our discussion on how we can reduce the sociotechnical gap by addressing the challenges that may exist within the KDP. We first discuss the emerging need for a translator. Subsequently, we will discuss the opportunity for tool design as well as the opportunity to educate future data professionals to reduce the sociotechnical gap and support the need of the translator role. We reflect on the ideas of combining domain and data expertise to augment the education of data professionals for helping data professionals succeed in the modern age of data. We will describe the benefits of conducting a CWA and thematic analysis on the same dataset. Finally, we will conclude by discussing the limitations of the study.

6.1 The need for a translator

The results from our study illustrate the need for translations throughout the KDP and the importance of communication. The role of the translator emerges to reduce the sociotechnical gap. As discussed in our findings, there are two types of translators that were observed in our interviews: 1) human as a translator, and 2) artifacts as a translator.

Figure 9 below illustrates the relationship between the environment, data, data professionals, and stakeholders. Both the data and the stakeholders function as the windows of data professionals into understanding the problem space. This results in two types of translations that need to occur: 1) between data and the data professional, and 2) between the stakeholder and data professional. By looking at this relationship, the data professional becomes a natural bridge and translator between the data and an organization. Data professionals must be able to translate business requirement into data requirements, and convert data patterns into actionable insights. If a data professional does not understand the business, they will be unable to identify and convert patterns into insights.

To become an effective translator, data professionals must understand the problem domain. Through the problem space WDA (Table 7), we can observe the 5 different types of domain knowledge that a data professional will need: 1) the functional purpose, 2) abstract function, 3)

generalized function, 4) physical function, and 5) physical form. In the following sections, we will discuss the opportunities of tool design to support these translators and opportunities to augmenting the education of data professionals to focus on developing the skills needed to become a translator.

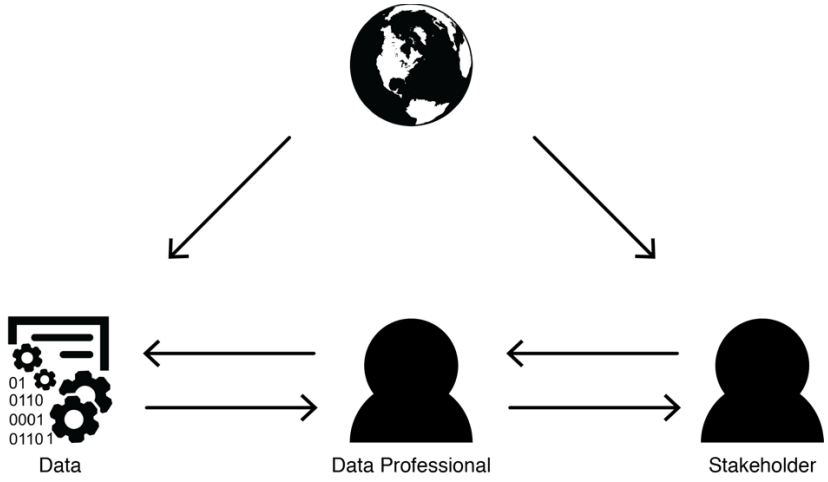

Figure 9: Communication structure

6.2 Design Implications

In this section, we discuss four design implications for future tools to data professionals based on the results found in our thematic analysis (Chapter 4) and CWA (Chapter 5).

6.2.1 Tools to support data professionals to "fill-in" the problem space WDA

Our study illustrates that data professionals must "fill-in" the problem space WDA (Table 7) to effectively navigate the different stages of the KDP. There is a need for tools that may assist data professionals to understand the functional purpose, abstract function, generalized function, physical function, and the physical form of a domain. When data professionals can fully understand the problem space, data professionals are better equipped to tackle two challenges that emerged from the thematic analysis: 1) dealing with lofty goals and 2) being unable to fully validate the data and the results. By understanding the functional purpose, data professionals will be able to help refine the problem space. Moreover, by understanding the functional purpose and physical form, data professionals will be better equipped to validate the data and the results. In particular, there should be a priority to help data professionals understand the higher-level functions (functional purpose) before the lower-level functions (physical form) because higher-level functions are the foundation for all

decisions made throughout the KDP. By understanding the higher-level functions, data professionals can act as better translators.

6.2.2 Developing structured ways to translate through artifacts

We found that data professionals often create artifacts to communicate data details. In our interviews, data professionals often create "data dictionaries" to convert data terminology into lay terms. In addition, artifacts often become stale and require time to create. One way to address this issue is to standardize these translation artifacts so that they can easily be created and understood by both data professionals and stakeholders. Standardized templates may potentially reduce the amount of time needed for data professionals to create these artifacts. Templates may also reduce the friction for data professionals and stakeholders to understand how a specific piece of data may represent the world. There is also a need for these artifacts to be constantly updated.

6.2.3 Capturing data provenance

Aligned with previous researchers [14,31,34], it is important to capture data provenance within a workflow. There is a need for tools to be created to capture data provenance. By capturing provenance, it will help data professional understand the origin each data point within a data set. Understanding the origin of a data point will help data professionals comprehend how a piece of data represents the world as well as increases their trust in the acquired data. This is critical as data professionals must trust their data before generating any insights. This further explains why data professionals need to understand the business processes, or the generalized function, in the problem space WDA (Table 7).

6.3 Augmenting data professional education

While tools can be created to assist with data communication and knowledge communication, augmenting the education of data professionals is the most effective way to teach data professionals to become a translator that understands stakeholders and the data at hand. By working with stakeholders, data professionals will be able to learn how to ask the right questions.

To cope with the lack of domain knowledge, we saw data professionals enroll in domain-specific training programs or embed themselves within the domain. The participants described that the need to understand the intricacies of a domain and the need to be a clear communicator as key attributes of a successful data professional. However, the combination of statistical skills, domain knowledge, and communication skills was rare. One method of approaching this problem is to augment the education

of data professionals by training data professionals to specialize in particular domains during their formal education. A second approach is to train data professionals to ask precise questions in order for them to "fill out" the problem space WDA (Table 7). The more information a data professional is able to "fill out" in the WDA, the more efficient a data professional will be as a translator and throughout the KDP. One method to train future data professionals is through practice and school projects.

Apart from training data professionals to be both data experts and domain experts, we reflect on current data professional training programs. We informally surveyed current training programs in top universities across the United States as well as training programs from the community such as the Microsoft Professional Program. These programs follow a general pattern of training future data professionals in statistics and computing skills. Only a small number of programs touch upon the legal, policy, and ethical implications of data as well as the social aspects of data science.

What is missing currently from the curriculum is emphasis on developing skills for working in specific domains and multidisciplinary teams. Students from these programs may be graduating with an extensive repertoire of data mining and machine learning methods, but they may not have enough exposure to examples from domains and working in a cross-functional team, where they need to translate discovered knowledge to stakeholders. Our participants stressed the importance of domain knowledge and effective communication in data professionals. Hence, training future data professionals with the appropriate communication skills and specific domain knowledge is a necessity, not just an enhancement to their training.

To experience and make effective use of partnerships, educational programs should place students' environments where they work in interdisciplinary teams focused on learning how to ask the right questions, learn about a specific domain, and communicate constraints and discovered knowledge. By learning domain knowledge and how to make effective use of partnerships, data professionals will be armed with the necessary tools to become translators within the discovery process.

6.4 Combining thematic analysis and cognitive work analysis

In this section, we will discuss the benefits of conducting thematic analysis and CWA on the same dataset. We compare these two approaches in terms of their methodological procedure and potential contribution to system design and evaluation. It is important to note that both types of analyses can be considered different methods of knowledge organization.

Fundamentally, the two types of analysis conducted: thematic analysis and CWA, approaches the analysis from two different perspectives. Thematic analysis focused on understanding the user from a *cognitivist* approach. On the other hand, CWA provided an *ecological* approach to the problem. In short, the thematic analysis approached the analysis from a user-first perspective, whereas CWA approached the problem from an environment-first perspective. In addition, thematic analysis focused on a descriptive analytical approach while CWA was formative.

As the approaches differed, the goals of the analyses differed as well. The goal of thematic analysis was to highlight the sociotechnical challenges and coping strategies of the KDP. Comparatively, the goal of CWA was to shed light on the constraints that affect data professionals from achieving their end goal.

Thematic analysis is a flexible method that allows for themes and results to inductively emerge with no pre-existing framework. Through its theoretical freedom, the method can still provide a rich and detailed account of the data [9]. One of the most time-consuming aspects of thematic analysis is the transcription and coding of the data set. The analysis can be used to generate unanticipated insights. One limitation is that thematic analysis has limited interpretative power beyond being descriptive, especially if it is not used within an existing framework that grounds the claims made [9].

The CWA framework shifted the focus from understanding collaborative aspects to understanding the environmental constraints that directed the behaviors of data professionals. CWA is a primarily a framework used to analyze complex sociotechnical systems by understanding the constraints that affect the who, what, when, where, and why of the system and activity.

There are a few studies that have leveraged both analytic techniques, CWA and thematic analysis. In previous studies, thematic analysis was most often used as a tool to identify and summarize the constraints at each level of CWA [24]. In comparison, CWA and thematic analysis are considered as two distinct analytical techniques within this study. Thematic analysis was independently applied first to the data set so the CWA framework would not interfere with the thematic analysis. By doing so, themes not related to environmental constraints could emerge from the data set.

CWA has a strong underlying framework of deductive reasoning. This is a strength, when analyzing domains where deductive reasoning is prevalent, such as data science. As the underlying framework of CWA is based on deductive reasoning, it does not presuppose themes or patterns in the domain. Therefore, similar themes emerged from CWA and thematic analysis. High-level themes such as the KDP and the need for domain knowledge emerged and is represented in both analyses.

The CWA, however, imposes a reasoning structure which uncovers each theme in a systematic manner.

Thematic analysis and CWA are highly complementary and provide valuable insights. Thematic analysis is used to describe specific activities that data professionals conducted. In contrast, CWA provided a systematic understanding of the domain in which the activities are embedded. For example, CWA provided a framework for understanding how the problem space may affect data professionals as seen in Table 7. In contrast, thematic analysis was focused on uncovering specific problems pertaining to each phase of the KDP. This is not surprising as the discovery process acted as the initial codes. Even though CWA may be a more efficient method of searching for workplace constraints, thematic analysis provided a method for a researcher to learn about a new environment without prior judgments.

However, by conducting both sets of analyses, limitations of CWA may be mitigated by the strengths of thematic analysis and vice-versa. Table 8, summarizes the two methods.

Table 8: Comparison between cognitive work analysis and thematic analysis

	CWA	Thematic analysis
Origin	Systems thinking, Ecological Psychology	Psychology
Major Application Domains	Cognitive Engineering	Psychology, sociology
Key Elements	1. Work domain analysis 2. Control task analysis 3. Strategies analysis 4. Social-organizational analysis 5. Worker competencies analysis	1. Familiarization with the data 2. Generating initial codes 3. Searching for themes 4. Reviewing themes 5. Defining and naming themes 6. Producing the report
View	Macro	Micro
Strengths	1. Systems perspective 2. Looking at problems where deductive reasoning is prevalent 3. Equipment with tools to model different organization aspects	1. Flexibility 2. Relativity easy and quick method to learn 3. Accessible to researchers with little to no experience 4. Useful to summarize key features of large body of data 5. Generate unanticipated insights 6. Allows for social and psychological interpretations of data

	CWA	Thematic analysis
Limitations	1. Lack of tools to describe specific, contextualized actions 2. Artificially separating activity from its actor, context, and tool	1. May have limited interpretative power beyond being descriptive, especially if it is not used within an existing framework that grounds the claims made

6.5 Study limitations

We generalize our results with some caution because of some limitations described in this section. When possible, the limitations were mitigated.

The study is limited by the chosen methodological approach. In this study, participants were asked to recall previous projects and describe complex interactions and tasks. Due to the complexity of the tasks, participants' recall of the actions may be inaccurate due to memory bias. Previous research has demonstrated that people remember pieces of an event without being able to remember the details [44]. However, consistency of challenges and coping strategies recalled across participants provides support for the validity of the findings.

Even though collaboration a large aspect of the KDP, the need to uphold confidentiality prevented the researcher from conducting any observations and viewing any conversations that participants described during the interviews. In addition, we were unable to obtain any documentation to analyze the content to observe how data was translated and relayed.

Another limitation of the qualitative interviews is the ability to qualify but not the prevalence of the challenges throughout the discovery process. The goal of the study was to understand *what*, *how*, and *why*, rather than *how much*, *how often*, and *how many*. As such, it was difficult to access the prevalence of the challenges. Also, we used a maximum variation sampling strategy to select a small number of cases that maximum the diversity relevant to the research question. Thus, our sample may not have been representative of the general population.

6.6 Chapter summary

This chapter discussed the findings presented in Chapter 4 and 5. The chapter discussed the implications of our key findings to the design of future tools and educational programs for data professionals. We also discussed the need for a translator throughout the workflow of data professionals. A comparison between thematic analysis and CWA illustrated the benefits of conducting both types of analysis on the same dataset. Finally, the limitations of the study were discussed.

Chapter 7
Conclusion

In today's data-centric world, data professionals are among the most sought-out role. Research has focused on understanding the technical challenges and skills that data professionals need and face throughout their workflows. However, little is known about understanding the sociotechnical challenges and skills that data professionals face. In this thesis, we conducted 20 semi-structured interviews with data professionals to understand the challenges that exist within the KDP. We conducted both a thematic analysis and a CWA on the data set. There are five challenges that emerged from our interviews:

1. Dealing with lofty goals, resulting in frequently changing project requirements that often affect the purpose and scope of the project.
2. Understanding the problem domain
3. Trusting and understanding acquired data
4. Unable to fully validate data and results
5. Understanding *how* and *what* to communicate

We also highlighted how the domain knowledge and organization can affect data professionals. We describe the five different types of domain knowledge that a data professional needs to obtain to seamlessly conduct the KDP. In addition, to address to sociotechnical gap within the KDP, the role of the translator, both humans and artifacts as translators, have emerged. Implications of this research will improve the productivity of data professionals through design recommendations for future tools and will have implications for training the next generation of data professionals.

7.1 Future work

This research suggests several opportunities for further investigation based on the insights gained from this initial exploration study, including future work to address the limitations of this thesis.

Future work may investigate into the prevalence of the challenges and the effectiveness of the coping strategies as discovered throughout this thesis. To measure the prevalence, a survey can be distributed amongst data professionals. This would allow us to understand the prevalence of the translator role.

This research employed semi-structured interviews as its main method of gathering data. As mentioned in the previous chapter, one limitation of this type of interview the reliance of accurate

recall of events. An observational study of data professionals may provide more depth into the minute conversations and challenges that data professionals may face. This would not only provide more accurate data than self-recall, but it may be possible to further study the interactions of data professionals. An observational study may answer what types of questions that data professionals ask, and what length of time data professionals spend in each phase of the discovery process.

As mentioned previously, future work can unpackage each phase of the KDP to help illustrate the finer processing steps within the decision ladder (DL). An expanding model can be helpful to determine where software support or decision-making support systems could be added to assist data professionals.

Our long-term research goal is to bring together different perspectives of various stakeholders throughout the discovery process. With a broader view of the challenges that modern data professionals face, the HCI and data professional communities can be better equipped to tackle the challenges and train the next generations of data professionals.

7.2 Contributions

As mentioned in Chapter 1, this thesis makes the following contributions:

1. Establishes an empirical understanding of the human aspect of the KDP, highlighting the sociotechnical gap in each of the phases.
2. Illustrates the emerging need of the translator within the KDP
3. Understands the role of domain knowledge within the KDP
4. Applies CWA to describe the data profession as a complex sociotechnical system
5. Compares two analytical techniques: thematic analysis and CWA on the same dataset by applying both a constructivist and ecological approach
6. Highlights the design opportunities that exist in building the next generation of tools for data professionals to consider the social and domain-specific aspects of the KDP
7. Illustrates the implications for training the next generation of data professionals so that they can cope with the sociotechnical challenges

Bibliography

1. Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human-Computer Interaction* 15, 2: 179–203.
2. Robert Amar, James Eagan, and John Stasko. 2005. Low-Level Components of Analytic Activity in Information Visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization* (INFOVIS '05), 15–. https://doi.org/10.1109/INFOVIS.2005.24
3. S.S. Anand and A.G. Büchner. 1998. *Decision Support Using Data Mining*. Financial Times Management. Retrieved from https://books.google.com/books?id=ESAdngEACAAJ
4. David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study. *Sociology* 31, 3: 597–606. https://doi.org/10.1177/0038038597031003015
5. Rosaline S Barbour. 2001. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ : British Medical Journal* 322, 7294: 1115–1117.
6. Christine A. Barry, Nicky Britten, Nick Barber, Colin Bradley, and Fiona Stevenson. 1999. Using Reflexivity to Optimize Teamwork in Qualitative Research. *Qualitative Health Research* 9, 1: 26–44. https://doi.org/10.1177/104973299129121677
7. Ann M Bisantz, Natalia Mazaeva, AM Bisantz, and CM Burns. 2009. Work domain analysis using the abstraction hierarchy: Two contrasting cases. *Applications of cognitive work analysis*: 15–47.
8. Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. 2017. How Data Workers Cope with Uncertainty: A Task Characterisation Study.
9. Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2: 77–101.
10. Virginia Braun, Nicola Gavey, and Kathryn McPhillips. 2003. The `Fair Deal'? Unpacking Accounts of Reciprocity in Heterosex. *Sexualities* 6, 2: 237–261. https://doi.org/10.1177/1363460703006002005
11. C. M. Burns, D. J. Bryant, and B. A. Chalmers. 2005. Boundary, Purpose, and Values in Work-Domain Models: Models of Naval Command and Control. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 35, 5: 603–616. https://doi.org/10.1109/TSMCA.2005.851153

12. Catherine M Burns and John Hajdukiewicz. 2004. *Ecological interface design*. CRC Press.

13. Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. 1998. *Discovering Data Mining: From Concept to Implementation*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

14. Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. 2006. VisTrails: Visualization Meets Data Management. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data* (SIGMOD '06), 745–747. https://doi.org/10.1145/1142473.1142574

15. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. 2000. CRISP-DM 1.0 Step-by-step data mining guide.

16. George Chin Jr., Olga A. Kuchar, and Katherine E. Wolf. 2009. Exploring the Analytical Processes of Intelligence Analysts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 11–20. https://doi.org/10.1145/1518701.1518704

17. Renee Wing Yee Chow. 2004. Generalizing ecological interface design to support emergency ambulance dispatching. Department of Mechanical and Industrial Engineering, University of Toronto.

18. Krzysztof J. Cios and Lukasz A. Kurgan. 2005. Trends in Data Mining and Knowledge Discovery. In *Advanced Techniques in Knowledge Discovery and Data Mining*, Nikhil R. Pal and Lakhmi Jain (eds.). Springer London, London, 1–26. https://doi.org/10.1007/1-84628-183-0_1

19. Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, and Lukasz A. Kurgan. 2007. *Data Mining: A Knowledge Discovery Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

20. Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology* 13, 1: 3–21. https://doi.org/10.1007/BF00988593

21. Thomas H. Davenport and D. J. Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. Retrieved March 26, 2017 from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

22. Pedro Domingos. 1999. The Role of Occam's Razor in Knowledge Discovery. *Data Min. Knowl. Discov.* 3, 4: 409–425. https://doi.org/10.1023/A:1009868929893

23. Brent Dykes. 2016. Actionable Insights: The Missing Link Between Data And Business Value. *Forbes*. Retrieved from http://www.forbes.com/sites/brentdykes/2016/04/26/actionable-insights-the-missing-link-between-data-and-business-value/

24. J. A. Effken, B. B. Brewer, M. D. Logue, S. M. Gephart, and J. A. Verran. 2011. Using Cognitive Work Analysis to fit decision support tools to nurse managers' work flow. *Int J Med Inform* 80, 10: 698–707.

25. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 19, 1: 1–16. https://doi.org/10.1109/TKDE.2007.250581

26. Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Advances in Knowledge Discovery and Data Mining. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy (eds.). American Association for Artificial Intelligence, Menlo Park, CA, USA, 1–34. Retrieved from http://dl.acm.org/citation.cfm?id=257938.257942

27. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (KDD'96), 82–88. Retrieved from http://dl.acm.org/citation.cfm?id=3001460.3001477

28. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM* 39, 11: 27–34. https://doi.org/10.1145/240455.240464

29. G. A. Fink, C. L. North, A. Endert, and S. Rose. 2009. Visualizing cyber security: Usable workspaces. In *2009 6th International Workshop on Visualization for Cyber Security*, 45–56. https://doi.org/10.1109/VIZSEC.2009.5375542

30. Danyel Fisher, Rob DeLine, Mary Czerwinski, and Steven Drucker. 2012. Interactions with Big Data Analytics. *interactions* 19, 3: 50–59. https://doi.org/10.1145/2168931.2168943

31. Juliana Freire, David Koop, Emanuele Santos, and Cláudio T. Silva. 2008. Provenance for Computational Tasks: A Survey. *Computing in Science and Engg.* 10, 3: 11–21. https://doi.org/10.1109/MCSE.2008.79

32. Fabien Girardin. 2007. Bridging the Social-technical Gap in Location-aware Computing. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '07), 1653–1656. https://doi.org/10.1145/1240866.1240875

33. B.G. Glaser and A.L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine. Retrieved from https://books.google.co.in/books?id=oUxEAQAAIAAJ

34. D. Gotz and M. X. Zhou. 2008. Characterizing users' visual analytic activity for insight provenance. In *2008 IEEE Symposium on Visual Analytics Science and Technology*, 123–130. https://doi.org/10.1109/VAST.2008.4677365

35. Jonathan Grudin. 1994. Computer-Supported Cooperative Work: History and Focus. *Computer* 27, 5: 19–26. https://doi.org/10.1109/2.291294

36. Harlan Harris, Sean Murphy, and Marck Vaisman. 2013. *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. O'Reilly Media, Inc.

37. Andreas Holzinger, Christof Stocker, Bernhard Ofner, Gottfried Prohaska, Alberto Brabenetz, and Rainer Hofmann-Wellenhof. 2013. Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data: Third International Workshop, HCI-KDD 2013, Held at SouthCHI 2013, Maribor, Slovenia, July 1-3, 2013. Proceedings*, Andreas Holzinger and Gabriella Pasi (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 13–24. https://doi.org/10.1007/978-3-642-39146-0_2

38. Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15, 9: 1277–1288. https://doi.org/10.1177/1049732305276687

39. Petra Isenberg, Anthony Tang, and Sheelagh Carpendale. 2008. An Exploratory Study of Visual Information Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), 1217–1226. https://doi.org/10.1145/1357054.1357245

40. Manyika James, Chui Michael, Brown Brad, Bughin Jacques, D Richard, R Charles, and H Angela. 2011. Big data: the next frontier for innovation, competition, and productivity. *The McKinsey Global Institute*.

41. Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12: 2917–2926.

42. Y. a Kang, C. Gorg, and J. Stasko. 2009. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, 139–146. https://doi.org/10.1109/VAST.2009.5333878

43. Y. a Kang and J. Stasko. 2011. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 21–30. https://doi.org/10.1109/VAST.2011.6102438

44. Elizabeth A. Kensinger. 2009. Remembering the Details: Effects of Emotion. *Emotion review* 1, 2: 99–113. https://doi.org/10.1177/1754073908100432

45. Ioannis Kopanas, Nikolaos M. Avouris, and Sophia Daskalaki. 2002. The Role of Domain Knowledge in a Large Scale Data Mining Project. In *Proceedings of the Second Hellenic Conference on AI: Methods and Applications of Artificial Intelligence* (SETN '02), 288–299. Retrieved from http://dl.acm.org/citation.cfm?id=645861.670288

46. B. c Kwon, B. Fisher, and J. S. Yi. 2011. Visual analytic roadblocks for novice investigators. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 3–11. https://doi.org/10.1109/VAST.2011.6102435

47. Charlotte P. Lee and Drew Paine. 2015. From The Matrix to a Model of Coordinated Action (MoCA): A Conceptual Framework of and for CSCW. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 179–194. https://doi.org/10.1145/2675133.2675161

48. Mark Mason. 2010. Sample Size and Saturation in PhD Studies Using Qualitative Interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 11, 3. Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1428

49. Viktor Mayer-Schnberger. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier*. John Murray Publishers, UK.

50. Sean A Munson, Hasan Cavusoglu, Larry Frisch, and Sidney Fels. 2013. Sociotechnical Challenges and Progress in Using Social Media for Health. *Journal of Medical Internet Research* 15, 10: e226. https://doi.org/10.2196/jmir.2792

51. Neelam Naikar, Robyn Hopcroft, and Anna Moylan. 2005. *Work domain analysis: Theoretical concepts and methodology*. DTIC Document.

52. A. Naweed. 2014. Investigations into the skills of modern and traditional train driving. *Applied Ergonomics* 45, 3: 462–470. https://doi.org/http://dx.doi.org/10.1016/j.apergo.2013.06.006

53. Michael Quinn Patton. 2005. Qualitative Research. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa514

54. M.Q. Patton. 2002. *Qualitative Research & Evaluation Methods*. SAGE Publications. Retrieved from https://books.google.com/books?id=FjBw2oi8El4C

55. W. James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* 27, 3: 258–284. https://doi.org/10.1080/00909889909365539

56. Michael G. Pratt. 2008. Fitting Oval Pegs Into Round Holes. *Organizational Research Methods* 11, 3: 481–509. https://doi.org/10.1177/1094428107303349

57. Priscilla M. Pyett. 2003. Validation of Qualitative Research in the "Real World." *Qualitative Health Research* 13, 8: 1170–1179. https://doi.org/10.1177/1049732303255686

58. J. Rasmussen. 1974. *The human data processor as a system component. Bits and pieces of a model*. Danish Atomic Energy Commission, Roskilde, Denmark.

59. Jens Rasmussen. 1980. The human as a systems component. In *Human interaction with computers*. Academic Press, Incorporated.

60. Jens Rasmussen, Annelise Mark Pejtersen, and L. P. Goodstein. 1994. *Cognitive Systems Engineering*. John Wiley & Sons, Inc., New York, NY, USA.

61. G. J. Read, P. M. Salmon, M. G. Lenne, and N. A. Stanton. 2015. Designing sociotechnical systems with cognitive work analysis: putting theory back into practice. *Ergonomics* 58, 5: 822–851.

62. A. Respício, F. Adam, G. Phillips-Wren, C. Teixeira, and J. Telhada. 2010. *Bridging the Socio-technical Gap in Decision Support Systems: Challenges for the Next Decade*. IOS Press. Retrieved from https://books.google.com/books?id=hFoOL46vcjMC

63. Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (CHI '93), 269–276. https://doi.org/10.1145/169059.169209

64. C. Seale. 1999. *The Quality of Qualitative Research*. SAGE Publications. Retrieved from https://books.google.com/books?id=eZAr3_-gCdQC

65. Michael Sedlmair, Petra Isenberg, Dominikus Baur, and Andreas Butz. 2010. Evaluating Information Visualization in Large Companies: Challenges, Experiences and Recommendations. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization* (BELIV '10), 79–86. https://doi.org/10.1145/2110192.2110204

66. Colin Shearer. 2000. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing* 5, 4.

67. Lisa Singh, Elisa Jayne Bienenstock, and Janet Mann. 2010. Perspectives on Social Network Analysis for Observational Scientific Data. In *Handbook of Social Network Technologies and Applications*, Borko Furht (ed.). Springer US, Boston, MA, 147–168. https://doi.org/10.1007/978-1-4419-7142-5_7

68. Cleidson R. de Souza, Stephen Quirk, Erik Trainer, and David F. Redmiles. 2007. Supporting Collaborative Software Development Through the Visualization of Socio-technical Dependencies. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (GROUP '07), 147–156. https://doi.org/10.1145/1316624.1316646

69. Kim J. Vicente. 1999. *Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-Based Work*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

70. Matthew A. Waller and Stanley E. Fawcett. 2013. Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics* 34, 2: 77–84. https://doi.org/10.1111/jbl.12010

71. Christopher D. Wickens. 1992. *Engineering psychology and human performance, 2nd ed.* HarperCollins Publishers, New York, NY, US.

72. Suk-Chung Yoon, Lawrence J. Henschen, E. K. Park, and Sam Makki. 1999. Using Domain Knowledge in Knowledge Discovery. In *Proceedings of the Eighth International Conference on Information and Knowledge Management* (CIKM '99), 243–250. https://doi.org/10.1145/319950.320008

Appendix A

Recruitment material

The following are the recruitment materials used to obtain the data professionals described in Chapter 3.

RECUIRMENT EMAIL

[DATE]

To [**NAME**]:

My name is Anson Ho a Master's student at the University of Waterloo and this letter is an invitation to consider participating in a study I am conducting on how data professionals and domain experts perform work in complex domains. I would like to provide you with more information about this project and what your involvement would entail if you decide to take part. Please note that participation in this study is voluntary.

The overall goal of this research is to explore solutions to improve productivity of data professionals working in complex domains. To inform the design of these tools, we must first understand how domain complexity affects data professionals. To understand this, we will first be carrying out an interview study with both data professionals and domain experts. We will be investigating (1) the skill sets and training of data professionals, (2) the communication between data professionals and domain experts, and (3) facets of current data analysis tools that may help or hinder this communication.

As part of this research, I am interested in talking to data professionals who have had some experience working in a complex domain, such as health care, finance, bioinformatics, quantum computing, supply chain management, software development, among others. Data professionals include individuals who are manipulating and obtaining insights from data. Data professionals job titles may include but not limited to data scientist, data analyst, and data engineer.

This study will involve an interview session in which we will ask you questions related to your job and your interactions with your team members and/or domain experts. You will receive an Amazon gift certificate for $15 for your participation. (The amount received is taxable. It is your responsibility to report this amount for income tax purposes.)

Participation in this study is voluntary. It will involve an interview of approximately 45-60 minutes in length to take place in a mutually agreed upon suitably private location. You may decline to answer any of the interview questions if you so wish. Further, you may decide to withdraw from this study at any time without any negative consequences by advising the researcher. With your permission, the interview will be audio recorded to facilitate collection of information, and later transcribed for analysis. All information you provide is considered completely confidential. Your name will not appear in any thesis or report resulting from this study, however, with your permission anonymous quotations may be used. Please note that the employer will not know who participated in the project

and that data from each participant will not be provided to the employer. Data collected during this study will be retained for 7 years on a secure server at the University of Waterloo only in a de-identified electronic format. Only researchers associated with this project will have access. There are no known or anticipated risks to you as a participant in this study.

If you have any questions regarding this study, or would like additional information to assist you in reaching a decision about participation, please contact Anson by email **ah3ho@uwaterloo.ca** or my supervisor Parmit Chilana by email **pchilana@uwaterloo.ca** or Professor Catherine Burns by email **c4burns@uwaterloo.ca**.

I would like to assure you that this study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee. However, the final decision about participation is yours. If you have any comments or concerns resulting from your participation in this study, please contact Dr. Maureen Nummelin in the Office of Research Ethics at 1-519-888-4567, Ext. 36005 or maureen.nummelin@uwaterloo.ca.

I look forward to speaking with you and thank you in advance for your assistance in this project.

Yours Sincerely,

Anson Ho
Department of Systems Design Engineering
University of Waterloo
Waterloo, ON N2L 3G1

CONSENT FORM

By signing this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

I have read the information presented in the information letter about a study being conducted Anson Ho of the Department of Systems Design Engineering at the University of Waterloo. I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted.

I am aware that I have the option of allowing my interview to be audio recorded to ensure an accurate recording of my responses (if I feel uncomfortable, I will let the researcher know in advance). I am also aware that excerpts from the interview may be included in a thesis and/or publications to come from this research, with the understanding that the quotations will be anonymous.

I was informed that I may withdraw my consent at any time without penalty by advising the researcher.

This project has been reviewed by, and received ethics clearance through a University of Waterloo Research Ethics Committee. I was informed that if I have any comments or concerns resulting from my participation in this study, I may contact the Director, Office of Research Ethics at 519-888-4567 ext. 36005.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

☐YES ☐NO

I agree to have my interview audio recorded.

☐YES ☐NO

I agree to the use of anonymous quotations in any thesis or publication that comes of this research.

☐YES ☐NO

Participant Name: _____ (Please print)

Participant Signature: _____

Witness Name: _____ (Please print)

Witness Signature: _____

Date: _____

FEEDBACK LETTER

Title of Project: **Understanding and Supporting Data Professionals in Complex Domains**

Student Investigator: **Anson Ho, ah3ho@uwaterloo.ca**

Faculty Supervisor: **Professor Parmit Chilina, pchilina@uwaterloo.ca**

Professor Catherine Burns, catherine.burns@uwaterloo.ca

We appreciate your participation in our study, and thank you for spending the time helping us with our research!

The overall goal of this research is to explore solutions to improve productivity of data professionals working in complex domains. To inform the design of these tools, we must first understand how domain complexity affects data professionals. To understand this, we will first be carrying out an interview study with both data professionals and domain experts. We will be investigating (1) the skill sets and training of data professionals, (2) the communication between data professionals and domain experts, and (3) facets of current data analysis tools that may help or hinder this communication.

All information you provided is considered completely confidential; indeed, your name will not be included or in any other way associated, with the data collected in the study. Furthermore, because the interest of this study is in the average responses of the entire group of participants, you will not be identified individually in any way in any written reports of this research. De-identified electronic data and audio recordings will be kept indefinitely on a secure password-protected server for 3 years, to which only researchers associated with this study have access. All identifying information will be removed from the records prior to storage.

This project has been reviewed by, and received ethics clearance through a University of Waterloo Research Ethics Committee. In the event, you have any comments or concerns resulting from your participation in this study, please contact Dr. Maureen Nummelin, the Director, Office of Research Ethics, at 1-519-888-4567, Ext. 36005 or maureen.nummelin@uwaterloo.ca.

If you think of some other questions regarding this study, please contact Anson by email ah3ho@uwaterloo.ca or my supervisor Parmit Chilana by email pchilana@uwaterloo.ca or Professor Catherine Burns by email c4burns@uwaterloo.ca.

We really appreciate your participation, and hope that this has been an interesting experience for you.

Appendix B

Interview questions

The following are the interview questions used the semi-structured interviews as described in Chapter 3.

Study Instrument: Sample Data Professional Interview Questions

Background Questions (ask via a printed questionnaire):

1. What is your current position?
2. What domain/ industry are you currently working in? (E.g., high-tech, biomedical, finance, etc.)
3. How much experience do you have in data analysis?
4. What relevant formal education/training do you have related to data analysis (if any)?
5. How many projects do you typically work on at a given time?
6. List some of the data analysis software tools and programming languages you currently use.

Main Interview Questions:

1. Describe your typical work day in your current role.
2. Please describe a data analysis project from a domain that required significant expertise which you did not have? [We will refer to this example ('X') in subsequent questions.]
    (a) What was your role?
    (b) How did you complete the project?
3. Describe how your process for analyzing data differs in a complex domain vs. working in a non-expert or familiar domain?
4. What would you say is the most difficult aspect of analyzing data for a complex domain ('X') system? Why?
5. How much time do you approximately spend on learning about the domain for a familiar (non-expert) problem? How does this change working with a complex domain ('X')?
6. How is your team structured?
    (a) Describe your typical communication with other team members. Is this different when working in a complex domain ('X')?
    (b) Describe your typical communication with domain experts. Is this different when working in a complex domain ('X')?
7. In your current role and team set-up, who is typically responsible for generating a data-related research question?
    (a) If you are responsible for generating questions, describe the process in which you generate and answer a research question in a complex domain ('X').
    (b) How is your process different in a complex domain than working in a non-expert domain?

8. Do you think there is any need to change the collaborative aspect of doing data analysis in complex domains vs. non-complex domains? In what way?
9. Have you ever faced any credibility issues due to your level of complex domain knowledge in ('X')? Please describe.
10. Have you tried to enroll in any specialized training related to the complex domain ('X')? If yes, can you describe how this was or was not helpful?
11. What kind of resources do you access to learn about a domain on your own? (i.e., search Google, Wikipedia, books, talk to experts?)
12. Can you share anything else from your experience which would help us better understand your work in complex domains?