DATA SCIENCE CAPABILITY MATURITY MODEL

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

MERT ONURALP GÖKALP

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

NOVEMBER 2021

DATA SCIENCE CAPABILITY MATURITY MODEL

Submitted by **MERT ONURALP GÖKALP** in partial fulfillment of the requirements for the degree of **Philosophy of Doctorate in Information Systems, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Prof. Dr. Sevgi Özkan Yıldırım
Head of Department, **Information Systems** _____

Assoc. Prof. Dr. Altan Koçyiğit
Supervisor, **Information Systems** _____

Examining Committee Members:

Assoc. Prof. Dr. Tuğba Taşkaya Temizel _____
Information Systems, Middle East Technical University

Assoc. Prof. Dr. Altan Koçyiğit _____
Information Systems, Middle East Technical University

Assoc. Prof. Dr. Oumout Chouseinoglou _____
Information Systems and Technologies, Bilkent University

Assoc. Prof. Dr. P. Erhan Eren _____
Information Systems, Middle East Technical University

Assoc. Prof. Dr. Ayça Kolukısa Tarhan _____
Computer Engineering Department, Hacettepe University

Date: 09.11.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name and Surname : Mert Onuralp Gökalp

Signature : _____

ABSTRACT

DATA SCIENCE CAPABILITY MATURITY MODEL

Gökalp, Mert Onuralp
PhD, Department of Information Systems
Supervisor: Assoc. Prof. Dr. Altan Koçyiğit

November 2021, 97 pages

Today, data science presents immense opportunities in attaining competitive advantage, generating business value, and driving revenue streams for organizations. Data science has also significantly changed our understanding of how businesses should operate. In order to survive, it is now indispensable for a contemporary organization to adopt data science as part of its business processes. However, organizations face difficulties in managing their data science endeavors for reaping these potential benefits. This has led to the need for a comprehensive and structured model to continuously assess and improve the maturity of data science capabilities of organizations. This thesis seeks to address this problem by proposing a theoretically grounded Data Science Capability Maturity Model (DSCMM) for organizations to assess their existing strengths and weaknesses, perform a gap analysis, and draw a roadmap for continuous improvements. DSCMM comprises six maturity levels from "Not Performed" to "Innovating" and twenty-eight data science processes categorized under six headings: Organization, Strategy Management, Data Analytics, Data Governance, Technology Management, and Supporting. The applicability and usefulness of DSCMM are validated through a multiple case study conducted in organizations of various sizes, industries, and countries. The case study results indicate that DSCMM is applicable in different settings, is able to reflect the organizations' current data science maturity levels and provide significant insights to improve their data science capabilities.

Keywords: Data Science, Maturity Model, Business Analytics, Big Data, Data-Driven Organization

ÖZ

VERİ BİLİMİ KABİLİYET OLGUNLUK MODELİ

Gökalp, Mert Onuralp
Doktora, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Doç. Dr. Altan Koçyiğit

Kasım 2021, 97 sayfa

Günümüzde veri bilimi, rekabet avantajı elde etme, iş değeri yaratma ve kuruluşlar için gelir akışlarını yönlendirme konusunda önemli fırsatlar sunuyor. Veri bilimi, işletmelerin nasıl çalışması gerektiğine dair anlayışımızı da önemli ölçüde değiştirdi. Çağdaş bir organizasyonun hayatta kalabilmesi için veri bilimini iş süreçlerinin bir parçası olarak benimsemesi artık vazgeçilmezdir. Ancak kuruluşlar, bu potansiyel faydaları elde etmek için yürüttükleri veri bilimi girişimlerini yönetmede zorluklarla karşılaşmaktadır. Bu, kuruluşların veri bilimi yeteneklerinin olgunluğunu sürekli olarak değerlendirmek ve geliştirmek için kapsamlı ve yapılandırılmış bir model ihtiyacı olduğunu göstermektedir. Bu tez, kuruluşların mevcut güçlü ve zayıf yönlerini değerlendirmeleri, fark analizi yapmaları ve ilerlemelerinde sürekli iyileştirmeye yönelik bir yol haritası çizmeleri için teorik olarak temellendirilmiş bir Veri Bilimi Yetenek Olgunluk Modeli (DSCMM) önererek bu sorunu çözmeyi araştırmaktadır. DSCMM, 'Uygulanmıyor' ile 'Yenilikçi' arasında altı olgunluk seviyesinden ve Organizasyon, Strateji Yönetimi, Veri Analitiği, Veri Yönetimi, Teknoloji Yönetimi ve Destek olarak altı grup altında sınıflandırılan yirmi sekiz veri bilimi sürecinden oluşur. DSCMM'nin uygulanabilirliği ve yararlılığı, çeşitli büyüklük, endüstri ve ülkelerdeki kuruluşlarda yürütülen çoklu vaka çalışmalarıyla geçerlenmiştir. Vaka çalışması sonuçları, DSCMM'nin farklı ortamlarda uygulanabilir olduğunu ve kuruluşların mevcut veri bilimi olgunluk düzeylerini yansıtabildiğini ve veri bilimi yeteneklerini geliştirmek için önemli bilgiler sağlayabildiğini göstermektedir.

Anahtar Kelimeler: Veri Bilimi, Olgunluk Modeli, İş Analitiği, Büyük Veri, Veri Odaklı Organizasyon

dedicated to my beloved family

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere appreciation to my supervisor, Assoc. Prof. Dr. Altan Koçyiğit. I am very grateful for his inspiring ideas. He always provides insightful and enlightening discussions about the research. He has given me warm supports, valuable guidance, helpful advice, time, and shown great patience.

I would like to thank Assist. Prof. Dr. P. Erhan Eren for his support, suggestions, comments, and sharing his knowledge through this research.

I would like to thank my wife, Selin, for her invaluable support and endless patience. I am lucky to have her in my life with her limitless love. Thank you for loving me endlessly and believing in me.

Special thanks to my sister Ebru Gökalp for her patience and for sharing her experience and knowledge while preparing this thesis study. Her endless encouragement and support through this thesis and the publications are so valuable and appreciated. This accomplishment would not have been possible without her.

I would like to thank my sister Emel Gökalp for her endless love, support, and advice throughout my life. I want to thank my brother Cenk Gökalp for supporting me all time in my life. I am thankful to my nephew Gökdeniz for making my life more enjoyable during this study.

Finally, I would like to express my very special gratitude to my parents, Saadet and Hidayet Gökalp, for their exceptional patience, love, encouragement, and endless support. Without their love, guidance, and trust in me, I would not accomplish this success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI: Artificial Intelligence
BP: Base Practice
CA: Cronbach Alpha
CL: Capability Level
CMMI: Capability Maturity Model Integration
CRISP-DM: Cross-Industry Standard Process for Data Mining
DMBOK: The Data Management Body of Knowledge
DSCMM: Data Science Capability Maturity Model
F: Fully Achieved
GP: Generic Practice
IEC: International Electrotechnical Commission
IT: Information Technology
IoT: Internet of Things
IS: Information Systems
ISO: International Organization for Standardization
KDD: Knowledge Discovery in Database
KMO: Kaiser-Mayer-Olkin
L: Largely Achieved
MM: Maturity Model
ML: Maturity Level
N: Not Achieved
P: Partially Achieved
PA: Process Attribute
RQ: Research Question
SD: Standard Deviation
SPICE: Software Process Improvement and Capability Determination
WoS: Web of Science

CHAPTER 1

INTRODUCTION

Data science offers opportunities for businesses to attain a competitive advantage by improving their operational efficiencies, supporting decision-making processes, and creating new revenue streams [1]. According to a research[1], 70 percent of senior business executives already indicate that incorporating data analytics is very or extremely important to sustain and improve their companies' competitiveness. However, organizations still fail to adopt data science and big data as part of their business processes despite these technologies' disruptive impact. According to a recent study[2], only four percent of companies with data science and big data initiatives consider themselves as successful, almost 43 percent of organizations attain tangible value from data, and 23 percent attain no value from data. This is essentially due to the fact that organizations' business strategies are not aligned with data science processes. Moreover, data science necessitates a broader focus to tackle organizational, managerial, strategical, and cultural challenges to exploit these promising benefits.

While data science and big data are regarded as groundbreaking technological breakthroughs by both practitioners and scholars, organizations generally experience a lower return on investment than expected in this field. According to Gartner [2], 80% of data science projects are not likely to produce any business value through 2020 because these projects are not managed and scaled with a standardized and systematic

[1] ZS, https://www.zs.com/publications/articles/broken-links-why-analytics-investments-have-yet-to-pay-off.aspx(Last Accessed, 15 April, 2020)

[2] CIO, https://www.cio.com/article/3003538/big data/study-reveals-that-most-companies-are-failing-at-big-data.html (Last Accessed, 15 April, 2020)

approach. The recent studies [3], [4] indicate that utilizing a structured and standard model to initiate data science adoption can increase return on investments. Moreover, establishing a data-driven culture with data science principles and extracting actionable insight from data to solve business problems can be treated systematically by following a framework with reasonably well-defined processes [5]. Thus, the endeavor of leveraging data science requires following systematically developed and standardized methodologies and frameworks with well-defined processes and evolutionary paths [6].

In this chapter, I start with a discussion of the statement of the problem. Then, I detail the research goal and main contributions of this thesis study. The following section describes the research strategy, and I conclude this chapter by outlining the thesis organization.

1.1 Problem Statement

Data science presents promising solutions for organizations in generating valuable insights to attain a competitive edge, advance operational productivity, improve data-driven decision-making management capabilities, and develop novel business models and revenue streams. Organizations still fail to adopt big data as part of their business processes despite its disruptive impact on today's competitive business environment [7], [8], not just because of the significant amount of data generated in a diverse set of sources but also because of the abundance of big data tools and platforms available to deal with the unique characteristics of big data. Hence, they need a standardized model and framework to adopt data science across their business units and manage their data science endeavors. Organizations can potentially reap significant benefits by regarding data and data science as their critical strategic assets to survive in the ever-evolving marketplace and to support decision-making and business operations.

Data science involves a set of processes that comprises technical and managerial skills and also domain knowledge. Thus, it is important to build a unified team that fosters collaboration among a diverse set of skills. Moreover, each team member should have some understanding of the fundamentals of the other's area of responsibility [9]. The success of data science in businesses highly depends on organizations' ability to support data-driven decision-making culture to manage those sets of skills effectively and to perform data science processes ubiquitously across the organization [5].

Implementing a data-driven culture and extracting valuable information from raw data to solve important business-related problems require a framework that is developed by following a standardized, systematic, and impartial methodology and comprises a well-defined set of data science processes [10]. In order to build a data-driven culture in an organization, processes of data science need to be defined, controlled, and measured with a structured method to assess the current level of maturity and to provide a road map for improvement of these processes in the form of a standardized model. To this end, there is a need to develop a standardized framework to assist businesses in establishing a data-driven culture in their organizations and performing data science processes efficiently and effectively.

Maturity Models (MMs) are increasingly applied in the Information Systems (IS) field to improve the software development processes continuously [11] and evaluate their strengths and weaknesses [12]. The MMs are utilized to evaluate and understand how to realize the benefits of relatively new technologies and capabilities in an organizational context. They describe fundamental patterns in the assessment of process capabilities and organizational maturity levels and provide directions for process capability and organizational maturity improvements. MMs can be used descriptively to assess the current process capability levels and the organizational maturity level. Additionally, they also have prescriptive objectives to provide the steps that organizations should undertake to improve their current process capabilities and organizational maturity level.

MMs have widespread adoption in the software engineering domain to appraise and improve process competence, for example, Capability Maturity Model Integration (CMMI) [13] and ISO/IEC 330xx series [14]. They have demonstrated their applicability and usability in software organizations worldwide by providing tangible benefits, including expense savings, improved process quality, predictable and consistent process outputs, and increased employee productivity. The success of this approach in the software engineering domain inspired researchers and scholars in customizing it into emerging domains. Hence, many different models based on Software Process Improvement and Capability Determination (SPICE) [15] are developed for various business domains, including automotive, [16], medical [17], and IT management [18]. Many organizations around the world utilize these MMs because of their proven practical advantages [19].

Despite these observed benefits of employing the MM approach to improve organizations, adopting the MM approach to define and assess data science from a multidisciplinary perspective for improving organizational capabilities in the data-driven organization still poses a critical research gap in practice and academic research. With the widespread adoption of Big Data, Internet of Things (IoT), and Industry 4.0 [20], [21], there is a growing need for a well-established MM to assist organizations in extensively and systematically planning their data science endeavors. There is, as a result, a need for a comprehensive and systematic model to assess organizations' current data science capabilities and evaluate organizational maturities to derive a gap analysis as well as provide an inclusive roadmap for continuously improving gains from data science. Accordingly, this thesis study is motivated by the following research questions;

- *RQ-1:* Do organizations need a maturity model (MM) for the data science domain?
- *RQ-2:* What are the main data science capabilities and processes that can serve as a basis for a standardized MM?
- *RQ-3:* How suitable and applicable is the application of an ISO/IEC 330xx based MM to be used with the purpose of identifying the current state of a data science process capability and organizational maturity?

To address these problems and research questions, the primary purpose of this thesis is to validate the need for a process capability maturity model for the data science domain and then develop a standardized, repeatable, and impartial Data Science Capability Maturity Model (DSCMM) to guide organizations in assessing their data science processes, evaluating their organizational data science maturities, revealing their strengths and limitations, performing a gap analysis, and providing an extensive roadmap for continuous improvement.

1.2 Significance of the Study

DSCMM represents the organizations' ability to adopt, integrate, govern, manage, and leverage diverse set of data sources to improve their data-driven decision-making capabilities. It also signifies developing a state-of-the-art digital transformation ecosystem, extracting insightful knowledge to generate a business value gain a competitive business advantage. Hence, DSCMM needs to provide a standardized, repeatable, and impartial assessment framework to evaluate an organization's current

data science capability levels as well as its organizational maturity level and provide a roadmap to assist the organization in moving its data science maturity to a higher level. The primary goal of DSCMM is to specify a structured method that can be reliably and efficiently adapted and repeated by different organizations to assess which capabilities need to be improved and to provide a generic roadmap to form a data-driven decision-making culture that allows evaluating and strengthening their businesses.

In this thesis, I assess the adequacy of the existing data science MMs, identify the strengths and weaknesses of these existing models and develop a structured, standardized, impartial, and repeatable MM for the data science domain. Towards these goals, the relevant academic and gray literature have been reviewed, and an exploratory survey has been employed to collect data from 183 practitioners, according to Hevner's information system research framework [22], then these industry data have been combined with scholarly data to reveal research gaps. DSCMM has been developed on the basis of the well-accepted ISO/IEC 330xx standard series [14] by following a theoretically grounded MM development methodology [23]. DSCMM defines six capability levels (CLs) and maturity levels (MLs) and covers twenty-eight data science processes in six process areas: Organization, Strategy Management, Data Analytics, Data Governance, Technology Management, and Supporting. The applicability and usefulness of DSCMM have been validated by conducting a multiple case study in three different organizations of various sizes. These companies operate in different industries in different countries, and they exhibit different data science adoption levels.

1.3 Research Strategy

In this thesis, I followed a systematic MM development methodology proposed by Becker et al. [23]. This research framework includes four main phases, as delineated in Figure 1-1.

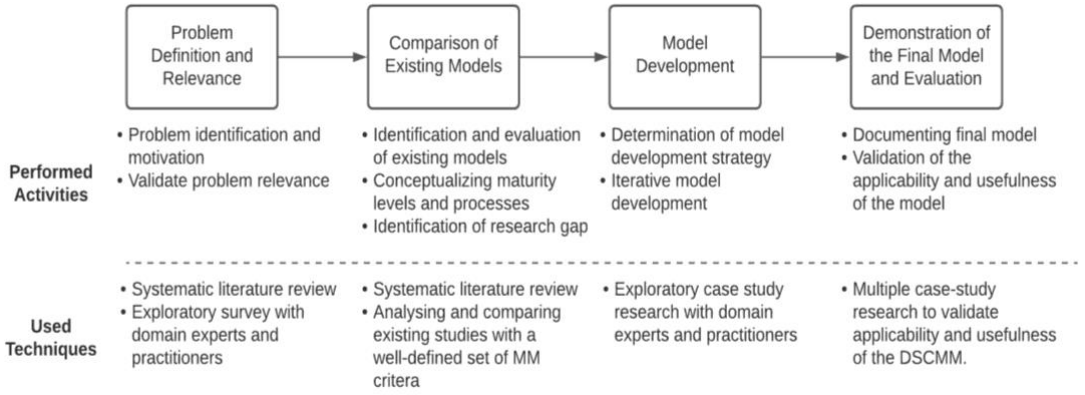

Figure 1-1. Research Methodology

The problem definition and relevance phase investigates the current state of the existing literature to validate the problem. In this phase, the literature has been reviewed (as presented in Chapter 2) to identify the research gap and validate the industrial need [24], [25]. According to the findings, the academic literature has only a handful of studies that are mostly at the development phase, and their empirical evaluations and validations are scarce. However, as organizations face a fierce competition with the widespread adoption of Web 2.0, Internet of Things (IoT) and, Industry 4.0, the need for an applicable and usable data science MM becomes prominent. In this phase, it is also important to validate the identified requirements and objectives relative to the business domain. Thus, the fundamental data science processes have been identified by analyzing existing MMs and relevant data science, big data, and data analytics studies in the literature. Then, an exploratory survey has been conducted to validate the problem, objectives, requirements as well as findings extracted from the literature and to understand the importance of these existing data science processes for practitioners.

Comparison of existing models phase includes an in-depth analysis of existing models, which contributes to collecting problem and domain requirements in the conceptualization of the MLs and processes. In this phase, a structured literature review has been conducted to identify prominent models and compare them objectively in terms of well-defined criteria [25], [26].

Model Development phase includes determining the model development strategy, understanding MM requirements, processes, and measurement attributes steps. These

steps require an iterative development approach to improve the generalizability and applicability of the final model continuously. In this phase, first the literature has been analyzed to reveal key data science processes and maturity indicators. Then, data have been collected from 183 expert practitioners in the data science domain by employing a survey-based research methodology, and these collected data were iteratively analyzed by a group of experts to determine core processes and process areas to reach a final consensus model [24], [25]. Then, the model development strategy and the applicability of the proposed approach have been validated by an exploratory case study, and improvement opportunities have been discovered. Accordingly, I developed the final model, DSCMM. According to the template provided by the study of Mettler [27], the decisions made when developing DSCMM are highlighted with gray boxes in Table 1-1. DSCMM needs to include both management-oriented and technology-oriented constructs and needs to focus on both management and technology-oriented processes. Moreover, I covered both theory-driven approaches from the literature and practitioners' feedbacks in the design and development process of DSCMM.

Table 1-1. Decisions made when developing DSCMM

Decision Parameter	**Characteristic**			
Novelty	Emerging	Pacing	Disruptive	Mature
Innovation	New	Variant	Version	
Breadth	General Issue		Specific Issue	
Depth	Individual/Group	Organizational	Inter-Organizational	Global/Society
Audience	Management Oriented	Technology Oriented	Both	
Maturity concept	Press-focused	Object-focused	People-focused	Combination
Goal function	One-dimensional		Multi-dimensional	
Design process	Theory Driven	Practitioner Based	Combination	
Design product	Textual description of form	Textual description of form and functioning	Instantiation (software)	Combination
Application method	Self-assessment	Third-party assisted	Certified professionals	
Respondents	Management	Staff	Business partners	Combination

Demonstration of the Final Model and Validation phase is devoted to documenting and sharing the final version of the model and evaluation of the final model. DSCMM

is based on the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 330xx family of standards [14], and it consists of two core dimensions: capability and process. This phase comprises the conception of transfer and evaluation, implementation of transfer media, and evaluation of the final model. In this phase, a multiple case study is employed to validate the applicability and usefulness of DSCMM, as detailed in Chapter 5. The multiple case study is conducted in three organizations with various data science adoption levels and from various countries and industries [26], [28], [29].

1.4 Thesis Structure

This thesis comprises six main chapters. After this introductory chapter, Chapter 2 gives an overview of MMs, data science, big data and investigates the related literature. In Chapter 3, I present the development stages of DSCMM and describe the application of the preliminary version of DSCMM. Chapter 4 introduces the final structure of the proposed DSCMM. Chapter 5 describes the application of the final version of the proposed DSCMM via case studies. Finally, the concluding remarks and directions for future researches are given in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

In this chapter, I give background information in the scope of this thesis study. Section 2.1. provides an overview of MMs (MMs). In Section 2.2., I briefly describe the structure of the ISO/IEC 330xx family of standards. I explain the data science and big data concepts in Section 2.3. Finally, I review existing studies related to DSCMM in Section 2.4.

2.1 Maturity Models (MMs)

MMs describe the fundamental practices, patterns and accordingly, maturity stages in assessing the quality of domain processes and provides suggestions and a roadmap for continuous improvement concerning a specific problem domain [13]. MMs are successfully exploited to evaluate and understand how to realize the promising value of emerging technologies, domains, and capabilities [30]. Moreover, MMs aim to describe organizational capability development patterns and indicate directions for process capability improvement. That is, MMs have both descriptive and prescriptive objectives. They can be used to assess the current capability level of each process and the current organizational maturity in relation to a certain technology or domain. MMs also can prescribe the actions that the organization should undertake to move their organizational maturity to the next level.

The MM approach can support organizations in managing complex and emerging capabilities and technological transformation [30]. MMs clearly define a sequence of discrete maturity stages and domain-related processes to identify the organizations' current process capabilities, as well as benchmarking, and determine the critical problems which may hinder the adoption of these processes [31]. Moreover, they provide an extensive and standardized roadmap for an organization to establish a

domain culture and management excellence throughout the organization [23]. The MM approach has been initially developed for the software engineering domain, and it has demonstrated noteworthy benefits in software development processes. Capability Maturity Model Integration (CMMI) [13] and Software Process Improvement and Capability Determination Model (SPICE), which is superseded by the ISO/IEC 330xx standards [14], are the prominent state-of-the-art MMs in the literature.

In the literature, there are also some studies [32], [33] that criticize the reliability and robustness of the MMs. These studies primarily argue that MMs oversimplify reality, and their staged hypothesis lacks an empirical foundation. Nevertheless, the benefits and positive impacts, such as expense savings, improved process quality, predictable/consistent process outputs, and increased employee productivity, of utilizing MMs have been demonstrated through several case studies [19], [34]. Accordingly, MMs have been applied by a multitude of software organizations worldwide due to such observed benefits of using them. This success of the MM approach in the software engineering domain attracted too much research interest and motivated the application of the approach to emerging domains [35]. Although there are numerous studies proposing MMs for various domains, the MM approach in the data science domain still has not gained widespread attention in research.

2.2 Software Process Improvement and Capability Determination Model

The ISO/IEC 330xx family of standards provides a structured methodology for determining and improving software processes. The ISO/IEC 330xx family of standards provides a two-dimensional process capability determination framework. The first dimension, also known as the process dimension, includes process definitions, outcomes, scopes, and base practices (BP) of these processes. The second dimension, also known as the capability dimension, includes Capability Levels (CLs), process attributes (PA), and rating methodology to determine these CLs. The structure of the ISO/IEC 330xx is depicted in Figure 2-1.

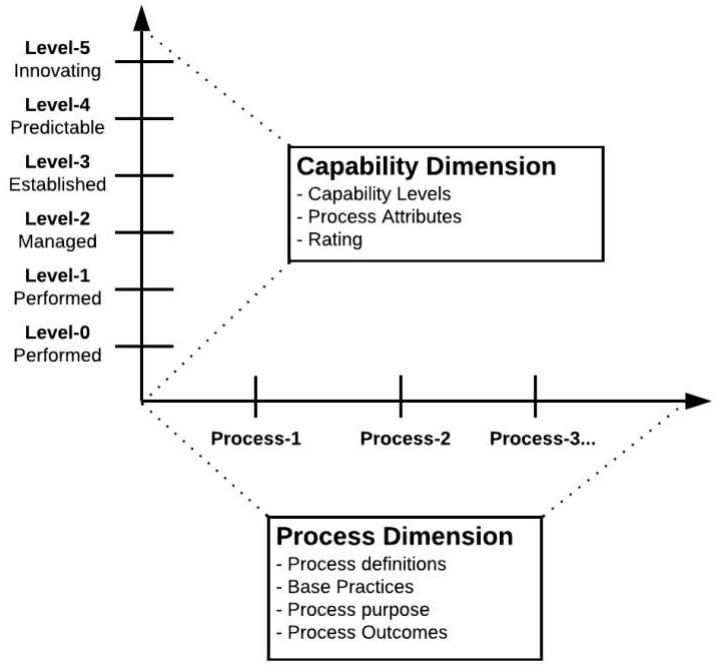

Figure 2-1. The Structure of ISO/IEC 330xx Family of Standards

The process CL determination model provided by ISO/IEC 330xx defines measurable PAs as requirements and rules to determine CL of any process. In other words, the PAs are the features of a process that can be evaluated on a scale of achievement to provide a measure of the capability of the process. Moreover, the PAs are applicable to all processes. A CL represents a well-defined set of PAs that significantly improve the process capability. The defined PAs in each level address specific needs of the CLs, and they progress through the improvement of the capability of any process. Each process is required to be at least a “Largely Achieved (L)” for the corresponding PAs for each CL and “Fully Achieved (F)” for any lower CLs PAs. The ISO/IEC 33020 Standard [36] defines six CLs from Level 0 Incomplete to Level 5 Innovating, and these CLs include nine PAs in total. The CLs and assigned PAs are defined in Table 2-1.

Table 2-1. ISO/IEC 33020 Capability Levels and Process Attributes

Capability Level (CL)	**Process Attributes (PAs)**	**Rating**
CL – 0 Incomplete	Not Applicable	Not Applicable
CL – 1 Performed	**PA-1.1: Process Performance**	**L or F**
CL – 2 Managed	PA-1.1: Process Performance	F
	PA-2.1: Work product management	**L or F**
	PA-2.2: Performance management	**L or F**
CL – 3 Established	PA-1.1: Process Performance	F
	PA-2.1: Work product management	F
	PA-2.2: Performance management	F
	PA-3.1: Process Definition	**L or F**
	PA-3.2: Process Deployment	**L or F**
CL – 4 Predictable	PA-1.1: Process Performance	F
	PA-2.1: Work product management	F
	PA-2.2: Performance management	F
	PA-3.1: Process Definition	F
	PA-3.2: Process Deployment	F
	PA-4.1: Process Control	**L or F**
	PA-4.2: Process Analysis	**L or F**
CL – 5 Innovating	PA-1.1: Process Performance	F
	PA-2.2: Work product management	F
	PA-2.3: Performance management	F
	PA-3.1: Process Definition	F
	PA-3.2: Process Deployment	F
	PA-4.1: Process Control	F
	PA-4.2: Process Analysis	F
	PA-5.1: Process Innovation	**L or F**
	PA-5.2: Process Self Optimization	**L or F**

2.3 Big Data and Data Science

Data science and big data provide valuable business insights and value to support strategic decision-making. The terms data science and big data are often used interchangeably. Even there are some substantial overlaps between data science and big data, these two concepts do not refer to the same thing. On the one hand, big data refers to the collection of large, complex, unstructured, and continuous data gathered from a large number of usually disparate data sources [37]. It focuses on capturing, storing, and processing data on highly scalable distributed architectures to efficiently

and effectively deal with big data's unique characteristics such as volume, variety, and velocity. Over time, an extensive literature has been developed to address the unique characteristics of such large collections of data and highly scalable and distributed infrastructures, and technologies have been proposed to gather, store and process big data in an efficient and effective manner. On the other hand, data science embraces the whole lifecycle, including business understanding, data collection, data preparation, data analytics, and deployment, to generate business value from raw data.

Data-driven organizations possess a culture of leveraging data-driven decision-making rather than depending on the intuitions of their managers in business activities. Data-driven organizations make each strategic decision based on the interpretation of data and analytics by utilizing the data science approaches. Thus, organizations are increasingly utilizing data science principles, algorithms, and methodologies to develop data analytics applications. Data collected from different operational stages can improve an organization's performance and create new business opportunities. Accordingly, most organizations primarily invest in data and technology capabilities to integrate data science into their daily operations. Data is essential to assess and improve business processes. A study by Manyika et al. [38] reveals that companies characterizing themselves as data-driven have better performance on financial and operational success measures. Exploiting big data and data science technologies provides promising solutions to enhance corporate, but organizations first need to change their decision-making culture to harness these potential benefits. Organizations need to be aware that being data-driven is not about leveraging big data to understand the past with business intelligence; it is a matter of predicting and shaping the organizations' future by embracing recent big data and data science approaches to reveal and tackle the business problems even before they arise. Today, organizations are trying to boost their data science capabilities to exploit big data and not to lag behind the hype. LinkedIn [39] reveals that the number of available data scientist jobs has increased by 650% since 2012. Moreover, it is expected that organizations that fail to adopt data science technologies and place data-driven culture in their organizations will face a competitive edge. Besides, a recent study by McKinsey [40] predicts that the potential impact of data science and artificial intelligence (AI) on the global economy will be around 13 trillion U.S. dollars by 2030, and this study also forecasts that non-adopters of data science and AI will experience a 20% decrease in their cash flow by 2030. To this end, businesses should place the data science and data-driven management culture as their core business assets to grasp the promising potential.

2.4 Related Work

In the literature, as far as our literature survey has shown, there is no study that comprehensively defines an MM for the data science domain. However, there are some studies in the academic and gray literature that develop an MM for the big data or data analytics domains.

In order to investigate existing academic literature systematically, I searched articles, proceedings, and book chapters in Web of Science (WoS), and Scopus databases with *("Data science" OR "big data" OR "data-driven") AND ("maturity" OR "capability" OR "roadmap")* search terms in abstracts, keywords, and titles. the searches were not limited to any year criteria. The search results cover studies up to December 2020. The research results have been documented in a spreadsheet to compare and merge duplicate studies. Initially, 458 journals and proceeding articles have been retrieved in total: 123 from WoS and 335 from Scopus databases. There were 111 duplicates that appeared in both WoS and Scopus databases. As a result, 347 studies are determined as potentially related studies.

These 347 studies were reviewed and evaluated according to an inclusion criterion: the study proposes a model or framework for the scope of the assessment, evaluation, maturity, or capability to leverage data science. According to our reviews of these relevant papers, there is no study that comprehensively defines an MM for the data science domain. After the review process, I identified eleven studies relevant to this thesis. Besides, I also traced the reference lists and citations of these studies as a complement to searching the databases and to discover related academic and gray literature extensively. Other studies that are not found in the systematic literature review and also studies that are not available in the reference lists and citations of the existing studies are not included in this study. By this means, I discovered five additional studies in the academic literature and six different studies in the gray literature. As a result, twenty-two existing MMs, listed in Table 2-3, were identified and analyzed.

MMs and their assessment methodology should be complete, comparable, unambiguous, consistent, and objective for widespread acceptance in practice or research [31]. Thus, I analyzed the existing studies based on a set of criteria adopted from well-known studies in the MM domain [31], [41], [42]. According to these

studies, I have set the assessment criteria outlined in Table 2-2 to evaluate the existing studies. Thus, the strengths and weaknesses of related works are investigated in a systematic way.

Table 2-2. Assessment Criteria for Gap Analysis

	Criteria	**Definitions**
C1	Fitness for Purpose	Evaluates the scope of MM. Specifically, it determines to what extent the data science domain is covered by the MM.
C2	Completeness of Processes	Assesses the completeness of processes and their definitions in comprising all or a subset of major data science processes.
C3	Granularity of Dimensions	The level of detail of explanations of the attributes in the corresponding dimensions.
C4	Definition of Measurement Attributes	Evaluates whether the model provides a detailed description of the measurement attributes or not.
C5	Description of Assessment Method	It questions if the study provides a complete description of the assessment method.
C6	Objectivity of the assessment method	The level of objectivity of maturity assessment method of the study, including definitions of the attributes and practices.

The analysis of existing MMs according to the defined rating scale is given in Table 2-3. In the evaluations, the achievement degree for each criterion was determined according to the rating scale percentage values defined in ISO/IEC 33002. In this scheme, “N” symbolizes the achievement degree from 0% to 15%; “P” means that the achievement is between 16% to 50%; “L” corresponds to the degree of achievement from 51% to 85%; “F” means that the degree of achievement is between 86% and 100%.

Table 2-3. Analysis of Existing MMs in the Context of Data Science

MMs	C1	C2	C3	C4	C5	C6
MM1 [43]	P	P	N	N	N	N
MM2 [44]	F	L	L	L	P	P
MM3 [45]	P	N	N	N	N	N
MM4 [46]	L	P	P	N	N	N
MM5 [47]	L	P	P	N	N	N
MM6 [48]	L	P	P	N	N	N
MM7 [49]	L	P	P	N	N	N
MM8 [50]	L	P	P	N	N	N
MM9 [51]	L	L	P	P	N	N
MM10 [52]	F	L	L	N	N	N
MM11 [53]	L	P	P	N	P	P

MM12 [54]	P	P	P	P	P	P
MM13 [55]	F	L	P	P	P	P
MM14 [56]	L	N	N	N	N	N
MM15 [57]	F	L	P	P	N	N
MM16 [58]	P	P	N	N	N	N
MM17 [59]	P	L	L	N	N	N
MM18 [60]	P	P	P	N	N	N
MM19 [61]	L	L	P	P	N	N
MM20 [62]	P	P	N	N	N	N
MM21 [63]	L	L	N	P	P	N
MM22 [64]	L	L	P	P	N	N

A big data MM is proposed in MM1. However, this study does not provide a comprehensive MM for the data science domain, it only investigates the technological readiness of an organization. However, organizational and environmental processes should also be considered for assessing the data science maturity of the organizations. MM2 integrates existing industry-developed MMs into one single coherent big data MM. It considers five processes and 6-main maturity levels from Level-0 to Level-5. However, the proposed MM does not provide a structured method to assess the maturity of the businesses. While a big data MM to gauge the readiness of zakat institutions to embark into the big data evolution is proposed in MM3, the scope of this MM is limited to a specific institution. Therefore, the proposed model does not entirely embrace the data science domain. MM4, MM5, and MM6 propose business intelligence MMs. Similarly, MM8 and MM7 provide data analytics maturity and capability indicators. However, MM7 and MM8 do not develop a complete assessment model to define how to objectively evaluate the data analytics maturity of an organization. Moreover, they do not develop their models based on an accepted framework, so they need to be validated across different organizations and industries. The most recent study, MM9, proposes an MM for big data analytics in airline network planning. MM10 investigates analytics maturity indicators. Nevertheless, MM9 and MM10 do not give an objective assessment approach, nor do they provide any detail about measurement attributes and assessment methods. M11 focuses on investigating big data management in intelligent manufacturing enterprises. However, this study does not provide a comprehensive MM as it does not address any organizational, strategic, and cultural processes. Besides, the description of measurement attributes and methods remains unclear. M12 proposes a lean MM for operational-level planning. M14 addresses assessing capability maturity levels of organizations in national data ecosystems. M16 develops an MM to support digitalization in the

manufacturing domain. The concepts provided in studies covered so far are related to our study's context, but they do not entirely cover the data science domain. Moreover, these researches essentially lack structured process definitions, fine-grained detail regarding dimensions, and descriptions of assessment attributes and methods. M13 investigates the processes and maturity levels for the data-driven manufacturing domain. Similarly, M15 presents an MM to assess and improve industrial analytics capabilities. Although the scopes of these studies coincide with the scope of our proposed approach, they generally fall short of proposing a theoretically grounded model to define processes and measurement attributes structurally. Moreover, the description and objectivity of the assessment methods remain unclear.

Besides the academic literature, practitioners have proposed MMs to address data science capabilities, including MM17, MM18, MM19, MM20, MM21, and MM22. Moreover, there are also some other studies [65]–[68] in the gray literature that are not found in the reference list and citations of the resulting academic studies from the systematic literature review. However, these studies mainly lack details of the model development process, assessment methodology, and validation. These shortcomings also hinder their widespread acceptance and adoption in academic literature and practice. Moreover, their theoretical and methodological foundations remain unclear as they do not detail the development approaches employed. More importantly, they do not validate their models. Besides, these models lack an unbiased academic view, as they are proposed by a consulting company or a technology vendor.

To sum up, the literature review shows a limited number of studies that focus on developing an MM for the data science domain. However, as detailed in Table 2-3, none of them fully satisfies the requirements of a well-defined MM. Moreover, none of these studies aims to improve the data science processes, and none of them is developed based on a well-established and standardized process capability MM. The existing studies mainly lack grounding their models on a theoretical development approach to provide complete, comparable, unambiguous, consistent, well-documented, and objective standard models. They do not provide any empirical results validating the applicability and usefulness of their proposed models, assessment attributes, and methods. Moreover, none of these studies provides an extensive roadmap to guide organizations for continuous improvement. Accordingly, none of these existing studies has gained widespread acceptance in practice or research.

Hence, there is a need to develop a complete, clear, unambiguous, objective, consistent, repeatable, and comparable MM as defined in [69] for the data science domain to address these critical research gaps in the literature. For this reason, this thesis study aims to fill this research gap by developing DSCMM based on the well-accepted family of standards, ISO/IEC 330xx.

CHAPTER 3

MODEL DEVELOPMENT

This chapter presents the development stages of DSCMM and the application of the preliminary version of DSCMM via an exploratory case study. In the development of DSCMM, an exploratory survey has been carried out to validate the need for a process capability/maturity model for the data science domain, investigate social and technical drivers of data science practices and define the main constructs of DSCMM by integrating literature findings and practitioners' considerations [25]. Moreover, an exploratory case study is conducted to assess the applicability and usefulness of adopting the ISO/IEC 330xx standard family in developing a capability maturity model for the data science domain [24].

In Section 3.1, I detail the exploratory survey conducted to develop the process dimension, present and discuss the survey results. I also analyze potential validity threats that may arise in conducting and analyzing the results of the survey. In Section 3.2., I detail the exploratory case study that was conducted with the preliminary version of DSCMM, present assessment results, and determine potential validity threats.

3.1 Development of Process Dimension

The development of the process dimension of DSCMM is methodologically grounded in design science research, which provides a solution-oriented approach in generating knowledge for the problem domain. This approach offers a suitable foundation for this study because the development of data science maturity is a relevant domain problem that requires new capabilities to be achieved by the organization to stay competitive.

In this study, both qualitative and quantitative data have been collected from the field to determine the need for DSCMM and its main building blocks in its natural settings

by conducting an exploratory survey. This section discusses the validity threats of the survey, data collection methods, and results.

3.1.1 Survey Development

The main rationale behind the exploratory survey is to answer two main research questions given below and to validate the findings in the literature about defined data science processes by combining scholarly data and real industry data, as indicated in Hevner's information system research framework [22]. The research questions for the exploratory survey are as follows:

- *RQ-1:* Do organizations need an MM for the data science domain?
- *RQ-2:* What are the main data science capabilities and processes that can serve as a baseline for an MM?

The survey-based research approach is a well-accepted method to understand the maturity constructs of an organization's data science capabilities [70]. The survey consisted of two main parts. The first part included questions to analyze the demographic and organizational information of respondents. The second part included questions to assess the importance of data science processes and capabilities. In this part, I defined data science processes and their corresponding explanations according to the literature review detailed in Section 2.4. These processes and their corresponding references are summarized in Table 3-1 to support the validity of each process.

The importance of each process was measured with a 5-point Likert scale ranging from "Not at all Important" (1) to "Extremely Important" (5), where higher values represent greater importance perceived by the respondents. The last part of the survey included both quantitative and qualitative questions to understand the importance of DSCMM, utilized metrics, and specialized project management techniques for data science projects. Moreover, feedbacks have been collected from the participants about the comprehensibility of process definitions and their suggestions for processes which are not included in the survey questions. However, none of the respondents requested any revision or extension in process definitions.

Table 3-1. Processes and their corresponding references

Process Area	Processes	References
Organization	Organizational Governance	[43], [44], [64], [45], [51], [52], [59]–[63]
	People Management	[43], [44], [52], [59]–[64]
	Project Management	[52], [61], [64]
	Communication	[52], [60], [61], [63], [64]
	Quality Assurance	[52], [60], [61], [64]
	Knowledge Management	[60], [61]
	Risk Management	[61], [63]
	Performance Management	[64]
Strategy Management	Vision & Strategy	[43]–[45], [50], [59]–[64]
	Sponsorship & Funding	[44], [59]–[61], [64]
	Strategic Alignment	[44], [51]
	Stakeholder Engagement	[44], [61]
	Innovation Management	[61], [64]
Data Analytics	Business Understanding	[60], [61], [63]
	Data Collection	[43], [44], [51], [59]–[64]
	Data Storage	[44], [45], [60]–[64]
	Data Preparation	[60], [61]
	Data Analysis	[44], [45], [59]–[64]
	Data Understanding	[60], [61], [63], [64]
	Model Deployment	[50], [61], [63]
	Data Visualization	[60], [61], [63]
Data Governance	Security and Privacy Management	[43]–[45], [60]–[64]
	Meta-data Management	[61], [63], [64]
	Data Quality Management	[44], [59], [60], [63], [64]
	Data Availability and Access	[43], [59]–[61], [63]
	Data Life-cycle Management	[51], [60]
	Architectural Approach & Standards	[59]–[61]
	Master Data Management	[59]–[61], [64]
	Data Integration	[59], [60], [64]
	Historical Data Management	[59], [61], [63]
Technology Management	Hardware Management	[44], [50], [59], [61]–[64]
	Software Management	[44], [45], [51], [59]–[61], [63], [64]
	Configuration Management	[59], [61], [63], [64]
	Deployment	[50], [61], [63], [64]

3.1.2 Data Collection and Results

The data collection took approximately five months, from March 2019 to July 2019. In total, 191 responses were collected by contacting relevant respondents in the data

science domain via face-to-face meetings and e-mail. Eight of these respondents did not give their consent to publish their answers publicly for scientific purposes, so the answers of these respondents have been excluded from the dataset. As a result, there are answers of 183 respondents in the final dataset. To ensure external validity, the survey covered a variety of industries, organizational sizes, and respondents' profiles, which are summarized in Table 3-2 and Table 3-3. The respondents were generally data scientists having more than three years of experience in the data science domain and having prior knowledge about the defined concepts.

Table 3-2. Descriptive Statistics of Respondents' Organizations

Factors	**Sample (N=183)**	**Percentage (%)**
Industry		
Software and Internet	41	22.5
Manufacturing	25	13.7
Finance & Insurance	17	9.3
Health	16	8.7
Telecommunication	14	7.6
Retail	13	7.1
Media	13	7.1
Research & Development	10	5.5
Defense Industry	9	4.9
Education & University	8	4.4
Government	5	2.7
Others (Transportation, EdTech, Energy)	12	6.6
Firm Size		
1-10	26	14.2
11-50	28	15.3
51-100	21	11.5
101-500	22	12.0
501-1000	35	19.1
1000+	51	27.9

The overall survey results are given in Appendix - A, including mean, standard deviation, percentiles, and intercorrelations among defined 34 processes. According to the results, data analysis, data collection, business understanding, data quality management, data understanding, data preparation, data integration, security & privacy, sponsorship & funding, and vision & strategy processes are considered to be the top ten critical factors for data science success in organizations. It has been observed that organizational considerations, strategy management, and technology management aspects gain higher importance as the firm size increases. Data

governance processes also gain more importance as firm size exceeds 51 employees. Similar to these findings, the study of [52] reports data quality and integration are the most common challenges for the adoption of big data.

Table 3-3. Descriptive Statistics of Respondents Background

Factors	**Sample (N=183)**	**Percentage (%)**
Experience in Data Science		
Less than 3 years	58	31.7
3-5 years	67	36.8
6-10 years	44	24.2
11-20 years	11	6.0
20+ years	2	1.1
No Experience	1	0.5
Highest Level of Education		
Bachelor's Degree	41	22.4
Master's Degree	89	48.6
Doctoral Degree	50	27.3
High School	1	0.5
College	2	1.1
Job Titles		
Data Scientist	66	36.0
C-Executive (CEO, CIO, CDO, Co-Founder)	26	14.2
Analyst	24	13.1
Manager	14	7.6
Engineer	12	6.6
Data Engineer	12	6.6
Developer	9	4.9
Scholar	8	4.4
Consultant	4	2.2
Others (Data Architect, R&D, etc.)	8	4.4

In order to understand the importance and value creation of each process, and to eliminate or merge correlated processes as well as to cluster these processes in different process areas, the survey results have been analyzed by six domain experts in data science, big data, and information systems domains. According to the analyses and judgements of the domain experts, the main constructs of the proposed model have been finalized. According to expert judgements, the overall 34 processes are clustered in five main process areas as Organization, Strategy Management, Data Analytics, Data Governance, and Technology Management. In order to assess the validity,

reliability, and consistency of defined process areas and to mitigate the reliability threats, a series of tests have been performed on the collected survey data. The results of these validity, reliability tests, and the correlations among processes areas are summarized in Table 3-4. At the process area level, all process areas have high correlations (>0.4) among each other. The Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy and Cronbach's Alpha-α for the consistency and reliability analysis are greater than 0.6 for all process areas deemed appropriate as the lower acceptability limits as indicated by Hair et al. [71].

Table 3-4. Validity, Reliability Test and Inter-Correlations

Process Areas	**Mean**	**S.D.**	**KMO**	**CA**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
(1) Organization	4.04	0.83	0.883	0.903	1.0				
(2) Strategy Management	4.09	0.67	0.612	0.625	0.669	1.0			
(3) Data Analytics	4.41	0.46	0.773	0.764	0.434	0.482	1.0		
(4) Data Governance	4.03	0.71	0.887	0.891	0.631	0.576	0.468	1.0	
(5) Technology Management	3.58	0.95	0.796	0.886	0.698	0.530	0.445	0.641	1.0
SD: Standard Deviation KMO: Kaiser-Mayer-Olkin CA: Cronbach Alpha									

3.1.3 Discussion

To validate the defined problem and answer RQ2, the opinions of the participants have been gathered about the need for an MM for the data science domain. According to the results, 88% of the respondents indicate that measuring their organization's capability for defined data science processes is "Extremely Important" or "Very Important" to improve the outcomes of data science projects, as depicted in Figure 3-1. In particular, 92% of the respondents who work in the manufacturing industry agree that measuring data science capabilities is "Extremely Important" or "Very Important". This is not surprising if we consider that digital transformation is expected to change the entire business and production models soon. Digital transformation is enabled by many technologies, including cyber-physical systems, digital twins, and the Internet of Things, closely tied to data science and pertinent concepts such as artificial intelligence and machine learning. Moreover, most of the respondents who gave the highest importance to measuring the organization's capability for data science processes work in large organizations with 500 or more workers. I observed that as the organization size grows, the need for process assessment and improvement roadmap also increases.

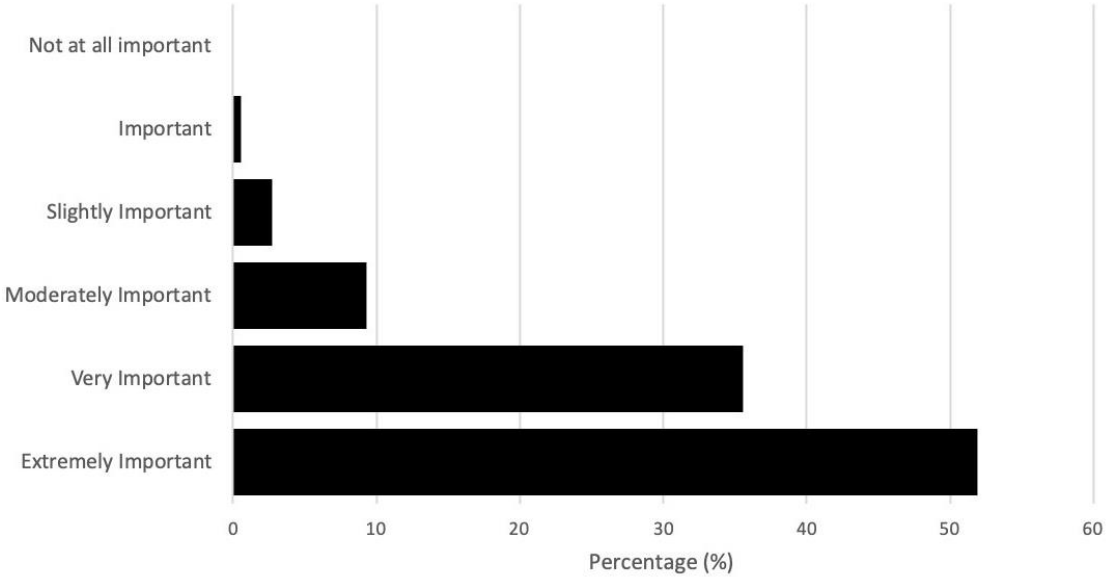

Figure 3-1. Do you think that measuring your organization's capability for data science processes is important or helpful?

In the survey results, there is clear evidence that organizations are attempting to improve their data science capabilities to extract measurable value and not to lag behind the hype. According to the results, 91% of the respondents state that they "Strongly Agree" or "Agree" to invest in data science operations in upcoming years, as depicted in Figure 3-2. Only 13% of the organizations indicate that they could integrate data science across the entire organization. There is a need to define a standardized guideline to increase the adoption of data science in organizations. To satisfy the necessity, DSCMM has been developed to assist organizations by providing a roadmap and a comprehensive strategy in assessing and improving their data science capabilities as described in the next section.

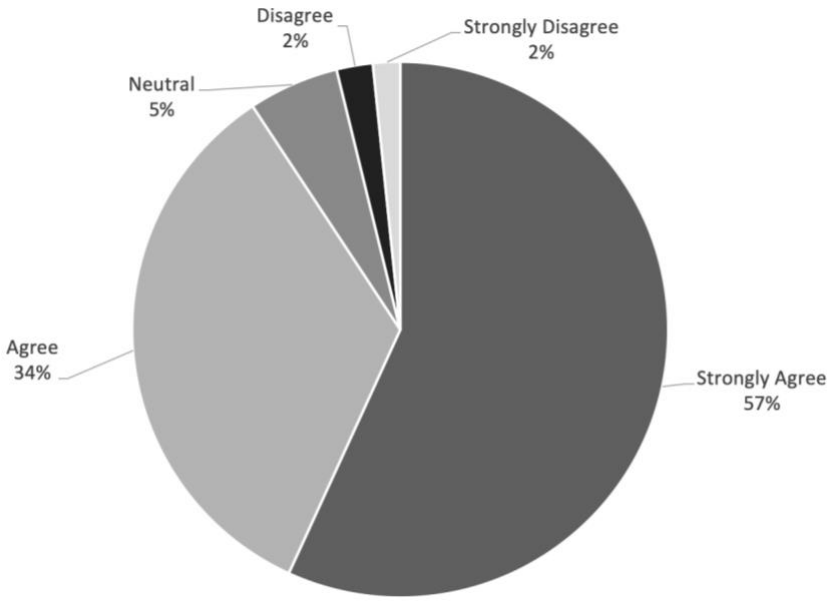

Figure 3-2. Do you think that your organization will invest to improve data science capabilities in upcoming years?

In the development of the process dimension of DSCMM, I have utilized the results of the exploratory survey. According to the findings of the exploratory survey, risk management, performance management, innovation management, and model deployment processes were excluded. Moreover, data life-cycle management, architectural approach and standards, master data management, and historical data management processes were combined in a single process. These modifications are also supported by the survey results. The mean and percentile distribution of risk management, performance management, and organizational governance processes significantly differ from the overall mean and standard deviation of the organization process area. However, the importance of organizational governance is increasing considerably as the organizational size grows. Since DSCMM needs to embrace larger organizations, the experts decided to keep the organizational governance process in the final model. In the strategy management process area, the innovation management and strategic alignment processes statistically have lower importance. Although the strategic alignment process does not have considerable importance, it seems very significant for the participants who work at large organizations that have more than 100 workers. In the data analytics process area, the model deployment process meaningfully deviates statistically from other processes for all organizational sizes.

Moreover, the model deployment process does not have any significant relationship with any other process to merge. In the data governance process area, data life-cycle management, architectural approach and standards, master data management, and historical data management do not provide any significant median, mean, and percentiles scores. However, there is a strong relationship that is larger than 0.75 among these processes. Thus, by taking into account the definitions of these processes, the experts decided to merge them into a single process as Data Management.

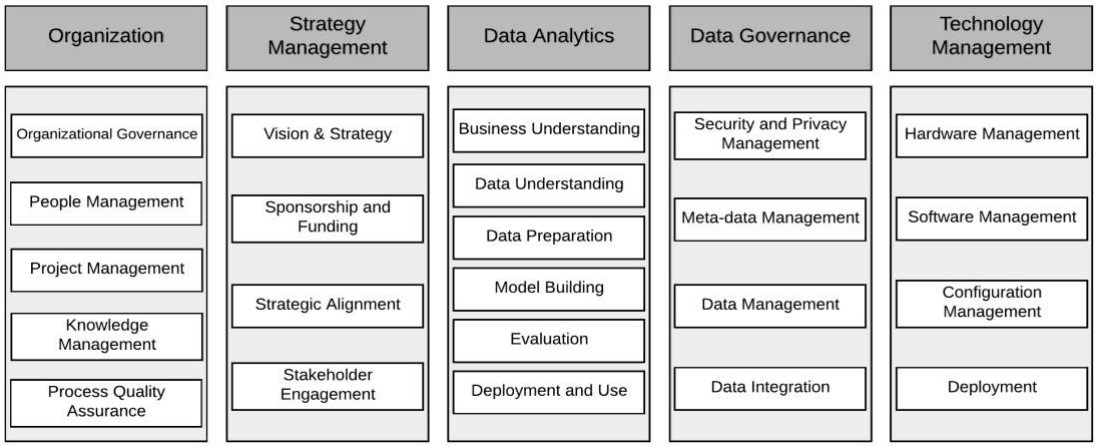

Figure 3-3. Process Areas and Corresponding Processes

Consequently, the process dimension of DSCMM is organized into five main process areas: Organization, Strategy Management, Data Analytics, Data Governance and Technology Management as depicted in Figure 3-3. These process areas and their relationships are also delineated in Figure 3-4. The process dimension comprises a reference model which includes definitions of purposes, BPs, and outcomes of each data science process. Hence, the defined data science processes can be evaluated whether the defined process is initiated and achieved its purpose in the organization or not by utilizing this reference model. I detailed the process dimension and each process in the following chapter.

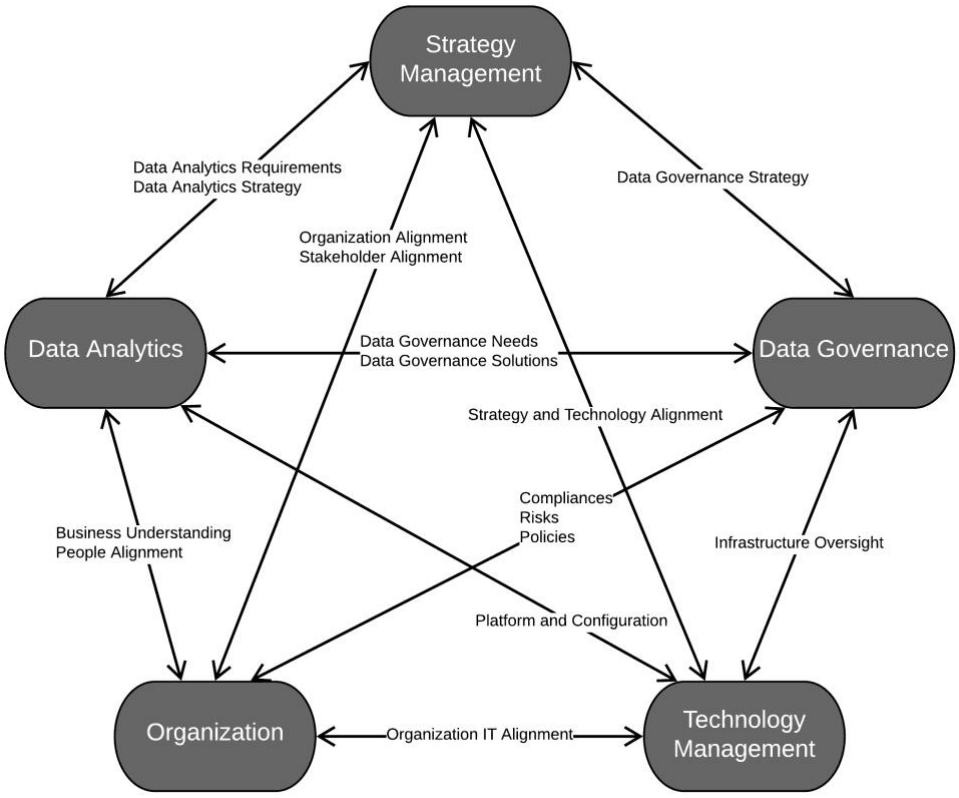

Figure 3-4. Relationship among process areas

3.1.4 Validity Threats

The survey approaches can provide valuable insights regarding the applied domain, but some validity concerns may arise in conducting and analyzing the results of the survey. Hence it is crucial to define validity threats and determine necessary measures in the planning phase. As suggested by Yin [72], I analyzed potential validity threats in four categories as follows;

Construct Validity threat for this survey is about ensuring the questionnaire really reflects the opinions that are relevant to the answers to the research questions. There might be issues regarding the interpretation of the terminology and constructs mentioned in the questionnaire. The researcher and the survey participants need to understand questions in the same way. To overcome this threat, I conducted a pre-test

survey with ten respondents, including data science experts and scholars. According to the feedbacks from pre-test respondents, survey constructs, questions, and definitions of processes were revised to increase understandability and generalizability. Since data science is an emerging research area, the description of data science and all of the defined processes were explicitly detailed to avoid misunderstanding among respondents.

Internal Validity threat is not a concern of this survey because the research questions do not include any "how" and "why" questions [72]. Nevertheless, in order to support our findings statistically, I performed a range of validity and reliability tests including, Kaiser-Mayer-Olkin (KMO) and Cronbach's Alpha, where applicable. The detailed results of these tests and measures are presented in Table 3-4.

External Validity threat of this study is about the generalizability of the results. To mitigate this threat, I collected data from respondents who work in different industries and organizational sizes (see Table 3-2 and Table 3-3). Moreover, I also collected data from different participants with a diverse set of qualifications and job titles, including executive members, managers, data scientists, analysts, data engineers, scholars, and developers.

Reliability evaluates that the operations of the research study, such as data analysis or data collection, can be repeated by another researcher and produce the same results. The protocol of our exploratory survey is given in detail throughout this chapter. The research questions, survey development methodology, data collection, and analysis methods are explained in detail.

3.2 Exploratory Case Study

In this thesis study, an exploratory case study was first performed to evaluate the applicability and usefulness of the main approaches that are utilized to develop DSCMM. This section explains the exploratory-case study design, potential validity threats, results, and measures taken against these threats at the design phase.

3.2.1 Exploratory Case Study Design

In this exploratory case study, I developed the "Business Understanding" process by following the requirements defined in ISO/IEC 33004 [69] to evaluate its capability

level. The process definition covers the process name, purpose, outcomes, base practices, and output work products. It also includes performance indicators of processes by defining a set of BPs to provide a definition of the tasks and activities that need to accomplish the process purpose and fulfill the process outcomes. I detailed the process definition of the "Business Understanding" process in Appendix B.

The main sources of evidence of this exploratory study are unstructured assessment interviews, audits, organizational documents, and observations. The assessment is planned, and the relevant data is gathered and validated by following ISO/IEC 33020: Process Measurement Framework for Assessment of Process Capability [36]. The research questions for this exploratory case study are as follows;

- *RQ-1:* How suitable and applicable is using an ISO/IEC 330xx based model with the purpose of identifying the current state of a data science process capability?
- *RQ-2:* How well does it provide a roadmap for improving the process capability of the organization's data science processes?
- *RQ-3:* What are the strengths and weaknesses of the proposed model for improvement?

In the exploratory case study, I have conducted a process capability assessment with a manufacturing organization that operates in the energy industry. This organization leverages data analytics to improve operational efficiencies of business units and to adopt data-driven decision-making and management culture. I held a two-hour assessment meeting with the executive managers that are mainly accountable for data analytics to collect data and evidence and a two-hour follow-up interview to discuss the applicability and usability of the model and assessment results. I collected data from multiple sources, including assessment interviews, organizational strategy documents, and observations. Organizational structures, hierarchy, strategy documents, and intra-communication channels among managers and employees are were also investigated. The assessment interviews were also recorded during the meetings with the consent of the participants.

Before conducting interviews, the scope of our research and the proposed model, including the process and capability dimensions and the assessment method, were presented to the participants in detail.

3.2.2 Exploratory Case Study Implementation

The CL assessment process plan for the exploratory case study is depicted in Figure 3-5. First, the assessment plan was created by defining stakeholders, process owners, and the expected interview schedule. Then, the plan was shared with the expected interview participants to determine the interview schedule. Finally, a three-hour semi-structured interview was held with the participants. The interview participants are primarily responsible for data analytics, IT, and management. In this interview, direct and indirect evidences were collected to rate PAs, determine CLs of the "Business Understanding" process. After considering the direct and indirect evidences, an assessment report that includes CL of the Business Understanding process and a list of suggestions to improve its capability to the next level was created and shared with the interview participants.

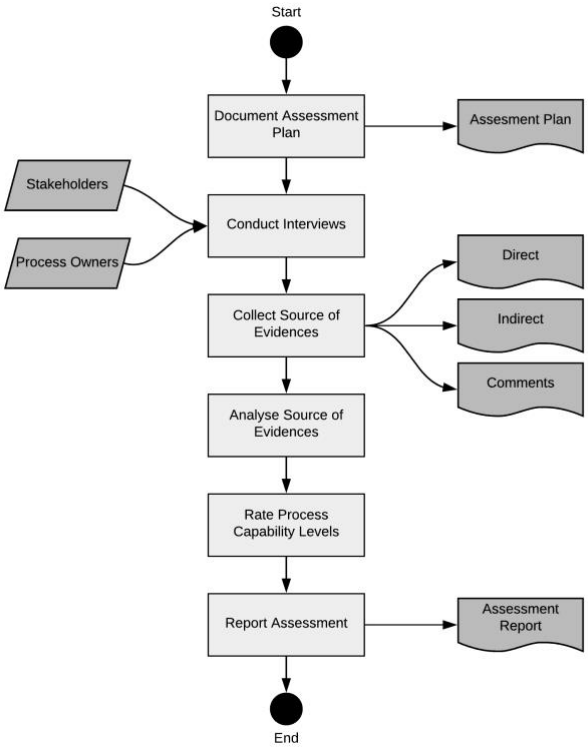

Figure 3-5. Assessment Process Plan

For assessments, I adopted the model and the process attributes (PAs) defined the ISO/IEC 330xx standard family, which extends the ISO/IEC 15504 Part – 5. Hence,

the assessments are based on PAs such as perform base practices, performance management, work product management, work product management, process definition, process deployment, process control, process analysis, process innovation, and process self-optimization implementation. In order to achieve CL – 1 for a process, an organization has to implement the corresponding base practices (BPs) of the process as PA-1.1. Process Performance evaluates whether the defined process is initiated and achieved its purpose in the organization or not. A process is rated as CL – 1 when the organization largely or fully performs the defined BPs. The other capability levels and related PAs are given in Section 4.3.

3.2.3 Results

I described the exploratory case study results and finding in Table 3-5. According to collected evidences, the "Business Understanding" process is determined as CL-2.

Table 3-5. Exploratory Case Study Assessment Results

Level	P.A.	Evidences	Result
Level–1	Perform Base Practices	The organization understands the business needs and requirements from a data point of view to formulate data science projects to achieve the objectives.	F
Level–2	Performance Management	The organization manages the process performance, but they need to explicitly define process performance and success metrics.	L
	Work Product Management	The organization clearly defines the output work product requirements. However, they mainly lack establishing quality and success metrics of the output work products and monitoring and evaluating them against these quality and success metrics.	P
Level–3	Process Definition	The organization has a standard definition for the "Business Understanding" process. However, they do not explicitly outline the interaction of the "Business Understanding" process among other business processes.	P
	Process Deployment	The organization develops, establishes, and maintains a guideline for process deployment and outlines the required infrastructure and work environment. Nevertheless, the organization mainly lacks defining and collecting data to evaluate the effectiveness and performance of the standardized "Business Understanding" process.	P

In order to reach CL – 3 in the Business Understanding process, the organization needs to improve their Performance, and Work Product Management attributes to Fully Achieved and Process Definition and Deployment at least Largely Achieved. In the assessment report, I provided some suggestions to reach Level – 3 in the Business Understanding process. These suggestions are as follows;

- Define the process performance and success metrics clearly.
- Outline quality criteria of the output work products; moreover, review and monitor output work products against these criteria.
- Establish a terminology glossary relevant to industrial terms and background information about data science projects to enable common understanding.
- Define the process (Responsibilities and authorities, key activities, sequence and interaction, milestones, work product templates, etc.) to perform it more robustly.
- Collect and establish a set of needs and requirements about infrastructure and work environment to deploy the process.
- Collect and analyze data to assess the progress and achievement of the business understanding process.
- Establish and maintain a draft project plan to analyze the project needs better and to estimate required IT and Human resources and budget to be needed throughout the project.

I also conducted follow-up interviews with the same participants after presenting the assessment results to find answers for the defined research question. In these interviews, structured questions which are answered by 5-point- Likert Scale (1: Strongly Disagree and 5: Strongly Agree) and open-ended questions, which are defined below, were utilized to discuss the applicability and usability of the proposed approach in determining capability levels of the sample data science process.

- Are process capability measurements and provided guidelines for improvement useful? (5-Point Likert Scale): Median: 4
- Do you think that applying these suggestions will improve the process performance? (5-Point Likert Scale): Median: 4
- Do you think that language and terminology in the questions are easy to understand? (5-Point Likert Scale): Median: 4
- Is there any missing item in the guideline for improvement list? (Open-ended): "*No"*
- Is there any information you want to add in the process definition? (Open-ended): "*No"*

The participants think that provided suggestions and guideline is applicable and useful to improve their current capability of the business understanding process, and they will apply these suggestions to improve their current capabilities. They also indicated that the language and terminology in the questions are easy to understand. Moreover, the process definition, guidelines, and suggestion list are complete and fully cover the business understanding process in the organization.

3.2.4 Validity Threats

The potential validity threats for the application of exploratory case study are investigated in four categories as proposed by Yin [72] to take corrective actions during the planning phase.

The possible *Construct Validity* threat for this exploratory case study research is identifying the correct source of evidences and interview participants to collect subjective judgements from both management and technical business units. The RQs of the exploratory case study are explicitly defined, and accordingly an assessment plan is prepared to find answers to these research questions.

Internal Validity threats are related to the relationships between input variables and produced outputs to understand and if the researchers can identify the key affecting factors in their study. To overcome internal validity threats, I first prepared an assessment results report and shared this report with the interview participants. Then, the assessment results and findings in the assessment report are reviewed and discussed to eliminate any deficient, bias, and error in the resulting findings.

External Validity threat concerns the generalizability of the case study results. In order to overcome this threat, I collected information and evidences from a diverse set of participants that have different responsibilities, including management, data analytics, and IT. Moreover, I also collected data from different multiple information sources, including interviews and observations. Moreover, the capability assessment process is explicitly detailed to participants. Each question and technical terms are also detailed to the participants by giving examples.

Reliability threats question the generalizability of the resulting findings and assess the applicability and validity of the same case study in different organizations. The main motivation of performing the exploratory case study is to understand the suitability

and applicability of the ISO/IEC 330xx based model in assessing the current process capability of data science processes. To this end, these threats are not considered in this exploratory case study.

CHAPTER 4

DATA SCIENCE CAPABILITY MATURITY MODEL (DSCMM)

This chapter presents the proposed Data Science Capability Maturity Model (DSCMM). I detail the final process and capability dimensions of DSCMM and present a method to evaluate organizational data science maturity in a staged manner.

In Section 4.1, the structure of DSCMM is presented. In Sections 4.2 and 4.3, the process capability determination and organizational data science maturity assessment methods are covered, respectively.

4.1 The Structure of DSCMM

The development of DSCMM was grounded on the ISO/IEC 330xx standard family [14], which comprises a set of standards to design a MM, define processes, plan and perform process capability, organizational maturity assessments, and define a roadmap for improvement according to the assessment results. The ISO/IEC 330xx standard family provides a well-established and commonly accepted structure. Hence, it is selected as a baseline for the development of DSCMM. In particular, in this thesis study, I utilized the ISO/IEC 33004 Standard [69] to define data science processes and maturity levels comprising DSCMM. The ISO/IEC 33003 [73] and ISO/IEC 33020 [36] standards were adopted to define the capability dimension of DSCMM. I utilized ISO/IEC 33002 [74] as a guideline to conduct process capability assessments, and ISO/IEC 33014 [75] was used to develop a standardized roadmap for process capability and organizational maturity improvements.

DSCMM has two core dimensions: the capability dimension and the process dimension. The capability dimension defines process Capability Levels (CLs), Process Attributes (PAs), and a rating scale. There are six main CLs from "CL – 0: Incomplete" to "CL – 5: Innovating". The process dimension embraces the Data Science Process

Reference Model (DS-PRM), data science process definitions, outcomes, and purposes. In the process dimension, 23 processes are grouped under five main process areas: Organization, Strategy Management, Data Analytics, Data Governance, and Technology Management. The high-level structure of the proposed model is depicted in Figure 4-1.

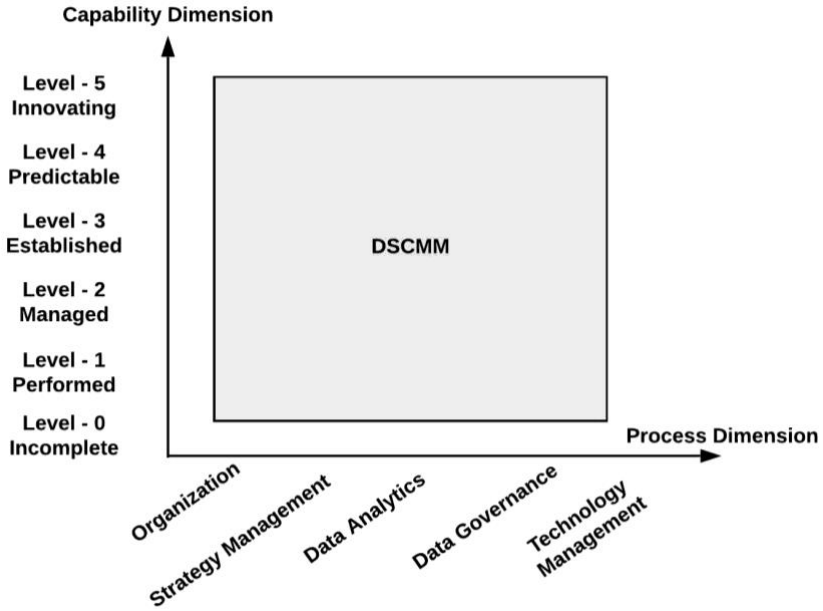

Figure 4-1. Structure of DSCMM

DSCMM provides a generic process capability and MM to support improvement from both organizational and process perspectives. Accordingly, it includes two main representations (i) continuous representation evaluating process CLs and (ii) staged representation assessing organizational data science maturity levels (MLs). In the continuous representation, each process is assessed individually in all dimensions of capability. The continuous representation allows an organization to select a specific data science process and make improvements based on it according to the business or specific project objectives. On the other hand, the staged model is designed to evaluate organizational data science maturity from a single source based on a data science process set corresponding to each maturity level. The staged representation demonstrates how an organization standardizes and consistently implements data science processes to achieve a desired maturity level throughout the organization.

4.2 Process Dimension

Data Science Process Reference Model [76], including process definitions of 28 data science processes, is developed by following the requirements defined in ISO/IEC 33004 [69]. The 23 of these processes, which are grouped under five process areas explained in the following subsections, are core data science processes to assess the process capability levels, and five of these processes are supporting processes, explained in Section 4.4.2 to define practices for improving overall performance, efficiency, and gains of data science-related activities in assessing organizational data science maturity.

Each process definition covers the process name, purpose, outcomes, base practices, and output work products as delineated in Table 4-1. It also comprises definitions and relationships among tasks, activities, and output work products to achieve the process purpose and realize the process outcomes. Hence, data science processes can be assessed according to these process definitions. A sample process, the Business Understanding, is detailed in Appendix - B. The detailed descriptions of all processes can be found in [76].

Table 4-1. Template for Process Definition

Process Reference Model	Process ID	The individual processes are described in terms of process name, process purpose, and process outcomes to define DSCMM reference model. Moreover, it also identifies processes with Process ID.
	Process Name	
	Process Purpose	
	Process Outcomes	
Process Performance Indicators	Base Practices (BPs)	Actions or action groups that affect the achievement of the process purpose.
	Output Work Products	Work products to realize the process purpose and outcomes.

Data science requires comprehensive management of organizational and technical process aspects. Thus, DSCMM broadly defines and assesses an organization's maturity from Organization, Strategy Management, Data Analytics, Data Governance, and Technology Management perspectives. These perspectives are discussed in the following sections.

4.2.1 Organization

Organizational structure, leadership, talent management, and culture aspects have a substantial impact on the success of data science projects [51]. Companies need to build a corporate culture, efficient leadership, and communication strategies among data science teams, stakeholders, and executive managers to become a data-driven organization [77]. Hence, the most critical requirements are developing, acquiring, and orchestrating human and organizational resources to support data-analytic thinking throughout the organization [78]. Thus, organizational maturity concerns how data-driven culture and management strategies are implemented and embraced among employees and executives. Accordingly, this process area defines BPs to evaluate organizational governance, people, project and knowledge management, and process quality assurance processes. The brief definitions and purposes of these processes are as follows:

- *Organizational Governance* process defines and develops roles and responsibilities within the organization as well as ensuring that the organization shares common goals and objectives to increase communication, collaboration, hierarchical alignment, and data-driven decision-making capabilities.
- *People Management* process focuses on acquiring, training, and integrating people skills to build the right team for data science.
- *Project Management* identifies, establishes, and controls the data science activities and organizational resources to manage the data science team, timeline, and troubleshooting in the context of the project's requirements and constraints.
- *Knowledge Management* process evaluates how the organization documents, stores, and transfers or shares the knowledge or know-how to demystify the data science processes across the organization and to develop work products more reliable.
- *Process Quality Assurance* process provides a standardized and repeatable roadmap and assurance for consistently producing high quality, availability, and value in whole data science lifecycle processes.

4.2.2 Strategy Management

Strategy management plays a significant role in managing data science investments in an efficient and effective manner. Organizations need a strongly and clearly articulated data science vision, strategy, and roadmap to tackle challenges towards becoming a

data-driven organization. Executive members, data scientists, and stakeholders need to be aligned in this strategy, vision, and roadmap to determine prior business cases and fund data science projects corresponding to the strategic business direction. A well-defined and aligned strategy and vision can move organizations forward on data science and is considered a significant success factor in future investments [51]. To define and assess the strategy management maturity, we can examine whether the organization has a well-defined vision & strategy, sponsorship and funding, strategic alignment, and stakeholder engagement processes. The brief definitions and purposes of these processes are as follows;

- *Vision & Strategy* process aims to define business strategies, goals, and vision clearly to make better business decisions, improve data-driven organization capabilities, reduce costs, increase profitability and optimize business processes.
- *Sponsorship and Funding* process focuses on managing internal and external data science investments to grasp optimal gain from strategically aligned investments at an affordable cost with a known and acceptable level of risk.
- *Strategic Alignment process* evaluates how data science initiatives are supported by top managers and be aligned at all levels of an organization includes executives, IT, data management, and stakeholders are aligned with bi-directional feedback.
- *Stakeholder Engagement* process evaluates how organizations embrace involving stakeholders in their decision-making and data-product development processes.

4.2.3 Data Analytics

This process area addresses the data analytics pipeline, including collecting, storing, transforming, and analyzing data and building predictive and prescriptive models [64]. The data analytics process lifecycle comprises processes to derive actionable insights from raw data [79]. It includes understanding the business requirements to formulate data analytics problems, cleaning incomplete, noisy, and inconsistent data, building descriptive, diagnostic, predictive, and prescriptive data analytics models, and evaluating them according to key performance indicators. Moreover, the data analytics process area also comprises enriching existing datasets by utilizing feature engineering approaches to derive, extract or, generate new attributes. There are various well-known frameworks or process models in data analytics, including SEMMA [80], Cross-Industry Standard Process for Data Mining (CRISP-DM) [79], and Knowledge Discovery in Database (KDD) [81]. These well-known analytics frameworks may differ in low-level details, but they mostly share a common understanding of the data

analytics lifecycle and essentially highlights six-core processes: business and data understanding, data preparation, model building, evaluation, and deployment and use as defined below.

- *Business Understanding* process identifies the business needs and requirements from a data point of view to formulate data science projects to achieve the objectives.
- *Data Understanding* process analyzes data needs and data sources, verify data quality, and analyzes data. This process represents the organizations' ability to understand the data with statistical and mathematical thinking.
- *Data Preparation* process assesses how organizations deal with data cleansing, data wrangling, data quality, and structuring.
- *Model Building* investigates how useful knowledge is extracted in a timely manner and to make data-driven organization capabilities pervasive in an organization, and how organizations embrace a diverse set of analytical techniques from descriptive analytics to prescriptive analytics.
- *Evaluation process* evaluates and validates the success of developed data products and services.
- *Deployment and Use process* focuses on utilizing the insights gained from the data science products into business actions as well as preparing a development, test environment, and IT infrastructure to leverage data science products and services.

4.2.4 Data Governance

Data governance can pose significant challenges that may affect the entire business processes in leveraging data science [82], and it needs to be elaborated at a corporate level [52]. The organizations collect data from a diverse set of data sources, which may include inconsistent, partial, and incorrect data, which mainly causes data integration and standardization issues. Data security, privacy, and quality problems should be addressed and managed before building models and making decisions based on these models [50]. The companies need to ensure that data is collected, stored, and managed securely, accurately, and effectively [44]. Data quality also plays a key role in improving data governance maturity since poor data quality and inadequate data may cause incorrect, missing, and unreliable information [83]. To improve information and data science product quality and to utilize relevant data to solve business problems, organizations need to develop and promote data quality awareness. The Data Management Body of Knowledge (DMBOK) [83] highlights crucial data governance

processes as security and privacy management, data and meta-data management, and data integration. Thus, the data governance process area defines and evaluates to what extent these processes are implemented. The brief definitions and purposes of these processes are given as follows:

- *Security and Privacy Management* process aims to monitor and manage permission, credentials, data, user, and project accesses continuously for ensuring consistency in the context of organizational data security needs and requirements.
- *Meta-data Management* process operationalizes data awareness and data lineage to enable data scientists to browse and understand data from a metadata perspective. Such metadata may include textual descriptions of tables and individual columns, key summary statistics, data quality metrics, among others.
- *Data Management* process evaluates how an organization defines data standards and architecture to manage data flow across sub-systems and map their data to business processes as data flows from one process to another.
- *Data Integration* process focuses on the integration and reconciliation of data from multiple sources into target destinations and standards to make data available at lower cost and complexity for leveraging shared data products and services.

4.2.5 Technology Management

Technology management is another prominent process area in the context of data science. The data science products and services require a scalable and distributed technology infrastructure to efficiently and effectively manage the unique characteristics of today's big data [82]. Organizations need to follow a systematic approach to tackle technology management challenges, including the complexity and pace of technology advancements and the wide range of technology sources to expedite their development and deployment processes [84]. The maturity of an organizations' technology management ensures that the available data and data products are leveraged with state-of-the-art hardware and software technologies most effectively and efficiently [78]. To address these challenges, the technology management process area assesses organizations' ability to identify, select, acquire, learn, exploit, and maintain their technology sources to develop and deploy data science products or services effectively. Accordingly, this process area is concerned with the hardware, software, configuration management, and deployment processes as detailed below.

- *Hardware Management* process manages on-premises, hosting, hybrid, and private hardware infrastructure solutions according to the security and privacy requirement.
- *Software Management* process evaluates the organizations' capability to reach, adopt, and integrate emerging data science software tools.
- *Configuration Management* process covers the definition of organization standards to install, configure and maintain the hardware and software infrastructures to optimize availability, speed, reliability, and security.
- *Deployment* process focuses on establishing and maintaining the integrity of data artifacts and products into existing infrastructure and management of deployment lifecycle, including version control.

4.3 Data Science Process Capability

The capability dimension describes the ability of an organization in a process and to what extent it provides existing and planned strategic business goals. It is adapted from the standard of ISO/IEC 33020-*Process assessment—Process measurement framework for assessment of process capability* [36]. It provides a well-defined evolutionary plateau describing the organization's capability relative to a particular process. It is applicable to all processes, independent of domain.

DSCMM has six CLs ranging from CL – 0 to CL – 5. All CLs except CL – 0 includes at least one PA to indicate an essential advancement in the capability of performing a process and address a specific strength of the assigned level. The PAs progress through the improvement of the capability of any process. A process CL reflects an organization's ability in a process to satisfy certain PAs. For a given process, in order to achieve a specific CL, the process is required to be at least "Largely Achieved (L)" for the corresponding level's PAs and "Fully Achieved (F)" for the lower CLs' PAs. Figure 4-2 presents the CLs and PAs assigned to each CL according to ISO/IEC 33020.

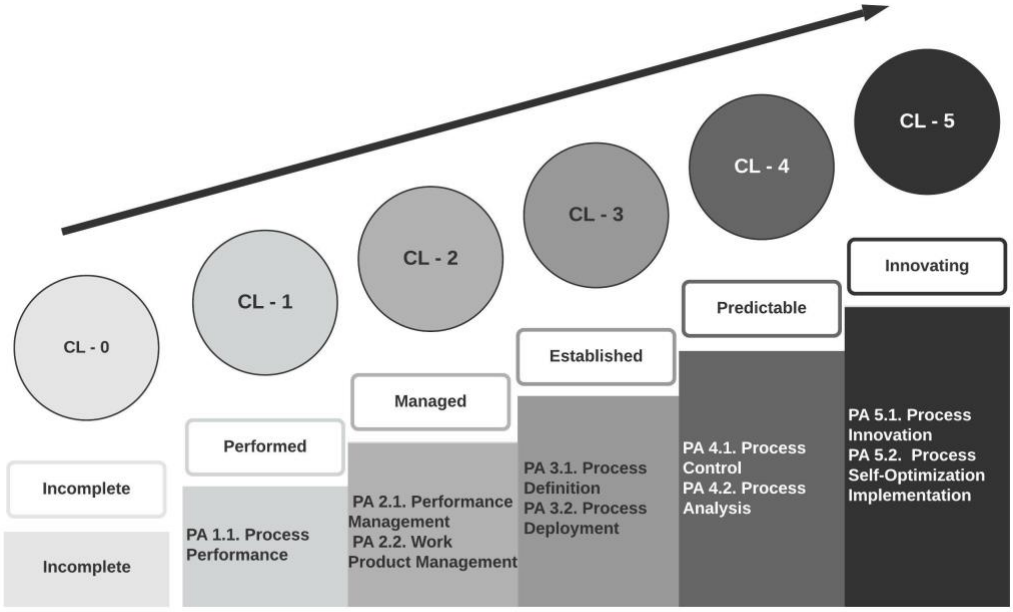

Figure 4-2. Capability Levels and Process Attributes

CL – 0 Incomplete: In this CL, the organization does not successfully establish any of the defined BPs. In other words, the organization does not have any initiative to adopt defined data science process.

CL – 1 Performed: This capability level assesses the PA 1.1. Perform Process Attributes, which structurally defines processes, their BPs, and output work products. To achieve Level 1, organizations should at least largely implement BPs of PA 1.1. Process Performance to demonstrate achievement in data science. The organization largely or fully performs the BPs of PA 1.1. Process Performance, but most of these processes are performed ad-hoc. In other words, there is not a consistent way of performing data science processes. At this CL, the processes are unpredictable, poorly controlled, and reactive.

CL – 2 Managed: The BPs are fully performed. The organization starts to recognize the business value of data science and starts focusing on improving process performance by defining performance objectives for each process. Moreover, organizations are expected to identify, manage, and control the work products of each process for performing processes consistently.

CL – 3 Established: This level comprises PA 3.1. Process Definition and PA 3.2. Process Deployment. PA 3.1. Process Definition assesses organizations' ability to establish a standardized and repeatable process definition for each process and PA 3.2. Process Deployment evaluates how organizations manage and maintain defined processes. In this CL, organizations need to fully achieve the PAs of CL 2 and also at least largely perform process definition and process deployment PAs to achieve level 3. At this level, organizations are expected to perform and maintain processes in a standardized way by defining and controlling each process.

CL – 4 Predictable: In this capability level, we assess how organizations establish quantitative objectives to monitor and control their process performance with PA 4.1. Process Control and organizations' ability to employ effective, practical, and quantitative process performance measures to reduce variation in process performance with PA 4.2. Process Analysis. The organization begins managing its processes through the quantitative data that describes the performance, and variations in performing best practices are reduced by controlling and analyzing processes. A controlled process is planned, performed, and monitored quantitatively. Thus, at this level, statistical and quantitative measures are collected to control and monitor the process against the plan and to take appropriate corrective action when it is needed. Moreover, organizations utilize outputs of data science in order to make all of their business decisions rather than intuitions of managers and process owners.

CL – 5 Innovating: In this capability level, we assess PA 5.1. Process Innovation to evaluate organizations' ability to identify potential improvements based on process performance and PA 5.2. Process Self-Optimization Implementation to assess how organizations learn from their quantitative process analysis to improve their processes continuously. At this CL, the organization fully realizes defined data science PAs and starts self-learning from collected measures to improve the performance of data science life-cycle continuously. Moreover, the business model is evolving into an innovative structure with the gained insights from data science.

Table 4-2. Rating scale percentage values according to ISO/IEC 33002

Rating	*Achievement*	*Explanation*
Not Achieved (N)	0% to ≤ 15% achievement	There are no or only very limited indications of PA fulfilment.
Partially Achieved (P)	> 15% to ≤ 50% achievement	There are some evidences that the PA is partially implemented. However, some processes, the process still remains unpredictable.
Largely Achieved (L)	> 50% to ≤ 85% achievement	There is a significant achievement of the defined PA in the assessed process, but process performance still has some weaknesses.
Fully Achieved (F)	> 85% to ≤ 100% achievement	The measured PA is implemented completely.

As it is delineated in Figure 4-2, the proposed DSCMM has six CLs from Level 0 to Level 5. A CL represents a well-defined set of PAs, which are defined as a measurable property of process capability, they represent the degree of the achievement of the BPs or PAs for the assessed process. The capability level of a process is determined based on the ratings of the PAs. Process capability indicators are the means of achieving the capabilities addressed by the considered PAs. Evidence of process capability indicators supports the judgment of the degree of achievement in the PA. The rating scale percentage values used for rating PAs according to ISO/IEC 33002 [36], which is depicted in Table 4-2, are utilized.

4.4 Organizational Data Science Maturity

Organizational data science maturity demonstrates how an organization standardizes and consistently implements data science processes to achieve a desired level throughout the organization. The maturity of an organization is evaluated by assessing its process capabilities in a staged manner to promote incremental improvements through standardization and optimization. The maturity levels (MLs) are defined in an ordinal scale to measure data science process capabilities, prioritize organizational improvement efforts, and support continuous improvement. As it is delineated in Figure 4-3, DSCMM comprises six maturity levels from ML-0 Incomplete to ML-5 Innovating. These maturity levels include certain goals to achieve better defined and more consistently implemented data science processes as the organization gains maturity. Each level is built upon the previous level, and assessments start from

essential data science processes and delve into intricate details to support iterative improvement and adoption.

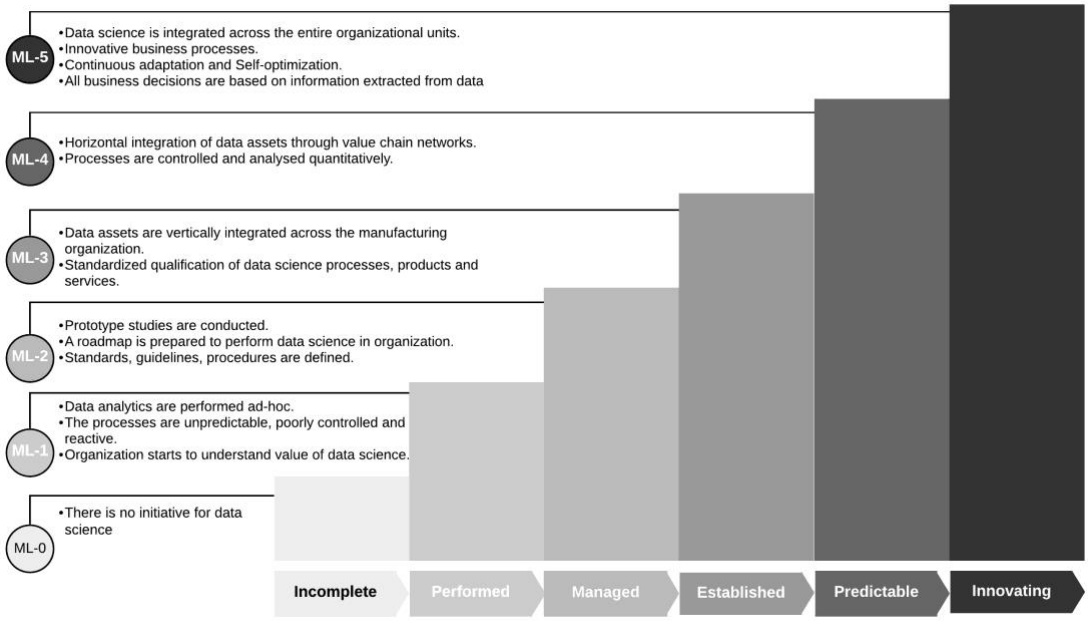

Figure 4-3. Organizational Data Science MLs

4.4.1 Maturity Assessment

In maturity assessment, DSCMM evaluates the capabilities for each process in four capability levels ranging from CL – 0 to CL – 3. Table 4-3 presents the process attributes and required PA ratings for each capability level to perform a maturity assessment. In order to achieve CL – 1 for a process, an organization has to implement the corresponding BPs of the process as PA-1.1. Process Performance evaluates whether the defined process is initiated and achieved its purpose in the organization or not. A process is determined as CL – 1 when the defined BPs are at least largely performed. PA-2.1 Work Product Management, which covers establishing, monitoring, and controlling output work products, as well as PA-2.2 Performance Management, covering defining, planning, and managing process performance, are evaluated for CL – 2. At CL – 3, the organizations should have a defined process which is deployed through the organization such that PA-3.1 Process Definition and PA-3.2.

Process Deployment are rated at least Largely Achieved, and lower PAs should be rated as Fully Achieved.

Table 4-3. CLs and PAs for Maturity Assessment (from ISO IEC 33020)

CL	*Process Attributes*	*Rating*	*Achievement*
CL – 0	Not Applicable	Not Applicable	There is no initiative to implement the process.
CL – 1	*PA-1.1: Process Performance*	*Largely Achieved*	The base practices of the process are performed in an ad-hoc manner.
CL – 2	PA-1.1: Process Performance *PA-2.1: Work product man.* *PA-2.2: Performance man.*	Fully Achieved *Largely Achieved* *Largely Achieved*	The process is implemented, planned, and monitored. In addition, the output work products are established, controlled, and maintained.
CL – 3	PA-1.1: Process Performance PA-2.1: Work product man. PA-2.2: Performance man. *PA-3.1: Process Definition* *PA-3.2: Process Deployment*	Fully Achieved Fully Achieved Fully Achieved *Largely Achieved* *Largely Achieved*	The process is performed in a standardized way by appropriately defining, establishing, and controlling process performance and output work products.

4.4.2 Supporting Process Area

In maturity assessment, the PAs defined in CL-4 and CL-5 are not evaluated. Thus, I defined a set of supporting processes to define practices for improving overall performance, efficiency, and gains of data science-related activities. The process areas and corresponding processes to assess organizational data science maturity is depicted in Figure 4-4. The supporting process area is only included in the organizational maturity assessment model as defined in ISO/IEC 330xx. This process set is named as Supporting Process area and includes Prototype Implementation, Vertical Integration, Horizontal Integration, Quantitative Performance Management, and Quantitative Process Improvement processes.

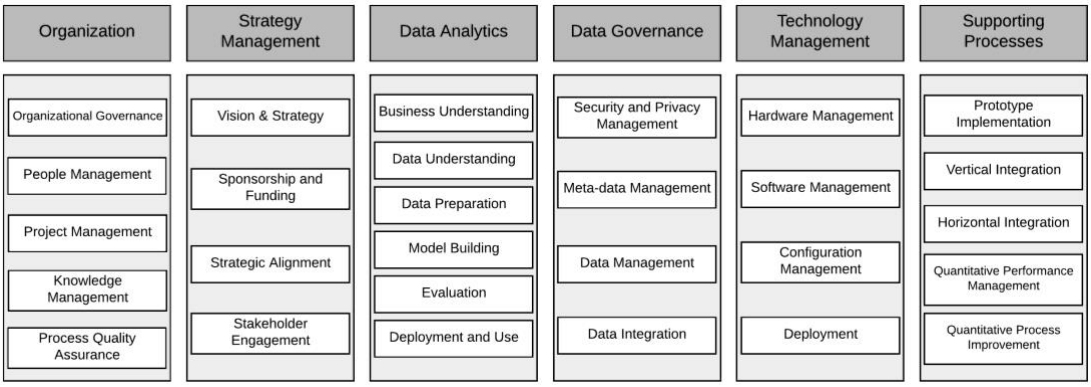

Figure 4-4. Maturity Assessment Process Areas and Corresponding Processes

Supporting: This process area evaluates to what extent the value chain of data science processes is digitized and integrated horizontally with external customers and suppliers and vertically with business units. Horizontal and vertical integrations of data and data products are crucial in improving data science [57], [85]. The prototype implementation process evaluates the applicability and usefulness of an innovative concept in a data science product before making investment decisions. The supporting processes area also involves performance management and improvement for each data science process through the use of measurements and quantitative techniques to ensure that the implemented processes support the achievement of the organization's relevant business goals. Thus, this process area includes prototype implementation, vertical integration, horizontal integration, quantitative performance management, and quantitative process improvement processes. The brief definitions and purposes of these supporting processes are as follows:

- *Prototype Implementation* process evaluates the applicability and usability of an innovative concept in a data-driven organization or a data science product before making investment decisions.
- *Vertical Integration* process aims to control and integrate the value chain of data science within the organization by identifying dependencies and touchpoints among organizational processes.
- *Horizontal Integration* process aims to control and integrate data science across the value chain of the organization by identifying dependencies and touchpoints among internal and external stakeholders as well as organizational units.

- *Quantitative Performance Management* process aims to evaluate process performances by leveraging quantitative measures and techniques to ensure process performance and support in accomplishing relevant business goals.
- *Quantitative Process Improvement* aims to improve the overall performance of the data science processes according to the analysis of the collected quantitative performance measures to become a data-driven organization in a systematic manner.

4.4.3 Maturity Levels

In this thesis, an expert panel comprising senior academicians, IT managers, senior executive members were formed to map the data science processes to organizational data science maturity levels. In line with the panel's suggestions, the MLs and, the corresponding processes and required CLs given in Figure 4-5 are determined. ML – 1 requires achieving at least CL – 1 for data science processes comprising the basic data analytics pipeline. For higher maturity levels, DSCMM requires achieving higher CLs for extended and more complex processes to support iterative improvement and integration. These maturity levels are as follows:

ML – 0 Incomplete: At this ML, there is no initiative for data science and an awareness of data science and its potential benefits in the organization.

ML – 1 Performed: The organization starts exploring the potential value of data science at this ML. However, the core data science processes are performed in an ad-hoc fashion, and their results are inconsistent and not standardized. Thus, the results of data science processes are unpredictable, poorly controlled, and reactive. Moreover, there is no strategy or vision to initiate data science investments according to the strategic business direction. The processes assigned to this ML should be at CL – 1 or higher.

ML – 2 Managed: The organization starts discovering the potential business value of data science by conducting prototype implementations. An organizational strategy and business directions are defined, and some technological investments are initiated to manage big data efficiently. The organization is motivated to advance its data science process performance by explicitly outlining standards, guidelines, and process performance goals. Additionally, it is expected to establish a data science roadmap to build a data-driven culture and strategic alignment throughout the organization. All

processes assigned to ML – 1 and ML – 2 should be at CL – 2 or higher in the organization at data science ML – 2.

ML – 3 Established: At this ML, organizations recognize data science as a core competency to attain competitive advantage. An organization at ML – 3 is expected to implement, define, and control data science processes in a standardized way. It is also needed to vertically integrate data science processes, services, and output work products across the organizational units to extend the application and benefits of data science. To achieve ML – 3, all processes assigned to MLs 1, 2, and 3 should be at CL – 3.

ML – 4 Predictable: The organization continuously analyzes its processes by collecting and monitoring statistical and quantitative measures to improve the process performances. At this level, an organization completely embraces a data-driven approach in decision-making processes rather than managers and process owners' intuitions. In order to achieve ML – 4, Horizontal Integration and Quantitative Performance Management processes should be at CL – 3, and processes assigned to lower MLs should also be at CL – 3.

ML – 5 Innovating: Data science and the capability to extract value from data are regarded as crucial strategic assets in the organization. The organization horizontally integrates its data science solutions across stakeholders and supporting organizations in the supply and value chains. The data science processes are improved continuously, and innovative business processes are defined by learning from collected quantitative and statistical measures. To achieve this ML, Quantitative Process Improvement and the rest of the processes should be at CL – 3.

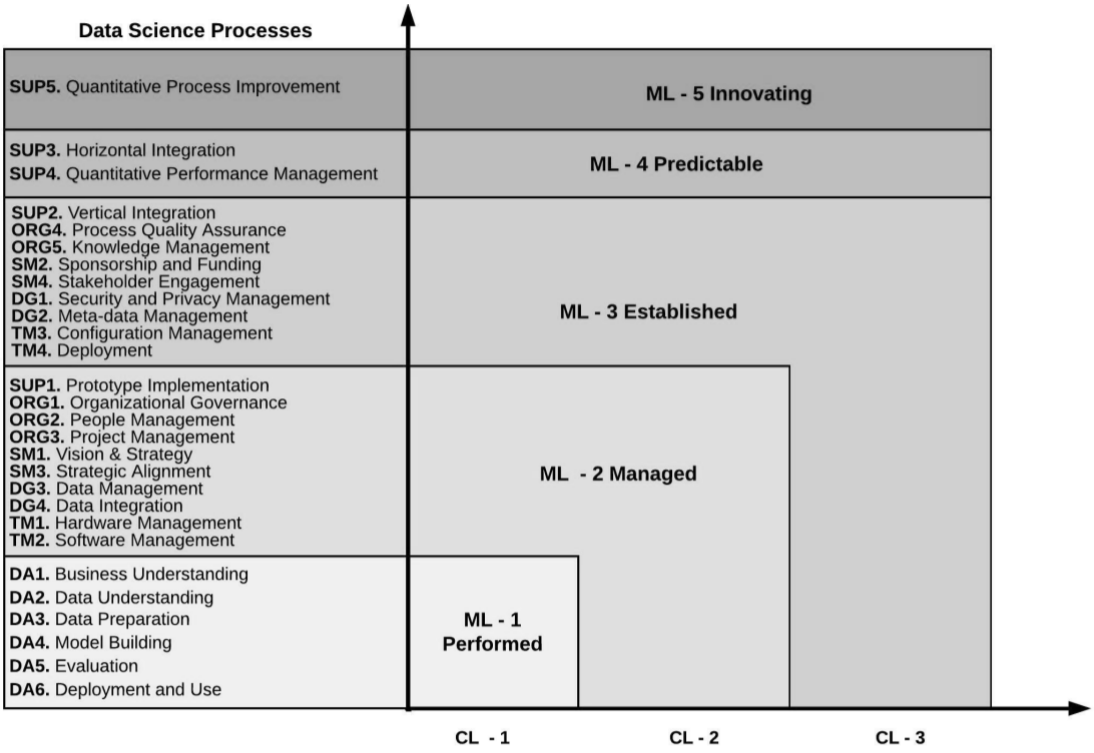

Figure 4-5. Relationship between CLs and MLs

CHAPTER 5

APPLICATION OF DSCMM

In this chapter, I present the application of DSCMM via case studies. In Section 5.1, I describe the multiple case study design, data collection methods, and potential validity threats and measures taken against these threats at the design phase. In Section 5.2, I present the assessment results. I discuss the assessment results and applicability and usefulness of DSCMM in Section 5.3.

5.1 Multiple Case Study Design

In this thesis study, a multiple case study is conducted to assess the applicability and usefulness of DSCMM [26]. This section explains the multiple case study design, potential validity threats, results, and measures taken against these threats at the design phase.

The multiple case study research is performed in three organizations of various organizational sizes, sectors, and data science adoption levels. Existing MM studies are strikingly limited in conducting empirical research to guide their application and validate their models' applicability and usefulness [19]. To this end, the multiple case study is employed to demonstrate and validate the proposed model's applicability and usefulness. Accordingly, the steps taken and actions performed to assess process capability and organizational maturity levels are explained in a detailed manner. In this thesis study, we followed the case study research methodology proposed by Yin [86] as follows;

- *Define the objective of the research:* The main motivation and objective in conducting a multiple case study is to evaluate the applicability and usefulness of proposed DSCMM.

- *Define Research Questions (RQs):* The RQs consistent with the objective of the thesis study are as follows:
    - *RQ-1:* How applicable and useful is DSCMM for identifying the current data science process CLs and organizational MLs?
    - *RQ-2:* How useful are the roadmap and suggestions provided by DSCMM evaluations in an organization for process capability and organizational data science maturity improvement?
- *Define Case Study Design Type:* We performed a multiple case study that involves three different organizations of different sizes and from different sectors.
- *Define the Measures Used in the Case Study:* The main measures of this multiple case study are the CLs of data science processes and the data science MLs of the organizations as defined in DSCMM.
- *Define Data Collection and Limitations:* Direct, indirect, and comments as evidence are collected from assessment interviews, follow-up interviews and information gathering documents, and observations.
- *Evaluate Objectivity of the Judgements:* We followed the ISO/IEC 33002 [74] to ensure the objectivity, reliability, and performance of this multiple case study research and data analysis for assessments. The ISO/IEC 33002 outlines standardized and impartial set of activities to plan assessment, collect data, report and share assessment results. It also provides a standardized rating scale.

5.1.1 Data Collection

To address the defined research questions, an assessment team assessed the data science process CLs and organizational MLs of three different organizations. assessment team including a lead assessor and four assessment team members. Each process attribute is rated with the consensus of assessment team as defined in the ISO/IEC 33002 [74]. The assessment team has competency in applying practices and guidelines provided by ISO/IEC 330xx. They also have experience in data science, big data, digital transformation, process quality assurance, and IT management domains.

The main reasons behind the selection of these organizations are that they have initiatives and experience in data science activities; they are willing to optimize their production processes, costs, manage their investments, and improve their data-driven manufacturing and management capabilities through data science.

The first organization operates in the energy industry, and in the rest of the thesis, I will refer to this case as The Energy Company for the sake of confidentiality. It produces energy materials in Europe and Asia regions with more than a thousand employees. The Energy Company carries out different data science projects to predict product quality, anomaly detection, and maintenance scheduling in its production environment. In the Energy Company, the assessment team conducted a three-hour assessment interview to collect direct and indirect evidence to rate PAs, determine capability levels of the data science processes, and a two-hour follow-up interview after presenting assessment results to discuss findings. A data scientist, a process owner, and a manager attended these interviews. The assessment interviews and follow-up interviews took 5 person-hours in total, and appraisal to evaluate and rate related PAs took 10 person-hours in total.

The second organization is a consumer goods company that produces daily care and cleaning products. It is a multinational company and has been operating worldwide in fifty countries for more than a hundred years. This organization has an extensive data science project portfolio, and some of its projects are related to customer analytics, predictive maintenance, and sales forecasting. In this case, the assessment team held a four-hour assessment interview and a two-hour follow-up interview. The interview participants are primarily responsible for digital transformation, data analytics, IT, and management. I will refer to this case as The Consumer Goods Company in the rest of the thesis. In this case study, the assessment team spent 6 person-hours in assessment and follow-up interview and 12 person-hours in total to evaluate collected data and information and rate PAs.

The third organization produces machines, and in the rest of the paper, I will refer to this case as The Machine Company. It is a medium-sized organization and has a limited budget and fewer employees compared to the Energy Company and Consumer Goods Company. It produces machines and spare parts in Europe. In the Machine Company, the assessment team conducted a four-hour interview with the head of data analytics, the IT manager, and the quality manager to collect data and evidences. The assessment team also conducted a two-hour follow-up interview with the same participants after presenting the assessment results and findings. In this case study, the assessment interviews and follow-up interviews took 6 person-hours in total, and spent 6 person-hours in total in evaluating and rating PAs.

The main prerequisites to apply DSCMM are that organizations need to be aware of the potential benefits and impact of data science and willing to optimize their production processes, reduce operational costs, manage their investments, and improve their data-driven manufacturing capabilities by means of data science. They have a project portfolio consisting of a diverse set of data science projects and assigned adequate people and technological resources to carry out these projects.

Intending to eliminate biases and provide consistent, reliable, and repeatable results, the assessment team followed the process capability maturity assessment guideline provided by ISO/IEC 33020 [36] throughout the assessment. The assessment team identified appropriate participants through contacting organizations and presenting the scope of DSCMM. The assessment team also recorded these interviews with the consent of the participants to transcript evidence and rated PAs impartially. DSCMM, assessment, data collection procedure, assessment plan, and capability/MLs are detailed to interview participants before assessments.

The assessment activities for the multiple case study research are outlined in Figure 5-1. The process starts with the documentation of the assessment plan and sharing it with the organizations and interview participants. The assessment team collected three sources of evidence: indirect, direct, and comments, to assess each PA during the interviews. They shared and presented assessment results as a final report that includes process capability levels, organizational ML, and a roadmap that comprises suggestions and recommendations to support continuous improvements in each process capability and organizational maturity.

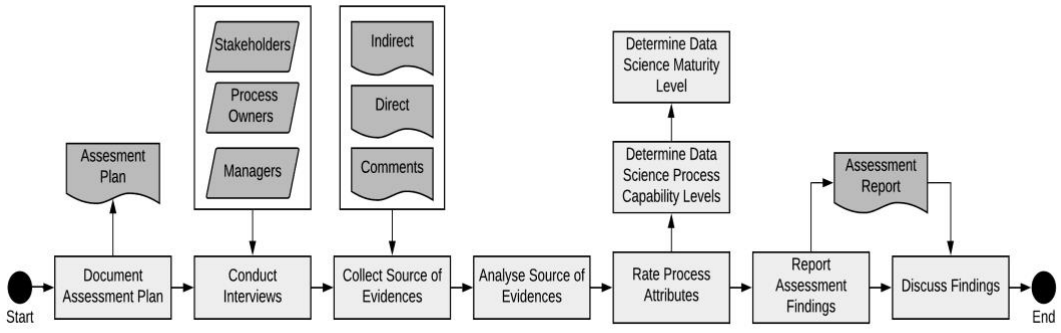

Figure 5-1. Assessment Activities for the Multiple Case Study Research

5.1.2 Data Analysis and Validity Threats

In the analysis of the collected data from the multiple case study and the evaluations of process capabilities of participant organizations in an objective and unbiased manner, the rating scale, assessment guidelines, and recommendations provided by ISO/IEC 33020 [36] are utilized. The case study researches may raise several validity concerns in collecting data and analyzing the results. Thus, the potential validity threats need to be determined in the planning phase to take corrective action at the early stages. As suggested by Yin [86], I analyzed potential validity threats for the multiple case study design in four categories as follows:

The main *Construct Validity* concerns for this multiple case study are about interpreting questions, collecting subjective judgments, and rating PAs objectively. The researcher and participants should interpret questions in the same way. To ensure this, the assessment team explicitly detailed each question and technical terms to the participants by giving real-world examples to avoid misunderstanding. The researcher needs to identify the correct source of evidence and participants to collect subjective judgments from both management and technical business units. To overcome potential threats, the assessment team collected data and information from different participants including process owners, data scientists, and executive members. The assessment team collected data from multiple sources, including interviews, organizational documents, and observations. The assessment team also took notes and recorded each interview to manage the assessment process objectively, impartial, and reliably. Each PA needs to be rated with a well-accepted method for objective and unbiased results. Thus, the rating scale provided by the ISO/IEC 33020 [36] as detailed in Table 4-2 is utilized.

Internal Validity is the extent to which causal relationships between input variables and produced outputs are explained by the factors considered in this study. To mitigate Internal Validity threats, the assessment team asked detailed questions about each data science process to understand its implementation and role in the business. After the assessment report is prepared, the assessment team also shared and discussed the findings with interview participants to eliminate assessment bias.

External Validity concerns the generalizability of the case study results. To improve the generalizability of the results and observe the applicability of DSCMM in different settings, a literal replication logic of the case study is employed in three different

organizations of different countries, organizational sizes, sectors, and data science adoption levels. Moreover, the assessment team collected data and evidence from a diverse set of data science projects in the organizations to support the generalizability of the case study results.

Reliability evaluates the extent to which the operations of the research study, such as data analysis or data collection, can be repeated by another researcher and produce the same results. To ensure reliability, we clearly developed and explained the case study materials, including research protocol, research questions, and assessment plan, as well as maintaining a chain of evidence. The ISO/IEC 33002 [74] standard is also utilized to ensure the reliability of the assessment activities and data analysis. We followed the same case study material in all of the case studies and produced consistent results and outputs.

5.2 Assessment Results

According to the collected direct evidences, indirect evidences, and comments during the assessment interviews, the capabilities of data science processes are determined by rating related PAs. The ratings of PAs and the corresponding capability levels of the data science processes assigned to each ML for The Energy Company, The Consumer Goods Company, and The Machine Company are respectively given in Appendix – C. ML – 1, ML – 2, and ML – 3 evaluation results are summarized in Figure 5-2, Figure 5-3, and Figure 5-4 respectively. In these tables, the black and gray boxes denote the critical processes that organizations should give priority to move their organizational maturities to the next higher level. According to these results, the organizational data science maturity of The Energy Company is determined as ML – 1, The Consumer Goods Company is determined as ML – 2, and the Machine Company is determined as ML – 0.

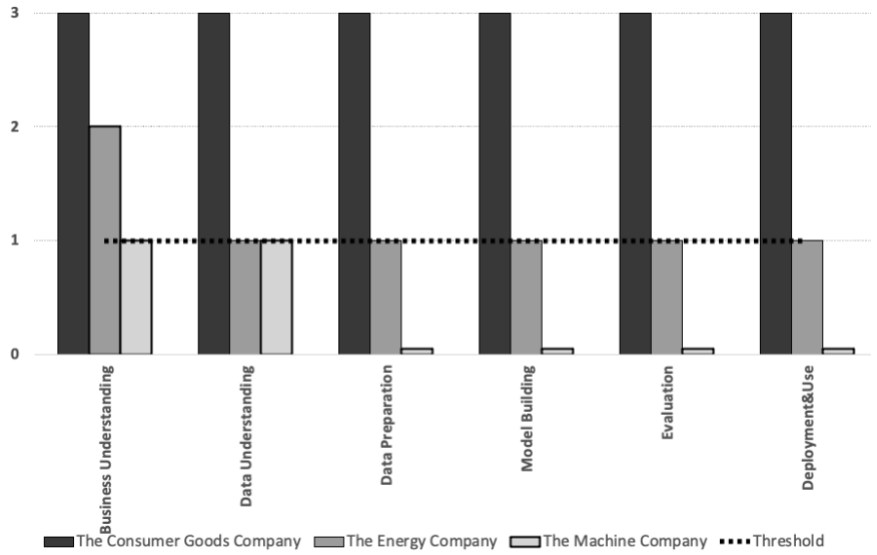

Figure 5-2. ML - 1 Assessment

The developed data science process reference model was used for the process capability level assessments. The base practices, outcomes, and purposes of the processes are checked to assess if the processes are performed. As a result of the assessment, it was observed that the companies initiated and implemented base practices of the data science processes: business understanding, data understanding, data preparation, model building, evaluation, and deployment and use. As each ML builds upon the previous level, the assessment team started with the processes of ML – 1. The CLs of the processes for ML – 1 for three companies are presented in Figure 5-2. To achieve organizational data science ML – 1, organizations need to achieve CL – 1 or higher in processes grouped under ML – 1. Both The Energy Company and The Consumer Goods Company satisfy the requirements of organizational ML – 1. However, the Machine Company could not achieve CL –1 in most of the processes. Accordingly, the organizational data science ML of The Machine Company is ML – 0. Consequently, higher levels are only evaluated for The Consumer Goods Company and The Energy Company.

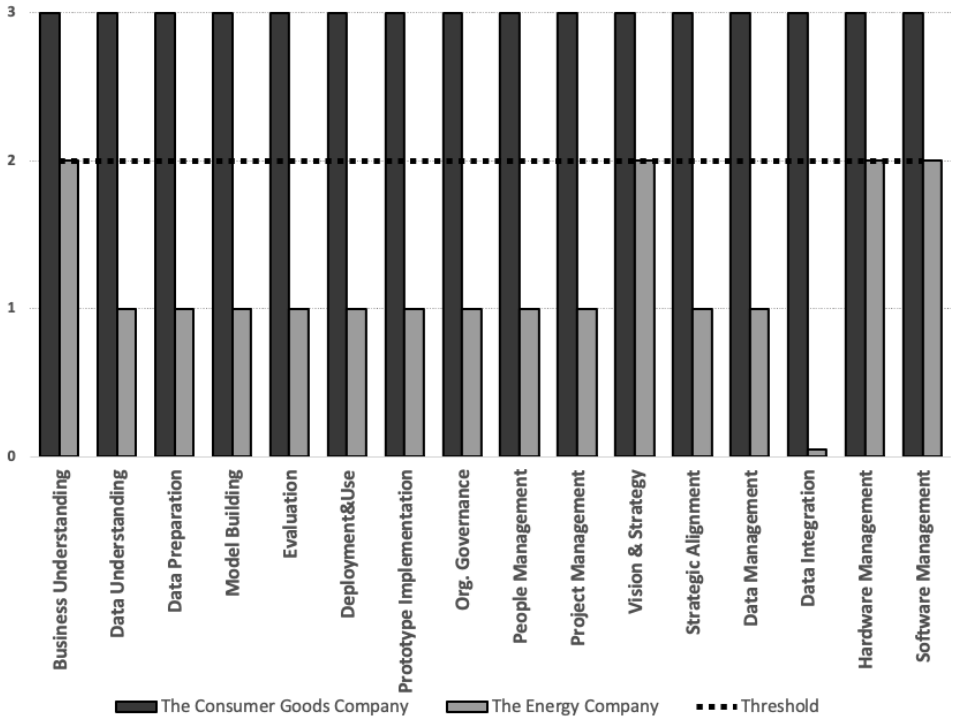

Figure 5-3. ML - 2 Assessment

Then, the data science processes assigned to ML – 2 were assessed in the remaining two companies. The CLs for the processes of The Consumer Goods Company and The Energy Company are delineated in Figure 5-3. In ML – 2, organizations need to achieve at least CL – 2 in all of the data science processes defined in ML – 1 and ML – 2. As the assessment results indicate, The Energy Company could not achieve CL – 2 in most of the processes. Thus, The Energy Company could not fulfill the ML – 2 requirements; accordingly, the organizational data science ML of the organization is determined as ML – 1. This is mainly because The Energy Company does not implement data science processes in a managed fashion. It performs these processes in an ad-hoc manner. On the other hand, The Consumer Goods Company satisfies the requirements of ML – 2 by achieving CL – 3 in all of the processes grouped under ML - 2.

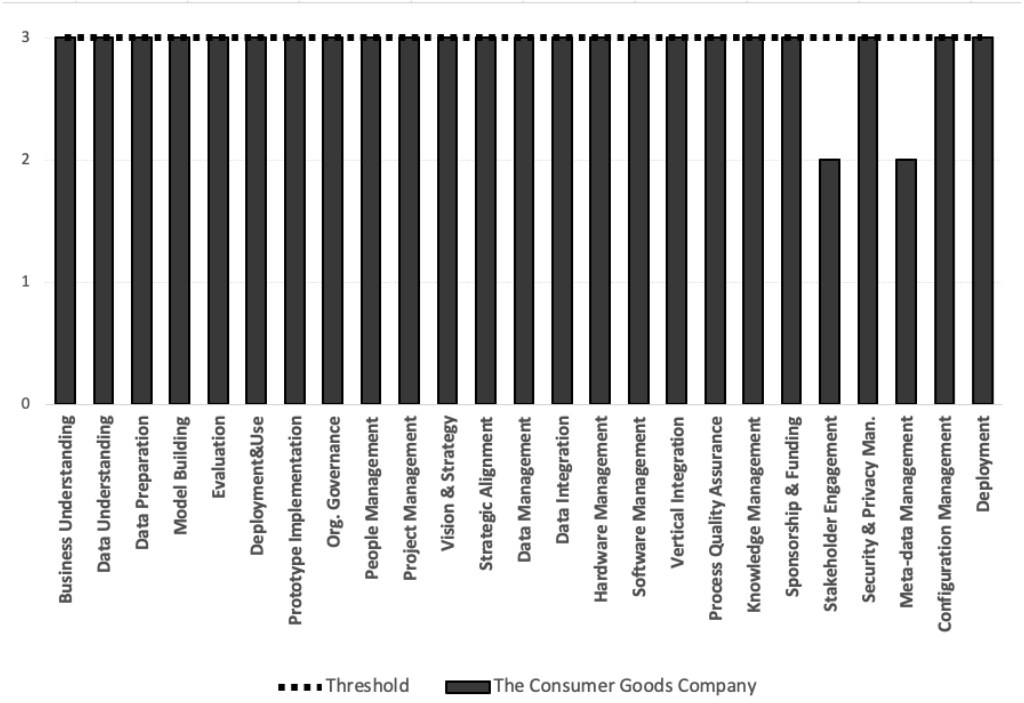

Figure 5-4. ML - 3 Assessment

The assessment team assessed organizational ML – 3 of The Consumer Goods Company. The Energy Company is not included in this assessment as it could not fulfill the requirements of ML – 2, and it is determined as ML – 1. In Figure 5-4, the assessment results of ML – 3 are detailed. In ML - 3, all processes assigned to ML – 1, 2, and 3 should be determined at least CL – 3. The Consumer Goods Company achieves CL – 3 in most of the data science processes. However, stakeholder engagement and meta-data management processes are determined as CL – 2. As the Consumer Goods Company does not satisfy the requirements of ML – 3, it is finally determined as organizational ML – 2.

5.3 Discussion

The Energy, Consumer Goods, and Machine companies have different process capabilities and organizational data science MLs, as summarized in Figure 5-5. The assessment team prepared and shared a detailed assessment report for each company as presented in the technical report [87]. These reports included their data science process capability levels, organizational data science MLs, and roadmaps that

comprise suggestions and recommendations to support continuous improvements in each process capability and their organizational data science maturities. In order to move to higher MLs, these organizations should give a high priority to the improvement of the processes that are denoted as black boxes and a medium priority to the improvement of the gray boxed processes in Appendix C.1, C.2, and C.3. In addition, the rest of the processes that are at CL – 3 are already performed in a standardized way, and the organizations may focus on these processes after achieving higher capability levels at the black and gray boxed processes.

The Energy Company has an initiative for data science, and it understands the importance of data science in its competitive business environment. However, The Energy Company performs basic data science processes in an ad-hoc manner, and these processes are poorly controlled, unpredictable, and reactive. Another critical issue in the Energy Company is that it has multiple manufacturing plants and collect big data from various smart devices in each plant. However, the integration and management of these data still constitute a significant challenge as they are stored in different data stores. This results in a lack of knowledge transfer among manufacturing plants and processes, leading to low-quality data products and services with missing and partial data. The Energy Company should give priority to improve basic data science processes for capturing, storing, cleaning, and analyzing data. The Energy Company should structurally define these processes, including key activities, responsibilities, interactions, and output work products, to perform them more robustly. Moreover, it also needs to define business success measures and key performance indicators to monitor and assess the effectiveness and achievements of the processes.

On the other hand, The Consumer Goods Company has a well-established and managed process asset library, including data science process definitions, workflows, and output work products. It has a clear business strategy, roadmap, and corresponding investments to vertically integrate data science across the organization to become a data-driven and expand generated business value. It also standardizes definitions and management of data science processes, products, and services. Besides these strengths, some weaknesses need to be taken into account. The Consumer Goods company has a weak relationship with stakeholders to adopt data science products in daily production processes. It should also prioritize horizontally integrating their data science processes and output work products across stakeholders and business partners in the value chain

at the business level to make better-informed business decisions. The Consumer Goods Company should also focus on data governance and meta-data management to operationalize data awareness and data lineage issues for browsing and understanding big data. After achieving these suggestions, it needs to focus on establishing a quantitative performance understanding to measure, monitor, and improve process performances with quantitative measures. This allows The Consumer Goods Company to ensure that the performance of data science processes support the achievement of the organizations' relevant business goals and achieve these goals in a systematically planned and predictable manner. I detailed an example roadmap in Table 5-1 for The Consumer Goods company for short, mid, and long-terms to move its organizational maturity to the next level. The suggestions for improvement for each process are determined according to the defined base practices and generic practices in the Data Science Reference Model. From the staged model perspective, The Consumer Goods company can achieve a next level of maturity in a short term by improving the Stakeholder Engagement and Meta-data Management processes by performing suggestions as listed in Table 5-1. On the other hand, from the continuous model perspective, it may also select a specific process for improvement according to its business or specific project objectives.

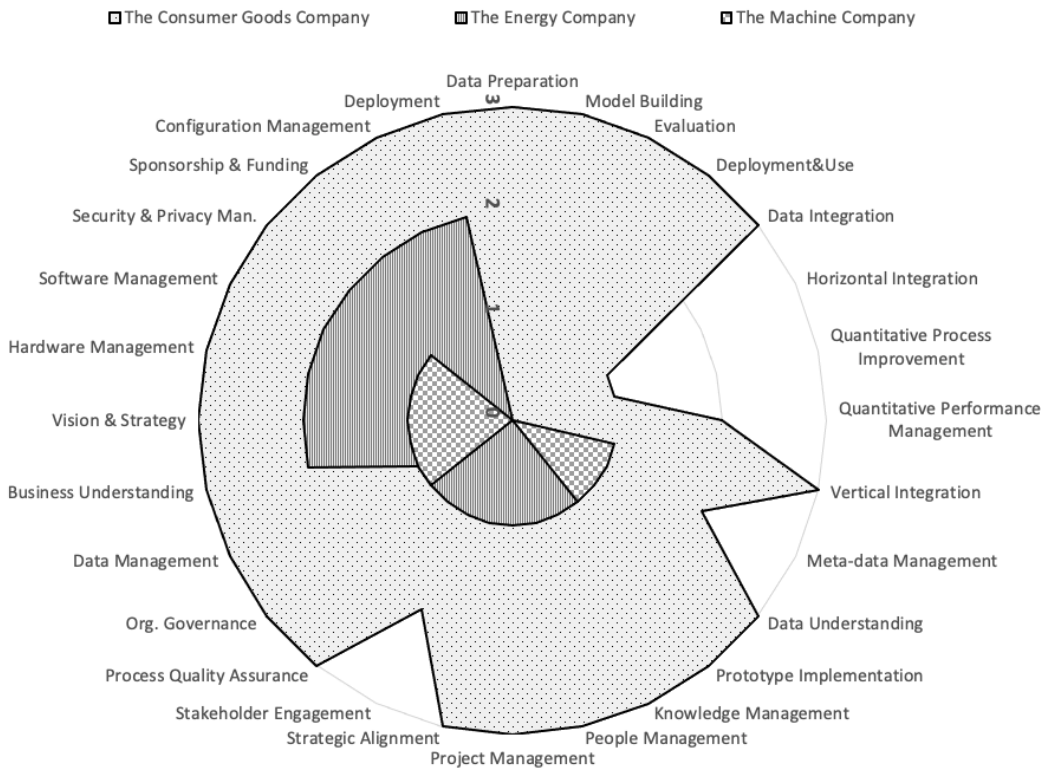

Figure 5-5. Comparisons of Data Science Process Capability Levels of The Cases

The Machine Company recognizes the importance of data science in attaining a competitive advantage. It has a short-term strategy and vision to adopt data science. Accordingly, The Machine Company also has some prototype implementations to determine if it can leverage data science to improve their manufacturing processes in spite of low budget and limited employees. It obtained satisfactory results from these prototype implementations. The Machine Company also has investments in its technological infrastructure to collect, store, and process data by leveraging state-of-the-art data science tools and platforms. Accordingly, The Machine Company is willing to adopt data science processes and improve their existing process capabilities. However, the Machine Company needs to start their data science journey by structurally defining the essential data science processes, including data preparation, model building, evaluation, and deployment and use. Thereafter, The Machine Company should clearly establish its business direction, policies, and leadership systems, and structures to adopt data-driven manufacturing and ensure the

organization shares common goals and objectives. Moreover, it needs to manage and control their data science activities, people, and technical resources to develop data science products and services with high quality, availability, and business value.

Table 5-1. An Example Roadmap for the Consumer Goods Company

Timeline	*Processes*	*Suggestions for Improvement*
Short Term	Stakeholder Engagement	- Establish and maintain a consultation and communication strategy and principle for external and internal stakeholders. - Continually examine stakeholder engagement and communication strategy. - Determine success measures to monitor the effectiveness and suitability of the stakeholder engagement process. - Provide adequate resources, information, and process infrastructure to improve the performance of the stakeholder engagement process.
	Meta-data Management	- Establish a meta-data management framework to support the search-and-retrieval functionality required for all guidance and management of data assets. - Collect and analyze data about the performance of the meta-data management framework to demonstrate its suitability and effectiveness. - Assign roles, responsibilities, and authorities to perform the meta-data management process.
Mid-term	Quantitative Performance Management	- Determine a relevant business goal and a standardized data science process to be addressed by quantitative management. - Determine measurement techniques to collect quantitative data from the standardized process. - Determine performance and conformance targets. - Continually monitor the performance and conformance targets with the collected quantitative measures. - Determine and implement corrective and preventative actions to address the causes of variations. - Periodically report performance of the standardized process against conformance targets.
	Horizontal Integration	- Define the horizontal integration requirements and objectives. - Determine assumptions, constraints, and risks of the horizontal integration. - Identify business units, their processes, and functions to be integrated horizontally. - Determine and prepare related applications, data, and technology to be integrated. - Define responsibilities and authorities to perform horizontal integration.

Long-term	Quantitative Process Improvement	- Identify potential improvement opportunities arising from emerging technologies, tools, platforms, data science concepts, and trends. - Select improvement opportunities based on their relevance and significance according to the business strategy and goals. - Select improvement opportunities by analyzing the costs, benefits, constraints, and risks. - Plan and implement the pilot improvements to collect early feedback about their potential benefits. - Review the results of pilot improvements to determine whether to proceed with organization-wide deployment. - Prioritize and select candidate improvements for deployment based on priorities and available resources. - Continually monitor the deployment of the improvements to take corrective or preventive actions when process improvements fail.

5.3.1 Analysis of DSCMM's Applicability and Usefulness

The data science process capability and organizational maturity assessment results, evidence, and the roadmap for improvement were shared with the participant organizations. Then, the assessment team conducted another interview with six participants from The Consumer Goods Company, three participants from The Energy Company, and four participants from The Machine Company to discuss the assessment results, collect feedback about DSCMM, and answer the research questions raised in Section 5.2.1.

In these interviews, the consistency and validity of the assessment results, the applicability of DSCMM, and the benefits of the provided roadmap are discussed and evaluated. To avoid any bias in responses, the assessment team collected data through semi-structured interviews and asked some open-ended questions rated according to the 5-point Likert scale from 1- Strongly Disagree to 5 - Strongly Agree. These open-ended questions and their median results are as follows;

- Is assessing the organizational data science maturity helpful for your organization? *Median: 4*
- Do you that the provided roadmap for improvements is applicable and helpful? *Median: 4*
- Do you think that implementing suggestions in the roadmap will increase the business value generated by data science? *Median: 4*
- Do you think that the language and terminology in the questions are easy to understand? *Median: 5*

The participants indicated that the assessment results and roadmap to improve data science processes provided in the reports are well prepared, reflecting their organizations' capabilities and gaps from a data science perspective. An executive manager of The Energy Company stated that the assessment results are fully compatible with their own findings and the provided suggestions for improvement include both previously realized and unrealized issues and weaknesses. The participant organizations indicated that the provided suggestions are applicable and helpful to discover their potential for improvement and they will initiate implementing these suggestions to improve the business value generated by data science. They also stated that the language and terminology utilized during the interviews were easy to understand. As a result, according to the feedbacks of the interview participants, the research questions are answered; the proposed DSCMM is applicable and useful with the purpose of identifying the current data science ML of organizations. It also provides a valuable roadmap and suggestions to improve organizational data science maturity.

The validity of DSCMM is also discussed based on a set of criteria defined in the studies [31], [41], [42] for an MM in Table 5-2.

Table 5-2 Evaluation of DSCMM

Criteria	**Analysis of DSCMM**
Purpose and Scope - *Fitness* for Purpose	The proposed model aims to guide organizations in assessing their data science process capabilities, evaluating their organizational data science MLs, revealing their strengths and limitations, performing a gap analysis, and providing an extensive roadmap for continuous improvement in a structured and repeatable way.
Processes - *Completeness of Processes* - *Granularity of Dimensions*	In the development of DSCMM, I first systematically reviewed the existing literature related to data science and employed a survey based-research method to understand practitioners' considerations. Then, I synthesized the finding in the existing literature related to data science and 183 practitioners' considerations into a well-accepted process capability maturity model standard to develop a complete model. Accordingly, DSCMM comprises twenty-eight processes in six process areas: Organization, Strategy

	Management, Data Analytics, Data Governance, Technology Management, and Supporting.
MLs - *Definition of Measurement Attributes* - *Description and objectivity of Assessment Method*	DSCMM defines six MLs from Level 0: Not Performed to Level 5: Innovating. Each MLs include goals to achieve better defined and more consistently implemented data science processes as the organization gains maturity. Data science ML is evaluated by assessing the process capabilities in a staged manner to promote incremental improvements through standardization and optimization. Each process CL comprises at least one well-defined measurement attribute to address a specific strength of the assigned CL to support continuous improvement. I grounded the development of DSCMM on ISO/IEC 330xx standard series for the objective, consistent, and unbiased model development and assessment results. I utilized the ISO/IEC 33004 to develop data science process reference model and MLs. The ISO/IEC 33003 and ISO/IEC 33020 standards were adopted to define capability dimension of DSCMM. The assessment team utilized ISO/IEC 33002 as a guideline to conduct process capability assessments and ISO/IEC 33014 was used to develop a standardized roadmap for process capability and organizational maturity improvements.
Verification and Validation	The applicability, validity, and usefulness of the proposed model are verified through a multiple case study. The model is also verified in a discussion interview with participants from management and technical domain experts, including process owners, stakeholders, data scientists, and executive members.

CHAPTER 6

CONCLUSION

The rapid technological advancements in Industry 4.0 and its enabling technologies, including the Industrial Internet of Things, big data, and Web 2.0, allow organizations to collect a huge volume of complex, real-time, and unstructured big data from a diverse set of data sources. Data science empowers organizations to transform this big data into valuable insights and business actions and realize the data-driven approaches. Thus, organizations can potentially reap significant opportunities by regarding data and data science as their critical strategic assets to strive. However, organizations face difficulties in benchmarking their current data science maturities, and identifying improvement areas and planning change due to a lack of standardized MM with well-defined processes and evolutionary paths. To help organizations assess their data science capabilities and determine improvement opportunities in a systematic way, I grounded the development of DSCMM based on the ISO/IEC 330xx standards family. The applicability and usefulness of the model are demonstrated by a multiple case study conducted in three different organizations.

In this chapter, I summarize the proposed model in Section 6.1, present the contributions and limitations in Section 6.2, and suggest future studies in Section 6.3.

6.1 Summary

There is a growing research interest regarding the application of data science in organizations. This interest is mainly motivated by the availability of a vast amount of potentially useful data from diverse sets of sources and improved technologies to deal with big data efficiently. However, there is limited research on applying the MM concept, determining the data science capability/maturity of an organization, and suggesting a roadmap for improvement. There are MMs proposed by academia and industry to address big data related activities specifically. However, big data mainly

focuses on the technical aspects of data analysis to deal with the unique characteristics of today's big data. While portraying the technical processes of data science, the proposed model, DSCMM, also considers managerial and organizational issues. To truly realize the potential of data science, an organization should focus on building the technical capabilities and bridging the gap between the technical capabilities and firm-specific softer resources, including the right set of managerial and labor skills, data-driven culture, and domain-specific knowledge.

The most important outcome of this thesis is developing a standardized, repeatable, and impartial DSCMM. The proposed model aims to guide organizations in assessing their data science processes, evaluating their organizational data science maturities, revealing their strengths and limitations, performing a gap analysis, and providing an extensive roadmap for continuous improvement. I developed DSCMM on the basis of the well-accepted ISO/IEC 330xx standard series [14] by following a theoretically grounded MM development methodology [23]. As I detailed in Chapter 3, an exploratory survey is conducted to discuss and validate the research gap in the MM for data science in organizations and to define the main constructs of DSCMM as defined in RQ-1 and RQ-2. The exploratory survey results were collected from 183 responders working in the industry. Then, the industry data is combined with scholarly data, as indicated in Hevner's information system research framework. Hence, DSCMM embodies the insights derived from scholarly studies and practitioner remarks to ensure its relevance, validation, and acceptability.

As I detailed in Section 4, DSCMM provides a generic process capability and MM to support continuous improvement from both organizational and process perspectives. It comprises two different representations as continuous and staged. On the one hand, the continuous representation assesses capability levels of each process individually. On the other hand, the staged representation evaluates organizational data science maturity to reveal how an organization standardizes and consistently implements data science processes. DSCMM defines six CLs and six MLs from Level – 0 Incomplete to Level – 5 Innovating based on ISO/IEC 330xx standard series. Moreover, it comprises 28 data science processes which are developed through a literature review and an exploratory survey with experienced practitioners, including data scientists, managers, engineers, and executives as I detailed in Section 3.1.

An iterative development approach is employed in the development of DSCMM. Accordingly, I first developed a preliminary version of DSCMM by defining the Data Analytics process area. Then, the applicability of the ISO/IEC 330xx based MM

development approach has been validated in an emerging data science domain with an exploratory case study to answer RQ-3 as detailed in Section 3.2. After that, the applicability and usefulness of the final version of DSCMM have been validated by multiple case study conducted in three different companies of various sizes as presented in Section 5. These companies operate in different industries in different countries, and they exhibit different data science adoption levels.

6.2 Contributions and Limitations

The main contributions of this study are (1) to provide a review of the available MMs from a specific data science perspective, (2) to close the research gap through the theoretically grounded, methodologically rigorous development of a holistic MM in the data science domain based on a well-accepted ISO/IEC 330xx standard, (3) to demonstrate the applicability and validity of DSCMM with a multiple case study research approach in organizations at different sizes and sectors. The assessment team conducted case studies in various settings to ensure the generalizability of the results. The results of these case studies indicate that DSCMM is applicable in assessing the organizational data science maturity of organizations; identifying organizations' strengths and weaknesses from the data science perspective; performing a gap analysis; developing a data science roadmap to improve each data science process capability and organizational data science maturity to support spur development of data-driven paradigm. The proposed model and findings of this research contribute to the growing body of knowledge on how organizations can leverage data science in an effective and efficient manner to strengthen their data-driven endeavors.

As a result of these observations, it may be possible to discover new processes and define new capability and maturity levels according to the organizations' necessities and advancements in the field. DSCMM includes generic practices instead of suggesting a specific technology. Nevertheless, as the data science concept is still evolving, the maturity-level requirements should be checked regularly to ensure that these levels are still reflecting the current trends and technological state of the art.

6.3 Future Work

In future studies, I am planning to conduct additional case studies in different organizations across different sectors to evaluate the generalizability of DSCMM. As a long-run project, I intend to observe the benefits and challenges of DSCMM in

organizational performance and data science capabilities. I also plan to implement a self-assessment software tool to ease maturity evaluation, allowing firms to assess and determine their own data science process capabilities and organizational data science ML.

DSCMM provides a generic and standardized model for all organizations, businesses and organizational sizes. However, I am also planning to customize DSCMM for different business domains, organizations, and organization sizes by extending the processes, process attributes, and CLs according to the specific business and organizational needs and objectives. I would like to interact with the ISO community and present DSCMM to transform it into an ISO standard for the data science community. DSCMM includes generic data science practices instead of suggesting a specific technology, framework, or methodology. Nevertheless, as the data science concept is still evolving, the maturity-level requirements should be checked regularly to ensure that these levels are still reflecting the current trends and technological state of the art. To this end, it may be possible to discover new processes and define new capability and maturity levels according to the organizations' necessities and advancements in the field.

REFERENCES

[1] D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," *International Journal of Information Management*, vol. 36, no. 5. Pergamon, pp. 700–710, Oct. 01, 2016, doi: 10.1016/j.ijinfomgt.2016.04.013.

[2] Gartner, "Our Top Data and Analytics Predicts for 2019," 2019. https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/ (accessed Nov. 19, 2020).

[3] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment," *Br. J. Manag.*, vol. 30, no. 2, pp. 272–298, 2019, doi: 10.1111/1467-8551.12343.

[4] C. Gröger, "Building an Industry 4.0 Analytics Platform," *Datenbank-Spektrum*, vol. 18, no. 1, pp. 5–14, 2018, doi: 10.1007/s13222-018-0273-1.

[5] F. Provost and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.," 2013.

[6] M. H. ur Rehman, I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, and C. Perera, "The role of big data analytics in industrial Internet of Things," *Futur. Gener. Comput. Syst.*, vol. 99, pp. 247–259, 2019, doi: 10.1016/j.future.2019.04.020.

[7] S. Mithas, A. Tafti, and W. Mitchell, "How a Firm's Competitive Environment and Digital Strategic Posture Influence Digital Business Strategy.," *MIS Q.*, vol. 37, no. 2, 2013.

[8] R. Sharma, S. Mithas, and A. Kankanhalli, "Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations," *Eur. J. Inf. Syst.*, vol. 23, no. 4, pp. 433–441, 2014.

[9] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management," *J. Bus. Logist.*, vol. 34, no. 2, pp. 77–84, 2013, doi: 10.1111/jbl.12010.

[10] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and

Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013, doi: 10.1089/big.2013.1508.

[11] M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V Weber, "Capability maturity model, version 1.1," *IEEE Softw.*, vol. 10, no. 4, pp. 18–27, 1993, doi: 10.1109/52.219617.

[12] P. Fraser, J. Moultrie, and M. Gregory, "The use of maturity models/grids as a tool in assessing product development capability," in *Engineering Management Conference, 2002. IEMC'02. 2002 IEEE International*, 2002, vol. 1, pp. 244–249.

[13] C. P. Team, "CMMI for Development, version 1.3," 2010. https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=9661 (accessed Oct. 12, 2020).

[14] ISO/IEC, "ISO/IEC 33001:2015 Information technology -- Process assessment -- Concepts and terminology," 2015. https://www.iso.org/standard/54175.html (accessed May 12, 2019).

[15] A. Dorling, "SPICE: Software process improvement and capability dEtermination," *Inf. Softw. Technol.*, vol. 35, no. 6–7, pp. 404–406, 1993, doi: 10.1016/0950-5849(93)90011-Q.

[16] S. I. G. Automotive, "Automotive SPICE process assessment model," *Final Release, v4*, vol. 4, p. 46, 2010.

[17] F. Mc Caffery and A. Dorling, "Medi SPICE development," in *Journal of Software Maintenance and Evolution*, 2010, vol. 22, no. 4, pp. 255–268, doi: 10.1002/spip.439.

[18] A. L. Mesquida, A. Mas, E. Amengual, and J. A. Calvo-Manzano, "IT Service Management Process Improvement based on ISO/IEC 15504: A systematic review," *Inf. Softw. Technol.*, vol. 54, no. 3, pp. 239–247, 2012.

[19] A. Tarhan, O. Turetken, and H. A. Reijers, "Business process maturity models: A systematic literature review," *Inf. Softw. Technol.*, vol. 75, pp. 122–134, 2016, doi: 10.1016/j.infsof.2016.01.010.

[20] T. M. Choi, S. W. Wallace, and Y. Wang, "Big Data Analytics in Operations Management," *Prod. Oper. Manag.*, vol. 27, no. 10, pp. 1868–1883, 2018, doi: 10.1111/poms.12838.

[21] C. Y. Lee and C. F. Chien, "Pitfalls and protocols of data science in

manufacturing practice," *J. Intell. Manuf.*, pp. 1–19, 2020, doi: 10.1007/s10845-020-01711-w.

[22] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[23] J. Becker, R. Knackstedt, and J. Pöppelbuß, "Developing Maturity Models for IT Management," *Business&Information Syst. Eng.*, vol. 1, no. 3, pp. 213–222, 2009, doi: 10.1007/s12599-009-0044-5.

[24] M. O. Gökalp, K. Kayabay, E. Gökalp, A. Koçyiğit, and P. E. Eren, "Towards a Model Based Process Assessment for Data Analytics: An Exploratory Case Study," in *Communications in Computer and Information Science*, 2020, vol. 1251 CCIS, pp. 617–628, doi: 10.1007/978-3-030-56441-4_46.

[25] M. O. Gökalp, E. Gökalp, K. Kayabay, A. Koçyiğit, and P. E. Eren, "The development of the data science capability maturity model: a survey-based research," *Online Inf. Rev.*, 2021, doi: 10.1108/OIR-10-2020-0469.

[26] M. O. Gökalp, E. Gökalp, K. Kayabay, A. Koçyiğit, and P. E. Eren, "Data-driven manufacturing: An assessment model for data science maturity," *J. Manuf. Syst.*, vol. 60, pp. 527–546, 2021, doi: 10.1016/j.jmsy.2021.07.011.

[27] T. Mettler, "Thinking in terms of design decisions when developing maturity models," *Int. J. Strateg. Decis. Sci.*, vol. 1, no. 4, pp. 76–87, 2010.

[28] C. M. Olszak and M. Mach-Król, "A conceptual framework for assessing an organization's readiness to adopt big data," *Sustain.*, vol. 10, no. 10, p. 3734, 2018, doi: 10.3390/su10103734.

[29] M. O. Gökalp, K. Kayabay, E. Gökalp, A. Koçyiğit, and P. E. Eren, "Assessment of process capabilities in transition to a data-driven organisation: A multidisciplinary approach," *IET Softw.*, 2021, doi: 10.1049/sfw2.12033.

[30] K. M. Hüner, M. Ofner, and B. Otto, "Towards a maturity model for corporate data quality management," in *Proceedings of the ACM Symposium on Applied Computing*, 2009, pp. 231–238, doi: 10.1145/1529282.1529334.

[31] J. Poeppelbuss, B. Niehaves, A. Simons, and J. Becker, "Maturity Models in Information Systems Research: Literature Search and Analysis," *Commun. Assoc. Inf. Syst.*, vol. 29, no. 1, p. 27, 2011, doi: 10.17705/1cais.02927.

[32] I. Benbasat, A. S. Dexter, D. H. Drury, and R. C. Goldstein, "A Critque of the

Stage Hypothesis: Theory and Empirical Evidence," *Commun. ACM*, vol. 27, no. 5, pp. 476–485, 1984, doi: 10.1145/358189.358076.

[33] J. L. King and K. L. Kraemer, "Evolution and Organizational Information Systems: An Assessment of Nolan's Stage Model," *Commun. ACM*, vol. 27, no. 5, pp. 466–475, 1984, doi: 10.1145/358189.358074.

[34] V. Isoherranen, M. K. Karkkainen, and P. Kess, "Operational excellence driven by process maturity reviews: A case study of the ABB corporation," in *IEEE International Conference on Industrial Engineering and Engineering Management*, 2016, vol. 2016-Janua, pp. 1372–1376, doi: 10.1109/IEEM.2015.7385872.

[35] S. Mittal, M. A. Khan, D. Romero, and T. Wuest, "A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs)," *J. Manuf. Syst.*, vol. 49, pp. 194–214, 2018, doi: https://doi.org/10.1016/j.jmsy.2018.10.005.

[36] ISO/IEC, "ISO/IEC 33020:2015 - Information technology - Process assessment - Process measurement framework for assessment of process capability," 2015. https://www.iso.org/standard/54195.html (accessed Dec. 10, 2020).

[37] M. O. Gökalp, A. Koçyiğit, and P. E. Eren, "A visual programming framework for distributed Internet of Things centric complex event processing," *Comput. Electr. Eng.*, vol. 74, pp. 581–604, 2019, doi: 10.1016/j.compeleceng.2018.02.007.

[38] A. McAfee, E. Brynjolfsson, and T. H. Davenport, "Big data: the management revolution," *Harv. Bus. Rev.*, vol. 90, no. 10, pp. 60–68, 2012.

[39] Economic Graph Team, "LinkedIn's 2017 U.S. Emerging Jobs Report," 2017. Accessed: Dec. 25, 2018. [Online]. Available: https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report.

[40] J. Manyika, M. Chui, and R. Joshi, "Modeling the global economic impact of AI," *McKinsey*, 2018. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact- (accessed Dec. 25, 2019).

[41] B. Rout, T., Tuffley, A., & Cahill, "CMMI Evaluation: Capability Maturity Model Integration Mapping to ISO/IEC 15504 2: 1998," *Softw. Qual. Institute, Griffith Univ.*, 2001.

[42] A. M. Maier, J. Moultrie, and P. J. Clarkson, “Assessing organizational capabilities: Reviewing and guiding the development of maturity grids,” *IEEE Trans. Eng. Manag.*, vol. 59, no. 1, pp. 138–159, 2012, doi: 10.1109/TEM.2010.2077289.

[43] S. Coleman, R. Göb, G. Manco, A. Pievatolo, X. Tort-Martorell, and M. S. Reis, “How Can SMEs Benefit from Big Data? Challenges and a Path Forward,” *Qual. Reliab. Eng. Int.*, vol. 32, no. 6, pp. 2151–2164, 2016, doi: 10.1002/qre.2008.

[44] M. Comuzzi and A. Patel, “How organisations leverage: Big Data: A maturity model,” *Ind. Manag. Data Syst.*, vol. 116, no. 8, pp. 1468–1492, 2016, doi: 10.1108/IMDS-12-2015-0495.

[45] H. Sulaiman, Z. C. Cob, and N. Ali, “Big data maturity model for Malaysian zakat institutions to embark on big data initiatives,” in *2015 4th International Conference on Software Engineering and Computer Systems, ICSECS 2015: Virtuous Software Solutions for Big Data*, 2015, pp. 61–66, doi: 10.1109/ICSECS.2015.7333084.

[46] T. Lukman, R. Hackney, A. Popovič, J. Jaklič, and Z. Irani, “Business intelligence maturity: The economic transitional context within Slovenia,” *Inf. Syst. Manag.*, vol. 28, no. 3, pp. 211–222, 2011, doi: 10.1080/10580530.2011.585583.

[47] R. Cosic, G. Shanks, and S. Maynard, “Towards a business analytics capability maturity model,” in *ACIS 2012 : Proceedings of the 23rd Australasian Conference on Information Systems*, 2012, pp. 1–11.

[48] D. Raber, R. Winter, and F. Wortmann, “Using quantitative analyses to construct a capability maturity model for Business Intelligence,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2012, pp. 4219–4228, doi: 10.1109/HICSS.2012.630.

[49] S. Lavalle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big Data, Analytics and the Path From Insights to Value,” *MIT Sloan Manag. Rev.*, vol. 52, no. 2, 2011, doi: 10.0000/PMID57750728.

[50] R. L. Grossman, “A framework for evaluating the analytic maturity of an organization,” *Int. J. Inf. Manage.*, vol. 38, no. 1, pp. 45–51, 2018, doi: 10.1016/j.ijinfomgt.2017.08.005.

[51] I. Hausladen and M. Schosser, “Towards a maturity model for big data analytics in airline network planning,” *J. Air Transp. Manag.*, vol. 82, p. 101721, 2020,

doi: 10.1016/j.jairtraman.2019.101721.

[52] J. Lismont, J. Vanthienen, B. Baesens, and W. Lemahieu, “Defining analytics maturity indicators: A survey approach,” *Int. J. Inf. Manage.*, vol. 37, no. 3, pp. 114–124, Jun. 2017, doi: 10.1016/j.ijinfomgt.2016.12.003.

[53] J. Ge, F. Wang, H. Sun, L. Fu, and M. Sun, “Research on the maturity of big data management capability of intelligent manufacturing enterprise,” *Syst. Res. Behav. Sci.*, vol. 37, no. 4, pp. 646–662, 2020, doi: 10.1002/sres.2707.

[54] M. A. Maasouman and K. Demirli, “Development of a lean maturity model for operational level planning,” *Int. J. Adv. Manuf. Technol.*, vol. 83, no. 5–8, pp. 1171–1188, 2016, doi: 10.1007/s00170-015-7513-4.

[55] P. O’Donovan, K. Bruton, and D. T. J. O’Sullivan, “IAMM: A maturity model for measuring industrial analytics capabilities in large-scale manufacturing facilities,” *Int. J. Progn. Heal. Manag.*, vol. 7, no. 32, pp. 1–11, 2016, doi: 10.36001/ijphm.2016.v7i4.2466.

[56] I. Marcovecchio, M. Thinyane, E. Estevez, and P. Fillottrani, “Capability maturity models towards improved quality of the sustainable development goals indicators data,” in *Proceedings of the 2017 ITU Kaleidoscope Academic Conference: Challenges for a Data-Driven Society, ITU K 2017*, 2017, vol. 2018-Janua, pp. 1–8, doi: 10.23919/ITU-WT.2017.8246989.

[57] C. Weber, J. Königsberger, L. Kassner, and B. Mitschang, “M2DDM–a maturity model for data-driven manufacturing,” *Procedia CIRP*, vol. 63, pp. 173–178, 2017.

[58] L. Canetta, A. Barni, and E. Montini, “Development of a Digitalization Maturity Model for the Manufacturing Sector,” in *2018 IEEE International Conference on Engineering, Technology and Innovation, ICE/ITMC 2018 - Proceedings*, 2018, pp. 1–7, doi: 10.1109/ICE.2018.8436292.

[59] B. F. Halper and D. Stodder, “A Guide to Achieving Big Data Analytics Maturity (TDWI Benchmark Guide),” 2016. https://tdwi.org/whitepapers/2018/01/aa-all-ms-a-guide-to-achieving-big-data-analytics-maturity.aspx (accessed Aug. 11, 2021).

[60] J. Radcliffe, “Leverage a Big Data Maturity Model to Build Your Big Data Roadmap,” *Radcliffe Advisory Services Ltd*, 2014. http://www.radcliffeadvisory.com/research/download.php?file=RAS_BD_MatMod.pdf (accessed Jan. 06, 2020).

[61] V. Dhanuka, “Hortonworks Big Data Maturity Model,” *Hortonworks*, 2016. http://hortonworks.com/wp-content/uploads/2016/04/Hortonworks-Big-Data-Maturity-Assessment.pdf (accessed Aug. 10, 2021).

[62] B. El-Darwiche, V. Koch, D. Meer, D. W. Tohme, A. Deckert, and R. T. Shehadi, “Big Data Maturity. An action plan for policymakers and executives,” 2014.

[63] M. Hornick, “Oracle Data Science Maturity Model,” 2018. https://blogs.oracle.com/r/data-science-maturity-model-summary-table-for-enterprise-assessment-part-12 (accessed Aug. 10, 2021).

[64] T. H. Davenport and J. G. Harris, *Competing on analytics: The new science of winning*. Harvard Business Press, 2007.

[65] Gartner, “How to Accelerate Analytics Adoption When Business Intelligence Maturity Is Low.” https://www.gartner.com/en/documents/3671218 (accessed Oct. 09, 2021).

[66] Infotech, “Big Data Maturity Assessment Tool,” 2014. https://www.infotech.com/research/ss/leverage-big-data-by-starting-small/it-big-data-maturity-assessment-tool (accessed May 06, 2020).

[67] Knowledgent, “Big Data Maturity Assessment,” 2014. https://bigdatamaturity.knowledgent.com/ (accessed May 06, 2019).

[68] IDC, “Big Data Maturity,” 2013. http://csc.bigdatamaturity.com/ (accessed May 06, 2019).

[69] ISO/IEC, “ISO/IEC 33004: Information technology - Process assessment - Requirements for process reference, process assessment and maturity models,” 2015. https://www.iso.org/standard/54178.html (accessed Dec. 10, 2020).

[70] D. Straub, M.-C. Boudreau, and D. Gefen, “Validation guidelines for IS positivist research,” *Commun. Assoc. Inf. Syst.*, vol. 13, no. 1, p. 24, 2004.

[71] R. E. HAIR JR, J.F.; BLACK, W.C.; BABIN, B.J; ANDERSON, *Análise Multivariada de Dados*. Bookman Editora, 2006.

[72] R. K. Yin, “Case study research: Design and methods (applied social research methods),” *London Singapore Sage*, 2009.

[73] ISO/IEC, “ISO/IEC 33003:2015, Information technology — Process assessment — Requirements for process measurement frameworks,” 2015.

https://www.iso.org/standard/54177.html (accessed Dec. 10, 2020).

[74] ISO/IEC, "ISO/IEC 33002:2015 Information technology — Process assessment — Requirements for performing process assessment," 2015. https://www.iso.org/standard/54176.html (accessed Dec. 10, 2020).

[75] ISO/IEC, "ISO/IEC 33014:2015 Information technology — Process assessment — Guide for process improvement," 2015. https://www.iso.org/standard/54186.html (accessed Dec. 10, 2020).

[76] M. O. Gökalp, "Data Science Process Reference Model Technical Report METU/II-TR-2021-176," 2021.

[77] D. R. Sjödin, V. Parida, M. Leksell, and A. Petrovic, "Smart Factory Implementation and Process Innovation: A Preliminary Maturity Model for Leveraging Digitalization in ManufacturingMoving to smart factories presents specific challenges that can be addressed through a structured approach focused on people, p," *Res. Technol. Manag.*, vol. 61, no. 5, pp. 22–31, 2018, doi: 10.1080/08956308.2018.1471277.

[78] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Data Sci. Big Data*, vol. 1, no. 1, pp. 51–59, 2013, doi: 10.1089/big.2013.1508.

[79] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.

[80] SAS, "SEMMA," 2017. https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en (accessed Nov. 25, 2020).

[81] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework.," *Int Conf Knowl. Discov. Data Min.*, vol. 96, pp. 82–88, 1996.

[82] M. O. Gökalp *et al.*, "Open-Source Big Data Analytics Architecture for Businesses," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Nov. 2019, pp. 1–6, doi: 10.1109/UBMYK48245.2019.8965572.

[83] M. Mosley, M. H. Brackett, S. Earley, and D. Henderson, *DAMA guide to the data management body of knowledge*. Technics Publications, 2010.

[84] D. Cetindamar, R. Phaal, and D. Probert, *Technology management: activities and tools*. Macmillan International Higher Education, 2016.

[85] D. Gürdür, J. El-khoury, and M. Törngren, “Digitalizing Swedish industry: What is next?: Data analytics readiness assessment of Swedish industry, according to survey results,” *Comput. Ind.*, vol. 105, pp. 153–163, 2019, doi: 10.1016/j.compind.2018.12.011.

[86] R. K. Yin, *Case study research and applications: Design and methods*. Sage publications, 2017.

[87] M. O. Gökalp, “Data Science Capability Maturity Model Case Studies Technical Report METU/II-TR-2021-175,” 2021.

APPENDICES

Appendix – A

Table 7-1. Overall Exploratory Survey Results

Process Areas	Processes	Mean	S.D.	Median	Percentiles			Rank
					25%	50%	75%	
Organization	Organizational Governance	3.934	1.189	4	3	4	5	22
	People Management	4.219	0.998	5	4	5	5	13
	Project Management	4.257	0.867	4	4	4	5	11
	Communication	4.175	1.182	5	4	5	5	14
	Quality Assurance	4.152	0.943	4	4	4	5	16
	Knowledge Management	4.142	0.921	4	4	5	5	18
	Risk Management	3.847	1.152	4	3	4	5	25
	Performance Man.	3.628	1.268	4	3	4	5	30
Strategy Management	Vision & Strategy	4.262	1.062	5	4	5	5	10
	Sponsorship & Funding	4.284	0.893	5	4	5	5	9
	Strategic Alignment	3.940	1.187	4	3	4	5	21
	Stakeholder Engagement	4.120	1.078	4	4	4	5	19
	Innovation Management	3.891	1.048	4	3	4	5	24
Data Analytics	Business Understanding	4.596	0.696	5	4	5	5	3
	Data Collection	4.656	0.599	5	4	5	5	2
	Data Storage	4.153	0.824	4	4	4	5	15
	Data Preparation	4.481	0.733	5	4	5	5	6
	Data Analysis	4.770	0.494	5	5	5	5	1
	Data Understanding	4.491	0.686	5	4	5	5	5
	Model Deployment	3.934	0.953	4	3	4	5	23
	Data Visualization	4.256	0.874	5	4	5	5	12
Data Governance	Security & Privacy Man.	4.344	0.869	5	4	5	5	8
	Meta-data Management	4.066	0.959	4	4	4	5	20
	Data Quality Management	4.492	0.777	5	4	5	5	4
	Data Availability & Access	4.148	0.969	4	4	4	5	17
	Data Life-cycle Man.	3.770	1.065	4	3	4	5	28
	Arch. Approach and Stand	3.568	1.150	4	3	4	5	32
	Master-Data Man.	3.645	1.084	4	3	4	5	29
	Data Integration	4.426	0.751	5	4	5	5	7
	Historical Data Man.	3.820	1.014	4	3	4	5	27
Technology Management	Hardware Management	3.393	1.167	3	3	3	4	34
	Software Management	3.831	0.977	4	3	4	5	26
	Configuration Management	3.475	1.123	3	3	3	4	33
	Deployment	3.607	1.133	4	3	4	5	31

Appendix – B

Table 7-2. Business Understanding Process Definition

Process Reference Model	**Process ID**	DA.1
	Process Name	Business Understanding
	Process Purpose	The purpose of the Business Understanding process is to understand the business needs and requirements from a data point of view to formulate data science projects to achieve the objectives.
	Process Outcomes	As a result of successful implementation of this process: 1) A gap analysis is conducted, and organizational issues are defined and formulated as a data science problem; 2) Objectives, scope, and business success criteria are defined; 3) Requirements, limitations, and assumptions are defined; 4) Project risks are analyzed and identified; 5) An action plan for each defined risk is prepared; 6) A business glossary document is published; 7) A sketch project plan is prepared; 8) Cost and benefit analysis is conducted according to the draft project plan; A document that includes all deliverables and steps in business understanding is published and shared with all related stakeholders.
Process Performance Indicators	**Base Practices (BPs)**	**DA.1.BP1: Conduct a Gap Analysis and Define Needs:** Conduct a gap analysis in your organization and define needs and formulate problems according to reach business strategy and vision. [Outcome: 1] **DA.1.BP2: Identify Objectives, Scope, and Business Success Criteria.** Define objectives, scope, and business success criteria for planning data products and services. [Outcome: 2] **NOTE 1:** The statement of business objectives, scope, and business success criteria is baselined and controlled using the practices of the Strategy & Vision Management. **DA.1.BP3: Define Requirements, Assumptions, and Constraints.** Identify all organizational and industrial assumptions, constraints for the data science project. Define sources of risks, areas of impact, and events that might cause to fail the project. [Outcome: 3] **DA.1.BP4: Identify Risks and Prepare Action Plans.** Identify and analyze risks to the project. Define an action plan for each identified risk. [Outcomes: 4, 5] **DA.1.BP5: Develop a Business Glossary.** Define a terminology glossary relevant to industrial terms and background information about data science projects to enable common understanding of the core business concepts and terminology and to reduce the risk that data will be misused due to inconsistent understanding of the business concepts [Outcome: 6] **DA.1.BP6: Prepare a Draft Project Plan**. Establish and maintain a draft project plan to analyze the project needs better and to estimate required IT and Human resources and budget to be needed throughout the project. [Outcome: 7] **NOTE 2:** This may include: • Estimating project cost,

		• Preparing a draft project time schedule, • Identify human and IT resource requirements, • The sketch of a work breakdown structure, • Assigning responsibilities. **NOTE 3:** The draft project plan may be utilized in Project Management process. **DA.1.BP7: Analyze Cost and Benefits.** Analyze cost and benefits of the project according to the draft project plan. Discuss the project benefits towards costs with all related stakeholders. [Outcome: 8] **DA.1.BP8: Publish a Business Understanding Document.** Prepare and publish a document that includes all steps, deliverables, and decisions in business understanding processes. [Outcome: 9]
	Output Work Products	Gap Analysis Results → [Outcome: 1] Requirement Specification → [Outcome: 2, 3] Risk Management Plan → [Outcomes: 4, 5] Business Glossary → [Outcome: 6] Schedule → [Outcome: 7] Work Breakdown Structure → [Outcome: 7] Cost and Benefit Analysis → [Outcome: 8] Business Understanding Document → [Outcome: 9]

Appendix – C

Table 7-3. Assessment Results of The Energy Company

Processes \ PAs		Level - 1	Level – 2		Level – 3		Process Capability Level
		PA 1.1	PA 2.1	PA 2.2	PA 3.1	PA 3.2	
ML – 1	Business Understanding	F	L	L	P	P	Level 2
	Data Understanding	F	P	P	-	-	Level 1
	Data Preparation	F	P	P	-	-	Level 1
	Model Building	L	N	P	-	-	Level 1
	Evaluation	L	N	N	-	-	Level 1
	Deployment and Use	L	N	N	-	-	Level 1
ML – 2	Prototype Implementation	L	-	-	-	-	Level 1
	Org. Governance	L	-	-	-	-	Level 1
	People Management	F	P	P	-	-	Level 1
	Project Management	F	P	L	-	-	Level 1
	Vision & Strategy	F	L	L	-	-	Level 2
	Strategic Alignment	L	-	-	-	-	Level 1
	Data Management	L	-	-	-	-	Level 1
	Data Integration	P	-	-	-	-	Level 0
	Hardware Management	F	L	L	-	-	Level 2
	Software Management	F	L	L	-	-	Level 2
ML – 3	Vertical Integration	N	-	-	-	-	Level 0
	Process Quality Assurance	F	P	P	-	-	Level 1
	Knowledge Management	P	-	-	-	-	Level 0
	Sponsorship & Funding	F	P	P	-	-	Level 1
	Stakeholder Engagement	P	-	-	-	-	Level 0
	Security & Privacy Man.	F	L	L	-	-	Level 2
	Meta-data Management	P	-	-	-	-	Level 0
	Configuration Manag.	F	L	L	N	N	Level 2
	Deployment	F	L	L	P	N	Level 2
ML – 4	Horizontal Integration	N	-	-	-	-	Level 0
	Quant. Performance Man.	N	-	-	-	-	Level 0
ML – 5	Quant. Process Impro.	N	-	-	-	-	Level 0

Table 7-4. Assessment Results of the Consumer Goods Company

Processes	PAs	Level - 1	Level – 2		Level – 3		Process Capability Level
		PA 1.1	PA 2.1	PA 2.2	PA 3.1	PA 3.2	
ML – 1	Business Understanding	F	F	F	L	L	Level 3
	Data Understanding	F	F	F	L	L	Level 3
	Data Preparation	F	F	F	L	L	Level 3
	Model Building	F	F	F	L	L	Level 3
	Evaluation	F	F	F	L	L	Level 3
	Deployment and Use	F	F	F	L	L	Level 3
ML – 2	Prototype Implementation	F	F	F	L	L	Level 3
	Org. Governance	F	F	F	F	L	Level 3
	People Management	F	F	F	L	L	Level 3
	Project Management	F	F	F	L	F	Level 3
	Vision & Strategy	F	F	F	L	L	Level 3
	Strategic Alignment	F	F	F	L	L	Level 3
	Data Management	F	F	F	L	L	Level 3
	Data Integration	F	F	F	L	L	Level 3
	Hardware Management	F	F	F	F	L	Level 3
	Software Management	F	F	F	F	L	Level 3
ML – 3	Vertical Integration	F	F	F	L	L	Level 3
	Process Quality Assurance	F	F	F	F	L	Level 3
	Knowledge Management	F	F	F	L	L	Level 3
	Sponsorship & Funding	F	F	F	L	L	Level 3
	Stakeholder Engagement	F	L	L	-	-	Level 2
	Security & Privacy Man.	F	F	F	L	L	Level 3
	Meta-data Management	F	L	L	-	-	Level 2
	Configuration Management	F	F	F	F	L	Level 3
	Deployment	F	F	F	F	L	Level 3
ML – 4	Horizontal Integration	F	L	P	-	-	Level 1
	Quant. Performance Man.	F	F	L	-	-	Level 2
ML – 5	Quant. Process Impro.	F	P	N	-	-	Level 1

Table 7-5. Assessment Results of The Machine Company

Processes	PAs	Level - 1	Level – 2		Level – 3		Process Capability Level
		PA 1.1	PA 2.1	PA 2.2	PA 3.1	PA 3.2	
ML – 1	Business Understanding	L	N	N	-	-	Level 1
	Data Understanding	L	N	N	-	-	Level 1
	Data Preparation	P	-	-	-	-	Level 0
	Model Building	P	-	-	-	-	Level 0
	Evaluation	P	-	-	-	-	Level 0
	Deployment and Use	P	-	-	-	-	Level 0
ML – 2	Prototype Implementation	L	P	N	-	-	Level 1
	Organizational Governance	L	P	P	-	-	Level 1
	People Management	P	-	-	-	-	Level 0
	Project Management	P	-	-	-	-	Level 0
	Vision & Strategy	L	P	N	-	-	Level 1
	Strategic Alignment	P	-	-	-	-	Level 0
	Data Management	L	N	N	-	-	Level 1
	Data Integration	N	-	-	-	-	Level 0
	Hardware Management	F	L	N	-	-	Level 1
	Software Management	L	L	P	-	-	Level 1
ML – 3	Vertical Integration	L	N	N	-	-	Level 1
	Process Quality Assurance	N	-	-	-	-	Level 0
	Knowledge Management	N	-	-	-	-	Level 0
	Sponsorship & Funding	N	-	-	-	-	Level 0
	Stakeholder Engagement	N	-	-	-	-	Level 0
	Security & Privacy Man.	L	P	P	-	-	Level 1
	Meta-data Management	L	P	P	-	-	Level 1
	Configuration Management	N	-	-	-	-	Level 0
	Deployment	N	-	-	-	-	Level 0
ML – 4	Horizontal Integration	N	-	-	-	-	Level 0
	Quant. Performance Man.	N	-	-	-	-	Level 0
ML – 5	Quant. Process Impro.	N	-	-	-	-	Level 0

CURRICULUM VITAE

Mert Onuralp Gökalp

E-mail: merttgokalp@gmail.com

Date of Birth: 25/06/1990

Education

Middle East Technical University, Ankara Turkey

Doctor of Philosophy in Information Systems

2015-2021

Thesis Title: Data Science Capability Maturity Model

Supervisor: Assoc. Prof. Dr. Altan Koçyiğit

Middle East Technical University, Ankara Turkey

Master of Science in Information Systems

2013-2015

Thesis Title: A cloud-based architecture for distributed real time processing of continuous queries

Supervisor: Assoc. Prof. Dr. Altan Koçyiğit

Yeditepe University, İstanbul, Turkey

Bachelor in Computer Science and Engineering

2008-2013

Areas of Interest

Data Science, Big Data, Machine Learning, Artificial Intelligence, Data Analytics, Internet of Things, Cloud Computing, Data-driven Organizations, Process Improvement, Process Capability and Maturity Models, Industry 4.0, Digital Transformation

Work Experience

10/2014 – Present

Research Assistant – Middle East Technical University, Ankara, Turkey

Teaching assistant in Big Data, Object-Oriented Analysis and Design, Software Architecture, and Software Management courses.

Worked as a researcher to better understand how data science shapes organizations and their management culture. My research interest lies in the field of data science, big data, machine learning, and their applications in the industry. My researches utilize an interdisciplinary approach to investigate applications and adoptions of these emerging fields in the industry to extend their organizational benefits.

Projects:

- Data Science Capability Maturity Model
- Data-driven Manufacturing: An Assessment Model for Data Science Maturity
- Assessment of Process Capabilities in Transition to a Data-Driven Organization: A Multidisciplinary Approach
- Data Science Roadmapping: An Architectural Framework Helping Organizations Become Data-Driven
- A visual big data tool for distributed IoT centric complex event processing
- Analysis of Big Data and Business Analytics Approaches for Enterprises and Solution Proposal
- A Process Modeling and Management Framework for Industry 4.0
- Big Data for Industry 4.0
- Predictive Maintenance in Healthcare Services with Big Data Technologies

Visiting Researcher – University of Cambridge, Cambridge, United Kingdom

Investigated open-source big data tools and developing a big data analytics architecture for businesses.

06/2017 – 09/2017

Developed a Machine Learning (ML) application to classify customer experience for the Caterpillar company by leveraging Natural Language Processing (NLP) techniques and state-of-the-art big data technologies.

The research outputs are the following publications;

- Big-Data Analytics Architecture for Businesses: a comprehensive review on new open-source big-data tools
- Open-Source Big Data Analytics Architecture for Business

06/2012 – 09/2012

Intern – Turkcell Technology, Research and Development Centre, Istanbul, Turkey

Worked as a Java Developer in Collection Development Team to develop J2EE applications for the call centers by leveraging Oracle DBMS, Java EE, Spring, PrimeFaces, MVC, and Hibernate technologies and tools.

Academic Publications

Indexed Journals

- **Gökalp, M. O.,** Gökalp, E., Kayabay, K., Koçyiğit, A., & Eren, P. E. (2021). Data-driven manufacturing: An assessment model for data science maturity. Journal of Manufacturing Systems (SCI-E, Q1), 60, 527-546.
- **Gökalp, M. O.,** Gökalp, E., Kayabay, Gökalp, S., K., Koçyiğit, A., & Eren, P. E. A Process Assessment Model for Big Data Analytics, Computer Standards and Interfaces (SCI-E, Q2)
- **Gökalp, M. O.,** Gökalp, E., Kayabay, K., Koçyiğit, A., & Eren, P. E. (2021). Data Science Maturity Assessment: Development of Data Science Capability Maturity Model. Online Information Review (SSCI-Q2)
- Kayabay, K., **Gökalp, M. O.,** Gökalp, E., Eren, P. E., & Koçyigit, A. (2021) Data Science Roadmapping: An Architectural Framework Helping

Organizations Become Data-Driven. Technological Forecasting & Social Change (SSCI, Q1)

- Gökalp, E., **Gökalp, M. O.,** & Çoban, S. (2020). Blockchain-Based Supply Chain Management: Understanding the Determinants of Adoption in the Context of Organizations. Information Systems Management (SCI-E, Q3), 1-22.
- **Gökalp, M. O.,** Koçyiğit, A., & Eren, P. E. (2019). A visual programming framework for distributed Internet of Things centric complex event processing. Computers & Electrical Engineering (SCI-E, Q2), *74*, 581-604.

ULAKBIM Indexed Journals

- Gökalp, E., **Gökalp, M. O.,** Çoban, S., & Eren, P. E. (2019) *Efficient Employment Management Under the Effect of Digital Transformation: Proposing a Roadmap*- Verimlilik Dergisi, (3), 201-222. Retrieved from http://dergipark.org.tr/verimlilik/issue/45968/498526
- Gökalp, E., **Gökalp, M. O.** & Eren, P. E. (2019) "Industry 4.0 Revolution in The Apparel Sector: Smart Apparel Factory" AJIT-e: Online Academic Journal of Information Technology, 2019; 10(37):73-96
- Çoban, S., **Gökalp, M. O.,** Gökalp, E., & Eren, P. E. (2017). "Cloud Computing Based Predictive Maintenance Framework for Medical Imaging Devices". Yönetim Bilişim Sistemleri Dergisi, 3(2),76-90.

Book Chapters

- **Gökalp, M. O.**, Akyol, M. A., Koçyiğit, A., & Eren, P. E. (2018). Big data in mHealth. In Current and Emerging mHealth Technologies (pp. 241-256). Springer, Cham.
- Gökalp, E., Kayabay, K., & **Gökalp, M. O.** (2021). Leveraging Digital Transformation Technologies to Tackle COVID-19: Proposing a Privacy-First Holistic Framework. Emerging Technologies During the Era of COVID-19 Pandemic, 348, 149.
- Gökalp E., **Gökalp, M. O.,** Eren, P. E. (2018) "Analyzing the Impact of a Decision Support System on a Medium Sized Apparel Company" Exploring Intelligent Decision Support Systems Current State and New Trend-Springer International Publishing.https://link.springer.com/chapter/10.1007/978-3-319-74002-7_4
- Gökalp, E., **Gökalp, M. O.,** & Eren, P. E. (2018). Industry 4.0 revolution in clothing and apparel factories: apparel 4.0. Industry 4.0 From the Management Information Systems Perspectives, 169-184.

International Conferences

- Kayabay, K., **Gökalp, M. O.,** Gökalp, E., Eren, P. E., & Koçyigit, A. (2021, March). Data Science Roadmapping: Towards an Architectural Framework. In 2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD) (pp. 1-6).
- **Gökalp, M. O.**, Gökalp, E., Koçyiğit, A., & Eren, P. E. (2020, September). Towards a Model Based Process Assessment for Data Analytics: An Exploratory Case Study. In European Conference on Software Process Improvement (pp. 617-628). Springer, Cham.
- **Gökalp, M. O.,** Kayabay, K., Zaki, M., Koçyiğit, A., Eren, P. E., & Neely, A. (2019, November). Open-source big data analytics architecture for businesses. In 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-6). IEEE.
- Çaldağ, M. T., **Gökalp, M. O.,** & Gökalp, E. (2019, November). Open Government Data: Analysing Benefits and Challenges. In 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-6). IEEE.
- Gökalp, E., Çoban, S., & **Gökalp, M. O.** (2019, November). Acceptance of blockchain based supply chain management system: Research model proposal. 2019 1st International Informatics and Software Engineering Conference (UBMYK). IEEE.
- Gökalp, E., **Gökalp, M. O.,** Çoban, S., & Eren, P. E. (2018, November). Dijital Dönüşümün İstihdama Etkisi: Mesleki Açıdan Fırsatlar ve Tehditler. 5th International Management Information Systems Conference.
- Gökalp, E., **Gökalp, M. O.,** Çoban, S., & Eren, P. E. (2018, September). Analysing opportunities and challenges of integrated blockchain technologies in healthcare. In Eurosymposium on systems analysis and design (pp. 174-183). Springer, Cham.
- Çoban, S., **Gökalp, M. O.,** Gökalp, E., Eren, P. E., & Koçyiğit, A. (2018, November). [WiP] Predictive maintenance in healthcare services with big data technologies. In 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA) (pp. 93-98). IEEE.
- Kayabay, K., **Gökalp, M. O.,** Eren, P. E., & Koçyiğit, A. (2018, November). [WiP] A Workflow and Cloud Based Service-Oriented Architecture for Distributed Manufacturing in Industry 4.0 Context. In 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA) (pp. 88-92). IEEE.
- **Gökalp, M. O.,** Kayabay, K., Çoban, S., Yandık, Y. B., & Eren, P. E. (2018, November). Büyük Veri Çağında İşletmelerde Veri Bilimi. In 5th International Management Information Systems Conference (pp. 94-97).

- Kayabay, K., Akyol, M. A., **Gökalp, M. O.,** Koçyiğit, A., & Eren, P. E. (2018). Current research topics in industry 4.0 and an analysis of prominent frameworks. Industry 4.0 from the MIS Perspective.
- Akyol, M. A., **Gökalp, M. O.,** Kayabay, K., Eren, P. E., & Koçyiğit, A. (2017, September). A Context-Aware Notification Architecture Based on Distributed Focused Crawling in the Big Data Era. In European, Mediterranean, and Middle Eastern Conference on Information Systems (pp. 29-39). Springer, Cham.
- Çoban, S., **Gökalp, M.O.,** Eren, P.E. & Koçyiğit, A. (2017, December) A Predictive Maintenance Architecture for Health Domain in Big-data Era, International Workshop Series on Digital Transformation and Artificial Intelligence (IDITAI) Workshop 1:The Bases of Digital Transformation and Key Players
- **Gökalp, M.O.,** (2017, December). Proposing a Holistic Framework for Smart City Initiatives, 1st International Workshop Series on Digital Transformation and Artificial Intelligence,
- Gökalp, E., **Gökalp, M. O.,** & Eren, P. E. (2019). Hazır giyim ve konfeksiyon sektöründe Endüstri 4.0 devrimi: akıllı konfeksiyon fabrikası,, 4th International Management Information Systems Conference.
- **Gokalp, M. O.,** Kayabay, K., Akyol, M. A., Eren, P. E., & Koçyiğit, A. (2016, December). Big data for industry 4.0: A conceptual framework. In 2016 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 431-434). IEEE.
- Kayabay, K., **Gökalp, M. O.,** Akyol, M. A., Koçyigit, A., & Eren, P. E. (2016). Big Data for Future Enterprises: Current State and Trends. In 3rd International Management Information Systems Conference, Izmir (pp. 298-307).
- Gökalp, E., **Gökalp, M. O**., & Eren, P. E. (2018). Industry 4.0 from the MIS Perspective.
- **Gökalp, M. O.,** Koçyigit, A., & Eren, P. E. (2015, August). A cloud-based architecture for distributed real time processing of continuous queries. In *2015 41st Euromicro Conference on Software Engineering and Advanced Applications* (pp. 459-462). IEEE.

Technical Reports

- **Gökalp, M. O.,** Kayabay, K., Zaki, M., Koçyiğit, A., Eren, P. E., & Neely, A. (2017). Big-Data Analytics Architecture for Businesses: a comprehensive review on new open-source big-data tools. Cambridge Service Alliance: Cambridge, UK.
- **Gökalp, M.O.,** (2021), Data Science Process Reference Model, Technical Report in Information Systems, Middle East Technical University/II-TR176

- **Gökalp, M.O.,** (2021), Data Science Capability Maturity Model: A Multiple Case Study, Technical Report in Information Systems, Middle East Technical University/II-TR175

Thesis

- **Gökalp, M. O. (2015)** *"A cloud-based architecture for distributed real time processing of continuous queries",* M.Sc. Thesis in Information Systems, Middle East Technical University (METU)