Bus Inf Syst Eng
https://doi.org/10.1007/s12599-025-00954-2

RESEARCH PAPER

Beyond the Unicorn? Job Roles in Data Science

Jonas Gunklach · Mario Nadj · Sven Michalczyk · Katharina Jacob · Christoph Gröger · Alexander Mädche

Received: 25 April 2022 / Accepted: 3 April 2025

Abstract As organizations collect ever-increasing amounts of data from more and more disparate sources, the demand for personnel with skills in data science continues to grow. Simultaneously, the field of data science is complex and has evolved over time, making it difficult for organizations to identify what job roles and associated skills they need to conduct data science successfully. This lack of clarity leads to the misconception that one person, the so-called data science unicorn, can do it all. Hence, as one job role alone cannot cover the whole spectrum of data science, this article offers clarity about the heterogeneous nature of job roles and skills required in data science by first conducting a systematic literature review on job roles in data science. Underscoring the notion that data science has become a team sport, we explore the proliferation and diffusion of data science over the past decade, tracing the shift from generalist Data Scientists to a landscape characterized by a variety of specialized roles. In a second step, we draw on 16.348 unique job postings from established online job platforms and extract and characterize nine job roles along their skill sets. Our research offers a comprehensive, data-driven perspective on the roles and skills essential in data science, empowering organizations to effectively staff and conduct data science initiatives to derive value from data and maintain a competitive edge.

Keywords Data science · Job roles · Text mining · Topic modeling · Clustering

Accepted after 3 revisions by Natalia Kliewer.

J. Gunklach (✉) · K. Jacob · A. Mädche
Institute for Information Systems (WIN), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: jonas.gunklach@kit.edu

M. Nadj
Rhine-Ruhr Institute of Information Systems, University of Duisburg-Essen, Essen, Germany

S. Michalczyk · C. Gröger
Robert Bosch GmbH, Stuttgart, Germany

1 Introduction

In today's dynamic and competitive business environment, organizations are increasingly acknowledging the strategic value of their human resources (Debortoli et al. 2014). As organizations gather ever-increasing amounts of data from a multitude of sources and seek to make sense of this data, they are reassessing their staffing strategies to more effectively perform data science tasks (Lismont et al. 2019; Kim et al. 2016; Zhang et al. 2020). In particular, they are trying to rapidly acquire the necessary skills in this field in order to remain competitive and leverage the potential of the collected data (De Mauro et al. 2018). According to Barney (1991), an organization's resources, including its human capital, play a critical role in shaping its ability to create a competitive advantage and drive business performance (Gerhart and Feng 2021). As a result, data science has become an in-demand field due to its potential to address complex problems and provide valuable insights from large and diverse datasets (Virkus and Garoufallou 2020). In this context, there is the misconception of the data science unicorn, a person capable of solving any type of problem, from business understanding to solution deployment (Miller 2019). However, one job role alone cannot cover the entire spectrum of data science. Instead, data science has become a "team sport" (Zhang 2019), and organizations have realized the benefits of hiring employees with complementary, specialized skills. In this line,

Published online: 11 July 2025

studies have also found an increase in the demand for personnel working at the intersection of various data-related disciplines (Alekseeva et al. 2021). Thus, it is important to note the complexity of data science as a discipline and of the skills required to be successful in the field. Its interdisciplinary nature requires a combination of skills, including mathematics and statistics, computer science, and the targeted application domain Mike and Hazzan (2023) "to study data and its environments [...] by following a data-to-knowledge-to-wisdom thinking and methodology" (Cao 2017, 8). Simultaneously, the field of data science is evolving over time. Together, these aspects make it difficult for organizations to identify the job roles and associated skills they need to perform data science tasks successfully.

Realizing the potential of data science requires organizations to focus on monetizing and extracting value from data effectively. In addition to driving operational efficiency and fostering innovation, data holds the key to unlocking new revenue streams and improving customer experiences (Wixom and Ross 2017). By leveraging data science, organizations can uncover hidden insights in their data, enabling them to anticipate market shifts, personalize customer interactions, and innovate products and services (Chen et al. 2012). The data value chain describes the process of transforming raw data into actionable insights to create value. This framework defines the sequential steps of the data lifecycle from data generation to application and usage, emphasizing the importance of diverse data science skills across different stages (Faroukhi et al. 2020). However, extracting the full potential value from data requires robust data governance practices (Otto 2011). Without such governance, data science efforts lack the necessary structure and reliability, diminishing their impact Gröger (2021). Thus, determining the roles in data science and aligning them with the skills required across data science disciplines (Mike and Hazzan 2023), supported by robust data governance, enables organizations to maximize the impact of their data science initiatives, creating sustainable value and competitive advantage.

While previous research (for an overview, see Table 1) has already applied text mining approaches such as bag-of-words in a variety of related fields (e.g., big data, business intelligence-BI) to extract skills and/or job roles from job advertisements (short: ads) (De Mauro et al. 2018; Debortoli et al. 2014), there is a lack of a systematic analysis how roles have developed over time to explore the diffusion of data science and, in particular, to better understand the evolving role of the Data Scientist. Such an analysis could provide valuable insights into historical trends and shifts in the job market, providing a macroscopic view of how the demand for specific skills and roles has changed over time. In addition, while existing research often categorizes skills and roles following a green field bottom-up approach, there is a need for a more structured approach based on established conceptual frameworks such as the data science Venn diagram (Mike and Hazzan 2023) and the data value chain (Faroukhi et al. 2020). This would allow for a more standardized categorization and easier comparison of roles and skills across studies.

Against this backdrop, we offer clarity about the heterogeneous nature of the job roles required in data science by (1) conducting a systematic literature review (SLR) on the diffusion of job roles in data science and (2) providing a state-of-the-art overview by preprocessing and analyzing a sample of 16.348 unique job ads published online on leading job platforms (i.e., Indeed, Monster, Glassdoor, and Stepstone). We thus formulate the following research questions: *(1) How did the job roles in data science diffuse into the scientific discourse? (2) What are important job roles in data science, and what skills make them?* We leverage state-of-the-art topic modeling techniques using word embeddings, as they take into account the context in which words appear and capture the meaning of a word based on its surrounding words in a sentence or document. Specifically, the topics consist of frequent word combinations from the crawled job ads in the form of skills, and job roles consist of a specific distribution of those skills. On this basis, we define and characterize nine important job roles in data science.

Table 1 Overview of related work

Article	Focus	Method	Sample	Portal	Results
De Mauro et al. (2018)	Big data	Topic modeling	2.786	Dice (US)	Job roles
Gottipati et al. (2021)	Data science	NLP with lexicon	2.804	Glassdoor	Job roles
Michalczyk et al. (2021)	Data science	Topic modeling	25.104	Multiple	Job roles
Debortoli et al. (2014)	Big data/BI	SVD	5.657	Not specified	Skills
Murawski and Bick (2017)	Big data	Topic modeling	500	Monster	Skills
Handali et al. (2020)	Ind. analytics	Topic modeling	17.282	Monster	Skills
Almgerbi et al. (2022)	Data analyics	Topic modeling	14.000	Multiple	Skills
Brauner et al. (2023)	AI	Topic modeling	1.159	LinkedIn	Skills

With our study, we contribute to the field theoretically (1) by providing a comprehensive review of the literature on roles in data science and understanding the proliferation of roles in data science over the past 12 years. We trace the evolution from Data Scientists as generalists to the emergence of specialized roles characterized by increased standardization. This shift reflects the maturation of data science as a discipline, where specialized roles support specific stages of the data value chain utilizing the diverse skills required to handle the growing complexity of data science projects. (2) Through our empirically grounded, state-of-the-art conceptualization of job roles and skills, we enrich the data science literature and provide insights for researchers and practitioners navigating the data science landscape. Our conceptualization of roles is theoretically grounded in the data science disciplines, bridging the gap between academic understanding and practical application. By exploring a collection of job ads from the end of 2023, our article contributes to the current understanding of roles and related skills in the rapidly changing landscape of data science (Mike and Hazzan 2023). In addition we contribute methodologically (3) by using state-of-the-art methods (i.e., word embeddings) that have not yet been used in existing work to infer job roles and skills. In contrast to bag-of-words, word embeddings capture contextualized word representations, enabling a deeper understanding of the semantic relationships within textual data (Liu et al. 2015). This approach allows us to uncover deeper insights from the extracted job postings and understand the topics in a more nuanced manner. Moreover, we explain our process transparently so that future research can reuse these methods to define job roles in other contexts. Finally, we contribute practically (4) by mapping employees' skills to their job roles giving existing employees transparency regarding their skill set and areas for development. For example, employees' current skills could be compared with ideal job role skills. Based on the comparison, personalized training programs could be suggested. In addition, our results help to outline the desired skills a new employee should possess in data science to complement the organization (e.g., when the company formulates job descriptions for new hires). Thus, such an understanding can help organizations define strategies and a common language for acquiring and cultivating the right skills to more effectively leverage data science.

The remainder of this article is structured as follows. The next section provides background on data science, as well as related work. Section 3 introduces the methods used, that is, the SLR conducted and the approach to extract skills and roles in data science. Section 4 illustrates the diffusion of roles in data science over the last 12 years, followed by Sect. 5 with an overview of the state-of-the-art skills and roles extracted. Finally, we discuss our results in Sect. 6 and offer a conclusion in Sect. 7.

2 Background

The proliferation of data science reflects the growing recognition of data as a critical organizational asset. However, unlocking the full value of data requires navigating an increasingly complex landscape of roles, tools, and organizational guidelines. Organizations must bridge technical and domain expertise while navigating the data value chain and establishing effective guidelines for data science. This section explores the interdisciplinary nature of data science, the process of transforming data into value, the role of data governance, and related research.

2.1 Data Science

Data science is an interdisciplinary field that involves extracting knowledge from data using techniques, tools, and technologies from various disciplines (Dhar 2013). Specifically, it integrates various skills from (1) mathematics and statistics, (2) computer science, and (3) the targeted application domain (Mike and Hazzan 2023) (see Fig. 1) to collect, organize, process, analyze, and interpret large and complex datasets in order to solve complex problems and make data-driven decisions (Cao 2017). Each of these domains contributes to different stages of the data value chain, which organizes the lifecycle of data from its generation to its application.

The data value chain is a structured process that describes the progression of data through various stages in order to create value (see Fig. 2), beginning with **data generation and acquisition** (Faroukhi et al. 2020). In this initial phase, organizations collect data from sources such as databases, social media, and IoT sensors (Chen et al. 2012). At this point, data is typically raw and unstructured, serving as the foundational input for subsequent processing. To ensure data quality and relevance, data profiling is employed early in the process to analyze the data set efficiently - this involves scanning database tables to collect metadata, and evaluating the data for types, patterns, and completeness (Naumann 2014). Next is **data processing**, which prepares the collected data for meaningful analysis. Real-world data often contains errors, missing values, outliers, and inconsistencies. Thus, data cleaning and preprocessing are required for tasks such as noise reduction, missing values imputation, and converting the data into a format suitable for analysis (Cui et al. 2019). Following preprocessing, the data moves to **data storage and management**, where it is organized and securely stored in scalable infrastructures. This stage involves leveraging

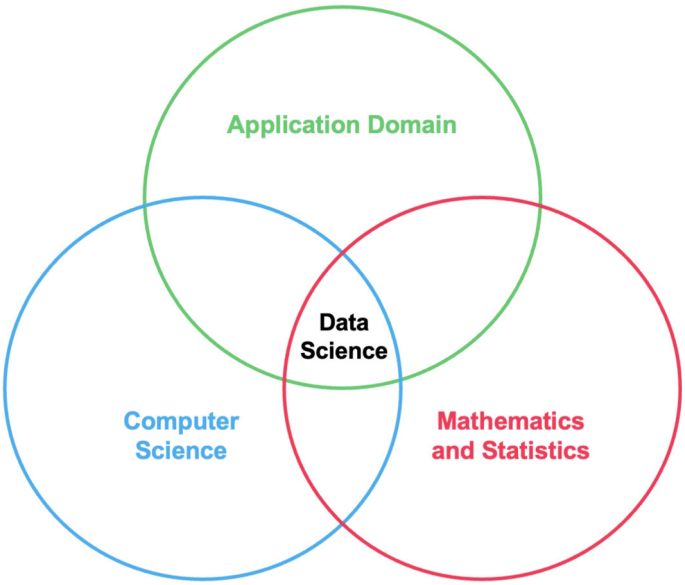


Fig. 1 Data science Venn diagram from Mike and Hazzan (2023)

storage systems, such as data lakes or distributed databases, designed to handle the high volume, variety, and velocity of modern datasets Fan and Geerts (2022). Effective data management ensures that data remains accessible, secure, and easy to retrieve for analysis, forming a solid foundation for downstream activities (Eichler et al. 2021). For **data analysis**, patterns, trends, and relationships are revealed (Krause et al. 2014). This often involves mathematical and statistical techniques. With data visualization tools, an initial data understanding is gained (Gunklach and Nadj 2023). On this basis, feature engineering is performed, the process of selecting, transforming, and creating meaningful features from available data (Krause et al. 2014). Next, various models are trained, for example, models that make predictions or uncover hidden patterns (Bani-Hani et al. 2019). This involves selecting appropriate algorithms, training the models using labeled data, and fine-tuning the model hyperparameters for optimal performance (Brath and Hagerman 2021). Once the models are built, their performance is assessed against the evaluation measures (Gunklach et al. 2024). For instance, how well the model generalizes to new, unseen data. Once insights are generated, they are refined and communicated during the **data visualization** stage. Visualization tools transform complex data into intuitive formats such as graphs, charts, and dashboards, making the information comprehensible to diverse stakeholders (Shi et al. 2021). By highlighting key findings, visualization supports decision-making processes and ensures that insights are actionable. Finally, the data value chain culminates in **application and usage**, where insights are used to address real world challenges (Faroukhi et al. 2020). Whether integrated into business processes, used to optimize operations, or leveraged for product development, this stage ensures the value of the data is realized.

The successful implementation of the data value chain depends on the interplay of three core disciplines (see Fig. 1). Each discipline contributes a unique perspective to the data value chain, supporting multiple stages in distinct ways. Mathematics and statistics provide the theoretical foundation for analyzing and modeling data, enabling techniques such as data profiling, feature engineering, and model evaluation to uncover meaningful patterns and insights. Computer science serves as the backbone of the process, offering the algorithms, infrastructures, and computational power required to manage, preprocess, and analyze vast datasets efficiently. Equally important, the targeted application domain ensures that the data value chain remains grounded in real world contexts, aligning data collection, processing, and application with domain-specific objectives and challenges. Together, these skills enable the data value chain to function cohesively, providing actionable insights and creating value.

Hereby, roles in data science work closely together to maximize value within the data value chain (Gunklach and Nadj 2023). For example, Data Scientists work with Data Engineers to design and implement ML models, while Business Analysts define key performance indicators and metrics for data-driven initiatives (Michalczyk et al. 2021). By fostering collaboration and alignment among these roles, organizations can effectively leverage their data assets to drive innovation, improve operational efficiency, and create value for customers and stakeholders to enhance competitive advantage (Elia et al. 2020; Gerhart and Feng 2021). In addition, successful collaboration of these roles is supported by a variety of tools and technologies, including programming languages like R and Python for data

analysis and machine learning (ML) (Michalczyk et al. 2020). Data visualization tools such as Power BI, Tableau, or Matplotlib enable the creation of interactive visualizations to understand and communicate insights (Gunklach et al. 2023). Moreover, big data frameworks such as Apache Spark and Apache Hadoop are used to handle and process large datasets (Mathis 2017). ML libraries such as TensorFlow, Scikit-learn, and PyTorch provide algorithms and prebuilt models, while database technologies such as NoSQL and SQL databases facilitate efficient data storage, retrieval, and management (Debortoli et al. 2014).

2.2 Data Governance

Without strong data governance, data science efforts would lack the necessary structure and reliability, which diminishing their impact and limiting their potential (Gröger 2021). Data governance serves as the foundation for ensuring data integrity, security, quality, and compliance. Data governance generally refers to organizational structures that treat data as an enterprise asset (Otto 2011). In essence, data governance "specifies decision rights and accountabilities for an organization's decision making about its data" (Abraham et al. 2019, 425). Unlike data management, which focuses on the daily execution of decisions, data governance establishes what decisions need to be made and who is responsible for them (Abraham et al. 2019). Therefore, the central elements of data governance are (1) decision areas (i.e., which decisions about data need to be made), (2) data-related roles (i.e., which roles are involved), and (3) authority (i.e., how are these roles involved in the data value chain) (Otto 2011).

Classical data governance approaches, sometimes called "data governance 1.0" (Legner et al. 2023), focus on operational IT systems and the compliant handling of corresponding data, especially master data (Gröger 2021). Ensuring data security, quality, integrity and privacy are the top priorities with data owners and data stewards as core data governance roles. This rather defensive perspective on data governance was challenged by the rise of data science and the explosion of data sources and led to a more offensive view, also called "data governance 2.0" (Legner et al. 2023). The focus shifts from solely ensuring control and integrity of data to sharing and providing data for a multitude of data science use cases to take advantage of its business value (Fadler and Legner 2022a). Consequently, data governance is a core element of a company's overall data strategy, requiring a balance between data offense and defense in alignment with its business strategy (DalleMule and Davenport 2017). As a result, classical data governance approaches have to be adapted to all data domains beyond master data. Without robust data governance practices, organizations risk compromising the reliability of their data assets, hindering the effectiveness of data science initiatives. Consequently, embracing strong data governance not only facilitates the responsible use of data but also allows organizations to unlock the full potential of their data resources (Fadler and Legner 2022b).

2.3 Related Work

Previous research has applied other text mining approaches in a variety of related fields, such as big data or BI, to extract skills and/or job roles from job ads. In the following section, we provide an overview of this related work by grouping them along (1) the targeted focus area, (2) the extraction method applied, (3) the sample size of the job ads, (4) the job portal used, and (5) the resulting outcomes. First, De Mauro et al. (2018) and Gottipati et al. (2021) focused on extracting job roles. For this purpose, De Mauro et al. (2018) crawled 2.786 job ads from the United States market published on the job portal Dice. They applied topic modeling and focused solely on the big data field. They carved out four job families, with each family comprising four roles: (1) the Business Analyst family (i.e., Project Manager, Business Analyst, Product Manager, and Program Manager), (2) the Data Scientist family (i.e., Data Engineer, Data Scientist, Data Analyst, and Data Consultant), (3) the Developer family (i.e., Software Engineer, Java Developer, Hadoop Developer, and Software Developer), and (4) the Engineering family (i.e., Data Architect, Devops Engineer, Solution Architect, and Systems Engineer). Gottipati et al. (2021), in turn, concentrated on data science in general by relying on 2.804 Glassdoor job ads from specific metropolitan areas (i.e., Singapore, Hong Kong, and London), using natural language processing (NLP) with lexicon creation. They identified three job roles, namely the Data Analyst, Data Engineer, and Data Scientist. In the same year, Michalczyk et al. (2021) performed LDA on 25.104 job ads, highlighting a complex interplay of business, technical, and analytical skills across multiple roles in data science (e.g., Data Analyst, Data Engineer, Data Scientist).

On the contrary, the studies by Handali et al. (2020) and Debortoli et al. (2014) did not define job roles but focused on extracting skills. To this end, Debortoli et al. (2014) compared 4.246 BI and 1.411 big data related job ads by applying a singular value decomposition (SVD) and classified the extracted skills in BI and big data analytics competencies. In addition, Murawski and Bick (2017) analyzed 500 job ads on Monster (with the search term: data analytics + Data Analyst) using LDA and clustered them into technical (e.g., SQL, Excel), business (e.g., management, communication), and system (e.g., development, problem-solving) skills. Next, Handali et al. (2020) crawled 17.282 job ads globally from the job portal

Monster; however, they focused solely on industry analytics by relying on topic modeling. They extracted 17 topics and categorized them according to business (e.g., communication, project management), analytical (e.g., business analysis, ML), and technical (e.g., web development, software development) knowledge domains. Almgerbi et al. (2022) applied topic modeling to 14.000 job ads (from smaller job platforms such as Careerbuilder or SimplyHired) and 3.600 analytics course descriptions, identifying core data analytics skills such as market analysis, BI, and project management. Finally, Brauner et al. (2023) used LDA to study 1.159 AI-related job ads from LinkedIn, revealing the following key skills required in the AI field: data science, AI software development, AI product development and management, AI client servicing, and AI research.

Although, numerous studies have examined the extraction of skills and job roles from data science job postings, there are notable gaps that could enrich our understanding of the evolving landscape of this field. First, despite the number of individual studies focusing on specific roles and skills, there is a lack of a SLR that analyzes these studies over time to explore the diffusion of data science and the evolving role of Data Scientists. Such an analysis could provide valuable insights into historical trends and shifts in the job market and provide a macroscopic view of how the demand for specific skills and roles has changed over time. Second, while existing research often categorizes skills and roles on a bottom-up, empirical basis, there is a clear need for a more structured approach based on established frameworks such as the data science Venn diagram (Mike and Hazzan 2023). This would allow for a more standardized categorization and easier comparison of roles and skills across studies. Finally, while the skills required in data science appear to be evolving relatively slowly, the roles that represent combinations of these skills are more dynamic. Current literature tends to focus predominantly on skill extraction without adequately contrasting these skills with the roles they define, overlooking how the importance and interrelation of specific skills can influence the definition and expectation of job roles within the field. This gap highlights the need for a more nuanced examination that not only delineates individual skills but also explores how they synergize to shape professional roles in data science.

3 Research Method

To address the research questions, we used a dual approach combining a SLR and data-driven analysis of job ads. The SLR provides a theoretical foundation by examining the evolution of data science roles and their diffusion in academic literature. The data-driven analysis complements this by analyzing 16.348 job advertisements using state-of-the-art topic modeling techniques to identify and conceptualize roles and associated skills. This dual approach bridges insights from academic discourse with contemporary trends, resulting in a comprehensive theoretical understanding of data science roles and their associated skills. The following section reflects on these methods in detail.

3.1 Systematic Literature Review

Theorizing and understanding concepts in information systems often requires the use of literature reviews to obtain a holistic, state-of-the-art overview of the concepts at hand (Webster and Watson 2002). Therefore, to identify conceptual work on roles and skills in data science, we conducted a SLR following the guidelines of Boell and Cecez-Kecmanovic (2015). Our initial reference was the influential Harvard Business Review article by Davenport and Patil (2012) which significantly shaped the discourse on the Data Scientist role. Therefore, we employed a dual approach in our SLR (see Fig. 3). Initially, we reviewed and analyzed all 1989 papers citing Davenport and Patil (2012). Subsequently, we used these papers to formulate a search string, applying it to databases such as AISeL, IEEE, ACM DL, and Web of Science, which are recognized in information systems research (Bandara et al. 2015; Haug and Maedche 2021). Our search string comprised two parts: (1) job roles proposed in literature akin to "*Data Scientist*", like "*Data Analyst*", "*Business Analyst*", "*Data Engineer*", "*Data Architect*", "*ML Engineer*", and "*Software Developer*", and (2) terms related to the investigation of roles, including "*role*" along with "*knowledge*", "*skill*", and "*ability*" as it is a commonly used framework in human resources (Stevens and Campion 1994). We also included "*task*" and "*job description*" to ensure that our search was thorough. Finally, we used wildcards to create the final search string:

("Data Scientist" OR "Data Analyst" OR "Business Analyst" OR "Data Engineer" OR "Data Architect" OR "Software Developer" OR "ML Engineer") AND ("Role" OR "Skill*" OR "Abilit*" OR "Knowledge" OR "Task*" OR "Job Description")*

Applying the established search string, we identified 401 studies from IEEE, 926 from WoS, 1.389 from ACM DL, and 1.062 from AISeL (2.207 in total). Our search was not confined to publications prior 2012, the publication year of Davenport & Patil's seminal paper. Including the 1.989 studies citing Davenport and Patil (2012), our initial pool expanded. However, many studies from the database search also referenced Davenport and Patil (2012), leading us to remove duplicates, resulting in 2.782 unique articles. In the

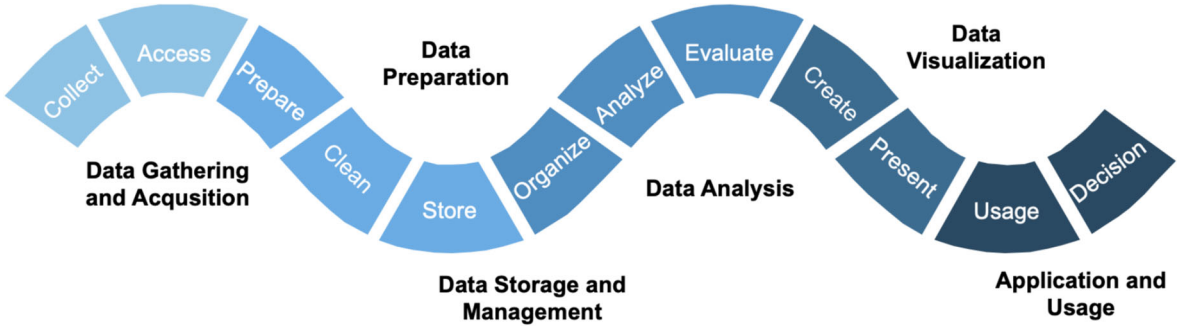


Fig. 2 Data value chain adopted from (Faroukhi et al. 2020; Gunklach and Nadj 2023; Open Data Watch 2018)

selection process, we manually excluded 2.593 articles by carefully scanning the title, abstract, and keyword section and by applying the following selection criteria: We included articles that investigated and described roles in data science. We excluded articles that designed data science curriculums and did not talk about the roles they wanted to target. In addition, we excluded articles that only covered the skills required in data science but did not relate these to roles. On this basis, 189 articles were left. Following the same criteria for a full-text review, 47 relevant articles remained. Lastly, we employed a forward and backward search and included another seven studies (in total 54).

3.2 Role Extraction

In the following, we describe our research process to extract the skills and roles in data science. Our research process entails (1) crawling job ads from four job platforms, (2) processing the job descriptions, (3) extracting topics using word embeddings, (4) clustering job roles in data science, and (5) qualitatively and quantitatively assessing the results.

3.2.1 Crawling

We collected job ads from four platforms, which are among the top ones according to job seekers and recruiters' visits worldwide, and developed our search string in several steps. In particular, we started with an exploratory search using general terms such as *data science*, *business intelligence*, and *big data*. After reviewing the results, we structured our search string into five parts: First, we relied on job roles in data science (e.g., Data Scientist, Data Analyst) as suggested by Debortoli et al. (2014). Second, we utilized general terms related to data science (e.g., big data, AI, BI). Third, following Cao (2017), we included data science terms such as ML, reinforcement learning, supervised learning, and unsupervised learning. Fourth, we added data analysis terms such as data understanding, preparation, and processing following Michalczyk et al. (2020). Finally, we further included BI terms (e.g., data lake, data warehouse, data mart) (Chen et al. 2012). Moreover, we observed that online job boards are limited in their ability to use logical operators. Thus, we combined all search terms separated by commas. If a job platform did not allow for that many search terms, we split the search string into parts and performed multiple searches. The final search string is in the Appendix. In particular, we crawled Indeed (15.990), Glassdoor (8.160), Monster (3.994), and Stepstone (2.253). Since job ads are often published and deleted as soon as a job is filled, we decided to run our crawler twice a week for two months from May to June 2023. We developed our crawler in Python, relying on Beautiful Soup as a crawling framework and ScraperAPI as an endpoint.

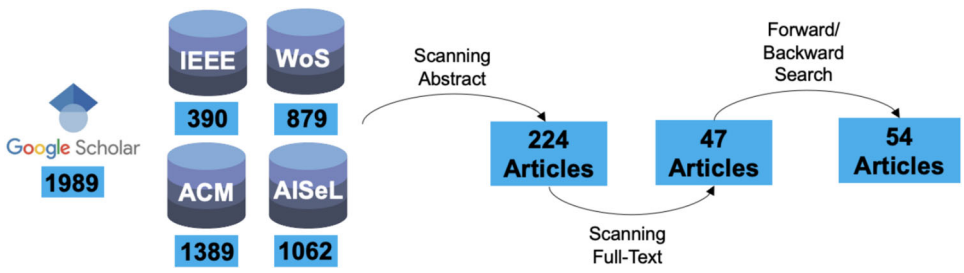


Fig. 3 Search results of the SLR

3.2.2 Processing

Next, we investigated the structure of our 30.397 crawled job ads. Thus, we removed HTML tags from the job description and performed a segmentation into individual sentences (Levy and Goldberg 2014). This approach allowed us to use the resulting sentences as input data for calculating word embeddings and modeling the topics. By splitting the documents into sentences, we further were able to filter out irrelevant information such as company descriptions.

Moreover, we kept English job ads only by using Google's Compact Language Detector (Python library langdetect) because (1) they make up the largest proportion of 26.762 job ads and (2) we require a common language for the word embeddings approach. The document processing is finished by removing duplicates and similar job ads. This step is necessary because we crawled platforms twice a week, which resulted in many duplicates. Thus, we compared exact matches and additionally identified similar job ads by calculating the cosine similarity. Cosine similarity represents the similarity as a vector of inner product space measuring the pairwise angle between job ads (Han et al. 2022). Such a course of action is suitable for our application as it does not consider the length of a job ad but the most common features. We needed to set a threshold above which we consider job ads to be similar for the deduplication. We started with a threshold of 99% and gradually decreased the measure by investigating the similarity of a representative sample of the candidates at each step. To assess similarity, we compared job titles, hiring companies, and, if necessary, the full text of the job ads. We finally defined a threshold of 80% that identifies reliable changes to a job ad irrelevant to our analysis, such as changes to contact persons, locations, or spelling corrections made after job ads have been republished. Eliminating duplicates and similar job ads was facilitated using the Python libraries *nltk* and *sklearn*. Lastly, we concluded the processing phase with the segmentation of each document into sentences, utilizing the Python library *spacy*.

3.2.3 Topic Modeling Using Word Embeddings

After document processing, we were left with 16.348 job ads, from which we extracted 219.966 unique sentences as the input dataset for our topic modeling approach. We developed an iterative process to prune sentences that were not part of the data science field. This process involved generating the topic representation model, identifying and filtering irrelevant topics, and adjusting the neighborhood parameter (i.e., specifying the size or extent of the context window used to capture the surrounding words when generating word embeddings). In each iteration, we progressively narrowed down the neighborhood parameter, increasing the number of smaller topics, and finding more local structures. After six iterations, our sentence dataset consisted of 76.744 sentences.

To extract topics, we relied on BERTopic, a topic modeling technique that takes advantage of the power of BERT (Bidirectional Encoder Representations from Transformers) embeddings for topic extraction (Grootendorst 2022). In contrast to traditional topic modeling techniques that use bag-of-words, BERTopic uses BERT embeddings to capture contextual information, enabling the extraction of more accurate and meaningful topics from job ads. BERTopic has shown particular efficacy when dealing with shorter texts like the segmented sentences at hand.

Based on BERTopic, we followed a three-step process to generate topics. First, we computed the embedding representation for each sentence extracted using a pre-trained language model (all-mpnet-base-v2). Second, we reduced the dimensionality of the resulting embeddings to optimize the subsequent clustering process. Third, we generated topic representations from the document clusters using a custom class-based variation of TF-IDF (C-TF-IDF), assigning each document to a specific topic. C-TF-IDF extends the concept of TF-IDF by incorporating class labels or categories of documents (Grootendorst 2022). Following Röder et al. (2015) and Debortoli et al. (2016), we calculated the semantic coherence for each number of topics and found the highest semantic coherence (0.5040) for 38 unique topics. To construct document vectors, we mapped the relevant sentences from the final sentence set back to the originating documents. We computed the mean of the probabilities calculated by the topic model for each relevant sentence representing the document. This mean computation was weighted by the sentence length to account for the fact that some sentences were longer and contained more information than others. Consequently, we obtained a matrix consisting of 38-dimensional probability vectors for each of the 16.348 job ads.

However, as Chang et al. (2009) demonstrate, metrics-based evaluation of unsupervised learning methods, such as semantic coherence and Silhouette scores, often fails to align with human judgments of topic quality. Although these metrics provide a useful quantitative benchmark to guide optimization, they do not always capture the nuanced interpretability and practical relevance that human evaluators seek. To address these limitations, we incorporated a qualitative validation step. Two researchers and one professional with working experience in data science labeled the 38 topics and classified them (Cohen's kappa 0.9) along (1) mathematics and statistics skills, (2) computer science skills, and (3) application domain skills. Conflicts were resolved by a fourth researcher. The inclusion of domain expertise during this labeling process mitigates the

misalignment often observed between computational evaluations and human judgments, emphasizing the importance of blending automated techniques with human expertise for robust topic modeling (Debortoli et al. 2016). Finally, to ensure that the results were meaningful and aligned with the conclusions drawn from the metrics, we presented the labeled topics and their classifications to two additional data scientists. This final review step helped validate the coherence of the topics and ensured that our conclusions were supported by both human evaluation and domain-specific expertise. By combining automated modeling, rigorous human validation, and external expert review, we aimed to balance computational efficiency with interpretability and practical relevance in our topic modeling approach. This multi-step evaluation process reflects best practices for enhancing interpretability and ensuring the practical relevance of the topics generated (Debortoli et al. 2016).

3.2.4 *Clustering Job Roles*

To define job roles, we used the topic assignment by the topic model as input to a clustering algorithm. Thus, we required a measure of pairwise dissimilarities between job ads, as clustering algorithms typically use the distance between objects to group them. We relied on the JSD because it is a dissimilarity measure between two probabilities distribution (Endres and Schindelin 2003). Our results fulfill this property because the topic model describes each job ad as a probability distribution of topics. The resulting dissimilarity matrix served as input for the clustering. We tried out the density-based clustering algorithm DBSCAN. However, this lead to one big cluster with some job ads labeled as noise. Thus, we decided to use K-Medoids as it is a robust alternative to k-means clustering and less sensitive to noise and outliers because it uses medoids as cluster-centers instead of means (Schubert and Rousseeuw 2019).

3.2.5 *Quantitative and Qualitative Assessment of Clustering Results*

Critical in the application of K-Medoids is to determine the number of clusters that describe the underlying data structure best. To increase objectivity, we followed a two-step assessment of the clustering result, consisting of a quantitative and a qualitative assessment. First, we trained for k = 3...13 models and plotted the average width of clustering silhouettes. This is a well-established approach to evaluate the validity because the silhouette "shows which objects lie well within their cluster" (Rousseeuw 1987, 1). The optimal number of clusters is indicated by a maximum in the curve, in our case, at k = 11 (see Fig. 4). Thus, 16.348 job ads were best separated into nine clusters. To explain these clusters (across 38 topics), we aggregated the topics from the job ads to the clusters' level. In the next step, we qualitatively assessed the clustering results. Two researchers and one professional with working experience in data science labeled the clusters independently as potential job roles. We discarded two clusters as they were only supported by 100 - 200 job ads thus below the average of 1816 job ads per cluster. In summary, we identified a total of nine job roles in the eleven clusters. The job roles are described by a combination of skills along the data

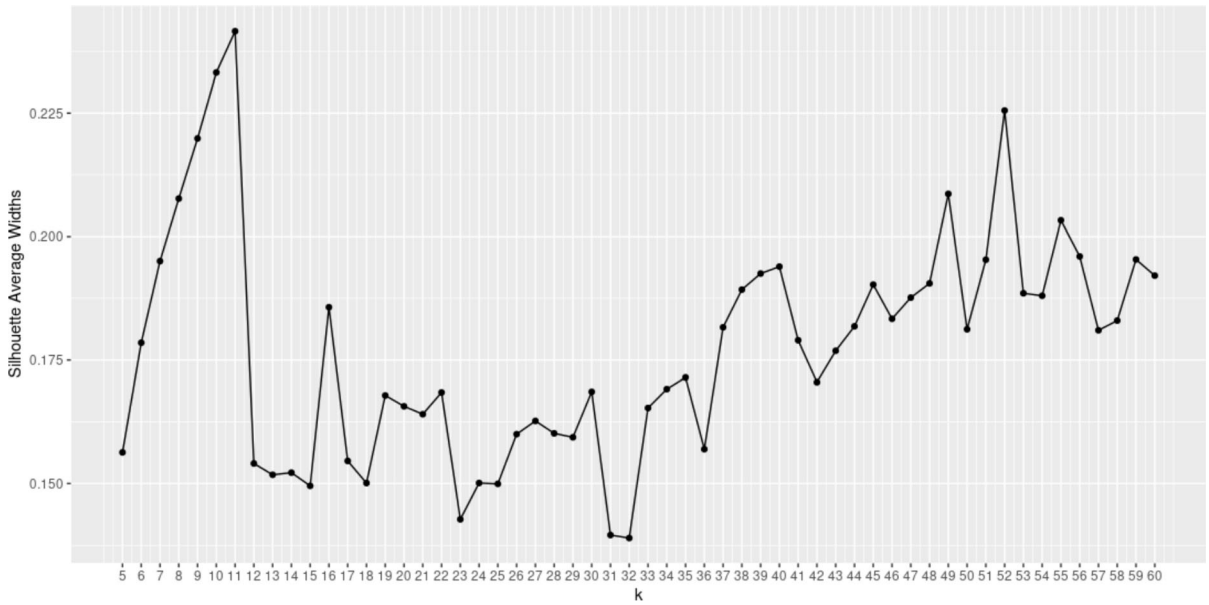


Fig. 4 Silhouette indicating optimal number of clusters

science disciplines (1) mathematics and statistics, (2) computer science, and (3) application domain (Mike and Hazzan 2023). In addition, we developed a 3D space visualization (see Fig. 8 that plots the identified job roles based on the data science disciplines. We calculated the percentage of skills from each domain to transform each job role into a three-dimensional vector, with values ranging from 0 to 1. These values represent the contribution of each domain to the role's skill set. To enhance visual clarity, roles dominated by computer science skills are highlighted in blue, those with a strong emphasis on mathematical and statistical skills are in red, and roles predominantly featuring application domain skills are in green. The results are elaborated on in the following sections.

4 Diffusion of Job Roles in Data Science

Over the past thirteen years, the field of data science has undergone a remarkable evolution, moving from a field dominated by generalists to one characterized by a variety of specialized (Davenport and Patil 2012; Mike and Hazzan 2023). This transformation, which has been documented through research and evolving industry practices, reflects the dynamic interplay between technological advancement and the ever-increasing impact of data on organizations (Chen et al. 2012). In the following, we describe their proliferation over time by outlining critical stages (see also Fig. 5) based on the papers identified in Sect. 3.1. Furthermore, we illustrate how the identified roles ideally contribute to different stages of the data value chain. This is visually represented in Fig. 5, where colored bubbles/dots highlight the alignment of specific roles with the corresponding stages of the data value chain based on the tasks for which they are typically responsible.

The Era of Traditional Analytics. Prior to 2012, the data landscape was dominated by Data Analysts, who were tasked with extracting insights from structured data using descriptive statistics. According to van der Aalst (2014), this foundational role laid the groundwork for the emergence of the Data Scientist. During this period, Data Analysts were expected to handle the entire data lifecycle, from data generation and preprocessing to analysis, visualization, and application. They relied on manual tools such as spreadsheets and SQL, providing descriptive insights and static reports to decision makers.

The Rise of the Data Scientist. The year 2012 marked a seminal moment with Davenport & Patil's declaration of the Data Scientist as the "sexiest job of the 21st century." This proclamation catalyzed a transformative shift, expanding the role of the Data Analyst. Data Scientists emerged as the new pioneers. Initially perceived as generalists, Data Scientists began taking over complex aspects of the data lifecycle, particularly in data preprocessing and analysis, by applying machine learning and advanced statistical techniques. This specialization allowed Data Analysts to focus more on visualization and business-driven decision-making.

Increasing Importance of Decision-Making. By 2013, the role of the Data Scientist had evolved beyond its original characteristics of being a generalist. Provost and Fawcett (2013) and Dhar (2013) highlighted this evolution, emphasizing the growing need for expertise in ML and AI.

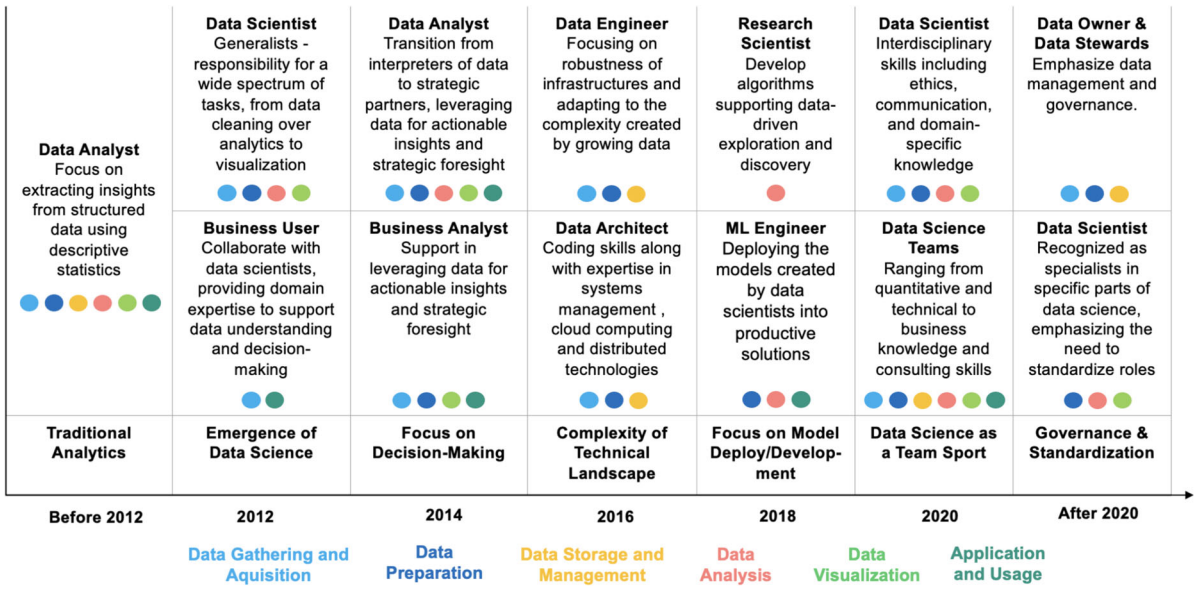


Fig. 5 Role diffusion in data science

During this period, Data Scientists expanded their set of skills to emphasize not only technical proficiency but also a deep understanding of the business context and strategic acumen to improve organizational decision-making (Chen et al. 2012). Viaene (2013) underscored the critical interplay between Data Scientists and Business Users, emphasizing the importance of cross-disciplinary collaboration. As a result, organizations have recognized the need to seamlessly integrate data-driven insights into their strategic frameworks and decision-making processes, elevating the role of Business Analysts in translating data into actionable insights (Abbasi et al. 2016).

Increasing Complexity of the Technical Landscape. The complexity of the technical landscape increased with time, necessitating the rise of the Data Engineer, a role highlighted by Debortoli et al. (2014) for its pivotal role in building the infrastructure that supports the expansive requirements of data science. The Data Engineer distinguished itself by skills in coding, database management, and ETL (De Mauro et al. 2018). De Mauro et al. (2018) identified a new role emerging from the convergence of the Data Engineer and Software Engineer: the Data Developer. Their expertise is primarily in coding, supported by knowledge of systems management, cloud computing, and distributed technologies. Furthermore, Murawski and Bick (2017) observed the emergence of the Data Architect role, originating from the Data Engineer and Data Developer roles, aimed at bridging the gap between Data Professionals and Business Analysts through proficiency in SQL creation, technical database support, and reporting skills.

Focus on Model Deployment and Development. Around 2018, a shift occurred with the distinct delineation of roles focusing on algorithm development and model deployment. Research Scientists primarily developed algorithms while ML Engineers supported the deployment of these models, ensuring that the transition from theory to practice was seamless and effective (Miller 2019). This phase highlighted the necessity for robust deployment pipelines that could handle the complexities of real-world applications, a challenge underscored by Cao (2019). Furthermore, the role of ML Engineers evolved to include the management of production environments, monitoring model performance, and continuously integrating feedback loops to refine algorithms, as discussed by Almgerbi et al. (2022), emphasizing the critical nature of operational stability in model deployment.

Data Science as a Team Sport. Gardiner et al. (2018) observed a widening range of roles in 2018. Lyon and Mattern (2017a) emphasized the growing importance of data storytelling skills, integrating them into the skill set of Business Analysts. Murawski and Bick (2017) introduced the Marketing Analyst role, further expanding the spectrum of roles within the field. Consequently, traditional roles such as Data Analysts and Business Analysts evolved to encompass skills in storytelling and market analysis (Murawski and Bick 2017). In addition to that, Verma et al. (2019a) identified the BI Analyst as having more skills in data engineering compared to the Business Analyst. This period also highlighted the increasing recognition of the interdisciplinary nature of data science education, as evidenced by the various focus areas of data science programs analyzed by Della Volpe and Esposito (2020). Academic institutions responded by adapting curricula to incorporate a wider range of competencies, ranging from technical skills to ethical considerations, reflecting recognition of the multifaceted nature of data science (Fayyad and Hamutcu 2022). Moreover, Cao (2019) and Davenport (2020) elucidated the different paths of Data Scientists and Data Engineers, signaling a trend toward specialization within these roles. Dong and Triche (2020) conducted a longitudinal analysis of job skills for Data Analysts and underscored the evolving technical requirements, highlighting the need for continuous learning and adaptation in the field. Moreover, collaboration among roles in data science became increasingly crucial, reflecting a trend toward seeing data science as a team sport (Zhang et al. 2020).

Governance and Standardization. By 2020, the roles within data science had begun to crystallize, with clear distinctions between Data Scientists, Data Analysts, and Data Engineers (Tamm et al. 2020). Michalczyk et al. (2021) outlined distinct responsibilities for each role, reflecting a maturation of the field. The growing emphasis on data governance introduced roles such as Data Owners and Stewards, signaling a shift toward standardization (Gröger 2021). Mildenberger et al. (2023) observed a maturing job market, increasingly inclined towards specific titles such as Data Scientist and Data Engineer, indicating a consolidation and standardization of roles over the previous decade.

5 Job Roles in Data Science: A State-of-the-Art Overview

This section presents the results of the analysis of 16.348 job advertisements, focusing on the derivation of skills and their subsequent organization into distinct roles. First, we identify key skills using topic modeling techniques, uncovering patterns that reflect the demands of the data science job market. These skills are then used to conceptualize and define distinct roles in data science.

5.1 Identified Skills

Based on Sect. 3.2, we have identified several skills that relate to computer science, mathematics and statistics, or

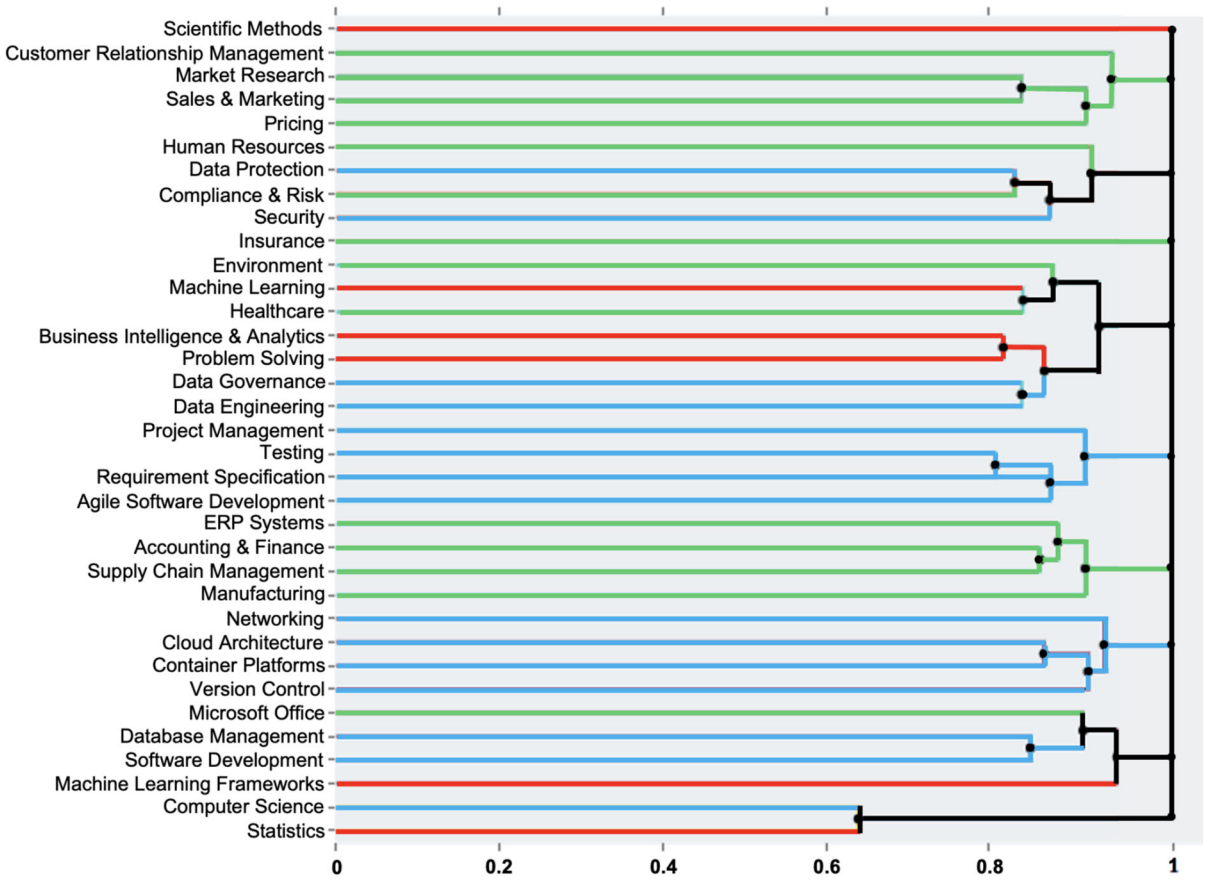


Fig. 6 Hierarchical clustering of skills in data science

represent an application domain. In addition, the dendrogram in Fig. 6 outlines a hierarchical clustering of skills relevant to data science, with each branch representing a cluster of related competencies. These clusters are organized based on their similarity, with closely related skills merging at lower levels on the horizontal axis. The colors of the branches highlight the alignment of these clusters with distinct data science disciplines, emphasizing the multifaceted nature of the field. The blue branches represent technical skills related to computer science, such as programming, software development, and database management. Green branches emphasize application-oriented disciplines, including domain knowledge in business, healthcare, and marketing. The red branches capture mathematics and statistics skills, such as probability, statistical modeling, and optimization (also see Fig. 1). Following the dendrogram, skills from different data science disciplines generally originate from distinct branches, indicating a strong alignment with the conceptual structure of the topic model. However, there are instances where the disciplines converge, reflecting the interconnected nature of data science. For example, the overlap between machine learning and healthcare highlights the integration of predictive analytics and data-driven decision-making into healthcare applications. These overlaps underscore both the strengths and limitations of topic modeling. While clustering algorithms effectively capture semantic and contextual relationships, revealing meaningful interdisciplinary connections, they may not fully distinguish domain-specific nuances without human oversight (Chang et al. 2009). For example, while machine learning in healthcare shares terminology with general-purpose machine learning, their applications differ significantly. To address this, we incorporated a qualitative validation step with two researchers and three data science professionals (see Sect. 3.2.3). This process ensured that the clustering results were not only algorithmically coherent but also meaningful and aligned with human expertise (Debortoli et al. 2016). In the following, we present the skills along (1) computer science, (2) mathematics and statistics, and (3) targeted application domains.

Computer Science Skills. In the discipline of computer science, we have identified 14 topics (see Table 2). The first topic (**CS1**) emphasizes the importance of mastering and obtaining a degree in **computer science**, highlighted by terms such as "*computer science*", "*degree*", and "*science*". **Microsoft Office (CS2)** focuses on proficiency in applications such as *Excel*, *Word*, and *PowerPoint*.

Table 2 Computer science skills

Topic Name	Count	Highest-loading features (C-TF-IDF)
CS1: Computer Science	1971	Computer science (0.35), computer (0.34), science (0.32), degree (0.31), degree computer (0.28), degree computer science (0.28), field (0.27), bachelors (0.27), statistics (0.26), years (0.26)
CS2: Microsoft Office	1274	Excel (0.63), word (0.50), microsoft (0.49), microsoft office (0.48), office (0.47), powerpoint (0.45), ms (0.44), word excel (0.40), ms office (0.39), excel powerpoint (0.37)
CS3: Data Engineering	1121	Etl (0.43), pipelines (0.41), migration (0.39), data pipelines (0.38), data migration (0.35), data (0.30), transformation loading (0.27), etl processes (0.27), extraction transformation (0.26), extraction transformation loading (0.26)
CS4: Data Governance	1094	Data quality (0.53), data governance (0.46), governance (0.43), quality (0.41), data (0.32), integrity (0.27), ensure data (0.27), accuracy (0.25), data quality issues (0.25), quality issues (0.24)
CS5: Database Management	1070	Sql (0.48), database (0.47), databases (0.37), oracle (0.33), relational (0.32), sql server (0.32), server (0.31), mysql (0.29), postgresql (0.27), relational databases (0.26)
CS6: Security	1040	Security (0.55), cyber (0.43), cyber security (0.35), threat (0.35), cybersecurity (0.33), information security (0.27), threats (0.27), vulnerability (0.25), vulnerabilities (0.25), incident (0.22)
CS7: Agile Software Development	983	Agile (0.75), scrum (0.43), working agile (0.38), methodologies (0.36), agile methodologies (0.36), experience agile (0.35), experience working agile (0.35), sprint (0.34), agile environment (0.32), agile development (0.31)
CS8: Software Development	903	Python (0.63), programming (0.44), languages (0.38), javascript (0.37), java (0.36), python developer (0.36), scripting (0.33), experience python (0.33), programming languages (0.32), developer (0.32)
CS9: Cloud Architecture	871	Aws (0.38), cloud (0.37), azure (0.34), spark (0.32), hadoop (0.27), gcp (0.27), kafka (0.26), apache (0.25), experience (0.24), airflow (0.23)
CS10: ERP Systems	659	Sap (0.66), erp (0.38), master data (0.29), master (0.28), experience sap (0.26), hana (0.25), knowledge sap (0.24), erp systems (0.23), s4 (0.23), s4 hana (0.20)
CS11: Networking	644	Cisco (0.50), network (0.44), routing (0.41), networking (0.38), switching (0.38)
CS12: Data Protection	519	Privacy (0.61), data protection (0.58), protection (0.55), data privacy (0.38), compliance (0.35)
CS13: Version Control	311	Git (0.63), devops (0.61), cicd (0.56), version control (0.52), version (0.49)
CS14: Container Platforms	238	Kubernetes (0.81), docker (0.70), terraform (0.49), container (0.48), docker kubernetes (0.46)

CS3 highlights skills in **data engineering**, encompassing *ETL* processes, *data pipelines*, and the management of *data warehouses*. **CS4** focuses on **data governance**, emphasizing *data quality*, *integrity*, and ensuring *data governance*. **CS5** revolves around **database management**, skills in SQL, database management, and relational databases are at the core of this topic, evident from terms such as "*SQL*", "*server*", and "*databases*". "**Security**" (**CS6**) emphasizes skills fighting *cybersecurity*, *threat management*, and *incident response*. **CS7** revolves around **agile software development** methodologies, including *Scrum*, *sprints*, and working in an *agile environment*. **CS8** centers around **software development**, emphasizing skills in *Python*, *scripting*, *languages* such as *Java* and *programming* in general. Next, **CS9** is centered around **cloud architecture**, involving skills in various cloud platforms such as *GCP*, *Azure*, and *AWS*. Following, **CS10** relates to **ERP systems**, particularly *SAP* and its various modules such as *HANA* and *s4*, which highlights the need to master SAP ERP systems. **CS11** delves into **networking**, covering vendors such as *Cisco* and areas such as *switching*, *firewalls*, *routing*, and *protocols*. **Data Protection** (**CS12**) involves skills in accounting for data *privacy*, *compliance*, and ensuring *GDPR*. Further, **CS13** covers skills in using **version control** by including tools such as *git*, *jira* or *cicd tools*. Lastly, **Container Platforms** (**CS14**) encompasses containerization technologies such as *Kubernetes* and *Docker*. These topics encompass a wide range of computer science skills, providing a comprehensive overview of the field and its various specializations.

Application Domain Skills. We have identified 18 topics that can relate either to skills specific to a particular application domain (e.g., healthcare) or to managerial skills (e.g., project management) (see Table 3). In the field of **Health-Care** (**AD1**), skills encompass *clinical* knowledge, *patient care*, *healthcare* practices, *medical* expertise, and *pharmaceutical* understanding. **Sales & Marketing** (AD2) involves skills in *sales* strategies, *marketing* campaigns, *brand* management, *social media*, *content* creation, and *customer* engagement. **Supply Chain Management** (**AD3**) covers skills in *supplier* management, *inventory* control, *procurement*, and smooth *supply chain* operations. The topic of **Testing** (**AD4**) *Testing* (*AD4*) encompasses skills in *test case* creation, *test* planning, *acceptance*

Table 3 Application domain skills

Topic name	Count	Highest-loading features (C-TF-IDF)
AD1: Healthcare	4473	Clinical (0.28), health (0.25), care (0.23), healthcare (0.22), patients (0.22), medical (0.22), patient (0.21), research (0.16), cancer (0.15), drug (0.15)
AD2: Sales & Marketing	2321	Marketing (0.34), sales (0.33), media (0.29), brand (0.26), social media (0.26), social (0.24), campaigns (0.24), content (0.23), customer (0.22), digital (0.19)
AD3: Supply Chain Management	1372	Supply (0.29), supply chain (0.27), chain (0.27), supplier (0.26), inventory (0.25), suppliers (0.24), manufacturing (0.23), quality (0.23), procurement (0.22), production (0.21)
AD4: Testing	1371	Test (0.59), testing (0.50), test cases (0.32), tests (0.29), test plans (0.29), acceptance (0.28), qa (0.26), user acceptance (0.26), acceptance testing (0.26), cases (0.25)
AD5: Environment	1331	Environmental (0.32), climate (0.30), sustainability (0.27), water (0.24), geotechnical (0.24), mining (0.22), marine (0.20), groundwater (0.20), biodiversity (0.19), environmental science (0.18)
AD6: Project Management	1275	Project (0.49), project management (0.30), projects (0.28), project management skills (0.27), project plans (0.25), scope (0.24), budget (0.24), plans (0.24), management skills (0.23), status (0.23)
AD7: Requirement Specification	1221	Requirements (0.39), business requirements (0.33), user stories (0.29), user (0.28), stories (0.28), business (0.26), functional (0.26), specifications (0.25), document (0.24), nonfunctional (0.21)
AD8: Compliance & Risk	1010	Compliance (0.43), audit (0.41), risk (0.39), audits (0.32), regulatory (0.30), risks (0.28), risk management (0.28), policies (0.28), internal (0.24), procedures (0.24)
AD9: Accounting & Finance	978	Accounting (0.34), finance (0.30), financial (0.29), degree (0.24), years (0.24), banking (0.22), experience (0.22), experience financial (0.22), years experience (0.21), financial services (0.21)
AD10: Manufacturing	906	Equipment (0.42), manufacturing (0.37), maintenance (0.31), quality (0.30), production (0.28), quality management (0.28), parts (0.27), assembly (0.26), standards (0.26)
AD11: Human Resources	895	Hr (0.58), talent (0.37), recruitment (0.31), human resources (0.30), human (0.30), hris (0.30), talent acquisition (0.26), hiring (0.24), resources (0.24), employee (0.24)
AD12: Customer Relationship Management	888	Stakeholder (0.55), stakeholder management (0.45), relationships (0.43), stakeholders (0.40), stakeholder engagement (0.40)
AD13: Energy	724	Energy (0.53), wind (0.36), renewable (0.34), electricity (0.33), solar (0.30), renewable energy (0.29), gas (0.25), power (0.25), transmission (0.24), energy industry (0.22)
AD14: Real-Estate	663	Real estate (0.45), estate (0.41), volunteering (0.38), communities (0.36), volunteer (0.35)
AD15: Market Research	509	Market (0.41), market research (0.35), competitor (0.32), trends (0.31), marketing (0.31)
AD16: Agriculture	289	Crop (0.48), seed (0.41), agriculture (0.41), agricultural (0.41), farming (0.38)
AD17: Pricing	270	Pricing (0.90), price (0.46), pricing strategies (0.35), pricing analyst (0.34), pricing strategy (0.32)
AD18: Insurance	226	Insurance (0.56), underwriting (0.39), reinsurance (0.38), insurance industry (0.34), experience insurance (0.25)

testing, *quality assurance*, and *automated* testing. In the **Environment** domain (**AD5**), skills relate to *environmental* science, *climate* studies, *sustainability*, *geotechnical* knowledge, and managing natural *resources*. **Project Management** (**AD6**) *Project Management* (*AD6*) includes skills in *project* planning, *scoping*, *budgeting*, and overseeing project progress and *deliverables*. **Requirement Specification** (**AD7**) focuses on gathering and documenting *user stories*, *functional* and *nonfunctional requirements*, and creating clear *specifications*. In **Compliance & Risk** (**AD8**), skills involve *audit* processes, *regulatory* compliance, risk management, and maintaining internal control *policies*. **Accounting & Finance** (**AD9**) involves *financial* management, accounts *payable/receivable*, *budgeting*, *accounting*, and invoice *processing*. **Manufacturing** (**AD10**) encompasses skills in equipment *maintenance*, *quality* management, *production* standards, and assembly line *operations*. For **Human Resources** (**AD11**), skills include HR management, HRIS (Human Resources Information Systems), employee data management, and HR team *coordination*. In **Customer Relationship Management** (**AD12**), skills center around *stakeholder* engagement, maintaining relationships with *internal* and *external stakeholders*, and *stakeholder* management. **Energy** (**AD13**) involves skills related to the energy industry, *renewable energy*, *transmission*, and knowledge of *wind*, *solar*, and other energy *sources*. Within the **Real-Estate** domain (**AD14**), skills pertain to property *management*, community *volunteering*, real estate transactions, and housing *initiatives*. **Market Research** (**AD15**) focuses on conducting *market research*, analyzing market *trends*, studying *competitors*, and implementing marketing *campaigns*. For **Agriculture** (**AD16**) , skills include crop *management*, *farming* techniques, seed selection, and

Table 4 Mathematics and statistic skills

Topic name	Count	Highest-loading features (C-TF-IDF)
MS1: Business Intelligence & Analytics	2286	Bi (0.38), power (0.33), power bi (0.33), tableau (0.32), dashboards (0.27), tools (0.23), analytics (0.22), powerbi (0.21), data (0.20), reports (0.20)
MS2: Machine Learning	1982	Machine learning (0.41), machine (0.41), ai (0.41), learning (0.40), ml (0.30), models (0.27), deep learning (0.26), learning models (0.25), machine learning models (0.25), deep (0.23)
MS3: Problem-Solving	1466	Ability (0.36), analytical (0.34), analytical skills (0.33), skills ability (0.32), strong analytical (0.32), skills (0.30), analytical skills ability (0.29), problemsolving (0.26), strong (0.26), strong analytical skills (0.25)
MS4: Scientific Methods	696	Research (0.50), scientists (0.33), interdisciplinary (0.29), scientific (0.27), researchers (0.25), science (0.24), big science (0.21), world (0.21), technology (0.21), discoveries (0.20)
MS5: Machine Learning Frameworks	644	Machine learning (0.39), machine (0.39), learning (0.38), deep learning (0.26), tensorflow (0.26), pytorch (0.25), ml (0.24), deep (0.24), science (0.23), computer (0.23)
MS6: Statistics	557	Statistical (0.52), spss (0.48), statistical skills (0.43), statistics (0.42), distributions (0.38), excel spss sas (0.38), excel spss (0.36), statistical tests (0.34)

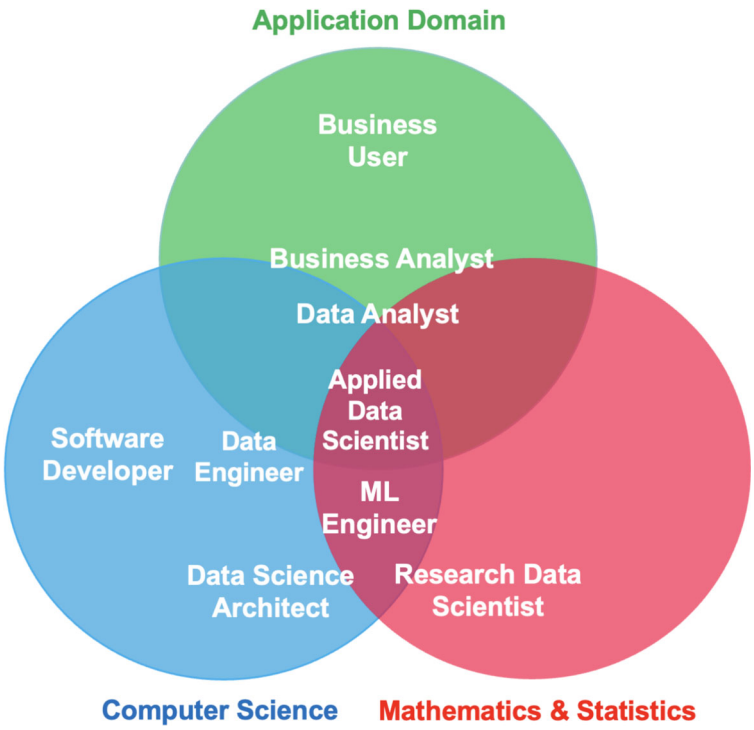
knowledge of *agricultural* practices. **Pricing** (**AD17**) involves skills in pricing *analysis*, revenue *management*, pricing strategies, and sales optimization. Lastly, **Insurance** (**AD18**) pertains to skills in insurance *underwriting*, *reinsurance*, *claims* management, and knowledge of the insurance *industry*. These topics span various application domains, offering a comprehensive overview of the skill sets required within each area.

Mathematics and Statistics Skills. We have identified six topics related to mathematics and statistics (see Table 4). In the field of **Business Intelligence & Analytics** (**MS1**), skills include *data analytics*, *understanding data*, *BI solutions*, working with tools like *Tableau* and *Power BI*, and creating interactive *dashboards*. **Machine Learning** (**MS2**) involves skills in *AI*, *machine learning*, *ML models*, *deep learning*, and understanding *artificial* intelligence concepts and algorithms. **Problem-Solving** (**MS3**) emphasizes *analytical* and *problem-solving* skills, including the ability to approach and solve complex problems using *strong analytical* and *problem-solving* abilities. **Scientific Methods** (**MS4**) focuses on skills related to *research*, working as *scientists* in an interdisciplinary setting,

Table 5 Job roles in data science

Role	Count	High-loading topics
Applied Data Scientist	5741	MS2: Machine Learning, MS5: Machine Learning Frameworks, MS6: Statistics, MS1: Business Intelligence & Analytics, CS5: Database Management, CS3: Data Engineering, CS1: Computer Science
Business User	912	AD12: Customer Relationship Management, AD2: Sales & Marketing, AD3: Supply Chain Management, CS11: ERP Systems, AD10: Manufacturing
Business Analyst	2517	MS1: Business Intelligence & Analytics, MS3: Problem-Solving, AD6: Project Management, CS11: ERP Systems, AD9: Accounting & Finance, AD12: Customer Relationship Management, AD3: Supply Chain Management
Data Analyst	539	MS1: Business Intelligence & Analytics, MS3: Problem Solving, MS4: Statistics, CS3: Data Engineering, MS5: Machine Learning Frameworks, AD6: Project Management, AD9: Accounting & Finance, AD10: Manufacturing
Data Engineer	3297	CS3: Data Engineering, CS5: Database Management, CS4: Data Governance, MS2: Machine Learning, MS5: Machine Learning Frameworks, CS9: Cloud Architecture, CS1: Computer Science, CS7: Agile Software De°velopment, CS12: Data Protection
Data Science Architect	453	CS1: Computer Science, CS9: Cloud Architecture, CS5: Database Management, CS7: Agile Software Development, MS6: Statistics, MS3: Problem-Solving
ML Engineer	846	MS2: Machine Learning, MS5: Machine Learning Framworks, CS3: Data Engineering, CS5: Database Management, CS4: Data Governance, CS9: Cloud Architecture, CS1: Computer Science
Research Data Scientist	256	MS2: Machine Learning, MS5: Machine Learning Frameworks, MS4: Scientific Methods, MS6: Statistics, MS1: Business Intelligence & Analytics, CS5: Database Management, CS3: Data Engineering, CS1: Computer Science
Software Developer	1723	CS1: Computer Science, CS8: Software Development, CS7: Agile Software Development, CS9: Cloud Architecture, CS5: Database Management, MS2: Machine Learning

Fig. 7 Job roles in data science



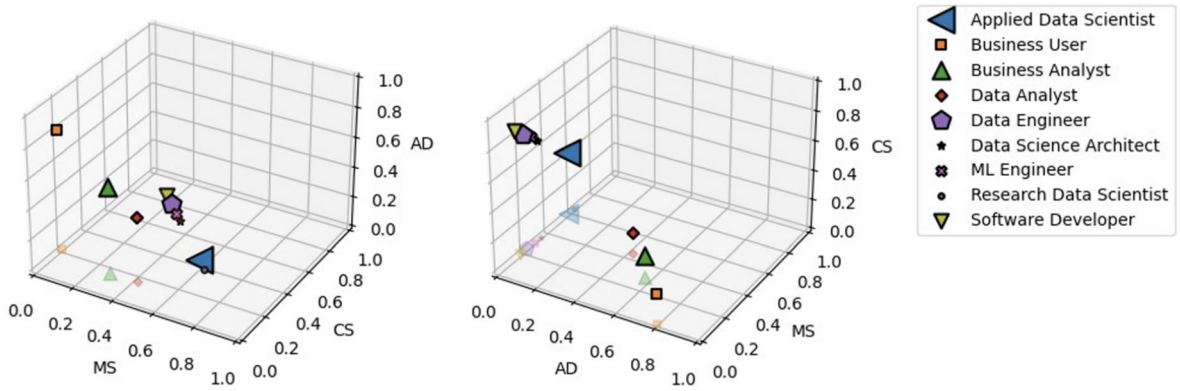


Fig. 8 3D vector space representing job roles in data science along two perspectives (MS = Mathematics and Statistics, CS = Computer Science, AD = Application Domain)

and contributing to scientific *discoveries* and advancements in various fields. **Machine Learning Frameworks** (**MS5**) encompasses skills in *ML* techniques, working with frameworks like *TensorFlow* and *PyTorch*, implementing *deep learning* algorithms, and having hands-on experience with *ML* projects. **Statistics** (**MS6**) involves skills in *statistical* analysis, using tools like *SPSS* and *SAS*, performing *statistical tests*, understanding *statistical distributions*, and working with data in *Excel*, *SPSS*, and *SAS*. These topics cover various aspects of data analysis, ML, problem-solving, and scientific research, providing a comprehensive overview of the skills required in each area.

5.2 Identified Job Roles

Based on our clustering approach, we identified a spectrum of nine job roles (see Table 5). These roles are defined by their associated skill sets, reflecting the diverse demands of the data science field. Figure 7 illustrates how these roles align with the core disciplines of data science by mapping identified skills to their respective domains, emphasizing the interdisciplinary nature of the field and the specific contributions of each role. Figure 8 represents the job roles as vectors in a three-dimensional space along two perspectives, showing how each role combines skills from these disciplines. Further, Fig. 9 shows how the identified roles support various stages of the data value chain, from

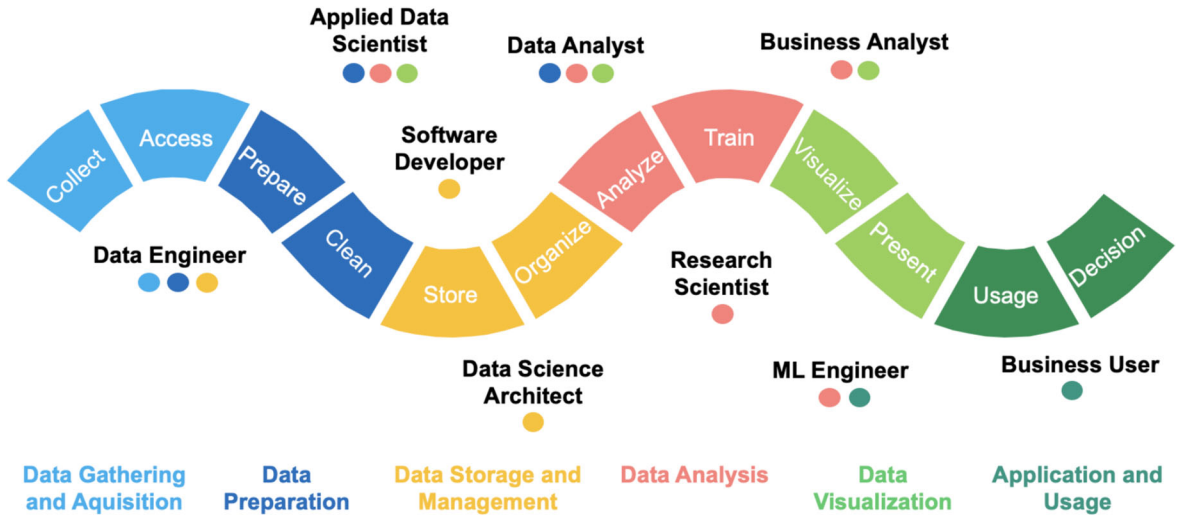


Fig. 9 Job roles in data science along data value chain

data generation to usage. By connecting these roles with the data value chain, this chapter bridges the findings of the systematic literature review with the empirical insights presented here, offering a comprehensive understanding of the roles in data science.

The first job role, the **Business User**, is composed of application domain skills without having statistical, mathematical, or computer science skills. Business Users are characterized by different degrees of domain knowledge and disciplinary responsibility. In particular, clerks and managers have only limited domain knowledge compared to domain experts. In this regard, Business Users provide domain knowledge to Data Scientists or Analysts. Our analysis revealed exemplary domain knowledge of Business Users in the field of supply chain management (**AD3**) and manufacturing (**AD10**) who therefore also know how to use ERP systems (**CS10**). Customer relationship management (**AD12**) or sales & marketing (**AD2**) make this role highly customer-oriented. For the Business User with higher disciplinary responsibility (i.e., managers) skills in project management (**AD4**) are necessary. This view aligns with the early 2010s literature, such as Provost and Fawcett (2013) and Viaene (2013), which often depicted domain knowledge as crucial but largely separate from the technical skills typically associated with data science.

For the next two job roles, we were able to distinguish two types of analysts: the Business Analyst and the Data Analyst. **Business Analysts** can be characterized as analytical Business Users, thus having strong domain knowledge, for instance, in customer relationship management (**AD12**), accounting & finance (**AD9**), or supply chain management (**AD3**). Because they have skills in business intelligence & analytics (**MS1**), Business Analysts can make analytical decisions and communicate them to Business Users. In this context, skills to solve problems (**MS3**) and project management (**AD6**) are notable. By having domain knowledge, Business Analysts know ERP systems well (**CS11**), which makes them a key job role at the beginning of a data science project, when an understanding of the data in a domain needs to be created. Thus, Business Analysts facilitate requirements specification (*AD8*) by using their domain knowledge. Although Business Analysts use tools such as Power BI or Tableau (**MS1**), relying on Excel (**CS2**) still dominates (Verma et al. 2019b). This role resonates with findings from Paul and Tan (2015), who describe Business Analysts as pivotal in bridging business and IT systems, an evolving role that has become increasingly analytical over the years. However, we see a strong development desire towards ML, which Business Analysts do not yet apply for their analyses today (Verma et al. 2021).

Data Analysts have less application domain knowledge (e.g., **AD9**, **AD10**) but more analytical skills compared to Business Analysts (De Mauro et al. 2018). Hence, skills in statistics (**MS6**) and problem-solving (**MS3**) are more common for Data Analysts (Verma et al. 2019b). In order to support analytical decision-making (**MS2**), they communicate and present findings to Business Users, for instance, by building dashboards for them. In contrast to Business Analysts, Data Analysts use simple ML approaches such as decision trees and k-means clustering (**MS5**). Although Data Analysts can take on a leadership role and manage projects (**AD6**), the Business Analyst with strong application domain knowledge is predestined for that. In addition, Data Analysts deal with technical topics such as data preparation (**CS13**) and cooperate with Data Engineers in this regard to prepare their analyses. In summary,

the Data Analyst still has a business understanding but with more mathematical and statistics skills. While literature such as Dhar (2013) emphasizes a broad skill set including ML and AI, Data Analysts in earlier literature were often not as technically intensive, suggesting a shift towards more sophisticated analytical capabilities over time.

In turn, the **Applied Data Scientist** has no specific business knowledge but relies strongly on ML (**MS2**, **MS5**) and statistics (**MS4**) to support analytical decision-making (**MS1**). Moreover, expertise with databases management (**CS5**), data engineering (**CS3**) is required. The latter requires proficiency in computer science (**CS1**) to understand the underlying concepts. However, one of the key challenges of this role, is contextualizing the results (e.g., of a classifier) and communicating the findings adequately and in a way that is appropriate to the target group. In summary, the Applied Data Scientist uses statistics and builds ML models by applying existing algorithms without dedicated business domain knowledge. This role has become increasingly specialized, focusing on deep technical expertise in specific analytical techniques and tools rather than a wider understanding of business contexts (van der Aalst 2014). The progression toward specialization highlights the field's shift from reliance on generalized data handling capabilities to specialized expertise in predictive modeling and algorithm development, allowing for more precise and impactful data-driven decisions (Michalczyk et al. 2021).

In contrast to the Applied Data Scientist, the **Research Data Scientist** takes on the role of a true scientist, concerned not so much with applying existing algorithms as with modifying them or developing new ones. Thus, "Scientific Methods" (**MS4**) as well as an academic degree in quantitative fields are required (typically at the doctoral level) to be able to "apply the scientific discovery research process, including hypothesis" (Saltz and Grady 2017, 620). For instance, Cao (2017) suggests that such roles are essential to advancing the technical frontiers of the field, consistent with the academic and innovative focus observed in recent years.

Lastly, we identified four highly technical job roles. The **Data Engineer** implements ETL pipelines (**CS3**) to extract data from source systems into other systems such as data warehouses (**CS5**). This job role makes data available, which is fostered by understanding how other job roles subsequently analyze data. Thus, this job role has knowledge in and ML (**MS2**) and ML techniques (**MS3**). Besides, expertise with cloud architecture is required (**CS9**), which frequently assumes computer science proficiency (**CS1**) but also work experience in "Agile Software Development" (**CS10**). The **ML Engineer** is frequently involved in diving deeper into the code and deploying the models created by Data Scientists into productive solutions. Hereby, this role uses ML frameworks (**MS2**, **MS5**) and has proficiency in computer science (**CS1**). Apart from the ML focus, they also have data engineering (**CS3**) skills. The **Software Developer** also has proficiency in computer science (**CS1**) and software development (**CS8**). The working style of this job role is agile (**CS7**). Technologies like the cloud (**CS9**) and big data (**CS5**) are important for Software Developers. Additionally, this job role is typically involved in ML activities (**MS2**) when models need to be deployed in production. The final job role, the **Data Science Architect**, has proficiency in computer science (**CS1**) and substantial experience with cloud architectures (e.g., **CS9**) and management of databases (**CS5**). Such roles can assist the corresponding team in cooperating in an agile manner (**CS7**). These roles align well with the ongoing emphasis in the literature due to (1) the increasing complexity of the technical landscape and (2) the focus on model deployment and development. The need for skills in cloud architecture and agile software development practices reflects the modern demands observed in job market analyses in recent years, reinforcing the consistent evolution of these roles toward the integration of more advanced technologies and methodologies. In particular, it reflects an increase in the demand for operational skills that bridge data science and software engineering (Miller 2019).

6 Discussion

In the following, we synthesize the findings from the literature analysis and present a state-of-the-art overview of job roles in data science, providing a comprehensive discussion of the implications for data science as a field.

6.1 Data Science as a Team Sport

First, our findings underscore the evolving concept of data science as a team sport. This highlights the importance of collaboration, which leads to increased standardization and governance within the field. This transformation reflects both a response and a driver of the growing specialization and segmentation of roles. (a) Regarding job roles located close to the application domains, we identified distinct roles tailored to varying degrees of business and "technical" skills. Unlike existing definitions, which often blur the distinctions between Business, BI, and Data Analysts (Verma et al. 2019b), our review did not find substantial support for the BI Analyst as a standalone role. Instead, we define these roles based on their technical capabilities and business involvement. Business Users predominantly focus on business operations, Business Analysts combine business understanding with foundational data analysis, and

Data Analysts engage more deeply with ML algorithms and data preparation, albeit with a reduced business focus. This stratification ensures clarity in role responsibilities and improves the integration of business insights with technical solutions. (b) Additionally, our concept expands the traditional responsibilities of data-centric roles. Data Engineers and Data Science Architects emerge as crucial for provisioning data and designing scalable infrastructures, equipped with advanced skills in technologies such as relational databases, big data, and cloud computing. This is distinct from the traditional Software Architect, as Data Science Architects must navigate existing systems and adapt them to new technologies, underscoring the necessity for a deep technological understanding that spans beyond mere software development. (c) Our concept acknowledges the need for a role that incorporates both software development and ML skills. We characterize the ML Engineer by being skills in ML with a Software Developer's technical profile. In contrast, some research (e.g., Gurcan and Cagiltay 2019) argues that today's Software Developers can be characterized as technically savvy Data Scientists because of their expertise in ML. Based on our concept, we argue that this is not the case due to the following key differences: (i) ML engineers are more trained in producing code (e.g., in Python), while Data Scientists tend to be less code-savvy, as their work primarily involves creating analytical models while also concentrating on the associated mathematics and statistics. (ii) It is in the ability to deploy a model that ML Engineers and Data Scientists probably differentiate the most. Whereas some Data Scientists know how to do this, we assume that ML engineers focus the bulk of their work on deployment.

The composition of data science teams plays a pivotal role in the transformation of data into actionable insights. The debate between generalists and specialists remains central to team structure Colson (2019). Generalists, valued for their versatility, can navigate multiple stages of the data lifecycle, making them ideal for smaller teams or projects with limited resources (Shi et al. 2023). However, their broad skill set may lack the depth needed for highly technical challenges. Specialists, on the contrary, offer in-depth expertise in areas such as data engineering, machine learning, or business analysis, enabling precise solutions for complex tasks (Nunez 2020). Critics of the specialization approach warn about silos and inefficiencies, while advocates highlight its role in delivering high-quality outcomes for advanced problems (Colson 2019; Nunez 2020). Striking the right balance between these approaches remains a critical challenge for organizations. Our research emphasizes a skill-based perspective on data science team composition, which highlights the importance of aligning roles with the demands of the data value chain while accounting for contextual factors. The data value chain provides a framework for identifying critical skills required at each stage of the data workflow (see Fig. 9). For instance, Data Engineers ensure high-quality infrastructure, Data Scientists and Analysts generate actionable insights through modeling and visualization, and Business Analysts and Business Users bridge data visualizations with strategic decision-making. Further, contextual factors such as budget, resource availability, and project scope significantly shape the composition of the team (Zhang et al. 2020). Smaller teams often rely on versatile generalists to handle multiple tasks, while larger teams benefit from specialists who can provide focused expertise at different stages of the data lifecycle (Lyon and Mattern 2017b). Ethical and regulatory constraints further necessitate roles such as Data Stewards to ensure compliance and governance. However, the success of any team structure depends on effective collaboration, which requires standardized tools, shared workflows, and seamless communication to integrate contributions across roles (Saltz et al. 2018)

6.2 Evolving Nature of Data Science

Second, our results reflect on and provide insights into the evolving nature of data science. After comparing our topics with the topics identified by Debortoli et al. (2014) and (Handali et al. 2020), we identified three emerging themes in the form of (1) healthcare, (2) self-service tools (e.g., Tableau, Microsoft PowerBI), and (3) data governance that are among the top topics in our analysis and therefore appear in almost all job ads. Specifically, with the digitization of healthcare systems and the widespread adoption of electronic health records, there is a vast amount of health data available. This includes patient medical records, clinical trial data, genetic data, and wearable device data (Rehman et al. 2022). The potential for improving healthcare delivery, and the opportunity to address public health challenges through data-driven approaches increases the need for people with such a skill set (Yeh et al. 2023). Similarly to organizations, healthcare facilities are trying to secure the necessary skills in data science and leverage the potential of the data collected. Moreover, more Business Users and their rich domain knowledge can be empowered with new branches of information systems such as Self-Service Business Intelligence & Analytics systems (e.g., Michalczyk et al. 2020) or interpretable decision support systems (Coussement and Benoit 2021). According to Lennerholt et al. (2021), Business Users "should be able to access and query data, use predefined reports, analyze data or create their own reports, in order to make decisions on time" (Lennerholt et al. 2021, 5056). In this regard, recent research has also highlighted the role that guidance can play in enabling Business Users with limited analytical skills (Gunklach and Nadj 2023). For instance, several BI

systems with guidance capabilities can be found in the literature that guide users through the data analysis process (e.g., Shi et al. 2021; Zschech et al. 2020). Moreover, collaboration within these roles is crucial for the successful application of data science. While Zhang et al. (2020) explored how technical roles like Data Scientists collaborate with non-technical roles such as Business Users, there is a pressing need to further investigate the interplay among all roles involved in data science. This includes not only the interactions within technical teams but also how these teams integrate with broader organizational roles to ensure a cohesive data strategy and execution. This holistic view is essential as data science continues to evolve as a team sport, where effective collaboration can significantly enhance outcomes and innovation.

In addition, our findings highlight the critical role of data science as a driver of organizational competitiveness, aligning with the Resource-Based View (RBV), which posits that organizations achieve sustained advantage by effectively utilizing unique and strategically valuable resources (Barney 1991). RBV offers a lens for understanding how organizations orchestrate human, tangible, and intangible resources to create value from increasingly complex data ecosystems (Gerhart and Feng 2021). By analyzing the diffusion and specialization of data science roles (Sect. 4) and the diverse skills required for their effective execution (Sect. 5), we demonstrate how these resources interact dynamically to enable business value and thus drive organizational competitiveness. Our findings resonate with the conceptualization of resources into human, tangible, and intangible categories from Mikalef et al. (2018). First, *human resources* are central to our findings, as we identified skills and topics from job ads and aligned them with the three core data science disciplines: (1) mathematics and statistics, (2) computer science, and (3) application domain expertise. These disciplines interact dynamically across the data value chain, where mathematics and statistics enable tasks such as data modeling and hypothesis testing, computer science drives data processing and model deployment, and application domain expertise ensures that insights are contextually relevant and actionable. *Tangible resources*, such as infrastructure, tools, and datasets, also emerge as foundational enablers in our study and are closely related to computer science skills. Roles like Data Science Architects, Software Developers, and Data Engineers contribute to building and maintaining the systems that facilitate scalable, efficient data operations, from data storage and management to analysis. *Intangible resources*, such as governance frameworks and a data-driven culture, further amplify the impact of these assets by ensuring alignment with organizational goals. Data governance ensures consistent data quality, integrity, and compliance across the organization. Meanwhile, a data-driven culture encourages employees to leverage insights in order to improve decision-making. Consistent with Dubey et al. (2019), we acknowledge that the effective bundling and deployment of resources are essential to improving competitive advantage. Our findings highlight that these resources interact dynamically throughout the data value chain, improving organizational competitiveness.

6.3 Importance of Data Governance for Data Science

Third, with respect to the development and application of data governance concepts and roles, it should be noted that they are rooted in the context of operational IT systems. For instance, following (Gröger 2021), key roles in data governance comprise data owners (i.e., take the overall responsibility for defined data and belong to the business), and data stewards (i.e., manage data on behalf of data owners and are responsible for implementing required policies and standards). To enhance the effectiveness of data governance roles, the formation of a data board (i.e., constitute central committees for decision-making on fundamental data governance concepts and principles) is recommended as this would bring together these key roles, facilitate coordinated governance efforts and ensure a unified approach to managing data integrity and compliance across the organization (Black et al. 2023). These aspects require rethinking and further developing data governance concepts and roles. On the one hand, established data governance roles for data ownership and data stewardship have to be harmonized in the context of data lakes to enable proactive and flexible use of raw data for data science. On the other hand, overlapping roles and areas of responsibility in data science and data governance have to be aligned. In particular, some roles have overlapping tasks with respect to data preparation, underlining the need for further research work (Gröger 2021). The skills related to data governance (e.g., data stewardship, regulatory compliance, data privacy and security expertise, data quality management, and metadata management) are vital for establishing a solid foundation for successful data science endeavors (Abraham et al. 2019).

6.4 The Impact of Generative AI on Job Roles in Data Science

Finally, the rapid advancement of Generative AI (GenAI) is fundamentally reshaping the field of data science (Gartner 2024). This development marks not just another step in technological progress, but a new critical stage in the ongoing evolution of roles in data science, a progression we have traced through previous technological waves in Sect. 4 of this paper (see Fig. 5). Unlike earlier stages,

which were characterized by the incremental addition of new tools and methods within data science, GenAI stands out as a foundational technology. It enables and transforms entire classes of tools, methods, and workflows, thus redefining what is possible in data-driven organizations (Abumalloh et al. 2024).

This shift is evidenced by the increasing adoption of GenAI across various occupations. For instance, the Anthropic Economic Index, based on millions of anonymized interactions with Claude - an advanced conversational AI assistant developed by Anthropic (2025b; Priyanshu et al. 2024) - demonstrates that GenAI adoption is highest among computer science, as well as mathematical and statistical professions. Approximately 37.2% of the GenAI queries analyzed were related to software modification, code debugging, and data engineering, underscoring GenAI's deep integration in these fields (Anthropic 2025a; Handa et al. 2025). In contrast, occupations involving physical labor or manual tasks remain least affected by GenAI (0.1% of queries), confirming that the main impact of GenAI is on knowledge-intensive work (Handa et al. 2025). GenAI introduces a dual dynamic for data science roles: it serves both as a tool for automation and as a collaborator for augmentation. The Anthrophic study found that 57% of AI-mediated tasks were augmentative – where AI collaborates with humans to validate, iterate, and enhance work – while 43% were fully automated (Handa et al. 2025). The usage of GenAI allows roles in data science to allocate more time to validating outputs, refining insights, and ensuring model reliability. Data Scientists, for instance, now spend less time writing boilerplate code and more time evaluating the validity and reliability of AI-generated outputs (Shih et al. 2024). Data Engineers use GenAI tools to assist with tasks such as writing transformation scripts, debugging, or documenting data pipelines, enabling them to focus on architectural decisions and scalability (Hassani and Silva 2023). Business Analysts interact more directly with data through natural language interfaces, narrowing the gap between technical complexity and business interpretation. Importantly, GenAI adoption is most pronounced among mid- to high-wage professions, such as Data Scientists and Software Engineers, while it remains less common in both the lowest- and highest-paid occupations (Handa et al. 2025). This pattern likely reflects the current limitations of AI capabilities and practical barriers to organizational adoption. Beyond reshaping existing roles, GenAI is also driving the emergence of new professional profiles. Notable examples include the Model Manager, who oversees the lifecycle management of large language models, and the Prompt Engineer, specializing in crafting and refining prompts to maximize AI effectiveness (Gartner 2024; Whiting 2023). These roles signal a broader redistribution of responsibilities, a growing emphasis on model governance, and the advent of new forms of human–AI collaboration (Via 2024).

At the same time, the integration of GenAI into data science practices introduces a range of new challenges. According to expert discussions at AWS re:Invent (Amazon Web Services 2023) and recent analyses (Yan et al. 2024; Sun et al. 2024), four main issues are particularly relevant. First, veracity: ensuring that AI-generated results are accurate and non-misleading is critical, given that GenAI systems can produce plausible but incorrect outputs Amazon Web Services (2023). This necessitates new validation workflows and heightened scrutiny from data professionals. Second, toxicity and safety: GenAI may inadvertently generate biased, offensive, or unsafe content, calling for robust governance frameworks and ongoing human oversight (Yan et al. 2024). Third, intellectual property: Legal uncertainty persists about the ownership of AI-generated content and the data used for model training (Yan et al. 2024). Finally, privacy: As GenAI systems process and generate outputs from sensitive organizational data, there is an increased risk of unintentional exposure or breaches of data protection regulation (Sun et al. 2024).

6.5 Practical Implications

Practically, we believe that our work can help organizations define strategies for acquiring and cultivating the right skills for a more effective use of data science. Following our concept, companies can capture the status quo of data science in their organization to determine current demands at the strategic level. Employees can be trained or new ones hired on this basis. Specifically, job roles with a strong analytical or technical background (e.g., Data Scientist, Data Engineer) would rather have to be recruited on the job market since a solid background in statistics or computer science is required, which is difficult to offer through on-the-job training. In this regard, our work can facilitate the hiring process for these roles because it is based on the most frequent terms in data science job ads (see Table 5). This information can help create relevant job ads and effectively target promising candidates in hiring campaigns. Thus, not all skills can be developed on the job. However, our 3D vector space visualization (see Fig. 8) revealed that especially analytics-affine employees could be trained to become Data Analysts, which would relieve the demand for highly sought-after job roles, such as Data Scientists, as these employees are the closest to Data Scientists. In this same vein, our work can supplement related training portfolios by demonstrating which skills are required for such training and by offering personalized options. Simultaneously, companies could establish mentoring programs. Here, senior Data Analysts could coach

Business Analysts to grow to the level of Data Analysts. At the operational level, our work can help staffing teams with data science projects by providing an understanding of the skills required for each job role and how the roles complement each other. For individuals, our work (a) is helpful to assess and develop their skill set in order to systematically accelerate career paths, and (b) serves as a reference that novices, scholars, and human resource managers can use for examinations.

6.6 Limitations

We are also aware of the fact that our work has limitations. In particular, any bias in the selection of the search string for the SLR process might result in a bias of the reviewed articles. To reduce this probability, our SLR and search process were based on well-established methodological guidelines (Boell and Cecez-Kecmanovic 2015). All decisions made during the planning, execution, and reporting stages are explicitly documented. With regard to the extraction of skills and roles, any bias in the algorithms' parameter settings could distort the results of the topic modeling and clustering approach. To reduce this possibility, we followed well-established methodological recommendations. All threshold and parameter settings are made transparent and defined clearly with a prior quantitative assessment. Furthermore, we conducted a qualitative expert assessment to select the parameters and suitable labels for our topics and clusters.

7 Conclusion

Organizations must develop the right skills among their workforce to effectively leverage data science. However, this goes far beyond hiring Data Scientists alone and leads to the need to fill various job roles in this field. In this article, we have offered clarity about the heterogeneous nature of job roles needed in data science by (1) providing a panoramic view of how data science roles have proliferated over the past 13 years, moving from generalist Data Scientists to a landscape characterized by a variety of specialized and standardized roles, and by (2) analyzing 16.348 job advertisements out of 30.397 collected from online platforms such as Indeed, Monster, Glassdoor, and Stepstone. Through this analysis, we identified nine job roles in demand by organizations (i.e., Business User, Business and Data Analyst, Data Engineer, Data Scientist, Research Data Scientist, ML Engineer, Software Developer, and Data Science Architect), and characterized each job role along their skill set. Methodologically, our text mining approach relied on word embeddings. In contrast to bag of words, word embeddings capture contextualized word representations, enabling a deeper understanding of the semantic relationships within the textual data. Our approach could also be reproduced to study job roles in other contexts. From a practical point of view, this understanding can help companies acquire and cultivate job roles to leverage data science more effectively. Thus, we hope that our findings can serve as an up-to-date reference for research and practice on the job role landscape of data science and the skills to consider.

Funding Open Access funding enabled and organized by Projekt DEAL.

Appendix

Search Terms

Job Roles. Data Analyst, Business Analyst, BI Analyst, Analyst, Data Scientist, Data Engineer, Data Steward, Data Manager, Data Officer, Data Architect.

General Terms. Big Data, Künstliche Intelligenz, KI, Business Intelligence, BI, Data Science, Analytics, Data Analytics, Artificial Intelligence, AI.

Data Science Terms. Machine Learning, Reinforcement learning, Unsupervised learning, Supervised Learning, Pattern recognition, Principal Component Analysis, Outlier Detection, Statistik, Statistical Analysis, Prediction, Extrapolation, Interpolation, Prescriptive, Statistical Analysis, Reporting, Deep Learning, Natural Language Processing, NLP, Text Mining, Topic Modelling, Generative AI, Automated Reasoning, Neural Network.

Data Analysis Terms. Data Preparation, Data Modeling, Data Model, Data Understanding, Data Wrangling, Data Cleaning, Data Processing, Data Preprocessing, Data Transformation, Data Visualization, Data Interpretation.

Business Intelligence Terms. Database, DB, Data Lake, Data Warehouse, Data Mart, Hadoop, MongoDB, Mapreduce, OLAP, Predective Modeling, Python, SQL, NoSQL, PowerBI, KNIME, Tableau, RapidMiner,

Cloudera, DataRobot, Databricks, Tibco Software, Dataiku, MathWorks, HVR, Mathlab, RStudio, SAS, Talend, ThoughtSpot, Qlik, Denodo, Informativa, Fivetran, Mastillion, SnapLogic.

References

Abbasi A, Sarker S, Chiang R (2016) Big data research in information systems: toward an inclusive research agenda. J AIS 17(2):1–32. https://doi.org/10.17705/1jais.00423

Abraham R, Schneider J, Vom Brocke J (2019) Data governance: a conceptual framework, structured review, and research agenda. Int J Inf Manag 49:424–438

Abumalloh RA, Nilashi M, Ooi KB, Tan G, Chan HK (2024) Impact of generative artificial intelligence models on the performance of citizen data scientists in retail firms. Comput Indust 161(104):128

Alekseeva L, Azar J, Gine M, Samila S, Taska B (2021) The demand for AI skills in the labor market. Labour Econ 71(102):002

Almgerbi M, De Mauro A, Kahlawi A, Poggioni V (2022) A systematic review of data analytics job requirements and online-courses. J Comput Inf Syst 62(2):422–434. https://doi.org/10.1080/08874417.2021.1971579 (https://www.tandfonline.com/doi/full/10.1080/08874417.2021.1971579)

Amazon Web Services (2023) Responsible AI in the generative era: science and practice. https://d1.awsstatic.com/events/Summits/reinvent2023/AIM220_Responsible-AI-in-the-generative-era-Science-and-practice.pdf. Accessed 25 Apr 2025

Anthropic (2025a) Introducing the anthropic economic index. https://www.anthropic.com/news/the-anthropic-economic-index. Accessed 25 Apr 2025

Antrophic (2025b) Claude. https://claude.ai. Accessed 25 Apr 2025

Bandara W, Furtmueller E, Gorbacheva E, Miskon S, Beekhuyzen J (2015) Achieving rigor in literature reviews: insights from qualitative data analysis and tool-support. Commun AIS 37(1):8

Bani-Hani I, Tona O, Carlsson S (2019) Modes of engagement in SSBA: a service dominant logic perspective. In: 25th Americas conference on information systems, AMCIS 2019. Association for Information Systems

Barney J (1991) Firm resources and sustained competitive advantage. J Manag 17(1):99–120. https://doi.org/10.1177/014920639101700108

Black S, Davern M, Maynard SB, Nasser H (2023) Data governance and the secondary use of data: the board influence. Inf Organ 33(2):100447

Boell SK, Cecez-Kecmanovic D (2015) On being 'systematic' in literature reviews. Formul Res Methods Inf Syst 2:48–78

Brath R, Hagerman C (2021) Automated insights on visualizations with natural language generation. In: 2021 25th International conference information visualisation (IV). IEEE, Sydney, Australia, pp 278–284. https://doi.org/10.1109/IV53921.2021.00052

Brauner S, Murawski M, Bick M (2023) The development of a competence framework for artificial intelligence professionals using probabilistic topic modelling. J Enterpr Inf Manag. https://doi.org/10.1108/JEIM-09-2022-0341

Cao L (2017) Data science: a comprehensive overview. ACM Comput Surv 50(3):1–42. https://doi.org/10.1145/3076253

Cao L (2019) Data science: profession and education. IEEE Intell Syst 34(5):35–44. https://doi.org/10.1109/MIS.2019.2936705

Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D (2009) Reading tea leaves: how humans interpret topic models. Adv Neural Inf Process Syst 22

Chen Chiang Storey (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165. https://doi.org/10.2307/41703503

Colson E (2019) Why data science teams need generalists, not specialists. Harvard Bus Rev. https://hbr.org/2019/03/why-data-science-teams-need-generalists-not-specialists. Accessed: 2025 Jan 14

Coussement K, Benoit DF (2021) Interpretable data science for decision making. Decis Support Syst 150(113):664. https://doi.org/10.1016/j.dss.2021.113664

Cui Z, Badam SK, Yalçin MA, Elmqvist N (2019) DataSite: proactive visual data exploration with computation of insight-based recommendations. Inf Visual 18(2):251–267. https://doi.org/10.1177/1473871618806555

DalleMule L, Davenport TH (2017) What's your data strategy. Harvard Bus Rev 95(3):112–121

Davenport T (2020) Beyond unicorns: educating, classifying, and certifying business data scientists. Harvard Data Sci Rev 2(2):5

Davenport TH, Patil D (2012) Data scientist. Harvard Bus Rev 90(5):70–76

De Mauro A, Greco M, Grimaldi M, Ritala P (2018) Human resources for big data professions: a systematic classification of job roles and required skill sets. Inf Process Manag 54(5):807–817. https://doi.org/10.1016/j.ipm.2017.05.004

Debortoli S, Müller O, Vom Brocke J (2014) Comparing business intelligence and big data skills: a text mining study using job advertisements. Bus Inf Syst Eng 6(5):289–300. https://doi.org/10.1007/s12599-014-0344-2

Debortoli S, Müller O, Junglas I, Vom Brocke J (2016) Text mining for information systems researchers: an annotated topic modeling tutorial. CAIS 39(1):7

Della Volpe M, Esposito F (2020) How universities fill the talent gap: the data scientist in the Italian case. Afr J Bus Manag 14(2):53–64

Dhar V (2013) Data science and prediction. Commun ACM 56(12):64–73. https://doi.org/10.1145/2500499

Dong T, Triche J (2020) A longitudinal analysis of job skills for entry-level data analysts 31

Dubey R, Gunasekaran A, Childe SJ, Blome C, Papadopoulos T (2019) Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. Br J Manag 30(2):341–361

Eichler R, Giebler C, Gröger C, Schwarz H, Mitschang B (2021) Modeling metadata in data lakes—a generic model. Data Knowl Eng 136(101):931

Elia G, Polimeno G, Solazzo G, Passiante G (2020) A multi-dimension framework for value creation through big data. Indust Market Manag 90:617–632

Endres DM, Schindelin JE (2003) A new metric for probability distributions. IEEE Trans Inf Theor 49(7):1858–1860

Fadler M, Legner C (2022a) Data ownership revisited: clarifying data accountabilities in times of big data and analytics. J Bus Anal 5(1):123–139

Fadler M, Legner C (2022b) Data ownership revisited: clarifying data accountabilities in times of big data and analytics. J Bus Anal 5(1):123–139. https://doi.org/10.1080/2573234X.2021.1945961

Fan W, Geerts F (2022) Foundations of data quality management. Springer, Berlin

Faroukhi AZ, El Alaoui I, Gahi Y, Amine A (2020) Big data monetization throughout big data value chain: a comprehensive review. J Big Data 7:1–22

Fayyad U, Hamutcu H (2022) From unicorn data scientist to key roles in data science: standardizing roles. Harvard Data Sci Rev 1:234. https://doi.org/10.1162/99608f92.008b5006

Gardiner A, Aasheim C, Rutner P, Williams S (2018) Skill requirements in big data: a content analysis of job advertisements. J Comput Inf Syst 58(4):374–384

Gartner (2024) Ai is creating new roles and skills in data & analytics. https://www.gartner.com/en/newsroom/press-releases/2024-05-14-artificial-intelligence-is-creating-new-roles-and-skills-in-data-and-analytics. Accessed 15 Apr 2025

Gerhart B, Feng J (2021) The resource-based view of the firm, human resources, and human capital: progress and prospects. J Manag 47(7):1796–1819. https://doi.org/10.1177/0149206320978799

Gottipati S, Shim KJ, Sahoo S (2021) Glassdoor job description analytics – analyzing data science professional roles and skills. In: 2021 IEEE global engineering education conference (EDUCON). IEEE, Vienna, pp 1329–1336. https://doi.org/10.1109/EDUCON46332.2021.9453931

Gröger C (2021) There is no AI without data. Commun ACM 64(11):98–108. https://doi.org/10.1145/3448247

Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794

Gunklach J, Nadj M (2023) Guidance in business intelligence and analytics systems: a review and research agenda. ECIS 2023 research papers 329

Gunklach J, Jacob K, Michalczyk S (2023) Beyond dashboards? designing data stories for effective use in business intelligence and analytics. ECIS 2023 research papers 327

Gunklach J, Nadj M, Knaeble M, Bragaglia I, Maedche A (2024) Designing for effective human-guided machine learning feasibility analysis. In: ICIS 2024 proceeding 1

Gurcan F, Cagiltay NE (2019) Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE Access 7:82541–82552. https://doi.org/10.1109/ACCESS.2019.2924075

Han J, Pei J, Tong H (2022) Data mining: concepts and techniques. Morgan Kaufmann, New York

Handa K, Tamkin A, McCain M, Huang S, Durmus E, Heck S, Mueller J, Hong J, Ritchie S, Belonax T, et al. (2025) Which economic tasks are performed with AI? Evidence from millions of claude conversations. arXiv:2503.04761

Handali JP, Schneider J, Dennehy D, Hoffmeister B, Conboy K, Becker J (2020) Industry demand for analytics: a longitudinal study. In: ECIS 2020 research papers

Hassani H, Silva ES (2023) The role of ChatGPT in data science: how AI-assisted conversational interfaces are revolutionizing the field. Big Data Cognit Comput 7(2):62

Haug S, Maedche A (2021) Crowd-feedback in information systems development: a state-of-the-art review. In: ICIS

Kim M, Zimmermann T, DeLine R, Begel A (2016) The emerging role of data scientists on software development teams. In: Proceedings of the 38th international conference on software engineering. ACM, Austin Texas, pp 96–107. https://doi.org/10.1145/2884781.2884783

Krause J, Perer A, Bertini E (2014) INFUSE: interactive feature selection for predictive modeling of high dimensional data. IEEE Trans Visual Comput Graph 20(12):1614–1623. https://doi.org/10.1109/TVCG.2014.2346482

Legner C, Fadler M, Pentek T (2023) Data governance methodologies: the CC CDQ reference model for data and analytics governance. Data governance: from the fundamentals to real cases. Springer, Heidelberg, pp 99–119

Lennerholt C, Van Laere J, Söderström E (2021) User-related challenges of self-service business intelligence. Inf Syst Manag 38(4):309–323. https://doi.org/10.1080/10580530.2020.1814458

Levy O, Goldberg Y (2014) Dependency-based word embeddings. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Baltimore, Maryland, pp 302–308. https://doi.org/10.3115/v1/P14-2050

Lismont J, Van Calster T, Óskarsdóttir M, vanden Broucke S, Baesens B, Lemahieu W, Vanthienen J, (2019) Closing the gap between experts and novices using analytics-as-a-service: an experimental study. Bus Inf Syst Eng 61(6):679–693. https://doi.org/10.1007/s12599-018-0539-z

Liu Y, Liu Z, Chua TS, Sun M (2015) Topical word embeddings. In: Proceedings of the AAAI conference on artificial intelligence, vol 29

Lyon L, Mattern E (2017a) Education for real-world data science roles (part 2): a translational approach to curriculum development. IJDC 11(2):13–26. https://doi.org/10.2218/ijdc.v11i2.417

Lyon L, Mattern E (2017b) Education for real-world data sciencen roles (part 2): a translational approach to curriculum development. Int J Digit Curat 11(2):13–26. https://doi.org/10.2218/ijdc.v11i2.417

Mathis C (2017) Data lakes. Datenbank Spektr 17(3):289–293. https://doi.org/10.1007/s13222-017-0272-7

Michalczyk S, Nadj M, Azarfar D, Maedche A, Gröger C (2020) A state-of-the-art overview and future research avenues of self-service business intelligence and analytics. In: European conference on information systems, p 21

Michalczyk S, Nadj M, Maedche A, Gröger C (2021) Demystifying job roles in data science: a text mining approach. In: ECIS 2021 research papers 115

Mikalef P, Pappas IO, Krogstie J, Giannakos M (2018) Big data analytics capabilities: a systematic literature review and research agenda. Inf Syst e-Bus Manag 16:547–578

Mike K, Hazzan O (2023) What is data science? Commun ACM 66(2):12–13

Mildenberger T, Braschler M, Ruckstuhl A, Vorburger R, Stockinger K (2023) The role of data scientists in modern enterprises-experience from data science education. ACM SIGMOD Record 52(2):48–52. https://doi.org/10.1145/3615952.3615966

Miller GJ (2019) The influence of big data competencies, team structures, and data scientists on project success. In: 2019 IEEE technology and engineering management conference (TEMSCON). IEEE, Atlanta, pp 1–8. https://doi.org/10.1109/TEMSCON.2019.8813604

Murawski M, Bick M (2017) Demanded and imparted big data competences: towards an integrative analysis. In: Proceedings of the 25th European conference on information systems (ECIS)

Naumann F (2014) Data profiling revisited. ACM SIGMOD Rec 42(4):40–49

Nunez G (2020) Generalists vs specialists in data science and analytics. Towards Data Science https://towardsdatascience.com/generalists-vs-specialists-in-data-science-and-analytics-fe5d8b55f1e6. Accessed 14 Jan 2025

Open Data Watch (2018) The data value chain: Moving from production to impact. The data value chain: moving from production to impact. https://opendatawatch.com/wp-content/uploads/2018/03/Data_Value_Chain-WR-1803126.pdf. Accessed 14 Jan 2025

Otto B (2011) Data governance. Bus Inf Syst Eng 3(4):241–244. https://doi.org/10.1007/s12599-011-0162-8

Paul D, Tan YL (2015) An investigation of the role of business analysts in is development. In: ECIS

Priyanshu A, Maurya Y, Hong Z (2024) AI governance and accountability: an analysis of anthropic's claude. arXiv:2407.01557

Provost F, Fawcett T (2013) Data science and its relationship to big data and data-driven decision making. Big Data 1(1):51–59. https://doi.org/10.1089/big.2013.1508

Rehman A, Naz S, Razzak I (2022) Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimed Syst 28(4):1339–1371

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth acm international conference on web search and data mining, pp 399–408

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Saltz JS, Grady NW (2017) The ambiguity of data science team roles and the need for a data science workforce framework. In: 2017 IEEE international conference on big data (Big Data). IEEE, Boston, pp 2355–2361. https://doi.org/10.1109/BigData.2017.8258190

Saltz J, Armour F, Sharda R (2018) Data science roles and the types of data science programs. CAIS 43:615–624. https://doi.org/10.17705/1CAIS.04333

Schubert E, Rousseeuw PJ (2019) Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In: Similarity search and applications: 12th international conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, proceedings 12. Springer, Heidelberg, pp 171–187

Shi D, Xu X, Sun F, Shi Y, Cao N (2021) Calliope: automatic visual data story generation from a spreadsheet. IEEE Trans Visual Comput Graph 27(2):453–463. https://doi.org/10.1109/TVCG.2020.3030403

Shi C, Su Y, Yang C, Yang Y, Cai D (2023) Specialist or generalist? instruction tuning for specific NLP tasks. arXiv:2310.15326

Shih JY, Mohanty V, Katsis Y, Subramonyam H (2024) Leveraging large language models to enhance domain expert inclusion in data science workflows. In: Extended abstracts of the chi conference on human factors in computing systems, pp 1–11

Stevens MJ, Campion MA (1994) The knowledge, skill, and ability requirements for teamwork: implications for human resource management. J Manag 20(2):503–530

Sun Y, Jang E, Ma F, Wang T (2024) Generative ai in the wild: prospects, challenges, and strategies. In: Proceedings of the 2024 chi conference on human factors in computing systems, pp 1–16

Tamm T, Seddon PB, The University of Melbourne, Shanks G, The University of Melbourne (2020) How do different types of BA users contribute to business value? CAIS 46(1):656–678. https://doi.org/10.17705/1CAIS.04628

van der Aalst WMP (2014) Data scientist: the engineer of the future. In: Enterprise interoperability VI, proceedings of the I-ESA conferences, vol 7, Springer, Cham, pp 13–26, https://doi.org/10.1007/978-3-319-04948-9_2, https://link.springer.com/10.1007/978-3-319-04948-9_2

Verma A, Yurov KM, Lane PL, Yurova YV (2019a) An investigation of skill requirements for business and data analytics positions: a content analysis of job advertisements. J Edu Bus 94(4):243–250. https://doi.org/10.1080/08832323.2018.1520685

Verma A, Yurov KM, Lane PL, Yurova YV (2019b) An investigation of skill requirements for business and data analytics positions: a content analysis of job advertisements. J Edu Bus 94(4):243–250. https://doi.org/10.1080/08832323.2018.1520685

Verma S, Singh V, Bhattacharyya SS (2021) Do big data-driven HR practices improve HR service quality and innovation competency of smes. Int J Organ Anal 29(4):950–973

Via A (2024) GenAI is reshaping data science teams. https://medium.com/data-science/genai-is-reshaping-data-science-teams-b4d5a419e0f6. Accessed 25 Apr 2025

Viaene S (2013) Data scientists aren't domain experts. IT Prof 15(6):12–17. https://doi.org/10.1109/MITP.2013.93, http://ieeexplore.ieee.org/document/6674007/

Virkus S, Garoufallou E (2020) Data science and its relationship to library and information science: a content analysis. DTA 54(5):643–663. https://doi.org/10.1108/DTA-07-2020-0167

Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Q xiii–xxiii

Whiting K (2023) The rise of the 'prompt engineer' and why it matters. https://www.weforum.org/stories/2023/05/growth-summit-2023-the-rise-of-the-prompt-engineer-and-why-it-matters/. Accessed 2025 Apr 15

Wixom BH, Ross JW (2017) How to monetize your data. MIT Sloan Manag Rev 58(3)

Yan L, Greiff S, Teuber Z, Gašević D (2024) Promises and challenges of generative artificial intelligence for human learning. Nat Hum Behav 8(10):1839–1850

Yeh YT, Eden R, Fielt E, Syed R, Donovan R, Eley R, Staib A (2023) Unveiling the value of big data analytics use: a digital hospital case study. In: ECIS 2023 research papers 318

Zhang V (2019) Stop searching for that data scientist unicorn. https://www.infoworld.com/article/3429185/stop-searching-for-that-data-science-unicorn.html. Accessed 15 Apr 2025

Zhang AX, Muller M, Wang D (2020) How do data science workers collaborate? Roles, workflows, and tools. Proc ACM Hum-Comput Interact 4(CSCW1):1–23. https://doi.org/10.1145/3392826

Zschech P, Horn R, Höschele D, Janiesch C, Heinrich K (2020) Intelligent user assistance for automated data mining method selection. Bus Inf Syst Eng 62(3):227–247. https://doi.org/10.1007/s12599-020-00642-3