Survey-Based Evaluation of Risks in Data Science Projects: Insights from a Literature Review and Considerations for Generative AI

Maike Holtkemper[1] and Christian Beecks[1]

[1] *University in Hagen, Chair of Data Science, Germany*

Abstract

Generative AI, often referred to as GenAI, has emerged as a transformative technology with the potential to modernize data management practices across various fields, including the application within data science projects. Building on a previous study that identified risks in data science projects through a comprehensive literature review, this study presents an evaluation of these risks through a survey administered to professionals in the field. The survey was designed to determine whether the risks identified in the literature are also observed in practical settings and to understand how they manifest in real-world scenarios. The survey results were then compared with the findings from the previous literature review to assess alignment and identify any emerging trends in practice. Additionally, this study explores the potential role of GenAI in mitigating the identified risks. By analyzing how GenAI tools can address specific challenges in data science projects, we provide actionable insights and recommendations for practitioners. The findings suggest that while there is significant overlap between the theoretical risks and those observed in practice, GenAI holds promise in effectively reducing certain project risks, thereby enhancing the overall success and efficiency of data science initiatives.

Keywords

Risks, Data Science Projects, Survey, Generative AI

1. Introduction

Data science and AI projects have the potential to deliver in-depth insights and significant competitive advantages for companies [1]. Nevertheless, the majority of these projects fail and do not make it to production [1,2], meaning that companies are unable to derive any economic benefit from their investments in big data, artificial intelligence and machine learning [3].

There are many reasons for project failure. The most common reasons include inadequate project management, poor data quality and the limited availability of data [4]. These problems can lead to projects not achieving their goals, running inefficiently or even failing completely.

These challenges can be partially addressed with the help of automation frameworks. The latter can automate many tasks along the data science lifecycle and thus improve the efficiency and quality of projects [5]. Popular examples for automation frameworks include AutoDS [5], AutoCure [6] and Auto-Prep [7], which can support tasks along the data science lifecycle such as data preprocessing. However, operating these frameworks often requires in-depth knowledge of programming to adapt them to individual problems for effective utilization. In addition, companies are facing an increasing shortage of data scientists on the labor market [8].

One possible solution for overcoming these challenges is the use of GenAI [9]. GenAI offers advanced methods for data processing and generation that can potentially improve the quality of data while also supporting project management [10,11]. As compared with many automation frameworks, GenAI is easier and more intuitive to use. In terms of data science projects, for example, large language models (LLMs), a special form of GenAI, can help to automate the entire ML workflow [12] which saves the data scientist's time.

In a previous study, a comprehensive literature review was conducted to identify the most common risks in data science projects [4]. These identified risks were then verified and validated through a survey of experts from industry and academia.

The aim of this study is to present the survey results in the context of the findings from the previous literature review and to explore how GenAI can be leveraged to support data science teams in mitigating risks that may lead to project failure. The study builds on earlier research by validating theoretical risks through practical, empirical data, thereby offering insights into both the alignment and discrepancies between these perspectives. Particular attention will be paid to minimizing problems related to data management, especially data quality.

Therefore, the paper aims to answer the following research question (RQ1)*: To what extent can the risks identified in the literature be validated in practice through the survey?* These findings will be then compared, which leads to the second research question (RQ2): *To what extent can GenAI contribute to minimizing the validated risks?*

The paper is structured as follows: First the background is described in Section 2. Second, the structure and implementation of the survey are described in the methodology, Section 3. Third, the findings are presented and discussed in Section 4 and finally, the conclusions and future research directions are given in Section 5.

2. Background

Van der Aalst [2016] describes Data Science as an interdisciplinary field [13], which, according to Cao, require special skills from the involved employees in their application [14]. As the original driver for the development of data science came from big data initiatives [15], these two fields are often seen as closely intertwined. Although complicated data is not a necessity for the discipline, it is more commonly used in data science compared to other data-driven analytical endeavors. Moreover, data science assessments usually require the application of advanced analytical techniques [14].

However, these analytical skills are just one of many skills that are required to successfully complete a data science project. During the last decade, the impression has aroused that a data scientist must have all the relevant competencies such as "hacking skills, math and stats knowledge, and substantive expertise" [16] for a complex data science project. According to Aßmann, "this shows that too many different skills are needed to be represented by one person" [17]. Although the competencies required for the successful implementation of data science in companies can be covered by the collaboration of several employees in a data science project, individual data science competencies are often lacking in companies [18]. Even if the necessary skills exist in the team, the project management of these interdisciplinary teams is a particularly difficult challenge [19].

At the end of November 2022, OpenAI introduced ChatGPT, a generative conversational agent that is designed for open domain interactions, as its new technology to the public [20]. Given the extensive range of potential applications for this technology, some believe it could represent a groundbreaking innovation comparable to the release of the iPhone in 2007 [21]. However, others argue that the potential risks may surpass the overall benefits [22].

According to Schuckart [2023], a generative pre-trained transformer (GPT) offers significant advantages in various professional fields, but many potential applications still need to be explored. For example, text generation with GenAI models such as GPT-3 can be used to generate high-quality texts that can be used in applications such as content creation, chatbots and automated writing [23]. In areas such as marketing, journalism and education, copywriting can automate repetitive tasks and provide new insights. It can also be used to create personalized content that improves communication and engagement.

Using short text inputs with image-generating AI such as DALL-E 2 enables advertising, media, education and arts professionals to create engaging content that combines visual and written elements to drive creativity and innovation. Image creation can bridge the boundaries between disciplines and encourage collaboration [20].

Conversational GenAI, such as ChatGPT, generates human-like responses that are useful for chatbots, virtual assistants and customer service and helps bridge communication gaps between experts in different fields by providing easy-to-understand explanations [20]. ChatGPT can also support personalized applications in healthcare and education, e.g., the technology is able to provide medical advice, even while emphasizing the need for professional supervision [20]. The technology's adaptability in generating specific and contextualized responses demonstrates its potential.

The rapid development of GenAI and the increasing availability of GenAI-based models has also found its way into computer science [23]. With regard to the educational sector, Smith et al. surveyed computer science majors at a small engineering-focused university to assess the immediate adoption of GenAI by aspiring computer scientists, their needs and concerns and its influence on Computer Science pedagogy, curriculum, culture and policy [24]. Vadaparty et al. [2024] explored the impact of incorporating Large Language Models (LLMs) into introductory programming courses (CS1) and provide insights and lessons for educators on this integration [25]. Motivated by the capabilities of LLMs to solve typical CS1 assignments and their increasing use by professional software engineers, Vadaparty et al. describe how a CS1 course at a large research-intensive university was redesigned to emphasize skills such as explaining and testing code, and decomposing problems, rather than traditional coding from scratch. Amer-Yahia et al. [2023] call for current curricula in Computer Science and Data Science to be reformed to include ChatGPT [23].

Regarding its use in the industrial sector, Chakraborty et al. [2024] developed Navigator, a GenAI-based question-answering system designed to enable business users to ask natural language questions and receive answers derived from domain-specific tabular datasets that have been integrated into the system [26].

Beasley and Abouzied [2024] aim to leverage LLMs to fully automate the descriptive analytics and visualization process using their tool, LLM4Vis. This tool enables users to specify their identity and provide a dataset, after which it establishes analysis objectives or metrics, produces code for data processing ana analysis, visualizes the results, and interprets these visualizations to highlight key insights for the user. By using LLM4Vis, the authors want to make data science accessible to non-technical users, allowing them to work with complex, multimodal datasets without needing specialized skills [27].

3. Methodology

In a prior study, an extensive literature review was conducted to systematically identify and categorize the prevalent risks in data science projects. This foundational work laid the groundwork for the current study, which aims to validate these risks in practical settings through empirical data collection [4].

To validate the identified risks in practical settings, a survey was administered to 80 experts from both industry and academia within the field of data science. The survey aimed to assess the relevance and impact of these risks in real-world scenarios.

Traditionally, surveys have been the primary method for gathering data in the field of information systems [28]. According to Recker [2021], a distinction is made between three types of survey: exploratory, descriptive and explanatory [29]. An exploratory survey is used to explore a topic and gain initial insights when little is known, helping to generate hypotheses and develop a basic understanding [29]. A descriptive survey is used to describe the current state of a phenomenon by collecting ana analyzing data to create a detailed picture of the topic, often involving the collection and evaluation of information from various sources such as literature [29]. An explanatory survey aims to identify and explain the causes and relationships between different variables, going beyond mere description to seek out causal relationships [29]. In this study, a descriptive survey is used.

3.1. Survey Design

The survey was implemented using a locally installed version of the open-source tool LimeSurvey, hosted by the University of Hagen. This platform facilitated the anonymous collection of responses while ensuring data integrity and confidentiality. To eliminate confusing wording, technical errors, and gaps in question coverage, the survey was thoroughly reviewed by the research team, resulting in a smooth and error-free deployment.

The identified risks in the literature research were grouped according to the Risk Breakdown Structure (RBS) of the PMBOK Guide, the Project Management Body of Knowledge [30]. The RBS, as a hierarchical representation of potential risk sources, supported the structuring of the survey by aligning with established project management practices [30].

The final survey consisted of 14 carefully structured questions, each designed to address specific aspects of the first research question. These questions were directly derived from the risks identified in the comprehensive literature review conducted in the previous study, ensuring that the survey captured a broad spectrum of risks relevant to data science projects.

Grounded in the PMBOK and the RBS framework, the questionnaire covered key areas such as informed consent, industry sector, company size, professional role, and the implementation of risk assessments within respondents' organizations.

Respondents were asked to evaluate the identified risks using a Likert scale, ranging from low to high, to assess their practical relevance. Additionally, the survey included questions on the importance of various skills in data science projects and the relevance of domain knowledge, with opportunities for free-text comments and suggestions. This design aimed to capture nuanced insights and comprehensively address the research questions.

3.2. Sample Selection

To attract as many participants as possible, the survey was published and advertised on several channels. A post with the link to the survey was created on LinkedIn and participation was encouraged in relevant data science groups. The survey was also shared via the "Work, Education, Digital" (ABD) working group and encouraged to participate via private networks. The survey is also available on the homepage of the University of Hagen in the survey section.

3.3. Recruitment

Despite extensive advertising efforts, only 80 participants have been recruited for the survey. Of these, 3 participants have only accessed the survey link, 35 have only partially completed, while 42 have completed the survey in full. Among these N = 42 participants, N = 10 (23,80 %) are engaged in scientific activities in the field of data science, N = 9 (21,43 %) work as project managers in data science, N = 8 (19,05 %) are professional data scientists, N = 4 (9,52 %) are data analysts, N = 3 (7,14 %) are business analysts, N = 2 (4,76 %) are data engineers, and N = 6 (14,29 %) hold different positions.

In terms of the sectors in which participants are employed, 10 work in the academic sector, 8 in the food industry, 4 in the energy industry, 3 in telecommunications, IT, and consumer electronics, 3 in the finance sector, 2 in the service industry, 2 in commerce, 2 in economics and politics, 1 in the metal industry, 1 in the consumer goods sector, 1 in the chemistry and raw materials sector, 1 in the pharmaceutical industry, and 4 in other sectors.

Regarding company size, 80.95% of respondents work for companies with more than 250 employees, 11.90% are employed by companies with fewer than 250 employees, and 7.14% work for companies with more than 50 employees.

3.4. Survey Analysis

The data were exported from LimeSurvey into a CSV file. Then the standard Python packages such as pandas, seaborn and matplotlib were used to compute descriptive statistics and perform statistical tests. Due to the low frequency of responses in the last question, which was aimed at further comments, a content analysis was not carried out.

It is acknowledged that the sample size is relatively small and may not be statistically representative, which limits the generalizability of the findings to all data science projects. However, despite the small sample size, a trend can be observed in the data, providing valuable insights that should be considered when interpreting the results. The usual limitations for surveys also apply, including opt-in bias and potentially inaccurate self-assessments.

4. Findings

4.1. Recap from previous literature review

Through an extensive literature review, risks associated with data science projects were identified and categorized [4]. Initially, 354 papers were identified, with 248 relevant risks documented in 40 papers. After refining the results, the most common risks were categorized, emphasizing the need for thorough risk assessments to support project management. In total 29 categories were identified, including: insufficient project management, poor team management, bad customer expectation management, bad data and information management, poor communication with customer/stakeholders, uncertainty about project outcome, organizational culture as barrier and lack of data science competence/skills, data security and privacy, poor data availability and quality, data inconsistency and incompleteness, data ownership unclear, high level of complexity, inaccuracy of the model, bad user management, lack of requirement management, lack of domain knowledge, bad documentation, bad maintenance planning, insufficient infrastructure, new technology, daily operational risks, inadequate technical development/deployment practices, market risks (i.e., competition) and regulatory risks (i.e., strengthening data protection).

The top ten risks of the literature review contain insufficient project management (35,08%), data security and privacy (10,48 %), poor data availability and quality (8,06%), high level of complexity (7,26%), lack of data science competence/skills (6,05%), inadequate technical development/deployment practices (5,64%), bad data and information management (4,44%), inaccuracy of the model (3,63%), poor communication with customer/stakeholder (2,82%) and organizational culture as barrier (2,42%).

4.2. Preparation for the survey

The identified risks through the literature review were mapped to the PMBOK Guide's RBS, which is suitable for categorizing risk sources. The Management risks dominate clearly, while external risks hardly play a role.

Table 1. Data science risks mapped to the RBS acc. to PMBOK

RBS Level 1	RBS Level 2	Frequency
Management Risk	Project management	90
Management Risk	Resourcing	16
Management Risk	Organization	6
Management Risk	Operations management	15
Management Risk	Communication	8
Total Management Risks		**135**
Technical Risk	Data	144
Technical Risk	Estimates, assumptions and constraints	22
Technical Risk	Technology	99
Technical Risk	Technical processes	1
Technical Risk	Requirements definition	4
Total Technical Risks		**111**
External Risk	Competition	1
External Risk	Regulatory	1
Total External Risks		**2**
Total of Risks		**248**

Subsequently, the risk categories were mapped to the RBS so that only 29 risk categories were included in the survey rather than 248 single risks. This means that the management risks include insufficient project management, poor team management, bad customer expectation management, poor data and information management, poor communication with customer/stakeholders, uncertainty about project outcome, organizational culture as barrier and lack of data science competence/skills; the technical risks contain data security and privacy, poor data availability and quality, data inconsistency and incompleteness, data ownership unclear, high level of complexity, inaccuracy of the model, inadequate user management, lack of requirements management, lack of domain knowledge, poor documentation, inadequate maintenance planning, insufficient infrastructure, new technology, operational risks, little technical development/deployment practices; the external risks include market risks (i.e., competition) and regulatory risks (i.e., strengthening data protection).

For the analysis, only participants (N = 42) who completed the survey in its entirety were considered. Following questions on general information such as profession and company size, participants were asked whether a risk assessment is conducted at the beginning of projects in their company. Of the respondents, 23.81% agreed, 35.71% disagreed, 38.10% partially agreed, and 2.38% did not provide an answer.

Regarding the relevance of risk assessments for data science projects, 16.67% of the participants rated it as having low relevance, 50% as medium relevance, 28.57% as high relevance, and 4.76% did not respond.

Participants were then asked for their opinions on four specific questions:

1. On the question of whether performing risk assessments during data science projects would reduce project failures, 40.48% tended to agree, 35.71% agreed, 16.67% were neutral, 4.76% tended to disagree, and 2.38% disagreed.
2. When asked if performing risk assessments would increase risk awareness, 64.29% agreed, 26.19% tended to agree, 4.76% tended to disagree, and 4.76% were neutral.
3. Regarding whether risk assessments would sensitize team members to potential risks, 54.76% agreed, 38.10% tended to agree, and 7.14% were neutral.
4. Finally, on whether risk assessments would enable project managers to react more quickly to potential risks, 42.86% agreed, 42.86% tended to agree, 7.14% were neutral, and 7.14% tended to disagree.

While risk assessments are not universally implemented at the beginning of data science projects, their perceived relevance and benefits are acknowledged by the majority. Participants generally believe that risk assessments can reduce project failures, increase risk awareness, sensitize team members, and enable quicker responses to risks. Given the high perceived benefits, there is a clear opportunity for more widespread and consistent implementation of risk assessments in data science projects.

4.2.1. RQ1: To what extent can the risks identified in the literature be validated in practice through the survey?

The analysis and comparison of the survey and literature review data reveal several key insights into the risks associated with data science projects (Figure 1). Firstly, data inconsistency and incompleteness emerge as the highest-rated risk in the survey data, with 38 instances (8.9%). This highlights a significant practical concern that is not as prominently mentioned in the literature, which only cites it twice (0.92%). This discrepancy suggests that practitioners encounter this issue more frequently than what is reflected in academic sources. Secondly, insufficient project management is the most frequently cited risk in the literature, with 87 instances (39.91%), compared to 35 instances (8.2%) in the survey data. This indicates that while this risk is recognized in both domains, it is emphasized more in academic discussions.

The survey data also highlights poor customer expectation management as a major concern, with 30 instances (7.03%), significantly higher than the 4 mentions (1.83%) in the literature. This suggests that managing customer expectations is a critical practical issue that may be underrepresented in academic discussions.

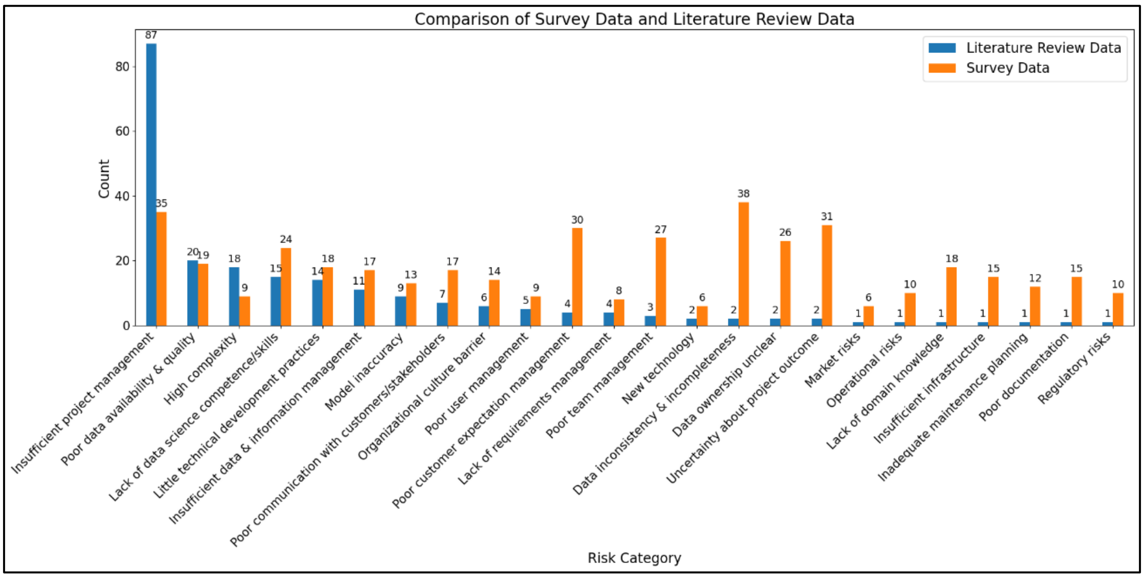

Figure 1. Comparison of Survey Data and Literature Review Data

Both sources identify the lack of data science competences and skills as a significant risk, though it is slightly more prevalent in the survey data (24 instances, 5.62%) compared to the literature (15 instances, 6.88%). Poor data availability and quality are consistently recognized as important risks in both sources, with the literature citing it 20 times (9.17%) and the survey data 19 times (4.45%). This consistency underscores the importance of data quality in the success of data science projects High complexity is another risk more frequently mentioned in the literature (18 instances, 8.26%) than in the survey data (9 instances, 2.11%), suggesting that complexity is a well-recognized challenge in academic research.

Little technical development practices are noted as significant in both sources, with the literature mentioning it 14 times (6.42%) and the survey data 18 times (4.22%). Insufficient data and information management are also identified as notable risks in both the literature (11 instances, 5.05%) and survey data (17 instances, 3.98%). Model inaccuracy is acknowledged as a risk across both sources, with the literature citing it 9 times (4.13%) and the survey data 13 times (3.04%). Organizational culture barriers are more frequently highlighted in the survey data (14 instances, 3.28%) than in the literature (6 instances, 2.75%), indicating its practical significance. In summary, while there is a general agreement between the literature and survey data on key risks, the survey data brings additional emphasis on specific practical concerns such as data inconsistency, customer expectation management, and organizational culture barriers. This comprehensive identification of risks across both sources underscores the complexity and multifaceted nature of managing data science projects, highlighting the need for robust and adaptive risk management strategies.

4.2.2. RQ2: To what extent can GenAI contribute to minimizing the validated risks?

GenAI can significantly contribute to minimizing these risks in data science projects. It can be used to automate the data cleaning process to address the risk of data inconsistency and incompleteness. For this purpose, Lan et al. [2023] propose GenAI-DAA, a data completeness enhancement algorithm based on GenAI [11]. Moreover, GenAI models can generate synthetic datasets to provide comprehensive datasets for analysis and model training [31].

To address the risk of insufficient project management, GenAI can decompose complex tasks into manageable steps, generate project requirements from past projects, create initial project plans, simulate potential risk scenarios based on historical data, monitor project progress and performance metrics in real-time to provide early warnings about potential issues, facilitate better communication with stakeholders through advanced Natural Language Processing (NLP) Tools and chatbots, and automate documentation to ensure better project management practices [10].

Regarding the risk of poor technical development and deployment practices, GenAI can automate and optimize key aspects of Continuous Integration and Continuous Delivery (CI/CD), ensuring seamless code integration and accelerating the deployment process with minimal manual intervention [32]. By utilizing AI-powered tools to monitor and manage the CI/CD pipeline, organizations can significantly enhance the quality, speed, and reliability of software development and deployment [32].

4.3. Survey Limitations

One of the primary limitations of this study is the relatively small number of survey participants, which impacts the overall generalizability of the findings. With only 42 complete responses, the sample size is not statistically representative of the broader population of professionals in the field of data science. This limitation may introduce bias and restrict the extent to which the results can be extrapolated to other contexts or industries.

The small sample size also limits the ability to conduct more robust statistical analyses, such as segmentation by industry or company size, which could have provided deeper insights into how risks vary across different settings. Additionally, the limited number of participants means that the findings should be interpreted as indicative trends rather than definitive conclusions about the prevalence and impact of the identified risks in practical environments.

Despite these limitations, the survey still provides valuable insights into the real-world relevance of risks identified in the literature, highlighting areas where further research with larger and more diverse samples could be beneficial. The trends observed in this study can serve as a starting point for more extensive investigations into the risks associated with data science projects and the role of Generative AI in mitigating these risks.

5. Conclusion

This study sought to determine whether the risks identified in previous literature on data science projects are also observed in practice, using a survey to gather empirical data. The survey, based on risks previously identified through a comprehensive literature review, provides valuable insights into how these risks manifest in real-world scenarios. Despite the small sample size, the survey results show a strong alignment with the literature, indicating that major risks such as insufficient project management, poor data availability and quality, and a lack of data science competences and skills are indeed prevalent in practical settings.

Additionally, the survey brought to light practical concerns that are less emphasized in the literature, such as data inconsistency, incompleteness, and poor customer expectation management. These findings suggest that while the theoretical risks discussed in the literature are relevant, practical applications introduce additional nuances.

GenAI holds substantial potential to address and minimize these risks in data science projects. For insufficient project management, GenAI can decompose complex tasks into manageable steps, monitor project progress and performance metrics in real-time, provide early warnings about potential issues, and automate documentation to enhance project management practices. Additionally, GenAI can analyze project requirements and historical data to optimize resource allocation, ensuring the right skills and tools are available when needed. Regarding data quality and consistency, GenAI can automate data cleaning and preprocessing, generate synthetic datasets to fill gaps, and improve overall data integrity. To manage customer expectations effectively, GenAI-powered natural language processing (NLP) tools and chatbots can facilitate better communication with stakeholders.

Furthermore, GenAI can enhance technical development practices by assisting in code generation and optimization, automating parts of the CI/CD pipeline, and generating comprehensive documentation. It can proactively identify emerging risks, simulate various scenarios to understand potential impacts, and develop mitigation strategies, thus improving risk awareness and management.

The findings from this study highlight the significant overlap between theoretical and practical risks, while also identifying areas where GenAI can play a crucial role in risk mitigation. By modernizing data management practices and addressing specific challenges in data science projects, GenAI can enhance the overall success and efficiency of data science initiatives. Despite the limitations of the survey, this study provides actionable insights and recommendations for practitioners, illustrating how GenAI can be integrated into a robust risk management framework that addresses both theoretical and practical concerns in data science projects.

6. References

[1] J. Westenberger, K. Schuler, D. Schlegel, Failure of AI projects: understanding the critical factors, Procedia Computer Science 196 (2022) 69–76. https://doi.org/10.1016/j.procs.2021.11.074.

[2] VentureBeat, Why do 87% of data science projects never make it into production?, 2019. https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/ (accessed 21 May 2024).

[3] R. Bean, T.H. Davenport, Companies Are Failing in Their Efforts to Become Data-Driven, 2019. https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven (accessed 27 June 2024).

[4] M. Holtkemper, M. Potanin, Oberst, Alexander, Beecks, Christian, Risk Identification of Data Science Projects: A Literature Review, in: LWDA 2023, Marburg, Germany, CEUR Workshop Proceedings, Marburg, 2023, pp. 1–13.

[5] D. Wang, J. Andres, J.D. Weisz, E. Oduor, C. Dugan, AutoDS: Towards Human-Centered Automation of Data Science, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama Japan, ACM, New York, USA, 2021, pp. 1–12.

[6] M. Abdelaal, R. Koparde, H. Schoening, AutoCure: Automated Tabular Data Curation Technique for ML Pipelines, in: Proceedings of the Sixth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management, Seattle WA USA, ACM, New York, NY, USA, 2023, pp. 1–11.

[7] M. Bilal, G. Ali, M.W. Iqbal, M. Anwar, M.S.A. Malik, R.A. Kadir, Auto-Prep: Efficient and Automated Data Preprocessing Pipeline, IEEE Access 10 (2022) 107764–107784. https://doi.org/10.1109/ACCESS.2022.3198662.

[8] F. Smaldone, A. Ippolito, J. Lagger, M. Pellicano, Employability skills: Profiling data scientists in the digital labour market, European Management Journal 40 (2022) 671–684. https://doi.org/10.1016/j.emj.2022.05.005.

[9] Z. Epstein, A. Hertzmann, M. Akten, H. Farid, J. Fjeld, M.R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, A.S. Pentland, O. Russakovsky, H. Schroeder, A. Smith, Art and the science of generative AI, Science 380 (2023) 1110–1111. https://doi.org/10.1126/science.adh4451.

[10] K. Bainey, AI-driven project management: Harnessing the power of artificial intelligence and ChatGPT to achieve peak productivity and success, John Wiley & Sons Inc, Hoboken, New Jersey, 2024.

[11] G. Lan, S. Xiao, J. Yang, J. Wen, M. Xi, Generative AI-based Data Completeness Augmentation Algorithm for Data-driven Smart Healthcare, IEEE J. Biomed. Health Inform. PP (2023). https://doi.org/10.1109/JBHI.2023.3327485.

[12] J. Xu, J. Li, Z. Liu, N.A.V. Suryanarayanan, G. Zhou, J. Guo, H. Iba, K. Tei, Large Language Models Synergize with Automated Machine Learning, 2024.

[13] W. van der Aalst, Data Science in Action, in: W. van der Aalst (Ed.), Process Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016, pp. 3–23.

[14] L. Cao, Data Science: Profession and Education, IEEE Intell. Syst. 34 (2019) 35–44. https://doi.org/10.1109/MIS.2019.2936705.

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

[16] D. Conway, The Data Science VENN Diagram, 2010. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (03/07/2024).

[17] J. Aßmann, J. Sauer, M. Schulz, Don't Be Afraid of Failure—Insights from a Survey on the Failure of Data Science Projects, in: T. Barton, C. Müller (Eds.), Apply Data Science, Springer Fachmedien Wiesbaden, Wiesbaden, 2023, pp. 65–76.

[18] P. Mikalef, M.N. Giannakos, I.O. Pappas, J. Krogstie, The human side of big data: Understanding the skills of the data scientist in education and industry, in: 2018 IEEE Global Engineering Education Conference (EDUCON), Tenerife, IEEE, 2018, pp. 503–512.

[19] M. Schulz, U. Neuhaus, J. Kaufmann, D. Badura, S. Kuehnel, W. Badwitz, D. Dann, S. Kloker, E.M. Alekozai, C. Lanquillon, Introducing DASC-PM: A Data Science Process Model, in: ACIS Proceedings.

[20] A. Schuckart, Introduction to work with GenAI, in: Proceedings of the 28th European Conference on Pattern Languages of Programs, Irsee Germany, ACM, New York, NY, USA, 2023, pp. 1–16.

[21] V. Savov, ChatGPT Could Be AI's iPhone Moment, 2022. https://www.bloomberg.com/news/newsletters/2022-12-12/chatgpt-the-gpt-3-chatbot-from-openai-microsoft-is-tech-magic (accessed 4 July 2024).

[22] U. Gal, ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned., 2023. https://theconversation.com/chatgpt-is-a-data-privacy-nightmare-if-youve-ever-posted-online-you-ought-to-be-concerned-199283 (accessed 4 July 2024).

[23] S. Amer-Yahia, A. Bonifati, L. Chen, G. Li, K. Shim, J. Xu, X. Yang, From Large Language Models to Databases and Back: A Discussion on Research and Education, SIGMOD Rec. 52 (2023) 49–56. https://doi.org/10.1145/3631504.3631518.

[24] C.E. Smith, K. Shiekh, H. Cooreman, S. Rahman, Y. Zhu, M.K. Siam, M. Ivanitskiy, A.M. Ahmed, M. Hallinan, A. Grisak, G. Fierro, Early Adoption of Generative Artificial Intelligence in Computing Education: Emergent Student Use Cases and Perspectives in 2023, in: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, Milan Italy, ACM, New York, NY, USA, 2024, pp. 3–9.

[25] A. Vadaparty, D. Zingaro, D.H. Smith IV, M. Padala, C. Alvarado, J. Gorson Benario, L. Porter, CS1-LLM: Integrating LLMs into CS1 Instruction, in: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, Milan Italy, ACM, New York, NY, USA, 2024, pp. 297–303.

[26] A. Chakraborty, A. Banerjee, S. Dasgupta, V. Raturi, A. Soni, A. Gupta, S. Harsola, V. Subrahmaniam, Navigator: A Gen-AI System for Discovery of Factual and Predictive Insights on Domain-Specific Tabular Datasets, in: Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD), Bangalore India, ACM, New York, NY, USA, 2024, pp. 528–532.

[27] C. Beasley, A. Abouzied, Pipe(line) Dreams: Fully Automated End-to-End Analysis and Visualization, in: Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics, Santiago AA Chile, ACM, New York, NY, USA, 2024, pp. 1–7.

[28] E. Mazaheri, M. Lagzian, Z. Hemmat, Research Directions in information Systems Field, Current Status and Future Trends, AJIS 24 (2020). https://doi.org/10.3127/ajis.v24i0.2045.

[29] J. Recker, Scientific Research in Information Systems, Springer International Publishing, Cham, 2021.

[30] Project Management Institute, A guide to the project management body of knowledge: PMBOK guide., sixth ed., PMI global standard, Pennsylvania, USA, 2017.

[31] C.Y. Chan, Data-Driven Innovation: The Potential of Synthetic Data through Generative AI. Bachelor of Engineering, 2024.

[32] P. Sharma, M.S. Kulkarni, A study on Unlocking the potential of different AI in Continuous Integration and Continuous Delivery (CI/CD), in: 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India, IEEE, 2024, pp. 1–6.