Journal of Information Technology Education: Research

An Official Publication of the Informing Science Institute
InformingScience.org

JITEResearch.org

Volume 22, 2023

DATA SCIENCE EDUCATION – A SCOPING REVIEW

Nkosikhona Theoren Msweli* University of Pretoria, Pretoria, South Africa u19401958@tuks.co.za

Tendani Mawela University of Pretoria, Pretoria, South Africa tendani.mawela@up.ac.za

Hossana Twinomurinzi University of Johannesburg, Johannesburg, South Africa hossanat@uj.ac.za

* Corresponding author

ABSTRACT

Aim/Purpose This study aimed to evaluate the extant research on data science education (DSE) to identify the existing gaps, opportunities, and challenges, and make recommendations for current and future DSE.

Background There has been an increase in the number of data science programs especially because of the increased appreciation of data as a multidisciplinary strategic resource. This has resulted in a greater need for skills in data science to extract meaningful insights from data. However, the data science programs are not enough to meet the demand for data science skills. While there is growth in data science programs, they appear more as a rebranding of existing engineering, computer science, mathematics, and statistics programs.

Methodology A scoping review was adopted for the period 2010–2021 using six scholarly multidisciplinary databases: Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and the AIS Basket of eight journals. The study was narrowed down to 91 research articles and adopted a classification coding framework and correlation analysis for analysis.

Contribution We theoretically contribute to the growing body of knowledge about the need to scale up data science through multidisciplinary pedagogies and disciplines as the demand grows. This paves the way for future research to understand which programs can provide current and future data scientists the skills and competencies relevant to societal needs.

Findings The key results revealed the limited emphasis on DSE, especially in non-STEM (Science, Technology, Engineering, and Mathematics) disciplines. In addition, the results identified the need to find a suitable pedagogy or a set of pedagogies

Accepting Editor Aaron M. Glassman │Received: May 19, 2023│ Revised: July 4, 2023 │
Accepted: July 10, 2023.
Cite as: Msweli, N. T., Mawela, T., & Twinomurinzi, H. (2023). Data Science Education – A Scoping Review. *Journal of Information Technology Education: Research, 22,* 263-294. https://doi.org/10.28945/5173

because of the multidisciplinary nature of DSE. Further, there is currently no existing framework to guide the design and development of DSE at various education levels, leading to sometimes inadequate programs. The study also noted the importance of various stakeholders who can contribute towards DSE and thus create opportunities in the DSE ecosystem. Most of the research studies reviewed were case studies that presented more STEM programs as compared to non-STEM.

Recommendations for Practitioners

We recommend CRoss Industry Standard Process for Data Mining (CRISP-DM) as a framework to adopt collaborative pedagogies to teach data science. This research implies that it is important for academia, policymakers, and data science content developers to work closely with organizations to understand their needs.

Recommendations for Researchers

We recommend future research into programs that can provide current and future data scientists the skills and competencies relevant to societal needs and how interdisciplinarity within these programs can be integrated.

Impact on Society

Data science expertise is essential for tackling societal issues and generating beneficial effects. The main problem is that data is diverse and always changing, necessitating ongoing (up)skilling. Academic institutions must therefore stay current with new advances, changing data, and organizational requirements. Industry experts might share views based on their practical knowledge. The DSE ecosystem can be shaped by collaborating with numerous stakeholders and being aware of each stakeholder's function in order to advance data science internationally.

Future Research

The study found that there are a number of research opportunities that can be explored to improve the implementation of DSE, for instance, how can CRISP-DM be integrated into collaborative pedagogies to provide a fully comprehensive data science curriculum?

Keywords

data science applications in education, pedagogy, teaching/learning strategies, transdisciplinary projects, data science education

INTRODUCTION

Data science offers actionable insights by mining structured and unstructured data using statistical and computational tools and methods to identify patterns. It is a growing field impacting various sectors, genres, and disciplines, and therefore places the spotlight on data science education (DSE) (Van Dusen et al., 2019). DSE is an umbrella term used to describe learning programs meant to equip data scientists with data science competencies and skills mainly from computer science, mathematics, statistics, engineering, psychology, and the domain of interest. This multidisciplinary nature of data science programs means that DSE is an integration of knowledge, methodologies, or techniques from different distinct disciplines into a unique and distinct discipline of its own. Nonetheless, data science is framed more as a Science, Technology, Engineering, and Mathematics (STEM) discipline (McMaster et al., 2011; Rosenthal & Chung, 2020; Twinomurinzi et al., 2022) with little emphasis on business domains.

The demand for data scientists with the appropriate skills is high (World Economic Forum, 2019) and is evident in the increasing number of data scientist job vacancies (Verma et al., 2019), and the mushrooming of many formal learning programs (at undergraduate and postgraduate levels) and short learning programs (Saltz, Armour, & Sharda, 2018). However, there is limited alignment between these learning programs; there is therefore a gap between academic data science and commercial data science (Berman et al., 2018). There are also inconsistencies among the existing learning programs. Organizing learning programs around data science process models has been suggested (Haynes et al., 2019; Jaggia et al., 2020).

The most consistent model for data science remains the CRISP-DM model (Saltz, 2021). CRISP-DM has been heavily adopted for data science projects and has been deemed useful in teaching data analytics (Jaggia et al., 2020; Kristoffersen et al., 2019). The major features of this model are its independence of technology and industry sectors (Ayele, 2020).

It is important to appreciate that data scientists support various sectors with a variety of data from different sources (Heinemann et al., 2018). Consequently, this raises the need to understand and create DSE curricula (Kross et al., 2020) that target all transdisciplinary competencies including practical skills that are linked to different domains (Dill-McFarland et al., 2021; Mokiy, 2019). The nature of data science demands different teaching and learning structures that are not constrained (Irizarry, 2020) but promote a collaborative environment to avoid teaching data science in silos (Mikroyannidis, Domingue, Bachler, & Quick, 2018). Nevertheless, the multiple disciplines that jointly form data science bring multiple opportunities and challenges to DSE (Danyluk et al., 2019).

There is therefore a growing call for standardising DSE (Heinemann et al., 2018; Irizarry, 2020), especially in the field of curriculum design (Chen, 2020; Finzer, 2013; Mikroyannidis et al., 2018; Song & Zhu, 2016). For instance, several academic workshops (panel sessions) and conferences have been hosted with the intent to discuss data science curriculum design (i.e., Danyluk et al., 2019; Howe et al., 2017; Mikroyannidis, Domingue, Phethean, et al., 2018; Oh et al., 2019; Van Dusen et al., 2019). However, these are still developing opportunities that might introduce some beneficial recommendations to improve DSE. Therefore, the following research question was formulated to understand the status of DSE:

> *How has DSE been investigated, and what are the gaps, opportunities for, and challenges associated with DSE?*

We theoretically contribute to the growing body of knowledge about the need to democratize data science, making it accessible to a broader range of individuals through multidisciplinary pedagogies and disciplines as the demand grows. Democratizing data science refers to the efforts aimed at making data science accessible and inclusive to a wider audience. Traditionally, data science has been associated with technical expertise and specialized skills, which have limited its accessibility. However, democratization seeks to break down these barriers and empower more people to participate in and benefit from the field of data science. This paves the way for future research to understand which programs can provide current and future data scientists the skills and competencies relevant to societal needs.

This study adopted a scoping review methodology to assess the status of DSE research since 2010 with specific attention paid to articles describing DSE, opportunities available in DSE, and challenges faced by DSE. The remainder of the paper is structured as follows: after presenting the methodology adopted and discussing the findings, this scoping review concludes with conclusions, implications, limitations, and areas for further research.

RESEARCH METHODOLOGY

Scoping reviews are conducted with the intent to identify pertinent published studies that address a specific research question. The primary purpose is to synthesize a body of knowledge related to the phenomena of interest (Siddaway et al., 2019). The sections that follow elaborate further on the broad criteria considered when conducting this scoping review.

Protocol

The study adopted the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to maintain transparency in reporting the findings of this study (Knobloch et al., 2011; Shamseer et al., 2015). To improve study quality and minimize biases, the inclusion and exclusion criteria were established *a priori* as suggested by Nightingale (2009).

Eligibility Criteria

The eligibility criteria were set as studies and academic reports published during the 12 years from 2010 to 2021. Papers reporting on working groups and panel sessions, and out of scope, were not included. Only papers published in the English language were eligible for inclusion.

Search Words and Data Sources

The keywords and data sources used to search for relevant and authoritative research papers for the systematic literature review are listed in Table 1.

Table 1. Search keywords and data sources

Search keywords	Data sources
DSE	Google Scholar
Big DSE	IEEE Xplore
Data Science Curriculum	ACM Digital Library
Data Science Curricular	ScienceDirect
Data Science Program	Scopus
Data Analytics Education	**AIS basket of eight journals**
Data Science Training	*Information Systems Journal*
Data Mining Education	*Journal of the Association for Information Systems*
Knowledge Discovery	*Journal of Information Technology*
Data Science Learning	*Journal of Management Information Systems*
	Journal of Strategic Information Systems
	Management Information Systems Quarterly
	European Journal of Information Systems
	Information Systems Research

Searching Process

The following search string was used across the data sources to retrieve the papers from the various sources listed in Table 1.

> ("Data science*" OR "Big Data*" OR "Data mining*" OR "Data analytics*" OR "Knowledge discovery*") AND ("Education*" OR "Curriculum*" OR "Training*" OR "Program*" OR "Learning*")

Selection of Studies

The selection criteria of the relevant papers are dependent upon the research question. The systematic literature review employed the pre-defined selection criteria for the selection of papers to be included in the review (Table 2).

Table 2. Inclusion, exclusion, and quality assessment criteria

Inclusion criteria	Exclusion criteria	Quality assessment of studies
The research paper is peer-reviewed. The research is related to the search string and area of "DSE". Research published between 2010 and 2021. The selected study must be a full-length published paper. Research publications must be written in the English language.	Papers published before 2010. Unpublished studies. Papers are not written in English. Papers not related directly to the research question (i.e., opportunities and barriers reported on DSE/curriculum).	As part of the quality assessment, each study was checked against the following questions: Is the study in relation to DSE? Does the paper provide a clear statement of findings? Is the paper peer-reviewed? Is the paper published in a reputable source?

Figure 1 depicts the process followed to select the final list of peer-reviewed research articles for inclusion in the systematic literature review. The preliminary search denotes the number of papers retrieved (research hits) after running the search string. The first order of selection was based on the review of paper keywords, title, and abstract. In the second order of selection, all duplicated research papers were eliminated. In the third order of selection, all papers that did not meet the eligibility criteria were discarded. The articles were accepted based on the selection criteria outlined in Table 3.

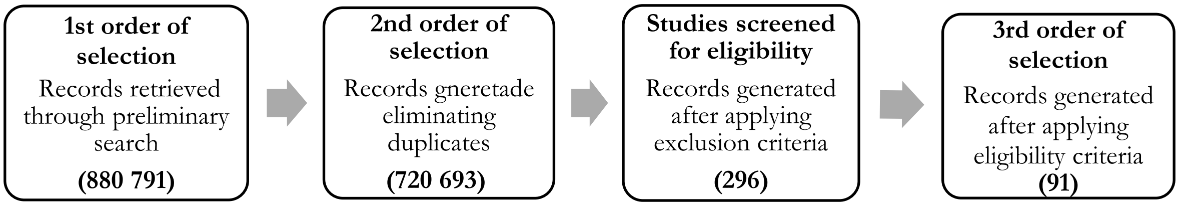
1st order of selection
Records retrieved through preliminary search
(880 791)
2nd order of selection
Records gneretade eliminating duplicates
(720 693)
Studies screened for eligibility
Records generated after applying exclusion criteria
(296)
3rd order of selection
Records generated after applying eligibility criteria
(91)

Figure 1. Order of selection of the papers for the systematic literature review

Table 3. Selection criteria

Acceptable in 1st order of selection	Acceptable in 2nd order of selection	Acceptable in 3rd order of selection
Abstract and keywords are accessible. Acceptable study types are journal or conference papers (peer-reviewed). Language is English Study publication date is within the 2010-2021 period.	Abstract and keywords are accessible. Acceptable study types are journal or conference papers (peer-reviewed). Paper is written in the English language. Study publication date is within the 2010-2021 period. Studies are unique (not duplicates).	Full text of the article is accessible. Acceptable study types are journal or conference papers (peer-reviewed). Paper is written in the English language. Study publication date is within the 2010-2021 period. The study focuses on DSE or is within the scope. Studies are unique (not duplicates).

DATA COLLECTION PROCESS

The systematic literature review of the selected research papers was conducted based on the inclusion and exclusion criteria. The data collection process was conducted during the period from July 2021 to September 2021. The data collection process was monitored and reviewed by the co-authors of this study. Data extraction included demographic details, origin (continent), methodology, focus, and other aspects.

FRAMEWORK FOR ANALYZING THE ELIGIBLE PAPERS

After collecting the eligible research papers, the study applied the classification and coding framework of Amui et al. (2017) to provide structure to the existing body of knowledge around the phenomena of interest. As shown in Appendix A, this framework uses numerical and letter codes to categorize the chosen papers.

FINDINGS

A descriptive and correlation analysis was performed to understand the relationships between the different classes tabled in Appendices A and B. A statistical correlation analysis was included because the number of papers reviewed was enough to draw statistical and inferential insights into the strengths and directions of relationships between the different aspects of DSE. Due to space constraints, only highly significant inferences are discussed.

DISTRIBUTION OF DSE PUBLICATION BY CONTINENT

The initial analysis focused on the distribution of the articles according to the regional geographical location or area in which the selected DSE articles were published. Figure 2 shows the distribution of the selected DSE articles based on the continent in which the studies were conducted.

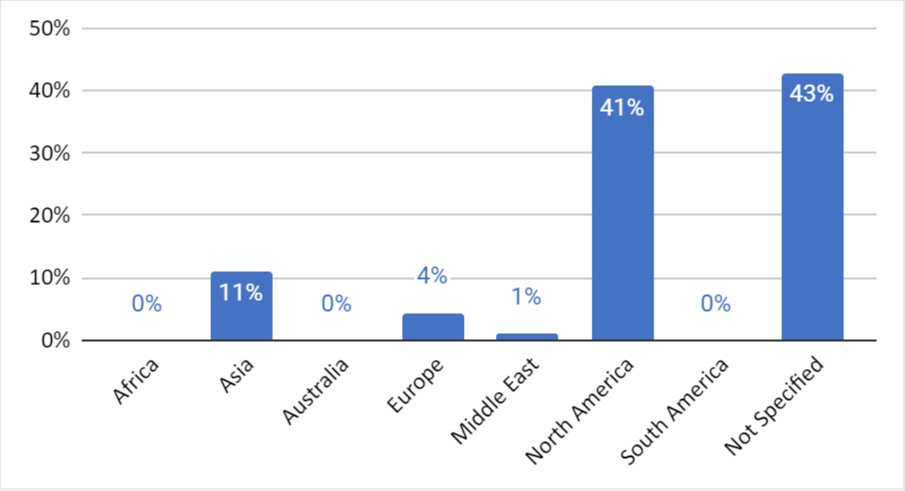

Figure 2. Paper distribution by continent

As can be seen from Figure 2, North America is the main contributor to DSE research, with 27 publications being published; this accounts for 41% of the papers published during the period 2010 to 2021. A similar trend has been observed in other studies (Farahi & Stroud, 2018; Hassan & Liu, 2020). A slightly higher number of studies (43%) did not specifically indicate the country of origin of their publications. The low number of DSE articles published in Asia (11%), Europe (4%), and the Middle East (1%) suggests that limited DSE research is being carried out in these continents. The low number of DSE papers emanating from Europe (4%) and Asia (11%) are surprising

(Mikroyannidis et al., 2018). No studies were recorded for Africa, Australia, and South America during the same period.

A correlation analysis (Appendix B) revealed that most of the research conducted in North America focused on project-based learning as a teaching strategy for DSE. North America is where most global technology companies such as Microsoft and Google (Luna et al., 2014) are located. Developing countries are not well-positioned to realize the need to derive benefits from data science (Hack-Polay et al., 2020; Shereni & Chambwe, 2020). Such countries often face various challenges such as poor infrastructure and the absence of skills thus putting the continents in which these countries are located at a disadvantage (Luna et al., 2014; Shereni & Chambwe, 2020; Takemura, 2018). It is also possible that the demand for data science has not advanced as much in these countries hence the limited research.

Classification According to (Based on) the Research Method

The use of appropriate research methods is important in any study to answer the research questions. Figure 3 shows that the research methods adopted in the selected papers ranged from experimental methods, surveys, action research, ethnography, and case studies to design science (design and creation) (Oates, 2006).

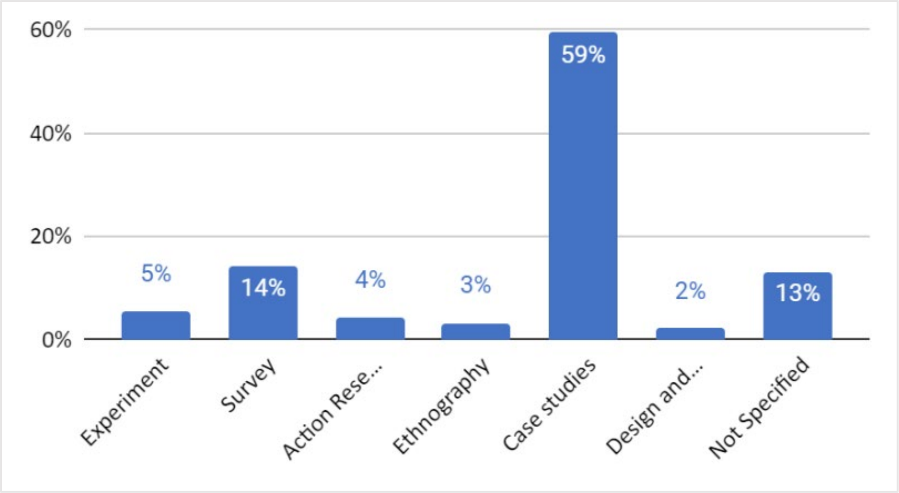

Figure 3. Article distribution based on the research methodology

The case study method was the most used research method (58%) followed by the survey method (14%), the experiment research method (5%), action research (4%), ethnography (3%), and design science (2%). According to Rowley (2002), case studies provide an appropriate platform to investigate emerging areas or projects that are in the exploratory phase. The preference for case studies as a research method suggests that DSE is indeed an emerging topic of interest.

It is evident from the findings of this scoping review that considerable attention was being paid to addressing the challenge of data science skills gap as several case studies reported on how modules can be adopted for use by data scientists (Buzydlowski, 2019; Çetinkaya-Rundel & Ellison, 2021; Facey-Shaw et al., 2018); other case studies focused on rebranding STEM courses (Bart et al., 2016; Buzydlowski, 2019; Kahn, 2020; Rao et al., 2018; Yadav & Debello, 2019). Only a single case study was reported that focused on non-STEM (Gil, 2014); another one was targeted at non-programmers (Jie et al., 2020). Furthermore, it is noteworthy that few trades involve DSE, such as the medical field (Garmire et al., 2017; Otero et al., 2014) and the engineering field (Qiang et al., 2019). The papers also recommended teaching practices and technologies suitable for DSE.

It is notable that none of the selected papers incorporated more than one research method. While case studies may be an appropriate method to investigate emerging areas or projects that are at the exploratory stage (Rowley, 2002), using a single method to investigate a particular problem may not be sufficiently rigorous (Chung et al., 2020). More specifically, multiple methods would offer different aspects of DSE; for example, the adequacy and effectiveness of DSE learning programs.

Analysis by Qualification Level at Which DSE is Offered

The two models available to the public for delivering content can be roughly classified as formal and non-formal. Formal DSE is offered from school to tertiary level, while informal DSE is often autodidactic. The distinction between formal and autodidactic depends on where they are offered and the content of the programs. Other DSE programs such as micro-credentials and short-learning programs can be offered formally or autodidactically at various levels. Figure 4 shows the various qualifications levels at which DSE is offered.

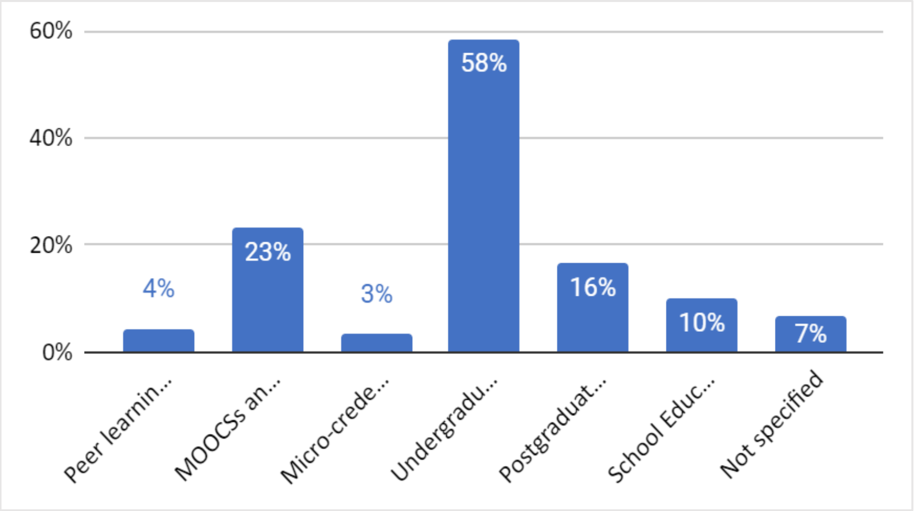

Figure 4. Article distribution per level of qualification

Figure 4 shows that the undergraduate program is the most researched DSE qualification making 58%. MOOCs and short learning programs were mentioned in less than half of the papers (23%), only, 16% and 10% of the reviewed papers mentioned postgraduate programs and school education as the most appropriate level for imparting DSE. The number of papers mentioning peer learning and micro-credentials for delivering DSE was significantly low at 4% and 3% respectively. Of the reviewed papers, 7% did not specify the level of qualification mooted for DSE. There is therefore an opportunity to consider different educational levels to introduce and offer DSE.

Analysis by DSE Discipline

Data science integrates different disciplines yet Figure 5 suggests that some disciplines appear to be more dominant than others, which makes it difficult to maintain the transdisciplinary trait of the learning program.

It is evident that the STEM (Science-Technology-Engineering-Mathematics) type disciplines account for the highest number of DSE research, mainly computer science (40%), followed by statistics (37%). About 33% of the publications did not specify the discipline of interest and instead just gave a broad description of "data science." For instance, Saltz, Dewar, and Heckman (2018) focus on teaching ethics in DSE while Wymbs (2016) looked at how data analytics can be incorporated into undergraduate business programs.

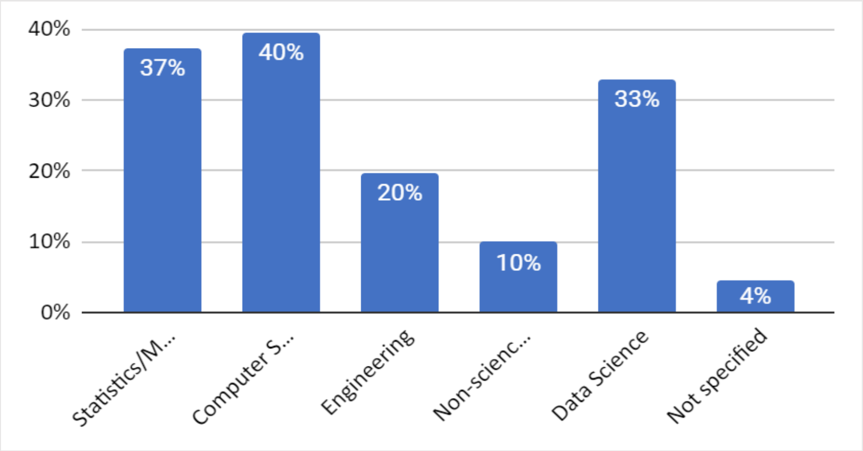

Figure 5. Article distribution per data science discipline

It was not clear how much of each element should be featured in DSE to balance the multidisciplinarity. However, the non-STEM domains appeared to be receiving less attention compared to other data science elements. This is possibly because DSE programs are often offered within the faculties of sciences and engineering (Gil, 2014), and little attention is given to the applicability to other disciplines.

The study showed a positive correlation between the engineering component of data science and the use of technology, presenting an opportunity for researchers and practitioners to propose strategies on how technology can be used to teach data science. This is informed by a lack of integrated platforms where students can develop hands-on experience (Zhang et al., 2017).

The study further showed that the CRISP-DM phase (Evaluation) and discipline-specific (non-STEM) domain are significantly correlated. This implies that rather than teaching students how to develop models, non-STEM education focuses on evaluating models to determine whether the suggested model is in line with the business objectives and actually solves the business problem.

Analysis of DSE Providers

DSE is offered by various educational service providers including public institutions, private institutions, industry organizations, and through collaborative partnerships (Figure 6).

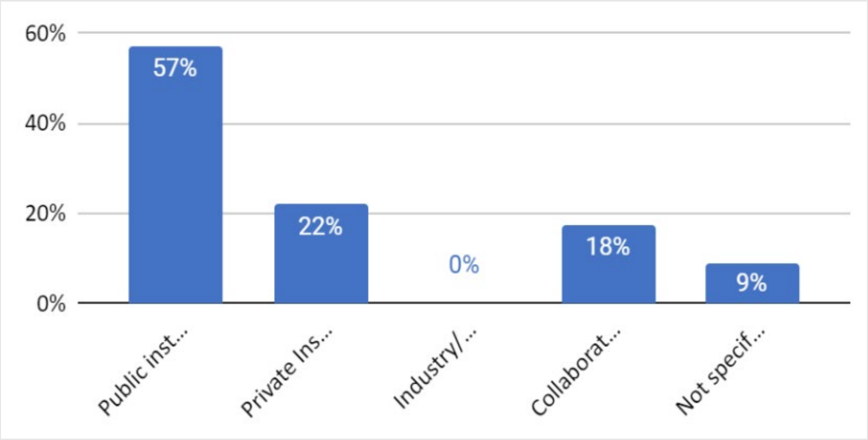

Figure 6. Article distribution per data science educational provider

As shown in Figure 6, most of the papers (57%) report data science programs being offered by public institutions of learning. Whereas 22% of the papers claim that DSE is offered by private institutions of learning, only 18% have reported DSE collaborative initiatives involving both the public and private sectors.

First, public institutions benefit from the teaching and learning funding model to support DSE through sourcing qualified lecturers and conducting research related to data science and DSE. For instance, Demchenko et al. (2017) presented a data science course that was funded by the European Commission, and Heinemann et al. (2018) presented a data science education for secondary schools that was funded by Deutsche Telekom Stiftung.

Second, public institutions typically have extensive interfaculty support systems in place, as well as external support from other institutions (Huppenkothen et al., 2018). Collaboration with international institutions is an important aspect of the external support system because it enables to access information and resources that are normally not easily accessible, and thus be part of ongoing studies that cover new trends in data science (X. Li et al., 2019).

Third, there is a high preference for public university qualifications among students over those offered by private institutions. Public universities are in a better position to implement DSE, however, there are challenges concerning peer learning results showing a negative correlation. A contributing factor may be the absence of policies that encourage and acknowledge peer learning as well as well-researched and widely accepted methods of student assessment.

The results further showed that public institutions of higher learning have less interest in MOOCs and short-learning programs. Private institutions have recently shifted their focus towards offering more data science short learning programs, MOOCs, and badges due to the high demand for data science programs. However, there is a need to focus more on the quality and relevance of the learning content rather than the number of programs offered. Collaboration may increase opportunities for developing collaborative DSE that captures the interests of various stakeholders.

Analysis of the Data Science Context Using the CRISP-DM Model

This study examined CRISP-DM as the most consistent transdisciplinary framework to guide data science projects and teaching. The purpose was to determine the extent to which DS programs are aligned with the model (Jaggia et al., 2020). Figure 7 shows the paper distribution across phases of the CRISP-DM model.

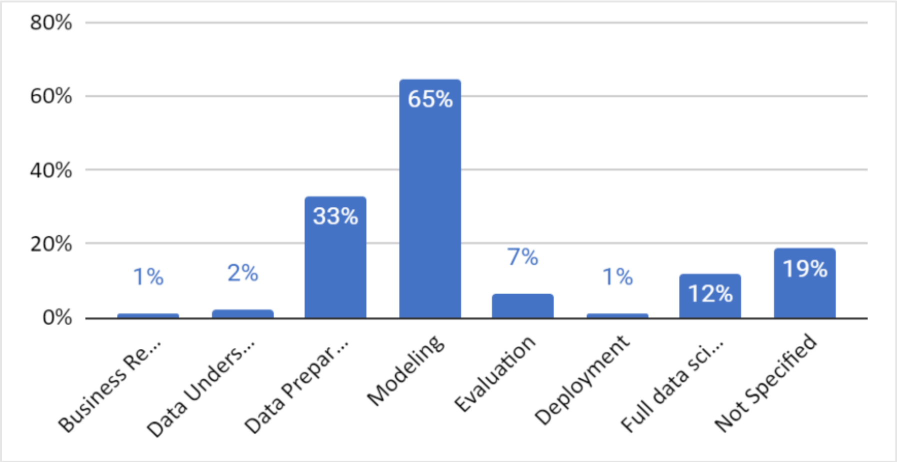

Figure 7. Article distribution per data science context based on CRISP-DM

According to Figure 7, an overwhelming majority of the papers (65%) emphasize the inclusion of data modeling in DSE, while slightly over half of the selected papers (33%) are inclined toward data preparation. Furthermore, only 12% of the selected papers appear to punt the inclusion of all phases of CRISP-DM in DSE. Evaluation was mentioned in 7% of the papers, while data understanding was mentioned in 2% of the papers. The idea of including evaluation (7%), data understanding (2%), and business understanding and deployment (1%) components in DSE does not appear to be favored by many researchers. A substantial number of papers (19%) did not express any preference for the inclusion of any specific CRISP-DM phase in DSE.

These results suggest that current DSE research does not give priority to all the CRISP-DM phases, and this affects the inclusion of these phases in the data science curriculum and limits the development of data science skills amongst students. These findings support Gil's (2014) argument that DSE focuses more on databases and machine learning contexts neglecting other elements.

Data science is applied across various industries; therefore, data scientists need to master the technical skills of data science (i.e., data mining and analysis, machine learning, and others) as well as business skills (i.e., marketing, data products, and others) (Qiang et al., 2019). Data science specialists should be able to participate in the whole data science lifecycle, mimicking the CRISP-DM model (Donoghue et al., 2021). Without these skills, organizations are deprived of the opportunity to use data to create a competitive advantage and to make smart decisions.

The results also demonstrated that business requirements as part of DSE can be offered through collaborations (Paul & Aithal, 2018). This finding suggests the existence of an opportunity for different stakeholders to work together and develop data science modules that focus on business understanding as the first phase in data science projects. This will allow data scientists to develop competencies to participate in the business requirement-gathering process and understand the business or economic side of data science before they can proceed with data wrangling. Organizations' focal points vary; therefore, collaboration with and amongst these organizations creates a setting where objectives and interests are shared while engaging a transdisciplinary DSE.

Analysis of Teaching Strategies

Data science is transdisciplinary and is therefore expected to adopt various teaching methods and tools. Figure 8, however, shows that there are mainly four teaching strategies adopted for DSE, namely competency-based learning (54%), use of technology (53%), teacher-led (49%), and project-based learning (44%). Flipped classroom (14%), student-led learning (9%), personalized learning (7%), and inquiry-based learning (4%) are not as popular. Only 13% of the papers did not mention any teaching strategy.

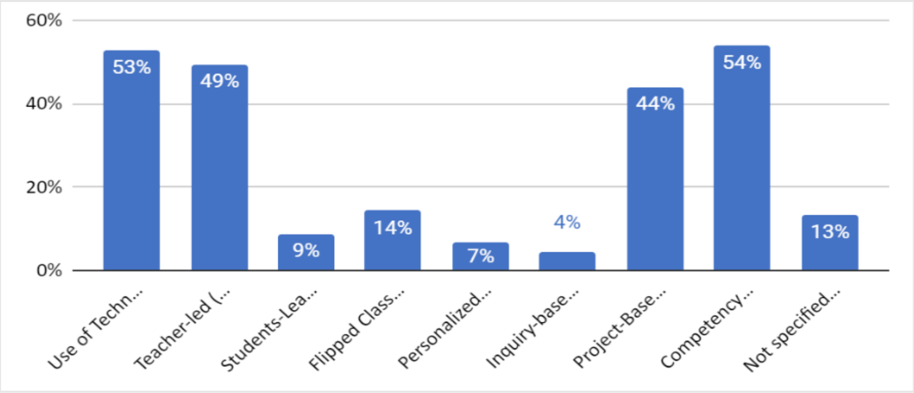

Figure 8. Distribution of publications by teaching strategies

There is a need to investigate more instructional approaches that will enable students to easily understand difficult concepts within DSE. For instance, the correlational analysis showed that flipped classrooms (especially pre-recorded videos) are mostly applied in micro-credential courses. Flipped or flexible classrooms and micro-credentials rank amongst the top new developments changing the education system (Klašnja-Milićević et al., 2017).

Strategies for improving DSE

The literature suggests that DSE opportunities have unfortunately not fully emerged (Finzer, 2013). Figure 9 shows the distribution of papers as per outlined opportunities based on the existing literature.

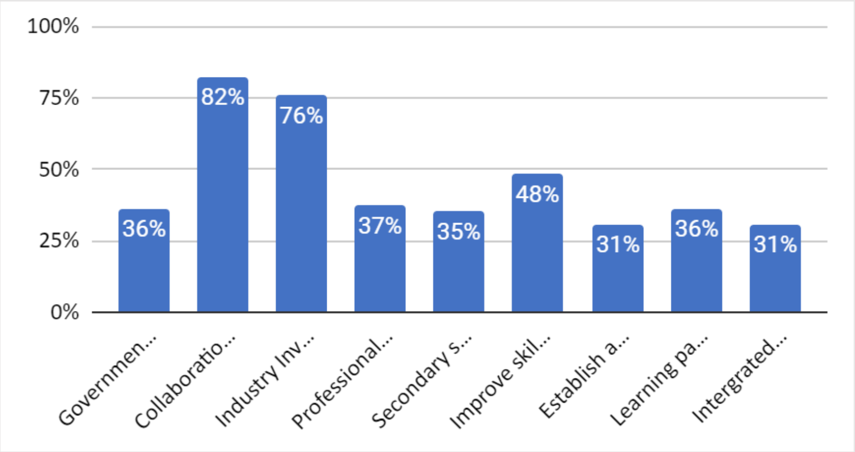

Figure 9. Article distribution as per outlined opportunities

Collaboration Between Universities (82% of the Papers)

Based on the data presented, the collaboration amongst university faculties presents an opportunity for the implementation of DSE. Some of the opportunities presented by collaboration amongst universities include sharing resources, such as lecturers, in cases where there is a lack of skills and capacity. Attwood et al. (2019) noted that it is often difficult to find suitably qualified candidates for DSE lecturing posts. Mostly, the technical aspects of data science can be crucial and can be immensely beneficial when more resources are available (Cleveland, 2001). Therefore, collaboration among universities can accelerate the creation of an environment where data science exists as a cross-campus endeavor that involves faculties and students in different departments (Van Dusen et al., 2019).

An opportunity also exists for universities to make available educational data that can be shared across different disciplines. However, such a collaborative approach will require regulating standards (both local and international) to address issues of ethics, security, and privacy (Daniel, 2019).

Involvement of Industry (76% of the Papers)

Organizations that have both data and data science skills have a competitive edge (Takemura, 2018), and understand the needs problems (Cybulski & Scheepers, 2021). Involvement and collaboration with organizations can provide academic institutions with some perspectives in terms of linking the teaching and learning content with real business scenarios. Furthermore, this provides opportunities for students and lecturers to access data for simulation purposes. With that, students get to be well prepared for the real working environment and organizations can recruit from a pool of well-qualified data scientists. Most importantly, DSE programs can be designed and developed with input from both nationally and internationally renowned industry experts and leading practitioners (Demchenko et al., 2015).

Improve Skills at the Tertiary Level (48% of the Papers)

Whereas a high increase in data science programs is being experienced, the challenge of skills complement among lecturing staff at universities remains. It is difficult to teach specialized data science skills when lecturing staff members do not, at the very least, have experience in the field. This limitation hinders DSE offering, especially at various levels of tertiary education. Academics teaching at this level may not be able to demonstrate all the techniques that data science students need to acquire (Paul & Aithal, 2018; Song & Zhu, 2016). Essentially, improving the skills of academic staff members will allow full data science participation from the secondary school level, and at the tertiary level. Not only would the availability of these resources support teaching, but also the development and continued review of DSE.

Professional Advancement (37% of the Papers)

Data science as an emerging field does not have many qualified professionals available with the requisite experience (Mikalef et al., 2018). There is therefore a need for re-skilling and upskilling the capabilities of those involved in DSE. The rapid change in technology means the modeling techniques are also rapidly evolving. This means that data scientists must adapt relevant skillsets continuously to suit business requirements. Staying relevant in a changing world is rewarding but it can also be time-consuming (Çetinkaya-Rundel & Ellison, 2021). To remain relevant, DSE needs to be flexible and agile enough to accommodate future developments in data science tools, models, and technologies for data science.

Government Involvement (36% of the Papers)

Government entities can take part in the implementation of DSE in a variety of ways, such as making available data for educational purposes. Open data is valuable when educating students about data concepts and, where possible, providing them with real business stories (Saddiqa et al., 2021). Integrating real data sets within data science courses could enable the development of data science skills, such as data collection, cleaning, analysis, and interpretation. Government can also benefit from these initiatives. Initiatives towards open government data can guide innovation and improve service delivery and involve citizens in decision-making processes.

Learning Pathways (36% of the Papers)

With unpredicted changes in the future of work and evolving technology, it is important to consider how students progress from the time they enroll, how they progress with their studies, and how their careers become real and change beyond studies (Iatrellis et al., 2020; Lyon et al., 2015; Miller & Hughes, 2017). The multidisciplinarity of data science offers options as a path for specialization, such as data engineering, machine learning, and algorithm development. In addition, data science serves multiple fields and there are key players and different career pathways in each field (Misnevs & Yatskiv, 2016).

Secondary School Curriculum (35% of the Papers)

Introducing data science into a secondary school curriculum was identified as an opportunity in 35% of the selected publications. This may assist students in acquiring some substantive data science competencies at a foundation level. However, the challenge lies in integrating data science into secondary school subjects so that students develop data science skills and the conceptual understanding needed to participate fully in society as citizens and workers (Finzer, 2013).

There is also the challenge of teachers with computational and mathematical skills to transfer knowledge to young aspiring data scientists. The majority of teachers are not trained nor have experience in DSE (X. Li et al., 2019). For instance, teachers are having challenges with programming languages such as R and this affects their statistical analysis capabilities (Gould et al., 2016). The improvement of DSE at the school level by ensuring that teachers are trained in data analytics and have experience working with data can potentially advance data proficiency and awareness (Biehler et al., 2018).

Establishing a DSE Governing Body (36% of the Papers)

This consortium can serve as the advisory board for content creation and review where necessary. So far, there are no guiding frameworks for DSE, hence the inconsistencies in the learning programs. Within organizations, some problems can be addressed fully or can be moderately solved, or automated through data science (Cybulski & Scheepers, 2021). These developments should be communicated with DSE institutes so that learning programs can focus more on the areas that cannot be automated. This calls for a guided process which can be achieved by having a governing body or a framework for implementation. Research on how this can be implemented is now significant. This can be local or global or, even better, it can be through a collaboration.

Integrated Digital Platforms (36% of the Papers)

Institutions need to investigate the implementation of integrated digital platforms for effective data science programs. Platforms add value by allowing students to have a simulated project, share resources, and execute data analysis. Cloud-based technologies are also a valuable tool for teaching data science, as they are quick to set up and allow an intuitive environment.

Analysis of Challenges in DSE

This category is aimed at identifying the challenges in DSE. Designing a transdisciplinary curriculum and training data scientists pose several challenges (Mikroyannidis et al., 2018). There were 11 themes on the challenges in DSE (Figure 10).

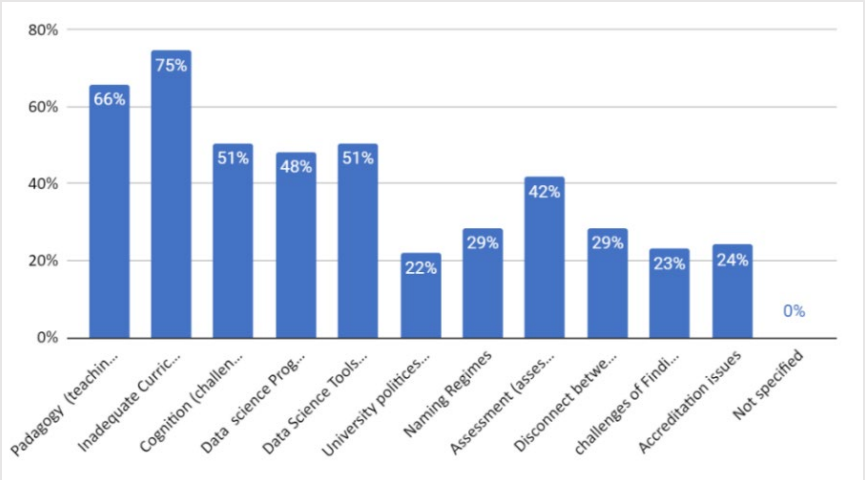

Figure 10. Article distribution based on DSE challenges

Inadequate Curriculum in DSE

The inadequate curriculum appears to be a major challenge in DSE and was noted by 75% of the publications. It is therefore quite clear that the issue of addressing inadequate curriculum in DSE is very critical. For instance, not developing data scientists with the competencies and skills to understand the domain as well as the business context presents an extra cost for organizations. Although data science is focused on statistical and computational thinking, it is also applied to solving domain-specific problems (Blei & Smyth, 2017). Therefore, it may prove difficult for data scientists to link data science outputs with organizational objectives. Inadequate curricula lead to inadequate competencies. While extant literature has also demonstrated a lack of consistency in DSE, specific recommendations to address these issues are scant.

Teaching Pedagogies

Challenges related to teaching pedagogies were highlighted by 66% of the selected papers It is noted that teaching modern data scientists is a challenge (Mikroyannidis et al., 2018; Oudshoorn et al., 2020). In sentiments shared by scholars, data skills cannot be taught using only direct instruction (Hardin et al., 2015; Mike, 2020; Takemura, 2018). Project-based pedagogies have been mentioned as one of the appropriate pedagogy for teaching data scientists (Donoghue et al., 2021; Saltz & Heckman, 2016; Takemura, 2018). Other teaching practices have been applied to promote data skills, such as gamification (Hee et al., 2016), and social student events like hackathons and datathons (Anslow et al., 2016; Huppenkothen et al., 2018). The common features among the mentioned teaching practices are that they are student-centered, and enforced hands-on learning that integrates real business scenarios and data (A. Y. Kim et al., 2018), and the ability to scale up data science (Donoghue et al., 2021). Topics on teaching pedagogies are not often initiated, yet so many individuals who graduate proceed to take teaching roles (Cleveland, 2001).

Cognitive Skills (Understanding of Challenging Concepts)

A lack of cognitive skills was mentioned as a challenge in 51% of the papers reviewed. In general, the reviewed papers pointed out statistics, mathematics, and programming as being challenging subjects where students have to apply their minds when solving problems that apply to these concepts.

Data Science Tools (Model Misuse, Misinterpretation of Models)

A significant number of the papers (51%) mentioned challenges associated with data science tools in DSE. As organizations adopt data science for various business practices, the models must be used appropriately to make practical predictions and well-informed business decisions (Blei & Smyth, 2017). Competencies and skills to work with data platforms, models, and tools to develop and operate data analytics applications effectively are of great significance and should be part of DSE (Wiktorski et al., 2019).

Data Science Program Structure

The structural issues of data science programs were mentioned in 48% of the papers. The findings of this study complement prior studies on DSE that have continuously mentioned the design of DSE as a problem (Clayton & Clopton, 2018; Cybulski & Scheepers, 2021; Twinomurinzi et al., 2022). Currently, only computer sciences and engineering dominate the current structure of DSE (Paul & Aithal, 2018). The dominance may indicate that universities are simply producing data scientists who are computer scientists with no real transdisciplinary expertise (Xia & Li, 2020). It needs to be understood that each industry has different needs, and they explore data science in different ways.

Assessment Issues

While there are various strategies for acquiring data science skills, assessing and validating competency remains a challenge. This challenge was mentioned in 42% of the papers. For instance, students can take part in hackathons or datathons where intensive learning opportunities and skills development exists (Dill-McFarland et al., 2021; Huppenkothen et al., 2018; Msweli, 2023). Although these events expose students to real-world data, it is often difficult to assess and validate the competency of the candidate in various areas of data science.

Naming Regimes

An estimated 29% of the papers registered the challenge that comes with diverse names of data science programs. The inconsistencies in data science program structures affect the identification of these programs (Saltz, Armour, & Sharda, 2018). For instance, Havill (2019) used "Data Analytics" instead of "Data Science" in learning programs to attract a diverse pool of students. Pettis et al. (2018) referred to the same as big data analytics programs, and Jafar et al. (2016) used "data analytics" to refer to both data and business analytics. All these programs differ in terms of programming competence and the degree of statistical abilities expected from students (Saltz, Armour, & Sharda, 2018).

The Disconnect Between Industry Practice and Data Science Learning Material

Based on the analysis of the reviewed papers, there is no shared framework for DSE. This makes the growth of data science learning programs unfocused due to the absence of agreed learning outcomes (Raj et al., 2019). This important element was mentioned in 29% of the papers. The implications of such a disconnection result in data education being driven from one side (often by the industry) (Farahi & Stroud, 2018). Having specific learning outcomes and competencies could help stakeholders such as lecturers, employers, and policy-makers, to have a mutual understanding of the specific skills, competencies, and knowledge that data science students should acquire (J. Kim, 2015).

Accreditation

The accreditation of data science learning programs is lacking and challenging (D. Li et al., 2021). Based on the reviewed papers, the accreditation issues were mentioned in 24% of the papers. As an emerging discipline, there is a mutual understanding that the DSE guidelines and the accreditation criteria are still under development; therefore, it can be assumed that the existing data science programs are built on emerging standards (Oudshoorn et al., 2020).

Finding Organizations Willing to Participate

While the involvement of industry in DSE can bring some structure and insights on relevant content, it is difficult to find industries that are willing to participate in curriculum development (Bohler et al., 2017); this was mentioned in 23% of the papers. To become involved in DSE, organizations will need to avail resources such as practitioners, specialists, or infrastructure. It is not easy to convince organizations to buy into developing initiatives where benefits are not guaranteed (Iatrellis et al., 2020). With the diversity of data science functions across different industries, the nature of skills and competencies required in each function also varies (Radovilsky et al., 2018) . Therefore, organizations need to participate in DSE to ensure the connection between education and competencies needed in the working environment.

University Politics

The transdisciplinary nature of data science exacerbates collaboration challenges. These challenges do not only exist in the workplace but also within tertiary institutions (Anderson et al., 2015; Finzer, 2013). Not only are the faculties affected by these university politics, but lecturers and students as individuals as well. Twenty-two percent (22%) of the papers confirmed the existence of politics within universities and faculties. For instance, with the high demand for data scientists, there is a risk of faculties losing their students to the data science field of study. This could result in an over-population of data scientists who do not appreciate the importance of other disciplines (Baumer, 2015). Several cases have been reported where transversal competencies and skills are not considered of primary importance (Demchenko, Wiktorski, et al., 2019; Gkamas et al., 2019; Takemura, 2018). The conflict between IT specialists and domain experts is usually caused by incongruities in their respective skill sets, processes, and terminologies which become a problem when training students to become data scientists. No study made suggestions on how this can be addressed.

Analysis of DSE Stakeholders

Generally, stakeholders can affect or be affected by business practices or policies. These practices can be internal or external, have interests, and can play various roles in organizations. Considering the nature of data science, it is important to identify the key stakeholders who can stand together to build DSE. Figure 11 shows the number of papers and their focus on different stakeholders.

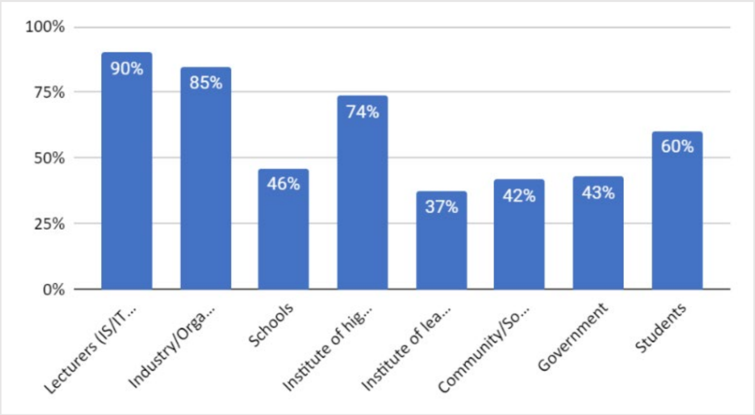

Figure 11. Article distribution per DSE stakeholders

It is observed from Figure 11 that each paper had more than one stakeholder representation, with lecturers having the highest number of representations (90%) followed by industry/organization (85%). Public institutes of higher learning were mentioned in 74% of the papers, while citations of private institutions of learning and schools were significantly lower at 37% of papers and 46% of the papers, respectively. In addition, students, government, and community were mentioned respectively in 60%, 43%, and 42% of the papers. The article distribution by DSE stakeholders shows a high interest in data science programs interest coming from different stakeholders, and thus suggest the importance of investigating and understanding the role of each stakeholder in DSE to maintain the transdisciplinary status.

Discussion

DSE is a growing academic area that is not being explored, especially in developing countries. Many developing countries, particularly in Africa, face various challenges that can put them at a disadvantage in the global economy. Poor infrastructure, internet connectivity, and affordability can make it difficult for businesses to operate efficiently and for individuals to access education and training opportunities (Malaka & Brown, 2015). Additionally, the absence of skilled employees in key areas like data science can limit a country's ability to innovate and compete in the global marketplace. Addressing these challenges requires a multi-faceted approach that includes investments in infrastructure, education and training programs, and policies that encourage economic growth and innovation.

With the little research that has been conducted, case studies are used to investigate DSE, often for an in-depth examination of a particular instance of DSE. The majority of the reviewed case studies focus on undergraduate programs, intending to redesign the current computer science and statistics curriculum to create programs in data science. Other research methods such as experimental studies, interviews, and face-by-depth information collection can be used to assess the effectiveness of different DSE interventions and to collect in-between DSE students' experiences. Different research methods are needed to explore different contexts of DSE.

The high number of studies on undergraduate-level programs could perhaps be resulting from data science undergraduate degrees being based on existing curricula. Nonetheless, undergraduate programs provide a solid ground for complex data science concepts (X. Li et al., 2019). There is also an opportunity for the integration of DSE at the foundational level. Countries in Europe and the Middle East are also in favor of this (Mikroyannidis et al., 2018; Takemura, 2018). This aim is to allow learners to develop and grasp the soft and cognitive skills that students need as they advance their careers in the data science field. Other scholars argue for postgraduate DSE qualifications (Cao, 2019; Hosack & Sagers, 2015; Paul & Aithal, 2018) noting the importance of the research component and advanced skills (Hassan & Liu, 2020; Shamir, 2020). Where there are no clear learning outcomes for each level of learning, a framework is needed to guide the structuring of programs, then the curriculum designers can decide on what skills need to be attained at a specific level. This is to further avoid overlaps. There is also a suggestion for DSE short learning programs to improve proficiency and accommodate new developments (Attwood et al., 2019; Garmire et al., 2017; Otero et al., 2014). These may include micro-credentials; however, the concept is still new and in need of proper conceptualization for effective usage. There is also no evidence of a framework that guides the structuring and development of these programs.

While multidisciplinarity is key in DSE (Twinomurinzi et al., 2022), the learning programs often focus on scientific domains (such as computer science and statistics) without looking at domain-specific areas like medicine, and finance among others. These sentiments have been shared by a number of scholars (Bohler et al., 2017; Schwab-McCoy et al., 2021). The inclusion of science and non-science disciplines is crucial to offer a balanced data science program. In addition, students in transdisciplinary programs need to be provided with opportunities to work together and gain knowledge from peers and professionals from various professions. This helps students develop a broader perspective and improves their capacity to collaborate across disciplines. There are a few initiatives that present such opportunities such as datathons and hackathons (Huppenkothen et al., 2018).

While public institutions have been taking the lead in data science offerings, the industry is not exploring data science programs. Many industries are looking into hiring candidates that already have data science skills rather than developing the skill in-house. Based on unique business needs, it is necessary to understand the influence industry has on data science programs, and how the industry can collaborate with other stakeholders in DSE. In-house training or micro-credentials can typically be tailored to the specific needs of the business and can be an effective way for employees to gain practical experience in data science while working on real-world problems (Msweli et al., 2022). When selecting a data science learning program, individuals consider the factors such as the program's cost, duration, and content, as well as the reputation of the provider and the availability of job placement services. Understanding the purpose of teaching data science and the intended audience is important.

The transdisciplinary nature expected in DSE has been ignored for the more technical component. Yet, in reality, these aspects are becoming much more accessible while the "business aspects" are what require a great deal of adaptation (Bohler et al., 2017). In the context of teaching data science, the CRISP-DM framework can be effectively used to give students an organized method of approaching data analysis (Heinemann et al., 2018). It is important to establish how CRISP-DM phases can be incorporated into a data science curriculum. For example, instructors can offer direction and assistance to students at any point in the process to help them comprehend the significance of each phase and how they all work together to generate insightful or accurate forecasts. DSE can be organized and allow for transdisciplinary inclusion of the non-technical aspects of data science by adopting the CRISP-DM framework into training and teaching pedagogies.

With regard to teaching data science, it is not clear which pedagogies are suitable for this field, especially the pedagogies that embrace transdisciplinary learning. As an emerging discipline, there is still a debate as to what content needs to be presented and how it should be presented (Cao, 2017; Shulman, 1986). Considering that data science is presented to a diverse group of students, teaching practices need to consider the targeted audience and their background. Essentially, technological, pedagogical, and content knowledge is necessary to understand how teaching practices influence the way students perceive DSE (Gudmundsdottir & Shulman, 1987). In addition, transdisciplinary learning needs to be encouraged together with additional teaching resources that may support DSE (Schwab-McCoy et al., 2021).

It is clear that DSE inherits some challenges from other disciplines especially those within STEM (Twinomurinzi et al., 2022). Below is a summary of challenges that need to be addressed as part of supporting data skills supply:

- *Absence of policy on resource and data sharing.* Ethics and privacy issues are some of the barriers to data sharing. Even though these issues exist, the lack of awareness of data science benefits especially among government makes it difficult for them to see the need for policies that support data sharing, in particular the public data. Data science, being an emerging discipline, availability of resources is a challenge. This includes teaching resources (i.e., learning content and qualified instructors). Since data science seems to be more technical and complex, it requires qualified and experienced human resources to teach in this field. As a new discipline, very few qualified individuals can teach data science concepts (Msweli, 2023). Resource sharing may be one of the solutions, however, it can only be achieved if there is an agreement among key stakeholders.

- *Lack of transdisciplinary teaching pedagogies.* New tools are continuously being developed to transform the data science landscape. Accordingly, data science teaching practices need to be reimagined. Acquiring data science skills needs to be supported by teaching practices that encourage continuous learning. Little knowledge is available on how this can be achieved. However, instructors in this field should have pedagogical content knowledge (Mike, 2020; Msweli, 2023).

- *Teaching diverse audiences.* Currently, DSE attracts students from different backgrounds, and preparing data science classes needs to consider these differences, particularly for students with minimal cognitive skills. In the discipline of data science, cognitive abilities including critical thinking, problem-solving, and decision making are crucial for success (Demchenko, Comminiello, & Reali, 2019). However, it is common for data science students to be lacking in these abilities, especially if they are new to the program. Despite their background, the student should be provided with hands-on experience that targets data analysis and visualization, exposure to real-world problems, and training in critical thinking and problem-solving. This will help them become effective contributors to the growing field of data science.

- *Standardization and inconsistencies are critical issues in DSE.* The lack of recognized standards for DSE can result in variations in the caliber and scope of data science programs at various learning institutions. Absence of a professional advisory board or accrediting body for data science programs it is difficult to say which disciplines are underrepresented or overrepresented (Schwab-McCoy et al., 2021). Establishing standards or guidelines for data science curricula can give institutions a framework to work within when creating their data science programs.

Data science is a transdisciplinary field that combines expertise from various areas such as statistics, mathematics, computer science, and domain-specific knowledge. Data analytics is also applied to various business and non-business domains (Bohler et al., 2017). Key stakeholders need to work together in building DSE. Establishing a solid ecosystem that supports both the technical and non-technical aspects of DSE is necessary. There is very little literature that focuses on DSE, particularly on the potential influence that different stakeholders may have on democratizing DSE, and data policy.

Conclusion, Implications, Limitations, and Areas for Further Research

This paper presents a scoping review of the status of DSE research, and the selected papers were classified and coded using a classification coding framework. The development of the data science field has prompted academia to see prospects of how to introduce different DSE programs to support the training of data scientists. Despite the growth in data science programs some gaps need to be investigated, and research into DSE is not advancing at the required pace.

The results reveal an emerging influential field that is fragmented. The fragmentation lies in the inconsistencies of DSE programs, types of programs, and teaching pedagogies. The multidisciplinarity of data science, much like information systems, makes it challenging to have a consistent curriculum. The information systems field has managed to build professional and academic bodies that have enabled it to have fairly standard curricula. We recommend a transdisciplinary professional body to guide curricula in data science. There are some which currently exist, but these mainly focus on STEM at the expense of non-STEM disciplines. The professional body would also assist with other important aspects such as naming conventions in data science because some areas of the discipline employ the same principles but use different names which is confusing for emerging data scientists.

The rapid change of technology today requires flexible curricula which therefore influences the pedagogies adopted in DSE. We found that project-based pedagogy is the dominant pedagogy in DSE, but we recommend a combination of pedagogies because of the multidisciplinary nature of the field. There have been some developments in teaching strategies and tools that improve the teaching of STEM subjects such as gamification and metaverse which have been shown to improve science education (Hee et al., 2016). These are some strategies that may be considered for DSE.

A pertinent question also remains about the regions that were noted as having very little or no investment in DSE research – what may be the implications of this on skills availability, potential brain drain, or opportunities for skills development?

We also suggest that there is a need for more research to be conducted on DSE as the main theme, with various research methodologies such as experiments, action research, ethnography, and design science being adopted. Scholars need to establish how they can apply different theories and philosophies when researching DSE. Research coalitions between countries, industry, and academia are also an important step for future studies in DSE to build the knowledge base and reference repository. Furthermore, an opportunity exists to investigate the data science skills and competencies applicable in each sector. Industry practitioners within various sectors can contribute by serving as advisory or review boards for academic institutions. This will offer a better understanding of the industry needs especially those in the non-STEM domains. Working with various stakeholders and understanding each stakeholder's role can shape the DSE ecosystem that can be shared globally to grow data science. The study identified a lack of balance concerning the inclusion of data science concepts. Concepts within STEM are put at the forefront, while research on business-related applications of data science is limited. Essentially, there is a need for researchers to compile guiding principles or develop frameworks that will guide how each element contributes to data science and how to ensure a balance of these across DSE programs.

DSE needs to serve various business practices and simulate CRISP-DM. We, therefore, recommend CRISP-DM as a framework to adopt collaborative pedagogies to teach DS. This research implies that it is important for academia, policymakers, and data science content developers to work closely with organizations to understand their needs. The primary issue is that the nature of data is diverse and changes at a rapid rate, thus demanding continuous (up)skilling. Essentially, academic institutions need to be up to date with new developments, evolving data, and organizational needs. Industry practitioners can offer insights based on their experience in the field.

This work provided a systematic and in-depth analysis of the existing literature on DSE, offering valuable insights into best practices, specifically highlighting the CRISP-DM framework and its significance in guiding data analysis and problem-solving in various domains. With that being said, this study contributes to the growing literature on DSE. The identification of challenges in DSE is a step towards building learning programs that are fit for purpose and address various stakeholders' needs. This paves the way for future research to understand which programs can provide current and future data scientists the skills and competencies relevant to societal needs.

Areas for Further Research

The study found that there are a number of research opportunities that can be explored to improve the implementation of DSE. Below are some of the questions proposed for future research:

Research global representation

a. How can DSE research be promoted in developing countries?
b. How can cross-continental DSE knowledge sharing be implemented?

Research methodology

a. How can multiple methods be incorporated into DSE research?

Levels of qualification

a. What is the impact of DSE offered as short learning programs?
b. How can DSE be introduced at pre-tertiary levels?

Transdisciplinary teaching pedagogy

a. How CRISP-DM can be integrated into collaborative pedagogies to provide fully comprehensive DS curricula?

Collaboration

a. What is the impact of DSE programs that are jointly developed between academics and practice?
b. How can collaboration be fostered across the disparate disciplines of DSE?
c. How can industry/practitioners be encouraged to share datasets for DSE?
d. What role does government policy play in opening data for DSE?

DSE curriculum and governance

a. How can we conceptualize an effective DSE curriculum for higher education?
b. What experiences or preparation do lecturers need for teaching and learning in DSE?
c. What are the elements of a sustainable DSE ecosystem?
d. What would a DSE accreditation framework look like?

The disconnect between practical application and data science learning material

a. How can DSE meet the needs of organizations and society at large?
b. How can we equip data scientists with the skills and tools for reasoning with various types of data?

Limitations

This study was limited to DSE research published between 2010 and 2021. The search was limited to the following databases: Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and a Basket of eight IS Journals. The study was initiated from an information systems perspective and as such databases focusing on psychology and education reviews were not included owing to the scope and target audience of the paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Amui, L. B. L., Jabbour, C. J. C., de Sousa Jabbour, A. B. L., & Kannan, D. (2017). Sustainability as a dynamic organizational capability: A systematic review and a future agenda toward a sustainable transition. *Journal of Cleaner Production*, *142*, 308–322. https://doi.org/10.1016/j.jclepro.2016.07.103

Anderson, P. E., Turner, C., Dierksheide, J., & McCauley, R. (2015, October). An extensible online environment for teaching data science concepts through gamification. *Proceedings of the Frontiers in Education Conference, Madrid, Spain*, 1-8. https://doi.org/10.1109/FIE.2014.7044205

Anslow, C., Brosz, J., Maurer, F., & Boyes, M. (2016, February). Datathons: An experience report of data hackathons for data science education. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, *Memphis, Tennessee, USA*, 615–620. https://doi.org/10.1145/2839509.2844568

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, *20*(2), 398–404. https://doi.org/10.1093/bib/bbx100

Ayele, W. Y. (2020). Adapting CRISP-DM for idea mining: A data mining process for generating ideas using a textual dataset. *International Journal of Advanced Computer Science and Applications*, *11*(6). https://doi.org/10.14569/IJACSA.2020.0110603

Bart, A. C., Tibau, J., Tilevich, E., Shaffer, C. A., & Kafura, D. (2016, June). Implementing an open-access, data science programming environment for learners. *Proceedings of the IEEE 40th Annual Computer Software and Applications Conference*, *Atlanta, Georgia, USA,* 728–737. https://doi.org/10.1109/compsac.2016.132

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician*, *69*(4), 334–342. https://doi.org/10.1080/00031305.2015.1081105

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. S. (2018). Realizing the potential of data science. *Communication of the ACM*, *61*(4), 67–72. https://doi.org/10.1145/3188721

Biehler, R., Frischemeier, D., Podworny, S., Wassong, T., Budde, L., Heinemann, B., & Schulte, C. (2018). Data science and big data in upper secondary schools: A module to build up first components of statistical thinking in a data science curriculum. *Archives of Data Science, Series A*, *5*(1), 1–19. https://doi.org/10.5445/KSP/1000087327/28

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, *114*(33), 8689–8692. https://doi.org/10.1073/pnas.1702076114

Bohler, J., Krishnamoorthy, A., & Larson, B. (2017). The financial and non-financial aspects of developing a data-driven decision-making mindset in an undergraduate business curriculum. *Journal of Business Education & Scholarship of Teaching*, *11*(1), 85–96.

Buzydlowski, J. W. (2019, March). Hip, hip, array: Teaching programming for data science is the same as computer science – Just different. *Proceedings of the 9th IEEE Integrated STEM Education Conference,* 308–310. https://doi.org/10.1109/ISECon.2019.8882101

Cao, L. (2017). Data science: Challenges and directions. *Communications of the ACM*, *60*(8), 59–68. https://doi.org/10.1145/3015456

Cao, L. (2019). Data science: Profession and education. *IEEE Intelligent Systems*, *34*(5), 35–44. https://doi.org/10.1109/MIS.2019.2936705

Çetinkaya-Rundel, M., & Ellison, V. (2021). A fresh look at introductory data science. *Journal of Statistics and Data Science Education*, *29*(1), 16–26. https://doi.org/10.1080/10691898.2020.1804497

Chen, A. (2020). High school data science review: Why data science education should be reformed. *Harvard Data Science Review*, *2*(4), 1–5. https://doi.org/10.1162/99608f92.1e28ce9e

Chung, S. H., Ma, H. L., Hansen, M., & Choi, T. M. (2020). Data science and analytics in aviation. *Transportation Research Part E: Logistics and Transportation Review*, *134*, 101837. https://doi.org/10.1016/j.tre.2020.101837

Clayton, P. R., & Clopton, J. (2018). Business curriculum redesign: Integrating data analytics. *Journal of Education for Business*, *94*(1), 57-63. https://doi.org/10.1080/08832323.2018.1502142

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. https://doi.org/10.1111/J.1751-5823.2001.TB00477.X

Cybulski, J. L., & Scheepers, R. (2021). Data science in organizations: Conceptualizing its breakthroughs and blind spots. *Journal of Information Technology*, *36*(2), 154–175. https://doi.org/10.1177/0268396220988539

Daniel, B. K. (2019). Big data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, *50*(1), 101–113. https://doi.org/10.1111/bjet.12595

Danyluk, A., Cassel, L., Leidig, P., & Servin, C. (2019, February). ACM task force on data science education draft report and opportunity for feedback. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, *Minneapolis, MN, USA,* 496–497. https://doi.org/10.1145/3287324.3287522

Demchenko, Y., Belloum, A., deLatt, C., Loomis, C., Wiktorski, T., & Spekschoor, E. (2017, December). Customisable data science educational environment: From competences management and curriculum design to virtual labs on-demand. *Proceedings of the IEEE International Conference on Cloud Computing Technology and Science, Hong Kong, China*, 363–368. https://doi.org/10.1109/CloudCom.2017.59

Demchenko, Y., Comminiello, L., & Reali, G. (2019, March). Designing customisable data science curriculum using ontology for data science competences and body of knowledge. *Proceedings of the International Conference on Big Data and Education, London, UK*, 124–128. https://doi.org/10.1145/3322134.3322143

Demchenko, Y., Gruengard, E., & Klous, S. (2015, February). Instructional model for building effective big data curricula for online and campus education. *Proceedings of the 6th International Conference on Cloud Computing Technology and Science, Singapore,* 935–941. https://doi.org/10.1109/CloudCom.2014.162

Demchenko, Y., Wiktorski, T., Cuadrado Gallego, J., & Brewer, S. (2019, September). EDISON Data Science Framework (EDSF) extension to address transversal skills required by emerging Industry 4.0 transformation. *Proceedings of the 15th International Conference on EScience*, *San Diego, CA, USA*, 553–559. https://doi.org/10.1109/eScience.2019.00076

Dill-McFarland, K. A., Konig, S. G., Mazel, F., Oliver, D. C., McEwen, L. M., Hong, K. Y., & Hallam, S. J. (2021). An integrated, modular approach to data science education in microbiology. *PLoS Computational Biology*, *17*(2). https://doi.org/10.1371/journal.pcbi.1008661

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education, 29*(sup1), S27-S39. https://doi.org/10.1080/10691898.2020.1860725

Facey-Shaw, L., Specht, M., & Bartley-Bryan, J. (2018, October). Digital badges for motivating introductory programmers: Qualitative findings from focus groups. *Proceedings of the Frontiers in Education Conference, San Jose, CA, USA,* 1-7. https://doi.org/10.1109/FIE.2018.8659227

Farahi, A., & Stroud, J. C. (2018, June). The Michigan Data Science Team: A data science education program with significant social impact. Proceedings of the *IEEE Data Science Workshop, Lausanne, Switzerland*, 120–124. https://doi.org/10.1109/DSW.2018.8439915

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, *7*(2). https://doi.org/10.5070/T572013891

Garmire, L. X., Gliske, S., Nguyen, Q. C., Chen, J. H., Nemati, S., Van Horn, J. D., Moore, J. H., Shreffler, C., & Dunn, M. (2017, January). The training of next generation data scientists in biomedicine. *Proceedings of the Pacific Symposium on Biocomputing, Hawaii, HI, USA,* 640–645. https://doi.org/10.1142/9789813207813_0059

Gil, Y. (2014, November). Teaching parallelism without programming: A data science curriculum for non-CS students. *Proceedings of the Workshop on Education for High Performance Computing, New Orleans, LA, USA,* 42–48. https://doi.org/10.1109/EduHPC.2014.12

Gkamas, V., Rigoum, M., Paraskevas, M., Zarouchas, T., Perikos, I., Vassiliou, V., Varbanov, P., Sharkov, G., Todorova, C., & Sotiropoulou, A. (2019). Bridging the skills gap in the data science and internet of things domains: A vocational education and training curriculum. *Proceedings of the ICDE World Conference on Online Learning*, 312–320.

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & Zanontian, L. (2016, July). Teaching data science to secondary students: The Mobilize Introduction to Data Science Curriculum. *Proceedings of the Roundtable Conference of the International Association of Statistics Education, Berlin, Germany*. https://doi.org/10.52041/SRAP.16402

Gudmundsdottir, S., & Shulman, L. (1987). Pedagogical content knowledge in social studies. *Scandinavian Journal of Educational Research*, *31*(2), 59–70. https://doi.org/10.1080/0031383870310201

Hack-Polay, D., Rahman, M., Billah, M. M., & Al-Sabbahy, H. Z. (2020). Big data analytics and sustainable textile manufacturing: Decision-making about the applications of biotechnologies in developing countries. *Management Decision*, *58*(8), 1699–1717. https://doi.org/10.1108/MD-09-2019-1323

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "think with data". *The American Statistician*, *69*(4), 343–353. https://doi.org/10.1080/00031305.2015.1077729

Hassan, I. B., & Liu, J. (2020, July). A comparative study of the academic programs between informatics/bioinformatics and data science in the U.S. *Proceedings 44th Annual Computers, Software, and Applications Conference, Madrid, Spain,* 165–171. https://doi.org/10.1109/COMPSAC48688.2020.00030

Havill, J. (2019, February). Embracing the liberal arts in an interdisciplinary data analytics program. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, *Minneapolis, MN, USA*, 9–14. https://doi.org/10.1145/3287324.3287436

Haynes, M., Groen, J., Sturzinger, E., Zhu, D., Shafer, J., & Mcgee, T. (2019, September). Integrating data science into a general education information technology course: An approach to developing data savvy undergraduates. *Proceedings of the 20th Annual SIG Conference on Information Technology Education, Tacoma, WA, USA,* 183–188. https://doi.org/10.1145/3349266.3351417

Hee, K., Zicari, R. V., Tolle, K., & Manieri, A. (2016, December). Tailored data science education using gamification. *Proceedings of the IEEE International Conference on Cloud Computing Technology and Science*, *Luxembourg City*, *Luxembourg*, 627–632. https://doi.org/10.1109/CloudCom.2016.0108

Heinemann, B., Opel, S., Budde, L., Schulte, C., Frischemeier, D., Biehler, R., Podworny, S., & Wassong, T. (2018, November). Drafting a data science curriculum for secondary schools. *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*, *Koli, Finland,* 1–5. https://doi.org/10.1145/3279720.3279737

Hosack, B., & Sagers, G. (2015). Applied doctorate in IT: A case for designing data science graduate programs. *Journal of the Midwest Association for Information Systems*, *1*(1), Article 5. https://doi.org/10.17705/3jmwa.00006

Howe, B., Franklin, M., Haas, L., Kraska, T., & Ullman, J. (2017, April). Data science education: We're missing the boat, again. *Proceedings of the IEEE 33rd International Conference on Data Engineering, San Diego, CA, USA,* 1473–1474. https://doi.org/10.1109/ICDE.2017.215

Huppenkothen, D., Arendt, A., Hogg, D. W., Ram, K., Vanderplas, J. T., Rokem, A., Designed, A. R., & Performed, A. R. (2018). Hack weeks as a model for data science education and collaboration. *PNAS*, *115*(36), 8872–8877. https://doi.org/10.1073/pnas.1717196115

Iatrellis, O., Savvas, I. K., Kameas, A., & Fitsilis, P. (2020). Integrated learning pathways in higher education: A framework enhanced with machine learning and semantics. *Education and Information Technologies*, *25*(4), 3109–3129. https://doi.org/10.1007/s10639-020-10105-7

Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review*, *2*(1), 1–8. https://doi.org/10.1162/99608f92.dd363929

Jafar, M. J., Babb, J., & Abdullat, A. (2016). Emergence of data analytics in the information systems curriculum. *Proceedings of the EDSIG Conference, Las Vegas, Nevada, USA*. http://proc.iscap.info/2016/pdf/4051.pdf

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2020). Applying the CRISP-DM framework for teaching business analytics. *Decision Sciences Journal of Innovative Education*, *18*(4), 612–634. https://doi.org/10.1111/dsji.12222

Jie, Z., Ying, J., & Ye, T. (2020, August). Exploration on the construction of big data basic course for non computer major. *Proceedings of the 15th International Conference on Computer Science and Education, Delft, Netherlands,* 623–626. https://doi.org/10.1109/ICCSE49874.2020.9201894

Kahn, J. (2020). Learning at the intersection of self and society: The family geobiography as a context for data science education. *Journal of the Learning Sciences*, *29*(1), 57–80. https://doi.org/10.1080/10508406.2019.1693377

Kim, A. Y., Ismay, C., & Chunn, J. (2018). The fivethirtyeight R package: "Tame Data" principles for introductory statistics and data science courses. *Technology Innovations in Statistics Education*, *11*(1). https://doi.org/10.5070/T5111035892

Kim, J. (2015). Competency-based curriculum: An effective approach to digital curation education. *Journal of Education for Library and Information Science*, *56*(4), 283–297. https://files.eric.ed.gov/fulltext/EJ1082900.pdf

Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2017). Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education*, *25*(6), 1066–1078. https://doi.org/10.1002/cae.21844

Knobloch, K., Yoon, U., & Vogt, P. M. (2011). Preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement and publication bias. *Journal of Cranio-Maxillofacial Surgery*, *39*(2), 91–92. https://doi.org/10.1016/j.jcms.2010.11.001

Kristoffersen, E., Aremu, O. O., Blomsma, F., Mikalef, P., & Li, J. (2019). Exploring the relationship between data science and circular economy: An enhanced CRISP-DM process model. In I. O. Pappas, P. Mikalef, Y. K. Dwivedi, L. Jaccheri, J. Krogstie, & M. Mäntymäki (Eds.), *Digital transformation for a sustainable society in the 21st century* (pp. 177–189). Springer. https://doi.org/10.1007/978-3-030-29374-1_15

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., & Leek, J. T. (2020). The democratization of data science education. *The American Statistician*, *74*(1), 1–7. https://doi.org/10.1080/00031305.2019.1668849

Li, D., Milonas, E., & Zhang, Q. (2021, August 20). Content Analysis of Data Science Graduate Programs in the U.S. *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--36841

Li, X., Fan, X., Qu, X., Sun, G., Yang, C., Zuo, B., & Liao, Z. (2019). Curriculum reform in big data education at applied technical colleges and universities in China. *IEEE Access*, *7*, 125511–125521. https://doi.org/10.1109/ACCESS.2019.2939196

Luna, D. R., Mayan, J. C., García, M. J., Almerares, A. A., & Househ, M. (2014). Challenges and potential solutions for big data implementations in developing countries. *Yearbook of Medical Informatics*, *23*(1), 36–41. https://doi.org/10.15265/IY-2014-0012

Lyon, L., Mattern, E., Acker, A., & Langmead, A. (2015). Applying translational principles to data science curriculum development. *Proceedings of the International Conference on Digital Preservation*. https://www.semanticscholar.org/paper/Applying-Translational-Principles-to-Data-Science-Lyon-Mattern/e4fd3239f0edac113870e431aeba4e30fe8c6b8c

Malaka, I., & Brown, I. (2015, September). Challenges to the organisational adoption of big data analytics: A case study in the South African telecommunications industry. *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists*, *Stellenbosch, South Africa*. https://doi.org/10.1145/2815782.2815793

McMaster, K., Sambasivam, S., Rague, B., & Wolthuis, S. (2011). A data science enhanced framework for applied and computational math. *Issues in Informing Science and Information Technology*, *15*, 191–206. https://doi.org/10.28945/4032

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018, April). The human side of big data: Understanding the skills of the data scientist in education and industry. *Proceedings of the IEEE Global Engineering Education Conference, Santa Cruz de Tenerife, Spain*, 503–512. https://doi.org/10.1109/EDUCON.2018.8363273

Mike, K. (2020, August). Data science education: Curriculum and pedagogy. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 324–325. https://doi.org/10.1145/3372782.3407110

Mikroyannidis, A., Domingue, J., Bachler, M., & Quick, K. (2018, October). Smart blockchain badges for data science education. *Proceedings of the IEEE Frontiers in Education Conference*, *San Jose, CA, USA,* 1–5. https://doi.org/10.1109/FIE.2018.8659012

Mikroyannidis, A., Domingue, J., Phethean, C., Beeston, G., & Simperl, E. (2018). Designing and delivering a curriculum for data science education across Europe. In M. Auer, D. Guralnick, & I. Simonics (Eds.), *Teaching and learning in a digital world* (pp. 540–550). Springer. https://doi.org/10.1007/978-3-319-73204-6_59

Miller, S. M., & Hughes, D. (2017). *The quant crunch: How the demand for data science skills is disrupting the job market.* Burning Glass Technologies. https://www.bhef.com/sites/default/files/bhef_2017_quant_crunch.pdf

Misnevs, B., & Yatskiv, I. (2016). Data science: Professional requirements and competence evaluation. *Baltic Journal of Modern Computing*, *4*(3), 441-453. https://www.researchgate.net/publication/311909668

Mokiy, V. (2019). International standard of transdisciplinary education and transdisciplinary competence. *Informing Science: The International Journal of an Emerging Transdiscipline*, *22*, 73–90. https://doi.org/10.28945/4480

Msweli, N. (2023). Instructors' perception of the competencies required to teach DS in HEI. In H. Twinomurinzi, N. Msweli, T. Mawela, & S. Thakur (Eds.), *Proceedings of the NEMISA Digital Skills Conference: Scaling data skills for multidisciplinary impact education*, *5*, 89–103.

Msweli, N. T., Twinomurinzi, H., & Ismail, M. (2022). The international case for micro-credentials for life-wide and life-long learning: A systematic literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, *17*, 151–190. https://doi.org/10.28945/4954

Nightingale, A. (2009). A guide to systematic literature reviews. *Surgery*, *27*(9), 381–384. https://doi.org/10.1016/j.mpsur.2009.07.005

Oates, B. J. (2006). New frontiers for information systems research: Computer art as an information system. *European Journal of Information Systems*, *15*(6), 617–626. https://doi.org/10.1057/palgrave.ejis.3000649

Oh, S., Song, I. Y., Mostafa, J., Zhang, Y., & Wu, D. (2019). Data science education in the iSchool context. *Proceedings of the Association for Information Science and Technology*, *56*(1), 558–560. https://doi.org/10.1002/PRA2.90

Otero, P., Hersh, W., & Ganesh, A. U. J. (2014). Big Data: Are biomedical and health informatics training programs ready? Contribution of the IMIA Working Group for Health and Medical Informatics Education. *Yearbook of Medical Informatics*, *23*(1), 177-181. https://doi.org/10.15265/IY-2014-0007

Oudshoorn, M. J., Titus, K. J., & Suchan, W. K. (2020, October). Building a new data science program based on an existing computer science program. *Proceedings of the IEEE Frontiers in Education Conference, Uppsala, Sweden*. https://doi.org/10.1109/FIE44824.2020.9273934

Paul, P. K., & Aithal, P. S. (2018). Computing academics into new age programs and fields: Big data analytics & data sciences in Indian academics – An academic investigation of private universities. *IRA – International Journal of Management & Social Sciences*, *10*(3), 107–118. https://doi.org/10.21013/jmss.v10.n3.p3

Pettis, C., Swamidurai, R., Abebe, A., & Shannon, D. (2018, April). Infusing big data concepts in undergraduate CS mathematics courses through active learning. *Proceedings of the IEEE SoutheastCon*, *St Petersburg, FL, USA*, 31–35. https://doi.org/10.1109/SECON.2018.8479114

Qiang, Z., Dai, F., Lin, H., & Dong, Y. (2019, August). Research on the course system of data science and engineering major. *Proceedings of the IEEE International Conference on Computer Science and Educational Informatization, Kunming, China*, 90–93. https://doi.org/10.1109/CSEI47661.2019.8938944

Radovilsky, Z., Hegde, V., Acharya, A., & Uma, U. (2018). Skills requirements of business data analytics and data science jobs: A comparative analysis. *Journal of Supply Chain and Operations Management*, *16*(1), 82-101.

Raj, R. K., Parrish, A., Impagliazzo, J., Romanowski, C. J., Aly Ahmed, S., Bennett, C. C., Davis, K. C., McGettrick, A., Pereira, T. S. M., & Sundin, L. (2019, July). Data science education: Global perspectives and convergence. *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education, Aberdeen, UK,* 265–266. https://doi.org/10.1145/3304221.3325533

Rao, A., Bihani, A., & Nair, M. (2018, October). Milo: A visual programming environment for data science education. *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing, Lisbon, Portugal,* 211–215. https://doi.org/10.1109/VLHCC.2018.8506504

Rosenthal, S., & Chung, T. (2020, February). A data science major: Building skills and confidence. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA,* 178-184. https://doi.org/10.1145/3328778.3366791

Rowley, J. (2002). Using case studies in research. *Management Research News*, *25*(1), 16–27. https://doi.org/10.1108/01409170210782990

Saddiqa, M., Magnussen, R., Larsen, B., & Pedersen, J. M. (2021). Open Data Interface (ODI) for secondary school education. *Computers and Education*, *174*, 104294. https://doi.org/10.1016/J.COMPEDU.2021.104294

Saltz, J. S. (2021). CRISP-DM for data science: Strengths, weaknesses and potential next steps. *International Conference on Big Data (Big Data) 2021*, 2337–2344. https://doi.org/10.1109/BigData52589.2021.9671634

Saltz, J., Armour, F., & Sharda, R. (2018). Data science roles and the types of data science programs. *Communications of the Association for Information Systems*, *43*, 33. https://doi.org/10.17705/1CAIS.04333

Saltz, J. S., Dewar, N. I., & Heckman, R. (2018). Key concepts for a data science ethics curriculum. *ACM Reference*. https://doi.org/10.1145/3159450.3159483

Saltz, J., & Heckman, R. (2016). Big data science education: A case study of a project-focused introductory course. *Themes in Science & Technology Education*, *8*(2), 85–94. https://www.learntechlib.org/p/171521/

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2021). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education*, *29*(S1), S40–S50. https://doi.org/10.1080/10691898.2020.1851159

Shamir, L. (2020). Eliminating self-selection: Using data science for authentic undergraduate research in a first-year introductory course. *Proceedings of the AAAI Workshop on Artificial Intelligence Diversity, Belonging, Equity and Inclusion.*

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *7647*(January), 1–25. https://doi.org/10.1136/bmj.g7647

Shereni, N. C., & Chambwe, M. (2020). Hospitality big data analytics in developing countries. *Journal of Quality Assurance in Hospitality & Tourism*, *21*(2), 361–369. https://doi.org/10.1080/1528008X.2019.1672233

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *14*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Song, I. Y., & Zhu, Y. (2016). Big data and data science: What should we teach? *Expert Systems*, *33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Takemura, A. (2018). A new era of statistics and data science education in Japanese universities. *Japanese Journal of Statistics and Data Science*, *1*, 109–116. https://doi.org/10.1007/s42081-018-0005-7

Twinomurinzi, H., Mhlongo, S., Bwalya, K. J., Bokaba, T., & Mbeya, S. (2022). Multidisciplinarity in data science curricula. *African Conference on Information Systems and Technology*. https://www.researchgate.net/publication/363436241_Multidisciplinarity_in_Data_Science_Curricula

Van Dusen, E., Suen, A., Liang, A., & Bhatnagar, A. (2019). Accelerating the advancement of data science education. *Proceedings of the Association for Information Science and Technology*, *56*(1), 601–603. https://doi.org/10.25080/Majora-7ddc1dd1-000

Verma, A., Yurov, K. M., Lane, P. L., & Yurova, Y. V. (2019). An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *Journal of Education for Business*, *94*(4), 243–250. https://doi.org/10.1080/08832323.2018.1520685

Wiktorski, T., Demchenko, Y., & Chertov, O. (2019, September). Data science model curriculum implementation for various types of big data infrastructure courses. *Proceedings of the IEEE 15th International Conference on EScience, San Diego, CA, USA*, 541–547. https://doi.org/10.1109/eScience.2019.00074

World Economic Forum. (2019). *Data science in the new economy: A new race for talent in the Fourth Industrial Revolution*. https://www3.weforum.org/docs/WEF_Data_Science_In_the_New_Economy.pdf

Wymbs, C. (2016). Managing the innovation process: Infusing data analytics into the undergraduate business curriculum (lessons learned and next steps). *Journal of Information Systems Education*, *27*(1), 61–74.

Xia, K., & Li, Y. (2020, December). Exploration on big data education for computer majors in applied colleges and universities. *Proceedings of the 5th International Conference on Mechanical, Control and Computer Engineering, Harbin, China*, 1846–1849. https://doi.org/10.1109/ICMCCE51767.2020.00405

Yadav, N., & Debello, J. E. (2019). Recommended practices for Python pedagogy in graduate data science courses. *Proceedings - Frontiers in Education Conference, FIE*, *2019-Octob*. https://doi.org/10.1109/FIE43999.2019.9028449

Zhang, Y., Zhang, T., Jia, Y., Sun, J., Xu, F., & Xu, W. (2017, May). DataLab: Introducing software engineering thinking into data science education at scale. *Proceedings of the IEEE/ACM 39th International Conference on Software Engineering: Software Engineering and Education Track, Buenos Aires, Argentina*, 47–56. https://doi.org/10.1109/ICSE-SEET.2017.7

Appendix A: Classification Coding Framework

(Appendix A can also be downloaded from https://bit.ly/DSECodingFramework

Theme	Description (initial coding framework)	Code
Continent (Origin)	Africa	1A
	Asia	1B
	Australia	1C
	Europe	1D
	Middle East	1E
	North America	1F
	South America	1G
	Not Specified	1H
Level (or type) of Qualification	Peer learning (i.e., outreach programs, hackathons, datathons, and bootcamps)	2A
	MOOCs and short learning programs	2B
	Micro-credentials, digital badges (badging system, digital platforms)	2C
	Undergraduate programs (degrees, diplomas, certificates)	2D
	Postgraduate (honors, masters and doctoral)	2E
	School education (i.e., primary, secondary)	2F
	Not specified	2G
Discipline-specific (Data Science Element)	Statistics/Mathematics	3A
	Computer science	3B
	Engineering	3C
	Non-science domain (non-STEM)	3D
	Data science	3E
	Not specified	3F
Data Science Education Provider	Public institutions of learning (universities, colleges, vocational education, and training)	4A
	Private institution	4B
	Industry/Organization	4C
	Collaborated	4D
	Not specified	4E
Data Science Context (using CRISP-DM)	Business requirement/understanding	5A
	Data understanding	5B
	Data preparation	5C
	Modeling	5D
	Evaluation	5E
	Deployment	5F
	Full data science lifecycle	5G
	Not specified	5H

Theme	**Description (initial coding framework)**	**Code**
Teaching Strategies	Use of technology (collaboration using digital platforms (apps), social media, or other digital communities)	6A
	Teacher-led (direct instruction)	6B
	Students-led learning/Game-based learning (extension of formal learning e.g., hackathons, game-based/competitions, community-driven)	6C
	Flipped classrooms (pre-recorded videos)	6D
	Personalized learning	6E
	Inquiry-based learning	6F
	Project-based learning	6G
	Competency-based learning	6H
	Not specified/other	6I
Opportunities/ Recommendations	Government involvement (policy, funding model, accreditation, open data)	7A
	Collaboration between university faculties (to maintain the multi-disciplinary nature of data science)	7B
	Industry involvement (live data/modern data streams/data expo, co-develop courses)	7C
	Professional advancement	7D
	Secondary school curriculum	7E
	Improve skills at schools and tertiary level/Capacity building (lecturers and school teachers)	7F
	Establish a data science governing body/Committee	7G
	Learning paths for data science	7H
	Integrated digital platforms (learning platforms/curriculum systems)	7I
Challenges	Pedagogy (teaching approaches)	8A
	Inadequate curriculum (e.g., aata ethics, business understanding, deployment)	8B
	Cognition (challenging concepts, i.e., statistics, programming)	8C
	Data science program structure	8D
	Data science tools (model misuse, misinterpretation of models)	8E
	University policies (regulatory frameworks across different disciplines, e.g., student recruitment and enrolment, limited resources)	8F
	Naming regimes	8G
	Assessment (assessing student achievement)	8H
	A disconnect between industry practice and data science learning material	8I
	Challenges of finding organizations willing to participate	8J
	Accreditation issues	8K
	Not specified	8L

Theme	Description (initial coding framework)	Code
Theory	Theory driven	9A
	No theory guiding the study	9B
Philosophy	Positivism	10A
	Interpretivism	10B
	Pragmatism	10C
	Critical realism	10D
	Not specified	10E
Research Method (Oates, 2006)	Experiment	11A
	Survey	11B
	Action research	11C
	Ethnography	11D
	Case studies	11E
	Design and creation (design science)	11F
	Not specified	11G
Data Science Education Stakeholders	Lecturers (IS/IT, mathematics/statistics, engineering, domain/business)	12A
	Industry/organizations (live data/modern data streams/data expo, internships)	12B
	Schools	12C
	Institute of higher learning – public	12D
	Institute of learning – private	12E
	Community/society/alumni (e.g., outreach programs)	12F
	Government	12G
	Students	12H

APPENDIX B: CORRELATIONAL ANALYSIS

(Appendix B is a spreadsheet that can be downloaded from this papers' publication page. It can also be downloaded from https://bit.ly/DSECorrelationalanalysis)

AUTHORS

Nkosikhona Theoren Msweli, MIT (Information Systems) is a Lecturer at the University of South Africa, College of Science, Engineering, and Technology. She is currently a Ph.D. candidate at the University of Pretoria pursuing a Ph.D. degree in information systems. Her research interests include, among other things, data science education, digital and mobile technologies, and online learning. She is a co-editor of NEMISA research colloquium proceedings.

Tendani Mawela is an Associate Professor at the University of Pretoria in the Department of Informatics. Prior to joining academia, she worked as an IT and management consultant on projects across the private and public sectors. She received an MBA from the Wits Business School and a Ph.D. in Informatics from the University of Pretoria. She is a Y2 NRF rated researcher. Her primary research interests are in ICTs for sustainable development, digital government, digital skills as well the ethical design, use and governance of technology and information systems.

Hossana Twinomurinzi is a C2 South Africa NRF Rated Researcher and 4IR Professor with the Department of Applied Information Systems, University of Johannesburg. He is currently Vice Dean for Research, Innovation & Internationalisation at the College of Business and Economics. He is also supporting the efforts in the College on infusing data science into research, teaching/learning and community efforts. He is also an Associate Editor for the *African Journal of Information and Communication*, past Associate Editor for the *African Journal of Information Systems*, the Immediate Past Chairperson for the ICT4D Flagship at the University of South Africa, and the immediate Past Secretary for SAICSIT. His primary research interests are in Applied Data Science, Digital Skills, Digital Government, Digital Innovation and ICT for development. He has supervised 15+ Masters and Doctoral students in the areas of data analytics, digital government and ICT for development. He serves on the editorial boards of several academic publications and has served as a convener and technical chair at several international and national conferences. He has led national research projects of ICT and done contract research in various sectors ranging from the military, government, non-profit organisations and banking. He has management and executive experience, having served in chief executive and senior management positions in South Africa, England, Swaziland and Uganda. He is a professional facilitator and is occasionally involved in social enterprise activities information systems.