Journal of Digital Educational Technology
2026, 6(2), ep2611
e-ISSN: 2752-5503
https://www.jdet.net/

Review Article

OPEN ACCESS

A review of data science definitions and competencies in higher education

Bahar Memarian [1]*, Tenzin Doleck [1]

[1] Simon Fraser University, Burnaby, BC, CANADA
***Corresponding Author:** bmemaria@sfu.ca

Citation: Memarian, B., & Doleck, T. (2026). A review of data science definitions and competencies in higher education. *Journal of Digital Educational Technology, 6*(2), Article ep2611. https://doi.org/10.29333/jdet/18527

ARTICLE INFO

Received: 19 Jan. 2026

Accepted: 21 Mar. 2026

ABSTRACT

Data science is expanding as a discipline and profession, yet its conceptual and philosophical foundations—particularly within higher education—remain underexamined. This study addresses this gap through a systematic literature review of peer-reviewed publications indexed in Scopus and Web of Science, focusing on how data science is defined and how its competencies and learning outcomes are articulated. Drawing on formal conceptual analysis, we examine whether definitions are structured as extensional or intensional, and whether learning outcomes are framed as nouns or verb-noun pairs. Using an interpretive framework, we evaluate the quality, strengths, and weaknesses of reported data science definitions and learning outcomes in higher education contexts. Findings indicate that most data science definitions adopt an intensional structure, with clearer insight achieved when both category and differentia are explicitly specified. Learning outcomes are predominantly expressed as verb-noun pairs and are more meaningful when they emphasize adaptive, timeless skills. The review also highlights ongoing tensions between university-based programs and certificate offerings, questions surrounding curriculum design, accreditation, stakeholder involvement, and the evolving role of industry and artificial intelligence in shaping the field. Overall, this work provides conceptual understanding and critical insights into the benefits, challenges, and future implications for defining data science and designing robust, inclusive data science curricula in higher education.

Keywords: Data science, learning outcomes, systematic review

INTRODUCTION

Over the past decade, data science has expanded from a technique to a discipline of its own (Donoho, 2017). Nearly every area in post-secondary education or industry uses data to model and solve real-world problems. Data science has become a hotspot for streamlining algorithmic decision-making and optimizations. It is thus critical for today's graduates to have an awareness of data science (Dichev & Dicheva, 2017). Yet, the literature presents little work on what data science means and the learning outcomes it entails.

The goal of this study is to review the literature to examine the type and quality of definitions (here also studied synonyms such as meaning or means, mean, define, defining, interpret, interpretation, understand, understanding, explanation, essence, context, spirit, content, sense, connotation, or nuance) and competencies (here also studied synonyms such as learning outcomes, educational outcomes, outcomes, proficiency, expertise, skill, mastery, aptitude, or capability) surrounding data science. Through a systematic literature review in Scopus and Web of Science (WoS) from peer-reviewed manuscripts, this study aims to extract and summarize descriptions surrounding data science.

Gap

Data science is often considered a multistep process or lifecycle that requires working with data of some form to turn it into insight (Stodden, 2020). Such a process needs understanding, scoping, processing, analyzing, validating, and sharing data within a context and problem of interest. A key feature of data science is to use data credibly to reach a finding. This makes data science a prevailing and comprehensive process, as any method or practice with data that is found credible can become part of data science.

The methods used in data science are wide-ranging (Dedge Parks, 2017). Various fields and subject terms have been attributed to or found analogous to data science. As a result, various definitions and learning outcomes of data science exist in the literature. For example, the use of statistical methods has been endorsed in data science literature (Cleveland, 2001; Donoho, 2017; Engel, 2017). Alternatively, applying computational methods has been suggested (Cao, 2017). Not to neglect unforeseen events, such as the pandemic, may

change the course of science curricula, including that of data science (Onyema et al., 2023; Vittorini & Galassi, 2021). The notion of data science may therefore vary across contexts and research studies.

Given the diverse disciplines that use data science in higher education, we find it important to mine the literature and conceptualize data science definitions and learning outcomes. With the proliferation of artificial intelligence and machine learning algorithms in higher education, the need to have an understanding of data science is becoming more necessary. A starting point would be to understand what data science means and what learning outcomes it entails for pedagogy. Definitions are perceived to progress from individually experiential to socially shared (Litowitz, 1977). Learning outcomes are needed that are actionable and signal what a student should be able to do (Baume, 2009; Nilson, 2016). Our work is motivated by the need to learn more about data science skills and learning outcomes (Blei & Smyth, 2017; Cao, 2016; Finzer, 2013). In conducting this systematic literature review, hence we seek to examine the state of data science definitions and learning outcomes in the higher education literature.

There have been debates on whether data science should expand more as a body of knowledge or skills (Donoho, 2017; Zhu & Xiong, 2015). With access to data becoming increasingly pervasive and cheap, efforts have been made to define the science of dealing with data, giving birth to data science (Kross et al., 2020). A comprehensive review of work, for example, reveals the great variability and extent of concepts associated with the field of data science (Cao, 2017).

Prior work has examined conceptions of data science in data science curricula. Work by Friedman (2019), for example, has examined and analyzed 40 data science syllabi used in private and public academic institutions. Less work, however, has been done to map the state of data science definitions and learning outcomes in the research literature, particularly at the higher education level.

Goal and Research Questions

An analysis of data science definitions and learning outcomes together may provide a picture of the strengths and limitations of the discipline, and the caveats each conception may have attached to them. The definitions and learning outcomes are two integral indicators for defining and pivoting a discipline (Adams, 2006; Education Development Center, 2003). Understanding data science definitions and learning outcomes at this time may therefore be a guiding factor in the direction data science is headed and its gaps and needs.

To achieve this, we study the following two research questions (RQs):

RQ1. What is the type and quality of reported data science definitions in the higher education literature?

RQ2. What is the type and quality of reported data science learning outcomes in the higher education literature?

Following an interpretive paradigm and best practices informed by the literature informed in the theoretical framework, we aim to find responses for the two abovementioned RQs.

PRELIMINARY LITERATURE REVIEW AND FRAMEWORKS ADOPTED

Research in cognitive science recognizes the significance of both intensional and extensional mechanisms (Whitney et al., 2009). Both types of intensional and extensional representations are regularly used in systems analysis and data management books (Elmasri & Navathe, 2010; Gillenson, 2012; Hoffer et al., 2012). High-level data models are abstractions close to the way the users see the information (Niinimäki, 2000). As Niinimäki (2004) denotes: "An extensional language uses the terminology of extensional entities (things in the domain of application) and relationships among them. In an intensional language, concepts have a central role, and the language has means of expressing relationships between concepts" (p. 27).

For a summary of data science definitions, we depend on extensional versus intensional classification. "Cognitive research suggests that understanding the semantics, or the meaning, of representations involves rising from concrete concepts denoting specific observations (that is, extension) to abstract concepts denoting the number of observations (that is, intension), and vice versa" (Samuel et al., 2018, p. 1). "According to the classical and prototype theories of concepts, the meaning of a concept is characterized by its gist or properties (i.e., intension). Exemplar theory suggests that the meaning of a concept is determined by a set of instances of things that exist in the referring domain or its extension" (Samuel et al., 2018).

We use our interpretive understanding of the best noted practices to evaluate and codify the noted definitions (either intensional or extensional) of data science as either being insightful or needing improvement.

For a summary of data science skills, we depend on noun versus verb-noun classification. Using verbs at the center of learning outcomes development is deemed a better practice (Adelman, 2015). Moreover, making use of operational skills is highlighted as favored (Adelman, 2015). When evaluating the data science skills, we probe into the noun and verb-noun quality and our interpretation of what skill is more operational, that is, pedagogically achievable, and with more longevity in learning. The interpretive data science conceptualization achieved in this paper thus assumes the subjective nature of being and meaning-making for the new field.

Best Practices to Assess Data Science Definitions

While different interpretations of the definition exist in the literature, a common approach is to view the definition through two recognized classifications, namely an intensional and extensional lens (Peregrin, 2007). As denoted by the Encyclopedia Britannica, "intension" indicates the internal content of a term or concept that constitutes its formal definition, and "extension" indicates its range of applicability by naming the particular objects that it denotes (Britannica, 2023). In other words, an intensional definition may provide the minimum viable description of a term, while an extensional definition creates a thorough list of everything that falls under that term. An intensional definition is further

decomposed into two classifiers, namely a "category or genus" and "differentia". Rooted in Aristotle's way of reasoning (Studtmann, 2021), the category specifies the class of things that have common characteristics and that can be divided into more granular kinds. The differentia, on the other hand, is anything that is not part of the category/genus and helps in discerning the term further.

We classify the extracted definitions around data science as either an extensional or intensional definition. We further assess the quality of data science definitions based on the completeness of the description and best practices reported in the literature. Examples include:

- "Signposting the logical/discourse structure of the subject/lecture or helping to maintain comprehension as the discourse progresses" (Flowerdew, 1992, p. 1).
- Reduce jargon and explain in the simplest terms (Podsakoff et al., 2016).
- Definition does not use circular reasoning, ambiguous, or overlapping terms (Johner Institute, 2023).
- Note the general function of the asset (Stanford University, 2023).
- The new term is placed into context with other terms through semantic relationships (UCF, 2023).
- Avoid "X is when" or "X is where" (Purdue Online Writing Lab, 2023).
- A new term cannot be understood only using old terms we understood beforehand (Lewis, 1970).

Best Practices to Assess Data Science Learning Outcomes

We also classify the science learning outcomes. One way to achieve this is to see if the description presented uses verb-noun pair(s) or noun(s) only. It is customary to describe a knowledge set through nouns and competencies through verb-noun pairs. Knowledge is often considered to be lower-order and learning outcomes higher-order in educational taxonomies (Anderson et al., 2001). We classify the extracted learning outcomes around data science as either a noun or a verb-noun pair. We further assess the quality of data science learning outcomes based on the completeness of the description and best practices reported in the literature. Examples include:

- "Express what learners are expected to achieve and how they are expected to demonstrate that achievement," and use action verbs (Kennedy et al., 2006, p. 1).
- Both intended and emergent learning outcomes are embraced (Hussey & Smith, 2003).
- "Disseminate up-to-date knowledge, develop the capability to use ideas and information, develop the student's ability to test ideas and evidence, develop the student's ability to generate ideas and evidence, facilitate the personal development of students, develop the capacity of the student to plan and manage their own learning" (Bourner, 1997, p. 2).
- Shift in focus from what the instructor can, or should, teach, to what the achieved or expected level of understanding of the student is (Attard, 2010).
- Enable a description of a university degree in terms of its outcomes, essentially, what a graduate knows, can do, and understands (Heywood, 2000).
- Provide consistency and transparency in the assessment processes for both student and educator and emphasize students' active role through an experiential student-centered approach (Ellis, 2004).
- Display of actual achievement, rather than being based on the time, or a specified period, that a student has engaged in study (Ecclestone, 1999, 2010).
- Empower student learning because they are not content-based, but outcome-based (Gibbs, 1995).
- Facilitate critical thinking, problem-solving, judgment and insight, research and scholarship, communication, creativity and design, self-regulation, and professional competence (HEQCO, 2005).

MATERIALS AND METHODS

Our goal with this work is to examine the research done by individuals, groups, and organizations predominantly at the higher education level that has informed data science definitions and learning outcomes. This work can further aid in informing higher-quality programming and assessments, outlining challenges in existing work, summarizing potentials not yet considered, and making recommendations for future work.

Our research is guided by an interpretivist epistemological framework. The interpretivist approach assumes that reality is subjective and socially constructed (Schwandt, 1994). We follow this framework because prior research reveals that interpretivism and social constructionism contribute to the development of research and thinking about data science, specifically on how objective new data are and on the role of knowledge in new data use and construction (Aragona, 2017). To help disseminate our interpretation, we follow established frameworks to chart and assess the quality of data science definitions and learning outcomes in the literature.

Literature Search Strategy

An overview of the search process is summarized in **Figure 1**. Additionally, inclusion and exclusion criteria can be found in **Table 1**.

Inclusion and Exclusion Criteria

We systematically searched the literature using the following databases: Scopus and WoS. Our search string comprised: "data science" AND (literacy or education) AND ((meaning or means or mean or definition or define or defining or interpret or interpretation or understand or understanding or explanation or essence or context or spirit or content or sense or connotation or nuance) OR (competence or competencies or competency or assessment or assess or learning outcomes or educational outcomes or outcomes or proficiency or expertise or skill or mastery or aptitude or capability)).

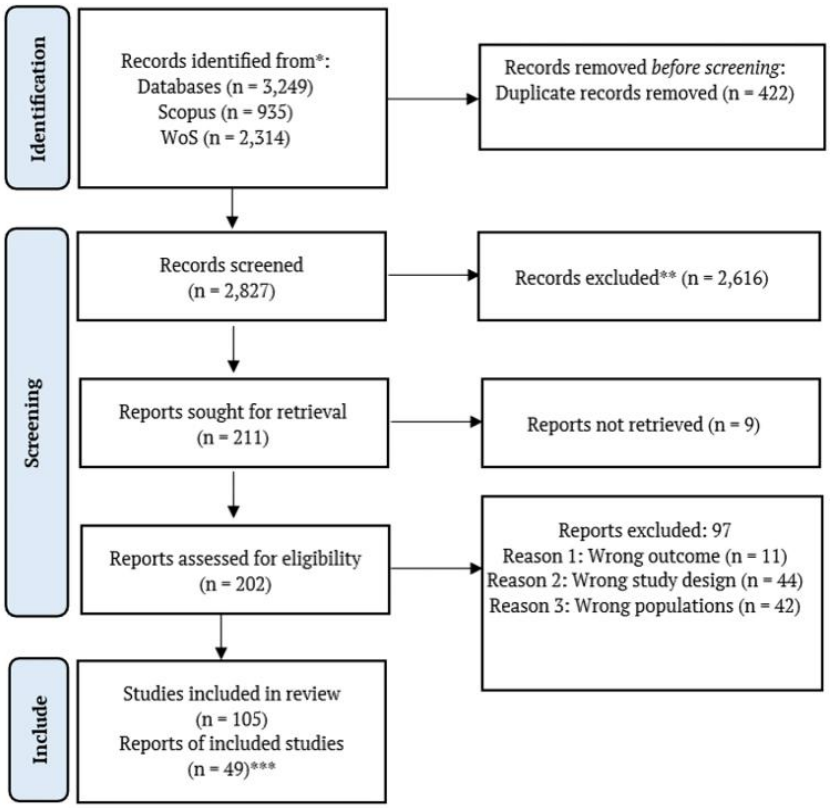

Figure 1. PRISMA flowchart of primary study selection (*excluded if search terms were not targeted in the title or abstract of the article, **excluded if the study was focused on (1) specialized industries [e.g., finance, healthcare], (2) At the secondary level, ***excluded if the study was focused on specialized components of pedagogy [e.g., dataset structures] and neither provided a definition nor a skill list) (Adapted from Moher et al., 2009)

Table 1. Inclusion and exclusion criteria

Criteria	Inclusion	Exclusion
Primary	Studies disseminated in English Peer-reviewed publications	Studies not disseminated in English Not peer-reviewed publications
Secondary	Studies focused on AI and higher education	Studies outside of AI and higher education, or exploring one of AI or higher education independently

To improve the consistency and comprehensiveness of our search, we included search strings for each of our RQs separately and connected them with the "OR" operator. Synonyms were also derived from the thesaurus, and terms related to our research focus were included in our research by using the "OR" operator. To further improve the quality of our information-seeking, we included peer-reviewed journals and conference proceedings focused on the higher education level and used an unlimited time frame. However, articles not in English were not included in our research. We, therefore, acknowledge that the summaries represent the countries and regions present in the reported literature.

We generally found most English studies to stem from North America and Europe, with a few collaborations in Japan and West Asia. We anticipate a great body of work done by countries such as China to be left out of our systematic search, and hence, the findings may not represent such regions. Our initial screening's inclusion criteria contained articles that answered both or at least one of the two RQs, were published in English, contained the term "data science" in the title or abstract, and focused on data science post-secondary education or closely relevant industry settings, as our focus is more on data science in higher education. Our initial exclusion criteria contained articles that were broadly unrelated, duplicated, unavailable full texts, or abstract-only papers that were not published in English, partial terms like "data literacy" or "science literacy," and focused on non-K-12 or pre-collegiate settings. For our secondary screening, we included articles that focused on ways researchers and educators have formulated and assessed data science and excluded articles that explored avenues impacted by data science and educational data science, such as using data science from MOOCs.

Of the 3,249 studies gathered from WoS (n = 2,314) and Scopus (n = 935), a total of 49 publications were included in this review, see **Figure 1** for the PRISMA flow chart. Of the 49 publications reviewed, 38 (78%) were from WoS and 11 (22%) from Scopus. Of the 49, 35 (71%) were conference papers, and 14 (29%) were articles. The majority of the publications introduced the concept of data science. The publications then continued with explaining their proposed data science initiative, along with mentions of learning objectives or outcomes. Our focus was on the quality of the definition provided for data science and associated learning outcomes.

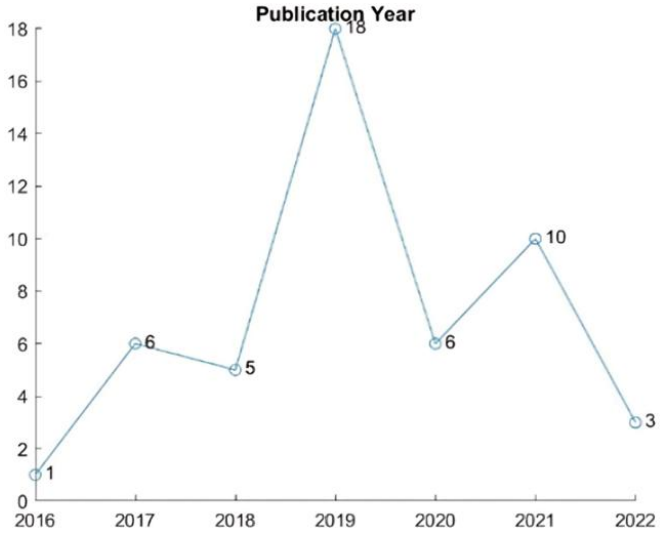

Figure 2. Articles published per year (Source: Authors' own elaboration)

Figure 3. Predominant countries of affiliation (Source: Authors' own elaboration)

We anticipated that the definition or learning outcomes provided for data science may not always be influenced by the quality of the study proposed, and therefore, evaluated each of the definitions and learning outcomes separately.

The number of studies published per year from 2016 to 2022 ranged from 1 to 18, with a notable increase in 2019 (18) and a decline afterward (**Figure 2**). This may highlight the perturbations that world events, such as the global pandemic, may have on research and discipline of data science. Although we had set no time limit for our search, no studies from before 2016 remained after our screening process. This may suggest that data science is more recently examined in the literature, which may in part be due to the proliferation of technologies like AI, which use data science as their cornerstone.

We also examined word cloud summaries of the demographic data. The United States, followed by Bulgaria, Norway, and the Netherlands, were the predominant affiliation countries of the authors (**Figure 3**).

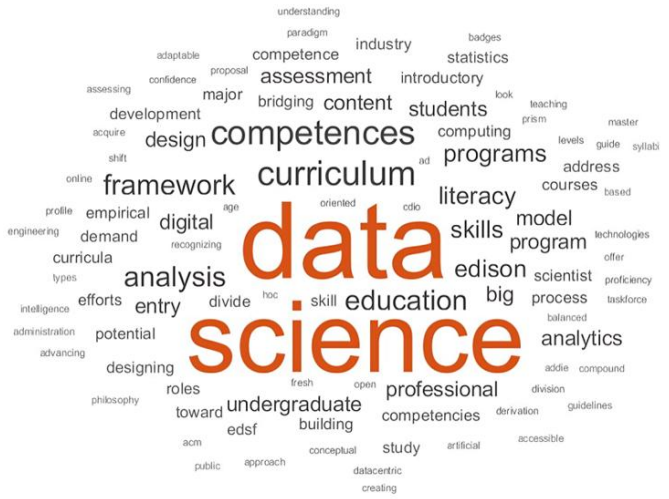

Figure 4. Predominant words in the article title (Source: Authors' own elaboration)

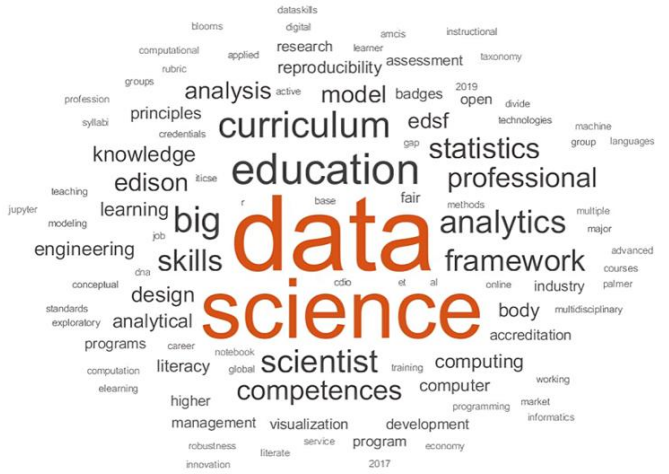

Figure 5. Predominant author keyword (Source: Authors' own elaboration)

As expected, the predominant words present in the titles of the publications were datum, science, competence, curriculum, program, and so on (**Figure 4**).

When examining the predominant author's keywords, we again found datum and science to be taking the lead, followed by curriculum, big (often referring to big data), education, and so on (**Figure 5**).

Of the reviewed 49 publications, 34 (69%) included a definition for data science, and 39 (80%) presented one or a combination of learning outcomes specific to data science. In the sections that follow, we explore references for each RQ independently.

Data Collection and Charting

The search string was input into WoS and Scopus, and the full record of all studies was downloaded and exported into the Covidence software. Then each paper was reviewed for initial and secondary screening. The title and abstract screening and studies selected for extraction were carried out in Covidence. A total of 105 included studies were then downloaded and considered, of which 49 articles were included in the review. Examples of codes used were whether the definition provided

Table 2. Overview of data science definition types presented in the 49 reviewed studies

Type of definition	Reference
Extensional (a comprehensive list of elements) (n = 9)	Carmichael and Marron (2018), Costa and Santos (2017), Danyluk and Buck (2019), Dinov (2019), Fekete et al. (2021), Hicks and Irizarry (2018), Petrova et al. (2018), Rosenthal and Chung (2020), Toleva-Stoimenova et al. (2019)
Intensional (a category and differentia) (n = 25)	Adams (2020), Cao (2017), Cetinkaya-Rundel and Ellison (2021), De Veaux et al. (2017), Demchenko et al. (2016, 2019a, 2021a), Dichev and Dicheva (2017), Friedman (2019), Inc and Wang (2017), Kostadinova et al. (2018), Leidig and Cassel (2020), Overton and Kleinschmit (2022), Pedersen and Caviglia (2019), Raj et al. (2019), Rasheva-Yordanova et al. (2019a, 2019b), Rossi (2021), Salas-Rueda (2020), Saltz et al. (2018), Sedelmaier et al. (2021), Wu (2017), Xiao et al. (2021), Zhou et al. (2020)
No direct note of a data science definition provided (n = 15)	Bargagliotti (2019), Belloum et al. (2019), Blair et al. (2021), Christozov et al. (2019), De Veaux et al. (2022), DeMasi et al. (2020), Demchenko et al. (2019a), Gottipati et al. (2021), Kakeshita et al. (2022), Menasalvas et al. (2019), Mikalef and Krogstie (2019), Rasheva-Yordanova et al. (2019a, 2019b), Schmitt et al. (2021), van der Aalst (2019)

Table 3. Overview of quality scored for the 9 extensional definitions

Extensional definition	Reference
Insightful (n = 2)	Hicks and Irizarry (2018), Petrova et al. (2018)
Needs improvement (n = 7)	Carmichael and Marron (2018), Costa and Santos (2017), Danyluk and Buck (2019), Dinov (2019), Fekete et al. (2021), Rosenthal and Chung (2020), Toleva-Stoimenova et al. (2019)

for data science was intensional (and if so, whether category and differentia were presented) or extensional, or whether the learning outcomes(s) noted followed a verb-noun or noun structure. We used MATLAB to explore and summarize the demographic (exported from full-records citation data) and extracted (e.g., coded) data. Mendeley was also used to keep track of references.

For the analysis of qualitative demographic data, we used tabulated summaries in descriptive, graphical, or tabular form. We also derived predominant word clouds of text (e.g., exploring the titles of publications) by changing text to lowercase, tokenizing, erasing punctuation, removing stop words, and normalizing words, see MATLAB guidelines (Mathworks, 2023). The data science definitions and learning outcomes were coded based on the theoretical framework adopted, and the references for each classification are summarized in tabular form.

Once the charting was done, we went through each data science definition and data science learning outcomes and summarized whether they presented a definition that was insightful or needed improvement based on best practices outlined in the theoretical framework subsection. In the next section that follows, we provide a summary and evaluation of reviewed articles for each of the data science definitions and learning outcomes by following the best practices reported in the literature. In the discussion section, we provide a summary of findings, followed by challenges and future potentials in conceptualizing data science definitions and learning outcomes.

The protocol included the following steps:

- Used the input string of search in both Scopus and WoS.
- Downloaded the full records of all the papers.
- Uploaded the full records to Covidence.
- Used Covidence to identify duplicate studies and remove them.
- Review the title and abstract of each article and decide if relevant or not (following initial inclusion and exclusion criteria as noted before).
- Download and review the full text of each article and decide if it is relevant or not (following secondary inclusion and exclusion criteria as noted before).
- Gather the full demographic records of studies selected for review.
- Extract and thematically chart definitions and learning outcomes of data science terms in each study in Excel.
- Create a synthesis of future recommendations and challenges in the discussion section based on the full review of each article.

RESULTS

RQ1. Data Science Definitions

Of the 34 publications, 9 (26%) presented an extensional, and 25 (74%) presented an intensional form of definition for data science, see **Table 2** for a summary of references found per type of definition. A definition of data science could not be found in the remaining 15 studies.

Extensional data science definitions

We reviewed the 9 extensional definitions and attempted to score the quality of the definition, as shown in **Table 3**. Our coding of definitions led to classifying each as having a more insightful versus a needs improvement definition quality. Below, we describe and justify our selection for each.

Insightful extensional data science definitions: Instead of presenting a list of topics, Petrova et al. (2018) point to what disciplines are employed and techniques are incorporated in data science. Their definition acknowledges that data science is not just a set of techniques (e.g., machine learning) but also requires working with systems and understanding how to extract or utilize data from them. In a similar vein, Hicks and Irizarry (2018) Take that data science is used differently in different contexts. Such a realization is effective because it can shift our focus from a static to a dynamic and causal definition of data science. This shift is needed in light of ongoing changes and upgrades made to technologies (e.g., AI) that utilize and require an understanding of data science.

Table 4. Overview of quality scored for the 25 intentsional definitions

Intensional definition	Reference
Insightful (n = 8)	Cao (2017), Cetinkaya-Rundel and Ellison (2021), De Veaux et al. (2017), Kostadinova et al. (2018), Leidig and Cassel (2020), Pedersen and Caviglia (2019), Raj et al. (2019), Xiao et al. (2021)
Needs improvement (n = 17)	Adams (2020), Demchenko et al. (2016, 2019a, 2021a), Dichev and Dicheva (2017), Friedman (2019), Inc and Wang (2017), Overton and Kleinschmit (2022), Rasheva-Yordanova et al. (2019a, 2019b), Rossi (2021), Salas-Rueda (2020), Saltz et al. (2018), Sedelmaier et al. (2021), Wu (2017), Zhou et al. (2020)

Needs improvement: extensional data science definitions: We found extensional definitions that were too broad, resulting in a definition that needs improvement. For example, Rosenthal and Chung (2020) propose that data science is at the intersection of multiple fields. It may be difficult to surmise if intersection implies a common denominator (e.g., machine learning is a method in all), or practices that go hand in hand (e.g., machine learning and artificial intelligence complement each other). Costa and Santos (2017) define data science as the coupling of two fields. Such an extensional definition does not provide much insight into what is being coupled from the two fields and why coupling is possible or necessary. Similarly, Fekete et al. (2021) posits that data science has elements from computer science and statistics. Such a definition may add complexity and dependency to data science. An analogy to this definition of data science is the definition of light being both a particle and a wave. Understanding light is thus dependent on learning what makes both a particle and a wave and keeping track of their evolving definitions.

There were also extensional definitions that had some level of ambiguity, which led to confusion. Danyluk and Buck (2019) Make the case that data science and AI are not similar but mention that many areas of AI are relevant to data science. We find there to be a flaw in the logic of having two areas (e.g., AI and data science) that are by no means synonymous yet have many areas of overlap. One can argue that for areas of overlap, at least, the two terms are synonymous. Toleva-Stoimenova et al. (2019) suggest that data science utilizes knowledge developed from different areas, such as statistics and informatics. We find there to be an implicit assumption that data science cannot develop knowledge of its own, which may not be true, especially in the case of artificial intelligence. Dinov (2019) posits that data science has an "intrinsic dual reliance on quantitative basic sciences techniques as well as qualitative artistry" (p. 1). It is unclear from such a definition what qualitative artistry would entail. Also, data science may employ techniques that go beyond basic science techniques, which are not acknowledged in the authors' definition. Carmichael and Marron (2018) note a "union of six areas of greater data science which are borrowed from David Donoho's article titled "50 years of data science" (Donoho, 2017, p. 3):

1. Data gathering, preparation, and exploration
2. Data representation and transformation
3. Computing with data
4. Data modeling
5. Data visualization and presentation
6. Science about data science

We find that their extensional definition of data science may have some level of circular reasoning by including 6. Science about data science.

Intensional Data Science Definitions

We also reviewed the 25 intensional definitions and attempted to score the quality of the definition, see **Table 4**. Our coding of definitions led to classifying each as having a vs. needing improvement in definition quality. Below, we describe and justify our selection for each.

Insightful intensional data science definitions: Definitions with a specific category and/or differentia often resulted in a more intuitive and robust understanding of data science. Xiao et al. (2021) define data science as "the ability to extract meaningful insights from data using a combination of domain knowledge, programming skills, and statistics" (p. 290). While the term ability may be a bit misleading, as science is more than an ability, the remainder of the definition provides good insight into the essence of data science, which requires a combination of knowledge and skills.

Similarly, Leidig and Cassel (2020) consider data science as "the field that brings together domain data, computer science, and the statistical tools for interrogating the data and extracting useful information." (p. 10). Kostadinova et al. (2018) also offer a rationalist definition of data science: "Data Science is the science of pooling, processing, and analyzing vast streams of (non-structured) data to understand and analyze events related to them." (p. 1). Cetinkaya-Rundel and Ellison (2021) and De Veaux et al. (2017) adopt the definition presented by the National Science Foundation (2014): "science of planning for, acquisition, management, analysis of, and inference from data" from the literature. These authors highlight that data is used to analyze and understand events related to that data. So, data science is not just dependent on the analysis but also the proper identification and extraction of data.

Besides being a science, data science has also been categorized as a cross-disciplinary field. Both Raj et al.(2019) Cao (2017), for example, consider data science as a multi-disciplinary field and Pedersen and Caviglia (2019) Note data science as the domain of specialists.

Needs improvement in intensional data science definitions: A vague category or differentia often resulted in an intensional description of data science. Zhou et al. (2020) perceive data science to: "draw conclusions from data of all walks of life, and all areas of study by combining computational and inferential reasoning". We find the category "from all walks of life and areas of study" (p. 1) to be vague. In Rasheva-Yordanova et al. (2019a), the category for their definition uses multiple terms that may seem superfluous, such as tools, methods, and systems. Salas-Rueda (2020) presents that data science is a broad field. We find these categories do not add value to a specific definition of data science. In a similar vein, Dichev and Dicheva (2017) present a differentia from Wikipedia: "about scientific processes and

Table 5. Overview of data science learning outcome types presented in the 49 reviewed studies

Learning outcomes description	Reference
Verb-noun pair (n = 29)	Adams (2020), Bargagliotti (2019), Blair et al. (2021), Cetinkaya-Rundel and Ellison (2021), Christozov et al. (2019), Costa and Santos (2017), Danyluk and Buck (2019), Demchenko et al. (2016, 2019a, 2019b, 2021a, 2021b), Dichev and Dicheva (2017), Fekete et al. (2021), Inc and Wang (2017), Kostadinova et al. (2018), Menasalvas et al. (2019), Mikalef and Krogstie (2019), Petrova et al. (2018), Raj et al. (2019), Rasheva-Yordanova et al. (2019a, 2019b, 2019c, 2019d), Rosenthal and Chung (2020), Rossi (2021), Saltz et al. (2018), Sedelmaier et al. (2021), Wu (2017), Xiao et al. (2021)
Noun (n = 10)	Belloum et al. (2019), De Veaux et al. (2017), DeMasi et al. (2020), Dinov (2019), Gottipati et al. (2021), Hicks and Irizarry (2018), Leidig and Cassel (2020), Overton and Kleinschmit (2022), Schmitt et al. (2021), Zhou et al. (2020)
No direct note of data science learning outcomes provided (n = 10)	Cao (2017), Carmichael and Marron (2018), De Veaux et al. (2022), Friedman (2019), Kakeshita et al. (2022), Overton and Kleinschmit (2022), Pedersen and Caviglia (2019), Salas-Rueda (2020), Toleva-Stoimenova et al. (2019), van der Aalst (2019)

Table 6. Overview of quality scored for the 38 data science learning outcomes

Learning outcomes description	Reference
Insightful (n = 31)	Adams (2020), Bargagliotti (2019), Belloum et al. (2019), Blair et al. (2021), Cetinkaya-Rundel and Ellison (2021), Costa and Santos (2017), Danyluk and Buck (2019), Demchenko et al. (2016, 2019a, 2019b, 2021a, 2021b), Dichev and Dicheva (2017), Fekete et al. (2021), Gottipati et al. (2021), Kostadinova et al. (2018), Leidig and Cassel (2020), Menasalvas et al. (2019), Mikalef and Krogstie (2019), Petrova et al. (2018), Raj et al. (2019), Rasheva-Yordanova et al. (2019a, 2019b, 2019c, 2019d), Rosenthal and Chung (2020), Rossi (2021), Saltz et al. (2018), Schmitt et al. (2021), Sedelmaier et al. (2021), Wu (2017)
Needs improvement (n = 8)	De Veaux et al. (2017), DeMasi et al. (2020), Dinov (2019), Hicks and Irizarry (2018), Inc and Wang (2017), Overton and Kleinschmit (2022), Xiao et al. (2021), Zhou et al. (2020)

systems to extract knowledge or insights from data in various forms" (p. 1). It may be difficult to decipher between methods, processes, and algorithms, and some may call them analogous.

When links were made to other domains, the intensional definitions detracted from data science as a field. Overton and Kleinschmit (2022), for example, focus on public administrator programs, and Inc and Wang (2017) on the business aspects of data science. Demchenko et al. (2016) associate a strong link between data science and emerging Big Data and data-driven technologies. In Demchenko et al. (2019a, 2019b, 2021a, 2021b), data science is also considered a research and academic discipline that provides a basis for data analytics and ML/AI applications.

An intensional definition that needs improvement also emerged from descriptions that lacked a differentia. Rasheva-Yordanova et al. (2019b) note that "Data science has emerged as a new inter-multi and even transdisciplinary area of knowledge" (p. 1). A strong feature of this definition is in acknowledging data science as a fluid concept. The definition, however, does not offer a differentia. In Friedman (2019), we also find the differentia to be missing. The authors define data science based on the description of Wu (2017) who made a direct connection between data science and statistics.

There were also intensional definitions that were either too long or used phrases that resulted in misunderstanding data science. In Rossi (2021), Saltz et al. (2018), and Sedelmaier et al. (2021), for example, definitions are provided that may be too long and contain too large a category and differentia. A similar trend can be seen in Wu (2017), where the definition also includes many overtures: "Applied data science and analytics (DSA) has been an emerging and developing discipline, which applies modern, data-driven analytical methods over massive data to application-oriented problem-solving and decision making based on evidence-based data analytical results" (p. 1). Adams (2020) proposes that data science finds its meaning when individuals apply their skills to data collected within some context. We find such a definition to assume a level of exceptional specialized skill from individuals when working with data. While talent and natural ability to work with data make some individuals stand out, they may not be considered attainable skills, and so may not be useful for defining data science.

RQ2. Data Science Learning Outcomes

Of the 39 publications, 10 (26%) had used a noun and 29 (74%) a verb-noun pair; see **Table 5** for a summary of references found per definition type. We reviewed the 38 data science learning outcomes and attempted to score the quality of the definition, as shown in **Table 6**. Our coding of learning outcomes led to classifying each as having an insightful versus a needs improvement description quality. We found that most of the studies that presented a verb-noun definition of data science learning outcomes were relatively more insightful than the ones presented as a noun.

Insightful data science learning outcomes

We found studies that utilized initiatives or skills mentioned in the Association for Computing Machinery (ACM) draft report to have a higher quality (Danyluk & Buck, 2019; Fekete et al., 2021; Leidig & Cassel, 2020). The ACM draft was first published by a large pool of experts, part of a data science task force committee in 2017 and summarized various distinguished projects that worked towards the development and accreditation of data science curricula. In addition, the report shared an extensive list of knowledge and skills for the various topics that are often taught in the data science curriculum. ABET, which offers accreditation for engineering programs in the United States, has also attempted to develop accreditation criteria for undergraduate data science programs

(Blair et al., 2021). The effort is led by members of a joint data science criteria subcommittee appointed by ABET's Computing Accreditation Commission and CSAB (the lead society for computing accreditation).

A common framework employed by several studies, which is also noted in the ACM draft report, was the EDISON data science framework, which, among other things, offers a comprehensive set of data science knowledge and skills (Demchenko et al., 2016, 2019a, 2019b, 2021a, 2021b). Demchenko et al. (2016) anchor the EDISON data science framework against industry standards and the literature to offer an updated description of data science skills. Schmitt et al. (2021) highlight what EDISON's skills have persisted, while Rossi (2021) further contextualized the EDISON framework within the analyze, design, develop, implement, and evaluate model to facilitate the design and delivery of data science curricula.

Some studies made an extensive review of literature and industry to inform a set of high-level data science skills. From an industry perspective, Mikalef and Krogstie (2019) gathered 229 surveys and manager interviews from executives from Norwegian firms to create a ranked list of skills for each of the technical, business, project management, and soft skills. Gottipati et al. (2021) mined the Glassdoor website to outline

(a) technical skills,

(b) Analysis by software, programming languages, and algorithms, and

(c) soft skills for the role of the data scientist.

From an academic perspective, Wu (2017) reviewed 70 programs in data science and analytics (DSA) and informed an extensive list of skills upon completion of DSA graduate programs. Costa and Santos (2017) reviewed and summarized the knowledge and skills needed from eight master's programs from academia and three certificate programs from the industry. Their skill set provides several learning outcomes required of a data scientist. Saltz et al (2018) provide a high-level summary of key learning outcomes seen in different data science/analytics programs. The work further presents the level of depth/focus (i.e., deep focus, some depth, and not a program focus) of these learning outcomes for three programs, namely data analytics, applied data science, and foundational data science.

Dichev and Dicheva (2017) conceptualize data science skills as the

- Ability to formulate productive questions,
- Ability to think computationally,
- Ability to think analytically,
- Ability to visualize and report summary data.

Rosenthal and Chung (2020), reviewed a list of courses and offered a few underlying objectives of the data science program, namely:

- "create effective mathematical solutions to analytical problems,
- create effective solutions to computing challenges in analytical projects,
- effectively organize and manage datasets for analytical projects,
- critically analyze problems and identify analytical solutions,
- communicate analytical problems, methods, and findings effectively orally, visually, and in writing,
- critically evaluate ethical, privacy, and security challenges in data analytics, and
- learning objectives that identify the action, such as communicating, evaluating, and analyzing, offer insight into how a skill is being practiced" (p. 3).

What sets this skill definition apart from others is that it does not just include a verb-noun pair but goes deeper to justify where and how data science skills would be used. Further, Adams (2020) and Belloum et al. (2019) emphasize the need for students to gain some training in a cognate or domain area.

Adaptive and timeless definitions were also deemed useful in defining data science skills. Petrova et al. (2018) use flexible terminology to make the skills of a data scientist relevant over time. The authors refer to data as unruly rather than unstructured, to pinpoint that data may often be beyond formless, it may be convoluted, noisy, and conflicting. They further suggest that data scientists need to stay on top of analytical techniques such as machine learning and work with a variety of programming languages. Sedelmaier et al. (2021) adopt a timeless definition as well: "A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions" (p. 1). Similarly, Menasalvas et al. (2019), used language that is adaptive over time: e.g., "Identify existing requirements to choose and execute the most appropriate data discovery techniques to solve a problem depending on the nature of the data and the goals to be achieved" (p. 1). Such definitions acknowledge that techniques are causal and change over time. And so, data scientists need to use relevant resources present in their time appropriately.

Some studies informed quality data science skills tailored to their offering. Bargagliotti (2019) defined data science skills based on their course units. Cetinkaya-Rundel and Ellison (2021) divide the underlying objectives into the course units and provide a set of skills needed for each unit (e.g., unit 1: exploring data which contains 6 broad learning objectives/skills). Kostadinova et al. (2018) divided data scientists into two groups: 1) active data science specialists and 2) those who are training in data science. Active data science specialists are proposed to refresh their knowledge through training courses in a field they believe they need. Many Universities offer courses for study through various MOOC platforms like Coursera, eDX, and others. Those who will train in data science are proposed to engage in a curriculum and take individual courses in parallel as well.

A few studies proposed their data science frameworks, which merit further study. Raj et al. (2019) provide a review of statistics in the UK and worldwide surrounding the courses, skills, professions, tools, and concepts covered in data science. The authors hypothesize that data-oriented competencies can be broken into data activities and their associated areas of application. Their framework consists of a two-dimensional

data-oriented competency space, where the x-axis shows data activities, and the y-axis shows specific areas of application. In their series of work, Rasheva et al. (2019a, 2019b, 2019c, 2019d) also present a new competency model with two dimensions. One dimension outlines the types of skills: hard skills, soft skills, and analytical skills. The other dimension breaks down each skill into either tasks, skills, or competencies. They also create a survey to elicit students' understanding of their data science skills.

Data science learning outcomes that need improvement

We found definitions that offered too broad a skill to lack a data science focus. Xiao et al. (2021), for example, note the following broad skills: "(1) Basic understanding: Gaining a basic understanding of the subject (e.g., factual knowledge, methods, principles, generalizations, theories) (2) Application: Learning to apply course material (to improve thinking, problem-solving, and decisions) (3) Develop skills: Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course (4) Numerical methods: Learning appropriate methods for collecting, …" (p. 4). De Veaux et al. (2017) note competencies for an undergraduate data science major using terms that may result in a superfluous list: computational and statistical thinking, mathematical foundations, Model building and assessment, algorithms and software foundation, data curation, knowledge transference—communication and responsibility. In this DeMasi et al. (2020), the authors raise and share generic efforts on co-curricular and ad-hoc methods in data science education for a range of audience levels (undergrad to faculty). They note: "goals of ad hoc efforts: improve knowledge of methods, mentoring or career development, exposure to research, community building, and other goals" (p. 14).

There were also data science definitions that were narrowly focused on a topic or technique. Inc and Wang (2017), for example, summarize challenges and actions to be taken with big data dimensions, namely volume, velocity, variety, and veracity. The author posits that a data science paradigm today encompasses big data analytics that goes beyond traditional databases.

The learning objectives/skills, therefore, are associated with big data dimensions. Hicks and Irizarry (2018) adopt the view of Nolan and Temple Lang and argue that data science is an extension of statistics. To the authors, data science is a renovation made to classical statistics curricula. Overton and Kleinschmit (2022) primarily focus on the integration of data science into Public Administrator curricula.

Studies that attributed data science skills to a set of courses also offered little insight into what competencies students need to build. In Dinov (2019), data science skills were presented as a set of required courses and knowledge bases. Zhou et al. (2020) suggest that the data science curriculum should be divided into lower and upper divisions and list a set of courses and credits. While the list provides an understanding of the order of courses that need to be taken, the course names provided in the paper alone offer little insight into what learning objectives are sought to be achieved.

DISCUSSION

The goal of this systematic literature review paper was to conceptualize data science by exploring definitions and learning outcomes surrounding data science. In total, 49 studies were reviewed from journals and conference papers. We analyzed the following factors of the studies selected: whether the data science definition was an intensional or extensional type, and which definitions were insightful vs. need improvement, and whether data science learning outcomes were a noun, or a verb-noun list, and which definitions were insightful vs. need improvement.

Summary of Findings

We conducted a systematic review of 49 research publications on data science definitions and learning outcomes, focused on the higher education level from WoS and Scopus databases. The majority of the publications were from Wos and published as conference proceedings. The number of studies published per year from 2016 to 2022 ranged from 1 to 18, with a notable increase in 2019 (18) and a decline afterward. This may be in part due to the increased publication rate of one or more authors at an instance in time (as an example, see Rasheva-Yordanova et al., 2019a). This may also highlight the perturbations world events such as the global pandemic may have on the research and discipline of data science. This may also be in part due to the research trends and attention a topic may receive at certain time intervals. Although we had set no time limit for our search, no studies from before 2016 remained after our screening process. This may suggest that data science has been more recently examined in the literature, which may be caused by the proliferation of technologies such as AI, which use data science as their cornerstone.

Of the 49 reviewed studies, 34 (69%) had included a definition for data science, and 39 (80%) had presented one or a combination of learning outcomes specific to data science. Of the 34 publications, 9 (26%) presented an extensional, and 25 (74%) presented an intensional form of definition for data science. Of the 39 publications, 10 (26%) had used a noun and 29 (74%) a verb-noun pair; see **Table 4** for a summary of references found per definition type.

We found that definitions that were too extensional often need improvement, as they lack clarity on what data science means. Instead, definitions that point to what disciplines are employed and techniques are incorporated in data science result in a more insightful extensional definition.

Definitions with a specific category and/or differentia often resulted in a more intuitive and robust understanding of data science and resulted in a more insightful intensional definition. A vague category or differentia, on the other hand, often results in an intensional definition that needs improvement.

We also found studies that utilized initiatives or skills mentioned in the ACM draft report to have learning outcomes of higher credibility and hence quality. Some studies also made an extensive review of literature and industry to inform a set of high-level data science learning outcomes. Adaptive and timeless definitions were deemed useful in defining data

science learning outcomes, given the ongoing changes to technology employed by data science.

There were also data science definitions that were narrowly focused on a topic or technique.

For example, studies that attributed data science skills to a set of courses and nouns often offered little insight into what competencies students need to build.

Our review does not aim to suggest that an intensional or extensional definition is better. Based on our interpretive coding and evaluation outlined in the conceptual framework section, we came to find extensional (7/9, 78%) and intentsional (17/25, 68%) definitions needed improvements, highlighting the work that remains to make clear and robust data science definitions.

Our review does aim to suggest that a verb-noun, as opposed to a noun learning outcome, is better. This is because verbs are at the heart of orchestrating effective learning outcomes (Adelman, 2015). Contrary to the data science dentitions that often required improvements, we found that learning outcomes were mostly insightful (31/38, 82%). This could be largely because. Established institutions, organizations, and bodies have made efforts to conceptualize data science learning outcomes as part of their instructional design and delivery.

Our findings thus highlight the diversity of concepts and skills that exist in the discipline of data science, making it easy to formulate a set of learning outcomes, but hard to provide a holistic picture of the meaning of data science.

Challenges For Conceptualizing Data Science

Overall, we find various conceptions and applications of data science to exist in the literature. How data science is understood as a discipline varies across studies. To provide an example, studies may consider data science to be:

- interchangeably used with other terms such as big data (Demchenko et al., 2019a, 2021a, 2021b)
- a separate field but shares similarities with other fields (Danyluk & Buck, 2019)
- a newly created discipline, and may further highlight that data science is a multidisciplinary field (Cao, 2017; Raj et al., 2019)
- a component of the data science programs (Rosenthal & Chung, 2020)
- deeply rooted in statistics (Hicks & Irizarry, 2018)
- highly dependent on expertise in the context of the study (Adams, 2020; Belloum et al., 2019)

Given that many technical and closely related domains may be used as part of data science, it becomes difficult to tell if data science is a subset or umbrella of the many technical domains. But this is not the only issue data science is facing. Below we further summarize some of today's challenges for conceptualizing data science.

There may be implications with augmenting data science into every discipline. Experts in every discipline must become data science literate (Belloum et al., 2019). Further, the experts need to foresee and account for the repercussions of augmenting data science into their discipline. A common issue in defining data science is in finding general categories of definitions. Studies often used an array of terms such as field, domain, concept, topic, discipline, context, and problem. We find many of these terms to be used interchangeably in literature and so not to be a good fit for making a clearcut definition. Perceiving the class of data science as a science or multidisciplinary discipline might be more productive because it highlights that data science is born out of many theories and practices. Another commonality seen across the reviewed studies is that when explaining disciplines involved in data science, they offer several fields that are varied in depth and breadth. Some of the mentioned disciplines are quite broad and vague (e.g., domain expertise), and some terms like data engineering are newly coined terms and may have roots in math and other fields (Petrova et al., 2018). The disciplines in data science, as such, and contrary to what some of the studies may suggest, are not conceptually linear but rather branched and hierarchical.

Assigning a focus and set framework (e.g., dimensions of data analytics) may improve student understanding as it makes learning outcomes more situated (Inc & Wang, 2017). However, such definitions run the risk of becoming outdated. In the event the notions and practices adopted in a focused area such as big data analytics change or lend themselves to a new field, the previously set frameworks may no longer benefit data scientists.

We find that an explanation of the time needed for training is often missing from a definition of data science and its competency (Pedersen & Caviglia, 2019). As a result, students may develop data science knowledge and skills with or without attending college. The lack of a time scale is perhaps a weakness of data science as we find there to be an abundance of certificate and degree programs (especially in online mode) and little insight into which provides a complete representation of data science.

The boundaries for the prerequisites to learning advanced data science are unclear or inconsistent. This can be because data science often has multiple streams, and by the time learners reach seniority, they may have taken a different assortment of courses from the streams (Rosenthal & Chung, 2020). This can bring benefits because each learner provides a new lens for problem-solving in complex courses. But this can also bring added challenges to the instructor and institution since the assessment of student learning outcomes becomes decentralized.

Our review of data science definitions and learning outcomes across 49 studies revealed how data science is influenced by different constructs. We find that at first glance, the most underlying influence on our conception of data science is language and the way definitions and learning outcomes for data science are formulated. Our classification of intensional or extensional definitions and noun or verb-noun learning outcomes, for example, showed an important role the semantics of language play in our understanding and dissemination of data science. Our analysis of the quality of studies highlighted that language (e.g., how data science definitions and learning outcomes are provided) sets the scene for the problem and disciplinary context, and the context dictates the constraints with technology used and/or available. And so data science is highly dependent on the language,

context, and availability of technology. These influencers can subsequently impact the understanding of the teacher, learner, or both. Contrary to popular belief that data science is a science, so it is objective and accurate, our review shows that the teaching and learning of objective scientific constructs, such as data science, are still largely influenced culturally and socially by the language and scope or context where a problem is being solved.

Future Potentials For Conceptualizing Data Science

The systematic review also informed us of several important considerations for conceptualizing data science, which are summarized below.

As more and more universities create data science programs, it is worth exploring the roles and outcomes of those who graduated from these programs against those who resorted to certificate programs only. While many certificate programs may offer the essence of data science and come to be less expensive, they may fail to offer students the experience offered by the undergraduate higher education curriculum. Unless otherwise emulated in certificate programs, students with certificates gain little exposure to curricular and co-curricular data science experiences. Generally, though, the types of skills listed across the certificate and institutional programs seem to be closely similar (Costa & Santos, 2017). This makes the need and efforts made to create higher education data science programs questionable. That is, if students can achieve similar outcomes from free or less expensive certificate programs, why should we care, or how can we justify the need for data science curricula in a higher education setting?

A data scientist is often involved in every step of the data manipulation process. Such a prominent role can bring both benefits and challenges to the outcomes of the work being carried out. A benefit can be that the data scientist is aware of and available to respond to challenges faced in every step of the data manipulation process. This is contrary to work done by multidisciplinary teams, where each stage is done by different disciplinary groups with little knowledge of the work done in the prior or following stages. A challenge can be that since a data scientist is the "one man band" in the data manipulation process, they may make errors or tend to carry out tasks in their set (preferred and familiar) way, which may not be most appropriate.

The essence of many data science courses and programs is about helping students progress through the stages of the data lifecycle. Often, the studies report on curriculum guidelines that are either newly created or borrowed from statistics or data science (Cetinkaya-Rundel & Ellison, 2021). There are, however, different considerations when creating a new curriculum versus integrating courses from other curricula. And there is still relatively little work done reporting on the accreditation of data science as a discipline. Another important consideration is what stakeholders should be involved in the discussion of data science (Leidig & Cassel, 2020). Data science is an applied discipline, and so both industry and academia can impact the notions and skills for data science. More importantly, we need to consider how the landscape may change if artificially intelligent programs themselves had a say and vote in this discussion.

Understanding modern and fair practices in data science is needed. AI, for example, has been around for many years, but it is still deemed a new field. Disciplines, even ones that may seem remotely related to data science (e.g., liberal arts), are racing to create a sort of data science initiative (e.g., program, course, workshop, etc.). A feature of introductory data science courses is to be able to involve a wide range of learners and require little to no mathematical prerequisites. Creating many introductory programs is one factor in increasing access; however, it is engagement that determines the outcomes of access. Students' predispositions and beliefs, the perceived usefulness of programs, and the efforts put into creating programs may all impact engagement and learning development. Attention thus needs to be paid to the development of data science as a discipline (Rosenthal & Chung, 2020). More importantly, accepting diverse ideas and research and resisting monopolization of research in data science literature and community can greatly benefit an inclusive and non-biased representation of data science.

CONCLUSION

The goal of this systematic review paper is to summarize and assess the quality of definitions presented and the types of learning outcomes promoted for data science. Grounded in an interpretive framework, we summarize and classify data science definitions and skills, present our analysis of insightful versus needs improvement descriptions, and offer our justifications. We find that most of the data science definitions follow an intensional format (containing a category and differentia) and become more insightful definitions when having a specific category and differentia. Further, most of the data science skill definitions follow a verb-noun format and become more insightful definitions when they have an adaptive and timeless description. The contribution of this work is in socially constructing a classification and evaluation method for data science definitions and skills, which can ultimately provide insight into how data science conceptualization may progress as a new field.

Author contributions: BM: conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, visualization, writing–original draft, and writing–review & editing & **TD:** conceptualization, funding acquisition, resources, software, supervision, validation, and writing–review & editing. Both authors approved the final version of the article.

Funding: This article was supported by the Canada Research Chair Program and the Canada Foundation for Innovation.

Ethics declaration: The authors stated that no ethical approval was required for this study. This study is based on a systematic review of previously published literature and does not involve human participants or identifiable personal data.

AI statement: The authors stated that they used AI-based tools to assist with language editing and clarity. All intellectual contributions remain those of the authors.

Declaration of interest: The authors declared no competing interest.

Availability of data and materials: All data generated or analyzed during this study are available for sharing when appropriate request is directed to corresponding author.

REFERENCES

Adams, J. C. (2020). Creating a balanced data science program. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 185-191). ACM. https://doi.org/10.1145/3328778.3366800

Adams, S. (2006). *An introduction to learning outcomes: A consideration of the nature, function and position of learning outcomes in the creation of the European Higher Education Area*. European University Association.

Adelman, C. (2015). *To imagine a verb: The language and syntax of learning outcomes statements*. National Institute for Learning Outcomes Assessment.

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Pearson.

Aragona, B. (2017). New data science: The sociological point of view. In N. Lauro, E. Amaturo, M. Grassia, B. Aragona, & M. Marino (Eds.), *Data science and social research. Studies in classification, data analysis, and knowledge organization* (pp. 17-24). Springer. https://doi.org/10.1007/978-3-319-55477-8_3

Attard, A. (2010). *Student centred learning: An insight into theory and practice*. European Students Union.

Bargagliotti, A. (2019). Integrating data analysis and statistics across disciplines. In *Proceedings of the 5th International Conference on Higher Education Advances* (pp. 341-352). https://doi.org/10.4995/HEAd19.2019.9236

Baume, D. (2009). *Writing and using good learning outcomes*. Leeds Metropolitan University.

Belloum, A. S. Z., Koulouzis, S., Wiktorski, T., & Manieri, A. (2019). Bridging the demand and the offer in data science. *Concurrency and Computation–Practice & Experience, 31*(17), Article e5200. https://doi.org/10.1002/cpe.5200

Blair, J. R. S., Jones, L., Leidig, P., Murray, S., Raj, R. K., & Romanowski, C. J. (2021). Establishing ABET accreditation criteria for data science. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 535-540). ACM. https://doi.org/10.1145/3408877.3432445

Blei, D. M., & Smyth, P. (2017). Science and data science. *PNAS, 114*(33), 8689-8692. https://doi.org/10.1073/pnas.1702076114

Bourner, T. (1997). Teaching methods for learning outcomes. *Education+ Training, 30*(9), 344-348. https://doi.org/10.1108/00400919710192377

Britannica. (2023). Intension and extension. *Britannica*. https://www.britannica.com/topic/intension

Cao, L. (2016). Data science: Nature and pitfalls. *IEEE Intelligent Systems, 31*(5), 66-75. https://doi.org/10.1109/MIS.2016.86

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys, 50*(3), Article 43. https://doi.org/10.1145/3076253

Carmichael, I., & Marron, J. S. (2018). Data science vs. statistics: Two cultures? *Japanese Journal of Statistics and Data Science, 1*, 117-138. https://doi.org/10.1007/s42081-018-0009-3

Cetinkaya-Rundel, M., & Ellison, V. (2021). A fresh look at introductory data science. *Journal of Statistics and Data Science Education, 29*(Suppl 1), S16-S26. https://doi.org/10.1080/10691898.2020.1804497

Christozov, D., Rasheva-Yordanova, K., & Toleva-Stoimenova, S. (2019). Designing data science curriculum in a way to address expected students entry competences. In *Proceedings of the 13th International Technology, Education and Development Conference* (pp. 2635-2640). https://doi.org/10.21125/inted.2019.0708

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/10.2307/1403527

Costa, C., & Santos, M. Y. (2017). A conceptual model for the professional profile of a data scientist. In Á. Rocha, A. Correia, H. Adeli, L. Reis, & S. Costanzo (Eds.), *Recent advances in information systems and technologies. WorldCIST 2017. Advances in intelligent systems and computing, vol 570* (pp. 453-463). Springer. https://doi.org/10.1007/978-3-319-56538-5_46

Danyluk, A., & Buck, S. (2019). Artificial intelligence competencies for data science undergraduate curricula. *AAAI Conference on Artificial Intelligence, 33*(1), 9746-9747. https://doi.org/10.1609/aaai.v33i01.33019746

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., ... Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application, 4*, 15-30. https://doi.org/10.1146/annurev-statistics-060116-053930

De Veaux, R., Hoerl, R., Snee, R., & Velleman, P. (2022). Toward holistic data science education. *Statistics Education Research Journal, 21*(2), Article 2. https://doi.org/10.52041/serj.v21i2.40

Dedge Parks, D. M. (2017). *Defining data science and data scientist* [Master's thesis, University of South Florida].

DeMasi, O., Paxton, A., & Koy, K. (2020). Ad hoc efforts for advancing data science education. *PLoS Computational Biology, 16*(5), Article e1007695. https://doi.org/10.1371/journal.pcbi.1007695

Demchenko, Y., Belloum, A. S. Z., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., & Brewer, S. (2016). EDISON data science framework: A foundation for building data science profession for research and industry. In *Proceedings of the 2016 8th IEEE International Conference on Cloud Computing Technology and Science* (pp. 620-626). IEEE. https://doi.org/10.1109/CloudCom.2016.0107

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

Demchenko, Y., Jose, C. G. J., Brewer, S., & Wiktorski, T. (2021a). EDISON data science framework (EDSF): Addressing demand for data science and analytics competences for the data driven digital economy. In *Proceedings of the EDUCON* (pp. 1688-1693). https://doi.org/10.1109/EDUCON46332.2021.9453997

Demchenko, Y., Maijer, M., & Comminiello, L. (2021b). Data scientist professional re-visited: Competences definition and assessment, curriculum and education path design. In *Proceedings of the 2021 4th International Conference on Big Data and Education* (pp. 52-62). ACM. https://doi.org/10.1145/3451400.3451409

Demchenko, Y., Wiktorski, T., Cuadrado Gallego, J., & Brewer, S. (2019b). EDISON data science framework (EDSF) extension to address transversal skills required by emerging Industry 4.0 transformation. In *Proceedings of the 2019 15th International Conference on eScience* (pp. 553-559). IEEE. https://doi.org/10.1109/eScience.2019.00076

Dichev, C., & Dicheva, D. (2017). Towards data science literacy. In *Proceedings of the International Conference on Computational Science* (pp. 2151-2160). https://doi.org/10.1016/j.procs.2017.05.240

Dinov, I. D. (2019). Quant data science meets dexterous artistry. *International Journal of Data Science and Analytics, 7*(2), 81-86. https://doi.org/10.1007/s41060-018-0138-6

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

Ecclestone, K. (1999). Empowering or ensnaring?: The implications of outcomes-based assessment in higher education. *Higher Education Quarterly, 53*(1), 29-48. https://doi.org/10.1111/1468-2273.00111

Ecclestone, K. (2010). *Transforming formative assessment in lifelong learning*. Open University Press.

Education Development Center. (2003). Understanding and creating definitions. *EDC*. https://www2.edc.org/makingmath/handbook/teacher/definitions/definitions.asp

Ellis, G. (2004). *Rough guide to learning outcomes*. The University of Teesside-Center for Learning and Quality Enhancement.

Elmasri, R., & Navathe, S. (2010). *Fundamentals of database systems*. Addison-Wesley.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44-49. https://doi.org/10.52041/serj.v16i1.213

Fekete, A., Kay, J., & Röhm, U. (2021). A data-centric computing curriculum for a data science major. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 865-871). ACM. https://doi.org/10.1145/3408877.3432457

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

Flowerdew, J. (1992). Definitions in science lectures. *Applied Linguistics, 13*(2), 202-221. https://doi.org/10.1093/applin/13.2.202

Friedman, A. (2019). Data science syllabi measuring its content. *Education and Information Technologies, 24*(6), 3467-3481. https://doi.org/10.1007/s10639-019-09935-x

Gibbs, G. (1995). *Assessing student centred courses*. Oxford Centre for Staff Learning and Development.

Gillenson, M. L. (2012). *Fundamentals of database management systems*. John Wiley & Sons.

Gottipati, S., Shim, K. J., & Sahoo, S. (2021). Glassdoor job description analytics–Analyzing data science professional roles and skills. In *Proceedings of the 2021 IEEE Global Engineering Education Conference* (pp. 1335-1342). https://doi.org/10.1109/EDUCON46332.2021.9453931

HEQCO. (2005). Learning outcomes. *HEQCO*. http://www.heqco.ca/en-ca/OurPriorities/LearningOutcomes/Pages/Home.aspx

Heywood, J. (2000). *Assessment in higher education: Student learning, teaching, programmes and institutions (vol. 56)*. Jessica Kingsley Publishers.

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *The American Statistician, 72*(4), 382-391. https://doi.org/10.1080/00031305.2017.1356747

Hoffer, J. A., Ramesh, V., & Topi, H. (2012). *Modern database management* (11th ed.). Prentice Hall.

Hussey, T., & Smith, P. (2003). The uses of learning outcomes. *Teaching in Higher Education, 8*(3), 357-368. https://doi.org/10.1080/13562510309399

Inc, D. P., & Wang, J. P. (2017). A data science paradigm shift in the age of big data. In *Proceedings of the International Conference on Modern Education and Information Technology* (pp. 402-406).

Johner Institute. (2023). How do you write a good definition? *Johner Institute*. https://www.johner-institute.com/articles/regulatory-affairs/and-more/how-do-you-write-a-good-definition/

Kakeshita, T., Ishii, K., Ishikawa, Y., Matsubara, H., Matsuo, Y., Murata, T., Nakano, M., Nakatani, T., Okumura, H., Takahashi, N., Uchida, G., Uematsu, E., Saeki, S., & Kato, H. (2022). Development of IPSJ data science curriculum standard. In D. Passey, D. Leahy, L. Williams, J. Holvikivi, & M. Ruohonen (Eds.), *Digital transformation of education and learning–Past, present and future. OCCE 2021. IFIP advances in information and communication technology, vol 642* (pp. 156-167). Springer. https://doi.org/10.1007/978-3-030-97986-7_13

Kennedy, D., Hyland, A., & Ryan, N. (2006). *Writing and using learning outcomes, a practical guide*. European University Association.

Kostadinova, I., Petrova, P., & Iliev, E. (2018). Analysis of data science courses through the prism of the digital divide. In *Proceedings of the EDULEARN18: 10th International Conference on Education and New Learning Technologies* (pp. 3903-3911). https://doi.org/10.21125/edulearn.2018.0993

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

Leidig, P. M., & Cassel, L. (2020). ACM taskforce efforts on computing competencies for undergraduate data science curricula. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 519-520). ACM. https://doi.org/10.1145/3341525.3393962

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy, 67*(13), 427-446. https://doi.org/10.2307/2023861

Litowitz, B. (1977). Learning to make definitions. *Journal of Child Language, 4*(2), 289-304. https://doi.org/10.1017/S0305000900001665

Mathworks. (2023). Analyze text data using multiword phrases. *The MathWorks, Inc*. https://www.mathworks.com/help/textanalytics/ug/analyze-text-data-using-multi-word-phrases.html

Menasalvas, E., Moreno, A. M., & Swoboda, N. (2019). A proposal for recognizing skills in data science using open badges. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (p. 316). ACM. https://doi.org/10.1145/3304221.3325561

Mikalef, P., & Krogstie, J. (2019). Investigating the data science skill Gap: An empirical analysis. In *Proceedings of the 2019 IEEE Global Engineering Education Conference* (pp. 1275-1284). IEEE. https://doi.org/10.1109/EDUCON.2019.8725066

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

National Science Foundation. (2014). Data science at NSF draft report of StatSNSF Committee: Revisions since january MPSAC meeting. *NSF*. https://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf

Niinimäki, M. (2000). Intensional and extensional languages in conceptual modelling. *ResearchGate*. https://www.researchgate.net/publication/221013931_Intensional_and_Extensional_Languages_in_Conceptual_Modelling

Niinimäki, M. (2004). *Conceptual modelling languages* [PhD thesis, University of Tampere].

Nilson, L. B. (2016). *Teaching at its best: A research-based resource for college instructors*. John Wiley & Sons.

Onyema, E. M., Khan, R., Eucheria, N. C., & Kumar, T. (2023). Impact of mobile technology and use of big data in physics education during coronavirus lockdown. *Big Data Mining and Analytics, 6*(3), 381-389. https://doi.org/10.26599/BDMA.2022.9020013

Overton, M., & Kleinschmit, S. (2022). Data science literacy: Toward a philosophy of accessible and adaptable data science skill development in public administration programs. *Teaching Public Administration, 40*(3), 354-365. https://doi.org/10.1177/01447394211004990

Pedersen, A. Y., & Caviglia, F. (2019). Data literacy as a compound competence. *Digital Science, 850*(173), 166-173. https://doi.org/10.1007/978-3-030-02351-5_21

Peregrin, J. (2007). Extensional vs. intensional logic. In D. M. Gabbay, P. Thagard, & J. Woods (Eds.), *Handbook of the philosophy of science. Volume 5: Philosophy of logic* (pp. 913-942). Elsevier. https://doi.org/10.1016/B978-044451541-4/50024-5

Petrova, P., Kostadinova, I., & Chantov, V. (2018). Analysis of online courses to acquire data science literacy. In *Proceedings of the EDULEARN18: 10th International Conference on Education and New Learning Technologies* (pp. 4073-4080). https://doi.org/10.21125/edulearn.2018.1031

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods, 19*(2), 159-203. https://doi.org/10.1177/1094428115624965

Purdue Online Writing Lab. (2023). Writing definitions. *Purdue Online Writing Lab*. https://owl.purdue.edu/owl/general_writing/common_writing_assignments/definitions.html

Raj, R. K., Parrish, A., Impagliazzo, J., Romanowski, C. J., Aly, S. G., Bennett, C. C., Davis, K. C., McGettrick, A., Pereira, T. S. M., & Sundin, L. (2019). An empirical approach to understanding data science and engineering education. In *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education* (pp. 73-87). ACM. https://doi.org/10.1145/3344429.3372503

Rasheva-Yordanova, K., Chantov, V., Kostadinova, I., Iliev, E., Petrova, P., & Nikolova, B. (2019a). Forming of data science competence for bridging the digital divide. In *Proceedings of the Future of Education* (pp. 174-179).

Rasheva-Yordanova, K., Toleva-Stoimenova, S., & Christozov, D. (2019b). Data science: Challenges and trends. In *Proceedings of the 12th Annual International Conference of Education, Research and Innovation* (pp. 10935-10943). https://doi.org/10.21125/iceri.2019.2689

Rasheva-Yordanova, K., Toleva-Stoimenova, S., Christozov, D., & Kostadinova, I. (2019c). Road map in developing Data science competences. In *Proceedings of INTED2019 Conference* (pp. 6643-6650). https://doi.org/10.21125/inted.2019.1614

Rasheva-Yordanova, K., Toleva-Stoimenova, S., Kostadinova, I., Christozov, D., & Nikolova, B. (2019d). Assessing the entry competences of data science master program potential students. In EDULEARN19: *11th International Conference on Education and New Learning Technologies* (pp. 4389-4394). https://doi.org/10.21125/edulearn.2019.1106

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

Rossi, R. (2021). Data science education based on ADDIE model and the EDISON framework. In *Proceedings of the 2021 International Conference on Big Data Engineering and Education* (pp. 40-45). https://doi.org/10.1109/BDEE52938.2021.00013

Salas-Rueda, R. A. (2020). TPACK: Technological, pedagogical and content model necessary to improve the educational process on mathematics through a web application? *International Electronic Journal of Mathematics Education, 15*(1), Article em0551. https://doi.org/10.29333/iejme/5887

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*, 615-624. https://doi.org/10.17705/1CAIS.04333

Samuel, B. M., Khatri, V., & Ramesh, V. (2018). Exploring the effects of extensional versus intensional representations on domain understanding. *MIS Quarterly, 42*(4), 1187-1210. https://doi.org/10.25300/MISQ/2018/13255

Schmitt, K. R., Clark, L., Kinnaird, K. M., Wertz, R. E., & Sandstede, B. (2021). Evaluation of EDISON's data science competency framework through a comparative literature analysis. *Foundations of Data Science, 5*(2), 177-198. https://doi.org/10.3934/fods.2021031

Schwandt, T. A. (1994). Constructivist, interpretivist approaches to human inquiry. *Handbook of Qualitative Research, 1*(1994), 118-137.

Sedelmaier, Y., Landes, D., & Erculei, E. (2021). How to design a competence-oriented study program for data scientists? In *Proceedings of the 2021 IEEE Global Engineering Education Conference* (pp. 1588-1592). https://doi.org/10.1109/EDUCON46332.2021.9453949

Stanford University. (2023). Guides & standards–Data governance. *Stanford University*. https://datagovernance.stanford.edu/guides-standards

Stodden, V. (2020). The data science life cycle: A disciplined approach to advancing data science as a science. *Communications of the ACM, 63*(7), 58-66. https://doi.org/10.1145/3360646

Studtmann, P. (2021). Aristotle's categories. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*.

Toleva-Stoimenova, S., Christozov, D., & Rasheva-Yordanova, K. (2019). Entry competences assessment of data science potential students. In *Proceedings of INTED2019 Conference* (pp. 4248-4256). https://doi.org/10.21125/inted.2019.1066

UCF. (2023). The definitions book: How to write definitions. *UCF*. https://www.unifiedcompliance.com/education/how-to-write-definitions/#What-is-a-Definition?

van der Aalst, W. M. P. (2019). Responsible data science in a dynamic world: The four essential elements of data science. In L. Strous, & V. Cerf (Eds.), *Internet of things. Information processing in an increasingly connected world. IFIPIoT 2018. IFIP advances in information and communication technology, vol 548* (pp. 3-10). Springer. https://doi.org/10.1007/978-3-030-15651-0_1

Vittorini, P., & Galassi, A. (2021). From blended to online due to the COVID outbreak: The case study of a data science course. *Open Learning: The Journal of Open Distance and e-Learning, 36*(3), 212-230. https://doi.org/10.1080/02680513.2021.1973399

Whitney, C., Grossman, M., and Kircher, T. T. J. (2009). The influence of multiple primes on bottom-up and top-down regulation during meaning retrieval: Evidence for 2 distinct neural networks. *Cerebral Cortex, 19*(11), 2548-2560. https://doi.org/10.1093/cercor/bhp007

Wu, H. (2017). Systematic study of big data science and analytics programs. In *Proceedings of the ASEE Annual Conference and Exposition*. ASEE. https://doi.org/10.18260/1-2--28900

Xiao, T., Greenberg, R. I., & Albert, M. V. (2021). Design and assessment of a task-driven introductory data science course taught concurrently in multiple languages: Python, R, and MATLAB. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (pp. 290-295). ACM. https://doi.org/10.1145/3430665.3456364

Zhou, T., Jiang, D., Wang, F., Li, X., & Zheng, L. (2020). A CDIO oriented curriculum for division of data science and big data technologies: The content, process of derivation, and levels of proficiency. In *Proceedings of the 2020 8th International Conference on Digital Home* (pp. 172-177). IEEE. https://doi.org/10.1109/ICDH51081.2020.00037

Zhu, Y., & Xiong, Y. (2015). *Defining data science*. arXiv. https://doi.org/10.48550/arXiv.1501.05039