Seeing the Whole Elephant — A Comprehensive Framework for Data Education

Iain J. Cruickshank
iain.cruickshank@westpoint.edu
United States Military Academy
West Point, New York, USA

Nathaniel D. Bastian
nathaniel.bastian@westpoint.edu
United States Military Academy
West Point, New York, USA

Jean R.S. Blair
jean.blair@westpoint.edu
United States Military Academy
West Point, New York, USA

Christa M. Chewar
christa.chewar@westpoint.edu
United States Military Academy
West Point, New York, USA

Edward Sobiesk
edward.sobiesk@westpoint.edu
United States Military Academy
West Point, New York, USA

ABSTRACT

While there has been exciting recent progress in developing curricula for data education, more work is needed to establish connection points between data science, computer science, and other disciplines. This position paper argues for a broader, more all-encompassing perspective on data education to ensure opportunities are not missed. Our primary contribution is a comprehensive framework to visualize the data education landscape with the goal of improving understanding of how the various data education disciplines, work roles, core competencies, and skills fit together. Students and educators could benefit from such a framework, and all constituents of data education might better communicate requirements and more effectively make use of data and the data workforce.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; *Model curricula*; **Computing education programs**; *Computing literacy*; Employment issues; • **Mathematics of computing**; • **Computing methodologies** → Machine learning;

KEYWORDS

Data Education, Computer Science Education, Curriculum Development, Data Science, Data Literacy

ACM Reference Format:
Iain J. Cruickshank, Nathaniel D. Bastian, Jean R.S. Blair, Christa M. Chewar, and Edward Sobiesk. 2024. Seeing the Whole Elephant — A Comprehensive Framework for Data Education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024), March 20–23, 2024, Portland, OR, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3626252.3630922

1 INTRODUCTION

The famous parable "The Blind Men and the Elephant" [32] describes six blind men each encountering an elephant. One approaches the elephant's 'broad and sturdy side' and thinks it like a wall; another touches only the tusk, which reminds of a spear; others, feeling only one distinct part, observe the elephant as a snake, tree, fan, or rope. Although the argument between the blind men based on their narrow perspectives is laughable, we worry this parable resonates too closely with the fractured perspective of data education.

As a remedy, this position paper proposes a framework that comprehensively envisions the landscape of computing-based data education, aspiring to leave us better than the outcome of our metaphorical parable. We are heartened to see compelling proposals for unifying, inclusive views of data science [26]. Many recent educational research works on data science, data engineering, and other computing-related topics [8, 12, 13, 28, 36, 37] focus on narrowed elements of data education, such as the development of a curriculum for a data science major or what curricula content is needed to introduce machine learning. However, despite previous efforts [12, 14, 28], we believe the computing education community would benefit from further work that ties together all aspects of the data education landscape into one conceptual perspective.

Such a framework is necessary for understanding how various disciplines, work roles, core competencies, and skills fit together so that everyone's diverse contributions are best appreciated and leveraged. With this, educators, students, employers, and all other constituents of data education are better enabled to improve professionalization for the varied components of a data-related workforce, who must integrate effectively to succeed. Urgency for this grows daily, as endeavors across all disciplines increasingly rely on effective use of data to support core functions, propagating to data-driven industries and even a data-driven society [27] and prompting contemplation of organizational change or new policies to facilitate evolution into a more data-driven civilization [5, 11, 27, 35].

2 RELATED RESEARCH

Many related works address various aspects of data education. Most focus on higher education approaches to Data Science [8, 12, 13, 28, 36, 37] or Data Analytics [22, 23, 31]. Recent initiatives promote Data Literacy, which aim to teach fundamental data and statistics concepts to a general audience so that all are prepared for a future that will prominently feature the use of digital data and data-driven tools in most, if not all, aspects of life [15, 16, 18, 29]. Notably, the Computing Competencies for Undergraduate Data Science Curricula 2021 (CCDS2021) [12] articulates the role of

SIGCSE 2024, March 20-23, 2024, Portland, Oregon, United States
2024. ACM ISBN 979-8-4007-0423-9/24/03.
https://doi.org/10.1145/3626252.3630922

computing discipline-specific contributions to Data Science. Further, the ITiCSE Working Group report on this subject identified several Data Science and Data Engineering education programs, resources, and challenges [28]. That report recommended categorizing Data Science and Data Engineering into the more cohesive discipline of Data Science and Engineering, since the two are complementary in nature and have significant overlap in desired competencies and skills. Seeing these efforts as ripe for extension, our paper proposes an all-encompassing framework to showcase the intersection of disciplines, competencies, and work roles across the data landscape.

For many aspects of a data-driven workflow, both free and proprietary tools are available. Although rapidly improving in usability and prevalence, configuration and application to situational needs may require programming skills and an understanding of various tasks in a data value chain [34]. Recent studies found several issues resulting from attempts by academics in various fields to implement machine learning with data tools [19, 30]. Many of these issues stem from incorrect data handling and/or not applying the appropriate tool for the problem. This challenge of using data tools in research has also been noted by other authors who seek to develop practices around their use for non-computer science researchers [20]. The proper use of data tools is emerging as a significant educational task that would benefit from a unifying framework.

Finally, it is helpful to consider how other computing disciplines recently wrestled with defining their field. For example, defining Cyber Education faced similar challenges with understanding what "Cyber" constituted and how cyber-related to existing educational offerings and paradigms [6, 7, 33]. An important result of these works was defining cyber education in a way that touches every discipline and is significantly broader than a computing-only field. Data education requires this same perspective.

3 A FRAMEWORK FOR DATA EDUCATION

With the hope of seeing the metaphorical whole elephant and not just its tusk, we start by introducing the organizing structures of our framework for data education. The primary artifact of this position paper appears later as Figure 1.

3.1 Core Competency Categories

Core competency categories combine and build on previous works that introduce and address competencies associated with data related work roles [9, 10, 12, 15, 25, 28]. Before delving into our specific competency categories, we highlight two important notes. First, all of the competencies involve (to some extent) a human using computing tools. Invariably, the extent of computing contributions to the activity will grow, and in some situations, there may be tools that fully replace human participation for a respective competency. This highlights the fact that education needs to be agile and sensitive to context. Second, the order that the core competencies are applied can vary and is often iterative in nature. For example, a data scientist typically performs Exploratory Data Analysis *before* Model Building whereas a machine learning engineer will often engage in Model Building first, then using Exploratory Data Analysis as a diagnostic tool.

Table 1 previews the core competency categories we consider imperative for data education, with a few representative knowledge and skill areas for each. When the competency categories are laid out in Figure 1 along the x-axis, they are further organized along a "prepare," "build/maintain," and "use" spectrum — inspired by [28] — based on the activities of data work that the core competency is primarily used for (e.g., data management is primarily needed to prepare for either building a data-enabled technology or data analysis). We note and depict that there is some overlap within this activity spectrum, so we discuss each competency category where we feel it is most prevalent.

Table 1: Data Work Core Competency Categories

Core Competency	Representative Knowledge and Skills
Domain Knowledge	• Stakeholder interaction • Expressing needs • Explaining domain • Articulating user feedback
Data Problem Formulation	• Requirements development and revision • Solution architecture and design • Accounting for law and policy • Data governance
Data Management	• Data collection (e.g., sensors, IoT, purchasing) • Collection ethics • Databases • Privacy and security • Data pipelines • Data quality assurance and governance
Infrastructure Building and Management	• Infrastructure engineering and administration (e.g., operating systems, networks, platform development, security) • Cloud computing • Internet of Things (IoT)
Model Building	• Machine Learning (ML) • Simulation and Optimization • Artificial Intelligence (AI) • Model selection
Model Production and Management	• Model tuning and packaging • MLOps (i.e., deploy, monitor, maintain)
Exploratory Data Analysis	• Descriptive statistics • Data quality evaluation • Data visualization
Communicating Data Results	• Results interpretation • Data story telling • Dash-boarding • Presentation ethics
Data Product Use	• Tool training • Appropriate tool use • Assess and provide feedback

One feature of our framework is its breadth; it includes competencies that are frequently left out or not accounted for when thinking about data work roles or disciplines. We feel that the functions of some competencies, such as infrastructure engineering or data problem formulation, are too often not addressed despite being vital to successful data work [21].

3.1.1 Prepare. The core competencies within the "prepare" activities of the spectrum primarily collect data and set conditions for

creating data-enabled products or conducting data analyses. Computer scientists play a critical role but must partner effectively with domain users and other disciplinary experts.

The **domain knowledge** category centers around determining which elements of an organization's operations one wishes to be data-enabled. Domain knowledge is needed to identify proper use cases for desired data-driven analysis, as well as the key decisions around collecting the appropriate data. For the purposes of our framework, we assume that developers of data systems are not typically the customers, and hence do not provide domain knowledge.

Data problem formulation includes identifying problems that can be solved by data, identifying potential data-enabled opportunities, and designing solutions. Simply put, one must determine the extent to which a problem is amenable to an analytical solution. It requires identifying relevant laws and policies for the particular domain of data and making sure that potential ethical considerations are addressed. Further, problem formulation can take into account user feedback from existing data products or analyses in order to refine existing problem formulations and iterate through solutions.

Data management includes data sourcing, collection, processing, curation, and maintenance, as well as designing and following a governance plan to ensure data provenance. Data acquisition methods may include using logs, sensors, Internet of Things (IoT) devices, web scraping and even purchasing data. A key consideration is applying appropriate ethics around data collection and privacy. Having acquired the data, it is also necessary to curate it to make it available for use in a data-enabled product or data analysis. Data curation is all of the steps for securely cleaning, storing, and transmitting data to appropriate devices; curation includes database design and implementation, big data tasks, creating/managing data pipelines, and Extract-Transform-Load (ETL) procedures.

Infrastructure building & management is a critical enabler of any data project and involves creating a computational infrastructure. Any data analysis or data-enabled product requires hardware, software, and other appropriate components to run. This infrastructure must be set up before any data work can be done and must be continuously secured and maintained over the course of any data product development, for the lifetime of the product. Without the computational infrastructure in place, a data-enabled product or data analysis becomes impossible.

3.1.2 Build/Maintain. The core competencies that fall primarily in the "build" and "maintain" activities of the spectrum are those most associated with data science work.

Model building includes the skills of applying machine learning, simulation, optimization, model design, and implementation ethics. The outcome of model building is frequently a set of featurized data combined with a computational model for that data.

Model production and management is another important core competency that is involved in the creation of a data-enabled product. Model production requires the skills needed to transfer a model from a development environment to a "live" production environment, thus making it a fully data-enabled product. Model management includes the skills coalescing into the concept of "MLOps," like model monitoring and model updating [3]. This also includes the maintenance, improvement, and retirement of computational models and their supporting data pipelines. Much of the difficulties around using machine learning models in the real world have driven the need for skills within this category.

Exploratory data analysis is most associated with data science or data analysis work and consists of holistically understanding a data set. This includes understanding its patterns, relationships, quality, and outliers. Exploratory data analysis includes the skills of effectively using descriptive statistics and data visualization.

3.1.3 Use. This activity starts with exploratory data analysis and is continued into other activities that directly deliver value and enable effective human-machine teaming. The domain experts, facilitated by product/project managers, are most important here.

Communicating data results core competencies are mostly about presenting data results and includes skills in results interpretation, data storytelling, dashboarding, and adhering to presentation ethics. This includes multi-modal approaches (visual, audio, etc.) to communicate findings from data and provide information about the data itself (data provenance, data quality, etc.).

Data product use typically requires domain knowledge and some skills in understanding at a high level how the data product works and how to use it for appropriate and optimal results. Thus, while some data-enabled technologies will require minimal additional skills to use, it is still important from an educational perspective to ensure students are equipped with the knowledge and ability to use, and know when to use, a data-enabled technology.

3.2 Higher-Order Competencies

Higher-order Knowledge, Skills, and Attitudes (KSAs) cross all competency categories and are essential for any successful data work. These fall into three broad categories.

Professional considerations are related to ethics, governance, law, and policy. Examples appear as representative KSAs in Table 1.

Inter- and multi-disciplinary competencies are more necessary than in other academic/professional endeavors since data work necessarily spans multiple disciplines. These interdisciplinary (an individual's use of personal knowledge of more than one discipline) and multidisciplinary (a collaborative effort between individuals with different disciplinary backgrounds) KSAs include many of the well-articulated, higher-education learning outcomes described in the Association of American Colleges and Universities VALUE rubrics for: (1) integrative learning, (2) intercultural knowledge and competence, (3) teamwork, and (4) oral communication [24].

Critical thinking and problem-solving competencies are also important and have their own VALUE rubrics [24].

3.3 Educational Disciplines

While we intuitively felt that no discipline sufficiently encompasses all activities in a data pipeline, seeing a list of relevant disciplines compared across the competencies was telling. Aspects of nearly every discipline either currently require — or will likely require in the very near future — some knowledge of and interaction with data-enabled technologies [15, 27, 35]. The top portion of the y-axis in Figure 1 lists groupings of the many disciplines that contribute directly to the data education landscape, which are introduced in more detail here. Note, however, that the framework can be scoped

to whatever level of abstraction best fits its use; for example, another version could include rows for every major offered at a college.

Mathematics, Statistics, and Operations Research represent the disciplines that largely focus on starting with a problem and (hopefully) a clean set of data within a given domain, then using that data to formulate and build appropriate quantitative models and perform analysis. These disciplines cover how to properly frame problems through a mathematical lens, explore and prepare data for modeling, develop and evaluate models, and then ensure the modeling results or analysis are well communicated and not biased.

Computer Science, Software Engineering, and Other Computing represent disciplines that contribute algorithms for statistical model implementations, machine learning, information retrieval and visualization, as well as developing infrastructures and software applications to collect, process, store, secure, and reliably transmit data. These disciplines cover how to analyze problems and then design, build, test, and deliver the computing services through which all data processing occurs. Students in these programs must also consider ethical, legal, privacy, and security principles governing data handling as they design, evaluate, and tune artificial intelligence algorithms and machine learning systems.

Data Science represents a recently formalized discipline to more directly address the growing need for data-savvy professionals. Data Science is typically more "data-" than "problem"-driven in mindset, when compared to Mathematics, Statistics, and Operations Research [31]. This discipline covers data discovery (searching for different sources of data and capturing it), data preparation (converting structured and unstructured data into a useful format, to include cleaning and validating), devising algorithms and/or building mathematical models (using variables and equations to establish relationships), analyzing the modeling results to identify patterns/trends/insights, operationalizing models by getting them into action, and communicating findings to stakeholders using visualization and other means. This discipline is so broad that typically academic offerings will be focused more on one of two possible disciplinary categories it overlaps with: Mathematics, Statistics, and Operations Research and Computer Science, Software Engineering, and Other Computing. This dual focus is in part why ABET and its supporting professional societies created program criteria for Data Science in both the Computing Accreditation Commission and the Applied Natural Science Commission[1, 2]; in fact Data Science is the first discipline to have program criteria in multiple commissions.

The remaining discipline rows interact with the data landscape but may not directly contribute to the data workforce. **Electrical Engineering and Computer Engineering** professionals develop hardware for data collection, storage, processing, and display. **Other Engineering and Science** includes disciplines that apply the engineering design process or the scientific method to data. This set comprises traditional disciplines such as civil, mechanical, environmental, systems, and industrial engineering, as well as geography, physics, chemistry, biology, economics, psychology, sociology, and political science. **All Others** covers disciplines traditionally less quantitative but focused on humans, cultures, and society, like languages, law, literature, history, philosophy, religion, and the arts. Some already teach rudimentary data skills and are researching how to adopt data techniques in their domains (e.g., [4], [17]).

3.4 Data Work Roles

Similar to the educational disciplines, examining competencies from work role perspectives provides important insights. The lower rows (y-axis) in Figure 1 represent current common work roles and are consistent with those mentioned in [10, 28].

Data Scientists use a scientific process to extract insights, discover patterns, and explore trends in data, while also developing data-driven models and/or theories based on data. On the other hand, a **Data Engineer** creates, acquires, curates, and constructs systems for handling data; they build and maintain tools, frameworks, and services with a focus on constructing data pipelines with an understanding of data models, database design, information flow, query execution and optimization, and comparative analysis of data stores. Based upon the observations from [28] in which the main work role of Data Science is specializing into sub-work roles, we have also included specialized data work roles such as Machine Learning Engineer [9].

Machine Learning Engineers are generally those personnel who design and develop the tools, systems, and processes that enable machine learning implementation and maintenance, driving the incorporation of developed machine learning models and pipelines into production-level machine learning systems in both on-premise and cloud-based delivery models. We also included the work role of Infrastructure Engineering, where an **Infrastructure Engineer** is generally responsible for the engineering of the information technology infrastructure, hardware (on-premise and cloud), and software that supports data storage, data analysis, and data tool construction and maintenance. An example of the tasks for someone in this work role would be to set up and maintain a multi-GPU server and install an operating system, which could then be used by people from the other work roles.

Another major data work role is **Product/Project Management**, where a *Product Manager* drives the development of a data product, serving as the one who develops the data product vision, defines data product development strategy, and produces a data product road map to deliver value for product users. A *Project Manager*, however, typically manages schedules and resources to get things done as it relates to data product development, but he or she often has little input into the business objectives and goals. Finally, based upon the recognition that nearly every domain will use data and/or data-driven tools, we have also included **Domain Experts** as another key work role in the data education landscape.

3.5 The Whole Framework

Figure 1 shows the full framework as a single unifying visualization of the data education landscape.

The columns represent core competency categories (detailed in Section 3.1) needed for working with data. While the ordering across columns may make sense as one possible view of the data life cycle, common practices are more agile in their use; they should **not** be considered as steps in a waterfall process.

The rows represent relevant disciplines and work roles. As mentioned earlier, the level of abstraction used to define the rows can vary depending on what you are trying to view. In our rendition, the top six non-header rows represent categories of relevant educational disciplines (see Section 3.3). The bottom six rows represent

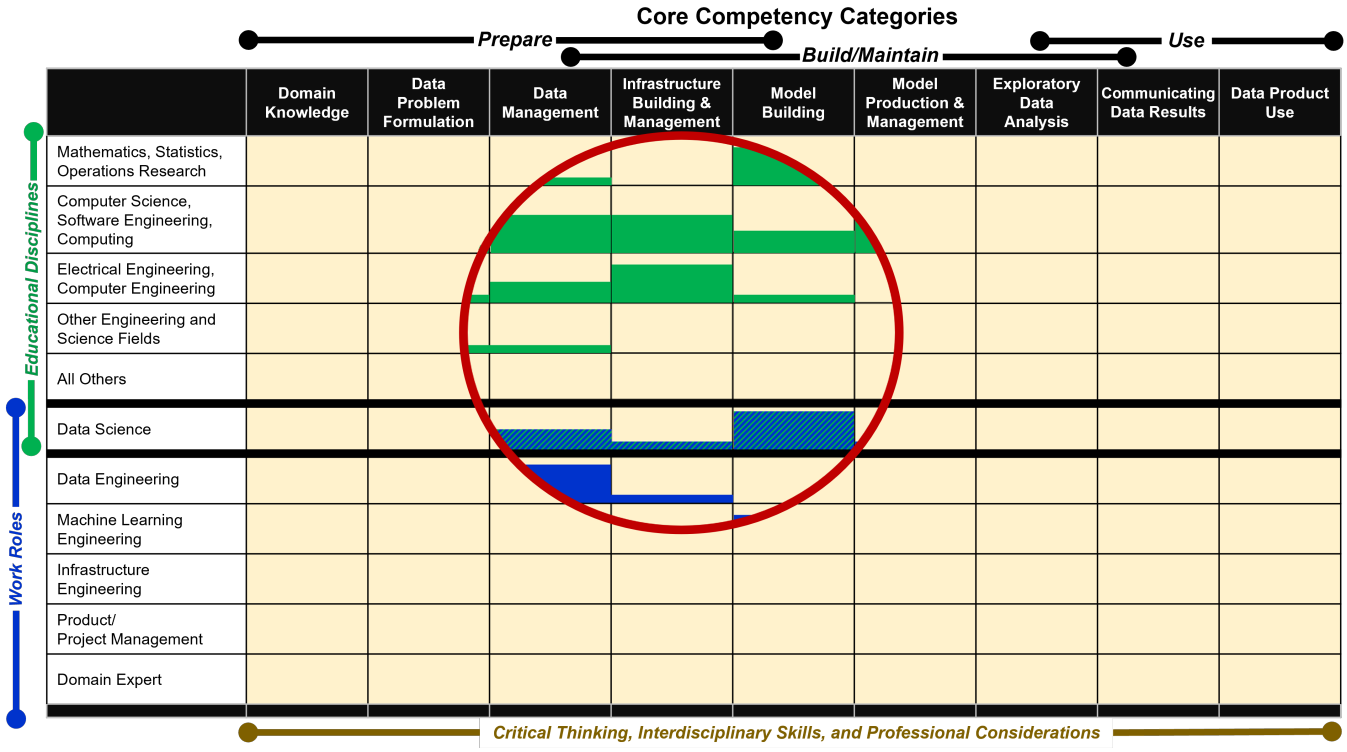

Figure 1: A Comprehensive Framework for Data Education. *Columns are core competency categories covering the landscape of data activities (no ordering is implied). Rows are disciplines/work roles; Data Science is both a discipline and a work role. The entire framework is shaded in gold depicting higher-order KSAs that touch all competency categories. See Section 3 for more detail. The round cutout and fill levels within illustrate an application of the framework to visualize relative needs for educational emphasis, fully discussed in Section 4.1.*

high-level data work roles (see Section 3.4). Data Science is both a discipline and a work role. In a finer-grained version of this framework, educational rows might correspond to academic majors while work role rows might depict positions within a particular project.

What one chooses to put in the cells will vary depending on the application. In the education rows, you could, for example, list learning outcomes or courses while listing skill identifiers in the work role rows. Alternatively, as we show in the Figure 1 cutout and describe in the next section, you can fill the cells with a quantity of color to visually demonstrate the territory of the competency category that the educational discipline or work role typically covers. The entire framework is shaded in gold, depicting higher-order KSAs (Section 3.2) that cross all competency categories, disciplines, and work roles.

4 APPLYING THE FRAMEWORK

In this section, we discuss several possible ways to use the framework, such as visualizing the relationship between disciplines and work roles, curricular or workforce planning, and others.

4.1 Visualizing Educational Needs

Building on fundamental empirical findings from [10, 28] and our own experience, the cutout in Figure 1 adds fill to a portion of the framework. The amount of fill in a given cell loosely depicts the territory that discipline or work role contributes in that competency category–more fill indicates higher contribution levels. This expected contribution is based only on activities that relate to data, including areas such as collection, curation, exploration, modeling, usage, etc., and not on all activities in the given discipline or work role. In this high-level abstract view of data-related activities, we use the magnitude of territory to represent either depth (being an expert in a few activities) or breadth (being able to do a little across many activities) within the competency category.

Although we have developed this example for Core Competency Categories, Educational Disciplines, and Work Roles (and not just the cutout region), the point in this position paper is to simply illustrate use of the framework. Thus, we leave for future work the task of defending a methodology for assigning specific fill levels. In that spirit and to more fully develop the example, we invite consideration of a few cells within the Data Science row included in the cutout. Previous work in the computing competencies for Data Science has identified that a Data Science graduate needs mastery in skills of creating a machine learning model, including supervised, unsupervised, and mixed methods (i.e., Model Building) but while some programming aptitude is expected, a high level understanding is sufficient for Big Data Systems problems of scale, structures of operating systems, file systems, networks, and security issues [12]. As such, we have given a full fill to the core competency

of Model Building, an intermediate fill to the competency of Data Management, and a minimal fill to Infrastructure Building and Management.

4.2 Other Uses of the Framework

We envision other impactful uses of the framework within education, industry, and government, such as:

Visualizing evolving educational needs as the machine capabilities in human-machine teams increase. Every competency already involves, to some extent, humans using computing tools. Invariably, the extent of tool (machine) contributions to the activity will grow and change the nature of needed human competencies. Using multiple instances of the framework, one can visualize changes over time, which can guide how data education needs to change.

Organizing educational offerings. The interdisciplinary nature of data education causes course-related offerings at an institution to multiply across different departments as each discipline's faculty recognizes how their curriculum should evolve to accommodate an increasingly data-centric society. Using our framework to map data-focused courses from across an educational institution can help facilitate conversations to coalesce into fewer, more broadly impactful courses. The framework could also be used to view coverage across currently offered majors, and to identify gaps and/or significant overlap across the majors.

Planning extracurricular activities. This framework will facilitate an institution planning its outside-the-classroom developmental opportunities to complement and reinforce their already-present educational offerings. Extracurricular activities might include data science competitions (i.e., "Kaggle"), the pursuit of internships, and/or student research topics that deepen their desired expertise.

Enabling better communication with constituents. Building upon insights from previous work (e.g., [28]), our framework facilitates more accurate and specific interactions between and within Academia and constituents, like Industry and Government, on how various academic offerings prepare students to engage with data and how those disciplines relate to different work roles. We believe our framework also builds on and complements the CCDS2021 [12] and CC2020 reports, which sought to establish guidelines [14] to "gain insight on the expectations of computing baccalaureate-degree graduates for the next decade."

Planning the workforce. From an industry or government perspective, the Chief Data Officer (CDO) and other leaders could use the framework for workforce planning as part of a strategic initiative to improve organizational data maturity. Many organizations struggle with implementing data strategies. An important element in data maturity is human capital management which recruits the right data personnel and enables workforce development. This approach could also be used to view the composition of a faculty.

5 THE ROAD AHEAD

This paper offers a comprehensive framework for data education, allowing all constituents to see the whole data education landscape. Our framework takes into account data-related academic disciplines as well as the full spectrum of data core competency categories. The framework also integrates data work roles as they exist today so that one can see contribution-level relationships with core competencies and educational disciplines. The framework is unique in that it incorporates a broad range of perspectives associated with data education – yet much work is still needed.

Approaches to data education bear greater consideration. Our framework is meant to inspire more holistic thinking about data education and associated curriculum development. As partially illustrated in Figure 1, we believe that no educational discipline or work role sufficiently covers the entire spectrum of the core competencies. Rather, considering breadth and depth in the core competencies makes it unlikely that any degree program will have sufficient student time to teach it all. Data education is inherently both interdisciplinary and multi-disciplinary, which suggests that a multi-departmental approach to data education could allow an academic institution to economize resources on courses and yet still deliver offerings that cover most of the data education landscape.

This paper is also a call for the computer science educational community to reflect on and debate what role(s) it should fill in the bigger picture.

Further work is needed to potentially extend the framework and demonstrate its use. Additional rows need to be added to include a wider variety of educational disciplines (e.g., bioinformatics, economics) or specialty work roles (e.g. data privacy advocates or legal personnel). The example application that applies relative fill levels to combinations of competencies, education disciplines/work roles needs a rigorous methodology behind it, informed by various stakeholder perspectives. The framework can also be used in a completely different way to highlight where the potential for AI tools and human-machine teaming has the highest probability of compensating for weaker areas on a team.

We believe that integrating data education into curricula will follow a similar path to that taken when computing and cyberspace were integrated into education. In those cases, every discipline eventually adopted computing and cyberspace competencies as critical elements. Similarly, the accessibility of data and tools will become an integral part of every academic discipline, but this is not (yet) facilitated in all cases. We should expect a future in which nearly every individual will use data-enabled technologies that require elements of data literacy. We postulate that non-data disciplines will work their way into data disciplines. Adopting a more holistic view of data education will be critical to producing the data-ready workforce of the future that can successfully and ethically use data to improve the outcomes of society as a whole.

ACKNOWLEDGMENTS

The authors thank the reviewers and our colleagues at the U.S. Military Academy for their insightful comments and recommendations that helped make this a better paper.

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, the Department of the Army, the Department of Defense, or the U.S. Government.

REFERENCES

[1] ABET, Inc. 2023. Program Criteria for Data Science, Data Analytics and Similarly Named Computing Programs. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-computing-programs-2023-2024/

[2] ABET, Inc. 2023. Program Criteria for Data Science, Data Analytics, and Similarly Named Programs. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-applied-and-natural-science-programs-2023-2024/

[3] Sridhar Alla and Suman Kalyan Adari. 2021. What is mlops? In *Beginning MLOps with MLFlow*. Apress, Berkeley, CA, 79–124.

[4] David Bamman. 2018. Big Data Meets Literary Analysis: Digital Humanities Research at the I School. https://www.ischool.berkeley.edu/news/2018/big-data-meets-literary-analysis-digital-humanities-research-i-school

[5] Nathaniel Bastian. 2020. Building the Army's Artificial Intelligence Workforce. *The Cyber Defense Review* 5, 2 (2020), 59–63.

[6] Jean R.S. Blair, Andrew O. Hall, and Edward Sobiesk. 2019. Educating future multidisciplinary cybersecurity teams. *Computer* 52, 3 (2019), 58–66. https://doi.org/10.1109/MC.2018.2884190

[7] Jean R.S. Blair, Andrew O. Hall, and Edward Sobiesk. 2020. Holistic cyber education. In *Cyber Security Education: Principles and Policies*, Greg Austin (Ed.). Routledge, New York, Chapter 10, 160–172.

[8] Jean R. S. Blair, Lawrence Jones, Paul Leidig, Scott Murray, Rajendra K. Raj, and Carol J. Romanowski. 2021. Establishing ABET Accreditation Criteria for Data Science. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 535–540. https://doi.org/10.1145/3408877.3432445

[9] Andriy Burkov. 2020. *Machine learning engineering*. Vol. 1. True Positive Incorporated.

[10] Joint Artificial Intelligence Center. 2020. DoD AI Education Strategy. https://www.ai.mil/docs/2020_DoD_AI_Training_and_Education_Strategy_and_Infographic_10_27_20.pdf

[11] Iain J. Cruickshank. 2023. An AI-Ready Military Workforce. *Joint Forces Quarterly* 110 (2023). https://doi.org/10.13140/RG.2.2.28909.87526

[12] Andrea Danyluk, Paul Leidig, Andrew McGettrick, Lillian Cassel, Maureen Doyle, Christian Servin, Karl Schmitt, and Andreas Stefik. 2021. *Computing Competencies for Undergraduate Data Science Curricula*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3453538

[13] Stefania Druga, Nancy Otero, and Amy J. Ko. 2022. The Landscape of Teaching Resources for AI Education. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1* (Dublin, Ireland) *(ITiCSE '22)*. Association for Computing Machinery, New York, NY, USA, 96–102. https://doi.org/10.1145/3502718.3524782

[14] CC2020 Task Force. 2020. *Computing Curricula 2020: Paradigms for Global Computing Education*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3467967

[15] Mark Frank, Johanna Walker, Judie Attard, and Alan Tygel. 2016. Data Literacy-What is it and how can we make it happen? *The Journal of Community Informatics* 12, 3 (2016), 4–8.

[16] Robert Gould. 2017. Data literacy is statistical literacy. *Statistics Education Research Journal* 16, 1 (2017), 22–25.

[17] James Grossman. 2012. "Big Data": an Opportunity for Historians? https://www.historians.org/research-and-publications/perspectives-on-history/march-2012/big-data-an-opportunity-for-historians

[18] Edith S Gummer and Ellen B Mandinach. 2015. Building a conceptual framework for data literacy. *Teachers College Record* 117, 4 (2015), 1–22.

[19] Sayash Kapoor and Arvind Narayanan. 2022. Leakage and the Reproducibility Crisis in ML-based Science. https://doi.org/10.48550/ARXIV.2207.07048

[20] Michael A. Lones. 2021. How to avoid machine learning pitfalls: a guide for academic researchers. https://doi.org/10.48550/ARXIV.2108.02497

[21] Andrew W Moore, Martial Hebert, and Shane Shaneman. 2018. The AI stack: a blueprint for developing and deploying artificial intelligence. In *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*, Vol. 10635. SPIE, 45–54.

[22] Julius Murumba and Elyjoy Micheni. 2017. Big data analytics in higher education: a review. *The International Journal of Engineering and Science* 6, 06 (2017), 14–21.

[23] Andy Nguyen, Lesley Gardner, and Don Sheridan. 2020. Data analytics in higher education: An integrated view. *Journal of Information Systems Education* 31, 1 (2020), 61.

[24] Association of American Colleges and Universities. 2009. *Valid Assessment of Learning in Undergraduate Education (VALUE)*. https://www.aacu.org/initiatives/value

[25] O*NET OnLine. 2023. *Data Scientist Occupation 15-2051.00*. https://www.onetonline.org/link/summary/15-2051.00

[26] M. Tamer Özsu. 2023. Data science—a systematic treatment. *Commun. ACM* 66, 7 (2023), 106–116. https://doi.org/10.1145/3582491

[27] Alex "Sandy" Pentland. 2013. The data-driven society. *Scientific American* 309, 4 (2013), 78–83.

[28] Rajendra K. Raj, Allen Parrish, John Impagliazzo, Carol J. Romanowski, Sherif G. Aly, Casey C. Bennett, Karen C. Davis, Andrew McGettrick, Teresa Susana Mendes Pereira, and Lovisa Sundin. 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* (Aberdeen, Scotland Uk) *(ITiCSE-WGR '19)*. Association for Computing Machinery, New York, NY, USA, 73–87. https://doi.org/10.1145/3344429.3372503

[29] Chantel Ridsdale, James Rothwell, Michael Smit, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, and Bradley Wuetherick. 2015. *Strategies and best practices for data literacy education: Knowledge synthesis report*. Technical Report. Dalhousie University. https://dalspace.library.dal.ca/bitstream/handle/10222/64578/StrategiesandBestPracticesforDataLiteracyEducation.pdf

[30] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3 (2021), 199 – 217. https://doi.org/10.1038/s42256-021-00307-0

[31] Robert Rose. 2016. Defining analytics: a conceptual framework: analytics' rapid emergence a decade ago created a great deal of corporate interest, as well as confusion regarding its meaning. *Or/MS Today* 43, 3 (2016), 34–39.

[32] John Godfrey Saxe. 2023. The Blind Man and the Elephant. https://www.commonlit.org/en/texts/the-blind-men-and-the-elephant

[33] Edward Sobiesk, Jean Blair, Gregory Conti, Michael Lanham, and Howard Taylor. 2015. Cyber education: a multi-level, multi-discipline approach. In *Proceedings of the 16th Annual Conference on Information Technology Education (SIGITE15)* (Chicago Illinois). ACM digital library, https://dl.acm.org/, 43–47.

[34] "Open Data Watch". 2022. The Data Value Chain: Moving from Production to Impact. https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/

[35] Annika Wolff, Daniel Gooch, Jose J Cavero Montaner, Umar Rashid, and Gerd Kortuem. 2016. Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics* 12, 3 (2016), 9–26. https://doi.org/10.15353/joci.v12i3.3275

[36] Wensheng Wu. 2022. Data Science Course Projects with Peer Challenges: An Experience Report. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1* (Dublin, Ireland) *(ITiCSE '22)*. Association for Computing Machinery, New York, NY, USA, 89–95. https://doi.org/10.1145/3502718.3524743

[37] Ting Xiao, Ronald I. Greenberg, and Mark V. Albert. 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* (Virtual Event, Germany) *(ITiCSE '21)*. Association for Computing Machinery, New York, NY, USA, 290–295. https://doi.org/10.1145/3430665.3456364