Foundations of Data Science
doi:10.3934/fods.2021031

EVALUATION OF EDISON'S DATA SCIENCE COMPETENCY FRAMEWORK THROUGH A COMPARATIVE LITERATURE ANALYSIS

Karl R. B. Schmitt*

Trinity Christian College
6601 West College Dr.
Palos Heights, Illinois 60463, USA

Linda Clark

Brown University
164 Angell St.
Providence, RI 02912, USA

Katherine M. Kinnaird

Smith College
44 College Lane
Northampton, MA 01063, USA

Ruth E. H. Wertz

Valparaiso University
1900 Chapel Dr.
Valparaiso, IN 46383-6493, USA

Björn Sandstede

Brown University
182 George St., Box F
Providence, RI 02912, USA

Abstract. During the emergence of Data Science as a distinct discipline, discussions of what exactly constitutes Data Science have been a source of contention, with no clear resolution. These disagreements have been exacerbated by the lack of a clear single disciplinary 'parent.' Many early efforts at defining curricula and courses exist, with the EDISON Project's Data Science Framework (EDISON-DSF) from the European Union being the most complete. The EDISON-DSF includes both a Data Science Body of Knowledge (DS-BoK) and Competency Framework (CF-DS). This paper takes a critical look at how EDISON's CF-DS compares to recent work and other published curricular or course materials. We identify areas of strong agreement and disagreement with the framework. Results from the literature analysis provide strong insights into

2020 *Mathematics Subject Classification.* Primary: 97K99, 97B10; Secondary: 97B70, 68T09.
Key words and phrases. Data Science, Data Science Education, Curricular Recommendations, Data Science Body of Knowledge, EDISON.
The authors were supported by NSF grants # 1839257, # 1839259, # 1839270, and by the Luce Foundation under the Clare Boothe Luce Program.
* Corresponding author: Karl R. B. Schmitt.

what topics the broader community see as belonging in (or not in) Data Science, both at curricular and course levels. This analysis can provide important guidance for groups working to formalize the discipline and any college or university looking to build their own undergraduate Data Science degree or programs.

1. **Introduction.** The rapid emergence and growth of Data Science[1] as its own discipline has been directly fueled by increased societal demands for combinations of advanced quantitative skills across many domains [8, 30, 34]. Given the interdisciplinarity of Data Science, and its co-emergence from many established disciplines, the absence of a single authoritative voice has presented challenges to defining core concepts and skills that are essential to professional competence and success. What is more, the scope of what is included in Data Science industry applications is evolving at a pace that is difficult for the academy and its institutions to match. The result is a new discipline that many are eager to claim and/or join, but that lacks a clear identity to serve as the central basis of educational curricula, research, and development. Identifying the core concepts of a well-defined Data Science discipline is further complicated by the highly varied pathways by which current industry leaders have acquired their skill sets and/or job titles. The practical skills that are now being included in Data Science by industry have long been essential in a diverse range of professions. Current "data scientist" job postings call for a mix of skills that have often been taught for decades, though spread out across several domains. What may be striking to realize is that few of the current mid-to-senior level "Data Scientist" ever received formal training in a full range of Data Science skills under an established Data Science degree or program, given that these degrees did not formally exist before 2012 [4, 5]. Rather, they took on the nomenclature of "data scientist" as a new title for the mix of skills they had acquired through their work and professional development, while their original degrees were from parental disciplines like computer science, electrical engineering, mathematics, physics, or statistics.

While rapid growth has presented challenges, broad expansion of Data Science job descriptions, degree programs, and published curricula also present an opportunity. It is possible to assess emergent patterns of what is considered to be central to the discipline from multiple stakeholders and perspectives [9, 15, 29, 31]. These patterns can provide both a framework and a common lexicon that guide the prioritization and development of educational curricula and discipline-based education research within Data Science. To better understand the scope and importance of this problem, we look to three primary levers of influence in the formation and identity of disciplines. These levers include professional societies, institutions of higher education, and national/international organizations that operate outside of the boundaries of any single institution.

1.1. **Professional societies.** Most academic disciplines include mechanisms by which a standardized core of knowledge has been established. For some, there are professional societies that produce curriculum guidelines for degree programs such as [3] for statistics and [38] for mathematics, while other disciplines have defined

[1]The authors recognize that there is difference between 'Data Science' and 'Data Analytics', where analytics typically involves less machine learning and artificial intelligence topics/skills, but these distinctions are not yet clear or broadly recognized. For the purpose of this article we use the term Data Science to be inclusive of Data Analytics for the sake of parsimony, while still recognizing that these terms are not interchangeable.

formal bodies of knowledge (BoKs), for example Project Management [28], Civil Engineering [12], and Finance/Accounting [2, 25]. Some guidelines and BoKs, such as the ACM/IEEE Computing Curricula 2020 report [10], are even collaboratively developed between multiple professional societies of closely related fields that have broad areas of cross-disciplinary overlap.

Curricular products generated by professional societies provide an important scaffolding for educational priorities that are grounded in professional practice. While BoKs fall into this category, it is worth noting that a formal BoK is distinct from curricular recommendations. A BoK can be thought of as a map that defines the breadth and depth of knowledge, skills, and attitudes recognized as belonging in the discipline. Curricular recommendations can then specify from the discipline's BoK how much, and in what combination, topics should be included for a given course, degree track, or program. For example, in the ACM/IEEE curricular recommendations for Computer Science (CS-2013), the guidelines specify the number of instruction hours student should receive in/on various knowledge areas [26].

In the case of Data Science, most of the current curricular guidance for Data Science has come from existing societies that have disciplinary overlap with Data Science, mainly those in Business, Computer Science, and Statistics. While there are a few professional societies that focus solely on the discipline, e.g. the Data Science Association [14], they have minimal membership overlap with academia. The professional societies who have contributed to curricular guidance often approach Data Science from the unique, and sometimes narrow, perspectives and identities of their respective disciplines. This has created a quagmire of curricular recommendations, with no obvious authoritative voice to help design a path forward. The discipline's disparate societal voices are contrasted by the several smaller, more complete efforts at defining full degree curricula. Examples of these more constrained effort are the Workshop on Data Science Education Report [9], the South Big Data Hub's *Keeping Data Science Broad* series [31] and the Park City Report [15].

1.2. **Institutions of higher education.** Institutions of higher education have been racing to meet the growing demand for data professionals. Hence the focus has been on repackaging existing courses and developing new courses that equip students with the necessary Data Science skills. Combined, these repackaged and new courses have been grouped into a sharply increasing set of new Data Science programs at the undergraduate and graduate levels.

Collectively, these degree programs and courses have produced a source of peer data that can be examined to identify variations in scope, and areas of consensus. Two early examples of this approach are Wu [36] and Hardin et al. [23], who identified program-level competencies and compared early introductory Data Science courses, respectively. To illustrate the growth in this area, we note that the 2017 Wu paper states the existence of approximately 100 programs, while the list of Data Science Colleges and Universities contains over 600 programs as of January 1st, 2021. The latter list provides an excellent source for accessing peer data and performing fine-grained comparisons. Yet that information may be incomplete due to self-selection bias, difficult to organize and mine due to varied formats, or have search capabilities that do not customize well to specific needs. We are aware of at least two projects currently in development that use natural language processing and text mining to examine these sources, but neither project has been completed.

1.3. **National and international organizations.** Other bodies with significant influence in defining the identity of Data Science include committees, task forces, and funding agencies that operate on broad national and international levels. Project reports and agendas at this scale tend to have lasting impact by driving research agendas and disciplinary knowledge development. For example, the EU has sponsored two such projects, the EDISON Project (www.edison-project.eu) [19] and the European Data Science Academy (EDSA, www.edsa-project.eu) [18], that significantly contributed to defining the field [16, 17, 27]. We can see evidence of their influence in the citations of the ACM Curricular guide to the EDISON report and the subsequent data programs that have been created throughout Europe. Similarly, in the US, the National Science Foundation has had multiple calls for Data Science Educational development (e.g. NSF 18-542 and NSF 19-518/21-523).

Three other recent and current efforts are particularly noteworthy. First, in 2018 the U.S. National Academy of Sciences, Engineering, and Medicine [22] released broad curriculum recommendations and maintains an ongoing working group exploring the topic. Second, during the spring of 2021 the Association of Computing Machinery (ACM) Education Board approved the final report from their Data Science Task Force [1]. Third, in the summer and fall of 2021 two Data Science programs will undergo a preliminary accreditation process through ABET via the Computing Accreditation Commission (CAC). Many of these groups are working on producing recommendations and requirements. Others are simply looking for guidance on implementing their own programs or supporting the broad development of the academic ecosystem for teaching Data Science. These organizations include the Academic Data Science Alliance [35] and individual colleges and universities.

1.4. **Problem statement.** Partly despite, and partly because of the rapid and distributed expansion of Data Science, the discipline still lacks cohesion. This lack encompasses everything from defining the basic scope of the field to detailed lists of foundational skills and knowledge that Data Science practitioners should have. Such cohesion is important to effectively transition discipline-based educational research in Data Science from experience reports and case-studies to formalized observational or experimental educational research. The value of such a transition is evident in the many other disciplines that made this shift in the past[32]. A fully 'comprehensive' research design to develop such a cohesive framework in a timely manner would be near impossible given the rate of change in practice here. Instead, we evaluated the most comprehensive body of work available at the time of our research, namely the EDISON Data Science Framework (EDSF)[2], as a potential foundation for educational research in Data Science within the broader international community.

Our premise is that, without a clear authoritative guide, community consensus among multiple stakeholders can provide this evaluation. To evaluate the EDSF as a foundation for educational curricula and research we investigated patterns

[2]Throughout this paper we reference three distinct products from the EDISON Project, their 'Data Science Framework' or EDSF, their 'Competency Framework for Data Science' or CF-DS, and their 'Data Science Body of Knowledge' or DS-BoK. The most recent documentation for these products can be found at https://github.com/EDISONcommunity/EDSF, while a more public-friendly version can be found at https://edison-project.eu/. Because the sub-documents are difficult to reference as they do not have associated meta-data, the formal citations used throughout the paper are simply to the project homepage. A journal article overview from 2016 is also available, see [16].

of consensus among key sources of disciplinary influence through a review of the literature. Guiding our investigations are two questions:

1. Which of the topics identified in the EDISON project have persisted since its publication?
2. Which of the topics identified in the EDISON project are included in a first Data Science course?

We have chosen to include the second research question for two reasons. First, within discipline-based educational research the 'first' course in a discipline is studied far more often than advanced topics, as it generally enrolls the most students and often serves as non-majors' first or only engagement with the field. Second, a first course in a discipline often focuses on skill sets that define the identity of the discipline.

To answer our study questions, we perform a comparison of relevant public documents, including formal reports, peer-reviewed publications, and openly disseminated course materials.

2. Motivation and background.

As part of an EU-funded effort to accelerate the creation of the Data Science profession, the EDSF project produced a comprehensive competency framework.[3] The framework included both a defined body of knowledge (DS-BoK) as well as a Competency Framework for Data Science (CF-DS). In generating this framework, the EDISON project team went through a lengthy data collection process, multiple refinements, and developed a complete model curriculum for a degree [19]. The CF-DS was developed around the following five major knowledge area groups:

- Data Analytics (DSDA)
- Data Science Engineering (DSENG)
- Data Management (DSDM)
- Research Methods and Project Management (DSRM)
- Domain Related Competencies and Business Analytics Competencies (DSBA).

These knowledge areas are defined by the DS-BoK as the core areas required for developing Data Science curricula, which in turn corresponds to the CF-DS which defines the explicit skills and knowledge that exemplify competence in these areas. While a BoK is ideal for directing nuanced educational research, most published literature outside of educational research (such as curricular guides, case study reports, etc) usually refers to student learning outcomes or competencies rather than knowledge [17]. Within the CF-DS the five major competency groups were further expanded, based on a job-market analysis, to include specific skills and knowledge topics required to support the competencies. We use the term EDISON Data Science Framework (EDSF) when referring to the overall EDISON project, which includes the both the CF-DS and DS-BoK among other components that extend beyond identification of core competencies. For the purpose of our analysis, we pulled competency items from the CF-DS. We refer to Tables 4.2 and 4.4 of their Competency Framework documentation [19] for the list of skills as "the EDISON Core Data Science Skills Table" and the knowledge units as "the EDISON Knowledge Table."

[3] Additional background information on the EDISON Data Science Framework (EDFS), including its scope and development process, is succinctly summarized in "EDISON Data Science Framework: Building the Data Science Profession" presented by Marian Bubak at the SKG 2016 Conference in Beijing, China, September 15-17, 2016. The slideshare of this presentation can be accessed at https://slideplayer.com/slide/11823347/.

The EDISON project [19] was selected as the basis of our analysis. It was chosen because, to our knowledge, it represents the first published set of materials which attempt to fully articulate what knowledge and skills belong in the field of Data Science. We considered other frameworks, but found these to be less comprehensive in the case of ACM [1],[4] while in the case of the Initiative for Analytics and Data Science (IADSS) the body of knowledge was still a work-in-progress [20].

2.1. **Preprocessing EDISON Data Science Framework.** Rather than using the full Data Science Body of Knowledge (DS-BoK), our study focuses a modified version of EDISON's Competency Framework for Data Science (CF-DS). This decision was motivated by:

1. Concerns that practitioners or specialized instructors would be unfamiliar with, or use alternative terms for, many of the more fine-grained technical BoK topics.
2. Recognition that, with the exception of the ACM report, the literature sources present knowledge topics at a significantly more abstract level than the detailed level of a BoK.
3. A desire to collect information on first-course topics while recognizing that many technical BoK topics would never occur in a first course.

The EDISON CF-DS distinguishes between Core Skills (in the EDISON Core Data Science Skills Table) and Knowledge (in the EDISON Knowledge Table) [19]. Using the complete list of both skills and knowledge would have introduced a significant amount of duplication, as the originals were not intended to be directly combined. On the other hand, many elements were unique to one of the two tables. We therefore generated a list of unique knowledge and skills from the EDISON CF-DS relying on our disciplinary knowledge. We examined both tables for any knowledge areas or skills that appeared as near-duplicates, which were then were combined into one item to create our final list of competencies. An example is shown in Figure 1. A breakdown of the number of topics and duplicates in these two tables and the count of unique topics in our final list are shown in Table 1. The resulting list of competencies is the content we wish to explore and compare against existing literature. The full list of competencies used is provided in the Appendix A.

3. **Study methodology and analysis.** We conduct a comparison of the EDISON CF-DS to other bodies of curricular literature. In this section, we interpret our overarching research questions from Section 1.4 as: *To what extent do scholars and researchers perceive, as encoded in published, peer-reviewed documents, the skills and knowledge identified by the EDISON Project as...*

- *Central to Data Science?*
- *Appropriate for a first course in Data Science?*

We consider two distinct corpora: (1) curricular level competencies, and (2) introductory course topics. Both corpora encode the opinions of instructors who are shaping Data Science as an academic discipline. We view these corpora as proxies for the collective opinions of departments and programs. We note the limitation that these sources are limited to those developed largely in European and English

[4]When this analysis was conducted, only Draft 2 of the ACM report was available. The preliminary analysis was shared with the ACM task-force at that time and may have influenced their finalization and approval of Draft 3.

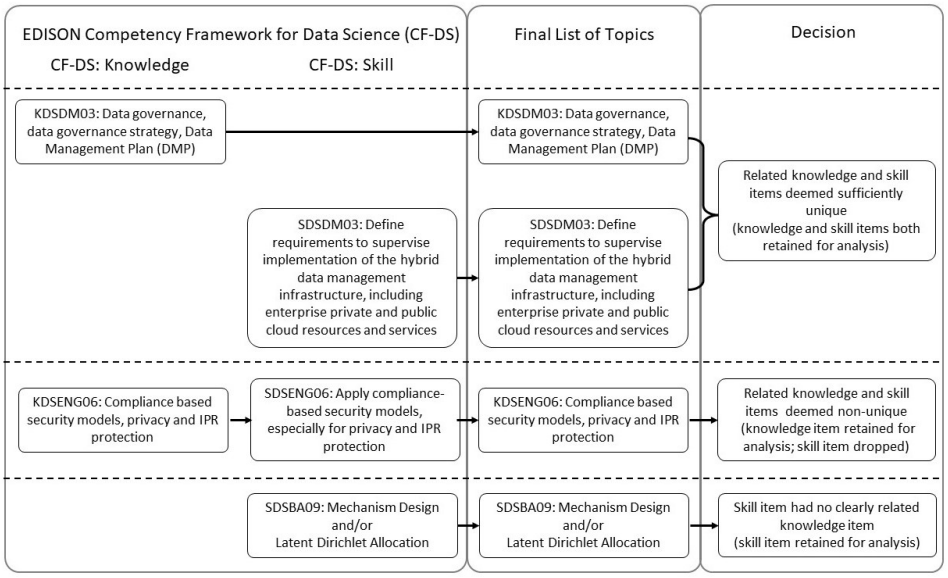

FIGURE 1. Examples of the reduction process and resulting item counts for merging items from EDISON Core Data Science Skills Table & EDISON Knowledge Table from [19] into the *List of Topics* that were investigated in our combined studies.

TABLE 1. Unique Item Counts by Knowledge/Skill in each competency category. The merged row indicates how often a Knowledge and Skill in that competency are merged. There are five acronyms in this table: DSDA stands for Data Science Analytics, DSENG stands for Data Science Engineering, DSDM stands for Data Management, DSRM stands for Research Methods and Project Management, and finally DSBA stands for Domain Related Competencies and Business Analytics Competencies

	DSDA	DSENG	DSDM	DSRM	DSBA	Total
Skill	6	4	6	5	2	23
Knowledge	7	2	6	5	1	21
Merged	8	8	3	1	7	27
Total	21	14	15	11	10	71

speaking countries. There may be inherent biases in these literature sources that a multilingual and broader cultural inclusion criteria may illuminate.

3.1. **Curricular level competencies.** To identify potential data sources for this comparison, we sought documents that provided a high-level list of competencies that undergraduate Data Science students should learn. Our first two sources were the incomplete BoK's from ACM [1] and IDASS [20]. In both cases competencies directly comparable to EDISON's CF-DS are explicitly listed. In addition to these documents, the Business Higher Education Forum (BHEF) has published a Data

Science and Analytics Competency Map [7]. This map is a list of Data Science concepts and principles tiered into when and where these concepts are learned, such as in college or in a work experience.

A much less structured source of competencies are the curricular contents of the many Data Science programs and majors that have been formed [13]. Typically these have followed curricular recommendations like those from De Veaux et al. [15], the National Research Council [13], Anderson et al. [4, 5], Blitzstein [6], and Hicks and Irizarry [24]. These recommendations typically provide broad statements of the content or courses that should be included while usually glossing over the details such as specific sub-classes of algorithms. For our analysis we will use the most commonly referenced recommendations known as "the Park City Report" [15].

The existence of a large number of successful programs also allows us to adopt a reverse approach. Rather than stating what should be in a curriculum, the existing programs can be analyzed to identify what is common across them. Wu [36] conducted such a systematic study of existing Data Science and Analytics programs, where he reported out a list of competencies that all graduates of those degrees are expected to have. For our purposes, these results will be treated as a set of curricular level 'recommendations' for what should be included in a program.

3.2. **Introductory course topics.** For this comparison, we considered a variety of published materials about first courses in Data Science. Rather than examining individual syllabi, we identify five well-cited papers whose models have been regularly adopted at other institutions. Our curriculum comparisons against the EDISON CF-DS included the courses in Hardin et al. [23] (except for the 36-week 'Data Science Specialization' program) and four more recent courses:

- Data-Science-in-a-Box by Dr. Cetinkaya-Rundel [11]
- Foundations in Data Science (Data 8), originating at U.C. Berkeley [33]
- DSC101 at Univ. Massachusetts, Dartmouth [37]
- Foundations of Data Science & Foundations of Big Data by the European Data Science Academy (EDSA) [18]

3.3. **Corpora analysis methods.** We treated the topics listed in the EDISON CF-DS as 'ground truth.' For each topic in this competency framework, we cross checked which of the documents in Section 3.1 and courses from Section 3.2 included, or not included, it. To avoid undercounting, we erred in favor of inclusion; that is, if it could be assumed that the document examined included at least partial coverage of the EDISON topic, it was marked as included. We then quantified the level of agreement in the resulting tables as follows.

3.3.1. *Measure 1: Percent coverage.* The first measure we used was a percent coverage of topics. This provides a simple assessment of how similar an individual document source is to the EDISON CF-DS. Given our approach, a score of 100% indicates that all of the topics in the EDISON CF-DS appear in the source. It does *not* tell us what or if any additional topics not already included in the EDISON CF-DS occur in the source. An important caveat with this measure is to recognize that it is not appropriate to combine these values into an average coverage score, though it is reasonable to compare them across different sources.

3.3.2. *Measure 2: Krippendorff-Alpha.* To assess overall averaged agreement across all courses, we compute a Krippendorff-Alpha (k_α) score using the online RelCal site [21]. Typically, k_α is used to measure the agreement between multiple 'coders' or

qualitative ratings of items. In this work, the 'coders' are the source documents that indicate whether a topic in the CF-DS belongs in Data Science or not, according to their perspective (as stated in the source document). The k_α is a similarity measure. A $k_\alpha = 0$ indicates systematic disagreement, while a $k_\alpha = 1$ indicates perfect agreement of either inclusion or exclusion.

While the k_α score takes into account any amount of partial positive or negative agreement, it does not provide more insight into which items caused a decrease in the score. To assist our identification of these points of disagreement we generated two derivative encodings, defining two types of 'agreement.' For topics included in at least four of the five curricular sources (noted in Section 3.1), we say that it has 'positive agreement' that the topic should persist (research question 1) or be included (research question 2) within a Data Science curriculum or course, respectively. Similarly any topic excluded in at least four of the curricular sources is considered to have 'negative agreement', meaning that a topic should *neither* persist nor be included within a Data Science curriculum or course, respectively. Topics that have neither kind of agreements are labeled as indeterminate. This derivative encoding provides more insight into specific k_α scores.

4. Corpora analysis results.

In the results of our corpora analysis, we provide the number of unique topics from the EDISON CF-DS that are covered by each document. We investigate the amount of inter-document agreement within each level. What is interesting (though not surprising) is that there is less consensus around what should be included in a Data Science curriculum as a whole (Section 4.1) than what should not be included in a first course about Data Science (Section 4.2).

4.1. Comparison of curricular-level competencies to the CF-DS.

Overall, in Table 2 we can see that the coverage by the current ACM guidelines is best aligned with the EDISON competencies. This was to be expected, as they referenced the EDISON framework and the other documents as background work [7, 15, 19]. Furthermore, we can immediately see that current curriculum implementations (from Wu [36]) cover more topics (61% vs. 44%) than originally specified in the Park City Report [15] but less than the potential specifications coming from ACM (61% vs. 82%).

TABLE 2. Counts & Coverage Percentage of Topics from EDISON DSF for each Curricular-Level data source

	Park City	IADSS	Wu	BHEF	ACM
Count of Topics	31	43	43	45	58
% Coverage of EDISON CF-DS	44%	61%	61%	63%	82%

When considering all five curricular sources, we find $k_\alpha = 0.288$ indicating a relatively low level of agreement among the sources. Since the ACM curriculum had a dramatically higher coverage of EDISON, we were curious if the other four documents were in more agreement with each other. Therefore we also computed k_α excluding the ACM mapping. Here we find a slightly higher agreement of $k_\alpha = 0.316$. However, this value still indicates a high level of disagreement on what should or should not be covered in the curriculum. To break this down further, we provide in Table 3 a count of positive or negative agreement, and indeterminate topics based

TABLE 3. Summary of topic agreement from literature analysis of EDISON CF-DS items that should persist (or not) in Data Science.

	Agreement		Indeterminate	Total
	Positive	Negative		Consensus
Count	34	14	23	48
Percent	48%	20%	32%	68%

on our definitions in the methods section. We also consolidate positive and negative agreement to indicate overall consensus in the community.

Thus based on the literature analyzed here, we observe that the community has reached consensus on 48 curricular topics, including the persistence of 33 topics from the EDISON CF-DS and the removal of 14 topics. The curricular literature did not reach a consensus on the remaining 23 topics, where the dissenting sources seem to mostly follow disciplinary lines. There are several examples of singular dissenting opinions. For instance, the Park City report and the IADSS framework had one topic which they each explicitly included that the other sources did not. Park City included competencies on 'simulation' and the IADSS included 'operations research.' The Park City report also included a significant description of various mathematical topics. However, the EDISON CF-DS makes a distinction between optimization, which the Park City report explicitly mentions, and operations research, which it did not explicitly mention. Both the ACM and BHEF frameworks had multiple items that were uniquely included, largely reflecting their disciplinary origins in Computing and Business respectively. Not surprisingly, Wu [36] based on realized programs had no unique inclusions.

4.2. **Comparison of introductory course literature.** Conducting our analysis for agreement on the first-course corpus produced the results found in Appendix C and summarized in Table 4. Examining the overall amount of coverage in each course produces a fairly consistent value of slightly over 30%. The exception to this level of coverage is the course from the University of Auckland. However, Hardin et al. [23] explicitly call out that course as being late in a student's experience, rather than a 'first course' like the others presented here. Additionally, even though the EDSA's foundational courses together achieve a slightly higher coverage, that occurs largely because we have chosen to include two courses rather than one. Since the EDSA curriculum was designed from the ground up to map to the EDISON CF-DS, they are able to more explicitly differentiate within their courses when specific topics were covered. This led each individual course to have a far lower coverage (not shown).

Overall, we found that there was fairly high agreement on what should be covered in a first course, with Krippendorff's Alpha-Reliability score being $k_\alpha = 0.57$. Excluding the Auckland course noticeably raises the agreement with a $k_\alpha = 0.68$. Some disagreement is to be expected since courses may adopt different approaches. For example, the course may be project driven, might be heavily statistical, or be heavily programming oriented. However, it seems that, regardless of the approach, many of the same fundamental topics do get covered (or not covered). To illustrate this in more detail, in Table 5, we observe a count of positive or negative agreement, and indeterminate topics based on our definitions in the methods section. We also use positive or negative agreement to indicate consensus from the community.

TABLE 4. Counts & Coverage Percentage of Topics from EDISON DSF for each Course-level data source

Source	Hardin et al. Smith	Auckland	UC B/D	St. Olaf	Purdue	Dusen et al. Data-8 (various schools)
Count of Topics	22	7	25	20	24	23
% Coverage	31%	10%	35%	28%	34%	32%

Source	Cetinkaya-Rundel Data-Science-Box (various schools)	Yan and Davis U.Massachusetts Dartmouth	European DSA Foundations of Data Science Big Data (2 courses)
Count of Topics	24	23	28
% Coverage	34%	32%	39%

TABLE 5. Summary of topic agreement from literature analysis of EDISON CF-DS items that should be included (or not) in an introductory in Data Science course.

	Agreement Positive	Agreement Negative	Indeterminate	Total Consensus
Count	15	41	15	56
Percent	21%	58%	21%	79%

It is clear that there is a central set of topics that are being taught in introductory Data Science courses. We find 15 topics (21%) regularly included and 41 topics (58%) regularly excluded. In total, that gives 56 topics (79%) which the community has strong agreement on and only 15 topics (21%) on which there is not a strong consensus. The details of the excluded list is provided in Appendix C, with the included topics listed below.

- Machine Learning (supervised)
- Machine Learning (unsupervised)
- Qualitative analytics
- Data preparation and pre-processing
- Performance and accuracy metrics
- Modeling and simulation, theory and systems
- Data Architecture, data types and formats, data modeling and design, including related technologies
- Data lifecycle and organizational workflow, data provenance and linked data
- Research methods, research cycle, hypothesis definition and testing
- Data lifecycle and data collection, data quality evaluation
- Use Machine Learning technology, algorithms, tools
- Use Data Mining techniques
- Apply analytics and statistics methods for data preparation and pre-processing
- Be able to use performance and accuracy metrics for data analytics assessment and validation
- Use effective visualization and storytelling methods to create dashboards and data analytics reports.

Within the 15 topics without a strong consensus, several stand out as instructional choices, including text data mining, cloud or big data systems, conducting

a full-fledged experiment process (design, collect data, and analyze), and database technologies. Given these results, it would be interesting to survey Data Science instructors and industry practitioners to see which of these topics students should be exposed to in their first, and often only Data Science course. We have one such survey in early stages.

5. Discussion and future work.

This work provides a preliminary evaluation of the Data Science Framework produced by the EDISON project [19]. The comparison with various curricular guidelines at both the discipline level and that of a first course leads us to conclude that the EDSF largely capture the current scope of the discipline of Data Science.

The most significant discrepancies are in the domain areas, specifically business applications. Originally Data Science, especially as a label, grew out of business analytics' need for a particular combination of skills. Now though, the world is seeing the value of those skills in far more than just business. Over the last few years, there has been explosive growth of Data Science into technical domains and other application areas beyond business. Therefore, it seems logical that a revision of what could be domain knowledge would be required.

In addition to the application of Data Science, these findings also impact curriculum development and teaching. For faculty who wish to 'teach' data science, these results point to the kinds of skills and knowledge that can be identified as Data Science, reducing the confusion as the definition of Data Science continues to evolve. For academic programs developing programmatic Data Science endeavors (at the undergraduate or graduate level), the identification of these skills, particularly those identified in the first course, allow faculty to selectively integrate common building blocks of data science into lower level introductory courses, building a planned deliberate pattern of courses into a mature curriculum. Finally, as higher education continues to emphasize assessment and accreditation, coming to a consensus on what a Data Science curriculum should consist of will facilitate the ability to evaluate programs.

Furthermore, from the corpora analysis review we identified a 'core' of topics that are typically included in an introductory Data Science course. As an example, an introductory course should include not only technical knowledge, such as machine learning, data types, and data mining among others, but also critical thought processes related to Data Science such as the data science workflow, research methods, and developing hypotheses. There remains significant variability in other topics though, notably topics such as 'Natural Language Processing', and 'Big Data.' One delimitation of this study was that our scope focused on comparisons among published literature that was generated through or by professional societies, scholars within higher education, and national/international organizations and consortia. What is not well represented within these groups, however, is the influence of industry corporations and not-for-profit learning platforms that also have significant influence on the development and identity of Data Science as a discipline. Future work that is currently in progress will address this gap with a more direct validation initiative by surveying both industry practitioners and academics involved in working as or teaching Data Science. Specifically, we present participants with a random subset of 10 of the CF-DS topics analyzed in the current study. Analogous to the corpora analysis, participants are asked to respond 'Agree' or 'Disagree' to

each presented topic as belonging 'within Data Science' and as 'appropriate for coverage during first course in Data Science.' Our preliminary results suggest a high correspondence with the corpora analysis results.

Combined, these efforts toward comparative evaluation of the EDSF will support important next steps in the development of Data Science as a discipline. Having a common consensus of the skills and knowledge in Data Science allows the intentional design and execution of educational research to assess how well courses are teaching Data Science topics. It allows the easier identification of courses being taught outside the umbrella of Data Science that provide broader coverage of necessary knowledge. Similarly, this commonality allows programs to articulate learning outcomes in a consistent language making it simpler for those hiring data scientists to understand their capabilities. For the areas in which there is disagreement, the community can focus on understanding for whom, or why, those topics might be important. Should they be specializations available within some degrees, should they be moved to advanced courses? Do these variations help to distinguish among important sub-disciplinary domains?

While it is exciting to see some consensus forming in the community, it is also clear that a significant amount of work is still needed to help Data Science coalesce into a cohesive and comprehensible discipline. The plethora of Venn diagrams defining what Data Science is have not gone away, but the EDISON Project, and this work, have more clearly mapped the interior boundaries of overlap.

Disclosure statement. The authors have no conflicts of interest to declare.

Acknowledgments. The authors gratefully acknowledge funding from the National Science Foundation through DMS # 1839257, 1839259, 1839270. Kinnaird is the Clare Boothe Luce Assistant Professor of Computer Science and Statistical and Data Science at Smith College and as such, is supported by Henry Luce Foundation's Clare Boothe Luce Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Luce Foundation or the National Science Foundation (NSF).

Appendix A. **Merging knowledge and skills of EDISON CF-DS.** Items where a knowledge and skill were merged are indicated by the "Merged" column in the table. Only one of the two source items is shown for brevity.

EDISON Competency Framework for Data Science			
Subdomain	Specific Label	Merged Skill	Brief Description
Data Science Data Analytics (DSDA)	KDSDA01		Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
	KDSDA02		Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
	KDSDA03		Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms
	KDSDA04		Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)
	KDSDA05		Text Data Mining: statistical methods, NLP, feature selection, apriori algorithm, etc.
	KDSDA06	SDSDA04	Predictive Analytics
	KDSDA07	SDSDA05	Prescriptive Analytics
	KDSDA08	SDSDA06	Graph Data Analytics: path analysis, connectivity analysis, community analysis, centrality analysis, etc.
	KDSDA09	SDSDA07	Qualitative analytics
	KDSDA10	SDSDA11	Natural language processing
	KDSDA11		Data preparation and pre-processing
	KDSDA12		Performance and accuracy metrics
	KDSDA13	SDSDA12	Operations Research
	KDSDA14	SDSDA13	Optimisation
	KDSDA15	SDSDA14	Simulation
	SDSDA01		Use Machine Learning technology, algorithms, tools (including
	SDSDA02		Use Data Mining techniques
	SDSDA03		Use Text Data Mining techniques
	SDSDA08		Apply analytics and statistics methods for data preparation and
	SDSDA09		Be able to use performance and accuracy metrics for data
	SDSDA10		Use effective visualiation and storytelling methods to create

EDISON Competency Framework for Data Science			
Subdomain	Specific Label	Merged Skill	Brief Description
Data Science Engineering (DSENG)	KDSENG01	SDSENG01	Systems Engineering and Software Engineering principles, methods and models, distributed systems design and organisation
	KDSENG02		Cloud Computing, cloud based services and cloud powered services design
	KDSENG03	SDSENG03	Big Data technologies for large datasets processing: batch, parallel, streaming systems, in particular cloud based
	KDSENG04	SDSENG04	Applications software requirements and design, agile development technologies, DevOps and continuous improvement
	KDSENG05'	SDSENG05'	Systems and data security, data access, including data
	KDSENG06	SDSENG06	Compliance based security models, privacy and IPR protection
	KDSENG07	SDSENG07	Relational, non- relational databases (SQL and NoSQL), Data
	KDSENG08	SDSENG08	Big Data infrastructures, high-performance networks,
	KDSENG09	SDSENG09	Modeling and simulation, theory and systems
	KDSENG10		Information systems, collaborative systems
	SDSENG02		Use Cloud Computing technologies and cloud powered services
	SDSENG10		Use and integrate with the organisational Information systems,
	SDSENG11		Design efficient algorithms for accessing and analysing large
	SDSENG12		Use of Recommender or Ranking system

EDISON Competency Framework for Data Science

Subdomain	Specific Label	Merged Skill	Brief Description
Data Management (DSDM)	KDSDM01		Data management and enterprise data infrastructure, private and
	KDSDM02	SDSDM02	Data storage systems, data archive services, digital libraries, and
	KDSDM03		Data governance, data governance strategy, Data Management
	KDSDM04		Data Architecture, data types and data formats, data modeling
	KDSDM05	SDSDM05	Data lifecycle and organisational workflow, data provenance and
	KDSDM06		Data curation and data quality, data integration and
	KDSDM07		Data protection, backup, privacy, IPR, ethics and responsible data
	KDSDM08	SDSDM08	Metadata, PID, data registries, data factories, standards and
	KDSDM09		Open Data, Open Science, research data archives/repositories,
	SDSDM01		Specify, develop and implement enterprise data management and
	SDSDM03		Define requirements to and supervise implementation of the
	SDSDM04		Develop and implement data architecture, data types and data
	SDSDM06		Consistently implement data curation and data quality controls,
	SDSDM07		Implement data protection, backup, privacy, mechanisms/
	SDSDM09		Adhere to the principles of the Open Data, Open Science, Open
Research Methods and Project Management (DSRM)	KDSRM01		Research methods, research cycle, hypothesis definition and
	KDSRM02		Experiment design, modelling and planning
	KDSRM03	SDSRM03	Data lifecycle and data collection, data quality evaluation
	KDSRM04		Use cases analysis: research infrastructure and projects
	KDSRM05		Research Data Management Plan (DMP) and data stewardship
	KDSRM06		Project management: scope, planning, assessment, quality and
	SDSRM01		Use research methods principles in developing data driven
	SDSRM02		Design experiment, develop and implement data collection
	SDSRM04		Apply structured approach to use cases analysis
	SDSRM05		Develop and implement Research Data Management Plan (DMP),
	SDSRM06		Consistently apply project management workflow: scope,
Domain Related Competencies and Business Analytics Competencies (DSBA)	KDSBA01	SDSBA01	Business Analytics (BA) and Business Intelligence (BI); methods
	KDSBA02	SDSBA02	Business Processes Management (BPM), general business
	KDSBA03	SDSBA03	Agile Data Driven methodologies, processes and enterprises
	KDSBA04	SDSBA04	Use Econometrics for data analysis and applications
	KDSBA05	SDSBA05	Data driven Customer Relations Management (CRP), User
	KDSBA06		Use cases analysis: business and industry
	KDSBA07	SDSBA07	Data Warehouses technologies, data integration and analytics
	KSDSBA08	SDSBA08	Use data driven marketing technologies
	SDSBA06		Apply structured approach to use cases analysis in business and
	SDSBA09		Mechanism Design and/or Latent Dirichlet Allocation
		Count	71
		Percent	100%

Appendix B. **Curriculum mapping.** Analysis of the merged knowledge and skill items at the curricular level. Summary values are provided at the end of the table.

EDISON Competency Framework for Data Science							Positive	Negative	Indeterminate
Specific Label	Park City	Wu	BHEF	IADSS	ACM	Percent Including	Agreement AGREE >= 80%	Agreement AGREE <= 20%	80% < A < 20%
KDSDA01	X	X	X	X	X	100%	X		
KDSDA02	X	X	X	X	X	100%	X		
KDSDA03				X	X	40%			X
KDSDA04		X		X	X	60%			X
KDSDA05				X	X	40%			X
KDSDA06		X	X	X	X	80%	X		
KDSDA07			X			20%		X	
KDSDA08			X		X	40%			X
KDSDA09	X	X	X		X	80%	X		
KDSDA10				X	X	40%			X
KDSDA11	X	X	X	X	X	100%	X		
KDSDA12	X	X	X	X	X	100%	X		
KDSDA13				X		20%		X	
KDSDA14	X			X	X	60%			X
KDSDA15	X					20%		X	
SDSDA01	X	X		X	X	80%	X		
SDSDA02	X			X	X	60%			X
SDSDA03				X	X	40%			X
SDSDA08	X	X	X	X	X	100%	X		
SDSDA09			X	X	X	60%			X
SDSDA10	X	X	X	X	X	100%	X		
KDSENG01		X	X	X	X	80%	X		
KDSENG02		X		X	X	60%			X
KDSENG03		X	X	X	X	80%	X		
KDSENG04		X		X	X	60%			X
KDSENG05'	X			X	X	60%			X
KDSENG06					X	20%		X	
KDSENG07	X	X	X	X	X	100%	X		
KDSENG08	X	X	X	X	X	100%	X		
KDSENG09	X			X		40%			X
KDSENG10			X	X	X	60%			X
SDSENG02	X	X	X	X	X	100%	X		
SDSENG10		X			X	40%			X
SDSENG11	X	X	X	X	X	100%	X		
SDSENG12					X	20%		X	

EDISON Competency Framework for Data Science							Positive	Negative	Indeterminate
Specific Label	Park City	Wu	BHEF	IADSS	ACM	Percent Including	Agreement AGREE >= 80%	Agreement AGREE <= 20%	80% < A < 20%
KDSDM01	X	X	X		X	100%	X		
KDSDM02		X	X	X	X	80%	X		
KDSDM03		X	X		X	60%			X
KDSDM04	X	X	X		X	100%	X		
KDSDM05	X	X	X		X	100%	X		
KDSDM06	X	X	X		X	100%	X		
KDSDM07	X	X	X		X	100%	X		
KDSDM08			X			20%		X	
KDSDM09			X			20%		X	
SDSDM01		X	X	X	X	80%	X		
SDSDM03		X		X	X	60%			X
SDSDM04	X	X	X	X	X	100%	X		
SDSDM06	X	X	X	X	X	100%	X		
SDSDM07	X	X	X		X	80%	X		
SDSDM09						0%		X	
KDSRM01	X	X	X	X		80%	X		
KDSRM02	X	X	X	X	X	100%	X		
KDSRM03	X	X	X	X	X	100%	X		
KDSRM04		X	X	X	X	80%	X		
KDSRM05		X	X		X	60%			X
KDSRM06		X		X	X	60%			X
SDSRM01		X	X	X	X	80%	X		
SDSRM02	X		X	X	X	80%	X		
SDSRM04						0%		X	
SDSRM05	X		X		X	60%			X
SDSRM06		X	X	X	X	80%	X		
KDSBA01			?			20%		X	
KDSBA02		X	X			40%			X
KDSBA03		X		X	X	60%			X
KDSBA04						0%		X	
KDSBA05			X		X	40%			X
KDSBA06	X	X	X		X	100%	X		
KDSBA07		X	X	X	X	80%	X		
KSDSBA08						0%		X	
SDSBA06					X	20%		X	
SDSBA09					X	20%		X	
Count	32	44	46	44	59		34	14	23
Percent	45%	62%	65%	62%	83%		48%	20%	32%

Appendix C. **Course mapping.** The same merged knowledge and skill items that were analyzed at the curricular level (depicted in Appendix A) were also analyzed at the course level. In this appendix, we present two tables. The first shows when topics are included in each of our studied courses. The second table aggregates the data from the first table.

EDISON Competency Framework for Data Science

Specific Label	Hardin et al.				Dusen et al. data8.org	Cetinkaya-Rundel	Yan and Davis	EDSA	Percent Agree
	Smith	UC B/D**	St. Olaf	Purdue	Data-8	Data-Science-in-a-Box	U.Mass-Dartmouth	Foundations of Data Science & Big Data (2 courses)	
KDSDA01	x	x	x		x	x	x	x	88%
KDSDA02	x	x	x	x		x	x	x	88%
KDSDA03									0%
KDSDA04	x	x	x	x	x		x		75%
KDSDA05	x		x						25%
KDSDA06	x	x			x	x	x		63%
KDSDA07									0%
KDSDA08	x								13%
KDSDA09	x	x	x	x	x	x	x	x	100%
KDSDA10									0%
KDSDA11	x	x	x	x	x	x	x	x	100%
KDSDA12	x	x	x	x	x	x	x	x	100%
KDSDA13									0%
KDSDA14									0%
KDSDA15	x	x		x		x			50%
SDSDA01	x	x	x	x	x	x	x	x	100%
SDSDA02	x	x	x	x	x	x	x	x	100%
SDSDA03									0%
SDSDA08	x	x	x	x	x	x	x	x	100%
SDSDA09	x	x		x	x	x	x	x	88%
SDSDA10	x	x	x	x	x	x	x	x	100%
KDSENG01									0%
KDSENG02									0%
KDSENG03				x				x	25%
KDSENG04					x				13%
KDSENG05									0%
KDSENG06									0%
KDSENG07	x	x	x	x				x	63%
KDSENG08								x	13%
KDSENG09	x	x	x	x	x	x	x	x	100%
KDSENG10									0%
SDSENG02					x	x		x	38%
SDSENG10									0%
SDSENG11		x		x	x			x	50%
SDSENG12								x	13%

EDISON Competency Framework for Data Science

Specific Label	Smith	Hardin et al. UC B/D**	Hardin et al. St. Olaf	Hardin et al. Purdue	Dusen et al. data8.org Data-8	Cetinkaya-Rundel Data-Science-in-a-Box	Yan and Davis U.Mass-Dartmouth	EDSA Foundations of Data Science & Big Data (2 courses)	Percent Agree
KDSDM01	x								13%
KDSDM02						x			13%
KDSDM03									0%
KDSDM04	x	x	x	x	x	x	x	x	100%
KDSDM05	x	x	x	x	x	x	x	x	100%
KDSDM06		x	x			x	x	x	63%
KDSDM07									0%
KDSDM08									0%
KDSDM09									0%
SDSDM01									0%
SDSDM03									0%
SDSDM04			x	x	x	x	x	x	75%
SDSDM06		x		x					25%
SDSDM07									0%
SDSDM09									0%
KDSRM01	x	x	x	x	x	x	x	x	100%
KDSRM02		x		x	x	x	x	x	75%
KDSRM03	x	x	x	x	x	x	x	x	100%
KDSRM04								x	13%
KDSRM05									0%
KDSRM06									0%
SDSRM01		x	x		x	x	x	x	75%
SDSRM02		x		x		x	x	x	63%
SDSRM04									0%
SDSRM05									0%
SDSRM06									0%
KDSBA01									0%
KDSBA02									0%
KDSBA03									0%
KDSBA04									0%
KDSBA05					x		x	x	38%
KDSBA06				x					13%
KDSBA07									0%
KSDSBA08									0%
SDSBA06									0%
SDSBA09									0%
Count	24	26	21	25	25	26	25	30	
Percent	33%	36%	29%	35%	35%	36%	35%	42%	

EDISON Competency Framework for Data Science

Intro Course	Positive Agreement	Negative Agreement	Indeterminate
Specific Label	AGREE >= 80%	AGREE <= 20%	80% < AGREE < 20%
KDSDA01	X		
KDSDA02	X		
KDSDA03		X	
KDSDA04			X
KDSDA05			X
KDSDA06			X
KDSDA07		X	
KDSDA08		X	
KDSDA09	X		
KDSDA10		X	
KDSDA11	X		
KDSDA12	X		
KDSDA13		X	
KDSDA14		X	
KDSDA15			X
SDSDA01	X		
SDSDA02	X		
SDSDA03		X	
SDSDA08	X		
SDSDA09	X		
SDSDA10	X		
KDSENG01		X	
KDSENG02		X	
KDSENG03			X
KDSENG04		X	
KDSENG05		X	
KDSENG06		X	
KDSENG07			X
KDSENG08		X	
KDSENG09	X		
KDSENG10		X	
SDSENG02			X
SDSENG10		X	
SDSENG11			X
SDSENG12		X	

REFERENCES

[1] ACM Data Science Task Force. Available from: http://dstf.acm.org/.
[2] AICPA, PFP Body of Knowledge. Available from: https://www.aicpa.org/interestareas/personalfinancialplanning/membership/pfsbodyofknowledge.html.

[3] American Statistical Association, Curriculum guidelines for undergraduate programs in statistical science. Available from: http://www.amstat.org/education/curriculumguidelines.cfm.
[4] P. Anderson, J. Bowring, R. McCauley, G. Pothering and C. Starr, An undergraduate degree in data science: Curriculum and a decade of implementation experience, *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, ACM, New York, NY, USA, 2014, 145–150.
[5] P. Anderson, J. McGuffee and D. Uminsky, Data science as an undergraduate degree, *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, ACM, New York, NY, USA, 2014, 705–706.
[6] J. Blitzstein, Teaching data science and storytelling, in *The Data Science Handbook*, Data Science Bookshelf, 2015, 174–187.
[7] Business-Higher Education Forum (BHEF, Webinar: Data science and analytics (dsa)-enabled graduate competency map | BHEF, 2019. Available from: https://s3.goeshow.com/dream/DataSummit/Data%20Summit%202018/BHEF_2016_DSA_competency_map_1.pdf.
[8] I. Cárdenas-Navia and B. K. Fitzgerald, The broad application of data science and analytics: Essential tools for the liberal arts graduate, *Change: The Magazine of Higher Learning*, **47** (2015), 25–32.
[9] B. Cassel and H. Topi, *Strengthening Data Science Education Through Collaboration*, Workshop on Data Science Education Workshop Report, 2015.
[10] CC2020 Task Force, *Computing Curricula 2020: Paradigms for Global Computing Education*, ACM, New York, NY, USA, 2020.
[11] M. Cetinkaya-Rundel, Computing infrastructure and curriculum design for introductory data science, *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, SIGCSE '19, ACM, New York, NY, USA, 2019, 1236–1236.
[12] Civil Engineering Body of Knowledge 3 Task Committee, *Civil Engineering Body of Knowledge: Preparing the Future Civil Engineer*, 3^{rd} edition, American Society of Civil Engineers, Reston, VA, 2019.
[13] N. R. Council, *Training Students to Extract Value from Big Data: Summary of a Workshop*, The National Academies Press, Washington, DC, 2014. Available from: https://www.nap.edu/catalog/18981/training-students-to-extract-value-from-big-data-summary-of.
[14] Data Science Association, About the Data Science Association. Available from: https://www.datascienceassn.org/.
[15] R. D. De Veaux, M. Agarwal, M. Averett, B. S. Baumer and A. Bray, et al., Curriculum guidelines for undergraduate programs in data science, *Ann. Rev. Statist. Appl.*, **4** (2017), 15–30.
[16] Y. Demchenko, A. Belloum, W. Los, T. Wiktorski and A. Manieri, et al.,EDISON data science framework: A foundation for building data science profession for research and industry, IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Luxembourg, Luxembourg, 2016.
[17] Y. Demchenko, L. Comminiello and G. Reali, Designing customisable data science curriculum using ontology for data science competences and body of knowledge, *Proceedings of the 2019 International Conference on Big Data and Education - ICBDE'19*, ACM Press, London, United Kingdom, 2019, 124–128.
[18] EDISON Project, Data science training and data science education - EU. Available from: http://edsa-project.eu/.
[19] EDISON Project, EDISON: Building the data science profession.Available from: https://edison-project.eu/.
[20] U. Fayyad and H. Hamutcu, Toward foundations for data science and analytics: A knowledge framework for professional standards, *Harvard Data Science Review*. Available from: https://hdsr.mitpress.mit.edu/pub/6wx0qmkl/release/2.
[21] D. G. Freelon, ReCal: Intercoder reliability calculation as a Web service, *Internat. J. Internet Sci.*, **5** (2010), 20–33. Available from: http://dfreelon.org/publications/2010_ReCal_Intercoder_reliability_calculation_as_a_web_service.pdf.
[22] L. Haas, A. Hero and R. A. Lue, Highlights of the national academies report on "Undergraduate data science: Opportunities and options", *Harvard Data Science Review*, **1**. Available from: https://hdsr.mitpress.mit.edu/pub/z4sb5j9l/release/3.
[23] J. Hardin, R. Hoerl and N. J. Horton, et al.,Data science in statistics curricula: Preparing students to "think with data", *Amer. Statist.*, **69** (2015), 343–353.

[24] S. C. Hicks and R. A. Irizarry, A guide to teaching data science, *Amer. Statist.*, **72** (2018), 382–391.
[25] T. K. Hira, Personal finance: Past, present and future, *Networks Financial Institute Policy Brief*, (2009), 23pp.
[26] Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) and IEEE Computer Society, *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*, ACM, New York, NY, USA, 2013. Available from: https://www.acm.org/binaries/content/assets/education/cs2013_web_final.pdf.
[27] A. Manieri, S. Brewer, R. Riestra, Y. Demchenko and M. Hemmje, et al.,Data science professional uncovered: How the EDISON Project will contribute to a widely accepted profile for data scientists, IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), Vancouver, BC, Canada, 2015.
[28] P. W. G. Morris, L. Crawford, D. Hodgson, M. M. Shepherd and J. Thomas, Exploring the role of formal bodies of knowledge in defining a profession - The case of project management, *Internat. J. Project Management*, **24** (2006), 710–721.
[29] National Academies of Sciences, *Data Science for Undergraduates: Opportunities and Options*, National Academies Press, 2018. Available from: https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options.
[30] C. Pompa and T. Burke, *Data science and analytics skills shortage: equipping the APEC workforce with the competencies demanded by employers*, Asia-Pacific Economic Cooperation Secretariat, Singapore, 2017, https://www.apec.org/Publications/2017/11/Data-Science-and-Analytics-Skills-Shortage.
[31] R. Rawlings-Goss, L. Cassel, M. Cragin, C. Cramer and A. Dingle, et al.,*Keeping Data Science Broad: Negotiating the Digital & Data Divide*, Technical report, South Big Data Hub, 2018. Available from: https://par.nsf.gov/biblio/10075971.
[32] S. R. Singer, N. R. Nielsen and H. A. Schweingruber, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, National Academies Press, 2012.
[33] E. Van Dusen, A. Suen, A. Liang and A. Bhatnagar, Accelerating the advancement of data science education, *Proceedings of the 18th Python in Science Conference*, 2019, 1–4.
[34] M. A. Waller and S. E. Fawcett, Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management, *J. Business Logistics*, **34** (2013), 77–84.
[35] J. M. Wing and D. Banks, Highlights of the inaugural data science leadership summit, *Harvard Data Science Review*, **1**.
[36] H. Wu, Systematic study of big data science and analytics programs, ASEE Annual Conference & Exposition Proceedings, ASEE Conferences, Columbus, Ohio, 2017.
[37] D. Yan and G. E. Davis, A first course in data science, *J. Statist. Education*, **27** (2019), 99–109.
[38] P. Zorn, C. S. Schumacher and M. J. Siegel,2015 CUPM Curriculum Guide to Majors in the Mathematical Sciences, The Mathematical Association of America, 2015. Available from: https://www.maa.org/sites/default/files/pdf/CUPM/pdf/CUPMguide_print.pdf.

Received July 2021; 1st and 2nd revisions October 2021; early access November 2021.

E-mail address: karl.schmitt@trnty.edu
E-mail address: linda_clark@brown.edu
E-mail address: kkinnaird@smith.edu
E-mail address: ruth.wertz@valpo.edu
E-mail address: bjorn_sandstede@brown.edu