Evaluating undergraduate data science curricula: what does a degree in data science actually mean?

Jeffrey C Oliver [Corresp., 1] , **Torbet McNeil** [1, 2]

[1] Office of Digital Innovation & Stewardship, University Libraries, University of Arizona, Tucson, Arizona, United States

[2] Department of Educational Policy Studies and Practice, University of Arizona, Tucson, Arizona, United States

Corresponding Author: Jeffrey C Oliver
Email address: jcoliver@email.arizona.edu

The interdisciplinary field of data science, which applies techniques from computer science and statistics to address questions across domains, has enjoyed recent considerable growth and interest. This emergence also extends to undergraduate education, whereby a growing number of institutions now offer degree programs in data science. However, there is considerable variation in what the field actually entails and, by extension, differences in how undergraduate programs prepare students for data-intensive careers. We used two seminal frameworks for data science education to evaluate undergraduate data science programs at a subset of four-year institutions in the United States; developing and applying a rubric, we assessed how well each program met the guidelines of each of the frameworks. Most programs scored high in statistics and computer science and low in domain-specific education, ethics, and areas of communication. Moreover, the academic unit administering the degree program significantly influenced the course-load distribution of computer science and statistics/mathematics courses. We conclude that current data science undergraduate programs provide solid grounding in computational and statistical approaches, yet may not deliver sufficient context in terms of domain knowledge and ethical considerations necessary for appropriate data science applications. Additional refinement of the expectations for undergraduate data science education is warranted.

1   **Evaluating undergraduate data science curricula: what does a degree in data science**

2   **actually mean?**

3   Jeffrey C. Oliver[1], Torbet McNeil[1,2]

4   [1] Office of Digital Innovation & Stewardship, University Libraries, University of Arizona,

5   Tucson, AZ, USA

6   [2] Department of Educational Policy Studies and Practice, University of Arizona, Tucson, AZ,

7   USA

9   Corresponding Author:

10   Jeffrey C. Oliver[1]

11   1510 E. University Blvd., Tucson, AZ, 85721, USA

12   Email address: jcoliver@arizona.edu

18 **Abstract**

19 The interdisciplinary field of data science, which applies techniques from computer

20 science and statistics to address questions across domains, has enjoyed recent considerable

21 growth and interest. This emergence also extends to undergraduate education, whereby a growing

22 number of institutions now offer degree programs in data science. However, there is considerable

23 variation in what the field actually entails and, by extension, differences in how undergraduate

24 programs prepare students for data-intensive careers. We used two seminal frameworks for data

25 science education to evaluate undergraduate data science programs at a small subset of four-year

26 institutions in the United States; developing and applying a rubric, we assessed how well each

27 program met the guidelines of each of the frameworks. Most programs scored high in statistics

28 and computer science and low in domain-specific education, ethics, and areas of communication.

29 Moreover, the academic unit administering the degree program significantly influenced the

30 course-load distribution of computer science and statistics/mathematics courses. We conclude

31 that current data science undergraduate programs provide solid grounding in computational and

32 statistical approaches, yet may not deliver sufficient context in terms of domain knowledge and

33 ethical considerations necessary for appropriate data science applications. Additional refinement

34 of the expectations for undergraduate data science education is warranted.

35

36 Keywords: computer science, curricula, data science, education, ethics, machine learning,

37 statistics

38

39

Background

40

41   Data-intensive work and the desire for data-driven decisions increasingly fuel interest in

42   the field of data science. According to the National Academies of Sciences, Engineering, and

43   Medicine (NASEM, 2018), employers across disciplines are demanding that employees have

44   skills in working with and extracting knowledge from data. Additionally, the report suggests

45   every undergraduate student should graduate with at least beginning competency for working

46   with data. However, many undergraduates are not obtaining the necessary training to prosper in

47   the new economy. College administrators have reacted, and the number of undergraduate data

48   science programs has multiplied. In 2014, there were approximately 13 undergraduate major

49   programs (Aasheim et al., 2015), while at least 50 undergraduate programs existed in the United

50   States as of September 2020 (Swanstrom, 2020). Interest in data science in undergraduate

51   education is also evidenced by the growth of courses such as the University of California,

52   Berkeley's lower-division Foundations of Data Science course, which increased from 100

53   students in 2013 to over 1,000 students in 2018 (Kafka, 2018). Similar growth can be expected

54   internationally, particularly in more developed, knowledge-based economies striving to improve

55   their universities, as the U.S. higher education model is often emulated (Bok, 2013).

56   Despite the attention data science is receiving, there is a lack of a clear definition of what

57   data science actually entails (Donoho, 2017; Irizarry, 2020). While an exhaustive history of the

58   term and definitions is beyond the scope of this work, we adhere to the following general

59   description: Data science draws on statistics and computer science to address questions in various

60   domains, such as biology, education, physics, business, linguistics, or medicine (Donoho, 2017;

61   National Academies of Sciences, Engineering, and Medicine, 2018). In addition to quantitative

62   skills and domain expertise, this definition of data science also includes "soft skills," such as

63   communication and a working cognizance of the ethics of data use and reuse (Irizarry, 2020).

64   This definition of data science is relatively broad and includes the narrower fields of data

65   engineering (development and management of data infrastructure for subsequent interrogation

66   and analyses) and data analytics (application of statistical and predictive analyses to address

67   unknowns in a particular domain). The training students receive when pursuing a data science

68   education, both what is included and what is excluded, is worthy of investigation.

69   While data science undergraduate degree programs are relatively new, previous work has

70   highlighted early trends. In a survey of five data science degree programs, Aasheim et al. (2015)

71   found an emphasis on statistics and mathematics coursework. All programs required coursework

72   in linear algebra and data mining as well as multiple programming and statistics courses. In

73   contrast, none of the five programs required coursework in ethical considerations of data science,

74   and only one program required coursework in communication skills. Additional descriptions of

75   individual data science undergraduate programs are quite varied and include those that are

76   business-focused (Anderson, McGuffee, & Uminsky, 2014) and those that require significant

77   coursework in domains outside of mathematics, statistics, and computer science (Anderson et al.,

78   2014). The variation among programs illustrates a potentially confusing landscape for students to

79   navigate and uncertainty for employers when assessing recent graduates' preparation (Parry,

80   2018).

81   Quantitative evaluation of data science programs requires an explicit framework

82   describing the components of data science education. Several frameworks exist, including general

83   frameworks focused on mathematical and computational foundations (De Veaux et al., 2017),

84   frameworks based on individual programs (Anderson et al., 2014), and emergent frameworks

85   developed for comparative analyses (Aasheim et al., 2015). For this work, we used two

86   frameworks for evaluation: the broad, explicit framework presented in the NASEM (2018) report

87   and the narrower, more conceptual framework of Donoho (2017). The framework of the NASEM

88   (2018), hereafter referred to as the NASEM framework, focuses on developing undergraduate

89   data acumen through a curriculum including important data science concepts, applications to real-

90   world problems with an understanding of limitations, and ethical concerns involved in data

91   science (Table 1, Supplemental Table 1). In addition to "traditional" data science competencies in

92   computer science and statistics, this framework gives guidelines for training in communication

93   skills, domain-specific knowledge, and ethical considerations. The NASEM framework lists ten

94   key concept areas ("areas" hereafter) that are further divided into specialized topics, skills, or

95   concepts ("sub-areas" hereafter), providing comprehensive expectations for undergraduate

96   training. This framework facilitates an extensive evaluation of how well undergraduate degree

97   programs meet the expectations set forth by the NASEM.

98   The framework presented by Donoho (2017) as Greater Data Science, hereafter the GDS

99   framework, describes programs that prepare professionals for gaining insights from data while

100   applying best practices (Table 2, Supplemental Table 2). The six areas in the GDS framework are

101   relatively high-level, and the framework has notably little discussion of the knowledge, skills, or

102   abilities necessary to apply such foundational knowledge to domain-specific questions. While not

103   explicitly designed for undergraduate education, the GDS framework furnishes a metric to assess

104   undergraduate programs' potential for preparing future professionals in data science careers.

105   Here we take the opportunity to evaluate undergraduate data science degree programs in a

106   comparative analysis using the two frameworks described above. Applying an evaluation rubric

107   we developed for each of the two frameworks, we investigated data science programs from a

108   sample of doctoral-granting universities. We reviewed major requirements and corresponding

109   course descriptions to assess how well each program addressed elements of each framework. We

110   also quantified the relative amount of coursework in three categories: computer science, statistics/

111   mathematics, and domain knowledge. Using evaluations and quantification of coursework, we

112   provide an overview of how well each of the two frameworks is being implemented and an

113   evaluation of training strengths and weaknesses in data science undergraduate degree programs.

114   **Methods**

115   In an attempt to make appropriate comparisons among undergraduate data science

116   programs, we chose institutions comparable to our home institution, the University of Arizona.

117   We used two means of inclusion: institutions recognized as peers by the University of Arizona

118   and institutions in the Pac-12 Conference (https://uair.arizona.edu/content/ua-peers). The union

119   of University of Arizona peers and Pac-12 institutions resulted in a total of 25 universities, all of

120   which are Research I universities (Carnegie Classification of Institutions of Higher Education,

121   2018) and four-year, doctoral-granting institutions. While half of these 25 institutions are in the

122   western United States, this sample also includes public universities in the midwestern, eastern,

123   and southern United States. This sample includes members of the Association of American

124   Universities (https://www.aau.edu/) and land-grant institutions. At the time of inception of this

125   work, roughly 50% of these institutions (53% of University of Arizona peer institutions and 50%

126   of Pac-12 institutions) offered an undergraduate major or minor in a data science-related field. 10

127   institutions had at least one undergraduate major in data science. Variation in the names of

128   programs required careful consideration. For example, Ohio State University offered a bachelor

129   of science in data analytics; however, the curriculum was similar to programs with the term "data

130   science" in the name of the degree, so this program was included in our evaluation. In contrast,

131   the bachelor of science in business data analytics at Arizona State University had a course

132   curriculum that was very different from other data science degrees and was, thus, excluded from

133   evaluation. In cases when an institution offered more than one degree in data science (e.g.,

134   University of Washington), we scored each of the degree programs independently. This selection

135   process resulted in a total of 18 scored programs (Table 3).

136   We evaluated how well each undergraduate data science program aligned with

137   recommendations in the NA and GDS frameworks. For each of the two frameworks, we

138   developed a rubric and coded the undergraduate data science curricula on a four-point scale using

139   direct survey methodology, specifically content analysis of course descriptions (Stefanidis &

140   Fitzgerald, 2014; Aasheim et al., 2015), indicating the familiarity with a topic that could be

141   expected from a student graduating from the program in question. The creation of the rubric used

142   for scoring was an iterative process. We created an initial rubric and used it to score each item in

143   the two frameworks. For nine of the ten areas of the NASEM framework, an item corresponded

144   to one sub-area within the larger area. For example, in the area of computational foundations, five

145   sub-areas were listed: basic abstractions, algorithmic thinking, programming concepts, data

146   structures, and simulations. The ability of a program to address each of these five sub-areas was

147   assessed separately. One area in the NASEM framework, domain-specific considerations, did not

148   list any additional specifics; so, in this case, the item scored was the area of domain-specific

149   considerations itself. Similarly, in the GDS framework, most areas lacked additional descriptions

150   of sub-areas, so items largely corresponded to the particular area. The two exceptions were data

151   representation and transformation and data modeling; each of these areas had two scored sub-

152   areas.

153   In general, a score of "1" indicated no expectation that a student graduating from the

154   program would have familiarity with the area/sub-area; a score of "4" indicated a student is well-

155   versed in the area/sub-area, with at least one required course covering the topic, often at length.

156   Scores were based on course titles and descriptions only (links to each programs' web page are

157   available in Supplemental Table 3 and lists of course requirements are available in Supplemental

158   File 1); a lack of standardization among institutions required close reading of all course

159   descriptions listed in posted curricula. We did not include information from course syllabi as

160   there was considerable variation in which courses had publicly available syllabi. There was

161   considerable revision to the rubric throughout the coding process as nuances in scoring and

162   inconsistencies were noticed. We scored each program for each framework independently and

163   then discussed discrepancies to reach agreements on a common score. Full details of the final

164   coding rubric are available in Appendix A.

165   In addition to scoring programs for the two frameworks, we quantified the total number of

166   credits required for each program in three categories: computer science, statistics/ mathematics,

167   and domain-specific courses. Domain-specific courses are those outside of computer science,

168   statistics, and mathematics, such as in biology, economics, or psychology. In cases where it was

169   difficult to categorize a course as computer science or statistics/mathematics, we used the identity

170   of the home department to inform the categorization. For example, if an ambiguous course was

171   offered by a computer science department, it was categorized as a computer science course. Such

172   cases were rare and only affected the categorization of five or fewer courses. Given the flexibility

173   in course choice in some programs, we recorded the minimum and maximum number of credits

174   for each of the categories as well as the minimum and maximum total credits required for each

175   program. The total credits for a program also included units that were not categorized into any of

176   the three categories, such as internships and senior capstone projects.

177   *Statistical Analyses*

178   In all subsequent analyses, we excluded programs that were not data science majors,

179   primarily to afford appropriate comparisons in coursework coverage and requirements. This

180   resulted in the exclusion of Stanford University's data science minor and the University of

181   Illinois at Urbana-Champaign's data science certificate.

182   To compare coverage among areas within each of the two frameworks, we first estimated

183   an ordinal mixed-effects model, treating area as a fixed effect and the program as a random-

184   intercept effect. We estimated separate models for each of the two frameworks. Based on the

185   ordinal mixed-effects models, we performed post-hoc pairwise comparisons to assess significant

186   differences between areas. All analyses were performed with the R programming language (R

187   Core Team, 2020) with the aid of the tidyverse (Wickham, 2017) and ordinal (Christensen, 2018)

188   packages. All data and R code are available at https://github.com/jcoliver/ data-sci-curricula.

189   We evaluated the portion of each undergraduate major programs' total credits dedicated

190   to computer science, statistics/mathematics, and domain-specific coursework. Given that there

191   was variation in the required credits within programs, we used the midpoint between the

192   minimum and maximum for each of the categories and total credits in subsequent analyses. For

193   example, the University of Iowa's bachelor of science in data science required between 17 and 23

194   credits in computer science coursework, so we used 20 as the expected number of computer

195   science credits for this program. We tested two hypotheses with these data using one-tailed

196   Student's *t*-tests:

197   1. Programs housed in computer science units have more required coursework in computer

198   science than programs housed in other academic units.

199   2. Programs housed in statistics or mathematics units have more required coursework in statistics

200   and mathematics than programs housed in other academic units.

201   *Limitations*

202   We based program assessments solely on course titles and descriptions rather than on

203   syllabi. Although they often provide detailed descriptions of course material, syllabi availability

204   is highly variable, and course content at the level of syllabi may vary by term and instructor. By

205   focusing on course descriptions, which were accessible online for all institutions and programs

206   investigated (Supplemental Table 3), we were able to consistently assess program performance.

207   Course titles and descriptions themselves were variable within and among institutions. For

208   example, some course descriptions consisted of a short enumeration of topics (e.g., the

209   description for STAT 102, Data, Inferences, and Decisions at the University of California,

210   Berkeley), while others included course content as well as format (e.g., DS 340W, Applied Data

211   Sciences at Pennsylvania State University). There remains potential for bias among the different

212   areas/sub-areas described in the two frameworks. That is, areas/sub-areas we characterized as

213   generally poorly covered (see Results) may reflect systematic poor representation in course

214   descriptions because the topics were not mentioned in descriptions even though they may have

215   been taught in the course. However, with only five exceptions, all areas/sub-areas scored a

216   minimum of 3 for at least one program, indicating that nearly all areas/sub-areas could be

217   described in a course description with enough detail to warrant the highest score possible.

218   Finally, the focus of the current work is undergraduate degree programs offered at doctoral-

219   granting universities in the United States; thus, care is needed in extrapolating the implications to

220   other types of institutions (primarily undergraduate institutions, liberal arts colleges, universities

221   outside the United States, etc.).

222   **Results**

223   For the guidelines set forth by the NASEM (NASEM framework), some areas/sub-areas

224   were generally well covered in the programs we evaluated, while other areas/sub-areas received

225   little to no attention in formal coursework (Figure 1). The highest scoring area was data

226   description and visualization, which primarily involves quality assessment and exploratory data

227   analysis (mean: 3.5, median: 3.7). Most programs also paid substantial attention to computational

228   foundations, which includes abstraction, algorithmic thinking, and programming concepts (mean:

229   3.4, median: 3.4). In contrast, areas/sub-areas focusing on reproducibility and ethics in data

230   science were generally not covered in undergraduate curricula. Reproducibility, including design

231   of workflows and reproducible analyses, was rarely indicated in course descriptions (mean: 1.6,

232   median: 1.0). Ethics of data science, covering privacy, confidentiality, and misrepresentations of

233   data and results, received the lowest average score (mean: 1.6, median: 1.6). Post-hoc pairwise

234   comparisons among areas illustrated computational foundations, statistical foundations, and data

235   description and visualization all had significantly higher scores than reproducibility, ethics, and

236   domain expertise (Supplemental Table 4).

237   In regards to the GDS framework, programs scored, on average, high in all areas/sub-

238   areas except science about data science (mean: 1.9, median: 2.0), which is the explicit

239   investigation of data science as a field (Figure 2). In post-hoc pairwise comparisons among area

240   scores, science about data science scored significantly lower than all other areas except data

241   gathering, preparation, and exploration (Supplemental Table 5).

242   The majority of programs investigated were characterized by coursework focused on

243   mathematical and statistical foundations (Figure 3). In 12 programs, courses in mathematics or

244   statistics made up over 50% of the required coursework. In all but two programs, computer

245   science courses accounted for less than 50% of the required coursework. Required coursework in

246   domains outside of computer science, statistics, and mathematics was relatively low, and in only

247   one program did domain coursework exceed 25% of required coursework. The academic unit

248   administering the degree program significantly influenced the proportion of the coursework

249   dedicated to computer science and to statistics and mathematics. Programs administered by

250   computer science units had significantly more required computer science courses than did

251   programs administered by other units (mean percentage of computer science coursework in

252   programs housed in computer science units: 39%, other units: 18%; $t = 2.899$, p = 0.016).

253   Similarly, programs housed in statistics or mathematics units required more coursework in

254   statistics and mathematics than did programs housed in other units (mean percentage of statistics/

255   mathematics coursework in programs housed in statistics or mathematics units: 70%, other units:

256   52%; $t = 2.441$, p = 0.015).

257   **Discussion**

258   Our comparison of undergraduate curricula to two seminal data science frameworks

259   reflects a focus on theoretical foundations and quantitative skills (Figure 1 and Figure 2). Under

260   both frameworks, training in computational foundations was among the highest scoring areas.

261   Indeed, computational applications addressing big, complex problems have been a hallmark of

262   data science. Most programs also scored well in statistics and mathematics training, as well as

263   data modeling. The emphasis on quantitative skills training, coupled with theoretical concepts

264 underlying applications, indicates a focus on the statistical and computational underpinnings of

265 data science and echoes similar findings in prior work on undergraduate (Aasheim et al., 2015)

266 and graduate (Tang & Sae-Lim, 2016) data science programs. While curricula also scored well in

267 data description and visualization (NASEM framework) and data visualization and presentation

268 (GDS framework), the competencies described in these areas are primarily concerned with

269 exploratory data analysis and quality assurance processes rather than using visual representations

270 of data to communicate ideas. This statistical and computational focus is further evidenced by a

271 heavy bias towards the number of course units in computer science, statistics, and mathematics in

272 undergraduate data science programs (Figure 3).

273   Our evaluation of programs presents a *sensu stricto* (*s.s.*) definition of data science

274 education for undergraduates, aligning with the description of the field in the GDS framework

275 (Donoho, 2017), whereby most programs emphasize the "hard skills" associated with computer

276 science, statistics, and mathematics. Accompanying the GDS framework, Donoho (2017) posited

277 that academic data science degree programs focus on statistics and machine learning, with some

278 attention given to the technology required to compute on big data. This view of data science *s.s.*

279 does little to include substantial training in domain knowledge outside of computer science,

280 statistics, and mathematics. For example, the undergraduate data science curricula described by

281 De Veaux et al. (2017) implies one or two domain-specific courses provide sufficient formal

282 training in a domain. For the most part, the programs evaluated here likely equip graduates with a

283 varied skill set for analyses and prediction, but graduates may lack the appropriate context for

284 designing and evaluating domain-specific data science applications.

285   In contrast, data science *sensu lato* (*s.l.*) would include substantial training in

286 communication, ethical considerations, and knowledge in the domain to which analyses and

287 predictive modeling are applied. The NASEM framework called out the importance of domain

288 knowledge for effective application of data science, yet few programs went beyond requiring two

289   additional courses outside of computer science, statistics, and mathematics. One notable

290   exception is the human centered design and engineering data science option at the University of

291   Washington, although the emphasis on domain-specific education may come at the cost of

292   reduced training in computational skills and statistics (Figure 3). The curriculum described in

293   Anderson et al. (2014) provides another example of undergraduate training in data science *s.l.*:

294   Students were required to take substantial coursework (15-22 units) within a "cognate," such as

295   biomechanics, geoinformatics, or sociology. This view of data science training supports the

296   notion that sufficient domain background is required to understand the context of models and

297   analyses (Provost & Fawcett, 2013), justifying substantial consideration of domain knowledge in

298   data science training (Berthold, 2019; Irizarry, 2020).

299   The degree to which data science education includes domain specialization is likely

300   influenced by the academic affiliations of those who set the criteria. This reflects a common

301   phenomenon within data science: The definition of data science (and by extension, data science

302   education) depends on who is doing the defining (Provost & Fawcett, 2013). Just as we found

303   that the academic unit in which the undergraduate program was housed had a significant effect on

304   the amount of computer science and statistics coursework, the academic fields of those people

305   creating curricula may affect how much emphasis there is on domain knowledge. For example,

306   the undergraduate data science curricula of De Veaux et al. (2017) includes, at most, three

307   domain-specific courses (one introductory, one intermediate, and one capstone), and the authors

308   of this curriculum are all from departments of computer science, statistics, or mathematics. In

309   contrast, the curriculum presented by Anderson et al. (2014) includes substantially more

310   coursework in a domain other than computer science, statistics, or mathematics and was created

311   by faculty from biology as well as computer science and mathematics. Such differences among

312   curricula further illustrate data science as an evolving field and demonstrate considerable

313   heterogeneity in what can be expected from recent graduates of undergraduate data science

314   programs.

315   Similar to the paucity of attention to domain knowledge, most programs did not explicitly

316   provide training in workflows, reproducibility practices, and the ethics of data use and reuse. The

317   dearth of training dedicated to ethical problem-solving is similar to earlier comparative findings

318   (Aasheim et al., 2015). Best practices in reproducibility and ethics are critical for maintaining

319   quality of data science applications (Saltz et al., 2018), and their omission from undergraduate

320   data science programs potentially creates a *Promethean* workforce prepared to use a variety of

321   computational and statistical tools in socially inappropriate ways. A growing body of examples

322   illustrate bias in data science applications (O'Neil, 2017); such biases have real-world impact in

323   criminal justice (Isaac & Lum, 2016), employment (Dastin, 2018), and healthcare (Obermeyer et

324   al., 2019). These impacts further reinforce the necessity of appropriate training in ethical

325   considerations in data science.

326   Areas marked by deficiencies in the two frameworks may also reflect the relative youth of

327   the field of data science. For example, the GDS framework includes the area science about data

328   science, which received relatively low scores; one contributing factor could be that there remains

329   discussion of what data science actually entails, precluding a formal study and circumscription of

330   the field (Donoho, 2017; Irizarry, 2020). Similarly, the lack of attention to ethical considerations

331   in undergraduate data science programs could be a consequence of the recent rise of data science.

332   For example, the qualitative and quantitative changes to analyses and predictions brought on by

333   the big data revolution have created a new landscape for ethical considerations, and training in

334   ethical issues in data science remains a growth area (Saltz et al., 2018). In contrast to "traditional"

335   data science topics, such as linear algebra, which have long been recognized as important for

336   statistical analyses, ethical precepts of data science may take more time to become integrated into

337   undergraduate data science curricula. The low scores in the ethical problem solving area of the

338   NASEM framework may also be due, in part, to our means of assessment. Course descriptions

339   rarely went beyond mentioning the "ethics of data science," while the sub-areas described in the

340   NASEM framework, and thus our evaluation rubric, included specifics such as "the ability to

341   identify 'junk' science" and "the ability to detect algorithmic bias." If these important ethical

342   topics are included in data science curricula, course descriptions would do well to call them out

343   explicitly.

344   **Conclusions**

345   Our assessments of undergraduate data science curricula demonstrate a focus on

346   theoretical foundations and quantitative skills with relatively little preparation in domains outside

347   of computer science, statistics, and mathematics. This generally aligns with the "greater data

348   science" definition provided by Donoho (2017). The work here suggests that data science

349   undergraduate students receive training similar to those enrolled in a statistics program, although

350   additional work formally comparing statistics degree programs to data science degree programs is

351   needed. Additionally, an evaluation of the factors influencing data science curricula is beyond the

352   scope of this work, but future work should consider the possibility that current data science

353   programs are reflexive responses to market demands *à la* academic capitalism (Slaughter &

354   Leslie, 1999; Slaughter & Rhoades, 2004) or isomorphic processes (Dimaggio & Powell, 1983).

355   Many programs fell short of guidelines put forth by the NASEM (2018). One possibility

356   is that the areas that were not well covered (e.g., reproducibility, ethics, domain knowledge) are

357   not recognized by the data science community as warranting substantial training. Alternatively, it

358   remains an open question of how realistic it is to expect an undergraduate program to effectively

359   cover all areas described in the NASEM framework. While our assessments generally treated all

360   competencies with equal weight, the NASEM's recommendations may afford interpretations

361   whereby some sub-areas require significant coursework, such as a year of linear algebra, while

362   other sub-areas merit one or a few class periods, such as ethical consideration of data science.

363   Our reliance on course titles and course descriptions may bias against topics that receive minimal,

364   albeit potentially impactful, coverage in curricula. Future work comparing curricular content

365   through in-depth program reviews or exhaustive syllabi sampling could demonstrate more nuance

366   in the variation among undergraduate data science degree programs. Additional revisions to

367   undergraduate data science education guidelines, including required levels of competencies,

368   could prove helpful in defining what should be expected from a recent graduate of an

369   undergraduate data science program.

370

371   **Acknowledgments**

372   We thank Harry Hochheiser and two anonymous reviewers for insightful feedback on an earlier

373   version of this work.

374

375   **Data Availability**

376   All data are available on Zenodo DOI 10.5281/zenodo.4329722.

377

378   **References**

379   Aasheim CL, Williams S, Rutner P, Gardiner A. 2015. Data analytics vs. data science: A study of

380   similarities and differences in undergraduate programs based on course descriptions. *Journal of*

381   *Information Systems Education* 26:103–115.

382   Anderson P, Bowring J, McCauley R, Pothering G, Starr C. 2014. An undergraduate degree in

383   data science: Curriculum and a decade of implementation experience. In: *SIGCSE 2014 -*

384   *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* 145–150.

385   DOI: 10.1145/2538862.2538936.

386   Anderson P, McGuffee J, Uminsky D. 2014. Data science as an undergraduate degree. In:

387   *SIGCSE 2014 - Proceedings of the 45th ACM Technical Symposium on Computer Science*

388   *Education*. Association for Computing Machinery 705–706. DOI: 10.1145/2538862.2538868.

389   Berthold MR. 2019. What Does It Take to be a Successful Data Scientist? *Harvard Data Science*

390   *Review* 1:1–6. DOI: 10.1162/99608f92.e0eaabfc.

391   Bok D. 2013. *Higher education in America*. Princeton, NJ: Princeton University Press.

392   The Carnegie Classification of Institutions of Higher Education. 2018. About Carnegie

393   Classification. *Available at http://carnegieclassifications.iu.edu/* (accessed February 23, 2020).

394   Christensen RHB. 2019. ordinal---Regression Models for Ordinal Data. R package version

395   2019.12-10. https://CRAN.R-project.org/package=ordinal.

396   Dastin J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women.

397   Reuters. *Available at www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-*

398   *scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G* (accessed

399   December 14, 2020).

400   De Veaux RD, Agarwal M, Averett M, Baumer BS, Bray A, Bressoud TC, Bryant L, Cheng LZ,

401   Francis A, Gould R, Kim AY, Kretchmar M, Lu Q, Moskol A, Nolan D, Pelayo R, Raleigh S,

402   Sethi RJ, Sondjaja M, Tiruviluamala N, Uhlig PX, Washington TM, Wesley CL, White D, Ye P.

403   2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of*

404   *Statistics and Its Application* 4:15–30. DOI: 10.1146/annurev-statistics-060116-053930.

405   DiMaggio PJ, Powell WW. 1983. The Iron Cage Revisited: Institutional Isomorphism and

406   Collective Rationality in Organizational Fields. *American Sociological Review* 48:147–160. DOI:

407   10.2307/2095101.

408   Donoho D. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*

409   26:745–766. DOI: 10.1080/10618600.2017.1384734.

410   Irizarry RA. 2020. The Role of Academia in Data Science Education. *Harvard Data Science*

411   *Review* 2. DOI: 10.1162/99608f92.dd363929.

412   Kafka AC. 2018. With Student Interest Soaring, Berkeley Creates New Data-Sciences Division.

413   *The Chronicle of Higher Education. Available at https://www.chronicle.com/article/with-student-*

414   *interest-soaring-berkeley-creates-new-data-sciences-division/* (accessed April 5, 2019).

415   Lambrecht A, Tucker C. 2019. Algorithmic bias? An empirical study of apparent gender-based

416   discrimination in the display of stem career ads. *Management Science* 65:2966–2981. DOI:

417   10.1287/mnsc.2018.3093.

418   Lum K, Isaac W. 2016. To predict and serve? *Significance* 13:14–19. DOI: 10.1111/j.1740-

419   9713.2016.00960.x.

420   National Academies of Sciences, Engineering and Medicine. 2018. *Data Science for*

421   *Undergraduates*. DOI: 10.17226/25104.

422   O'Neil C. 2017. *Weapons of Math Destruction*. New York, NY: Broadway Books.

423   Obermeyer Z, Powers B, Vogeli C, Mullainathan S. 2019. Dissecting racial bias in an algorithm

424   used to manage the health of populations. *Science* 366:447–453. DOI: 10.1126/science.aax2342.

425   Parry M. 2018. Data Scientists in Demand: New programs train students to make honest sense of

426   numbers.(THE TRENDS REPORT). *The Chronicle of Higher Education* 64:B33.

427   Provost F, Fawcett T. 2013. Data Science and its Relationship to Big Data and Data-Driven

428   Decision Making. *Big Data* 1:51–59. DOI: 10.1089/big.2013.1508.

429   R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria.

430   https://www.r-project.org/.

431   Saltz JS, Dewar NI, Heckman R. 2018. Key concepts for a data science ethics curriculum.

432   *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science*

433   *Education* 2018-January:952–957. DOI: 10.1145/3159450.3159483.

434   Slaughter S, Leslie LL. 1997. *Academic Capitalism : Politics, Policies, and the Entrepreneurial*

435   *University*. Baltimore, Md.: Johns Hopkins University Press.

436   Slaughter S, Rhoades G. 2004. *Academic Capitalism and the New Economy : Markets, State, and*

437   *Higher Education*. Baltimore, Md.: Johns Hopkins University Press.

438   Stefanidis A, Fitzgerald G. 2014. IS Education Research: Review of Methods of Surveying the IS

439   Curriculum to Support Future Development of IS Courses. In: Wrycza S ed. *Information*

440   *Systems: Education, Applications, Research*. Cham: Springer International Publishing, 1–11.

441   Swanstrom R. 2020.College & University Data Science Degrees. *Available at*

442   *http://datascience.community/colleges* (accessed February 23, 2020).

443   Tang R, Sae-Lim W. 2016. Data science programs in U.S. higher education: An exploratory

444   content analysis of program description, curriculum structure, and course focus. *Education for*

445   *Information* 32:269–290. DOI: 10.3233/EFI-160977.

446   Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A,

447   Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D,

448   Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to

449   the {tidyverse}. *Journal of Open Source Software* 4:1686. DOI: 10.21105/joss.01686.

Figure 1

Figure 1. Undergraduate data science programs' scores for the NASEM framework.

Open circles show the average score for each area across all programs (±2 SE). Filled circles show mean scores in each area of the framework for individual programs.
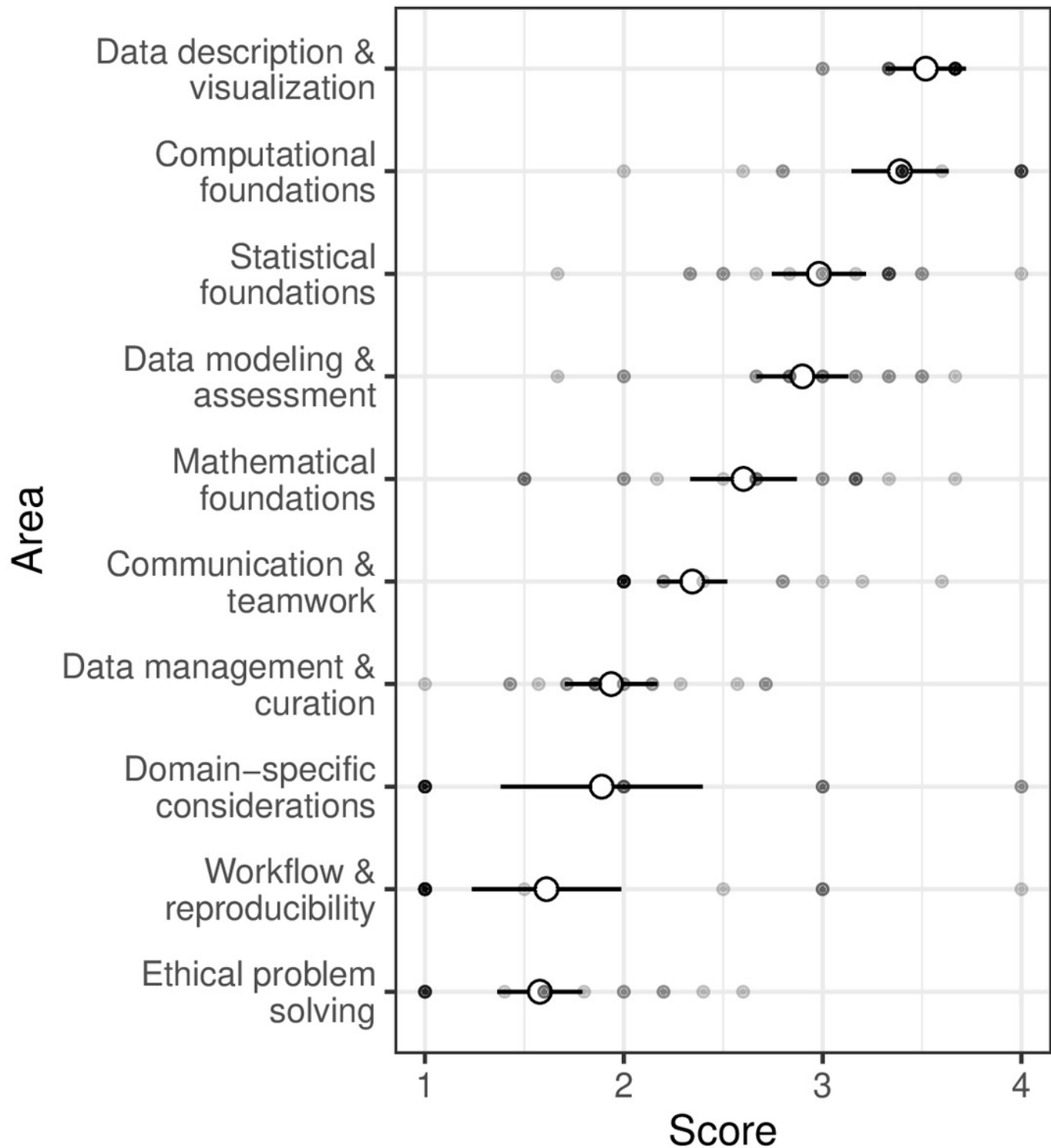
Figure 2

Figure 2. Undergraduate data science programs' scores for the GDS framework.

Open circles show the average score for each area across all programs (±2 SE). Filled circles show mean scores in each area of the framework for individual programs.
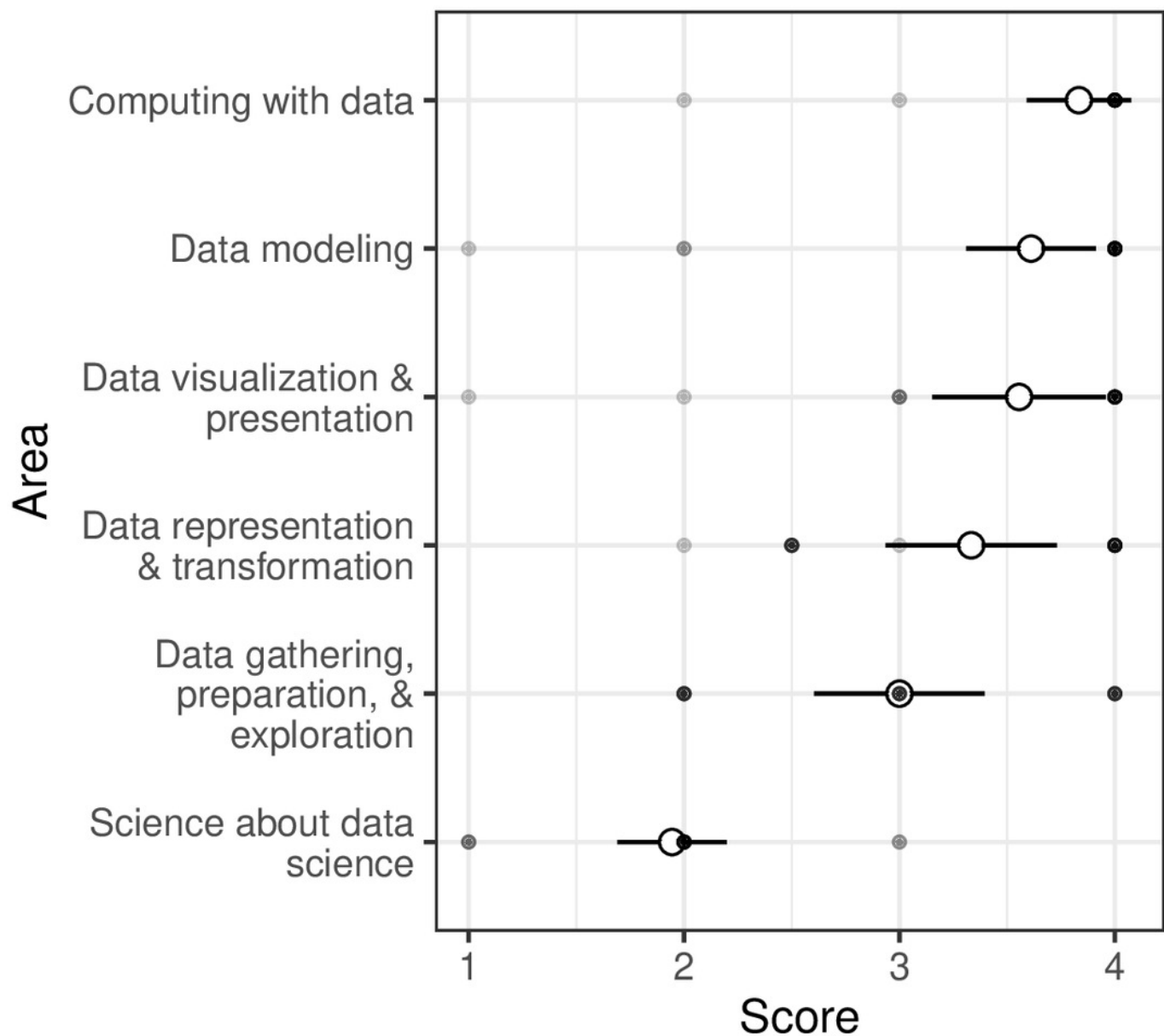
Figure 3

Figure 3. Proportion of coursework in computer science, statistics/mathematics, and domain-specific courses for each data science degree program.

Values are the proportion of total units/credits required for a degree in the data science program; see Methods for calculation details. Shaded rectangles indicate affiliation of academic unit(s) that administer the degree program.
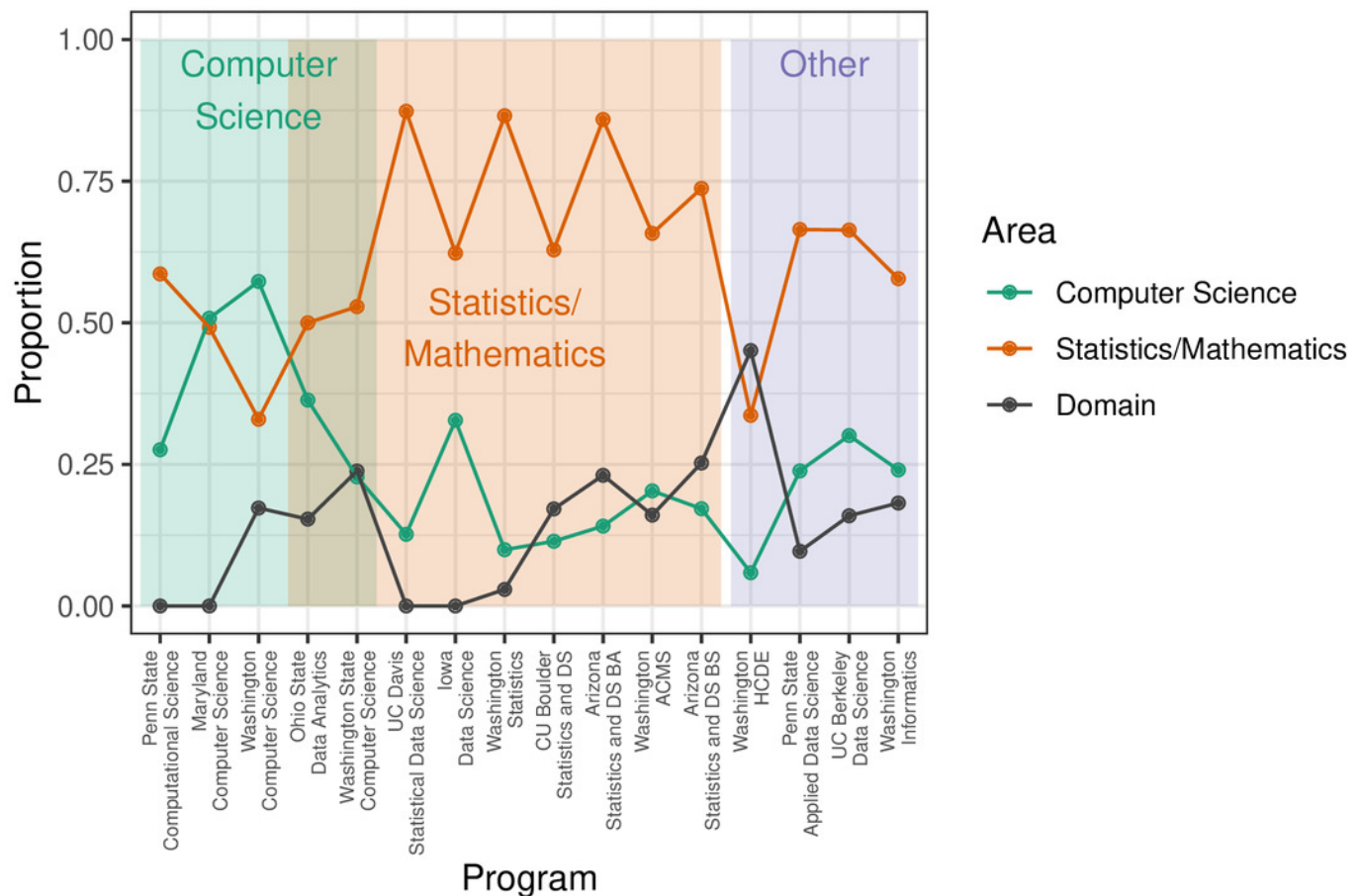
Table 1. 10 areas of emphasis in the NASEM framework.

Table 1. 10 areas of emphasis in the NASEM framework.

Mathematical foundations

Computational foundations

Statistical foundations

Data management and curation

Data description and visualization

Data modeling and assessment

Workflow and reproducibility

Communication and teamwork

Domain-specific considerations

Ethical problem solving

Table 2. Six areas of emphasis in the GDS framework.

Table 2. Six areas of emphasis in the GDS framework.

Data gathering, preparation, and exploration
Data representation and transformation
Computing with data
Data modeling
Data visualization and presentation
Science about data science

Table 3. Undergraduate data science programs considered in this work.

Administering unit lists the academic unit(s) responsible for the major, minor, or certificate program.

Table 3. Undergraduate data science programs considered in this work. Administering unit lists the academic unit(s) responsible for the major, minor, or certificate program.

Institution	Program	Administering unit
Ohio State University - Main Campus	B.S. in Data Analytics	Computer Science & Engineering; Statistics
Pennsylvania State University - Main Campus	B.S. in Data Sciences, Applied Data Science Option	Information
Pennsylvania State University - Main Campus	B.S. in Data Sciences, Computational Science Option	Computer Science & Engineering
Stanford University	Minor in Data Science	Statistics
University of Arizona	B.A. in Statistics and Data Science	Mathematics
University of Arizona	B.S. in Statistics and Data Science	Mathematics
University of California - Berkeley	B.A. in Data Science	Data Science
University of California - Davis	B.S. in Statistics, Statistical Data Science Track	Statistics
University of Colorado - Boulder	B.A. in Statistics and Data Science	Mathematics
University of Illinois at Urbana-Champaign	Certificate in Data Science	Statistics
University of Iowa	B.S. in Data Science	Statistics & Actuarial Science
University of Maryland - College Park	B.S. in Computer Science, Data Science Specialization	Computer Science
University of Washington - Seattle	B.S. in Applied & Computational Mathematical Sciences, Data Sciences & Statistics Track	Applied & Computational Mathematical Sciences
University of Washington - Seattle	B.S. in Computer Science, Data Science Option	Computer Science & Engineering
University of Washington - Seattle	B.S. in Human Centered Design & Engineering, Data Science Option	Human Centered Design & Engineering
University of Washington - Seattle	B.S. in Informatics, Data Science Track	Information
University of Washington - Seattle	B.S. in Statistics, Data Science Option	Statistics
Washington State University	B.S. in Data Analytics	Computer Science & Engineering; Mathematics & Statistics