Data Science in Undergraduate Life Science Education: A Need for Instructor Skills Training

NATHAN C. EMERY, ERIKA CRISPO, SARAH R. SUPP, KAITLIN J. FARRELL, ANDREW J. KERKHOFF, ELLEN K. BLEDSOE, KELLY L. O'DONNELL, ANDREW C. MCCALL, AND MATTHEW E. AIELLO-LAMMENS

There is a clear demand for quantitative literacy in the life sciences, necessitating competent instructors in higher education. However, not all instructors are versed in data science skills or research-based teaching practices. We surveyed biological and environmental science instructors (n = 106) about the teaching of data science in higher education, identifying instructor needs and illuminating barriers to instruction. Our results indicate that instructors use, teach, and view data management, analysis, and visualization as important data science skills. Coding, modeling, and reproducibility were less valued by the instructors, although this differed according to institution type and career stage. The greatest barriers were instructor and student background and space in the curriculum. The instructors were most interested in training on how to teach coding and data analysis. Our study provides an important window into how data science is taught in higher education biology programs and how we can best move forward to empower instructors across disciplines.

Keywords: biology, environmental science, data science, pedagogy, higher education

The natural and social science fields are increasingly using data science approaches to answer important questions (Marx 2013). Technological advances that allow for the acquisition of increasingly large amounts of data are becoming common across disciplines such as ecology (Michener and Jones 2012, Hampton et al. 2013), wildlife biology (Lewis et al. 2018), evolutionary biology (Muñoz and Price 2019), environmental science (Gibert et al. 2018), genomics (Stephens et al. 2015), and neurobiology (Dierick and Gabbiani 2015). Data science is inherently interdisciplinary (De Veaux et al. 2017), and data science skills are valuable for students to learn before graduating from colleges and universities (Johnson 2018, National Academies of Sciences, Engineering, and Medicine 2018). Providing quality data science instruction for undergraduates can have numerous benefits for students' careers and for society. As examples, workers with data science competencies are needed to perform the phylogenetic surveillance necessary to track pandemics (Hodcroft et al. 2021) and to help alleviate the reproducibility crisis associated in part with an increasing abundance of data (Peng 2015, Lewis et al. 2018). A 2016 survey of National Science Foundation–funded principal investigators revealed that there is an unmet need with respect to a skilled workforce capable of handling the vast quantities of data produced in the sciences (Barone et al. 2017). Teaching undergraduates data science, especially across disciplines, can better prepare them for careers in a data-rich world.

The push for more data science instruction across disciplines, and in biology in particular, has led to two strategies. Some instructors use the approach of introducing problems in biology to computer science students (LeBlanc and Dyer 2004, Berger-Wolf et al. 2018, Oesper and Vostinar 2020), whereas other instructors have incorporated computational skills into biology courses (Madlung 2018, Wilson Sayres et al. 2018, Wright et al. 2020). Although both strategies are beneficial for teaching applicable data science concepts and skills, bringing life science content into computer science curricula should not replace quantitative instruction in life science courses. If, for example, data science skills were primarily offered in computer science courses, life science students would have to opt in to these courses to learn data science skills, possibly exacerbating self-selection bias and reduced retention of underrepresented groups in the sciences (Stephenson et al. 2018). Furthermore, integration of important learning concepts across several courses helps to decrease knowledge compartmentalization students tend to do, where they

BioScience 71: 1274–1287.
https://doi.org/10.1093/biosci/biab107 Advance Access publication 27 October 2021

have difficulty applying knowledge gained in one course to another course or situation (Ambrose et al. 2010).

Although there are promising curricular innovations for computer science programs (Sahami et al. 2010, Karbasian and Johri 2020), there can be advantages to incorporating data science instruction into disciplinary courses in the life and environmental sciences. Directly integrating data science skills into biology coursework can benefit students by improving overall quantitative literacy, providing new ways to explore biological concepts, repeating the use of skills throughout the curriculum, highlighting the applicability of data science skills to biology, and building important workforce skill sets. By embedding hands-on data science practices into biology and environmental science curricula, students learn important skills that will carry into their careers in fields that desperately need data-savvy biologists (Rubinstein and Chor 2014, Barone et al. 2017, Mariano et al. 2019, Robeva et al. 2020). Quantitative literacy is identified within the core concepts for biological literacy of the AAAS *Vision and Change* call to action for undergraduate biology reform (Brewer and Smith 2011). Specifically, computational tools were identified as an important component within the core concept of the interconnectedness of living systems, and computational tools for modeling and simulation were identified as core competencies.

Critical challenges for transforming higher education courses in any discipline are instructor training and access to resources (Brownell and Tanner 2012). Although broad barriers to pedagogical change are well established, less is known about discipline-specific barriers to incorporating data science skills into undergraduate biology and environmental science courses. Key barriers might include the perceived lack of space in curricula to sufficiently teach computing skills while simultaneously teaching biology content (Guzman et al. 2019) and a lack of teaching resources (Strasser and Hampton 2012). Known barriers to integrating bioinformatics skills into biology courses include instructor training, curricular space, and a perceived lack of student interest or preparation (Williams et al. 2019). Despite the recognized importance of teaching data science skills early and often in disciplinary curricula (Wilson Sayres et al. 2018, Wright et al. 2020), the practice is still not widespread across institutions.

To more fully integrate data science skills in undergraduate biology and environmental science curricula, instructors should be trained on how to both use and teach modern computational data science skills. Key data science skills for biology students and future researchers include data management, analysis, visualization, modeling, workflow reproducibility, and coding (Hernandez et al. 2012, Strasser and Hampton 2012, Hampton et al. 2013, National Academies of Sciences, Engineering, and Medicine 2018, Guzman et al. 2019), as well as being able to scale analyses for high-performance computing (Barone et al. 2017), write scripts, and use command line interfaces, version control, and high-performance computing clusters (Loman and Watson 2013). Because computational capabilities and corresponding skills change rapidly, many of these techniques may not have been emphasized when current biology and environmental science instructors were receiving their education, and, therefore, educators may need to upgrade their skill sets to provide up-to-date instruction. To help bridge this gap, a number of networks and consortia have been established to promote data science education in biology and environmental science fields, some with the explicit goal of training educators (supplemental file S1).

The Biological and Environmental Data Education (BEDE) Network is a community dedicated to providing professional development and training specific to undergraduate biology educators, with the goal of advancing confidence in data science skills and a framework for including these skills within current biology and environmental science curricula (https://qubeshub.org/community/groups/bede). The BEDE Network was formed on the basis of a recognition of the need to increase data science skills among undergraduate students, and a concern that instructors may not have received the training or support needed to navigate curricular changes. The network identified 17 participants across different institution types, career stages, and instructional perspectives, for the inaugural meeting, which convened in June 2019. The participants collaborated on a discussion to develop and refine a survey to assess the state of data science education in undergraduate biology and environmental science programs. The main goal of the survey was to identify critical gaps in current undergraduate data science education, to identify potential barriers to incorporating data science skills in courses, and to prioritize instructional training needs. The survey tool would also be used as an initial step to fostering growth of the biology and data science education-focused BEDE Network, and to build stronger support and training structures for undergraduate instructors.

To identify specific needs for instructor training in data science pedagogy across biology and environmental science courses, we designed and implemented a survey for undergraduate educators. Our objectives were to assess which data science skills are perceived by instructors as important for undergraduates to learn, how frequently instructors are both teaching different data science skills and using them outside the classroom, which perceived barriers exist for teaching data science skills, and what training instructors feel would better prepare them to teach data science skills. The survey also collected demographic and institutional data to assess where educator-training initiatives are most crucially needed. We hypothesized that teaching and use of data science skills would differ across institution types, because more research-oriented institutions may provide different pedagogical models or opportunities to access data science skills than baccalaureate or teaching-oriented institutions. We also predicted that instructors are likely to value—and, therefore, teach—data science skills that they use regularly in their own research. Finally, because the use of data science approaches has recently and rapidly expanded in biological

disciplines, we predicted that instructors who more recently earned their terminal degree may be more likely to value, use, and teach data science skills to their own students. The ultimate goal of this study was to identify areas of need for data science skills and pedagogical training for undergraduate biology instructors.

Survey development and distribution

To identify the main opportunities and obstacles for integrating data science into undergraduate biology and environmental science curricula, we developed a survey to assess the attitudes, interests, and expertise of instructors who teach undergraduates in biology and environmental science from a wide range of institutions. The survey was collaboratively developed by coprincipal investigators of the project, discussed at the BEDE Network group meeting held at Denison University in June 2019 and further refined on the basis of that feedback. Specifically, we queried instructors about six fundamental data science skills: data management, data analysis, modeling, writing code, data visualization, and reproducible workflows. These areas were chosen on the basis of the authors' experiences and recent literature, including curricular recommendations made by the National Academies of Sciences, Engineering, and Medicine (Loman and Watson 2013, Barone et al. 2017, Hampton et al. 2017, National Academies of Sciences, Engineering, and Medicine 2018). Several questions were used to assess how each skill fits into the respondent's institutional curriculum, how instructors perceive the importance of data science skills for their undergraduate students, and to assess each respondent's own pedagogical approach. Several additional questions were addressed, including how frequently instructors use each of the data science skills in their own research, perceived barriers to teaching these skills to undergraduates, and instructor interest in pedagogical training in the data science skills. The questions were structured on a five-point Likert scale, ranked options, or a select-one basis. The full survey can be found in supplemental file S2.

A final set of questions gathered the instructors' characteristics, including their academic appointment type, their highest degree earned, year in which their highest degree was earned, their racial identity, their ethnicity, and their gender identity. For each participant we asked for information about their current institution including its Carnegie classification, total student body size, and whether it is a minority-serving institution. No personal identifying information was gathered, and the survey was given exempt status by institutional review boards at Kenyon College (IRB 20,190,024), Denison University, and Pace University.

The survey was created using Qualtrics, and the data were securely hosted and housed on Kenyon College servers. An invitation to participate and a link to the survey were emailed to department chairs in the biological and environmental sciences at 536 US colleges and universities, determined on the basis of the institutions included in the *US News and World Report* lists of national universities and liberal arts colleges. This list included baccalaureate, master's, and doctoral granting institutions. In addition, the survey was shared on social media, shared within the personal contact networks of the project members, and survey recipients were encouraged to share the survey invitation with interested colleagues. The initial survey email was sent out 8 October 2019, and responses were collected until 10 December 2019.

Data analysis: Survey processing and data preparation

We downloaded all of the survey responses and used R (R Core Team 2020) to conduct statistical analysis of the results and generate data visualizations. The data were filtered to only include complete submissions, where the participant finished the survey, resulting in 106 responses. Our survey analysis was divided into two broad categories of predicted differences in data science instruction among institutions and among instructor characteristics. Before each statistical comparison, we assessed the number of responses for each of our three institutional characteristics (Carnegie classification, total student body size, and minority-serving institution), and each of our four instructor characteristics (academic appointment type, year of terminal degree completion, gender identity, and racial or ethnic identity); we omitted responses from categories that had fewer than five respondents or that included "I don't know" or "prefer not to answer." All of the code and analyses are available in a publicly archived repository (https://doi.org/10.5281/zenodo.4898117).

For the analyses comparing institutional differences, we examined Carnegie classification (four categories: associate's college, baccalaureate college, master's college or university, doctoral university) and total student body size (undergraduate and graduate, fewer than 5000, between 5000 and 15,000, and over 15,000). For the analyses comparing instructor characteristics, we examined academic appointment type (four categories: full-time staff, tenure-track faculty, tenured faculty, and visiting, temporary, or adjunct faculty) and the number of years since their highest degree earned. For analyses based on academic appointment type, we excluded responses from faculty with appointment types that were too rare for comparison, including one each of the following five appointment types: part-time staff, postdoctoral fellow, professor emerita, teaching assistant, and teaching-track faculty. For visualization and analytical purposes, the respondents' use of different data skills were grouped. We used *frequently* to represent "daily use" and "once to twice per week," *often* to represent "once to twice per month," and *rarely* to represent "once to twice per term" and "less than once per term."

Data analysis: Analytical comparisons of instructor and institution characteristics

The statistical test that we used to compare across our instructor and institution characteristics differed depending

on the response and predictor variable type. Several response variables were included in analyses comparing data science education metrics among institution types and instructor demographics. The first set included the ranked perceived importance of each of the six data science skills (on a scale of 1 to 5, from *not important at all* to *extremely important*), which we treated as a categorical variable. Separate analyses were conducted for each of the six data science skills: data management, data analysis, modeling, coding, data visualization, and reproducibility. For each skill, we performed a χ^2 test for independence between the importance levels assigned and the institution, or instructor, characteristic. The second set of response variables included the intention to teach each of the six data science skills. For each of the six data science skills, the respondents chose among four responses: "I don't teach or intend to teach this," "I intend to teach this," "I teach this," or "I want to teach this but don't know how." For each skill, we performed a χ^2 test for independence to compare response tallies among the levels of the predictor variables.

We additionally evaluated where students were most likely to learn data science skills, comparing these responses among institution types. We then evaluated perceived barriers (as ranks) to teaching data science skills and ranked interest in receiving training in each of the six data science skills, comparing these responses among institution types and among instructor demographic groups. For these analyses, we performed Friedman tests separately across the levels of each of the institution type variables, followed by paired Wilcoxon tests to investigate pairwise differences in rankings among the sources for learning data science or barriers to teaching data science.

Finally, we used Kendall's rank correlation tests separately for each data science skill to examine whether there was a relationship between the number of years since the instructors had earned their highest degree and the ranked categorical response related to their own data science skill use and their importance ranking of each skill. For these analyses, we converted categorical values to their corresponding ranked ordinal values (i.e., integers 1 to 5 or 1 to 3, depending on the variable).

For each set of analyses described above, we adjusted *p* values using a false discovery rate correction (Benjamini and Hochberg 1995). For example, when performing six separate χ^2 tests for independence between perceived importance of each data science skill examined and the Carnegie classification of the survey participants' institution, a false discovery rate correction was applied to the six *p* values associated with those χ^2 tests. The exception to this approach was that Bonferonni correction was used in the Wilcoxon tests applied to ranking data.

Respondent characteristics

The survey respondents earned their highest degree in years ranging from 1968 to 2019, with the mean year being 2004 and the median year being 2006 (figure 1a). Many of the survey participants originated from the United States (85%), but several responses came from European instructors, and one response was from China (figure 1b). The majority of the respondents indicated that they were tenured faculty (46%), pretenure tenure-track faculty (25%), full-time staff (17%), or temporary, visiting, or adjunct faculty (7%; figure 1c). Approximately 1% of the respondents were part-time staff, postdoctoral researchers, teaching track faculty, professor emerita, or teaching assistants. The vast majority (94%) indicated that the highest degree they had earned was a PhD or equivalent degree, one respondent's highest degree earned was a BS or equivalent, one respondent's highest degree earned was a professional degree such as an MD, and four respondents chose not to answer (figure 1c).

Of the 106 respondents, 41% indicated that they taught at doctoral universities, 27% at baccalaureate colleges, and 25% at master's colleges and universities (8% of the respondents indicated they did not know their institutions' Carnegie classification; figure 1d). None of the respondents indicated they taught at associate's colleges. In addition, none of the respondents indicated that they taught at minority serving institutions, although 44% indicated that they did not know whether their institution was minority serving. The responses were fairly evenly distributed among institutions of different sizes, with 40% of the respondents indicating that their institutions had fewer than 5000 students, 33% were from institutions with 5000–15,000 students, and 26% were from institutions with more than 15,000 students (one respondent indicated they did not know their institution size; figure 1d). Carnegie classification and institution size were strongly related ($\chi^2(6) = 38.568$, df = 4, $p < .001$; see supplementary file S3), with doctoral institutions being larger, baccalaureate colleges smaller, and master's institutions a more evenly spread mixture of institution sizes.

The fields of study that best described the respondents' departmental affiliations included biology (66%), ecology or evolution (12%), environmental science or studies (5%), plant biology (5%), cell and molecular biology (4%). Approximately 1% of the respondents indicated that they work in entomology, chemistry, biochemistry, natural science, science, or science and mathematics departments.

The majority of the respondents identified as White (85%), and the other respondents identified as Asian (3%), Black or African American (2%), and Alaska Native or Native American (1%); 9% of the respondents chose not to answer. Three percent of the respondents identified as Hispanic or Latino. Fifty percent of the respondents identified as men, 44% as women, and 2% as gender variant, nonconforming, or self-specified (and 4% chose not to answer).

Data science use, importance, and instruction

Understanding how instructors teach data science may depend on how familiar they are with using data science skills in their research and nonteaching activities. The most frequently used data science skills across categories were data analysis, data visualization, and data management

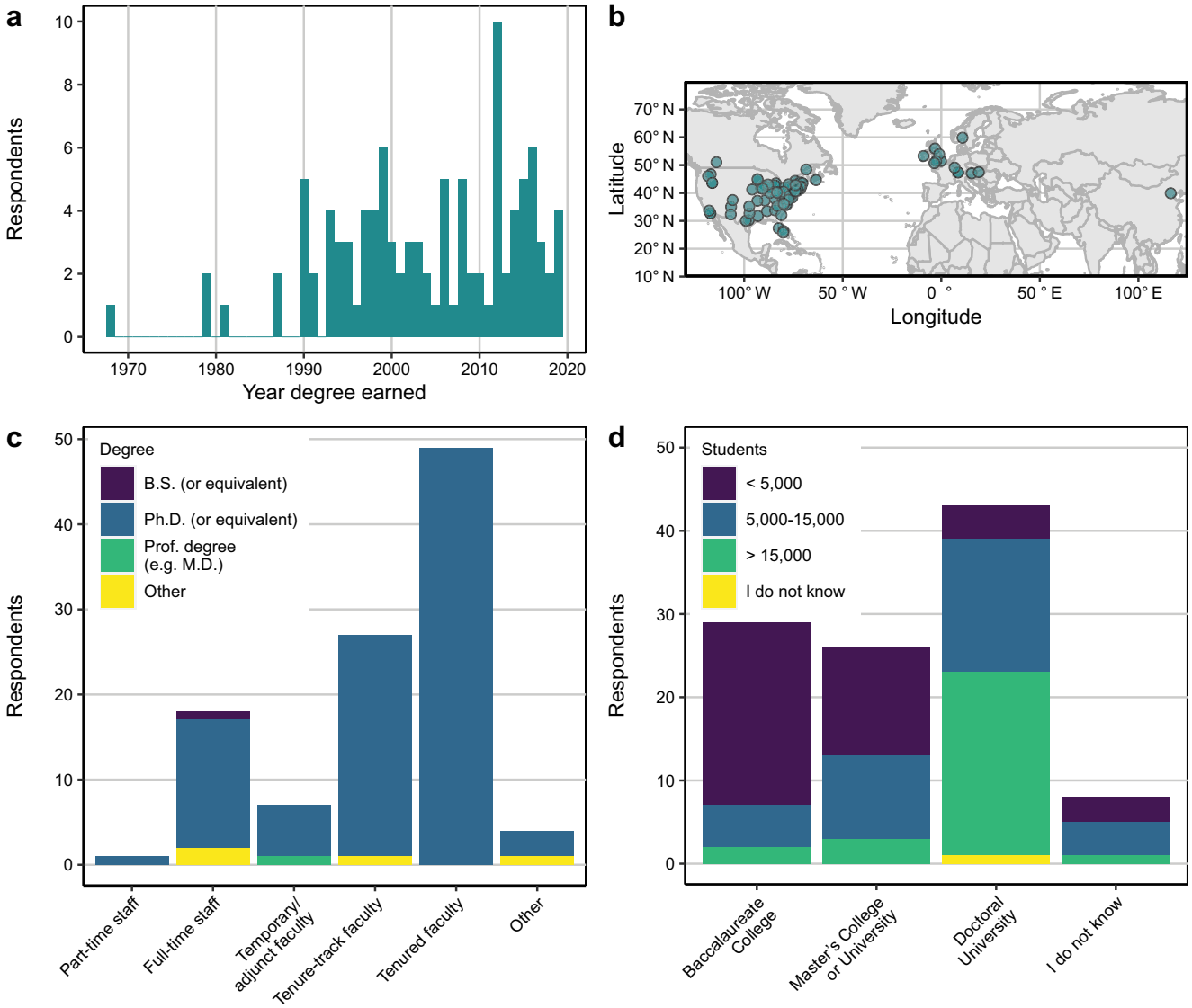

Figure 1. Summary of respondent characteristic data as a function of the year a degree was earned (a), location (b), appointment and degree type (c), and institution type and size (d).

(figure 2). Coding use differed by institution type with doctoral institutions having the highest frequency of use (44% daily or once or twice per week of use), followed by baccalaureate (25% daily or once or twice per week of use) and master's institutions (4% daily or once or twice per week of use). This pattern of coding use was mirrored across institution sizes, with instructors at larger institutions using code more frequently than midsize and smaller institutions (supplemental file S4.2). For both Carnegie classification and institution size, the interaction with coding use was significant on the basis of chi-squared analyses ($\chi^2(6) = 14.91$ and 15.11, and $p = .017$ and .027, respectively; see supplemental files S4.1 and S4.3). In general, instructors from doctoral and large institutions reported a significantly higher frequency of data science skills use than other institution types and sizes (the chi-squared analyses are presented in supplemental files S41 and S4.3). Modeling and reproducibility were used infrequently by all instructors (18% daily or once or twice per week of use; supplemental file S4.4). Interestingly, there was no indication that some data science skills are more frequently used than others by instructors on the basis of institution types. For example, relative to instructors at baccalaureate colleges and master's college or universities, doctoral university instructors score the use of reproducibility and modeling similarly low. There were no significant interactions between appointment types and data science skill use, and use of specific skills was consistent with the institutional level findings (supplemental file S4.5).

A further examination of data science skill use as a function of year of terminal degree suggest that early career instructors tend to use more code, reproducibility, and data visualization than more senior instructors (figure 3). The

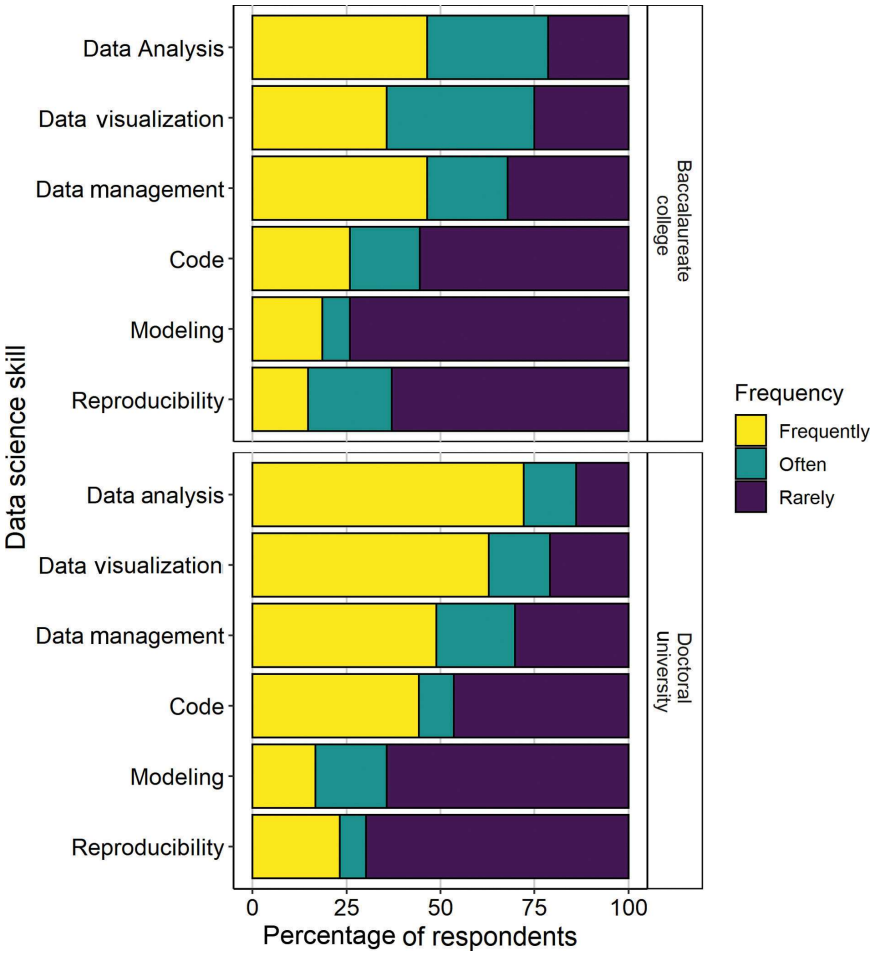

Figure 2. Frequency of use of data science skills by instructors in their research and nonteaching activities across institution types. The responses were grouped as* frequently *for "daily use" and "once to twice per week,"* often *for "once to twice per month," and* rarely *for "once to twice per term" and "less than once per term."

median year of terminal degree for frequent use for code, reproducibility, and data visualization were 2012, 2010, and 2006, respectively. The instructors who rarely use code, reproducibility, and data visualization, tended to be more senior, with median years of terminal degree of 1999, 2002, and 2001, respectively. However, on the basis of Kendall's rank correlation tests carried out for each skill separately, where we calculated Kendall's tau correlation coefficient between years since degree and rankings for frequency of use, significant negative correlations were found for code and reproducibility only ($\tau = -.327$, $p = .00003$, and $\tau = -.189$, $p = .0181$, respectively). Furthermore, correcting for multiple testing using false discovery rate, only effects for code remained significant (supplemental file S4.6).

Overall, the instructors perceived data analysis (91%), data visualization (87%), and data management (62%) as being either extremely or very important (figure 4). None of the respondents perceived data analysis and visualization to be unimportant at all. Greater than 50% of the instructors perceived code, modeling, and reproducibility as moderately important to not at all important. This pattern of the relative importance of data analysis, management, and visualization was consistent across institution types, sizes, and instructor appointment type (all χ^2 tests for independence were nonsignificant; supplemental files S5.2, S5.4, and S5.6).

An analysis of the rankings of where students learn their data science skills suggested significant differences among the ranks for each of the Carnegie classifications considered (on the basis of Friedman's test results; supplemental file S6.1). Furthermore, analysis of pairwise differences in rankings between learning locations showed variation between classification levels as well. For doctoral and master's granting institutions, students are more likely to learn data science skills in a required course, in an elective course, or in a course in another department, and unlikely learn these skills beyond the institution (figure 5, supplemental file S6.2). However, in baccalaureate colleges, students were about as likely to learn data science skills outside of coursework as compared with learning in courses offered by their institution (supplemental file S6.2). When comparing across institution size, again there were significant differences among rankings for the learning locations. However, further pairwise difference analysis suggests that any significant effects of learning location are driven primarily by the high ranking score (i.e., low likelihood score) of students learning data science skills external to the institution (supplemental file S6.5).

Over half of all instructors teach or intend to teach data analysis (87%), visualization (77%), and management (58%; figure 6). Code and modeling were similar and less likely to be taught or intended to be taught in undergraduate courses (47% teach or intend to teach code, 42% modeling). Many of the instructors reported not teaching or intending to teach reproducibility to students (48%). This pattern was consistent across institution types, sizes, and instructor appointment type, as was reflected in the nonsignificant χ^2 tests of independence between teaching intentions and institution characteristics (supplemental files S7.2, S7.4, and S7.6). With regards to who is teaching data science skills, especially for code, reproducibility, and data management, it appears that these skills are more likely to be taught by early career instructors compared with more senior instructors (figure 7), however we only found significant negative

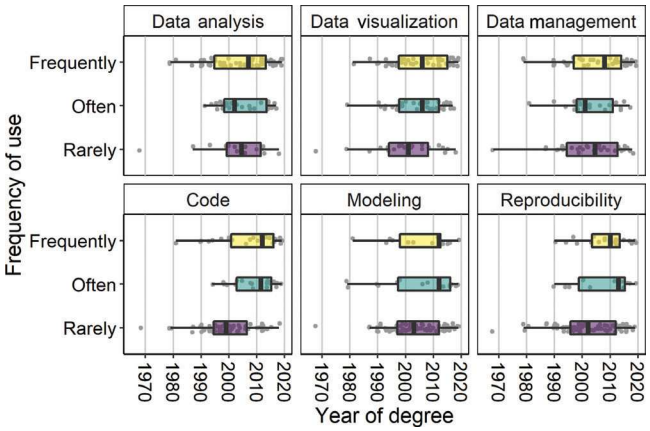

Figure 3. The frequency of use of data science skills by the year of degree of the respondent.* Frequently *represents "daily use" and "once to twice per week,"* often *represents "once to twice per month," and* rarely *represents "once to twice per term" and "less than once per term." The boxplots represent the median, first and third quartiles, and 1.5 times the interquartile range; the points indicate individual responses.

correlations between years since degree and level of teaching intention for code and reproducibility ($\tau = -.152$, $p = .0445$ and $\tau = -.246$, $p = .0013$, respectively), with only the correlation with reproducibility remaining significant following false discovery rate correction (supplemental file S7.7). The median years of terminal degree for instructors who teach or intend to teach code, reproducibility, and data management were 2008.5, 2011, and 2008, respectively. The median years of terminal degree for instructors who don't teach or intend to teach code, reproducibility, and data management were 1999, 1999, and 1999.5, respectively.

Across the survey participants and data science topics, the instructors used a variety of sources for their teaching materials. As no distinguishable patterns existed within a data science skill, we pooled together all data science skills (Data management, data analysis, etc.) and screened out the participants who responded that they don't teach or intend to teach a given data science skill. There were similar proportions of the instructors that used each source of teaching material across institution types ($\chi^2(6) = 9.297$, df = 6, $p = .16$; figure 8). The majority of the instructors, regardless of institution type, used their own materials (32%) or open source online materials (57%) for teaching data science in undergraduate courses. The sources of teaching materials did differ significantly when considering institution size ($\chi^2(6) = 25.412$, $p = .0003$), and approached significance when considering appointment types ($\chi^2(9) = 15.627$, $p = .075$). Compared with all instructors combined, the instructors at large institutions were slightly more likely to use materials developed at their institutions or that they themselves had developed versus the instructors at small institutions, who were more likely to use open source online and proprietary materials (supplemental file S8.1). Considering appointment type, the full-time staff and temporary or adjunct faculty were more likely to use proprietary materials (supplemental file S8.2).

Barriers to data science integration

Overall, the three biggest barriers to integrating data science skills into the curriculum were identified as a lack of instructor and student background in necessary skills and knowledge and space in the curriculum (table 1). Student background appears to be a bigger barrier at baccalaureate and master's colleges or universities compared with doctoral universities. Overall, the instructors across institution types tended to rank the lack of support and the lack of access to resources as relatively low barriers to teaching data science in undergraduate classrooms (supplemental file S9.3). Interestingly, when considering either Carnegie classifications or institution sizes, Friedman's tests suggest that there are significant differences among the rankings of the barriers, consistent with the above observations, but paired Wilcoxon tests for pairwise comparisons show very few differences among the barrier rankings (supplemental files S9.1, S9.2, S9.5, and S9.6). Considering the instructor appointment types, there were only significant differences in the rankings of barriers for tenured faculty, who ranked instructor background in necessary skills as the biggest barrier.

The respondents had the option of writing any additional barriers to data science integration that we did not include as response options in our survey. Some of these barriers were related to the ranked choices. In particular, the respondents mentioned barriers related to a lack of institutional and departmental support, such as a lack of incentives for course innovation, a resistance to change, or a lack of cooperation among departments and with the administration. Many of the respondents mentioned the lack of time instructors are given to innovate in their courses and plan new curricular activities involving data science, and some mentioned the balance between teaching course content and skills. Two respondents indicated that their colleagues' perceptions of and willingness to teach data science might often be a barrier. Although several of the respondents indicated that others (e.g., students, colleagues and administrators) felt that data science skills were unimportant for students to learn, only one respondent indicated that it should not be a priority for undergraduate coursework.

Student-specific attributes were also identified as barriers. One respondent indicated that although learning data science is important regardless of what students end up doing in the future, helping students identify the applicability of data science was a challenge. Another respondent identified a lack of student patience or grit as a barrier that might additionally be related to students' shortfall in necessary background skills. Several of the respondents indicated that a lack of student access to equipment and technology was a barrier, and one respondent specifically mentioned inequity, although the respondent did not provide details about the inequity.

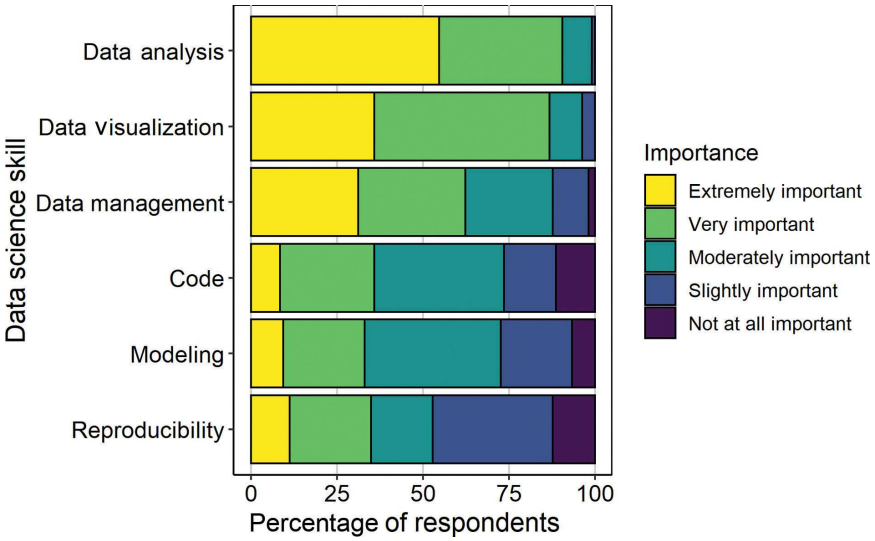

Figure 4. Instructors' perceived importance ratings, as percentages, of students learning data science skills in undergraduate courses.

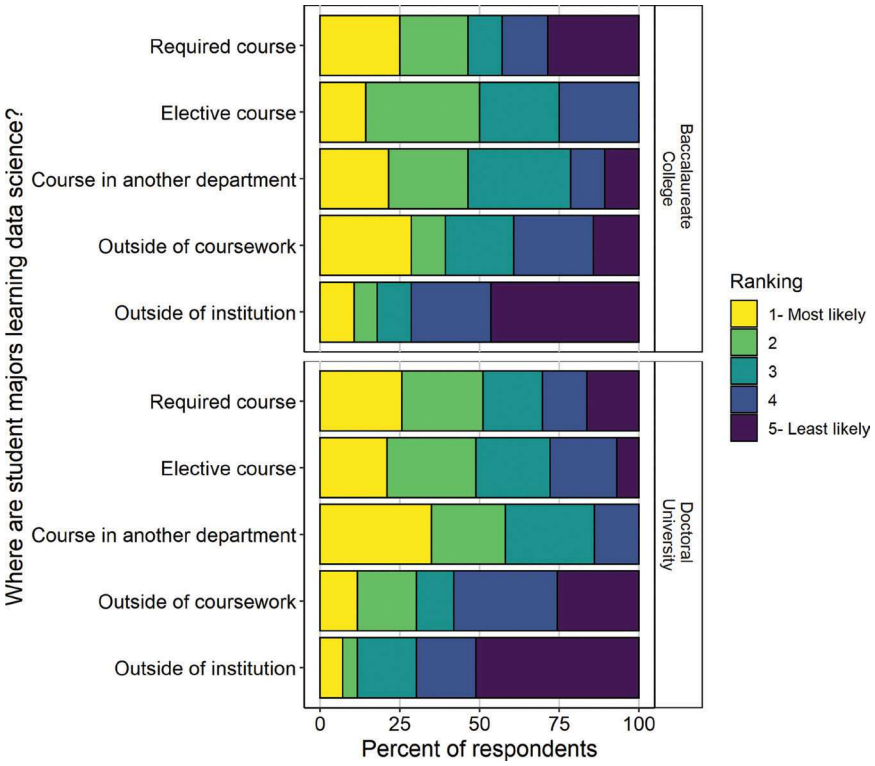

*Figure 5. The percentage of instructors who ranked where undergraduate majors are most likely to learn data science skills on the basis of Carnegie classification. The ranking ranges from 1 (***most likely***) to 5 (***least likely***).*

One respondent noted that it is difficult to communicate the importance of a skill when it is not related to the course content, and another respondent indicated that data science skills lack application toward degree programs, which are difficult to categorize as barriers at the student, instructor, or institutional levels. One respondent indicated that barriers are expected to differ among subfields, and another indicated that barriers are likely to differ among data science skills. Interestingly, one respondent indicated that data science education occurring across the university was a barrier to data science integration in the major, perhaps because of a lack of cohesion and consistency in pedagogy.

Interest in data science training

In general, the greatest interest in data science skill training was in data analysis, coding, and data visualization. Of lowest interest was reproducibility (figure 9). This pattern was similar across institution types, and sizes, where rankings for interest in reproducibility were significantly lower than several other data science skills (supplemental files S10.1, S10.4, and S10.6). Instructors at master's institutions ranked modeling relatively high, and instructors at doctoral institutions ranked data management relatively high, these patterns were not strongly supported by the Wilcoxon pairwise analyses (supplemental file S10.3). Among faculty appointment types, tenured (n = 49) and tenure-track faculty (n = 27) ranked the data science skills similarly, whereas full time staff (n = 18) were interested in data management training and temporary faculty (n = 8) had less interest in receiving training for data visualization than all other instructors (supplemental file S10.7). However, differences in rankings were only significant for tenured instructors, who scored interest in reproducibility as low (supplemental files S10.8 and S10.9).

Outside of the survey options for interest in data science training, instructors described their preferred mode of training. The respondents were fairly split on how they prefer to be trained in data science teaching with 45% preferring a self-guided tutorial, 20% preferring a webinar format, and 35% preferring a workshop setting, ideally in person. Several of the respondents selected more than one preferred mode of training.

Potential gaps in instructor training were identified by comparing the perceived importance of a data skill with how often it was reported as being taught by instructors. Of note were gaps in data visualization, data management, and

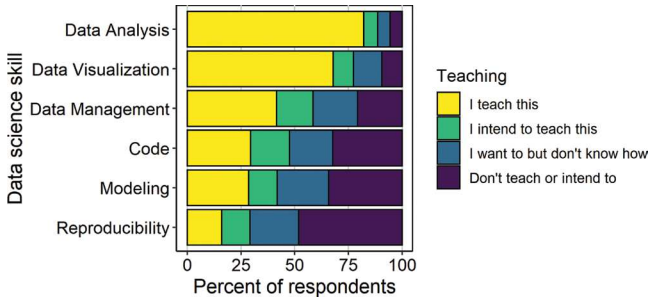

Figure 6. The percentage of all instructors who teach various data science skills at their institution.

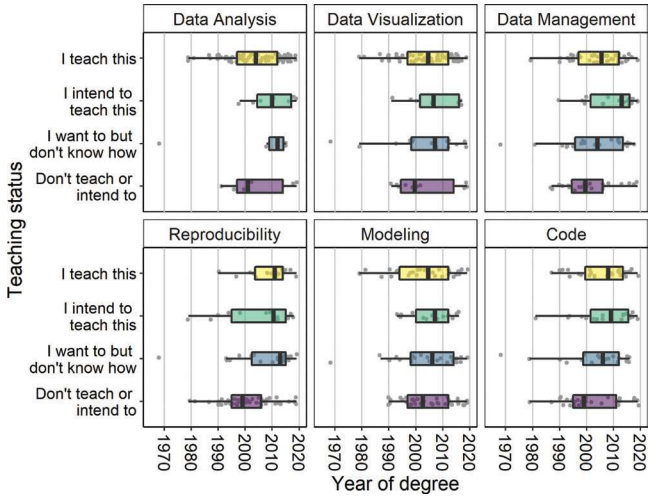

Figure 7. Instructors' teaching status as a function of year of degree for a given data science skill. The boxplots represent the median, first and third quartiles, and 1.5 times the interquartile range.

reproducibility, with a greater percentage of instructors valuing each skill as extremely or very important compared with how frequently it was reported taught by instructors (figure 10).

Teaching and use of data science by life science instructors

Our assessment of the state of data science education for biology and environmental science instructors confirmed the findings of previous studies (Strasser and Hampton 2012, Hampton et al. 2017, Wilson Sayres et al. 2018, Williams et al. 2019) while providing details about the relative importance of data science skills, trends in using and teaching data science, barriers to entry, and reception to training opportunities. Data management, analysis, and visualization were consistently highlighted as important for students to learn (figure 4), commonly taught by instructors (figure 6), and used by instructors frequently outside of teaching (figure 2). Relative to these skills, coding and reproducibility were not viewed as important for students, although there was a distinct gap in the time since degree for instructors that frequently teach these data skills and use them outside of teaching (figures 3 and 7). Across all of the respondents and institutions, the instructors tended to use their own materials or open source materials for teaching data science in undergraduate classrooms (figure 8), highlighting the importance of online resources and regularly maintained open educational resources. The main perceived barriers to teaching data science were instructors and students lacking background knowledge, and insufficient space in the curriculum (table 1), consistent with previous findings (Tenopir et al. 2016, Williams et al. 2019). Instructors were primarily interested in receiving training in data analysis, visualization, and coding (figure 9), although full-time staff were also very interested in data management (supplemental file S10.10). Our findings provide valuable insight on the data science skills that are valued and taught across institutions as well as informing future professional development initiatives.

Across institution types, student majors were very likely to learn data science in a required (median rank = 2) or elective course (median rank = 2) offered by their institution (figure 5). In contrast, students were still often likely to learn data science skills mainly through courses in other departments (median rank = 3) or in other avenues outside of courses (median rank = 3) or outside of their institution (median rank = 4). These "outside" experiences may include courses or tutorials not hosted by the institution, or skills learned through research or internship experience, but our question did not cover that level of detail. That at least half of the respondents indicated a high likelihood for students to learn skills in required or elective courses is encouraging as it suggests that instructors managing undergraduate biology and environmental science programs acknowledge the importance of data science skills in life science curricula (Madlung 2018, Wilson Sayres et al. 2018, Wright et al. 2020). However, our findings suggest that data science skills are not taught equally across courses, with perhaps greater emphasis on data analysis, management and visualization, and relatively less emphasis on modeling, code, and reproducibility (figure 6). Only in baccalaureate institutions did the instructors report that students were nearly as likely to learn data science skills outside of coursework (noncourse median rank = 3, beyond institution median rank = 4) as in their home department (required course median rank = 3, elective course median rank = 2.5). This result could possibly be explained by reduced course offerings at institutions without graduate programs, because there was no pattern across institution sizes (supplemental file S6.3). Although the results are encouraging, there is still room for institutional improvement, because instructors face numerous barriers to integrating data science skills into life science courses.

Instructor and student background as barriers to teaching data science

In our study, we found that the respondents perceived instructor background, student background, and the lack of space in curricula to be the largest barriers to integrating data science into undergraduate courses (table 1). This result is consistent with previous studies on teaching

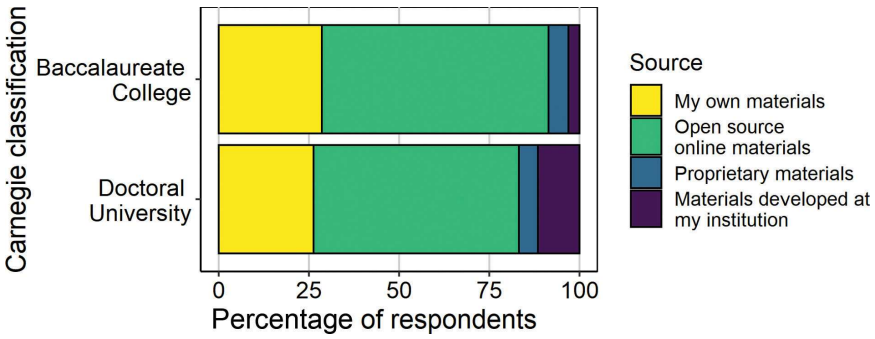

Figure 8. The source of teaching materials across all data science skills for instructors who teach or want to teach data science in undergraduate courses. The data are plotted by the percentage of instructors who use a given source of teaching materials split up by their institution type.

bioinformatics (Williams et al. 2019) and data management (Tenopir et al. 2016). Student background was slightly less often perceived as a barrier at doctoral institutions (supplemental file S9.3), which could be an artifact of which courses were taught by the survey participants from those schools or a reflection of the breadth of course offerings at schools with graduate programs. Although finding space in the curricula is a commonly identified barrier (Guzman et al. 2019, Williams et al. 2019) and would require structural change to address, training in data science skills and evidence-based pedagogy could surmount the barrier of instructor background. Instructor self-efficacy might increase with gaining adequate background knowledge to teach basic data science skills, and to reach students who may not have previous exposure to data science skills, alleviating the barrier of insufficient student background.

As instructors gain confidence in their abilities, they may feel empowered to integrate data science into multiple courses within a department. Encountering data science skills across multiple courses is ideal for students, because learning spread over long time periods (i.e., scaffolded across multiple semesters or courses) is a tactic previously shown to increase retention (Rohrer 2015). Our survey did not inquire about the extent of teaching data science across courses, and it is unclear how frequently students need exposure to data science skills to improve student learning outcomes. There are likely multiple pathways to improving data science education for undergraduates majoring in the life sciences (Robeva et al. 2020). Implementing data science skills across early to advanced course levels, as opposed to within a single course or outsourced to a suite of computer science courses, is most likely to improve overall student learning outcomes (Ambrose et al. 2010).

Perceived importance is linked to the use of data science by instructors

Data science skills are increasingly recognized as important for students to realize careers in the life sciences (Hampton et al. 2013, Barone et al. 2017, Gibert et al. 2018, National Academies of Sciences, Engineering, and Medicine 2018). Our study found that data analysis, management, and visualization were predominantly taught by instructors across institutions (figure 6, supplemental file S7.1). Perhaps unsurprisingly, these three skills were also perceived to be the most important for students to learn (figure 4). This result is concurrent with previous work that also stressed the importance of data management (Strasser and Hampton 2012), analysis, and visualization (Hampton et al. 2017) for undergraduate education. This congruence belies the importance of these skills across disciplines and potentially the ease of incorporating these skills into undergraduate curricula. Instructors reported frequent use of data management, analysis, and visualization outside of their teaching obligations (figure 2), and data analysis and visualization ranked highly as desired further training areas (figure 9). If instructors are often using these skills, it follows that they may also be comfortable teaching these skills in their courses.

Coding, although taught or intended to be taught by almost half of the instructors surveyed (figure 6), was not frequently perceived to be an important skill for students to learn despite the importance placed on data analysis skills (figure 4). A possible explanation is that coding may be a skill that has been more recently emphasized in doctoral training (Hernandez et al. 2012), and modern coding software or tools may not be as accessible to more senior instructors or to instructors at nondoctoral institutions. A greater proportion of instructors at baccalaureate and master's institutions (Rarely = 60%) reported infrequent use of coding compared with instructors from doctoral institutions (rarely, 47%; figure 2) and coding was used less frequently and less likely to be taught by more instructors further from the time of degree (figure 3). Incorporating coding into courses may be novel for instructors, many of whom are challenged by a lack of methods and tools for teaching code (Medeiros et al. 2019). Coding was most frequently prioritized as a skill where training was desired (figure 9), which may indicate that biology instructors recognize its importance and potential to benefit biological research, even if they are not yet convinced that it should be prioritized as a skill for undergraduate students. But increasingly, calls are being made to include computational literacy and code as fundamental skills for biology undergraduates (Mariano et al. 2019, Auker and Barthelmess 2020), a necessary skill for working in a research lab or in careers in biology—a call that has not yet caught up to biology educational practice (Robeva et al. 2020).

Modeling was another data science skill that was not commonly taught (41% of the respondents teach or intend to teach; figure 6), was ranked low in perceived importance for undergraduate students (figure 4), and for which

Table 1. Mean ranks for perceived barriers to teaching data science skills, according to institutional Carnegie classification.

	Instructors lack the necessary background		Students lack the necessary background		A lack of space in the curriculum		A lack of student interest in learning data science		A lack of instutional and departmental support		The instructors lack access to resources	
Institution	Mean rank (M)	Standard deviation (SD)	M	SD	M	SD	M	SD	M	SD	M	SD
All institutions	**3.06**	1.59	3.07	1.68	3.28	1.70	3.77	1.80	3.80	1.68	4.03	1.57
Baccalaureate colleges	3.21	1.42	**2.76**	1.77	3.07	1.75	3.76	1.77	4.28	1.49	3.93	1.69
Master's colleges and universities	3.20	1.63	**2.48**	1.69	3.60	1.63	3.88	1.76	4.08	1.73	3.76	1.48
Doctoral universities	**2.81**	1.64	3.73	1.50	3.05	1.68	3.83	1.90	3.43	1.67	4.14	1.57

Note: Ranks ranged from 1 (*biggest barrier*) to 6 (*smallest barrier*). n = 98. The values in bold represent the biggest barrier within an institution type.

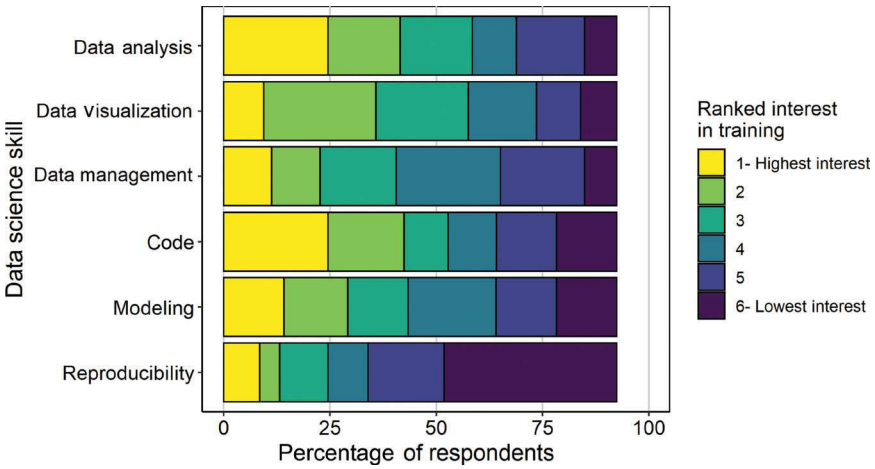

Figure 9. Ranked instructor interest in receiving training among data science skills (1,* highest interest*; 6,* lowest interest*).

few instructors prioritized a desire for continued training (figure 9). It is perhaps unsurprising that modeling ranked similarly low compared with coding, since most modeling approaches require instructors and students to have at least a basic understanding and ability to use code, in addition to statistics. Despite a push for increased quantitative learning and requirements in biology departments, many programs still do not require undergraduate students to take advanced math or statistics courses, or if these courses are required, they may be outsourced to another department that may not use examples from the life sciences (Robeva et al. 2020). Our survey did not collect enough detail to determine why modeling was consistently ranked relatively lower in perceived importance than other data science skills, but the pattern persisted across different institution types (supplemental file S5.1). Our findings suggest that modeling is another area in which biology education practice may not be meeting biology research and career needs.

Scientists may be in the middle of a reproducibility crisis (Peng 2015), but instructors consistently downplayed the importance of teaching reproducibility in undergraduate courses. Nearly half of instructors responded that reproducibility was only slightly important or not at all important to teach (47%; figure 4) and that they did not intend to teach reproducibility (48%) in their undergraduate courses (figure 6). It's possible that instructors are unfamiliar with the concept and value of reproducibility as they rarely use it outside of teaching, with 71% stating they use it once or twice per term or less (figure 2), and it was the lowest ranked priority for desired future training (41% ranked as lowest training priority; figure 9). At postbaccalaureate levels, Hernandez and colleagues (2012) found that graduate advisers in environmental science disciplines emphasized the importance of reproducibility but were concerned that graduate students did not have the skills to follow reproducible workflows. Reproducibility is a critical concept to apply to the scientific process, which allows others to recreate research findings through code, analysis, visuals, or data. Through a data science lens, reproducibility can be implemented using tools such as version control (e.g., git and GitHub), but it can also entail documenting and sharing data, or sharing methodologies online (Sandve et al. 2013, White et al. 2013). Furthermore, reproducibility seems difficult to disentangle from ethical frameworks on how science works, that undergraduate students should be learning throughout their coursework. Our survey suggests that among the data science skills we targeted, reproducibility is the least valued in the classroom and by the instructors teaching them—potentially indicating an important missed

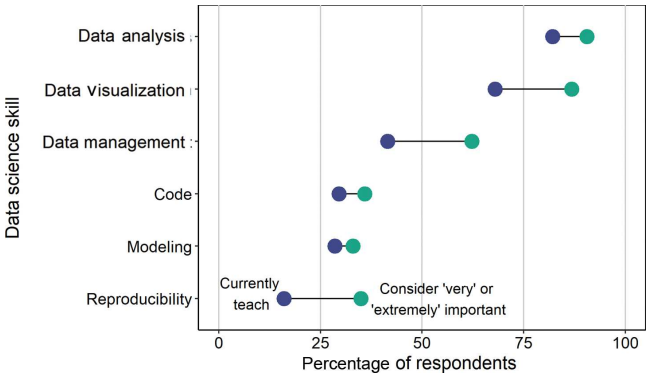

Figure 10. The gaps in data skill importance and whether it is taught by instructors. The teal points are the percentage of the respondents who said a particular skill was "extremely" or "very" important for undergraduate biology students to learn. The dark blue points are the percentage of the respondents who responded, "I teach this" for each skill. The distance between those may represent an aspirational gap in items acknowledged as important for students, but that aren't being taught.

opportunity. It is also possible that the wording of the survey item may have contributed to the respondents' interpretation of reproducibility. The skill was listed as reproducibility of computational workflows (e.g., version control) when reproducibility in data science is multifaceted. Future researchers could investigate whether our findings are particular to our sample, represent a misunderstanding of what reproducibility means, suggest a lack of knowledge of how to use reproducible processes and tools, or truly indicate that life science instructors do not see reproducibility as important for undergraduates to learn. Interestingly, there seemed to be a temporal trend in how familiar the respondents were in using reproducibility that indicates recent prior training may be important. Most of the respondents who used reproducibility in their own work frequently (n = 19; median degree year = 2010) or often (n = 12; median degree year = 2013) received their terminal degrees more recently than the respondents who used it rarely (n = 74; median degree year = 2002; figure 3).

Instructors seek training in data science skills

Although instructors recognize the importance of data science skills, they are not always prepared to incorporate these skills into their courses or teach them in the classroom. For many of the data science skills, notably data management, coding, modeling and reproducibility, approximately one-fifth of the respondents (22%, 21%, 24%, and 23%, respectively) indicated that they didn't know how they would teach such skills. A viable solution is to provide instructors with the appropriate tools and training, allowing them to adapt data science skills into their respective curricula and course materials. The instructors in our study expressed interest in training for a range of data science skills with the exception of reproducibility (figure 9). Coding and data analysis were the top choices for training and to be expected given the instructors' relative unfamiliarity with coding (figure 2), and the high proportion of instructors that teach data analysis (figure 6). It's evident that there are some data science skills that instructors view as important, but do not currently teach (figure 10). Training and resources for these skills have the potential to overcome one of the indicated barriers, the lack of instructor background, to teaching data science skills in the classroom (table 1).

Increased instructor training will require investments from both individuals and institutions to build confidence in the core skill sets, and a framework for implementing them in the classroom. The respondents were split on their stated preferred format of data science training, which included self-paced resources, recorded webinars, short-format workshops, guided peer instruction and in-person events. To reach the broadest audience, future efforts may need to provide multiple opportunities for instructors to learn data science skills, including shorter self-paced materials and longer-term in-person events or mentored guidance. Professional development programs that assist instructors will have the added benefit of using open source resources and helping instructors develop their own material, because the vast majority of the survey's participants (more than 75% across all skills) used these two sets of resources, as opposed to proprietary or institutionally developed materials (less than 15% across all skills), to teach data science skills in their courses (figure 8). Although there are many open source resources for creating course content and modules or labs that include data science skills (e.g., http://datanuggets.org), other organizations provide free or low cost opportunities for training (supplemental table S1), although many of these are not specific to undergraduate biology instructors or may be more focused on training for research purposes. Short-format workshops and self-guided materials can be an important stopgap in helping instructors learn and update skills, but they are often not enough (Henderson et al. 2011, Stes and Hoekstra 2015, Emery et al. 2020). The BEDE Network plans to use the results of this survey to target training modules to the identified curricular gaps, and undergraduate biology instructor training needs. A pilot workshop has already been developed and delivered (https://qubeshub.org/community/groups/bede), and BEDE plans to continue this work through ongoing workshops, faculty mentoring networks, curricular maps, and open training resources.

Limitations and future work

Of the 106 survey responses, instructors were almost evenly distributed across different Carnegie classifications (except for associate's colleges) and across institution sizes (from less than 5000 for small institutions to more than 15,000 students for the large institutions; figure 1d). Despite this even distribution, there was low racial and ethnic diversity among the

respondents, not unlike the demographics of a similar study in bioinformatics (Wilson Sayres et al. 2018). The majority of the respondents were also tenured or tenure-track faculty, potentially limiting the applicability of our results to other appointment types. Instructors came from a variety of life science departments, although the majority were biology based. Data science is taught in numerous disciplines, and conclusions drawn from our study results may disproportionately represent responses from instructors in general biology departments. As this was a self-reported survey, it is possible that our results do not completely capture the perspectives and reality of teaching data science in life science courses. More work is needed to understand the details of how data science is being taught within departments (e.g., modes of instruction, cross-listed courses, multiple entry points versus hierarchical linear structure) and which students and faculty have access to data science learning (are students accessing data science skills representative of prior social, economic, or educational privilege?).

Conclusions

Undergraduate students increasingly need exposure to data science skills to compete for modern careers in the life and environmental sciences or to prepare them for graduate study (Hernandez et al. 2012, Hampton et al. 2017, Robeva et al. 2020). Our survey results suggest that there are important differences in how frequently different data science skills are taught in undergraduate biology courses, indicating potentially critical gaps in student learning and preparation, and a missed opportunity to update the curriculum (Robeva et al. 2020). Although instructors do appear to value and be teaching skills such as data management, analysis, and visualization, it is concerning that coding, modeling, and reproducibility skills are not more frequently emphasized in undergraduate coursework, or are perceived to be of relatively low value. In reenvisioning biology and environmental undergraduate learning outcomes that include data science skills (Brewer and Smith 2011, Johnson 2018, National Academies of Sciences, Engineering, and Medicine 2018) instructors represent a key link in achieving educational goals. We acknowledge that addressing the multiple barriers to teaching data science is complicated as it likely requires institutions to free up instructor time, support continued training opportunities, and to recognize the importance of quantitative data science education across disciplines. Without this departmental and institutional level recognition, eager educators might be willing and able to effectively upgrade their pedagogical skills, but be stymied by financial or time-availability barriers or lack of support. Ultimately, external resources or organizations such as the BEDE Network may provide support for instructors who are interested in learning how to best integrate data science skills into life science courses. Such training initiatives can supplement institutional efforts and fill an important gap in instructional development for instructors around the world. Through this work, we can better and more rapidly align biology education practices with biology education recommendations and career needs.

Acknowledgments

The authors would like to thank fellow BEDE Network members for thoughtful discussions and supporting the mission of the network. Other workshop participants included Sam Donovan, Susy Escheverria-Lodoño (who also helped secure the funding), Linda Forester, Justin Kitzes, Joslyn Lee, Jen McMahon, Lina Yoo, and Ethan White. This article was possible thanks to the many survey participants who volunteered their perspectives on data science skills in the life sciences. Also, many thanks to Erika Farfan and Kenyon College staff, who helped coordinate the survey and IRB process, and to staff who facilitated the in-person BEDE Network June 2019 meeting at Denison University. Funding provided by Nation Science Foundation grant no. DUE-1820782.

Supplemental material

Supplemental data are available at *BIOSCI* online.

References cited

Ambrose S, Bridges M, Lovett M, DiPietro M, Norman M. 2010. How Learning Works: 7 Research-Based Principles for Smart Teaching. Jossey Bass.

Auker LA, Barthelmess EL. 2020. Teaching R in the undergraduate ecology classroom: Approaches, lessons learned, and recommendations. Ecosphere 11: e03060.

Barone L, Williams J, Micklos D. 2017. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLOS Computational Biology 13: e1005755.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57: 289–300.

Berger-Wolf T, Igic B, Taylor C, Sloan R, Poretsky R. 2018. A biology-themed introductory CS course at a large, diverse public university. Pages 233–238 in Barnes T, Garcia D, Hawthorne EK, Pérez-Quiñones MA, eds. SIGCSE '18: Proceedings of the 49th ACM Technical Symposium on Computer Science Education. Association for Computing Machinery.

Brewer CA, Smith D. 2011. Vision and Change in Undergraduate Biology Education: A Call to Action. American Association for the Advancement of Science.

Brownell SE, Tanner KD. 2012. Barriers to faculty pedagogical change: Lack of training, time, incentives, and… tensions with professional identity? CBE—Life Sciences Education 11: 339–346.

De Veaux RD, et al. 2017. Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application 4: 15–30.

Dierick HA, Gabbiani F. 2015. Drosophila neurobiology: No escape from "big data" science. Current Biology 25: R606–R608.

Emery NC, Maher JM, Ebert-May D. 2020. Early career faculty practice learner-centered teaching up to 9 years after postdoctoral professional development. Science Advances 6: eaba2091.

Gibert K, Horsburgh JS, Athanasiadis IN, Holmes G. 2018. Environmental data science. Environmental Modelling and Software 106: 4–12.

Guzman LM, Pennell MW, Nikelski E, Srivastava DS. 2019. Successful integration of data science in undergraduate biostatistics courses using cognitive load theory. CBE—Life Sciences Education 18: 49.

Hampton SE, et al. 2017. Skills and knowledge for data-intensive environmental research. BioScience 67: 546–557.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162.

Henderson C, Beach A, Finkelstein N. 2011. Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. Journal of Research in Science Teaching 48: 952–984.

Hernandez RR, Mayernik MS, Murphy-Mariscal ML, Allen MF. 2012. Advanced technologies and data management practices in environmental science: Lessons from academia. BioScience 62: 1067–1076.

Hodcroft EB, De Maio N, Lanfear R, MacCannell DR, Minh BQ, Schmidt HA, Stamatakis A, Goldman N, Dessimoz C. 2021. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature 591: 30–33.

Johnson JW. 2018. Scaling up: Introducing undergraduates to data science early in their college careers. Journal of Computing Sciences in Colleges 33: 76–85.

Karbasian H, Johri A. 2020. Insights for curriculum development: Identifying emerging data science topics through analysis of Q&A communities. Pages 192–198 in Zhang J, Mark Sherriff M, Heckman S, Cutter P, Monge A, eds. SIGCSE '20: Proceedings of the 51st ACM Technical Symposium on Computer Science Education. Association for Computing Machinery.

LeBlanc MD, Dyer BD. 2004. Bioinformatics and computing curricula 2001: Why computer science is well positioned in a post-genomic world. ACM SIGCSE Bulletin 36: 64–68.

Lewis KP, Wal EV, Fifield DA. 2018. Wildlife biology, big data, and reproducible research. Wildlife Society Bulletin 42: 172–179.

Loman N, Watson M. 2013. So you want to be a computational biologist? Nature Biotechnology 31: 996–998.

Madlung A. 2018. Assessing an effective undergraduate module teaching applied bioinformatics to biology students. PLOS Computational Biology 14: e1005872.

Mariano D, Martins P, Santos LH, Melo-Minardi RC de. 2019. Introducing programming skills for life science students. Biochemistry and Molecular Biology Education 47: 288–295.

Marx V. 2013. The big challenges of big data. Nature 498: 255–260.

Medeiros RP, Ramalho GL, Falcão TP. 2019. A systematic literature review on teaching and learning introductory programming in higher education. IEEE Transactions on Education 62: 77–90.

Michener WK, Jones MB. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology and Evolution 27: 85–93.

Muñoz MM, Price SA. 2019. The future is bright for evolutionary morphology and biomechanics in the era of big data. Integrative and Comparative Biology 59: 599–603.

National Academies of Sciences, Engineering, and Medicine. 2018. Data Science for Undergraduates: Opportunities and Options. National Academies Press.

Oesper L, Vostinar A. 2020. Expanding undergraduate exposure to computer science subfields: Resources and lessons from a hands-on computational biology workshop. Pages 1214–1219 in Zhang J, Mark Sherriff M, Heckman S, Cutter P, Monge A, eds. SIGCSE '20: Proceedings of the 51st ACM Technical Symposium on Computer Science Education. Association for Computing Machinery.

Peng R. 2015. The reproducibility crisis in science: A statistical counterattack. Significance 12: 30–32.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Robeva RS, Jungck JR, Gross LJ. 2020. Changing the nature of quantitative biology education: Data science as a driver. Bulletin of Mathematical Biology 82: 127.

Rohrer D. 2015. Student instruction should be distributed over long time periods. Educational Psychology Review 27: 635–643.

Rubinstein A, Chor B. 2014. Computational thinking in life science education. PLOS Computational Biology 10: e1003897.

Sahami M, Aiken A, Zelenski J. 2010. Expanding the frontiers of computer science: Designing a curriculum to reflect a diverse field. Pages 47–51 in Lewandowski G, Wolfman S, Cortina TJ, Walker EL, eds. SIGCSE '10: Proceedings of the 41st ACM Technical Symposium on Computer Science Education. Association for Computing Machinery.

Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013. Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology 9: e1003285.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big data: Astronomical or genomical? PLOS Biology 13: e1002195.

Stephenson C, et al. 2018. Retention in Computer Science Undergraduate Programs in the U.S.: Data Challenges and Promising Interventions. Association for Computing Machinery.

Stes A, Hoekstra A. 2015. Convergence in diversity: Evaluating faculty development across the globe. Studies in Educational Evaluation Complete 46: 1–3.

Strasser CA, Hampton SE. 2012. The fractured lab notebook: Undergraduates and ecological data management training in the United States. Ecosphere 3: 116.

Tenopir C, Allard S, Sinha P, Pollock D, Newman J, Dalton E, Frame M, Baird L. 2016. Data management education from the perspective of science educators. International Journal of Digital Curation 11: 232–251.

White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. 2013. Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution 6: 1–10.

Williams JJ, et al. 2019. Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. PLOS ONE 14: e0224288.

Wilson Sayres MA, et al. 2018. Bioinformatics core competencies for undergraduate life sciences education. PLOS ONE 13: e0196878.

Wright AM, Schwartz RS, Oaks JR, Newman CE, Flanagan SP. 2020. The why, when, and how of computing in biology classrooms. F1000Research 8: 1854.

Nathan C. Emery (emeryna1@msu.edu) is a research associate with the Great Lakes Bioenergy Research Center and the Plant Biology Department at Michigan State University, in East Lansing, Michigan, in the United States. Erika Crispo is an associate professor in the Biology Department at Pace University, in New York City, New York, in the United States. Sarah R. Supp is an assistant professor of data analytics, an undergraduate interdisciplinary program at Denison University, in Granville, Ohio. Kaitlin J. Farrell is the ecology lab coordinator for the Odum School of Ecology at the University of Georgia, in Athens, Georgia, in the United States. Andrew J. Kerkhoff is associate provost and a professor of biology at Kenyon College, in Gambier, Ohio, in the United States. Ellen K. Bledsoe is a postdoctoral teaching and research fellow at the University of Regina with CIEE's Living Data Project, in Regina, Saskatchewan, Canada. Kelly L. O'Donnell is the director of Science Forward at Macaulay Honors College, part of the City University of New York, in New York, New York, in the United States. Andrew C. McCall is an associate professor of data analytics at Denison University, in Granville, Ohio. Matthew E. Aiello-Lammens is an associate professor in the Environmental Studies and Science Department and director of the Environmental Science Graduate Program at Pace University, in New York City, New York, in the United States.