LANDSCAPE OF DATA SCIENCE IN UNITED STATES COLLEGES AND UNIVERSITIES

Michael A. Posner[1], April Kerby-Helm[2], Marjorie E. Bond[3], Alana Unfried[4], Amira Colón[4], Douglas Whitaker[5], and Leyla Batkci[6]
[1]Department of Mathematics and Statistics, Villanova University, Villanova, PA, USA
[2]Department of Mathematics and Statistics, Winona State University, Winona, MN, USA
[3]Department of Statistics, Pennsylvania State University, State College, PA, USA
[4]Department of Mathematics and Statistics, California State University, Monterey Bay, Seaside, CA, USA
[5]Department of Mathematics and Statistics, Mount Saint Vincent University, Halifax, NS, Canada
[6]Department of Mathematical Sciences, Elizabethtown College, Elizabethtown, PA, USA
michael.posner@villanova.edu

As colleges and universities rapidly expand offerings in data science, fundamental questions remain. WHAT should be in the data science curriculum? WHO is teaching introductory data science (IntroDS) courses? WHAT content are those faculty teaching? This paper summarizes data gathered to address these questions. Using a nationally representative sample of 147 U.S. colleges and universities, we estimate the prevalence of IntroDS courses and data science majors. We identify data science program topics developed from professional organization recommendations and expert consultants, and present them via an interactive Shiny app. Lastly, survey responses from instructors across diverse disciplines provide insight into who teaches IntroDS, course structures, and content coverage. Together, these findings offer a snapshot of data science education and the IntroDS course in U.S. higher education.

INTRODUCTION

Data science remains one of today's most in-demand and rapidly evolving disciplines. This was true in 2012 when *Data Scientist* was dubbed "the sexiest job of the 21st century" (Davenport & Patil), in 2017 when *The Economist* declared that "data is the new oil," and it remains true now as nations race to build the data centers that fuel the explosive growth and transformative potential of generative AI. For academics, persistent questions remain: how do we teach data literacy to all citizens, prepare data-savvy professionals for the workforce of tomorrow, and cultivate deep thinkers who will lead in an uncertain future? Progress on effective data science education is hindered by the absence of an agreed-upon definition of a data science curriculum and differing opinions on what should be taught in introductory data science courses.

This paper presents work by the Motivational Attitudes toward Statistics and Data Science Education Research (MASDER) team, funded by the U.S. National Science Foundation (DUE 2013392). Our broader project centers on developing six instruments to measure student attitudes toward statistics and data science, instructor attitudes toward teaching, and the features of the learning environment (Unfried et al., 2025). The Student Survey of Motivational Attitudes towards Data Science was developed to assess students' attitudes toward data science (Kerby-Helm et. al, 2022). The instrument includes an expectancy construct (Eccles & Wigfield, 2020; Wigfield & Eccles, 2019) measuring how students believe they will perform in the data science course. To generate items for this construct, it was necessary to identify the core skills and concepts that students are expected to learn in an introductory data science course, which led to this work.

Here, we provide national estimates of the prevalence of data science programs and introductory data science (IntroDS) courses in the United States based on a nationally representative institutional sample. We also summarize survey findings on who teaches IntroDS, how these courses are structured, and what content they include. Lastly, we synthesize recommendations from professional organizations and data science professionals across various disciplines into a set of data science program topics.

DATA AND METHODS

We used three data sources – a national database of statistics and data science programs, a survey of instructors who teach an introductory data science course, and recommendations from professional organizations for data science programs, curricula, or content. Each of these is explained in more detail below.

National Database of Statistics and Data Science Programs

A primary goal of this component of the MASDER grant was to estimate the proportion of U.S. schools offering majors/minors in statistics and/or data science, as well as the prevalence of undergraduate introductory statistics and data science courses. To construct a nationally representative sampling frame of U.S. colleges and universities, we merged institutional characteristics from the College Scorecard (U.S. Dept. of Education, 2022) and the Carnegie Classification Database (Carnegie Database, 2021). We excluded U.S. Service schools, graduate-only institutions, schools with missing enrollment data or fewer than 300 undergraduates, and some special focus schools (e.g., faith-related, music, and technical profession schools).

Given expected differences across institution types, we employed a stratified random sampling design to increase efficiency and enable comparisons across strata. Institutions were classified by highest degree offered (Associate, Bachelor, or Graduate), selectivity (Selective or Highly Selective ("Selective+") vs. Inclusive; not measured for Associate-degree institutions, which were listed as "All"), and Minority Serving Institution (MSI) status – defined to include Historically Black Colleges and Universities, Hispanic-Serving Institutions, and Tribal Colleges. The final sampling frame included 2,751 institutions. Based on power calculations for an overall margin of error of $\sigma/3$, we selected a stratified random sample of 147 schools (Table 1).

Table 1: Sampling Strata and Sizes

Stratum	N	n	Example School
Grad, Selective+, not MSI	438	19	Duke University
Grad, Inclusive, not MSI	386	18	University of New England
Bach/Grad, Selective+, MSI	64	12	Spelman College
Bach/Grad, Inclusive, MSI	207	12	Fayetteville State University
Bach, Selective+, not MSI	232	11	Swarthmore College
Bach, Inclusive, not MSI	315	16	Miami University - Middletown
Associate, All, MSI	259	11	La Positas College
Associate, All, not MSI	850	48	Cape Cod Community College
TOTAL	2,751	147	

For each sampled institution, we examined publicly available information on institutional websites to determine whether an introductory course in data science was offered and whether a major in data science existed. Courses were classified as IntroDS if they carried titles such as *Data Analytics*, *Data Science*, *Machine Learning*, *Introduction to Data*, or *Data Visualization*. Courses such as *Introduction to Computer Science*, *Data Structures*, or *Introductory Statistics* were excluded. Course descriptions keywords including *data modeling*, *big data*, *algorithmic problem solving*, *data wrangling*, and *data-driven decision making* were used to support classification decisions. Additional data related to data science minors and statistics programs and courses were presented by Colón et al. (2024). A detailed methodological paper describing the sampling design and database access is forthcoming.

Introductory Data Science Course Survey

We created an introductory data science course survey designed to collect detailed information about IntroDS courses taught across the United States. The survey captures instructor demographics, course characteristics (e.g., modality, technologies used, topics covered), and the definition of data science provided to students. Instructors were also invited to upload a syllabus from their IntroDS course. See Posner et al. (2024) and Posner and Kerby-Helm (2024) for more details.

The survey was distributed to instructors of statistics and data science courses through distribution lists of the American Statistical Association, Isolated Statisticians, and the Special Interest Group of Computer Science Educators (SIGCSE), as well as instructors participating in the MASDER project survey development. Incomplete survey responses were excluded from analysis, resulting in 144 valid responses from instructors. These responses are used to learn who is teaching IntroDS and what is being taught. A full exposition of methods and results will appear in a forthcoming paper.

Data Science Program Topics

Despite the inconsistency in what is taught across data science programs, several professional organizations have issued curricular recommendations. We compiled an initial set of competencies and concepts from four major sources: the *ABET Accreditation Requirements* (ABET, 2022), the *Association for Computing Machinery's Computing Competencies for Undergraduate Data Science Programs* (ACM Data Science Task Force, 2021), the *Two-Year College Data Science Summit* report (ASA, 2018), and *Data Science for Undergraduates: Opportunities and Options* (National Academies, 2018). These recommendations were augmented with early findings from the IntroDS course survey, and a presentation by Iain Cruickshank at SIGCSE 2024 (Cruickshank et al., 2024).

Topics were organized into overarching areas: Data, Visualization, Data Processing, Modeling Principles, Modeling Methods, Computational Foundations, Statistical Foundations, Mathematical Foundations, and Data Science Skills. Within each area, we identified subtopics and then enumerated the associated competencies and concepts. Through consultations with subject-matter experts in computer science, data science, statistics, and related disciplines, we refined terminology and clarified the alignment of competencies with their corresponding topics and subtopics. To reduce confusion arising from discipline-specific language, we also added alternate terms, descriptions, and common coding packages, libraries, and functions.

RESULTS

Prevalence of Data Science Programs and Courses

In the representative sample drawn from the National Database of Statistics and Data Science Programs, 32% of institutions offer a data science major (95% CI: 0.23–0.43) and 31% of institutions offer an introductory data science (IntroDS) course (95% CI: 0.24–0.38). The highest prevalence in both cases was among Bachelor-granting, Selective+, non-MSI institutions, with 45% of such institutions having a data science major and 64% having a data science course. The lowest prevalence was among Associate-granting non-MSI institutions, with 25% of those institutions having majors and 15% having courses. Figure 1 displays the percentage of institutions offering a data science major and the percentage offering an IntroDS course across all strata. Majors are chosen only for bachelor's degrees, so two-year colleges in the U.S., which do not offer bachelor's degrees, were excluded when summarizing majors.

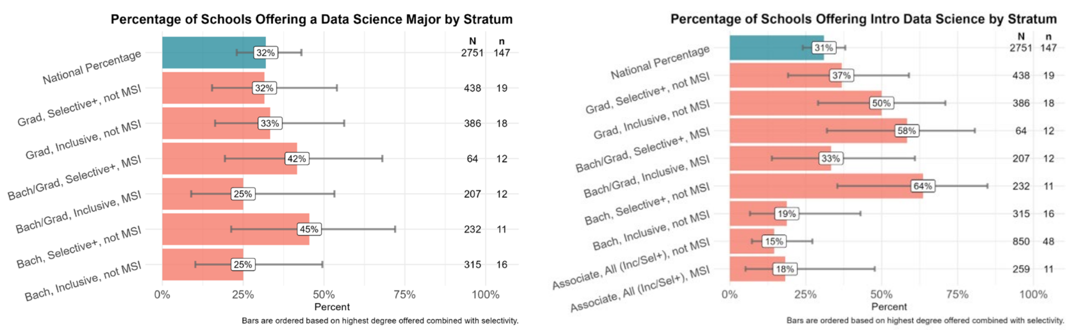

Figure 1: Prevalence of DS Major and IntroDS Course by Stratum

DS Program Topics

The synthesized information regarding data science program topics is available through [an interactive Shiny app]. Figure 2 illustrates how the information is displayed using the app. The user can click on each node of the collapsible tree to expand or collapse it, and hovering over the node allows the viewer to see additional descriptions and information. A spreadsheet version of the material is linked from the Shiny app for ease of view / use.

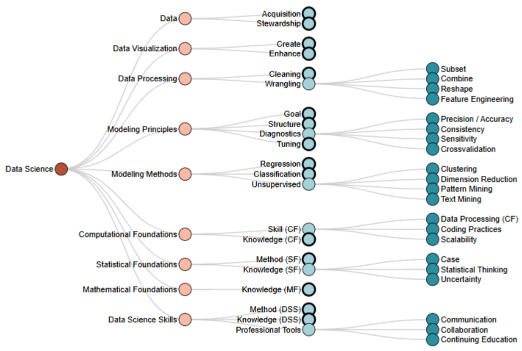

Figure 2: Interactive Shiny app for DS Program Topics

WHO is Teaching IntroDS?

Because of the inherently interdisciplinary nature of data science, IntroDS courses can be housed in various academic units. For the analyses below, we applied the following classification hierarchy: the course was categorized as "DS" if it was taught through a department of data science or if <u>both</u> statistics and computer science were included in the department name (e.g. "statistics and data science"). Next, if computer science or statistics were included in the department name, those disciplines were selected (e.g. "computer and information sciences" would be "CS" and "mathematics and statistics" would be "Stat"). Finally, if some form of mathematics (e.g. "mathematical sciences") was included in the department name, it was recorded as "Math". If none of those apply, but a response was given, it was listed as "Other".

Figure 3 shows the distribution of departments offering IntroDS courses. Data science departments (including departments of "statistics and data science" or "statistics and computing sciences", due the hierarchy described above) have the largest number of offerings, while statistics and mathematics departments contribute similar numbers. We recognize that instructors from CS departments are underrepresented in analyses of the IntroDS Course Survey. Additionally, 22% of survey respondents did not answer the question identifying the department in which their IntroDS course is taught.

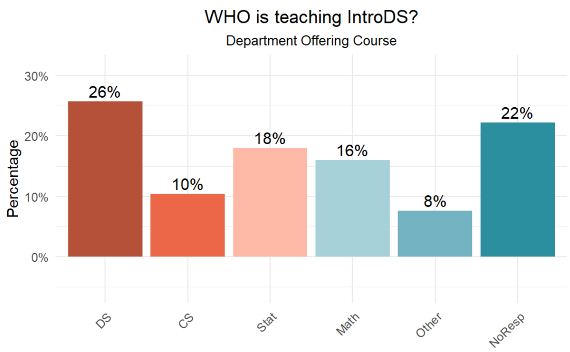

Figure 3: Department in Which DS Course is Offered

Instructors were asked to indicate a primary, secondary, and tertiary professional identity. The same hierarchy used for department classification was applied here. For example, respondents who listed "Stat" and "CS" as co-primary fields were recoded as "DS" for both primary and secondary identities. Respondents listing "Stat" and "Math" as co-primary fields were coded as "Stat" for primary and "Math" for secondary, with any additional field assigned to tertiary.

Figure 4 presents the cumulative distributions of instructor identities. Although more respondents listed statistics than data science as their primary identity (33% vs. 19%), data science becomes more prevalent when secondary identities are incorporated (54% DS vs. 50% Stat) and even more so when tertiary identities are included (69% DS vs. 54% Stat). These findings suggest that many instructors without formal training in data science nonetheless identify with the field, likely due to skills acquired through experience or additional training.

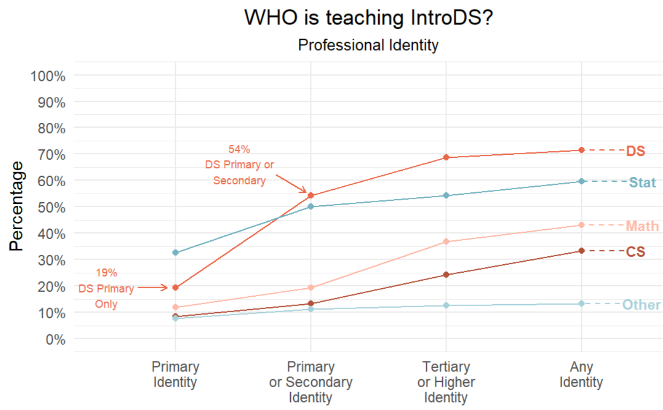

Figure 4: Professional Identity of Instructor (based on DS hierarchy)

WHAT is Taught in IntroDS?

In the IntroDS course survey, instructors were asked, "How do you define data science to your students?" Responses were then categorized by thematic emphasis. The most common definition (39%) described data science as using data to draw insights or knowledge, tell a story, or solve problems. The second most common category (29%) consisted of discipline-specific definitions (some combination of statistics, computing, mathematics, and subject matter expertise). Definitions emphasizing data wrangling and visualization (10%) and those focused on analysis or prediction (10%) were also mentioned.

Respondents were also asked about course prerequisites. A majority of IntroDS courses (58%) require no prerequisite. When a prerequisite was specified, statistics was the most common (17%), followed by computer science (10%), both statistics and computer science (8%), and statistics recommended (8%). Because prerequisites often shape course content, the results presented below are stratified by prerequisite requirements.

A central design decision in any data science course is the choice of technology used for data acquisition, wrangling, analysis, and visualization. As expected, R (https://cran.r-project.org/) is preferred in courses with a statistics prerequisite (83%) (or where statistics is recommended (82%)), whereas Python (https://www.python.org/) is used more often in courses with a computer science prerequisite (86%). Courses requiring both statistics and computer science tend to employ both (50% primarily Python, 33% primarily R, 8% primarily both). Those with no prerequisites tended to favor R (35%) over Python (13%), consistent with the underrepresentation of computer science response, with 43% using none.

Respondents were given a list of potential data science topics covered in their course—data curation and management, data visualization, modeling, data wrangling, data science concepts, statistical foundations, computational foundations, professional skills, and an "other" category. (See the *Data Science Program Topics* section below for more details on each topic.) For each respondent, we calculated the percentage of total course time devoted to each topic. Figure 6 summarizes these results stratified by course prerequisite: instructors spend the most course time on data visualization and the least on professional skills, data curation and management, and other topics. Courses requiring (or recommending) only statistics devote more time to modeling, whereas courses with a computer science prerequisite allocate more time to computational foundations.

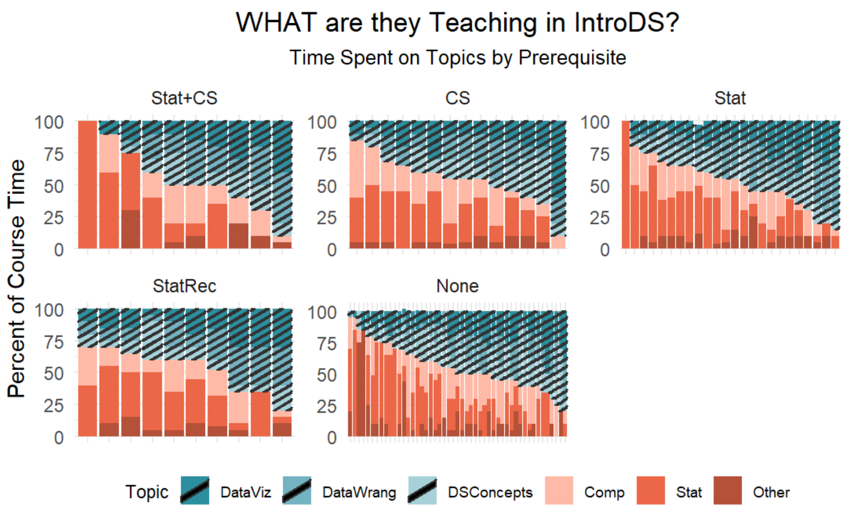

Figure 6: Average Time Spent by Prerequisite

CONCLUSION

As data science continues to evolve rapidly, with AI just beginning to introduce substantial disruptions, there remains a pressing need for clear guidance for faculty initiating new programs or courses, as well as for the administrators who support them. This paper endeavors to describe key elements of the current landscape of data science in the United States. Although data science programs and courses are increasingly common, institutions with fewer resources and those granting lower level degrees have been slower to develop new offerings. We also present a synthesized set of data science program topics, integrating recommendations from several professional organizations.

There were some similarities in what is being taught in IntroDS across the disciplines. Over 75% of the respondents to the IntroDS Course Survey teach courses that contain some common elements. One primary focus is data wrangling which includes creating new variables/features and importing data from common formats like .csv. The other primary focus is data visualization where students create and interpret graphs with one or two variables. Other skills commonly taught include an emphasis on making data-driven decisions, identifying bias, as well as model use and interpretation. Additional topics covered by many respondents include algorithmic bias, data quality, identifying outliers, and ensuring that data visualizations are consistent with the chosen models and that they effectively tell the desired story.

There were some differences in topics taught across disciplines, and more specifically based on the prerequisite for the course, beyond the commonalities mentioned above. Courses with a CS prerequisite leaned more heavily on computational foundations (functions, loops, etc.), using advanced data types (such as JSON), and included machine learning techniques such as decision trees, random forests, clustering, and cross-validation. CS courses tended to focus less than others on both written and oral communication. In contrast, courses with a statistics prerequisite focused more on using data in multiple formats, creating multivariable (3+ variables) data visualizations, and workflow reproducibility. The prominent topics taught in a course without a Stat/CS prerequisite are data privacy and web scraping, as well as some courses focusing on introductory modeling and statistical inference, and such courses do not cover APIs; it appears that some of these courses serve as a substitute for (or as a modification of) the typical introductory statistical methods course.

How to best teach data literacy to all and to prepare data-savvy professionals to succeed in the unpredictable workforce of tomorrow remains unclear. While some disciplinary differences in course design emerged in our data, additional work is needed to better understand and quantify these distinctions. Our hope is that this work will support both institutions and faculty engaged in developing or revising data science curricula.

ACKNOWLEDGEMENTS

This material is based upon work supported by the United States' National Science Foundation (NSF) (DUE 2013392). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES

ABET Computing Accreditation Commission. (2022). *Criteria for Accrediting Computing Programs*. https://www.abet.org/wp-content/uploads/2023/05/C001_CAC-Criteria_2023-2024.pdf

ACM Data Science Task Force (2021). Computing Competencies for Undergraduate Data Science Curricula. New York: Association for Computing Machinery. https://doi.org/10.1145/3453538

American Statistical Association (2018), The Two-Year College Data Science Summit Report. https://www.amstat.org/education/two-year-college-data-science-summit

American Statistical Association (2014), Curriculum Guidelines for Undergraduate Programs in Statistical Science. https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

Carnegie Database (*Carnegie Classifications 2021 Public Data File*, 2021) https://carnegieclassifications.acenet.edu/carnegie-classification/resources/

College Scorecard database (US Department of Education, 2022). https://collegescorecard.ed.gov

College & University Data Science Degrees. (2017). http://datascience.community/colleges. Accessed: 2017-10-12.

Colón, A., Unfried, A., Posner, M.A., Bond, M.E., Whitaker, D., Kerby-Helm, A., & Batakci, L. (2024). National Trends in Undergraduate Statistics and Data Science Courses from a Stratified Sample. Joint Statistical Meetings. (Portland, OR, August 2024) https://sdsattitudes.com/wp-content/uploads/2024/08/jsm-2024-masder-sampling-presentation.pdf

Cruickshank, Iain, J., Bastian, Nathaniel, D., Blair, Jean, R.S., Chewar, Christa, M. & Sobiesk, Edward. (2024) Seeing the Whole Elephant - A Comprehensive Framework for Data Education. *SIGCSE 2024: Proceedings of the 55th ACM Technical Symposium on Computer Science Education.* V.1. pages 248-254. https://doi.org/10.1145/3626252.3630922

Davenport, T.H. and Patil, D.J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review, 90* (10), 70–76.

Eccles, J.S. & Wigfield, A. (2020). From Expectancy-Value Theory to Situated Expectancy-Value Theory: A Developmental, Social Cognitive, and Sociocultural Perspective on Motivation. *Contemporary Educational Psychology*, *61*. https://doi.org/10.1016/j.cedpsych.2020.101859

Kerby-Helm, A., Posner, M.A., Unfried, A., Whitaker, D., Bond, M.E., Batakci, L., & Bolon, W. (2022). S-SOMADS: A Family of Instruments to Measure Data Science Attitudes. In S.A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the Eleventh International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina*. Voorburg, The Netherlands: International Statistical Institute. https://icots.info/icots/11/proceedings/pdfs/ICOTS11_258_POSNER.pdf

National Academies of Sciences, Engineering, & Medicine (2018). *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104

Posner, M.A., Kerby-Helm, A., Unfried, A., Bond, M.E., Whitaker, D., & Batakci, L. (2024, Abstract Only). Topics for an Introductory Data Science Course. Birds of a Feather, *SIGCSE 2024: Proceedings of the 55th ACM Technical Symposium on Computer Science Education V 2*, March 2024, p. 1920. Portland, OR, USA. https://doi.org/10.1145/3626253.3635384

Posner, M.A. & Kerby-Helm, A. (2024). Topics for an Introductory Data Science Course. Poster presentation. *Electronic Conference on Teaching Statistics 2024*. Online. (June 2024)

The Economist. (2017, May 6). The world's most valuable resource is no longer oil, but data. *The Economist*. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Unfried, A., Bond, M. E., Kerby-Helm, A., Posner, M. A., & Whitaker, D. (2025). MASDER Statistics and Data Science Attitudes: A Platform for Collecting Quality Data and Visualizing Results. *2025 Joint Mathematics Meetings*. Seattle, WA, USA. (January 2025)

Wigfield, A. & Eccles, J.S. (2020). Chapter Five - 35 years of research on students' subjective task values and motivation: A look back and a look forward. *Advances in Motivation Science, 7*, 161-198. http://dx.doi.org/10.1016/bs.adms.2019.05.002