Issues in Information Systems
Volume 22, Issue 3, pp. 52-61, 2021

DOI: https://doi.org/10.48009/3_iis_2021_58-68

Developing a master's degree program in data science

John C. Stewart, *Robert Morris University, stewartj@rmu.edu*
G. A. Davis, *Robert Morris University, davis@rmu.edu*
Diane Igoche, *Robert Morris University, igoche@rmu.edu*

Abstract

The term Data Science has been used frequently in the media and in advertisements in recent years. Yet, few can offer a clear and concise definition of the term. This lack of a definition complicates the communication and understanding among entities that focus on the collaboration, use, and application of Data Science. Further complicating the landscape is the fact that advances in technology have altered the perception of Data Science, and consequently, the ways in which it is defined and applied. The goals of this paper are to derive definitions that capture the current collective academic and industry interpretations and perceptions of Data Science (and its components), map these views to emerging areas in Data Science, and determine how well this mapping translates to an applied curriculum for academia. Specific questions addressed include: 1) are we adequately defining Data Science within the confines of how it is currently applied, and how we aspire to apply it?, and 2) are we adequately preparing students to make contributions to, and lead in, this field in terms of the application and advancement of Data Science, based on the current programs being offered? Finally, this study explores what a Master of Science in Data Science should look like, based on the perceived requirements from industry, and discusses the constraints on creating a comprehensive, distinctive, and high-quality Master's-level program in Data Science.

Keywords: data science, curriculum development, graduate programs

Introduction

Data Science (DS) has been an "in vogue" term for some time. However, no fixed or universally accepted definition seems to encompass the specific background, capability, or set of skills that describes a Data Scientist. Further, no institute of higher learning has thoroughly defined the academic discipline, and consequently, the curriculum that would produce competent Data Scientists (Davenport, et al., 2012). It has even been proposed that DS is not an academic discipline at all, but rather a conglomeration or amalgamation of several different knowledge bases and skills that would be difficult to find in a single individual (Irizarry, 2020). The current study explores the numerous definitions of Data Science, from the perspective of industry and academia. The study then concludes with a proposal for a Master of Science Degree in Data Science.

Defining Data Science

Data Science (DS) has been defined as "the study of extracting value from data" (Wing, 2019, p. 2) and "extracting knowledge from data to solve business problems" (Provost and Fawcett, 2013, p. BD52). DS has also been defined as "the theory, method, and technology of studying data science" (Zhu and Xiong, 2015, p. 2). Finally, Irizarry defines DS as an "umbrella term to describe the entire complex and multistep processes used to extract value from data" (Irizarry, 2020, p. 4). The general idea behind DS is using "any means necessary" to leverage value from complex, untidy, and disorganized massive datasets. This broad idea, however, opens the door to a potentially nonspecific and unwieldy all-encompassing definition that includes a number of sub-disciplines.

The problem for academia is that these broad and general definitions do little to provide clues to the skills and expertise necessary to be a Data Scientist. However, a specific definition is required if higher education is to offer degree programs in DS and determine the specifics of a curriculum. In the absence of more detailed definitions, one can examine the skills necessary to accomplish the more general definition of extracting value from a large, disorganized dataset that, again, may be too broad for any single individual to possess.

What constitutes the core knowledge in DS, and the resulting curriculum that prepares data scientists for contribution to industry and society, might be derived from employer expectations alongside academic leadership. The partnership of industry and academia would be an ideal place to begin the development of a DS curriculum.

Literature Review

Much of the research on DS curricula has focused on undergraduate courses and program development (Devaux, et al., 2017; Donoghue, 2020; Anderson et al., 2014; Yan and Davis, 2019; Brunner and Kim 2016), or on the contents of introductory or overview courses (Dicheva and Dicheva, 2017; Stander and Valle, 2017, Asamoah, 2015).

A number of programs have simply used existing courses to create a DS major. This approach is cumbersome and inefficient. The redesign of existing courses and the creation of new courses is necessary to create a quality program, given any limitations required on the number of credits (Deveaux, et al., 2016). Further, some institutions have simply changed the course names to reflect a DS focus. What is needed is a logical approach and deep reflection on the objective we have in creating a Data Science program. As Tang (2017, p. 2) inquired, “What is the fundamental core and intellectual foundation of DS as an academic discipline?”

Several studies proposed that course content should comprise data wrangling, data visualization, and statistical modeling (Hicks and Irizarry, 2018; Zeng, 2017; Yavuz and Ward, 2018). There have been attempts to structure the DS curriculum to incorporate the areas of programming, statistics, and data technologies (Hardin, et al., 2015).

One shift in recent years has been more focus on computing and data competence in statistical analysis. Researchers have argued that the traditional field of statistics has not embraced computing as a discipline, and has, therefore stifled the contribution of the field of statistics and its capability (Nolan and Temple Lang, 2010; Baumer 2015). Some researchers advocate the expansion of the statistics curriculum to use the technologies of DS (Loy, et al., 2019).

One study proposed that DS students should possess competencies in statistics, machine learning, visualization, and ethics (Dicheva and Dicheva, 2017). Within such an approach, the focus of DS curricula has been seen as data wrangling, exploring, and transforming data, computing, modeling, and visualization. (Donoghue, 2017). DS curricula have also focused strongly on communicating the results of such analyses (Dhar, 2013). The National Academy of Sciences suggests that DS practitioners should have computational and statistical skills, data management, the ability to describe and visualize data, modeling, communication skills, ethics in problem solving, domain-specific knowledge, and workflow reproducibility (NAS, 2018).

Other experts have advocated a focus on "Big Data" technologies and model building techniques (Song and Zhu, 2016).

There is, however, no consensus on the components of a professional Master's-level DS program that will reflect the expectation of knowledge and skills that graduates should possess. There is an obvious need for standards, given the number of professional Master's degree programs in DS (Wing et al., 2018).

Despite a very wide range of sub-disciplines within DS, and the accompanying absence of a definitive universal agreement on curriculum standards, there does appear to be an emergence of commonality of topics that includes data visualization, wrangling, statistical modeling, machine learning, programming, and reproducible research that have become the norm in many programs (Schwab-McCoy, et al., 2020).

Industry Requirements and the Data Science Curriculum

Many questions arise from industry regarding DS. For example, how does the potential convergence on a standardization of DS topics match with the expectations that prospective employers have of graduates from academic programs? How do the perspectives from industry and academia map to the makeup and structure of courses offered in flexible and dynamic higher education programs? Finally, are colleges and universities preparing students to meet the challenges of industry, while teaching students to lead, explore, and innovate in this very diverse and the rapidly developing and evolving DS discipline?

A 2020 survey by *365 Data Science* found that the most sought-after skills listed in job descriptions for Data Scientists are machine learning, statistics, and programming skills in Python (Krastev, 2020). The remaining most-quoted skills were computing, data visualization, statistical modeling, and predictive modeling. There was no specific mention of data cleaning or wrangling.

Under the Machine Learning techniques, the most often quoted skills in the survey results were deep learning, clustering, and natural language processing (NLP). The next skills (in rank order) are predictive modeling and expertise in various Python libraries with Machine Learning capability. Data Scientists, then, need practical expertise in computer programming to support and enhance Machine Learning expertise.

Therefore, the 365 DS survey purports that the core of DS, at least in the minds of industry, is the focus should be on the application of Machine Learning, statistical analysis, and programming to solve problems. Specialized skill requirements are related specifically to supporting Machine Learning disciplines. Not surprisingly then, the indication by this survey is that the focus of the Data Scientist should be on application, as opposed to theory and concepts (Hicks and Irizarry, 2018).

Currently, the programming languages offering the greatest capability and use in developing Machine Learning models, include R (with its capability for rapid exploration and plotting), and Python with its libraries like TensorFlow and Keras. This need for R and Python is seen in the expectations of employers. The 365 DS survey found that proficiency in Python was the top computing capability that employers expect, followed closely by R and SQL (Krastev, 2020). Additionally, the most listed database and cloud storage technologies were Spark, Amazon Web Services (AWS), and Hadoop. The two main data visualization tools required were Tableau and Microsoft Power BI.

Clearly, along with the computational, statistical, programming, and model development competencies, an effective Master's level program in DS must include the use and training in the appropriate tools that graduates would expect to encounter in their initial jobs and in their careers. This finding does not imply that academia cannot lead in new disciplines and software proficiency. However, it would be perilous to structure the foundation of a program while ignoring the demands of the marketplace.

In summary, industry requires an understanding of skills in Machine Learning (including Deep Learning, clustering, and NLP), computing, statistical and predictive modeling, and in the necessary toolset of open-source and vendor-supported software. Now that the needs of industry are fully understood, an effective curriculum for a graduate level program in DS can be discussed.

Inherent Problems in Data Science Curricula Development

Creating a graduate-level program in DS is fraught with a number of challenges. Aside from the unresolved issue of which competencies to include, one of the main problems with developing a DS curriculum, particularly at the Master's degree level, is fitting the broad topic areas into a manageable timeframe for prospective students. Additionally, while faculty might propose a rigorous program with comprehensive coverage of the DS sub-disciplines, the administration could favor a limitation on the number of credits, and the use of existing courses that may have limited applicability to outcomes envisioned by the faculty.

One of the often-quoted issues in the literature is the difficulty instructors have in balancing the varied backgrounds and expectations of students (Kross and Guo, 2019; Schwab-McCoy, et al., 2020). DS inherently attracts students from diverse backgrounds. Hence, a flexible program, with electives (or other options), where students with a strong background in an area can have courses waived (or replaced with alternate courses), and where students with a weaker background can take foundation courses (to raise their competency level) are preferable to programs with rigid core requirements.

While the current and future demand for Data Scientists is well documented, this type of shortage in the market does not necessarily translate into a surge of students applying for admission into a DS program. Many faculty believe the placement of DS belongs at the Graduate Student level. Few high school seniors or college freshmen understand the discipline or implications of DS. Conversely, those in the workplace, with some experience, will know the value of leveraging data, see the opportunity, and seek additional education. Further, students with work experience gain more from a program after having an understanding and perspective from working with data in the workplace.

There appears to be some synergy between the skills and competencies sought by industry and what academia is proposing, in terms of curricula. What often is seen by academia as critical curriculum components in DS does not always match with what industry demands (and what industry lists as needed skills in job postings) (Schwab-McCoy, et al., 2020; Krastev, 2020; Mitchel, 2019).

Obviously, the approach to the development of any specific DS course reflects the instructor's view of the world and what he/she views as the priorities. This approach may involve some disconnect between the course structure and the proposed course within the curriculum. Although, this disconnect may not necessarily be a bad thing. Improvement, enhancement, or augmentation of what is proposed, and/or changes as technology or algorithm advancements occur is ideal and to be expected. However, adherence to a strategy or structure outlined in a curriculum would generally be the basis or foundation upon which to build courses (Schwab-McCoy, et al., 2020).

Further, while the above input on DS topics from researchers does give one a basis from which to begin the development of a curriculum, and provides some direction on specific courses to include, it lacks the detail provided by surveys of competency and skill demands by industry. Machine learning is a broad field. What aspects of Machine Learning are the most in demand? What should be the focus? What about programming? What language(s)? What level of programming skill? Machine learning, including Deep Learning and NLP are at the top of the skills list for Data Scientists in at least one survey (Krastev, 2020). Competency in these areas will not be accomplished in a single Machine Learning course. The specific software tools must also be addressed. For example, what specific software tools are the most conducive for learning the concepts, and which tools will be most in demand by employers?

A survey of DS instructors found that the top two programming languages (by a wide margin) used in introductory DS courses were R and Python (Schwab-McCoy, et al., 2020). Software resources used in academia appear to be generally aligned with expectations in industry, based on the survey. The use of R, R Studio, and Python, with their supportive libraries, seem to be the tools of choice within industry (Dichev and Dicheva, 2017: Loy, et al., 2019, Hicks and Irizarry, 2018; Broatch, 2019, Stander and Valle, 2017).

Research Methodology

An exploratory content analysis of 30 Master of DS programs was conducted in a prior study. This content analysis found markedly different core requirements, in terms of courses and credits (Bukari, 2020). The average number of required core courses was 9.7, with NYU having the highest number of total required courses at 17, and MIT requiring the highest number of credits at 84. On the low end of required core courses and credits was 4 and 7, respectively. Another study of 30 Master's-level programs found an average of 40 required credit hours, with 30 in the core and 10 electives (Tang, 2017). This wide variability in graduate-level programs could be the result of either a credit total of core course requirements, or of the level of flexibility inherent in the programs that allows students to choose courses among the different sub-disciplines in DS.

The tremendous challenge in developing DS curricula, it appears, is to offer the right mixture of breadth and depth to adequately prepare students, while keeping the requirements to a manageable and competitive level, as perceived by higher education administrations. The authors of this study have found that efforts to structure a flexible Master's Level program (i.e., one that encompasses all the current sub-disciplines of DS) become muted and stifled in administrative channels that favor benchmarking against competing programs. The authors also found that administrations frequently wants to limit total required credits to a predetermined level.

The current study looked at the course offerings from the top 20 DS schools (see Table 1) as determined by U.S. News and World Report. While the research focused on Master's Level programs, the authors looked at all DS courses offered at each school, without differentiating between graduate and undergraduate schools. From Table 1, one can see that most schools do not offer specific courses in data cleaning/data wrangling to prepare data for analysis. It is important to note that data cleaning/wrangling is one of the most time-consuming tasks for a data scientist, even with the automated tools that are currently available. It is possible that the important skills of data cleaning/data wrangling may be covered in other courses like Machine Learning, or perhaps in some of the introductory courses.

Only 35% of the schools in the study offer courses in Deep Learning. However, Deep Learning may be incorporated into a separate Machine Learning course. Nearly all schools in the study offer at least one course (most offer several) in Machine Learning and/or Data Mining. Not surprisingly, almost all of the schools offer more than one course in Statistics. The percentages in areas of Data Visualization, Big Data, and Database courses (i.e., 65%, 55%, and 60% respectively) show that more than half of the programs value these components. The lower number of course offerings in Natural Language Processing and Forecasting point to a lack of emphasis in these areas. This lack of emphasis reveals a glaring inconsistency with the 365 DS survey, which lists NLP as one of the top areas of demand (Krastev, 2020). Forty-five percent of the schools listed in Table 1 offer at least one course in Optimization and/or Simulation.

Clearly, none of the schools is offering all of the potential courses that would cover all components of the current, and very broad, view of what encompasses the capabilities of a Data Scientist. A number of schools appear to show an emphasis in specific areas. Based on the information in Table 1, offering several courses in areas like Big Data, Optimization/Simulation, or Advanced Statistics promotes a stronger foundation in these sub-disciplines.

Table 1: Data Science Courses at Top-Ranked Schools

	Intro Data Science	Machine Learning /Data Mining	Data Cleaning/ Wrangling	NLP	Visualize /Explore	Data-base	Optimization/ Simulation	Deep Learning	Programming	Massive Datasets	Stats	Forecast
Purdue	1	3		3		1	2	1	1	1		
Rochester	2	4		3		1	2	1	1	1	3	1
NYU	3	2		2		1	2	1	2	1	2	2
CMU	4	3				4				3		
Columbia	2	2			1			2	1		3	
Colorado State	3		1		1		1				2	
Iowa		2			1	1					6	
NC State	1	4				3			4			
Georgia Tech	4	3		1	2	1	4	1		4	6	1
DePaul	2	3			2	1		1	2	1	2	1
USC LA		2			2	4	4			1	2	
MIT		1						1	1		1	
Boston Univ	3	3			1	1			3			
Northwestern	1	3		1	1	1	2	1	1		2	1
Brown	1				1			1			2	
Chicago		3			1		1		1	1	4	1
VA	1	3			1				2		1	
Seattle	1	2		1	1	1			2	1	1	
Kennesaw		3							1	2	4	1
OK		3			1		2		2	2	2	
Totals	29	49	1	11	16	20	20	10	24	18	43	8

A Proposed M.S. in Data Science

As mentioned previously, graduate level DS programs attract students from diverse backgrounds, and with various abilities, expectations, and prior experience. In an improperly developed and inflexible curriculum, the frontline instructor is left to sort out and manage the diversity at the course level (Kross and Guo, 2019; Schwab-McCoy, et al., 2020). Teaching within such a curriculum is a cumbersome and ineffective approach. Such an ineffective approach has the very real potential for capable students to under-perform, and for unprepared students to become frustrated.

The authors of this study believe that a Master's-level program must offer all the components and sub-disciplines of DS, as they are currently perceived. These courses should be offered as required foundation courses or via electives. The expectations of skills coming into the program should be clearly articulated to students. If students lack the requisite skills or knowledge, then foundation courses may be offered. Conversely, a stand-alone program might ideally start all cohorts with the assumption of little or no experience and waive certain courses on an individual basis (i.e., if the courses have been completed as similar courses at the undergraduate level).

Based on the demands of prospective employers, full length courses in Deep Learning, NLP, and Big Data (working with tools needed for analysis of massive datasets) should be part of the curriculum. All of the current "high demand" tools like R, Python, SQL, Spark, Hadoop, Tableau, and MS-Power BI should be incorporated into the coursework, as tools to enhance learning.

A proposed program would require 42 Master's-level credits and would encompass all of the current sub-disciplines, as currently perceived by industry and academia. The credit hours can be altered to allow for flexibility in the program and make use of electives to allow students to be more focused on a sub-discipline rather than a rigid adherence to core course requirements. However, to assure well-prepared and competent graduates, students should have in-depth exposure to all of the course components that have been outlined previously in this study.

Table 2 outlines the proposed curriculum for a Master's Degree in DS. As seen in Table 2, this proposed curriculum addresses the needs of industry, but also allows flexibility to accommodate a diverse range of student backgrounds. The proposed curriculum in Table 2 contains Statistics and Deep Learning, but also addresses the need for in-demand software skills, like R, Python, Hadoop, and Spark. Finally, the proposed curriculum is designed to allow for new concepts and technologies to be conveyed in (easily adaptable) seminar courses.

Table 2: Proposed Curriculum for a Master's Degree in Data Science

Course Title	General Topics	Software Tools	Credits
Statistics	Through Multiple Regression	R and R Studio	3
Data Mining (Introductory Machine Learning)	Classification, Neural Networks, k-NN, Decision Trees, etc.	R, Python	3
Introductory Python			3
R Programming			3
Database Management	Overview, Basic Course	SQL	3
Advanced Database Systems	NoSQL, NewSQL, Cloud (AWS)	AWS	3
Advanced Machine Learning	Gradient Boosting, Interpretable Machine Learning, Autoencoders	R, Python	3
Optimization & Simulation	Spreadsheet Modeling, Linear Programming	Excel, R	3
Forecasting	Time Series, etc.	Excel, R	3
Natural Language Processing (NLP)	Text mining, Parsing, Stylometry	Python	3
Big Data Technologies	Massive Datasets	Hadoop, Spark, etc.	3
Data Visualization	Plots and Graphics	Tableau, R, Python	3
Deep Learning	Convolutional Neural Networks, Recurrent Neural Networks, Generative Adversarial NN	R and Python	3
Data Science Seminar(s)	To evaluate new information in the continually changing field of Data Science		1 credit per semester or 3 credits total
Total Credits			42 credits

Conclusions

In consideration of the current study's findings on the expectations of the broad expertise of a Data Scientist, one can conclude that a valid, adequate, and distinctive program should include the course components outlined in Table 2. In order to offer a Master's Degree in DS that can be completed in a timely manner (within a prescribed number of credit hours), certain prerequisite skills will be required prior to entering the program. In addition, the depth and breadth of courses in the degree program will need to be such that students are exposed to the major areas of DS but have available avenues for further exploration of topics. This course structure would mean that one 3-credit course may not be allocated to each of the DS sub-disciplines. Further, an alternate way of proportioning time to each sub-discipline would need to be determined and developed to create a broad exposure of these varying topics to students. Students can also be introduced to sub-disciplines by collaborating with faculty on research endeavors via independent study, or via seminar courses. This approach to DS Master's degree curricula will give students both a hands-on and a conceptual learning experience.

There are numerous limitations to the current study that should be noted. First and foremost, the current study only involved a sampling of 20 schools. A more extensive study could sample a larger number of schools that offer DS courses and programs. In addition, the current study examined schools that offer DS courses at the undergraduate and graduate levels. A more focused study could concentrate only on schools that offer Master's-level degree programs in DS. Subsequent studies in this area could survey academia and/or industry to determine the optimal combination of courses in a Master of Data Science program.

References

Asamoah, D., Doran D., Schiller, S. (2015). Teaching the Foundations of Data Science: An Interdisciplinary Approach. Pre-ICIS SGDSA Workshop. Retrieved May 11, 2021 from https://works.bepress.com/daniel_asamoah/ 14/ Bastian M., H.

Anderson, Paul & McGuffee, James & Uminsky, David. (2014). Data science as an undergraduate degree. 705-706. 10.1145/2538862.2538868. Proceedings of the 45th ACM technical symposium on Computer science education.

Ben Baumer (2015) A Data Science Course for Undergraduates: Thinking With Data, The American Statistician, 69:4, 334-342, DOI: 10.1080/00031305.2015.1081105

Bhatnagar, S.; Alexandrova, A.; Avin, S.; Cave, S.; Cheke, L.; Crosby, M.; Feyereisl, J.; Halina, M.;Loe, B. S.; h ´ Eigeartaigh, S. O.; Martnez-Plumed, F.; Price, H.; Shevlin, H.; Weller, A.; Winfield, A.; and Hern´andez-Orallo, J. 2018. Mapping Intelligence: Requirements and Possibilities. In M¨uller, V. C., ed., Philosophy and Theory of Artificial Intelligence 2017. Berlin: Springer. 117–135.

Broatch, J. E., Dietrich, S., & Goelman, D. (2019). Introducing Data Science Techniques by Connecting Database Concepts and dplyr. Journal of Statistics Education, 27(3), 147-153. https://doi.org/10.1080/10691898.2019.1647768

Brunner, R. J., & Kim, E. J. (2016). Teaching data science. Procedia Computer Science, 80, 1947-1956. https://doi.org/10.1016/j.procs.2016.05.513

Bukhari, Duaa, Data Science Curriculum: Current Scenario (2020). International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.10, No.3, May 2020 , DOI:10.5121/ijdkp.2020.10301 1, Available at SSRN: https://ssrn.com/abstract=3616600

Davenport, Thomas H., and D. J. Patil. "Data Scientist: The Sexiest Job of the 21st Century." Harvard Business Review 90, no. 10 (October 2012): 70–76.

Richard D. De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, Lei Z. Cheng, Amanda Francis, Robert Gould, Albert Y. Kim, Matt Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruviluamala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, Ping Ye. Annual Review of Statistics and Its Application 2017 4:1, 15-30.

Dhar, Vasant. (2012). Data Science and Prediction. Communications of the ACM. 56. 10.2139/ssrn.2086734.

Dichev, Christo & Dicheva, Darina. (2017). Towards Data Science Literacy. Procedia Computer Science. 108. 2151-2160. 10.1016/j.procs.2017.05.240. Donoghue, 2017.

Donoghue, T., Voytek, B. and Ellis, S.E. (2021) Teaching Creative and Practical Data Science at Scale, Journal of Statistics and Data Science Education, 29:sup1, S27-S39, DOI: 10.1080/10691898.2020.1860725

Hardin, J. Hoerl, R., Horton, D. Baumer, O. Hall-Holt, P. Murrell, R. Peng, P. Roback, D. Temple Lang & M. D. Ward (2015) Data Science in Statistics Curricula: Preparing Students to “Think with Data”, The American Statistician, 69:4, 343-353, DOI: 10.1080/00031305.2015.1077729

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. The American Statistician, 72, 382–391. https://doi.org/10.1080/00031305.2017.1356747

Irizarry, R. A. (2020). The Role of Academia in Data Science Education . Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.dd363929

Krastev, N., (2020). Study: What Are the Requirements for Data Scientist Jobs in 2020? Oracle AI and Data Science Blog, October 22, 2020. (https://blogs.oracle.com/ai-and-datascience/post/study-what-are-the-requirements-for-data-scientist-jobs-in-2020)

Kross, S. and Guo, P.J.: Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. CHI 2019: 263 Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019. ACM 2019, ISBN 978-1-4503-5970-2

Loy, A., Kuiper, S. and Chihara, L. (2019): Supporting Data Science in the Statistics Curriculum, Journal of Statistics Education, DOI: 10.1080/10691898.2018.1564638Mitchel 2019

National Academies of Sciences, Engineering, and Medicine. (2018). Data Science for Undergraduates: Opportunities and Options. Retrieved from The National Academies Press website: https://doi.org/10.17226/25104

Provost, F., and Fawcett, T. Data science and its relationship to big data and data-driven decision making. Big Data 1, 1 (2013), 51-59; http://online.liebertpub.com/doi/pdfplus/ 10.1089/big.2013.1508

Schwab-McCoy, A., Baker, C.M. and Gasper, R.E. (2020): Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years, Journal of Statistics Education, DOI: 10.1080/10691898.2020.1851159

Song and Zhu, 2016 Song, I.-Y., & Zhu, Y. (2016). Big data and data science: What should we teach? Expert Systems,33(4), 364–373. https://doi.org/10.1111/exsy.12130

Stander, J. and Dalla Valle, L. (2017) On Enthusing Students About Big Data and Social Media Visualization and Analysis Using R, RStudio, and RMarkdown, Journal of Statistics Education, 25:2, 60-67, DOI: 10.1080/10691898.2017.1322474

Tang R, Sae-Lim W. Data Science Programs in U.S. Higher Education: An Interview with the Authors. Journal of eScience Librarianship 2017;6(1): e1105. https://doi.org/10.7191/jeslib.2017.1105. Retrieved from https://escholarship.umassmed.edu/jeslib/vol6/iss1/4

Tang, R., & Sae-Lim, W. (2016). Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. Education for Information, 32(3), 269–290. https://doi.org/10.3233/EFI-160977

Wing, Jeannette & Janeja, Vandana & Kloefkorn, Tyler & Erickson, Lucy. (2018). Data Science Leadership Summit Summary Report. 10.13140/RG.2.2.13710.61764.

Wing, J. M. (2019). The data life cycle. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e26845b4

Yan, D. and Davis, G.E. (2019) A First Course in Data Science, Journal of Statistics Education, 27:2, 99-109, DOI: 10.1080/10691898.2019.1623136

Yavuz, F. G., & Ward, M. D. (2018). Fostering Undergraduate Data Science. American Statistician, 0(0), 1–9. https://doi.org/10.1080/00031305.2017.1407360

Zheng, T. (2017). Teaching Data Science in a Statistical Curriculum: Can We Teach More by Teaching Less? Journal of Computational and Graphical Statistics, 26(4), 772–774. https://doi.org/10.1080/10618600.2017.1385473

Zhu, Y.and Xiong, Y. “Defining data science,” arXiv:1501.05039 [cs], Jan. 20, 2015. arXiv: 1501 . 05039. [Online]. Available: http://arxiv.org/abs/1501. 05039 (visited on 02/23/2021).