Building a New Data Science Program Based on an Existing Computer Science Program

Michael J. Oudshoorn
School of Engineering
High Point University
High Point, NC, USA
moudshoo@highpoint.edu

Kimberly J. Titus
Department of Computer Science
High Point University
High Point, NC, USA
ktitus@highpoint.edu

William K. Suchan
Department of Computer Science
High Point University
High Point, NC, USA
wsuchan@highpoint.edu

***Abstract*—Data Science is an emerging program of study which, by its very nature, is interdisciplinary. This paper explores the challenge of introducing a data science program, given an established computer science program, within a liberal arts institution which has a large general education core. In addition to the challenge of creating a new program, there are several other major challenges surrounding the fact that the discipline of data science continues to evolve, and accreditation criteria are currently being developed.**

***Index Terms*—data science, computer science, curriculum development**

I. Introduction

Data science is an emerging discipline [1]–[6] for which many universities are scrambling to create degree programs. Concurrently, the Computing Accreditation Commission (CAC) of ABET is beginning the process of developing program accreditation criteria for data science [7]. This will likely be similar to other computing criteria [8] which will require programs to satisfy the general criteria as well as additional program specific criteria. To add to the challenge, the computer science curriculum guidelines [9] are in the process of being revised [10], [11].

Despite the changes afoot, there is a growing demand for graduates of data science programs and employers are fiercely competing for graduates of the small, but growing, number of data science degree programs that currently exist. The challenge for universities is how to develop data science degrees to support the needs of industry even as data science is trying to define itself. Likewise, finding faculty with domain expertise is difficult. While many early pioneers in data science programs opted to offer Masters-level programs, the current demand for graduates, the accreditation activities, and the papers discussing data science curriculum [12]–[19], suggest that there is also a strong demand for undergraduate programs.

At High Point University (HPU), a bachelor's degree in computer science is in place. A concentration in cybersecurity is also available within the degree and it is currently servicing about 150 majors with 5 faculty members. The department is fortunate enough to have several new faculty lines allocated to it and is looking to hire faculty to support and develop a data science program. The degree will be based on existing documents that attempt to define what data science is [1]–[3], [5], and the accreditation criteria for data science that is being developed by ABET [7]. Using these guidelines, the Department of Computer Science is developing a 4-year data science bachelor's degree program with well-defined computer science and math/statistics requirements as well as a domain area in which data science is being applied. The application domain will follow a plug-and-play model to support a variety of disciplines including business, physics, biology, health sciences, and others. A capstone course, taught by data science faculty using mentors from the various domains, will have students apply data science techniques to application domain specific data sets in meaningful ways.

While this approach provides depth in computer science, and substantive coverage of mathematics and statistics in line with employer needs, it does not offer significant depth in the application domain. Within the context of a liberal arts institution with a large general education requirement (50 semester credit hours), it was felt that the coverage of domain knowledge was minimal at best. To address this, students have an option to complete a double degree (minimum 160 semester credits) over 5 years to provide depth and substance in the application domain in order to more expertly and meaningfully apply data science techniques to large, domain-specific data sets. This paper describes the degree structure and curriculum under development and the rationale for decisions made. In particular, it addresses how the introduction of a data science program impacts an existing program in computer science and what new courses are necessary to support the data science program.

II. Liberal Arts

HPU is a liberal arts institution with a general education core of 50 semester credits which provides breadth across numerous disciplines. HPU brands itself as a ”life skills university” which, in part, refers to the broad liberal arts foundation that students receive, but also to the fact the this general education core also delivers soft or professional skills such as critical thinking, problem solving, communication skills, research methods, ethics, entrepreneurship, and leadership. These are all valuable and necessary skills which are highly sought after by employers.

In developing a data science program at HPU, it is necessary to work within the constraints derived from such a large general education/liberal arts core with 4-credit hour courses [20]. Fortunately, a few courses in the general education core, such as the Quantitative Reasoning requirement, also count toward the data science program. However, the large core restricts the number of credits available to implement the data science program.

III. Accreditation

While there are no accreditation requirements for data science at this time, work is underway to develop ABET program criteria [7]. The criteria are in an embryonic stage at this time and are unlikely to be completed and made available for public comment until July 2020 at the earliest. It will be at least another year after that before the criteria are finalized and ready to be applied to programs. However, the current thinking is that the data science criteria will fall under the CAC and, as with other program-specific criteria, will be required to satisfy the general criteria [8] as well as the emergent data science criteria. The general criteria are stable and unlikely to change as the data science criteria are developed. To further complicate matters, the current computing criteria are based on the 2013 ACM/IEEE computing curriculum guidelines [9], but new computing curriculum guidelines are currently under development [10], [11], [21].

With the understanding that the computing curriculum guidelines and the accreditation criteria are both currently under development, HPU is electing to build a data science program based on what we believe are the emerging standards. That said, the program must be flexible and adapt to changes that will inevitably occur within the data science accreditation guidelines over the next 2 years. The current plan is to satisfy the ABET general computing criteria and ensure students have at least 30 semester credit hours of computer science, including data structures and algorithms, software development, machine learning and artificial intelligence, information management, and an exposure to computer architecture and organization, networking and communication, and distributed and high-performance computing. Furthermore, students must be able to apply data science theory and techniques to satisfy stakeholder needs. Students also need a minimum of 18 semester credit hours of mathematics and statistics that must include discrete mathematics and substantial coverage of probability and statistics. Students must also have a minimum of 6 semester credit hours of domain knowledge that provides context for the data science activities, and must complete a major project that requires integration and application of knowledge from earlier course work. While topics such as statistics, domain knowledge, and algorithms are mandated by the proposed accreditation standards, the actual content within those topics is not, and there is an expectation that the content appropriately supports the program and the institutional objectives for the program.

Taking these computing and data science accreditation projections, and combining them with the large liberal arts general education requirements at HPU, the program comes close to exceeding the maximum number of credits permitted in an undergraduate degree. The challenge is further complicated when HPU’s 4-credit course model is also considered since the majority of the accreditation requirements are multiples of 3-credit courses.

IV. CS + X

There has been significant work in exploring CS+X where X can be almost any other discipline [22]–[24]. Typically, these programs look to see what happens when the intersection of Computer Science and discipline X is explored, such as Computer Science and law [25], or biology [26], or behavioral and social sciences [27]. The proposed data science program at HPU shares some similarities with these initiatives in that it seeks to apply computer science to an application domain. However, it also differs from the CS+X programs in that it also brings statistics and probability into the mix, focuses the computing topics covered by way of the accreditation requirements, and requires an integrative and substantial project that brings together computing, statistics, and the application domain through the use of data science theory and techniques. In short, data science is treated as a new discipline with specific learning outcomes to be achieved, rather than just an interesting intersection of sometimes seemingly disparate disciplines.

V. Degree structure

Table I shows the typical 4-year degree plan for Data Science as proposed at HPU. It satisfies all of the

Table I
TYPICAL DATA SCIENCE 4-YEAR PLAN OF STUDY (128 CREDITS).

BS in Data Science

Freshman Year - Fall		**Freshman year - Spring**	
Course name	*Credits*	*Course name*	*Credits*
Introduction to Data Science (DS)	4	Python Programming (DS)	4
Calculus I (M, C)	4	Calculus II (M)	4
First Year Seminar (C)	4	College Writing and Public Life (C)	4
General education course - literature (C)	4	General education course - social science (C)	4
President's Seminar (C)	1		

Sophomore Year - Fall		**Sophomore year - Spring**	
Course name	*Credits*	*Course name*	*Credits*
Advanced Programming & Data Structures (DS)	4	Database Systems (DS)	4
Statistics for Data Science (DS, M)	4	Data Visualization (DS)	4
Discrete Structures (DS, M)	4	Domain area course (D)	4
Domain area & general education course (D, C)†	4	General education course - foreign language (C)	4

Junior Year - Fall		**Junior year - Spring**	
Course name	*Credits*	*Course name*	*Credits*
Computer Systems (DS)	4	AI & Machine Learning (DS)	4
Mathematics or Statistics elective (M)	4	Algorithms (DS)	4
Domain area course (D)	4	Engineering and Technology Ethics (C)	4
General education course - natural science (C)	4	General education course - religion (C)	4

Senior Year - Fall		**Senior year - Spring**	
Course name	*Credits*	*Course name*	*Credits*
Data Management (DS)	4	Software Engineering (DS)	4
Cloud Computing (DS)	4	Data Science Capstone Experience (DS)	2
General education course - history (C)	4	General education course - fine arts (C)	4
elective	4	elective	4
		Physical Education course (C)	1

(DS) = Data Science/Computer Science course   (D) Domain area course
(M) = Mathematics or Science course   (C) = General Education/core course
† The most likely application domains to pair with data science are in the natural and social sciences.

constraints identified above. It leverages the existing computer science courses to the fullest extent possible and requires the introduction of only 5 new data science/computing courses (Intro to DS, Data Visualization, Data Management, Cloud Computing, and the Capstone course) and a specialized statistics course for data science. Keeping the proposed ABET criteria in mind, the plan includes 46 hours of data science/computing courses, 20 hours of mathematics and statistics, and 12 hours in the student's domain area of choice. Two free electives in year 4 allow the student to take additional courses of their choosing, either in computing, in their domain area, or in any other area of interest.

The application domain, similar to the CS+X paradigm, can be almost any data-rich discipline (STEM-disciplines, health sciences, pharmacy, business, economics, meteorology, etc). The data science major focuses on data science theory and techniques and draws examples from a wide range of companion domains.

Students pursue statistics somewhat independently from their computer and data science courses, but several data science courses (e.g., data visualization) leverage the statistics that the students have been exposed to in prior courses. Similarly, the application domain is pursued somewhat independently from the data science courses throughout much of the degree. This was a deliberate choice to permit a plug-and-play model for the application domain. This allows data science to partner with many disciplines across campus but comes at the cost of a less integrated model in comparison to when the data science program is bound to just one or two application domains. This is a decision that will need to be monitored as the program is implemented to ensure that the students are well-served. The capstone experience integrates the statistics, application domain, and computer/data science content. Students propose a research topic in the fall of their senior year and implement it the following Spring. The course is taught by data science faculty in conjunction with application domain faculty who will help mentor the students. While the capstone experience may heavily focus on the application of data science theory and techniques, the work undertaken must be relevant and useful to the application domain and produce results that are statistically significant. This could, for example, include working with large data sets related to health records, geographic information systems, astrophysics data, stock market information, manufacturing, and much more.

The 4-year degree plan (see Table I) may satisfy the proposed ABET accreditation criteria [7], but it does not provide significant exposure to the application domain. Ideally, graduates should have a much greater exposure to the domain than is achievable within a 4-year program given the large core at HPU. The solution is to also offer students an opportunity to pursue a 5-year, 160+ credit program, which provides students a bachelor of science degree in data science and a bachelors degree in the corresponding application domain. While many students will not choose the 5-year option, it is important that it be available to provide student choice, and satisfy what we perceive to be industry demand for domain expertise to accompany the data science sills. By spreading the 4-year program across 5 years, prerequisites can be preserved and space for additional topics in the application domain can be interleaved throughout the curriculum. There is also a mechanism for students to revert to a 4-year degree in either data science or in the application domain should they decide the 5-year program is not for them, as long as students make this decision by the end of the Fall semester in their junior (3^{rd}) year.

VI. Infrastructure Support

HPU is fortunate enough to be able to support the acquisition of computing equipment infrastructure to support the development of a data science program. Currently the department has ample resources to support the existing programs in computer science and cybersecurity. The new data science program can leverage the existing programs to a significant extent, especially given that the data science program utilizes a number of existing computer science courses. However, to create an environment that fully supports the objectives of the data science program to prepare students for a professional life as a data scientist, experience with cloud computing and large data sets is necessary [7]. While we have a high performance computing cluster to support research, we have elected to build a cloud computing environment to support the data science program and researchers across campus. The cloud environment will support the courses and also host the large data sets used in the program and the domain areas. Researchers across campus will be encouraged to use the facility and store their data sets there and make them available to students within the program. This helps build the number of domain areas that the students can pair with the data science program, and provides students with a research opportunity.

VII. Conclusion

Data science is an emerging, interdisciplinary program which already has significant and growing demand for graduates in industry. HPU is developing a 4-year BS in data science and a 5-year double degree in data science and an application domain such as physics, biology, health sciences, or business. Both of these options are within the context of a strong liberal arts tradition that provides the graduates with soft/professional skills to complement their technical abilities. The challenge facing HPU, and other institutions currently exploring the development of data science programs at the bachelors level, is the uncertainty regarding the definition of the discipline itself, the accreditation requirements currently under development, and the changes that will likely occur in the model ACM/IEEE Computer Science curriculum.

References

[1] A. Danyluk, P. Leidig, S. Buck, L. Cassel, A. McGettrick, W. Qian, C. Servin, and H. Wang. (2019, January) Computing competencies for undergraduate data science curricula. [Online]. Available: http://dstf.acm.org/DSReportInitialFull.pdf

[2] A. Danyluk, P. Leidig, and L. Cassel, "An update on the ACM data science taskforce," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 635–636. [Online]. Available: https://doi.org/10.1145/3328778.3366995

[3] R. D. De Veaux, M. Agarwal, M. Averett, B. S. Baumer, A. Bray, T. C. Bressoud, L. Bryant, L. Z. Cheng, A. Francis, R. Gould, and et al., "Curriculum guidelines for undergraduate programs in data science," *Annual Review of Statistics and Its Application*, vol. 4, no. 1, p. 15–30, Mar 2017. [Online]. Available: http://dx.doi.org/10.1146/annurev-statistics-060116-053930

[4] Y. Demchenko, A. Belloum, W. Los, T. Wiktorski, A. Manieri, H. Brocks, J. Becker, D. Heutelbeck, M. L. Hemmje, and S. Brewer, "Edison data science framework: A foundation for building data science profession for research and industry," in *17th IEEE International Conference on Cloud Computing Technology and Science*. IEEE, November 2016, pp. 588–593.

[5] H. Karbasian and A. Johri, "Insights for curriculum development: Identifying emerging data science topics through analysis of Q&A communities," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 192–198. [Online]. Available: https://doi.org/10.1145/3328778.3366817

[6] National Academies of Sciences, Engineering, and Medicine, *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: The National Academies Press, 2018. [Online]. Available: https://www.nap.edu/catalog/24886/envisioning-the-data-science-discipline-the-undergraduate-perspective-interim-report

[7] A. Parrish, L. Jones, P. Leidig, and R. Raj, "Sculpting accreditation criteria for data science - a working session," Proceedings of the 51st ACM Technical Symposium on Computer Science Education, New York, NY, USA, 2020.

[8] ABET, Inc., "Criteria for accrediting computing programs. effective for review during the 2019-20 accreditation cycle," 2018, accessed: November 20, 2019. [Online]. Available: https://www.abet.org/wp-content/uploads/2018/11/C001-19-20-CAC-Criteria-11-24-18.pdf

[9] ACM/IEEE-CS Joint Task Force on Computing Curricula, "Computer science curricula 2013," ACM Press and IEEE Computer Society Press, Tech. Rep., December 2013, accessed: April 3, 2018. [Online]. Available: http://dx.doi.org/10.1145/2534860

[10] A. Clear, A. S. Parrish, J. Impagliazzo, and M. Zhang, "Computing curricula 2020: Introduction and community engagement," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 653–654. [Online]. Available: https://doi.org/10.1145/3287324.3287517

[11] J. Impagliazzo and A. N. Pears, "The CC2020 project — computing curricula guidelines for the 2020s," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, April 2018, pp. 2021–2024.

[12] J. C. Adams, "Creating a balanced data science program," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 185–191. [Online]. Available: https://doi.org/10.1145/3328778.3366800

[13] University of California Berkeley, "Data science academic resource kit," 2020. [Online]. Available: https://data.berkeley.edu/education/ark

[14] D. Deb, R. M. Smith, and M. Fuad, "Infusing data science across disciplines," in *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 302. [Online]. Available: https://doi.org/10.1145/3304221.3325579

[15] D. Deb and E. Jones, "University-wide adoption of data science," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1300. [Online]. Available: https://doi.org/10.1145/3328778.3372657

[16] I. B. Hassan and J. Liu, "Data science academic programs in the U.S," *J. Comput. Sci. Coll.*, vol. 34, no. 7, p. 56–63, Apr. 2019.

[17] B. Ramamurthy, "A practical and sustainable model for learning and teaching data science," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, ser. SIGCSE '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 169–174. [Online]. Available: https://doi.org/10.1145/2839509.2844603

[18] S. Rosenthal and T. R. Chung, "A data science major: Building skills and confidence," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 178–184. [Online]. Available: https://doi.org/10.1145/3328778.3366791

[19] E. Van Dusen, A. Suen, and C. Carson, "Innovation in undergraduate data science education," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1397. [Online]. Available: https://doi.org/10.1145/3328778.3367015

[20] J. Havill, "Embracing the liberal arts in an interdisciplinary data analytics program," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 9–14. [Online]. Available: https://doi.org/10.1145/3287324.3287436

[21] A. Clear, A. Parrish, S. Frezza, J. Impagliazzo, A. Pears, A. Takada, H. Topi, G. Van Der Veer, E. Cuadros-Vargas, A. Vichare, L. Waguespack, P. Wang, M. Zhang, H. Alrumaih, O. Bogoyavlenskaya, A. Decker, E. Durant, M. Exter, S. Gordon, E. Hayashiguchi, R. Le Blanc, P. Leidig, B. Lunt, B. McMillin, T. McVeety, L. Marshall, N. Mead, M. Moro, Simon, T. Bermadino, P. Tymann, B. Viola, and S. Zwebrn, "Computing curricula 2020." [Online]. Available: https://www.cc2020.net/

[22] A. Gautam, W. Bortz, and D. Tatar, "Abstraction through multiple representations in an integrated computational thinking environment," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 393–399. [Online]. Available: https://doi.org/10.1145/3328778.3366892

[23] D. J. Mir, S. Mishra, P. Ruvolo, L. Pollock, and S. Engen, "How do faculty partner while teaching interdisciplinary cs+x courses: Models and experiences," *J. Comput. Sci. Coll.*, vol. 32, no. 6, p. 24–33, Jun. 2017.

[24] R. H. Sloan, V. Barr, H. Bort, M. Guzdial, R. Libeskind-Hadas, and R. Warner, "CS + X meets CS 1: Strongly themed intro courses," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 960–961. [Online]. Available: https://doi.org/10.1145/3328778.3366975

[25] R. H. Sloan, C. Taylor, and R. Warner, "Initial experiences with a CS + law introduction to computer science (cs 1)," in *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 40–45. [Online]. Available: https://doi.org/10.1145/3059009.3059029

[26] T. Berger-Wolf, B. Igic, C. Taylor, R. Sloan, and R. Poretsky, "A biology-themed introductory CS course at a large, diverse public university," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 233–238. [Online]. Available: https://doi.org/10.1145/3159450.3159538

[27] V. Carr, M. Jones, and B. Wei, "Interdisciplinary computing: Applied computing for behavioral and social sciences," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 400–406. [Online]. Available: https://doi.org/10.1145/3328778.3366799