Fostering and Simplifying Data Science Transfer Pathways in Massachusetts

A Preprint

Benjamin S. Baumer *
Statistical & Data Sciences
Smith College
Northampton, MA 01063
bbaumer@smith.edu

Nicholas J. Horton
Mathematics & Statistics
Amherst College
Amherst, MA 01002
nhorton@amherst.edu

July 19, 2022

Abstract

A substantial fraction of students who complete their college education at a public university in the United States begin their journey at a two-year college. While the number of four-year colleges offering bachelor's degrees in data science continues to increase, data science instruction at many two-year colleges lags behind. A major impediment is the relative paucity of introductory data science courses that serve multiple student audiences and can easily transfer. In addition, the lack of pre-defined transfer pathways (or articulation agreements) for data science creates a growing disconnect that leaves students who want to study data science at a disadvantage. We present a "State of the Commonwealth" that describes transfer pathways among public colleges in Massachusetts. Five points of curricular friction merit attention: 1) the need for a first course in data science, 2) a second course in data science, 3) a course in scientific computing, data science workflow, and/or reproducible computing, 4) lab sciences, and 5) navigating communication, ethics, and application domain requirements in the context of general education and liberal arts course mappings. We explore barriers and opportunities to facilitate transfer options for students. We review existing transfer pathways in related disciplines, efforts to align curricula across institutions, and obstacles to overcome. We describe approaches to foster data science pathways and propose minimally-disruptive solutions. Improvements in these areas are critically important to ensure that a broad and diverse set of students are able to engage and succeed in these programs.

Keywords articulation · associate's programs · bachelor's programs · community colleges · course design · curriculum · data acumen · data analytics · two-year colleges

1 Introduction

Two-year colleges (also known as community colleges) play a critical role in higher education in Massachusetts, as well as within the United States generally. As of 2020, these 15 institutions with 29 campuses enroll more than 67,000 undergraduate students annually, similar in number to the University of Massachusetts system (57,000 during the same year). Two-year colleges provide associate's degrees that lead directly to employment, as well as options to transfer to bachelor's programs. With average annual tuition of $4,424 for in-state students, two-year colleges are the most effective and affordable option for many students. Blumenstyk (2021) describes two-year colleges as:

> "the keystone for the nation's plan to help more people earn a postsecondary credential."

*More information about the Massachusetts Data Science Pathways project can be found at https://dsc-wav.github.io/ma-ds-pathways

By all accounts, job prospects in data science are excellent, due to high salaries, expansive job growth, and comfortable working conditions. According to Glassdoor, data scientist is the #2 job in America for 2021, and has ranked among the top three every year since 2016. The US Bureau of Labor Statistics reports a mean annual wage of $103,930 for data scientists, and estimates that jobs will grow 22% for Computer and Information Research Scientists and 33% percent for Mathematicians and Statisticians over the next ten years. A number of companies have reported that they can't find sufficient skilled candidates for these positions (`http://oceansofdata.org/projects/mentoring-new-data-pathways-community-colleges`). Due to the nature of the work, data scientists have adapted smoothly to working remotely, an increasingly relevant factor that should only improve employment prospects. The high probability of financial success for graduates in data science stands in stark contrast to the increasingly dim prospects for many master's students in other fields. Korn and Fuller (2021) conclude that 38% of master's programs at top-tier private universities in the U.S. aren't worth the price of admission.

Providing equitable access to these desirable jobs is a challenge that is symptomatic of larger issues of class and income inequality in the United States. Several national reports (e.g., Rawlings-Goss et al. (2018), National Academies of Science, Engineering, and Medicine (2018), and Engineering, National Academies of Sciences, and Medicine (2016)) recognize this challenge and call for tighter partnerships between two- and four-year colleges. If the field of data science is serious about diversifying its workforce, then there must be paths to high-paying jobs in data science that begin at two-year colleges, which enroll a much larger fraction of historically under-served students than four-year colleges.

The recent National Science Foundation's (NSF) Data Science Corps (DSC) program focuses on creative approaches to developing a competitive and diverse workforce in data science. Through our roles as leaders of the NSF-funded DSC-WAV (Wrangle, Analyze, Visualize) program we have had the opportunity to engage in data science projects with community organizations and to work with partners at three local two-year colleges to foster new courses and programs. This work included organizing a symposium on Data Science at Massachusetts Two Year Colleges for academic leaders on June 13, 2022 and faculty development workshops in 2020, 2021, and 2022.

1.1 Our contribution

The purpose of this paper is to help to blaze a trail toward the development and implementation of multiple academic pathways to a bachelor's degree in data science. We ground this work on the perspective of a student entering one of the Commonwealth of Massachusetts' fifteen two-year colleges. We have focused on one state because the landscape of two-year colleges varies dramatically by state and Massachusetts is where we live and work. However, we believe that the insights and approaches we suggest may be useful to other states.

We begin by surveying the landscape of data science in higher education nationally and in Massachusetts (Section 2). At the time of this writing, the only full-blown bachelor's degree in data science offered by a *public* university in Massachusetts is the bachelor's of science degree in data science at UMass-Dartmouth. This program is described in Section A.3. At UMass-Amherst, students pursuing a bachelor of science degree in informatics can choose data science as one of two concentrations (see Section A.1) with other curricular offerings in other units in development.[2]

While we have these and other programs in mind, it is certain that many more data science bachelor's programs will arise in the coming years. Thus, our analysis focuses on a generic bachelor's program in data science that we see as likely to represent a curricular consensus.

In Section 3, we analyze the current state of pathways to bachelor's programs in data science beginning at two-year colleges in Massachusetts, with a particular emphasis on pathways that are facilitated by the MassTransfer program (operated by the Massachusetts Department of Higher Education, MDHE). The data analytics associate of science program at Bunker Hill Community College is the only current data science associate's program in Massachusetts (see also Section D.5). However, this program does not yet lead directly to a MassTransfer.

The lack of transfer pathways make a bachelor's degree in data science burdensome for a two-year college student to achieve without significant—and probably unreasonable—foresight and perseverance through

[2] A new bachelor of science in data science was approved at Westfield State University (Section A.4) in the summer of 2021. In addition, we understand that other data science related programs are in the planning stages at various campuses of the University of Massachusetts.

administrative and bureaucratic obstacles. Our analysis leads directly to recommendations that could provide explicit pathways in data science with relatively few new courses and modest impact on existing programs.

In Section 4, we mock-up several of the most promising pathways. Our short-term goal is to get at least one of these pathways approved and added to the MassTransfer website. Longer-term, we hope that this example pathway could serve as a proof-of-concept for future pathways from other two-year colleges in Massachusetts to offerings at other campuses of the University of Massachusetts and the nine State Universities. We believe that this initiative may help various entities coordinate their efforts as new data science bachelor's programs and pathways are created.

We conclude with final thoughts in Section 5.

Additional appendices describe data science programs in Massachusetts (Appendix A), learning outcomes for introductory data science courses (Appendix B), learning outcomes for data science programs (Appendix C), and related resources including the Dana Center Design Principles (Appendix D.1) and the Expanding Computing Education Pathways project (Appendix D.2).

2 Background and related work

2.1 Data science programs in higher education

Since Cleveland (2001)'s action plan for data science, the field has continued to blossom within academia. Academic data science can be aspirationally described using a pyramid, with doctoral degrees rare but important for leadership and research in the field. Master's degrees are the next level, with larger numbers and considerable job opportunities. For established disciplines, bachelor's programs (offered at four-year colleges) and associate's programs (offered at two-year colleges), make up the third and fourth levels of the pyramid, with larger and larger numbers of students obtaining these degrees. Jobs are available at each level, with the potential for interested students to pursue more advanced degrees to deepen skills and expand their work opportunities. However, workforce opportunities remain opaque to too many students.

As an emerging discipline, data science has not yet matured to that extent, with master's programs leading the way, bachelor's programs on the rise, and associate's program lagging behind.

Colleges and universities in Massachusetts have been significant drivers of innovation in this area. While several doctoral programs in data science now exist in the United States, one of the first was established at Worcester Polytechnic Institute. Innovative programs like the Data Science Internship Program through the Massachusetts Life Science Center are likely to increase workforce development opportunities for graduates at a variety of levels.

Far more common are master's programs in data science and data analytics, which are offered by many universities (both online and in-person), including Harvard, Northeastern, Tufts, and the University of Massachusetts-Amherst, which offers a master's degree in computer science with a concentration in data science through the Center for Data Science, which is itself located within the College of Computer Science.

Nationally, the growth in the number of master's degrees granted in analytics and data science is dramatic, with more than 45,000 degrees reported in 2020 by the Institute of Advanced Analytics.[3]

While the study of data science at the graduate level continues to evolve, its footprint is already substantial. The growth of these programs has made it possible for students at the undergraduate level to more easily identify future programs of graduate study. What undergraduate majors should best prepare a student for graduate study in data science? Computer science, statistics, and mathematics are the closest cognate disciplines, and while statistics is not always available as an undergraduate major, it is taught everywhere and can be folded into either a computer science or mathematics major, both of which are available at virtually any institution.

Historically rarer (but increasingly less so) are bachelor's degrees in data science and related fields (e.g., data analytics). These programs make up the next level of the pyramid, with larger numbers of students potentially entering the workforce. In Massachusetts, Boston University's program (which began enrolling students in the fall of 2021) may be one of the latest, but students at private colleges (e.g., Smith College, Mount Holyoke College) and public universities (e.g., UMass-Dartmouth) can now major in data science. (Appendix A catalogs bachelor's degrees in data science offered by public institutions in Massachusetts.) The

[3] According to Statista, there were more than a million associate's degree recipients in the United States during the 2018-2019 academic year.

options—which are certain to grow in the coming years—already provide two-year college students who are interested in data science with visible future programs of study.

Many workforce roles for data scientists exist at the bachelor's level (De Veaux et al. 2017; National Academies of Science, Engineering, and Medicine 2018), and the number is growing (Gould et al. 2018).

The bachelor's-to-master's transition is characterized by flexibility and adaptation, because graduate schools know that they will receive applications from students who attended a wide variety of undergraduate schools, and who studied highly variable subjects therein. Moreover, bachelor's programs typically involve at least 120 credit hours of study, which often provides ample flexibility for a student to deviate from any pre-defined curricular path. From our own experiences, we know that it is not uncommon for a traditional bachelor's student to major in say, economics, only to then decide before their senior year that they want to pursue a master's degree in data science, load up on statistics and computer science courses in their senior year, and still put together a competitive graduate school application.

It is important to remember that dramatically less flexibility is available for the associate's-to-bachelor's transition, since for two-year college students, every credit counts. **We recognize that for most two-year college students, any credit that doesn't count towards their associate's degree program or their pre-defined transfer pathway may be considered a "waste" of both time and money.** While the MassTransfer system provides a clear solution for existing pathways in Massachusetts, the larger difficulties with transfer pathways are longstanding (Blumenstyk 2021).

Longer-term, alternative options, including associate's-to-workforce programs (Rawlings-Goss et al. 2018; Gould et al. 2018) are desirable but outside the scope of this paper. Associate's programs in cybersecurity, information technology, and web development—designed as terminal degrees—have proven effective in workforce development and the same potential exists for data science.

2.2 Data science as an academic discipline

The field of data science continues to accrue markers of an established academic discipline, in addition to the various degree programs mentioned above.

De Veaux et al. (2017) provide curriculum guidelines for undergraduate majors in data science that are endorsed by the American Statistical Association. The "Data Science for Undergraduates: Opportunities and Options" consensus study (National Academies of Science, Engineering, and Medicine 2018) provided a number of recommendations and findings relevant to undergraduate data science programs and outlined key aspects of *data acumen*. The Association for Computing Machinery (ACM) Data Science Task Force enumerated computing competencies for undergraduate data science curricula (Danyluk et al. 2021), and syllabi from example courses. Comprehensive textbooks (Wickham and Grolemund 2016; Baumer, Kaplan, and Horton 2021) and course materials (Çetinkaya-Rundel 2020) support the teaching of a variety of different data science courses. Donoho (2017) ruminates on the nature of data science as a standalone scientific discipline.

In 2019, the National Center for Education Statistics unveiled a new series of Classification of Instructional Programs (CIP) codes for data science (30.70). These new codes allow the federal government to track the growth of programs in data science and should result in an improved ability to quantify how many students are studying data science.[4]

In what might be an important stamp of legitimacy, ABET (Accreditation Board for Engineering and Technology) has begun accrediting its first undergraduate data science programs, with plans to expand to the graduate and associate's levels.

2.3 The DSC-WAV project

While our interest in data science education is longstanding and well-documented, our specific interest in two-year college pathways in data science is motivated by our involvement in the Data Science Corps (DSC): Wrangle, Analyze, Visualize (WAV) project (Horton et al. 2021). The first arm of the NSF-funded program links teams of undergraduate students (often data science majors) at the Five Colleges (Amherst, Hampshire, Mount Holyoke, and Smith Colleges plus the University of Massachusetts-Amherst) with local,

[4]Until recently, the new CIP codes were not classified as STEM disciplines, which had negative implications for the immigration status of international students. Efforts by the Academic Data Science Alliance and others led to reclassification of the data science CIP code.

community-based organizations in the service of a real-world data science problem. Legacy et al. (2022) details how this program supports the growth of DSC-WAV student participants.

As the Data Science Corps is a workforce development initiative, the DSC-WAV project has an additional goal of growing and diversifying the data science workforce. In this fast-growing segment of the economy, highly-satisfying, high-paying jobs are plentiful. After several years of working closely with our partners at Holyoke, Greenfield, and Springfield Technical Community Colleges on a variety of curricular- and student-focused issues, our attention is now centered on the pathway predicament. *We believe that while the obstacles to transfer pathways in data science are formidable, we can overcome them with relatively non-disruptive changes.* Our current focus is to help broker transfer agreements between two-year colleges and public universities in Massachusetts.

3 "State of the Commonwealth"

Nationally, the "Data Science for Undergraduates" consensus report (National Academies of Science, Engineering, and Medicine 2018) recommended that "Academic institutions should provide and evolve a range of educational pathways to prepare students for an array of data science roles in the workplace." In addition, Gould et al. (2018) provided curricular guidelines for two-year college programs in data science, and identified six associate's degree programs in other states. Many others have been created.[5]

Where does this leave programs in Massachusetts? As noted in Section 1, as of late 2021, the data analytics option within the associate of science program at Bunker Hill Community College is the only two-year degree program in data science in Massachusetts. Students in this program take multiple courses in computer science, receive foundational training in statistics, linear algebra, and college writing, and are exposed to R, Python, and SQL. To the best of our knowledge, no other Massachusetts associate's degree programs exist in data science. Moreover, as of late 2021, no other two-year colleges in the Commonwealth offered even a formal course in data science. In this Section, we detail the major obstacles to transfer pathways to bachelor's programs in data science.

While informal agreements between individual programs may permit direct transfer from a two-year college to a four-year program, the gold standard is the MassTransfer system operated by the Massachusetts Department of Higher Education (MDHE). Their website allows any student to select one of the 15 two-year colleges, one of the 13 public universities, and an intended bachelor's degree field. The website will then return a list of "A2B Mapped Pathways," which have been pre-approved by MDHE for transfer. This approval sends important signals to prospective students that their academic plan is sound. A screenshot of the website shown in Figure 1, reveals that there are no approved pathways from Bunker Hill Community College (BHCC) to UMass-Dartmouth in "Other Sciences," even though we know from personal communication that students from BHCC's data analytics program have successfully transferred to UMD's data science program in recent years.

3.1 What's working

Transfer pathways for mature data science adjacent disciplines like mathematics and computer science are well established (see the MassTransfer A2B Degree maps at `https://www.mass.edu/masstransfer/a2b`). While we haven't investigated all $2 \cdot 13 \cdot 15 = 390$ possible pathways, many of them already exist with clearly described requirements and course mappings.

As a result of these pathways for adjacent disciplines, many courses exist and are easy to transfer. This includes mathematics and statistics courses relevant to data science (e.g., statistics, calculus, linear algebra, and discrete math) along with computer science courses (e.g., computer science I and II, data structures and algorithms). These existing course mappings provide a solid foundation for a transfer pathway in data science—but they are not enough.

For example, students at BHCC interested in mathematics can choose the Mathematics Concentration associate's program and find pathways to either the Applied and Computational Mathematics bachelor of science or the Mathematics bachelor of arts degrees at UMass-Dartmouth. While no transfer pathways in computer science are mapped between BHCC and UMass-Amherst or UMass-Dartmouth, students at Cape

[5]The Academic Data Science Alliance Data Science Institution Updates and the American Mathematical Association of Two-Year Colleges (AMATYC) Data Resources have created listings that are likely incomplete. Unfortunately, no comprehensive census of programs is readily available.

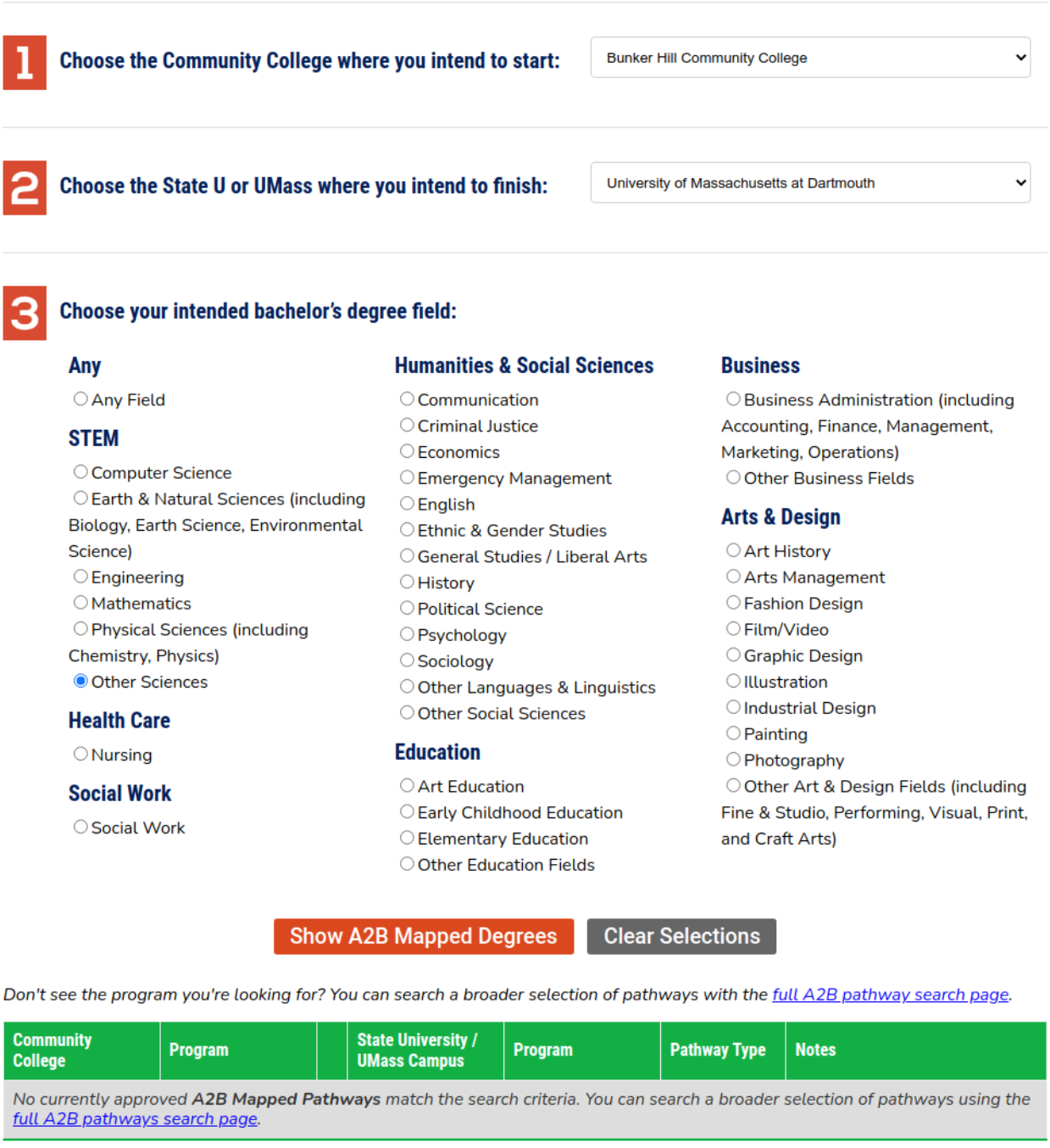

Figure 1: A screenshot from the MassTransfer website. Note that 'Data Science' does not appear as a bachelor's degree option. Moreover, selecting 'Other Sciences' results in no approved mapped pathways.

Cod Community College have three different liberal arts concentration options that map to the Computer Science BS degree at either the Amherst or Dartmouth campus. This does not imply that it is *impossible* for a student to attend BHCC and end up with a computer science bachelor's degree from UMass. It does, however, mean that a student pursuing that path would have to forge their own pathway, which might mean taking courses at BHCC that were outside of the requirements of their associate's degree program, taking catch-up courses at UMass once they arrive, and/or obtaining explicit transfer credit for courses that are not already mapped by MassTransfer. **All of these obstacles add unnecessary friction, cost, and time that students and society cannot afford.**

Another option for a two-year college student is to pursue a liberal arts transfer pathway. This strategy focuses the student experience at the two-year college level on obtaining university credit for general courses in broadly applicable fields. Once the student has transferred to a four-year institution, the experience becomes much more focused on their major. This may be an appropriate alternative pathway to consider in the future, though there may be disadvantages for students who decided not to transfer when entering the workforce with such a general degree.

3.2 Obstacles to transfer pathways

While data science lacks a national curriculum analogous to those in more established disciplines like computer science, mathematics, and statistics, a general framework for a bachelor's program in data science is taking shape. We have been involved in several efforts to shape such curriculum and accreditation guidelines at the national level, most notably including De Veaux et al. (2017); National Academies of Science, Engineering, and Medicine (2018); Gould et al. (2018). We have also been leaders in developing courses (Baumer (2015)), undergraduate majors, and textbooks (Baumer, Kaplan, and Horton (2021)) that support instruction in data science.

These experiences give us some standing to anticipate what the general landscape of bachelor's programs in data science will look like over the next ten years. The handful of bachelor's programs that we describe in Section A will likely double at least once in that time, and one of the major goals of this paper is to coordinate efforts across all parties such that the MassTransfer system is prepared to adapt to this changing landscape.

The bachelor's program in data science at UMass-Dartmouth (Figure 3) is a good target because it exists and it conforms reasonably well to other curricular guidelines in data science. However, our analysis is not specific to this program. To the contrary, it is based on the sum of our knowledge about generic data science bachelor's programs as they are likely to be in the coming years, given our understanding today. Our design is such that our analysis should be relevant to other programs that may be proposed (including the new program at Westfield State, any forthcoming programs at UMass-Amherst, or those developed at other institutions).

Generally, existing pathways in mathematics (which typically includes statistics as an elective), computer science, and the liberal arts provide usable mappings between two-year college courses and university courses that cover the majority of the credits and knowledge needed to transfer to a bachelor's program in data science. However, we have identified **five points of curricular friction**, which are listed here in decreasing priority and explored in further detail in the Sections that follow:

1. A first course in data science (Data Science I)
2. A second course in data science (Data Science II)
3. A course in scientific computing, data science workflow, and/or reproducible computing
4. Lab sciences
5. Navigating communication, ethics, and application domain requirements in the context of general education and liberal arts course mappings

Other sources of friction, such as institutional inertia, faculty development and retention, and technology, are no less real, but are not our focus in this paper.

3.2.1 Data Science I

Students at UMass-Dartmouth (Yan and Davis 2019), consistent with the recommendations of De Veaux et al. (2017) and Gould et al. (2018), take a first course in data science in the first semester of their first year. Unfortunately, with the exception of CIT-137 at BHCC, no two-year college in Massachusetts offers a

first course in data science.[6] It is not possible to imagine a sensible transfer pathway in which students are not exposed to the key ideas in data science until their junior year. Irrespective of pathways to degrees it is critically important that two-year college students have the opportunity to develop these skills.

Note that such a course is not simply a grab bag of existing material from existing courses in statistics and computer science, but rather focuses on new components of data acumen including the data science lifecycle, and historically underdeveloped skills like data wrangling and data visualization that support—but are not subsumed within—those existing courses. Some courses (e.g., Data 8 at Berkeley and Çetinkaya-Rundel (2020)) include elements of statistical modeling and inferential statistics, while others (e.g., SDS 192 at Smith) do not. In either case, a first course in statistics is a separate requirement that exists under many existing transfer pathways in mathematics.

In order for data science transfer pathways to work, two-year colleges must offer a first course in data science. This is by far the largest obstacle to bringing these pathways online and the place where the biggest gain will be achieved in helping institutions to make data science accessible to their students. This will help to partially address the recommendation from National Academies of Science, Engineering, and Medicine (2018) that: "To prepare their graduates for this new data-driven era, academic institutions should encourage the development of a basic understanding of data science in all undergraduates."

Many new introductory data science courses will be developed in the coming years, and it is vital that faculty at the bachelor's and associate's levels coordinate their efforts to ensure that explicit course mappings are created that will facilitate transfer.

> **Our recommendation** is that institutions develop a flexible, shared understanding of what constitutes a first course in data science, and that any new courses developed at any institution are designed with transfer mappings in mind.

Ideally, such a first course would:

1. have minimal prerequisites;
2. dovetail in useful ways with introductory computer science and statistics courses to allow students to take these foundational courses in any order;
3. transfer to a variety of programs at the bachelor's level, and;
4. satisfy a variety of distribution requirements, including the R2 analytical reasoning designation at UMass-Amherst.

At a high level, such a course should prepare students to demonstrate the ability to:

- use a general-purpose computational environment (e.g., Python or R) to analyze data
- scrape, process, clean, and wrangle data from various sources, including relational databases
- visualize and interpret relationships between variables in multidimensional data
- design accurate, clear, and appropriate data graphics
- communicate the results of an analysis in a correct and comprehensible manner
- collaborate within reproducible workflow
- assess the ethical implications to society of data-based research, analyses, and technology in an informed manner.

Appendix B provides a set of learning outcomes for a sample of introductory data science courses. New course structures should facilitate an inclusive and engaging learning environment for students.[7]

3.2.2 Data Science II

Cultivating a rich facility in data science requires repeated exposure: a single course is not sufficient for students to develop mastery. To help students along this path, bachelor's programs in data science typically include a second course in data science, often taken during the sophomore year. This course is intended to reinforce and extend fundamental skills in data wrangling, data visualization, statistical modeling, and predictive analytics. A richer treatment of data technologies and database querying in SQL may arise in such

[6] An achievement of the DSC-WAV project is the creation of a pilot first course in data science at Holyoke Community College (MTH 190), which is to be offered in the fall 2022 semester.

[7] Appendix D.1 includes a set of course design principles from the Dana Center that we suggest be incorporated in the course development process.

a course. The second course may be taught in a different language (e.g., Python) than the first course (e.g., R). The focus of the second course will vary from institution to institution depending on the focus of the first course (see Section 3.2.1), but we expect the general content areas to be similar to those listed above. The Data 100 course at Berkeley and the DSC 201 course at UMass-Dartmouth are examples of second courses in data science.

Second courses in data science obviously depend on a first course, and often build upon on other core requirements, which may include: a first course in programming, a first course in statistics, and/or linear algebra. These prerequisites have an impact on student pathways and may necessitate delaying completion of this course to the sophomore year.[8]

Given the difficulty of launching a first course in data science at two-year colleges, it may be best, especially in the short-term, to leave the second course in data science to the universities. While not optimal, it may be feasible for transfer students to take their second course in data science during the first semester of their junior year, and while this will likely disrupt their path relative to non-transfer students, that disruption can be minimized.

With appropriate planning, transfer students should be able to take some of their junior-level (upper division) courses (e.g., more advanced computer programming) as well as complete their general education requirement at their two-year college in place of a second course in data science. (A computer science class taught in an appropriate language might help develop their computational foundation and may allow transfer students to be in a stronger position to excel in their subsequent courses in data science.)

> **Our recommendation** is that, for the next few years, second courses in data science are left to bachelor's programs, and the credits are replaced with another course with an existing mapping. Planning should begin on course designs and frameworks for such a course to be taught at both two- and four-year institutions since this would support both students planning to transfer as well as associate's-to-workforce programs.

3.2.3 A course in scientific computing, data science workflow, and reproducible computing

A generic bachelor's program in data science will include explicit instruction in how to advance science by computing with data in a reproducible, collaborative workflow. In some programs, this instruction will be woven into modules that permeate a series of courses. In others, there will be a standalone course that focuses on these issues. It is important that the technologies to support workflow and reproducible analysis as a component of data acumen (National Academies of Science, Engineering, and Medicine 2018) should not be assumed to be known by students or left for them to learn outside of a course, lest existing disparities in background are exacerbated.

Topics in this area include version control systems (e.g., `git`), collaboration and project management tools (e.g., GitHub, Trello), software development paradigms (e.g., Agile/Scrum), document authoring software (e.g., variants of `markdown`, quarto, LaTeX), command line scripting (e.g., UNIX), cloud computing, as well as further exposure to R, Python, and/or SQL.

While there are existing models of such courses at two-year colleges in Massachusetts, they are less likely to have existing MassTransfer course mappings. Given the variety of topics in these courses and the difficulty of coordinating the content across institutions, these credits will probably have to be mapped on a one-to-one basis. One promising avenue is a course in R or Python that is outside of the main computer science sequence (which is often taught in Java or C++). An example of such a course is CSE 160 at Springfield Technical Community College (see Section 4.2).

> **Our recommendation** is that individual programs map credits where reasonably equivalent options exist, and replace them with general education or liberal arts credits where they don't.

3.2.4 Lab sciences

Many of the existing transfer options in computer science (and other STEM disciplines) require two semesters of lab sciences (e.g., physics, biology, or chemistry) as a component of their general education requirements.

[8] Some two-year college students may need to complete additional developmental math courses. Ideally, co-requisite approaches (see efforts by the Dana Center) could allow them to complete these requisites in a timely fashion without delaying their progress towards their associates degree.

Requiring a student pursuing a bachelor's degree in data science to take two semesters of physics, biology, or chemistry provides an opportunity for them to learn important aspects of the scientific process as well as the collection and analysis of data. At present, many of these courses may be less germane for data science students, but there is considerable potential for them to reinforce and build basic data sciences skills for all students while building domain knowledge.

As an alternative to explore, we can imagine that a future data science infused lab course could be developed as a way to provide more exposure to key data science topics while meeting the learning outcomes for a lab course.

> **Our recommendation** is that students use existing pathways for lab sciences, choosing courses when possible that incorporate aspects of scientific data (e.g., Greenfield Community College's BIO 120 Introduction to Environmental Science)[9].

3.2.5 Communication, ethics, and application domains

Bachelor's programs in data science include training in communication (how do we transfer knowledge gained from data analysis from data scientist to a broader audience? (Parke 2008)) and ethics (what responsibilities to data scientists have to their users, customers, and society as a whole? (Baumer et al. 2022)). In addition, a domain of application is valuable (how does data science enhance our understanding of another subject?). These vital aspects of a data science curriculum cannot wait entirely until the junior year, and thus, two-year college students must find ways to build skills in these areas before they transfer.

Most two-year colleges offer courses in communication. If any of those courses focus on *communicating with data*, they should be taken. Courses that focus on more general writing skills are still valuable, and are already part of the general education requirements for any associate's degree. Where courses in ethics, or preferably, data ethics are available, they should be taken at the two-year college level, as this will help to infuse ethics early in a student's education.

For those students whose application domain will intersect with the lab sciences mentioned in Section 3.2.4, that requirement might provide a helpful synergy. We imagine that this might be particularly beneficial for students interested in public health, biostatistics, or bioinformatics.

One challenge here will be ensuring that whatever these courses are, they count towards the associate's degree program.

> **Our recommendation** is that institutions think carefully and holistically about how requirements for communication, ethics, and domain application can be used to accrue credits at two-year colleges and foster successful transfers.

3.3 Where to situate programs?

Unfortunately, the interdisciplinary nature of data science is in conflict with the siloing of programs within departments. The NASEM 2018 report found that many bachelor's degree programs in data science are housed in a college or school of business, a mathematics or statistics department, or a computer science department (see pages 3–5 of National Academies of Science, Engineering, and Medicine (2018)). A few undergraduate data science majors were described as hybrids of these three models, with joint administration/programmatic coordination. We believe that such hybrid models are better suited to ensure that students develop a deep foundation in all aspects of data acumen.

When considering where to situate associate's degree programs within departments at two-year colleges, the compressed timeline given the two-year nature of the degree only compounds the problem. As a result, until there are associate's degree programs in data science, even explicit transfer pathways (such as the ones we are trying to create) may force students to choose between two potentially undesirable options: obtaining an associate's degree in liberal arts studies that may not be as marketable as a degree in a more technical field, or supplementing a degree in mathematics or computer science with several additional courses. Our hope is to provide guidance about flexible pathways that could soften these rough edges that exist at present.

[9] Shodor (`https://http://shodor.org`) has worked to incorporate data science into various STEM majors, including biology, chemistry, and physics.

4 New transfer pathways in data science

In this section, we describe three proposed pathways in data science at specific pairs of institutions.

We hope that these proposals will help to identify generic pathways that can be operationalized at many two-year colleges and universities.[10]

4.1 Bunker Hill to UMass-Dartmouth

We believe all the pieces for a data science transfer pathway from Bunker Hill Community College (BHCC) to UMass-Dartmouth are already in place. Although students who attend BHCC may be more interested in staying in Boston and transferring to Northeastern, a public option at UMD would be the first potential MassTransfer pathway in data science. Figure 2 presents a mock-up of this pathway. Most of the course mappings in Figure 2 are already approved by the MassTransfer system. In what follows, we provide detail about the exceptions.

The mapping of two courses in data science are not approved. As noted previously, the data science courses at BHCC are unique among two-year colleges. We hope that this paper will lead directly to conversations among relevant faculty and MDHE that will result in an approved mapping for these courses.

The mapping of CSC 125 to MTH 280 is not approved. However, it appears to us that the content of the two courses is similar enough that a mapping could be approved.

The mapping for discrete math is not approved. It is not entirely clear that the MAT 171 course at BHCC meets the requirements of MTH 181 at UMass-Dartmouth. If this mapping is not viable, there are numerous other math courses at BHCC that might suffice.

Using only courses that already exist, we see this as the best candidate to be the first data science transfer pathway. We must note, however, that the collection of courses taken at BHCC is quite different than the existing data analytics concentration, and thus a new associate's degree (data science transfer option) at BHCC might need to be created.

4.2 Springfield Technical to UMass-Dartmouth

Figure 4 shows a mock-up of what a MassTransfer pathway in data science from Springfield Technical Community College to UMass-Dartmouth might look like. Because STCC has existing MassTransfer pathways in computer science, mathematics, and liberal arts/general studies, most of these course mappings are already approved.

The three exceptions are:

1. **a first course in data science**. In this case, DSC 101 at UMass-Dartmouth (see Figure 3), which is taught in the first semester of the first year. We describe in Section 3.2.1 our recommendation is that STCC simply has to create a new course that will map to DSC 101 in order for any pathway to work.
2. **a second course in data science**. In this case, DSC 201 at UMass-Dartmouth, which is taught in the fall semester of the second year. Since this course is only offered in the fall, and in light of our reasoning in Section 3.2.2, our recommendation is that this requirement be replaced by another course. There are several options. ENG 104 (Technical Report Writing) and ENG 110 (English Composition 2: Journalism) are existing courses at STCC that are already approved for general education transfer. The former seems most directly comparable to ENL 266 (Technical Communications) at UMass-Dartmouth, and the latter might provide a relevant alternative for students interested in data journalism. Either of these courses would replace the three missing credits. Although DSC 201 is a prerequisite for CIS 360, faculty at UMass-Dartmouth are confident that transfer students could take DSC 201 and CIS 360 concurrently in the fall of their junior year as long as they have prior experience with Python. One way to ensure this would be to have the new first course in data science at STCC taught in Python. The point becomes moot if the scientific computing course is taught in Python (see below).

[10] In the fall of 2021, Bunker Hill Community College signed an articulation agreement with Northeastern University from their associate of science in data analytics degree to Northeastern's bachelor of science in analytics degree program. This agreement serves as a proof-of-concept that transfer pathways to universities in Massachusetts can be created.

Area of Study	Credits	Bunker Hill Community College	To	University of Massachusetts Dartmouth	Degree Requirements Fulfilled by Course(s)
Mathematics Credits					
Calculus I	4	**MAT 281** Calculus I *(4 credits)*	=	**MTH 151** Calculus I *(4 credits awarded)*	MAJOR GEN ED
Calculus II	4	**MAT 282** Calculus II *(4 credits)*	=	**MTH 152** Calculus II *(4 credits awarded)*	MAJOR
Discrete Mathematics	3	**MAT 171** Finite Mathematics *(3 credits)*	=	**MTH 181** Discrete Mathematics I *(3 credits awarded)*	MAJOR
** Calculus I will be waived as a requirement for students who complete the approved Calculus II course and 60 transferrable credits prior to transfer, and Calculus II will be waived for students who complete the approved Calculus III course and 60 transferrable credits prior to transfer.*					
Statistics	3	**MAT 181** Statistics *(3 credits)*	=	**MTH 231** Elementary Statistics I *(3 credits awarded)*	MAJOR
Linear Algebra	4	**MAT 291** Linear Algebra *(4 credits)*	=	**MTH 221** Linear Algebra *(3 credits awarded)*	MAJOR
Note: Students must earn a grade of "C" or higher in all of the above courses to transfer to the UMass Dartmouth mathematics program					
Scientific Computing	3	**CSC 125** Python Programming*(3 credits)*	=	**MTH 280** Introduction to Scientific Computation *(3 credits awarded)*	MAJOR
Lab Science Credits					
Physics I & II (Calculus-Based)	8	**PHY 251** College Physics I/Lab *(4 credits)* & **PHY 252** College Physics II/Lab *(4 credits)*	=	**PHY 113** Classical Physics I *(4 credits awarded)* & **PHY 114** Classical Physics II *(4 credits awarded)*	MAJOR GEN ED
Computer Science Credits					
Computer Science Core Block	11	**CSC 120** Intro to Computer Science and Object Oriented Programming *(4 credits)* & (**CSC 237** C++ Programming *(4 credits)* OR **CSC 239** Java Programming *(4 credits)*) & **CSC 242** Data Structures *(3 credits)*	=	**CIS 180** Object-Oriented Prog I *(4 credits)* & **CIS 181** Object-Oriented Prog II *(4 credits)* & **CIS 280** Software Specifications and Design *(4 credits)* &	MAJOR
Data Science Credits					
Data Science I & II	7	**CIT 137** Introduction to Big Data with R and R Studio *(4 credits)* & **CIT 187** Data Analytics and Predictive Analysis *(3 credits)*	=	**DSC 101** Introduction to Data Science *(3 credits awarded)* & **DSC 201** Data Analysis and Visualization *(3 credits awarded)*	MAJOR GEN ED
Other Credits					
English Composition/Writing	6	>> View eligible courses			GEN ED
Behavioral/Social Science	6	>> View eligible courses			GEN ED
Humanities/Fine Arts	3	>> View eligible courses			GEN ED
Any Area	Remaining Balance	Students should talk to the advising contacts listed above to ensure that their course selections fulfill both MassTransfer Mapped and associate degree requirements. Minimum grade requirements are in accordance with MassTransfer and institutional requirements.			MAJOR
62 credits completed at Bunker Hill Community College Data Science Transfer (AA)			=	**61 credits transferring to University of Massachusetts Dartmouth** Guaranteed admission *(with 2.50+ GPA and space permitting)* to Bachelor of Science in Mathematics	

Figure 2: Mock-up of a transfer pathway in data science from Bunker Hill Community College to UMass-Dartmouth.

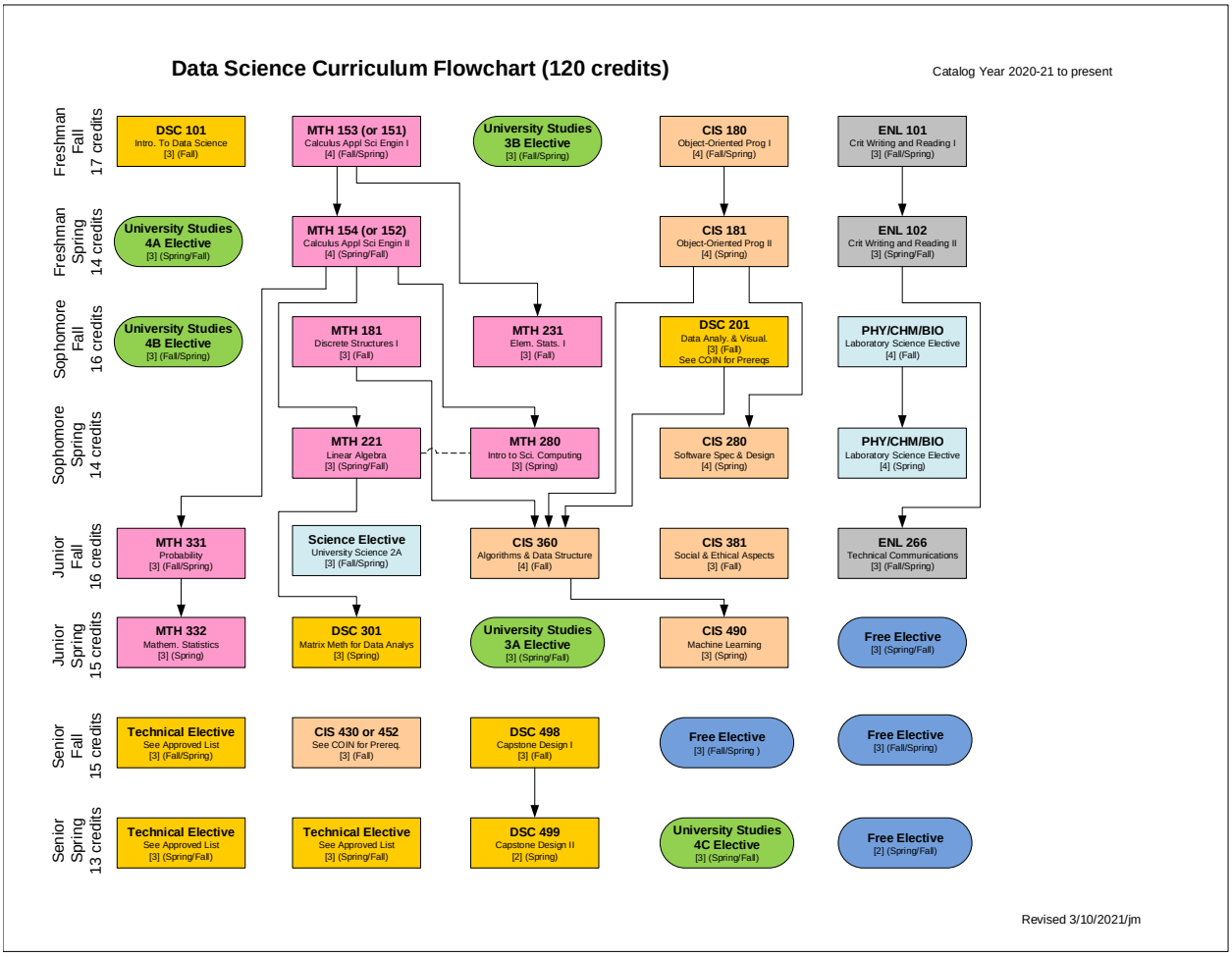

Figure 3: Flowchart illustrating progress through the data science bachelor's program at UMass-Dartmouth.

3. **a course in scientific computing**. In this case, MTH 280 at UMass-Dartmouth focuses on scientific computing in Python, and includes learning outcomes involving `git`, LaTeX, and Jupyter notebooks (see Section 3.2.3). STCC has a similar course on the books already: CSE 160 Introduction to Programming Using Python. While this mapping is not currently approved, we hope that it could be.

Thus, we are reasonably confident that a data transfer pathway from STCC to UMass-Dartmouth could be approved if STCC were to offer a suitable first course in data science.

4.3 Holyoke to UMass-Amherst

Figure 5 shows a flowchart of progression through the Informatics degree at UMass-Amherst.

Figure 6 shows a mock-up of a transfer pathway from Holyoke Community College to the informatics program with a data science concentration at UMass-Amherst. Such a pathway would allow transfer students access to the new S-STEM funded program Boosting Access to Data Science Scholars.

4.4 Transfer pathways in other states and disciplines

While we have focused on pathways in Massachusetts, considerable progress has been made in other states. Similar efforts are underway in California. While the scale of the California system—which includes both the UC and Cal State constellations—provides obvious challenges, there have been encouraging developments. Models for addressing similar challenges in related disciplines (e.g., engineering) exist (Enriquez et al. 2018). We are confident that the smaller scale of pathways in Massachusetts will be more tractable.

Area of Study	Credits	Springfield Technical Community College	To	University of Massachusetts Dartmouth	Degree Requirements Fulfilled by Course(s)
Mathematics Credits					
Calculus I & II	8	**MAT 131** Calculus 1 *(4 credits)* & **MAT 132** Calculus 2 *(4 credits)*	=	**MTH 151** Calculus I *(4 credits awarded)* & **MTH 152** Calculus II *(4 credits awarded)*	MAJOR GEN ED
	**Calculus I will be waived as a requirement for students who complete the approved Calculus II course and 60 transferrable credits prior to transfer.*				
Discrete Mathematics	4	**MAT 220** Discrete Structures*(4 credits)*	=	**MTH 181** Discrete Mathematics I *(3 credits awarded)*	MAJOR GEN ED
Statistics	3	**MAT 115** Statistics*(3 credits)*	=	**MTH 231** Elementary Statistics I: Exploratory Data Analysis *(3 credits awarded)*	MAJOR GEN ED
Linear Algebra	3	**MAT 240** Linear Algebra *(3 credits)*	=	**MTH 221** Linear Algebra *(3 credits awarded)*	MAJOR
Scientific Computing	3	**CSE 160** Introduction to Programming Using Python*(3 credits)*	=	**MTH 280** Introduction to Scientific Computation *(3 credits awarded)*	MAJOR
Lab Science Credits					
Physics I & II (Calculus-Based)	8	**PHY 231/231L** Classical Physics 1 + Lab: Classical Physics 1 *(3+1 credits)* & **PHY 232/232L** Classical Physics 2 + Lab: Classical Physics 2 *(3+1 credits)*	=	**PHY 113** Classical Physics I *(4 credits awarded)* & **PHY 114** Classical Physics II *(4 credits awarded)*	MAJOR GEN ED
Computer Science Credits					
Computer Science Core Block	12	**CSC 111/111L** Introduction to the Java Programming Language + Lab: Introduction to the Java Programming Language *(3+1 credits)* & **CSC 112/112L** Intermediate Topics in Java Programming + Lab: Intermediate Topics in Java Programming *(3+1 credits)* & **CSC 220/220L** Data Structures and Algorithms + Lab: Data Structures and Algorithms *(3+1 credits)*	=	**CIS 180** Object-Oriented Prog I *(4 credits)* & **CIS 181** Object-Oriented Prog II *(4 credits)* & **CIS 280** Software Specifications and Design *(4 credits)* &	MAJOR
Data Science Credits					
Data Science I & II	6	**XXX ???** Introduction to Data Science *(3 credits)* & **XXX ???** Data Visualization Tools *(3 credits)*	=	**DSC 101** Introduction to Data Science *(3 credits awarded)* & **DSC 201** Data Analysis and Visualization *(3 credits awarded)*	MAJOR GEN ED
Other Credits					
English Composition/Writing I & II	6	>> View eligible courses			GEN ED
Behavioral/Social Science Electives	6	>> View eligible courses			GEN ED
Humanities/Fine Arts Electives	3	>> View eligible courses			GEN ED
Any Area	Remaining Balance	Students should talk to the advising contacts listed above to ensure that their course selections fulfill both MassTransfer Mapped and associate degree requirements. Minimum grade requirements are in accordance with MassTransfer and institutional requirements.			MAJOR
62 credits completed at Springfield Technical Community College Engineering and Science Transfer: Computer Science Transfer (AS)			=	**61 credits transferring to University of Massachusetts Dartmouth** Guaranteed admission *(with 2.50+ GPA and space permitting)* to Bachelor of Science in Data Science	

Figure 4: Mock-up of a transfer pathway in data science from Springfield Technical Community College to UMass-Dartmouth. Note that courses at STCC with course numbers XXX do not yet exist.

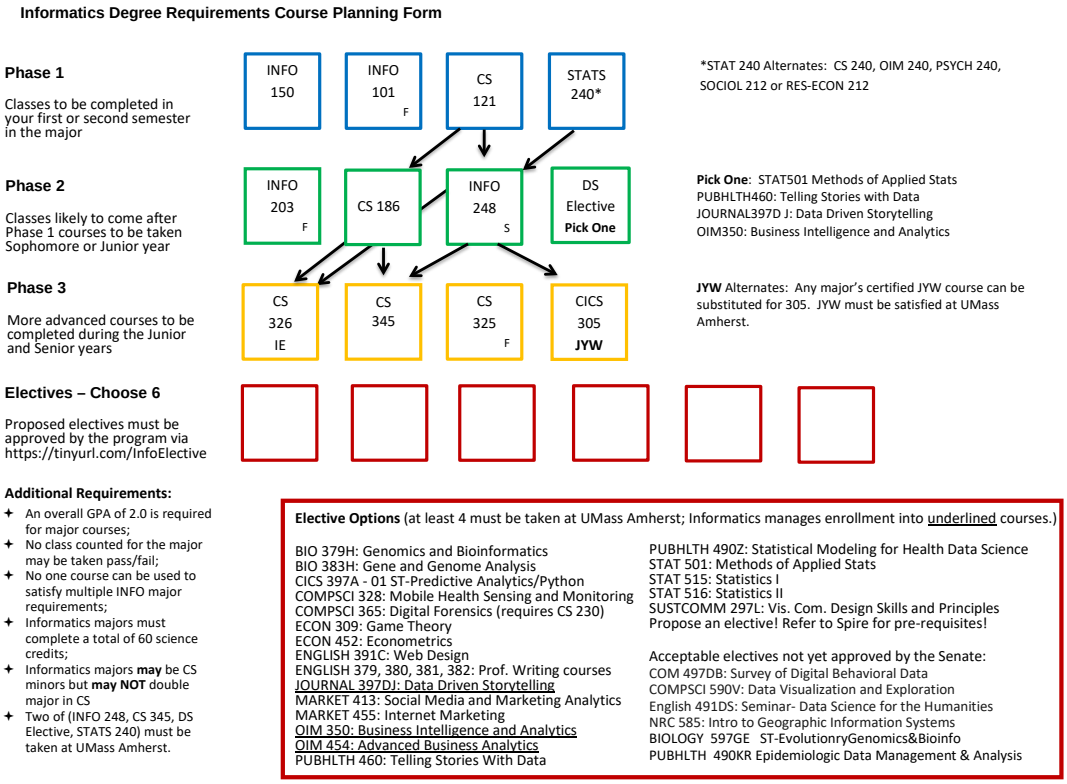

Figure 5: The Informatics major at UMass-Amherst.

5 Closing thoughts

At the December 2018 meeting of the National Academies Postsecondary Data Science Education Roundtable (National Academies of Sciences, Engineering, and Medicine 2020), D.J. Patil, former Chief Data Scientist in the White House Office of Science and Technology Policy, described how his experience from a two-year college bestowed upon him "three gifts": "a love of mathematics, an understanding of how to write in various genres, and confidence to succeed at the postsecondary level (page 158)". He expressed that his experience at two-year college provided a crucial "on-ramp" to his future success in data science.

Like Patil, we see two-year colleges as key players in developing the next generation of data science students. Our experience with the DSC-WAV project and other interactions have shown that our two-year college system has countless committed and engaged educators and administrators working to build better futures for their students, amidst time and resource constraints.

There's considerable work needed to foster sustainable courses, structures, and programs. We acknowledge that this will require focus and attention for many years. Efforts such as the NSF-funded EDC Oceans of Data "Mentoring New Data Pathways" project (`http://oceansofdata.org/projects/mentoring-new-data-pathways-community-colleges`) have engaged Bunker Hill Community College in an effort to support new data programs (see D.5).

Resource disparities at many two-year colleges and insufficient partnerships between two- and four-year institutions could hamper these efforts. As but one example, due to resources and other circumstances, two-year colleges could feel at a disadvantage and perhaps be reluctant to offer courses that are not included in guaranteed transfer systems such as MassTransfer, or courses not belonging to an already structured pathway.

Area of Study	Credits	Holyoke Community College	To	University of Massachusetts Amherst	Degree Requirements Fulfilled by Course(s)
Major Foundational Credits					
Calculus I	4	**MTH 113** Calculus I *(4 credits)*	=	**MATH 131** Calculus I *(4 credits awarded)*	MAJOR GEN ED
Calculus II	4	**MTH 114** Calculus II *(4 credits)*	=	**MATH 132** Calculus II *(4 credits awarded)*	MAJOR
Calculus III	4	**MTH 213** Calculus III *(4 credits)*	=	**MATH 233** Multivariate Calculus *(4 credits awarded)*	MAJOR
Discrete Math	4	**MTH 230** Discrete Mathematics *(4 credits)*	=	**INFO 150** A Mathematical Foundation for Informatics *(3 credits awarded)*	MAJOR
Linear Algebra	4	**MTH 205** Linear Algebra *(4 credits)*	=	**MATH 235** Intro to Linear Algebra *(4 credits awarded)*	MAJOR
Statistics	3	**MTH 245** Probability and Statistics for Engineers & Scientists*(3 credits)*	=	**STAT 240** Introduction to Statistics *(4 credits awarded)*	MAJOR GEN ED
Computer Science Credits					
Computer Science Core Block	8	**CSI 106** Programming Fundamentals I *(4 credits)* & **CSI 258** Data Structures *(4 credits)*	=	**COMPSCI 121** Introduction to Problem Solving with Computers *(4 credits)* & **COMPSCI 186** Using Data Structures *(4 credits)* &	MAJOR
Informatics	8	**CSI 111** Computer Concepts with Applications *(4 credits)* & **CSI 218** Programming Fundamentals II *(4 credits)*	=	**INFO 101** Introduction to Informatics *(3 credits)* & **INFO 203** A Networked World *(3 credits)*	MAJOR
Data Science Credits					
Data Science I & II	6	**XXX ???** Introduction to Data Science *(3 credits)* & **XXX ???** Data Visualization Tools *(3 credits)*	=	**INFO 248** Introduction to Data Science *(3 credits awarded)* & **DSC 201** Data Analysis and Visualization *(3 credits awarded)*	MAJOR GEN ED
Other Credits					
English Composition/Writing	6	>> View eligible courses			GEN ED
Behavioral/Social Science	6	>> View eligible courses			GEN ED
Humanities/Fine Arts	3	>> View eligible courses			GEN ED
Any Area	Remaining Balance	Students should talk to the advising contacts listed above and *follow Holyoke Community College requirements in Degree Works* to ensure that their course selections fulfill both MassTransfer Mapped and associate degree requirements. Minimum grade requirements are in accordance with MassTransfer and institutional requirements.			MAJOR
		RECOMMENDED:			
Advising Notes: **1. Students need to earn a grade of C- or better in Math 132 before taking certain courses at the 300 level or higher once they transfer to UMass Amherst.**					
60 credits completed at Holyoke Community College Data Science Transfer Option (AS)			=	**58 credits transferring to University of Massachusetts Amherst** Guaranteed admission *(with 2.50+ GPA and space permitting)* to Bachelor of Science in Informatics (Data Science Concentration)	

Figure 6: Mock-up of a transfer pathway in data science from Holyoke Community College to UMass-Amherst informatics. Note that courses at HCC with course numbers XXX do not exist.

There are some useful models that we can consider. In Ohio, 36 public institutions of higher education, 27 two-year colleges, and 9 four-year colleges approved a set of learning outcomes for a general education data science course developed by faculty from two- and four-year institutions (Ricardo Moena, personal communication). We see this as a necessary but not sufficient step.

Faculty development is another critical issue. At a time when data science positions are challenging for employers to fill, where will the next generation of instructors come from? This is another area where partnerships between two- and four-year institutions as well as industry will be critical (see Enriquez et al. (n.d.) for strategies for engineering transfer programs).

The changing preK-12 landscape raises important questions. As states are reviewing and revising their mathematics, science, and computing standards, statistics and data science are being elevated and made more explicit. We believe that this will impact the knowledge, skills, and abilities students bring to their post-secondary education. These changes may impact the future of pathways, potentially in positive ways.

There are many other issues that we could address at this juncture, including aspects of associate's to workforce programs, challenges and opportunities of dual enrollment, and the pressing need for improved computational infrastructure. But we intentionally limit our primary focus to fostering pathways, which needs to begin by identifying barriers and resources to the widespread teaching of accessible and pedagogically sound introductory data science courses.

5.1 A call to broaden participation

We close with some reflections on the critical role that two-year colleges provide in terms of **affordable** options that are accessible to a **diverse** population.

The Broadening Data Science Education (Rawlings-Goss et al. 2018) report notes that:

> Many individuals in today's data science workforce are coming from doctoral or master's degree programs, which have seen a dramatic increase in recent years. While these advanced degrees are valuable, it is not economically feasible for all data scientists to complete four years of an undergraduate degree, then a one- or two-year master's program before they can undertake useful work. Ensuring the future growth of the workforce requires an expansion to four-year and two-year degrees (page 45).

At the June 2019 NASEM Roundtable meeting, Uri Treisman of the University of Texas-Austin and the Dana Center described data science programs as "powerful resources for students seeking upward mobility (page 165)." (National Academies of Sciences, Engineering, and Medicine 2020)

Moreover, the Broadening Data Science Education (Rawlings-Goss et al. 2018) report suggests that: "the potential impact of the Data Divide is no less dire for our institutions of higher education" (page 7). Such concerns lead to the finding that: "Data science would particularly benefit from broad participation by underrepresented minorities because of the many applications to problems of interest to diverse populations." (National Academies of Science, Engineering, and Medicine 2018) The California Alliance for Data Science Education notes that "increasing access to data science as a career option for all students is key to making data science a more diverse and inclusive field." The Broadening Data Science Education (Rawlings-Goss et al. 2018) report states this even more directly:

> If we do not make diversity and inclusion a priority now, we will not have it in the future. We do not want to repeat the mistakes of the past, so we must reverse the trend for the growing divide to make and keep data science broad. Diversity will bring a lot of ideas and voices to the table, which may lead to significantly fewer models producing biased results when trained using algorithms on biased data sets. (page 30).

We agree that two-year colleges are the only affordable game in town and serve a key role in data science education now and in the future.

6 About the authors

Benjamin S. Baumer and Nicholas J. Horton serve as Principal Investigators of the NSF-funded DSC-WAV project. They have written papers on data science education and a textbook: *Modern Data Science with R*.

Benjamin S. Baumer is an Associate Professor of Statistical and Data Sciences at Smith College. He works with ABET/CSAB on the development of accreditation for data science programs. Ben was a member of the team which developed the Park City Mathematics Institute curriculum guidelines for data science programs, which are endorsed by the American Statistical Association.

Nicholas J. Horton is Beitzel Professor of Technology and Society (Statistics and Data Science) at Amherst College. He is vice-president of the American Statistical Association and co-chair of the National Academies Committee on Applied and Theoretical Statistics. He served as a member of the National Academies Data Science Roundtable, the "Data Science for Undergraduates" consensus report, the "Keeping Data Science Broad" report, and the "Two-Year College Data Science" report.

7 Acknowledgements

We acknowledge the many efforts of DSC-WAV Project Coordinator Andrea Dustin, our many collaborators and students on the project, as well as financial support from NSF grants HDR DSC-1923388 and HDR DSC-1924017. We appreciate the input and efforts of the co-PIs from our local two-year colleges: Ileana Vasu (Holyoke), Ebenezer Afarikumah (Greenfield), and Brian Candido (Springfield Technical). We thank Brant Cheikes, Matthew Rattigan, Tom Bernadin, Michelle Trim, Scott Field, and Iren Valova for sharing their thoughts and suggestions. Sarah Dunton, Jenn Halbleib, Michael Harris, Tyler Kloefkorn, Kate Kozak, Donna LaLonde, Sears Merritt, Ricardo Moena, Roxy Peck, Josh Recio, Rachel Saidi, and Rebecca Wong provided many helpful comments and suggestions on an earlier draft of the manuscript.

8 References

Baumer, Ben. 2015. "A Data Science Course for Undergraduates: Thinking with Data." *The American Statistician* 69 (4): 334–42. https://doi.org/10.1080/00031305.2015.1081105.

Baumer, Benjamin S., Randi L. Garcia, Albert Y. Kim, Katherine M. Kinnaird, and Miles Q. Ott. 2022. "Integrating Data Science Ethics into an Undergraduate Major: A Case Study." *Journal of Statistics and Data Science Education* 30 (1). https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2038041.

Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. 2021. *Modern Data Science with R*. 2nd ed. Chapman; Hall/CRC Press: Boca Raton. https://www.routledge.com/Modern-Data-Science-with-R/Baumer-Kaplan-Horton/p/book/9780367191498.

Blumenstyk, Goldie. 2021. "The Edge: The 'Dirty Secret' That Obstructs Transfer." The Chronicle of Higher Education. https://www.chronicle.com/newsletter/the-edge/2021-11-10.

Cleveland, William S. 2001. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *International Statistical Review* 69 (1): 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x.

Çetinkaya-Rundel, Mine. 2020. "Data Science in a Box." https://datasciencebox.org/. https://datasciencebox.org/.

Danyluk, A, P Leidig, S Buck, L Cassel, A McGettrick, W Qian, C Servin, and H Wang. 2021. "Computing Competencies for Undergraduate Data Science Curricula." Association for Computing Machinery; Association for Computing Machinery. https://dstf.acm.org/DSTF/_Final/_Report.pdf.

De Veaux, Richard D., Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, et al. 2017. "Curriculum Guidelines for Undergraduate Programs in Data Science." *Annual Review of Statistics and Its Application* 4 (1): 1–16. https://doi.org/10.1146/annurev-statistics-060116-053930.

Donoho, David. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26 (4): 745–66. https://doi.org/10.1080/10618600.2017.1384734.

Engineering, National Academy of, National Academies of Sciences Engineering, and Medicine. 2016. Edited by Shirley Malcom and Michael Feder. Washington, DC: The National Academies Press. https://doi.org/10.17226/21739.

Enriquez, A., N. Langhoff, E. Dunmire, T Rebold, and W. Pong. n.d. "Strategies for Developing, Expanding, and Strengthening Community College Engineering Transfer Programs." *American Society for Engineering Education* 2018. https://par.nsf.gov/biblio/10063235.

Enriquez, A, Nicholas Langhoff, E Dunmire, Thomas Rebold, and Wenshen Pong. 2018. "Strategies for Developing, Expanding, and Strengthening Community College Engineering Transfer Programs." In *American Society for Engineering Education*, 2018:16. https://doi.org/10.18260/1-2--30995.

Gould, Rob, R Peck, J Hanson, N J Horton, Brian Kotz, K Kubo, J Malyn-Smith, et al. 2018. "The Two-Year College Data Science Summit." American Statistical Association. https://www.amstat.org/asa/files/pdfs/2018TYCDS-Final-Report.pdf.

Horton, Nicholas J., Benjamin S. Baumer, Andrew Zieffler, and Valerie Barr. 2021. "The Data Science Corps Wrangle-Analyze-Visualize Program: Building Data Acumen for Undergraduate Students." *Harvard Data Science Review* 3 (1): 1–8. https://doi.org/10.1162/99608f92.8233428d.

Korn, Melissa, and Andrea Fuller. 2021. " 'Financially Hobbled for Life': The Elite Master's Degrees That Don't Pay Off." The Wall Street Journal. https://www.wsj.com/articles/financially-hobbled-for-life-the-elite-masters-degrees-that-dont-pay-off-11625752773.

Legacy, Chelsey, Andrew Zieffler, Benjamin S. Baumer, Valerie Barr, and Nicholas J. Horton. 2022. "Facilitating Team-Based Data Science: Lessons Learned from the DSC-WAV Project." *Foundations of Data Science*. https://doi.org/10.3934/fods.2022003.

National Academies of Science, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. National Academies Press: Washington, DC. https://nas.edu/envisioningds.

National Academies of Sciences, Engineering, and Medicine. 2020. "Roundtable on Data Science Postsecondary Education." American Statistical Association. https://www.nap.edu/25804.

Parke, Carol S. 2008. "Reasoning and Communicating in the Language of Statistics." *Journal of Statistics Education* 16 (1). https://doi.org/10.1080/10691898.2008.11889555.

Rawlings-Goss, R, L Cassel, M Cragin, C Cramer, A Dingle, S Friday-Stroud, N J Herron A Horton, et al. 2018. "Keeping Data Science Broad: Negotiating the Digital and Data Divide Among Higher Education Institutions." South Big Data Hub. https://southbigdatahub.org/resources/newsblog/keeping-data-science-broad-program.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc.: Sebastopol, CA. https://r4ds.had.co.nz.

Yan, Donghui, and Gary E Davis. 2019. "A First Course in Data Science." *Journal of Statistics Education* 27 (2): 99–109. https://doi.org/10.1080/10691898.2019.1623136.

A Data science at public universities

In this section, we provide a high-level summary, circa early 2022, of the state of data science programs at public universities in Massachusetts.

A.1 UMass-Amherst

At its flagship campus in Amherst, the University of Massachusetts offers several degree options that involve data science. Many of these are coordinated by the Center for Data Science. However, none as of yet result in a bachelor's degree in data science.

A.1.1 Computer science and informatics

Data science is one of three optional concentrations within in the master's degree in computer science. The College also offers a master's-level certificate in Statistical and Computational Data Science that is offered jointly with the Department of Mathematics and Statistics.

At the bachelor's level, neither the bachelor of science nor the bachelor of arts programs in computer science offers an explicit data science designation. The bachelor of science in informatics—which is similarly housed within the Manning College of Information and Computer Sciences—offers data science as one of two concentrations (the other being "health and life sciences").

While all students in the informatics program take foundational courses in data science (INFO 248) and statistics (STAT 240), data science concentrators also take courses in data analytics (CICS 397A), data management (COMPSCI 345), and a statistics elective. The informatics degree provides excellent breadth in computing writ-large, offering courses like Social Issues in Computing (CICS 305) and Web Programming (COMPSCI 326). However, by design it lacks the upper-level computer science courses that characterize the computer science bachelor's, as well as the data science capstone experiences that characterize the data science bachelor's at UMass-Dartmouth.

A.2 Mount Ida

At its new Mount Ida campus in Newton, UMass offers three graduate programs that are data science-adjacent: business analytics, statistics, and geographic information science and technology. Many courses in these programs are offered online or in the evenings.

A.3 UMass-Dartmouth

The bachelor of science in data science program at UMass-Dartmouth is offered jointly by the departments of mathematics and computer science. The current requirements include a mixture of mathematics, statistics, and computer science courses, along with some integrative courses in data science, university electives, and a capstone. Figure 3 shows a flowchart of the recommended sequence of courses through the data science bachelor's program at UMass-Dartmouth.

Notable elements of the UMass-Dartmouth curriculum include two data science courses taken during the first two years: an introduction to data science (DSC 101) and data analysis and visualization (DSC 201).

DSC 101 is described by Yan and Davis (2019) and centers around the data science life cycle popularized by Wickham and Grolemund (2016) (see also Section B.3). Students in this course learn to program in R, and get brief exposure to descriptive statistics, data visualization, data wrangling, regression modeling, and inferential statistics (hypothesis testing). We discuss first courses in data science more broadly in Section 3.2.1.

DSC 201 is a deeper dive into data wrangling and visualization, with a brief appeal to machine learning at the end. This course is purposefully taught in Python, although student are not expected to have extensive previous Python experience. We discuss second courses in data science more broadly in Section 3.2.2.

A.4 Westfield State

The board of trustees approved a bachelor of science degree in data science in June of 2021. A Letter of Intent was sent from Westfield State's president to MDHE for a bachelor of science in data science in the fall of 2021. As of January 2022, no mention of the program appears on the college's website.

The program is designed to achieve the following goals:

- provide students with solid theoretical knowledge of math, statistics, and computer science
- train students to develop relevant programming abilities and execute statistical analyses with Python, R, SQL, and other popular software
- equip students with the ability to build and assess statistical models
- train students to design, build, and use a relational database
- train students to design and create computer information systems in a real-world environment
- equip students with the ability to solve practical problems with data science and present their solutions effectively.

The school does offer a first course in data science that appears to focus on computational statistics, taught using R and Python.

B Learning outcomes for a selection of introductory data science courses

B.1 Data Science in a Box learning goals

For more details, see `https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1804497`

B.1.1 Unit 1: Exploring Data

This unit has three main foci: data visualization, data wrangling, and data import. The learning goals of the unit are as follows:

1. Introduce the R statistical programming language via building simple data visualizations.
2. Build graphs displaying the relationship between multiple variables using data visualization best practices.
3. Perform data wrangling, tidying, and visualization using packages from the tidyverse.
4. Import data from various sources (e.g., CSV, Excel), including by scraping data off the web.
5. Create reproducible reports with R Markdown, version tracked with Git and hosted on GitHub.
6. Collaborate on assignments with teammates and resolve any merge conflicts that arise.

B.1.2 Unit 2: Making Rigorous Conclusions

In Unit 1 students develop their skills for describing relationships between variables, and the transition to Unit 2 is done via the desire to quantify these relationships and to make predictions. This unit is designed to achieve the following learning goals:

1. Quantify and interpret relationships between multiple variables.
2. Predict numerical outcomes and evaluate model fit using graphical diagnostics.
3. Predict binary outcomes, identify decision errors, and build basic intuition around loss functions.
4. Perform model building and feature evaluation, including stepwise model selection.
5. Evaluate the performance of models using cross-validation techniques.
6. Quantify uncertainty around estimates using bootstrapping techniques.

B.1.3 Unit 3: Looking Forward

This unit is designed to shrink or expand as needed depending on time left in the semester. Each module is designed to cover one class period and aims to provide a brief introduction to a topic students might explore in higher level courses. One exception to this is an ethics module, which kicks off the unit and is the only required component. In this module, we introduce ethical considerations around misrepresentation in data visualizations and reporting of analysis results, p-hacking, privacy, and algorithmic bias.

The remaining topics in the unit vary from semester to semester depending on interests of the students and the instructor.

B.2 Data 8 learning outcomes

These appear courtesy of John DeNero (personal communication).

Upon completion of CS/STAT/INFO C8, students should be able to:

1. Write correct small programs that manipulate and combine data sets and carry out iterative procedures.
2. Extend a program with multiple functions so that it runs correctly with additional functionality.
3. Calculate specified statistics of a given dataset.
4. Identify the sources of randomness in an experiment.
5. Formulate a null hypothesis that relates to a given question, which can be assessed using a statistical test.
6. Carry out statistical analyses including computing confidence intervals and performing hypothesis tests in a variety of data settings.
7. Given the result of a statistical analysis from the course, form correct conclusions about a question based on its meaning.
8. Given a question and an analysis, explain whether the analysis addresses the question and how the analysis could change and still address the question.
9. Articulate the benefits and limits of computing technology for analyzing data and answering questions.
10. Correctly generate and interpret histograms, bar charts, and box plots.
11. Correctly make predictions using regression and classification techniques.
12. Assess the accuracy and variability of a prediction.

B.3 University of Massachusetts Dartmouth DSC 101 course goals

See `https://www.tandfonline.com/doi/full/10.1080/10691898.2019.1623136` for more details.

1. It introduces to students the notion that data entails value, thus helping motivate students to the study of data science.
2. It provides students with a big picture and basic concepts of data science, as well as the main ingredients of data science.
3. Students will learn some practical techniques and tools that they can apply later in more advanced courses or when they start work after their degree program.

B.4 Introduction to Data Science (DATA 601) from Ramapo College of New Jersey

See `https://www.ramapo.edu/data-science/data-600-introduction-to-data-science/` for more details.

By the end of this course, students will:

1. Demonstrate advanced skills in data acquisition and management.
2. Demonstrate advanced skills in data analysis techniques using mathematics and statistical principles.
3. Demonstrate advanced skills in data presentation, communication, and visualization.
4. Demonstrate the ability to make data-driven decisions.

C Learning outcomes for data science programs at the associate's level

C.1 Learning outcomes from the Montgomery College (MD) Associate's of Science in Data Science

1. Students will be able to assess different analysis and data management techniques and justify the selection of a particular model or technique for a given task.
2. Students will be able to execute analyses of large and disparate datasets and construct models necessary for these analyses.
3. Students will be able to demonstrate competency with programming languages and environments for data analysis.

4. Students will be able to summarize findings of complex analyses in a concise way for a target audience using both graphics and statistical measures.
5. Understand, evaluate, and apply ethical principles and practices in the data lifecycle.

D Related resources

D.1 Dana Center Design Principles

The Charles A. Dana Center developed design principles for their introductory data science course as part of their Launch Years initiative (`https://www.utdanacenter.org/sites/default/files/2021-05/data_science_course_framework_2021_final.pdf`). We believe that these principles are important for all courses to ensure that they are engaging, coherent, and accessible. These include:

Active Learning: The course provides regular opportunities for students to actively engage in data explorations using a variety of different instructional strategies (e.g., hands-on and technology-based activities, projects, small group collaborative work, facilitated student discourse, interactive lectures).

Growth Mindset: The course supports students in developing the tenacity, persistence, and perseverance necessary for learning data science, for using mathematics and statistics to tackle authentic problems, and for being successful in post-high school endeavors.

Problem Solving: The course provides opportunities for students to engage in the entire statistical problem-solving process.

Authenticity: The course presents data explorations that allow students to address relevant questions that arise in their communities.

Context and Interdisciplinary Connections: The course presents data science in context and connect data science to various disciplines and everyday experiences.

Communication: The course develops students' ability to communicate insights from their data explorations and findings in varied ways, including with words, data visualizations and numbers.

Technology: The course introduces students to current technologies appropriate for data exploration and visualization, and prepares them to learn and use new ones.

Assessment: The course uses project- based assessments both as formative assessments and to evaluate student progress.

More details can be found at the Dana Center website.

D.2 Expanding Computing Education Pathways

The goal of the ECEP (Expanding Computing Education Pathways) project is to improve and broaden participation in computing education (see `https://ecepalliance.org/about`).

They have worked to engage K-20 groups of educational stakeholders to:

- define high school computing curricula
- increase the number of well-trained, certified computing teachers
- improve post-secondary degree programs
- properly align curriculum
- offer comprehensive advising to underrepresented students
- assist in retention efforts
- increase recruitment of underrepresented students
- promote K-20 computing education reform

D.3 Boosting Access to Data Science Scholars

The Boosting Access to Data Science Scholars program (Michelle Trim, PI) will provide scholarships to 40 unique full-time students who are pursuing bachelor's degrees in Informatics or Computer Science in the College of Information and Computer Sciences (CICS). The project aims to increase student persistence in STEM fields by linking scholarships with evidence-based supports, including faculty and near-peer mentoring, faculty led advising, mentored research experiences, graduate school and career preparation, and participation in discipline-specific conferences. Scholars will work with faculty and near-peer mentors to develop individual academic plans outlining their areas of interest and steps toward achieving their goals. The project will also support curriculum improvements aimed at increasing first-year student retention and decreasing time to completion in STEM. The data science focus of the informatics major at UMass Amherst and the diverse population of students served by the institution will contribute to broadening participation in a critical workforce area.

D.4 Position classification flysheet for data science series (1560)

Office of Personnel Management, United States Government: `https://www.opm.gov/policy-data-oversight/classification-qualifications/news/2021/12/`

Series Definition: This series covers professional positions which primarily involve work related to identifying the methods, processes, algorithms, tools, and systems to extract and interpret findings from varied structured and unstructured data sets related to the data science lifecycle. Work also involves the development of algorithms and/or tools to support data manipulation and processing as well as the use of data visualization techniques to articulate findings. The primary requirements of the work are applying professional knowledge of computer science and mathematical and statistical theories, techniques, and methods to gather, analyze, design and construct new processes for modeling, interpret, and/or report quantitative information, trends, relationships and correlations among or within data sets.

The work requires knowledge of, or skills and abilities related but not limited to the following:

- Algorithms
- Application of fact-finding and investigative techniques
- Artificial intelligence
- Big data principles
- Communicate findings both orally and in writing
- Computer science
- Data analytics
- Data modeling
- Data governance
- Data visualization
- Machine learning
- Mathematics
- Natural Language Processing
- Optimization Methods
- Programming languages
- Simulation
- Statistical methods and techniques
- Statistical software and computer programs to perform computer analysis of statistical data and findings
- Statistical theory

D.5 EDC Oceans of Data

`http://oceansofdata.org/projects/mentoring-new-data-pathways-community-colleges`

Quote from summary of project:

As more sectors of the economy come to rely increasingly on data, the demand for skilled data workers is growing at a pace that outstrips the capacity of colleges to develop the programs needed to produce qualified employees. Workforce demand for data skills is disrupting the job market. By 2020, the number of jobs for all US data workers will increase by 364,000 openings to 2,720,000. Annual demand for the new roles of data

scientist, data developers, and data engineers will reach nearly 700,000 openings. Data Science and Analytics (DSA) jobs remain open an average of 45 days, five days longer than the market average. Companies report that they cannot find the qualified employees to fill their open data positions. Projections indicate that this gap will continue to grow. Meanwhile, as evidenced in a recent Data Science Summit held by the American Statistical Association, there is a surging awareness among two-year colleges of employer demands for data workers, a growing interest in developing data programs, and a rapidly growing population of students looking for new career opportunities. What is missing are strategies and supports that can enable colleges to rapidly respond to these opportunities and scale-up their efforts to train the next generation of data workers in a sustained and timely manner.

Mentoring New Data Pathways in Community Colleges, a partnership of EDC's Oceans of Data Institute, Bunker Hill, Sinclair, and Normandale Community Colleges, will strengthen the capacity of faculty in community colleges to design and launch new data programs. This project will train mentors who have participated as partners in ATE's previously funded Creating Pathways to Big Data Careers, to share their experiences in developing data programs with faculty from four mentee institutions, and guide mentees as they: 1) conduct an internal self-assessment of their college's data program development assets; 2) develop and implement a strategic plan to create their new data program; 3) work with employers to prioritize work tasks required for success in local industries; 4) align courses/curriculum to local industry skill demand; 5) establish broad, internal support for the new data program, and 6) develop a new program proposal for their college's curriculum committee. ODI will maintain a professional learning community to support mentors and mentees throughout this process, linking them to best practices learned from ATE's successful MentorLinks and Mentor-Connects projects. Mentors will meet monthly to plan, to discuss challenges and to share successes. Mentors and mentees will visit each other's institutions to engage college faculty and administrative staff in substantive fact-finding and problem solving conversations. Mentors and mentees will meet annually in advance of ATE's annual HI-TEC conference to engage in professional development, plan and problem solve. Mentoring New Data Pathways will deploy a variety of resources to support mentee intuitions in their pursuit of developing data science programs. These resources will also be made widely available to the public, and outreach and dissemination efforts will target community colleges who would most benefit from this information

The EDC Oceans of Data project is funded by the National Science Foundation, grant # 1902568. Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

See also `https://par.nsf.gov/servlets/purl/10297829`

D.6 California Alliance for Data Science Education

See `https://academicdatascience.org/adsa-meetings/2021-annual-meeting` and `https://www.youtube.com/watch?v=1yzQCovnqjA`

ADSA announcement: The California Alliance for Data Science Education is a new collaborative effort that spans across 150 campuses of the California Community Colleges (CCCs), California State Universities (CSUs), and the University of California (UC) systems. It aims to democratize and streamline access to Data Science Education throughout California, building access to Data Science curriculum in traditionally underserved communities in STEM. The alliance does so by synchronizing data science efforts across campuses to build a shared open-source infrastructure, create community college transfer pipelines, and empower leaders at partner campuses.

In partnership with a non-profit computing organization 2i2c and CloudBank, the team has facilitated access to fully-maintained JupyterHubs for California Community Colleges that have launched Data Science Curriculum. The board of the alliance has been focused on standardizing procedures for community college Data Science class articulation to four-year colleges in California, working with key stakeholders including campus articulation officers, the UC Office of the President, and the CCC Chancellor's Office. Lastly, the alliance has held multiple data science education workshops and regularly holds cross-campus discussions, elaborating on the specifics of building a successful undergraduate Data Science curriculum.

D.7 Massachusetts Board of Higher Education

The Massachusetts Board of Higher Education has developed a vision for a new undergraduate student experience that is intended to achieve racial equity in public higher education (`https://www.mass.edu/bhe/documents/09a_NUE%20Report_FINAL.pdf`). The report identifies admissions, enrollment, and transfer

issues, curriculum, equity-minded teaching, learning, and assessment, high impact practices, hiring, and holistic student support. See also `https://www.mass.edu/strategic/equity.asp`.