The Two-Year College Data Science Summit

A report on NSF DUE-1735199

TYCDSS Writing Team: Robert Gould (co-chair), Roxy Peck (co-chair),
Julie Hanson, Nicholas Horton, Brian Kotz, Kathy Kubo, Joyce Malyn-Smith,
Mary Rudis, Brad Thompson, Mark Daniel Ward, Rebecca Wong

Executive Summary

To explore the robust future for two-year colleges in data science education, the American Statistical Association (ASA), with funding from the National Science Foundation, hosted the Two-Year College Data Science Summit in Arlington, Virginia on May 10 – 11, 2018. The summit assembled 72 educators, researchers and practitioners in statistics, mathematics, computer science, and data science. Summit participants included faculty from two-year colleges, four-year colleges and representatives from industry, government and non-profits. The primary goal of the summit was to produce curricular guidelines to assist two-year colleges in establishing and maintaining data science programs. In addition to primary funding from the National Science Foundation (NSF), the summit also received support from Booz Allen Hamilton.

The participants considered three types of potential data science programs: (1) Associate degree programs for students who intend to transfer to a four-year institution, (2) associate degree programs for students aiming to go directly into the workforce, and (3) credit bearing certificate programs. These three different types of programs share many aspects, but differ in the emphasis placed on each of these program outcomes.

Based on input from two days of discussions among participants, and after comparing similar curricular guidelines and suggestions from the Park City Math Institute (PCMI) Data Science Initiative and the National Academies of Sciences (NAS) Committee on Envisioning the Data Science Discipline, the TYCDSS writing team produced the following set of general program outcomes which are broken into eight categories.

<u>Computational Foundations</u>

- Solve data-related problems using a programming language used in data science
- Solve data-related problems using an additional programming language
- Extract data from a database using a relational database language
- Implement basic statistical procedures using statistical software

<u>Computational Thinking</u>

- Solve problems that require the application of computational thinking as defined by Wing (2006)

<u>Statistical Foundations</u>

- Determine if conclusions are appropriate based on study design
- Produce and interpret data visualizations to describe, explore and communicate insights from data
- Produce and interpret numerical summaries to describe and explore data
- Produce and interpret confidence intervals
- Formulate statistical claims in terms of null and alternative hypotheses; carry out and interpret basic hypothesis tests
- Investigate and explore relationships for highly multivariate datasets

Statistical Thinking

Recognize questions and problems that can be investigated using data
Identify data appropriate to answer statistical questions
Explain the role of randomization and random sampling in data collection
Identify sources of variability when drawing conclusions

Statistical Modeling

Use statistical models to describe relationships between variables
Discern between modeling for prediction and modeling for inference
Fit, interpret and evaluate basic statistical models, including cross-validation

Data Management and Curation

Demonstrate the ability to acquire data in diverse formats and structures
Apply exploratory data analysis to identify potential problems in data
Demonstrate the ability to clean data and prepare data for analysis
Identify problems associated with missing data
Combine multiple data sources to address a given statistical goal
Manage databases
Identify and explain issues related to data privacy and security
Construct, maintain and share files as well as collaborate with others in a version control system

Mathematical Foundations

Calculus
Matrices and basic linear algebra
Basic probability

Productivity Foundations

Collaboration and communication

The three types of programs each face unique challenges. Programs intended for transfer students will, by necessity, emphasize courses required to transfer to four-year institutions and will require cooperation with those institutions to develop articulation agreements. Some two-year colleges may find there is little flexibility in the resulting program sequence to add data science specific courses. Programs for students looking to go directly into the workforce will have more flexibility in the courses they can offer, but they will need to identify and collaborate with local industry partners to create programs that will ensure employability of their graduates. Certificate programs will most likely show the greatest variation from college to college, since they are usually designed to meet the needs of local and regional employers. Students interested in certificate programs might also have great variability in their backgrounds, ranging from students with advanced degrees in fields such as computer science or business, who intend to learn more statistics to improve their career prospects, to students who have not yet earned their college degree and are seeking to boost their employability by learning important and relevant skills.

One major difference between the three types of programs is in the level of mathematics required. Certificate programs might not require further courses in mathematics. Transfer programs may be obligated to require lower-division calculus and linear algebra courses (and perhaps more) depending on the requirements at four-year institutions. Transfer programs can assume that students will complete the majority of data science coursework at a four-year institution, and may therefore address fewer of the above program outcomes (assuming they will be addressed at the four-year institution) or address some learning outcomes in less depth (assuming they will be revisited in four-year curriculum). Additionally, transfer programs will likely require more general education coursework than direct-to-work associate degree programs and certificate programs. Direct-to-work programs, on the other hand, have fewer external constraints (compared to transfer programs) and can therefore tailor themselves to industry and student needs. Additionally, this may allow for more flexibility and innovation with mathematical requirements.

Based on the summit discussions and deliberations, we recommend the following:

Recommendation 1: Create courses that provide students with a modern and compelling introduction to statistics that, in addition to traditional topics in inferential statistics, includes exploratory data analysis, the use of simulations, randomization-based inference, and an introduction to confounding and causal inference.

Recommendation 2: Ensure that students have ample opportunities to engage with realistic problems using real data so that they see statistics as an important investigative process useful for problem solving and decision-making.

Recommendation 3: Explore ways of reducing mathematics as a barrier to studying data science while addressing the needs of the target student populations and ensuring appropriate mathematical foundations. Consider a "math for data science" sequence which emphasizes applications and modeling.

Recommendation 4: Design courses so that students solve problems that require both algorithmic and statistical thinking. This includes frequent exposure to realistic problems that require engaging in the entire statistical investigative process and are based on real data.

Recommendation 5: All programs should (a) expose students to technology tools for reproducibility, collaboration, database query, data acquisition, data curation, and data storage; (b) require students to develop fluency in at least one programming language used in data science and encourage learning a second language.

Recommendation 6: Ethical issues and approaches should be infused throughout the curriculum in any program of data science.

Recommendation 7: Whenever possible, classroom pedagogy should foster active learning and use real data in realistic contexts and for realistic purposes. Programs should consider portfolios as summative and formative assessment tools that both improve and evaluate student learning.

1. Introduction

The rapid growth of data science programs at the undergraduate and graduate level, along with the demand for data science professionals in the workforce, has prompted many to think about the role of two-year colleges in data science education. To explore the opportunities for two-year colleges, the American Statistical Association (ASA), with funding from the National Science Foundation (NSF) and additional support from Booz Hamilton, hosted the Two-Year College Data Science Summit (TYCDSS) in Arlington, Virginia on May 10 – 11, 2018. This summit assembled 72 educators, researchers and practitioners in statistics, mathematics, data science, and computer science. Summit participants included faculty from two-year colleges, four-year colleges and universities, as well as non-profit, industry and government representatives. The primary goal of the summit was to produce curricular guidelines to assist two-year colleges in establishing data science programs.

Participants were asked to consider data science as a developing field that merges the disciplines of statistics, mathematics, and computer science in order to facilitate the ability to draw meaning and understanding from data. The ubiquity of data as well as the complexity and scale of these data drives a need for a workforce that can safely and securely store, maintain and provide data; that can access data from a variety of sources and prepare it for analysis; that can find meaningful patterns in large and complex data sets and communicate these findings along with the data limitations to diverse communities; and that can scale algorithms for data discovery, classification, and prediction. Equally important is the growth of a citizenry that is aware not only of the role that data play in a democracy, but also of the need to maintain and protect security and privacy.

Summit organizers recruited participants from stake holding communities, including professional associations such as the American Statistical Association (ASA), the Association of Computing Machinery (ACM), the American Mathematical Association of Two-Year Colleges (AMATYC), and the Mathematical Association of America (MAA). The summit also included many participants involved in previous and current attempts to establish data science education programs, including the Park City Math Institute (PCMI) Data Science initiative, the Oceans of Data Institute (ODI), the National Academy of Sciences (NAS) Committee on Envisioning the Data Science Discipline, organizers of the NSF-funded Data Science Education (DSE) workshop organized by the ACM Education Board, and organizers of the Data Science Education Technology group (DSET) hosted by the Concord Consortium.

Participants in the summit were asked to consider three types of programs: (1) programs offering associate degrees for students who intend to transfer to a four-year institution (2) programs offering associate degrees for students who intend to directly join the workforce and (3) credit bearing certificate programs. To accomplish this, participants were assigned to small groups to discuss one of these three program types. The groups were asked to consider desired program outcomes, challenges to implementing these outcomes, and resources to assist implementation. Plenary sessions allowed the small groups to discuss their progress and to receive feedback from other groups. Following the summit, three writing teams, one for each type of program, met for one-and-a-half days to capture the participants' recommendations.

This report summarizes the findings of the summit and also provides an overview of the current state of two-year college data science education. Section 2 provides background, including the current state of data science at two-year colleges as well as summaries of recent efforts to define standards and educational frameworks for data science programs.

Through the small- and large-group discussions about the nature of data science programs, a number of common themes and challenges arose, particularly around how such concepts such as statistical thinking and computational thinking might be taught in data science programs. Section 3 summarizes the common themes, challenges, and differences between the programs.

Section 4 introduces and defines the primary program outcomes for three different types of programs and indicates an appropriate level of depth for each outcome. Sample curricula and a discussion of curriculum maps appear in Section 5. (The curriculum maps themselves are provided online at https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-summit.aspx).

Section 6 looks ahead at some future goals, ideas, and research for data science education at two-year colleges. Section 7 highlights resources to assist those who aim to propose and implement data science programs at two-year colleges and offers suggestions for future directions.

2. Background

2.1 Data Science at Two-Year Colleges

Data science is an emerging field that uses methods from statistics, mathematics, and computer science in order to find and communicate meaning in data. Due to the rapidly increasing role of data in commerce and science, data science has gained wide attention in a relatively short amount of time in the private sector, government, and academia. In the private sector, job advertisements for data scientists have burgeoned. The National Institutes of Health and the NSF CISE Directorate created chief data scientist positions with dozens more such positions across the federal government. The Society for Human Resource Management reports that 82% of organizations "currently have or expect to have positions that require data analysis skills." They predict continued growth in this demand over the next five years (2016 SHRM).

To meet the growing demand for data scientists and data science skills, academia has responded: dozens of data science master's programs have been created, many bachelor's programs are established (for an example, see Shedden & Prakash, 2015) and individual courses are offered at many colleges and universities, including many online options. The 2018 National Academy of Sciences "Undergraduate Data Science Opportunities and Options" report provides a number of modalities through which data science education is currently offered, including two-year institutions (NAS, p.3-1).

Two-year colleges face unique challenges in establishing data science programs. Two-year colleges may lack full-time faculty with the required credentials and knowledge and often must rely heavily on adjuncts. Identifying and evaluating faculty qualified to teach in a data science program can be a challenge without administration and staff who themselves have an understanding of the requirements of the discipline. Two-year colleges might, depending on their location, have greater difficulty in partnering with local employers compared with four-year institutions. Some employers, on the other hand, may be unaware of the knowledge and skills that graduates of two-year college data science programs can bring to their organizations.

Two-year colleges also generally have a more diverse student population with respect to race, ethnicity, gender, and mathematical preparation than many four-year colleges. While this diversity may pose challenges for curriculum designers, it provides an opportunity to increase diversity in STEM by providing STEM career paths that students find exciting and engaging but which do not have many of the barriers to entry shared by other STEM disciplines. In particular, the mathematical prerequisites for a two-year college data science program are a subset of those required for programs that include the full three semester calculus sequence, and this allows access for students with a broader range of mathematics preparation.

Despite these challenges, a small number of two-year colleges have already created data science programs. The summit organizers and participants identified 18 programs implemented at two-year colleges as of May 2018. Developers from two of these programs, Brian Kotz at Montgomery College in Germantown, MD and Mary Rudis at Great Bay Community College in New Hampshire, were co-organizers of this summit.

2.2 Existing Two-Year College Data Science Programs, 2018

This section provides a list of existing data science programs at two-year colleges as of May 2018. While a solid education in data science could quite conceivably be available within a statistics, computer science, information technology, or mathematics program, our focus was on programs that explicitly mentioned data science, analytics, or data analysis in their title. We maintain a list of these programs at https://padlet.com/rgould2/r36z3yksg1ey.

We found 11 certificate programs, one "direct to workplace" associate degree program, and six associate degree programs intended for transfer to four-year colleges or universities. Note that some colleges offer more than one program. Table 2.1 lists these programs.

Table 2.1. Programs in Data Science at Two-Year Colleges as of May 2018

<u>Certificate Programs</u>

- Great Bay Community College, Portsmouth/Rochester, NH
  - Data - Practical Data Science
- Wake Tech Community College, Raleigh, NC
  - Business Analytics
- Montgomery College, Germantown, MD
  - Data Science Certificate Program
- Bunker Hill Community College, Boston, MA
  - Data Management (Fast-Track) Certificate Program
- Washtenaw Community College, Ann Arbor, MI
  - Applied Data Science Certificate
- Cerro Coso Community College, Ridgecrest, CA
  - Data Analyst Certificate
- Manchester Community College, Manchester, NH
- Normandale Community College, Bloomington, MN
  - Data Analysis Certificate
  - Applied Data Analytics Certificate
- Sinclair College, Dayton, OH
  - Data Analytics One-year Technical Certificate
- Johnson County Community College, Overland Park, KS
  - Data Analytics Certificate
- Seattle Central College, Seattle, WA
  - Certificate in Data Analysis

Associate Degrees Direct to Workplace

Great Bay Community College, Portsmouth, NH

Analytics Associate in Science (designed for both transfers and direct-to-workplace)

Associate Degrees for Transfer to Four-Year

Great Bay Community College, Portsmouth, NH

Analytics Associate in Science

Community College of Allegheny County, Pittsburgh, PA

Data Analytics Technology, Associate of Science

Nashua Community College, Nashua, NH

Foundations of Data Analytics, Associate in Science

Manchester Community College, Manchester, NH

Analytics and Data Science Public Pathways Program

Community College of Allegheny County, Pittsburgh, PA

Data Analytics Technology Associate of Science Program

Normandale Community College, Bloomington, MN

Data Analytics, Associate in Science Program

In order to better compare these programs, we examined the mathematics requirements, the statistics requirements, and programming languages taught. To do this, we counted the number of mathematics and statistics courses listed as requirements. We did not review syllabi for these or other courses; it is quite likely that many courses not explicitly listed as mathematics or statistics courses might still contain considerable mathematical and statistical content. Course level information was not available for every program. Tables 2.2-2.4 compare associate degree programs with certificate programs in terms of mathematics and statistics content as well as software used.

Table 2.2. Number of mathematics courses (beyond prerequisites) in the programs from Table 2.1.

Required Math	**Certificate**	**Associate Degree**
None	9	0
Finite Math	1	0
Pre-Calculus	1	0
Calculus I	0	3
Calculus II	0	2
Quantitative Literacy	0	1
Intermediate Algebra	0	1
Linear Algebra	0	1

Table 2.3. Number of statistics courses required from programs in Table 2.1.

Required Statistics	**Certificate**	**Associate Degree**
None	3	1
Introductory Statistics	3	1
Probability and Statistics	2	1
Business Statistics	2	0
Statistics I and II	0	1
Data Analysis I, II, and III	3	0

Table 2.4. Software languages taught in the programs from Table 2.1.

Software	Certificate	Associate Degree
SQL	6	0
R	3	1
Excel	3	0
Python	2	1
C++	0	1
Java	0	2
Object-oriented programming	0	1
Unspecified statistical programming language	1	0
Unspecified data storage/database language	1	0
Not explicitly mentioned	2	0

Interestingly, three certificate programs and one associate degree program do not require any statistics courses. One reason for the lack of statistics courses among the certificate programs may be that statistical content is embedded in other courses, such as data management or programming courses.

2.3 Previous Efforts

ACM Educational Board Workshop on Data Science Education

In 2015, the NSF funded a workshop on Data Science Education, organized by the Association of Computing Machinery and led by Boots Cassel and Heikki Topi (Cassel & Topi 2015). The workshop debated the issue of the how to define data science and noted that it could be considered either an emerging discipline of its own or as a synthesis of existing disciplines. Goals of this meeting included recommendations for establishing undergraduate curricula in this emerging field. Key recommendations were to establish an interdisciplinary task force to develop curriculum guidance and to establish an "infrastructure and culture of sharing materials and experience" to support the development of data science programs.

Oceans of Data Institute

Participants of the Two-Year College Data Science Summit considered two documents produced by The Oceans of Data Institute (ODI) relevant to this discussion. The first is the *Profile of the Data Practitioner* (http://oceansofdata.org/our-work/profile-data-practitioner). Based on a panel of ten "big data experts", and drawing on previous survey responses from over 100 data science practitioners, this profile identifies six key "duties" of a data practitioner and breaks down the individual tasks that comprise each duty. The duties are the following: initiates the project; sources the data; transforms the data; analyzes the data; closes out the project; engages in professional development. This report also identifies important software and hardware tools, skills and knowledge, and behavior.

The second ODI report included a more general call for "data literacy" at all educational levels and across all countries. A panel of educators, data scientists, statisticians, and computer scientists drafted a definition of data literacy that could be useful for the development of general education curriculum design:

> The data-literate individual understands, explains, and documents the utility and limitations of data by becoming a critical consumer of data, controlling his/her personal data trail, finding meaning in data, and taking action based on data. The data-literate individual can identify, collect, evaluate, analyze, interpret, present, and protect data.

(http://oceansofdata.org/our-work/building-global-interest-data-literacy-dialogue-workshop-report)

The Oceans of Data Institute carries out several projects that directly impact data science at community colleges. The Data Analytics Technical Advancement Program partners with Columbus State Community College to establish pathways in central Ohio to increase the supply of data analytics technicians. The Creating Pathways for Big Data Careers initiative partners with Bunker Hill Community College, Normandale Community College, Johnson County Community College, and Sinclair Community College to establish career pathways in data science.

Park City Math Institute

The Park City Math Institute (PCMI) 2016 Undergraduate Faculty Program composed a set of curriculum guidelines for undergraduate programs in data science (De Veaux et al., 2017). The committee, chaired by statistician Richard De Veaux, included 25 faculty from mathematics, computer science, and statistics departments. Although these guidelines were intended to guide the development of undergraduate majors, they should also prove useful for two-year colleges preparing programs for students who intend to transfer to a four-year program or for students seeking a direct-to-work associate degree.

The PCMI guidelines identified six key competencies for undergraduate data science: computational thinking and statistical thinking, mathematical foundations, model building and assessment, algorithms and software, data curation, and communication and responsibility.

The PCMI guidelines also provide curriculum recommendations and suggestions for course outlines and include a description of the design principles that guided the creation of the recommendations in this document. These design principles should be of particular interest to those designing programs for two-year colleges. First, the guidelines note that technology changes rapidly, and students who learn a technological tool are less likely to have long-run success than those whose study includes the more conceptual foundations of computational thinking, which require "thinking at multiple levels of abstraction" (Wing, 2006). Second, courses should strive to synthesize computational thinking and statistical thinking. Whenever possible, students should be encouraged to create and recognize connections between computation and statistics. Third, programs should consider replacing the traditional "calculus sequence" with a sequence of courses which more directly serves the needs of data science: "Although the tools needed...include calculus, linear algebra, probability theory and discrete mathematics, we envision a substantial realignment of the topics within these courses and a corresponding reduction in the time students will spend to acquire them" (De Veaux et al., 2016, p. 8).

The PCMI guidelines propose two sequences for students in the first year of study of a B.S. degree, with both consisting of two courses. These sequences could be useful for two-year colleges preparing students to transfer. The first is an "Introduction to Data Science" sequence, and the second is a "Mathematical Foundations" sequence. The Introduction to Data Science courses would include the following: a high level programming language useful in data science; exploring and manipulating data; functions and basic coding; introduction to deterministic and stochastic modeling; concepts of projects and code management; an overview of databases; and an introduction to data collection and statistical inference. The mathematical foundations sequence would focus on mathematical tools to assist in solving data-oriented, real-world problems and would help students develop an intuitive, geometric and visual way of thinking. Topics would include the following: mathematical structures (functions, sets, etc.); linear modeling and matrix computation (eigenvalues/eigenvectors, projection/least-squares); optimization (calculus concepts related to differentiation); multivariate thinking (concepts and numerical computation of multivariate derivatives and integrals); and probabilistic thinking and modeling (including counting principles, distributions, independence).

Keeping Data Science Broad: Negotiating the Digital and Data Divide Among Higher-Education Institutions.

The goal of this project was to garner community input into pathways for keeping data science as a broadly inclusive discipline. Representatives from a number of traditional and alternative data science programs, and from a range of institution types – including minority-serving institutions, community colleges, liberal arts colleges, tribal colleges, universities, and industry partners – participated in a series of virtual and in-person meetings which led to a report (available at https://southbigdatahub.org/publications-and-resources/) describing opportunities and challenges for future data science programs to remain diverse and inclusive.

National Academies of Sciences Envisioning the Data Science Discipline Consensus Report ("*Data Science for Undergraduates: Opportunities and Options*")

The National Academies of Sciences' (NAS) consensus report "Data Science for Undergraduates: Opportunities and Options" (https://nas.edu/envisioningds) provides a vision for data science education at the undergraduate level. The report is comprehensive, in the sense that it explicitly addresses two-year and four-year institutions and acknowledges the need to include secondary level education in our understanding of the data science education landscape.

The report makes several recommendations which two-year colleges are particularly well-suited to address. The first is that data science programs should "focus on attracting students with varied backgrounds and degrees of preparation" (Recommendation 4.1, p.19). Two-year colleges serve a diverse study body, particularly when compared to private four-year institutions and to the four-year student population as a whole (Franklin 2014). The second is that academic institutions "should provide...a range of educational pathways" to data science. (Recommendation 2.2, p. 19). The Two-Year College Data Science Summit explicitly addressed this latter recommendation by considering three populations of students (students intending to transfer to four-year institutions, students looking to go directly to work with an associate degree, and students seeking certificates) and also by considering the need to remove barriers of entry, particularly in mathematics.

The report also calls (Recommendation 3.1) for the establishment of a forum for four-year and two-year institutions to engage across institutions on all aspects of data science education, training, and workforce development.

Finally, a useful and interesting product of the NAS report was the establishment of the notion of "data acumen", which is defined as "the ability to understand data, to make good judgments about and good decisions with data, and to use data analysis tools responsibly and effectively" (p. 9). The goal of a data science education can be viewed as the need to build data acumen within the students, and the report identifies several key areas which are crucial for building data acumen: mathematical foundations, computational foundations, statistical foundations, data management and curation, data description and visualization, data modeling and assessment, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving (see Finding 2.3).

2.4 The ASA Two-Year College Data Science Summit

The primary goal of the summit was to produce curricular guidelines for two-year college data science programs serving each of these three communities:

(1) students seeking associate degrees for direct entry to the workforce;
(2) students seeking associate degrees for transfer to four-year institutions;
(3) students seeking credit bearing data science certificates.

Additionally, participants were asked to contribute towards a directory of resources to help two-year colleges establish new data science programs and to discuss the challenges faced by new programs.

The 72 summit participants included faculty from two-year and four-year institutions, representatives of professional societies involved in data science and data science education, researchers in data science and related fields, as well as educators experienced in computer science and statistics education. In order to build upon the experiences of other data science education efforts, representatives from these efforts were also invited.

The participants were organized into six workgroups, created such that each contained representation from as many of the stakeholder communities as possible. Two workgroups were assigned to each of the three types of programs, and each group was charged with discussing desired program outcomes, concerns, challenges and known resources. Each workgroup contained at least one member from the writing team that prepared this report.

Details regarding the summit organization and structure, including a complete list of participants, appear in the appendix.

3. Common Themes and Differences

In the discussions described above, several common themes emerged and were further explored by the writing groups. The following section summarizes themes that played a prominent role in the meeting's discussions.

3.1 Statistics

The summit participants concurred with the NAS report that knowledge of statistics and the development of statistical thinking was of key importance for all three program types.
The NAS report states that an important foundation in statistics would include:

- variability, uncertainty, sampling error, inference
- multivariate thinking
- nonsampling error, design, experiments, biases, confounding, and causal inference
- exploratory data analysis
- statistical modeling and model assessment
- simulations and experiments

(National Academy of Science 2018, p. 2-10)

In addition, many of these foundational topics are included (explicitly or implicitly) in some of the other "key concepts" for data acumen. These include data management and curation, data description and visualization, and data modeling and assessment.

While many of the concepts listed above are intended to pertain to bachelor degree programs, a two-year program may also introduce many of these concepts. A consensus view of the summit, however, is that all three types of programs should focus on providing students with frequent opportunities to engage in statistics as an investigative process (GAISE College Report 2016) and to encourage students to see statistical methods as part of a foundational approach rather than a collection of unrelated tools.

In determining the level of depth required, participants felt two-year colleges should be guided by the intended career paths of their students. Students in a transfer program or a certificate program might suffice with a traditional introductory statistics course, as long as that course has an emphasis on data analysis (although see cautions in the next paragraph). On the other hand, students intending to go directly into the workforce, either after an associate degree or a certificate, might instead focus on the interpretation of statistical output and an understanding of common interpretive pitfalls.

Although many (but not all) introductory statistics courses at least nominally include the topics listed in the NAS report, participants felt that some areas required further emphasis and were of particular interest to data science. These include the study of exploratory data analysis, the use of simulations for solving probability problems and for statistical inference, multivariate thinking, and causal inference. It was emphasized that simply covering a list of topics at a surface level is not sufficient. These topics must be embedded within an approach that requires students to solve real-world problems using real-world data.

Multivariate thinking is increasingly important for statistics and data science students at all levels. Few, if any, real-world problems depend on a single variable because few, if any, real-world systems consist of a single variable. For this reason, students must learn to analyze systems containing multiple variables. In early levels, this may mean reading and interpreting data visualizations for multivariate systems. (For example, see the New York Times and American Statistical Association's *What's Going On in This Graph* discussion forum, https://www.nytimes.com/column/whats-going-on-in-this-graph). Students need experience working with problems that do not provide them solely with the variables needed for the problem; instead, students should be exposed to multi-variable datasets and learn to choose which variables (and which subsets of the rows) are necessary to answer investigative questions. Understanding the role of confounding factors (and Simpson's paradox) and how experiments can be designed to minimize their role are additional important components of multivariate thinking.

Recommendation 1: Create courses that provide students with a modern and compelling introduction to statistics that, in addition to traditional topics in inferential statistics, includes exploratory data analysis, the use of simulations, randomization-based inference, and an introduction to confounding and causal inference.

Recommendation 2: Ensure that students have ample opportunities to engage with realistic problems using real data so that they see statistics as an important investigative process useful for problem solving and decision-making.

3.2 Mathematics

All groups encouraged minimizing barriers for entry into data science programs, and this most often translates to a recommendation that institutions carefully consider mathematics prerequisites. Data science is a rigorous and demanding science, but the workplace requires data scientists with many different levels of expertise. New programs should take this into account by carefully tailoring the mathematics requisites and courses to the program outcomes.

One reason for concern over mathematics offerings is that all too often, completion rates for traditional mathematics courses are very low. Data strongly suggest that long developmental mathematics course sequences are a leading factor in student attrition (for example, Burdman, 2015abc). For this reason, institutions should consider alternative approaches that provide students with the needed mathematics without providing unintended barriers. For example, co-requisite models of mathematics instruction are growing as a national trend to assist student progress and provide broader access while maintaining a high degree of rigor and ensuring that students have appropriate foundations in mathematics and abstraction. In such models, incoming students who have been assessed as academically underprepared are no longer placed in a series of "remedial" courses that do not count for college-level credit. Co-requisite models aim to meet this challenge by placing underprepared students directly into credit-bearing classes while simultaneously providing them rigorous, "just in time" support. (http://www.utdanacenter.org/dc-helps-launch-co-requisites-in-cali/)

The summit participants acknowledge the need to continue creating more flexible math pathways, as mentioned by the NAS report's acknowledgement of the work of the Dana Center at the University of Texas (www.utdanacenter.org/higher-education/dcmp) and also by paring down the mathematical foundations to better address the needs of data science students (as recommended by the PCMI guidelines).

We note that the level of required mathematics for the three types of programs vary considerably. For instance, programs for students intending to transfer to four-year institutions will be bound by their articulation agreements with those institutions and so may be required to offer a more traditional set of calculus-driven mathematics courses. On the other hand, associate degree programs intended for direct-to-work may have more flexibility (depending on state and local requirements) than those for transfer students. If this is the case, then these programs could better tailor the mathematics required to suit the needs of their students. Certificate programs, on the other hand, might not require any additional mathematics courses, either because their pool of potential students might be assumed to already possess a sufficient proficiency in mathematics or because the area of concentration of the certificate does not require additional mathematics beyond program prerequisites.

The NAS report highlights the following topics as fundamental to data science mathematical foundations: set theory and basic logic, multivariate thinking via functions and graphical displays, basic probability theory and randomness, optimization, and matrices and basic linear algebra. Two-year institutions should not feel obligated to offer courses in all of these areas, but should plan carefully to meet the needs of their student populations and their desired program outcomes. The PCMI guidelines recommends culling from the topics typically taught in the first two years of collegiate mathematics those that are not necessary for data science. The guidelines also recommend a more radical rethinking of these topics into a first-year mathematics course that emphasizes mathematical modeling and applications of mathematics rather than the traditional calculus sequence. Hardin and Horton (2017) propose a two-course mathematics sequence for undergraduate data science students.

Recommendation 3: Explore ways of reducing mathematics as a barrier to studying data science while addressing the needs of the target student populations and ensuring appropriate mathematical foundations. Consider a "math for data science" sequence which emphasizes applications and modeling.

3.3 Computational and Statistical Thinking

A key aspect of the construct of computational thinking is algorithmic thinking. Summit participants discussed the relationship between algorithmic thinking and statistical thinking. Listing program outcomes under each of these large-scale types of "thinking" results in a long list and might lead one to fear that completing these outcomes in two years or fewer is impossible. However, these two types of "thinking" are more similar than they are different. Both primarily emphasize the development of critical and analytical thinking skills in the context of solving problems that require data. Both involve some aspect of abstraction. Algorithmic thinking also includes the ability to craft solutions into algorithms (where appropriate) and to evaluate and assess these algorithms. Statistical thinking focuses on understanding the role of variability in sampling, measurement, and decision-making, and includes the ability to craft solutions into statistical models and to evaluate and assess these models.

Futschek (2006) defines algorithmic thinking as a set of the following abilities that are related to constructing and understanding algorithms:

- the ability to analyze a given problem
- the ability to precisely specify a problem
- the ability to find the basic actions that are adequate to the given problem
- the ability to construct a correct algorithm to a given problem using basic actions
- the ability to think about all possible special and normal cases of a problem
- the ability to improve the efficiency of an algorithm

Teaching statistical thinking is one of the primary recommendations of the 2005 and 2016 Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (GAISE 2016). Statistical thinking requires considering statistics as an investigative process of problem-solving and decision-making. The notion of statistical thinking can be summarized by the "PPDAC" cycle of Wild & Pfannkuch (1999). Based on observations of real-life statisticians solving problems, Wild & Pfannkuch identified several stages of problem solving, which they labeled as "Problem, Plan, Data, Analysis, and Conclusion".

Both algorithmic thinking and statistical thinking should be developed through engaging students in real and realistic problems that require working with real and realistic data. Students must be given opportunities to engage in realistic problems that require frequent engagement with the entire investigative process in every data science course, although the level of complexity would naturally be altered to adjust for the level of the course. (For example, see Kim et al 2018 for a discussion for how data can be "tamed" to provide students with a realistic and meaningful data-analytic problem at different stages of their learning.)

The PCMI guidelines remark that it is not sufficient to teach computational/algorithmic thinking as separate from statistical thinking. Instead, students need to learn to synthesize these types of thinking in order to engage in problems centered on extracting insight from data. For this reason, we encourage programs to take every opportunity to enhance statistical understandings with computing throughout the curriculum. For example, inference can be understood with randomization-based testing, which can also help develop programming skills and algorithmic thinking. Bootstrapping methods can be easily implemented at the beginning and intermediate levels (although deeper statistical understandings may require more experience). And simulations can be used both to improve understanding of probability as well as to enhance programming skills to solve complex probability problems.

<u>Recommendation 4</u>: Design courses so that students solve problems that require both algorithmic and statistical thinking. This includes frequent exposure to realistic problems that require engaging in the entire statistical investigative process and are based on real data.

3.4 Additional Themes

Discussions at the summit identified additional themes that two-year colleges must consider when developing programs.

Data Skills

Students in all types of programs need some exposure to fundamental problems involving the acquisition, storage, curation, and dissemination of data. These include:

(a) methods of data acquisition (web scraping, data base queries)
(b) data storage and curation (database management, data privacy/security concerns)
(c) data preparation for analysis (cleaning, transforming, dealing with missing values, recoding)
(d) version control

Students' depth of exposure to these topics will depend on the goals of their program, but at a minimum these students need some experience with the structure of databases, a database query language such as SQL, and experience working in a setting that requires version control to manage the workflow.

Programming

Students in all programs should gain proficiency with at least one programming language commonly used to analyze data, for example, R or Python. Programs with more time (such as a program for transfer students that is, in effect, one half of a four-year program) should consider providing proficiency in one language and an introduction to at least one other language to establish a foundation in computational literacy. Additional languages might be database query languages, languages for parallel processing such as Hadoop or Spark, languages for regular expressions, languages for modeling (e.g., Stan or Julia) or languages at a lower level than R and Python, such as C++.

Ethics

Awareness of the importance of ethical data science practice continues to grow as stories of the violation of ethical practices continue to make news. O'Neil (2016) reminds us that problem solving with algorithms and "big data" can amplify cultural biases and disparities. Lane et al. (2014) caution about big data's threats to privacy, but also offer constructive approaches to analyzing data for social good. The GAISE College report recommends that students of statistics learn proper ethical practices for collecting data (GAISE Report 2016, p 11) and the ethical implications of "p-hacking" and multiple testing. Summit participants were in agreement that programs should frequently expose students to case studies that examine ethical issues and, whenever possible, instill best practices for safe and conscientious data management, storage, collection and analysis. This can be accomplished by requiring a class in ethics, but even so, participants felt that examples of sound ethical practice and further consideration of ethical issues should be reinforced in multiple courses within a program.

Pedagogy

The GAISE College Report (GAISE Report 2016) offers advice for statistical pedagogy that is particularly relevant for data science, which we paraphrase as:

- integrate real data with a context and purpose
- foster active learning
- use technology to explore concepts
- use assessments to both improve and evaluate student learning

Students should frequently be working in situations that require them to apply and develop statistical and computational thinking skills through exposure to open-ended, real problems working with real and complex data. Courses should require projects, and the projects should require that students work in groups to investigate vaguely-stated problems from beginning to end, including presentations of findings. The level of complexity should be tailored to the pedagogical needs and readiness of the student.

One approach to assessment that is consistent with the GAISE recommendations is to require students to develop and maintain a portfolio. Summit participants recommend that programs follow this practice. The portfolio could include examples of student projects, both completed and on-going, and could be used to demonstrate a student's abilities and understandings. Such artifacts might also be helpful for formative and summative assessment of programs. Portfolios need not be physical and might in fact be spread across several digital repositories. For example, git repositories are commonly used within the data science community to highlight work products and also to share ideas. A student portfolio might be a collection of links pointing to a git repository, to specific projects on a project competition site such as kaggle.com, or to capstone projects deployed on a college server. An added benefit is that these portfolios can also potentially provide potential employers with a better understanding of a student's strengths, their understanding of workflow and reproducibility, and the general quality of a student’s work.

Recommendation 5: All programs should (a) expose students to technology tools for reproducibility, collaboration, database query, data acquisition, data curation, and data storage; (b) require students to develop fluency in at least one programming language used in data science and encourage learning a second language.

Recommendation 6: Ethical issues and approaches should be infused throughout the curriculum in any program of data science.

Recommendation 7: Whenever possible, classroom pedagogy should foster active learning and use real data in realistic contexts and for realistic purposes. Programs should consider portfolios as summative and formative assessment tools that both improve and evaluate student learning.

4. Recommended Program Outcomes

Our recommended program outcomes are organized into four categories: Computational, Statistical, Data Management and Curation, and Mathematical. Some of these are further broken down into *Foundations*, *Thinking*, and *Modeling* outcomes. *Foundations* are fundamental concepts and ideas, while *Thinking* outcomes are typically high-level outcomes that require students to synthesize foundational outcomes to solve new problems. The *Modeling* outcomes explicitly address concepts, understandings, and skills needed work with statistical and computational models.

For each program, the outcome is labeled as requiring "exposure", "working knowledge", or "mastery." These indicate increasing levels of complexity. "Exposure" suggests that students should have seen content related to that outcome and are aware of its existence and can describe what it means. “Working knowledge” means that students are able to apply this outcome to routine problems, but may need some assistance and support when working in novel situations. “Mastery” indicates that students have a deep understanding of the topic and can apply it in novel situations and consider the topic from multiple perspectives.

With respect to programming languages and software use, "exposure" would indicate that students have seen code in a particular programming language and perhaps have some knowledge about how that language compares to others. "Working knowledge" means that the student can code using the programming language, although perhaps not using sophisticated algorithms and possibly requiring assistance from support manuals or online forums. "Mastery" means that the student is comfortable with the language, and can develop efficient, readable code and constructively critique the code of others.

Program outcomes were difficult to establish within the context of a general certificate program. In practice, the mission — and therefore, the outcomes — for individual certificate programs will vary greatly. This makes it difficult to definitively assign a level of mastery for many of the listed certificate program outcomes (and a few of the outcomes for other program types as well). To indicate variability, we use a slash, which indicates a range of mastery levels for a particular outcome. For example, a certificate program emphasizing relational databases would strive for "mastery" in that area, while a certificate program emphasizing data analysis might be satisfied with a "working knowledge" of relational databases. In these cases, we emphasize the possible range of knowledge with "working knowledge/mastery".

Section 5 provides sample curricula for programs for transfer students and programs for direct-to-work students. Certificate programs are sufficiently varied that the authors felt that sample curricula would not be helpful. (See Table 2.1 for links to existing curricula.)

4.1 Computational

1. Foundations

Outcome	**To Work**	**Transfer**	**Certificate**
CF.A. Solve problems using a programming language used in data science.	mastery	mastery	mastery
CF.B. Solve problems using an additional programming language(s).	working knowledge	exposure	exposure
CF.C. Extract data from a database using a relational database language.	working knowledge	exposure	working knowledge/mastery
CF.D. Implement basic statistical procedures using statistical software.	mastery	working knowledge	working knowledge

2. Thinking

Outcome	**To Work**	**Transfer**	**Certificate**
CT.A. Solve problems that require the application of computational thinking. (Analyze problem and identify basic actions required for	working knowledge	working knowledge	exposure

Outcome	To Work	Transfer	Certificate
solution; construct an algorithm using the basic actions; consider special and routine cases; evaluate the efficiency of the algorithm)			

4.2 Statistical

1. Foundations

Outcome	To Work	Transfer	Certificate
SF.A. Determine if conclusions are appropriate for a study based on study design (observational or controlled experiment), including identifying potential confounding factors and appropriate controls.	working knowledge	working knowledge	working knowledge
SF.B. Produce and interpret data visualizations, including dashboards, graphs and charts to describe and explore data and communicate findings.	working knowledge	working knowledge	working knowledge/ mastery
SF.C. Produce and interpret numerical summaries to describe and explore data.	working knowledge	working knowledge	working knowledge
SF.D. Produce and interpret confidence intervals	working knowledge	working knowledge	exposure/working knowledge
SF.E. Formulate statistical claims in terms of null and alternative hypotheses, carry out and interpret basic hypothesis tests.	working knowledge	working knowledge	exposure
SF.F. Investigate and explore relationships between more than two variables.	working knowledge	working knowledge	exposure/working knowledge

2. Thinking

Outcome	To Work	Transfer	Certificate
ST.A. Recognize questions and problems that can be investigated. using data	working knowledge	working knowledge	exposure
ST.B. Identify data appropriate to answer statistical questions.	working knowledge	working knowledge	exposure

Outcome	To Work	Transfer	Certificate
ST.C. Explain the role of randomization and random sampling in data collection.	working knowledge	working knowledge	exposure
ST.D. Identify sources of variability including sampling variability, when drawing conclusions from data.	working knowledge	working knowledge	exposure

3. Statistical Modeling

Outcome	To Work	Transfer	Certificate
SM.A. Use statistical models to describe relationships between variables.	working knowledge	working knowledge	exposure
SM.B. Discern between modeling for prediction and modeling for inference.	working knowledge	working knowledge	exposure
SM.C. Fit, interpret and evaluate basic statistical models (e.g. linear, logistic, exponential).	working knowledge	working knowledge	exposure/working knowledge
SM.D. Explain the purpose of cross-validation.	exposure	exposure	exposure

4.3 Data Management and Curation

Outcome	To Work	Transfer	Certificate
DMC.A. Demonstrate ability to acquire and represent data in diverse formats and structures, such as databases, web pages, JSON, etc.	working knowledge/mastery	exposure	working knowledge/mastery
DMC.B. Apply exploratory data analysis to identify problems in the data.	working knowledge	exposure	exposure/working knowledge
DMC.C. Clean data and prepare data for analysis and identify problems that might arise from assumptions made during this process.	working knowledge	working knowledge	exposure/working knowledge
DMC.D. Identify problems associated with missing data.	exposure	exposure	exposure/working knowledge

Outcome	To Work	Transfer	Certificate
DMC.E. Combine multiple data sources to address a given statistical goal.	working knowledge	exposure	working knowledge
DMC.F. Manage databases.	working knowledge/mastery	exposure	working knowledge/mastery
DMC.G. Explain issues related to data privacy and security.	working knowledge	exposure	working knowledge/mastery
DMC.H. Construct, maintain, and share files in a version control system.	working knowledge	exposure	working knowledge/mastery

4.4 Mathematical Foundations

Recent recommendations on undergraduate data science programs acknowledge that the mathematics for data science need not be a long sequence of courses and may be a subset of the mathematics typically required for other STEM disciplines (see for example, National Academy of Science Committee on Envisioning the Data Science Discipline (2018), Hardin and Horton (2017)). The traditional three semesters of calculus followed by linear algebra is more than is necessary for exposure and working knowledge in data science, as noted in the PCMI report (De Veaux et. al, (2017)). This means that programs should consider developing “mathematics for data science” courses that provide the foundation needed for the study of data science. Topics included in such a course might be set theory and basic logic, matrices and basic linear algebra, and enough calculus for understanding of functions and optimization. One example of such a course is the Data Science Math Skills course offered by Duke University online through Coursera (https://www.coursera.org/learn/datasciencemathskills#syllabus). Mathematics courses for data science programs should also emphasize applications and the ties between theory and application.

It is also noted that due to local and state-wide system constraints, data science programs that offer an associate degree may be required to include the mathematics required for transfer to a four-year program, regardless of whether the mathematics courses serve the primary purpose of the data science program.

Topic	To Work	Transfer	Certificate
Math.A. Calculus	not required	working knowledge/mastery	not required
Math.B. Matrices and basic linear algebra	exposure	exposure	not required
Math.C. Basic probability	working knowledge	working knowledge	not required

4.5 Productivity Foundations

Data science is inherently a collaborative enterprise, and so interpersonal and communication skills are of vital importance. Students will work in teams in the workplace; this is a given. And so programs should teach students to work effectively in groups. Collaborators must communicate effectively with people who share their technical knowledge, but also with colleagues who may be unfamiliar with the language and tools of data science and sometimes with the public. These communications will likely range from informal notes (emails or conversations) to formal multimedia presentations, white papers, and publications.

In addition to these soft skills, there are desired affective outcomes, or "habits of mind", that all programs should strive to cultivate within their students. Students should be prepared to be life-time learners, since technology changes rapidly and these changes affect how data science is conducted. Students should graduate from the program with an empirical mindset, able to recognize when data are required or helpful for decision making, supporting claims, and furthering discovery. A capstone experience can be a good context in which to deepen (and assess) these skills.

4.6 Notes Concerning Outcomes

These notes reiterate and further develop the recommendations of Section 3. They address a wide range of areas, including staffing, pedagogy, and content.

For all programs

- Capstone experiences are necessary for students to achieve mastery. We define a capstone experience as a long-term project that requires students to creatively solve novel and challenging problems and to synthesize understanding and skills acquired across multiple courses. Capstones should produce a tangible product, for example a written report or documented software, and should require students to work in a team where they must communicate both within their team and with those outside of their project.
- Ethics should be explicitly included in the outcome goals of any data science program and should be embedded as a course objective in a number of courses in the curriculum.
- We suggest that introduction and reinforcement of computational and algorithmic thinking be done in the context of data and using a language that is appropriate for data science applications (e.g., Python and/or R).
- Data should be a focus of all courses in the curriculum. This means that more theory-oriented courses (such as mathematics) should be focused on applications of finding knowledge from data.
- Because learning to become a productive member of a team is essential, developing collaboration skills should be integrated throughout the curriculum.
- Classroom environments should be as active as possible, including hands-on experience working with data within the classroom setting.
- Faculty development will be a challenge – and an opportunity – at all colleges that aspire for their faculty to evolve and their institution to build new undergraduate data science programs. The ASA/MAA guidelines for instructors of statistics (http://magazine.amstat.org/blog/2014/04/01/asamaaguidelines/) may be relevant in terms of appropriate background to teach the data science courses that we describe. However, we note that the suggestion that instructors have at least a master's degree in statistics (or computer science) may not be possible at many two-year colleges.

- Provide opportunities for faculty (re-)development, particularly across disciplines. A computer science faculty might have little prior experience in analyzing data, and likewise a mathematics or statistics faculty might have little prior experience with a programming language.
- Programs would benefit from strong partnerships with local employers to ensure graduates are prepared for the workforce.

<u>For Transfer Programs</u>

Associate degree programs for students intending to transfer to a four-year program are often constrained by the transfer requirements demanded of the destination school(s). Still, we recommend that these programs provide students with sufficient skills to earn an internship related to data science at the end of their program, or to at least have the same qualifications for an internship as a rising junior who started at a four-year school.

<u>For Certificate Programs and Direct-to-Work Programs</u>

Certificate programs and direct-to-work programs are often designed to be mindful of the needs of local employers. Therefore two-year colleges should identify potential industry partners to assist in the many facets of developing course content and capstone experiences.

In order to minimize barriers for students who might be working demanding jobs or juggling careers with family and with their education, programs might consider alternative structures to semester-length courses. By modularizing offerings, courses can be broken into six- or eight-week segments. Online learning options should also be considered.

Colleges with multiple certificate programs might increase efficiencies by identifying a set of common core courses that could serve multiple certificate programs.

4.7 Challenges

Two-year colleges face many challenges in designing and implementing data science programs. Many of the challenges identified below will be no surprise to two-year college administrators or faculty. We include them here because they generated a great deal of discussion during the summit, and this list may be useful during the planning stages of a new program.

The consensus of the summit was that these challenges are surmountable, and should in no way inhibit two-year colleges from establishing data science programs.

<u>For all programs</u>

- A successful data science program will require strong relationships with local employers. These relationships will help in providing realistic and meaningful capstone experiences and internships as well as valuable advice on selecting program outcomes. Developing such relationships may require dedicated personnel, both to establish these partnerships and also to serve as liaisons between industry partners, students, and faculty.

- As it is not always clear which department should house a new data science program, two-year colleges should consider the many factors that will determine whether the best implementation will be as a part of an existing program or as a new program.
- Some two-year colleges may find it challenging to find qualified faculty to teach data science courses and/or to find and pay for professional development opportunities for existing faculty. Some two-year schools may need new structures to support faculty development, e.g., if the faculty are mid-career. Given the myriad job opportunities for data scientists, recruitment and retention of faculty with the appropriate skills and backgrounds to teach these courses may be another obstacle. (We note that this is also an ongoing issue for many four-year institutions as well.)
- Computation is at the heart of data science and will require special consideration when planning a program. Cloud-based systems, such as JupyterHub and Rstudio.cloud, provide a potentially useful pathway. The big three cloud providers (Amazon, Microsoft, Google) now provide a plethora of options for instructors and students. Many of these are utilized at four-year institutions but not (yet) at two-year institutions. The separation between administrative computing (managing servers) and academic computing (computer labs and software) at many campuses add further complications to navigate.
- The emergence of a new field of STEM with fewer barriers to entry and that appeals to a broad range of student interests provides an opportunity to create equitable pathways and curricula and to create a profession that reflects the diverse society it serves. Equity is a major challenge for data science programs, and one that should be addressed explicitly and in the planning and assessment stages. For instance, the simple fact that students with greater experience and familiarity with computers are more likely to be drawn to data science poses challenges for equity. Institutions must actively seek pathways to recruit, prepare, and retain students from diverse backgrounds.
- Data science is a developing field. Research in data science education and curriculum is an on-going process. Administrators of data science programs should keep abreast of developments in statistics, computer science, and data science pedagogy.

For Certificate Programs

- Federal financial aid requires minimum credit loads (six credits per term). Students enrolling in certificate programs may be disqualified from financial aid if the program does not require students to enroll in at least six credits.
- Some institutions might prefer to venture into data science education on a smaller scale with a certificate program, and building on success, then establish a two-year degree program around the certificate program. One could imagine a certificate program that is similar to a two-year degree except it lacks the general education requirements.
- Graduates of certificate programs should have ample opportunities to demonstrate their work-readiness to potential employers. A capstone experience provides one such opportunity, as does including real-world problem-solving projects throughout the other courses in the program.
- Student portfolios of completed work and work in progress can provide evidence of skills, achievements, and employability for students.
- There was general agreement about the need to think quite carefully about the role of prerequisites in certificate programs. If courses have a mix of students with advanced degrees and

students with only high school diplomas, the courses might end up being useful to neither group, with the former group not learning the necessary and expected skills and the latter group being too challenged to engage in the material.

- Institutions should take special care to prevent superficial coverage of topics. Emphasize depth over breadth in short programs.
- Certificate programs, perhaps more than the others, will have students with diverse backgrounds concerning the students' mathematical, computational and statistical backgrounds. Care must be taken to provide a challenging yet accessible program.

For Transfer Programs

- An ideal associate degree program would enable students to transfer to data science, statistics, or related programs at a four-year institution, with sufficient credits to achieve junior status upon transfer. This may be challenging to achieve in practice for all students and will require close cooperation with local four-year institutions to draft and regularly update articulation agreements.
- Ideally, an associate degree program would provide students with some job skills immediately after completing the degree, as well as providing the ability to transfer.
- Some students may plan to join the workforce (possibly part-time) upon completing an associate degree, and attend a four-year school while also working. The recent development of associate programs in cybersecurity, along with the existing associate programs in information technology, may serve as useful models for associate programs in data science. Marketable skills include experience in communication, computation, statistics, and solving problems that require students to have engaged in the entire data analysis cycle – from gathering data to communicating of results. After completing an associate degree, students would ideally have comparable skills and qualifications to a rising junior who began at a four-year school.
- There is considerable variability in the level of mathematics required for a bachelor's degree program in data science. Historically, multivariable calculus and linear algebra were prerequisites for probability and theoretical statistics. One might expect that four-year college programs where data science is housed within statistics departments or engineering programs might continue to require high levels of mathematics. On the other hand, the PCMI guidelines recommend a more carefully tailored selection of mathematical topics, and programs that adopt this approach will likely have fewer prerequisites. Data science programs associated with business programs might also be examples of programs with fewer mathematics prerequisites. Articulation agreements may dictate the level of mathematics required for the associate degree program and may limit the student's ability to transfer to a different four-year school's program. If possible, two-year colleges may want to develop multiple pathways for students, especially with respect to mathematics.
- Within a two-year college, the various departments that house the data science courses could impact transferability of courses and/or staffing of instructors. Ensuring that courses and programs have clearly delineated learning outcomes may help make transfer options more transparent to students and advisors.
- Research-based evidence has demonstrated the value of early research opportunities (for example, see the Association of American Colleges and Universities High Impact Practices resources https://www.aacu.org/resources/high-impact-practices). Many four-year colleges have prioritized

support for students to do research projects during their first two years of study. However, such research opportunities, while quite valuable, are not the norm at two-year colleges. Students may be less likely than their four-year counterparts to have opportunities to work on applied projects in large research centers. As a result, students who transfer from two-year colleges may be on unequal footing with their peers in this regard.

- Programs should aim to be as flexible and adaptable as possible. Some students want to pursue major programs of study in data science, but others will only want to pursue a minor or to obtain some data science competencies through completion of a relatively small set of courses. Some students may not decide to study data science until their sophomore year.

5. Sample Curricula

In this section, we propose sample curricula that meet the program outcomes of Section 4. We intend these both as "proof of concept" and also as a preliminary guide to those designing new programs. Each curriculum and its corresponding curriculum map (which maps program outcomes to specific courses) was developed by an independent group of summit participants (and co-authors of this document). Because curricula for certificate programs will vary considerably based a variety of factors, such as industry partners, we felt that a sample curriculum would be less helpful. Instead, we refer interested readers to the websites of the programs listed in Table 2.1.

Tables 5.1 and 5.2 list courses for two-year programs. The courses are given generic names, and a brief description is provided. A more detailed understanding of course content can be gleaned by referring to the curriculum maps. Due to space limitations, these maps are provided online only (https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx). The maps indicate which outcomes are addressed within each course, and can therefore be used to generate content for a course.

5.1 Sample Curriculum for Transfer Students

Table 5.1. Sample curriculum for programs for transfer to a four-year program.

	Semester 1	**Semester 2**	**Semester 3**	**Semester 4**
Course 1	Intro to Data Science	Statistics Foundations	Math for Data Science	Predictive Analytics
Course 2	Calculus I	Data Models and Technologies	GE (Tech. Writing)	Data Structure and Algorithms
Course 3	Data Ethics	Calculus II	Calculus III	Capstone
Course 4	Intro to Computer Science	GE (Public Speaking)	GE	Linear Algebra
Course 5	GE	GE	GE	GE

Introduction to Data Science:	An opportunity for students to experience a data-science project in its entirety and to be exposed to foundational concepts.
Statistics Foundations:	An introductory statistics course that emphasizes exploratory data analysis, inferential statistics, and statistical modeling that includes predictive modeling.
Data Models and Technologies:	A deeper dive into modeling that includes the use of data models and technologies to acquire, clean and manage data.
Data Ethics:	A course discussing issues involving privacy, data security, societal impacts, etc.
Introduction to Computer Science:	Fundamentals of programming, including control structures, data structures (lists, arrays, pointers), and algorithms (e.g., recursion).
Predictive Analytics:	Statistical modeling and machine learning with the aim of developing models to predict future observations in both supervised and unsupervised contexts.
Data Structures:	How data are stored in a variety of formats in order to curate, store, and access data of any type.

5.2 Sample Curriculum for Direct-to-Work Students

Table 5.2 provides a sample curriculum of core courses for programs intended to prepare students to go directly into the workforce after earning their associate degree in data science. Note that most two-year degrees require more courses. This table is intended to show a possible programming sequence for the core courses in a data science program.

Table 5.2. Sample Curriculum for Direct-to-Work Associate Degree Programs

	Semester 1	**Semester 2**	**Semester 3**	**Semester 4**
Course 1	Intro to Data Science	Database Design and Programming	Math for Data Science	Applied Predictive Modeling
Course 2	Intro to Programming	Data Visualization	Intro to Predictive Modeling	Data Analytics
Course 3	Intro to Statistics	Applied Data Programming	Data Structures	Capstone

Introduction to Data Science:	An opportunity for students to experience a data-science project in its entirety and to be exposed to foundational concepts.
Introduction to Computer Science:	Fundamentals of programming, including control structures, data structures (lists, arrays, pointers), and algorithms (e.g. recursion).

Introduction to Statistics:	An introductory statistics course that emphasizes exploratory data analysis, inferential statistics, and statistical modeling that includes predictive modeling.
Database Design/Programming:	Designing databases to meet specific needs of an organization; using query language to access data and manage project workflow.
Data Visualization:	Covers the design and development of a wide variety of data visualizations beyond (but including) the use of "traditional" pre-packaged graphics. This includes interactive graphics such as dashboards and basic animations. Students learn to design visualizations to solve problems and to communicate findings to a variety of audiences.
Applied Data Programming:	Advanced programming in a language used in data science, such as R or Python. Includes data wrangling and preparation as well as algorithms necessary to program statistical routines and fit models.
Math for Data Science:	Applied differential and integral calculus, mathematical modeling, matrices.
Intro to Predictive Modeling:	An introduction to supervised and unsupervised machine learning.
Data Structures:	Experience understanding how data are stored in a variety of formats in order to curate, store, and access data of any type.
Applied Predictive Modeling:	Deeper dive into supervised and unsupervised learning, including cross-validation and resampling methods for goodness of fit, development of the bias-variance tradeoff as a guide for developing models. This course provides an opportunity for students to apply approaches learned in the "introduction" course to real-world problems.
Data Analytics:	Deeper dive into topics introduced in introductory statistics and data science courses with emphasis on application to real-world problems.
Ethics:	In this model, no single class teaches ethics. Instead, it is taught within each course.

5.3 Curriculum Maps

We provide curriculum maps for the sample curricula at https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx. These maps indicate which outcomes are addressed for each course in the sample curricula and the depth of coverage for that outcome: *introduced*, *revisited*, *applied*. The *applied* label was used only in the direct-to-work curriculum to emphasize the need for students to apply concepts and skills learned in earlier courses.

For a learning outcome designated as “exposure” in the curriculum map, we expect that the outcome is addressed in at least one course, and it might even be introduced in one course and revisited in another course. For a learning outcome designated as “working knowledge” in the curriculum map, we expect that the outcome is attained with repeated exposure. While outcomes that are revisited in one or more courses are likely to lead to working knowledge, it is also possible that a single course could provide in depth coverage of a particular outcome and lead to both exposure and working knowledge.

Courses where a learning outcome is labeled as *applied* or a set of courses in which a learning outcome is first *introduced* and then *revisited* move students towards mastery level understanding. Outcomes that are revisited in the context of a capstone course are also seen as moving students towards mastery. It is not possible to say at exactly which point working knowledge becomes mastery, but as a guide, the writing groups felt that mastery requires project work or a capstone experience in which students can achieve mastery by applying knowledge in novel and complex situations.

5.4 Comparisons of Curricula

A comparison of the sample curricula and curriculum maps for transfer and to-work programs show many similarities. Both programs offer introductory statistics as well as introductory data science, and both have a course that provides an introduction to the fundamentals of programming. An examination of the curriculum maps for these courses shows that the groups' conception of these courses differ only in minor ways. Still, there are important differences in these programs.

The primary purpose of programs designed for transfer students is to enable students to transfer to a four-year college or university with junior status. This places considerable restrictions on the program. In particular, at least for the near term, programs for transfer students will most likely be providing a "traditional" series of math courses leading to multivariable calculus or linear algebra. Nonetheless, we recommend a "math for data science" course that would help students see how mathematical theory is applied to the context of data-oriented problems and to emphasize the particular mathematical skills needed for data scientists. This course would emphasize applications over proofs and derivations and would most likely focus on mathematical modeling and matrix algebra.

Transfer programs can also assume that students will have two additional years to study data science and thus might put more weight on achieving "working knowledge" than "mastery" for some outcomes. Additionally, transfer programs will likely require more general education coursework than other types of programs.

Programs preparing to move students directly to the work place can fill their curricula with courses more directly tied to data science. For example, our sample curriculum goes deeply into predictive modeling (machine learning), applications, and programming. Precisely which areas a Direct to Workforce program emphasizes would depend on local employers.

We have not offered sample curricula for certificate programs because there are so many possibilities for what different programs might emphasize and because two-year institutions conceivably have great latitude in designing such programs. The only true constraint is to ensure that graduates of these programs are employable. The Educational Development Center's Oceans of Data Institute has developed a number of resources that should be of use to those designing certificate programs in data science. Specifically, they are engaged in a number of "workforce preparation" projects that address data science pathways and pipelines (http://oceansofdata.org/workforce-prep-projects). A recommended resource is the Profile of a Big-data-enabled Specialist (http://oceansofdata.org/our-work/profile-big-data-enabled-specialist).

6. Looking Ahead

Participants reported being excited and energized by the summit, and were eager to move forward in establishing new programs and in assisting others in establishing new programs. A strong desire was expressed to continue the conversation and hold future summits. In this spirit, the summit concluded by discussing future plans. Most of these plans address some of the challenges outlined in Section 4.6 and were often phrased by participants in terms of content for future national summits. Without arguing whether summits are the best format for achieving future goals, we report on some of the ideas brought up for discussion:

- Professional development for mathematics, computer science, data science, and statistics professors at two-year institutions is vital.
- Local industries and governments are generally not well informed about how graduates with associate degrees can help them. There need to be more conversations between potential employers and course and program designers. Some two-year colleges do not have strong industry contacts and may need assistance in both approaching local employers and involving them in developing programs, hiring students as interns, and hiring graduates as permanent employees.
- Data science degrees at four-year institutions are relatively new, and so new articulation agreements must be forged between two-year colleges and their local four-year institutions. Is there a role for national professional organizations or a national summit/workshop to play in assisting with these relationships?
- How can the emerging data science education community best advise and assist new data science programs in developing introductory data science courses and mathematics-for-data-science courses?

7. Resources

We provide a list of resources suggested by summit participants. This list contains websites, core documents, supplementary documents, a list of some existing introductory data science courses and a list of software resources. The References of this report should also be consulted for additional resources.

Resources and additional information about the summit can be found at
https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx

Compendium Websites
AMATYC Data Science Resources Page
    https://amatyc.site-ym.com/page/DataResources

Core Readings and Websites
The following documents are available at the amstat.org website:

    List of links to two-year college programs as of 5/18
        https://padlet.com/rgould2/r36z3yksg1ey

ACM DSE report (Cassel & Topi 2015)
https://dl.acm.org/citation.cfm?id=2875438
ODI Profile of the Big Data Practitioner
http://oceansofdata.org/our-work/profile-data-practitioner
ODI Workforce Preparation Projects
http://oceansofdata.org/workforce-prep-projects
Park City Math Institute Curriculum Guidelines for Undergraduate Programs in Data Science
https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-060116-053930
NAS Data Science for Undergraduates Report
http://sites.nationalacademies.org/cstb/currentprojects/cstb_175246
GAISE College Report, K-12 Report
https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx
South Big Data Hub's "Keeping Data Science Broad"
https://southbigdatahub.org/programs/keeping-data-science-broad

Supplementary Documents and Websites

The following documents, or links to them, are provided on the amstat.org website
Burdman Degrees of Freedom (Mathematics Pathways at Two-Years)
http://www.edpolicyinca.org/publications/degrees-freedom-varying-routes-math-readiness-and-challenge-intersegmental-alignment-report-2-3-part-series
Dana Center Mathematics Pathways (Mathematics Pathways at Two-Years)
https://www.utdanacenter.org/our-work/higher-education/dana-center-mathematics-pathways
Remedial Education Reforms at California's Community Colleges
https://www.ppic.org/wp-content/uploads/remedial-education-reforms-at-californias-community-colleges-august-2018.pdf

Existing Intro to Data Science Courses

This is not intended to be an exhaustive listing. Several existing two-year college programs list introductory courses but do not provide online syllabi, course outcomes, or topic lists.

Institution	**Course**
Amherst College	https://www.amherst.edu/academiclife/departments/courses/1718S/STAT/STAT-231-1718S
University California, Berkeley ,Computer Science	https://bcourses.berkeley.edu/courses/1267848
University California, Berkeley, Data 8	https://berkeleydsep.gitbook.io/zero-to-data-8
Duke University	http://www2.stat.duke.edu/courses/Fall18/sta112.01
Great Bay CC	http://greatbay.edu/courses/elements-of-data-science
Johnson County CC	http://catalog.jccc.edu/coursedescriptions/ds/#DS_210

Institution	Course
Montgomery College	https://catalog.montgomerycollege.edu/preview_course_nopop.php?catoid=8&coid=11413
Smith College	https://rudeboybert.github.io/SDS192/syllabus.html and https://beanumber.github.io/sds192/syllabus.html
Villanova University, Winston Salem State University	https://www.admiusa.org/admi2016/Papers_Faculty/ADMI2016_DichevEtAll.pdf
Rstudio education project	https://datasciencebox.org

Computing:

R	Statistical Programming Language	https://cran.r-project.org
Rstudio	Programming and Analysis environment for R	https://www.rstudio.com https://rstudio.cloud
Jupyter	"Notebook" environment for Python programming and data analysis	https://jupyterlab.readthedocs.io/en/stable
Anaconda	Environment for Python and R notebooks	https://www.anaconda.com/what-is-anaconda

8. References

Burdman, P. 2015a. DEGREES OF FREEDOM: Diversifying Math Requirements for College Readiness and Graduation (Report 1 of a 3-part series). April 2015. Policy Analysis for California Education. http://www.edpolicyinca.org/publications/degrees-freedom-varying-routes-math-readiness-and-challenge-intersegmental-alignment-report-2-3-part-series

Burdman, P. 2015b. DEGREES OF FREEDOM: Varying Routes to Math Readiness and The Challenge of Intersegmental Alignment (Report 2 of a 3-part series). May 2015. Policy Analysis for California Education. http://www.edpolicyinca.org/publications/degrees-freedom-varying-routes-math-readiness-and-challenge-intersegmental-alignment-report-2-3-part-series

Burdman, P. 2015c. DEGREES OF FREEDOM: Probing Math Placement Policies at California Colleges and Universities (Report 3 of a 3-part series). June 2015. Policy Analysis for California Education. http://www.edpolicyinca.org/publications/degrees-freedom-probing-math-placement-policies-california-colleges-and-universities-report-3-3-part-series

Cassel, B. & Topi, H. 2015, Strengthening Data Science Education Through Collaboration: Report on a Workshop on Data Science Education. NSF Award # DOE 1545135. http://computingportal.org/sites/default/files/Data%20Science%20Education%20Workshop%20Report%201.0_0.pdf

De Veaux, R., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R.J., Sondjaja, M., Tiruviluamala, N., Uhlig, P.X., Washington, T.M., Wesley, C.L., White, D., Ye, P. 2017. Curriculum Guidelines for Undergraduate Programs in Data Science. The Annual Review of Statistics and Its Application, 4:2.1-2.16. doi: 10.1146/annurev-statistics-060116-053930.

Franklin, Rachel S., 2014. Benchmarking Student Diversity at Public Universities in the United States: Accounting for State Population Composition, Ann. Reg Sci. 2012 Oct: 49(2): 355-372. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4262928/#

Futschek, G. (2006). Algorithmic Thinking: The Key for Understanding Computer Science, in *Informatics Education - The Bridge between Using and Understanding Computers: International Conference in Informatics in Secondary Schools - Evolution and Perspectives.* ISSEP 2006, Vilnius, Lithuania, November 7-11.

GAISE College Report, Everson, M. (co-chair), Mocko, M. (co-chair), Carver, R., Gabrosek, J., Horton, N. J., Lock, R., Rossman, A., Holmes Rowell, G., Velleman, P., Witmer, J., Wood, B. 2016. Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report. American Statistical Association. https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx

Hardin, J. & Horton, N. J. (2017). Ensuring That Mathematics is Relevant in a World of Data Science. Notices of the American Mathematical Society. 64. 986-990. 10.1090/noti1572.

Kim, et al fivethirtyeight R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses. *Technology Innovations in Statistics Education*, 11(1). Retrieved from https://escholarship.org/uc/item/0rx1231m

Lane, J., Stodden, V., Bender, S., Nissenbaum, H., (Eds.) (2014). Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press, New York, New York.

National Academy of Science Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective, 2018. "Data Science for Undergraduates: Opportunities and Options." http://nap.edu/25104

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York.

Rawlings-Goss, R., Cassel, L., Cragin, M., Cramer, C., Dingle, A., Friday-Stroud, S., Herron, A., Horton, N.J., Inniss, T.R., Jordan, K., Ordóñez, P. Rudis, M., Rwebangira, R., Schmitt, K., Smith, D., and Stephens, S. (2018) "Keeping Data Science Broad: Negotiating the Digital and Data Divide Among Higher-Education Institutions", https://southbigdatahub.org/publications-and-resources.

SHRM, 2016. Jobs of the Future: Data Analysis Skills. *Society for Human Resource Management Survey Findings*, November 14, 2016. https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/Data-Analysis-Skills.pdf.

Shedden, K. & Prakash, A., 2015. "New Undergraduate Data Science Programs", *Amstat News*, August 2015. http://magazine.amstat.org/wp-content/uploads/2015an/AMSTAT_Augfinal.pdf.

Wild, C. & Pfannkuch, M. (1999). "Statistical Thinking in Empirical Enquiry", *International Statistical Review*, 67(3) 223-265. https://iase-web.org/documents/intstatreview/99.Wild.Pfannkuch.pdf.

Wing, J. (2006), Computational Thinking, *Communications of the ACM* 49(3) 33-35.d.

Appendix A.1 : Summit Planners and Participants

The summit was organized by a steering committee representing stake-holder communities. The steering committee held monthly conference calls which included invited guests from industry, as well as D.J. Patil, former Chief Data Scientist for the White House. Members of the steering committee are listed in Table A.1.

Table A.1 TYCDSS Steering Committee

Name	Affiliations
Rob Gould (co-chair)	Four-Year, PCMI, ASA, AMATYC
Roxy Peck (co-chair)	Four-Year, ASA, AMATYC
Nicholas Horton	Four-year, ASA, NAS
Steve Pierson	ASA
Donna LaLonde	ASA
Randy Kochevar	EDC ODI
Brian Kotz	Two-Year, NAS, AMATYC, ASA
Mary Rudis	Two-Year, Four-Year, ASA
Brad Thompson	Two-Year
Beth Hawthorn	Two-Year, ACM
Heikki Topi	Four-Year, ACM

Key:
PCMI = Park City Math Institute Data Science Curriculum Guidelines
Four-Year = Faculty at four-year university or college
Two-year = Faculty at two-year college
ASA = American Statistical Association
AMATYC = Association of Mathematicians at Two Year Colleges
EDC ODI = Education Development Center Oceans of Data Institute
ACM = Association of Computing and Machine Learning
NAS = National Academies of Sciences Undergraduate Data Science Report

Participants were chosen to include faculty from two-year and four-year institutions, professional societies involved in data science and data science education, researchers in data science and related fields, as well as educators experienced in computer science, data science, and statistics education. In order to build upon the experiences of other data science education efforts, representatives from these efforts were also invited.

Initially, the plan was for a small summit with 30 or so participants. But interest was quite strong and the organizers decided to provide for increased inclusivity and representation. Seventy-eight invitations were sent, and 72 people participated.

Table A.2 illustrates attendees' affiliations, as indicated by the participants themselves. (Participants were asked to state up to three affiliations.) Some participants wore many hats, and so the total number of affiliations exceeds the number of participants. Table A.3 shows primary affiliations only.

Table A.2. Total Affiliations (Participants listed up to 3)

Two-year college	40
Four-year college	16
ACM	3
AMATYC	2
EDC	3
Government	1
High School	1
Industry	5
MAA	2
Big Data Hubs (South and Midwest)	2
NAS	2
NSF	6
ODI	7
PCMI	1

Table A.3. Primary affiliations only.

Primary Affiliations Only	Frequency
Two-year college	39
Four-year college	15
ASA	1
EDC	1
Government	1
High School	1
Industry	5
MAA	1
NSF	6
ODI	2
Total	72

Complete List of Participants

A complete list of participants is provided at https://www.amstat.org/asa/files/pdfs/2018-TYCDSS-Participants.pdf

Appendix A.2 Pre-Summit Participant Preparation

To prepare for the summit, a pre-summit webinar was held on April 18, 2018. The purpose of the webinar was to acquaint participants of the primary goals of the summit and to inform them of the previous work of NAS, PCMI, ODI, and to give them an overview of two established two-year college data science programs. Short presentations were given by Mary Rudis (Great Bay Community College), Brian Kotz (Montgomery College), Randy Kochevar (ODI), Nicholas Horton (NAS), Rob Gould (PCMI), and Roxy Peck (moderator).

Prior to the workshop, participants from two-year colleges were invited to respond to a short survey intended to help us better understand the challenges facing two-year colleges that choose to design data science programs. The survey asked if their institution has a data science program or if their institution is planning to have one. If they did have a data science program, they were asked to describe the type of program, what motivated their institution to create the program, and what obstacles they faced and how they were overcome. If the institution was planning to create a data science program, they were asked what they expected the major challenges to be. Finally, all were asked to identify any currently existing two-year college data science programs they were aware of.

The response rate was high; all two-year college representatives who attended the summit responded, as did one other who ended up not attending. Of those, 10 came from colleges with existing programs and 12 came from colleges currently designing programs.

The 10 colleges with existing data science programs collectively offered five Associate programs intended to transfer to four-year colleges, two AA programs for direct-to-workforce, and nine certificate programs. Two of these colleges offered programs in all three categories.

Appendix A.3 Structure of Summit

Participants gathered at Booz Allen Hamilton offices in Arlington, VA. The day began with a keynote by Sallie Keller, Past-President of the American Statistical Association speaking on "The Need for Data Scientists and the Role of Two-Year Colleges". The student perspective was provided by very recent Great Bay Community College alumna Antonella Duni who told us about her "Journey to Analytics: A Community College Student's Story".

Next, participants were introduced to three "straw curricula". Each was a short description of an example of the three types of programs (Transfer, Workforce, Certificate). We used the term "straw curricula" to indicate that these were starting points in a discussion and were intended, like the "strawman argument", to be knocked down.

Participants were then organized into working groups, and working groups were dedicated to one of the three types of programs. Each group had approximately 12 members, and there were two groups focusing on each type of program. A moderator was appointed for each group, and a member of the writing team served as a note-taker. (Some groups had more than one note-taker.) Work groups were asked to consider:

What are ideal outcomes for a graduate of this program?
What are feasible outcomes?
What prevents the feasible from becoming the ideal?
What should be in a successful student's portfolio?

The overall structure of the summit was to (a) put as many ideas onto the table, without worrying about a consensus, (b) share ideas between the thematic groups, and (c) refine ideas, identify program outcomes with consensus and identify outcomes that were not common to the three types of programs.

Writing Teams
We formed three writing teams, each including members of two-year college and four-year college faculty. The writing teams consisted of Brian Kotz, Mary Rudis, and Joyce Malyn-Smith (Certificate programs); Nicholas Horton, Mark Daniel Ward, and Julie Hanson (associate degree for Transfer programs), Kathy Kubo, Rebecca Wong, Brad Thompson, and Roxy Peck (associate degree for Direct-to-Work). Rob Gould served as an editor and supported the three groups.

Members of the writing team were assigned to the corresponding working groups during the summit and charged with taking notes on the discussion. The writing team then met for one-and-a-half days immediately following the summit to draft the current report. The report was refined through subsequent online discussions and conference calls.

Appendix A.4 Recommendation Summary

Recommendation 1: Create courses that provide students with a modern and compelling introduction to statistics that, in addition to traditional topics in inferential statistics, includes exploratory data analysis, the use of simulations, randomization-based inference, and an introduction to confounding and causal inference.

Recommendation 2: Ensure that students have ample opportunities to engage with realistic problems using real data so that they see statistics as an important investigative process useful for problem solving and decision-making.

Recommendation 3: Explore ways of reducing mathematics as a barrier to studying data science while addressing the needs of the target student populations and ensuring appropriate mathematical foundations. Consider a "math for data science" sequence which emphasizes applications and modeling.

Recommendation 4: Design courses so that students solve problems that require both algorithmic and statistical thinking. This includes frequent exposure to realistic problems that require engaging in the entire statistical investigative process and are based on real data.

Recommendation 5: All programs should (a) expose students to technology tools for reproducibility, collaboration, database query, data acquisition, data curation, and data storage; (b) require students to develop fluency in at least one programming language used in data science and encourage learning a second language.

Recommendation 6: Ethical issues and approaches should be infused throughout the curriculum in any program of data science.

Recommendation 7: Whenever possible, classroom pedagogy should foster active learning and use real data in realistic contexts and for realistic purposes. Programs should consider portfolios as summative and formative assessment tools that both improve and evaluate student learning.