Curriculum Guidelines for Undergraduate Programs in Data Science*

Richard D. De Veaux,[1] Mahesh Agarwal,[2] Maia Averett,[3] Benjamin S. Baumer,[4] Andrew Bray,[5] Thomas C. Bressoud,[6] Lance Bryant,[7] Lei Z. Cheng,[8] Amanda Francis,[9] Robert Gould,[10] Albert Y. Kim,[11] Matt Kretchmar,[12] Qin Lu,[13] Ann Moskol,[14] Deborah Nolan,[15] Roberto Pelayo,[16] Sean Raleigh,[17] Ricky J. Sethi,[18] Mutiara Sondjaja,[19] Neelesh Tiruviluamala,[20] Paul X. Uhlig,[21] Talitha M. Washington,[22] Curtis L. Wesley,[23] David White,[24] and Ping Ye[25]

Annu. Rev. Stat. Appl. 2017. 4:15–30

First published online as a Review in Advance on December 23, 2016

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-060116-053930

*Author affiliations can be found in the Acknowledgments section.

Keywords

curriculum, statistics education, computer science education

Abstract

The Park City Math Institute 2016 Summer Undergraduate Faculty Program met for the purpose of composing guidelines for undergraduate programs in data science. The group consisted of 25 undergraduate faculty from a variety of institutions in the United States, primarily from the disciplines of mathematics, statistics, and computer science. These guidelines are meant to provide some structure for institutions planning for or revising a major in data science.

1. INTRODUCTION

Data science is experiencing rapid and unplanned growth, spurred by the proliferation of complex and rich data in science, industry, and government. Fueled in part by reports, such as the widely cited McKinsey report (McKinsey Global Inst. 2011), that forecast a need for hundreds of thousands of data science jobs in the next decade, data science programs have exploded in academics as university administrators have rushed to meet the demand. The website **http://datascience.community/colleges** currently lists 530 programs in data science, analytics, and related fields at more than 200 universities around the world. The vast majority of these are master's degree and certificate programs offered both traditionally and online. Although PhD programs in data science (or data analytics) are still relatively rare, there has been rapid growth of undergraduate programs at both research institutions and liberal arts colleges. We expect this number to increase significantly in the near future.

The 2016 Park City Mathematics Institute (PCMI), sponsored by the National Science Foundation (NSF) and the Institute for Advanced Study at Princeton, held a workshop focused on the task of producing curriculum guidelines for an undergraduate degree in data science. Twenty-five faculty, comprised of computer scientists, statisticians, and mathematicians from a variety of liberal arts colleges and research universities, met for three weeks to discuss our vision for data science in an undergraduate context, what activities and skills we thought would be necessary for a data science program, and how we could imagine implementing such a major both currently and in the future. These guidelines are the product of that effort.

We have based our guidelines for an undergraduate data science major on a ten semester-course major common among the liberal arts colleges, realizing that research universities typically add several courses to that. We do not intend that these guidelines be prescriptive, but rather we hope that they will serve to inform and enumerate the core skills that a data science major should have before graduation. We started with the reports from the NSF Workshop on Data Science Education (Cassel & Topi 2015), the AALAC (Alliance to Advance Liberal Arts Colleges) conference "Teaching Big Data in the Liberal Arts Context," and the guidelines for undergraduate majors in mathematics, statistics, and computer science (see the sidebar Curriculum Guidelines in Related Disciplines).

We begin by discussing the background and some guiding principles that informed our thinking in Section 2, then consider skills that students should develop while pursuing the major in Section 3, and finally summarize key curriculum topics in Sections 4 and 5. We show a possible selection of current courses that cover most of the basics of our identified skills in Section 6. However, it is important to point out that this smorgasbord approach to course selection is less than ideal. We believe that many of the courses traditionally found in computer science, statistics, and mathematics offerings should be redesigned for the data science major in the interests of

CURRICULUM GUIDELINES IN RELATED DISCIPLINES

- *2015 CUPM Curriculum Guide to Majors in the Mathematical Sciences* (MAA 2015): **http://www.maa.org/sites/default/files/pdf/CUPM/pdf/CUPMguide_print.pdf**
- *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science* (ACM/IEEE 2013): **https://www.acm.org/education/CS2013-final-report.pdf**
- *Curriculum Guidelines for Undergraduate Programs in Statistical Science* (ASA 2014b): **http://www.amstat.org/education/pdfs/guidelines2014-11-15.pdf**

efficiency and the potential synergy that integrated courses would offer. Relying on existing courses at most institutions, a student might have to take 14 or more courses in order to obtain all the skills one would expect from a data science major. With some significant course redesign, we think that this number could be substantially reduced to fit into the constraints of a typical ten-course liberal arts major. Details of those courses are found in the **Supplemental Appendix** (provided in the supplemental material, follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**).

Supplemental Material

2. BACKGROUND AND GUIDING PRINCIPLES

Our curriculum guidelines are guided both by recent work on the relation of data science to the other sciences and the need for a workforce better able to meet the demands of the data-driven economy of the future.

2.1. Data Science as Science

Even though an exact definition of data science remains elusive, we have taken as our starting point a view that seems to have emerged as a consensus from the StatSNSF (National Science Foundation Directorate for Mathematical and Physical Sciences Support for the Statistical Sciences at NSF—a subcommittee of the Mathematical and Physical Sciences Advisory Committee) committee statement that data science comprises the "science of planning for, acquisition, management, analysis of, and inference from data" (NSF 2014, p. 4). At the undergraduate level, we conceive of data science as an applied field akin to engineering, with its emphasis on using data to describe the world. At present, the theoretical foundations are drawn primarily from established strains in statistics, computer science, and mathematics. The practical real-world meanings come from interpreting the data in the context of the domain in which the data arose. For an undergraduate program, we envision a case-based focus and hands-on approach, as is common in fields such as engineering and computer science.

2.2. Interdisciplinary Nature of Data Science

Data science is inherently interdisciplinary. Working with data requires the mastery of a variety of skills and concepts, including many traditionally associated with the fields of statistics, computer science, and mathematics. Data science blends much of the pedagogical content from all three disciplines, but it is neither the simple intersection nor the superset of the three (see the sidebar Is Data Science a Science?). By applying the concepts needed from each discipline in the context of

IS DATA SCIENCE A SCIENCE?

There is still considerable debate about exactly what the science of data science is, but prominent scientists such as David Donoho, Michael Jordan, and others suggest that there is a science at the core and that it will continue to evolve. As Donoho says, "Fortunately, there *is* a solid case for *some entity* called 'Data Science' to be created, which would be a true science: facing essential questions of a lasting nature and using scientifically rigorous techniques to attack those questions" (Donoho 2015, p. 10). Regardless of the consensus (or lack thereof) surrounding the evolution of the science of data science, a data science program at the undergraduate level provides a synergistic approach to problem solving, one that leverages the content in all three disciplines. We believe that a data science program will serve students well whether they join the marketplace or continue on to more advanced study.

data, the curriculum can be significantly streamlined and enhanced. The integration of courses, focused on data, is a fundamental feature of an effective data science program and results in a synergistic approach to problem solving.

This document outlines the core knowledge and methods that data science students should master. Our position is that, ideally, new courses should be developed to take advantage of the efficiencies and synergies that an integrated approach to data science would provide. However, because not all institutions will be able to create many new courses immediately, we suggest which traditional courses might provide coverage of the basic topics of the major. We also propose a model of an integrated curriculum to serve as a possible blueprint for the future.

2.3. Data at the Core

The recursive data cycle of obtaining, wrangling, curating, managing and processing data, exploring data, defining questions, performing analyses, and communicating the results lies at the core of the data science experience. Undergraduates need understanding of, and practice in performing, all steps of this data cycle in order to engage in substantive research questions. In the words of Google's Diane Lambert, students need the ability to "think with data" (Horton & Hardin 2015, p. 259; see also ASA 2014a and Shron 2014). Data experiences need to play a central role in all courses from the introductory course to the advanced elective/capstone. These experiences should include raw data from a variety of sources and should involve the process of cleaning, transforming, and structuring data for analysis. They should also include the topic of data provenance and how it informs the conclusions one can draw from data. Data science is necessarily highly experiential; it is a practiced art and a developed skill. Students of data science must encounter frequent project-based, real-world applications with real data to complement the foundational algorithms and models. The Committee on the Undergraduate Program in Mathematics curriculum guides (MAA 2004, 2015) reinforced the importance of real applications and data analysis for all mathematical science majors. They stated that

> the analysis of data provides an opportunity for students to gain experience with the interplay between abstraction and context that is critical for the mathematical sciences major to master. Experience with data analysis is particularly important for majors entering the workforce directly after graduation, for students with interests in allied disciplines, and for students preparing to teach secondary mathematics. (MAA 2004, p. 45)

Statisticians, naturally, feel the same. In 2002, a report by the ASA recommended that undergraduate statistics curriculum include a heavy emphasis on data analysis (perhaps more weight should be given to the data than the analysis) (ASA 2002). By the same logic, students learn data science by doing data science. The recursive data cycle should be a featured component of most data science learning experiences, and projects involving group analysis and presentation should be common throughout the curriculum. Capstone projects are also an essential component of the experience and internships fit naturally in a data science program.

2.4. Analytical (Computational and Statistical) Thinking

Breiman (2001) spoke of the two cultures of algorithmic (computational) and data (statistical) models (renamed "predictive" and "inferential," respectively, by Donoho 2015). Data science offers the opportunity to integrate and use both computational and statistical thinking to solve problems rather than emphasizing one over the other. Interestingly, these are not new ideas. As Wilkinson (2008) pointed out, many of Tukey's ideas are now conventional wisdom. For example, Tukey accorded algorithmic models the same foundational status as the algebraic (data) models that statisticians had favored in the previous half-century. The two pillars of computational and statistical thinking should not be taught separately. The balance between them may change from one course to another, but both should be present for the most effective and efficient teaching.

2.5. Mathematical Foundations

Data scientists employ models to understand the world and mathematics provides the language for these models, so a working data scientist requires a firm foundation in mathematics. However, traditional mathematics curricula often delay the connection between abstract mathematics and messy, possibly ill-posed real-world problems, especially with respect to those involving data. We propose a fresh approach that distills the essential aspects of mathematics needed for data science at the undergraduate level. Rather than requiring four or more courses in mathematics, an efficient data science major should present these mathematical concepts in two courses, in the context of modeling for data-driven problems. This will streamline the mathematical curriculum to focus on data science rather than theory, derivations, or proofs. In particular, we propose modeling (both algorithmic and statistical) as a motivator for mathematical tool development, introducing concepts as they become necessary in order to solve our real-world problems. Matrix algebra is motivated by solving linear systems, derivatives are motivated by optimization and sensitivity analysis, and integration is motivated by probabilistic applications. Although this shift toward modeling and applications is in line with current trends in calculus reform, our approach is revolutionary in its approach to the mathematics, its ordering of topics, and the selection of topics.

A possible caveat is that our approach is not meant to serve as an alternative pathway for the mathematics major, though students introduced to mathematics this way may want to further their study. Such students would need additional theoretical foundations before entering traditional upper-division mathematics courses.

2.6. Flexibility

We must prepare students to learn new techniques and methods that may not exist today. They will need to work with increasingly varied forms of data, or they will not be prepared for the jobs of the future. We need to pay attention to the core foundations of mathematical, computational, and statistical thinking and practice while incorporating the practical and important data science skills.

Data science, at all levels, is evolving and changing quickly. Most institutions will implement a data science major from current courses in existing disciplines, perhaps transitioning to more fully integrated courses as outlined in the **Supplemental Appendix** (provided in the supplemental material, follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**) at a future date. Our hope is that institutions use these guidelines in their planning to meet the needs of their students both now and in the future. We fully expect that institutions will regularly review their programs to reflect new developments in this fast-evolving field.

Supplemental Material

3. KEY COMPETENCIES AND FEATURES OF A DATA SCIENCE MAJOR

A graduate with an undergraduate data science degree should be prepared to interact with data at all stages of an investigation (see the sidebar Key Competencies for an Undergraduate Data Science Major) and will be expected to work within a team environment.

KEY COMPETENCIES FOR AN UNDERGRADUATE DATA SCIENCE MAJOR

- Computational and statistical thinking
- Mathematical foundations
- Model building and assessment
- Algorithms and software foundation
- Data curation
- Knowledge transference—communication and responsibility

3.1. Analytical (Computational and Statistical) Thinking

Data science consists of a problem-solving approach for working within empirical settings in which meaning must be extracted from data. This approach is a synthesis of modes of thought in statistics, computer science, and mathematics.

1. **Statistical thinking in a data-rich environment.** Statistical thinking is an approach to understanding the world through data, and involves everything "from problem formulation to conclusions" (Wild & Pfannkuch 1999, p. 1). The data scientist needs an understanding of basic statistical theory. Students should understand the basic statistical concepts of data analysis, data collection, modeling, and inference. A sound knowledge of basic theoretical foundations will help inform their analyses and the limits to their models. Successful graduates will be able to apply statistical knowledge and computational skills to formulate problems, plan data collection campaigns or identify and gather relevant existing data, and then analyze the data to provide insights.
2. **Computational thinking.** Working with data requires extensive computing skills. A data science student must be prepared to work with data as they are commonly found in the workplace and research labs. For example, accessing and organizing data in databases, scraping data from websites, processing text into data that can be analyzed, and ensuring secure and confidential data storage all require extensive computing skills. These computational problem-solving skills recur throughout the workflow of the data scientist. As Wing (2006, p. 34) put it, "thinking like a computer scientist means more than being able to program a computer. It requires thinking at multiple levels of abstraction." Data science graduates should be proficient in many of the foundational software skills and the associated algorithmic, computational problem solving of the discipline of computer science. To be prepared for careers in data science, students also need facility with, and exposure to, professional statistical analysis software packages, and an understanding of the principles of programming and algorithmic problem solving that underlie these packages.
3. **Integration of approaches.** Data science at the undergraduate level combines computational and statistical thinking practices, relying on mathematical foundations. In addition to their competence in the areas noted below, data science graduates should demonstrate an understanding of the connections between these knowledge domains. They should be able to bring a wide array of different skills and problem-solving approaches to bear on any particular problem and should make informed choices about which skills are appropriate in a given setting. They should have facility for working with a diverse collection of tools, as well as for learning new tools and—in some cases—contributing to the development of new tools themselves.

   Computing environments, both software and hardware, change rapidly and frequently, and these changes have consequences for data structure, data storage, and computational efficiency. Data scientists must be capable of adapting smoothly to such changes. Data scientists should understand both the computational and modeling challenges in their work, and how they might be intertwined. For example, data scientists should recognize that fitting a given model on a particular set of data will engender computational challenges, and they should have some facility for implementing a solution that may involve either a modification of the model or a change in the computing environment, or both. When integrated with statistical thinking, computational thinking greatly amplifies the ability of data scientists to distribute solutions to clients, understand many modern statistical modeling approaches, and achieve scientific reproducibility.

3.2. Mathematical Foundations

Mathematically speaking, the emphasis of an undergraduate data science degree should be on choosing, fitting, and using mathematical models. Because data-driven problems are often messy and imprecise, students should be able to impose mathematical structure on these problems by developing structured mathematical problem-solving skills. Students should have enough mathematics to understand the underlying structure of common models used in statistical and machine learning as well as the issues of optimization and convergence of the associated algorithms. Although the tools needed for these 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.

3.3. Model Building and Assessment

1. **Informal modeling.** Statistical models are used to describe, predict, and explain processes, but they are also used to communicate understandings and lay foundations for future models. Informal modeling involves identifying potential sources of variation, discerning between stochastic and deterministic variation, and understanding how these might be modeled mathematically and computationally. Graduates must also be adept at data visualization, which is an important tool in informal modeling, as it can be used to communicate with others and identify weaknesses in proposed models.
2. **Formal modeling.** Graduates should be able to build and assess statistical and machine learning models, employ a variety of formal inference procedures, and draw conclusions of appropriate scope from the analysis. This includes understanding how data issues (e.g., collection methods, sources of bias and variance) impact the analysis, interpretation, and generalization of statistical findings. Graduates should also be able to bring computational considerations to bear in the analysis of data, including issues of scale.

3.4. Algorithms and Software Foundations

The data science graduate should be able to employ algorithmic problem-solving skills to the task at hand. These include defining clear requirements to a problem, decomposing the problem, using efficient strategies to arrive at an algorithmic solution, and implementing solutions through programming in a suitable high-level language. Graduates should understand the memory and execution performance of the structures and software they create, and that of the libraries and packages they use. They should know and utilize good practices in documentation and structure and be able to use appropriate tools for maintaining their software. They should be able to leverage existing packages and tools to solve their computational problems.

3.5. Data Curation

Data curation involves managing data through the entire problem-solving process. The two main steps are outlined below.

1. **Data preparation.** Graduates should be able to work with data from a variety of sources and formats. Data may come from a web page, a database, or a stream, and may consist of images, sounds, or video, as well as numbers or text. These data may have been collected through a controlled experiment or an observational study, or may be opportunistic data collected through sensors or an automated procedure. Given a particular data set, graduates should be able to prepare the data for use with a variety of statistical methods and models and should recognize how the quality of the data and the means of data collection may affect conclusions.
2. **Data management**. Data scientists must not only prepare data for analyses, but also ensure the integrity of the data while it passes through all stages of the analysis. This requires working with relational databases [such as a Structured Query Language (SQL) database], maintaining version control, and tracking data provenance as data from multiple sources are merged.

3.6. Knowledge Transference

Data do not exist in a vacuum, but arise from a particular context. Knowledge of that context is necessary to analyze the data, and thus undergraduates need experience applying their discipline outside the core of statistics, computing, and mathematics. A data science program should feature data-based investigations in a complementary discipline, such as a physical science, a life science, business, a social science, the humanities, or the fine arts. There are two important areas that training should address.

1. **Communication.** Effective communication is a core skill of the data scientist. As members of a team, data scientists must communicate to teammates as well as to those with less intimate knowledge of the project particulars. Increasingly, data scientists communicate directly to the public via both static and interactive data visualizations. A thoughtful data science program integrates communication-based opportunities and learning development throughout the whole of the curriculum rather than partitioning them into separate classes. Students should gain experience using oral, written, and visual modes to communicate effectively to a variety of audiences.
2. **Ethics and reproducibility.** The capabilities of data science introduce new ethical questions. Programs in data science should feature exposure to and ethical training in areas such as citation and data ownership, security and sensitivity of data, consequences and privacy concerns of data analysis, and the professionalism of transparency and reproducibility.

4. CURRICULAR CONTENT FOR DATA SCIENCE MAJORS

The goal of our curriculum is to repeatedly engage students in the full cycle by which we learn from data and to help them acquire the skills listed in the previous section. This necessitates the interweaving and integration of traditionally siloed topics and tools into a cohesive presentation. Our data science program includes six main subject areas, which are described below and in the sidebar Six Main Subject Areas of a Data Science Major. We then describe suggested courses that prepare students in all six areas (also see the sidebar An Outline of the Data Science Major). In the **Supplemental Appendix**, we show in more detail how courses in each of these cycles might be constructed. A summary of the courses designed for these subject areas is found in the following.

Supplemental Material

SIX MAIN SUBJECT AREAS OF A DATA SCIENCE MAJOR

- Data description and curation
- Mathematical foundations
- Computational thinking
- Statistical thinking
- Data modeling
- Communication, reproducibility, and ethics

AN OUTLINE OF THE DATA SCIENCE MAJOR

1. Introduction to data science
   - Introduction to Data Science I
   - Introduction to Data Science II
2. Mathematical foundations
   - Mathematics for Data Science I
   - Mathematics for Data Science II
3. Computational thinking
   - Algorithms and Software Foundations
   - Data Curation—Databases and Data Management
4. Statistical thinking
   - Introduction to Statistical Models
   - Statistical and Machine Learning
5. Course in an outside discipline
6. Capstone course

4.1. Overview of Course Sequence

- **Introduction to Data Science I and II.** Students will immediately use a high-level language to explore, visualize, and pose questions about data. In the second semester, a more algorithmic language may be introduced to help students understand the thinking and structure behind the higher-level functions they experienced in the first semester.
  - Introduction to high-level language
  - Exploring and manipulating data
  - Functions and basic coding
  - Introduction to modeling, both deterministic and stochastic
  - Concepts of projects and code management
  - Databases
  - Introduction to data collection and statistical inference
- **Mathematical Foundations I and II.** Data science students connect mathematical tools to real-world problems. Unlike pure mathematics, which seeks to build theory and prove propositions, data science is about seeing the value of mathematical methods while understanding their limitations. Data science students should also develop a geometric, intuitive,

visual way of thinking throughout their mathematical training. We propose a two-semester Mathematics for Data Science sequence that begins with students who would have placed into Calculus 1. This sequence emphasizes mathematical modeling, especially linear and polynomial models (see the **Supplemental Appendix**), and would include the following topics.

Supplemental Material

- Mathematical structures (e.g., functions, sets, relations, and logic)
- Linear modeling and matrix computation (e.g., matrix algebra and factorization, eigenvalues/eigenvectors, and projection/least-squares)
- Optimization (e.g., calculus concepts related to differentiation)
- Multivariate thinking (e.g., concepts and numerical computation of multivariate derivatives and integrals)
- Probabilistic thinking and modeling (e.g., counting principles, univariate and multivariate distributions, and independence, relying often on computational simulations)

- **Algorithms and Software Foundations.** To develop a grounded computational ability, a data science undergraduate should study foundational computer science topics and build facility in algorithmic problem solving and development of software/programming.
  - Algorithm design: Students must develop the skill set to understand the problem, break it into manageable pieces, assess alternative problem-solving strategies, and arrive at an algorithm that efficiently solves the problem.
  - Programming concepts and data structures: Students should have the knowledge to implement their algorithms using procedural and functional programming techniques and their associated data structures, including lists, vectors, data frames, dictionaries, trees, and graphs.
  - Tools and environments: Students should understand the appropriate use of tools and packages available. Such packages enable programmatic access to data services and input/output; perform data transformations, explorations, visualization, and analysis; and assist in the development and maintenance of software, including development environments, and tools for versioning and tracking.
  - Scaling for big data: As the data and processing associated with data science continue to scale, data science undergraduates should develop the capacity to work with larger data sets. They should be able to apply techniques in concurrent programming to build systems that perform parallel processing of data. They must also be able to work with current and new forms of distributed data storage as a part of the data management areas discussed above. They should be knowledgeable in how to work with streaming data.
- **Data Curation—Databases and Data Management.** A data science undergraduate major must understand and be able to effectively apply principles of data management. This is much broader than traditional database management and must include systems supporting the volume and velocity attributed to big data. Thus a data science major must apply knowledge of data query languages to relational databases and emerging large store NoSQL (not only SQL) data systems, and must be able to access data from less-structured systems through web services, lower-level access to data available across the Internet, and data sourced from streams. Once data are collected, data management includes cleaning and initial structuring, using the software knowledge and skills outlined above, and then transforming data into structured forms required for exploration, visualization, and analysis.
- **Introduction to Statistical Models.** This serves to introduce students to the statistical analysis of data and the elements of a framework for inference. The foundation is linear models, which are then compared to nonlinear approaches. The course builds on important

concepts introduced in the first-year data science courses that form the foundation of any statistical analysis. All the ideas are firmly grounded in and inspired from real-world data.

- Exploratory data analysis approaches and graphical data analysis methods
- Estimation and testing: exposure to statistical (e.g., basic central limit theory and law of large numbers) and algorithmic (e.g., bootstrap resampling methods) approaches to point and interval estimation and hypothesis testing; likelihood theory; and Bayesian methods
- Simulation and resampling: Monte Carlo simulation of stochastic systems; resampling-based inference (e.g., bootstrap, jackknife, permutations); basic understanding of design of studies, surveys, and experiments (e.g., random assignment, random selection, data collection, and efficiency) and issues of bias, causality, confounding, and coincidence
- Introduction to models: simple linear, multivariate, and generalized linear models; algorithmic models (e.g., regression trees and nearest neighbors); and unsupervised learning (e.g., clustering)
- Introduction to model selection and performance: regularization, parsimony and bias/variance tradeoff; loss functions and model selection (e.g., cross-validation, penalized regression, and ridge regression)

■ **Statistical and Machine Learning.** This course blends the algorithmic perspective of machine learning in computer science and the predictive perspective of statistical thinking. Its focus is on the common machine learning methods and their application to problems in various disciplines. The student will gain not only an understanding of the theoretical foundations of statistical learning, but also the practical skills necessary for their successful application to new problems in science and industry.

- Further exploration of alternatives to classical regression and classification
- Algorithmic analysis of models, addressing issues of scalability and implementation
- Performance metrics and prediction, and cross validation
- Data transformations: re-expression of variables and feature creation, techniques of dimension reduction (e.g., principal component analysis), and smoothing and aggregating
- Supervised learning versus unsupervised learning
- Ensemble methods (e.g., boosting, bagging, and model averaging)

■ **Data in Context—Capstone Experience.** A capstone experience in which students consider scientific questions, collect and analyze data, and communicate the results.

A possible path through the major is shown in **Figure 1**.

5. ADDITIONAL CONSIDERATIONS

1. **Graduate study.** Students interested in graduate study in mathematics, statistics, or computer science may consider taking more advanced courses in theoretical foundations. The courses in mathematics for data science will not likely prepare a student for immediate acceptance into a PhD program in one of the three disciplines.
2. **Articulation with community colleges.** Community colleges attempt to prepare students for many different purposes and institutions and, as a result, institutional change may be slow. In the meantime, given the existing course structure, students can prepare themselves to transfer to a college or university data science degree program.

   ■ Students can prepare by taking Calculus 1 and 2 as well as an Introduction to Computer Science course. Additional computer science courses, if offered, would be very helpful preparation. A few community colleges teach introductory statistics courses that

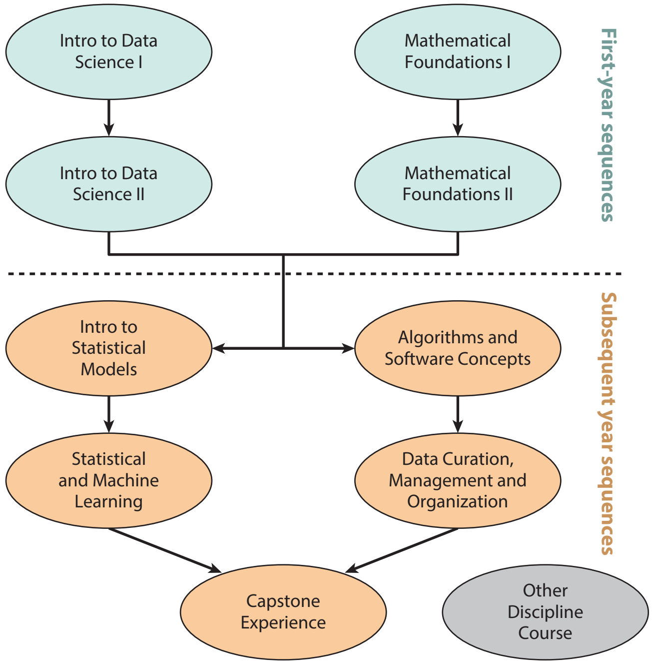

Figure 1

A flow chart displaying a possible path through the data science major.

emphasize data analysis and statistical thinking, and such courses should be required for transfer. More mathematical statistics courses that emphasize a rote, methods-based approach to statistics may not be an optimal preparation for data science.

- Not all community colleges will have the resources within a single department to develop a course such as the Data Science I and II courses proposed here. However, institutions should encourage collaboration between departments of mathematics and computer science in order to develop introductory statistics courses that (*a*) emphasize statistical thinking in the context of real and complex data sets, (*b*) develop fundamental computational thinking through learning and using statistical software, and (*c*) develop basic data handling skills, such as creating new variables through transformations, uploading data with different delimiter types and different basic row/column structures, and developing habits of reproducibility.
- The Statway (**http://www.carnegiefoundation.org/resources/videos/introducing-statway/**) and the New Mathways (**http://www.utdanacenter.org/higher-education/new-mathways-project/**) course sequences offered at some institutions may, depending on the local implementation, provide students with a strong grounding in statistical thinking and, if students learn to explore statistical concepts through simulations, also develop basic computational thinking.

3. **Prerequisites and preparation in high school.** As students exposed to Common Core standards for statistics enter college, some introductory material may need to be reexamined. To be prepared for the data science curriculum, students should

- Be calculus ready (i.e., have a good precalculus course)
- Have an understanding of basic matrix algebra (solving systems of linear equations)

- Understand the basics of a relationship between variables, including scatterplots, the idea of correlation, and the line of best fit
- Have experience with descriptive statistics: notions of center, spread, and skew

4. **Internship and applied experiences.** Internships and other applied experiences are a significant part of a data science program. Practical projects should be implemented often throughout the curriculum and provide the central experience of a capstone course.

6. TRANSITIONING TO A DATA SCIENCE MAJOR USING TYPICAL EXISTING COURSES

The curriculum we have outlined is founded on an integrated set of courses that span topics in three disciplines: mathematics, computer science, and statistics. Many of these topics are covered in traditional courses found in those disciplines. The courses shown in bold are the ten courses that cover the bare minimum of the basic skills needed for data science. However, it is important to note that students will not be exposed to the richness of the interactions of these areas by taking only this set.

6.1. Courses in Mathematics

In order to best serve data science majors, math courses should emphasize connecting the mathematics to real-world problems, especially data-driven problems.

- **Calculus 1**
- **Calculus 2**
- Calculus 3
- **Linear Algebra**
- Probability Theory
- Discrete Math

6.2. Courses in Computer Science

The content in computational thinking courses is distributed across several traditional computer science courses, but not entirely contained in the three required courses listed.

- **Introduction to Computer Science**
- **Computer Science 2: Data Structures/Algorithms**
- Computer Systems and Architecture
- Advanced Algorithms
- **Databases**
- Software Engineering

6.3. Courses in Statistics

Content in the Introduction to Statistics course should follow the revised Guidelines for Assessment and Instruction in Statistics Education for college courses (**http://www.amstat.org/education/gaise**).

- **Introduction to Statistics**
- **Statistical Modeling/Regression**
- **Machine Learning/Data Mining**
- Theory of Statistics (requires Probability Theory)

6.4. Related Courses

- Introduction to partner discipline
- Intermediate course in partner discipline
- **Capstone Course with Data Experience and Projects**
- Two courses in writing, preferably one in technical writing
- Public Speaking
- Ethics

The ten bold courses cover the bare necessities of the material required for a data science major. We suspect many programs will want to add more courses to provide additional content and experience. For a more integrated major with ten newly designed courses, please see the **Supplemental Appendix**.

Supplemental Material

7. SUMMARY AND NEXT STEPS

We hope that these guidelines can serve as a starting point for discussion for building new programs and transitioning existing programs.

SUMMARY POINTS

In summary, the key points of our proposal involve:

1. Data science is a fast evolving discipline centered on the acquisition, curation, and analysis of data.
2. Courses from the traditional disciplines of mathematics, statistics, and computer science provide the basic infrastructure for the major at present.
3. A redesign of the curriculum, integrating the elements of mathematical foundations and computational and statistical thinking at all levels, will provide a rich and effective series of courses to prepare graduates for a career in data science.

We realize that the field is evolving rapidly but hope that the basic areas we have outlined will be useful. During our discussions, several issues arose that were outside the scope of our meeting, including the following.

FUTURE ISSUES

1. **Faculty development.** The courses outlined in the **Supplemental Appendix** are clearly bold steps toward a new integrated program in data science. To be effective they will require many iterations. Resources for faculty including notes, examples, case studies, and—perhaps most importantly—new textbooks will be essential.
2. **Engagement with two-year colleges and high schools.** The data science major will be attractive to many students coming from both high school and two-year colleges. Interactions with these institutions will be crucial in order to coordinate courses and instruction to facilitate transfer to four-year institutions.

3. **Periodic revision.** This is a first attempt at providing concrete guidelines for this emerging field. We realize that revisions will be necessary as the field continues to evolve, and we welcome feedback on these guidelines.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to thank the PCMI for supporting us in this effort. In addition, we would like to thank the National Science Foundation and the Institute for Advanced Study for supporting PCMI. Affiliations for the authors of this article are as follows:

[1]Department of Mathematics and Statistics, Williams College, Williamstown, Massachusetts 01267

[2]Department of Mathematics and Statistics, University of Michigan, Dearborn, Michigan 48128-2406

[3]Department of Mathematics and Computer Science, Mills College, Oakland, California 94613

[4]Department of Statistical & Data Sciences, Smith College, Northampton, Massachusetts 01063

[5]Department of Mathematics, Reed College, Portland, Oregon 97202

[6]Department of Mathematics and Computer Science, Denison University, Granville, Ohio 43023

[7]Department of Mathematics, Shippensburg University, Shippensburg, Pennsylvania 17257

[8]Department of Mathematics, Olivet Nazarene University, Bourbonnais, Illinois 60914

[9]Department of Mathematics, Brigham Young University, Provo, Utah 84601

[10]Department of Statistics, University of California, Los Angeles, Los Angeles, California 90095-1554

[11]Department of Mathematics, Middlebury College, Middlebury, Vermont 05753

[12]Department of Mathematics and Computer Science, Denison University, Granville, Ohio 43023

[13]Department of Mathematics, Lafayette College, Easton, Pennsylvania 18042-1780

[14]Department of Mathematics and Computer Science, Rhode Island College, Providence, Rhode Island 02908

[15]Department of Statistics, University of California, Berkeley, California 94720

[16]Department of Mathematics, University of Hawaii, Hilo, Hawaii 96720-4091

[17]Department of Mathematics, Westminster College, Salt Lake City, Utah 84105

[18]Department of Computer Science, Fitchburg State University, Fitchburg, Massachusetts 01420

[19]Department of Mathematics, New York University, New York, New York 10012

[20]Department of Mathematics, University of Southern California, Los Angeles, California 90089

[21]Department of Mathematics, St. Mary's University, San Antonio, Texas 78228

[22]Department of Mathematics, Howard University, Washington, DC 20059

[23]Department of Mathematics, LeTourneau University, Longview, Texas 75602
[24]Department of Mathematics and Computer Science, Denison University, Granville, Ohio 43023
[25]Department of Mathematics, University of North Georgia, Oakwood, Georgia 30566

LITERATURE CITED

ACM/IEEE (Assoc. Comput. Mach./Inst. Electr. Electron. Eng.). 2013. *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*. New York: ACM. **https://www.acm.org/education/CS2013-final-report.pdf**

ASA (Am. Stat. Assoc.). 2002. Curriculum guidelines for bachelor of arts degrees in statistical science. *J. Stat. Educ.* 10(2). **http://ww2.amstat.org/publications/jse/v10n2/tarpey.html**

ASA (Am. Stat. Assoc.). 2014a. *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society*. Arlington, VA: ASA. **http://www.amstat.org/asa/files/pdfs/POL-BigDataStatisticsJune2014.pdf**

ASA (Am. Stat. Assoc.). 2014b. *Curriculum Guidelines for Undergraduate Programs in Statistical Science*. Alexandria, VA: ASA. **http://www.amstat.org/education/pdfs/guidelines2014-11-15.pdf**

Breiman L. 2001. Statistical modeling: the two cultures. *Stat. Sci.* 16(3):199–231

Cassel B, Topi H. 2015. *Strengthening Data Science Education Through Collaboration*. Rep. on Workshop on Data Science Education Funded by the Natl. Sci. Found., Oct. 3–5, Arlington, VA

Donoho D. 2015. *50 Years of Data Science*. Presented at Tukey Centennial Worksh., Princeton, NJ, Sept. 18

Horton NJ, Hardin JS. 2015. Teaching the next generation of statistics students to "think with data": special issue on statistics and the undergraduate curriculum. *Am. Stat.* 69:259–65

MAA (Math. Assoc. Am.). 2004. *Undergraduate Programs and Courses in the Mathematical Sciences: CUPM Curriculum Guide 2004*. Washington, DC: MAA. **http://www.maa.org/programs/faculty-and-departments/curriculum-department-guidelines-recommendations/cupm/cupm-guide-2004**

MAA (Math. Assoc. Am.) 2015. *2015 CUPM Curriculum Guide to Majors in the Mathematical Sciences*. Washington, DC: MAA. **http://www.maa.org/sites/default/files/pdf/CUPM/pdf/CUPMguiderint.pdf**

McKinsey Global Inst. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. New York: McKinsey & Co. **http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation**

NSF (Natl. Sci. Found.). 2014. *Data Science at NSF. Draft Report of StatSNSF Committee: Revisions Since January MPSAC Meeting*. April. **https://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf**

Shron M. 2014. *Thinking with Data: How to Turn Information into Insights*. Sebastopol, CA: O'Reilly Media Inc.

Wild CJ, Pfannkuch M. 1999. Statistical thinking in empirical enquiry. *Int. Stat. Rev.* 67(3):223–65

Wilkinson L. 2008. The future of statistical computing. *Technometrics* 50(4):418–35

Wing J. 2006. Computational thinking. *Comm. ACM* 49(3):33–35

Annual Review of Statistics and Its Application

Volume 4, 2017

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics