ENSURING THAT MATHEMATICS IS RELEVANT IN A WORLD OF DATA SCIENCE

Johanna S. Hardin (Pomona College) and Nicholas J. Horton (Amherst College)

The recent growth of data science has been remarkable. Analysts now have rich data and powerful computational tools to help answer important questions. Examples of ways that insights can be wrangled from this information abound in diverse areas. This has led some to dub computational thinking (or fluency) as the "new literacy" on par with writing and quantitative skills. A major unanswered question relates to the role of mathematics in the training of future data scientists. How can we be sure that data science is on a firm mathematical and statistical foundation? In the article, we will consider what courses in mathematics would best prepare future data scientists.

Background and brief history

Some institutions have responded to the development of data science by creating innovative new programs. At the University of California, Berkeley the Data 8 introductory course (http://data8.org/) is now offered to a large proportion of incoming students, with connector courses on topics such as genomics, neuroscience, cultural data, social data, demography, smart cities, ethics, and social networks (as well as courses in statistics and mathematics). Many (most?) other four-year colleges and universities are responding with their own initiatives.

While data science is often described as a new discipline, those in the mathematical sciences have been engaged with data science for decades. In a widely referenced call to action, Donoho (2017, in press), quotes noted statistician John Tukey from 1962 who presaged "an as-yet unrecognized science, whose subject of interest was learning from data, or 'data analysis' ". In his paper, Donoho describes the history of data science as a new field and speculates about a future that brings together statistics and machine learning by marrying computational and inferential methods. His proposed "Greater Data Science" (GDS) includes six main divisions (see Table 1).

INSERT TABLE 1 AROUND HERE

Table 1: David Donoho's Six Main Divisions for a "Greater Data Science" (Donoho, 2017)

1. Data exploration & preparation: addresses the 80% (or more) of data wrangling needed prior to analysis
2. Data representation and transformation: including modern databases and special types of data
3. Computing with data: multiple environments, high-performance computing, and workflow
4. Data visualization and presentation: as a way to explore and present results in static or dynamic form
5. Data modeling: including both generative (stochastic model) and predictive (modern machine learning)
6. Science of data analysis: described as one of the most complicated of all sciences

What mathematical preparation do future data scientists need?

What training is needed for data scientists to be able to extract meaning from data? This question is the topic for discussion by several working groups of the National Academy of Sciences as well as a working group from the 2016 Park City Mathematics Institute. The potential for missteps, overgeneralization, and inferential errors abound. One of the challenges in training the next generation of students to 'think with data' is to ensure that they have sufficient background in the mathematical sciences to provide a firm foundation for their future work in data science.

Unfortunately, many new data science programs have arisen that provide little or no formal preparation in the theoretical (mathematical, statistical and computational) underpinnings of this new field. While data science programs should appropriately focus on applications and practice, underlying many approaches is the use of modeling, a topic very familiar to the mathematical sciences, and abstraction, which underlies modern mathematics, statistics, and computational science. Practitioners need to understand when methods are applicable, where they are robust to underlying assumptions, and the potential for misbehavior. The danger is that students who skip out on math completely run the peril of "black box thinking", with no understanding of the *uncertainties* and *limitations* of models and algorithms. We argue that key concepts in statistics and mathematics undergird data science and that these essential aspects are needed as a foundation for data science. Additionally, we believe that mathematicians should take on the mantle of being directly involved in curricular decisions with respect to new data science programs.

What kind of training in mathematics would be ideal for a future data scientist? It is not, we argue, the same training as would be ideal for a future mathematician. The proposal we outline below (two new courses on mathematics for data scientists) creates a path for integration of mathematics into data science. These new courses would not replace existing paths, since different preparation is needed for students who will be pursuing graduate degrees in mathematics.

A gathering of computer scientists, statisticians, and mathematicians assembled at Park City Mathematics Institute (PCMI) during the summer of 2016 to propose guidelines for the discipline of Data Science (De Veaux et al, 2017). The group suggested that data science majors would indeed be well prepared by three semesters of calculus (including single and multivariable), Linear Algebra, Discrete Math, and Probability (in addition to several courses in statistics). They also noted, however, that such a course progression is *not feasible* for all students: it is *not realistic* for students to build a mathematical foundation that consists of such a long string of prerequisite courses *before* starting courses within their own “data science” curriculum (even if space could be made, the leakiness of lower-division pathways is a continuing problem, see http://www.tpsemath.org/).

Project INGeniOuS (Investing in the Next Generation through Innovative and Outstanding Strategies, http://www.maa.org/programs/faculty-and-departments/ingenious) focused on ways that the mathematical sciences could help prepare the next generation of STEM students (at the same time that the mathematical sciences remained a vibrant choice for students). The joint report by the AMS, MAA (Mathematical Association of America), SIAM (Society for Industrial and Applied Mathematics), and the ASA (American Statistical Association) highlighted the importance of alternative curricular pathways and new approaches to teaching to ensure that the mathematical sciences are not left out of the growth of data science and other innovative interdisciplinary programs: "Curricula in the mathematical sciences traditionally aim toward upper-level majors’ courses focused on theory. Shorter shrift is usually given to applications that reflect the complexity of problems typically faced in BIG

(Business/Industry/Government) environments, and to appropriate uses of standard BIG technology tools."

How can the mathematics community respond to the challenge being posed by the growth of data science? We don't have all the answers, but we see the mathematical sciences as a key component of a vibrant and useful data science curriculum that provides students with a solid theoretical foundation. We suggest that the solution is to make changes to the mathematics and data science curricula to give future data scientists a glimpse into the power of mathematics and statistics for modeling and understanding a larger quantitative framework. Our fear is that the important mathematical foundational ideas will get lost if alternate pathways are not developed.

Mathematics preparation

What then, is needed in terms of mathematical preparation? In order for students to be able to function effectively in the world of data science, we believe that that mathematics departments new to consider developing additional entry points as service courses.

We propose two new courses - one discrete and one continuous (other approaches with similar pedagogic goals would also be natural to consider) that intertwine abstraction, modeling, and problem-solving. The idea of two new courses comes directly from the PCMI report:

> 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 [ideas] on [data science] 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.

Proposed New Course 1 (Mathematical Foundations I: Discrete Mathematics):

The first proposed mathematics course formalizes the connections between mathematics and discrete model building (which leads naturally to more sophisticated topic and extensions in terms of continuous distributions, multivariate relationships, and causal inference). Combinatorial techniques can provide concrete pathways for explicitly conceptualizing models and their limitations. Linear algebra allows ideas of multivariate relationships, including independence. Many computer science departments teach a discrete course in their own departments. We suggest that those courses often focus more on algorithms as opposed to our suggestion that discrete models be used to conceptualize and model actual data and real world scenarios (and further develop the ability to problem solve using mathematics).

The discrete topics suggested below would help the data scientist communicate about the multivariate problems they will inevitably encounter on a regular basis. Key discrete mathematical topics that would help a data scientist to model data effectively include:

- Linear algebra (ideas of independence / invertibility, Markov models and eigenvalues)

- Counting principles (understanding of first principles related to randomness)
- Computational (discrete) simulations associated with continuous models
- Graph theory (understand confounding, causal inference and analysis of network data)

Proposed New Course 2 (Mathematical Foundations II: Continuous Mathematics):

A key aspect to modeling in data science is optimization. Part of what makes a model appropriate has to do with its boundaries, maximal values, and sensitivity to parameter choices -- all features that use mathematical optimization. In statistics, one foundational method is to find parameter estimates by maximizing the relevant likelihood. Alternatively, in other mathematical models, the goal might be nonlinear state-space system identification. In both cases, a solid foundation of calculus, differential equations, and numerical methods techniques will allow the data scientist to solve the problem at hand. However, we argue that understanding how to find simple minima and maxima (with ideas of local and global) acts as a vehicle for understanding what optimization means at a fundamentally intuitive level. We recognize that the ideas below are typically taught across many semesters. We are suggesting that much of the content will be removed or taught differently so as to emphasize the critical mathematical components necessary for data science. (For a model of such a course, see MATH 135, Applied Calculus, taught at Macalester College to a large fraction of the undergraduate population.)

To this end, the continuous mathematics course we suggest focuses on understanding the continuous mathematical ideas necessary for problem solving. Some key topics to be incorporated into such a course might include:

- Functions and basic mathematical logic
- Enough calculus to understand the ideas of partial derivatives (interactions in a model)
- Taylor expansion method of approximating functions
- Probability as area / integration
- Multivariate thinking (functions, optimization, integration)

The importance of computing

To be relevant to the broader data science curriculum, the proposed mathematics courses need to be heavily infused with computing. As the MAA CUPM guidelines recommend, mathematics students should not only learn to use technological tools (Cognitive Recommendation 3) but the mathematics programs should include methods which promote data analysis, computing, simulation, and mathematical modeling (Content Recommendation 3). We believe that their recommendations are even more important for future data scientists.

One aspect of integrating computing into the mathematics curriculum is a plea for mathematicians to connect more with computer scientists. If the computer scientists believe that mathematicians care only about theory (without understanding the challenges in the real world), it will be difficult to have a two-way exchange of information across the fields. Indeed, we believe that the computing world would do well to embrace theoretical constructs; but this will only come when the mathematical world is willing to embrace computation.

Integrating computing into the mathematics curriculum not only serves students by giving them computational skills, but additionally, technology in the mathematics classroom allows students to understand the *mathematical theory* more completely. As the CUPM guidelines state:

> In courses at all levels, substantial and realistic applications involve "messy" mathematics that makes calculation by hand onerous or infeasible. Using technology opens the door for students to set up solution strategies, justify their analyses, and interpret the results.

Using computational skills to simulate produces a deeper understanding of the model and complements analytic solutions. Additional computing will help develop better problem-solvers (and may yield additional mathematics majors drawn to the power and beauty of what they see in these courses).

While this article focuses on mathematical preparation, we believe that statistical preparation is also critically important. In recent years, the statistics community has taken on the challenge to improve their existing curriculum in order to ensure that statistical courses incorporate theoretical concepts, computation, and statistical practice (see for example the revised "Guidelines for Assessment and Instruction in Statistics Education [GAISE] College" report (ASA GAISE working group, 2016) (and the ASA revised "Guidelines for Undergraduate Programs in Statistics" (ASA Curriculum Guidelines working group, 2014). The latter report recommends that introductory and intermediate statistics courses (1) be an integral part of a data science curriculum, (2) incorporate reproducible research using statistical software (e.g., R Studio, Python notebooks, or GitHub), (3) use modern and relevant real data, possibly obtained through data scraping. While more work is needed by the statistics community, the article at hand primarily discusses the data science curriculum with respect to foundational courses in *mathematics*.

Closing thoughts

We see the world of data and modeling changing quickly. As mathematicians (and statisticians) we need to be proactive about what our disciplines have to offer. Mathematics will be better off if it is part of the solution. Data Science will be on a better foundational footing if it starts with mathematical first principles: abstraction and modeling. From students for many years, we understand at a visceral level how difficult it is for undergrads to grasp the benefits of generality and abstraction. Ensuring that they see the mathematical conceptual framework early and often will help make for better data scientists. In addition, abstraction is a key component of computer science and important linkages can be made.

We argue that mathematics needs to meet the growing data science community halfway so that the analysis and models leverage vital foundational mathematical concepts. If not, we run the risk that math will be left out. We have proposed one pathway to provide mathematical sophistication for beginning data scientists.

Our deliberately provocative suggestions, which build on the PCMI guidelines and the supplementary material therein, will not necessarily be easy to implement for many mathematics departments, given multiple competing interests and limited resources. However, we implore the community of mathematicians to take our suggestions seriously and engage in curricular discussions at their institutions so as to provide a strong theoretical framework to the world of data science and ensure that mathematics is not left behind. We look forward to working with our colleagues to develop multiple alternative approaches along the lines of those outlined by the Park City group in 2016.

References

ASA Curriculum Guidelines working group, Beth Chance, Steve Cohen, Scott Grimshaw, Johanna Hardin, Tim Hesterberg, Roger Hoerl, Nicholas Horton (chair), Chris Malone, Rebecca Nichols, and Deborah Nolan (2014), http://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

ASA GAISE working group, Rob Carver, M. Everson (co-chair), John Gabrosek, Ginger H. Rowell, Nicholas J. Horton, Robin Lock, M. Mocko (co-chair), Allan Rossman, Paul Velleman, Jeffrey Witmer, and Beverly Wood (2016), "Guidelines for assessment and instruction in statistics education revised college report". http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

David Donoho (2017), "50 years of data science", *International Statistical Review*, http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Richard D. De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, Lei Z. Cheng, Amanda Francis, Robert Gould, Albert Y. Kim, Matt Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruviluamala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, and Ping Ye (2017), "Curriculum guidelines for undergraduate programs in data science", *Annual Review of Statistics and its Application*, DOI: 10.1146/annurev-statistics-060116-053930