DATA SCIENCE FOR UNDERGRADUATES
Opportunities and Options

Nicholas Horton (Amherst College)
TPSEMath Partners Meeting
April 1, 2019

Study funded by the National Science Foundation
Overview

As our economy, society, and daily life become increasingly dependent on data, new college graduates entering the workforce need to have the skills to analyze data effectively.

This study explores what data science skills are essential for undergraduates now and in the future, and how academic institutions can structure their data science education programs to best meet those needs.

- Interim report released in Sept 2017
- Webinars and public input in Fall 2017
- Final report released on May 2, 2018
The Committee

LAURA HAAS, NAE, University of Massachusetts Amherst, Co-Chair
ALFRED O. HERO III, University of Michigan, Co-Chair
ANI ADHIKARI, University of California, Berkeley
DAVID CULLER, NAE, University of California, Berkeley
DAVID DONOHO, NAS, Stanford University
E. THOMAS EWING, Virginia Tech
LOUIS J. GROSS, University of Tennessee, Knoxville
NICHOLAS J. HORTON, Amherst College
JULIA LANE, New York University
ANDREW MCCALLUM, University of Massachusetts Amherst
RICHARD MCCULLOUGH, Harvard University
REBECCA NUGENT, Carnegie Mellon University
LEE RAINIE, Pew Research Center
ROB RUTENBAR, University of Pittsburgh
KRISTIN TOLLE, Microsoft Research
TALITHIA WILLIAMS, Harvey Mudd College
ANDREW ZIEFFLER, University of Minnesota, Minneapolis
National Academies' Groups and Staff

Groups:
- Computer Science and Telecommunications Board (CSTB)
- Board on Mathematical Sciences and Analytics (BMSA)
- Committee on Applied and Theoretical Statistics (CATS)
- Board on Science Education (BOSE)

Staff:
- MICHELLE K. SCHWALBE, Director, BMSA, *Study Director*
- JON EISENBERG, Director, CSTB
- BEN WENDER, Director, CATS
- AMY STEPHENS, BOSE, Program Officer
- LINDA CASOLA, BMSA, Associate Program Officer and Editor
- RENEE HAWKINS, CSTB, Financial Manager
- JANKI PATEL, CSTB, Senior Program Assistant
SUMMARY

1 INTRODUCTION
   A Look to the Future
   Report Overview
   References

2 KNOWLEDGE FOR DATA SCIENTISTS
   Data Science Personas of Today and Tomorrow
   Data Acumen
   A Code of Ethics for Data Science
   References

3 DATA SCIENCE EDUCATION
   Undergraduate Modalities
   Middle and High School Education
   References

4 STARTING A DATA SCIENCE PROGRAM
   Ensuring Broad Participation
   Academic Infrastructure
   Curriculum
   Faculty Resources
   Assessment
   References

5 EVOLUTION AND EVALUATION
   Evolution
   Evaluation
   Roles for Professional Societies
   References

6 CONCLUSIONS

APPENDIXES
   A Biographies of the Committee
   B Meetings and Presentations
   C Contributing Individuals
   D Data Science Oath
“Envisioning” the future is hard; as motivation we take a utopian point of view

- Every student knows some basics of data science (has some modicum of data acumen)
- Data science jobs of varying types are ubiquitous: all industries, all geographies
- Data scientists are diverse demographically (all groups are fairly represented) and educationally (from all domains and levels)

The report includes
- descriptions of needed competencies
- many examples of programs
- useful citations

The report recognizes that educational landscape will evolve
We are in the infancy of data science
There are many different data science roles
Data science is a unique field that borrows heavily from multiple other fields
   - A major/minor/certificate/etc. should not be the same as, e.g., a degree in statistics or in computer science
   - There will need to be educational opportunities to expose faculty to the breadth of the field
   - There will need to be ways to share educational resources
Coordination among professional societies could support the evolution of the undergraduate data science enterprise.
There must be **multiple pathways** for undergraduates as a result.

The undergraduate experience should cater to and **promote diversity** – demographic and intellectual – in the students it serves.

There are some core competencies that all data science students (and, ideally, all undergraduates) should have:

- They should develop **data acumen**
- Ethical problem-solving is a key component of data acumen

Evaluation of programs is critical.
The Findings and Recommendations
Finding 2.1 Data scientists today draw largely from extensions of the “analyst” of years past trained in traditional disciplines. As data science becomes an integral part of many industries and enriches research and development, there will be an increased demand for more holistic and more nuanced data science roles.

Recommendation 2.1 Academic institutions should embrace data science as a vital new field that requires specifically tailored instruction delivered through majors and minors in data science as well as the development of a cadre of faculty equipped to teach in this new field.

Recommendation 2.2 Academic institutions should provide and evolve a range of educational pathways to prepare students for an array of data science roles in the workplace.
Finding 2.3  A critical task in the education of future data scientists is to instill data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts involved in developing data acumen include the following:

- 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
- Ethical problem solving.
Key mathematical concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Set theory and basic logic,
- Multivariate thinking via functions and graphical displays,
- Basic probability theory and randomness,
- Matrices and basic linear algebra,
- Networks and graph theory, and
- Optimization.
While it would be ideal for all data scientists to have extensive coursework in computer science, new pathways may be needed to establish appropriate depth in algorithmic thinking and abstraction in a streamlined manner. This might include the following:

- Basic abstractions,
- Algorithmic thinking,
- Programming concepts,
- Data structures, and
- Simulations.
Important **statistical foundations** might include the following:

- Variability, uncertainty, sampling error, and inference;
- Multivariate thinking;
- Nonsampling error, design, experiments (e.g., A/B testing), biases, confounding, and causal inference;
- Exploratory data analysis;
- Statistical modeling and model assessment; and
- Simulations and experiments.
Key **data management and curation** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Data provenance;
- Data preparation, especially data cleansing and data transformation;
- Data management (of a variety of data types);
- Record retention policies;
- Data subject privacy;
- Missing and conflicting data; and
- Modern databases.
Key **data description and visualization** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Data consistency checking,
- Exploratory data analysis,
- Grammar of graphics,
- Attractive and sound static visualizations,
- Dynamic visualizations and dashboards.
Key data modeling and assessment concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Machine learning,
- Multivariate modeling and supervised learning,
- Dimension reduction techniques and unsupervised learning,
- Deep learning,
- Model assessment and sensitivity analysis, and
- Model interpretation (particularly for black box models).
Workflow and reproducibility concepts

Key workflow and reproducibility concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Workflows and workflow systems,
- Reproducible analysis,
- Documentation and code standards,
- Source code (version) control systems, and
- Collaboration.
Key communication and teamwork concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Ability to understand client needs,
- Clear and comprehensive reporting,
- Conflict resolution skills,
- Well-structured technical writing without jargon, and
- Effective presentation skills.
Key aspects of **ethics** needed for all data scientists (and for that matter, all educated citizens) include the following:

- Ethical precepts for data science and codes of conduct,
- Privacy and confidentiality,
- Responsible conduct of research,
- Ability to identify “junk” science, and
- Ability to detect algorithmic bias.
Recommendation 2.3 To prepare their graduates for this new data-driven era, academic institutions should encourage the development of a basic understanding of data science in all undergraduates.

Joy of Data (Data0)?

Berkeley Data 8 model (major focus on computation)
Finding 3.1  Undergraduate education in data science can be experienced in many forms. These include the following:

- Integrated introductory courses that can satisfy a general education requirement
- A major in data science, including advanced skills as primary field of study
- A minor or track in data science, where intermediate skills are connected to major field of study
- Two-year degrees and certificates
- Other certificates, often requiring fewer courses than a major but more than a minor
- Massive open online courses, which can engage large numbers of students at a variety of levels
- Summer programs and boot camps, which can serve to supplement academic or on-the-job training.
Finding 4.1  The nature of data science is such that it offers multiple pathways for students of different backgrounds to engage at levels ranging from basic to expert.

Finding 4.2  Data science would particularly benefit from broad participation by underrepresented minorities because of the many applications to problems of interest to diverse populations.

Recommendation 4.1  As data science programs develop, they should focus on attracting students with varied backgrounds and degrees of preparation and preparing them for success in a variety of careers.
Chapter 5: Evolution

Recommendation 5.2  During the development of data science programs, institutions should provide support so that the faculty can become more cognizant of the varied aspects of data science through discussion, co-teaching, sharing of materials, short courses, and other forms of training.

Finding 5.3  The data science community would benefit from the creation of websites and journals that document and make available best practices, curricula, education research findings, and other materials related to undergraduate data science education.
Next Steps and Related Work
Focus on taking the results from the NASEM study and fleshing out what this means for two year and four year institutions

NSF funded workshop in 2018 with representatives from key organizations and institutions

Next steps for the mathematical sciences (my opinions now)

- Ensure that mathematics is part of the data science discussion: key role of abstraction in mathematics, computer science, and statistics
- Double down on modeling in K-12
- Leverage co-requisite approaches to flatten curriculum
- Integrate computation early (let students factor $x^7 + 4x + 3 = 0$ and not just quadratic equations)
- Ensure that mathematical foundation courses for data science exist (e.g., Hardin and Horton, 2017, Notices of the AMS, “Ensuring Mathematics is Relevant in a World of Data Science”)
Next steps for the mathematical sciences (my opinions now)

- Interweave computation in statistics and mathematics courses early and often (key role of multivariate thinking)
- Embrace simplified interfaces and approaches to minimize cognitive load
- Embrace cloud computing to minimize barriers to technology
- Integrate and adopt high impact practices and active learning techniques (e.g., pair programming, group based learning)
- Leverage the considerable statistical and modeling foundations that many students have from Common Core