DATA SCIENCE FOR UNDERGRADUATES
Opportunities and Options

Study funded by the National Science Foundation

nas.edu/EnvisioningDS
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 to be 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
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
1. Introduction

2. Acquiring data science skills and knowledge
   • Foundational skills
   • Translational skills
   • Ethical skills
   • Professional skills

3. Data science education in the future
   • Innovative curriculum development
   • Suggestions for institutions

4. Broad participation in data science
   • Recruitment and retention strategies
   • Institutional partnerships
   • K-12 objectives
   • Public outreach
   • Evaluation and assessment

5. Reflections
   • Hippocratic Oath
   • Summary of findings and open questions

<table>
<thead>
<tr>
<th>Topic</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building Data Acumen</td>
<td>Nicole Lazar, University of Georgia</td>
</tr>
<tr>
<td></td>
<td>Mladen Vouk, North Carolina State University</td>
</tr>
<tr>
<td>Incorporating Real-World Applications</td>
<td>Cláudio T. Silva, New York University</td>
</tr>
<tr>
<td></td>
<td>Sears Merritt, Mass Mutual Financial Group</td>
</tr>
<tr>
<td>Faculty Training and Curriculum</td>
<td>Michael Posner: Villanova University</td>
</tr>
<tr>
<td></td>
<td>Robert Panoff, Shodor</td>
</tr>
<tr>
<td>Communication Skills and Teamwork</td>
<td>Madeleine Claire Elish, Data &amp; Society</td>
</tr>
<tr>
<td></td>
<td>Adam Hughes, Pew Research</td>
</tr>
<tr>
<td>Inter-Departmental Collaboration and</td>
<td>Mark Embree, Virginia Tech</td>
</tr>
<tr>
<td>Institutional Organization</td>
<td>Michael Franklin, University of Chicago</td>
</tr>
<tr>
<td>Ethics</td>
<td>Sorin Matei, Purdue University</td>
</tr>
<tr>
<td></td>
<td>Brittany Fiore-Gartland, University of Washington</td>
</tr>
</tbody>
</table>
Envisioning the DATA SCIENCE DISCIPLINE
The Undergraduate Perspective
Webinar Series (Sept.-Nov. 2017)

Assessment and Evaluation for Data Science Programs
Pamela Bishop, University of Tennessee, Knoxville
Kari Jordan, Data Carpentry

Diversity, Inclusion, and Increasing Participation
Talithia Williams, Harvey Mudd College
Allison Master, University of Washington

Two-Year Colleges and Institutional Partnerships
Brian Kotz, Montgomery College
Suzanne Smith, Johnson County Community College

View webinar recordings and slides at nas.edu/EnvisioningDS
Final Report Contents

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, at all degree levels)

The report will be highly referenced

- For the descriptions of needed competencies
- For the many examples of programs
- For the reference citations

The report recognizes in its recommendations that the nature of the field and the educational landscape will continue to evolve rapidly
Key Learnings: Data Science

- We are in the infancy of data science
- There are and in the future will continue to be 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 (e.g., course materials, etc.)
- Coordination among professional societies could usefully support the evolution of the undergraduate data science experience (as well as the evolution of the field)
Key Learnings: Undergraduate Data Science

- Education at all levels will need to evolve as the field evolves
- 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
  - To ensure they evolve as data science evolves
  - To ensure they meet the needs of the various roles students will take in the workplace.
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.

Finding 2.2 Data science programs that strive to meet the needs of their students will likely evolve to emphasize certain skills and capabilities. This will result in programs that prepare different types of data scientists.

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.
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.

Recommendation 2.4  Ethics is a topic that, given the nature of data science, students should learn and practice throughout their education. Academic institutions should ensure that ethics is woven into the data science curriculum from the beginning and throughout.

Recommendation 2.5  The data science community should adopt a code of ethics; such a code should be affirmed by members of professional societies, included in professional development programs and curricula, and conveyed through educational programs. The code should be reevaluated often in light of new developments.
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.

Recommendation 3.1  Four-year and two-year institutions should establish a forum for dialogue across institutions on all aspects of data science education, training, and workforce development.
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.

Finding 4.3  **Institutional flexibility will involve the development of curricula that take advantage of current course availability** and will potentially be constrained by the availability of teaching expertise. Whatever organizational or infrastructure model is adopted, incentives are needed to encourage faculty participation and to overcome barriers.

Finding 4.4  **The economics of developing programs has recently changed with the shift to cloud-based approaches and platforms.**
Finding 5.1  The evolution of data science programs at a particular institution will depend on the particular institution’s pedagogical style and the students’ backgrounds and goals, as well as the requirements of the job market and graduate schools.

Recommendation 5.1  Because these are early days for undergraduate data science education, academic institutions should be prepared to evolve programs over time. They should create and maintain the flexibility and incentives to facilitate the sharing of courses, materials, and faculty among departments and programs.

Finding 5.2  There is a need for broadening the perspective of faculty who are trained in particular areas of data science to be knowledgeable of the breadth of approaches to data science so that they can more effectively educate students at all levels.
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.
Finding 5.4  The evolution of undergraduate education in data science can be driven by data science. Exploiting administrative records, in conjunction with other data sources such as economic information and survey data, can enable effective transformation of programs to better serve their students.

Finding 5.5  Data science methods applied both to individual programs and comparatively across programs can be used for both evaluation and evolution of data science program components. It is essential that both processes are sustained as new pathways emerge at institutions.

Recommendation 5.3  Academic institutions should ensure that programs are continuously evaluated and should work together to develop professional approaches to evaluation. This should include developing and sharing measurement and evaluation frameworks, data sets, and a culture of evolution guided by high-quality evaluation. Efforts should be made to establish relationships with sector-specific professional societies to help align education evaluation with market impacts.
Finding 5.6  As professional societies adapt to data science, improved coordination could offer new opportunities for additional collaboration and cross-pollination. A group or conference with bridging capabilities would be helpful. Professional societies may find it useful to collaborate to offer such training and networking opportunities to their joint communities.

Recommendation 5.4  Existing professional societies should coordinate to enable regular convening sessions among their members. Peer review and discussion are essential to share ideas, best practices, and data.
DATA SCIENCE FOR UNDERGRADUATES
Opportunities and Options

Report Release Webinar
May 2, 2018
11am-12pm EDT

Study funded by the National Science Foundation

nas.edu/EnvisioningDS
Questions?

Thanks to the National Science Foundation (CISE, MPS/DMS, EHR, and SBE) for support of this study.