Envisioning the Data Science Discipline: The Undergraduate Perspective

Rebecca Nugent
Carnegie Mellon Statistics & Data Science

on behalf of the National Academies of Sciences Committee

May 3, 2019
NAS Committee

- Laura Haas, NAE (Univ of Massachusetts, Amherst) Co-Chair
- Alfred O. Hero III (Univ of Michigan) Co-Chair
- Ani Adhikari, NAE (Univ of California, Berkeley)
- David Culler, NAS (Univ of California, Berkeley)
- David Donoho, NAS (Stanford University)
- E. Thomas Ewing (Virginia Tech)
- Louis J. Gross (Univ of Tennessee, Knoxville)
- Nicholas J. Horton (Amherst College)
- Julia Lane (New York University)
- Andrew McCallum (Univ 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 (Univ of Minnesota, Minneapolis)
Almost all material presented here is culled from report and overview slides created by NAS Study Committee members and staff.
So what is it?

Data Science

Copyright © 2014 by Steven Geringer Raleigh, NC. Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact.
So what is it?

Figure 1
So what is it?
So what is it?

Relation between techniques of Big Data Analytics
NAS Overview

http://www.nas.edu/EnvisioningDS; Full/interim report; Webinar materials

▶ Building Data Acumen
Nicole Lazar (Univ of Georgia), Mladen Vouk (North Carolina State Univ)

▶ Incorporating Real-World Applications
Claudio T. Silva (NYU), Sears Merritt (Mass Mutual Financial Group)

▶ Faculty Training and Curriculum Development
Michael Posner (Villanova University), Robert Panoff (Shodor)

▶ Teamwork
Madeline Claire Elish (Data & Society), Adam Hughes (Pew Research)

▶ Interdepartmental Collaboration and Institutional Organization
Mark Embree (Virginia Tech), Michael Franklin (University of Chicago)

▶ Ethics
Sorin Matei (Purdue Univ), Brittany Fiore-Gartland (Univ of Washington)

▶ Assessment and Evaluation for Data Science Programs
Pamela Bishop (Univ of Tenn, Knoxville), Kari Jordan (Data Carpentry)

▶ Diversity, Inclusion, and Increasing Participation
Talithia Williams (Harvey Mudd), Allison Master (Univ of Washington)

▶ Two-Year Colleges and Institutional Partnerships
Brian Kotz (Montgomery Coll), Suzanne Smith (Johnson County Comm College)
Final Report Contents

▶ Introduction: A Look to the Future, Report Overview, References
▶ Knowledge for Data Scientists
  ▶ Data Science Personas of Today and Tomorrow
  ▶ Data Acumen
  ▶ A Code of Ethics for Data Science
▶ Data Science Education
  ▶ Undergraduate Modalities
  ▶ Middle and High School Education
▶ Starting a Data Science Program
  ▶ Ensuring Broad Participation
  ▶ Academic Infrastructure
  ▶ Curriculum
  ▶ Faculty Resources
  ▶ Assessment
▶ Evolution and Evaluation (incl. Roles for Prof. Societies)
Key Insights: Data Science

- We are in infancy of data science
- There are and will continue to be many different data science roles
- Data Science is a unique field that borrows heavily from other fields
  - DS Major/minor/certificate not same as a degree in statistics or computer science
  - Will need to be educational opportunities to expose faculty to breadth of field
  - Need ways to share educational resources (course materials, etc)
- Coordination across prof societies could support evolution of the undergraduate data science experience
Key Insights: Undergraduate Data Science

- Education at all levels will need to evolve as the field evolves
- Need multiple pathways
- Undergraduate experience should cater to and promote diversity (demographic and intellectual)
- Core competencies include data acumen and ethical problem-solving
- Evaluation of programs is critical
  - Ensure programs evolve as data science evolves
  - Ensure meet needs of changing workplace roles
Chapter 2: Knowledge for Data Scientists

- **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...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.
Chapter 2: Knowledge for Data Scientists

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 Concepts

Mathematical:
- Set theory and basic logic
- Multivariate thinking via functions and graphical displays
- Basic probability thinking and randomness
- Matrices and basic linear algebra
- Networks and graph theory
- Optimization

Computational:
- Basic abstractions
- Algorithmic thinking
- Programming concepts
- Data structures
- Simulations
Key Concepts

**Statistical:**
- Variability, uncertainty, sampling error, inference
- Multivariate thinking
- Nonsampling error, design, experiments, biases, confounding, causality
- Exploratory Data Analysis
- Statistical modeling and model assessment
- Simulations and experiments

**Data Management and Curation:**
- Data provenance
- Data preparation (incl. cleaning and transformation)
- Record retention policies
- Data subject privacy
- Missing and conflicting data
- Modern databases
Key Concepts

*Data Visualization:*
- Data consistency checking
- Exploratory Data Analysis
- Grammar of Graphics
- Attractive and sound static visualizations
- Dynamic visualizations and dashboards

*Data Modeling and Assessment:*
- Multivariate modeling and supervised learning
- Machine learning
- Dimension reduction techniques, unsupervised learning
- Deep learning
- Model assessment and sensitivity analysis
- Model interpretation

rnugent@stat.cmu.edu
Key Concepts

Workflow and reproducibility:
- Workflows and workflow systems
- Reproducible analysis
- Documentation and code standards
- Source code (version) control systems and collaboration

Communication and Teamwork:
- Ability to understand client needs
- Clear and comprehensive reporting
- Conflict resolution skills
- Well-structured technical writing without jargon
- Effective presentation skills
Key Concepts

**Ethics:**

- Ethical precepts for data sciences and codes of conduct
- Privacy and confidentiality
- Responsible conduct of research
- Ability to identify “junk” science
- Ability to detect algorithmic bias

Ethics should be woven throughout curriculum and also taught on its own
**BOX D.1**  
**Hippocratic Oath**

I swear to fulfill, to the best of my ability and judgment, this covenant:
I will respect the hard-won scientific gains of those physicians in whose steps I walk, and gladly share such knowledge as is mine with those who are to follow.
I will apply, for the benefit of the sick, all measures which are required, avoiding those twin traps of overtreatment and therapeutic nihilism.
I will remember that there is art to medicine as well as science, and that warmth, sympathy, and understanding may outweigh the surgeon’s knife or the chemist’s drug.
I will not be ashamed to say “I know not,” nor will I fail to call in my colleagues when the skills of another are needed for a patient’s recovery.
I will respect the privacy of my patients, for their problems are not disclosed to me that the world may know. Most especially must I tread with care in matters of life and death. If it is given to me to save a life, all thanks. But it may also be within my power to take a life; this awesome responsibility must be faced with great humility and awareness of my own frailty. Above all, I must not play at God.
I will remember that I do not treat a fever chart, a cancerous growth, but a sick human being, whose illness may affect the person’s family and economic stability. My responsibility includes these related problems, if I am to care adequately for the sick.
I will prevent disease whenever I can, for prevention is preferable to cure.
I will remember that I remain a member of society, with special obligations to all my fellow human beings, those sound of mind and body as well as the infirm.
If I do not violate this oath, may I enjoy life and art, respected while I live and remembered with affection thereafter. May I always act so as to preserve the finest traditions of my calling and may I long experience the joy of healing those who seek my help.

**SOURCE:** J.C. Laccour, 1964, *Hippocratic Oath Modern*

---

**BOX D.2**  
**Data Science Oath**

I swear to fulfill, to the best of my ability and judgment, this covenant:
I will respect the hard-won scientific gains of those data scientists in whose steps I walk and gladly share such knowledge as is mine with those who follow.
I will apply, for the benefit of society, all measures which are required, avoiding misrepresentations of data and analysis results.
I will remember that there is art to data science as well as science, and that consistency, candor, and compassion should outweigh the algorithm’s precision or the interventionist’s influence.
I will not be ashamed to say, “I know not,” nor will I fail to call in my colleagues when the skills of another are needed for solving a problem.
I will respect the privacy of my data subjects, for their data are not disclosed to me that the world may know, so I will tread with care in matters of privacy and security. If it is given to me to do good with my analyses, all thanks. But it may also be within my power to do harm and this responsibility must be faced with humility and awareness of my own limitations.
I will remember that my data are not just numbers without meaning or context, but represent real people and situations, and that my work may lead to unintended societal consequences, such as inequality, poverty, and disparities due to algorithmic bias. My responsibility must consider potential consequences of my extraction of meaning from data and ensure my analyses help make better decisions.
I will perform personalization when appropriate, but I will always look for a path to fair treatment and nondiscrimination.
I will remember that I remain a member of society, with special obligations to all my fellow human beings, those who need help and those who don’t.
If I do not violate this oath, may I enjoy vitality and virtuosity, respected for my contributions and remembered for my leadership thereafter. May I always act to preserve the finest traditions of my calling and may I long experience the joy of helping those who can benefit from my work.
Ch 3: Data Science Education

Variety of forms this can take:

- Integrated introductory course/general education requirement
- Major in data science, incl. advanced skills: Data Science (+ X)
- Minor or track in data science, incl. intermediate skills: X + Data Science
- Two year degrees and certificates
- Additional certificates
- Massive open online courses
- Summer programs and boot camps
Ch 4/5: Starting/Evolving a Data Science Program

- Multiple pathways for students of different backgrounds (good thing!)
- Benefit from broad participation by underrepresented minorities
- Instructional flexibility:
  - development of curricula that take advantage of current courses
  - constrained by availability of teaching expertise; need incentives
- Economics changing with shift to cloud-based approaches/platforms

- Depends on institution’s infrastructure and pedagogical style
- Must be prepared to continuously evaluate and evolve over time
- Faculty need to broaden perspective; provide support to do so
- Need to incentivize sharing of materials, courses, and faculty
- Create journals, websites, dissemination mechanisms to make available best practices
Other Data Science Education Reports:

- NAS Roundtable on Data Science Postsecondary Education (online)
- Park City Math Institute (2016): Curriculum Guidelines for Undergraduate Programs in Data Science Working Group led by Dick DeVeaux; white paper endorsed by ASA
- Guidelines for Assessment and Instruction in Statistics Education (2016)
- ACM Task Force on Data Science white paper/report (hearing about today)
- NSF Data Science Education Workshop: June 2019, Berkeley
- Data Science Leadership Summit (Moore/Sloan): November 2019, Santa Fe
- Several NSFs interdisciplinary calls under the Harnessing the Data Revolution (HDR) theme
RN Personal Takeaways

- Turns out this is pretty hard
- First wave of programs was largely driven by financial concerns
- Slow and steady wins the quality race
- Great opportunity to build data science pathways for everyone, particularly for non-STEM students and younger students (down to K)
- Industry knows that they want data scientists but are pretty inconsistent about the jobs, titles, descriptions, etc.
- Data Science should not divide people into “haves/have nots”. Anyone can contribute along the Data Science Pipeline. Everyone is a Data Scientist.