Evolution of the Undergraduate Statistics (& Data Science) Program

Rebecca Nugent

Carnegie Mellon University
Statistics & Data Science

July 29, 2018

The National Landscape

As we're all aware,

- Interest in Statistics is sky-rocketing at all levels (K-12, undergrad, grad, professional), both academia and industry
- Data Science has taken over the popular press and job postings
- Pressure to define and capitalize on data science; build related programs and build them quickly
- Data Science is an organic, growing community fully exploiting the democratization afforded through technology, internet, etc. The People's Science.

Immense attention has been paid to updating material for individual courses, primarily the first sequence (for good reason).

Here we'll take a look at the evolution of the undergraduate program as a whole and some of our most pressing challenges.

GAISE, 2016

Current recommendations from the Guidelines for Assessment and Instruction in Statistics Education (GAISE, 2016):

- Teach statistical thinking
  - Teach statistics as an investigative process of problem-solving and decision-making
  - Give students experience with multivariable thinking
- Focus on conceptual understanding
- Integrate real data with a context and a purpose
- Foster active learning
- Use technology to explore concepts and analyze data
- Use assessments to improve and evaluate student learning

Primarily used for intro stat, but could be applied to entire program

Program Curriculum Guidelines, 2014

ASA Working Group on Undergraduate Program Guidelines (2014):

- Focus on problem-solving
- Increased importance of data science
- Real applications
- More diverse models and approaches
- Ability to communicate
- Creative approaches to new curricular needs
- Flexibility, multiple pathways
- Curriculum
  - Statistical method and theory
  - Data Management and computation
  - Mathematical Foundations
  - Statistical Practice

NAS Study: Data Science for Undergrads, 2016-18

Final Report Recommendations:

- Encourage development of basic data science understanding in all undergraduates
- Embrace data science as field; build majors, minors; train faculty
- Attract and train students of varied backgrounds, preparation
- Build multiple related pathways
- Include ethics throughout curriculum; adopt a community code
- Develop programs with ability to evolve and be flexible; disseminate materials; support inter-department communication
- Evaluate programs; develop assessment frameworks
- Coordinate regularly convening related meetings through professional societies

NAS Study: Data Science for Undergrads, 2016-18

Data Acumen Topics (in no particular order):

- Mathematical, Computational, 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

What students think it is: looking at cool data sets (possibly through an app) and running algorithms on lots of data to make decisions (and possibly lots of money)

Where does theory belong?

The excitement and popularity of new data visualization and analysis tools pretty much overwhelms student desire to learn theory/math.

This is a problem. But a not well-defined one.

- A theoretical foundation is necessary and non-negotiable.
  So what should it be? Should it depend on what you do next?
- Given the realistic constraints of a degree program, if we could start from scratch, what would be the most important theoretical topics for students to learn? to use in their jobs? grad school?

  to correctly interpret the results of large-scale, black-box deep learning neural net algorithms that most students claim they understand but actually just run by pushing some buttons?
- What does it mean to apply statistical theory to data sets that were generated under (almost) no assumptions?

Programming/Coding

Data Analysis vs Statistical Computing

- Stat (& DS) majors should be coding, but how/what kind?
- Can't assume that knowing R(markdown)/tidyverse corresponds to knowing computing/algorithms (and vice versa)
- Cognitive load in course should shift accordingly
- Should be taught in conjunction with best practices in reproducibility and communication/dissemination
- Outsource; work with other departments; online materials (e.g., Data Science in a Box; Cetinkaya-Rundel, Duke); be efficient; teach Intro to Programming in R?

Ethics

In the age of big data, AI, reproducibility, replicability, data for social good, etc, it's never been more important to talk about ethics.

Too often it's sprinkled throughout courses but not a dedicated stand-alone topic. "Data Ethics" should be a separate course, seminar, workshop that is required for all Stat(DS) majors.

- Models perpetuating bias
- Analytics being skewed for profitability
- Implications of automation
- Academic elitism; isolating non-technical people

- Bloomberg: Data Hive, Data for Democracy, Data for Good
- Myriad of articles on reproducibility
- NAS report - Data Science "oath"

The Relationship between Research and Education

Bring the classroom into research; bring research into the classroom

- Textbooks are a wealth of information; they're all online
- All our homeworks, solutions, exams, etc are online.
  We can fight it, but we need to admit it.

  Does it matter? How can we use textbooks more as references? (recommended vs required)
- To fully engage in active learning, add open research questions; teach them to expect null or surprising results; look for the unexpected; criticize; brainstorm
- Students can provide new insights on faculty research, particularly applied/methods; far more engaged
- Leverage existing data set repos; build research problem repos

Corporate Engagement

Connecting with industry: mutually beneficial two-way street

- Internships - how much do we get involved?
  - Experiential learning
  - To give credit or not? Pre-OPT vs CPT
  - Leverage alumni, career center
- Capstone projects
  - Companies donate research problem, data sets, (some) oversight, money, pizza
  - One example: DePaul Corporate Affiliate Program
  - Leverage university-level Advancement group
- Competitions, DataFests, Info Sessions, etc
- Industry Affiliate/Adjunct Programs
  Smith College Statistics, MassMutual Fellows

Where do we start?

An incredible amount of work. To think about and build even some of these components is a daunting task in the face of non-trivial issues.

- Resources
  - Manpower
  - Finances
  - Access
- Infrastructure Constraints
- Communication

As Deb Nolan once told me, issues are really just challenges.
And challenges are meant to be won. *(Ok, that last part is me.)*

Challenges & Ideas

Constrained by a fixed curriculum and institutional infrastructure

- Difficulties with adopting new courses or new textbooks
- Hierarchical administrative structure; permissions, paperwork
- Who owns your curriculum?
  Who owns Statistics and Data Science on your campus?

- Don't use the textbooks (or change how you use them)
- Try bringing in 1-2 projects or real data sets instead
- Pilot a research group, seminar, working group;
  students will vote with their feet; NSF IUSE: EHR
- Reach out to typical and atypical advocates and collaborators

Challenges & Ideas

Curriculum already very full; no room to add; what do we take out?

- Hard question; we can only do so much over four years. Need to be creative/efficient. Also need to take a hard look at what we're teaching and not be "Stat Hoarders". #letitgo
- Statistics and Data is trickling all the way down to kindergarten. Need to adapt.
- Combine things; work that spans entire data analysis pipeline
- Add non-course components to the curriculum; experiential learning; workshops; hackathons; indep. studies

Challenges & Ideas

Too many students. How do we scale?

- Constrained by class size limits set by university
- Constrained by unrealistic expectations set by leadership
- Constrained by only having 24 hours a day

- The answer is not to turn people away
- Be thoughtful about what we assign and what we grade
- Leverage students; hire undergraduate teaching assistants
- Leverage technology; GradeScope, Canvas, automatic grading/feedback
- Set expectations that undergraduate courses are not the sole responsibility of a few

Challenges & Ideas

How do we recruit/retain a diverse student population?

- Engage with K-12, Two-year colleges
- Collaborate with non-STEM; humanities, social sciences
- Do not isolate non-technical people;
  not everyone wants to code in nine languages
- Incorporate diverse examples, data; let students design projects
- ASA StatFest (Amherst, September 2018)
- AMS Notices: Reflections on PCMI program on Increasing Minority Participation in Mathematics
- Women in Data Science, Women in Statistics & Data Science (ASA)
- Disseminate best practices; open source

Challenges & Ideas

Too many students. Not enough instructors. Not enough support.

- Stems from institutions and leadership not having bandwidth to adapt quickly to realities of today's education and research environment. Want great programs? Need to pay for them.
- Having one type of faculty is not enough. Comically not enough.
- Prioritizing one type of faculty productivity over all others creates tension, dilutes the pool, and handicaps the future.

  Where do new faculty and new researchers come from?
  Oh, that's right. Students.
- To offer instructors short-term (1-2 years, constant review, etc) contracts is terrible, short-sighted, and demeaning.

  If you are worried about your job, you are not innovating.

Challenges & Ideas

Everyone is valuable. Invest in everybody.

- Invest (as your institution allows) deeply in long-term growth/stability. Rome wasn't built in an academic year.
- Push for faculty lines associated with education and/or statistical practice. Give them time for research.
- More integration and collaboration among faculty (all kinds); less segregation; ideas can come from anywhere
- Increase contract lengths of non-tenure track/instructors; provide them professional opportunities and discretionary funds
- Do more to support tenure-track faculty for teaching and directing programs. Need to learn balance early.
- Worth of a good academic advisor far exceeds their salary

Acknowledgments

A whole-hearted thank you and appreciation to all of the Statistics and Data Science educators at all levels who get up every day and do amazing work training the literally millions of people each year who want to learn about our discipline. The impact of your energy, ideas, and experience is immeasureable. Let's change the game.

Rebecca Nugent

Carnegie Mellon Statistics & Data Science

rnugent@stat.cmu.edu; @CMU_Stats

http://www.stat.cmu.edu/∼rnugent

To learn more about CMU's new Interactive Statistics Learning Environment (ISLE) for Intro Stat, Thu 8/2 #649 10:30am, CC-W 212