Data Science Education - Successes and Challenges: Stories from the Classroom and Beyond

Nicholas J. Horton

Amherst College

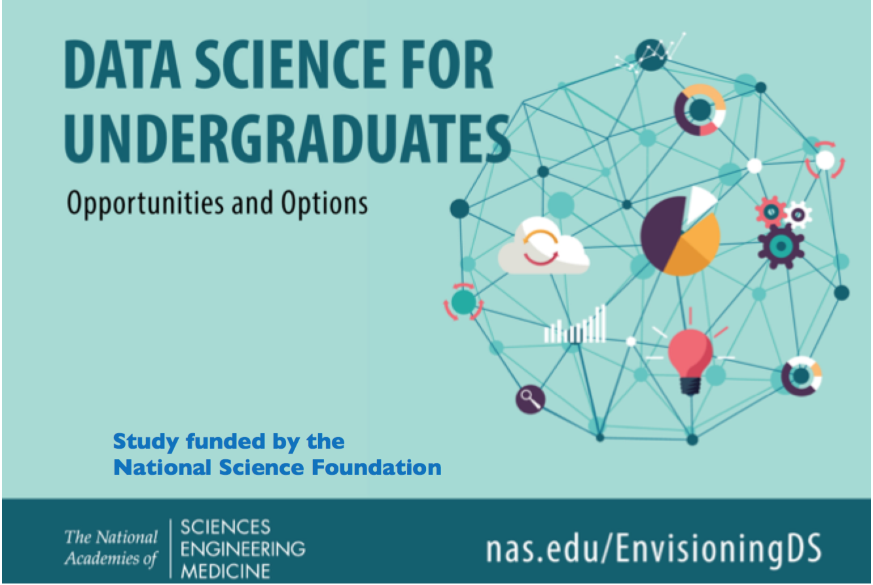
JSM, July 29, 2018

nhorton@amherst.edu
https://github.com/Amherst-Statistics/JSM2018

Plan

- NAS report and ‘data acumen‘
- data science analysis cycle and repeated exposure (Fisher and Warsinske)
- data technologies (Sun)
- undergrad vs. graduate (vs. associate!) (Jones)
- structures of programs (Fisher)
- diverse preparation

NAS consensus report



DATA SCIENCE FOR UNDERGRADUATES

Opportunities and Options

Study funded by the National Science Foundation

The National Academies of SCIENCES ENGINEERING MEDICINE

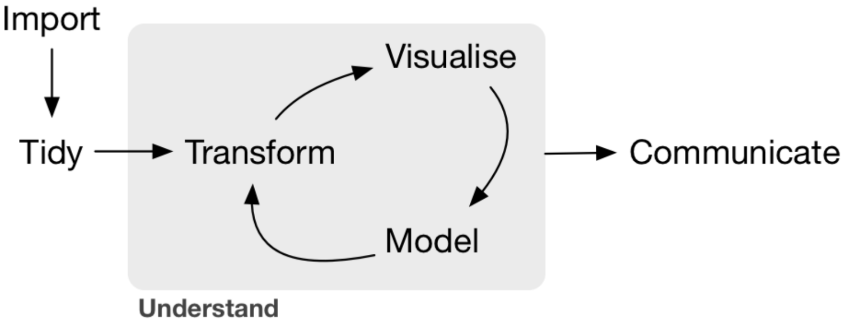
nas.edu/EnvisioningDS

NAS consensus report: Data acumen

1. Mathematical foundations
2. Computational foundations
3. Statistical foundations
4. Data management and curation
5. Data description and visualization
6. Data modeling and assessment
7. Workflow and reproducibility
8. Communication and teamwork
9. Domain-specific considerations
10. Ethical problem solving

`https://nas.edu/envisioningds`

Wickham and Grolemund cycle



repeated exposure

- I Introduce topics early on
- R Reinforce them in later courses and co-curricular experiences
- M move towards Mastery in capstone and integrative coures

ASA Guidelines for Undergraduate Programs in Statistics (2014)

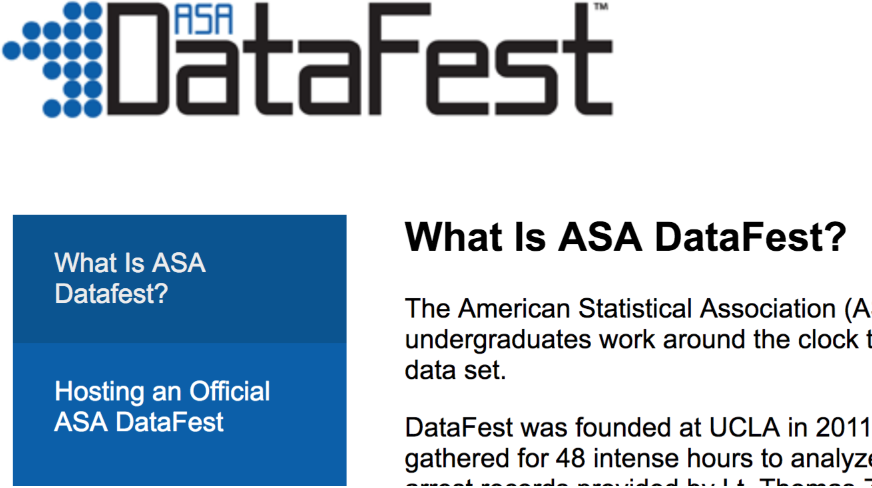
- Graduates should be expected to write clearly, speak fluently, and construct effective visual displays and compelling written summaries
- They should demonstrate ability to collaborate in teams and to organize and manage projects
- They should be able to communicate complex statistical methods in basic terms to managers and other audiences and visualize results in an accessible manner
- There is pedagogical value in having students practice communication to identify gaps in their understanding.
- Communication skills need to dovetail with students' technical and statistical knowledge: excellent communication of inappropriate or incorrect analyses is counterproductive.

Importance of pedagogy

AAC&U High impact practices (`https://www.aacu.org/leap/hips`)

Collaborative Assignments and Projects Collaborative learning combines two key goals: learning to work and solve problems in the company of others, and sharpening ones own understanding by listening seriously to the insights of others, especially those with different backgrounds and life experiences. Approaches range from study groups within a course, to team-based assignments and writing, to cooperative projects and research.

Co-curricular Experiences (cont.)



What Is ASA Datafest?

Hosting an Official ASA DataFest

What Is ASA DataFest?

The American Statistical Association (A
undergraduates work around the clock t
data set.

DataFest was founded at UCLA in 2011
gathered for 48 intense hours to analyze
arrest records provided by Lt. Thomas

Data technologies (NAS report key concepts)

- Data provenance
- Data preparation, especially data cleansing and data transformation
- Data consistency checking
- Data management (of a variety of data types)
- Record retention policies
- Data subject privacy
- Workflow and reproducibility
- Version control systems

Data technologies (how to teach?)

- need framework for entities, attributes, relations, and tables (data models)
- tidy-data (Wickham, JSS, 2014) as one possible organizing approach

undergrad vs. graduate (vs. associate!)

- can't afford to have all students complete masters to do useful work
- NAS report called for a continuum (data science gened, data science + X, data science minors, majors)
- rise of associates programs (Two Year College Data Science Summit)
- importance of flexible pathways

entry level pathways

- Data0 (“joy of data”)
- intro to data science with no prereqs (e.g., Data8.org)
- intro to data science building on some stat and some CS (Sun)

structures of programs

- depth and breadth
- theoretical foundations
- incorporate AAC&U high impact practices
- assessment baked in and revisited

diverse preparation and students

- K-12 disparities
- unstated and hidden expectations
- need for creative solutions
- potential for cloud-computing
- potential for student-led support structures

next steps

- coordination of efforts
- networks of institutions and associations
- faculty development (stats PhD doesn't prepare faculty for leadership)
- how to sustain?
- many opportunities and options

Data Science Education - Successes and Challenges: Stories from the Classroom and Beyond

Nicholas J. Horton

Amherst College

JSM, July 29, 2018

nhorton@amherst.edu
https://github.com/Amherst-Statistics/JSM2018