Growing Data Science Courses and Programs – The Need to Collaborate

AMATYC Annual Conference 2018 - Orlando, FL
November 15, 2018
Michael Posner, Rob Gould, Kate Kozak
slides at http://homepage.villanova.edu/michael.posner
We hope you will learn...

- What is Data Science? Why is it so popular?
- The relationship between Data Science and Statistics / Mathematics
- Challenges and opportunities collaborating across disciplines
- Resources available for building data science courses and programs
Michael A. Posner, PhD, PStat® - Villanova U
Associate Professor of Statistics and Director, Center for Statistics Education
Vice-Chair, Data Science Subcommittee, Statistics Committee, AMATYC
Chair-elect, Section on Statistics Education, ASA
TANGO Stat Ed, Principle Investigator

Rob Gould, PhD - UCLA
Lecturer with Security of Employment, Vice-Chair of UG Studies, and
Director of Center for Teaching Statistics
Former Chair, AMATYC/ASA committee
Chair, ASA DataFest Committee

Kathryn Kozak, MS – Coconino CC (AZ)
Mathematics Instructor
President-elect, AMATYC
StatPREP, Co-Principle Investigator
Panel Format

• Overview of Data Science (~5 minutes)
• Each panelist... (~10 minutes each)
  • Data Science – definition and path
  • Growing Data Science courses and programs
    • Creating and updating a program
    • Transferability and articulation agreements
    • Collaborating across disciplines
    • AMATYC’s Data Science Subcommittee, Statistics Committee
• Question & Answers / Discussion (~15 minutes)
• Resources
To show this poll

1. Install the app from pollev.com/app
2. Start the presentation

Still not working? Get help at pollev.com/app/help or Open poll in your web browser
To show this poll

1. Install the app from pollev.com/app
2. Start the presentation

Still not working? Get help at pollev.com/app/help or Open poll in your web browser
The Venn Diagram

- Hacking Skills
- Math & Statistics Knowledge
- Substantive Expertise

- Machine Learning
- Traditional Research

- Data Science
- Danger Zone!
Yet Another Venn Diagram

Data Science Venn Diagram v2.0

- Data Science
- Computer Science
- Machine Learning
- Math and Statistics
- Traditional Software
- Traditional Research
- Subject Matter Expertise

Unicorn

Copyright © 2014 by Steven Gerinker Raleigh, NC. Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact.
From IBM - http://www.ibmbigdatahub.com/infographic/four-vs-big-data
Your Data are Valuable!
Name That Company...

• What is the largest transportation company in the world?

- How many vehicles do they own?

• What company rents out more rooms in the world than any other?

- How many properties do they own?
Kozak – My Data Science Path

• My experience
  • Math and Statistics Teacher at Coconino Community College, Flagstaff, AZ – Considering creating a DS emphasis (degree)
  • President-Elect of AMATYC
  • Liaison to AMATYC Statistics Committee
  • Participated in South Big Data Summit
  • Participated in Two-Year College Data Science Summit
  • Participated in TPSE Math Data Science Meeting

• AMATYC Data Science Subcommittee of the Statistics Committee
AMATYC Data Science Subcommittee

• Goals
  • Support and maintain a community of individuals interested in increasing the presence of data science in two-year colleges (including program and curriculum development);
  • Develop and share best practices in two-year college data science;
  • Facilitate communication among two-year college faculty from various disciplines interested in data science;
  • Facilitate communication with external organizations involved in data science education efforts;
  • Actively participate in discussions defining the levels of mathematics and statistics needed in two-year college level data science curricula; and
  • Encourage professional development and support for the teaching and learning of data science.

• Officers – Ambika Silva (Chair), Michael Posner (Vice-Chair), Brian Kotz (past-Chair)
Gould - What is Statistics?

• Statistics is the science of learning from data and of measuring, controlling, and communicating uncertainty. (American Statistical Association, https://www.amstat.org/ASA/Newsroom.aspx)

• So what is Data Science?
  • In my view, an expansion of Statistics to include accessing, storing, “curating”, preparing data, and communicating results to a wide variety of audiences.
Gould—My Data Science Path

- UCLA Statistics program started in 1998 to emphasize computational statistics and data analysis.
- UCLA UG Major in 2004 emphasized same.
- Founded ASA DataFest in 2011 to give UG students experience with complex data
- Co-developed Mobilize Introduction to Data Science, the first DS course for HS students, in partnership with LAUSD
- Co-chaired ASA 2YDSS in May (report coming soon!)
- DS at UCLA within Stats major: intro and advanced programming in R, optimization theory, resampling methods, predictive modeling/statistical learning, EDA and data cleaning
- Coming soon: Data Theory major. Joint with Math Dept.
Gould - Preparing Students for DS at a 4YC

- If you don’t have the means to create a Data Science curriculum for students who wish to transfer, then a good first step is to offer a statistics course that uses R or Python and requires students to
  - “Wrangle” with data
  - Solve problems involving real data in realistic contexts
  - Engage in the entire four steps of the “statistical investigative process” aka the “Data Cycle” (Ask Questions, Consider Data, Analyze Data, Interpret) as many times as possible.
2YDSS Recommendations for DS Programs

1. Create courses that provide students with a strong introduction to statistics that includes exploratory data analysis, the use of simulations, randomization-based inference, and an introduction to causal inference.

2. Ensure students engage with realistic problems and real data.

3. Explore ways of removing mathematics as a barrier to studying data science.

4. Design courses that require both algorithmic and statistical thinking.

5. Expose students to technology tools for reproducibility, collaboration, data curation; develop fluency in at least one programming language used in data science.

6. Infuse ethical issues throughout the curriculum.

7. Foster active learning using real data. Consider the use of portfolios for assessment tools.
Posner – Data Science Path - Skills

• Trained as a Statistician
• Increasingly encountered data I couldn’t handle
  • Volume
    • ...but computing speed also increased
  • Variety – text mining, JSON?, Videos?!?
• Additional training needed!!!
  • Coursera Data Science courses (over sabbatical)
  • Mastering the Tidyverse (JSM)
  • DataFest and new projects (Castro Speeches, DS Course)
Posner – Data Science Path - Villanova

• Villanova has...
  • Master’s in Applied Statistics, but I focus today on first two years
  • Minor in Statistics, working on Major in Statistics with DS track
  • Computer Science department - also interested in DS
  • Business school analytics programs – Minor BA, MSA, CBA

• Data Science course/program development
  • Intro stat course – need to transform
    • Simulation-based inference
    • Data investigative process - excitement of discovery with data, not pessimism (Steve Feinberg)
  • Worked together with CS faculty, co-taught a course with CS faculty
  • My own course(s) – demand > supply
Posner – Data Science Path - Collaboration

• Opportunities (in working with CS)
  • Big ideas in statistics – uncertainty, bias (including coding choices), multivariable thinking
  • Big ideas in computer science – data lifecycle, different language

• Challenges (in working with CS)
  • Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write!
  • Software choice (and the definition of bias)

• Two year colleges
  • Professional development
  • Clear syllabus specification at 4YC
  • Identify and converse with potential local employers
Posner – Data Science Path - Future

- Data Science Subcommittee
- ASA’s Section on Statistics Education
  - Name change?
  - Free professional development
Posner – Educating Future Data Scientists

• Statistician ("Type A Data Scientist" – Analyze). Need to know...
  • Reading and wrangling data in various formats
  • Text mining
  • Statistical/Machine learning – black box methodologies
  • Algorithmic thinking, including data structures, computer engineering, etc.
  • Data Science Ethics

• Computer Scientist ("Type B Data Scientist" – Build). Need to know...
  • Uncertainty
  • Descriptive measures (beyond algorithm)
  • Bias, especially from processing the data (i.e. aggregation, missing data)
  • Multivariable thinking and causal inference (big data doesn’t solve all problems)
  • Design of Experiments
  • Data Science Ethics
Resources

- Data Science subcommittee website
  - Meeting on Sat @ 11:55am in Acapulco room
  - Also...join the subcommittee!!!
- Data Science for Undergraduates – Opportunities and Options - NAS
- Park City Math Institute Curriculum Guidelines for Data Science
- ASA 2Y Data Science Summit
- Lifelong Learning - ASA Section Courses, DataCamp, etc.
- People’s resource pages – Shonda Kuiper, ModernDive (Ismay and Kim), R4DS (Grolemund, Wickham), etc.
And Now for the Important Part...

Questions / Comments / Discussion