BUILDING BRIDGES FOR DATA SCIENCE EDUCATION: WHAT I LEARNED FROM WORKING WITH COMPUTER SCIENCE FACULTY

July 30, 2019
JSM 2019 – Denver

Michael A. Posner, PhD, PStat®
Villanova University
OVERVIEW

• Stat vs. CS: Why Stat is losing the DS battle!
• DS at Villanova
• What I Learned from Collaborating with CS colleagues
• What I Want my CS Colleagues to Know
• DS related Resources
OVERVIEW

• Stat vs. CS: Why Stat is losing the DS battle!
• DS at Villanova
• What I Learned from Collaborating with CS colleagues
• What I Want my CS Colleagues to Know
• DS related Resources
WHAT IS DATA SCIENCE?
STAT IS LOSING THE DS BATTLE!

- What’s my return on investment?
  - **Computer Scientist** - Fast, Simple
  - **Statistician** - Slower, Asks Qs, Conditional, Range
- Hitting our limiting factors...
  - **Statistician** – Data we can’t handle
  - **Computer Scientist**
    - Model we don’t know...Machine Learning!
    - Bias we don’t think about...who cares?
- Connection with industry
  - **Computer Scientist** - SIG C SE
  - **Statistician** - USCOTS
OVERVIEW

- Stat vs. CS: Why Stat is losing the DS battle!
- DS at Villanova
- What I Learned from Collaborating with CS Colleagues
- What I Want my CS Colleagues to Know
- DS related Resources
DS AT VILLANOVA – NSF GRANT

• NSF IUSE Grant (DUE 1432438), 2014 – 2018
  • “Data Computing for All: Developing an Introductory Data Science Course”
• Faculty: Villanova (1 Stat, 2 CS) and WSSU (2 CS)
• Intro DS Course - VU once (Stat + CS), WSSU twice
  • No prereqs – Freshman only
  • Team teaching had its challenges
  • Two languages – Python and JMP was challenging
• Presentations at 4 SIGCSE conferences
DS AT VILLANOVA - COURSES/CENTERS

- **Stat-based**
  - DS course - Junior/Seniors, Stat Prereq, Prog exp preferred
    - 1-2 sections/semester, based on Mastering the Tidyverse
    - Fills up fast, great comments from students
  - Grad DS course (updated Stat Programming)
  - New Stat Major – Two programming/DS courses required

- **CS-based**
  - Run twice over past three years, CS1 prereq

- **Collaborative CS/Stat - Master’s in Data Science**

- **Elsewhere at Villanova**
  - Business School – MS Analytics, minor, Center Bus Analytics
  - Engineering – Center for Analytics of Dynamic Systems
OVERVIEW

• Stat vs. CS: Why Stat is losing the DS battle!
• DS at Villanova
• What I Learned from Collaborating with CS colleagues
• What I Want my CS Colleagues to Know
• DS related Resources
WHAT I LEARNED FROM CS COLLEAGUES

- Data Life Cycle
DATA LIFE CYCLE

Generation
Collection
Storage
Security
Manipulation
Visualization
Analysis
Communication
Impact

Need to understand upstream and downstream implications
WHAT I LEARNED FROM CS COLLEAGUES

• Data Life Cycle...extends beyond statistics
• Language
• CS Ed – pseudo-code, scaffolding programming skills
• Speed of publication is much faster, including proceedings
• (Reminded) Sometimes...simple answers are the best
• Sometimes Python is better than R
OVERVIEW

• Stat vs. CS: Why Stat is losing the DS battle!
• DS at Villanova
• What I Learned from Collaborating with CS colleagues
• What I Want my CS Colleagues to Know
• DS related Resources
WHAT I WANT MY CS COLLEAGUES TO KNOW

• Long history in statistics (see Donoho)
• Big data does not solve all your problems
• Statistical thinking is more than just descriptive measures
  • Uncertainty
  • Inference is sometimes appropriate
• Bias
  • Missing data
  • Programming choices – aggregation, etc.
  • Example
• Multivariate Thinking - confounders
  • Data Viz is more than just plotting data
Know what you know. Know what you don’t know.

What you know, you know, what you don't know, you don't know. This is true wisdom.

--- Confucious
YOUR DS PROGRAM ... SHOULD BE DESIGNED LOCALLY

- DS programs are popping up all over the place
- Suggestions for designing a DS program
  - Recognize local resources
    - Faculty expertise – Stat, CS, domain
    - Industry partnerships – curriculum, job demand, support
  - Leverage existing infrastructure
    - ...as little as possible
  - Leverage existing resources
  - Focus on desired skillset
HOW IS DS DIFFERENT THAN STAT?

• Data life cycle and workflow
  • Some aspects don’t apply to statistics

• End product
  • Statistics – model
  • Data Science – scalable app
RESO URC ES

- NAS
  - Envisioning the Data Science Discipline
  - DS for UGs: Opportunities and Options
  - http://sites.nationalacademies.org/CSTB/CurrentProjects/CSTB_175246

- Curriculum Guidelines
  - Park City - https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf
  - ACM – Computing Competencies for UG DS Curricula
    - www.cs.williams.edu/~andrea/DSReportInitialFull.pdf

- NSF – Harnessing the Data Revolution (HDR) grants
  - https://www.nsf.gov/cise/hamessingdata/


- Harvard DS Review - https://review.datascience.harvard.edu/

- Teach DS Blog - https://teachdatascience.com/