Scaling a Data Science Curriculum to the Masses

Success and Failures in the Undergraduate Classroom

Thomas J. Fisher

MIAMI UNIVERSITY

OXFORD, OH • EST. 1809

Miami University

- Public University in Oxford, Ohio
- Chartered 1809, opened in 1824
- Modern expansion during 1950s-1970s
- Roughly 19,000 students
  - Just under 17,000 undergraduates
  - Approximately 2,300 graduate students
- Typically ranked in the top 5 universities for undergraduate teaching
- Data Science and Analytics efforts spread across 3 divisions (colleges)

Department of Statistics

- Came into existence in 2009, I joined in August 2013
- BS in Statistics (roughly 140 majors)
- BS in Mathematics & Statistics (another 140)
- Analytics Co-Major (roughly 100 students, ~40 double counted)
- MS in Statistics (~25 students)
- About 2600 students take Intro Stat every year (10 GAs)
- 12 continuing faculty
  - 3 in Administrative roles
  - 2 Lecturers/Clinical Faculty
  - 3 tenured non-admin faculty

Current Data Science Program

Predictive Analytics Co-Major

- Computing Core
  - Data structures & database access (SQL)
- Statistical Modeling
  - Regression (and related) modeling
  - Predictive Modeling
- Data Handling (STA402)
- Data Visualization (STA404)

Today we will concentrate on STA402 and STA404

STA402 - Statistical Programming

- Course concentrates on data handling and management
- Programming for Statistical problems (data, simulation, bootstrapping)
- Primarily in SAS (don't scoff!) with some R
- Handling data from the raw source (ex: http://www.aoml.noaa.gov/hrd/hurdat/hurdat2-nepac.html)
- Students are fairly well-prepared for SAS Base Certification Exam
- Use of Webscraping and connecting to SQL is also highlighted
- Analysis is secondary in this course (PROCs are treated as functions to help with data management, not the primary purpose)
- Homework and project based
  - Students complete an individual project in lieu of a final exam
  - Projects consists of data handling, programming and written report

STA404 - Advanced Data Visualization

- Concentrates on building effective visualizations of data
- Primarily taught in R (tidyverse) with some other tools (Tableau)
- Data management -- Tall-to-wide, mutating, etc...
- Static plots
  - Complexity of plot (multivariate)
  - Aesthetic choices, Formatting and interpretation
- Dynamic plots
  - Shiny and dashboards
  - Plotly for additional interactivity
- Largely project based
  - Students work in groups to complete projects
  - Many are for external clients: http://dataviz.miamioh.edu/

Successful arching theme

- Data oriented -- Let the data speak for itself
- Data handling
  - From raw source into a format usable for analysis
  - Students get their hands dirty
- Story telling!
  - Data is not just numbers, it provides insight
  - What is the story? How does the visualization aid in that story?
- These two classes get students jobs!!!
- Great publicity: http://miamioh.edu/news/top-stories/2018/05/overdose-web-app.html
- Successful DataFest participants typically have both classes under their belt
- All things we “want” in a data science program!!!

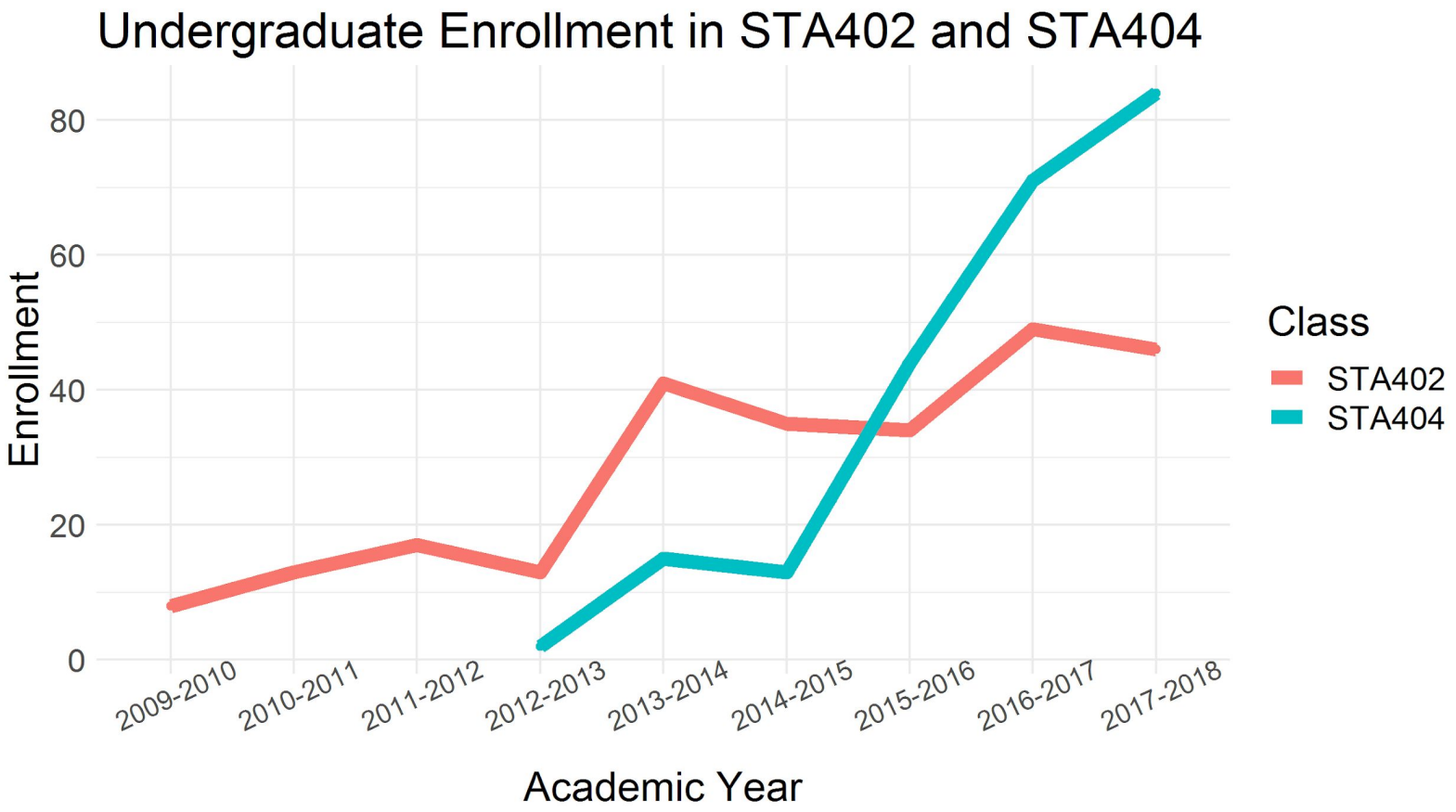
First set of challenges

- Historical Prerequisites and study body
  - Only an *Intro Stat* courses was all that is required for STA402 or STA404
  - Statistics, Mathematics, Computer Science, Business (broad range) students
  - Both classes also cross-listed with our MS Statistics degree
  - Diverse academic backgrounds and understanding entering the courses
- Implemented Changes
  - STA402 now requires (and soon STA404 will require) our Introduction to Statistical Modeling course (STA363)
  - STA363 is essentially Stat2 and is being redesigned to include an introduction to modern R
    - Applied multiple regression modeling and some experimental design
    - Data handling and plotting in the *tidyverse*
    - This will become the gateway course for the Statistics and/or Data Science program

Second set of challenges

- Both classes are Project based and involve heavy coding
  - Finding ‘fresh’ projects becomes overly burdensome
  - Sometimes they have been *Client-based* - competing against Stat Consulting course
  - Grading is burdensome, particularly during finals
  - Group work leads to academic dishonesty and grade inflation
  - Finding a balance is key
- Staffing
  - Traditional Statistics PhD does not prepare for these classes
  - Been slowing getting more and more faculty up-to-speed
    - Not just a new prep, but new material for many faculty
  - Due to the design of the course, classes are typically capped in the low 20s
  - Cannot have faculty just teach these two courses.

Third challenge



Third Challenge - Scaling

- Enrollment in this two classes has increased substantially since 2012-2013
- Classes are capped otherwise there would be more!
  - 20+ students are denied these classes each semester
- The plot does not include graduate student enrollment (another 15-20 per year)
- In Fall 2018
  - STA402 has 37 students enrolled with a waitlist of 12
  - STA404 has 36 students enrolled with a waitlist of 4
  - 20 Graduate students are spread across both classes as well
- How to scale to the masses?

By no means an all our fix but we can borrow from Computer Science

An attempt at scaling - STA402 Final Projects

Before:

- Students picked a topic of interest, approved by the instructor
  - Topics ranged from statistical theory, psychology, biology, everything
- Several milestones (proposal, progress report, rough draft and final draft)
- Grading the rough draft would take at least an hour (or more) per student
- Grading the final drafts would take ~30 minutes per student

Good:

- Meaningful experience for students: in theory its data they care about!
- Student lead - they made many decisions (and consequences!)

Bad:

- Variability in project difficulty
- More than ~20ish students is infeasible for one faculty member

An attempt at scaling - STA402 Final Projects

Spring 2018 (class of 22 students) I picked projects:

- Each student received a “randomly assigned” individual project
- Based on 16 datasets, everyone had a different task/different dataset (minimal overlap)
- Control of difficulty - all assignments had similar elements
- Not so exploratory - students were encouraged (and graded) on adding their own work but were given an explicit set of task

- Students submitted:
  - Written report
  - Executable Source code
- Explicit projects lead to a more straightforward grading rubric

An attempt at scaling - STA402 Final Projects

- Grading rubric
  - Code readability - meaningful comments, understandable functionality, etc…
  - Code execution - REPRODUCIBILITY of all reported results
  - Code robustness - I tested all functionality with different but similar data
  - Report details - Correctness, description, analysis and formatting

- Anecdotal evidence of improvement
  - Written reports were better (end goal, not so much rambling)
  - Code was much more polished
  - I could replicate the results for most students
  - Students ‘earned’ their final grades

Scaling creates new challenges

Findings of this ‘experiment’

- Ton of upfront work -- assigning projects took some time
  - I managed with 22 in the spring, confident I could handle ~30
  - With TA help (not available currently) could handle more (this is what Comp Sci does)
- Grading became manageable - no more work than a standard final exam (maybe less for some projects)
- Keeping projects “fresh” could be a challenge
  - Time-dependent data - always updating
  - More and more public datasets are available
- Something is lost in the experience