Data 8: *Foundations of Data Science*

- In the 2015-2016 academic year, we taught over 500 students, most at the freshman level, in our new Data Science course
  - a wide range of students, pre-disciplinary
  - it was a major success

- In the 2016-2017 academic year, we expect to teach over 1000 students

- The course is a conceptual blend of traditions in statistics and computer science, with a strong outward-looking component (“connector courses”)
Some shapers of our perspective

- The increasingly prominent use of “data science” in job descriptions in the tech industry
  - and our undergraduates’ awareness of the term
- The rise of the “CS for All” movement
  - and its partial failure at Berkeley
- Success in teaching inference at the graduate level via resampling and sampling ideas
  - surely this can work at the undergrad level too?
- The explosive growth of iPython (Jupyter)
  - and related browser-based computing environments

A job description, circa 2016

- Your Boss: “I need a Big Data system that will replace our classic service with a personalized service”
- “It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us”
- “It should run just as fast as our classic service”
- “It should only improve as we collect more data; in particular it shouldn’t slow down”
- “There are serious privacy concerns of course, and they vary across the clients”
Computational thinking

- Abstraction, modularity, scalability, robustness, etc.
  - ideas that are taught beautifully and successfully in freshman-level courses in computer science

- What’s missing in computational thinking:
  - much of what we teach in our best statistics courses

Inferential thinking

- We’ve coined the term “inferential thinking” to emphasize:
  - that our perspective is similar in depth and breadth to that of computer science
  - that we are complementary to computer science
  - that “data science” can be viewed, at least in part, as the confluence of these two styles of thinking
How to teach inferential thinking

Written in Berkeley in the 1970's, FPP transformed the way statistics is taught.

Analyzing data: three main steps

- The question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and calculations
- Interpretation of the results in the language of the domain, without statistical jargon
Stat instruction: a common approach

- The question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and formulas
- Interpretation of the results in the language of the domain, without statistical jargon

The FPP way

- The question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and formulas
- Interpretation of the results in the language of the domain, without statistical jargon
The Data 8 way

- The question, from some domain; reasonable assumptions about the data; choice of method
- Visualization and computation
- Interpretation of the results in the language of the domain, without statistical jargon

The approach

- A course for freshmen
- No prerequisites
- Teach data science as a way of thinking:
  - Computational thinking
  - Inferential thinking
The DataBears

Over 500 students took Foundations of Data Science in 2015-2016. We expect over 1000 in 2016-2017.

Data 8 Spring 2016 survey

- About 450 students, 418 responses
  - 42% women
  - 17% underrepresented ethnic or racial minority
  - 52 different intended majors

- Does the course fulfill a major requirement?
  - Yes: 8.6%
  - Maybe: 10.4%
  - I don’t know: 10.6%
  - No: 70.4%
Of the roughly 110 students in Foundations of Data Science in Fall 2015, about 30 worked on the Spring 2016 offering of the course and its connectors, as lab assistants, tutors, graders, and course developers.

Background for demo

- Demo is material from about Week 5
- At that point, the course has not covered
  - average, SD, percentiles
  - formulas for probability distributions
- But it has covered
  - Python expressions, arrays, tables
  - Defining functions; iterating
  - and a property of large random samples:
Sample resembles population

(Demo 1)
Syllabus for Inference in Data 8

- Confidence intervals, for estimation and for testing hypotheses
- Comparing two distributions
  - Categorical
  - Quantitative
- Predicting one variable based on others
  - Classification
  - Regression

Taught by using simulation and the bootstrap
Three 2-week projects

● **Water use in California:** mapping the water districts and overlaying IRS taxable income data by zip code

● **Murder rates and the death penalty:** nonparametric inference; the importance of visualization

● **Classification of song lyrics:** hip-hop or country?

“Connector” Courses

● Data science in a particular domain
● Much smaller: 5-50 students
● Half the time commitment of a typical undergraduate course
● Can be taken concurrently with Data 8, or later
Connector Courses, 2015-2016

- Data Science for Smart Cities
- Data Science and the Mind
- Data Science in Ecology and the Environment
- Exploring Geospatial Data
- How Does History Count?
- Data and Ethics
- Health, Human Behavior, and Data
- Race, Policing, and Data Science
- Literature and Data
- Matrices and Graphs in Data Science
- Computational Structures in Data Science
- Probability and Mathematical Statistics for Data Science

Conclusion

- A reinvention of statistical education in an era of pervasive computation
- Deep, substantive collaboration with Computer Science
- Collaborations with departments and faculty campus-wide, with the aim of making Data Science a core part of a liberal education
A conversation on Piazza

- Post by Anonymous
  Would someone please explain why this is:
  Using **Randomness** in **Inference** allows us to understand how “sure” we may be of an answer.

- The students’ answer
  When you use randomness many times and see a similar pattern, it becomes easier to make an inference on the data.

Environment and course materials

- Jupyter notebooks; Python 3
- JupyterHub (thanks to Jessica Hamrick)
  - Multi-user server for Jupyter notebooks
  - Browser-based computation in the cloud

Source: [https://github.com/data-8](https://github.com/data-8)
New Python datascience library: [https://pypi.python.org/pypi/datascience/](https://pypi.python.org/pypi/datascience/)

Fall 2016 course (available late August 2016): [data8.org](http://data8.org)