The data science revolution in education: a report from the front lines

Matthew Brett

Undergraduate data science

> *. . . academic institutions should encourage the development of a basic understanding of data science in all undergraduates.*

National Academies of Sciences and Medicine (2018)

> *I think it likely that in ten years' time every undergraduate programme will have to include some teaching in data science.*

Professor Sir Adrian Smith, Director of Turing Institute, October 2019.

But:

- What is data science?
- What should we teach?
- What is it going to look like?

Data science in Berkeley

- February 2013: Supporting Data Science Workshop
- 2013: Berkeley Institute of Data Science
- 2015: Foundations of data science course
- 2018: National workshop on data science education
- 2018: Division of data science: announced

Berkeley teaching programme

- Massive (~1500 student) course Foundations of data science - "Data 8". No requirements in mathematics or programming. Running since 2015.
- Large (~900 student) intermediate course Principles and techniques of data science with further requirements in Python programming and linear algebra.
- 27 Connector courses: domain applications of teaching methods from the foundation course.
- "... embracing a reinvention of statistical education in the era of pervasive computation." Report by Data science education rapid reaction team
- The greatest change in undergraduate teaching in a generation.

Data 8 elements

- ▶ **No Prerequisites** - There are no prerequisites besides the high school math that it took you to get to UC:
  - ▶ No Math or Programming Background Required
- ▶ **Easy to Access** - There is an easy-to-use computing platform that is integrated into every aspect of the course and works on any browser with no technical requirements
  - ▶ No specific computer, software license, install needed
- ▶ **Coding and Statistics** - There are advantages to learning coding at the same time as learning statistics - *Computational Thinking, Inferential Thinking*
- ▶ **Visualize First** - represent the data graphically in order to motivate questions about inference and concepts of statistics
- ▶ **Minimize Formulas** - Don't show or depend on formulas, delay formulas until the concept is explained, intuition and interpretation are more important
- ▶ **Re-Sampling** - guide students towards non-parametric approaches by motivating resampling for hypothesis testing

Eric Van Dusen - slide deck from 2019 conference

Principles of the course

- Teaching statistics "assuming computers exist, rather than assuming they don't exist."
- "Express in code what we would otherwise express in equations."

John DeNero, 2018 Webinar

Analyzing data: the 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

Ani Adhikari - slide deck from 2019 conference

The old approach

- 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

- 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

A demo

To follow along: http://bit.ly/swains_jury

Students, Spring 2019

- 49% first-years, 35% second-years
- 55% female
- 21% consider themselves to be a member of an underrepresented ethnic or racial minority within UC Berkeley
- Over 60 different majors
- At the start of the term, 38% said, "I have no skill at programming"

Ani Adhikari - slide deck from 2019 conference

Students, Spring 2017

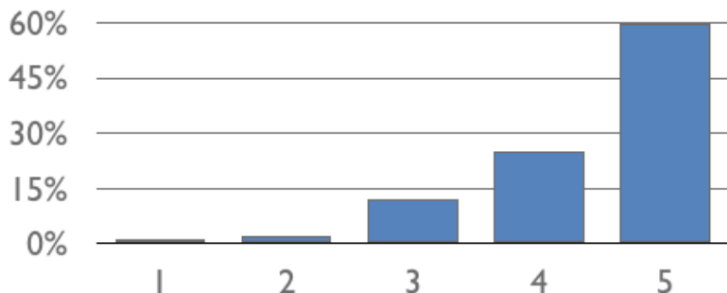What is your declared or intended major? (618 responses)

16.9%

14.5%
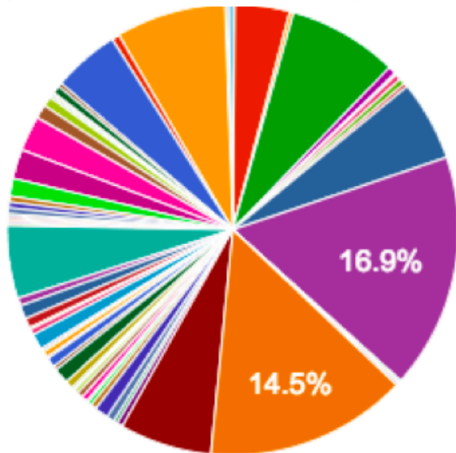
Wide range of majors, $> 14\%$ slices are economics, computer science – John DeNero, 2018 Webinar materials

Student feedback

How happy are you about your decision to take Data 8?
(418 responses)

John DeNero, 2018 Webinar materials

Student feedback, Spring 2019

- I never thought I would ever code or program but this class made it really approachable.
- Learn[ed] to code in a way that I feel will actually be useful for me in the future, even as someone in a social sciences major.
- Loved the problem-solving skills this class taught me and how Data 8 showed me the various ways data science could be applied to multiple disciplines!
- Data manipulation helped change literally how I see the world.

Ani Adhikari - slide deck from 2019 conference

Coding is not a specialist skill

- understand what algorithms are, how they are implemented as programs on digital devices, and that programs execute by following precise and unambiguous instructions;
- create and debug simple programs;
- use logical reasoning to predict the behaviour of simple programs
- use technology purposefully to create, organise, store, manipulate and retrieve digital content

Coding is not a specialist skill

- understand what algorithms are, how they are implemented as programs on digital devices, and that programs execute by following precise and unambiguous instructions;
- create and debug simple programs;
- use logical reasoning to predict the behaviour of simple programs
- use technology purposefully to create, organise, store, manipulate and retrieve digital content

National curriculum in computing: Key stage 1 (5-7 year olds).

Undergraduate data science at the University of Birmingham

- 2017: Bioscience second year undergraduates; Introduction to R programming language.
- 2018: "Data science for everyone" WHM
- 2018: Geography first year undergraduates; Introduction to R.
- 2019: Introduction to programming second-year undergraduate / MSc module (10 UK credits, run over one term).
- Plans to extend this approach across modules and schools.

What do we need?

- debate
- vision
- mechanism

What can we do?

- CPD to explain pedagogy.
- Harness commitment from faculty and students.

The end

Materials at https://github.com/matthew-brett/ds-rev