Cross-discipline teaching in data analysis

Matthew Brett

Data science

. . . academic institutions should encourage the development of a basic understanding of data science in all undergraduates – A 2018 report from the US National Academies of Sciences, Engineering and Medicine.

Data science

. . . the really important intellectual event of the next fifty years" – Donoho (2015) "50 years of data science".

Really though?

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Harvard Business Review, October 2012.

What is it?

- Radical change in culture of data analysis.
- Code as foundation;
- Analysis becomes:
  - versatile (big, messy, mixed, complex);
  - realistic about the work of analysis (cleaning, exploring . . .);
  - collaborative;
  - reproducible;
- Emphasis on algorithms over mathematics for analysis and explanation.

Reproducible

Computing results are now being presented in a very loose, "breezy" way—in journal articles, in conferences, and in books. All too often one simply takes computations at face value. This is spectacularly against the evidence of my own experience. I would much rather that at talks and in referee reports, the possibility of such error were seriously examined.

– David L. Donoho (2010). An invitation to reproducible computational research. Biostatistics Volume 11, Issue 3 Pp. 385-388

Realistic

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. J. W. Tukey (1962) "The future of data analysis".

Emphasis on algorithms

The Introductory Statistics Course: A Ptolemaic Curriculum

George W. Cobb
Mount Holyoke College

Resampling for statistical explanation

Julian Simon (1997) “Resampling, the new statistics” at http://www.resample.com/intro-text-online.

Trials of teaching method in high school students (1969) and undergraduates (1976) including those with “low skills and little interest in mathematics”.

http://www.jstor.org/stable/pdf/27958125.pdf

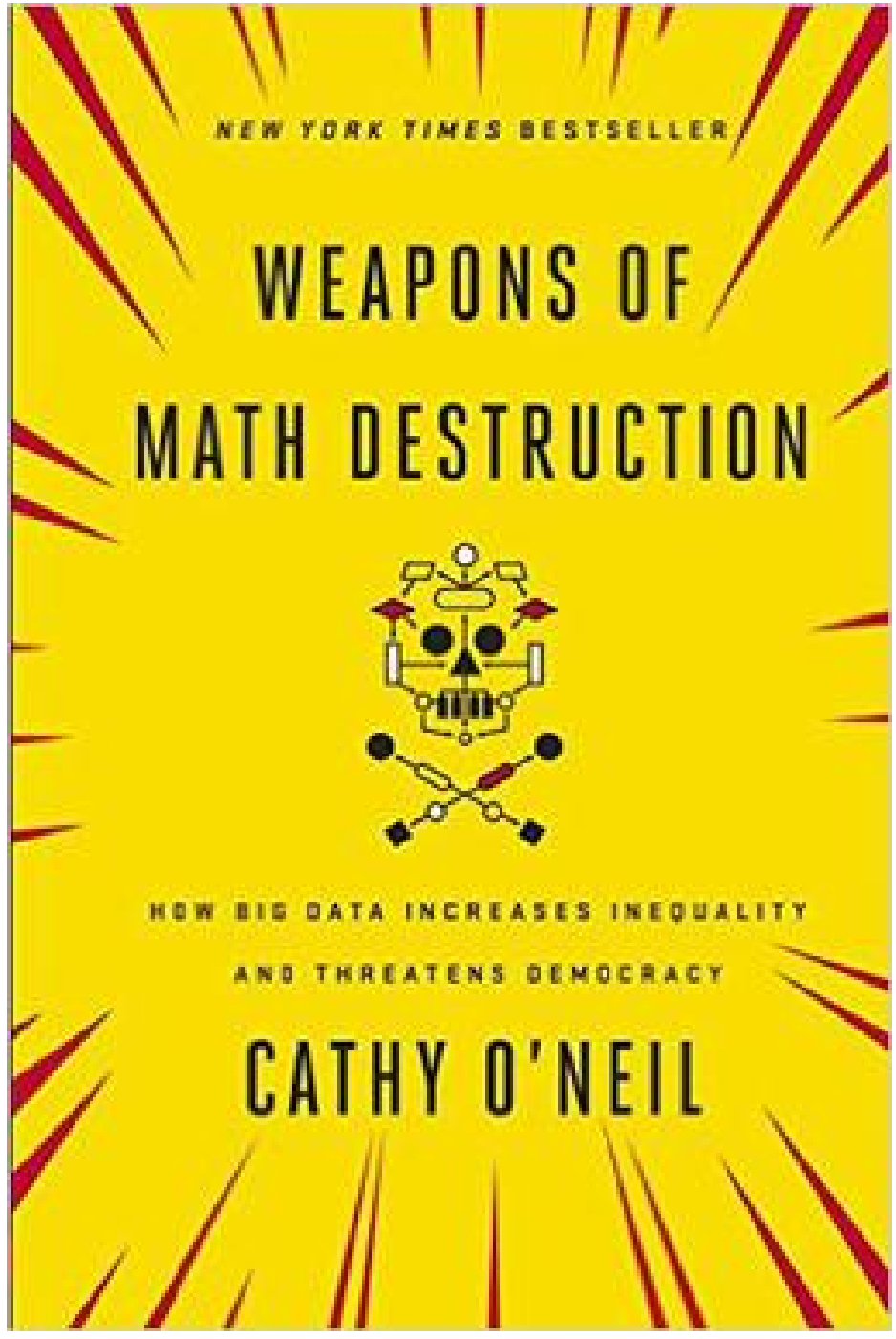
Why did this happen now?

- Industry has to get the answer right.
- Demands of data drove the
- Development of tools - Python, R, machine learning.

Introducing the Berkeley course

- http://data8.org
- "The course is designed for entry-level students from any major. It is designed specifically for students who have not previously taken statistics or computer science courses." (link).
- Fastest growing program in the University's history, with 2000 students projected to enroll for 2017-18.
- Textbook is Inferential and computational thinking.
- Source for textbook on Github.

Algorithms


NEW YORK TIMES BESTSELLER
WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL

Data journalism

Full analysis at https://github.com/propublica/compas-analysis

A short course on statistics with almost no mathematics

- permutation test;

A prediction

- in 10 years, all traditional "statistics" courses will be data science courses;
- all undergraduates will take these courses, to teach quantitative methods across fields;
- we will have much more cross-discipline teaching, because of shared courses and shared tools.
- our students will be more thoughtful and effective as academics and citizens.

Is this the end?

Yes, it's the end of the talk.

All material for this talk at:
https://github.com/matthew-brett/collab-teach

References

Simon, Julian L, David T Atkinson, and Carolyn Shevokas. 1976. "Probability and Statistics: Experimental Results of a Radically Different Teaching Method." *The American Mathematical Monthly* 83 (9). JSTOR:733–39.

Simon, Julian L, and Allen Holmes. 1969. "A New Way to Teach Probability Statistics." *The Mathematics Teacher* 62 (4). JSTOR:283–88.

Simon, Julian Lincoln. 1997. "Resampling: The New Statistics." Resampling Stats Arlington, Virginia.