Why data science?

What is data science?

> *Data Science is about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference.*

Berkeley textbook for Foundations of Data Science

Why is it important?

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

What does it involve?

> *. . . what data scientists do is make discoveries while swimming in data . . . At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. . . . Data scientists' most basic, universal skill is the ability to write code.*

Davenport & Patil (2012) "Data Scientist: The Sexiest Job of the 21st Century" Havard Business Review.

It's about getting the answer right

Scientists at Amgen (a drug company) tried to reproduce findings from 53 "landmark" studies.

> *. . . when findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors' direction, occasionally even in the laboratory of the original investigator.*

Of 53 studies, only 6 replicated (11%).

Glenn Begley and Lee Ellis (2012) "Raise standards for preclinical cancer research" Nature 483 pp. 531–533

It's about transparency

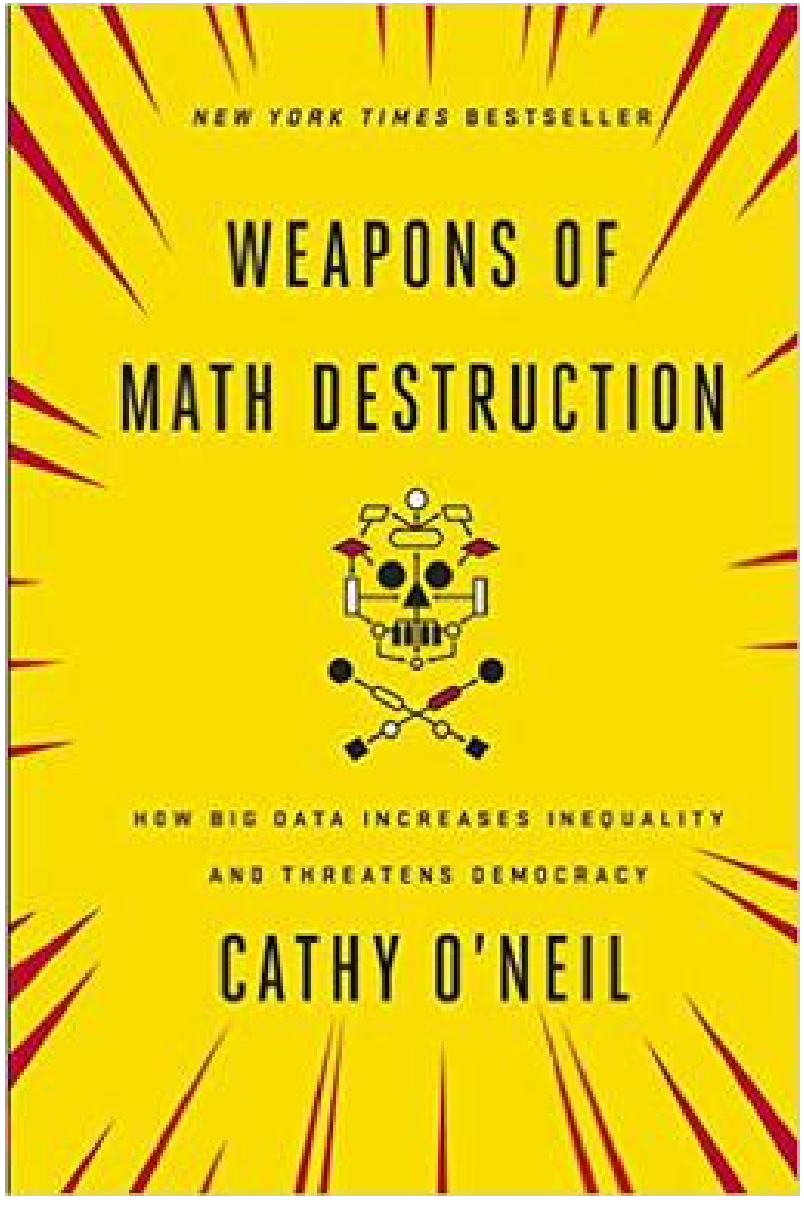
(by kladcat under CC BY 2.0, via Wikimedia Commons)

Transparency

> *The scientific method's central motivation is the ubiquity of error - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error."*

Donoho, David L, et al. 2009. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* 11, 8–18.

It's about engagement with modern data



Data journalism

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

Who is learning this?

- 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

Undergraduates

. . . 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.

What's the down side?

- Learning a new language (Python, or R)
- The value of frustration

Computers are confusing

- always work with someone else if you can;
- always get someone else to check what you did;
- force yourself to get up and walk away;
- always have a pencil and paper next to you.

Today

- pair coding: the navigator and the driver;
- if you know what to do, help your partner.
- ask me if you are stuck.