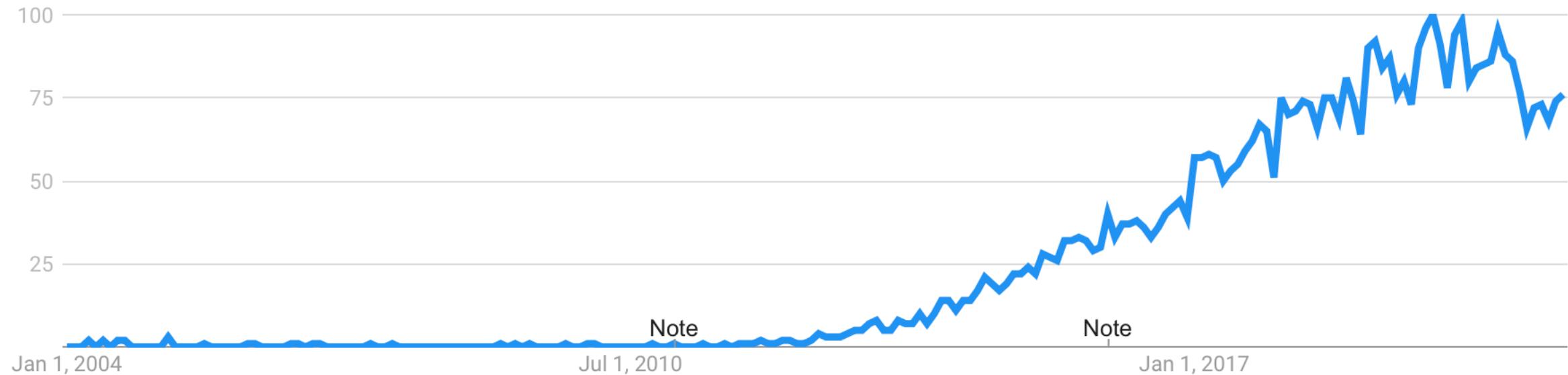


THE EMERGENCE AND FUTURE OF DATA SCIENCE

Jeff Goldsmith, PhD
Columbia Biostatistics



Data science is pretty new





Data science is pretty new





Coauthors

- The Emergence and Future of Public Health Data Science
 - Jeff Goldsmith, Yifei Sun, Linda P. Fried, Jeannette Wing, Gary W. Miller, Kiros Berhane

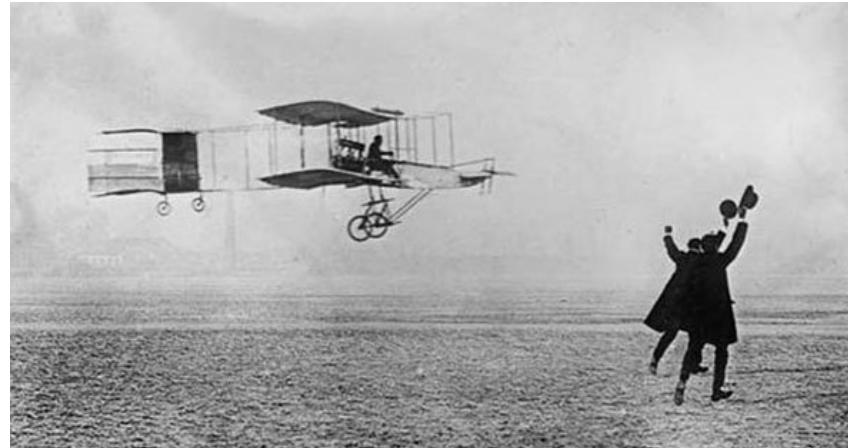


My background in data science

- I do functional data analysis motivated by
 - Wearable devices (accelerometers, mostly)
 - Motor control (stroke recovery; brain / behavior dynamics)
- I've taught P8105: Data Science I since 2017
 - Intended for MS students in biostatistics
 - Enrollment is now approx. 200
 - (That's more than 20, but less than a million)
 - Think "tidyverse as a service course"

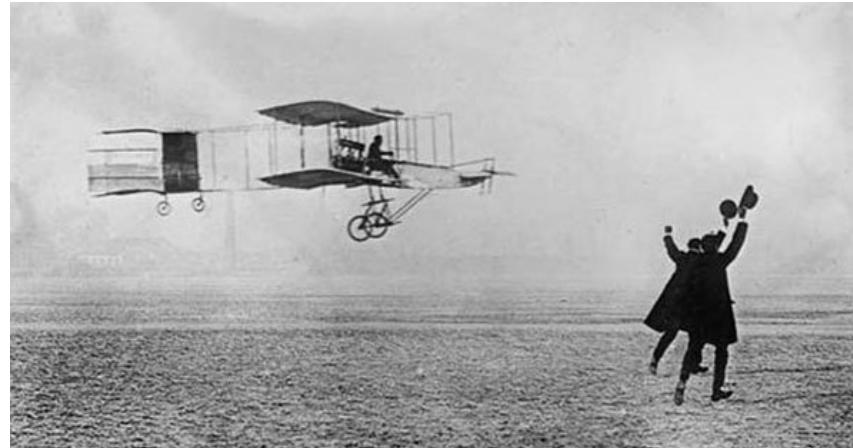
A data science analogy

1910s



A data science analogy

1910s



1969 / 1970





Defining data science

Data science is the study of extracting value from data.

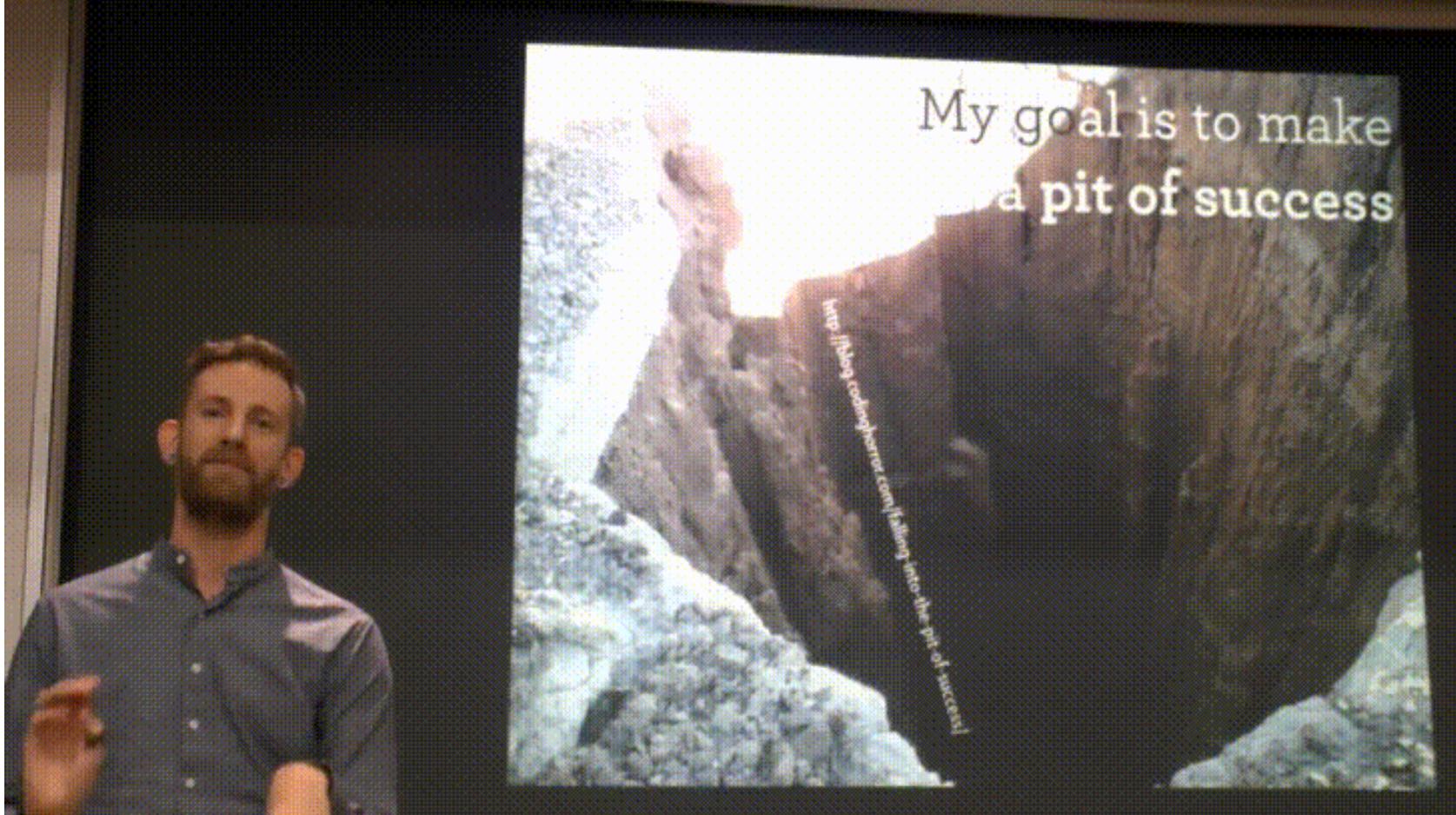


Another definition

Data science is the study of formulating and rigorously answering questions using a data-centric process that emphasizes clarity, reproducibility, effective communication, and ethical practices.

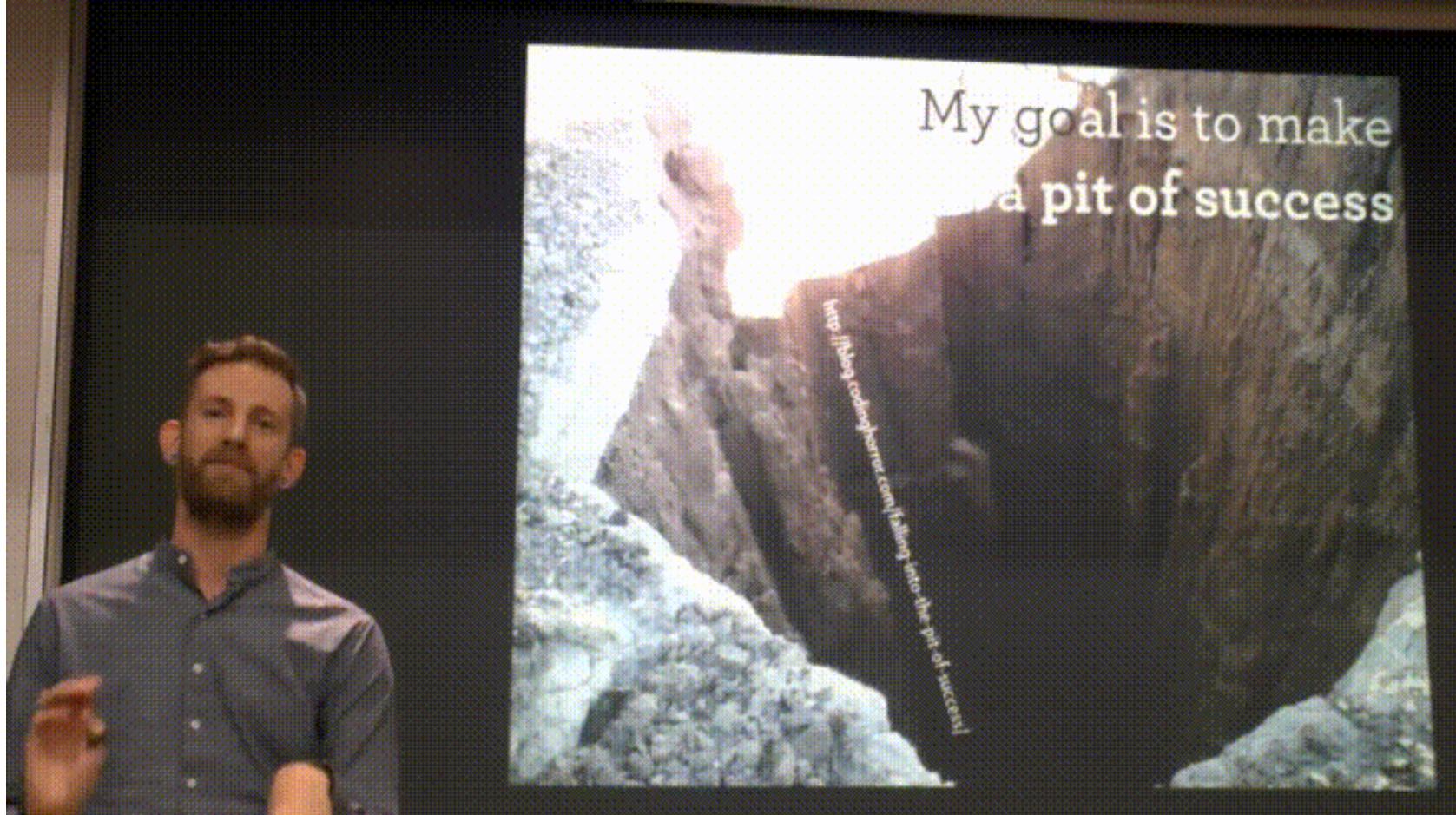


ISI 2017





ISI 2017





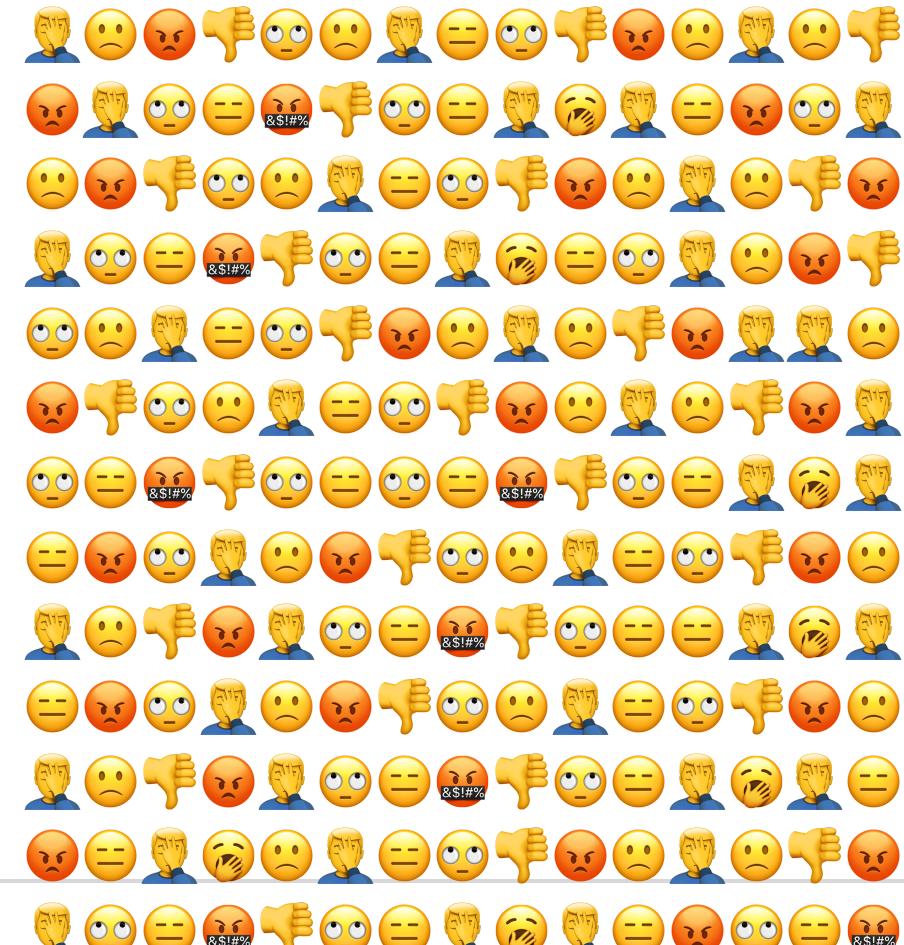
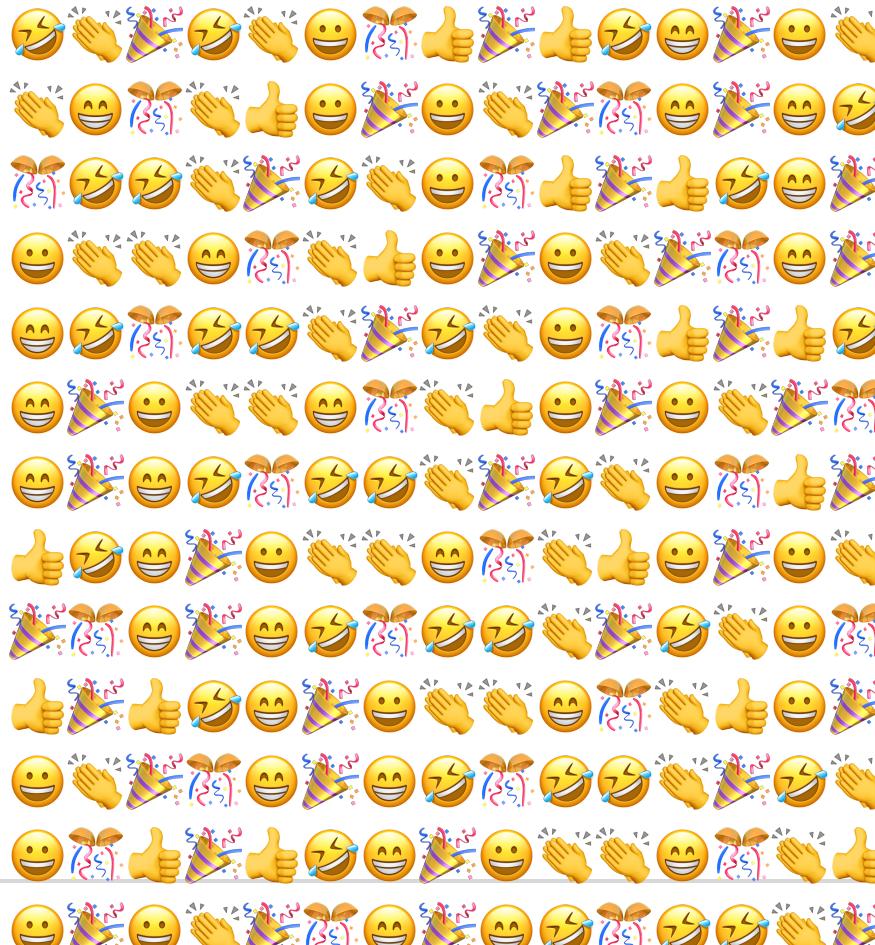
First question from the audience

“What is the point of ‘data science’? Aren’t we already data scientists?”



First question from the audience

“What is the point of ‘data science’? Aren’t we already data scientists?”





Response from Hadley Wickham (roughly)

“A data scientist is a statistician who’s useful”



Response from Hadley Wickham (roughly)

“A data scientist is a statistician who’s useful”





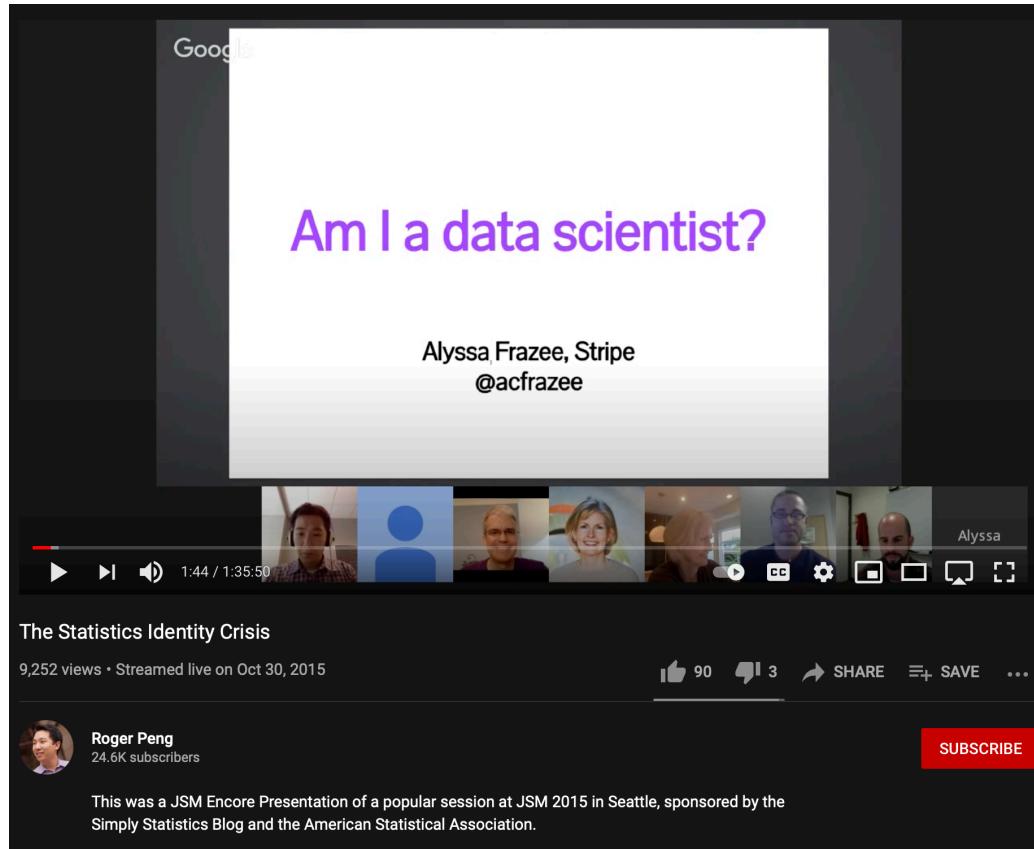
That question is understandable

- It's easy, in 2021, to forget what the statistical identity crisis phase was like
- But that was a whole thing, for quite a while



That question is understandable

- It's easy, in 2021, to forget what the statistical identity crisis phase was like
- But that was a whole thing, for quite a while





What made “data science” happen

- Data science emerged in parallel to six broad trends:
 - Big data
 - Emphasis on prediction
 - Reproducibility crisis in science
 - Interdisciplinary research
 - Diversity, equity, and inclusion
 - Everything should be on the internet
- These weren’t new in 2012 and aren’t unique to data science
- ... but they had a big impact on the “data science” perspective



Connotation >> definition

- Core data science values aren't built into the definition, but were critical to the valence of "data science"
- In statistics, "data science" mapped onto existing arguments about what matters to the field
 - Connotation seemed to resonate with a lot of vaguely disaffected applied statisticians



Data science as external validation

- The fact that data science caught on implied that
 - stated values ≠ demonstrated values
- Ideally, this would suggest a need to bring these into closer alignment
 - Not saying old values were bad – but that other things should be valued, too



Did that happen?

- Some, yeah.
 - More awareness of issues around equity and inclusion
 - Broader view of important / valid publication outlets
 - Techniques for working with data are explicitly taught
 - Slow shift towards expecting better code / reproducibility



Did that happen?

- Some, yeah.
 - More awareness of issues around equity and inclusion
 - Broader view of important / valid publication outlets
 - Techniques for working with data are explicitly taught
 - Slow shift towards expecting better code / reproducibility
 - (Exciting aside – reproducibility at JASA ...)



Did that happen?

- Some, yeah.
 - More awareness of issues around equity and inclusion
 - Broader view of important / valid publication outlets
 - Techniques for working with data are explicitly taught
 - Slow shift towards expecting better code / reproducibility
 - (Exciting aside – reproducibility at JASA ...)
- But also ... not in other ways.
 - “Find ways to get traditional academic products / credit” is the advice given to academic data scientists



So ... I think Jeannette is kinda right

- Data-oriented disciplines will slowly incorporate the values that “data science” implies in their own ways
- That’ll be true enough that “data science” will be a secondary / situational academic identity
 - “I’m a [...] and data scientist” not “I’m a data scientist”
 - “For this grant, I’m a data scientist”
- Upshot is that a maximalist definition of data science will win, in practice, over a definition that tries to create a clear boundary / distinct discipline
 - This is not a bad thing



Public Health Data Science

[Public health] data science is the study of formulating and rigorously answering questions [in order to advance health and well-being] using a data-centric process that emphasizes clarity, reproducibility, effective communication, and ethical practices.



DS \leftrightarrow PHDS

- “Data science” will evolve as it draws on existing domain skills and traditions
- PHDS will add some ways of thinking and tools from other quantitative disciplines



Some predictions about PHSD

- It'll follow the data science trajectory, just delayed a few years
 - A “PHDS is just ...” phase will happen and then be mostly over
 - Public health data scientists will be common outside academia
 - This is why people take my class ...
 - This requires academic and professional perspectives
 - ⇒ PHDS training programs will proliferate

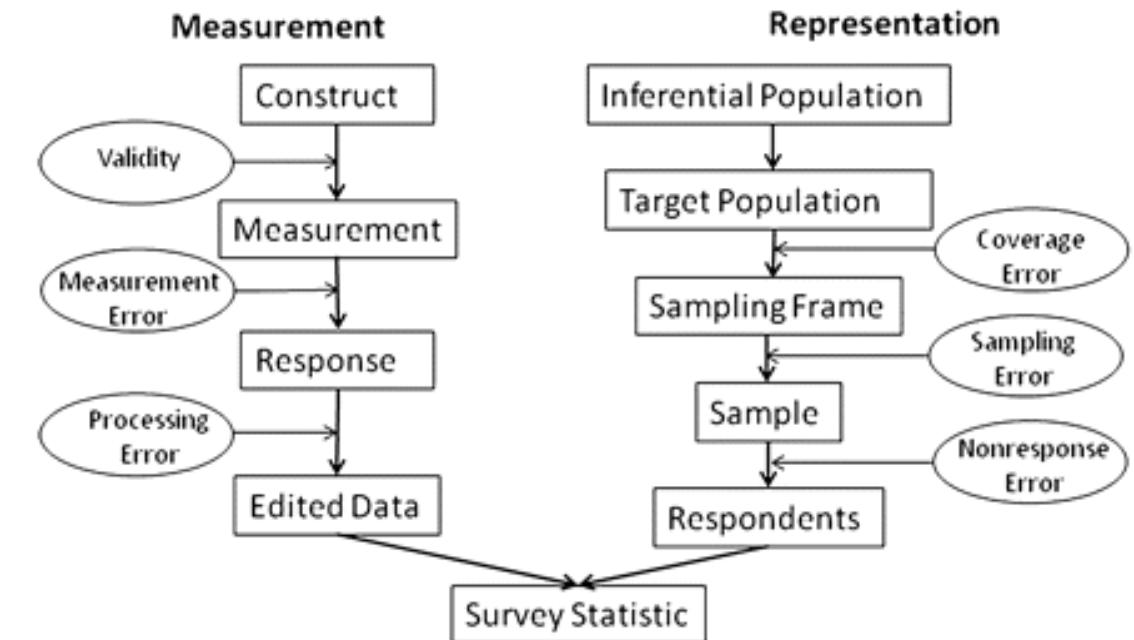


“Public Health” is important

- Public health training emphasizes some elements that are critical data science thinking and work:
 - Study design
 - Sampling process
 - Measurement process
 - Desire vs ability to infer causation
 - Cross-disciplinary collaboration
 - Engagement with data ethics
 - Public dissemination and dialog

“Public Health” is important

- Public health training emphasizes some elements that are critical data science thinking and work:
 - Study design
 - Sampling process
 - Measurement process
 - Desire vs ability to infer causation
 - Cross-disciplinary collaboration
 - Engagement with data ethics
 - Public dissemination and dialog



From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg)
via “Data Alone Isn’t Ground Truth” by Angela Bassa



Thanks!

- jeff.goldsmith@columbia.edu
- jeffgoldsmith.com
- github.com/jeff-goldsmith/

- P8105.com

