

data-ppf.github.io mar 30, 2021

data science, 1962-present

chris wiggins + matt jones, Columbia

student reactions

student reactions

237 science
69 statistics
41 donoho
35 scientists
35 neff
28 ethics/ethical
5 matter/mattered
5 wiggins
4 breiman
3 tukey
3 interpretability
2 gatekeeping/gatekeeper

“matter”?

- ▶ At first, while I was reading the Donoho piece I was unsure of why defining what data science really was mattered. To me, the argument appeared to be a conflict of egos, with statisticians feeling ripped off and data scientists wanting to feel cool.

“matter”?

- ▶ At first, while I was reading the Donoho piece I was unsure of why defining what data science really was mattered. To me, the argument appeared to be a conflict of egos, with statisticians feeling ripped off and data scientists wanting to feel cool.
- ▶ Does it matter if data science is the joint application of statistics and computer science? Does there need to be some binding element that sets it apart?

“matter”?

- ▶ At first, while I was reading the Donoho piece I was unsure of why defining what data science really was mattered. To me, the argument appeared to be a conflict of egos, with statisticians feeling ripped off and data scientists wanting to feel cool.
- ▶ Does it matter if data science is the joint application of statistics and computer science? Does there need to be some binding element that sets it apart?
- ▶ How would data science jobs be different if the field was considered a subdivision of math rather than science? Does this distinction really matter?

“matter”?

- ▶ At first, while I was reading the Donoho piece I was unsure of why defining what data science really was mattered. To me, the argument appeared to be a conflict of egos, with statisticians feeling ripped off and data scientists wanting to feel cool.
- ▶ Does it matter if data science is the joint application of statistics and computer science? Does there need to be some binding element that sets it apart?
- ▶ How would data science jobs be different if the field was considered a subdivision of math rather than science? Does this distinction really matter?
- ▶ I would echo some of the previous commenters in noting that it's very valid to ask the question, “Why does this even matter?” and who benefits from what we label a certain form of analysis?

gate-keeping and S-T-E-M

- ▶ The debate over what exactly constitutes data science and who can be considered to be a data scientist to me is reflective of the larger issue of 'gatekeeping' that often permeates STEM fields.

gate-keeping and S-T-E-M

- ▶ The debate over what exactly constitutes data science and who can be considered to be a data scientist to me is reflective of the larger issue of 'gatekeeping' that often permeates STEM fields.
- ▶ Donoho seems embittered about the funneling of resources and opportunities away from traditional statistics departments and towards shiny new data science initiatives, and he comes across at times as a kind of attempted gatekeeper (for instance he seems bothered by the fact that none of the faculty of the UC-Berkeley data science program come from traditional academic statistics backgrounds.)

gate-keeping and S-T-E-M

- ▶ The debate over what exactly constitutes data science and who can be considered to be a data scientist to me is reflective of the larger issue of 'gatekeeping' that often permeates STEM fields.
- ▶ Donoho seems embittered about the funneling of resources and opportunities away from traditional statistics departments and towards shiny new data science initiatives, and he comes across at times as a kind of attempted gatekeeper (for instance he seems bothered by the fact that none of the faculty of the UC-Berkeley data science program come from traditional academic statistics backgrounds.)
- ▶ Donoho references Professor Wiggins' assertion that data science is not a science but a form of engineering. I'm curious about the meaning of of this distinction, and would like to discuss it further in Tuesday's class.

reminder of course coarse themes

reminder of course coarse themes

- ▶ data and truth: labels and ideas

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
 - ▶ are dynamic (e.g., “ML”)

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
 - ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
 - ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011
- ▶ data and power:

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
 - ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011
- ▶ data and power:
 - ▶ fields are “moving targets”...

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
 - ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011
- ▶ data and power:
 - ▶ fields are “moving targets”...
 - ▶ ... because they are moved!

reminder of course coarse themes

- ▶ data and truth: labels and ideas
 - ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
 - ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011
- ▶ data and power:
 - ▶ fields are “moving targets”...
 - ▶ ... because they are moved!
 - ▶ including by pieces like our reading

ideas to look for today

outline of today

outline of today

1. math stats vs. applied stats post-WWII to today

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA
 - ▶ aftermath

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA
 - ▶ aftermath
2. the “data science” deluge 2009-2020

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA
 - ▶ aftermath
2. the “data science” deluge 2009-2020
 - ▶ industry

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA
 - ▶ aftermath
2. the “data science” deluge 2009-2020
 - ▶ industry
 - ▶ academia

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA
 - ▶ aftermath
2. the “data science” deluge 2009-2020
 - ▶ industry
 - ▶ academia
3. data science today: moving targets

outline of today

1. math stats vs. applied stats post-WWII to today
 - ▶ Tukey's FoDA
 - ▶ aftermath
2. the “data science” deluge 2009-2020
 - ▶ industry
 - ▶ academia
3. data science today: moving targets
4. data science future: broader impact, human impact

1. math stats vs. applied stats post-WWII to today

1. math stats vs. applied stats post-WWII to today

Recall from Part 1:

- ▶ Neyman/fisher 1955-1956 math-fight

1. math stats vs. applied stats post-WWII to today

Recall from Part 1:

- ▶ Neyman/fisher 1955-1956 math-fight
- ▶ “statistics” outside math stats (FDA, ETS, NIH)

1. math stats vs. applied stats post-WWII to today

Recall from Part 1:

- ▶ Neyman/fisher 1955-1956 math-fight
- ▶ “statistics” outside math stats (FDA, ETS, NIH)
- ▶ within stats:

1. math stats vs. applied stats post-WWII to today

Recall from Part 1:

- ▶ Neyman/fisher 1955-1956 math-fight
- ▶ “statistics” outside math stats (FDA, ETS, NIH)
- ▶ within stats:
- ▶ “statisticians have long focused on estimation to the exclusion of prediction” - D. Madigan

1.1 Tukey's FoDA

Tukey (1915-2000)

- ▶ math->applied math, stats



Figure 1: with computer

Tukey (1915-2000)

- ▶ math->applied math, stats
- ▶ consulting & intelligence work



Figure 1: with computer

Tukey (1915-2000)

- ▶ math->applied math, stats
- ▶ consulting & intelligence work
- ▶ split academia (Princeton) + industry (Bell) 1945-1995



Figure 1: with computer

Tukey (1915-2000): role of WWII

Previously in pure math, then. . .

John was indeed active in the analysis of the Enigma system and then of course was part of our force in the fifties which did the really historic work on the Soviet codes as well. So he was very effective in that whole operation. - W.O.Baker, former president of Bell Labs

Thanks to "war problems," "it was natural to regard statistics as something that had the purpose of being used on data—maybe not directly, but at most at some remove. Now, I can't believe that other people who had practical experience failed to have this view, but they certainly—I would say—failed to advertise it." - John Tukey interview, 1985.

Recall about Bell from prior weeks

it was quite clear that if you look at some of the work of Shannon in connection with cryptography, that must have stemmed from contacts that we had. We had very close contacts with the Bell Laboratories. They were very, let's say, willing to work along with us.

- ▶ S. Kullback 1982 NSA oral history interview, declassified 2015

Recall about Bell from prior weeks

it was quite clear that if you look at some of the work of Shannon in connection with cryptography, that must have stemmed from contacts that we had. We had very close contacts with the Bell Laboratories. They were very, let's say, willing to work along with us.

- ▶ S. Kullback 1982 NSA oral history interview, declassified 2015
- ▶ Shannon->Tukey->Chambers-> Donoho and Yu

Recall about Bell from prior weeks

it was quite clear that if you look at some of the work of Shannon in connection with cryptography, that must have stemmed from contacts that we had. We had very close contacts with the Bell Laboratories. They were very, let's say, willing to work along with us.

- ▶ S. Kullback 1982 NSA oral history interview, declassified 2015
- ▶ Shannon->Tukey->Chambers-> Donoho and Yu
- ▶ Communications problems as incubator of technologies

Recall about Bell from prior weeks

it was quite clear that if you look at some of the work of Shannon in connection with cryptography, that must have stemmed from contacts that we had. We had very close contacts with the Bell Laboratories. They were very, let's say, willing to work along with us.

- ▶ S. Kullback 1982 NSA oral history interview, declassified 2015
- ▶ Shannon->Tukey->Chambers-> Donoho and Yu
- ▶ Communications problems as incubator of technologies
- ▶ Key example of US Cold War R&D

e.g., Shannon: mathematical theory of crypto/communications

[24] ~~CONFIDENTIAL~~ (P7)

1-17810-1 (4-48)

COVER SHEET FOR TECHNICAL MEMORANDA
RESEARCH DEPARTMENT

SUBJECT: A Mathematical Theory of Cryptography - Case 20878 (4)

ROUTING:

- 1 - H.W.B.-Case Files
- 2 - CASE FILES
- 3 - J. W. McRae
- 4 - L. Espenschied
- 5 - H. S. Black
- 6 - F. B. Llewellyn
- 7 - H. Nyquist
- 8 - B. M. Oliver
- 9 - R. K. Potter
- 10 - C. B. H. Feldman
- 11 - R. C. Mathes
- 12 - R. V. L. Hartley
- 13 - J. R. Pierce
- 14 - H. W. Bode
- 15 - R. L. Dietzold
- 16 - L. A. MacCall
- 17 - W. A. Shewhart
- 18 - S. A. Shelkumoff
- 19 - C. E. Shannon
- 20 - Dept. 1000 Files

MM- 45-110-92
DATE September 1, 1945
AUTHOR C. E. Shannon
INDEX NO. P 0.4

~~CONFIDENTIAL~~

ABSTRACT

DOWNGRADED AT 3 YEAR INTERVALS
DECLASSIFIED AFTER 12 YEARS
GPO 1957-12

ABSTRACT

A mathematical theory of secrecy systems is developed. Three main problems are considered. (1) A logical formulation of the problem and a study of the mathematical structure of secrecy systems. (2) The problem of "theoretical secrecy," i.e., can a system be solved given unlimited time and how much material must be intercepted to obtain a unique solution to cryptograms. A secrecy measure called the "equivocation" is defined and its properties developed. (3) The problem of "practical secrecy." How can systems be made difficult to solve, even though a solution is theoretically possible.

THIS DOCUMENT CONTAINS INFORMATION AFFECTING THE NATIONAL DEFENSE OF THE UNITED STATES WITHIN THE MEANING OF THE ESPIONAGE LAWS, TITLE 18, U.S.C., SECTIONS 793 AND 794. ITS TRANSMISSION OR THE REVELATION OF ITS CONTENTS IN ANY MANNER TO AN UNAUTHORIZED PERSON IS PROHIBITED BY LAW.

e.g., Shannon: mathematical theory of communications

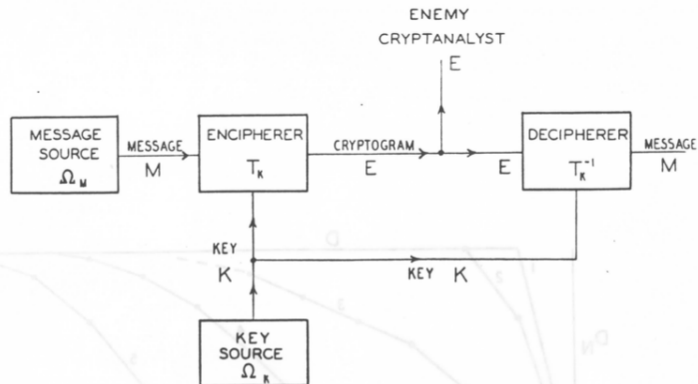


Figure 3: Shannon crypto diagram

1.1 Tukey's FoDA 1962

I. GENERAL CONSIDERATIONS

1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their "dealing with fluctuations" aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

2 (algo)
3., 1.

2 (thy)

Figure 4: opening

- opens with impostor/moving target line

1.1 Tukey's FoDA 1962

I. GENERAL CONSIDERATIONS

1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their "dealing with fluctuations" aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

2 (algo)
3., 1.

2 (thy)

Figure 4: opening

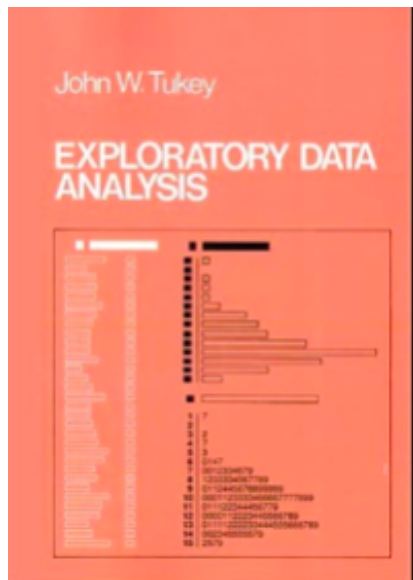
- ▶ opens with impostor/moving target line
- ▶ attack on mathematization

1.1 Tukey's call for data analysis "S" not "M"

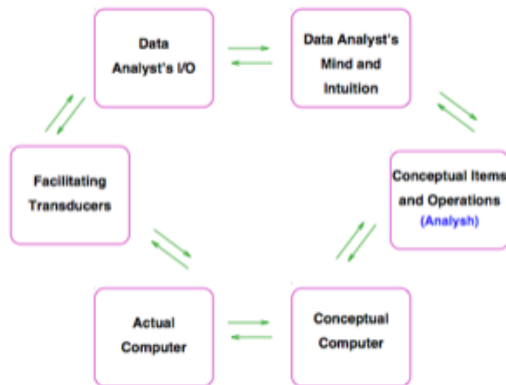
Data analysis, and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics, specifically: (1) Data analysis must seek for scope and usefulness rather than security. (2) Data analysis must be willing to err moderately often in order that inadequate evidence shall more often suggest the right answer. (3) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proofs or stamps of validity.

- ▶ John W. Tukey, "The Future of Data Analysis," *Annals of Mathematical Statistics*, no. 1 (Mar 1962): 1–67, <https://doi.org/10.1214/aoms/1177704711>. 6.

1.2 Tukey's FoDA : aftermath, e.g., EDA 1977



1.2 Tukey's FoDA : aftermath, e.g., proto-R



John W. Tukey (Feb. 3, 1965)

1.2 Tukey's FoDA : aftermath, e.g., GLS 1993

Greater or Lesser Statistics: A Choice for Future Research

John M. Chambers

AT&T Bell Laboratories, Murray Hill, New Jersey

Abstract

The statistics profession faces a choice in its future research between continuing concentration on traditional topics, based largely on data analysis supported by mathematical statistics, and a broader viewpoint, based on an inclusive concept of learning from data. The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal in activities to which it can make important contributions.

1.2 Tukey's FoDA : aftermath, e.g., GLS 1993

Greater statistics: learning from data

Three broad categories characterize work in greater statistics:

- ▶ preparing data, including planning, collection, organization, and validation

1.2 Tukey's FoDA : aftermath, e.g., GLS 1993

Greater statistics: learning from data

Three broad categories characterize work in greater statistics:

- ▶ preparing data, including planning, collection, organization, and validation
- ▶ analyzing data, by models or other summaries

1.2 Tukey's FoDA : aftermath, e.g., GLS 1993

Greater statistics: learning from data

Three broad categories characterize work in greater statistics:

- ▶ preparing data, including planning, collection, organization, and validation
- ▶ analyzing data, by models or other summaries
- ▶ presenting data in written, graphical or other form

1.2 Tukey's FoDA : aftermath, e.g., GLS 1993

Many mundane . . . activities generate large quantities of potentially valuable data. Examples . . . include retail sales, billing, and inventory management. The data were not generated for the purpose of learning; however, the potential for learning is great. . .

The data usually pass through a computer system nowadays, but aside from the enormous quantity the data are typically thrown away fairly quickly. The computational challenge of collecting and organizing such data is huge. A more clearly statistical challenge is that the data may represent only a portion of the conceptually relevant data; if so, the sample is often biased in crucial ways.

1.2 Tukey's FoDA : aftermath, e.g., DS 2001

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

Abstract

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

1.2 Tukey's FoDA : aftermath, e.g., DS 2001

The focus of the plan is the practicing data analyst. One outcome of the plan is that computer science joins mathematics as an area of competency for the field of data science. This enlarges the intellectual foundations. It implies partnerships with computer scientists just as there are now partnerships with mathematicians. The primary agents for change should be university departments themselves. But it is reasonable for departments to look both to university administrators and to funding agencies for resources to assist in bringing about the change.

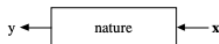
1.2 Tukey's FoDA : aftermath, e.g., 2 cultures 2001

- ▶ Leo Breiman on academic mathematical statistics: “Alice in Wonderland. That is, I knew what was going on out in industry and government in terms of uses of statistics, but what was going on in academic research seemed light years away. It was proceeding as though it were some branch of abstract mathematics.” (in Olsen, 2001)

1.2 Tukey's FoDA : aftermath, e.g., 2 cultures 2001

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables \mathbf{x} (independent variables) go in one side, and on the other side the response variables \mathbf{y} come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;

Information. To extract some information about how nature is associating the response variables to the input variables.

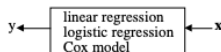
There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = $f(\text{predictor variables, random noise, parameters})$

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

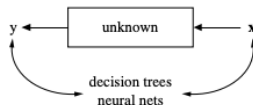


Model validation. Yes—no using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$ —an algorithm that operates on \mathbf{x} to predict the responses \mathbf{y} . Their black box looks like this:



Model validation. Measured by predictive accuracy.

Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the

1.2 Tukey's FoDA : aftermath, stats + math

- ▶ NSF/DMS doubles down on math (1952)

2. *Physical, Mathematical, and Engineering Sciences*. For this report (a) *physical sciences* are those sciences concerned primarily with the understanding of the natural phenomena associated with nonliving things; (b) mathematical sciences are those sciences which employ logical reasoning with the aid of symbols and which are concerned with the development of methods of operations employing such symbols, including mathematics, pure and applied; astronomy, theoretical mechanics, statistics, logistic research, and computer research exclusive of engineering; (c) *engineering sciences* are those sciences which are concerned with studies directed toward making specific scientific principles usable in engineering practice.

NSF on mathematical statistics

- ▶ NSF/DMS doubles down on math (2004)

A Report on the Future of Statistics

Bruce G. Lindsay, Jon Kettenring and David O. Siegmund

Abstract. In May 2002 a workshop was held at the National Science Foundation to discuss the future challenges and opportunities for the statistics community. After the workshop the scientific committee produced an extensive report that described the general consensus of the community. This article is an abridgment of the full report.

Figure 5: From NSF 2004

NSF on mathematical statistics

Would John Tukey agree that [development of statistical models, methods, and related theory] is the core activity of statistics? Given all that I know about John, I doubt it.

than outward. At a time when statistics is beginning to recover from its “overmathematization” in the post World War II years and engage in significant applications in many areas, the report is a step into the past and not into the future.

Figure 6: Breiman on NSF 2004

NSF on mathematical statistics

Statistics as a discipline exists to develop tools for analyzing data. As such, statistics is an engineering discipline and methodology.

Yet:

Statistics has primarily focused on squeezing the maximum amount of information out of limited data. This paradigm is rapidly diminishing in importance and statistics education finds itself out of step with reality.

- ▶ David Madigan and Werner Stuetzle, “[A Report on the Future of Statistics]: Comment,” *Statistical Science* 19, no. 3 (2004): 408.

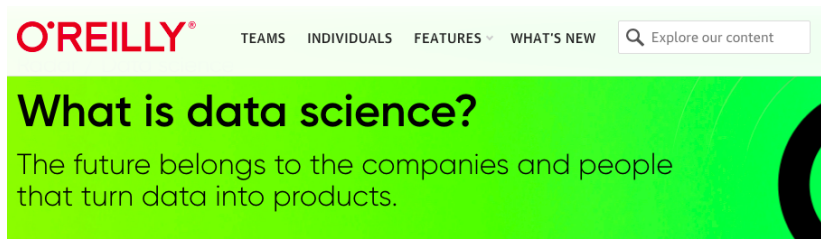
2 the “data science” deluge 2009-2020

2.1 Meanwhile in Web 2.0 (2008->)

At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn't quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization in a clear and concise fashion. To capture the skill set required to perform this multitude of tasks, we created the role of "Data Scientist."

Figure 7: 2009, ex-FB

2.1 Meanwhile in Web 2.0 (2008->)



By [Mike Loukides](#)

June 2, 2010

Figure 8: 2010, ORM (conferences+books)

2.1 Meanwhile in Web 2.0: 2010 infographic

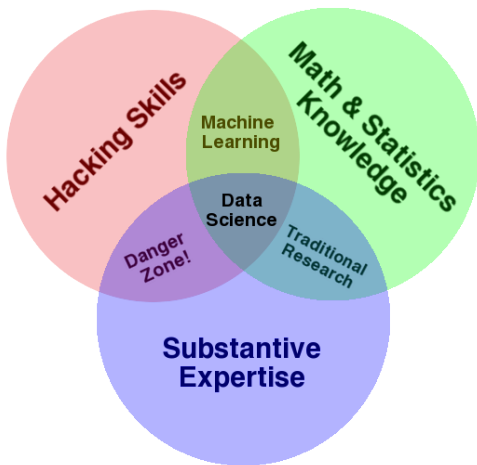


Figure 9: Drew Conway

2.1 Meanwhile in Web 2.0: jobs

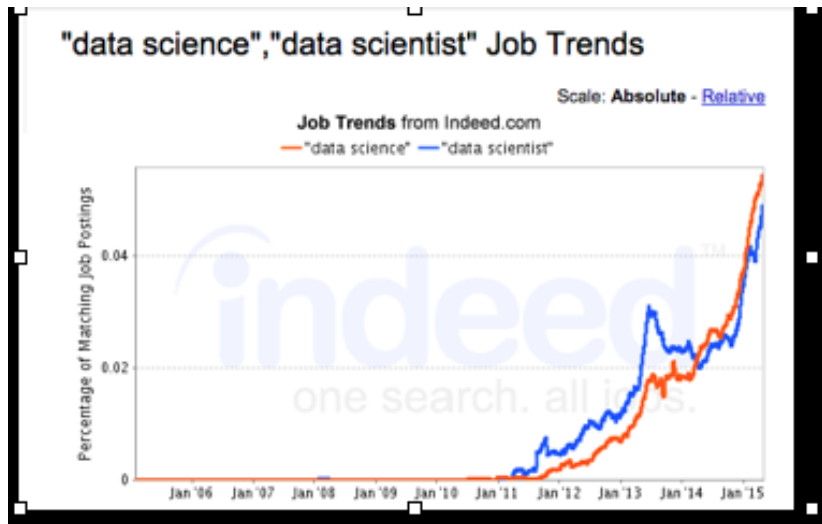


Figure 10: jobs

2.2 Academia, lagging indicator e.g., @ NYU & Columbia

The Center for Data Science Manifesto

Yann LeCun, November 2011, (rev4.1 2012-02-12)

History of this Document

A version of this document was first written in the spring of 2009, and refined over the years until the fall of 2011, in part to alert the NYU administration of the growing importance of data-driven science, and the emergence of the field of data science. In the fall of 2011, NYU launched a University-Wide Initiative in Data Science and Statistics, under the leadership of Gérard Ben Arous, the new director of the Courant Institute and the Vice Provost for Science and Engineering. I was asked to chair a working group to define what NYU should do in this area. The working group was composed of faculty from all over the university, and submitted a report in the summer of 2012 recommending the creation of the multi-disciplinary <http://cds.nyu.edu>, and the creation of a Master of Science in Data Science and a PhD program in Data Science. The launch of the CDS was publicly announced in February 2013, and I was named founding director. The first MSDS students starts in September 2013. The PhD program is pending approval.

Figure 11: Yann's data science manifesto, 2011; CU IDSE 2012

2.2 Academic funding, example of why it “matters”, 2013



The Data Science Environments Partnership includes New York University, the University of California, Berkeley, the University of Washington, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation. The goal of this partnership is to dramatically advance data-intensive scientific discovery, empowering researchers to be vastly more effective by utilizing new methods, new tools, new partnerships, and new career paths. We are accomplishing this via the creation of “Data Science Environments” at the three universities with a five-year \$37.8 million cross-institutional effort supported by the Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation.

2.2 Academia, lagging indicator e.g., Bin Yu 2014

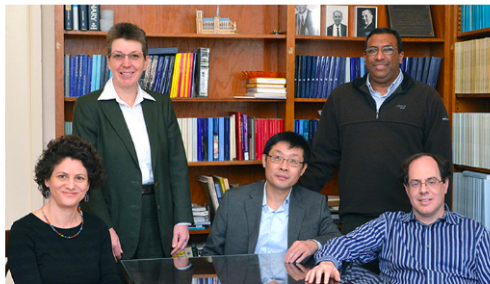
Data Science represents an inevitable (re)-merging of computational and statistical thinking in the big data era. We have to own data science, because domain problems don't differentiate computation from statistics or vice versa, and data science is the new accepted term to deal with a modern data problem in its entirety. Gains for the statistics community are many, and include attracting talent and resources, and securing jobs for our majors, MAs and PhDs.

- ▶ Yu, Let us Own Data Science

2.2 stats->math stats->stats+DS

Introducing DS2 — the future of data science at Yale

By Jim Shelton | MARCH 6, 2017



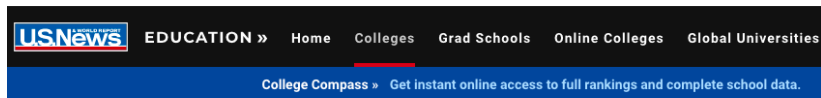
2.2 stats->math stats->stats+DS



October 16, 2017

Carnegie Mellon Changes Statistics Department's Name To Reflect Its Eminent Position in Data Science Research and Education

2.2 Academic “ranking”, example of why it “matters”,



The screenshot shows the top navigation bar of the US News & World Report website. It features the US News logo on the left, followed by a series of navigation links: EDUCATION » (highlighted with a red underline), Home, Colleges, Grad Schools, Online Colleges, and Global Universities. Below these links is a blue banner with the text "College Compass » Get instant online access to full rankings and complete school data."

[Home](#) / [Education](#) / [Colleges](#) / [Best Undergraduate Data Scie...](#)

Best Undergraduate Data Science Programs

Computer Science

3. what is an academic “moving target?”

Recall from G. Gigerenzer:

3.7 THE STATISTICAL PROFESSION: INSTITUTIONS AND INFLUENCE

- ▶ universities: website, Institute, MS, major, PhD, Department

3. what is an academic “moving target?”

Recall from G. Gigerenzer:

3.7 THE STATISTICAL PROFESSION: INSTITUTIONS AND INFLUENCE

- ▶ universities: website, Institute, MS, major, PhD, Department
- ▶ community:

3. what is an academic “moving target?”

Recall from G. Gigerenzer:

3.7 THE STATISTICAL PROFESSION: INSTITUTIONS AND INFLUENCE

- ▶ universities: website, Institute, MS, major, PhD, Department
- ▶ community:
 - ▶ IRL: workshop, conference, annual conference

3. what is an academic “moving target?”

Recall from G. Gigerenzer:

3.7 THE STATISTICAL PROFESSION: INSTITUTIONS AND INFLUENCE

- ▶ universities: website, Institute, MS, major, PhD, Department
- ▶ community:
 - ▶ IRL: workshop, conference, annual conference
 - ▶ publication: special issue, journal, textbooks

3. what is an academic “moving target?”

Recall from G. Gigerenzer:

3.7 THE STATISTICAL PROFESSION: INSTITUTIONS AND INFLUENCE

- ▶ universities: website, Institute, MS, major, PhD, Department
- ▶ community:
 - ▶ IRL: workshop, conference, annual conference
 - ▶ publication: special issue, journal, textbooks
- ▶ funders: special opportunity, regular program, office, division, agency. . .

3. in industry, target moves as well (2018, Reddit): why “da”->“ds”?



16



Posted by u/maxmoo **PhD | ML Engineer | IT** 3 years ago 

Are Data Scientists at Facebook really Data Analysts?

Does anyone know how data science research is structured at Facebook? I have been talking to a couple of recruiters and it seems like their "data scientists" are really SQL data analysts, and their ML engineers are really engineers (mostly working in C++). Do they have something that's more in the middle, like research scientists working on algorithms, but not necessarily stuck in the weeds of implementation details?

Figure 12: Reddit, 2018: why would they?

3. in industry, target moves as well (2018, medium)

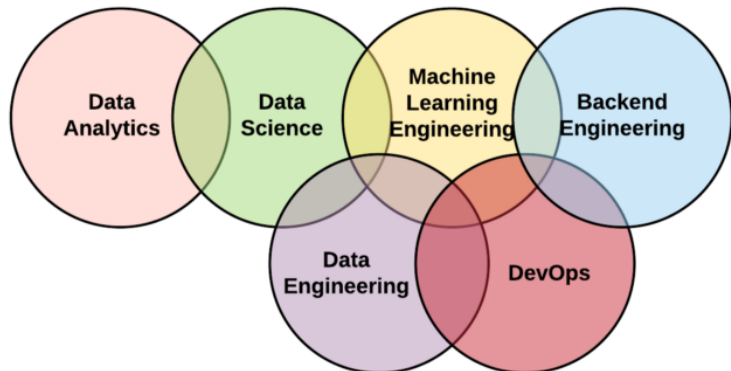


Figure 13: at least 3 of these are in 2009 data scientist via Hammerbacher

zooming out: not just US, not just ML

zooming out: not just US, not just ML

- ▶ the above was a US view, what about elsewhere?

zooming out: not just US, not just ML

- ▶ the above was a US view, what about elsewhere?
- ▶ data science as part of a big tent, with human impacts (Neff)

zooming out: not just US, not just ML

- ▶ the above was a US view, what about elsewhere?
- ▶ data science as part of a big tent, with human impacts (Neff)
 - ▶ more on this next week

Ecologies of instrumental prediction of high-dimension, messy data

► USA

Ecologies of instrumental prediction of high-dimension, messy data

- ▶ USA
- ▶ USSR

Ecologies of instrumental prediction of high-dimension, messy data

- ▶ USA
- ▶ USSR
- ▶ Japan

Ecologies of instrumental prediction of high-dimension, messy data

- ▶ USA
- ▶ USSR
- ▶ Japan
- ▶ France

Ecologies of instrumental prediction of high-dimension, messy data

- ▶ USA
- ▶ USSR
- ▶ Japan
- ▶ France
- ▶ Netherlands . . . *Lots of stories yet tell*

4. data science as “big tent”

4. data science as “big tent”

- ▶ but what else is there than stats or CS? Neff et al.: Critical Data Studies

Neff: bio

► CC'93!

Neff: bio

- ▶ CC'93!
- ▶ PhD ethnography

Neff: abstract, 4 critiques

We summarize four critiques that are commonly made in critical data studies:

1. data are inherently interpretive,

Neff: abstract, 4 critiques

We summarize four critiques that are commonly made in critical data studies:

1. data are inherently interpretive,
2. data are inextricable from context,

Neff: abstract, 4 critiques

We summarize four critiques that are commonly made in critical data studies:

1. data are inherently interpretive,
2. data are inextricable from context,
3. data are mediated through the sociomaterial arrangements that produce them, and

Neff: abstract, 4 critiques

We summarize four critiques that are commonly made in critical data studies:

1. data are inherently interpretive,
2. data are inextricable from context,
3. data are mediated through the sociomaterial arrangements that produce them, and
4. data serve as a medium for the negotiation and communication of values.

Neff: abstract, 4 propositions

1. communication is central to the data science endeavor;

Neff: abstract, 4 propositions

1. communication is central to the data science endeavor;
2. making sense of data is a collective process;

Neff: abstract, 4 propositions

1. communication is central to the data science endeavor;
2. making sense of data is a collective process;
3. data are starting, not end points, and

Neff: abstract, 4 propositions

1. communication is central to the data science endeavor;
2. making sense of data is a collective process;
3. data are starting, not end points, and
4. data are sets of stories.

Neff: abstract, 2 calls to action

1. bringing social scientific and humanistic expertise into data science practice

“more ethical” 3x in abstract

Neff: abstract, 2 calls to action

1. bringing social scientific and humanistic expertise into data science practice
 - ▶ will advance data science and

“more ethical” 3x in abstract

Neff: abstract, 2 calls to action

1. bringing social scientific and humanistic expertise into data science practice
 - ▶ will advance data science and
 - ▶ will advance critical data studies.

“more ethical” 3x in abstract

Neff: abstract, 2 calls to action

1. bringing social scientific and humanistic expertise into data science practice
 - ▶ will advance data science and
 - ▶ will advance critical data studies.
2. leverage the insights from critical data studies to build new kinds of organizational arrangements

“more ethical” 3x in abstract

Neff: abstract, 2 calls to action

1. bringing social scientific and humanistic expertise into data science practice
 - ▶ will advance data science and
 - ▶ will advance critical data studies.
2. leverage the insights from critical data studies to build new kinds of organizational arrangements
 - ▶ will help advance a more ethical data science.

“more ethical” 3x in abstract

Neff: abstract, 2 calls to action

1. bringing social scientific and humanistic expertise into data science practice
 - ▶ will advance data science and
 - ▶ will advance critical data studies.
2. leverage the insights from critical data studies to build new kinds of organizational arrangements
 - ▶ will help advance a more ethical data science.

“more ethical” 3x in abstract

- ▶ ok but what *is* ethical. . . . tune in next week

Neff: messages

- ▶ critiques benefit from understanding process

Neff: messages

- ▶ critiques benefit from understanding process
 - ▶ e.g., not reflecting reflexive data scientists

Neff: messages

- ▶ critiques benefit from understanding process
 - ▶ e.g., not reflecting reflexive data scientists
- ▶ data scientists benefit from critical data studies

Neff: messages

- ▶ critiques benefit from understanding process
 - ▶ e.g., not reflecting reflexive data scientists
- ▶ data scientists benefit from critical data studies
- ▶ “A key insight of *critical data studies* is that interpretation is cooked into the very structures of data and that the work of claiming something as data or a dataset becomes a “rhetorical” move.”

Neff: messages

- ▶ critiques benefit from understanding process
 - ▶ e.g., not reflecting reflexive data scientists
- ▶ data scientists benefit from critical data studies
- ▶ “A key insight of *critical data studies* is that interpretation is cooked into the very structures of data and that the work of claiming something as data or a dataset becomes a “rhetorical” move.”
- ▶ “Data, as a word, although ends up sounding more authoritative than perhaps those who produce it ever intended”

Neff: observations

- ▶ subjective design choices: acknowledge rather than eliminate

Neff: observations

- ▶ subjective design choices: acknowledge rather than eliminate
- ▶ problem choice $>$ problem solution (cf. JWT)

Neff: observations

- ▶ subjective design choices: acknowledge rather than eliminate
- ▶ problem choice $>$ problem solution (cf. JWT)
- ▶ communication is everything

Neff: observations

- ▶ subjective design choices: acknowledge rather than eliminate
- ▶ problem choice $>$ problem solution (cf. JWT)
- ▶ communication is everything
- ▶ prominence of ethics

Neff: observations

- ▶ subjective design choices: acknowledge rather than eliminate
- ▶ problem choice $>$ problem solution (cf. JWT)
- ▶ communication is everything
- ▶ prominence of ethics
 - ▶ ok but what *is* ethics. . . . tune in next week

re-reminders

re-reminder 1: data and truth: labels and ideas

- ▶ are contested (e.g., “AI”)

re-reminder 1: data and truth: labels and ideas

- ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning

re-reminder 1: data and truth: labels and ideas

- ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
- ▶ are dynamic (e.g., “ML”)

re-reminder 1: data and truth: labels and ideas

- ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
- ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011

re-reminder 1: data and truth: labels and ideas

- ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
- ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011
- ▶ are dynamic (e.g., “DS”)

re-reminder 1: data and truth: labels and ideas

- ▶ are contested (e.g., “AI”)
 - ▶ rules vs learning
- ▶ are dynamic (e.g., “ML”)
 - ▶ 1959 vs 1983 vs 2011
- ▶ are dynamic (e.g., “DS”)
 - ▶ 1977,2001,2008,2010,2021...

re-reminder 2: data and power:

- ▶ fields are “moving targets”...

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading
 - ▶ by people w/interests e.g., reputation, resources

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading
 - ▶ by people w/interests e.g., reputation, resources
 - ▶ academia: grants, tuition, philanthropy

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading
 - ▶ by people w/interests e.g., reputation, resources
 - ▶ academia: grants, tuition, philanthropy
 - ▶ industry: recruit / retain talent (at price)

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading
 - ▶ by people w/interests e.g., reputation, resources
 - ▶ academia: grants, tuition, philanthropy
 - ▶ industry: recruit / retain talent (at price)
 - ▶ the above are coupled! (among disciplines, groups, companies, each other)

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading
 - ▶ by people w/interests e.g., reputation, resources
 - ▶ academia: grants, tuition, philanthropy
 - ▶ industry: recruit / retain talent (at price)
 - ▶ the above are coupled! (among disciplines, groups, companies, each other)
- ▶ power ascends via innovation...

re-reminder 2: data and power:

- ▶ fields are “moving targets”...
- ▶ ... because they are moved!
 - ▶ including by pieces like our reading
 - ▶ by people w/interests e.g., reputation, resources
 - ▶ academia: grants, tuition, philanthropy
 - ▶ industry: recruit / retain talent (at price)
 - ▶ the above are coupled! (among disciplines, groups, companies, each other)
- ▶ power ascends via innovation. . .
- ▶ ... power accepted via “ethics”/consent of the stakeholders

Appendix

- ▶ 2021-01-12: intro to course

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI
- ▶ 2021-03-16: big data, old school (1958-1980)

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI
- ▶ 2021-03-16: big data, old school (1958-1980)
- ▶ 2021-03-23: AI2.0

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI
- ▶ 2021-03-16: big data, old school (1958-1980)
- ▶ 2021-03-23: AI2.0
- ▶ 2021-03-30: data science, 1962-present

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI
- ▶ 2021-03-16: big data, old school (1958-1980)
- ▶ 2021-03-23: AI2.0
- ▶ 2021-03-30: data science, 1962-present
- ▶ 2021-04-06: ethics

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI
- ▶ 2021-03-16: big data, old school (1958-1980)
- ▶ 2021-03-23: AI2.0
- ▶ 2021-03-30: data science, 1962-present
- ▶ 2021-04-06: ethics
- ▶ 2021-04-13: present problems: attention economy+VC=dumpsterfire

Appendix

- ▶ 2021-01-12: intro to course
- ▶ 2021-01-19: setting the stakes
- ▶ 2021-01-26: risk and social physics
- ▶ 2021-02-02: statecraft and quantitative racism
- ▶ 2021-02-09: intelligence, causality, and policy
- ▶ 2021-02-16: data gets real: mathematical baptism
- ▶ 2021-02-23: WWII, dawn of digital computation
- ▶ 2021-03-09: birth and death of AI
- ▶ 2021-03-16: big data, old school (1958-1980)
- ▶ 2021-03-23: AI2.0
- ▶ 2021-03-30: data science, 1962-present
- ▶ 2021-04-06: ethics
- ▶ 2021-04-13: present problems: attention economy+VC=dumpsterfire
- ▶ 2021-04-15: future solutions