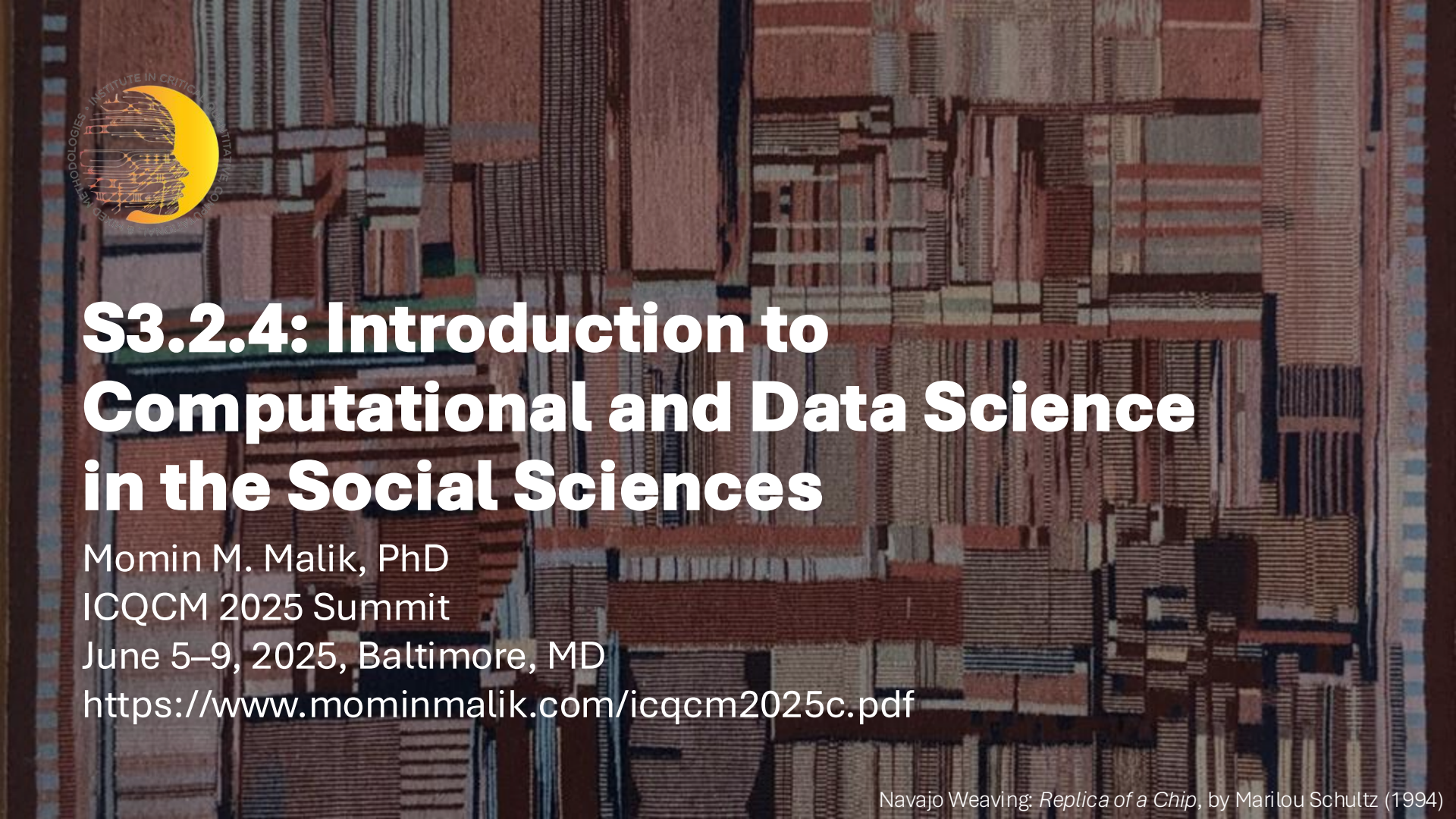


S3.2.4: Introduction to Computational and Data Science in the Social Sciences

Momin M. Malik, PhD
ICQCM 2025 Summit
June 5–9, 2025, Baltimore, MD
https://www.mominmalik.com/icqcm2025c.pdf

Navajo Weaving: *Replica of a Chip*, by Marilou Schultz (1994)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

What would *you* count as a "computational method"?

- Using a calculator?
- Using Excel?
- A linear regression in Excel?
- A linear regression in R?
- A Bayesian hierarchical model in SAS?
- A hierarchical mixture model in MPLUS?
- A regular expression (string matching) script in Python?
- A decision tree in Python?
- A topic model with Latent Dirichlet Allocation in MATLAB?
- A Generalized Additive Mixed Model in R?
- An agent-based simulation in Netlogo?
- Cross-tabs from a billion phone call records?
- Using ChatGPT online?
- Using ChatGPT's API* in Python?

*API = Application Programming Interface

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Outline

- History and nature of machine learning
- Data and machine learning in the social sciences
- What in data science, and what computational methods, are substantive and worth using?

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Learning Goals

- Know what is the actual content of data science, “computational science”, machine learning, and artificial intelligence as compared to (traditional) statistics
- Identify use cases, non-use cases, and weaknesses for these
- Identify anxieties within social science around data and computation, and determine the extent to which they are justified

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

History and nature of machine learning

“So, it’s not real AI?” (Broussard, 2018)

Introduction

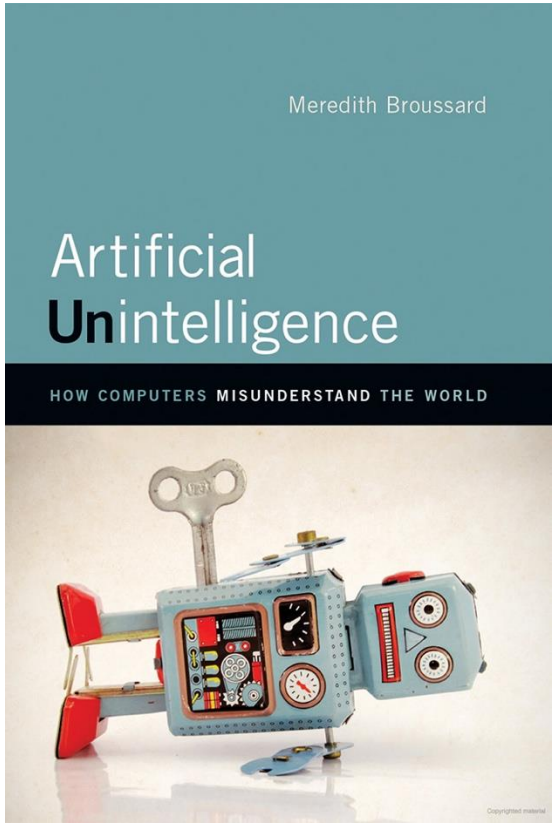
History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References



- “So, it’s not real AI?” he asked.
- “Oh, it’s real,” I said. “And it’s spectacular. But you know, don’t you, that there’s no simulated person inside the machine? Nothing like that exists. It’s computationally impossible.”
- His face fell. “I thought that’s what AI meant,” he said. “I heard about IBM Watson, and the computer that beat the champion at Go, and self-driving cars. I thought they invented real AI.”

Machine learning vs. statistics

Introduction

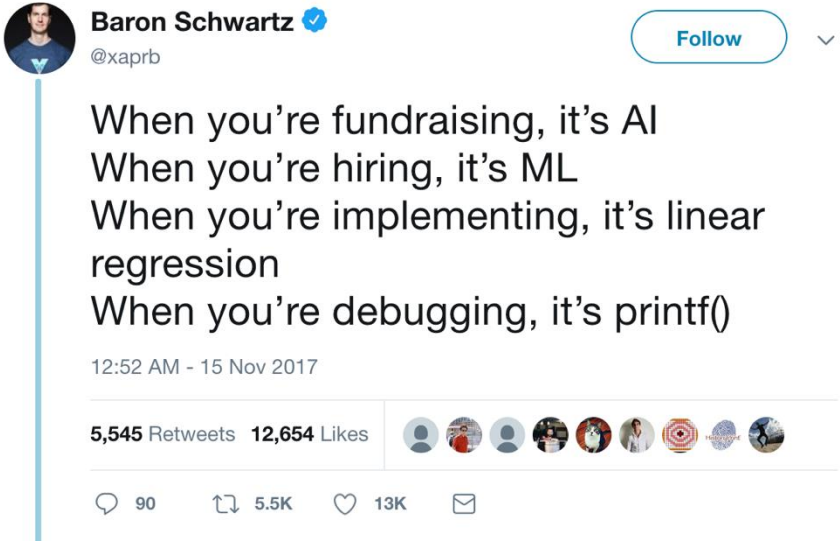
History and nature of machine learning

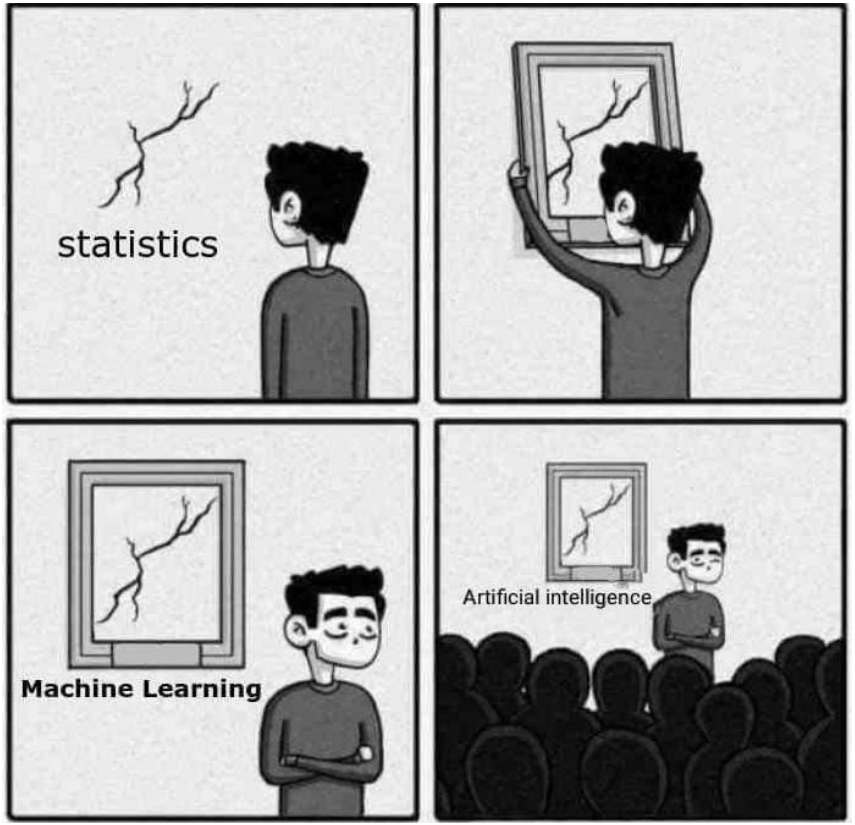
Data and machine learning in the social sciences

What is worth using?

Summary

References





Introduction

History and nature of machine learning

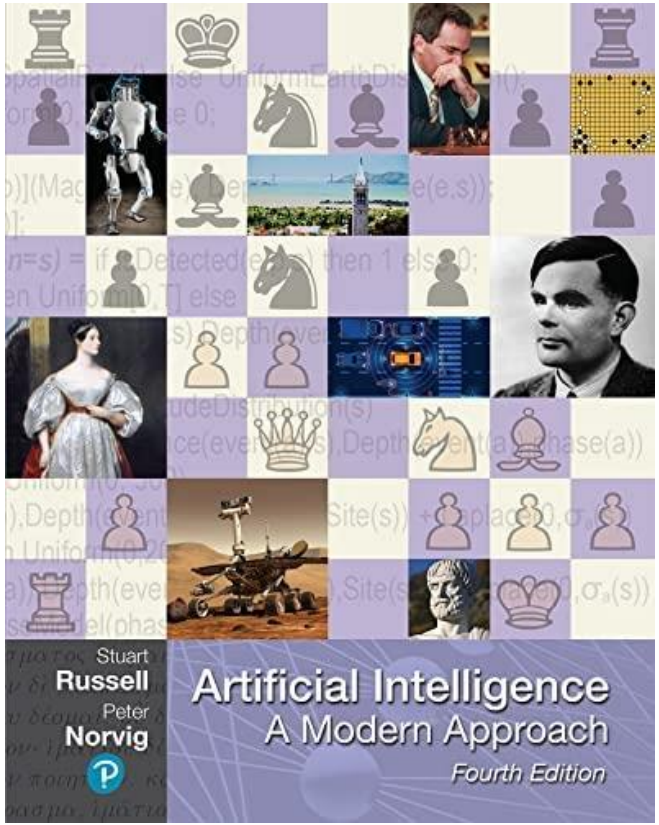
Data and machine learning in the social sciences

What is worth using?

Summary

References

How/why/when? View from AI



"As Steve Abney wrote in 1996, 'In the space of the last ten years, statistical methods have gone from being virtually unknown in computational linguistics to being a fundamental given.'… **after about 14 years of trying to get language models to work using logical rules, I started to adopt probabilistic approaches".**

– Norvig, "On Chomsky", 2010

How/when/why? View from cognitive science

Introduction

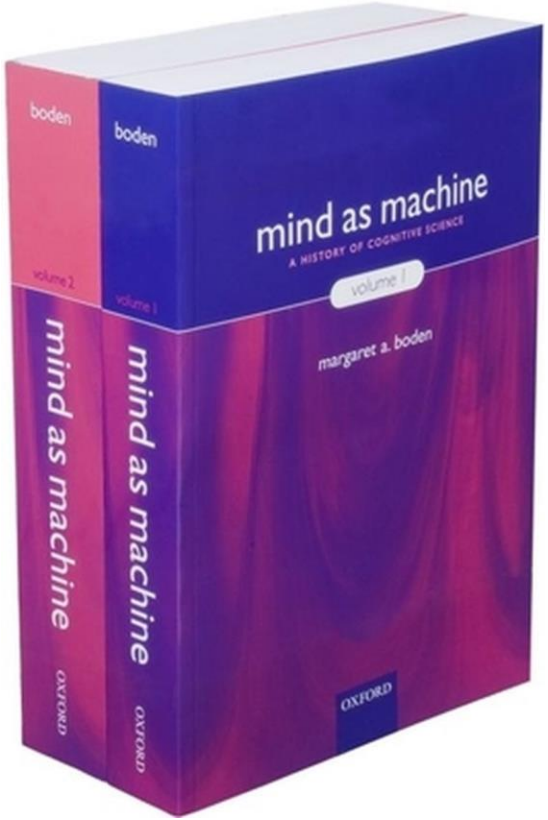
History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References



"1980s–1990s work in machine learning often replayed insights available in traditional statistics… Indeed, it became increasingly clear through the 1990s that **many 'insights' of connectionism were differently named versions of statistical techniques."**

– Boden, *Mind as Machine*, 2006

How/when/why? View from statistics (retrospective)

Introduction

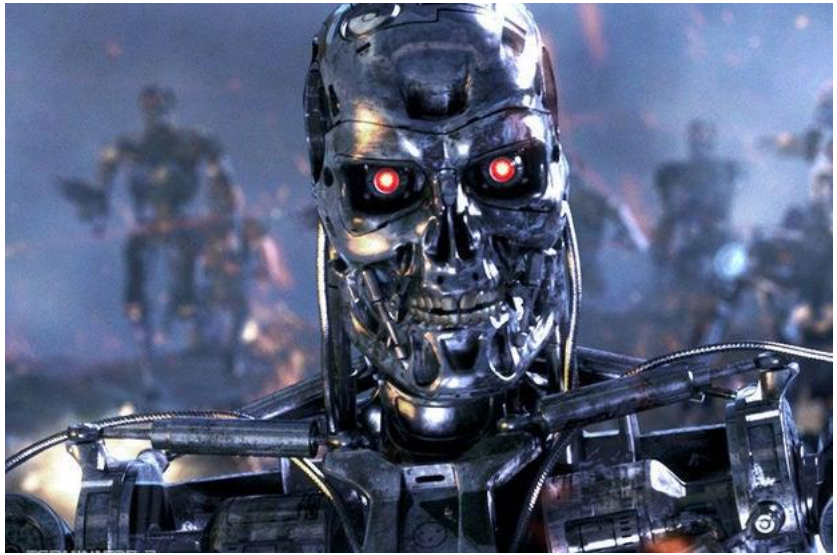
History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References



"At first, ML researchers developed… a collection of rather primitive (yet clever) set of methods to do classification… that eschewed probability. But very quickly they adopted advanced statistical concepts like empirical process theory and concentration of measure. **This transition happened in a matter of a few years."**

– Wasserman, "Rise of the Machines", 2014

Introduction

History and nature of machine learning

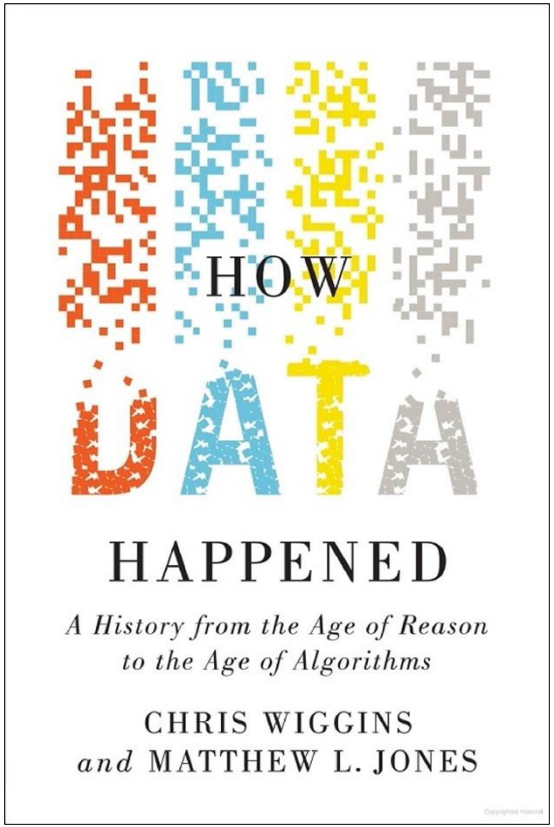
Data and machine learning in the social sciences

What is worth using?

Summary

References

How/when/why? Finally (2023), a perspective from history of science



- "Some fields, like biology, are named after the object of study; others like calculus are named after a methodology. Artificial intelligence and machine learning, however, are named after an aspiration: **the fields are defined by the goal**, not the method used to get there."
- "By the 1960s, practitioners argued, **pattern recognition succeeded in large part because it had abandoned the effort to simulate human perception**: 'Whatever successes we have had… have been the result of an effective transformation of a perception-recognition problem into a 'classification problem.' And pattern recognition researchers cared little about the symbolic side of artificial intelligence."
- "In a moment of profound irony, machine learning, a little-respected relative of artificial intelligence, would come in the new millennium to become the greatest success, even savior of AI, to such an extent that **after 2013 machine learning came largely to displace the far more ambitious goals of traditional AI**, and the terms came to be used interchangeably."

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

A *preemptive* view from statistics

- Leo Breiman (1928–2005) was a statistician who worked outside of academia for 20 years (as a consultant, including for the US military). In that time, he came across (and contributed to) machine learning. He noticed that the "algorithms" of machine learning were actually a style of statistical model (to which he gave the name "algorithmic modeling"); and understanding them *as* models could help illuminate and improve their performance.
- He wrote a landmark paper, published in 2001, to advocate for statisticians to do more "algorithmic modeling."

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Breiman's (2001) prescient diagnosis

"the focus in the statistical community on data models has:

- "Led to irrelevant theory and questionable scientific conclusions;
- "Kept statisticians from using more suitable algorithmic models;
- "Prevented statisticians from working on exciting new problems".

"In the past fifteen years, the growth in algorithmic modeling applications and methodology has been rapid. It has occurred **largely outside statistics in a new community—often called machine learning—that is mostly young computer scientists** (Section 7). The advances, particularly over the last five years, have been startling."

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Breiman's (2001) ominous prognosis

"Perhaps the damaging consequence of the insistence on data models is that **statisticians have ruled themselves out of some of the most interesting and challenging statistical problems that have arisen out of the rapidly increasing ability of computers to store and manipulate data**. These problems are increasingly present in many fields, both scientific and commercial, and solutions are being found by nonstatisticians."

"Over the last ten years, there has been a noticeable move toward statistical work on real world problems and reaching out by statisticians toward collaborative work with other disciplines. I believe this trend will continue and, in fact, has to continue **if we are to survive as an energetic and creative field."**

Introduction

History and nature of machine learning

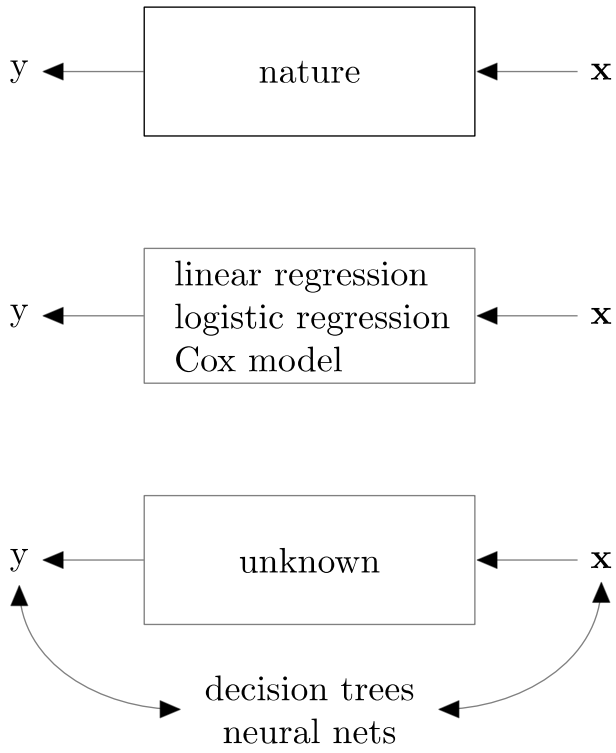
Data and machine learning in the social sciences

What is worth using?

Summary

References

Defining machine learning



- From Breiman, I get a "realist" definition of machine learning: *An instrumental use of correlations to try and mimic the outputs of a target system (rather than trying to understand causal relationships between inputs and outputs)*. Focus on highly flexible "curve fitting" methods. (Diagram: Breiman, 2001. See also Jones, 2018)
- Key: define ML by contrast to *statistics*, not programming.
  - E.g., if we say "learn from examples rather than being programmed with rules", okay, but how? Answer: correlations!!

Instrumentalist (Jones, 2018) orientation of ML

Introduction

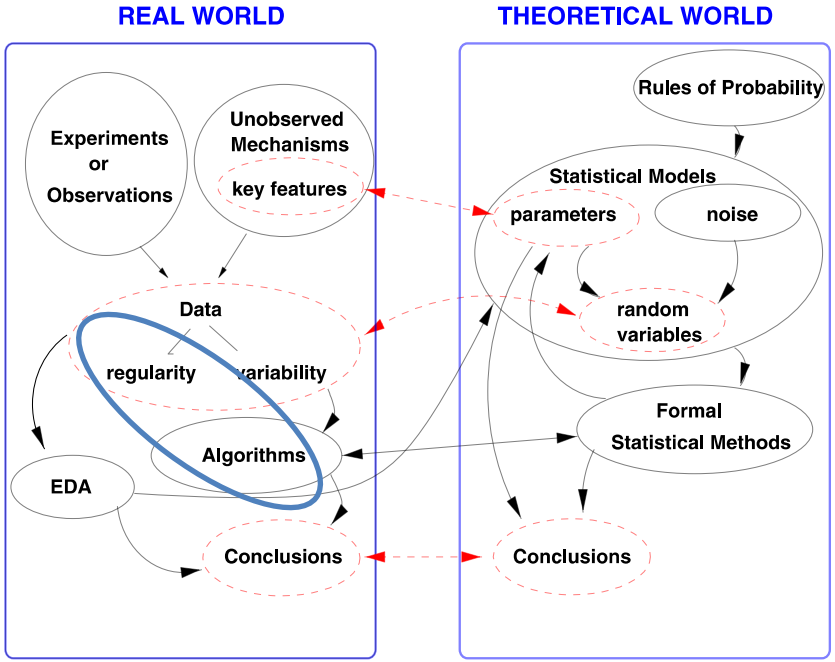
History and nature of machine learning

Data and machine learning in the social sciences

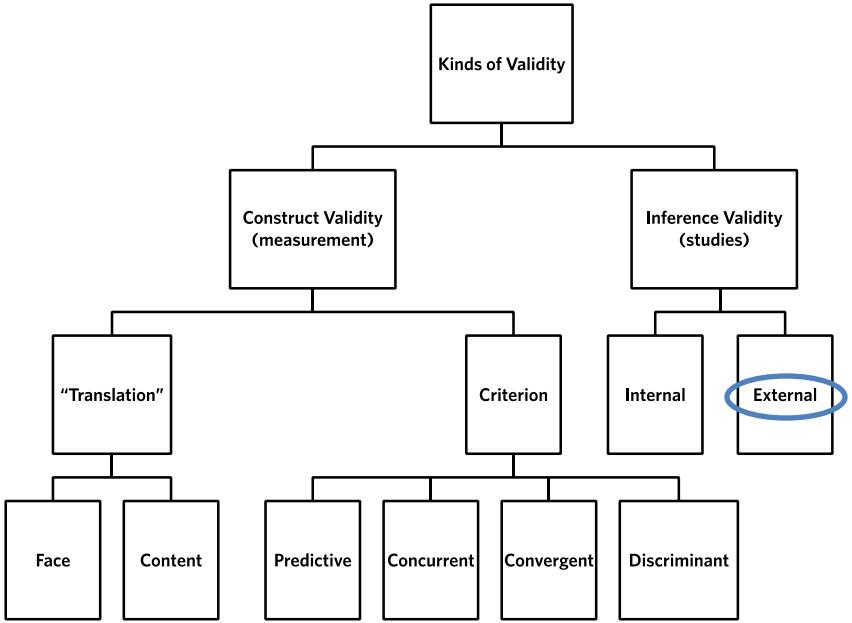
What is worth using?

Summary

References



Kass, 2011



Adapted from Borgatti, 2012

AI/ML as an *illusion of statistics* (using correlations)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Caroline Chan, “Everybody Dance Now: Motion Retargeting Video Subjects.” https://youtu.be/PCBTZh41Ris

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

When are correlations alone sufficient?

- In 2008, in "End of Theory", *Wired*'s Chris Anderson argued that when data are big enough, causality doesn't matter, and correlations alone are sufficient (and we don't need "models").
- He was trivially wrong (see Harford, 2014; Meng, 2018: and anyway, correlation is a model. Also, no amount of data is ever enough, without requiring independence assumptions and/or parametric assumptions; Opsomer et al., 2001)
- But Leo Breiman's paper was a legitimate version of this argument much earlier: *sometimes, for some use cases, correlations alone are sufficient* (and this has nothing to do with the size of the data).

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

ML use cases: Building systems

- Recommend/narrow people's choices to "relevant" ones (friend connections, search results, products)
- Detection (facial, fraud)
- Anticipation (customer demand, equipment failure)
- It "works"...

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

(Large language models and "generative" AI)

- Most machine learning up until now has been targeted, and for a specific purpose. Large language models were done for getting numerical representations of words for downstream tasks
- "Generative" AI software (better called "imitative"), built on large language models (or large "multi-modal" models), does not have a specific purpose. It's not a search engine; it's not "truth"; it's not content; it's just *realistic synthetic text*, and taking it seriously is harmful (Bender, 2024)
- Its outputs are, in principle, a form of `predict()`; but from a model we can't reason from, or even use for a specific purpose
- Not having a specific use case means that there's no clear way to benchmark its performance. On some potential benchmarks (e.g., answering professional licensing test questions) many LLMs do well; on others (e.g., doing math) all LLMs do poorly

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

(Large language models and "generative" AI)

- The software gives an impressive illusion, but what can we *reliably* build on top of this?
  - Is it worth the centralization of power; energy costs; and water consumption? Should we come to rely on a potentially unsustainable business model?
- It's annoying to have to deal with this, because it's more burdensome than exciting from a principled scientific and engineering standpoint
- It's being pushed by marketing, and bought into by non-experts, who are forcing staff scientists and engineers to figure out how to use it as less-badly as possible
- Can't even do reliability testing, because there are constant, drastic changes even within a single sub-version of the software. "…it is impossible to make falsifiable assertions. A system that you cannot debug through a logical, Socratic process is a vulnerability that exploitative tech tycoons will use to do what they always do, undermine the vulnerable" (Dash, 2023)
- I can't wait until the hype collapses, and I can go back to fundamentals of modeling, and dealing with the problems even in purpose-built ML

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References`

Problems even from Breiman

- Breiman did not mention in his article the extensive consulting work he did for the US military (Jones, 2018), which is another major part of the story.
- One of the discussants to his article, Bruce Hoadley, contributed to the development of FICO scores in the 1980s: while these weren't continuous with machine learning, Hoadley's description details how insurance independently came up with things like decision trees and Generalized Additive Models, and were approaching problems in a very machine learning style. Breiman reacted enthusiastically in his rejoinder.
- But insurance is perhaps the most harmful place where correlations (alone) have become totally acceptable and accepted in law and practice (Ochigame, 2020; Kiviat, 2019; Fergus, 2013; Fourcade and Healy, 2013), at least in the US.

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Correlations can go wrong

- Treating people based on correlations denies agency and individuality
- Do we know if a *specific* output is right or wrong?
- Correlations are *proxies*, which can be gamed
- Correlations optimize to the average, leaving out those who are not "average" (as measured!) (Keyes, 2018)
- Mistakes can be unequally distributed across groups
- Correlations are fragile, and can be a poor basis for prediction

Introduction

History and nature of machine learning

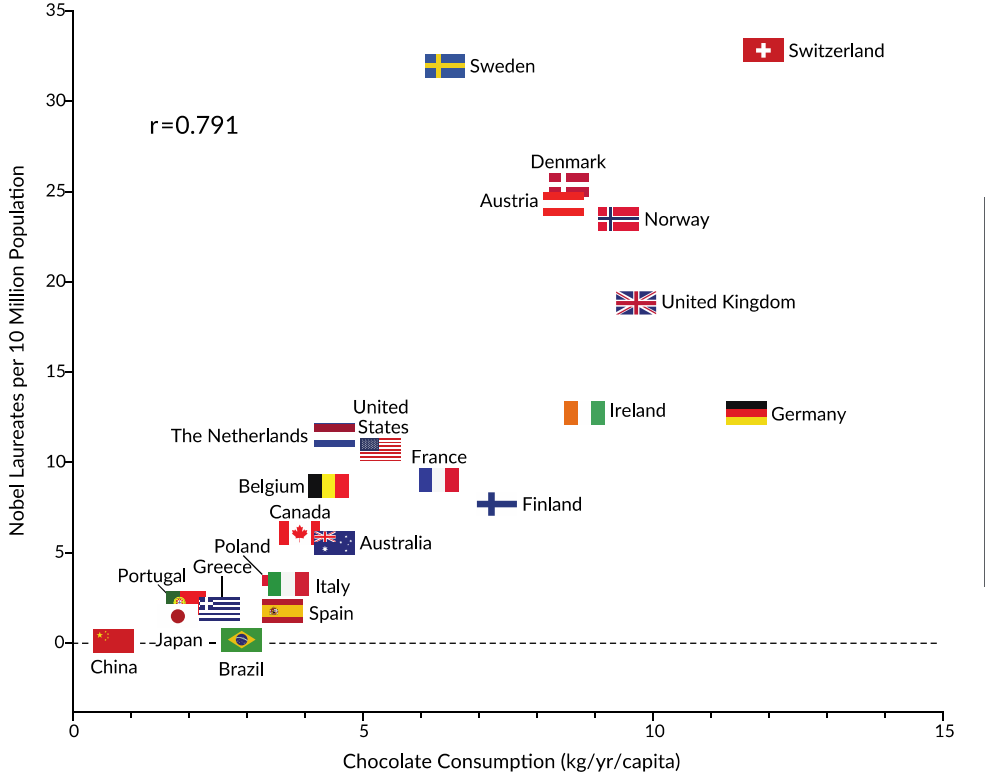
Data and machine learning in the social sciences

What is worth using?

Summary

References

Ex: Chocolate and Nobel prizes?



(Messerli, 2012)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

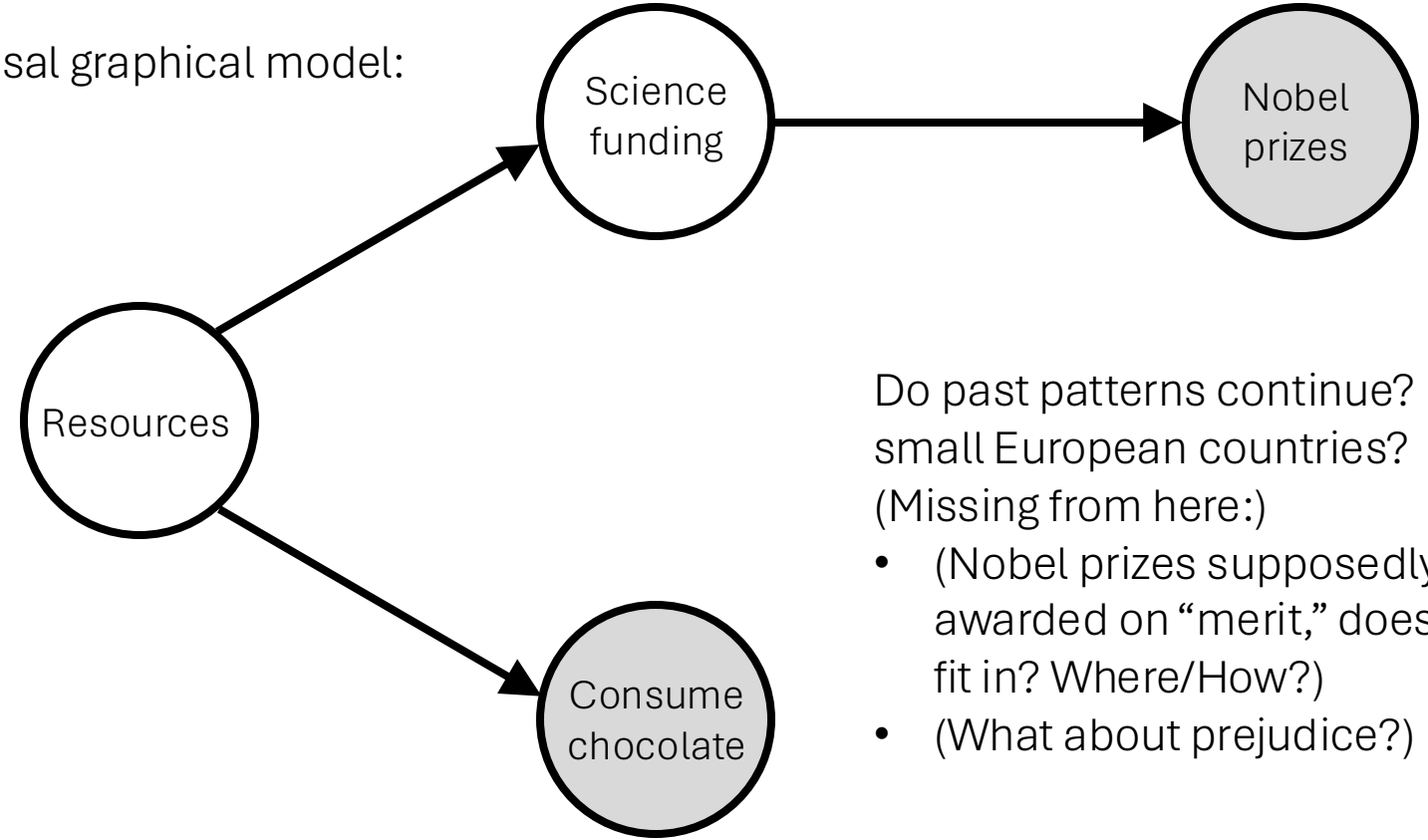
What is worth using?

Summary

References

Correlated, but *cause* is resources

A causal graphical model:



Do past patterns continue? E.g., small European countries? (Missing from here:)

- (Nobel prizes supposedly awarded on "merit," does that fit in? Where/How?)
- (What about prejudice?)

Introduction

History and nature of machine learning

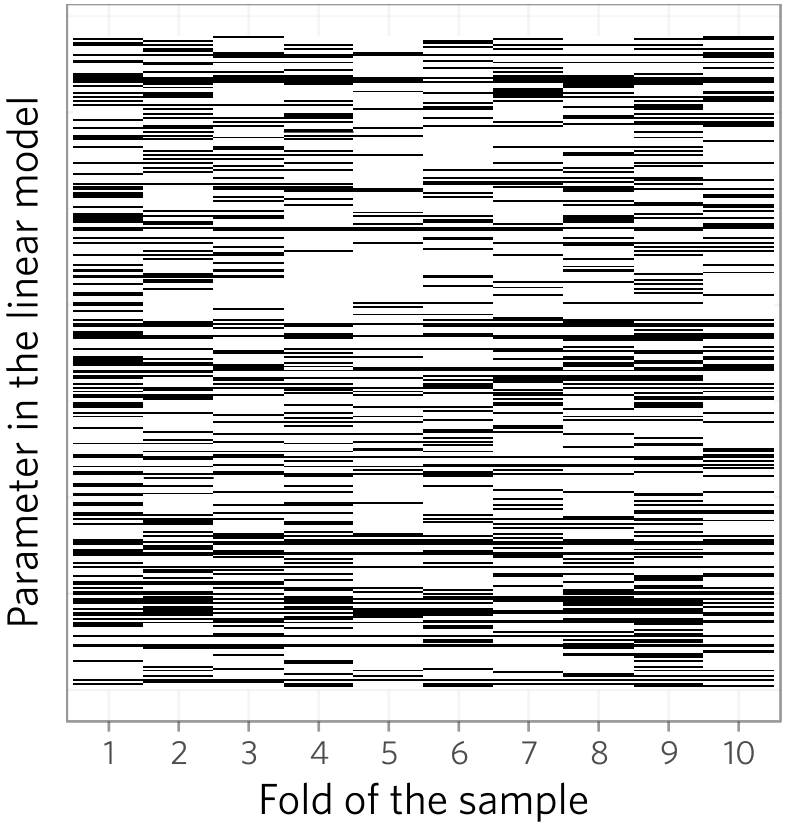
Data and machine learning in the social sciences

What is worth using?

Summary

References

Can't *intervene* based on correlations



- Probably won't win more Nobel prizes by feeding population more chocolate
- Very different sets of correlations can "predict" equally well (Mullainathan & Spiess, 2017)

Introduction

History and nature of machine learning

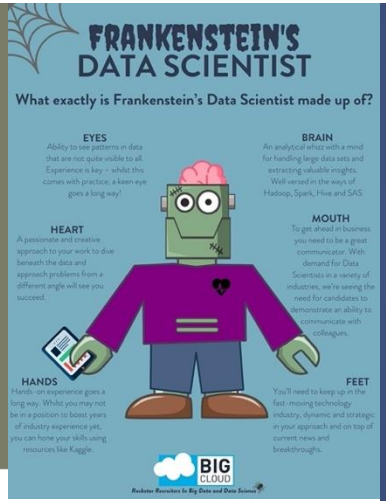
Data and machine learning in the social sciences

What is worth using?

Summary

References

What is “data science”?



- Modern usage attributed to DJ Patil and Jeff Hammerbacher in 2008 (recorded use of term in 1985; John Tukey described something like it in 1962)
  - Aspiration: What statistics could/should have been (consulting, communication, viz, etc.)
- Practically: applied machine learning and a bit of statistics, mostly in business
- Statistics is increasingly rebranding as “statistics and data science” (from 2014 onwards)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Data and machine learning in the social sciences

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

“The coming crisis of empirical sociology” (2007)

“In 2004 when [Savage] attended the ESRC Research Methods festival… he was enrolled in a session designed to popularize social network methods. He talked about an ESRC-funded research project which mapped the personal connections and ties of members of three voluntary organizations using social network analysis. **The project had proved time consuming and intensive.** A lot of time had been spent finding three organizations prepared to participate, a postal questionnaire had been sent to 320 members in total, with a very high response rate. Many members had been interviewed face-to-face to ask detailed questions about their social networks. Thirty life histories had been conducted. **The resulting intensive study of the members’ social ties was amongst the most detailed ever carried out in the UK** (see Ray et al., 2003; Warde et al., 2005).

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

“The coming crisis of empirical sociology” (2007)

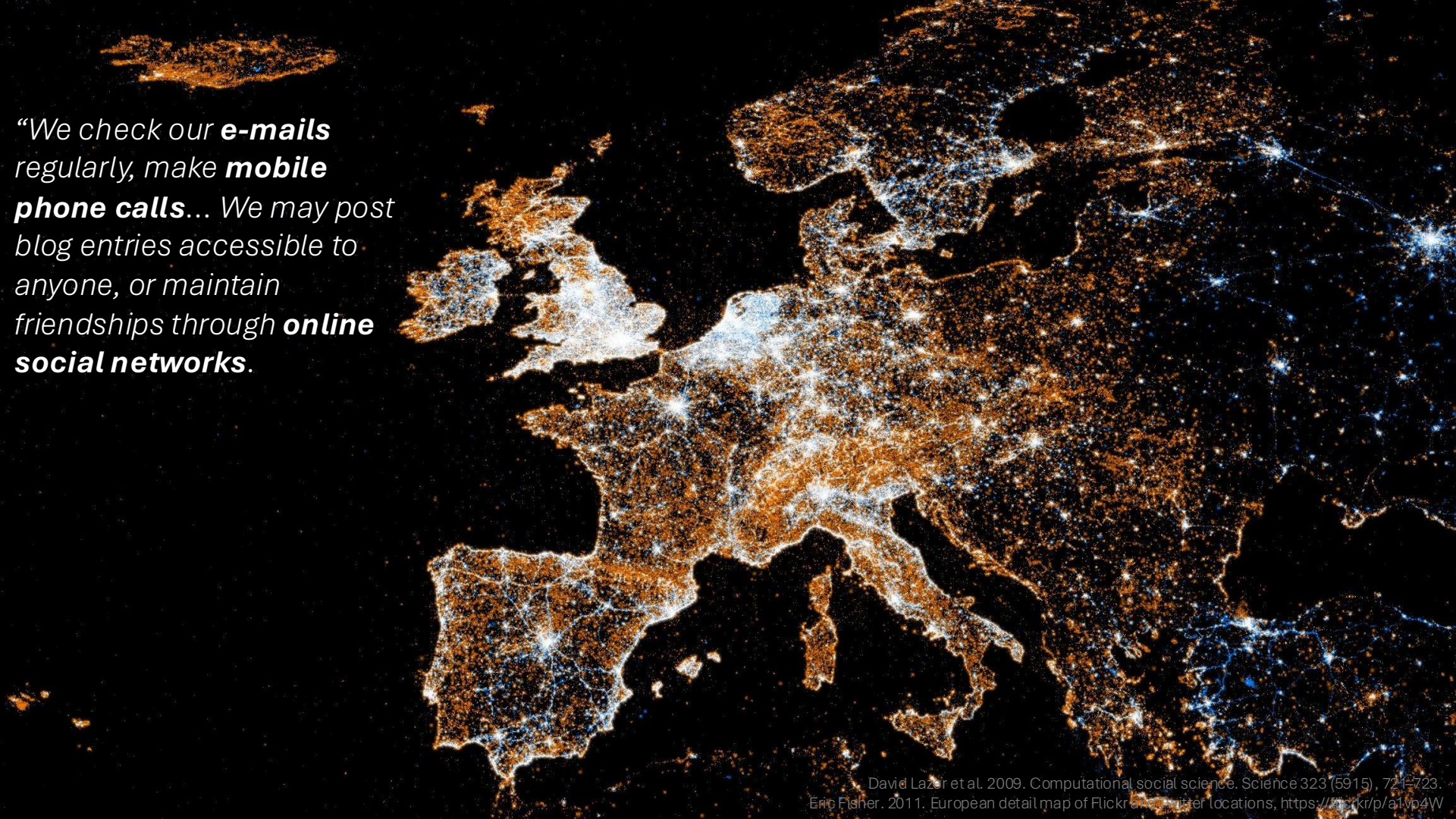
“During the Festival Savage talked to other participants interested in social network methods. It turned out that **one enthusiast was not an academic but worked in a research unit attached to a leading telecommunications company**. When asked what data he used for his social network studies, **he shyly replied that he had the entire records of every phone call made on his system over several years, amounting to several billion ties**. This is data which dwarves anything that an academic social scientist could garner. Crucially, it was data that did not require a special effort to collect, but was the digital by-product of the routine operations of a large capitalist institution. It is also private data to which most academics have no access. To be sure, we can cavil about its limits. It does not tell us what the callers actually talked about. We can emphasize our superior reflexivity, theoretical sophistication, or critical edge. Fair enough – up to a point. Yet the danger is that this response involves taking refuge in the reassurance of our own internal world, our own assumed abilities to be more ‘sophisticated’, and thereby we chose to ignore the huge swathes of ‘social data’ that now proliferate.”

David Lazer et al. 2009. Computational social science. Science 323 (5915) , 721-723.
Eric Fisher. 2011. European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*"We check our **e-mails** regularly, make **mobile phone calls**...*

David Lazer et al. 2009. Computational social science. Science 323 (5915), 721-723.
Eric Fisher. 2011. European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*"We check our **e-mails** regularly, make **mobile phone calls**... We may post blog entries accessible to anyone, or maintain friendships through **online social networks**.*



David Lazer et al. 2009. Computational social science. Science 323 (5915), 721-723.
Eric Fisher. 2011. European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*"We check our **e-mails** regularly, make **mobile phone calls**... We may post blog entries accessible to anyone, or maintain friendships through **online social networks**. Each of these transactions leaves **digital traces** that can be compiled into comprehensive pictures of both individual and group behavior,*

David Lazer et al. 2009. Computational social science. Science 323 (5915), 721-723.
Eric Fisher. 2011. European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*"We check our **e-mails** regularly, make **mobile phone calls**... We may post blog entries accessible to anyone, or maintain friendships through **online social networks**. Each of these transactions leaves **digital traces** that can be compiled into comprehensive pictures of both individual and group behavior, with the **potential to transform our understanding of our lives, organizations, and societies.**"*

David Lazer et al. 2009. Computational social science. Science 323 (5915), 721-723.
Eric Fisher. 2011. European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

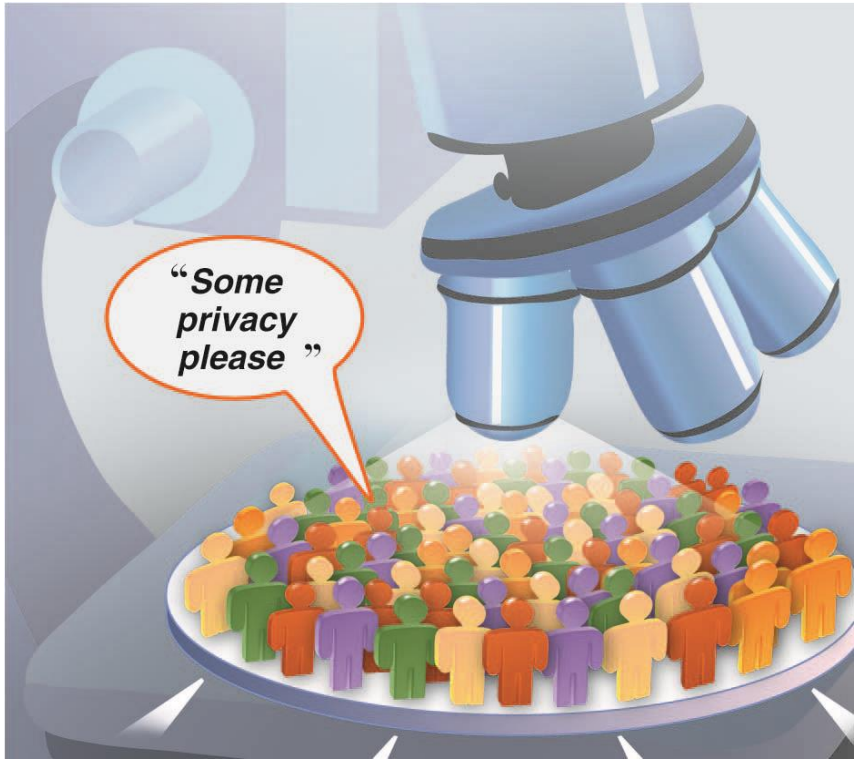
What is worth using?

Summary

References

A “microscope” for social science?

“Disciplines are revolutionized by the development of novel tools: the telescope for astronomers, the **microscope for biologists**, the particle accelerator for physicists, and brain imaging for cognitive psychologists. **Social media provide a high-powered lens into the details of human behavior and social interaction** that may prove to be equally transformative.”



Golder and Macy (2012)

King (2011)

Cells *described* in 1665; cell *theory* in 1830s!

Introduction

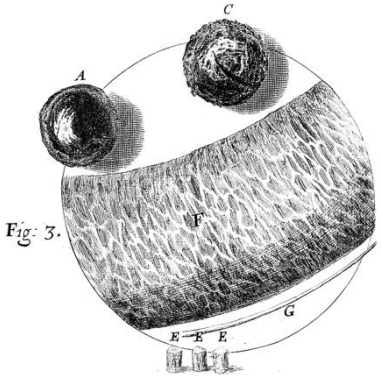
History and nature of machine learning

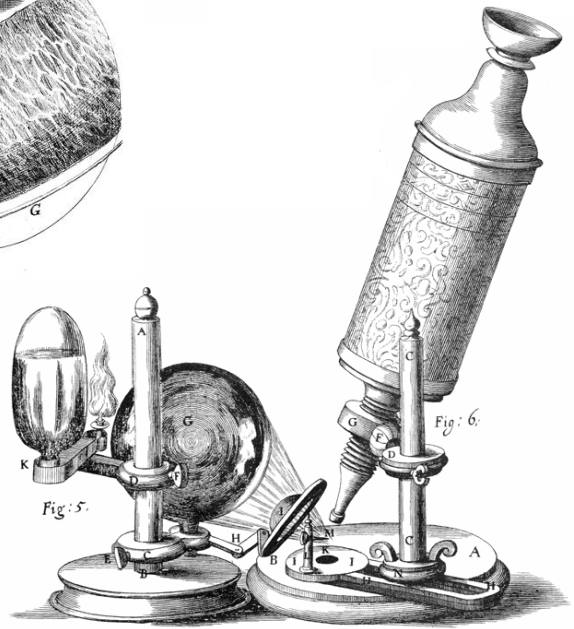
Data and machine learning in the social sciences

What is worth using?

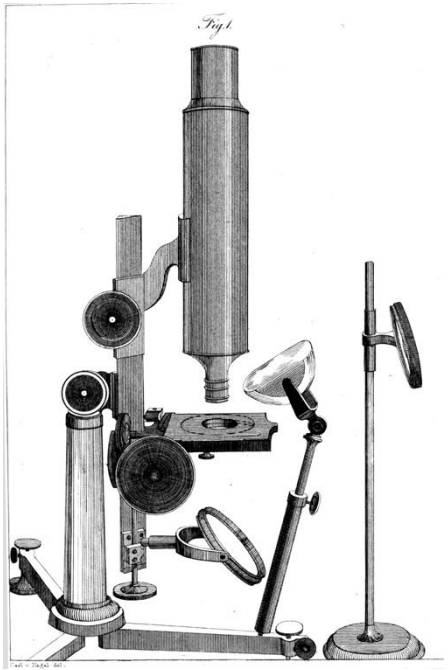
Summary

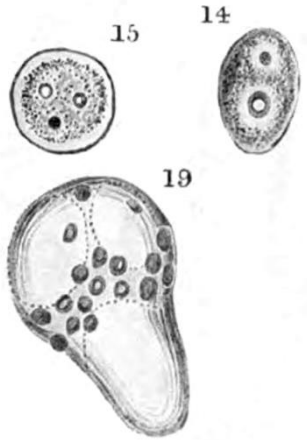
References





Robert Hooke (1665). *Micrographia: or some phyfiological defcriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*.





Theodor Schwann (1839). *Mikroskopische Untersuchungern uber die Uebereinstimmung in der Stuktur und dem wachsthum der Thiere und Pfanzen.* https://wellcomecollection.org/works/mjpkz6zb.
Joseph Berres (1837). *Anatomie der mikroskopischen Gebilde des menschlichen Körpers*.

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

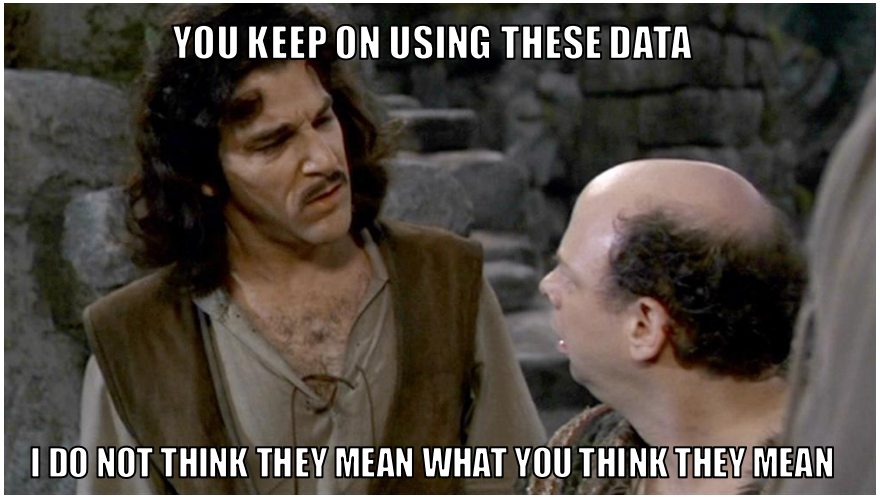
What is worth using?

Summary

References

What happened to social science big data?

- I did my dissertation (2018) critiquing claim-making with digital trace data; and from what I've seen, it never went past a bunch of lofty claims (and on the other side, anxiety), or bragging about this or that paper that never really impress me.
- There's no major findings or studies I can point to that are accepted as furthering social science. Are arguments that nothing happened (Maxmen, 2019). Also now things are just swamped out by hype around AI and LLMs
- Social Science One (project to distribute Facebook data) failed



Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

ML: Fragile Families Challenge

- What about machine learning?
- Matt Salganik organized a "common task" challenge around the Fragile Families dataset
- Tried to make machine learning models for "predicting" life outcomes: material hardship, GPA, grit, eviction, job training, and layoff
- Out of 160 teams, even the "winning" models had an R^2 of 0.2 for material hardship and GPA, and close to 0 for everything else (Salganik et al., 2020)
- R^2 (or any other retrospective goodness-of-fit) doesn't capture causality; but ML couldn't even get a high R^2!

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

What is worth using?

Introduction

History and nature of machine learning

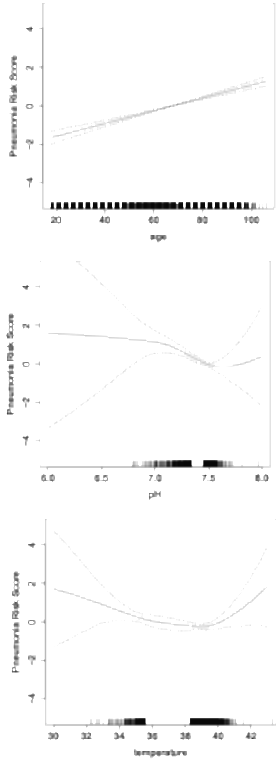
Data and machine learning in the social sciences

What is worth using?

Summary

References

Nonparametrics



- Note: even most “nonlinear regressions” are “linear models”
- *Nonparametrics* (which are still linear models) are what are fancy: No predetermined functional form (but still predetermined constraints)
- The simplest examples: histograms and LOWESS curves. But standard for an actual model: “spline smoothing.”
- The “curse of dimensionality” means that nonparametrics aren’t feasible without assuming independence between variables
- “Generalized additive modeling” (GAM) is a separate nonparametric fit (usually splines) for each covariate. Big achievement for modern statistics. (Hierarchical version: GAMMs). But like histograms and LOWESS, for splines and in GAMs, **at the end of the day, you’re looking at plots; there are no coefficients to interpret**

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Nonparametrics

- For really good but not very informative fits: random forests, or other ensembles of trees (XGBoost) are the best off-the-shelf classifier for tabular data (Caruana et al., 2008; Fernández-Delgado et al., 2014).
- Neural networks only matter for certain types of "unstructured" data

Image: Random Forests, in UC Business Analytics R Programming Guide, via University of Cincinnati. https://uc-r.github.io/random_forests

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Semiparametrics

- This is partitioning out your problem into a parametric part, such as inferring a causal treatment effect, and a nonparametric part, such as calculating propensity scores (rather than using more parametric models for propensity scores)
- Examples: Targeted Maximum Likelihood Estimation (next slide), Double Machine Learning. Random forest a common choice for ML part
- Matching-based methods are not robust to omitted variable bias from unobserved confounders (Arceneux, Gerber, & Green, 2010); but if you care about causality, this may better than not trying to control for selection

Introduction

History and nature of machine learning

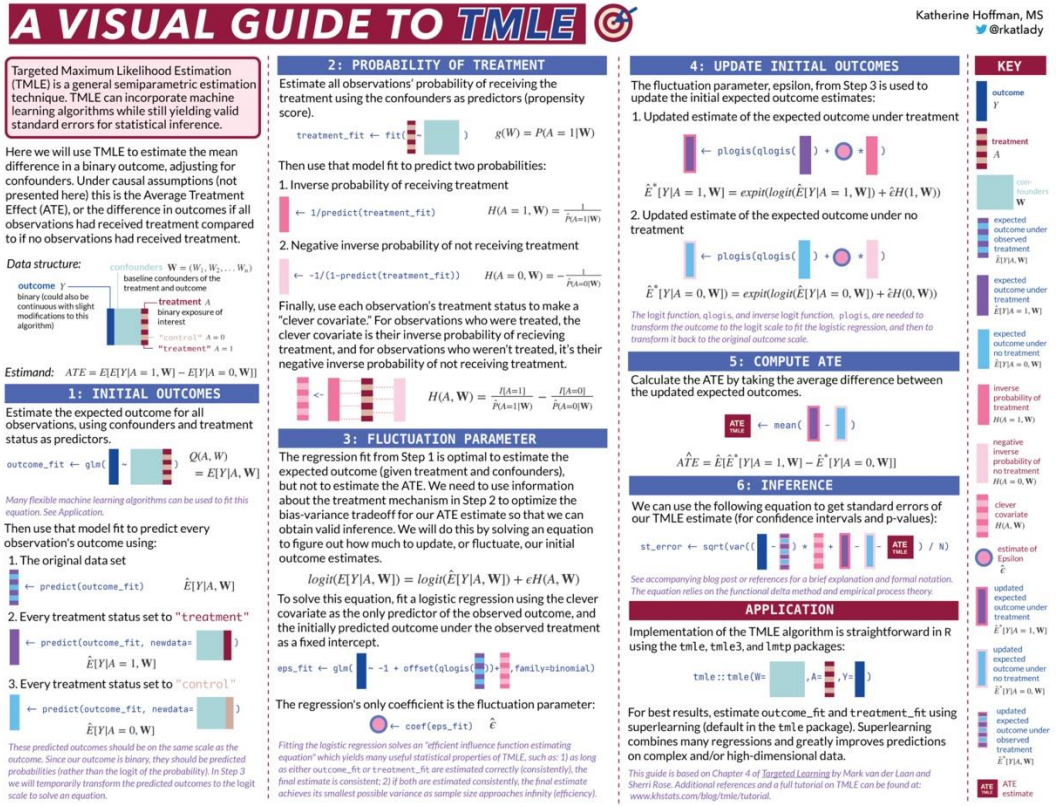
Data and machine learning in the social sciences

What is worth using?

Summary

References

Targeted Maximum Likelihood Estimation



TMLE is an example of what are called semiparametrics: use a random forest to estimate propensity scores, then have a parametric model to estimate Average Treatment Effect

Hoffman, Kat. 2020. "Visual Guides for Causal Inference." *KHstats*. https://www.khstats.com/art/illustrations_viz

Introduction

History and nature of machine learning

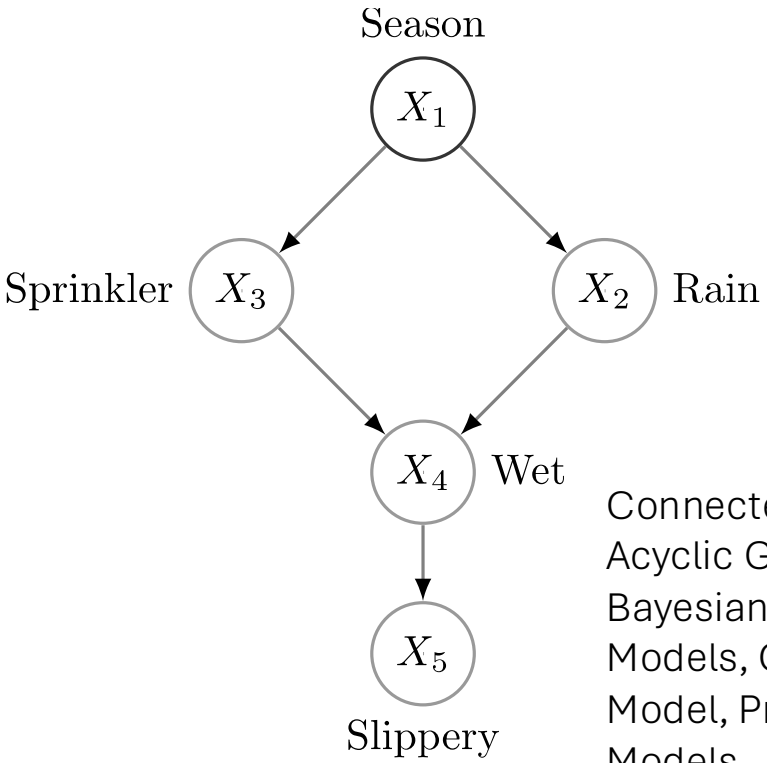
Data and machine learning in the social sciences

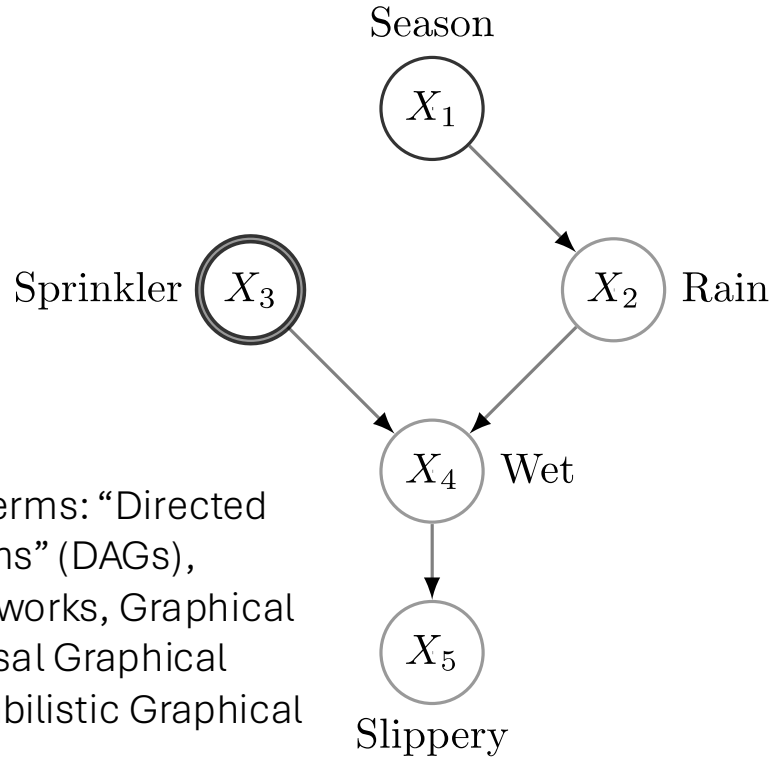
What is worth using?

Summary

References

Structural causal models (SCM)





Connected terms: "Directed Acyclic Graphs" (DAGs), Bayesian networks, Graphical Models, Causal Graphical Model, Probabilistic Graphical Models

Pearl, 2009

SCMs for identifying causal "estimand" (Lundberg et al., 2021)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Table 1. Unit-Specific Quantities Defined in Potential Outcomes Unlock Many Causal Estimands for Inquiry

Estimand name	Mathematical statement	DAG	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n}\sum_i \left(Y_i(d') - Y_i(d)\right)$	$D \longrightarrow Y$	Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x}\sum_{i:X_i=x} \left(Y_i(d') - Y_i(d)\right)$	$X \longrightarrow D \longrightarrow Y$	Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n}\sum_i \Big(\left(Y_i(a',d') - Y_i(a',d)\right) - \left(Y_i(a,d') - Y_i(a,d)\right)\Big)$	$A \to Y \leftarrow D$	Vanderweele (2015)	Joint treatment effect
Controlled direct effect	$\frac{1}{n}\sum_i \left(Y_i(d',m) - Y_i(d,m)\right)$	$D \to Y$, $M \to Y$	Acharya et al. (2016)	Mediation (Illustrations: Example 2)
Natural direct effect	$\frac{1}{n}\sum_i \Big(Y_i(d', M_i(d)) - Y_i(d, M_i(d))\Big)$	$D \to Y$, $M \to Y$	Imai et al. (2011)	Mediation (Part B of the Online Supplement)
Effect of time-varying treatment	$\frac{1}{n}\sum_i \left(Y_i(d'_1,d'_2) - Y_i(d_1,d_2)\right)$	$D_1 \to D_2 \to Y$	Wodtke et al. (2011)	Cumulative effect

Note: Social scientists who define the research goal before moving to regression uncover more possible questions than those who confine themselves to regression parameters. The table provides a non-exhaustive list of common causal estimands. The mathematical statement of each estimand involves counterfactuals—potential outcomes under unobserved treatment assignments—and is the parameter the quantitative analysis would hope to estimate. The DAG depicts one potential set of identification assumptions to link unobservable quantities to observable data. Y indicates the outcome, D indicates the treatment, M indicates a mediator, X indicates pre-treatment covariates, capital letters indicate random variables, and lowercase letters indicate fixed values. Controlled direct effects and other mediation-based estimands appear in sociology, although not always labeled as such (see Part B of the Online Supplement).

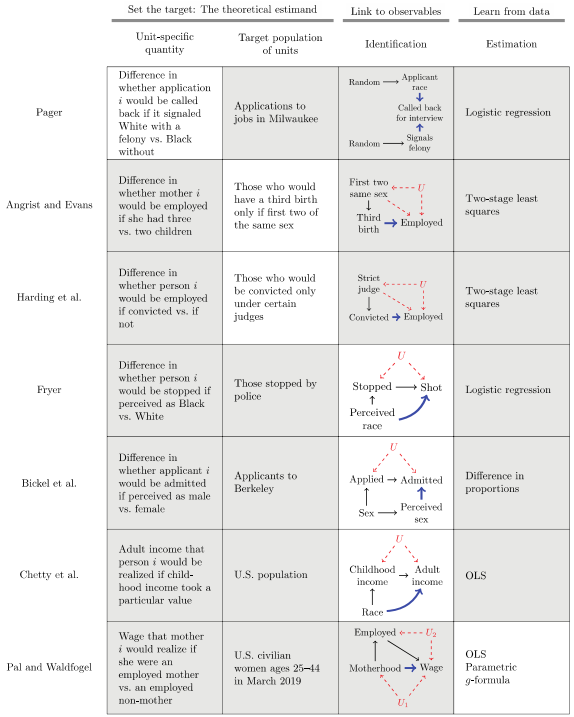


Figure 2. Estimands Are Relevant to a Broad Range of Social Science Studies
Note: White boxes on the diagonal are the focus of the main text, but every study implicitly involves all four steps. Some steps (e.g., DAGs for identification) are simplified to fit in the table. In the identification step, thick arrows represent the causal effect at the center of the paper and dashed edges represent threats to identification.

SCMs for articulating disputes and misleading conclusions (Ibid.)

Introduction

History and nature of machine learning

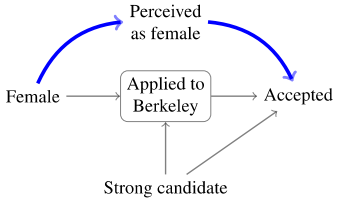
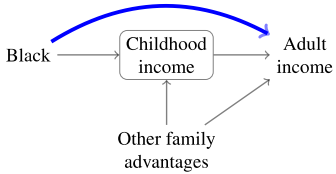
Data and machine learning in the social sciences

What is worth using?

Summary

References

Table 2. Empirical Regularities Can Be Misleading without Estimands

Study	Empirical Regularity	Misleading Conclusion	Directed Acyclic Graph
Fryer (2019)	Among those they stop, police shoot the same proportion of Black individuals as White individuals.	Police do not discriminate against Black individuals when using lethal force.	
Bickel et al. (1975)	Among those who apply, Berkeley departments admit a higher proportion of women than of men.	Admissions committees do not discriminate against women.	
Chetty et al. (2020)	Among those with equal childhood incomes, Black and White women earn similar amounts as adults.	Equalizing childhood incomes would eliminate the racial gap in women's adult incomes.	

"Note: Each example reports an empirical regularity with a vague connection to a theoretical claim. The empirical regularity supports the misleading conclusion only under identification assumptions that the node at the bottom of each Directed Acyclic Graph (DAG; Pearl 2009) does not affect both the variable that the researchers hold constant (boxed) and the outcome (at right). We draw the Fryer (2019) example from a critique by Knox and colleagues (2020) that highlights this and other issues with the original paper. In the first row, **equal use of lethal force against Black individuals stopped by police may stem from the fact that** being stopped is a collider: **among those stopped, the behavior of Black individuals is likely to be less dangerous**. In the second row, **equal or higher acceptance rates among female candidates who apply to Berkeley could result because** applying to Berkeley is a collider: **among women, only the strong candidates apply**. In the third row, childhood income is a collider: **Black families who overcome discrimination to attain incomes comparable to those of White families likely have other advantages that may contribute to their children's incomes in adulthood**. When we state the theoretical and empirical estimands, the DAG makes clear they are not equal and thus **the descriptive quantity does not support the conclusions drawn**."

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

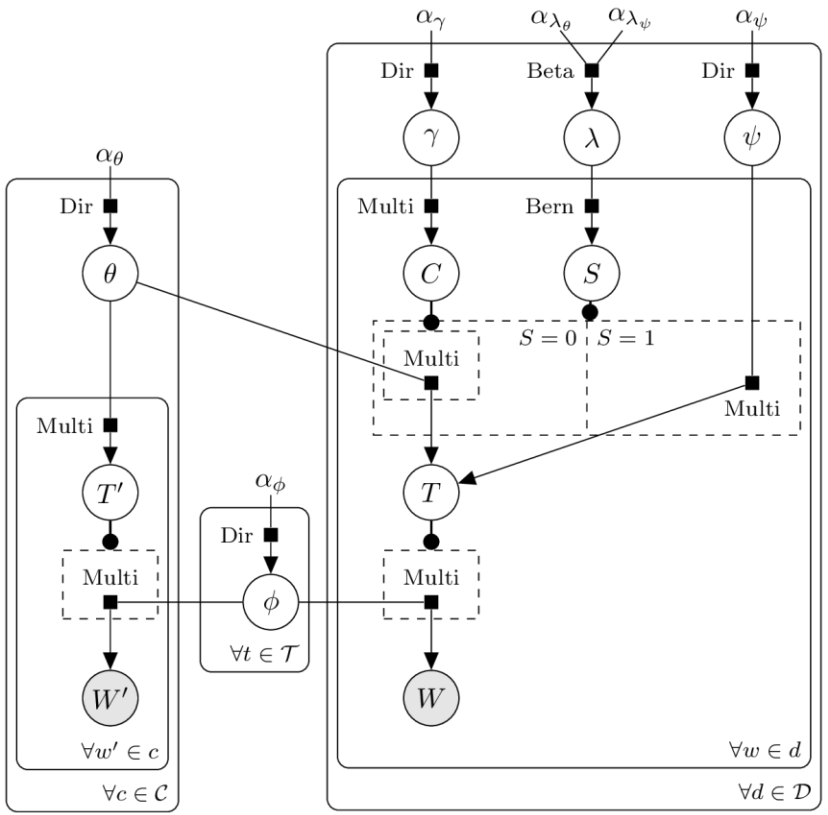
What is worth using?

Summary

References

Exploration via [Structural] Topic Models (STM)

- "Latent Dirichlet Allocation" in a graphical model, but not a causal one, for clustering (calculating latent classes) in textual data
- STMs allows including covariates
- Ultimately an exploratory technique; clusters require interpretation, and it's easy to interpret pure noise; and is based on co-occurrence, so can only find patterns that appear as co-occurrences



Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Variable selection for exploration

- Traditional statistics never came up with good methods for model/variable/feature selection. Stepwise regression was never really a good idea
- Model/Variable/Feature selection works well for “prediction”, but it can also be used as an exploratory step
- E.g., fit a random forest, then look at variable importance. Or fit a lasso, and see what is selected in. BUT IT IS NOT CAUSAL, nor is it even the only set of variables that fit equally well (Mullainathan & Spiess, 2017)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Scaling (human) labels

- Have your human coders (annotators, labelers) label a small part of a corpus, then using ML models based on correlations with word co-occurrence frequency to scale up those labels. All that matters is the quality of the final label, which you can check
- Can propagate standard errors using technique in “prediction-powered inference” (Angelopoulos et al., 2023)

Introduction

History and nature of machine learning

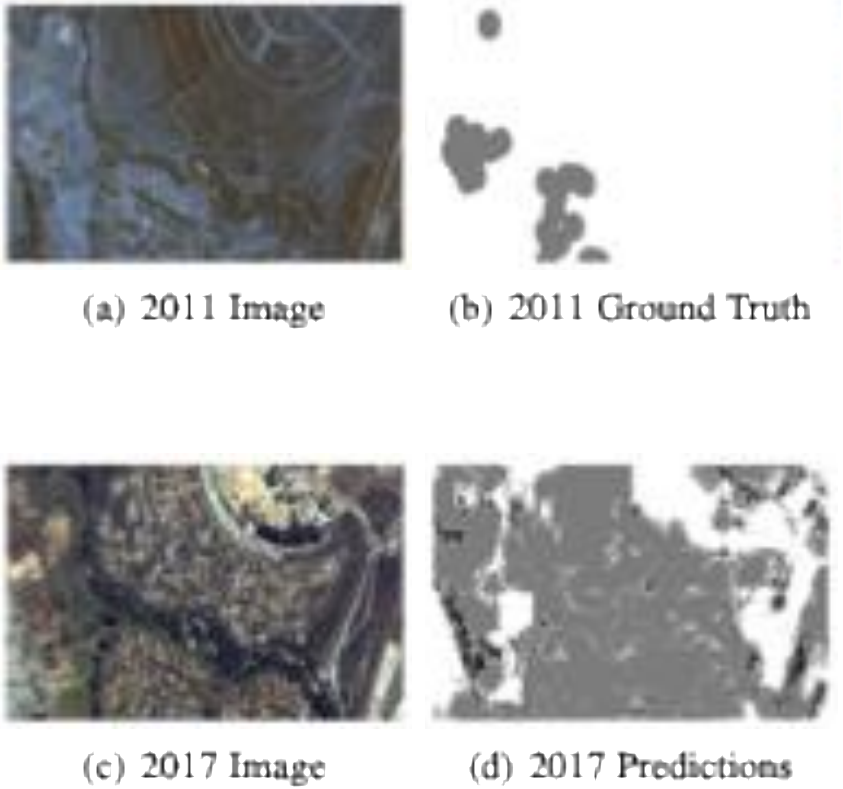
Data and machine learning in the social sciences

What is worth using?

Summary

References

Neural networks for feature extraction, dimension reduction, and labeling



- If you do have image data, audio data, video data, then neural networks can help do something with it
- Left: spatial apartheid in South Africa, work from Timnit Gebru's DAIR Institute (Sefala et al., 2021)
- "Edge detection" that happens as a part of neural networks can be useful
- Text data; language models can help, but the embeddings they produce are based only on co-occurrence and nothing else

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Summary

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

Summary

- What is data science? Mostly, applied machine learning
  - What is machine learning? Correlation-only statistics, no attempt to get at causality but more flexible at getting correlations
- What is a “computational method” or “computational science”? Unclear; but computation is a core part of everything now
- What is worthwhile?
  - Nonparametrics, DAGs for representing causality, limited use of techniques for unstructured data, and scaling human labeling/annotation/[qual] coding
- “All models are wrong…” (Box, 1979)

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

References 1 of 2

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-Powered Inference." *Science* 382 (6671): 669–674. https://doi.org/10.1126/science.adi6000

Arceneux, Kevin, Alan S. Gerber, and Donald P. Green. 2010. "A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark." *Sociological Methods & Research* 39 (2): 256-282. https://doi.org/10.1177/0049124110378098

Bender, Emily. 2024. "Resisting Dehumanization in the Age of 'AI'." *Current Directions in Psychological Science* 33 (2): 114-120. https://doi.org/10.1177/09637214231217286

Box, George E. P. 1979. "Robustness in the Strategy of Scientific Model Building." In Robustness in Statistics, edited by Robert L. Launer and Graham N. Wilkinson, 201–236. Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press.

Burrows, Roger and Mike Savage. 2014. "After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology". *Big Data & Society* 1 (1): 1–6. https://doi.org/10.1177/2053951714540280

Burrows, Roger and Nicholas Gane. 2006. "Geodemographics, Software, and Class." *Sociology* 40 (5): 793–812. https://doi.org/10.1177/00380385060675O7

Caruana, Rich. Nikos Karampatziakis, and Ainur Yessenalina. 2008. "An Empirical Evaluation of Supervised Learning in High Dimensions." In *ICML '08: Proceedings of the 25th International Conference on Machine Learning*. https://doi.org/10.1145/1390156.1390169

Dash, Anil. 2023. "Today's AI is unreasonable." June 8. https://www.anildash.com/2023/06/08/ai-is-unreasonable/

Fernández-Delgado, Manuel, Eva Cernadas, and Senén Barro. "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 5 (1): 3133–3181. https://doi.org/10.5555/2627435.2697065

Fergus, Devin. 2013. "The Ghetto Tax: Auto Insurance, Postal Code Profiling, and the Hidden History of Wealth Transfer." In *Beyond Discrimination: Racial Inequality in a Postracial Era*, edited by Fredrick C. Harris and Robert C. Lieberman, 277–316. Russel Sage Foundation.

Fourcade, Marion and Kieran Healy. 2013. "Classification Situations: Life-Chances in the Neoliberal Era." *Accounting, Organizations and Society* 38 (8): 559–572. https://doi.org/10.1016/j.aos.2013.11.002

Frey, Frey, Bruno S., David A. Savage, and Benno Torgler. 2011. "Behavior under Extreme Conditions: The *Titanic* Disaster." *Journal of Economic Perspectives* 25 (1): 209–222. https://doi.org/10.1257/jep.25.1.209

Frey, Bruno S., David A. Savage, and Benno Torgler. 2010. "Noblesse Oblige? Determinants of Survival in a Life-or-Death Situation." *Journal of Economic Behavior & Organization* 74 (1–2): 1–11. https://doi.org/10.1016/j.jebo.2010.02.005

Golder, Scott and Michael Macy .2012. "Social Science with Social Media." *ASA footnotes* 40 (1).

Harford, Tim. 2014. "Big Data: Are We Making a Big Mistake?" *Significance* 11 (5): 14-19. https://doi.org/10.1111/j.1740-9713.2014.00778.x

Jones, Matthew L. 2018. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48 (5): 673–684. https://doi.org/10.1525/hsns.2018.48.5.673

Keyes, Os. 2018. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." In *Proceedings of the ACM on Human-Computer Interaction*, 2 (CSCW), 88:1–88:22.

King, Gary. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331 (6018), 719–721. https://doi.org/10.1126/science.1197872

Introduction

History and nature of machine learning

Data and machine learning in the social sciences

What is worth using?

Summary

References

References 2 of 2

Kiviat, Barbara. 2019. "The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores." *American Sociological Review* 84 (6): 1134–1158. https://doi.org/10.1177/0003122419884917

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–1205. https://doi.org/10.1126/science.1248506

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review* 86 (3): 532–565. https://doi.org/10.1177/00031224211004187

Maxmen, Amy. 2019. "Can Tracking People Through Phone-Call Data Improve Lives?" *Nature* 569 (7758): 614–617. https://doi.org/10.1038/d41586-019-01679-5

Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *The Annals of Applied Statistics* 12 (2): 685–726. https://doi.org/10.1214/18-AOAS1161SF

Messerli, Franz H. 2012. "Chocolate Consumption, Cognitive Function, and Nobel Laureates." *The New England Journal of Medicine* 367: 1562–1564. https://doi.org/10.1056/NEJMon1211064

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87

Ochigame, Rodrigo. 2020. "The Long History of Algorithmic Fairness: Fair Algorithms from the Seventeenth Century to the Present." *Phenomenal World*, January 30, 2020. https://www.phenomenalworld.org/analysis/long-history-algorithmic-fairness/

Opsomer, Jean, Yuedong Wang, and Yuhong Yang. 2001. "Nonparametric Regression with Correlated Errors." *Statistical Science* 16 (2):134–153. https://doi.org/10.1214/ss/1009213287

Ogburn, Elizabeth L., and Ilya Shpitser. 2021. "Causal Modelling: The Two Cultures." *Observational Studies* 7 (1): 179–183. http://doi.org/10.1353/obs.2021.0006

Pearl, Judea. 2009. *Causality* (2nd ed.). Cambridge University Press. Epilogue: http://bayes.cs.ucla.edu/BOOK-2K/causality2-epilogue.pdf

Rodolfa, Kit T., Pedro Saleiro, and Rayid Ghani. 2021. "Bias and Fairness". In *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*, 2nd ed., edited by Bylan Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, 281–312. New York: Chapman and Hall/CRC.

Salganik, Matt, Laren Maffeo, and Cynthia Rudin. "Prediction, Machine Learning, and Individual Lives: An Interview with Matthew Salganik." *Harvard Data Science Review* 2 (3). https://doi.org/10.1162/99608f92.eecdfa4e

Savage, Mike and Roger Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41 (5): 885–899. https://doi.org/10.1177/0038038507080443

Savage, Mike and Roger Burrows. 2009. "Some Further Reflections on the Coming Crisis of Empirical Sociology." *Sociology* 43 (4): 762–772. https://doi.org/10.1177/0038038509105420

Sefala, Raesetje, Timnit Gebru., Luzango Moorosi, and Richard Klein. 2021. "Constructing a Visual Dataset to Study the Effects of Spatial Apartheid in South Africa." In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track* (Round 2). https://openreview.net/forum?id=WV0waZz9dTF

Wasserman, Larry. 2014. "Rise of the Machines." In *Past, Present and Future of Statistical Science*, edited by Xihong Lin, Christian Genest, David Banks, Geert Molenberghs, David Scott, and Jane-Ling Wang, 525–536. CRC Press.

Wiggins, Chris, and Matthew L. Jones. 2023. *How Data Happened: A History from the Age of Reason to the Age of Algorithms*. National Geographic Books.