CONCEPTUALIZING PROGRESS: BEYOND THE "COMMON TASK FRAMEWORK"

Part of "Beyond the data and model: Integration, enrichment, and progress", Webinar 3 in support of the NIH/NCATS Bias Detection Tools for Clinical Decision Making Challenge. With Shauna M. Overgaard, Young J. Juhn, and Chung Il Wi. National Center for Advancing Translational Sciences, National Institutes of Health.

Momin M. Malik, PhD
Senior Data Science Analyst – AI Ethics
Center for Digital Health
Mayo Clinic

Slides at https://www.mominmalik.com/ncats2023.pdf; video at https://vimeo.com/799944018

KENTARO TOYAMA, "GEEK HERESY"

"In the course of five years [at Microsoft Research in India], I oversaw at least ten different technology-for-education projects… Each time, **we thought we were addressing a real problem.** But… in the end it didn't matter—**technology never made up** for a lack of good teachers or good principals. Indifferent administrators didn't suddenly care more because their schools gained clever gadgets… and school budgets didn't expand no matter how many 'cost-saving' machines the schools purchased. If anything, **these problems were exacerbated by the technology, which brought its own burdens.**"

KENTARO TOYAMA, “GEEK HERESY”

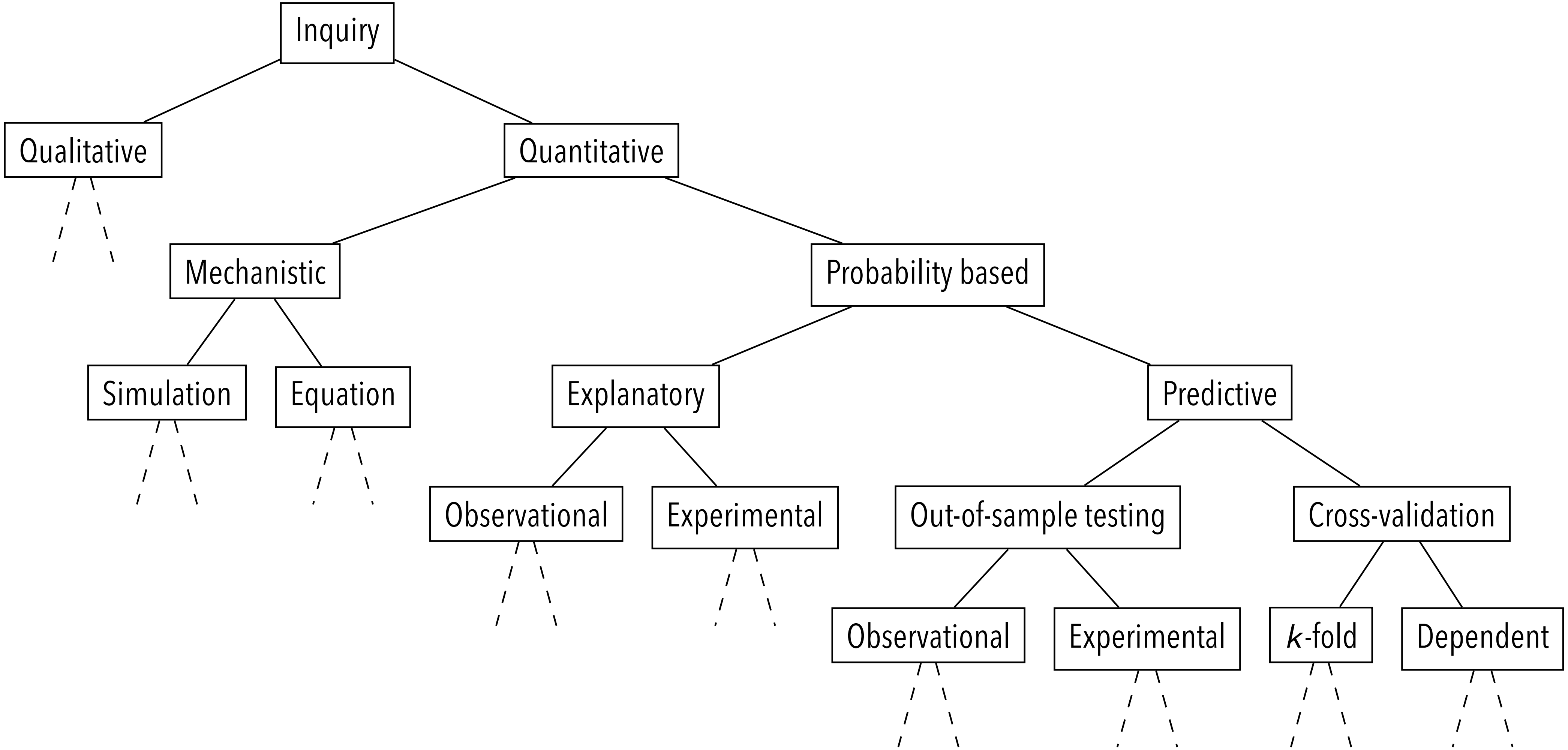
“**These revelations were hard to take.** I was a computer scientist, a Microsoft employee, and the head of a group that aimed to find digital solutions for the developing world. **I wanted nothing more than to see innovation triumph**, just as it always did in the engineering papers I was immersed in. But **exactly where the need was greatest, technology seemed unable to make a difference.**”

KEY POINTS

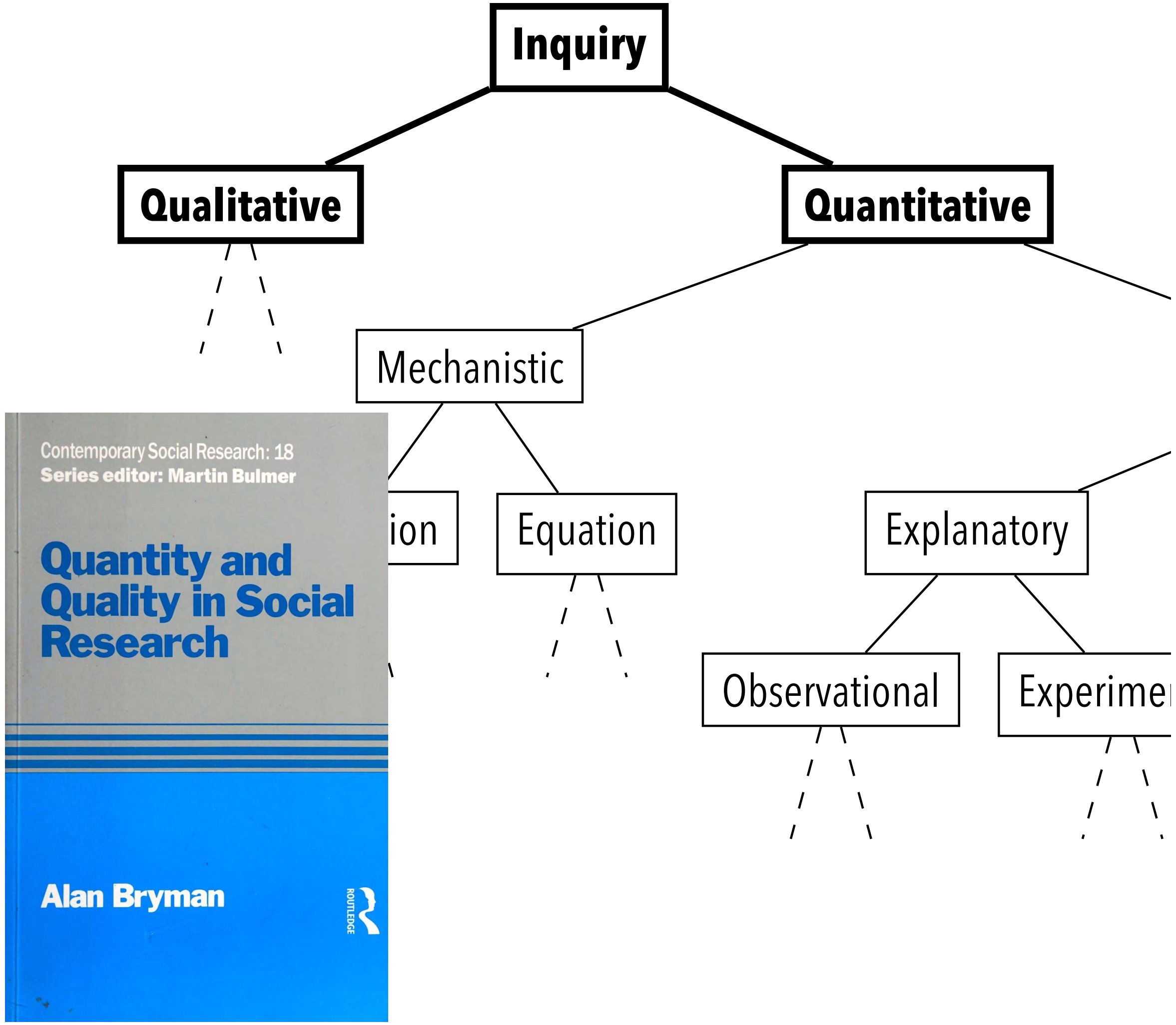
- There are numerous critiques of “technological solutionism”: belief that social (or sociotechnical) problems can have *technical* solutions
  - Dramatic practical and moral failings of such technical approaches can sometimes drive practitioners to seek alternatives (Malik & Malik, 2021); but it shouldn’t have to come to that!
- Instead, recognize: Measurement is always imperfect, which always limits what quantification can do
  - Uncertainty quantification is one useful tool within quantitative frameworks
- We can’t just look at differential model performance, but need to consider underlying causal dynamics, both quantitative and qualitative
  - We want to avoid a repeat of “actuarial fairness”

HIERARCHY OF METHODOLOGICAL LIMITATIONS

MALIK (2020)

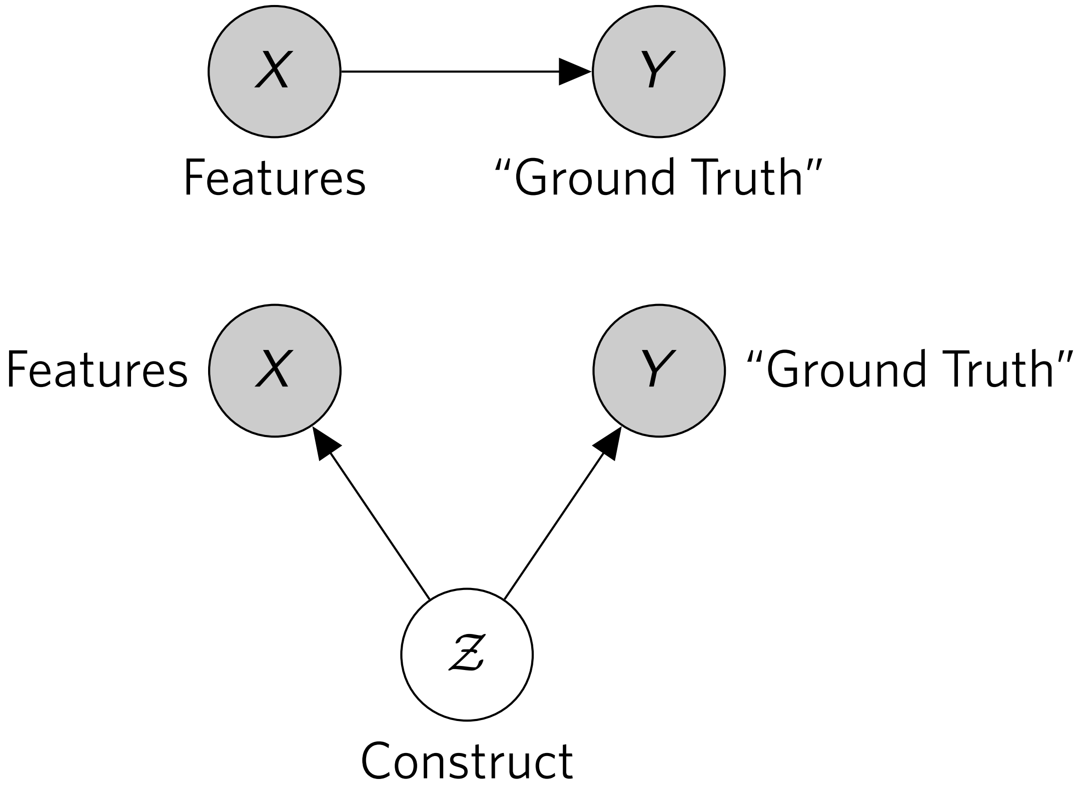


QUALITATIVE VS. QUANTITATIVE



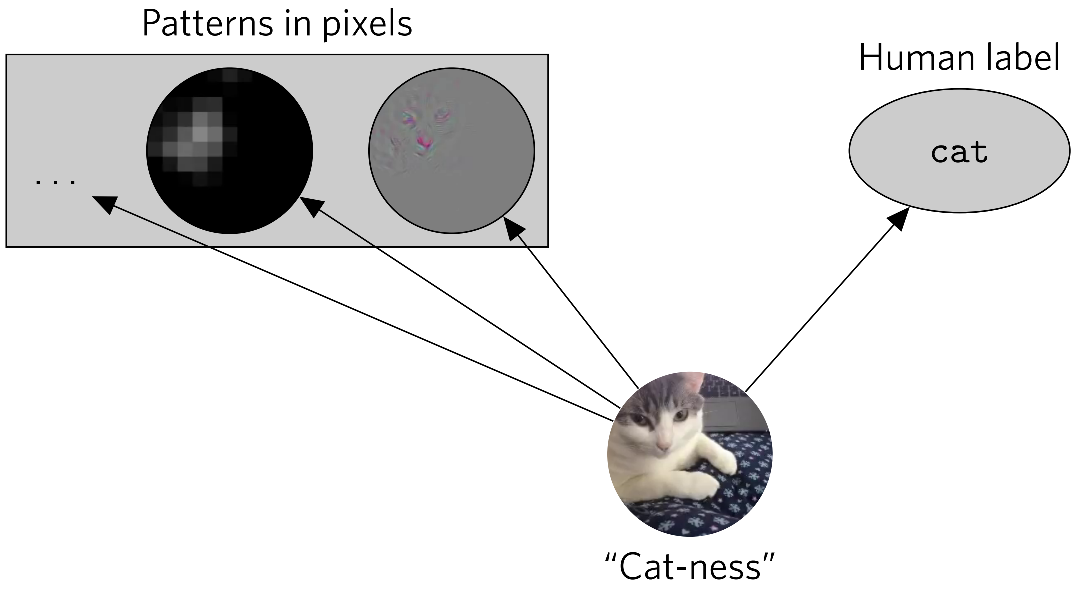
- Qualitative research can get directly at how things are multifaceted, heterogeneous, intersubjective
- Quantification/measurements lock in one meaning; and always are *proxies,* which are imperfect (“all models are wrong;” Box, 1979)

CHALLENGES OF QUANTIFICATION/ MEASUREMENT



- *Constructs*: primitives of social science
  - What we care about (Jacobs & Wallach, 2020)
  - Often unobservable and hypothetical/subjective (e.g., friendship). Constructs are sometimes even unobservable in physics! (Chang, 2004)
  - Proxies always give errors (for binary constructs: false negatives and false positives), and even can be gamed (Campbell, 1975)

CONSTRUCTS: SUBJECTIVE, MULTIFACETED



Patterns in pixels

...

Human label

cat

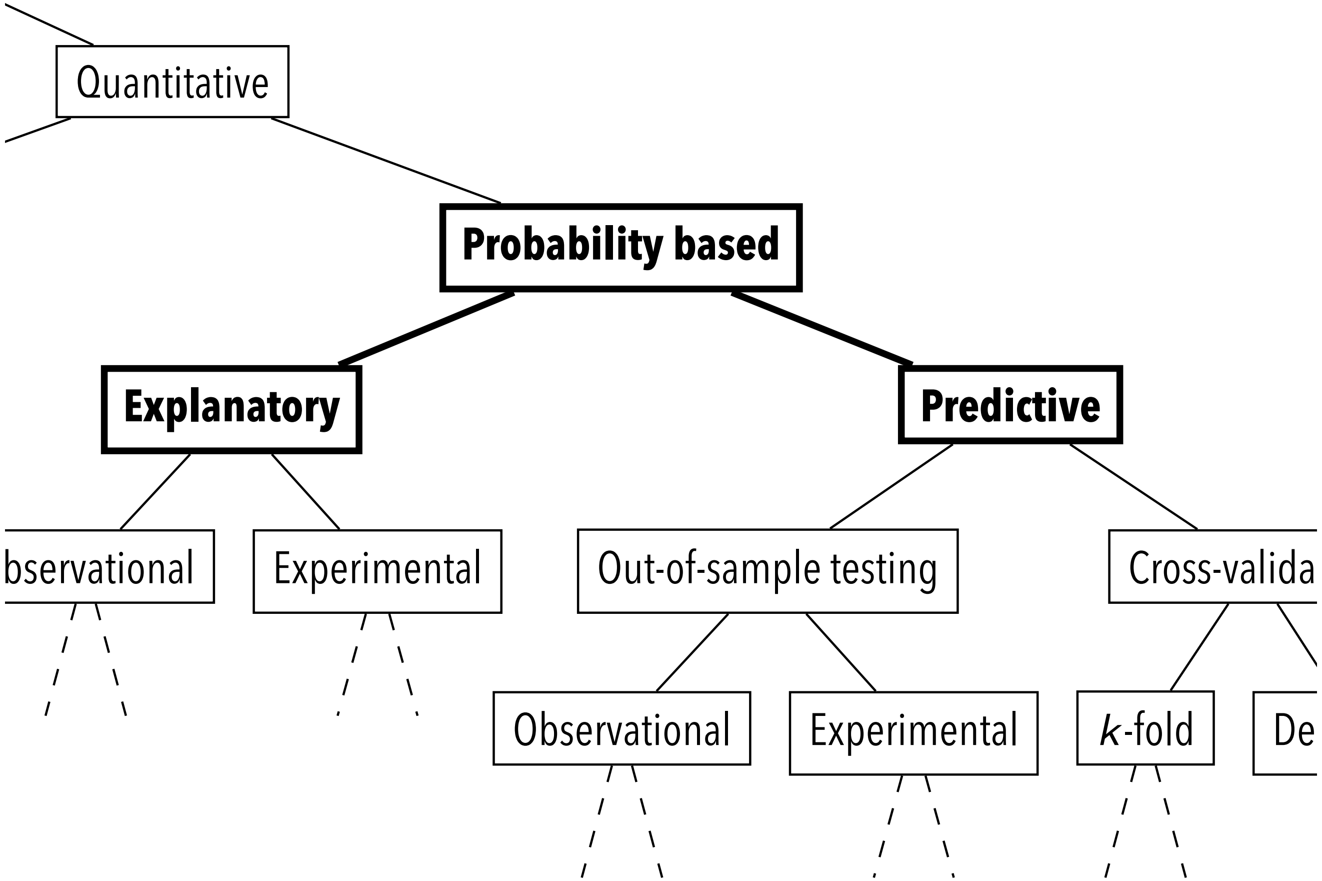
"Cat-ness"

WAYS OF UNDERSTANDING PEOPLE

	As a case (quant)	**In narrative (qual)**
Context/circumstance	Stripped away	Key
Mental states	Absent (for the most part)	Crucial; constitutive
Relevant features	Determined in advance	Emergent
Orientation to time	Atemporal	Chronological
Ordering of features	Unimportant	Meaningful
Other actors	Invisible	Often present
Causal logic	Mathematical	Theoretical
Boost predictive validity	Add cases	Know person better

Slide from Barbara Kiviat (work in progress), based on “Bowker and Star 2000; Bruner 1986; Desrosières 1998; Espeland 1998; Espeland and Stevens 1998, 2008; Fourcade and Healy 2017; Hacking 1990; Porter 1994, 1995; Ricouer 1998; White 1980, 1984”. I would add: Abbott, 1988

ML IS “PREDICTION” ONLY



- “Predictions” are defined as what minimizes loss *within a predetermined frame*
  - *Correlations* do this

- Non-causal correlations can sometimes predict well within a frame, but they frequently don’t explain, and can fail outside (Shmueli, 2010; Mullainathan & Spiess, 2017)

- ML is also fundamentally statistical; so we should estimate uncertainty

UNCERTAINTY QUANTIFICATION

- Prediction-only means that we can ignore uncertainty quantification like *p*-values. Right?
- No! Metrics are *estimators* (specifically, point estimates) of out-of-sample performance
    - They have distributions, which we should also estimate and study to know more about out-of-sample performance (i.e., is a difference in AUC significant or not? If not, unlikely the finding reproduces)
- *k*-fold cross-validation turns out to not be a valid way of estimating out-of-sample distributions! (Wager, 2019) Instead, use completely held-out data

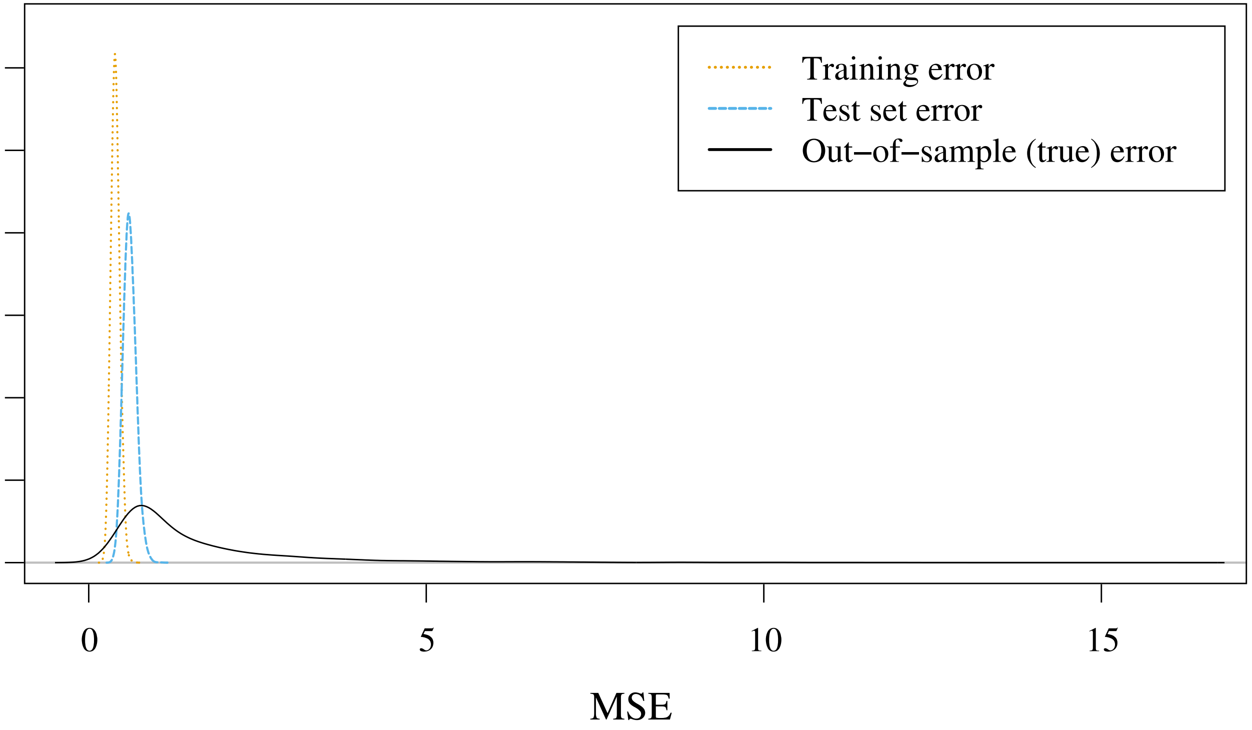
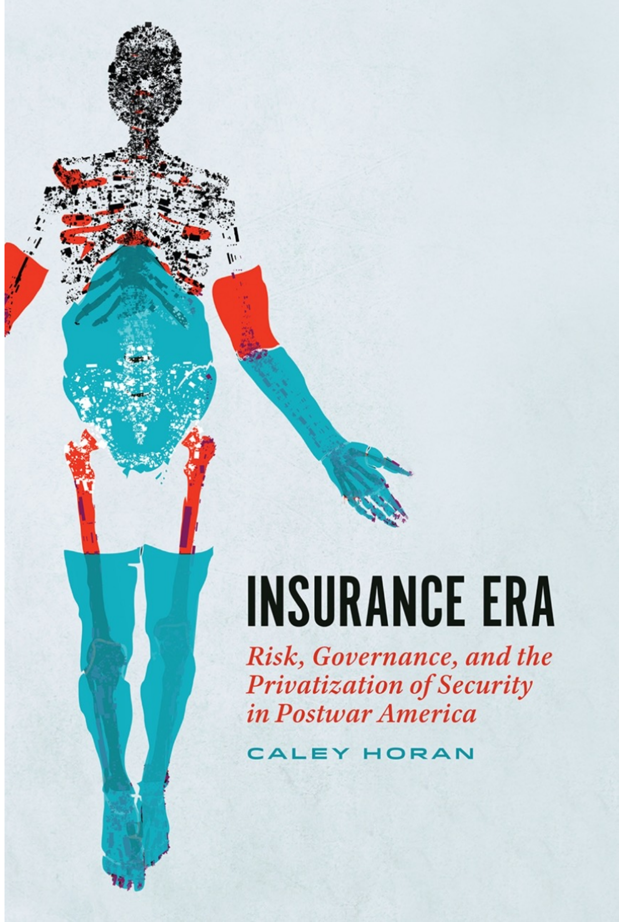


Figure: How *dependencies* bias cross-validation estimates of out-of-sample performance, from Malik (2020)

“ACTUARIAL FAIRNESS”

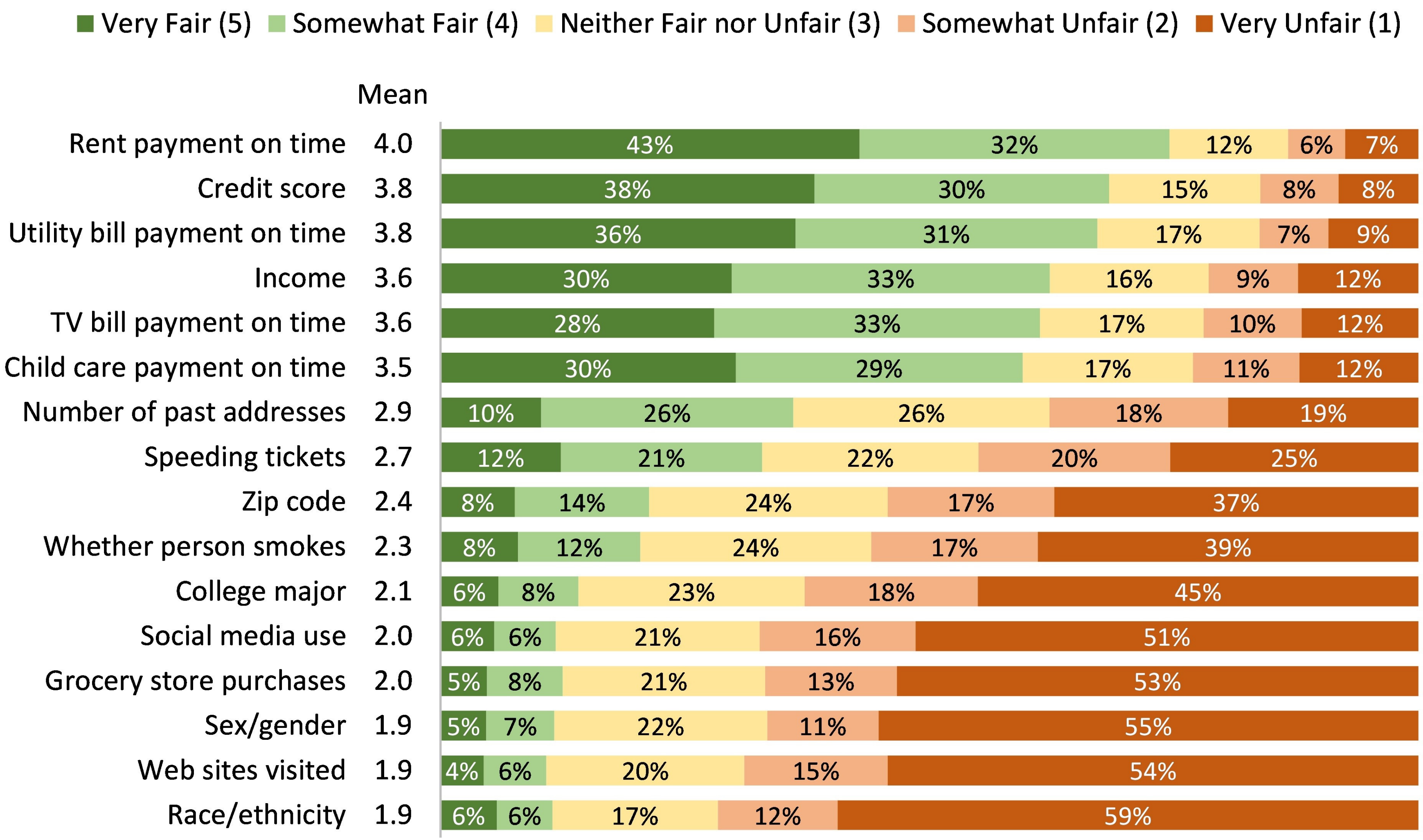
OCHIGAME (2020)

- “Actuarial fairness” was a concept invented in the 1970s in response to civil rights campaigns, organizing against insurance redlining and for nationalizing insurance, and feminist campaigners, organizing against higher health insurance premiums for women (Horan, 2021)
- Actuarial fairness is: if it correlates, it is “fair” to use. This moved the debate into a technical realm. But people find the *outcomes* of actuarial fairness deeply unfair (Kiviat, 2019; Heras et al., 2019; Landes 2015)
- E.g., ZIP Code is a legal input for car insurance rates; indeed, empirically, car insurance payouts are higher in areas with more Black and brown residents.
  - Why? *Segregation* and *redlining* correlates race with geography, and underinvestment correlates geography with car insurance claims
  - *Even if race itself is excluded, any correlates of it have the same effect.* Similarly, *excluding correlates leads to worse performance*
- Even if not used for the *purpose* of discrimination, using race or its correlates effectively further punishes the victims of injustice and cruelty (Hellman, 2008), amounting to a “tax” on Black and brown people (Fergus, 2013)


INSURANCE ERA
Risk, Governance, and the Privatization of Security in Postwar America
CALEY HORAN

WHAT IS FAIR TO USE FOR CONSUMER LENDING?

NATIONAL SURVEY BY KIVIAT (2021)



DISCUSSION

- Not everything should be formalized, but also, not everything *can* be formalized (Selbst et al., 2019); some things can only be addressed qualitatively
  - Try to have better measurements of more things that matter
  - Uncertainty quantification does help for quantitative work
- What can be put into a “common task framework” already presupposes too much ethically (LaCroix & Luccioni, 2022) and can be unrealistic (Wagstaff, 2012)
  - Have we decided on optimizing to an unjust status quo, or whether we try to change it, e.g., through affirmative, reparative actions?
- AI risks recreating the harms of “actuarial fairness” by optimizing to outcomes regardless of how or why they come about (Ochigame, 2020)

REFERENCES

Bouk D. How our days became numbered: risk and the rise of the statistical individual. Chicago: University of Chicago Press; 2015.

Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, editors. Robustness in statistics. New York: Academic Press; 1979. p. 201-236. doi: 10.1016/B978-0-12-438150-6.50018-2.

Campbell T. Assessing the impact of planned social change. In: Lyons GM, editor. Social research and public policies: the Dartmouth / OECD Conference. Dartmouth College: The Public Affairs Center; 1975. p. 3-45.

Chang H. Inventing temperature: measurement and scientific progress. Oxford (UK): Oxford University Press; 2004. doi: 10.1093/0195171276.001.0001.

Fergus D. The ghetto tax: auto insurance, postal code profiling, and the hidden history of wealth transfer. In: Harris FC, Lieberman RC, editors. Beyond discrimination: racial inequality in a postracial era. New York: Russel Sage Foundation; 2013. p. 277-316.

Hellman D. When is discrimination wrong? Cambridge (MA): Harvard University Press; 2008.

Heras AJ, Pradier P-C, Teira D. What was fair in actuarial fairness? Hist Human Sci. 2019 Sep 15;33(2):91-114. doi: 10.1177/0952695119856292.

Horan C. Insurance era: risk, governance, and the privatization of security in postwar America. Chicago: University of Chicago Press; 2021.

Jacobs AZ, Wallach H. Measurement and fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21); 2021 Mar 3-10. New York: ACM Press; 2021. p. 375-85. doi: 10.1145/3442188.3445901.

Kiviat B. The moral limits of predictive practices: The case of credit-based insurance scores. Am Sociol Rev. 2019;84(6):1134-58. doi: 10.1177/0003122419884917.

Kiviat B. Which data fairly differentiate? American views on the use of personal data in two market settings. Sociol Sci. 2021 Jan 13;8(2):26-47. doi: 10.15195/v8.a2.

LaCroix T, Luccioni AS. Metaethical perspectives on 'benchmarking' AI ethics. arXiv: 2204.05151 [Preprint]. 2022 Apr 11: [39 p.]. Available from: https://arxiv.org/abs/2204.05151.

Landes X. How fair is actuarial fairness? J Bus Ethics. 2015;128:519-33. doi: 10.1007/s10551-014-2120-0.

Malik MM. A hierarchy of limitations in machine learning. arXiv: 2002.05193 [Preprint]. 2020 Feb 29: [68 p.]. Available from: https://arxiv.org/abs/2002.05193.

Malik M, Malik MM. Critical technical awakenings. J Social Comput. 2021 Dec; 4(4):365-384. doi: 10.23919/JSC.2021.0035.

Mullainathan S, Spiess J. Machine learning: An applied econometric approach. J Econ Perspect. 2017;31(2):87-106. doi: 10.1257/jep.31.2.87.

Ochigame R. The long history of algorithmic fairness. Phenomenal World [Internet]. 2020 Jan 30. Available from: https://www.phenomenalworld.org/analysis/long-history-algorithmic-fairness.

Toyama K. Geek heresy: rescuing social change from the cult of technology. Public Affairs; 2015.

Selbst AD, boyd d, Freidler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* '19); 2019 Jan 29-31. New York: ACM Press; 2019. p. 59-68. doi: 10.1145/3287560.3287598.

Shmueli G. To explain or to predict? Stat Sci. 2010 Aug;25(3):289-310. doi: 10.1214/10-STS330.

Wagstaff ML. Machine learning that matters. Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML'12); 2012 Jun 26-Jul 1. New York: ACM Press; 2012. p. 1851–6. Available from: https://icml.cc/2012/papers/298.pdf

Wagner S. Cross-validation, risk estimation, and model selection: comment on a paper by Rosset and Tibshirani. J Am Stat Assoc. 2020;115(529):157-60. doi: 10.1080/01621459.2020.1727235