ICQCM

CRITICAL DATA SCIENCE FOR A DIVERSE WORLD

Critical Approaches to Machine Learning

Momin M. Malik

Sunday, 27 March 2022

ICQCM 2022 Summit

Baltimore, MD

Manzel Bowman, *Station No.99* (2018)

Overview

- Background/review
- Scope; review of some critiques *of* machine learning
- Examples of work that take, or enable, a critical approach
- Possibilities

Goals

- To separate out possibilities “external” to ML to those “internal” to ML (while also critiquing, and blurring, that boundary)
- To give neat examples of ML
- To connect some specific technical constructs in ML to possibilities for critique
- To get a conversation/sharing going!

Background/review

Building on two remote sessions

- "Defining Critical Quantitative and Computational Methodologies." William T. Grant AQC SCHOLARS Virtual Seminar Series, 27 May 2021. https://www.mominmalik.com/icqcm2021.pdf
- "Computational Approaches III: Applications." ICQCM 2021 Seminar Series, 22 July 2021. https://www.mominmalik.com/icqcm2021b.pdf
- If you didn't know and/or didn't attend those sessions: this builds on them, but doesn't require them. Also feel free to check out the slides!

“Defining QCM” key points: What

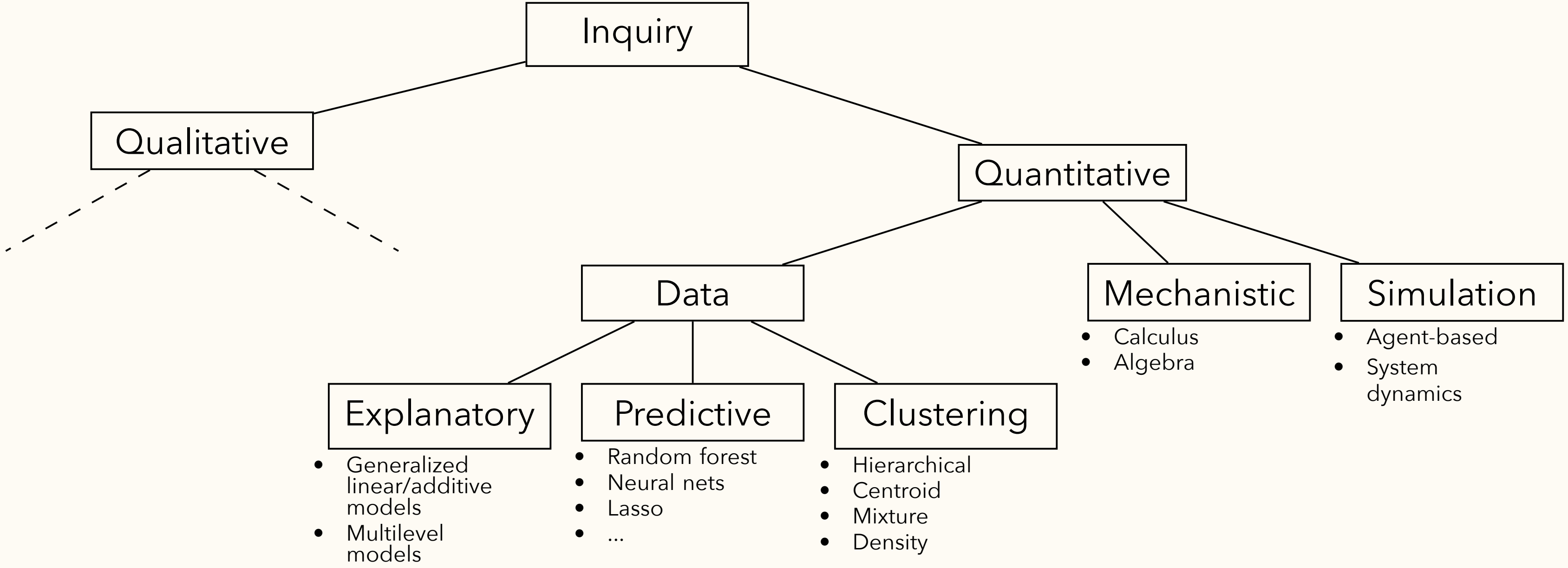
- Incorporating ready-made quant methods into a critical approach is most straightforward
  - “Minimal” critical QCM: quant demonstrations of disparities *that links to theory* about the source of those disparities (e.g., white supremacy, dehumanization)
- But more intellectually interesting for me is integrating the logic of modeling with the logic of critical theory at a fundamental level
- Much harder—requiring dual training—but a rich intellectual project
  - On the other hand, maybe useless practically, and the “minimal” version of critical QCM is most useful and important

"Defining QCM" key points: Why

- Strategic quantification/strategic modeling (after Spivak's "strategic essentialism") to demonstrate inequality? (Strategic positivism; Wyly, 2009)
  - Rhetorical use: convince power-brokers?
- "Counterhegemonic modeling" (Richardson 2020): modeling ironically to reveal the absurdity of modeling?
- Alternatively: just because quantification is currently associated with power does not mean it is essentially so. Qualitative inquiry can be just as or more oppressive, it just isn't currently in power

ICQCM

"Computational Approaches" key points



"Computational Approaches" key points

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

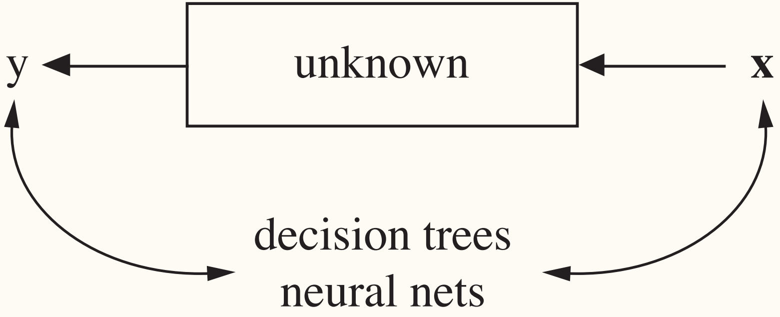
Possibilities

Conclusion and discussion

References

Statistics:

Machine learning:



Machine learning: An instrumental use of statistical correlations to *mimic* the output of a target process, rather than understand the *relationship* between inputs and outputs. Involves finding expressions that maximize correlation.

Breiman 2001. See also Jones 2018.

“Computational Approaches” key points

“Source subject”: Marquese Scott

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

Scope; Review of some critiques

ML is bad for making claims about the world!!

- *The best-fitting* (most accurate*) *model does not necessarily reflect how the world works*
- This has been shocking in statistics for decades (Stein's paradox, Leo Breiman's "two cultures" [2001]), but little known outside
- Why: one reason is the "bias-variance tradeoff"
  - Even when available, the "true" covariates may be noisy, in which case proxies (or even just going with the mean) sometimes does better
- Another reason: narrowing in to get one causal relationship "correct" might require sacrificing the rest of the model
- So: we can use correlations to "predict" without "explaining" (knowing causality)!

* Or other relevant metric of success

Machine learning to "prove"?

- Unlike statistics, machine learning isn't well-suited for making claims about the way the world works
- Maybe incidental outputs of models are revealing (e.g., "Word embeddings quantify 100 years of gender and ethnic stereotypes"; Garg et al., 2017)… but that is indirect, and reifies ML

ICQCM

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

Critiques of ML from "outside"?

- Safiya Noble
- Ruha Benjamin
- Meredith Broussard
- Virginia Eubanks
- Cathy O'Neil
- Adrian Mackenzie
- Dan McQuillan
- Matteo Pasquinelli
- Yarden Katz
- *Data Feminism*
- Coalition for Critical Technology

- (Some of these people are "insiders", but these are critiques that aren't making use of machine learning, or anything more than straightforward quantitative methods, and so outside the scope I'm setting)
- **This type of work is probably societally more important than the scope I consider**, and there's lots more work to do of this type of work, but I am interested in a less explored space of critical approaches *with/within/through* ML

ICQCM

Applied STS?

- Incorporating science studies lessons (e.g., "responsible research and innovation") into data science education
  - "Integrating FATE/critical data studies into data science curricula: Where are we going and how do we get there?" (Bates et al., 2020)
  - "Critique and contribute: A practice-based framework for improving critical data studies and data science" (Neff et al., 2017)
  - "'You social scientists love mind games': Experimenting in the 'divide' between data science and critical algorithm studies" (Moats & Seaver, 2019)

ICQCM

Overview

Background/ review

Scope; review of some critiques

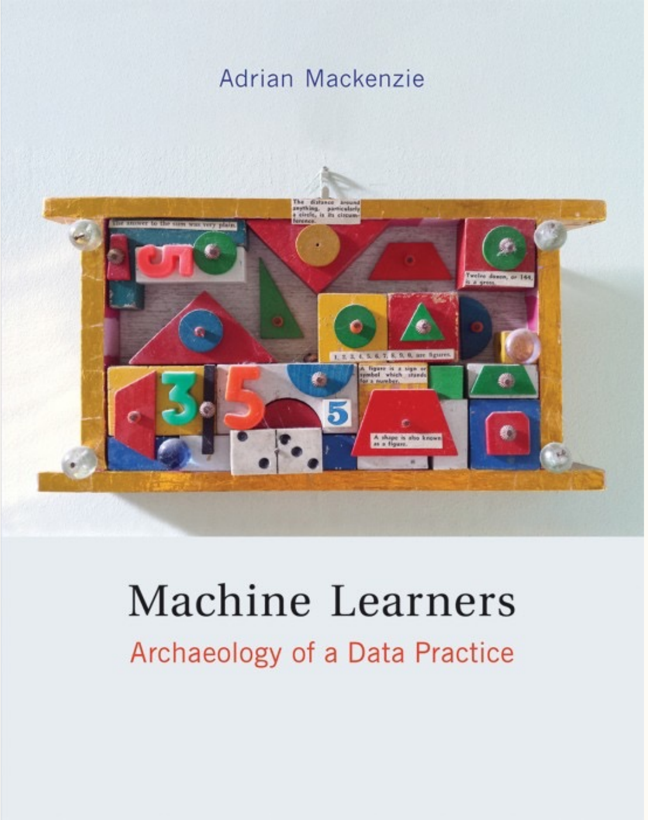
Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

Can critique and quantification mix?



9 We might also approach the epistopic fault line in machine learning topologically. More than a decade ago, the cultural theorist Brian Massumi wrote that "the space of experience is really, literally, physically a topological hyperspace of transformation" (Massumi 2002, 184). Much earlier, Gilles Deleuze had conceptualized Michel Foucault's philosophy as a topology, or "thought of the outside" (Deleuze 1988b), as a set of movements that sought to map the diagrams that generated a "kind of reality, a new model of truth" (Deleuze 1988b, 35). More recently, this topological thinking has been extended and developed by Celia Lury among others. In "The Becoming Topological of Culture," Lury, Luciana Parisi, and Tiziana Terranova suggest that "a new rationality is emerging: the moving ratio of a topological culture" (Lury, Parisi, and Terranova 2012, 4). In this new rationality, practices of ordering, modeling, networking, and mapping co-constitute culture, technology, and science (Lury, Parisi, and Terranova 2012, 5). At the core of this new rationality, however, lies a new ordering of continuity. The "ordering of continuity," Lury, Parisi, and Terranova propose, takes shape "in practices of sorting, naming, numbering, comparing, listing, and calculating" (4). The phrase "ordering of continuity" is interesting because we don't normally

Vectorization and Its Consequences 65

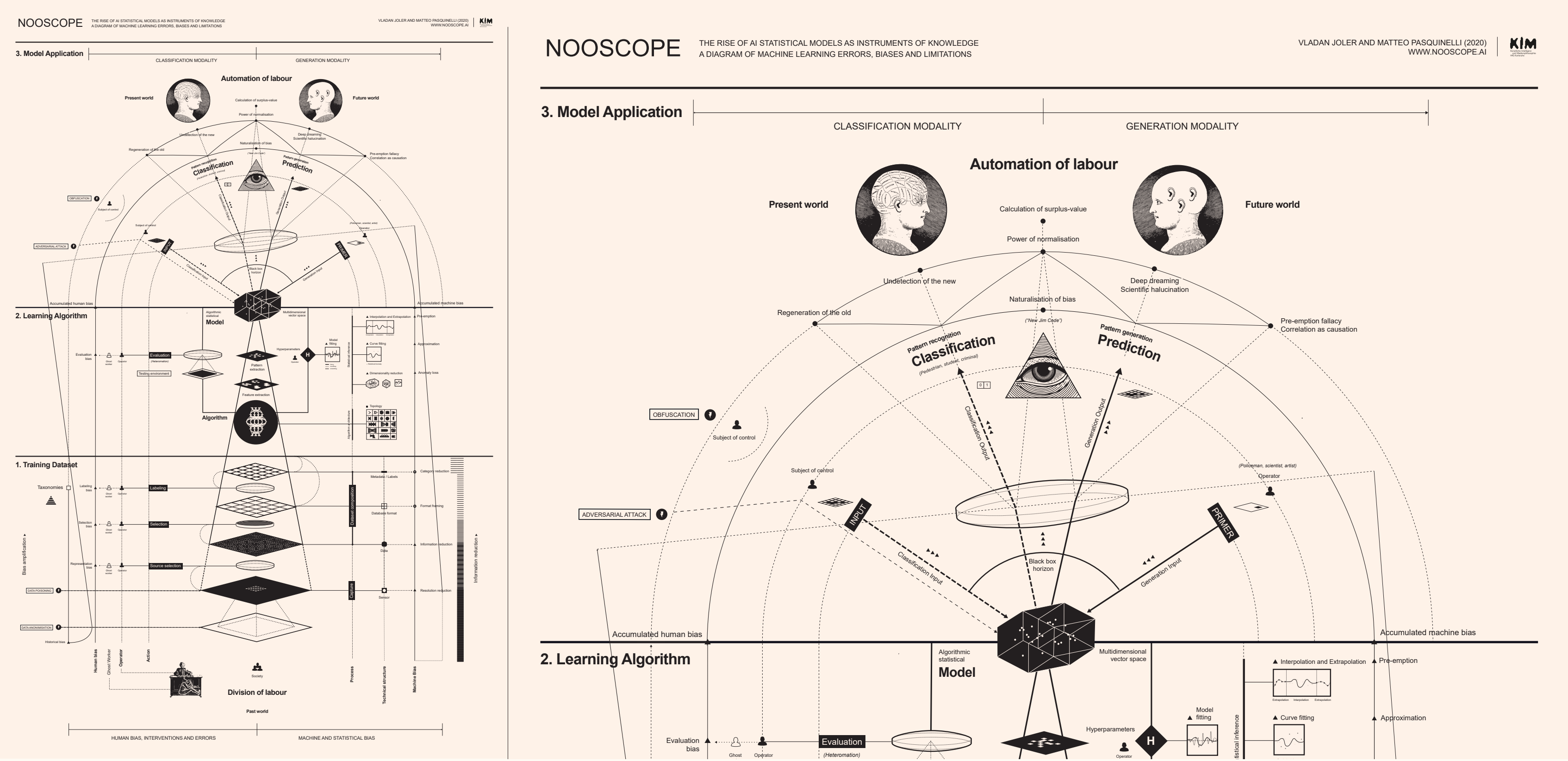
and Andrew Ng advocate returning often to equations). The mainstay of statistics, the linear regression model, usually appears in a more or less algebraic form:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \tag{3.1}$$

$$\hat{Y} = X_T \hat{\beta} \tag{3.2}$$

Equations 3.1 and 3.2 express a plane (or hyperplane) in increasingly diagrammatic abstraction. The possibility of diagramming a high-dimensional space derives largely from linear algebra. Reading equation 3.1 from left to right, the expression $\hat{Y}$ already

Artistic approaches?



ICQCM

General methodological critique?

- “Troubling trends in machine learning scholarship” (Lipton & Steinhardt, 2018)
- “Machine learning that matters” (Wagstaff, 2012)
- “Reliance on metrics is a fundamental challenge for AI” (Thomas & Uminsky 2020)
- “Underspecification presents challenges for credibility in modern machine learning” (D’Amour et al., 2020)

Examples of work that take, or enable, a critical approach

Auditing

- "Gender shades: Intersectional accuracy disparities in commercial gender classification" (Buolamwini & Gebru, 2019)
- Follow-up work: "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products" (Raji & Buolamwini, 2019), "Lessons from archives: strategies for collecting sociocultural data in machine learning" (Jo & Gebru, 2020)

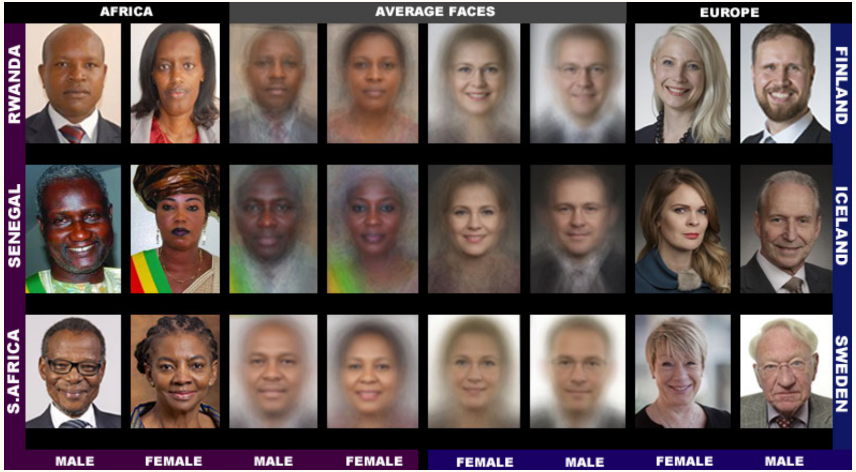
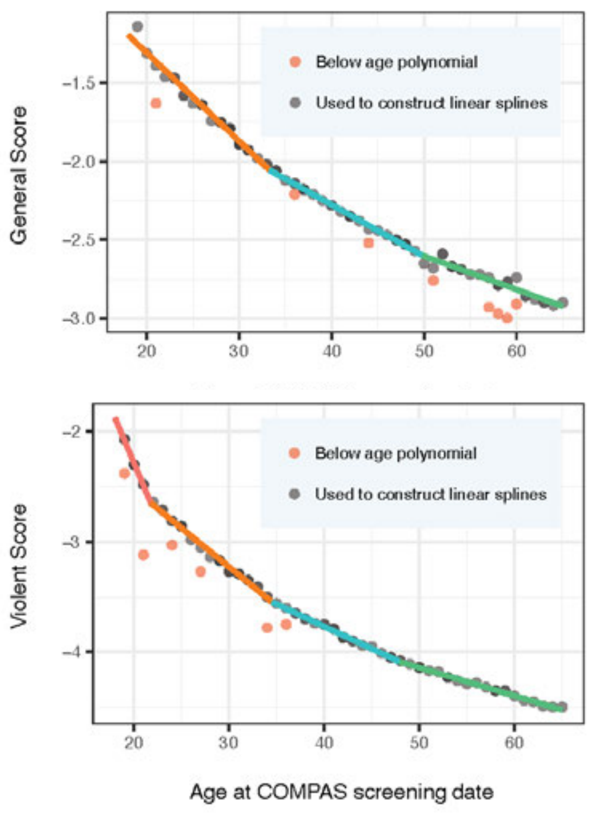


Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

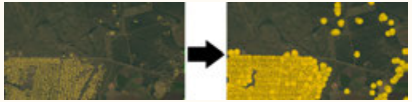
Reverse-engineering?

- "The age of secrecy and unfairness in recidivism prediction" (Rudin et al., 2020)
- Convincingly and impressively reverse-engineers the functional form and specification of the (proprietary) COMPAS model
- Allows for much finer-grained analysis and critique (and counter-modeling, although questionable)

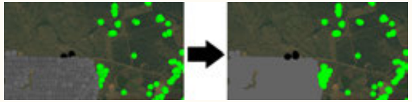


ML as scalable measurement?

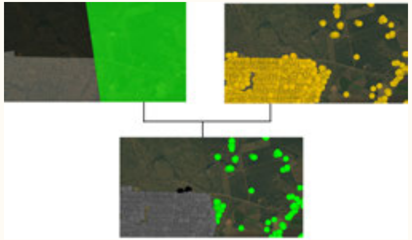
- "Constructing a visual dataset to study the effects of spatial apartheid in South Africa" (Sefala et al., 2021)
- (See also: "Measuring urban social diversity using interconnected geo-social networks"; Hristova et al., 2016)



(a) Converting building point data to polygons using a buffer algorithm to approximate the space covered by the buildings.



(c) Dissolving overlapping building polygons by neighborhood. Images are of size 786 x 386 pixels and resolution 2.5m per pixel.



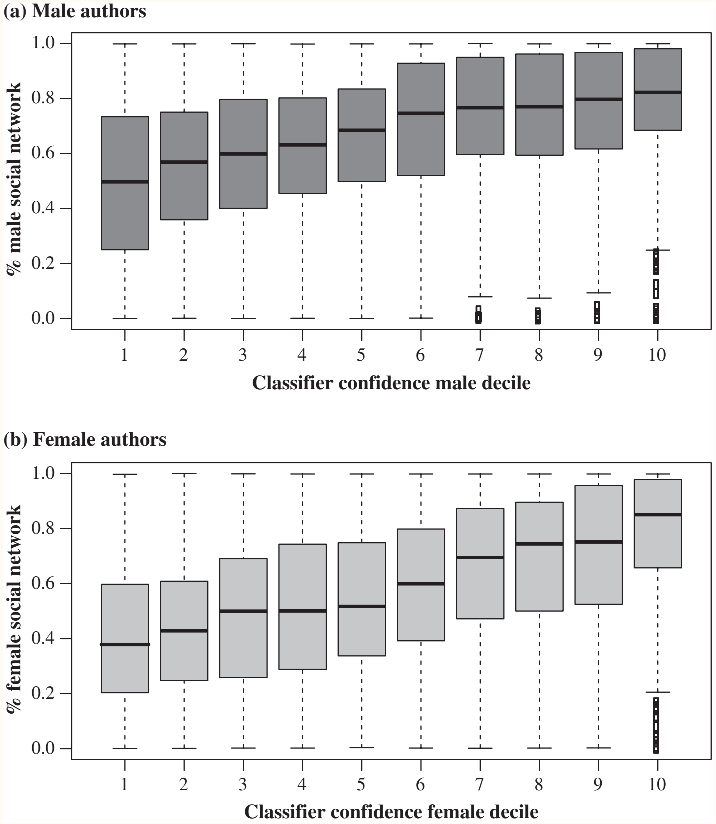
(b) Computing the spatial intersection between the land use labels from the EA dataset and the buffed building polygons so that we can label the types of neighborhoods these houses are in.

(a) 2011 Image (b) 2011 Ground Truth (c) 2017 Image (d) 2017 Predictions

Figure 7: Examples of the change detected between 2011 and 2017 images in a wealthy neighborhood near a big mall. Dark gray: Wealthy Neighborhood, White: Background.

Qualitative follow-up?

- “Gender identity and lexical variation in social media” (Bamman et al., 2014)
- Classified along a gender binary in line with existing work, but then looks at people who are “misclassified” qualitatively, in recognition of gender as a performance
- (See also: “Gender recognition or gender reductionism? The social implications of automatic gender recognition systems”, Hamidi et al., 2018; “The misgendering machines: Trans/HCI implications of automatic gender recognition”, Keyes 2018)



ICQCM

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

Qualitative evaluation?

- “Predictive learning analytics ‘at scale’: Towards guidelines to successful implementation in higher education based on the case of the Open University UK” (Herodotou et al., 2019a)
- (See also: “A large-scale implementation of predictive learning analytics in higher education: The teachers’ role and perspective”, Herodotou et al., 2019b)

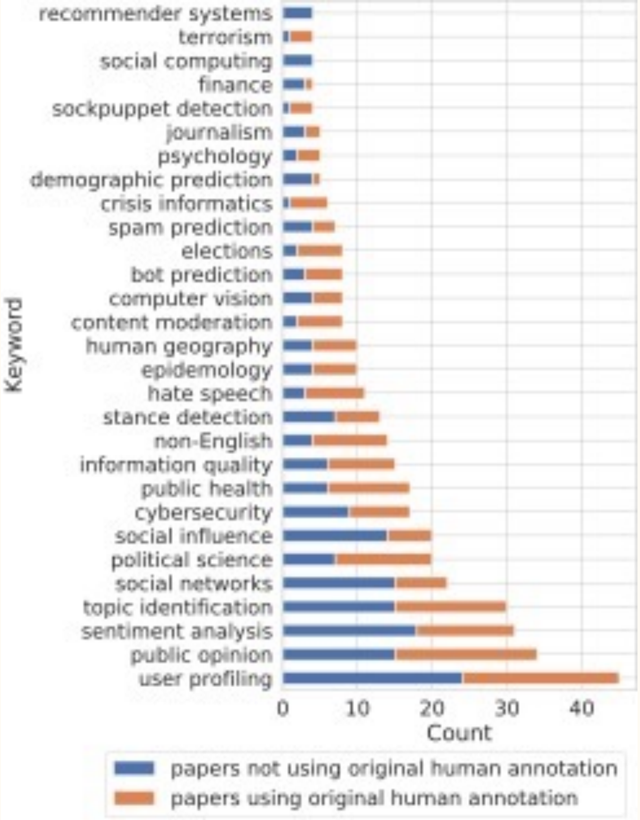
Table 1: Themes from Interviews with 20 Education Managers (Organized by Faculty)

Themes	RQs	Science faculty	Business Faculty	Teaching and Learning Unit	Student engagement and support	Tuition delivery	Data professionals
General perceptions	RQ1	Positive	Positive	Positive	Positive	Positive	Positive
Perceived challenges	RQ2	Lack of evaluation (Rec. 1)*; Lack of understanding as to how to use PLA (Rec. 2; Rec. 7)	Lack of evaluation (Rec. 1); Lack of understanding as to how to use PLA (Rec. 2; Rec. 7)	Lack of systematic evaluation (contradictory outcomes) (Rec. 1; Rec. 7)	Lack of evaluation (Rec. 1; Rec. 7)	Course design should define PLA use	Alignment across stakeholders (Rec. 3); Development of digital skills
Factors explaining slow uptake	RQ2	Teachers’ workload (Rec. 5); Varied course designs	Institutional changes impacting teachers’ work (Rec. 5); Lack of evidence about PLA effectiveness (Rec. 1)	Lack of ongoing support; Development of relevant skills	Management priorities (Rec. 4); Investment in staff	Lack of evidence (Rec. 1); Voluntary nature of participation; Lack of training; Teachers’ contracts (Rec. 5)	Lack of a common vision about PLA (Rec. 6)

ICQCM

Qualitative rigor?

- "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?" (Geiger et al., 2020)
- Where labels come from matter!



Qualitative rigor in context, too?

- “Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing” (Patton et al., 2020a)

Labels	Loss			Other			Aggression			Macro F1
	p	r	f	p	r	f	p	r	f	
Gold	77.08	56.92	65.49	88.04	95.76	91.74	50	27.59	35.56	64.26
Distant	50.00	48.46	49.22	85.63	84.50	85.06	19.72	24.14	21.71	52.00

Table 2: SVM performance trained on hand-labeled vs distantly-labeled data. The difference between F1 scores is statistically significant with p=0.001.

- “VATAS: An open-source web platform for visual and textual analysis of social media” (Patton et al., 2020b): “The lack of expertise is more relevant for marginalized communities or other critical domains that require specific training or background knowledge that are common in the field of social work. Poor performance is subsequently propagated into machine-learning models, as the models statistically fit the resulting data set with the purpose of learning to label samples the same way it was done by annotators. As a result, unreliable annotations can lead to models with low classification accuracy and biased predictions. This issue is why social work should drive social media annotation and interpretation of data and results, particularly when it relates to the most challenging social problems.”

Adding complexity?

- “Inherent disagreements in human textual inferences” (Pavlick & Kwiatkowski, 2019)
- We can always add complexity, but it is seldom a good idea for the purposes of modeling: when, and how, is it worth the extra bother?

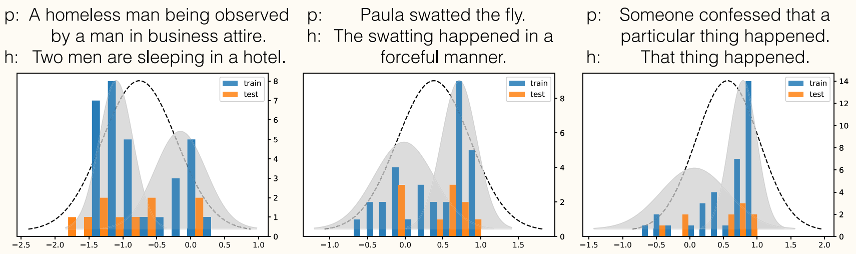


Figure 4: Examples of sentence pairs with bi-modal human judgment distributions. Examples are drawn from SNLI, the VerbNet portion of DNC, and the MegaVerdicality portion of DNC (from left to right). Training distribution is in blue; test in orange. Dotted black line shows the model fit when using a single component; shaded gray shows the model learned when allowed to fit k components. Distributions are over z-normalized scores in which 0 roughly corresponds to neutral ($p \not\to h$) but not precisely (§3.3).

ICQCM

Investigating context?

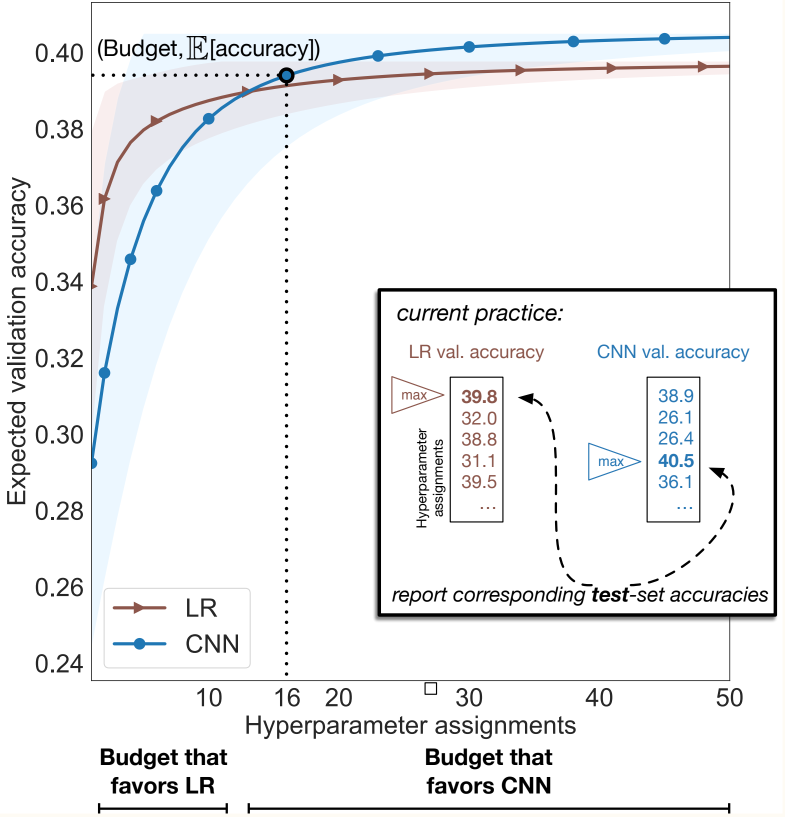
- "Energy and policy considerations for deep learning in NLP" (Strubell et al., 2019)
- The paper that first pointed out the enormous energy cost (and therefore, CO_2 emissions) of deep learning models

Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

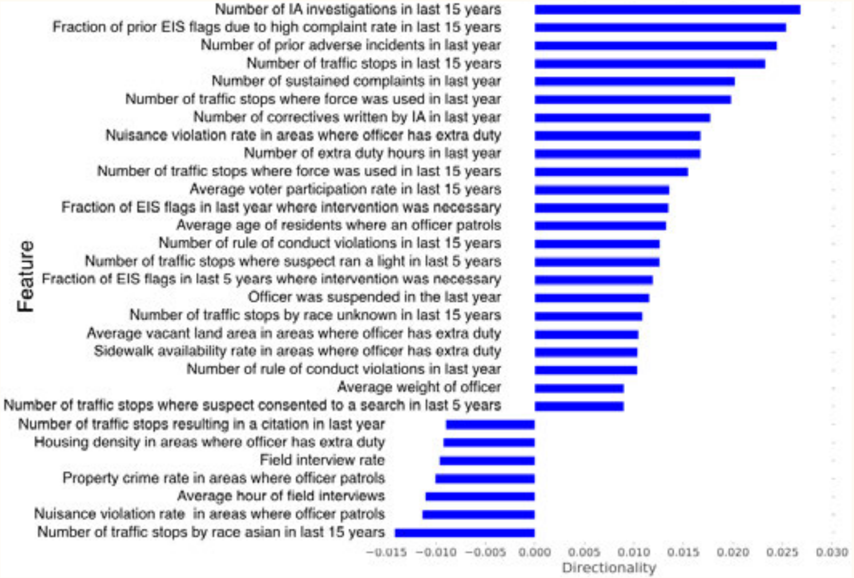
Parameterizing energy/computation budget?

- "Show your work: Improved reporting of experimental results" (Dodge et al., 2019)
- If you optimize for accuracy *and* energy usage, different models come out on top



Studying up?

- “Identifying police officers at risk of adverse events” (Carton et al., 2016)
  - I.e., police brutality
- Why not subject those in power to the same “algorithmic” surveillance?
- (See also: “Studying up: reorienting the study of algorithmic fairness around issues of power”; Barabas et al., 2020)



Studying the whole system?

- "Algorithmic risk assessment in the hand of humans" (Stevenson & Doleac, 2019)
- Judges can ignore recommendations. They do so… but not at random
- Gives justification for giving harsher sentences to Black people ("the 'algorithm' says so!"
- But judges exercise discretion with younger defendants (the innocence of youth)
- (Also looks at the aggregate impact of the system)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Diverted | risk = low						
Alternative risk score	0.013					0.010
	(0.010)					(0.010)
Black		-0.015				-0.014
		(0.015)				(0.016)
Unemployed			0.025			0.009
			(0.017)			(0.018)
Female				0.040**		0.038**
				(0.016)		(0.017)
Age<23					0.069****	0.065***
					(0.020)	(0.020)
Observations	3943	3943	3943	3943	3943	3943
R^2	0.204	0.204	0.204	0.205	0.206	0.280
Mean DV	0.44	0.44	0.44	0.44	0.44	0.44
Panel B: Diverted | risk = high						
Alternative risk score	-0.004					-0.007
	(0.005)					(0.005)
Black		-0.029***				-0.045****
		(0.010)				(0.012)
Unemployed			0.043****			0.018
			(0.012)			(0.012)
Female				0.038***		0.040***
				(0.013)		(0.014)
Age<23					0.065****	0.058****
					(0.011)	(0.011)
Observations	7598	7598	7598	7598	7598	7598
R^2	0.142	0.143	0.144	0.143	0.146	0.197
Mean DV	0.16	0.16	0.16	0.16	0.16	0.16

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

Irony?

- "Predicting financial crime: Augmenting the predictive policing arsenal" (Clifton et al., 2017)
- Not something that we'd expect will actually be taken up: but this is what it would look like if we treated white people's crimes as symmetric with those of Black people

Fig. 6: The WCCEWS user interface. The map shows downtown Manhattan in New York City, NY. Color indicates the predicted density of white collar criminal activity. The left-hand panel shows the Top Risk Likelihoods of the listed crimes occurring within the selected geohash. Below is a histogram indicating predicted Approximate Crime Severity associated with discrete brackets of violation amount in $USD. Finally, the panel lists Potential Offenders operating within the selected geohash, and a generalized white collar criminal subject.

Re-application?

- "To predict and serve?" (Lum & Isaac, 2016)
- Re-applying PredPol on the basis of national surveys of drug use, rather than prior "crime" data

FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

Accidental absurdity? *(Or is it?)*

- "Pigeons as trainable observers of pathology and radiology breast cancer images" (Levenson et al. 2015)
- Pigeons as competitive with deep learning. What does that mean?
  - Can say that it's because there is a connection (Lisa Zhang, "From pigeons to Artificial Neural Networks")
- Or, ML as equally theoretically absurd ("Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI"; Alvarado, 2022)

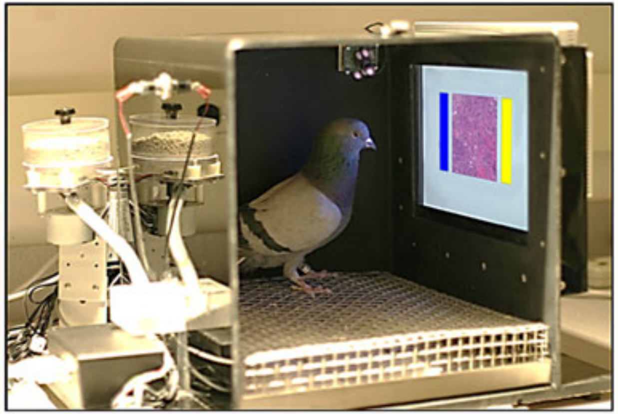
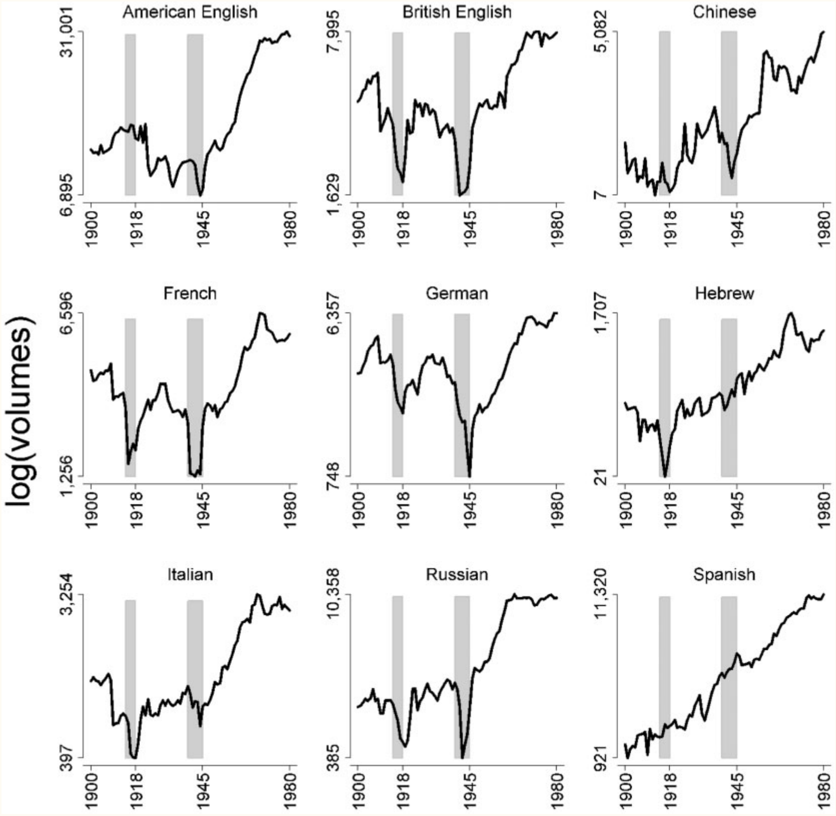


Fig 1. The pigeons' training environment. The operant conditioning chamber was equipped with a food pellet dispenser, and a touch-sensitive screen upon which the medical image (center) and choice buttons (blue and yellow rectangles) were presented.

Clean up?

- "The impact of lacking metadata for the measurement of cultural and linguistic changes using the Google Ngrams data sets—Reconstructing the composition of the German corpus in times of WWII" (Koplenig, 2015)
- Critique of "culturomics" paper that uses Ngram frequencies to make sociological and historical claims
- (See also: "Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution"; Pechenick et al., 2015)



Accidental reflexivity?

- "Human Rademacher Complexity" (Zhu et al., 2009)
- Had people try to remember/"learn" random labels: quantifies how we interpret signal in pure noise
- Can use to say in an ML-"legitimate" way, for example, how much people might interpret any clustering output as valid, if primed to do so

Theorem 1. *Let $\mathcal{F}$ be a set of functions mapping to $[-1,1]$. For any integers n, m,*

$$P\left\{\left|R(\mathcal{F},\mathcal{X},P_X,n) - \frac{1}{m}\sum_{j=1}^{m} \sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n} \sigma_i^{(j)} f(x_i^{(j)})\right|\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{\epsilon^2 nm}{8}\right) \qquad (2)$$

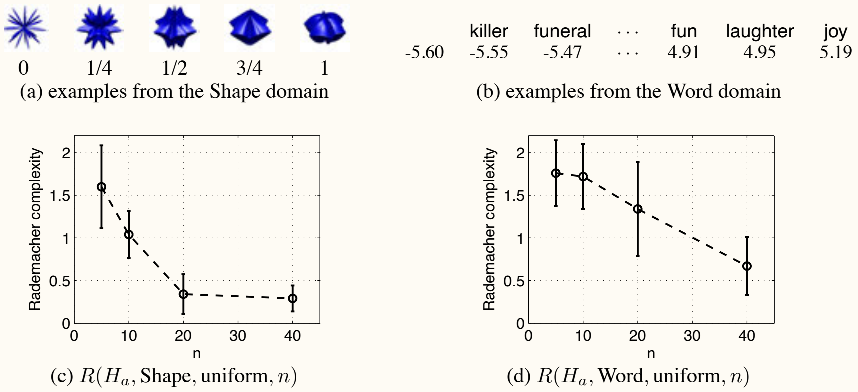


Figure 1: Human Rademacher complexity on the "Shape" and "Word" domains.

Comparative study?

- "The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods" (Nelson et al., 2018)

Table 1. Evaluation Metrics for Each Automated Method.

Method (Coding Scheme)	Inequality/Relevant			Not Inequality/Irrelevant			Weighted Average[a]			Time Trends		Support
	Precision (1)	Recall (2)	F1-Score (3)	Precision (4)	Recall (5)	F1-Score (6)	Precision[b] (7)	Recall[b] (8)	F1 Score[b] (9)	Corr (Two-Year MA) (10)	Correlation (11)	*N* (12)
(1) Supervised ML[c]												
Relevant versus irrelevant (A)	.85	.90	.87	.81	.74	.77	.83 (.81–.86)	.84 (.81–.86)	.83 (.81–.86)	.75	.74	745
Inequality versus not inequality (B)	.73	.60	.66	.80	.88	.84	.78 (.74–.80)	.78 (.75–.80)	.78 (.74–.80)	.69	.63	745
Inequality versus economic versus irrelevant (C)	.67	.70	.69	.76	.84	.80	.68 (.64–.71)	.69 (.65–.71)	.69 (.64–.71)	.72	.69	745
(2) Dictionary												
Levay-Enns (D)	.91	.25	.40	.83	.99	.91	.85	.84	.80	.42	.59	1,253
McCall (B)	.48	.84	.61	.86	.52	.65	.73	.63	.64	.66	.44	1,253
(3) Unsupervised ML												
Topic model versus explicit (D)	.63	.45	.53	.86	.93	.90	.82	.83	.82	.58	.68	1,253
k-means versus explicit (D)	.88	.14	.24	.81	.99	.89	.83	.81	.76	N/A	N/A	1,253

Exploration?

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

- "Race, writing, and computation: Racial difference and the US novel, 1880-2000" (So et al., 2019)
- "Computational methods demand the quantification of one's objects of study. It's likely easier to accept measuring a novel's popularity by sales figures or classifying its genre by diction than labeling it according to discrete racial identifiers. Such labeling is an affront to critical race studies, which has taken as its very mission the deconstruction of racial categories."
- "Critical suspicion, of course, can also lead to critical adaptation."
- Carefully done, but underwhelming in the end

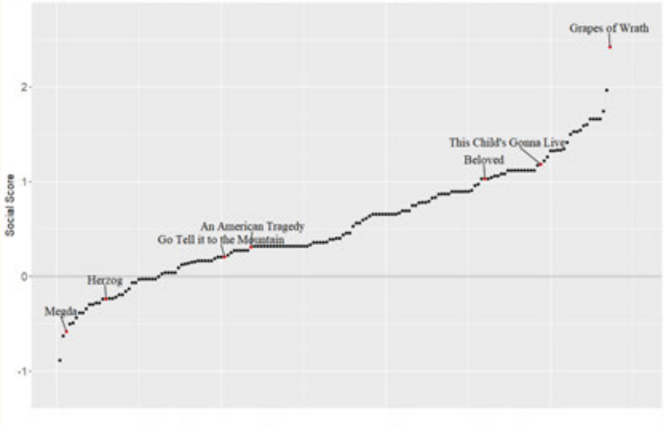
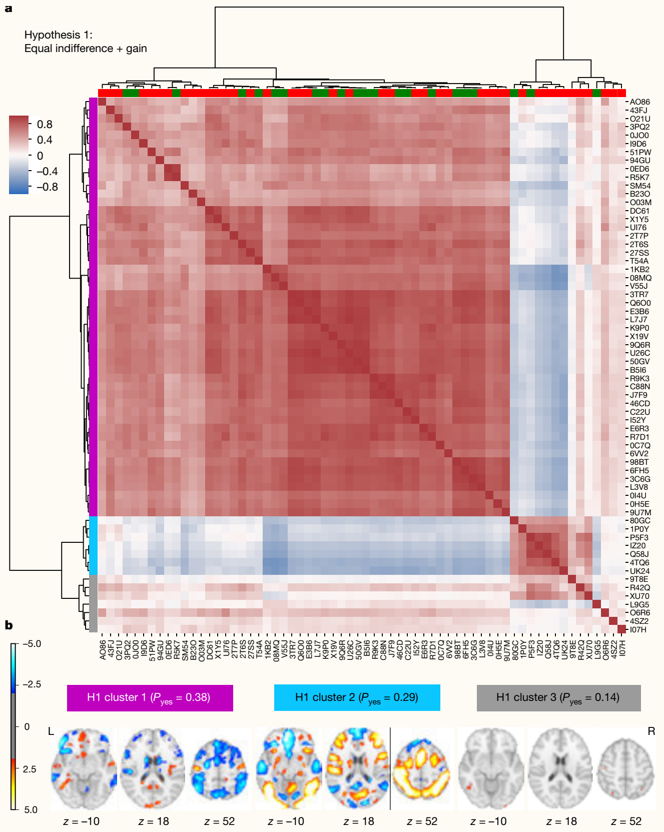


Figure 3. Plot showing the "social" score for all novels containing alignments with the Bible. Lower scores indicate novels where the Bible is less frequently cited in a "social" way, as we define the term. Scores closer to zero indicate novels where the "social" and "non-social" contexts are split evenly, as in James Baldwin's Go Tell it to the Mountain.

Investigating inconsistency?

- “Variability in the analysis of a single neuroimaging dataset by many teams” (2020)
- Lots of variation in analytic pipeline, and in results (but they suggest meta-analysis can resolve)
- (See also: “Many analysts, one data set: Making transparent how variations in analytic choices affect results”; Silberzahn et al., 2018)



Using math where appropriate

- "Recombination: A family of Markov chains for redistricting" (DeFord, Duchin, & Solomon, 2020)
- (Not actually "machine learning", but <u>actually</u> algorithms, specifically MCMC)
- Compare proposed redistricting to distributions of random valid redistricting

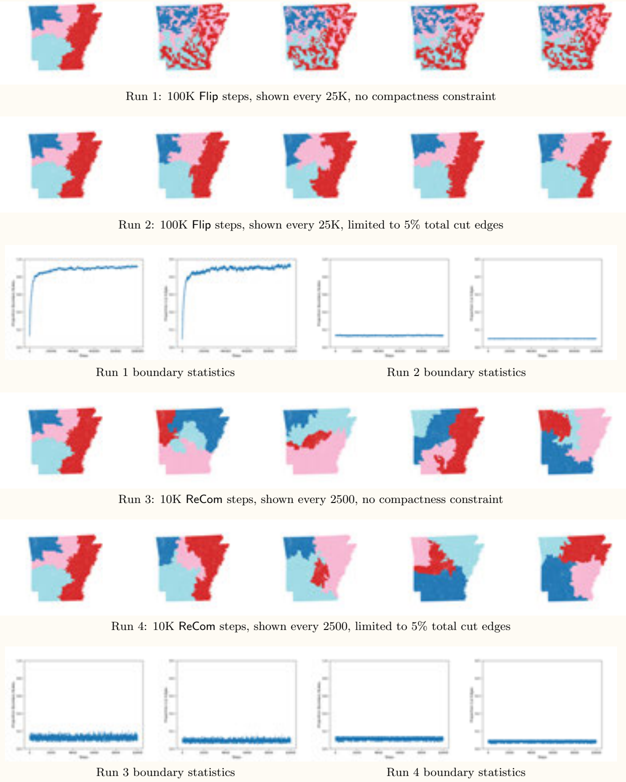
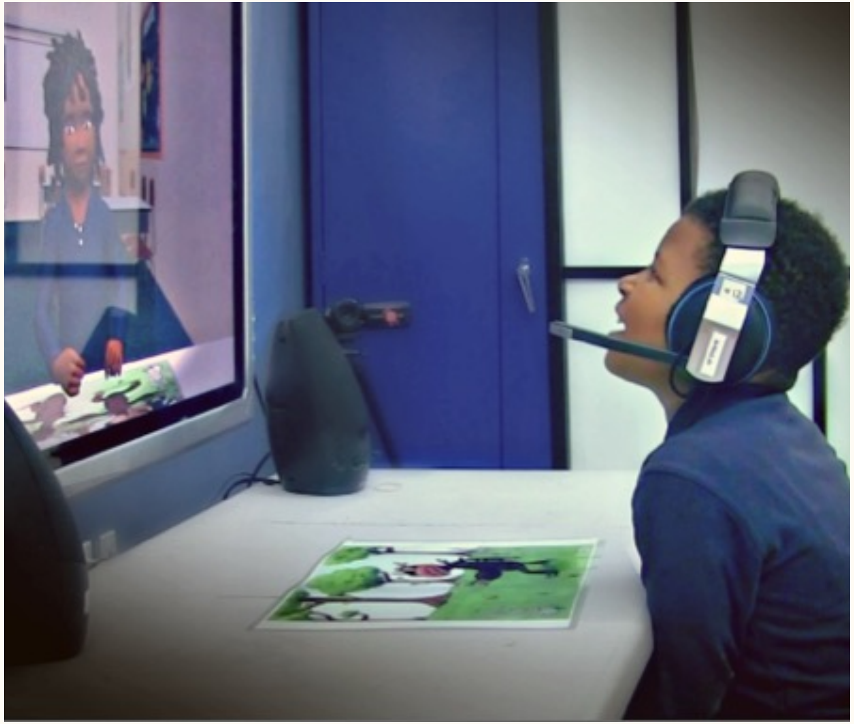


Figure 7: Arkansas block groups partitioned into $k = 4$ districts, with population deviation limited to 5% from ideal. Imposing a compactness constraint makes the Flip chain unable to move very far.

ICQCM

Building systems?

- "'Alex speaks with my voice!' Promoting science discourse with bidialectal virtual peers" (Finkelstein, 2017)
- An AAVE virtual conversational agent



ICQCM

Statistical critique?

- "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US Presidential Election" (Meng, 2018)
- Derives a new fundamental statistical equation quantifying the effect of data quality/bias

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}}$$

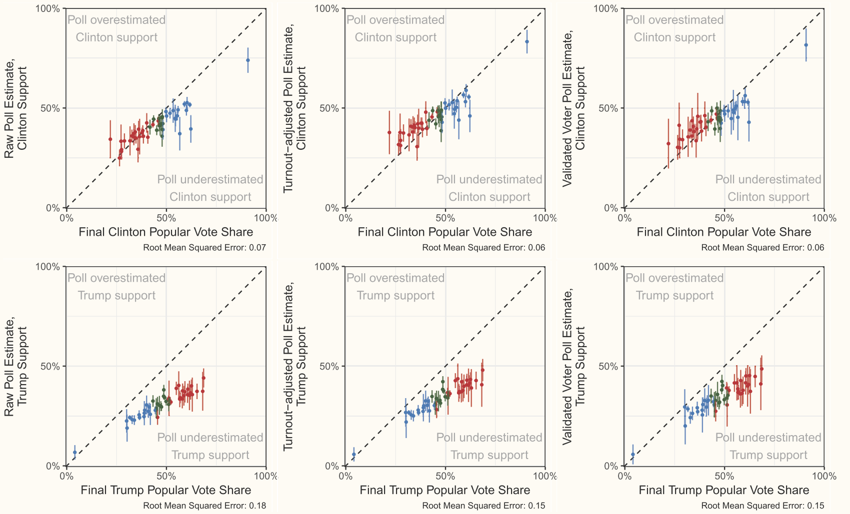


FIG. 4. *Comparison of actual vote shares with CCES estimates (and 95% confidence interval) across 50 states and DC. Top row for Clinton; bottom row for Trump. Color indicates a state's partisan leanings in 2016 election: solidly Democratic (blue), solidly Republican (red), or swing state (green). The left plot uses sample averages of the raw data ($n = 64{,}600$) as estimates; the middle plot uses estimates weighted to likely voters according to turnout intent (estimated turnout $\hat{n} = 48{,}106$); and the right plot uses sample averages among the subsample of validated voters (subsample size $n = 35{,}829$). Confidence intervals based on unweighted sample proportions are computed following (3.9), where the use of SRS variances can be conservative given the stratified design of the survey, and yet they still do not provide any realistic protection against the increased MSE caused by the non-response bias. For the turnout adjusted estimate, which is in a ratio form, a δ-method is employed to approximate its variance, which is then used to construct confidence intervals.*

Statistical critique?

- "The grand leap" (Humphreys & Freedman, 1996)
- (See also: "Graphical models for causation, and the identification problem"; Freedman, 2004)
- Specifically a critique of a rather narrow subfield of "causal learning", but it's a subfield that is gaining some attention

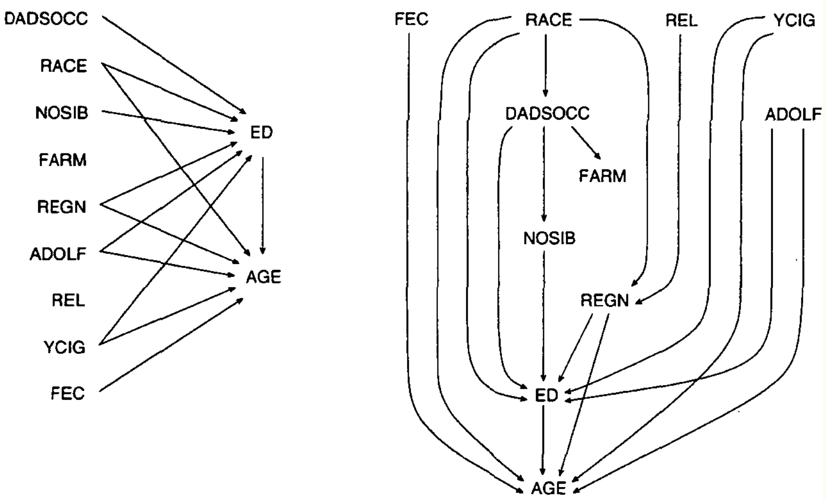
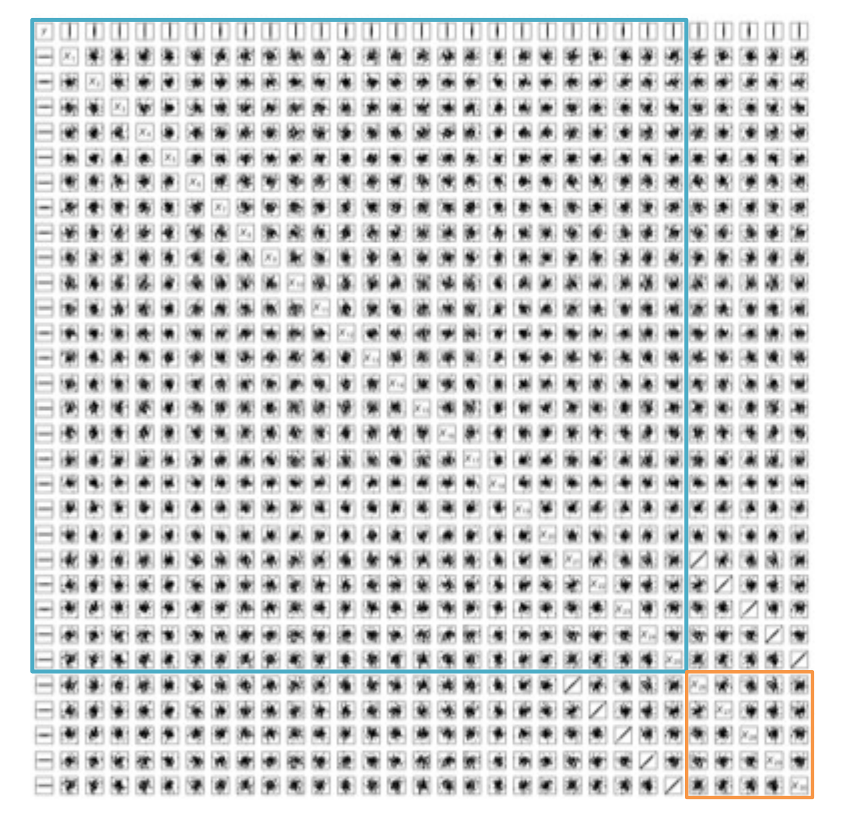
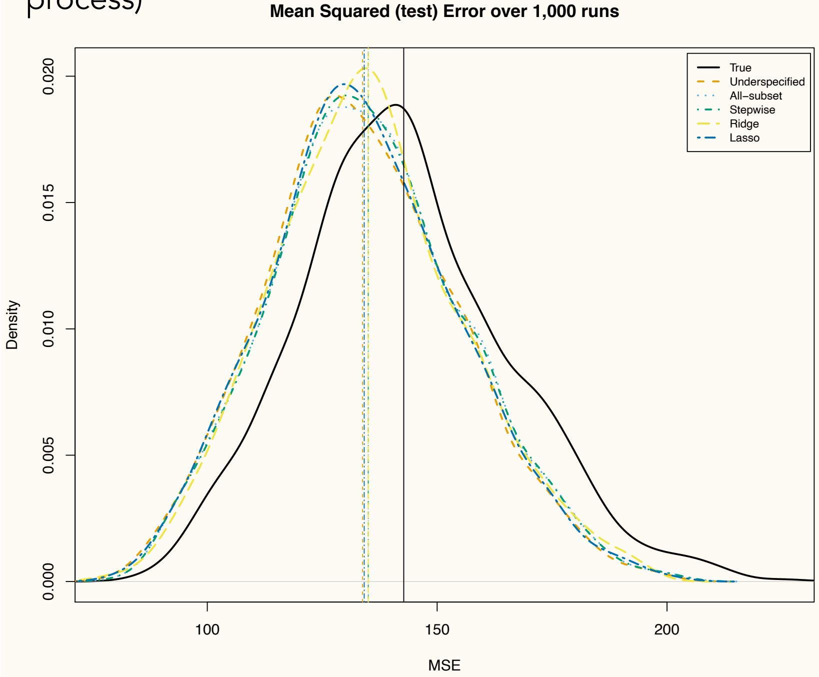


Fig.3. The left-hand panel shows the model reported by SGS (p.140).The right-hand panel shows the whole graph produced by the SGS search program TETRAD.[16]

Statistical critique?

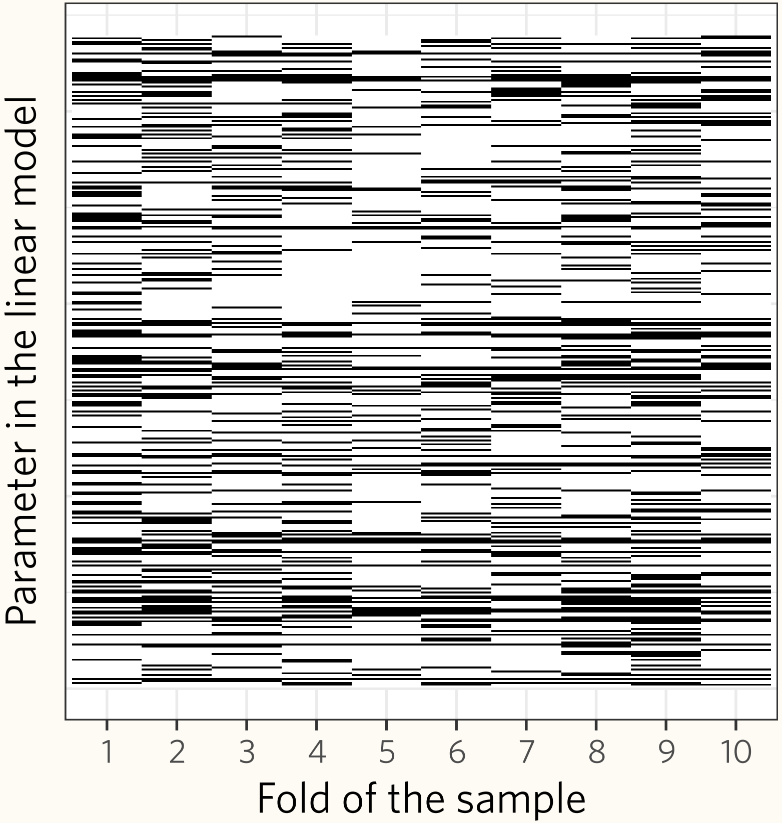


Simulation of a point in "To explain or to predict?" (Shmueli, 2010): that 'false' models can "predict" (i.e., fit) better than the 'true' model (the data-generating process)



Statistical clarification?

- "Machine learning: An applied econometric approach" (Mullainathan & Spiess, 2017)
- Very different sets of correlations can "predict" (fit) equally well
  - Leo Breiman (2001) called this the "Rashomon Effect"
- But different fits suggest very different interventions



Possibilities

Graphical models to express

- "Racial categories in machine learning" (Benthall & Haynes, 2018): graphical model to express causes of ~~race~~ racialization
- (See also: "Towards a critical race methodology in algorithmic fairness"; Hanna et al., 2020)
- When used for expression, are basically path diagrams, but calling them "graphical models" (and following additional rules) makes them more formal

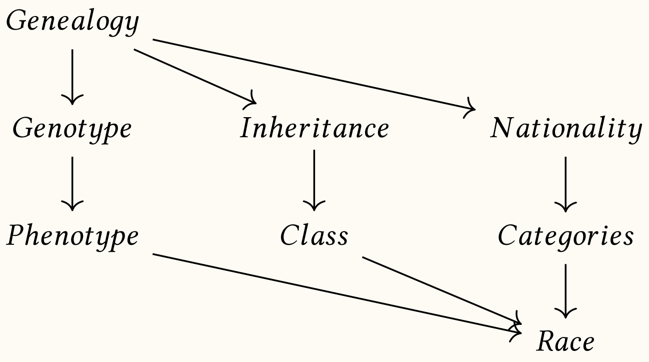
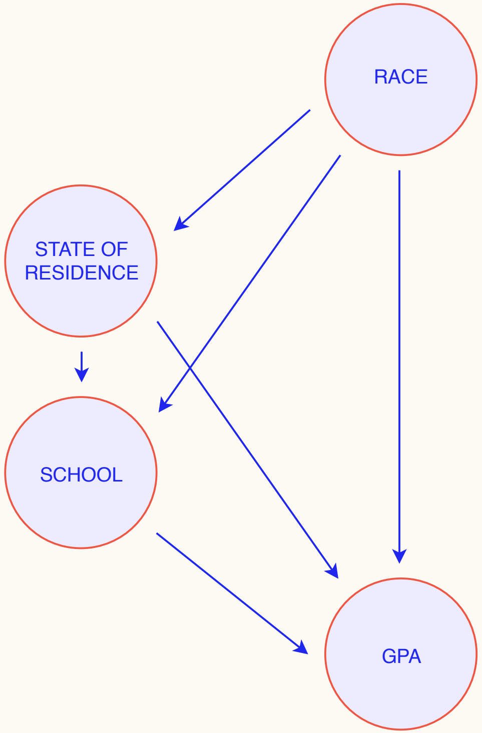


Figure 1: A model of how individual biological properties (genealogy, genotype, and phenotype) are racialized through national political categories and associations with socioeconomic class. Here inheritance refers to all forms of capital, including economic and social, passed from parents to children. Broadly speaking, genealogy is a strong determiner of race, but importantly as a common cause of phenotype, class, and nationally recognized racial categories, which are separate components of racial classification.

Graphical models as guide to critique

- "Disparate causes, pt. I" and "Disparate causes, pt. II" (Hu, 2019ab)
  - Would need to backtrack through entire life history to really account for race as a counterfactual
  - Expressed with reference to graphical models
- (See also: "Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination"; Kohler-Hausmann, 2019)

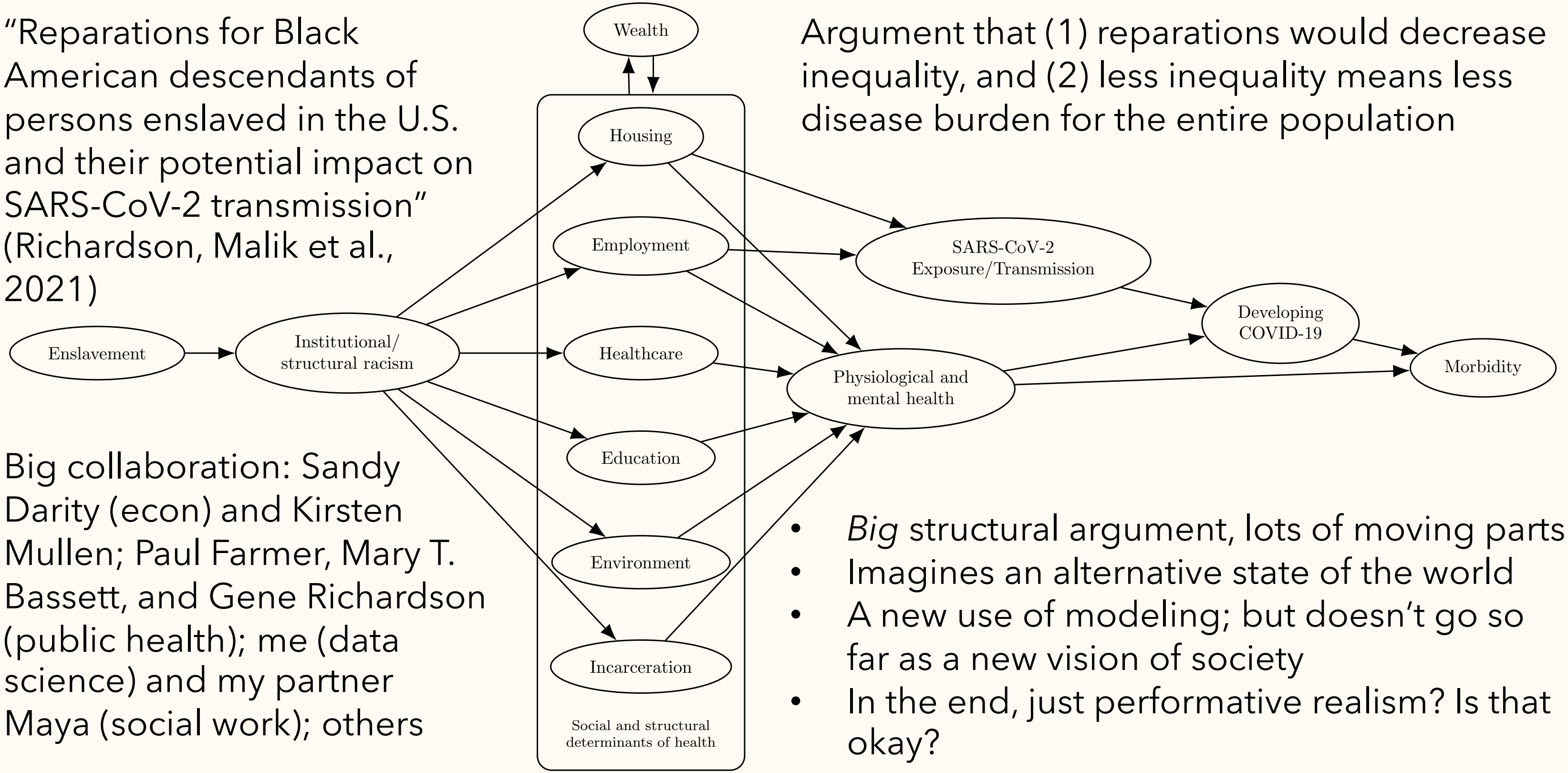


Hu 2019b

Reparations and COVID-19 paper

"Reparations for Black American descendants of persons enslaved in the U.S. and their potential impact on SARS-CoV-2 transmission" (Richardson, Malik et al., 2021)

Argument that (1) reparations would decrease inequality, and (2) less inequality means less disease burden for the entire population



Big collaboration: Sandy Darity (econ) and Kirsten Mullen; Paul Farmer, Mary T. Bassett, and Gene Richardson (public health); me (data science) and my partner Maya (social work); others

- *Big* structural argument, lots of moving parts
- Imagines an alternative state of the world
- A new use of modeling; but doesn't go so far as a new vision of society
- In the end, just performative realism? Is that okay?

Math to make a point (Malik, 2020)

$$
\begin{aligned}
\mathrm{err}(\hat{\mu}) &= \frac{1}{n}\mathbb{E}_f\|Y - \widehat{Y}\|_2^2 \\
&= \frac{1}{n}\left[\mathbb{E}_f\|Y\|_2^2 + \mathbb{E}_f\|\widehat{Y}\|_2^2 - 2\mathbb{E}_f(Y^T\widehat{Y})\right] \\
&= \frac{1}{n}\left[\mathbb{E}_f\|Y\|_2^2 + \mathbb{E}_f\|\widehat{Y}\|_2^2 - 2\,\mathrm{tr}\,\mathbb{E}_f(Y\widehat{Y}^T)\right] \\
&\quad + \frac{1}{n}\left[\ \mu^T\mu + \mathbb{E}_f(\widehat{Y})^T\mathbb{E}_f(\widehat{Y}) + 2\,\mathrm{tr}\,\mu\mathbb{E}_f(\widehat{Y})^T\right] \\
&\quad + \frac{1}{n}\left[-\mu^T\mu - \mathbb{E}_f(\widehat{Y})\mathbb{E}_f(\widehat{Y})^T - 2\mu^T\mathbb{E}_f(\widehat{Y})\right] \\
&= \frac{1}{n}\Big[\underbrace{\mathrm{tr}\,\Sigma}_{\text{Irreducible error}} + \underbrace{\|\mu - \mathbb{E}(\widehat{Y})\|_2^2}_{\text{Bias squared}} + \underbrace{\mathrm{tr}\,\mathrm{Var}_f(\widehat{Y})}_{\text{Variance}} - \underbrace{2\,\mathrm{tr}\,\mathrm{Cov}_f(Y,\widehat{Y})}_{\text{“Optimism”}}\Big]
\end{aligned}
$$

ICQCM

Apply this to non-iid data

- Imagine we have, for $\mathbf{\Sigma}_{ii} = \sigma^2$ and $\mathbf{\Sigma}_{ij} = \rho\sigma^2, \quad i \neq j$

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix}\beta, \begin{bmatrix} \mathbf{\Sigma} & \rho\sigma^2\mathbf{1}\mathbf{1}^T \\ \rho\sigma^2\mathbf{1}\mathbf{1}^T & \mathbf{\Sigma} \end{bmatrix}\right)$$

- Then, optimism in the training set is:

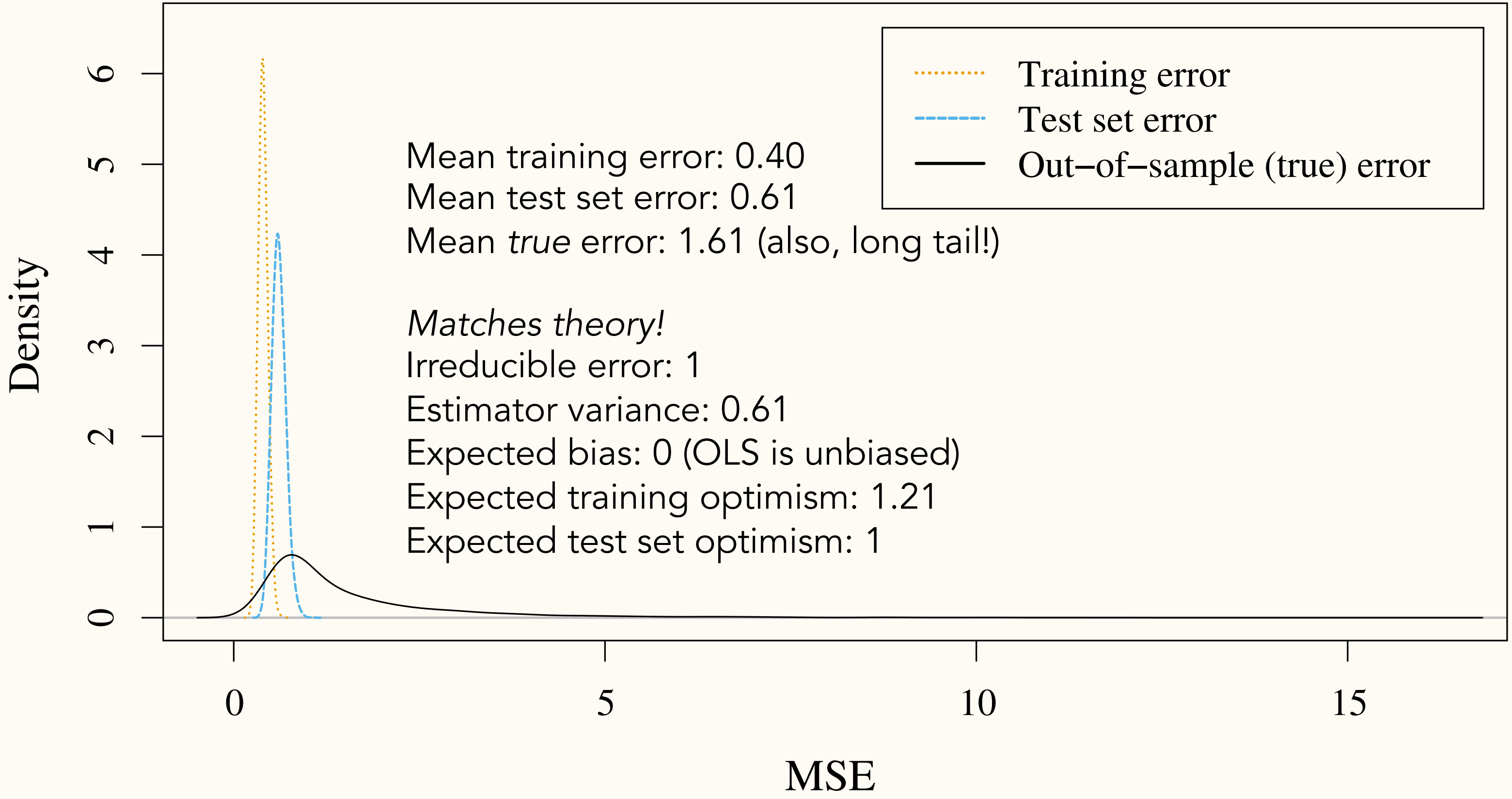
$$\frac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_1, \widehat{Y}_1) = \frac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_1, \mathbf{H}Y_1) = \frac{2}{n}\operatorname{tr}\mathbf{H}\operatorname{Var}_f(Y_1) = \frac{2}{n}\operatorname{tr}\mathbf{H\Sigma}$$

- But test set also has nonzero optimism!

$$\frac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_2, \widehat{Y}_1) = \frac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_2, \mathbf{H}Y_1) = \frac{2\rho\sigma^2}{n}\operatorname{tr}\mathbf{H}\mathbf{1}\mathbf{1}^T = 2\rho\sigma^2$$

Simulating the toy example

Out-of-sample MSE: *much* worse!



Implications

- *No amount of data is ever enough!*
- We need to make independence/ exchangeability assumptions to model any data (we can test these assumptions with additional data… but only if we assume that the additional data are independent)
- (Opsomer et al., 2001: mean function and covariance function are not jointly identifiable nonparametrically)
- Using statistical theory to make a critical point about the limitations of machine learning (and statistics)

Conclusion and discussion

Some things we covered

- Auditing
- Reverse-engineering
- Scalable measurement
- Qualitative follow-up
- Qualitative evaluation
- Qualitative rigor (in context)
- Adding appropriate complexity
- Investigating context
- Parameterizing budget
- Studying up
- Studying the whole system
- Re-application
- Absurdity
- Clean-up
- Reflexivity
- Comparative study
- Exploration
- Investigating inconsistency
- Using math where appropriate
- Building systems
- Statistical critique/clarification

ICQCM

Reminder: What's the point?

With increasing depth:

- Rigor. Just doing things more carefully and outlining limitations (including sociotechnical)
- Critiquing/destabilizing machine learning: listing limitations, maybe absurdity (!)
- Strategic positivism: using machine learning to add legitimacy (risks reifying)
  - (See also: "Roles for computing in social change", Abebe et al., 2020; but maybe not strategic, maybe just positivism)
- Imagining, mixed methods

References (1/5)

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 252–260). https://doi.org/10.1145/3351095.3372871

Alvarado, R. (2022). Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics, 36*(2), 121–133. https://doi.org/10.1111/bioe.12959

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160). https://doi.org/10.1111/josl.12080

Barabas, C., Doyle, C., Rubinovitz, J. B., & Dinakar, K. (2020). Studying up: Reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 167–176). https://dl.acm.org/doi/abs/10.1145/3351095.3372859

Bates, J., Cameron, D., Checco, A., Clough, P., Hopfgartner, F., Mazumdar, S., Sbaffi, L., Stordy, P., & de la Vega de León, A. (2020). Integrating FATE/critical data studies into data science curricula: Where are we going and how do we get there? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 425–435). https://doi.org/10.1145/3351095.3372832

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (pp. 289–298). https://doi.org/10.1145/3287560.3287575

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R. G., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature, 582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3). https://doi.org/10.1214/ss/1009213726

Buolamwini, J., & Gebru, T. (2019). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). https://proceedings.mlr.press/v81/buolamwini18a.html

Carton, S., Helsby, J., Joseph, K., Mahmud, A., Park, Y., Walsh, J., Cody, C., Patterson, C. E., Haynes, L., & Ghani, R. (2016). Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. https://doi.org/10.1145/2939672.2939698

Clifton, B., Lavigne, S., & Tseng, F. (2017). Predicting financial crime: Augmenting the predictive policing arsenal. For *The New Inquiry Magazine* (Vol. 59: ABOLISH). https://whitecollar.thenewinquiry.com/

ICQCM

References (2/5)

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., … Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. http://arxiv.org/abs/2011.03395

DeFord, D., Duchin, M., & Solomon, J. (2021). Recombination: A family of Markov Chains for redistricting. *Harvard Data Science Review, 3*(1). https://doi.org/10.1162/99608f92.eb30390f

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 2185–2194). https://aclanthology.org/D19-1224.pdf

Finkelstein, S. (2017). *'Alex speaks with my voice!' Promoting science discourse with bidialectal virtual peers*. [Doctoral dissertation, Carnegie Mellon University].

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review, 28*(4), 267–293. https://doi.org/10.1177/0193841x04266432

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16). https://doi.org/10.1073/pnas.1720347115

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. https://doi.org/10.1145/3351095.3372862

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender recognition or gender reductionism? The social implications of automatic gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. https://doi.org/10.1145/3173574.3173582

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 501–512). https://doi.org/10.1145/3351095.3372826

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers' role and perspective. *Educational Technology Research and Development, 67*(5), 1273–1306. https://doi.org/10.1007/s11423-019-09685-0

Herodotou, C., Rienties, B., Verdin, B., & Boroowa, A. (2019). Predictive learning analytics "at scale": Guidelines to successful implementation in higher education. *Journal of Learning Analytics, 6*(1). https://doi.org/10.18608/jla.2019.61.5

References (3/5)

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., & Mascolo, C. (2016). Measuring urban social diversity using interconnected geo-social networks. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16).* https://doi.org/10.1145/2872427.2883065

Hu, L. (2019a). Disparate causes, pt. I. *Phenomenological World*. https://www.phenomenalworld.org/analysis/disparate-causes-i/

Hu, L. (2019b). Disparate causes, pt. II. *Phenomenological World*. https://www.phenomenalworld.org/analysis/disparate-causes-pt-ii/

Humphreys, P., & Freedman, D. (1996). The grand leap [Review of *Causation, prediction, and search*, by P. Spirtes, C. Glymour, & R. Scheines]. *The British Journal for the Philosophy of Science, 47*(1), 113–123. http://www.jstor.org/stable/1215829

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* https://doi.org/10.1145/3351095.3372829

Jones, M. L. (2018). How we became instrumentalists (again): Data positivism since WWII. *Historical Studies in the Natural Sciences, 48*(5), 673–684. https://doi.org/10.1525/hsns.2018.48.5.673

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. In *Proceedings of the ACM on Human-Computer Interaction* (Vol. 2, Issue CSCW, pp. 1–22). https://doi.org/10.1145/3274357

Kohler-Hausmann, I. (2019). Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review, 113*(5), 1163–1228. https://scholarlycommons.law.northwestern.edu/nulr/vol113/iss5/6/

Koplenig, A. (2015). The impact of lacking metadata for the measurement of cultural and linguistic changes using the Google Ngrams data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities, 32*(1), 169–188. https://doi.org/10.1093/llc/fqv037

Levenson, R. M., Krupinski, E. A., Navarro, V. M., & Wasserman, E. A. (2015). Pigeons (*Columba livia*) as trainable observers of pathology and radiology breast cancer images. *PLOS ONE, 10*(11), e0141357. https://doi.org/10.1371/journal.pone.0141357

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship. *Queue, 17*(1), 45–77. https://doi.org/10.1145/3317287.332853

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mackenzie, A. (2017). *Machine learners: Archaeology of a data practice.* MIT Press.

Malik, M. M. (2020). A hierarchy of limitations in machine learning. https://arxiv.org/abs/2002.05193

ICQCM

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

References (4/5)

Meng, X.-L. (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

Moats, D., & Seaver, N. (2019). "You social scientists love mind games": Experimenting in the "divide" between data science and critical algorithm studies. *Big Data & Society, 6*(1), 205395171983340. https://doi.org/10.1177/2053951719833404

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big Data, 5*(2), 85–97. https://doi.org/10.1089/big.2016.0050

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2018). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research, 50*(1), 202–237. https://doi.org/10.1177/0049124118769114

Opsomer, J., Wang, Y., & Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science, 16*(2), 134–153. https://doi.org/10.1214/ss/1009213287

Pasquinelli, M. , & Joler, V. (2020). The Nooscope manifested: Artificial Intelligence as instrument of knowledge extractivism. https://nooscope.ai

Patton, D. U., Blandfort, P., Frey, W. R., Schifanella, R., McGregor, K., & Chang, S.-F. U. (2020). VATAS: An open-source web platform for visual and textual analysis of social media. *Journal of the Society for Social Work and Research, 11*(1), 133–155. https://doi.org/10.1086/707667

Patton, D. U., Frey, W. R., McGregor, K. A., Lee, F.-T., McKeown, K., & Moss, E. (2020). Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20).* https://doi.org/10.1145/3375627.3375841

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics, 7*, 677–694. https://doi.org/10.1162/tacl_a_00293

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE, 10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19).* https://doi.org/10.1145/3306618.3314244

Richardson, E. T. (2020). *Epidemic illusions: On the coloniality of global public health*. MIT Press.

ICQCM

Overview

Background/ review

Scope; review of some critiques

Examples taking, or enabling, a critical approach

Possibilities

Conclusion and discussion

References

References (5/5)

Richardson, E. T., Malik, M. M., Darity, W. A., Jr., Mullen, A. K., Morse, M. E., Malik, M., Maybank, A., Bassett, M. T., Farmer, P. E., Worden, L., & Jones, J. H. (2021). Reparations for Black American descendants of persons enslaved in the U.S. and their potential impact on SARS-CoV-2 transmission. *Social Science & Medicine, 276*, 113741). https://doi.org/10.1016/j.socscimed.2021.113741

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.6ed64b30

Sefala, S., Gebru, T., Moorosi, N., & Klein, R. (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

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-sts330

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

So, R., Long, H., & Zhu, Y. (2019). Race, writing, and computation: Racial difference and the US novel, 1880-2000. *Journal of Cultural Analytics, 3*(2). https://doi.org/10.22148/16.031

Stevenson, M., & Doleac, J. L. (2021). Algorithmic risk assessment in the hands of humans. *SSRN*. http://dx.doi.org/10.2139/ssrn.3489440

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th *Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (pp. 3645–3650). https://aclanthology.org/P19-1355.pdf

Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. In *Ethics of Data and Analytics* (pp. 342–349). Auerbach Publications. https://doi.org/10.1201/9781003278290-50

Wagstaff, K. (2012). Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)* (pp. 529-536). https://arxiv.org/abs/1206.4656

Wyly, E. (2009). Strategic positivism. *The Professional Geographer, 61*(3), 310–322. https://doi.org/10.1080/00330120902931952

Zhang, L. (n.d.). From pigeons to artificial neural networks. https://www.cs.toronto.edu/~lczhang/360/lec/w01/pigeon.html

Zhu, J., Gibson, B., & Rogers, T. T. (2009). Human Rademacher Complexity. In *Advances in Neural Information Processing Systems 22 (NIPS 2009).* https://papers.nips.cc/paper/2009/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html