A critical perspective on measurement in digital trace data and machine learning, and implications for demography

Momin M. Malik

Senior Data Science Analyst – AI Ethics, Mayo Clinic

Fellow, Institute in Critical Quantitative, Computational, & Mixed Methodologies

Instructor, University of Pennsylvania School of Social Policy & Practice

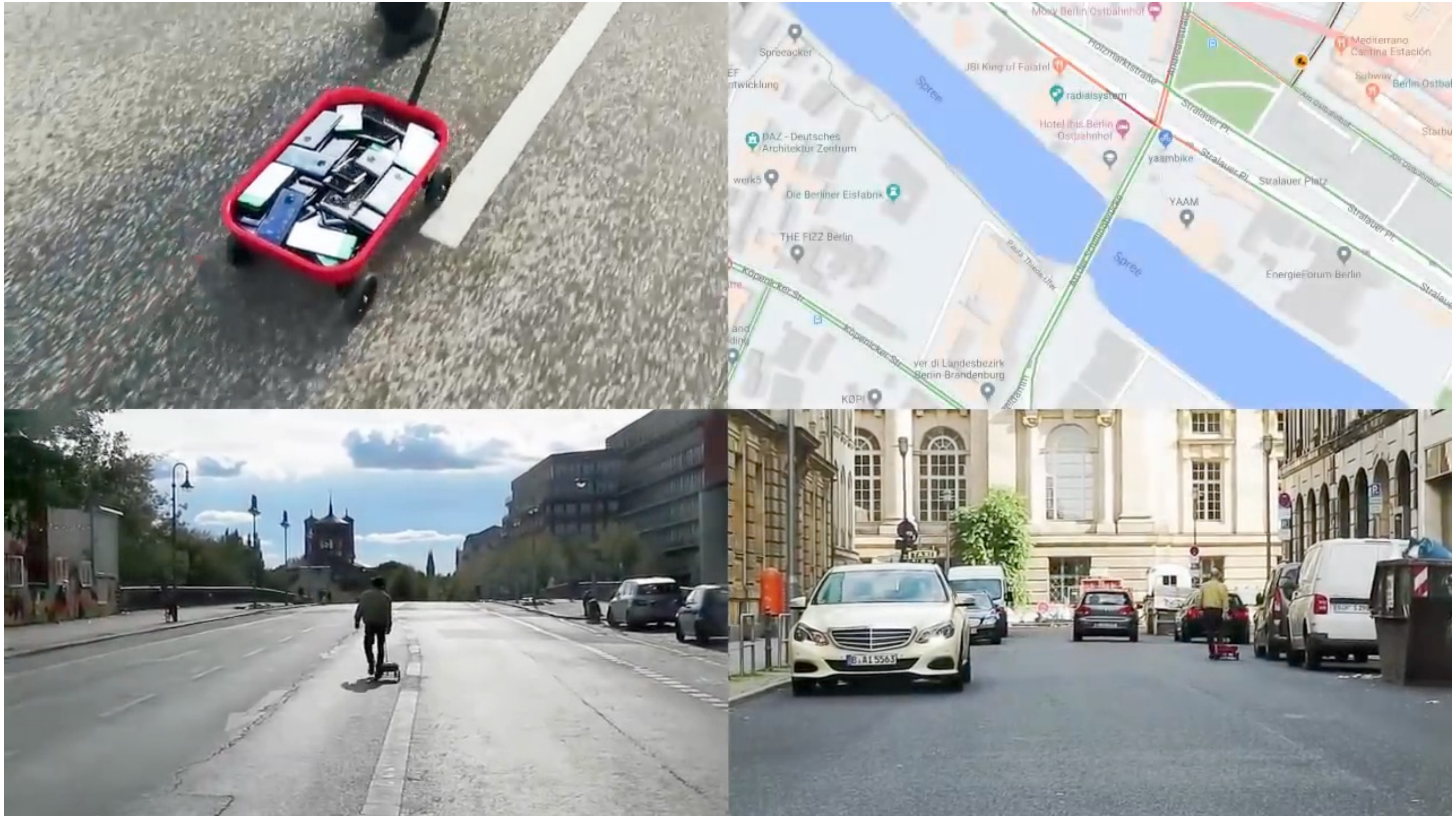
Tuesday, 26 April 2022

Max Planck Institute for Demographic Research

MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Simon Weckert, “Google Maps Hack”



This shows larger themes

- Available data are often only a *proxy*
- So long as the proxy is never the thing itself, it can fail
  - But by interrogating proxies, especially ones we did not construct, we can better understand them
- *Models* of relationships and processes, too, are not the things themselves
- Box (1979): *"[For] a model there is no need to ask the question 'Is the model true?'. If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is 'Is the model illuminating and useful?'."*

Quick survey

- How many people know of Savage and Burrows (2007)? Breiman (2001)?
- What disciplinary backgrounds?
  - Computer science?
  - Statistics? (Math/economics?)
  - Social science?
- How much do you know what machine learning is (or use it)?
  - How is it different from statistics?

Goals and outline

- Brief historical tour: Savage and Burrows (2007) and Breiman (2001)
  - About me
- Bias in geotagged tweets (ICWSM-2015 SPSM)
- Platform effects (ICWSM-2016)
- Hierarchy of limitations in machine learning (2020)
- Problems of cross-validation
- Summary and conclusion

Introduction

Brief historical tour

Bias in geotagged tweets

Platform effects

Hierarchy of limitations in machine learning

Problems of cross-validation

Summary and conclusion

References

Brief historical tour

Two key historical pieces

- Savage & Burrows (2007): "The coming crisis of empirical sociology"
  - Before Anderson's "End of theory" (2008) and Lazer et al.'s "Computational social science" (2009)
- Breiman (2001): "Statistical modeling: The two cultures"
  - Even earlier
  - Includes seeds of things we aren't even fully talked about yet: from problems with interpretability, to limits of cross-validation, to multiplicity of models

"Coming crisis of empirical sociology" (2007)

"In 2004, [Savage] was enrolled in a [ESRC Research Methods festival] session designed to popularize social network methods. He talked about an ESRC-funded research project [on volunteer organizations]… **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… The resulting intensive study of the members' social ties was amongst the most detailed ever carried out in the UK."

"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.**"

“Statistical modeling: The two cultures” (2001)

“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.”

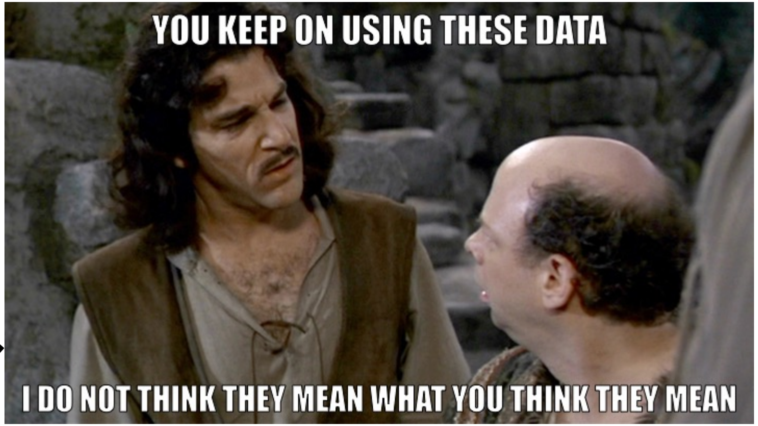
“Statistical modeling: The two cultures” (2001)

“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.”

About me

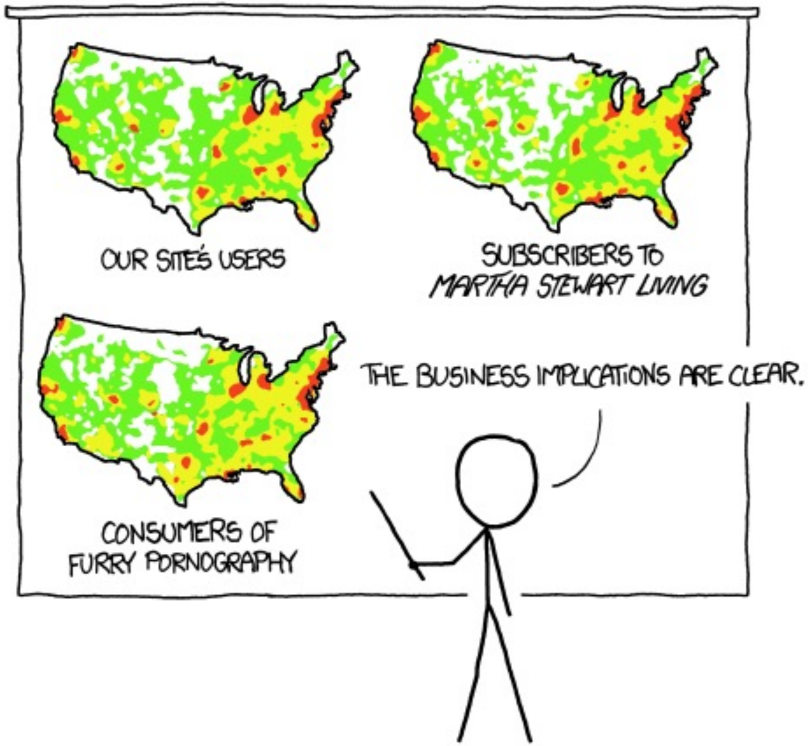
- UG: Department of the History of Science, Harvard University
- MSc: Oxford Internet Institute, University of Oxford
- PhD: Carnegie Mellon University School of Computer Science / Carnegie Mellon University Societal Computing
  - During: ML Machine Learning Department; Data Science For Social Good Summer Fellowship
- Post-doc: Berkman Klein Center for Internet & Society at Harvard University
- Previously: Avant-garde Health
- Currently: Mayo Clinic Center for Digital Health / Penn Social Policy & Practice, University of Pennsylvania / ICQCM Critical Data Science for a Diverse World



Bias in geotagged tweets

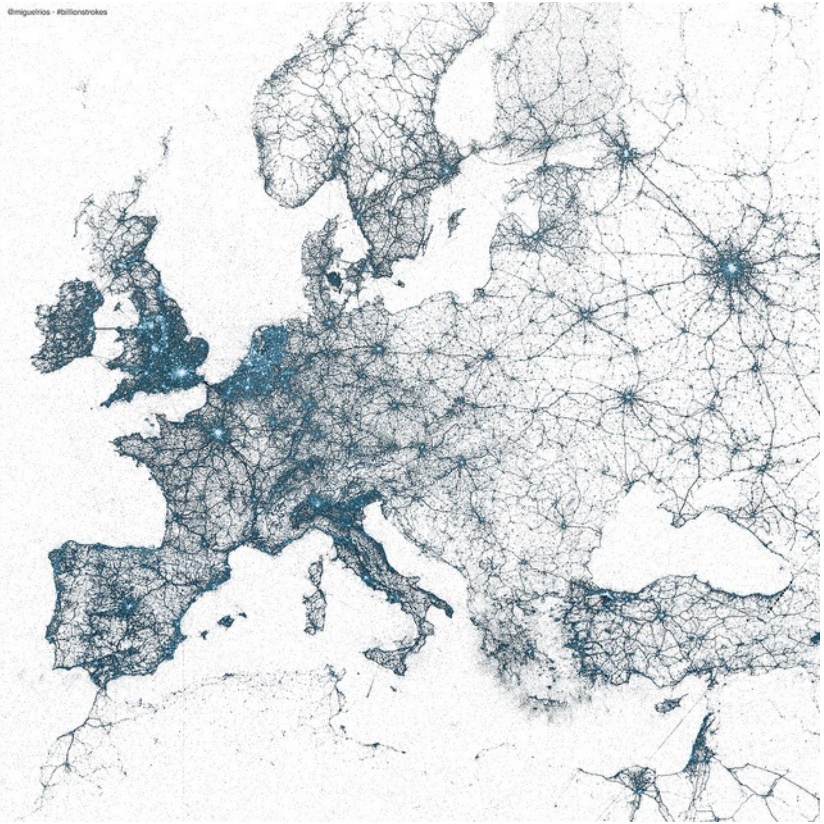
Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research* (ICWSM-15 SPSM), pages 18–27. May 26, 2015, Oxford, UK. Updated version (2018): https://www.mominmalik.com/malik_chapter1.pdf

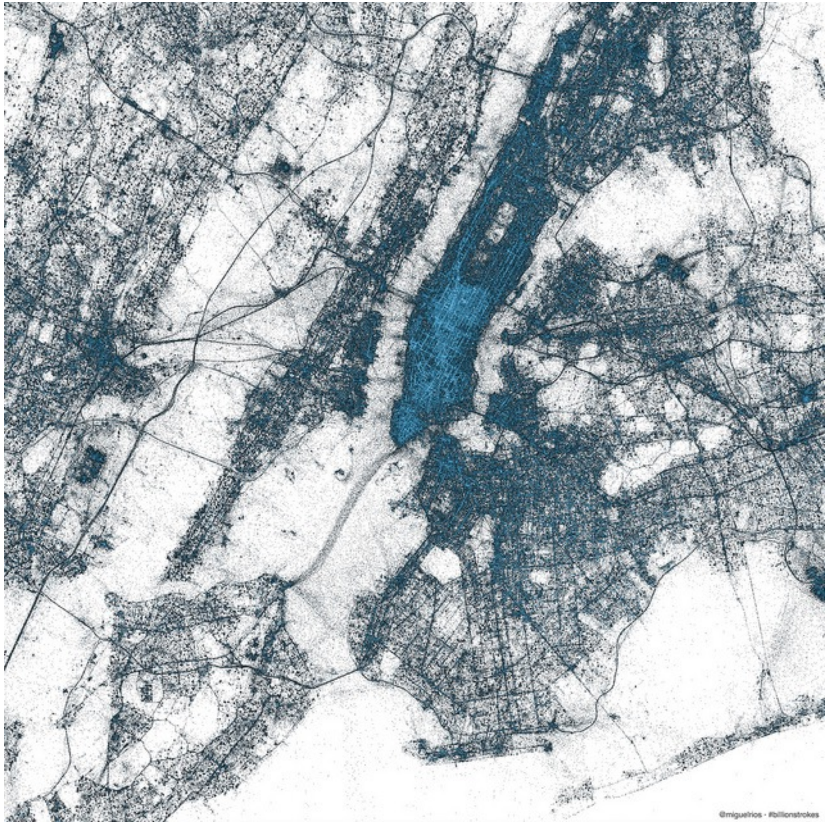
Many maps just show population



Randall Munroe. 2012. Heatmap. https://xkcd.com/1138/

But maybe we can use this?

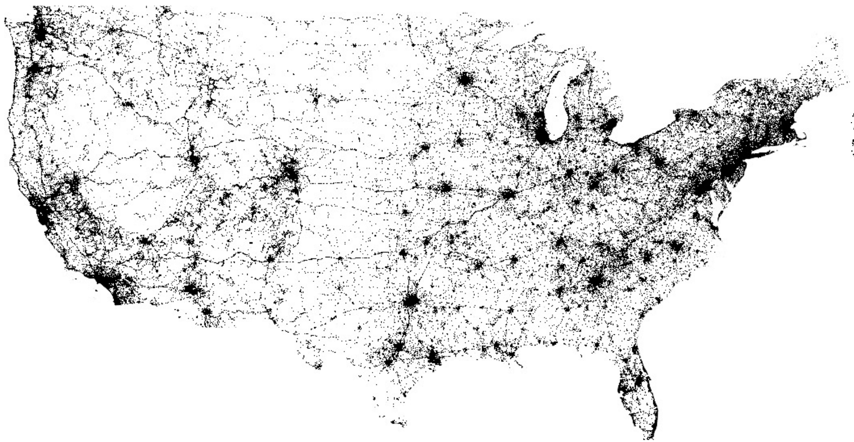




- Introduction
- Brief historical tour
- **Bias in geotagged tweets**
- Platform effects
- Hierarchy of limitations in machine learning
- Problems of cross-validation
- Summary and conclusion
- References

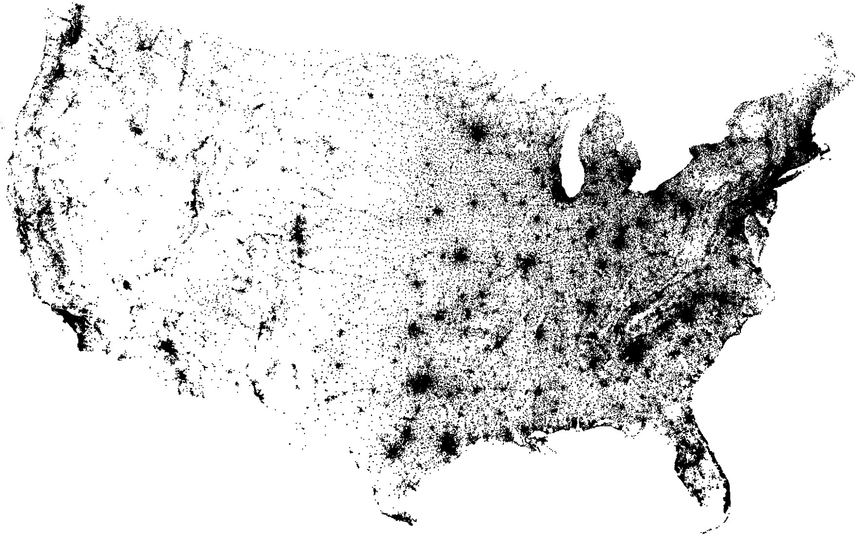
Do tweets measure population?

Geotagged tweets



Adapted from Eric Fischer, 2009, Contiguous United States geotag map. https://flic.kr/p/a7WMWS.

Population



Population density in 2010 US Census. Each square represents 1,000 people. Adapted from Geography Division, U.S. Department of Commerce / Economics and Statistics Administration / U.S. Census Bureau, Nighttime Population Distribution Wall Map.

Modeling population vs. users

- Users proportional to population:

$$U_i = \alpha P_i + \varepsilon_i P_i$$

- Take a log transformation (+Taylor):

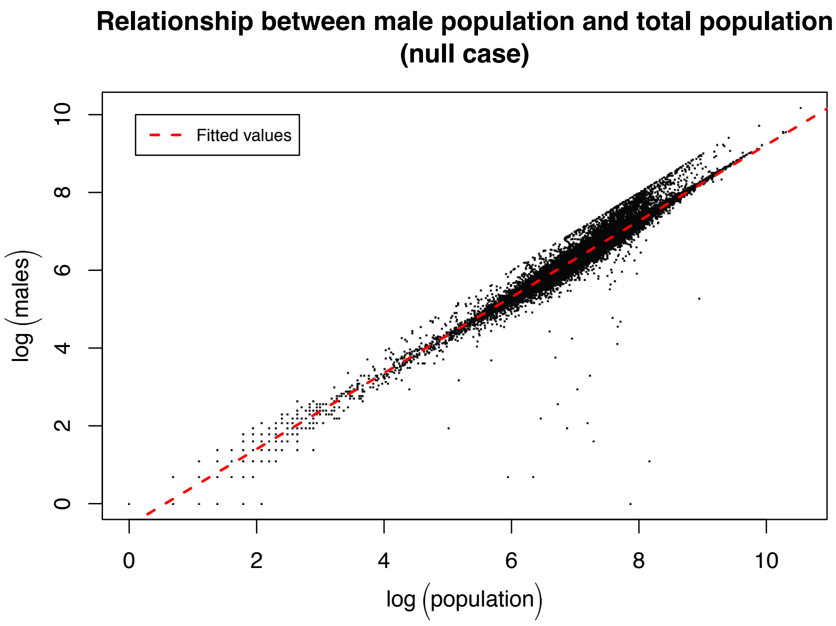
$$\log U_i = \log \alpha + \log P_i + \varepsilon_i'$$

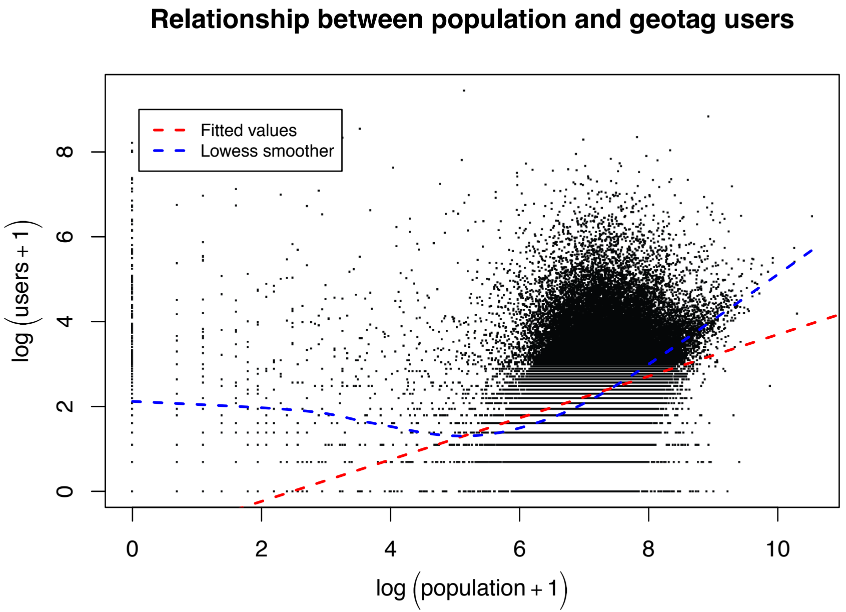
- Compare to a linear model:

$$\log U_i = \beta_0 + \beta_1 \log P_i + \varepsilon_i'$$

Result: Not proportional

(Each dot is a Census *block group*)



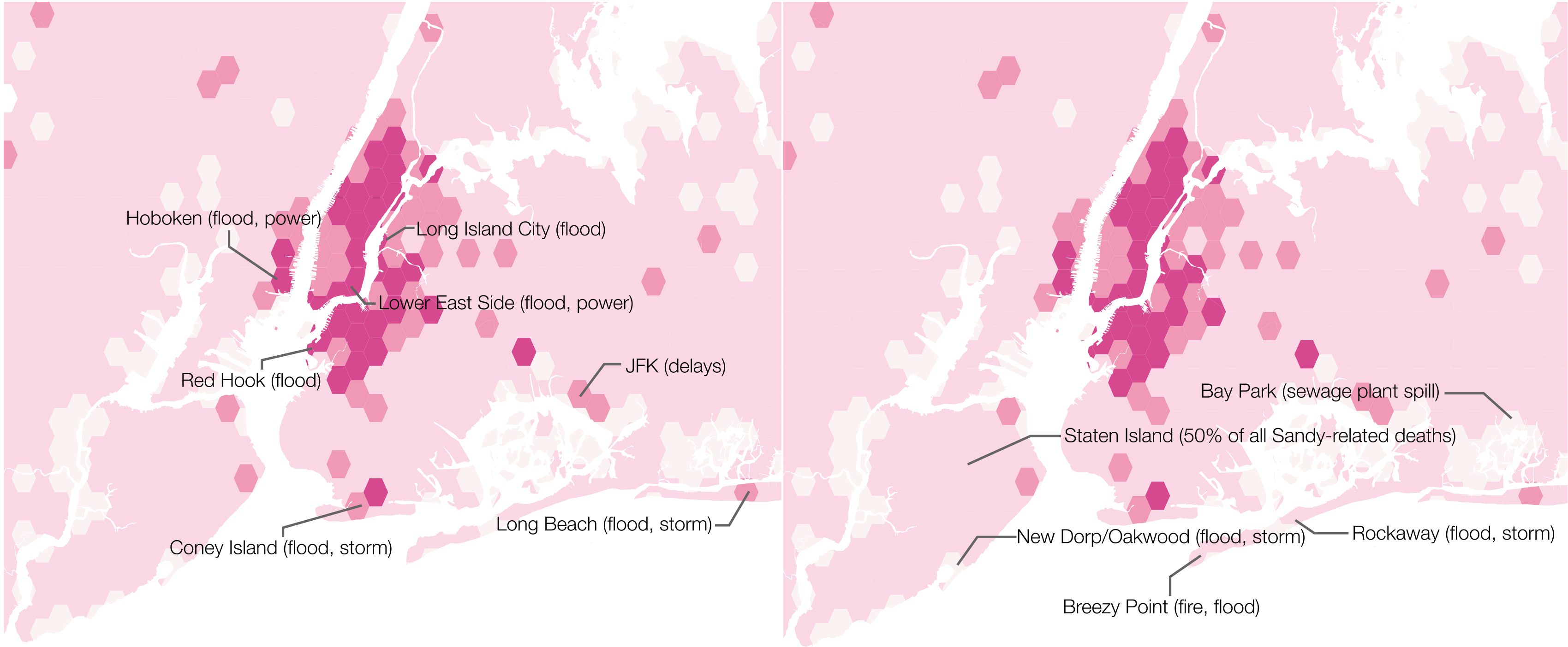


Identifying specifics

- Spatial multivariate modeling of biases
  Geotagged tweet users associated with:
  - □ Rural, poor, elderly, non-coastal
  - □ Asian, Hispanic, black
- …but these are only the demographics we can access. E.g., harassment of women on Twitter likely discourages geotag use

Why it matters: Some uses are bad

Hurricane Sandy, tweets vs. damage/deaths (Shelton et al., 2014)



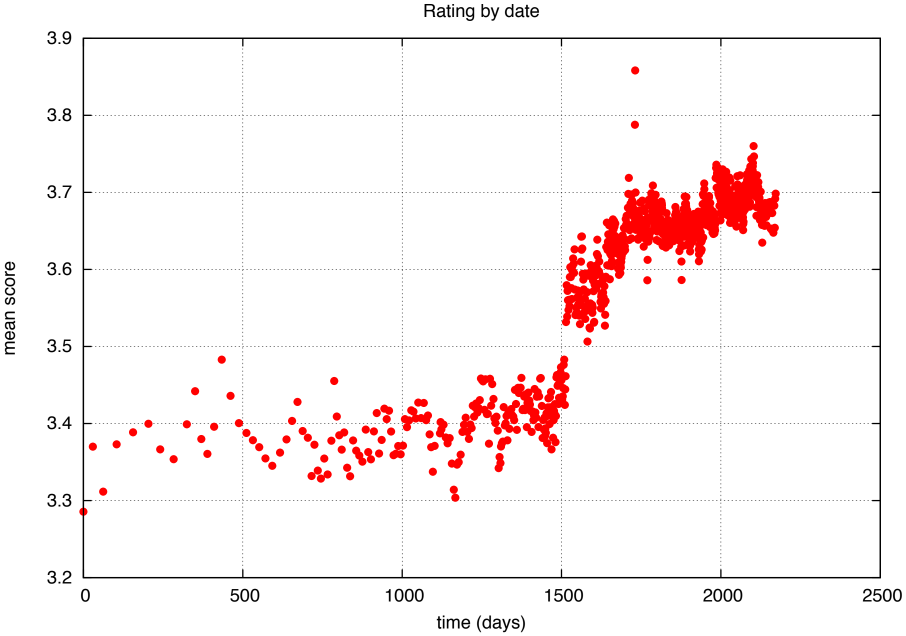
Responses to demographic bias

- Model the specific biases!
- Calibration and weighting (Zagheni & Weber, 2015)
- **Use data for appropriate questions**
  - "Postcards, not ticket stubs" (Tasse et al., 2017)
- Find clever study designs or data comparisons, establish *panels*, etc.

Platform effects in social media

Momin M. Malik and Jürgen Pfeffer. 2016. Identifying platform effects in social media data. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (ICWSM-16), pages 241–249. May 18–20, 2016, Cologne, Germany. Expanded version (2018): https://www.mominmalik.com/malik_chapter2.pdf

Design can cause/change behavior



Average Netflix movie ratings over time. Each point averages 100,000 rating instances.

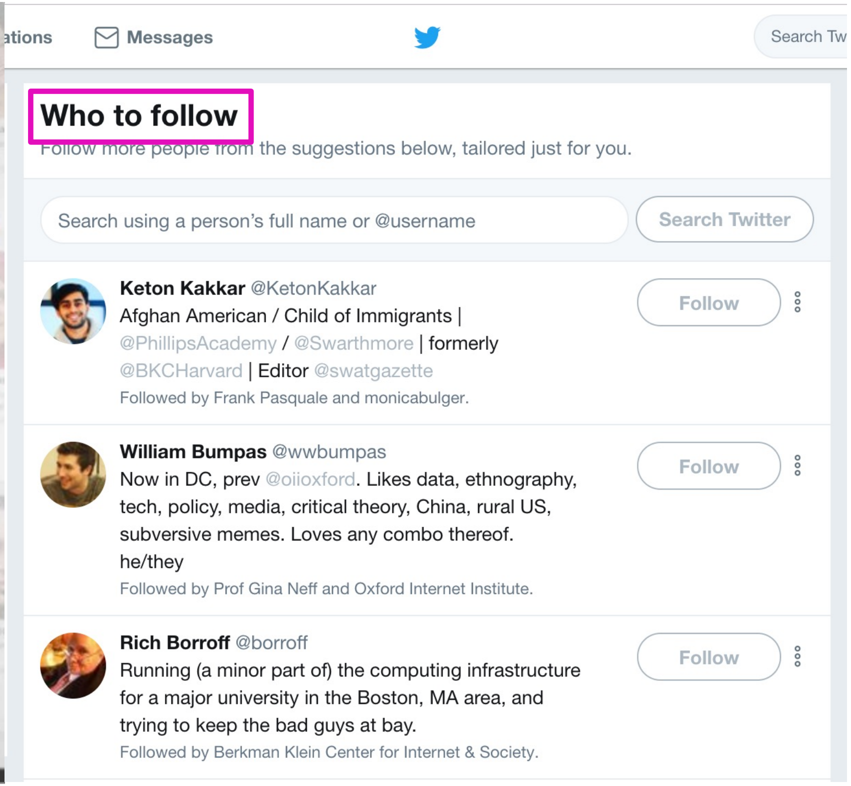
Koren, 2009

Social media platforms are businesses

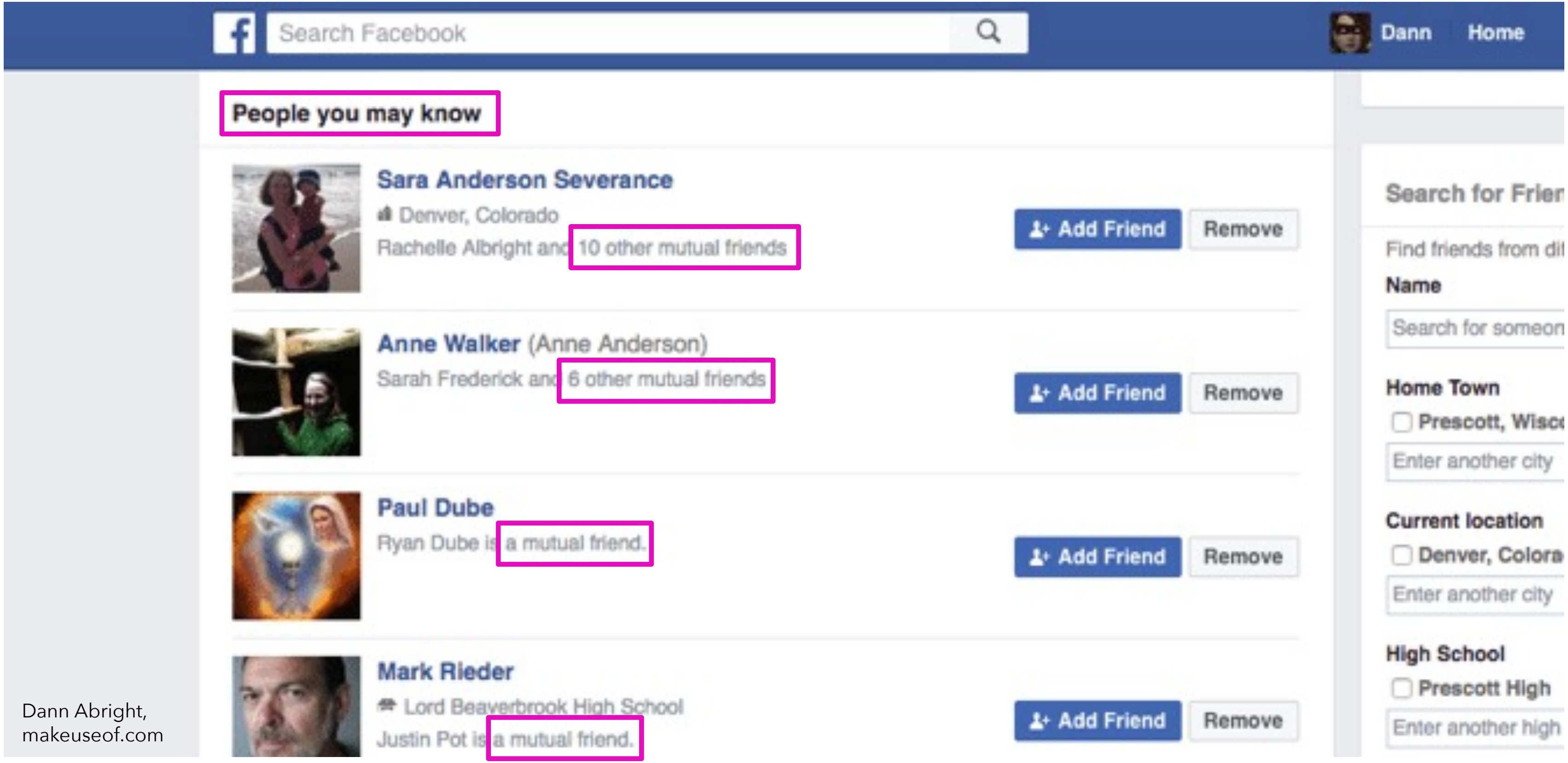
Markets Insider, Business Insider (2018)

- Not neutral utilities or research environments
- Platform engineers try to shape user behavior towards desirable ends

Sites try to grow their users' networks



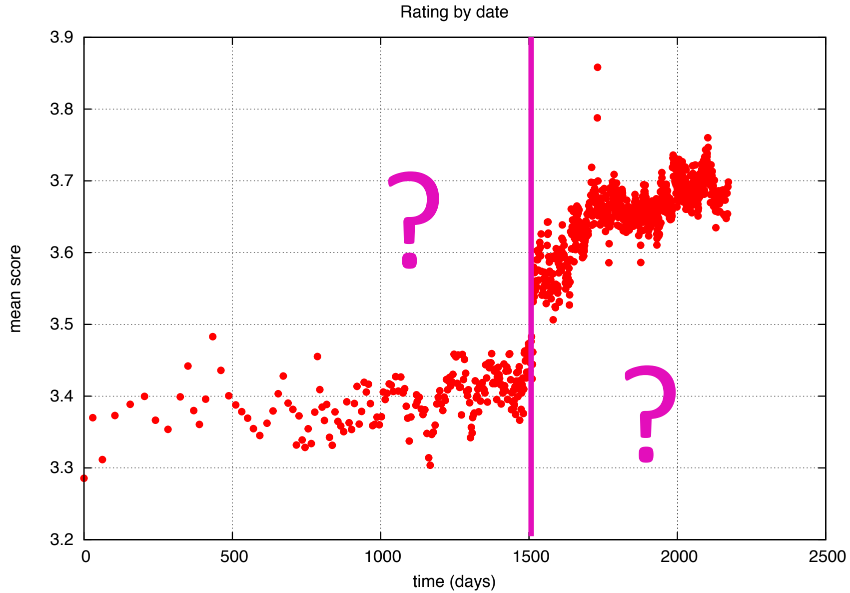
Recommending "friend-of-a-friend"



Dann Abright, makeuseof.com

Behavior, or platform effects?

- When we measure behavior, what are we really measuring? People's behavior, or platform effects?
- How, as outsiders, can we find out?



Average Netflix movie ratings over time. Each point averages 100,000 rating instances.

Data artifacts can reveal inner workings

The Matrix (1999) “déjà vu” scene

Data artifacts as natural experiments

- Regression Discontinuity (RD) Design (technically, Interrupted Time Series, ITS) estimates causality

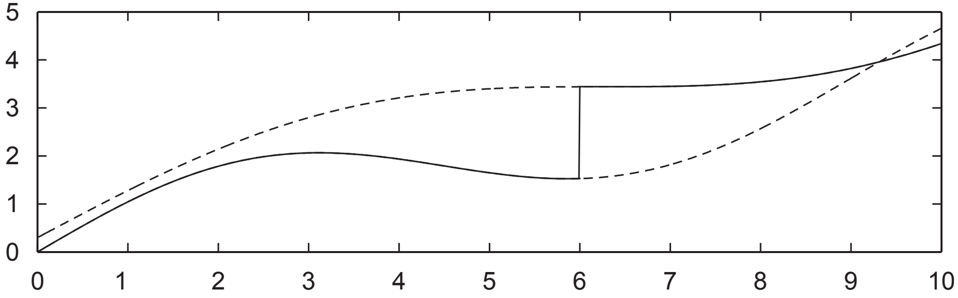
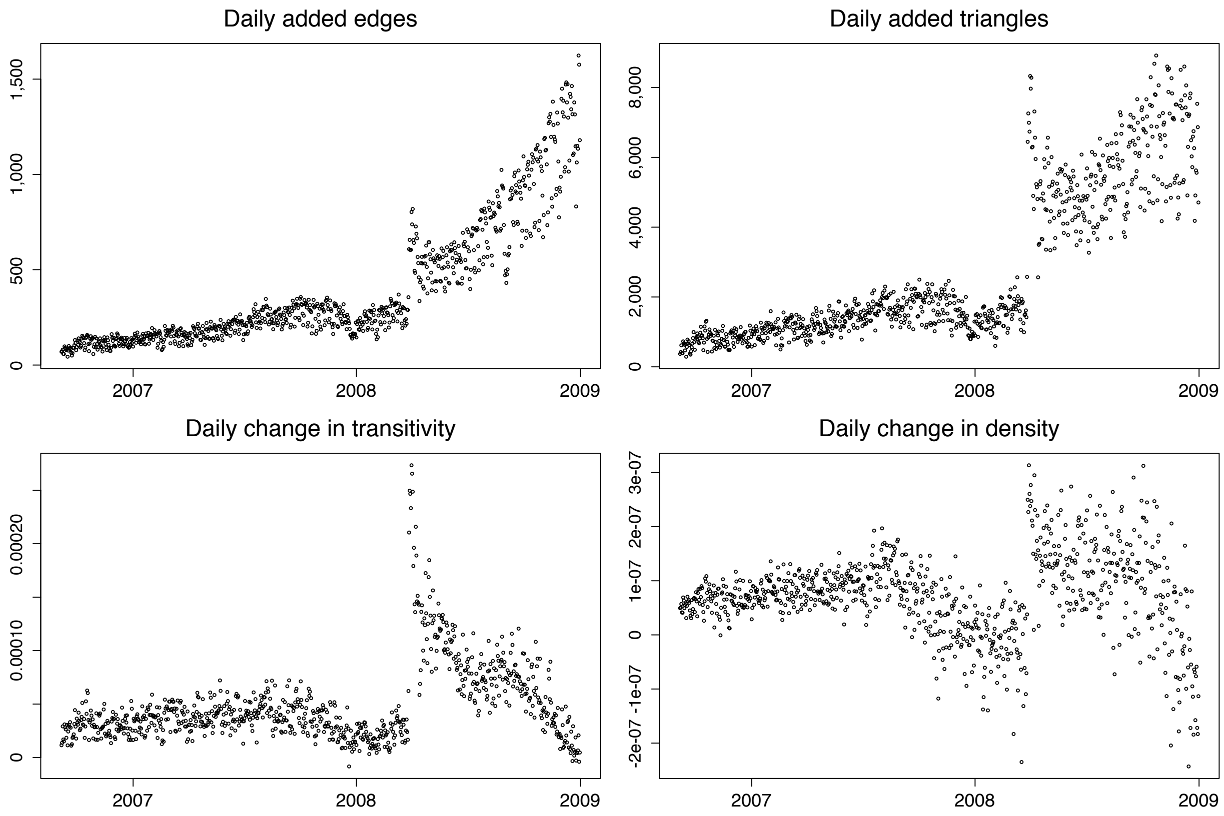


Fig. 2 from Imbens and Lemieux (2008): Potential and observed outcome regression functions.

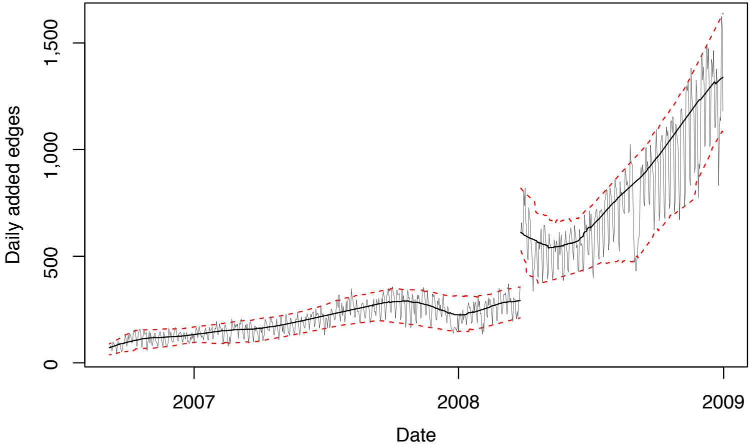
- The difference between “before” and “after” estimates the *local average treatment effect*

Case: Facebook's "People You May Know"

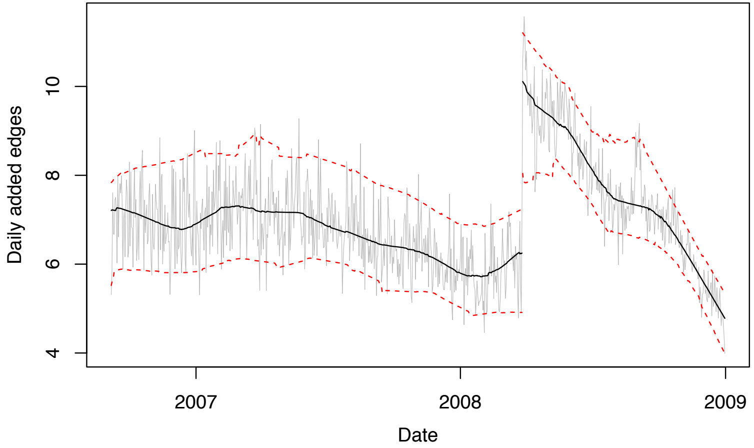


PYMK changed the Facebook network!

- Facebook links: +300 new edges per day (x2)



- Triangles: +3.8 triangles per edge (x1.62)



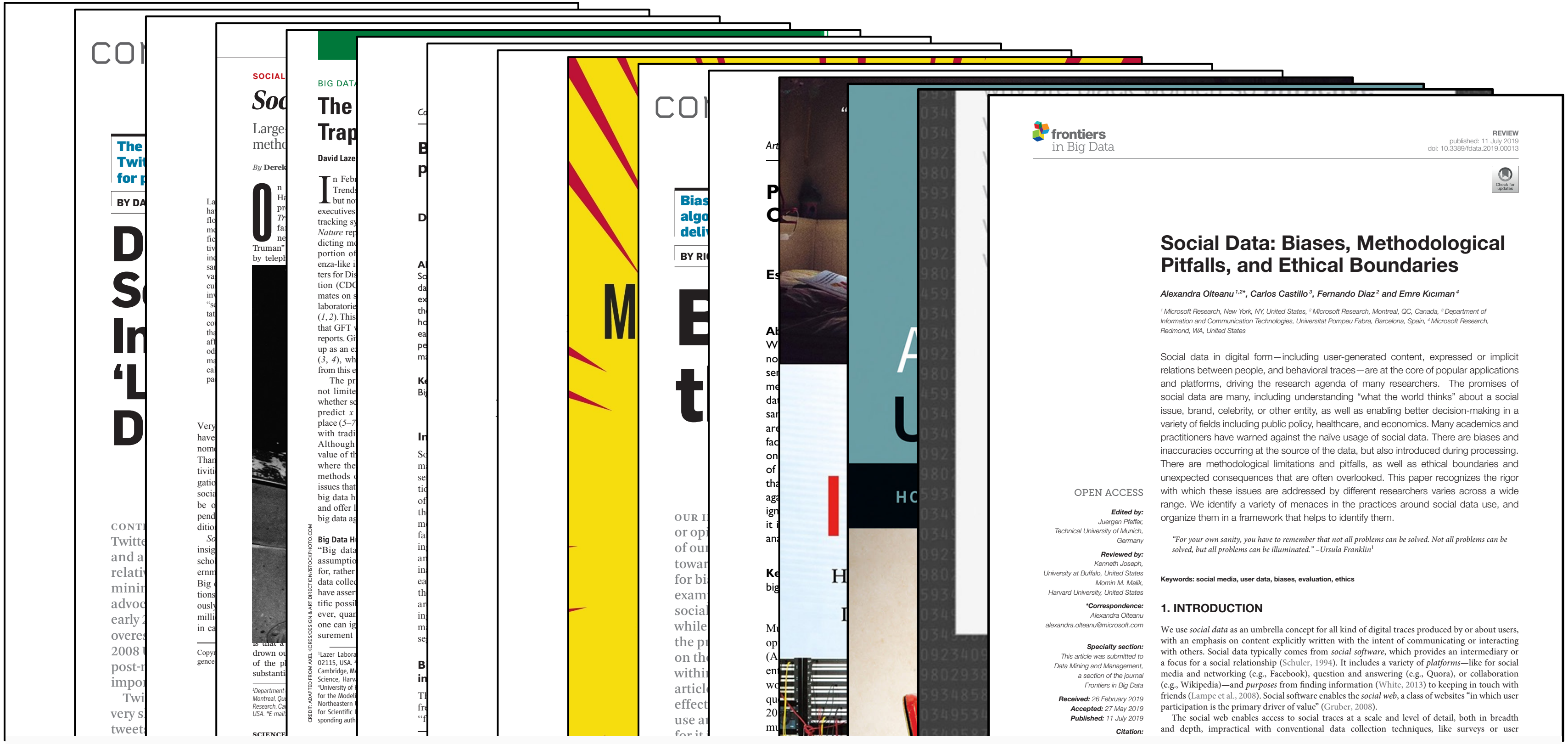
Responses to platform effects

- Investigate: how do Facebook "friendship" fail to generalize? What about the Facebook social network?
- Platform effects are phenomena to study in themselves!
- Data artifacts as natural experiments

Hierarchy of limitations in machine learning

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

Data well-considered; *models*, not so much



frontiers in Big Data

REVIEW
published: 11 July 2019
doi: 10.3389/fdata.2019.00013

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

[1] Microsoft Research, New York, NY, United States, [2] Microsoft Research, Montreal, QC, Canada, [3] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, [4] Microsoft Research, Redmond, WA, United States

Social data in digital form—including user-generated content, expressed or implicit relations between people, and behavioral traces—are at the core of popular applications and platforms, driving the research agenda of many researchers. The promises of social data are many, including understanding "what the world thinks" about a social issue, brand, celebrity, or other entity, as well as enabling better decision-making in a variety of fields including public policy, healthcare, and economics. Many academics and practitioners have warned against the naïve usage of social data. There are biases and inaccuracies occurring at the source of the data, but also introduced during processing. There are methodological limitations and pitfalls, as well as ethical boundaries and unexpected consequences that are often overlooked. This paper recognizes the rigor with which these issues are addressed by different researchers varies across a wide range. We identify a variety of menaces in the practices around social data use, and organize them in a framework that helps to identify them.

"For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated." –Ursula Franklin[1]

OPEN ACCESS

Edited by:
Juergen Pfeffer, Technical University of Munich, Germany

Reviewed by:
Kenneth Joseph, University at Buffalo, United States
Momin M. Malik, Harvard University, United States

***Correspondence:**
Alexandra Olteanu
alexandra.olteanu@microsoft.com

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *26 February 2019*
Accepted: *27 May 2019*
Published: *11 July 2019*

Citation:

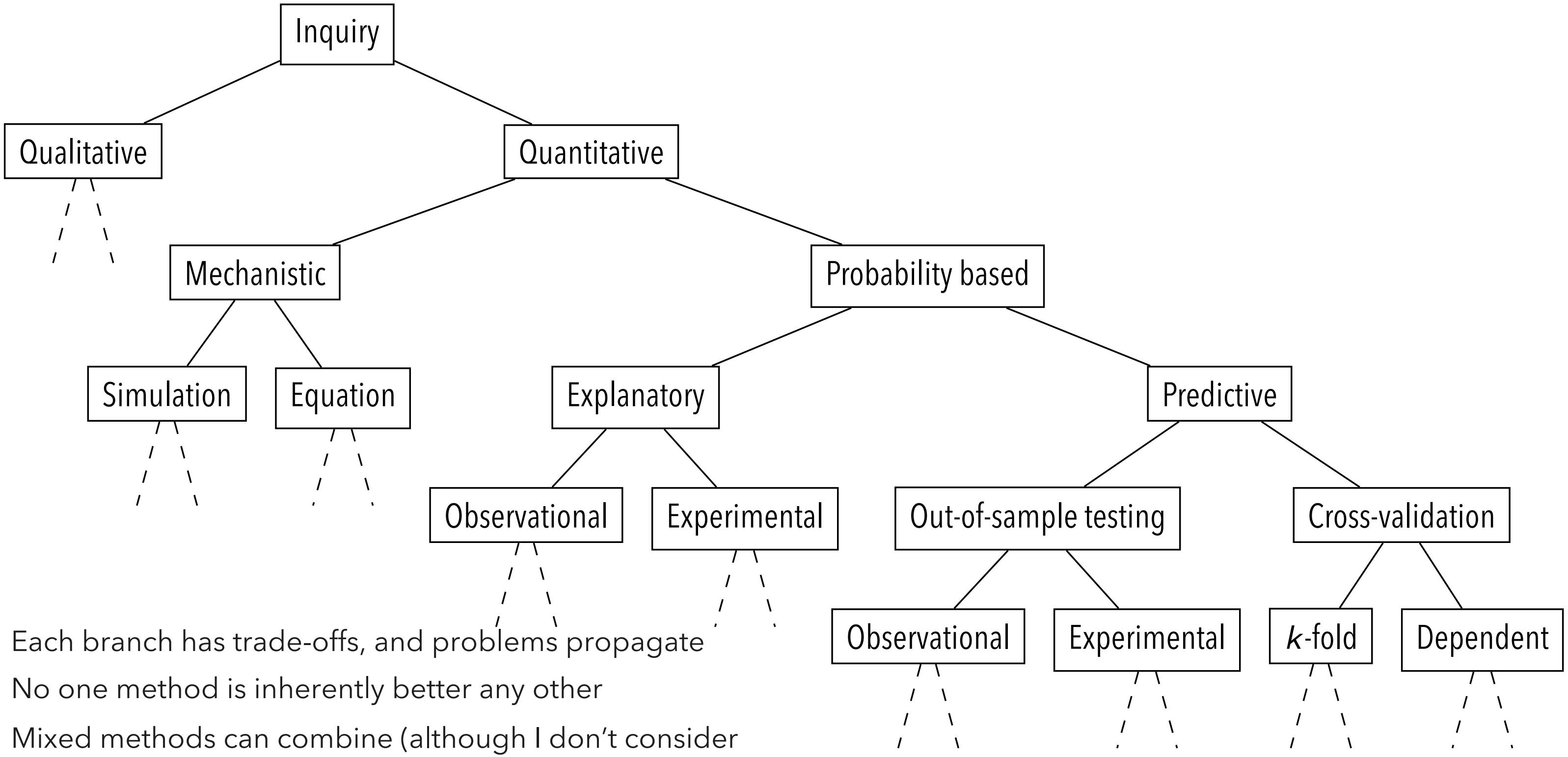
Keywords: social media, user data, biases, evaluation, ethics

1. INTRODUCTION

We use *social data* as an umbrella concept for all kind of digital traces produced by or about users, with an emphasis on content explicitly written with the intent of communicating or interacting with others. Social data typically comes from *social software*, which provides an intermediary or a focus for a social relationship (Schuler, 1994). It includes a variety of *platforms*—like for social media and networking (e.g., Facebook), question and answering (e.g., Quora), or collaboration (e.g., Wikipedia)—and *purposes* from finding information (White, 2013) to keeping in touch with friends (Lampe et al., 2008). Social software enables the *social web*, a class of websites "in which user participation is the primary driver of value" (Gruber, 2008).

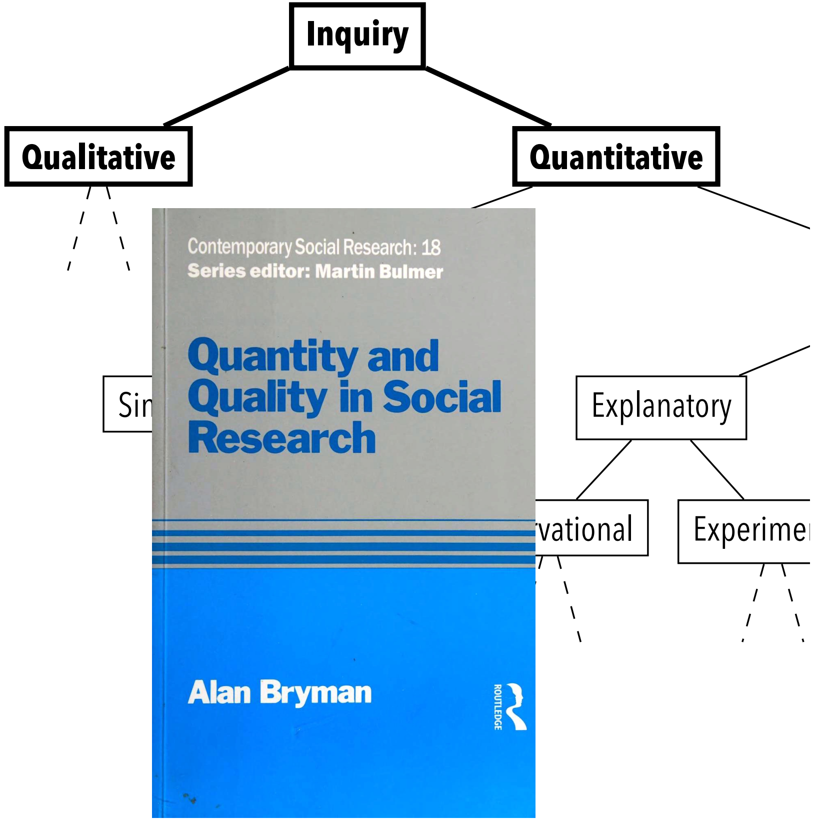
The social web enables access to social traces at a scale and level of detail, both in breadth and depth, impractical with conventional data collection techniques, like surveys or user

Approaches to research



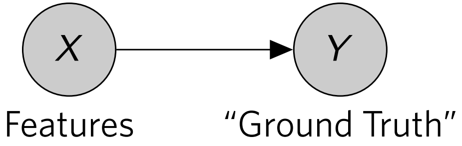
- Each branch has trade-offs, and problems propagate
- No one method is inherently better any other
- Mixed methods can combine (although I don't consider this in the paper)

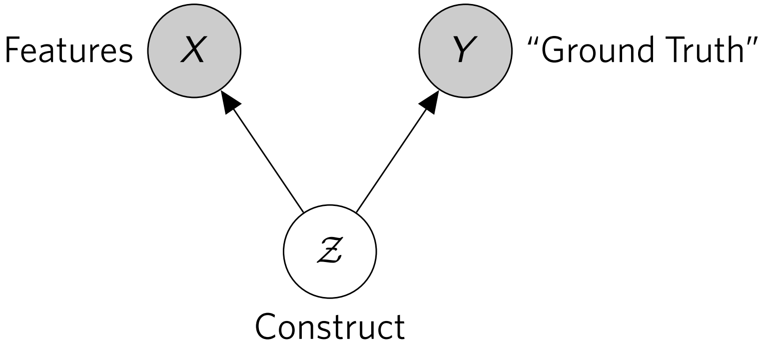
Quantification locks in meaning



- Qualitative research can get directly at how things are multifaceted, heterogeneous, intersubjective
- Quantification/ measurements lock in one meaning; and frequently are *proxies*, which are imperfect

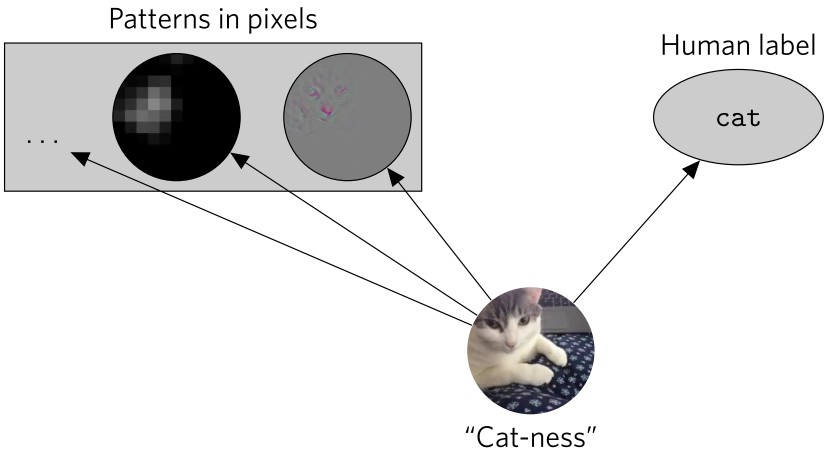
Challenges of quantification/ measurement



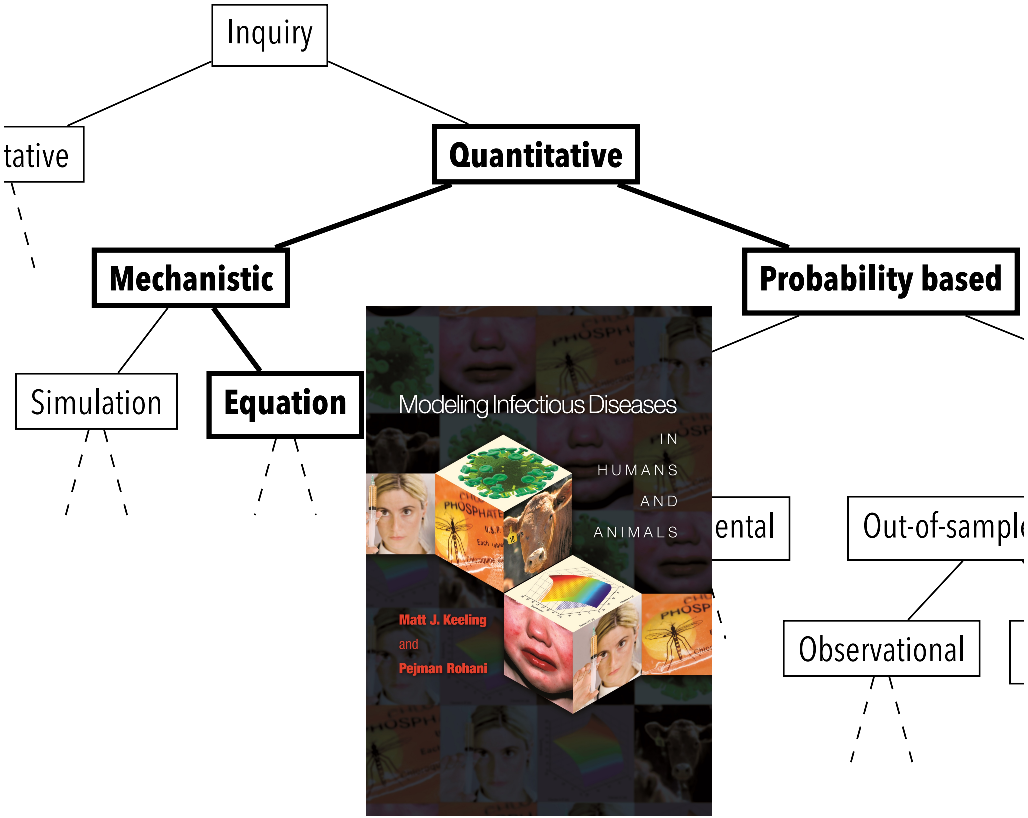


- *Constructs*: primitives of social science
  - What we care about
  - Often unobservable (and hypothetical/subjective, e.g. friendship)
  - Proxies always give errors (for binary constructs: false negatives and false positives)
  - E.g., Google maps usage is not traffic

Constructs: Subjective, multifaceted

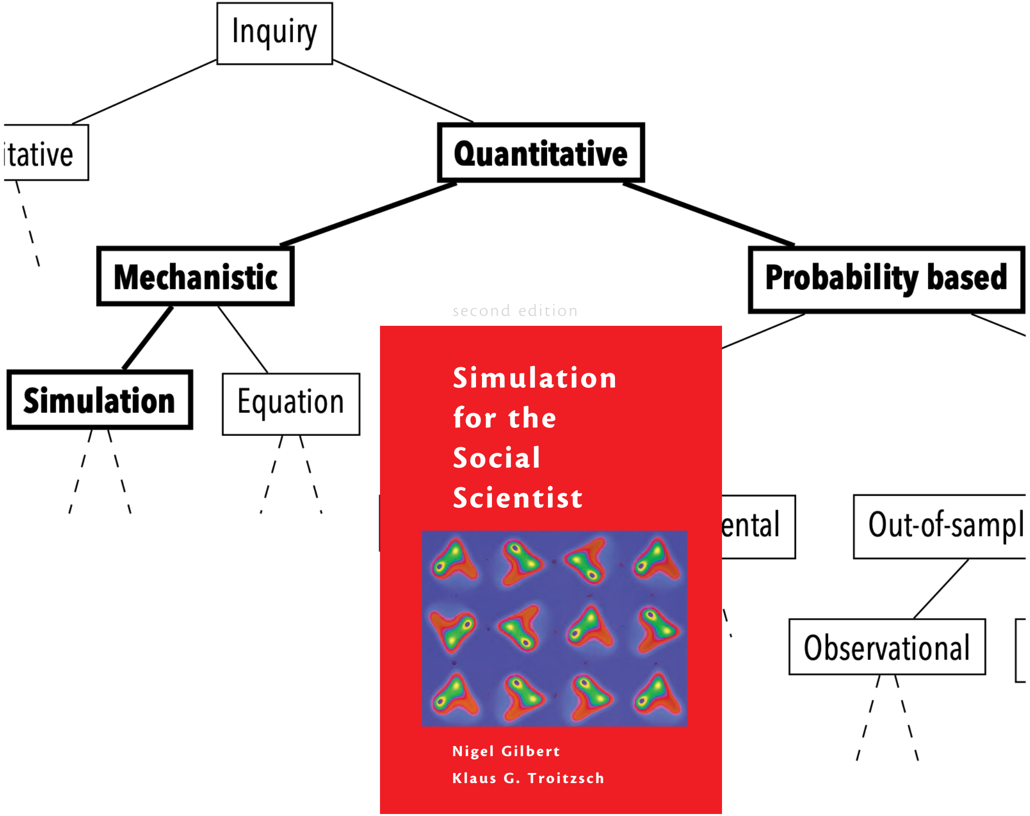


Stats and ML use central tendencies



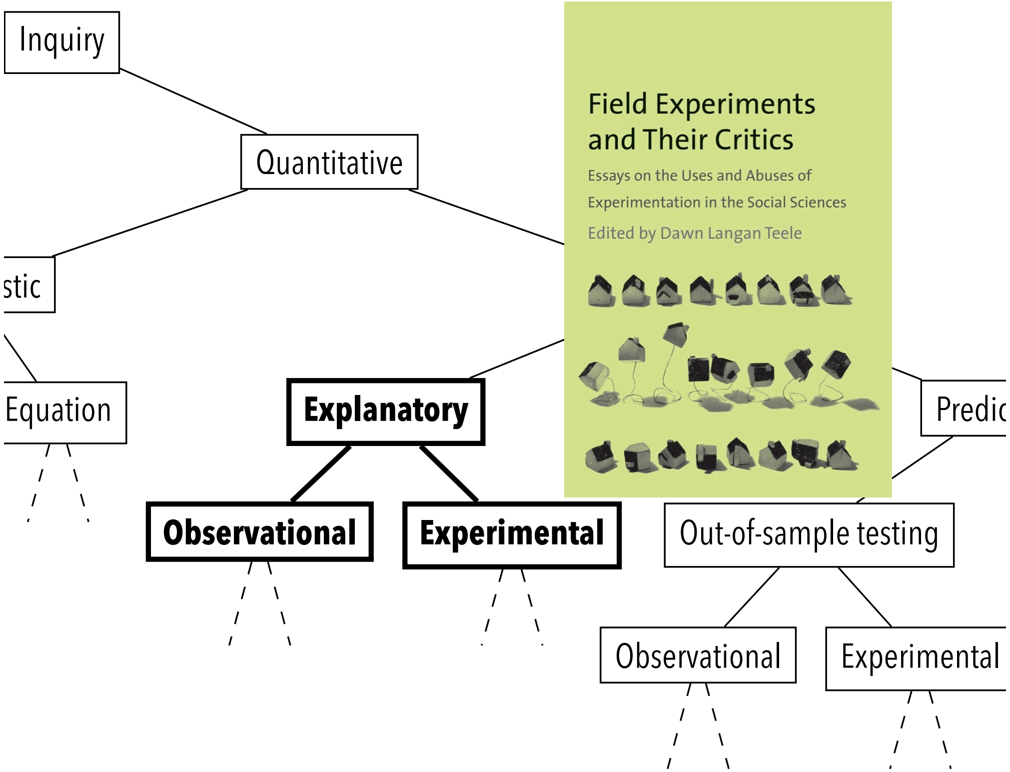
- Statistics and machine only option to both directly use data *and* account for variability
- They do so via *central tendency*
- This requires multiple observations, and independence assumptions

Stats and ML use central tendencies



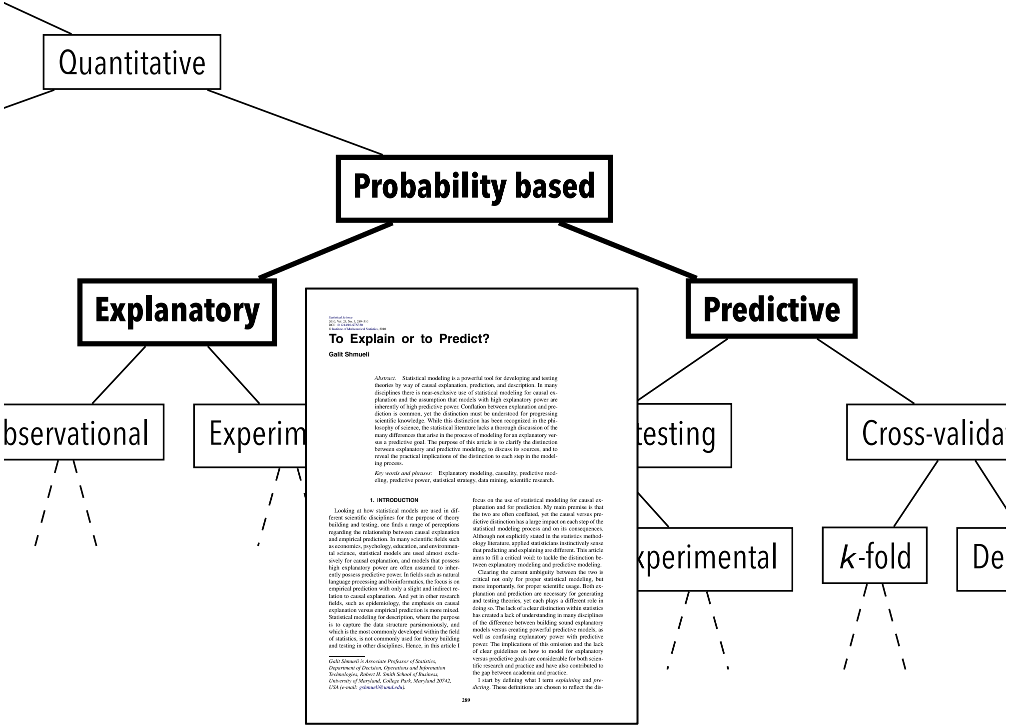
- (Statistics uses *numerical* simulations, and simulation modeling uses statistical *summaries*, but they are distinct types of models)
- (Agent-based simulation also ends up using central tendencies to summarize a response surface)
- (ABMs generally cannot be used for prediction, are only appropriate when we can't do statistics)

Causality is hard, maybe too hard



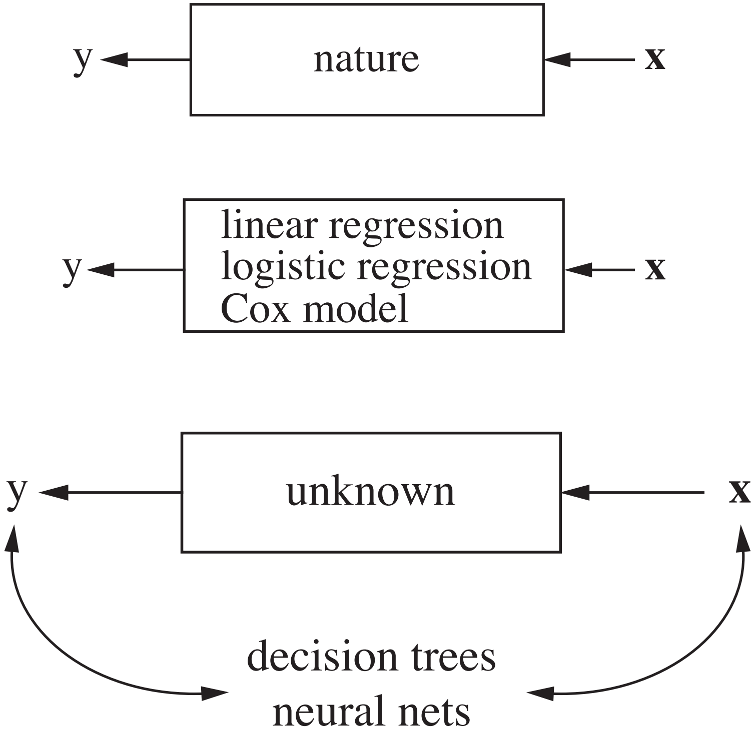
- Properly controlled experiments lack ecological validity
- Observational inference can never totally account for the possibility of hidden confounders, which can frustrate even the most perfect application of causal techniques (Arceneaux, Gerber, & Green, 2010)

ML is "prediction" only



- "Predictions" are defined as what minimizes loss
- I.e., *correlations*
- Non-causal correlations can sometimes predict well, but they frequently don't explain, and can fail unexpectedly

Defining machine learning



- 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)

Why are these different goals?

$$\widehat{y}$$

Spurious (non-causal) correlations may fit robustly

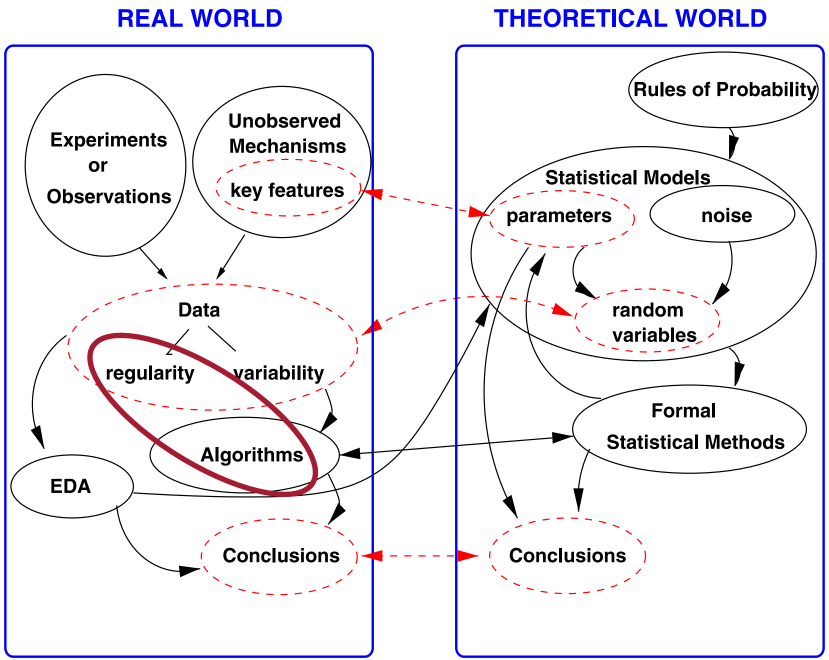
- Breiman 2001: Prediction problems
- Shmueli 2010: To predict
- Kleinberg et al. 2015: "Umbrella problems"
- Mullainathan and Spiess 2017: y-hat

$$\widehat{\beta}$$

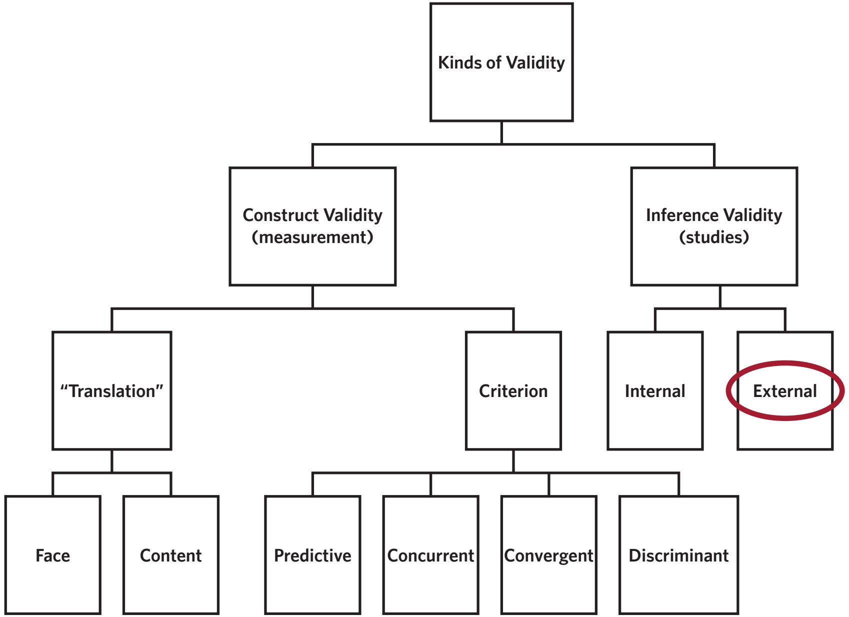
Carefully built models that capture causality (or "pure" associations) may fit poorly overall

- Breiman 2001: Information
- Shmueli 2010: To explain
- Kleinberg et al. 2015: "Rain dance problems"
- Mullainathan and Spiess 2017: beta-hat

ML: Only external validity

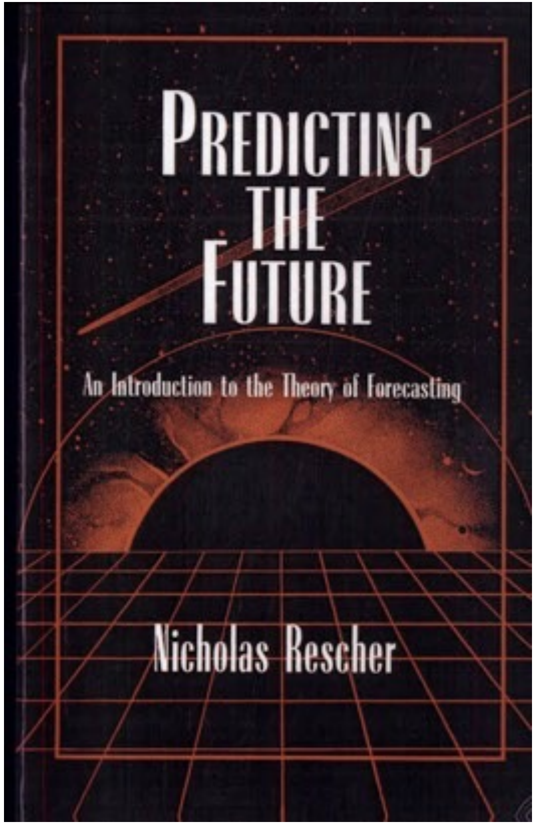


Kass, 2011



Adapted from Borgatti, 2012

Not an obvious usage of "predict"

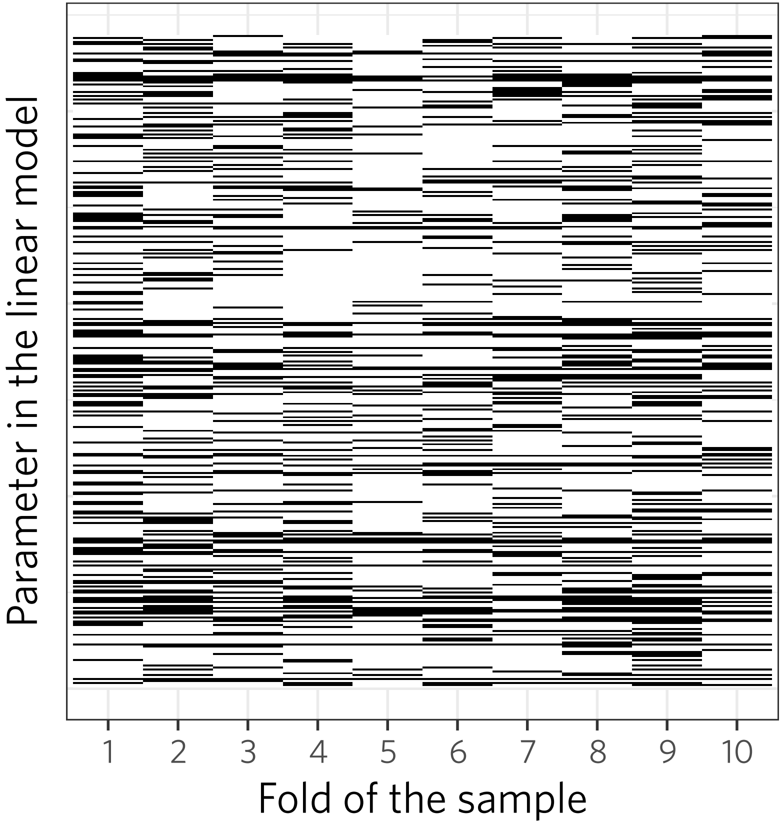


88 ■ PREDICTING THE FUTURE

TABLE 6.1: A SURVEY OF PREDICTIVE APPROACHES

Predictive Approaches	Linking Mechanism	Methodology Of Linkage
UNFORMALIZED/JUDGMENTAL		
judgmental estimation	expert informants	informed judgment
FORMALIZED/INFERENTIAL		
RUDIMENTARY (ELEMENTARY)		
trend projection	prevailing trends	projection of prevailing trends
curve fitting	geometric patterns	subsumption under an established pattern
circumstantial analogy	comparability groupings	assimilation to an analogous situation
SCIENTIFIC (SOPHISTICATED)		
indicator coordination	causal correlations	statistical subsumption into a correlation
law derivation (nomic)	accepted laws (deterministic or statistical)	inference from accepted laws
phenomenological modeling (analogical)	formal models (physical or mathematical)	analogizing of actual ("real-world") processes with presumably isomorphic model process

Can’t *intervene* based on correlations



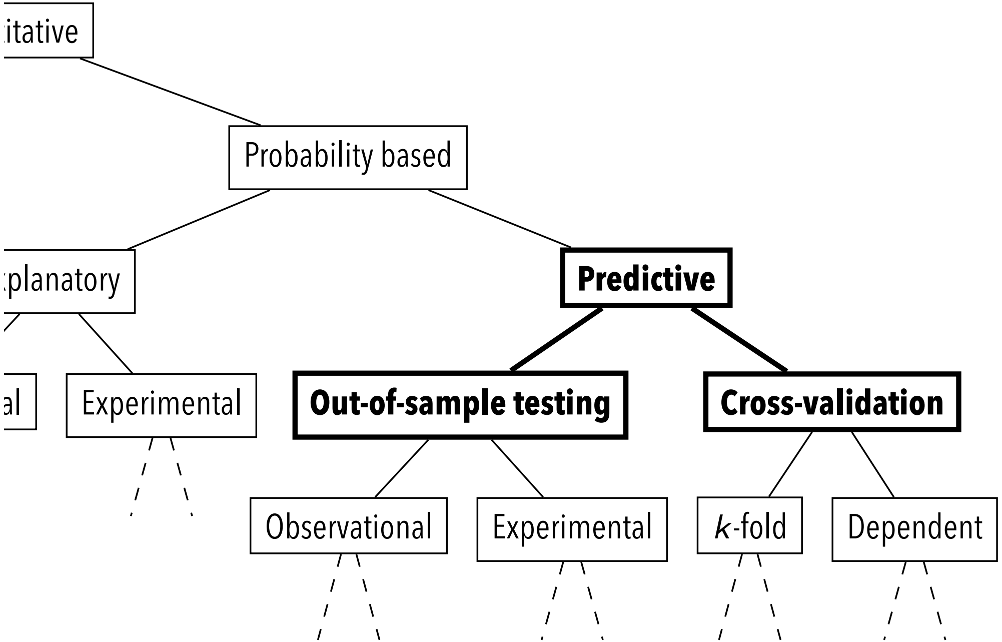
- Very different sets of correlations can “predict” (fit) equally well (Mullainathan and Spiess 2017)
  - Breiman (2001) called this the “Rashomon Effect”
- But different fits suggest very different interventions

Interpretability: A red herring?

"A project I worked on in the late 1970s was the analysis of delay in criminal cases in state court systems... A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, 'I knew those guys in District N were dragging their feet.'" (Breiman, 2001)

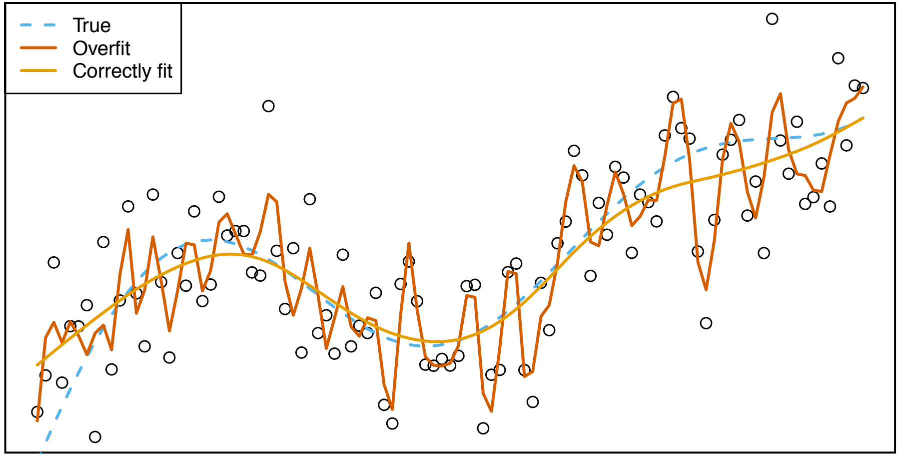
Problems of cross validation

ML performance claims are from cross-validation



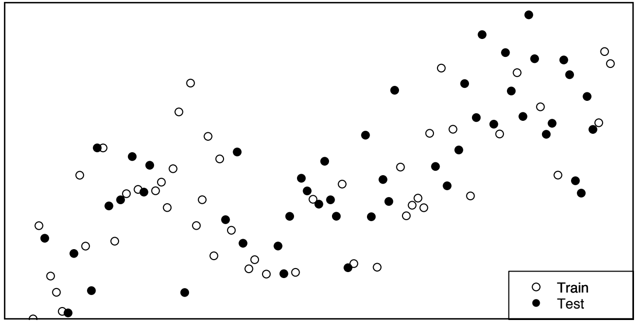
- Rescher (1998) notes every prediction involves a meta-prediction: predict whether the prediction works
- Cross-validation is meta-prediction for ML
- But, how well does cross-validation work?
  - "Professor Breiman emphasizes the importance of performance on the test sample. However, this can be overdone. The test sample is supposed to represent the population to be encountered in the future. But in reality, it is usually a random sample of the current population. High performance on the test sample does not guarantee high performance on future samples, **things do change**." (Hoadley 2001)

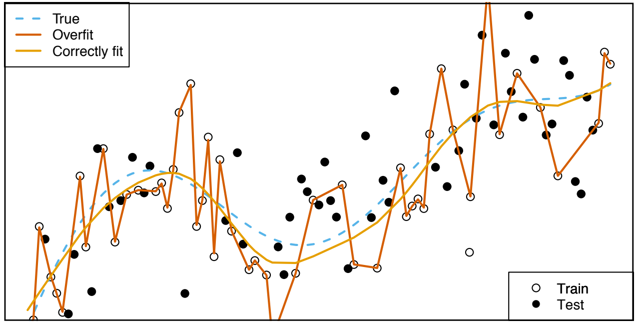
Purpose of cross-validation



- If we are no longer guided by theory, and use automatic methods, we risk overfitting: fitting to the the noise, not the data

Intuition for cross-validation

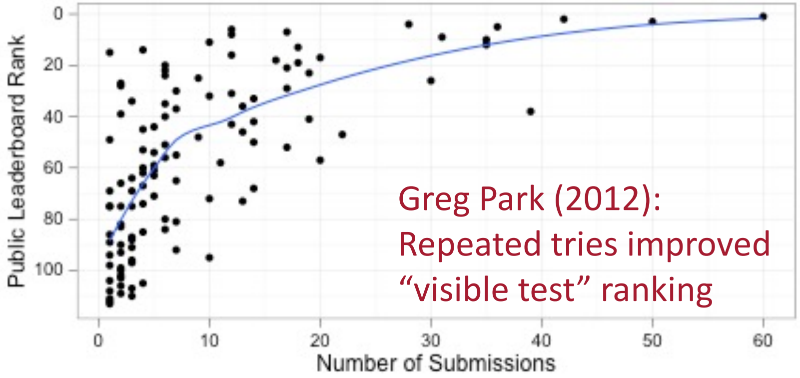


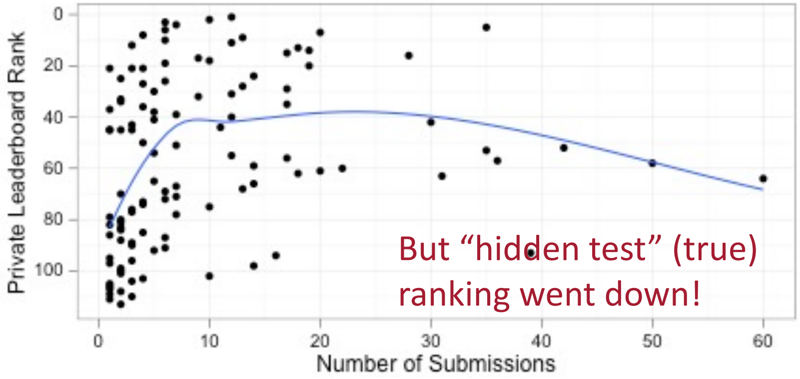


- Idea: if we split data into two parts, the signal should be the same but the noise would be different
- *Cross validation:* Fitting the model on one part of the data, and "testing" on the other

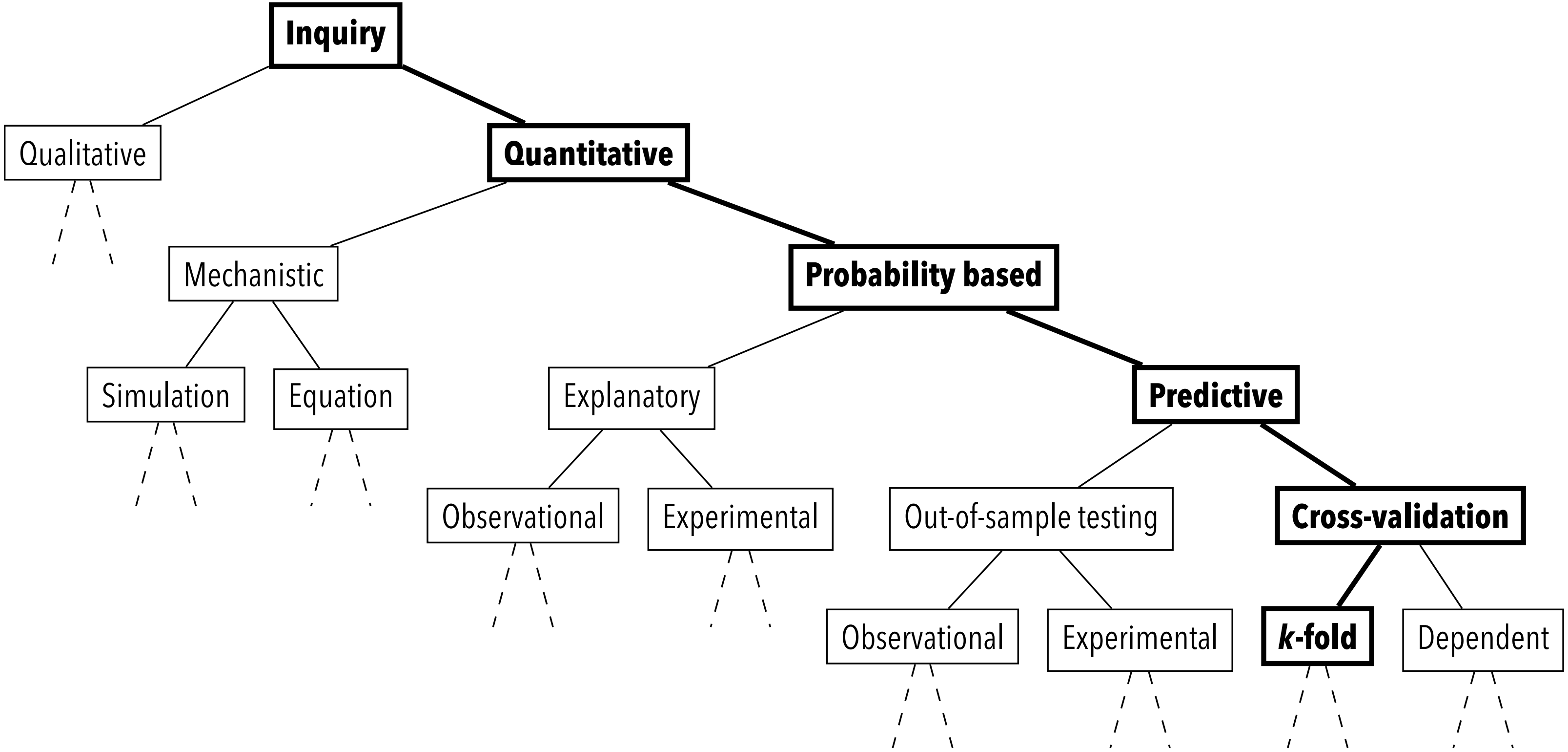
Overfitting on the test set

- Re-using a test set can overfit! (Dwork et al., 2015)
  - “in industry and academia, there is sometimes a little tinkering, which involves peeking at the test sample. The result is some bias in the test sample or cross-validation results. This is the same kind of tinkering that upsets test of fit pureness.” (Hoadley 2001, discussant of Breiman)
- Happens in Kaggle, which has public leaderboard (visible throughout) and private leaderboard (revealed only at end of competition)





Problems of dependencies



Classic argument for CV

$$
\begin{aligned}
\mathrm{Err}(\hat{\mu}) &= \tfrac{1}{n}\mathbb{E}_f \| Y^* - \widehat{Y} \|_2^2 \\
&= \frac{1}{n}\left[\mathbb{E}_f \| Y^* \|_2^2 + \mathbb{E}_f \| \widehat{Y} \|_2^2 - 2\mathbb{E}_f (Y^{*T} \widehat{Y})\right] \\
&= \frac{1}{n}\left[\mathbb{E}_f \| Y^* \|_2^2 + \mathbb{E}_f \| \widehat{Y} \|_2^2 - 2\,\mathrm{tr}\,\mathbb{E}_f (Y^* \widehat{Y}^T)\right] \\
&\quad + \frac{1}{n}\left[\ \mu^T\mu + \mathbb{E}_f(\widehat{Y})^T \mathbb{E}_f(\widehat{Y}) + 2\,\mathrm{tr}\,\mu\mathbb{E}_f(\widehat{Y})^T\right] \\
&\quad + \frac{1}{n}\left[-\mu^T\mu - \mathbb{E}_f(\widehat{Y})\mathbb{E}_f(\widehat{Y})^T - 2\mu^T\mathbb{E}_f(\widehat{Y})\right] \\
&= \frac{1}{n}\left[\mathrm{tr}\,\Sigma + \| \mu - \mathbb{E}(\widehat{Y}) \|_2^2 + \mathrm{tr}\,\mathrm{Var}_f(\widehat{Y}) - 2\,\mathrm{tr}\,\mathrm{Cov}_f(Y^*, \widehat{Y})\right] \\
&= \ \text{irreducible error} + \text{bias}^2 + \text{variance} - \text{optimism}
\end{aligned}
$$

Apply this to non-iid data

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

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

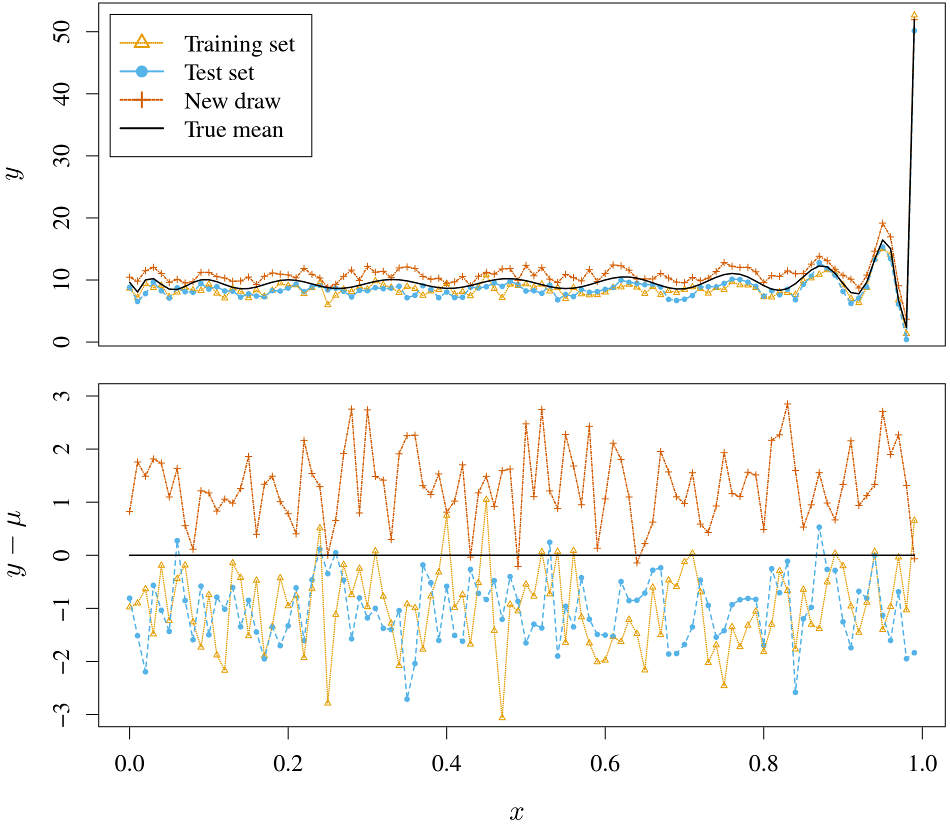
- Then, optimism (Efron, 2004) in the training set is:

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

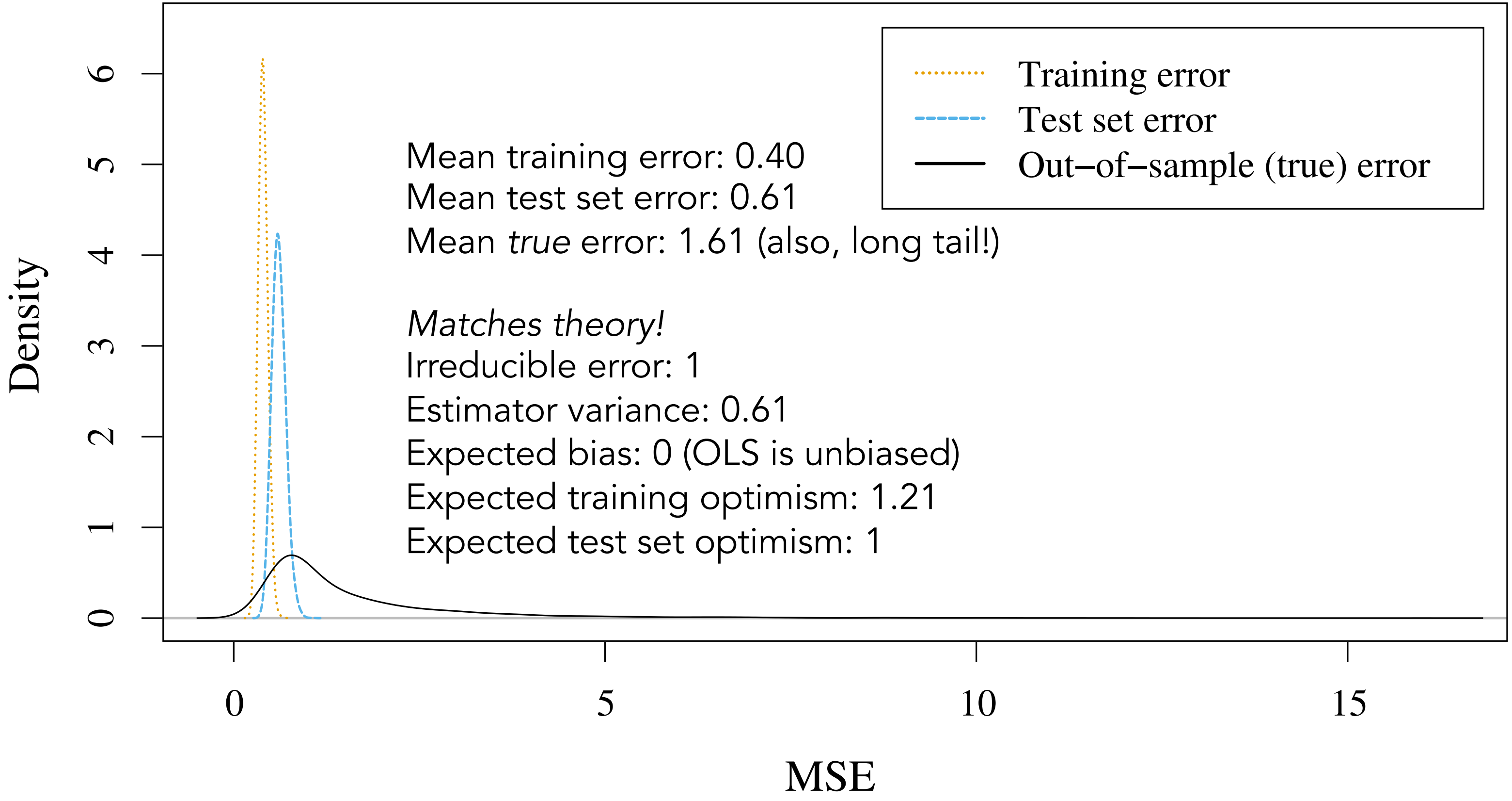
- But test set also has nonzero optimism!

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

Simulating the toy example



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



Many real-world examples

- There are indeed cases where cross-validation assessments of machine learning performance fail!
- Time series: do cross-validation in blocks
  - Otherwise, "time traveling," gives great performance
- Activity recognition: "leave one subject out" cross validation performs far worse (i.e., more honestly)
- Necessary but not sufficient; underlying causal processes can introduce unobserved variance, destroying previously-holding correlations

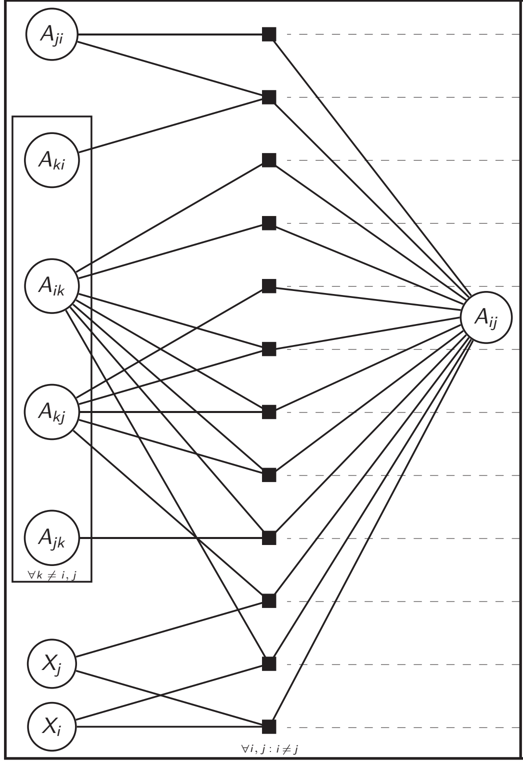
Application to networks

	Y	X_1	X_2	$\cdots$	X_d
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1d}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nd}

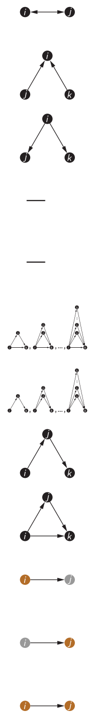
index	*from*	*to*	Y	W_1	W_2	W_3	$\cdots$
e_1	1	2	y_{12}	$\mathbf{1}(x_{11} = x_{21})$	$x_{12} - x_{22}$	x_{13}	$\cdots$
e_2	2	3	y_{23}	$\mathbf{1}(x_{11} = x_{31})$	$x_{12} - x_{32}$	x_{13}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
e_{n+1}	2	1	y_{21}	$\mathbf{1}(x_{21} = x_{11})$	$x_{22} - x_{12}$	x_{23}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$e_{2\binom{n}{2}}$	$n-1$	n	$y_{(n-1)n}$	$\mathbf{1}(x_{(n-1)1} = x_{n1})$	$x_{(n-1)2} - x_{n2}$	$x_{(n-1)3}$	$\cdots$

But dyads are dependent too!

Factor graph



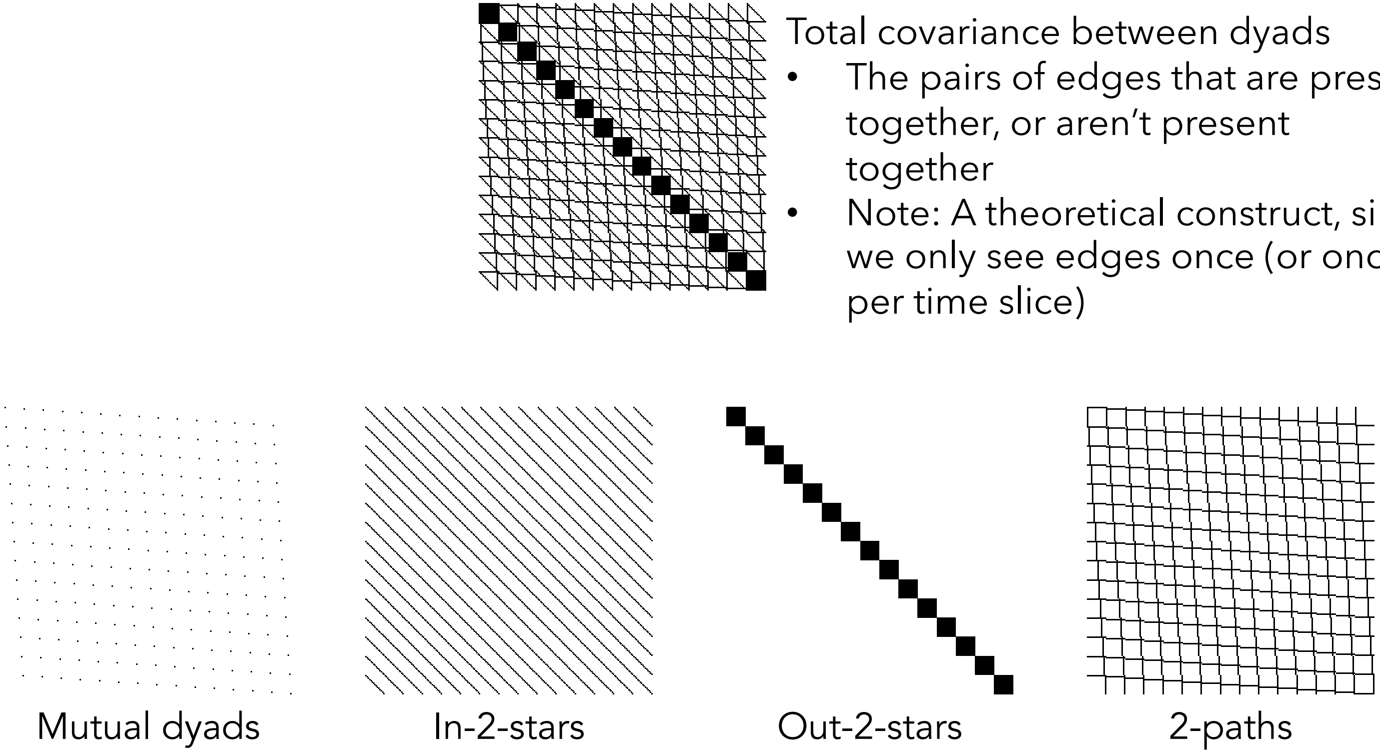
Network Motif



Parameter name	Network Motif	Parameterization	Matrix notation
mutual dyads		$\sum_{i<j} A_{ij}A_{ji}$	$\frac{1}{2}\,\mathrm{tr}\,(\mathbf{AA}^T)$
in-two-stars		$\sum_{(i,j,k)} A_{ji}A_{ki}$	$\mathrm{sum}\,(\mathbf{AA}^T) - \mathrm{tr}\,(\mathbf{AA}^T)$
out-two-stars		$\sum_{(i,j,k)} A_{ij}A_{ik}$	$\mathrm{sum}\,(\mathbf{A}^T\mathbf{A}) - \mathrm{tr}\,(\mathbf{A}^T\mathbf{A})$
geom. weighted out-degrees	—	$\sum_i \exp\left\{-\alpha \sum_k A_{ik}\right\}$	$\mathrm{sum}\,(\exp\{-\alpha\ \mathrm{rowsum}\,(\mathbf{A})\})$
geom. weighted in-degrees	—	$\sum_j \exp\left\{-\alpha \sum_k A_{kj}\right\}$	$\mathrm{sum}\,(\exp\{-\alpha\ \mathrm{colsum}\,(\mathbf{A})\})$
alternating transitive k-triplets		$\lambda \sum_{i,j} A_{ij} \left\{1 - \left(1 - \frac{1}{\lambda}\right)^{\sum_{k \neq i,j} A_{ik}A_{kj}}\right\}$	$\lambda\ \mathrm{sum}\left(\mathbf{A} \odot \left(1 - \left(1 - \frac{1}{\lambda}\right)^{\mathbf{AA} - \mathrm{diag}(\mathbf{AA})}\right)\right)$
alternating indep. two-paths		$\lambda \sum_{i,j} \left\{1 - \left(1 - \frac{1}{\lambda}\right)^{\sum_{k \neq i,j} A_{ik}A_{kj}}\right\}$	$\lambda\ \mathrm{sum}\left(1 - \left(1 - \frac{1}{\lambda}\right)^{\mathbf{AA} - \mathrm{diag}(\mathbf{AA})}\right)$
two-paths (mixed two-stars)		$\sum_{(i,k,j)} A_{ik}A_{kj}$	$\mathrm{sum}\,(\mathbf{AA}) - \mathrm{tr}\,(\mathbf{AA})$
transitive triads		$\sum_{(i,j,k)} A_{ij}A_{jk}A_{ik}$	$\mathrm{tr}\,(\mathbf{AAA}^T)$
activity effect		$\sum_i X_i \sum_j A_{ij}$	$\mathrm{sum}\,(\mathbf{X} \odot \mathrm{rowsum}\,(\mathbf{A}))$
popularity effect		$\sum_j X_j \sum_i A_{ij}$	$\mathrm{sum}\,(\mathbf{X} \odot \mathrm{colsum}\,(\mathbf{A}))$
similarity effect		$\sum_{i,j} A_{ij} \left(1 - \frac{\lvert X_i - X_j \rvert}{\max_{k,l} \lvert X_k - X_l \rvert}\right)$	$\mathrm{sum}\,(\mathbf{A} \odot \mathbf{S})$

Graphical model and matrix notations for ERGM specification terms given in: Snijders et al. 2006. Joint work with Antonis Manousis and Naji Shajarisales, 2018.

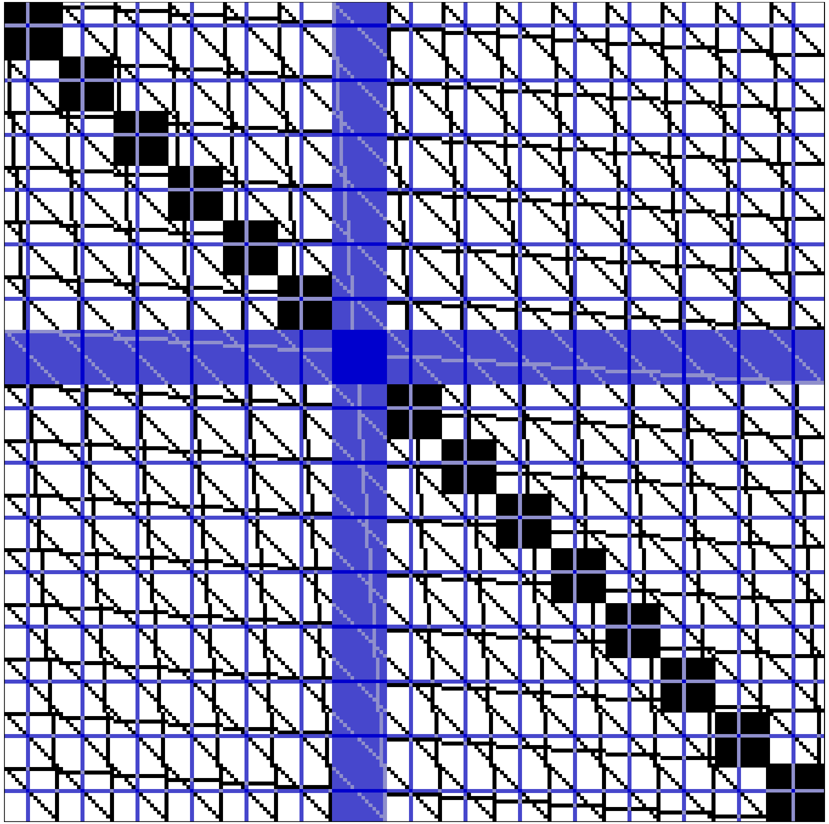
Covariance structure of *edges* (*n* = 15)



Total covariance between dyads

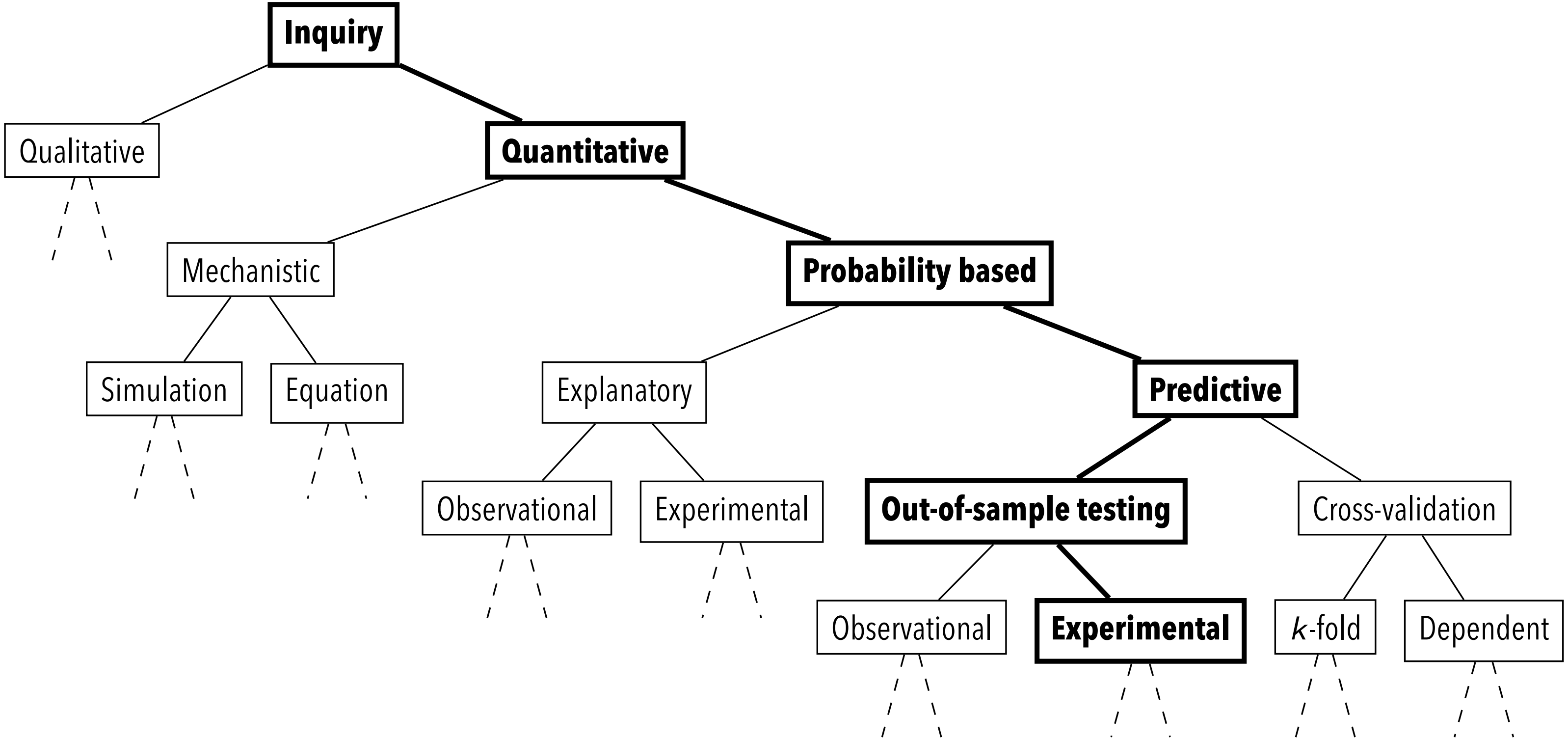
- The pairs of edges that are present together, or aren't present together
- Note: A theoretical construct, since we only see edges once (or once per time slice)

So, what to do?

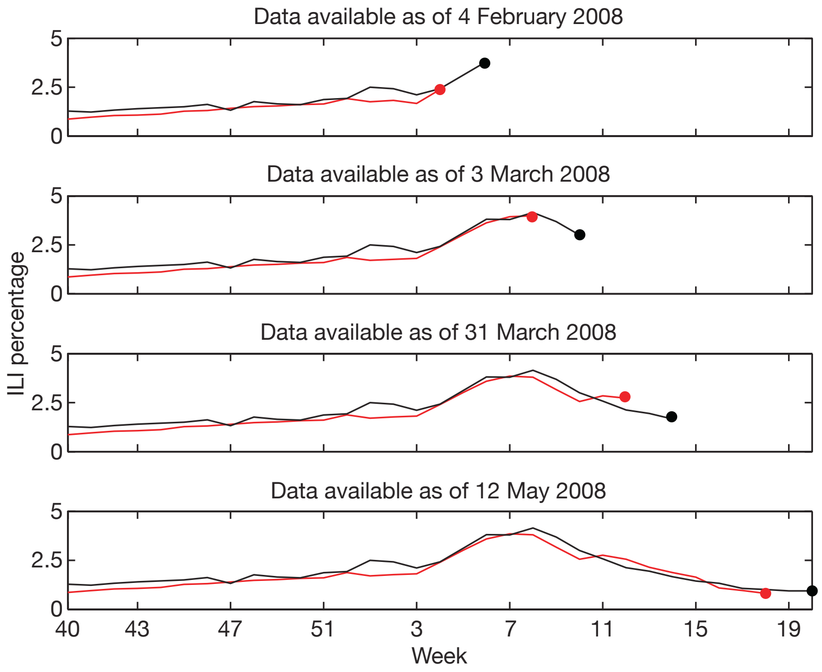


- Partition nodes into training and test sets?
  - Breaks up triads; omitted edges “share” information across training and test (diagram: blue are edges that include node 7)
- Partition dyads?
  - Breaks up nodes; even worse
- Can’t *eliminate*, but can *minimize* optimism by careful data splitting

Importance of out-of-sample testing



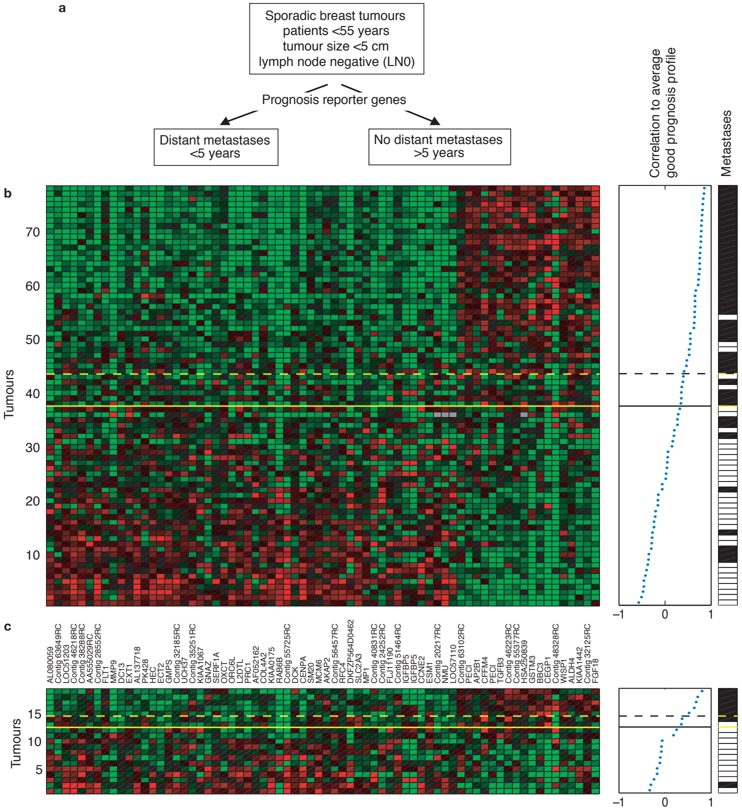
“Things do change”



Ginsberg et al., 2012

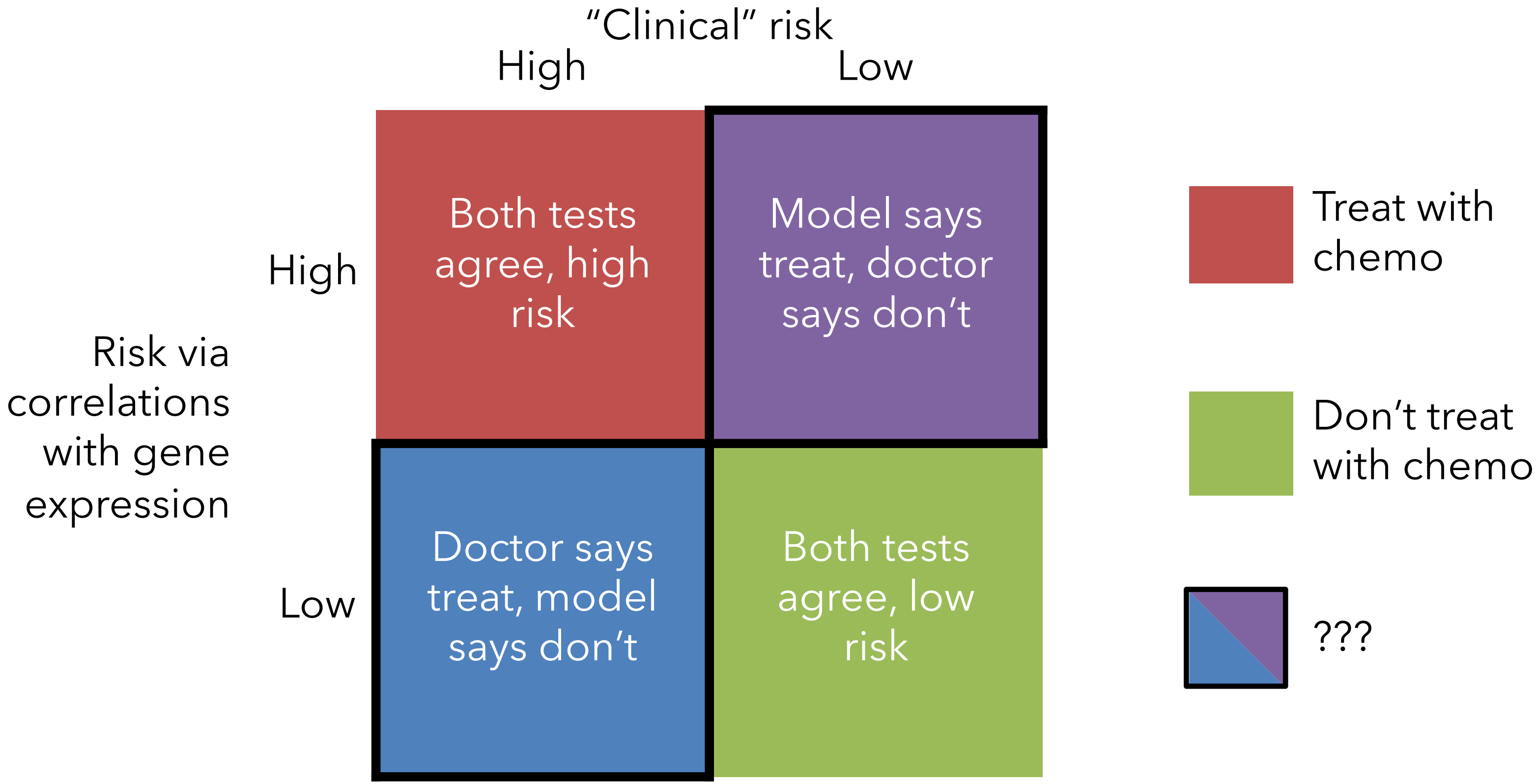
Santillana et al., 2014

Real-world testing of ML results



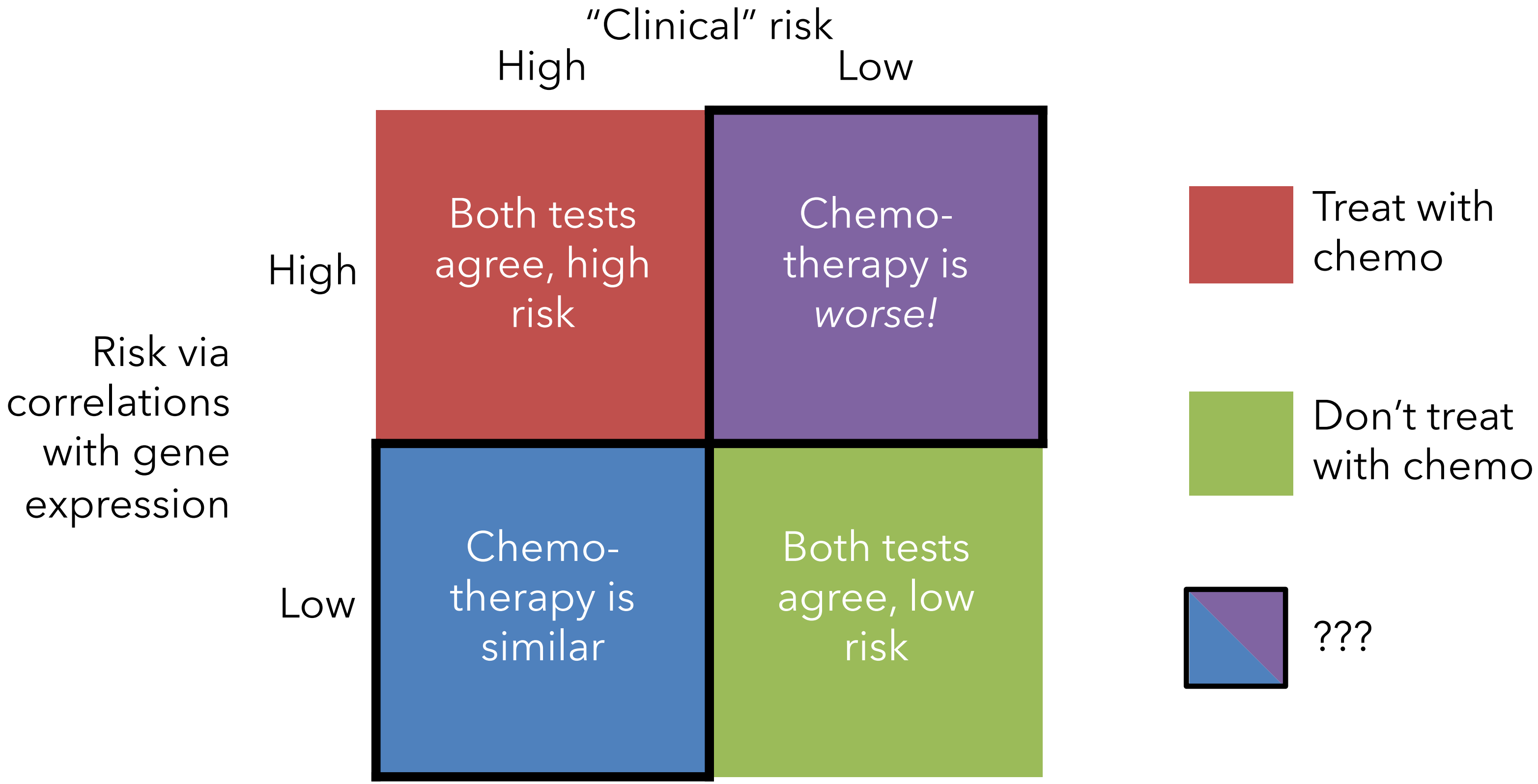
- van't Veer et al. (2002) found 70 genes correlated with developing breast cancer
- Of course the correlations were optimal, post-hoc. But did it generalize?

Implementation testing



Cardoso et al., 2016, *NEJM*

Implementation testing



Cardoso et al., 2016, *NEJM*

Implementation testing

Risk via correlations with gene expression	"Clinical" risk: High	"Clinical" risk: Low
High	Both tests agree, high risk	Chemo-therapy is *worse!*
Low	Chemo-therapy is similar	Both tests agree, low risk

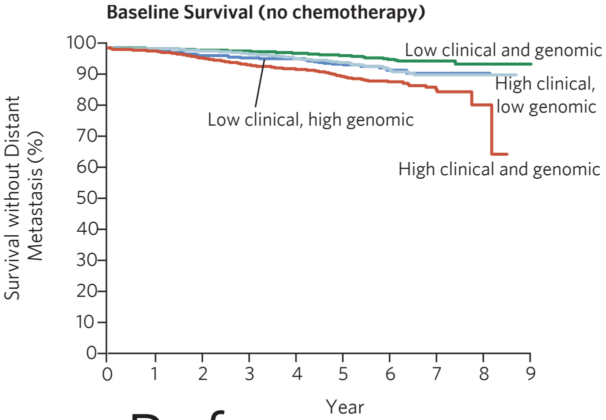
Treat with chemo (red)

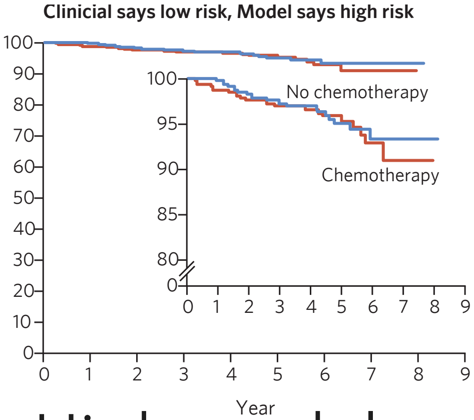
Don't treat with chemo (green)

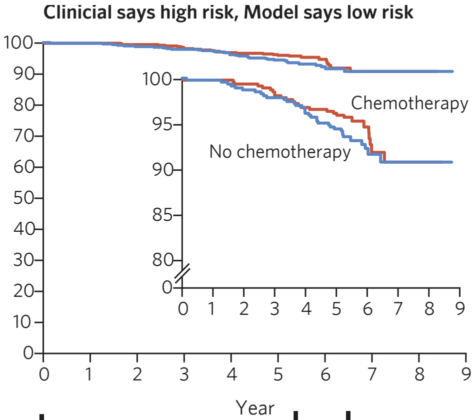
Finding: Machine learning *alone* would make things worse. But as a *secondary* diagnosis, on average it catches false positives and avoids unhelpful chemo!

Cardoso et al., 2016, *NEJM*

Implementation testing: Details







- Before experiment (training data)

(Note: still limitations in how experimental subjects may be unrepresentative.)

- High model risk, low clinical risk: randomize. Chemo *worse*!

- Low model risk, high clinical risk: chemo makes no difference

Summary and conclusion

Summary

- Biases exist, but are not simple, and may be unknowable in general
  - When we have a comparison source, we can calibrate, but better may be to find appropriate use cases
- Commercial platforms are not always fit for research; but we can try to investigate how their design and incentives
- Machine learning presents new opportunities, but has multiple failure points (often corresponding to long-recognized problems) that must be recognized and dealt with
  - Prediction, and cross validation, have fragilities
  - Out-of-sample testing is always a good idea

Conclusion

- There are problems with data; these have been widely recognized, and we are making progress on how to work with new forms of data, including opaque secondary and commercial data
- There are still fundamental problems limitations of different modeling approaches, and how modeling relates to the world: machine learning is supercharging these, forgetting lessons of the past

Other work of mine

- This is what I think is most interesting to this audience
- I have other work:
  - Trying to use modeling to *imagine* alternative states of the world (Richardson, Malik et al., 2021)
  - Why do "technical" people have such a narrow view of the world, and how do some come to change? (Malik & Malik, 2021)
- My current work is on "AI ethics", which I take to mean, how do we rigorously and responsibly develop and deploy (or choose not to develop or deploy) modeling, applied to large-scale data? How do we choose what modeling approach is appropriate (if any)?
  - This is largely aimed at biomedical use cases, but developed guidance will, I hope, apply to any field of social science

References (1/3)

Arceneaux, 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://dx.doi.org/10.1177/0049124110378098

Borgatti, Steve. 2012. Types of validity. BA 762: Research Methods. Gatton College of Business & Engineering, University of Kentucky. https://sites.google.com/site/ba762researchmethods/reference/handouts/types-of-validity

Box, George E. P. 1979. *Robustness in the strategy of scientific model building*. Technical Report #1954. Mathematics Research Center, University of Wisconsin-Madison.

Breiman, Leo. 2001. Statistical modeling: The two cultures. *Statistical Science* 16 (3): 199–231. https://dx.doi.org/10.1214/ss/1009213726

Bryman, Alan. 1988. *Quantity and quality in social research*. Routledge. https://doi.org/10.4324/9780203410028

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, Sylvain Causeret, Mauro DeLorenzi, Annuska M. Glas, Vassilis Golfinopoulos, Theodora Goulioti, Susan Knox, Erika Matos, Bart Meulemans, Peter A. Neijenhuis, Ulrike Nitz, Rodolfo Passalacqua, Peter Ravdin, Isabel T. Rubio, Mahasti Saghatchian, Tineke J. Smilde, Christos Sotiriou, Lisette Stork, Carolyn Straehle, Geraldine Thomas, Alastair M. Thompson, Jacobus M. van der Hoeven, Peter Vuylsteke, René Bernards, Konstantinos Tryfonidis, Emiel Rutgers, and Martine Piccart. 2016. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine* 375 (8): 717–729. https://dx.doi.org/10.1056/NEJMoa1602253

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis." *Science* 349 (6248): 636–638. https://dx.doi.org/10.1126/science.aaa9375

Efron, Bradley. 2004. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association* 99 (467): 619–632. https://dx.doi.org/10.1198/016214504000000692

Gilbert, Nigel, and Klaus Troitzsch. 2005. *Simulation for the social scientist*. 2nd edition. Open University Press.

References (2/3)

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1015. https://dx.doi.org/10.1038/nature07634

Hoadley, Bruce. 2001. [Statistical modeling: The two cultures]: Comment. *Statistical Science* 16 (3): 220-224.

Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142 (2): 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Jacobs, Abigail Z., and Hanna Wallach. 2021.Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 375-85. https://doi.org10.1145/3442188.3445901

Jones, Matthew L. 2015. How we became instrumentalists (again): Data positivism since World War II. *Historical Studies in the Natural Sciences* 48 (5): 673–684. https://dx.doi.org/10.1525/hsns.2018.48.5.673

Kass, Robert E. 2011. Statistical inference: The big picture. *Statistical Science* 26 (1): 1–9. https://dx.doi.org/10.1214/10-STS337

Keeling, Matt J., and Pejman Rohani. 2008. *Modeling infectious diseases in humans and animals*. Princeton University Press. https://doi.org/10.1515/9781400841035

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105 (5): 491–495. https://dx.doi.org/10.1257/aer.p20151023

Koren, Yehuda. 2009. Collaborative filtering with temporal dynamics.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google Flu: Traps in big data analysis. *Science* 343 (6176): 1203–1205. https://dx.doi.org/10.1126/science.1248506

Malik, Maya and Momin M. Malik. 2021. Critical technical awakenings. Journal of Social Computing 2 (4): 365–384. https://doi.org/10.23919/JSC.2021.0035

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

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2(13): 1–33. https://doi.org/10.3389/fdata.2019.00013

References (3/3)

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2(13): 1–33. https://doi.org/10.3389/fdata.2019.00013

Park, Greg. 2012. The dangers of overfitting: A Kaggle postmortem. http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/

Rescher, Nicholas. 1998. *Predicting the future: An introduction to the theory of forecasting*. State University of New York Press.

Richardson, Eugene T., Momin M. Malik, William A. Darity, Jr., A. Kirsten Mullen, Michelle E. Morse, Maya Malik, Adia Benton, Mary T. Bassett, Paul E. Farmer, Lee Worden, and James Holland Jones. 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

Santillana, Mauricio, Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *American Journal of Preventive Medicine* 47 (3): 341–347. http://dx.doi.org/10.1016/j.amepre.2014.05.020

Savage, Mike, and Roger Burrows. 2007. The coming crisis of empirical sociology. *Sociology* 41 (5): 885–899. https://doi.org/10.1177/0038038507080443

Shmueli, Galit. 2010. To explain or to predict? *Statistical Science* 25 (3): 289–310. https://dx.doi.org/10.1214/10-STS330

Tasse, Dan, Zichen Liu, Alex Sciuto, and Jason I. Hong. 2017. State of the geotags: Motivations and recent changes. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (ICWSM-2017), 250–259.

Teele, Dawn Langan. 2014. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. Yale University Press.

van't Veer, Laura J., Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, no. 6871 (2002): 530–536. https://dx.doi.org/10.1038/415530a

Zagheni, Emilio, and Ingmar Weber. 2015. Demographic research with non-representative internet data. *International Journal of Manpower* 36 (1): 13-25. https://doi.org/10.1108/IJM-12-2014-0261