

model classification is in any case not a place
because this road is endless*.

Revisiting

'ALL MODELS ARE WRONG':

**Addressing Limitations in Big Data,
Machine Learning, and Computational
Social Science**

Momin M. Malik

Data Science Postdoctoral Fellow

Berkman Klein Center for Internet & Society at Harvard University

Wednesdays@NICO, 05 February 2020

Northwestern Institute on Complex Systems

Northwestern University, Evanston, Illinois

model elaboration is in any case not a pleasure
because this road is endless*.

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

Now it would be very remarkable if any model in the real world could be exactly represented by a mathematical model. However, cunningly chosen parsimonious

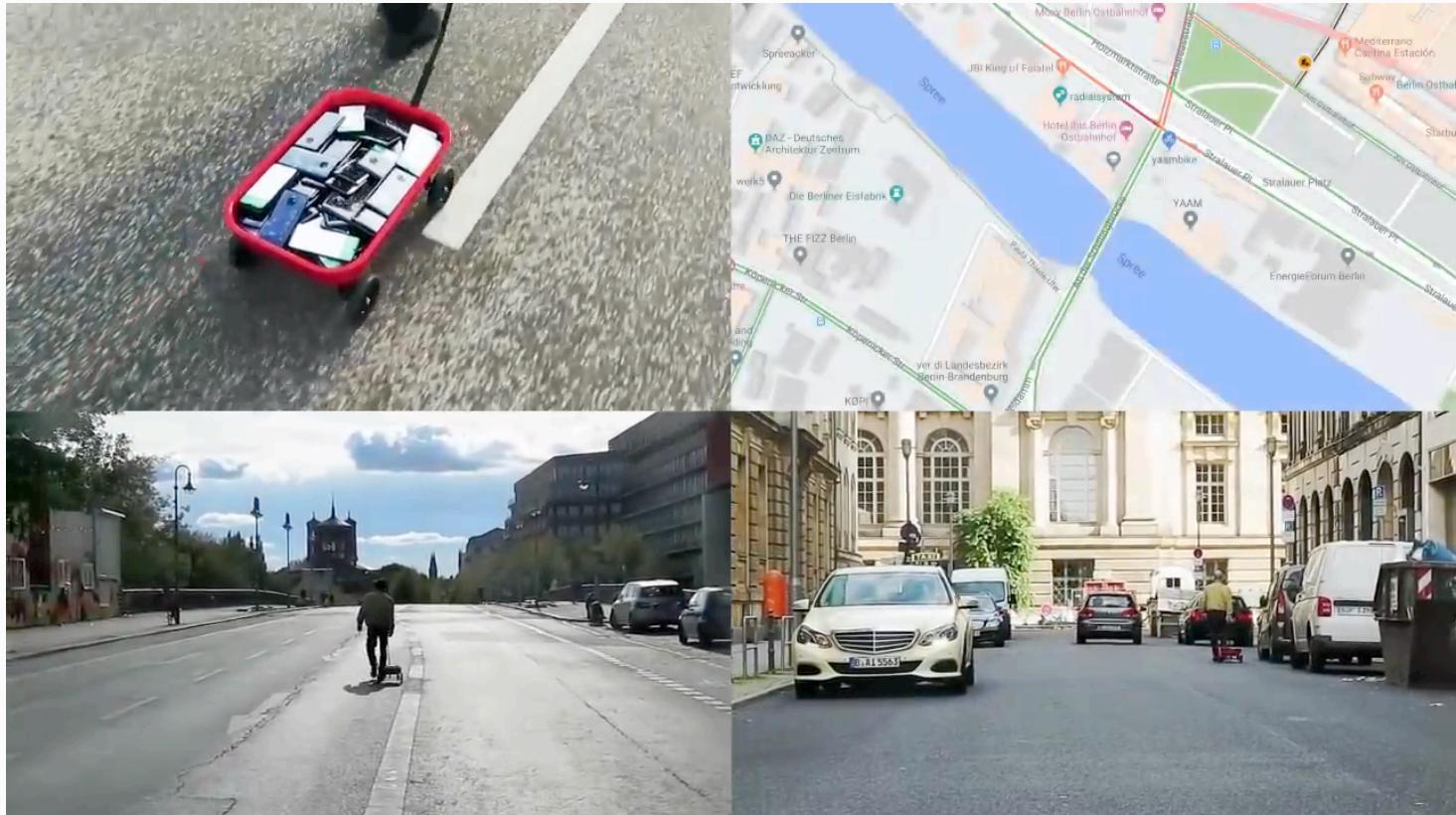
* Suppose for example that in advance of any model of the form of (1) with the usual normal distribution it might be objected that the distribution

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



› Simon Weckert, "Google Maps Hack"

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References



➤ This shows larger themes

- Available data are often only a *proxy*
- So long as the proxy is never the thing itself, it can fail
- Models of relationships and processes, too, are not the things themselves
- Box: “[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?’.”

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

- **When, how data/models are *wrong***
- **When and how it matters**
- **What we can do**

> Outline

- > Introduction
- > Bias in geotagged tweets
- > Platform effects in social media
- > Tradeoffs in types of modeling
- > Dependencies and cross validation
- > Discussion and conclusion
- > References

- > Introduction
- > Bias in geotagged tweets
- > Platform effects in social media
- > Trade-offs in types of modeling
- > Dependencies and cross validation
- > Discussion and conclusion

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

› About me

- ›  DEPARTMENT OF THE
HISTORY OF SCIENCE
HARVARD UNIVERSITY
- ›   Berkman
The Berkman Center for Internet & Society
at Harvard University
- ›  
- › **Carnegie Mellon University**
School of Computer Science
- › **Data Science For Social Good**
Summer Fellowship
- ›  **BERKMAN KLEIN CENTER**
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY



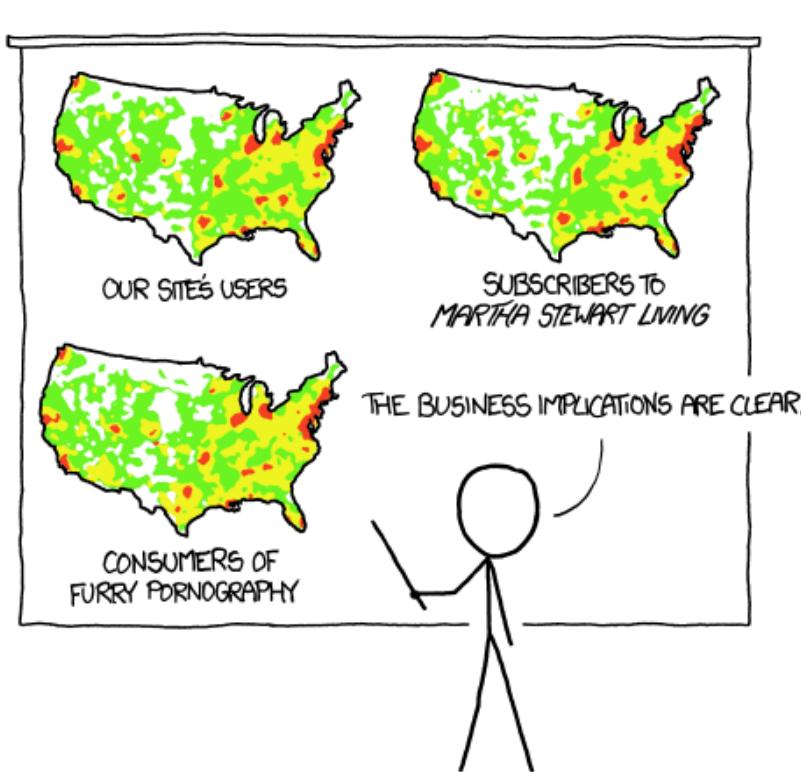
- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

› 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.
https://www.mominmalik.com/malik_chapter1.pdf

► Many maps just show population

- Introduction
- Bias in geotagged tweets
- Platform effects in social media
- Tradeoffs in types of modeling
- Dependencies and cross validation
- Discussion and conclusion
- References

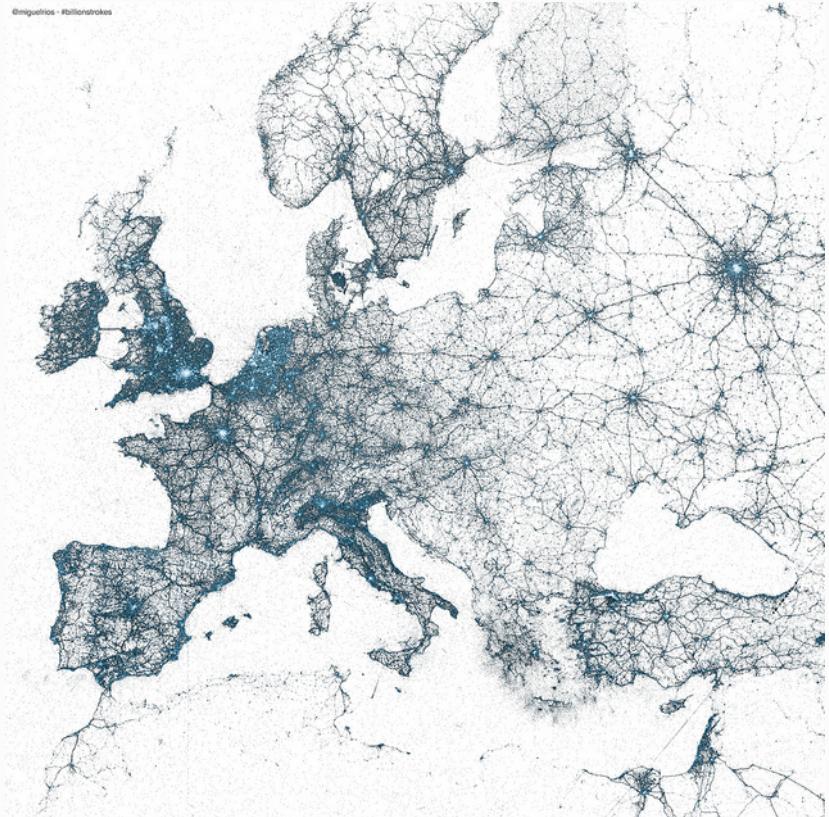


Randall Munroe. 2012. Heatmap. <https://xkcd.com/1138/>

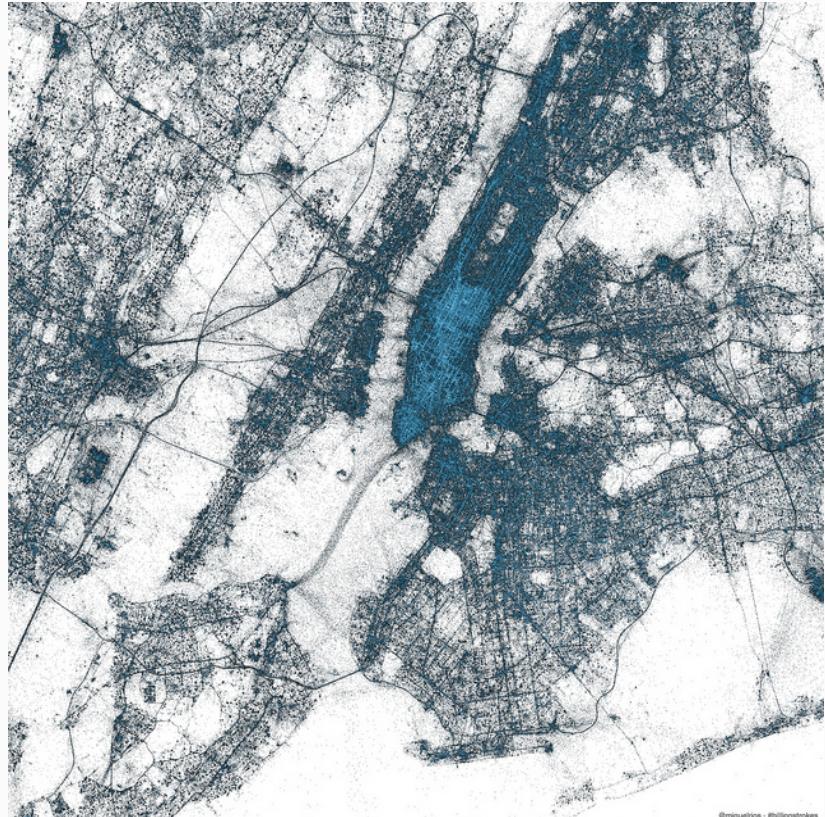
Revisiting “All Models are Wrong”

➤ But maybe we can use this?

- Introduction
- Bias in geotagged tweets
- Platform effects in social media
- Tradeoffs in types of modeling
- Dependencies and cross validation
- Discussion and conclusion
- References



Revisiting "All Models are Wrong"



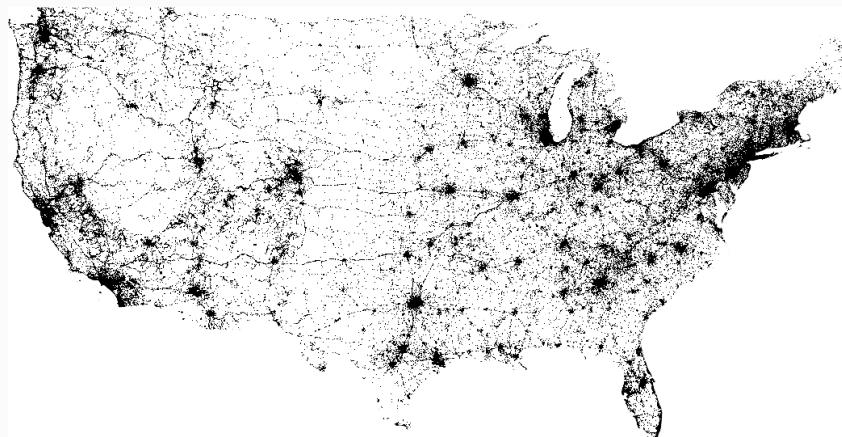
11 of 57

<https://MominMalik.com/nico2020.pdf>

› Do tweets measure population?

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

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:

$$\log U_i = \log \alpha + \log P_i + \varepsilon'_i$$

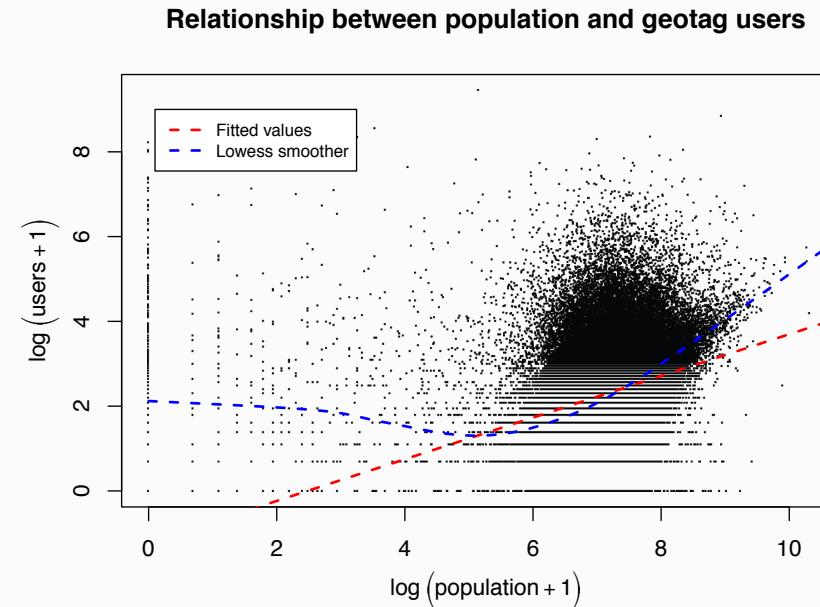
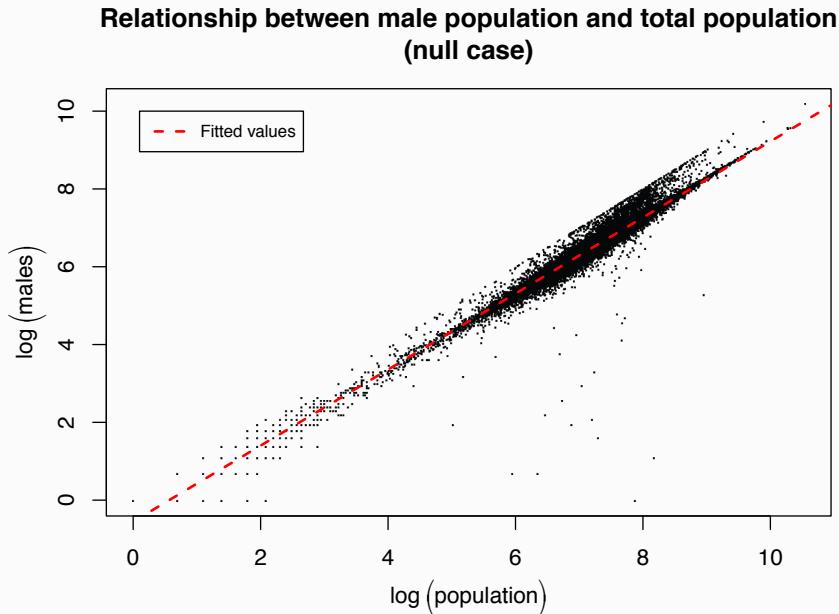
› Compare to a linear model:

$$\log U_i = \beta_0 + \beta_1 \log P_i + \varepsilon'_i$$

Result: Not proportional

- ▶ Introduction
- ▶ Bias in geotagged tweets
- ▶ Platform effects in social media
- ▶ Tradeoffs in types of modeling
- ▶ Dependencies and cross validation
- ▶ Discussion and conclusion
- ▶ References

(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

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

Hurricane Sandy, tweets vs. damage/deaths



Shelton et al., 2014.

Revisiting “All Models are Wrong”

› Responses to demographic bias

- › Model the biases!
- › Calibration and weighting
- › Use data for appropriate questions
 - “Postcards, not ticket stubs” (Tasse et al., 2017)
- › Find clever study designs or data comparisons, establish *panels*, etc.

› Introduction

› Bias in
retweeted
tweets› Platform
effects in
social media› Tradeoffs in
types of
modeling› Dependencies
and cross
validation› Discussion
and
conclusion

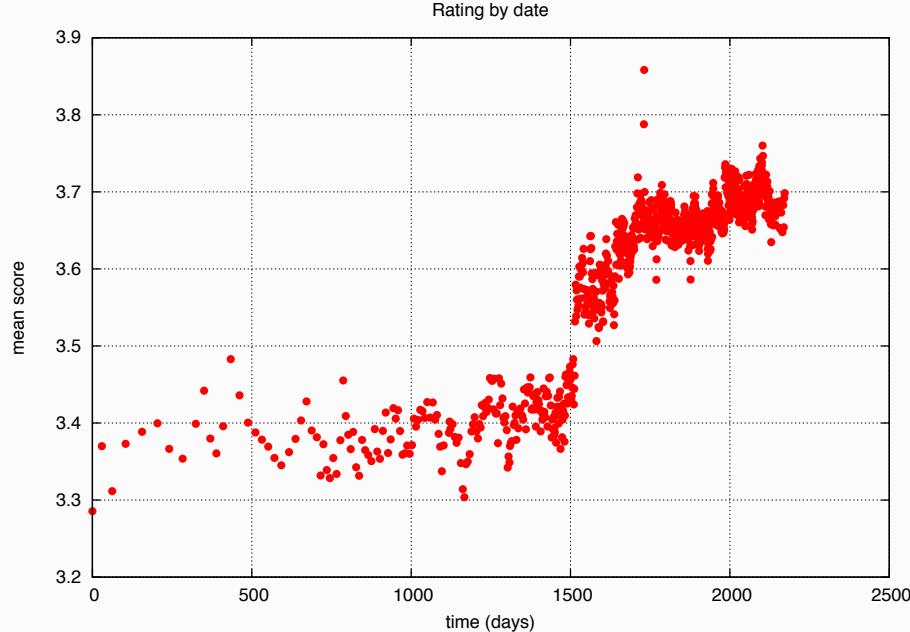
› References

› 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. https://www.mominmalik.com/malik_chapter2.pdf

► Design can cause/change behavior

- Introduction
- Bias in eotagged tweets
- Platform effects in social media
- Tradeoffs in types of modeling
- Dependencies and cross validation
- Discussion and conclusion
- References



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

Koren, 2009.

- › Introduction
- › Bias in retagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

› Social media platforms are businesses



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

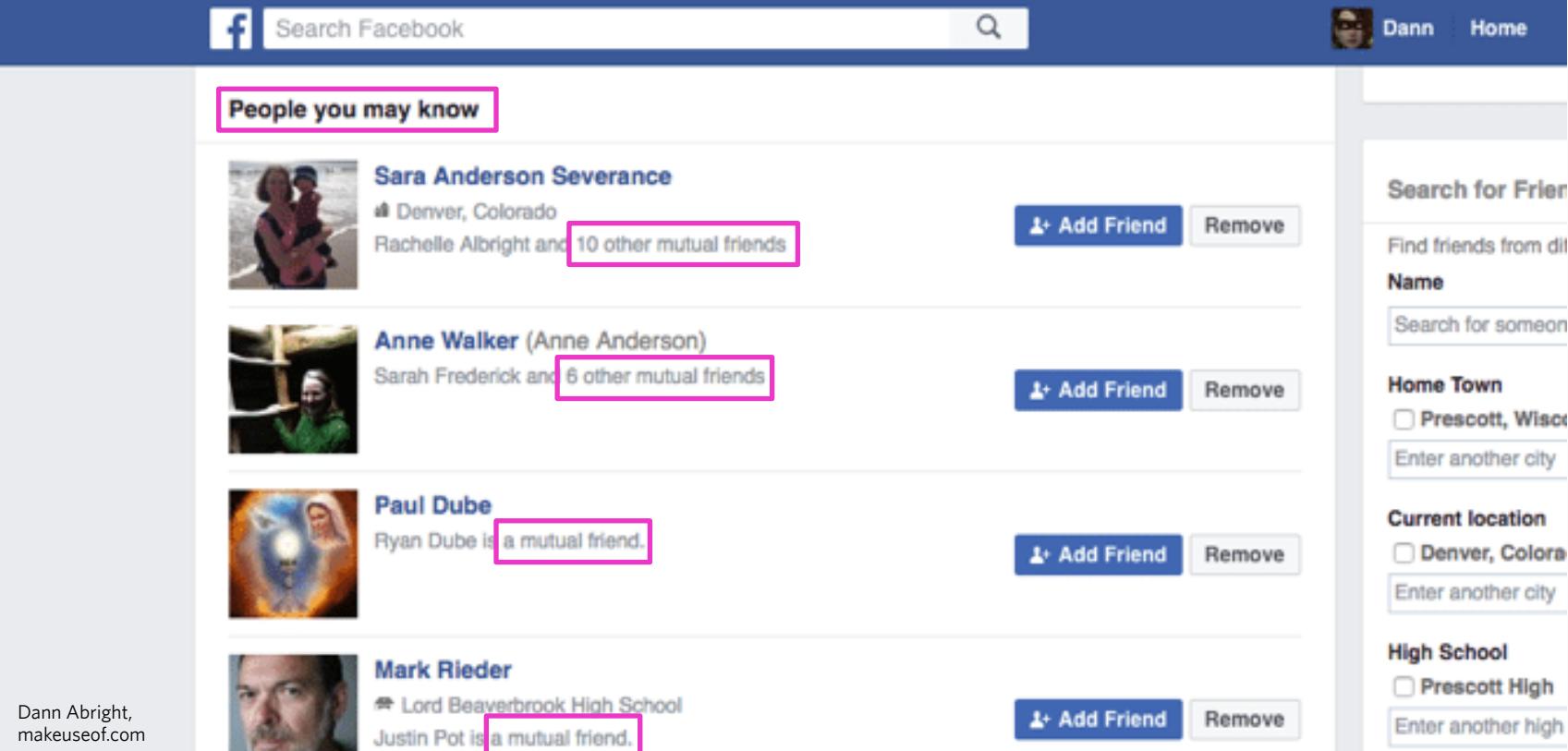
► Sites try to grow their users' networks

- Introduction
- Bias in retagged tweets
- Platform effects in social media
- Tradeoffs in types of modeling
- Dependencies and cross validation
- Discussion and conclusion
- References

The image is a composite of two screenshots from social networking platforms. On the left, the LinkedIn homepage is shown with a banner that reads "It's easier than ever to grow your professional network". Below this, there is a section titled "INTRODUCING THE NEW" with a box around the "People You May Know" feature. On the right, a screenshot of the Twitter interface shows a "Who to follow" section. This section includes a search bar, a list of recommended users (Keton Kakkar, William Bumpas, Rich Boroff), and "Follow" buttons next to each profile. The entire image is framed by a thick black border.

➤ Recommending “friend-of-a-friend”

- Introduction
- Bias in eotagged tweets
- Platform effects in social media
- Tradeoffs in types of modeling
- Dependencies and cross validation
- Discussion and conclusion
- References



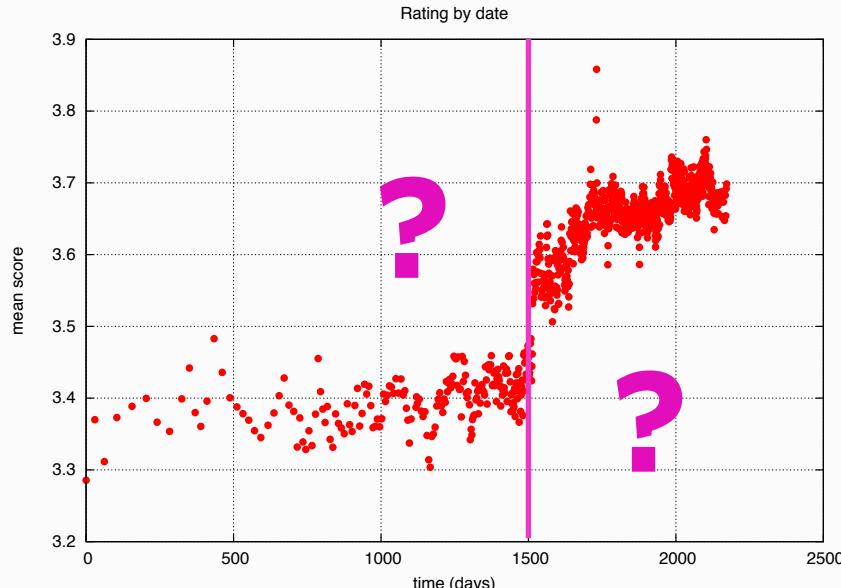
The screenshot shows a Facebook interface with a blue header bar containing the 'f' logo, a search bar, and a user profile for 'Dann'. Below the header, a sidebar on the left lists navigation items: 'Introduction', 'Bias in eotagged tweets', 'Platform effects in social media', 'Tradeoffs in types of modeling', 'Dependencies and cross validation', 'Discussion and conclusion', and 'References'. The main content area displays a 'People you may know' section with four profiles:

- Sara Anderson Severance**: Located in Denver, Colorado. Shows 'Rachelle Albright and 10 other mutual friends'. A pink box highlights the '10 other mutual friends' text.
- Anne Walker (Anne Anderson)**: Shows 'Sarah Frederick and 6 other mutual friends'. A pink box highlights the '6 other mutual friends' text.
- Paul Dube**: Shows 'Ryan Dube is a mutual friend.' A pink box highlights the 'Ryan Dube is a mutual friend.' text.
- Mark Rieder**: Shows 'Lord Beaverbrook High School' and 'Justin Pot is a mutual friend.' A pink box highlights the 'Justin Pot is a mutual friend.' text.

On the right side of the main content, there are search fields for 'Search for Friend' (with placeholder 'Find friends from all'), 'Name', and 'Search for someone' (with placeholder 'Search for someone'). Below these are sections for 'Home Town' (with placeholder 'Prescott, Wisconsin') and 'Current location' (with placeholder 'Denver, Colorado'). At the bottom, there is a 'High School' section (with placeholder 'Prescott High') and a page footer with the URL 'https://MominMalik.com/nico2020.pdf'.

➤ 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*

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References



The Matrix (1999) "déjà vu" scene

› Data artifacts as natural experiments

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

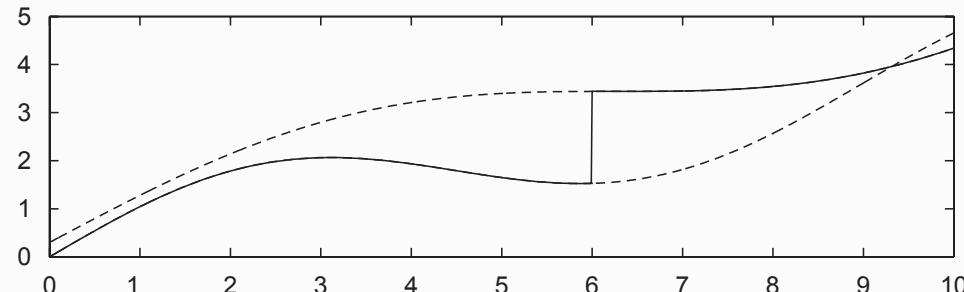
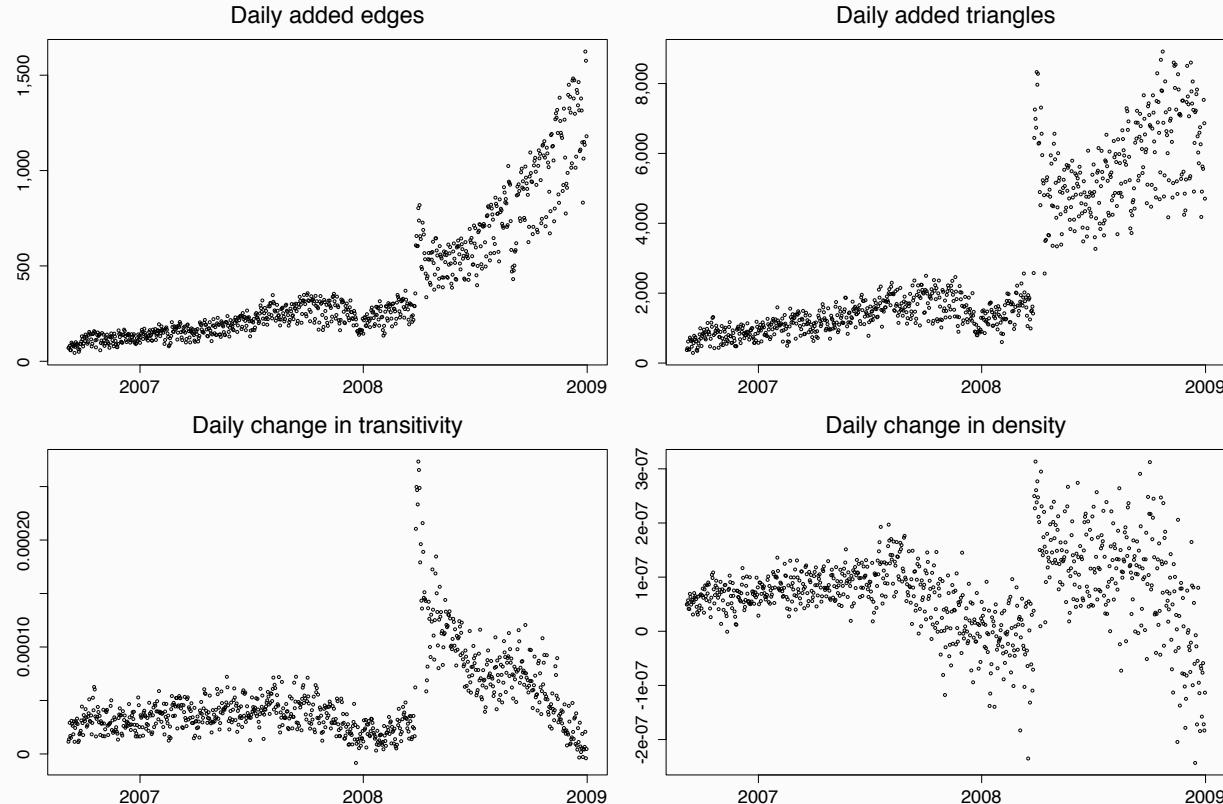


Fig. 2 from Imbens and Lemieux (2008): Potential and observed outcome regression functions.

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

Case: Facebook's "People You May Know"

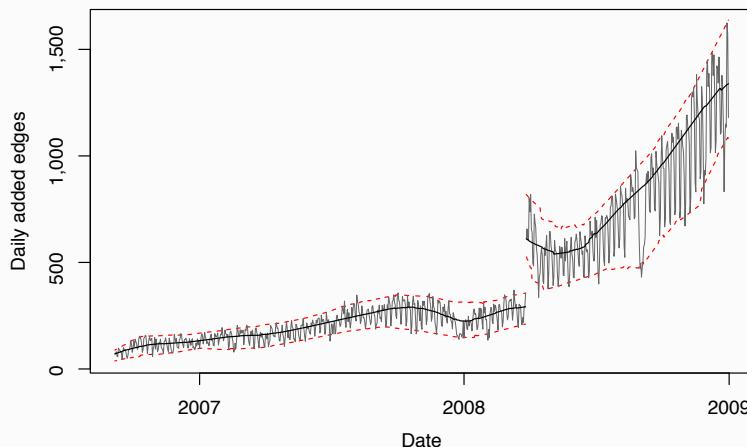
- ▶ Introduction
- ▶ Bias in eotagged tweets
- ▶ Platform effects in social media
- ▶ Tradeoffs in types of modeling
- ▶ Dependencies and cross validation
- ▶ Discussion and conclusion
- ▶ References



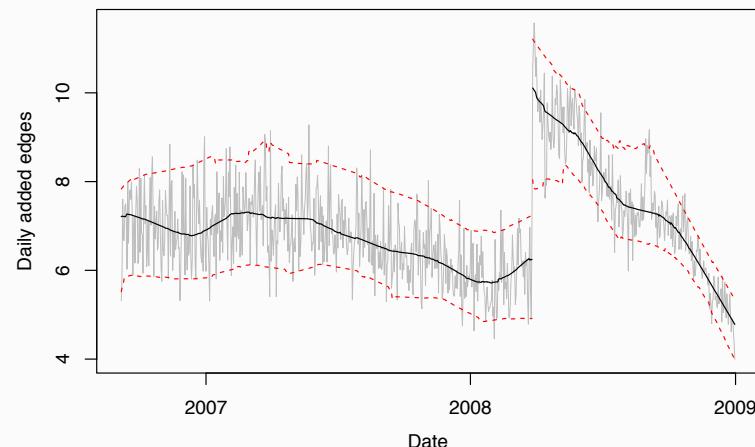
› PYMK changed the Facebook network!

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

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



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

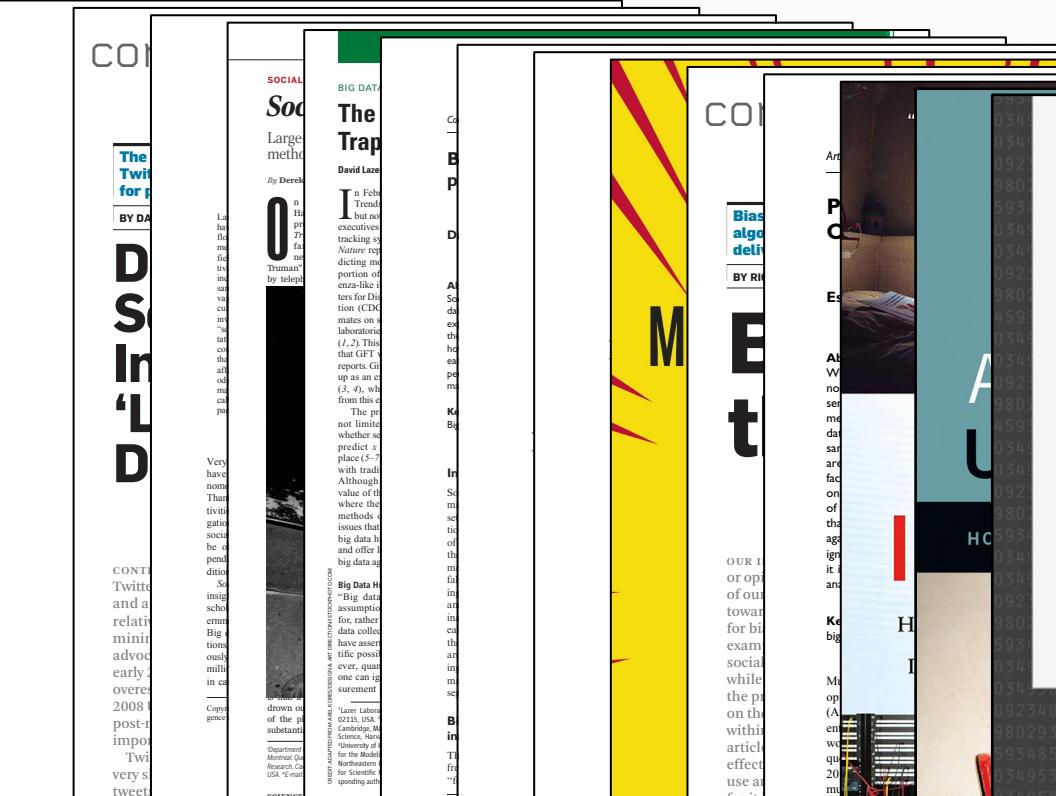


› Responses to platform effects

- › 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

- ▶ Introduction
- ▶ Bias in retagged tweets
- ▶ Platform effects in social media
- ▶ Tradeoffs in types of modeling
- ▶ Dependencies and cross validation
- ▶ Discussion and conclusion
- ▶ References

► Data well-studied; *models*, not yet



Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Oltanu^{1,*}, Carlos Castillo², Fernando Diaz² and Emre Kiciman¹

¹Microsoft Research, New York, NY, United States, ²Microsoft Research, Montreal, QC, Canada, ³Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, ⁴Microsoft 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 Franklin¹

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 or interaction. It includes a variety of platforms—for social media and communication (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

¹Correspondence: Alexandra Oltanu, alexandra.oltanu@microsoft.com

²Specialty section: This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

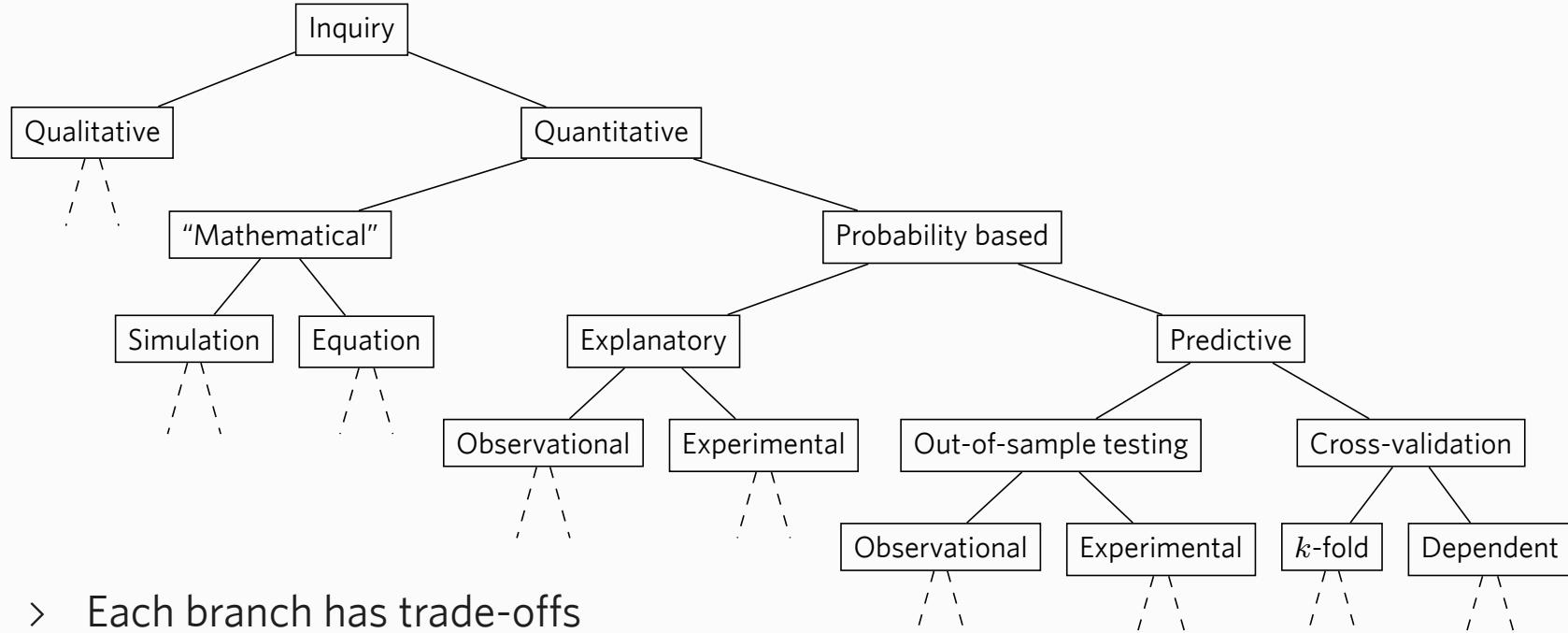
³Received: 26 February 2019; Accepted: 27 May 2019; Published: 11 July 2019

- › Introduction
- › Bias in eotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

➤ Tradeoffs in types of modeling

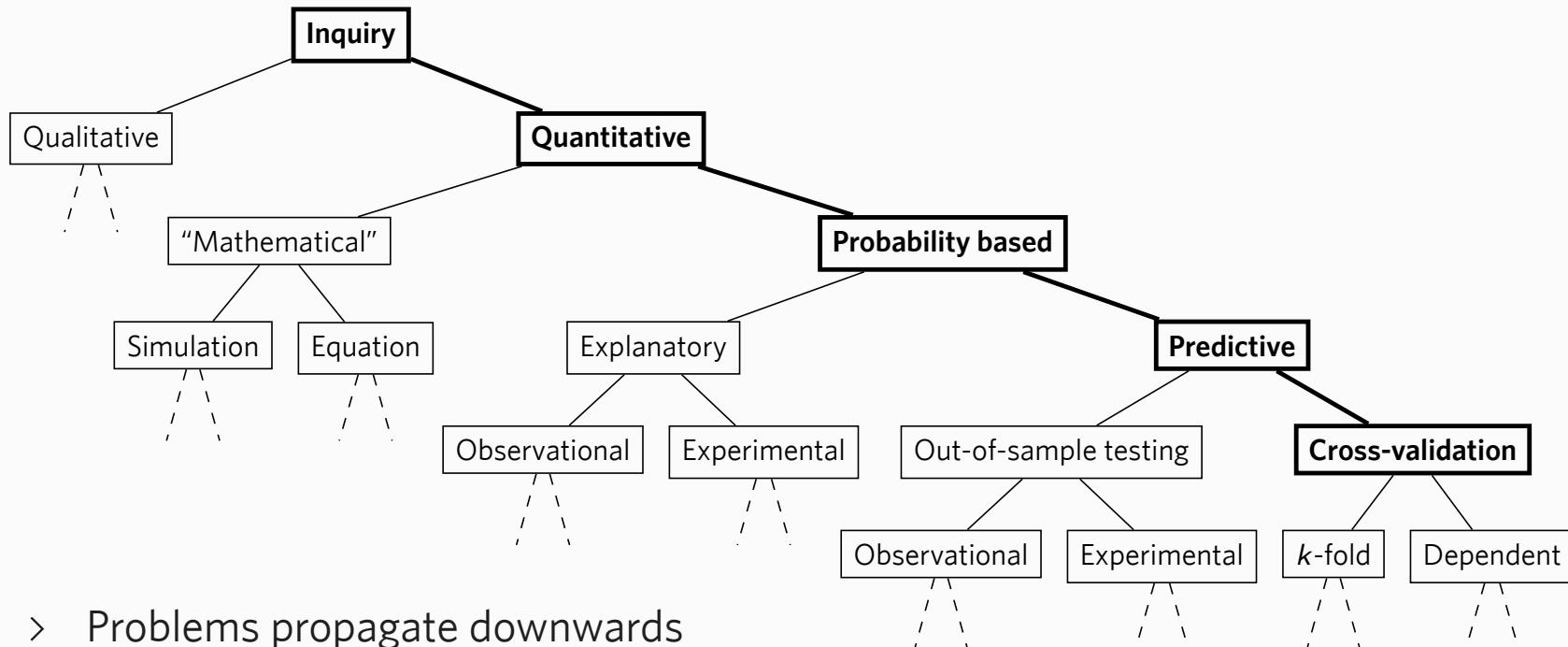
Momin M. Malik. 2020. A hierarchy of limitations in machine learning. In submission.
https://www.mominmalik.com/hierarchy_draft.pdf

> Approaches to research



- > Each branch has trade-offs
- > No one method is better than any other
- > Mixed methods can combine

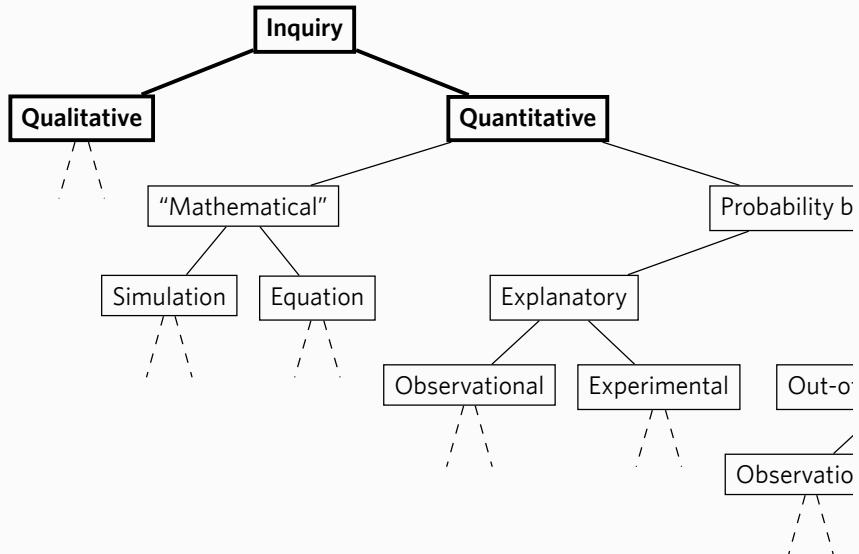
> Typical machine learning



- > Problems propagate downwards
- > E.g., quantification affects everything below

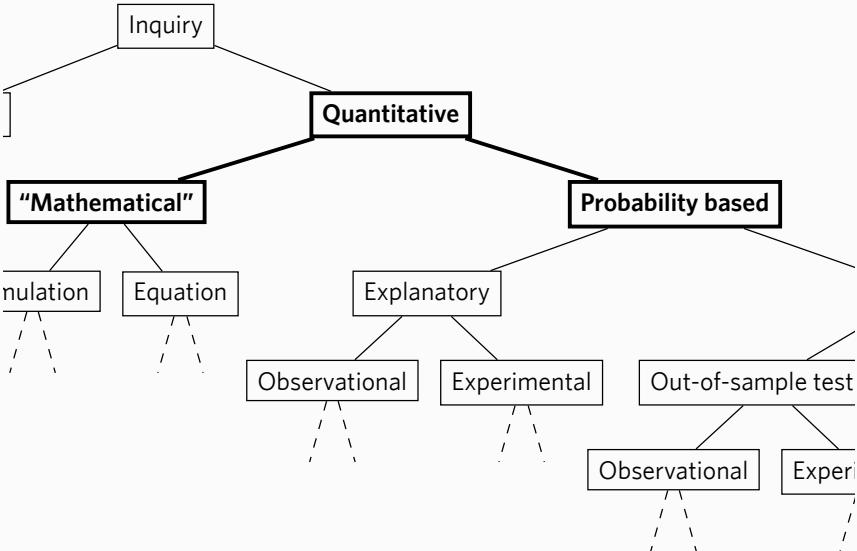
- › Introduction
- › Bias in retagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References

› 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

- › Introduction
- › Bias in geotagged tweets
- › Platform effects in social media
- › Tradeoffs in types of modeling
- › Dependencies and cross validation
- › Discussion and conclusion
- › References



- › Statistics and machine learning are the only options to both directly use data and account for variability
- › They do so via central tendency
- › This requires multiple observations, and independence assumptions

› Introduction

› Bias in
retweeted
tweets

› Platform
effects in
social media

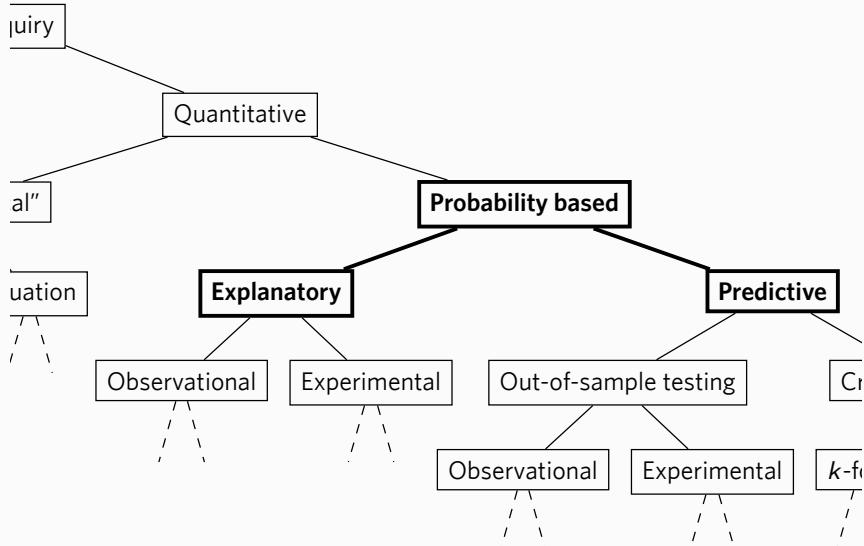
› Tradeoffs in
types of
modeling

› Dependencies
and cross
validation

› Discussion
and
conclusion

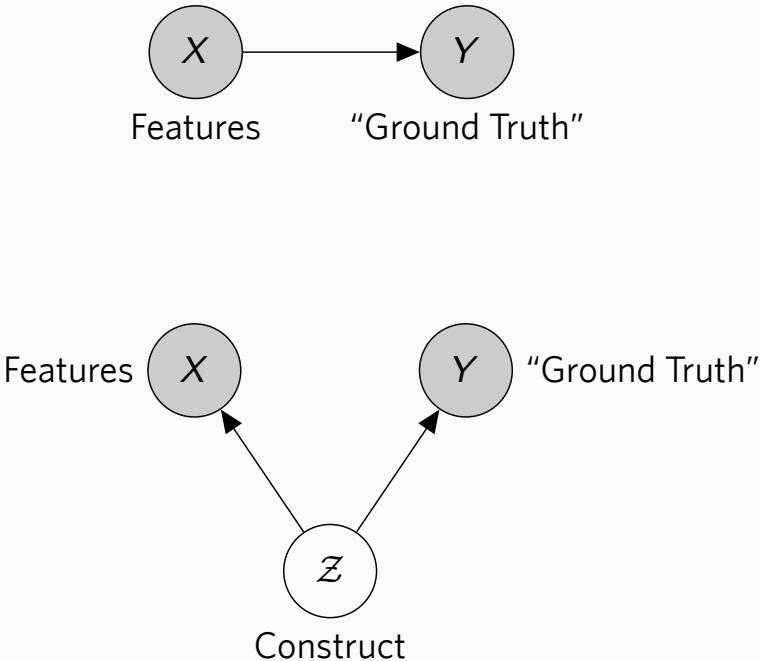
› References

› 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

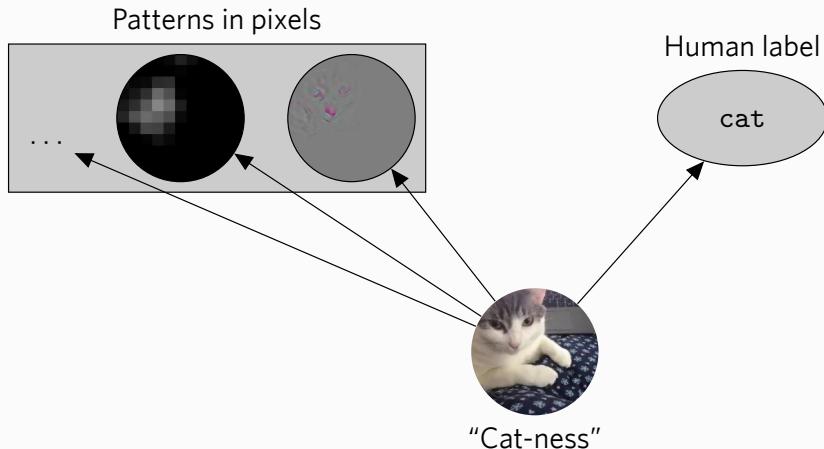
› Prediction misses constructs



- › 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

- Introduction
- Bias in eotagged tweets
- Platform effects in social media
- Tradeoffs in types of modeling
- Dependencies and cross validation
- Discussion and conclusion
- References



› Responses to problems of proxy

- › Identify/define the underlying construct
- › How does the correlation work? Where does it fail?
- › Treat “ground truth” labels as *measurements*; investigate validity
- › Use machine learning for *scaling subjective human judgments*, rather than thinking it uncovers underlying “truth”



» Introduction

» Bias in
retweeted
tweets

» Platform
effects in
social media

» Tradeoffs in
types of
modeling

» Dependencies
and cross
validation

» Discussion
and
conclusion

» References

► Dependencies and cross validation

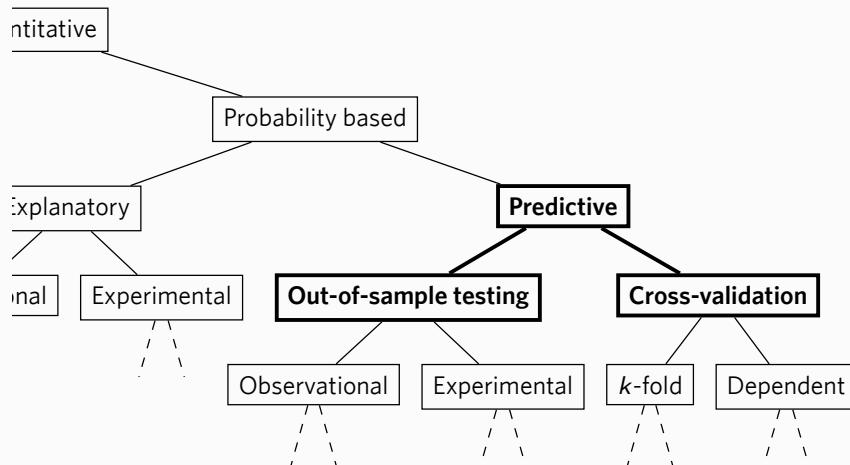
Momin M. Malik. 2020. A hierarchy of limitations in machine learning. In submission.
https://www.mominmalik.com/hierarchy_draft.pdf

▶ Introduction

▶ Bias in
retweeted
tweets▶ Platform
effects in
social media▶ Tradeoffs in
types of
modeling▶ Dependencies
and cross
validation▶ Discussion
and
conclusion

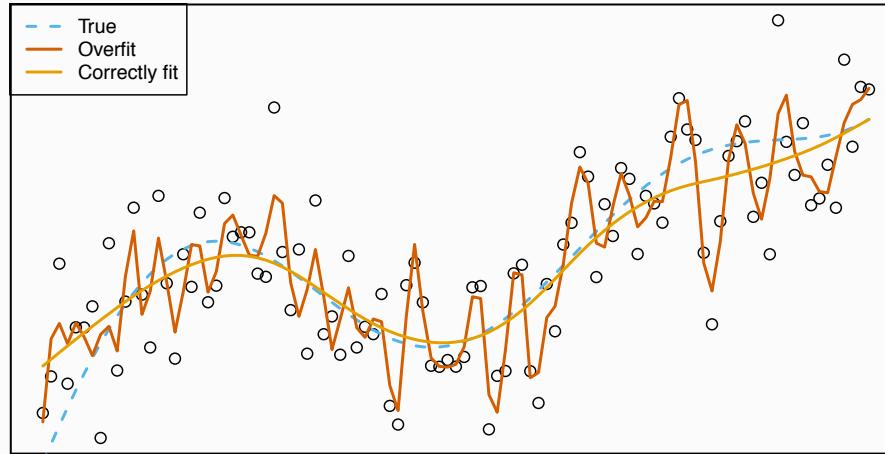
▶ References

► Performance claims are from cross-validation



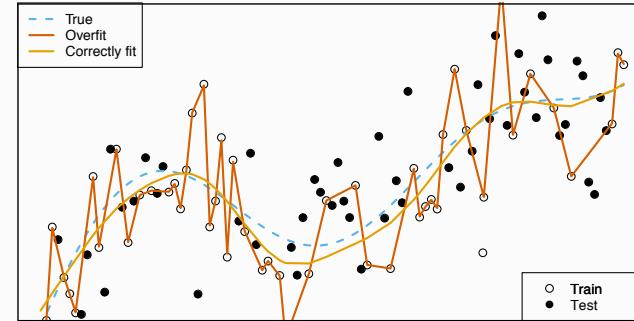
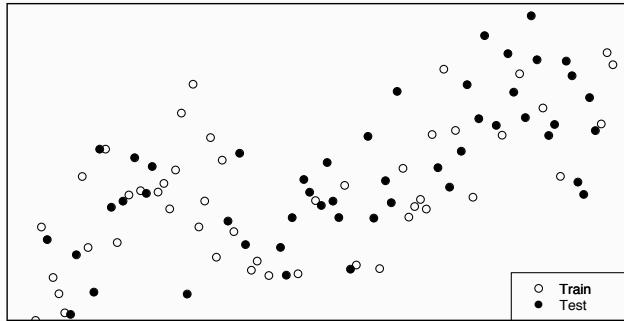
- Rescher (1998) notes every prediction involves a metaprediction: do we think the prediction works?
- Cross-validation is metaprediction for ML
- But, how well does cross-validation work?

› Purpose of cross-validation



- › If we are no longer guided by theory, and use automatic methods, we risk overfitting: fitting to 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

- » Introduction
- » Bias in eotagged tweets
- » Platform effects in social media
- » Tradeoffs in types of modeling
- » Dependencies and cross validation
- » Discussion and conclusion
- » References

➤ Classic argument for CV

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

› Apply this to non-iid data

- › Imagine we have, for $\Sigma_{ii} = \sigma^2$ and $\Sigma_{ij} = \rho\sigma^2$, $i \neq j$

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

- › Then, optimism in the training set is:

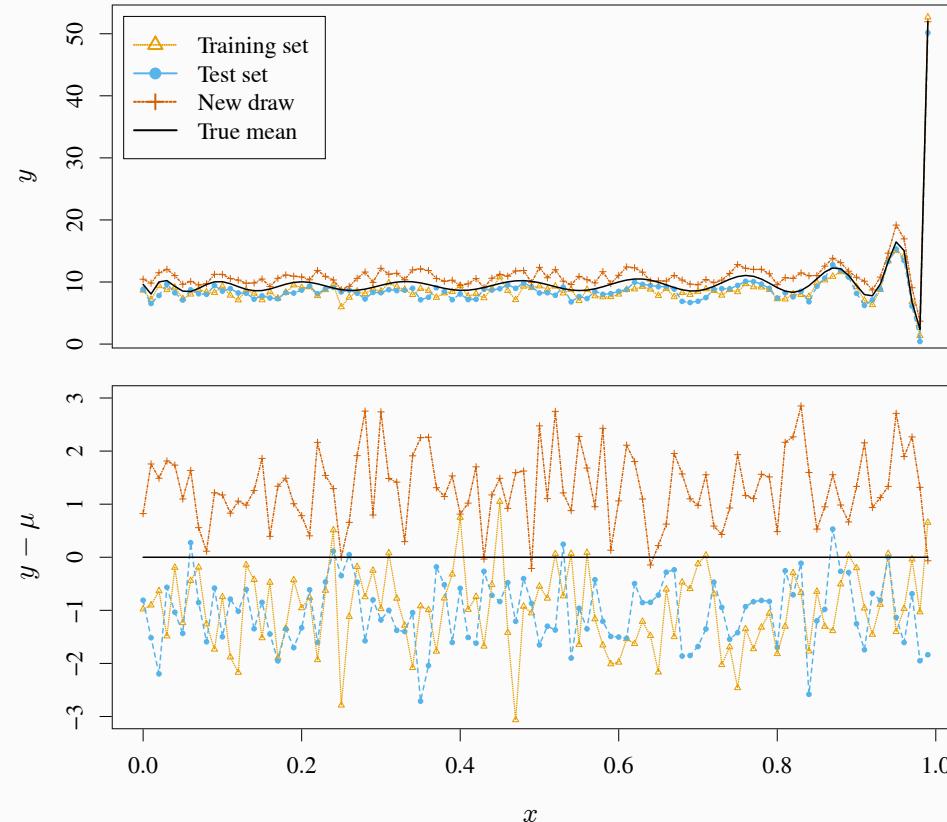
$$\frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_1, \hat{Y}_1) = \frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_1, \mathbf{H}Y_1) = \frac{2}{n} \operatorname{tr} \mathbf{H} \operatorname{Var}_f(Y_1) = \frac{2}{n} \operatorname{tr} \mathbf{H} \Sigma$$

- › But test set also has nonzero optimism!

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

- » Introduction
- » Bias in eotagged tweets
- » Platform effects in social media
- » Tradeoffs in types of modeling
- » Dependencies and cross validation
- » Discussion and conclusion
- » References

» Simulating the toy example



▶ Introduction

▶ Bias in
retweeted
tweets

▶ Platform
effects in
social media

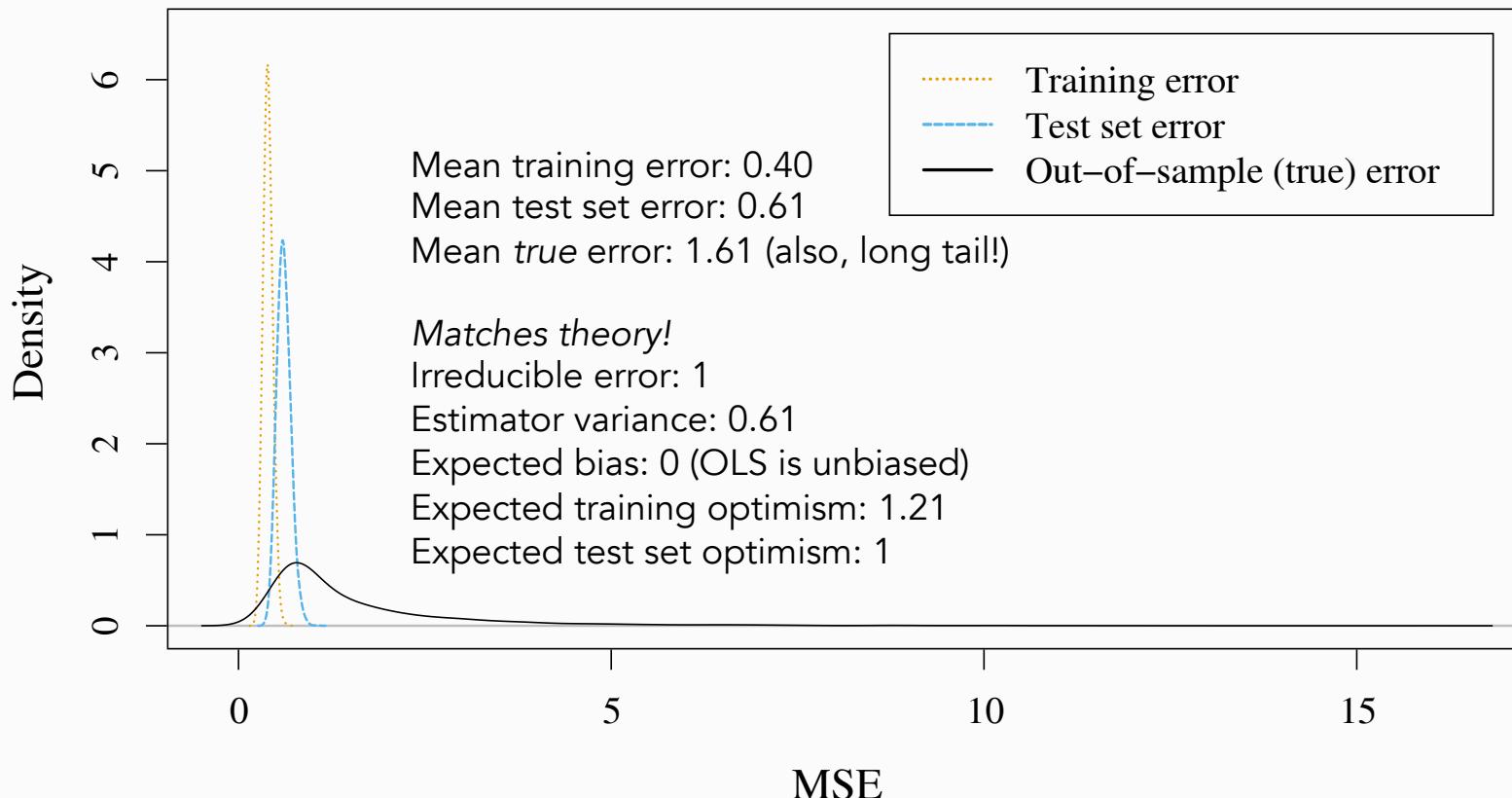
▶ Tradeoffs in
types of
modeling

▶ Dependencies
and cross
validation

▶ Discussion
and
conclusion

▶ References

➤ 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

› Responses to failures in CV

- › Do *true* out-of-sample testing
- › Do experimental testing if predictions used for decisions (Cardoso et al., 2014)
- › All performance claims are preliminary until such testing
- › Language: maybe use “retrodiction” and “back-testing,” or simply “correlation,” instead of “prediction” to not mislead
- › For robustness, maybe do statistics instead



- » Introduction
- » Bias in eotagged tweets
- » Platform effects in social media
- » Tradeoffs in types of modeling
- » Dependencies and cross validation
- » Discussion and conclusion
- » References

► Discussion and conclusion

➤ Larger themes for this work

- “Confirmation holism,” and “experimenter’s regress”: if we don’t like a result, we can always find *something* to challenge
- We should do this even when we *do* like a result
- Box: “this road is endless...”
- Qualitative, critical, and theoretical social science can guide, especially around where and how claims of universalism and objectivity support injustice
- Data and models should *reflect* understandings of the world, not *define* them

➤ The work to be done

- We have a good idea of where biases are; but work remains in quantifying them
- Modelers should be trained with clear articulations of limitations of data and modeling
- Mixed methods probably the most promising way forward for research
 - Qualitative annotation for “ground truth” (Patton et al., 2019)
 - Experimental design for testing machine learning

► References (1/6)

- Abbott, Andrew. 1988. Transcending general linear reality. *Sociological Theory* 6 (2): 169–186.
<https://dx.doi.org/10.2307/202114>.
- Agre, Philip E. 1997. Towards a critical technical practice: Lessons learned from trying to reform AI. In *Social science, technical systems, and cooperative work: Beyond the great divide*, edited by Geoffrey C. Bowker, Susan Leigh Star, Will Turner, and Les Gasser, 131–158. Lawrence Erlbaum Associates.
<https://web.archive.org/web/20040203070641/http://polaris.gseis.ucla.edu/pagre/critical.html>.
- Agre, Philip E. 2000, July 12. Notes on critical thinking, Microsoft, and eBay, along with a bunch of recommendations and some URL's. *Red Rock Eater Newsletter*.
<https://pages.gseis.ucla.edu/faculty/agre/notes/00-7-12.html>.
- Bailey, David H., Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2014. Pseudo-mathematics and financial charlatanism: The effects of backtest

- overfitting on out-of-sample performance. *Notices of the AMS* 61 (5): 458–471.
<https://dx.doi.org/10.1090/noti1105>.
- Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis* 120: 70–83.
<https://dx.doi.org/10.1016/j.csda.2017.11.003>.
- Borgatti, Steve. 2019. Types of validity. BA 762: Research Methods. Gatton College of Business & Engineering, University of Kentucky.
<https://sites.google.com/site/ba762researchmethods/materials/handouts/typesofvalidity>.
- 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>

► References (2/6)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

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 Saghatelian, 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>.

Chatfield, Chris. 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158 (3): 419-466.
<https://dx.doi.org/10.2307/2983440>.

Cox, David R. 1990. Role of models in statistical analysis.

Statistical Science 5 (2): 169-174.
<https://dx.doi.org/10.1214/ss/1177012165>

Doshi-Velez, Finale and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.
<https://arxiv.org/abs/1702.08608>.

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>.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222: 309-368.
<https://dx.doi.org/10.1098/rsta.1922.0009>.

► References (3/6)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

Gayo-Avello, Daniel. 2012. "I wanted to predict elections with Twitter and all I got was this lousy paper": A balanced survey on election prediction using Twitter data.
<https://arxiv.org/abs/1204.6441>.

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>.

Hammerla, Nils Y., and Thomas Plötz. 2015. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp '15), 1041-1051.
<https://dx.doi.org/10.1145/2750858.2807551>.

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

Jones, Matthew L. 2018. How we became instrumentalists Revisiting "All Models are Wrong"

(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>.

Keyes, Os. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* (CSCW), 2, 88:1-88:22.

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. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '09), 447-456.
<https://dx.doi.org/10.1145/1557019.1557072>.

► References (4/6)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

- Lanius, Candice. 2015, January 15. Fact check: Your demand for statistical proof is racist. *Cyborgology*.
<https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/>.
- Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespiagnani. 2014. The parable of Google Flu: Traps in big data analysis. *Science* 343 (6176): 1203-1205.
<https://dx.doi.org/10.1126/science.1248506>.
- Lipton, Zachary C. 2015. The myth of model interpretability. *KDnuggets* 15 (13).
<https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html>.
- Lipton, Zachary C. and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship.
<https://arxiv.org/abs/1807.03341>.
- Messerli, Franz H. 2012. Chocolate consumption, cognitive function, and Nobel laureates. *The New England Journal of Medicine* 367: 1562-1564.
<https://dx.doi.org/10.1056/NEJMoa1211064>.
- 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.
<https://dx.doi.org/10.3389/fdata.2019.00013>.
- Opsomer, Jean, Yuedong Wang, and Yuhong Yang. 2001. Nonparametric regression with correlated errors. *Statistical Science* 16 (2): 134-153.
<https://dx.doi.org/10.1214/ss/1009213287>.
- Park, Greg. 2012. The dangers of overfitting: A Kaggle postmortem.
<http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/>.

► References (5/6)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

Patton, Desmond U., Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS-52)*, 2142–2151.
<https://dx.doi.org/10.24251/HICSS.2019.260>.

Patton, Michael Quinn. 2014. The nature, niche, value, and fruit of qualitative inquiry. In *Qualitative research & evaluation methods: Integrating theory and practice*, 4th edition, 2–44. SAGE Publications, Inc.
https://uk.sagepub.com/sites/default/files/upm-binaries/64990_Patton_Ch_01.pdf.

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

Rose, Todd. 2016. *The end of average: How we succeed in a world that values sameness*. HarperOne. See excerpt at
<https://www.thestar.com/news/insight/2016/01/16/>

[when-us-air-force-discovered-the-flaw-of-averages.html](#). Animated video:
<https://vimeo.com/237632676>.

Rosset, Saharon, and Ryan J. Tibshirani. 2019. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*.
<https://dx.doi.org/10.1080/01621459.2018.1424632>.

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>.

Shapiro, Ian. 2014. Methods are like people: If you focus only on what they can't do, you will always be disappointed. In *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, edited by Dawn Langan Teele, 228–241. Yale University Press.

► References (6/6)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. Mapping the data shadows of Hurricane Sandy: Uncovering the socio-spatial dimensions of 'big data.' *Geoforum* 52: 167–179.
<http://dx.doi.org/10.1016/j.geoforum.2014.01.006>.

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

Spirites, Peter and Kun Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* 3 (3): 1–28.
<https://dx.doi.org/10.1186/s40535-016-0018-x>.

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.
<https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15588>.

Tibshirani, Robert. 2015, December 6. Recent advances in post-selection inference. Breiman Lecture, NeurIPS 2015.
<http://statweb.stanford.edu/~tibs/ftp/nips2015.pdf>

Wallach, Hanna. 2018. Computational social science ≠ computer science + social data. *Communications of the ACM* 61 (3): 42–44.
<https://dx.doi.org/10.1145/3132698>.

Wasserman, Larry A. 2013. Rise of the machines. In *Past, present, and future of statistical science*, 525–536. Chapman and Hall/CRC.
<http://www.stat.cmu.edu/~larry/Wasserman.pdf>.