Generalizability, meaningfulness, and meaning: Machine learning in the social world

Momin M. Malik

Senior Data Science Analyst – AI Ethics, Center for Digital Health, Mayo Clinic
Instructor, School of Social Policy & Practice, University of Pennsylvania
Fellow, Institute in Critical Quantitative, Computational, & Mixed Methodologies

Seminario Conjunto de Estadística y Ciencia de Datos
Centro de Investigación en Matemáticas
[online], 2023 May 10.
Slides: https://www.MominMalik.com/cimat2023.pdf

CIMAT

Goals and summary

- This talk is aimed at students and researchers in statistics and data science who:
  - Want to make a positive impact on the social world using the probability-based modeling
  - May not have any background outside of modeling about how to do this
- (Arguably) unlike mechanical or natural systems, the ways in which we are a part of, and relate to, social systems is of enormous importance; and the details we ignore matter much more
  - Modeling is abstraction. The very power of abstraction is that it ignores "irrelevant" parts of a system and highlights "relevant" parts: but relevance is a judgement call (perhaps relevance is uniquely determined by a goal, but then that goal is the judgement call)
  - Mechanical and natural systems don't have inner lives. For social systems, that our modeling needs to either ignore it (which is behavioralism, and that has even empirical problems) or try to account for it (which requires "flattening" the infinite multiplicity of meanings)
  - The failure points of probability-based models often cannot be expressed only in internal terms (empirical inadequacy for stated goals), but in how they contradict competing goals
- This one talk is not enough to cover everything: but the goal is to help orient you to relevant concepts, debates, arguments, and literature

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Outline

- Motivation: Kentaro Toyama, *Geek Heresy*
- Reviewing the nature of machine learning and statistics versus other methodologies
  - Measurement and quantification
  - Causal mismatch
- Uncertainty quantification reveals internally relevant failure points
- Dealing with context: reflexivity and positionality

Motivation

MAYO CLINIC

- Introduction
- **Motivation: Empirical failure**
- Stats/ML and trade-offs in methodology
- Uncertainty quantification
- Reflexivity and positionality
- Summary and conclusion
- References

Kentaro Toyama

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

Kentaro Toyama, Flickr (July 22, 2011)

Kentaro Toyama, "David – teacher training 2", Flickr (November 6, 2013)

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Kentaro Toyama

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

Kentaro Toyama, Flickr (July 22, 2011)

Relevant works

- Agre, 1997, "Towards a critical technical practice"
  - Agre reflects on how his AI education produced intellectual narrowness, and how he broke out of it
- Freedman, 2009, *Statistical models and causal inference: A dialogue with the social science*
  - Collection of notable works by the late great statistician. Notable articles: "Statistical models and shoe leather", "What is the chance of an earthquake?", "On specifying graphical models for causation, and the identification problem", and many more
- Wagstaff, 2012, "Machine learning that matters"
  - The limitations of the common task framework and progress on benchmark datasets
- Morozov, 2013, *To save everything, click here: The folly of technological solutionism*
  - Intro and chapter 1 have a great argument against "solutionism", and for technological *enrichment* instead
- Toyama, 2015, *Geek heresy: Rescuing social change from the cult of technology*
- Selbst et al., 2019, "Fairness and abstraction in sociotechnical systems"
  - A notable review of the ways that abstraction breaks down in sociotechnical systems
- Jacobs & Wallach, 2020, "Measurement and fairness"
  - Review of measurement theory for machine learning
- Raji et al., 2022, "The fallacy of AI functionality"
  - Looks at how and why, among deployed AI systems, many do not work

MAYO CLINIC

- Introduction
- **Motivation: Empirical failure**
- Stats/ML and trade-offs in methodology
- Uncertainty quantification
- Reflexivity and positionality
- Summary and conclusion
- References

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

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

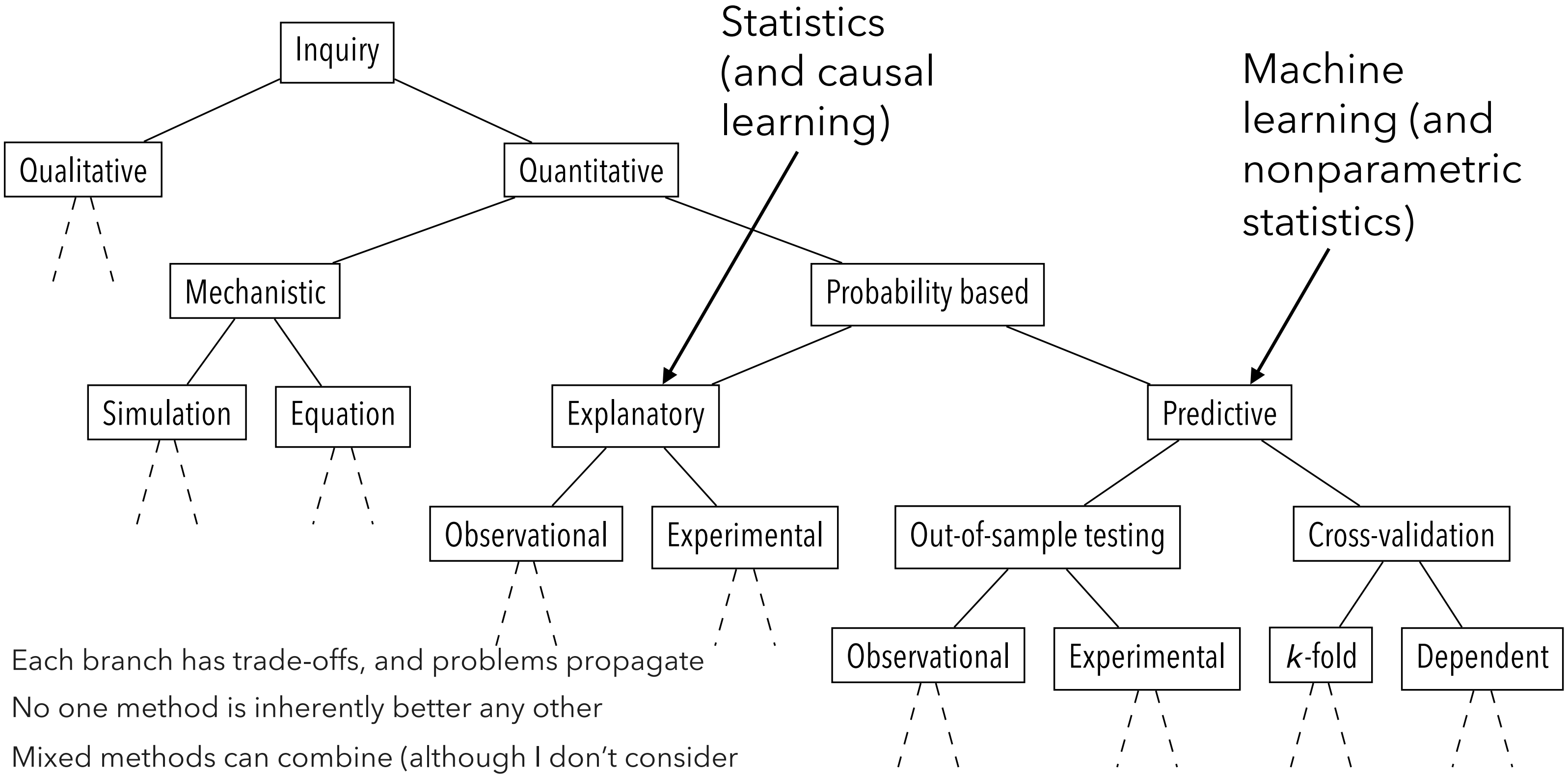
Reflexivity and positionality

Summary and conclusion

References

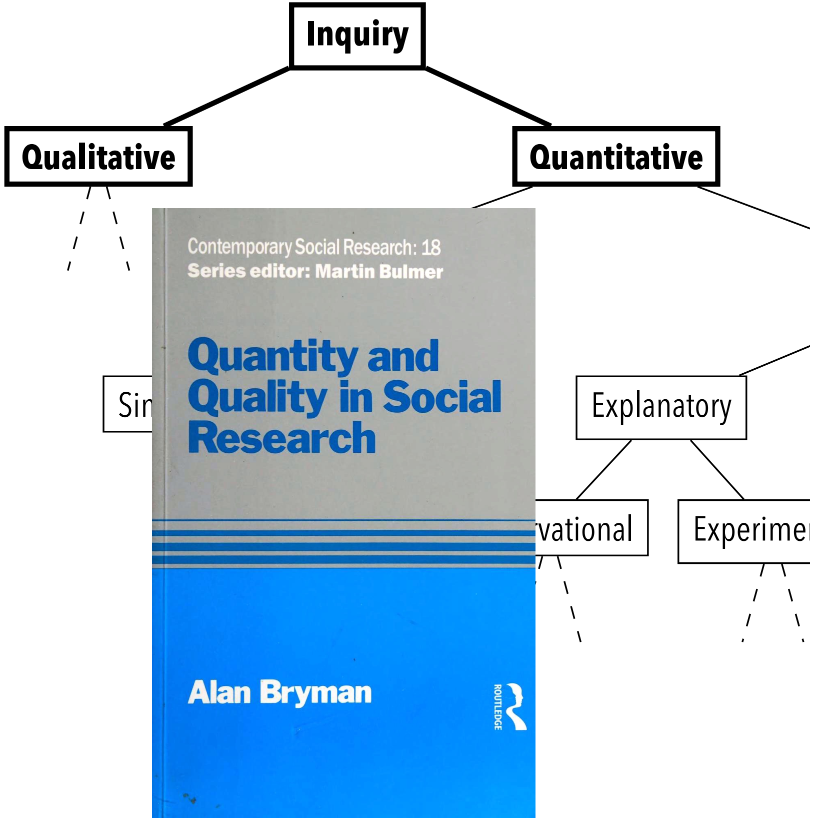
Stats/ML versus other methodologies: Limitations and trade-offs

Tree of methodologies (Malik, 2020)



- 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 ("all models are wrong;" Box, 1979)

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

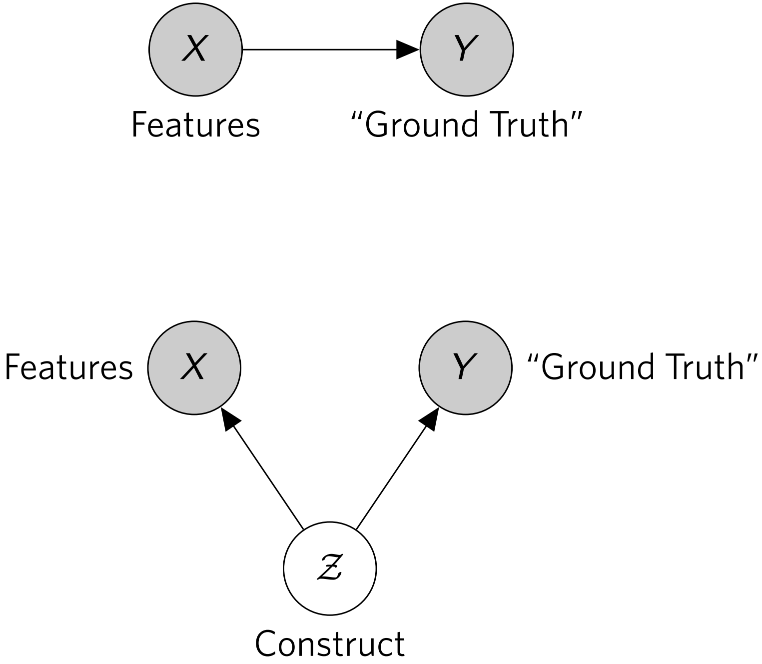
Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

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), and even can be gamed

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

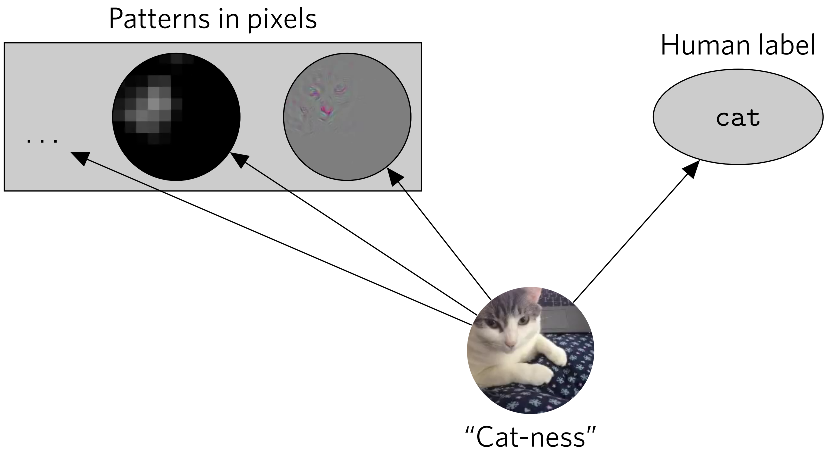
Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

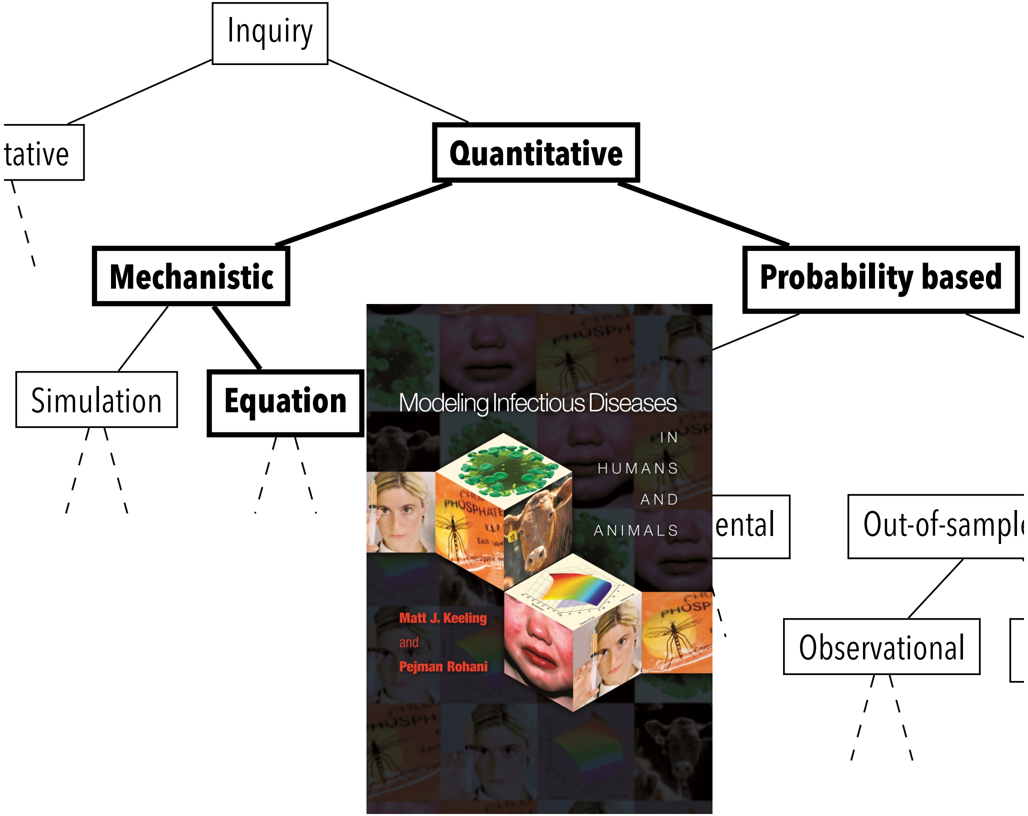
Constructs: Subjective, multifaceted



Example: Epic sepsis model

- Wong et al. (2021) found that a model to predict sepsis from the electronic health records company Epic worked far less well than claimed
  - AUC of .63, versus what Epic reported of .76 to .83
- One possible culprit: *different definitions*. Epic developed its model based on defining sepsis by the point where physicians intervened (what there was direct data for). Wong et al.'s evaluation was based on defining sepsis by meeting a certain number of CDC and ICD-10 criteria
- *Of course* the model as fitted wouldn't generalize! Maybe the same model, re-fitted on the "better" measure, would work; but also, *why* are there different definitions of sepsis?

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 (we cannot do anything with an *n* of 1!)

Importance of sampling frame

- Because ML uses the same fundamental mechanism as stats (reducing aggregates via central tendency), it has the same issue that *results will only generalize insofar as the sample is representative* (see also Meng, 2018)
  - Failures of Literary Digest poll of 1936 (Peverill, 1988) and "Dewey defeats Truman" in 1948 led to reforms in survey sampling
- The "patterns" we "recognize" are correlations, not necessarily universal regularity, so we can't ignore the sampling frame
- "Sampling on the dependent variable" is a classic problem: Cohen and Ruths (2013) have an amazing *mea culpa* where they note that they filtered Twitter users to only those who had a signal for political orientation. That was an unrealistic sampling frame

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

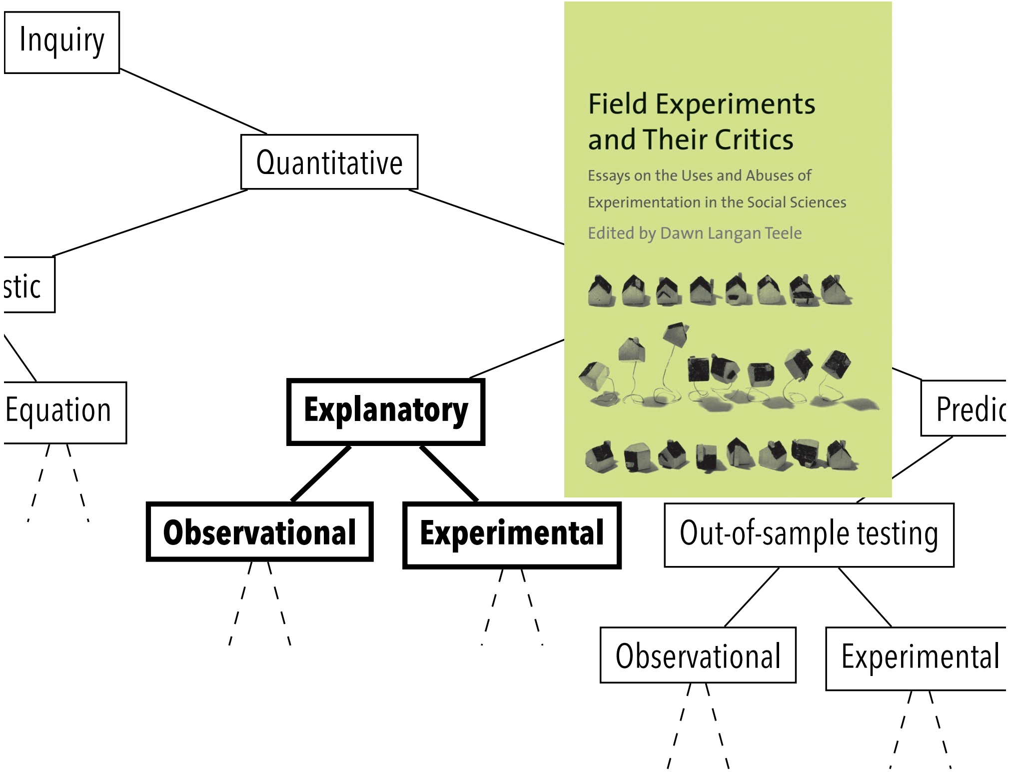
Summary and conclusion

References

Fixes: Study design (look at sampling frame and use measurement models)

- Sampling frame is typically taught in social sciences, not necessarily in machine learning
- Measurement models are the domain of psychometrics, and are almost completely unknown in ML (Jacobs & Wallach, 2019)
- These are a standard part of education that ML should make room for (will return to later under "culture")

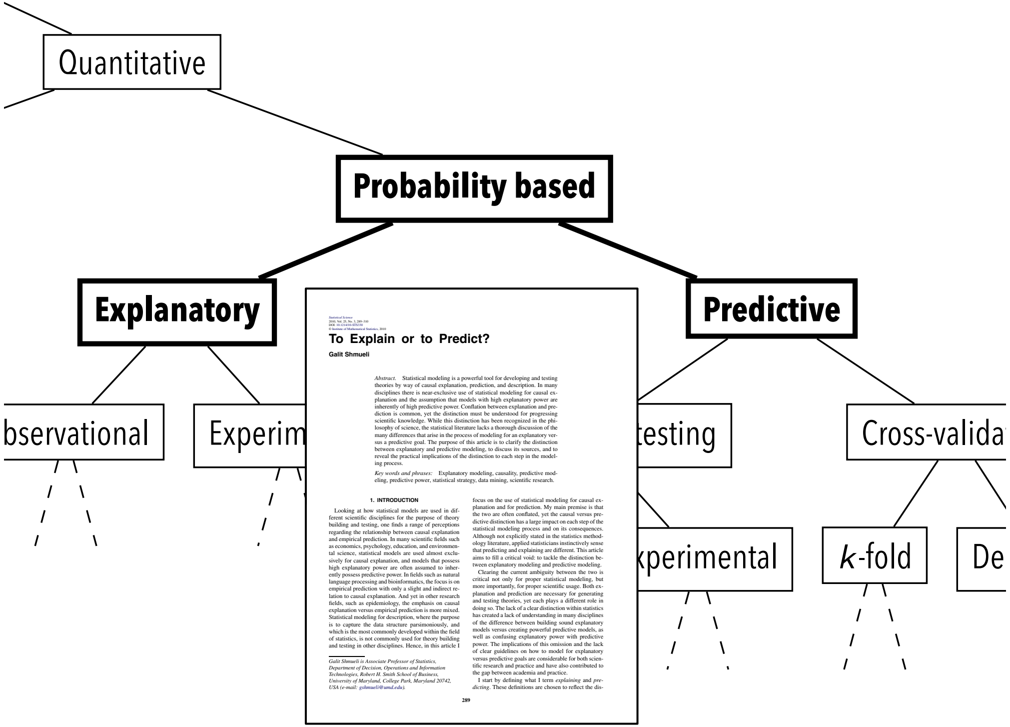
Causality is hard, maybe too hard



- 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)

Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

ML is "prediction" only



Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

1. INTRODUCTION

Looking at how statistical models are used in different scientific disciplines for the purpose of theory building and testing, one finds a range of perceptions regarding the relationship between causal explanation and empirical prediction. In many scientific fields such as economics, psychology, education, and environmental science, statistical models are used almost exclusively for causal explanation, and models that possess high explanatory power are often assumed to inherently possess predictive power. In fields such as natural language processing and bioinformatics, the focus is on empirical prediction with only a slight and indirect relation to causal explanation. And yet in other research fields, such as epidemiology, the emphasis on causal explanation versus empirical prediction is more mixed. Statistical modeling for description, where the purpose is to capture the data structure parsimoniously, and which is the most commonly developed within the field of statistics, is not commonly used for theory building and testing in other disciplines. Hence, in this article I focus on the use of statistical modeling for causal explanation and for prediction. My main premise is that the two are often conflated, yet the causal versus predictive distinction has a large impact on each step of the statistical modeling process and on its consequences. Although not explicitly stated in the statistics methodology literature, applied statisticians instinctively sense that predicting and explaining are different. This article aims to fill a critical void: to tackle the distinction between explanatory modeling and predictive modeling.

Clearing the current ambiguity between the two is critical not only for proper statistical modeling, but more importantly, for proper scientific usage. Both explanation and prediction are necessary for generating and testing theories, yet each plays a different role in doing so. The lack of a clear distinction within statistics has created a lack of understanding in many disciplines of the difference between building sound explanatory models versus creating powerful predictive models, as well as confusing explanatory power with predictive power. The implications of this omission and the lack of clear guidelines on how to model for explanatory versus predictive goals are considerable for both scientific research and practice and have also contributed to the gap between academia and practice.

I start by defining what I term *explaining* and *predicting*. These definitions are chosen to reflect the dis-

Galit Shmueli is Associate Professor of Statistics, Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, USA (e-mail: gshmueli@umd.edu).

- "Predictions" are defined as what minimizes loss *within a predetermined frame*
  - *Correlations* do this
- Non-causal correlations can sometimes predict well within a frame, but they frequently don't explain, and can fail outside
  - If that was the definition (Milton Friedman: "prediction in the presence of change"), correlations wouldn't work, but that is hard to formalize

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

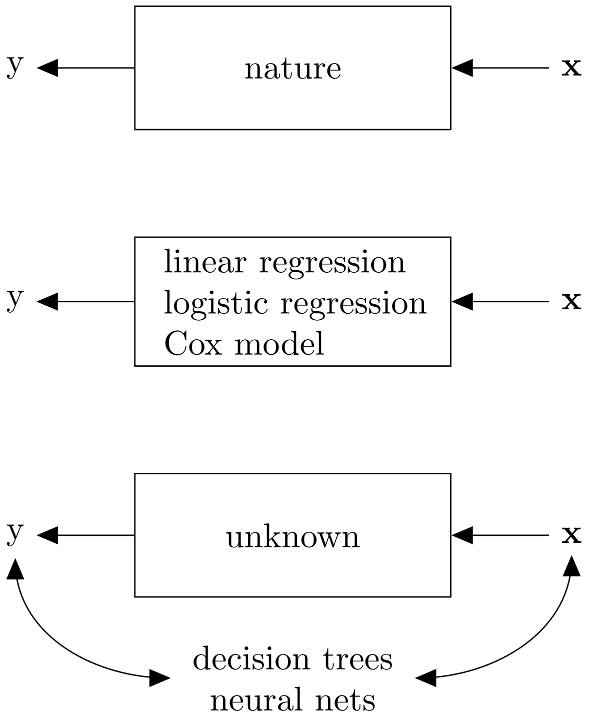
Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

A "realist" definition for machine learning



- (Realist definitions: what things are, rather than what they aspire to be)
- 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)
- Yes theory-agnostic modeling has its place, but there is a cost to abandoning many hard-won guardrails

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

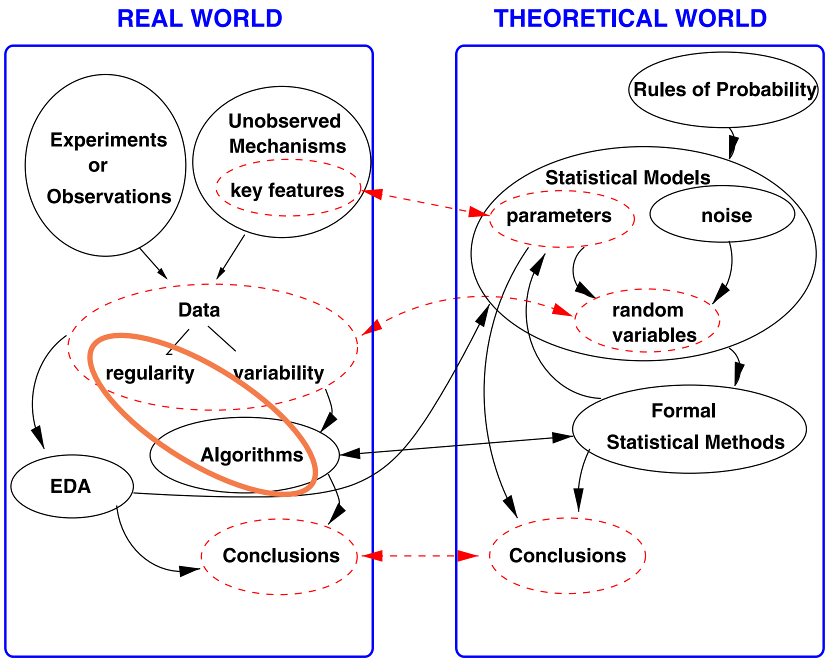
Uncertainty quantification

Reflexivity and positionality

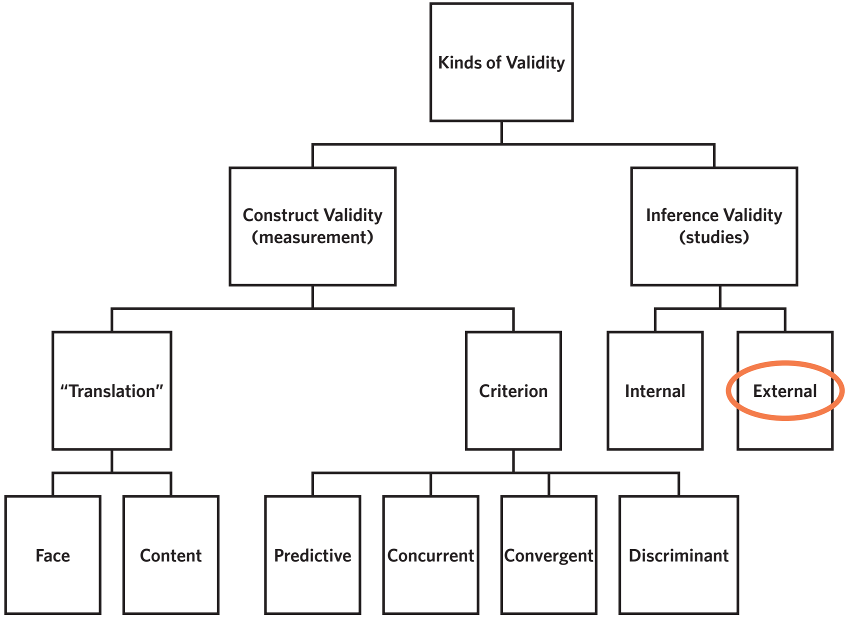
Summary and conclusion

References

ML: Only regularity and external validity

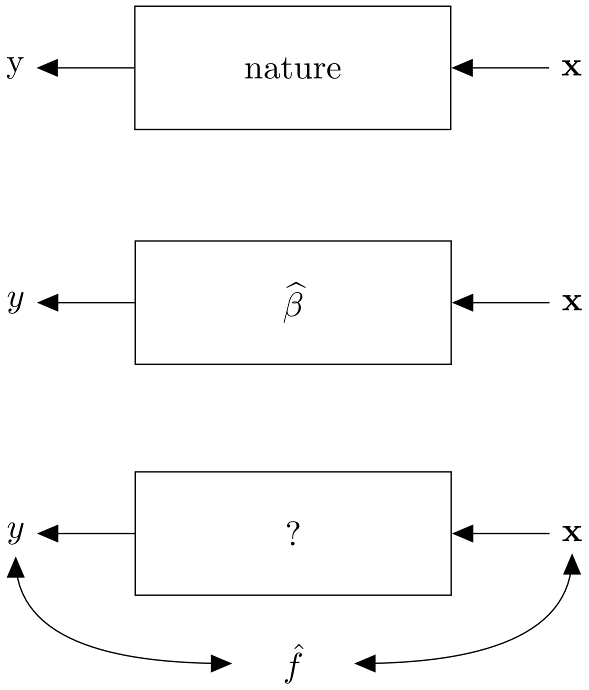


Kass, 2011



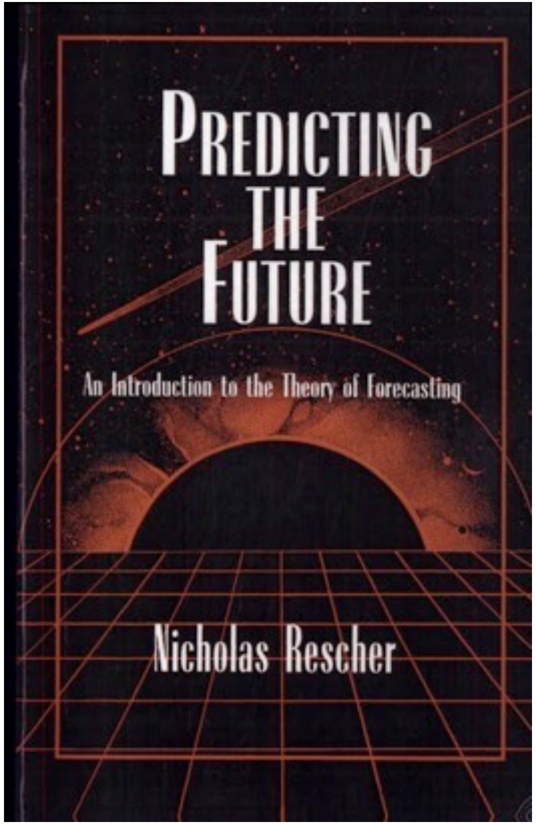
Adapted from Borgatti, 2012

Leads to two separate goals



- Non-causal ("spurious") correlations may fit robustly (e.g., latent common cause)
  - Breiman, 2001: "prediction problems"
  - Shmueli, 2010: "to predict"
  - Kleinberg et al., 2015: "umbrella problems"
  - Mullainathan & Spiess 2017: "**y-hat problems**"
- Carefully built models that capture causality (or "pure" associations) may fit poorly overall
  - Breiman: "information"
  - Shmueli: "to explain"
  - Kleinberg et al.: "rain dance problems"
  - Mullainathan & Spiess: "**beta-hat problems**"

Levels of prediction (Rescher, 1998)

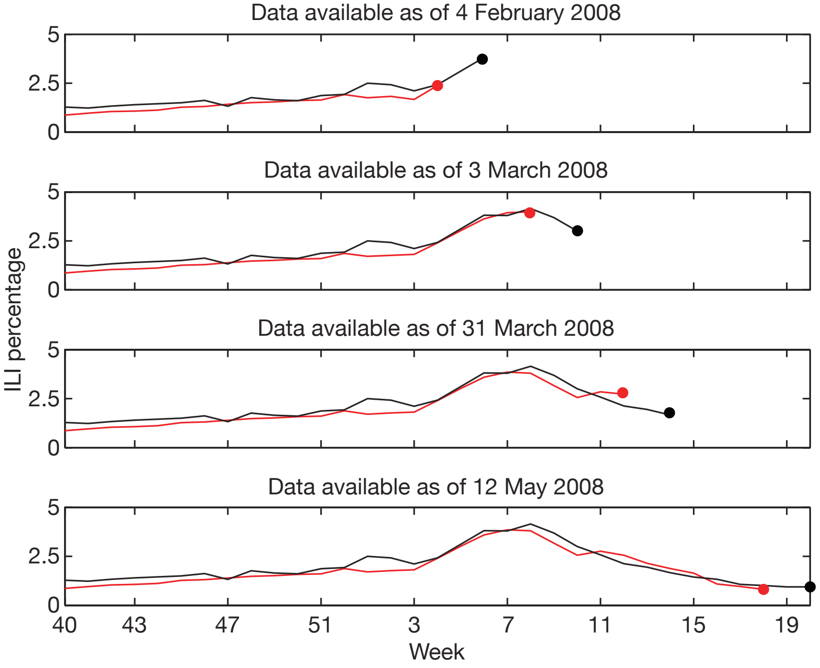


88 ■ PREDICTING THE FUTURE

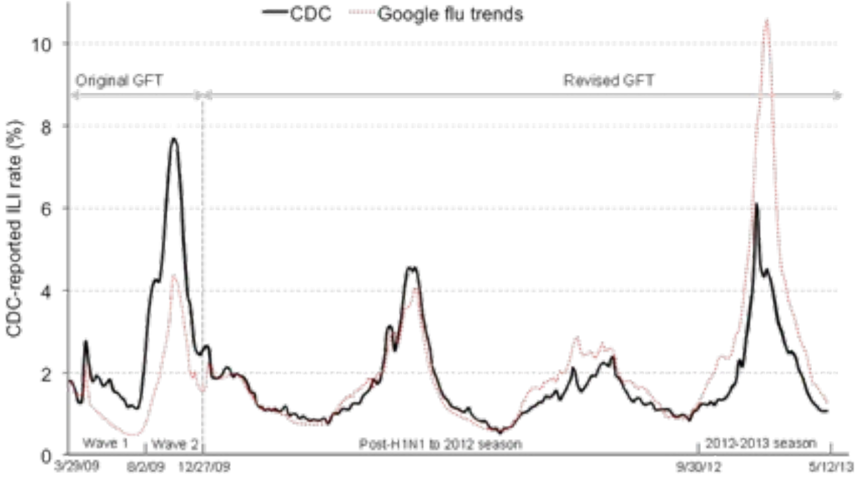
TABLE 6.1: A SURVEY OF PREDICTIVE APPROACHES

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

Google Flu Trends, or, "things do change" (Hoadley, 2001)

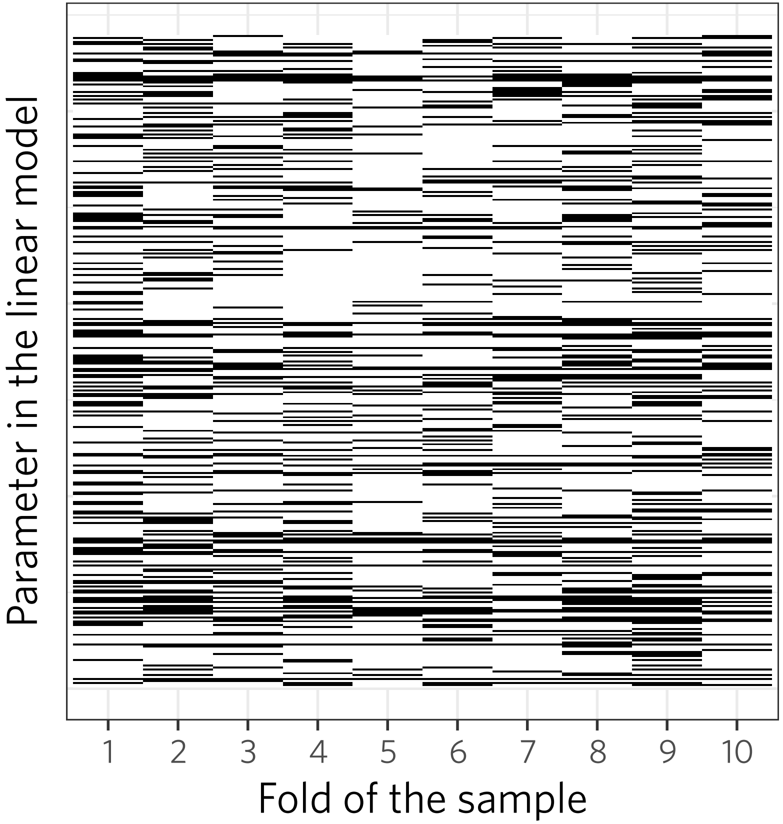


Ginsberg et al., 2012



Santillana et al., 2014

Correlations can't "predict in the presence of change" or of interventions



- Very different sets of correlations can "predict" (correlate) equally well (Mullainathan and Spiess 2017)
  - Breiman (2001) called this the "Rashomon effect" and saw it as a point in favor of prediction-only
- But different fits suggest very different outputs under covariate shift, and under interventions
- There is also heterogeneity in *amenability to intervention*; e.g., likely hospital readmissions are not the same as *preventable* likely hospital admissions (Marafino et al., 2020)

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

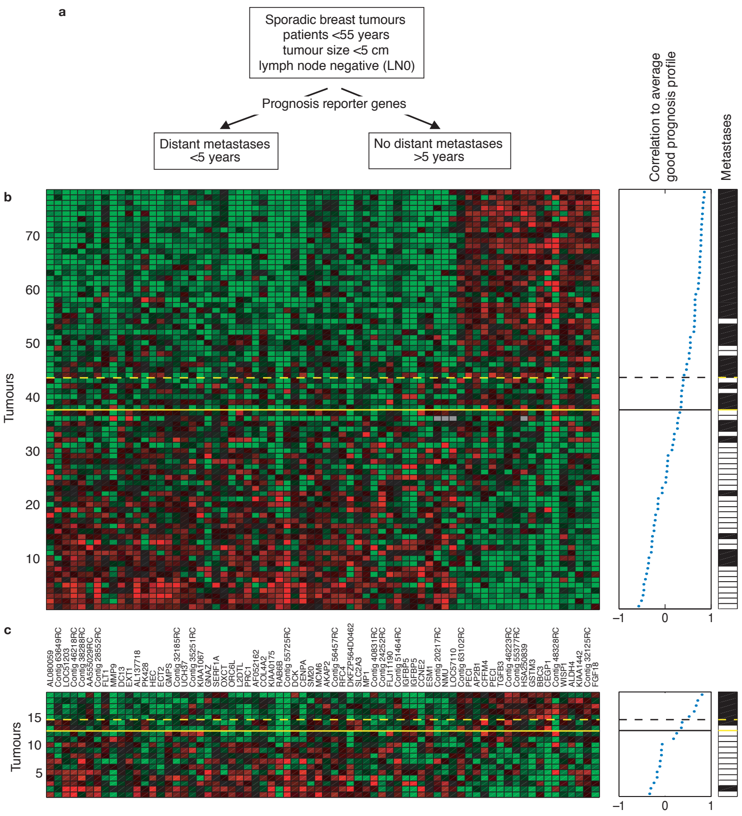
Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

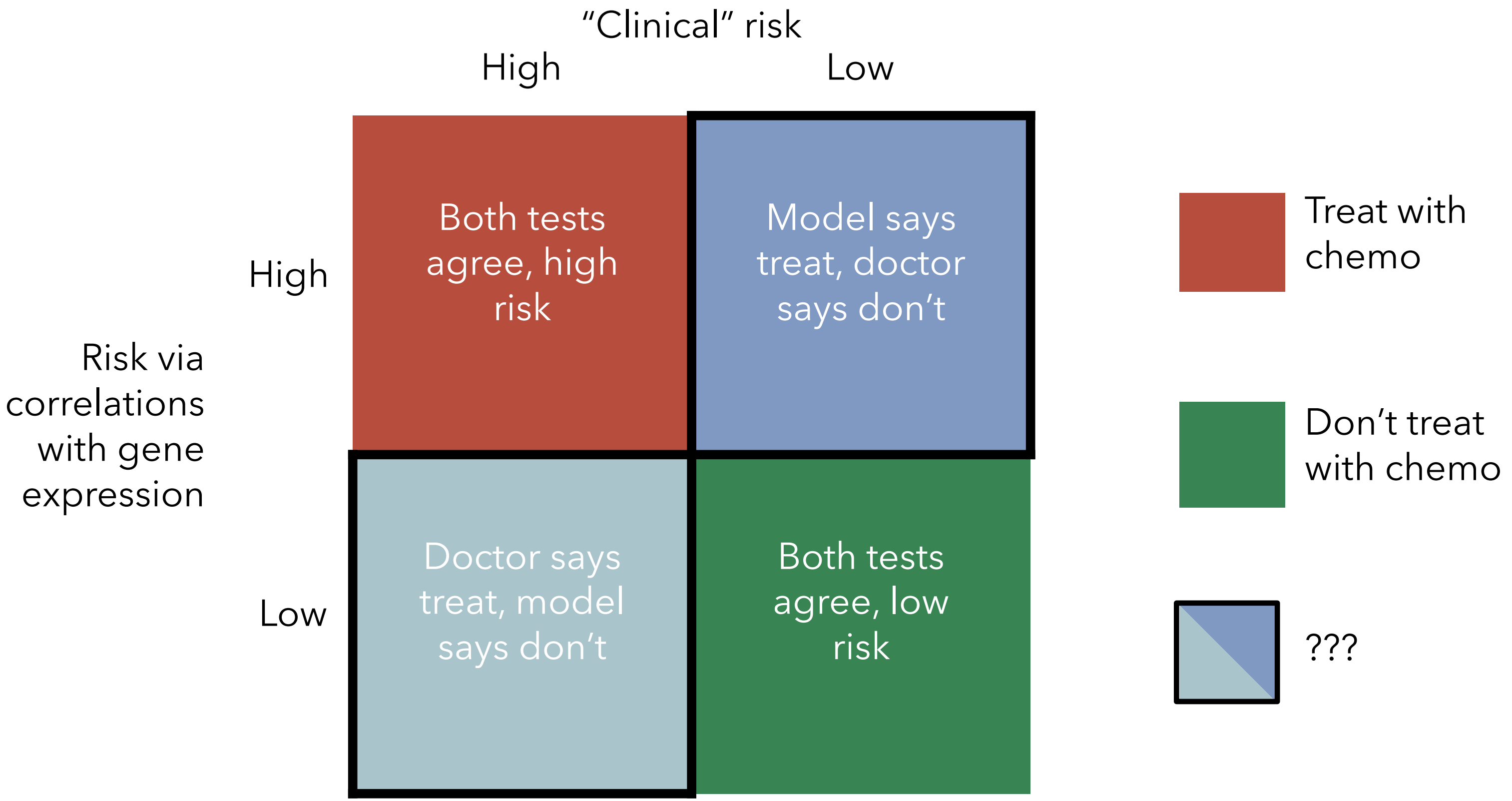
References

Positive example: Testing generalizability



- I really like the example of Cardoso et al. (2016). van't Veer et al. (2002) fit a model for genetic correlates of metastatic breast cancer. Of course it was optimal, *post-hoc*. But did it generalize?
  - (Probably could be re-done much better with more data and modern software: only trained on 98 breast tumors, done via a custom-implemented decision tree. But this was from 2002.)
- Cardoso et al. (2016) tested on 6,693 women in Europe

Testing generalizability

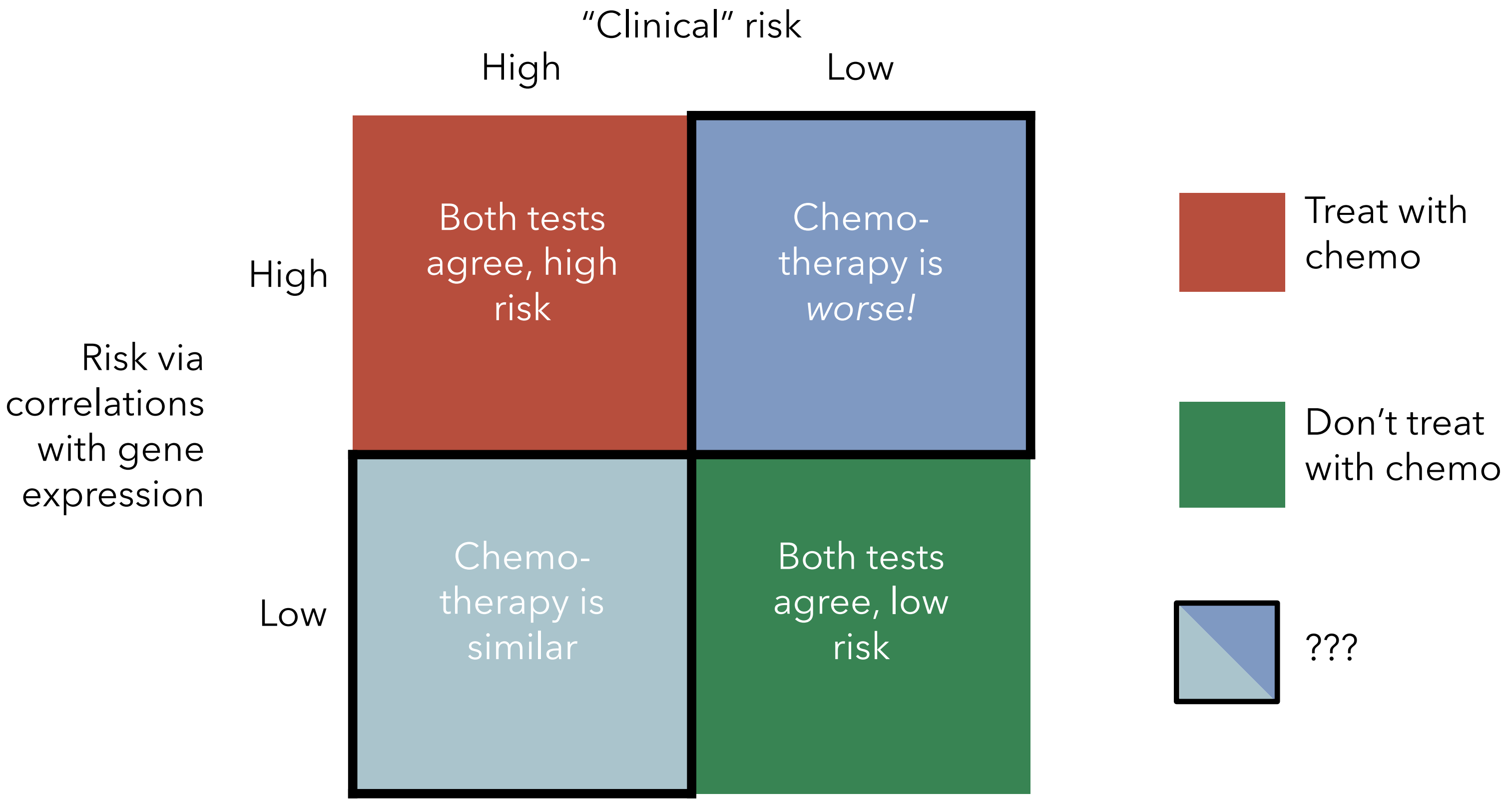


Cardoso et al., 2016, *NEJM*

MAYO CLINIC

Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

Testing generalizability



Cardoso et al., 2016, *NEJM*

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Testing generalizability

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

Don't treat with chemo

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*

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Pet peeve: language

- Communication: **stop saying “prediction” if it is really “correlation”**
  - **The use of ‘prediction’ leads to false, inflated expectations.** Instead of saying “prediction” for post-hoc demonstrations (Gayo-Avello, 2012), use “retrodiction”: it is awkward, but that’s what we need. For time series: nowcasting, back-testing (although better language is not enough: Bailey & de Prado, 2021)
  - Partial correlation (i.e., for “ceteris paribus” interpretations) can be described with “association”
- “Prediction” is overused as it is
  - Statements like ”predict the probability of risk”, or “calculate the probability of a likelihood” exist and are redundant if not nonsensical (akin to, “a probability of a probability [of a probability]”).
    - Probabilities and risks are always latent (and indeed, are hypothetical and metaphysical), so how can we “predict” them? We should say that we *estimate* probabilities and risk (say *estimated probabilities*, etc.), and not overload on synonyms for probability
  - Use “detection” or “classification” if labels are manifest but unknown. E.g., we don’t “predict” race; “detecting” and “predicting” cancer imply two very different tasks; etc.
- **Models, not algorithms** (unless you really do mean an optimization algorithm). Why? Specificity: logistic regression is a *model*, IRLS is an algorithm. Random forests are a *model*, CART is an algorithm. And: we already know “all models are wrong” (Box, 1979)

Fixes: Language, expectations, and claim-making

- If by "generalizability," we mean that a fitted model will apply to very different contexts, probably very few ML models will generalize (at least for the social world)
  - But if we mean that the ML *procedure*, allowing for different weights (and even different selected features) for a different context, then things are probably not as bad
  - Using Rescher's (1998) "level of prediction" can help be more precise
- Being more precise about language will help this, including setting expectations from ML being based on maximizing correlations [in a given sample] rather than achieving prophecy
- *Just because we can find a correlation doesn't mean we've advanced scientific understanding*: it hopefully can be used to make progress, but only if it successfully generalizes

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Uncertainty quantification and over-optimistic machine learning

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Model metrics as estimators

- If we make a commitment to a statistical view of the world (unobservable but inferable underlying regularity realized with haphazard variability), then the precision, recall, AUC, etc., are *estimators* of the underlying quantity of out-of-sample performance
  - Quantifying uncertainty provides a hedge on performance claims
- We can frame and study their properties statistically!
  - *Dependencies* cause test error to be biased (and, in a simple case, error has a generalized non-central chi-square distribution, which is heavily right-tailed, versus the symmetry of a binomial distribution)
  - Metrics other than accuracy (binomial) look like they have weird distributions. Somebody should look into this, and also design tests and power calculations
  - This view explains how it makes sense to use instrumental variables for estimating out-of-sample performance! (Kleinberg et al., 2018)

Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

Matrix bias-variance decomposition

$$
\begin{aligned}
\operatorname{err}(\hat{\mu}) &= \tfrac{1}{n}\mathbb{E}_f\|Y-\widehat{Y}\|_2^2 \\
&= \frac{1}{n}\left[\mathbb{E}_f\|Y\|_2^2+\mathbb{E}_f\|\widehat{Y}\|_2^2-2\mathbb{E}_f(Y^T\widehat{Y})\right] \\
&= \frac{1}{n}\left[\mathbb{E}_f\|Y\|_2^2+\mathbb{E}_f\|\widehat{Y}\|_2^2-2\operatorname{tr}\mathbb{E}_f(Y\widehat{Y}^T)\right] \\
&\quad+\frac{1}{n}\left[\mu^T\mu+\mathbb{E}_f(\widehat{Y})^T\mathbb{E}_f(\widehat{Y})+2\operatorname{tr}\mu\mathbb{E}_f(\widehat{Y})^T\right] \\
&\quad+\frac{1}{n}\left[-\mu^T\mu-\mathbb{E}_f(\widehat{Y})\mathbb{E}_f(\widehat{Y})^T-2\mu^T\mathbb{E}_f(\widehat{Y})\right] \\
&= \frac{1}{n}\Big[\underbrace{\operatorname{tr}\Sigma}_{\text{irreducible ("Bayes") error}}+\underbrace{\|\mu-\mathbb{E}(\widehat{Y})\|_2^2}_{\text{bias squared}}+\underbrace{\operatorname{tr}\operatorname{Var}_f(\widehat{Y})}_{\text{variance of the estimator}}-\underbrace{2\operatorname{tr}\operatorname{Cov}_f(Y,\widehat{Y})}_{\text{"optimism"}}\Big]
\end{aligned}
$$

Classic argument for CV

Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

Training:

$$\begin{aligned}\mathrm{err}(\hat{\mu}) &= \frac{1}{n}\mathbb{E}_f\|Y - \widehat{Y}\|_2^2 \\ &= \frac{1}{n}\left[\mathrm{tr}\,\Sigma + \|\mu - \mathbb{E}(\widehat{Y})\|_2^2 + \mathrm{tr}\,\mathrm{Var}_f(\widehat{Y}) - 2\,\mathrm{tr}\,\mathrm{Cov}_f(Y, \widehat{Y})\right]\end{aligned}$$

Testing:

$$\begin{aligned}\mathrm{Err}(\hat{\mu}) &= \frac{1}{n}\mathbb{E}_f\|Y^* - \widehat{Y}\|_2^2 \\ &= \frac{1}{n}\left[\mathrm{tr}\,\Sigma + \|\mu - \mathbb{E}(\widehat{Y})\|_2^2 + \mathrm{tr}\,\mathrm{Var}_f(\widehat{Y}) - \cancel{2\,\mathrm{tr}\,\mathrm{Cov}_f(Y^*, \widehat{Y})}\right]\end{aligned}$$

The difference is the *optimism* (Efron, 2004; Rosset & Tibshirani, 2020):

$$\mathrm{Opt}(\hat{\mu}) = \mathrm{Err}(\hat{\mu}) - \mathrm{err}(\hat{\mu}) = \frac{2}{n}\,\mathrm{tr}\,\mathrm{Cov}_f(Y, \widehat{Y})$$

Apply this to non-iid data

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

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

- Then, optimism in the training set is:

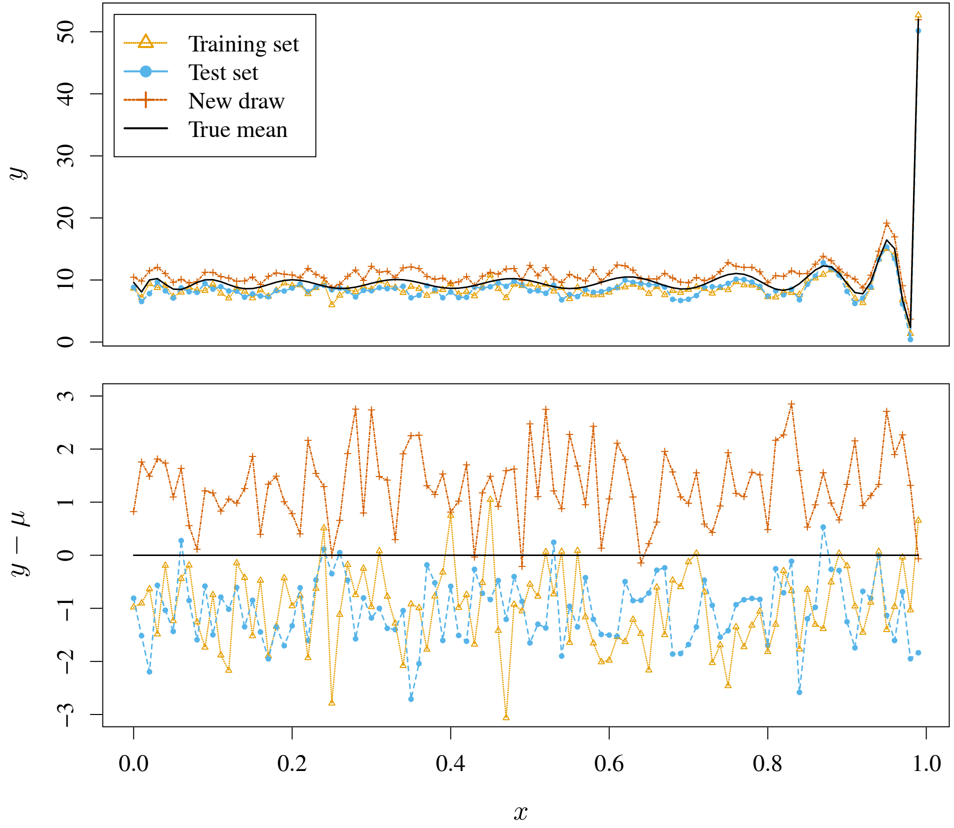
$$\frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_1, \widehat{Y}_1) = \frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_1, \mathbf{H} Y_1) = \frac{2}{n} \operatorname{tr} \mathbf{H} \operatorname{Var}_f(Y_1) = \frac{2}{n} \operatorname{tr} \mathbf{H\Sigma}$$

- But test set also has nonzero optimism!

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

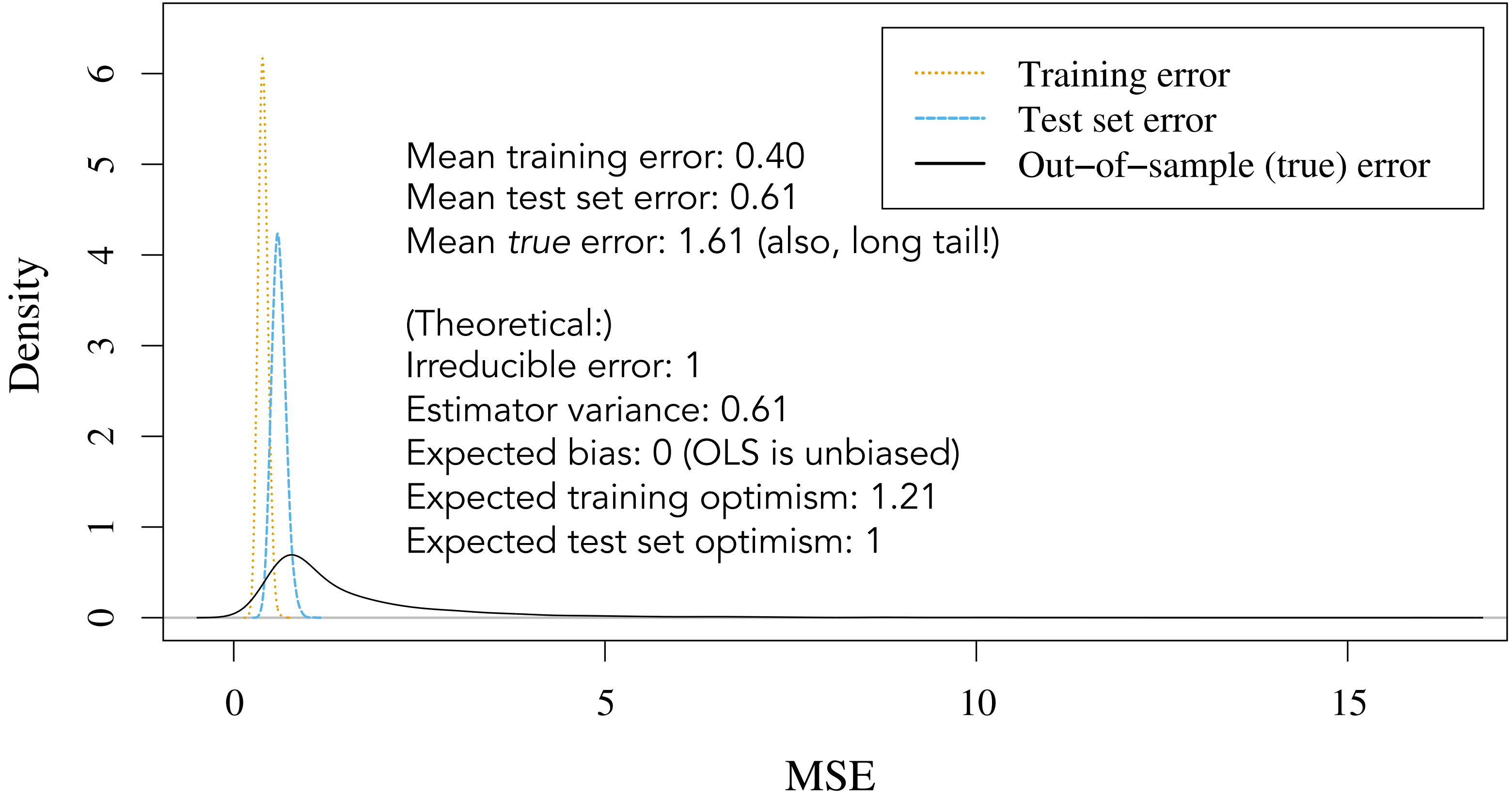
One draw as an example

Correlation between observations can pull training and test observations close to one another, but potentially far from an independent draw



Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

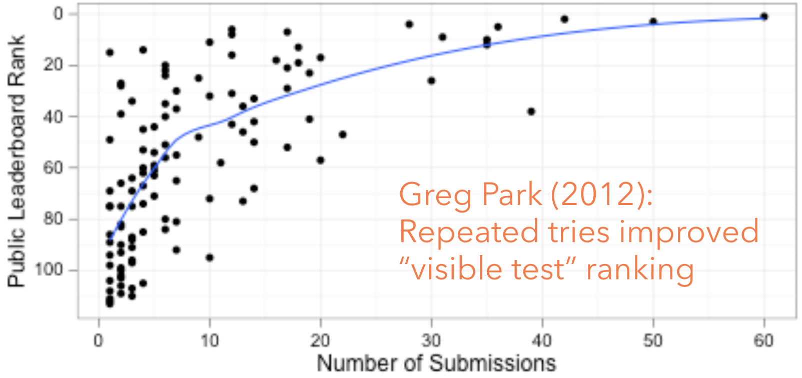
Simulated MSE

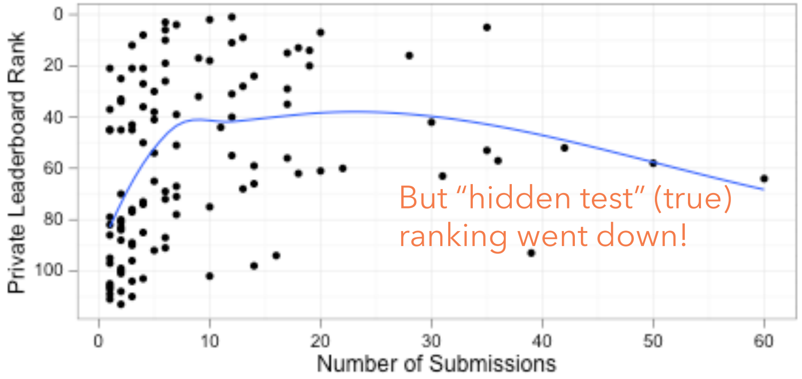


MAYO CLINIC

Quick examples

- "Twitter mood predicts the stock market" (Bollen et al., 2011) trains on future values, tests on past values: that is "time-traveling"! ("No limits to garbatrage," *Buy the Hype* blog, August 29, 2013; Lachanski & Pav, 2017)
- A colleague of mine trained a model to recognize birds on his windowsill in webcam images, splitting frames randomly…
- Park (2012) has a great example of overfitting to the test set in Kaggle. Having a "private leaderboard" helps catch overfitting in Kaggle
  - I agree with Wagstaff (2012) that in research, it's probably not worth having a test set we only use once (do we give up if performance is bad?). But we *should* temper our claims, and do out-of-sample testing





Lessons: Split by dependencies

- ML needs to contend with dependencies, because the iid assumption matters for estimates of model performance
  - Even statistical relational learning doesn't discuss
- Maybe we can't make a better *model*, but dependencies are a form of leakage between training and test sets
  - We can use the framework of "optimism" to understand and quantify this (**meta-meta-prediction** is useful; Rescher, 1998)
  - Test set re-use (Dwork et al., 2014) falls within this as well
  - Ideally, no dependencies between training and test sets
  - Unfortunately, the mean function and covariance function are jointly unidentifiable nonparametrically (Opsomer et al., 2001), so we will have to rely on theory and limited explorations (e.g., ACF, PACF)

How are metrics distributed? (Preliminary explorations)

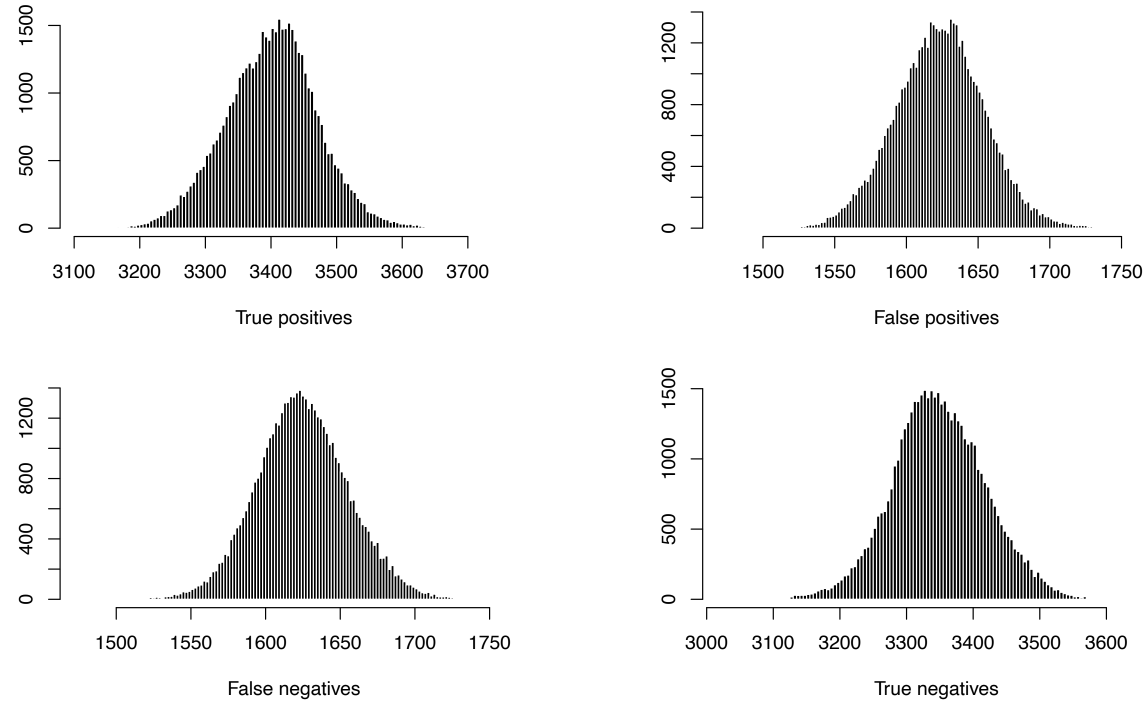
- Under this specification and DGP, the test error has a "generalized non-central chi-squared" distribution
- But even in the iid case, we know frighteningly little about distributions (in that I found no work other than around accuracy, which is binomial and gives McNemar's test) and the variability they might suggest
  - We should consider both asymptotics and convergence
- A quick simulation of a logistic fit of $X_i \sim \mathcal{N}(0, 1)$ and $Y_i \sim \text{Bin}(\text{logistic}(x_i))$ at $n = 10{,}000$ (large sample size) gives reasons for worry

Distributions of counts? $n = 10^4$ ($n_{\text{sim}} = 50{,}000$). Looks okay

Predicted positive | Predicted negative

Actual positive

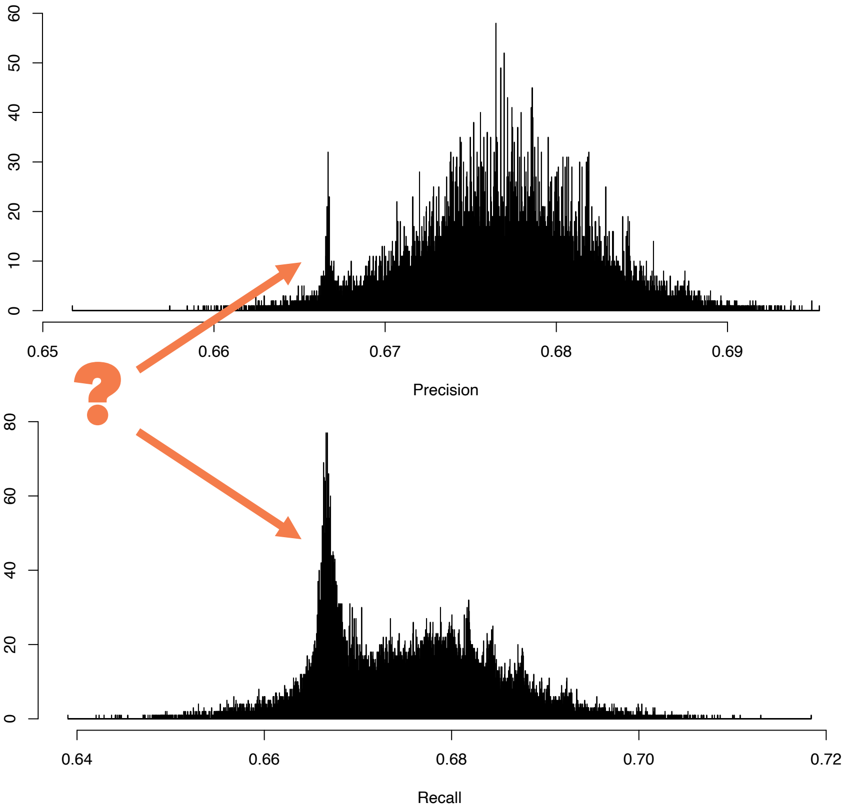
Actual negative

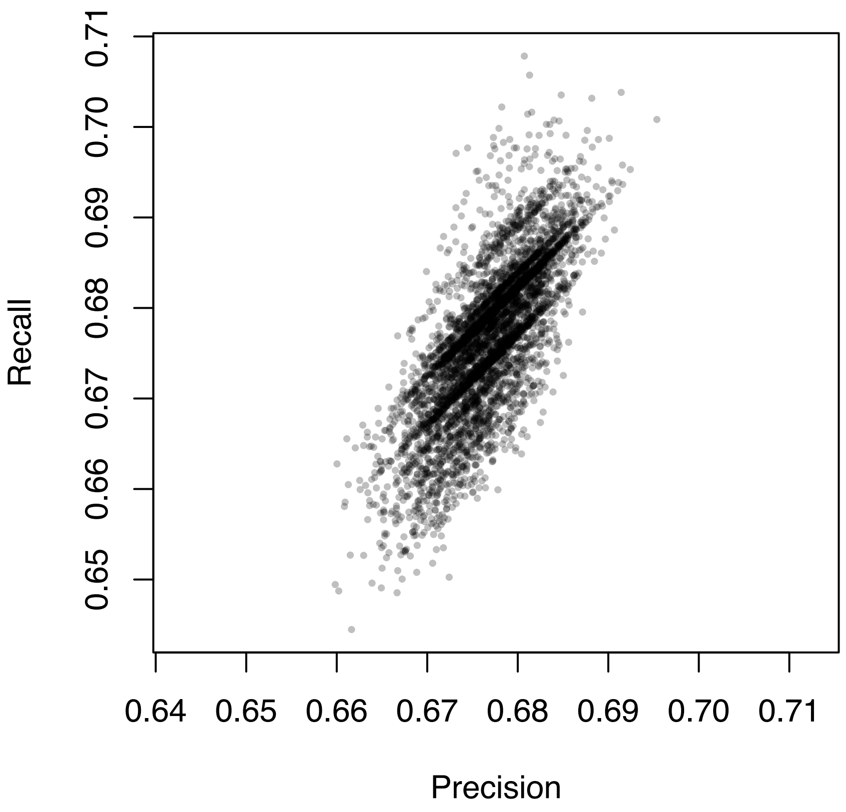


MAYO CLINIC

Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

Distributions of precision/recall? $n = 10^4$ (n_{sim} = 50,000). Looks weird…





Distributions of AUC/F_1? $n = 10^4$ (n_{sim} = 50,000). Also weird

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

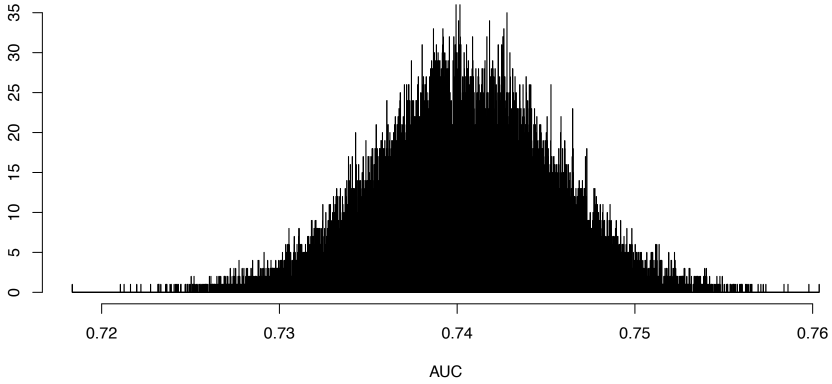
Uncertainty quantification

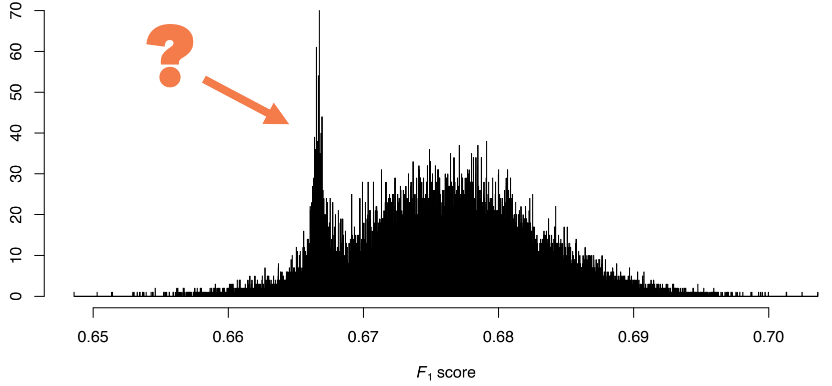
Reflexivity and positionality

Summary and conclusion

References

- 95% empirical confidence (tolerance) interval for AUC is [.731, .734], probably okay. (For n = 101, it is [.64, .83])
  - Other metrics? Small sample size? **Power!!**
- Distribution of estimators is stat theory 101!
  - I only found scattered, preliminary work (Lieli & Hsu, 2017; Delmer et al., 2017; Zhang et al., 2012)
  - Finite-sample performance may be weird too, and needs to be looked at
- Conclusion: even for large sample size, a simple DGP, and a "true" model, the distribution of common metrics is not so simple; we should certainly try to make inferences rigorously





MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

What do do? Quick notes

- **Do not use *k*-fold cross validation for assessing model performance!**
  - Wager (2020) has a great exploration that shows that CV has very different properties for model *selection*, versus model *evaluation*. *k*-fold CV consistently *selects* the best model, but is asymptotically uninformative about out-of-sample performance
- **For model *evaluation*, use a totally held-out test set** (contra Raschka, 2020)
- To get standard errors/confidence intervals, for now, we can always bootstrap on the test set (some of my current work)

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Fixes: We should study asymptotic distributions of metrics, and use them!

- Can somebody please find the distributions of ML model success metrics? (I started to try, via joint distribution of TP, FP, FN, TN as a multidimensional [3+1 dimensions] binomial, and then taking ratios of marginals, but it's a lot of algebra)
- With distributions, we could find asymptotic confidence intervals, and conduct significance testing of model results
  - Yes, *p*-values and hypothesis testing have done enormous damage, **but ignoring variance might be worse**
  - Also, start doing **power calculations** in ML
- Maybe, when studying asymptotic distributions, we'll find sufficient statistics for model success (like the parameters of a multivariate binomial) and good estimators thereof
  - We usually avoid ratio statistics, because they can have a Cauchy distribution

Reflexivity and positionality

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Narrow technical training

- Phil Agre (1997):
  - "My college did not require me to take many humanities courses, or learn to write in a professional register, and so I arrived in graduate school at MIT with little genuine knowledge beyond math and computers. This realization hit me with great force halfway through my first year of graduate school...
  - "I was unable to turn to other, nontechnical fields for inspiration… The problem was not exactly that I could not understand the vocabulary, but that I insisted on trying to read everything as a narration of the workings of a mechanism."
- Study design and measurement still partially fall under "technical" knowledge; the problem is far more profound

“Paradigms of inquiry”: Unknown in ML (even stats), but basic in social science

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Issue	**Positivism**	**Post-positivism**	**Critical theory et al.**	**Constructivism**	**Participatory**
Ontology	Naïve realism: Reality independent of and prior to human conception of it, apprehensible	Critical realism: Reality independent of and prior to human conception, but imperfectly and approx. apprehensible	Disenchantment theory: reality is secret/hidden, shaped by power structures and solidified over time	Relativism: multiple realities, constructed in history through social processes	Participative: multiple realities, co-constructed through interactions between specific people and environments
Epistemology	Reality knowable. Findings are singular, neutral, perspective-independent, atemporal, universally true	Findings provisionally true; multiple descriptions can be valid but are probably equivalent; findings can be affected/distorted by social + cultural factors	How we come to know something, or who knows it, matters for how meaningful it is	Relativistic: no neutral perspective to adjudicate competing claims	We come to know things, create new understandings, & transform world by involving other people in process of inquiry
Methodology	Hypotheses can be verified as true. Quant methods, math.	Falsification of hypotheses; primacy of quant, but some qual and mixed methods	Dialogic (conversation + debate) or dialectical (thesis$_1$ → antithesis$_1$ → synthesis$_2$:= thesis$_2$…)	Dialetical, or exegetical (reading between the lines)	Collaborative, action-focused; flattening hierarchies; engaging in self- and collective reflection, action
Axiology	Quant knowledge-holders have access to truth, and responsibility from it	Quant knowledge valuable but can be distorted; qual can help find and correct	Marginalizat*ion* provides unique insights, knowledge of marginaliz*ed* valuable	Understanding construction is valuable; value relative to given perspective	Reflexivity, co-created knowledge, and non- western ways of knowing are valuable and combat erasure and dehumanization

Malik & Malik (2021), via Guba and Lincoln (2005)

Empowerment, not charity

- Paulo Freire went in the 1940s to work with "illiterate farmers" in Brazil. He originally subscribed to a "banking model" of education, where he was a bank, holding knowledge, from which others would make "withdrawals"
  - He discovered himself learning from those with whom he worked: their experience of marginalization taught him things about how society worked that he, because of his privilege, was ignorant
- Framings of *charity* are unidirectional, removing autonomy from those who are supposedly helped; framings of *service* see privilege as creating obligations, but it still may permit the privileged to set the nature and scope of that service
- Better is *empowerment*, and best is co-creation, recognizing how everyone has something to contribute. This is hard with quantitative knowledge; but if we do not have humility, and if we do not trust people to themselves decide what is best for themselves, we heavily risk creating harm

MAYO CLINIC

Introduction
Motivation: Empirical failure
Stats/ML and trade-offs in methodology
Uncertainty quantification
Reflexivity and positionality
Summary and conclusion
References

"Ways of understanding a person": The quant view is strange and unnatural!

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

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

Why this matters: it's why we *expect* generalization

- We expect that models are picking up on signal, not noise
  - Statistics makes the assumption that we can treat the world as made up of entities that are distinct but are realizations of an underlying process. Machine learning shares this assumption, even if it is not explicit about it (e.g., theory about convergence to the "oracle predictor" rather than about convergence to a "true" parameter)
- If we define the "signal" as what is invariant, then failures of generalizability means we've failed to find the underlying regularity
- But is there really aggregate regularity? Or only *narrative*, if any?
  - E.g., Twitter and elections (Gayo-Avello, 2012)
  - Note: one explanation for stats working is that it *imposes* regularity

“What are we even doing?”

- “If science isn’t ‘true’, then what are we even doing? We might as well be doing English literature, or art criticism!”
  - Intellectual supremacy is probably a bad reason for doing science
- At least for the social world, I am skeptical of attempts to find underlying regularity in the [social] world as cases; both because only trivial things can have universal aggregate regularity, and because attempts to find social regularity can end up imposing it (“performativity”; Healy, 2019)
  - But neither can I imagine our civilization without the use of summary statistics for management, planning, and allocation…

"Sociotechnical" approaches

- The term "sociotechnical" is both helpful, and overused: it refers to considering the parts of a system that make a system succeed or fail apart from its technical content
  - E.g., usability; adoption; compliance/usage; "off-label" use; alignment; aggregate effects
- Including a qualitative component can help achieve and identify success in ways more meaningful than specific metrics can capture (e.g., Elish & Watkins, 2020; Farrell et al., 2017)

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Summary and conclusion

Methodological reform

- Sometimes, we should expect non-generalizability, when we reason about or have qualitative evidence for problems with the basic measurement and quantification (e.g., perhaps some "essential information" is not present anywhere in the loss landscape of any objective function we could define based on the data we have)
- We can avoid many basic errors by considering sampling frame; the appropriateness of non-causal modeling; and doing uncertainty quantification
  - (Forthcoming work: Kapoor et al., "Reporting Standards for ML-based Science" (REFORMS), https://reporting-standards.cs.princeton.edu/)

Encountering the social world

- "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning." (Wigner, 1960)
  - We still don't know why quantification/mathematics works for the natural world: so we shouldn't be surprised if/when it turns out to not extend to the social world. And certainly not if restricted to a relatively narrow (but supremely effective) subset of modeling around probability-based models
  - (Counterpoint: the adoption of quantification in the natural sciences actually came after, and because, of its effectiveness in finance and bureaucracy; Porter, 1995)
- Reflecting on our own relationship to goals, institutions, and affected communities can help us understand the relationship of models to the social world, and the impacts they may have (or not have)

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Methodological issues

- ML models will only generalize insofar as the **data are representative**
- **Selection on the dependent variable** is not something we can do when applying models
- If the **underlying measurements** are not consistent, the model can also fail to generalize
- Point estimates of **model metrics** don't give possible **variability** even with the same population
- **Dependencies** cause a form of **leakage**
- Unless models give unbiased estimates of partial correlation, **causal shifts** will make them invalid

Suggested fixes

- **Gather representative data** and/or make more limited claims
- **Include weak signal observations**, rather than filter them out
- Use a **measurement model** (for the response), or at least consider validity and reliability
- Get **confidence intervals** around all measures of model success, and **study asymptotics**
- **Split data by dependencies** (temporal block CV, leave-one-subject-out CV, network CV, etc.)
- **Change language** to temper expectations, and **sometimes, pursue causality**

MAYO CLINIC

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

References (1/2)

Agre, Philip. E. 1997. Toward a critical technical practice: Lessons learned in trying to reform AI. In *Bridging the great divide: Social science, technical systems, and cooperative work*, edited by Geoffrey C. Bowker, Susan Leigh Star, William Turner, and Les Gasser, 131–157. Mahwah, New Jersey: Erlbaum.

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://doi.org/10.1177/0049124110378098

Bailey, David H. and Mario Lopéz de Prado. 2021. How "backtest overfitting" in finance leads to false discoveries. *Significance* 18: 22-25. https://doi.org/10.1111/1740-9713.01588

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://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://doi.org/10.1056/NEJMoa1602253

Cohen, Raviv and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (ICWSM-13), 91–99. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128

Demler, Olga V., Michael J. Pencina, Nancy R. Cook, and Ralph B D'Agostino, Sr. 2017. Asymptotic distribution of ΔAUC, NRIs, and IDI based on theory of U-statistics. Statistics in Medicine 36 (21): 3334–3360. https://doi.org/10.1002/sim.7333

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://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://doi.org/10.1198/016214504000000692

Elish, Madeleine Clare and Elizabeth Anne Watkins. 2020. *Repairing innovation: A study of integrating AI in clinical care*. Data & Society, September 30, 2020. https://datasociety.net/library/repairing-innovation/

Farrell, Tracie, Alexander Mikroyannidis, and Harith Alani. 2017. "We're seeking relevance": Qualitative perspectives on the impact of learning analytics on teaching and learning. In *EC-TEL 2017: Data driven approaches in digital education*, edited by Élise Lavoué, Hendrik Drachsler, Katrien Verbert, Julien Broisin, and Mar Pérez-Sanagustín, 397-402. https://doi.org/10.1007/978-3-319-66610-5_33

Freedman, David A. (author), and David Collier, Jasjeet S. Sekhon, and Philip B. Stark (editors). 2009. *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815874

Friedman, Milton. 1953. Essays in positive economics. University of Chicago Press.

Gayo-Avello, Daniel. 2012. No, you cannot predict elections with Twitter. *IEEE Internet Computing* 16 (6): 91–94. https://doi.org/10.1109/MIC.2012.137

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://doi.org/10.1038/nature07634

Guba, Egon G. and Yvonna S. Lincoln. 2005. Paradigmatic controversies, contradictions, and emerging confluences. In *The SAGE Handbook of Qualitative Research*, edited by Norman K. Denzin and Yvonna S. Lincoln, 191–215. London: SAGE, 2005.

Healy, Kieran. 2015. The performativity of networks. *European Journal of Sociology* 56 (2): 175–205. https://doi.org/10.1017/S0003975615000107

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

Humphreys, Paul and David Freedman. 1996. The grand leap. *The British Journal for the Philosophy of Science* 47 (1): 113–123. https://doi.org/10.1093/bjps/47.1.113

Jacobs, Abigail Z., and Hanna Wallach. 2021. Measurement and fairness. In *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://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://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, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133 (1): 237–293, https://doi.org/10.1093/qje/qjx032

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

References (2/2)

Introduction

Motivation: Empirical failure

Stats/ML and trade-offs in methodology

Uncertainty quantification

Reflexivity and positionality

Summary and conclusion

References

Lachanski, Michael and Steven Pav. 2017. Shy of the character limit: "Twitter mood predicts the stock market" revisited. *Econ Journal Watch* 14 (3): 302–345. https://econjwatch.org/articles/shy-of-the-character-limit-twitter-mood-predicts-the-stock-market-revisited

Lazer, David. 2014. Mistaken analysis: It's too easy to be led astray by the lure of big data. *MIT Technology Review*, April 23. https://www.technologyreview.com/2014/04/23/173144/mistaken-analysis/

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://doi.org/10.1126/science.1248506

Lieli, Robert P. And Yu-Chin Hsu. The null distribution of the empirical AUC for classifiers with estimated parameters: A special case. IEAS Working Paper: academic research 16-A007, Institute of Economics, Academia Sinica, Taipei, Taiwan. http://www.personal.ceu.hu/staff/Robert_Lieli/AUC_submitted.pdf

Marafino, Ben J., Alejandro Schuler, Vincent X. Liu, Gabriel J. Escobar, and Mike Baiocchi. 2020. Predicting preventable hospital readmissions with causal machine learning. *Health Services Research* 55 (6): 993-1002. https://doi.org/10.1111/1475-6773.13586

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

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

Meng, Xiao-Li. 2018. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics* 12 (2): 685–726. https://dx.doi.org/10.1214/18-AOAS1161SF

Morozov, Evgeny,. 2013. *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

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

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

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

Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew D. Selbst. 2022. The fallacy of AI functionality. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 959–972. https://doi.org/10.1145/3531146.3533158

Raschka, Sebastian. 2020. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808. https://arxiv.org/abs/1811.12808

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

Rosset, Saharon and Ryan J. Tibshirani. 2020. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association* 115 (529): 138–151. https://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

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

Selbst, Andrew A., danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (FAT* '19), 59–68. https://doi.org/10.1145/3287560.3287598

Squire, Peverill. 1988. Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly* 52 (1): 125. https://doi.org/10.1086/269085

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. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415 (6871): 530–536. https://doi.org/10.1038/415530a

Wager, Stefan. 2020. Cross-validation, risk estimation, and model selection: Comment on a Paper by Rosset and Tibshirani. Journal of the American Statistical Association 115 (529): 157–160. https://doi.org/10.1080/01621459.2020.1727235

Wagstaff, Kiri L. 2012. Machine learning that matters. In Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML'12, 1851–1856. https://icml.cc/2012/papers/298.pdf

Wigner, Eugene. 1960. The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics* 13 (1).

Wong, Andrew, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine* 181 (8): 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

Zhang, Peng and Wanhua Su. 2012. Statistical inference on recall, precision and average precision under random selection. In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD 2012), 1348–1352. https://doi.org/10.1109/FSKD.2012.6234049

Appendix: Simulation code

```r
library(ModelMetrics)

# Rename for convenience
logistic <- function(x) plogis(x)
logit <- function(p) qlogis(p)

set.seed(20220728)
nsim <- 50000
results <- data.frame(accuracy = rep(NA, nsim),
                      ppv = rep(NA, nsim),
                      tp = rep(NA, nsim),
                      tn = rep(NA, nsim),
                      fp = rep(NA, nsim),
                      fn = rep(NA, nsim),
                      tpr = rep(NA, nsim),
                      tnr = rep(NA, nsim),
                      auc = rep(NA, nsim),
                      f1score = rep(NA, nsim))

# Either run with 97 or 101 (small sample size:
# these are prime number close to 100, so
# that accuracy and other fractions divided by
# a prime denominator), or 10k (large sample #
size)

# n <- 97
# n <- 101
n <- 10000
```

```r
# Draw X once ("fixed X" setting), then draw a new Y
# each simulation run, y ~ bernoulli(logistic(x))
x <- rnorm(n = n, mean = 0, sd = 1)


for (i in 1:nsim) {
  y <- rbinom(n = n, size = 1, prob = logistic(x))
  glm1 <- glm(y ~ x, family = "binomial")
  results$accuracy[i] <- mean(y==(predict.glm(glm1, type = "response") > .5)
  results$ppv[i] <- ppv(y, predict.glm(glm1, type = "response")) # Precision
  results$tp[i] <- sum(y==1 & (predict.glm(glm1, type = "response") >= .5))
  results$tn[i] <- sum(y==0 & (predict.glm(glm1, type = "response") < .5))
  results$fp[i] <- sum(y==0 & (predict.glm(glm1, type = "response") >= .5))
  results$fn[i] <- sum(y==1 & (predict.glm(glm1, type = "response") < .5))
  results$tpr[i] <- tpr(y, predict.glm(glm1, type = "response")) # Recall
  results$tnr[i] <- tnr(y, predict.glm(glm1, type = "response")) # Specificity
  results$auc[i] <- auc(y, predict.glm(glm1, type = "response"))
  results$f1score[i] <- f1Score(y, predict.glm(glm1, type = "response"))
  if (i%%1000==0) {print(i)}
}
```