BERKMAN
KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Interpretability is a Red Herring: Grappling with "Prediction Policy Problems"

Momin M. Malik, PhD <momin_malik@cyber.harvard.edu>
Data Science Postdoctoral Fellow
Berkman Klein Center for Internet & Society at Harvard University

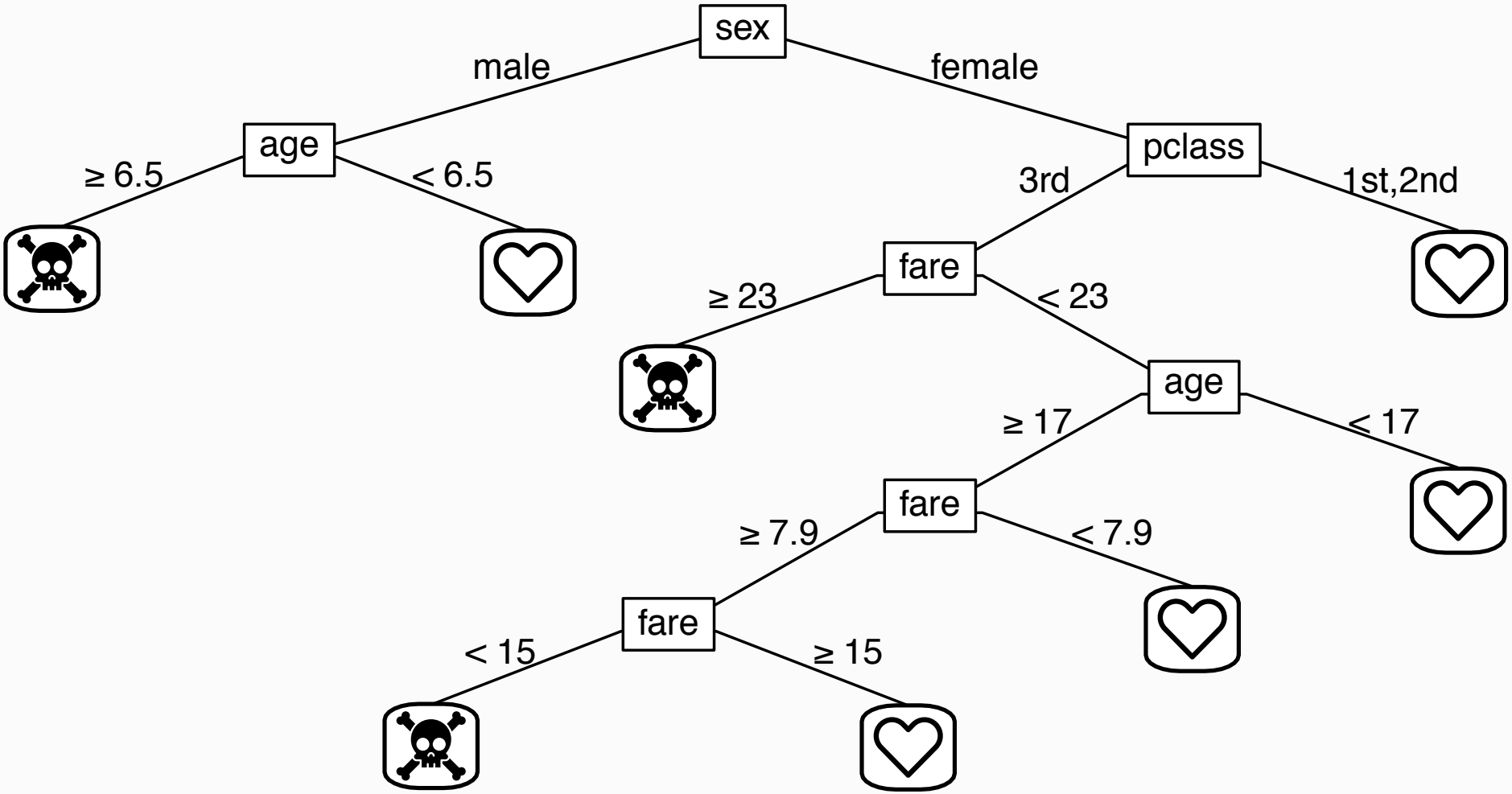
17th Annual Information Ethics Roundtable: Justice and Fairness in Data Use and Machine Learning
Northeastern Ethics Institute, Northeastern University
Boston, Massachusetts, April 5, 2019

Slides and paper draft: https://mominmalik.com/ier2019.pdf

Preliminaries

- I am not a philosopher
- Models, not algorithms
- Decision trees are both interpretable and explainable under any proposed definition
- Will use *Titanic* dataset: survival for each person aboard with covariates of age, sex, passenger class, fare

Decision tree for survival on *Titanic*



Overview

- Interpretability is the wrong conversation to be having for just use of machine learning. Causality is the real issue
- Interpretability of a non-causal model is actually useless
- "Prediction policy problems" would be cases for just use of non-causal models
- Maybe no such problems actually exist

Outline

- Machine learning vs. statistics
- The problem with explainable models
- A decision tree for *Titanic* survival
- The problem with "prediction"
- Prediction vs. causal explanation
- "Prediction policy problems"

Outline

- Machine learning vs. statistics
- The problem with explainable models
- A decision tree for *Titanic* survival
- The problem with "prediction"
- Prediction vs. causal explanation
- "Prediction policy problems"

Breiman and the "two cultures" (2001)

Google Scholar

Leo Breiman 1928-2005

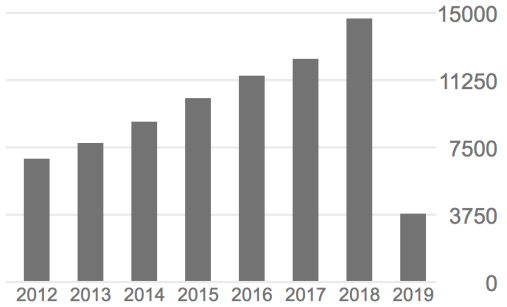
FOLLOW

Professor of Statistics, UC Berkeley
Verified email at stat.berkeley.edu - Homepage
Data Analysis Statistics Machine Learning

TITLE	CITED BY	YEAR
Random forests L Breiman Machine learning 45 (1), 5-32	44180	2001
Classification and regression trees L Breiman Chapman & Hall/CRC	40145 *	1984
Bagging predictors L Breiman Machine learning 24 (2), 123-140	20376	1996
Statistical modeling: The two cultures (with comments and a rejoinder by the author) L Breiman Statistical Science 16 (3), 199-231	~~2374~~	2001
Estimating optimal transformations for multiple regression and correlation L Breiman, JH Friedman Journal of the American Statistical Association, 580-598	1991	1985

Cited by VIEW ALL

	All	Since 2014
Citations	125205	61547
h-index	50	33
i10-index	78	46



Outline

- Machine learning vs. statistics
- **The problem with explainable models**
- A decision tree for *Titanic* survival
- The problem with "prediction"
- Prediction vs. causal explanation
- "Prediction policy problems"

Explanations of models seem to be about the world

if male **and** adult **then** *survival probability* 21% (19%–23%)
else if 3rd class **then** *survival probability* 44% (38%–51%)
else if 1st class **then** *survival probability* 96% (92%–99%)
else *survival probability* 88% (82%–94%)

- Decision list: interpretable and explainable
- Lethan, Rudin et al.: "For example, we predict that a passenger is less likely to survive than not *because* he or she was in the 3rd class."
- "Because" the model, or "because" the world?

But models are correlations, not causes

- Finale Doshi-Velez & Been Kim: "one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern."
- Rich Caruana et al.: "Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation."
- Zachary Lipton: "Another problem is that such an interpretation might explain the behavior of the model but not give deep insight into the causal associations in the underlying data... The real goal may be to discover potentially causal associations that can guide interventions."

E.g.: Trees are interpreted causally

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

But trees are not causal

- Is being in District N correlated with other variables?
- E.g., fewer resources?
- If so, maybe it's not the district but the resources that causes delays
- Causal question: would other districts have the same delay if given the same (lack of) resources?
- But decision trees split on whatever is optimal

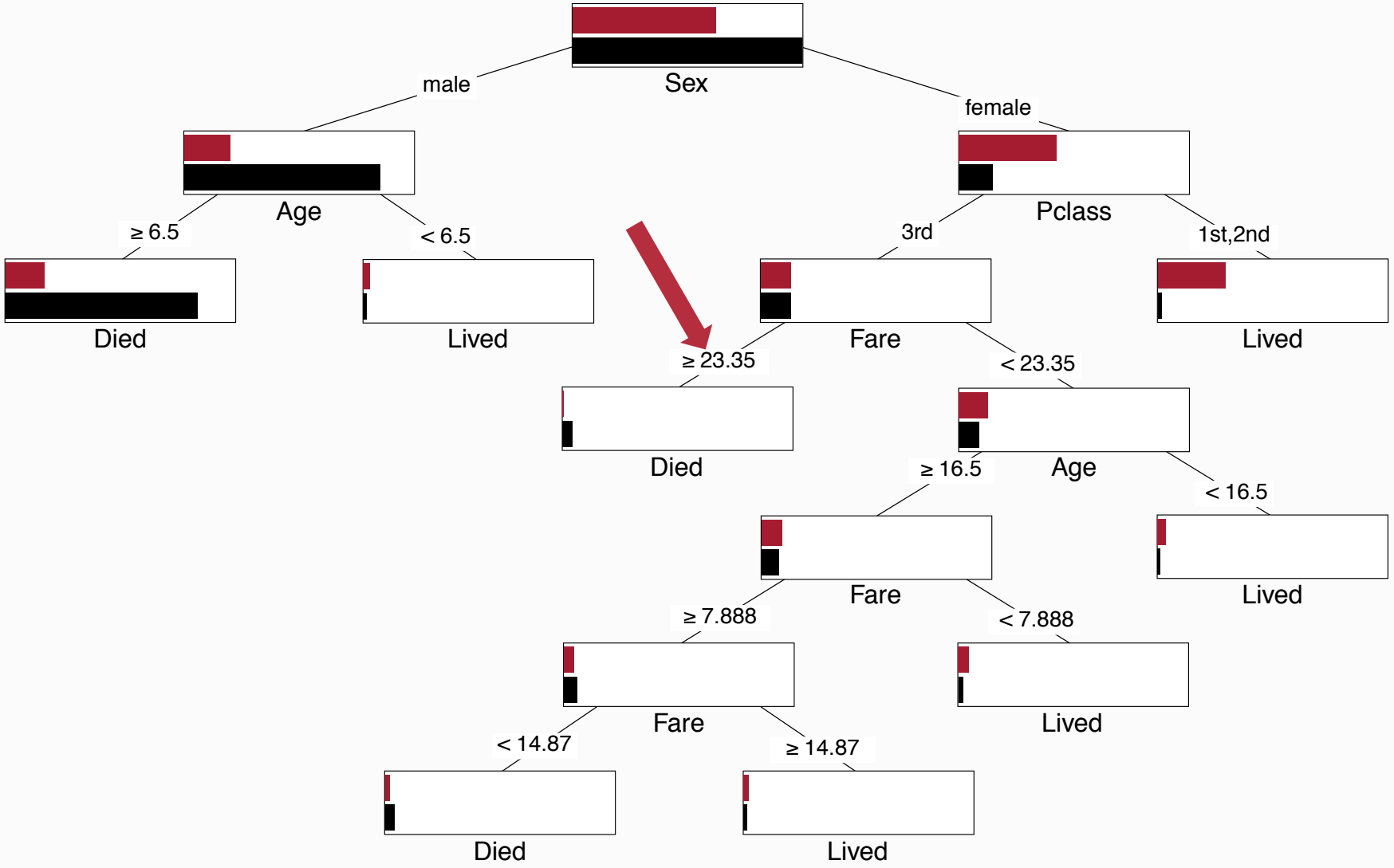
Outline

- Machine learning vs. statistics
- The problem with explainable models
- **A decision tree for *Titanic* survival**
- The problem with "prediction"
- Prediction vs. causal explanation
- "Prediction policy problems"

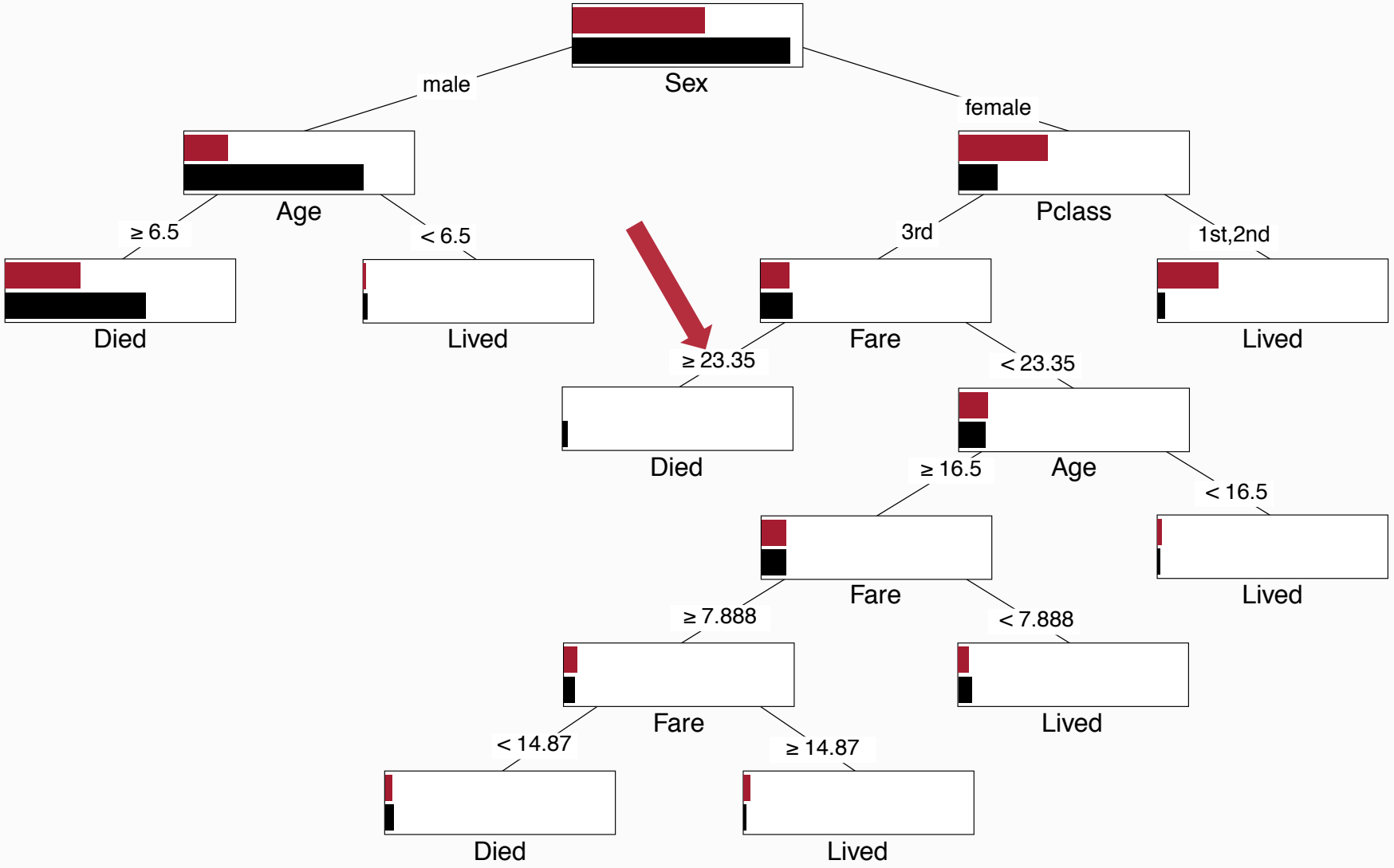
Wish list for interpretability

- Face validity as a way to check the model
- Anticipate where the model might break down (e.g., when it fails face validity)
- Use domain knowledge to 'fine-tune' the model

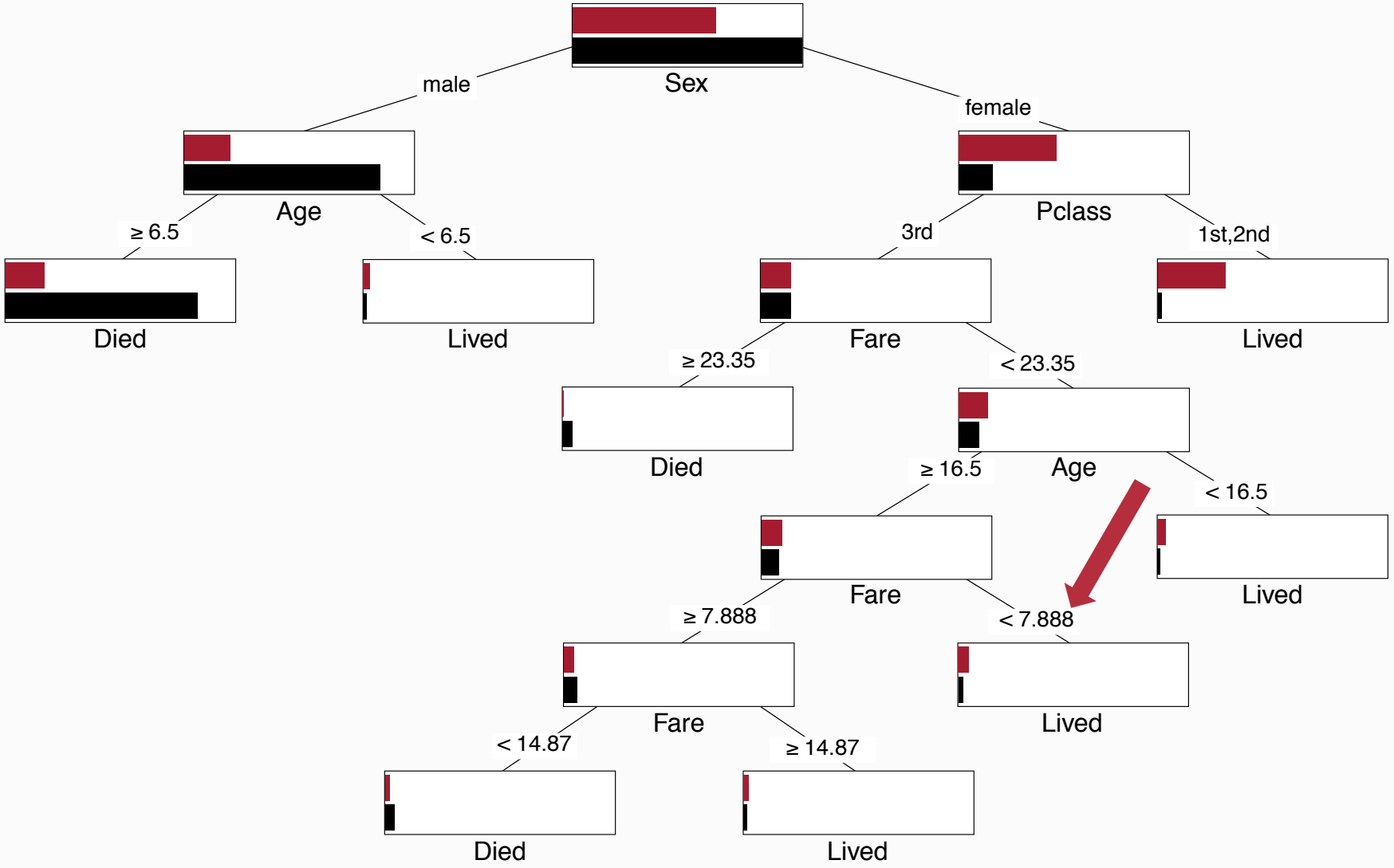
Female, 3rd class less likely to survive because of higher fare?



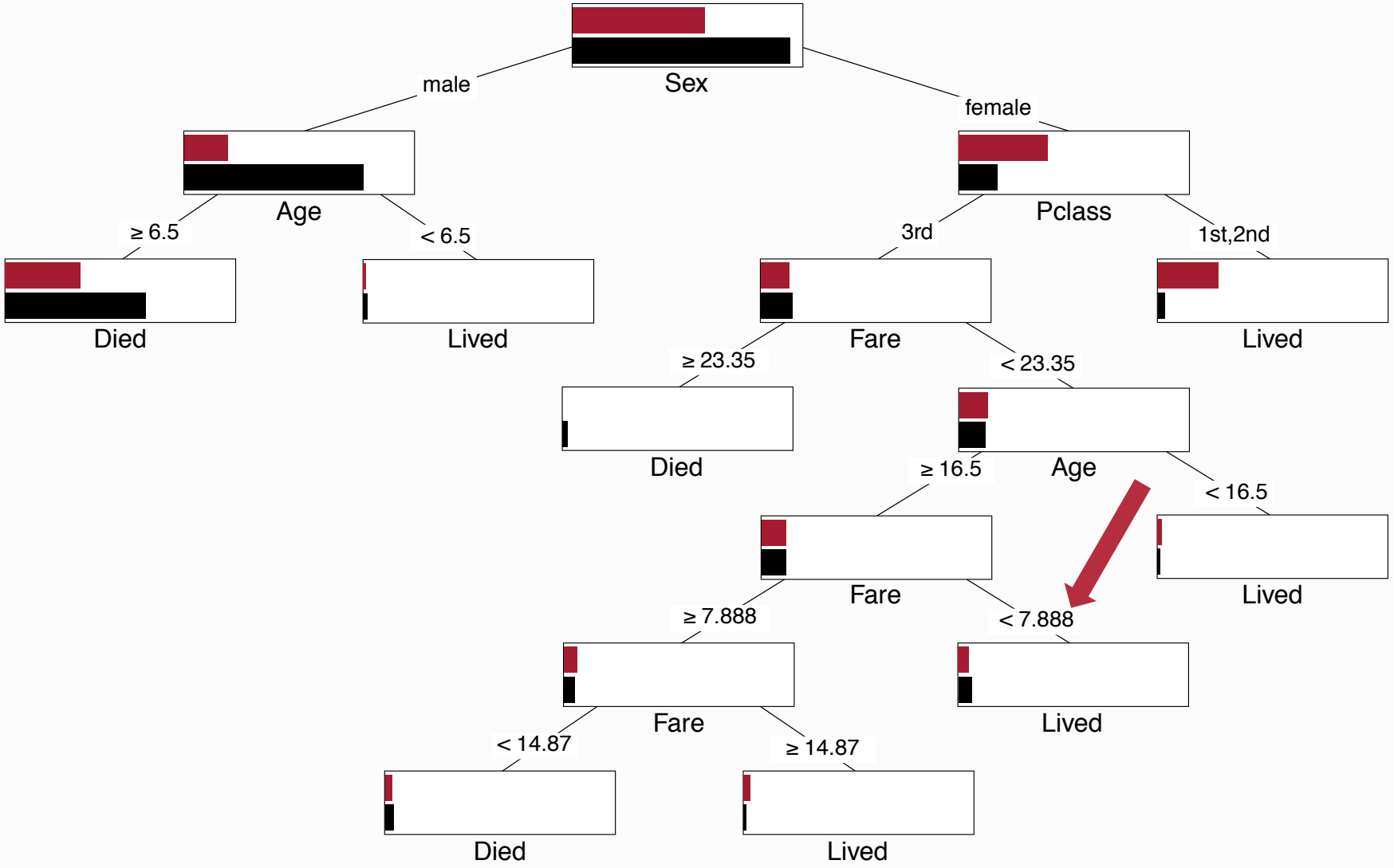
Lacks face validity, but holds on test data



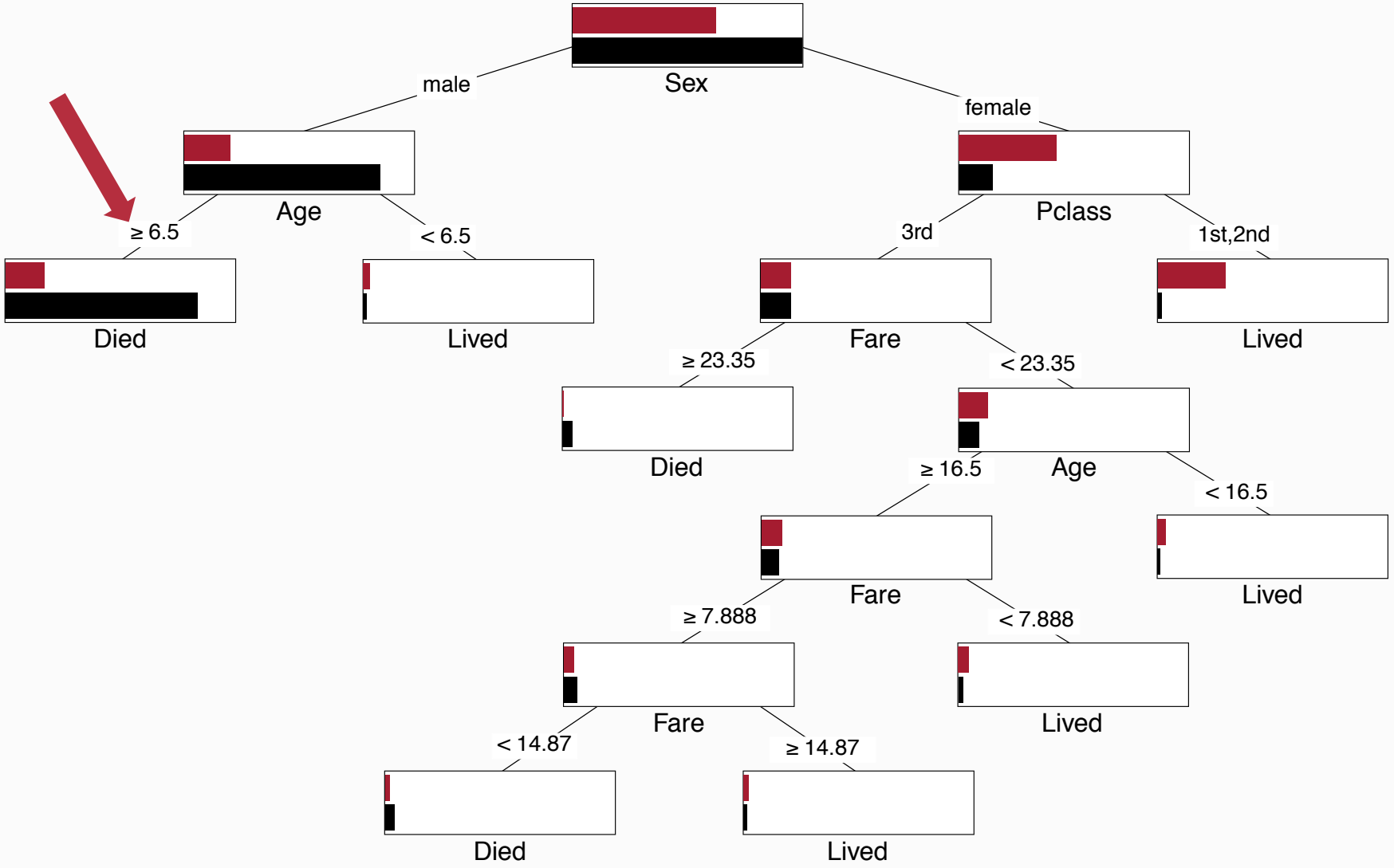
Converse: has face validity, but fails to generalize?



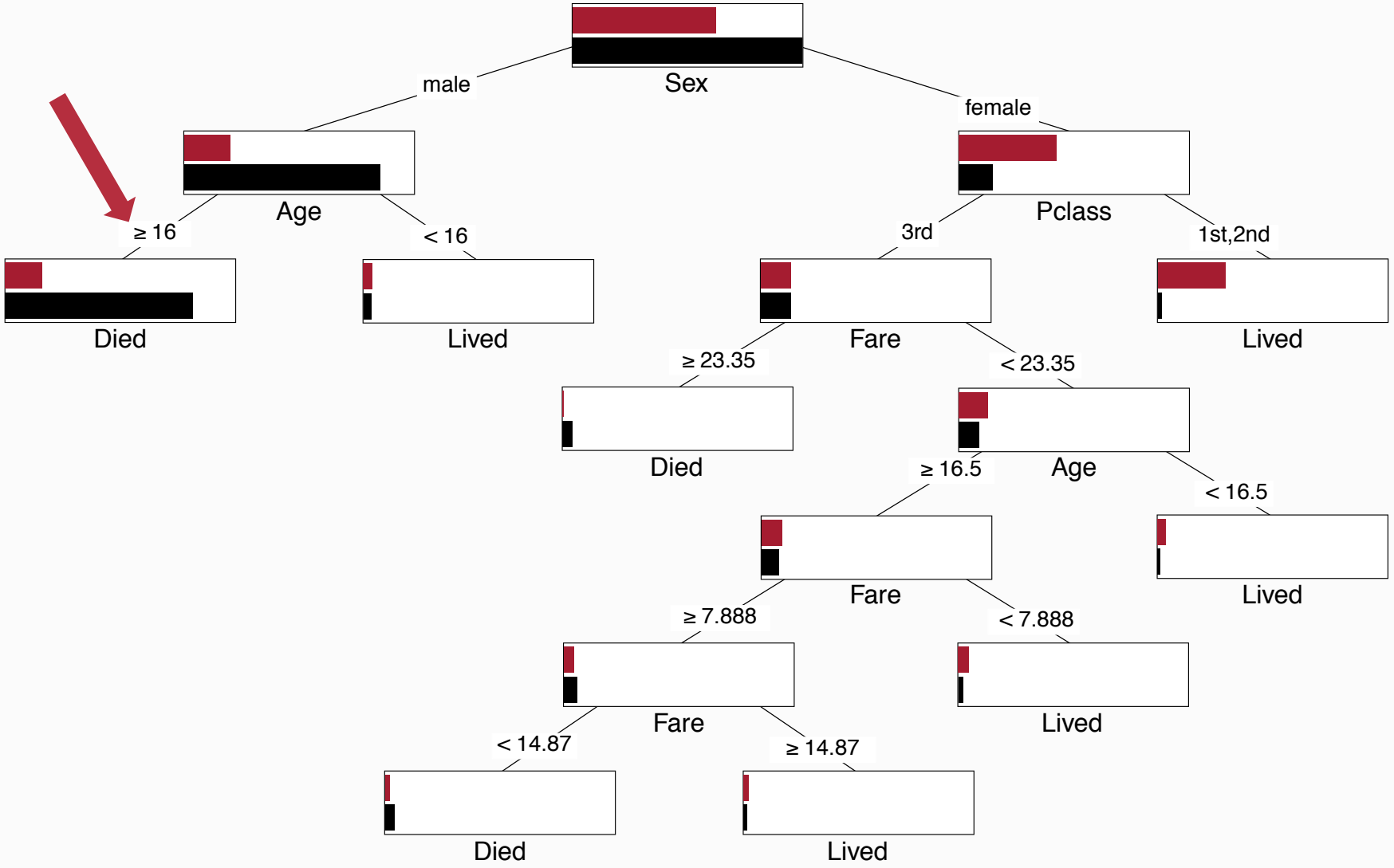
Yes. Interpretability doesn't help anticipate breakdowns



Interpretations to 'fine-tune' model?



Model is already optimally tuned



Outline

- Machine learning vs. statistics
- The problem with explainable models
- A decision tree for *Titanic* survival
- **The problem with "prediction"**
- Prediction vs. causal explanation
- "Prediction policy problems"

"Predict" the future?

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

***Abstract*—In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.**

I. Introduction

Social media has exploded as a category of online discourse where people create content, share it, bookmark it and network at a prodigious rate. Examples include Facebook, MySpace, Digg, Twitter and JISC listservs on the academic side. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry.

Since social media can also be construed as a form of collective wisdom, we decided to investigate its power at predicting real-world outcomes. Surprisingly, we discovered that the chatter of a community can indeed be used to make quantitative predictions that outperform those of artificial markets. These information markets generally involve the trading of state-contingent securities, and if large enough and properly designed, they are usually more accurate than other techniques for extracting diffuse information, such as surveys and opinions polls. Specifically, the prices in these markets have been shown to have strong correlations with observed outcome frequencies, and thus are good indicators of future outcomes [4], [5].

In the case of social media, the enormity and high variance of the information that propagates through large user communities presents an interesting opportunity for harnessing that data into a form that allows for specific predictions about particular outcomes, without having to institute market mechanisms. One can also build models to aggregate the opinions of the collective population and gain useful insights into their behavior, while predicting future trends. Moreover, gathering information on how people converse regarding particular products can be helpful when designing marketing and advertising campaigns [1], [3].

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

We have focused on movies in this study for two main reasons.

- The topic of movies is of considerable interest among the social media user community, characterized both by large number of users discussing movies, as well as a substantial variance in their opinions.
- The real-world outcomes can be easily observed from box-office revenue for movies.

Our goals in this paper are as follows. First, we assess how buzz and attention is created for different movies and how that changes over time. Movie producers spend a lot of effort and money in publicizing their movies, and have also embraced the Twitter medium for this purpose. We then focus on the mechanism of viral marketing and pre-release hype on Twitter, and the role that attention plays in forecasting real-world box-office performance. Our hypothesis is that movies that are well talked about will be well-watched.

Next, we study how sentiments are created, how positive and negative opinions propagate and how they influence people. For a bad movie, the initial reviews might be enough to discourage others from watching it, while on the other hand, it is possible for interest to be generated by positive reviews and opinions over time. For this purpose, we perform sentiment analysis on the data, using text classifiers to distinguish positively oriented tweets from negative.

Our chief conclusions are as follows:

- We show that social media feeds can be effective indicators of real-world performance.
- We discovered that the rate at which movie tweets are generated can be used to build a powerful model for predicting movie box-office revenue. Moreover our predictions are consistently better than those produced by an information market such as the Hollywood Stock Exchange, the gold standard in the industry [4].

[1] http://www.twitter.com

MIT Technology Review

Topics+ The Download Magazine Events

Intelligent Machines

Software Predicts Tomorrow's News by Analyzing Today's and Yesterday's

Prototype software can give early warnings of disease or violence outbreaks by spotting clues in news reports.

by Tom Simonite February 1, 2013

A method of using online information to accurately predict the future could transform many industries.

"Prediction" is not prediction!

- Daniel Gayo-Avello: "*It's not prediction at all!* I have not found a single paper predicting a future result. All of them claim that a prediction could have been made; i.e. they are *post-hoc* analysis and, needless to say, negative results are rare to find."

"Predictions" are correlations

- Lipton: "The real goal may be to discover potentially causal associations that can guide interventions, as with smoking and cancer. The optimization objective for most supervised learning models, however, is simply to minimize error, a feat that might be achieved in a purely correlative fashion."

Prediction as minimizing error not obvious or inevitable

- Milton Friedman: purpose of "positive economics" is "to provide a system of generalizations that can be used to make correct predictions *about the consequences of any change in circumstances*." [emphasis added]
- Physics: prediction is of the *results of an experiment*
- Causal inference: causality will help predictions be *robust*

Outline

- Machine learning vs. statistics
- The problem with explainable models
- A decision tree for *Titanic* survival
- The problem with "prediction"
- **Prediction vs. causal explanation**
- "Prediction policy problems"

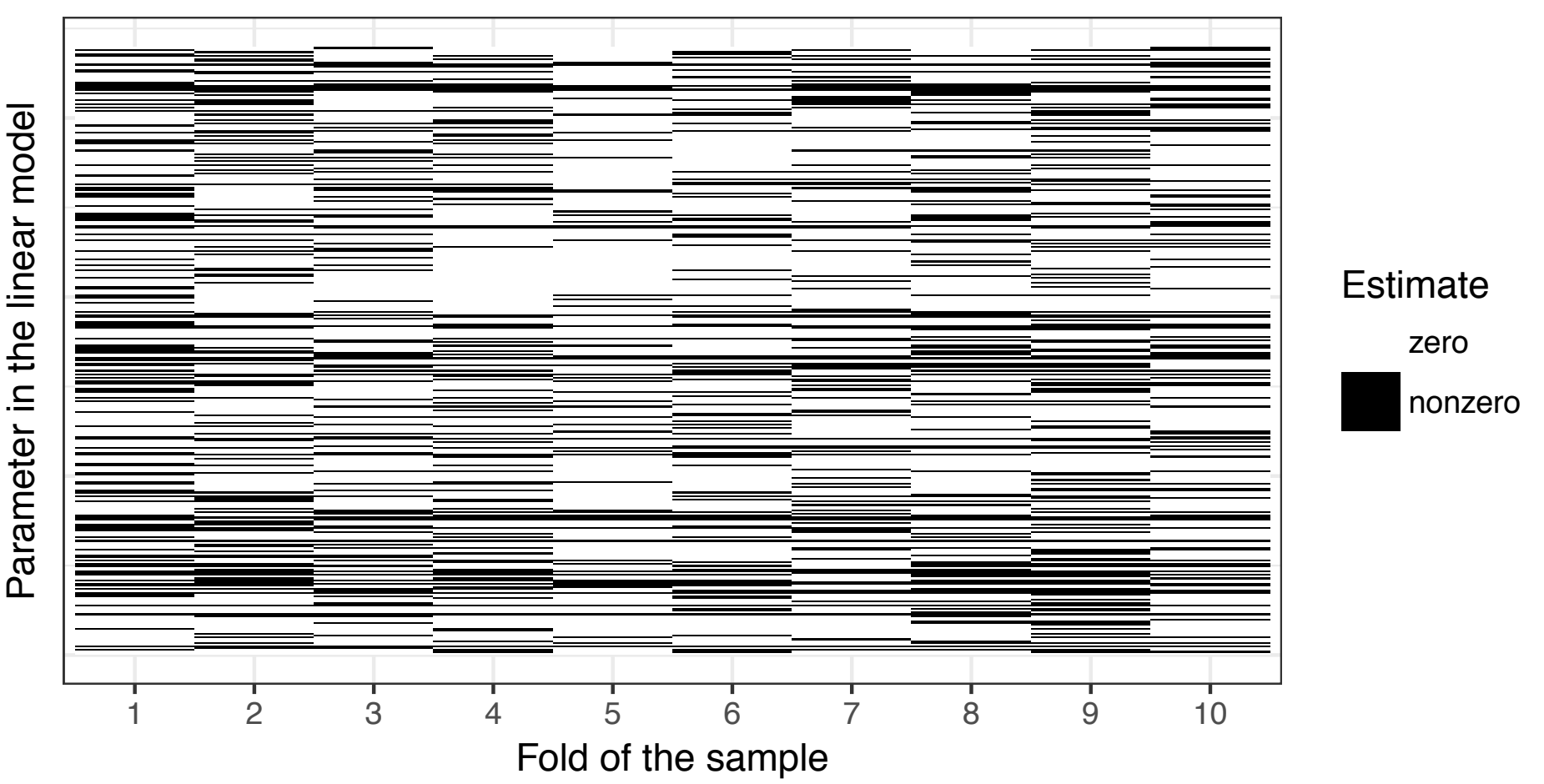
Two distinct modeling goals

- Leo Breiman: models for prediction, or models for information
- Galit Shmueli: prediction and (causal) explanation
- Sendhil Mullainathan (2014): "umbrella" problems and "rain dance" problems
- Sendhil Mullainathan (2017): $\hat{y}$ problems and $\hat{\beta}$ problems

They are in competition

- Shmueli: Because of regularization, a 'false' model can predict better than a 'true' one (see also: Stein's paradox)
- Mullainathan & Spiess: very different sets of variables give equivalent predictive performance

Same predictions, different implications



Outline

- Machine learning vs. statistics
- The problem with explainable models
- A decision tree for *Titanic* survival
- The problem with "prediction"
- Prediction vs. causal explanation
- **"Prediction policy problems"**

Definition: *y* only

- Benefits of decision depend on an outcome variable, *y*
- Decision is a function only of outcome, not of things that go into modeling the outcome
- Then, can use "prediction" only: i.e., correlations

Can using correlations alone be just?

- Failure points:
  - Counterfactual comparison ("better than")
  - Intervening on the system
  - Intervening on covariates
  - "Gaming the system"
  - (Negative) feedback loops
  - Unreliable metrics and data
  - ...probably other things
- If we would want to challenge predictions, one clue that the problem is causal

Claimed prediction policy problems

- "Other illustrative examples include:
  - (i) in education, predicting which teacher will have the greatest value added (Rockoff et al. 2011);
  - (ii) in labor market policy, predicting unemployment spell length to help workers decide on savings rates and job search strategies;
  - (iii) in regulation, targeting health inspections (Kang et al. 2013);
  - (iv) in social policy, predicting highest risk youth for targeting interventions (Chandler, Levitt, and List 2011); and
  - (v) in the finance sector, lenders identifying the underlying credit-worthiness of potential borrowers."

(i) Teacher with greatest value added

- "Value added" model has been critiqued
- The metric is a proxy for underlying goal
- Will a group of students learn better under this teacher? Counterfactual, so, causal

(ii) Unemployment spell length

- Purpose: "help workers decide on savings rates and job search strategies"
- Job search strategies, potentially saving rates as well, would affect unemployment length spell
- If the goal is to decrease unemployment spell, then is causal

(iii) Targeting health inspections

- Cited paper's results have recently been refuted (Daniel Ho and Kristen Altenburger, *The Web Conference* 2019)
- Goal is not finding violations, but encouraging systematic compliance
- What targeting strategy will cause restaurants to comply?

(iv) Highest risk youth

- Goal: decrease risk (i.e., frequency of negative outcomes)
- Will interventions be on any aspects that go into the model?
- Certainly, is intervening on the system

(v) Underlying creditworthiness

- Probably the best existing example of injustices from prediction-only
- Martha Poon, "Scorecards as Devices for Consumer Credit"
- Josh Lauer, *Creditworthy*
- Marion Fourcade and Kieran Healy, "Classification Situations"

Best candidate: Bail decisions

- Still ultimately fails because can be arrested/charged without committing a crime, and can commit a crime without being arrested/charged
- If that weren't the case, couldn't game the system
- Is a counterfactual comparison...

Best candidate: Bail decisions

- Mullainathan and colleagues: censored data problem. Don't know if a person, if released, would have been arrested/ charged
- Generalization error is an underlying quantity: cross-validated test error is a (biased) estimator
- Use *judge leniency as an instrumental variable* to get unbiased estimates of prediction

Best candidate: Bail decisions

- Still ultimately fails because of data problems
- But otherwise, would have worked
- Tremendous amount of work went into making the argument that it is a prediction policy problem
- That's what it *should* take to determine if the use of machine learning is just

Places where unjust use may be okay

- Less unjust than the status quo
  - E.g., no people denied bail who would have otherwise been released, only additional people released
  - (But: danger of institutionalization, and unequal distribution of the lesser injustice)
- Maybe unjust aspects are minor (e.g., minimal negative feedback loops, or small gains from gaming)
- When threats are to the people in power

Conclusion

- Interpretability is the wrong conversation to be having for just use of machine learning. Causality is the real issue
- Interpretability of a non-causal model is actually useless
- "Prediction policy problems" would be cases for just use of non-causal models
- Maybe no such problems actually exist

Interpretability is a Red Herring: Grappling with "Prediction Policy Problems"

Presented at the 17th Annual Information Ethics Roundtable:
Justice and Fairness in Data Use and Machine Learning
April 5–7, 2019, Northeastern University, Boston, MA

Momin M. Malik <momin_malik@cyber.harvard.edu>
Berkman Klein Center for Internet & Society at Harvard University

Version 1.1
April 5, 2019

Abstract

The goal of 'interpretability' fails to grapple with the core paradox of machine learning: that we can make effective predictions on the basis of non-causal correlations. If a machine learning model's correlations are interpretable but non-causal, then we will be systematically misled if we try to use prior knowledge or intuition about how the world works as a way of validating the model's operation, or if we try to anticipate when the model might break down under changing conditions of the world, or if we seek to 'fine tune' parts of the model that we may interpret as effectively unjust while retaining the model's integrity. Interpretability may be useful for model diagnostics and debugging, but not for ensuring just usage. For just usage, our focus should instead be on whether a situation is one in which correlations are sufficient: a 'prediction policy problem.' If we have such a problem, interpretability is not necessary. Conversely, if we do *not* have such a case, we should not be using machine learning at all, interpretable or not. But determining whether something is indeed a prediction policy problem may be so difficult as to leave little space for the just use of machine learning when it comes to human systems.

Introduction

In this paper, I endeavor to show that 'interpretability' is insufficient as a guide for using machine learning in just ways. My goal is not to engage in definitional debates, as I believe my critique will apply to any definition of interpretability that is not in terms of "capturing causality."[1]

Cynthia Rudin notes that "Interpretability is a domain-specific notion, so there cannot be an all-purpose definition."[2] She cites obeying causal knowledge of a domain as one possible aspect of interpretability, but also notes that interpretability may simply be that a model is "useful to someone". One of her works with colleagues gives rules derived from decision trees as a concrete example of an interpretable model:

> Our goal is to build predictive models that are highly accurate, yet are highly interpretable. These predictive models will be in the form of sparse *decision lists*, which consist of a series of *if...then...* statements where the *if* statements define a partition of a set of features and the *then* statements correspond to the predicted outcome of interest. Because of this form, a decision list model naturally provides a reason for each prediction that it makes.

Taking the example of the dataset of passenger survival aboard the *Titanic*, the same example I will use further below, they write,[3]

[1] Can a non-interpretable model capture causality? Maybe not, insofar as causality relates to human understanding of our ability to make and identify interventions. If we were to define interpretability in terms of causality, we might as well talk in terms of causality directly.

[2] Cynthia Rudin, "Please Stop Explaining Black Box Models for High Stakes Decisions," in *The NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning*, 2018 Conference on Neural Information Processing Systems (NeurIPS 2018) (2018), https://arxiv.org/abs/1811.10154.

[3] Benjamin Letham et al., "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model," *The Annals of Applied Statistics* 9, no. 3 (September 2015): 1350–1371, doi:10.1214/15-AOAS848.

> The goal is to predict whether the passenger survived based on his or her features. The list provides an explanation for each prediction that is made. For example, we predict that a passenger is less likely to survive than not *because* he or she was in the 3rd class.

The causality of this 'because' statement relates to the *model* and not to the underlying system, which illustrates the confusion I think arises with interpretability. That is, I believe this statement *should* be read as "being in 3rd class is the cause of *our model predicting* a passenger as less likely to survive" (the *model* "providing a reason for each prediction that it makes") but the statement *seems* to be saying, "being in 3rd class *caused* passengers to be less likely to survive" (which would be the *world*, or our knowledge of it, providing a reason). The second statement is likely true, and indeed both empirical evidence from the data as well as contextual historical information supports making such an argument. But as a general principle, decision trees or rules are constructed from correlations between features/variables and an outcome, correlations which may be causal—but also may not be.

Work discussing the difficulties of choosing meaningful definitions and standards for interpretability[4] do raise the problem of confusing interpretability and causality. I go further, and propose that when model is "interpretable," we end up conflating the logic of our interpretation (causal relationships in the world) with the logic of the model (optimal correlations). This makes interpretability a dangerous distraction. Instead, we have to ask when optimal correlations are sufficient for a problem at hand. In public policy, Jon Kleinberg, Sendhil Mullainathan, and colleagues call these "prediction policy problems."

Insofar as such prediction policy problems exist, and insofar as machine learning is more effective for these problems, what is required for the just use of machine learning is to *identify problems as prediction policy ones*. For those, we should use input/output testing in real-world settings to establish *reliability*, and not worry about interpretability. Interpretability may still be useful for model diagnosis and debugging, but not for ensuring just usage.

Historical background

In his famous 2001 paper, Leo Breiman argued that the commitment of statisticians to modeling for gaining information about modeled systems "has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.".[5] There are applied problems, Breiman argued, in which it is sufficient to have models that can reliably anticipate/match the outputs of a system from given inputs, without needing to know how those inputs are connected to the outputs. He describes his exposure to and examples of such problems:

> "After a seven-year stint as an academic probabilist, I resigned and went into full-time freelance consulting. After thirteen years of consulting I joined the Berkeley Statistics Department in 1980 and have been there since... As a consultant I designed and helped supervise surveys for the Environmental Protection Agency (EPA) and the state and federal court systems. Controlled experiments were designed for the EPA, and I analyzed traffic data for the U.S. Department of Transportation and the California Transportation Department.[6] Most of all, I worked on a diverse set of prediction projects. Here are some examples:
>
> - "Predicting next-day ozone levels.
> - "Using mass spectra to identify halogen-containing compounds.
> - "Predicting the class of a ship from high altitude radar returns.
> - "Using sonar returns to predict the class of a submarine.
> - "Identity of hand-sent Morse Code.

[4] Finale Doshi-Velez and Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, arXiv:1702.08608, https://arxiv.org/abs/1702.08608; Zachary C. Lipton, "The Mythos of Model Interpretability," *ACM Queue* 16, no. 3 (June 2018): 31–57, doi:10.1145/3236386.3241340.

[5] Leo Breiman, "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)," *Statistical Science* 16, no. 3 (2001): pp. 199, doi:10.1214/ss/1009213726.

[6] Breiman conspicuously but understandably leaves out the extensive, classified military work he did during this period for the US Department of Defense. The influence of these problems in the development of his thinking is discussed in Matthew L. Jones, "How We Became Instrumentalists (Again): Data Positivism since World War II," *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673–684, doi:10.1525/hsns.2018.48.5.673.

- "Toxicity of chemicals.
- "On-line [real-time] prediction of the cause of a freeway traffic breakdown.
- "Speech recognition[.]
- "The sources of delay in criminal trials in state court systems."

Breiman argued that what he labeled as "algorithmic modeling," as opposed to "data modeling," provides much more effective solutions to these problems than those of traditional statistics, and that "interesting new developments" of these models "has occurred largely outside statistics in a new community—often called machine learning—that is mostly young computer scientists."

More ominously, Breiman wrote that "the damaging consequence of the insistence on data models is that statisticians have ruled themselves out of some of the most interesting and challenging statistical problems that have arisen out of the rapidly increasing ability of computers to store and manipulate data. These problems are increasingly present in many fields, both scientific and commercial, and solutions are being found by nonstatisticians." The implication: statisticians were poised to become irrelevant, replaced by machine learning, and would had nobody to blame but themselves. This proved prophetic, validating his critique and giving it extra sting in retrospect.

He also explicitly brings up interpretability, noting that models that are the best "predictors" are often difficult for practitioners to make sense of. About decision trees, he says, "While trees rate an A+ on interpretability, they are good, but not great, predictors. Give them, say, a B on prediction."[7]

However, the exact meaning of "prediction" bears closer examination.

"Predictions" are post-hoc correlations

In writing about claims to predict election results with Twitter, Daniel Gayo-Avello writes,[8]

> "*It's not prediction at all!* I have not found a single paper predicting a future result. All of them claim that a prediction could have been made; i.e. they are *post-hoc* analysis and, needless to say, negative results are rare to find."[9]

In one sense, this critique is unfair, because a "prediction" is a technical term, defined in terms of something that minimizes a static loss function on previously observed data. A "predicted value" is synonymous with a "fitted value," and is not defined necessarily as saying anything about the future[10]—it is simply an assumption that correlations observed in the past will be effective for anticipating the future.

This assumption is often justified and will often hold, but in another sense, this is not a natural or inevitable way of using prediction and so Gayo-Avello's critique is more than fair. For example, Milton Friedman's view of "positive economics" sees its purpose as being "to provide a system of generalizations that can be used to make correct predictions *about the consequences of any change in circumstances* [emphasis added]."[11] And in statistical mechanics, taking a textbook[12] as an example, "prediction" is always spoken of in terms of predicting the *outcome of experiments*; that is, manipulations/interventions, which are causal. If statistics and machine learning were to define prediction as minimizing a loss function under changes and interventions, then correlations could not necessarily be used for predicting "the future" (i.e., fitted values would have far greater loss than anticipated

[7] The latter is contested in the above cited work by Rudin and colleagues, as they demonstrate decision rules with predictive accuracy "on par with the current top algorithms [models] for prediction in machine learning," but the important point here is that decision trees are firmly an example of an interpretable model.

[8] Daniel Gayo-Avello, "No, You Cannot Predict Elections with Twitter," *IEEE Internet Computing* 16, no. 6 (2012): 91–94, doi:10.1109/MIC.2012.137.

[9] Italics are original in the pre-print, Daniel Gayo-Avello, *'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper' — A Balanced Survey on Election Prediction using Twitter Data*, arXiv:1204.6441, May 2012, http://arxiv.org/abs/1204.6441v1.

[10] Hence the proliferation of seemingly self-redundant titles about "predicting the future", e.g., Sitaram Asur and Bernardo A. Huberman, "Predicting the Future with Social Media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '10 (2010), 492–499, doi:10.1109/WI-IAT.2010.63.

[11] Milton Friedman, *Essays in Positive Economics* (University of Chicago Press, 1953).

[12] Robert H. Swendsen, *An Introduction to Statistical Mechanics and Thermodynamics* (Oxford University Press, 2012), doi:10.1093/acprof:oso/9780199646944.001.0001.

on out-of-sample data[13]). Relatedly, the long-standing literature about causal inference/discovery[14] makes the case to those "algorithmic modelers" that, even if their only interest is in prediction, their "predictions" will fail to *actually* predict if something about the underlying system shifts (that is, fail to be robust).

When 'predictions' fail to explain

Understanding the specific way in which prediction is defined lets us see that just because a model "predicts" well (finds a combination of features that correlate very well with the outcome) does not mean it "explains" well in terms of capturing casual relationships, or even of capturing associations (the "pure" relationship of a variable on the outcome even if causal directionality is unknown, rather than the estimated relationship being confused by collinearity, unmodeled interaction effects, missing mediating/moderating variables, or other forms of model misspecification). But can using only correlations do *better* at capturing the variability of the outcome than trying to ascertain causal relationships?

Galit Shmueli raises a core paradox: models that explain well may not predict well, and conversely, models that predict well may be poor explanations (rather than Breiman's prediction vs. information, Shmueli uses prediction vs. [causal] "explanation").[15]

In the one direction, of models attempting explanation doing a poor job at predicting, we can (as Breiman also suggests) easily imagine the world being more complex than the forms of models we apply to it when forming explanations (e.g., through linear or generalized linear models, through additive models perhaps with up to two-way interactions, and through models that assume independence between observations). Under this way of thinking, we could continue and suppose that models that "fit to the shape of the data" in ways that entail fewer assumptions (i.e., nonparametric models, or other similarly flexible models from machine learning) will better capture the complexity of the world. It is less clear how a model attempting explanation can do a good job at explaining even if it fails at predicting, but consider a regression model where a 1-unit rise in X_p is robustly associated with a $\widehat{\beta}_p$ rise in y, but that overall R^2 is low (although R^2 is a poor measure of model fit[16]).

Given the definition of prediction, clearly non-causal correlations can achieve it. But the idea that they can do better than causal relationships remains deeply unintuitive. Shmueli provides a helpful example, drawn from a chemical engineering journal.[17] This shows that we can find the conditions under which an "underspecified model" (a model with fewer variables than the "truth") has lower error than if we fit the exact equation that originally generated the data (which we can take to be the "truth"). The conditions turn out to be sensible if pathological (that is, highly unlikely to happen in the world), and include high irreducible variance, collinearity between variables, and some variables having small magnitudes compared to the ones with which they are collinear. But it is an existence proof that shows, within the world of technical definitions, it is possible for a "false" model to predict better than a "true" model!

This example relates to *regularization*, and involves a paradox that goes back decades in statistics.[18] But we can also think about different sets of correlations, for which econometricians Sendhil Mullainathan and Jann Spiess provide a helpful example. They take data from the American Housing Survey, split the data into 10 parts, and apply a regularization technique that came from statisticians but frequently used in machine learning, the "lasso" (also known as sparse regression, or technically as "ℓ_1 regularized regression"), which effectively "selects" a subset of features that together achieve the best predictive performance. In each of the 10 subsets, a very different set of features are selected in (fig. 1), yet the "predictive" performance (the difference between the fitted values and the observed values) is about the same. The implication: because of collinearity, given slightly different realizations of the same underlying process, very different models—with vastly different implications for intervention—may perform equally well.[19]

[13] A good example of this happening is in the "Parable of Google Flu Trends." While phrased as overfitting, I think this is better understood of how using correlations to make predictions can fail if something about the underlying system changes, such as flu incidences not co-occurring with winter, or other less directly interpretable changes. David Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis," *Science* 343, no. 6176 (2014): 1203–1205, doi:10.1126/science.1248506.

[14] Peter Spirtes and Kun Zhang, "Causal Discovery and Inference: Concepts and Recent Methodological Advances," *Applied Informatics* 3, no. 3 (2016): 1–28, doi:10.1186/s40535-016-0018-x.

[15] Galit Shmueli, "To Explain or to Predict?," *Statistical Science* 25, no. 3 (2010): 289–310, doi:10.1214/10-STS330.

[16] Sanford Weisberg, *Applied Linear Regression*, 3rd ed. (Wiley, 2005), pp. 81–84.

[17] Shaohua Wu, T. J. Harris, and K. B. McAuley, "The Use of Simplified or Misspecified Models: Linear Case," *The Canadian Journal of Chemical Engineering* 85, no. 4 (2007): 386–398, doi:10.1002/cjce.5450850401.

[18] Bradley Efron and Carl Morris, "Stein's Paradox in Statistics," *Scientific American* 236, no. 5 (1977): 119–127, doi:10.1038/scientifica

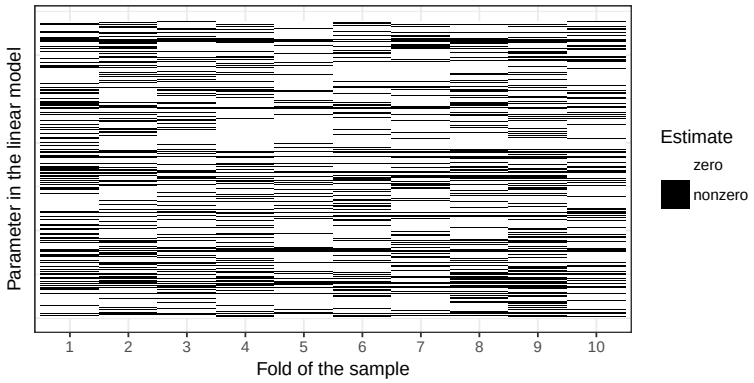


Figure 1: Features selected in across different subsets of the American Housing Survey. Figure reproduced from code from the supplementary materials of Sendhil Mullainathan and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106, doi:10.1257/jep.31.2.87.

Prediction vs. causation, and beyond

Mullainathan and colleagues pick up on some of the same themes as Breiman and Shmueli, although without citing either. In a 2015 paper[20] they argue almost identically to Breiman that there is a class of problems, which they formalize as "prediction policy problems", that are common and important, that have been neglected in empirical policy research, and which machine learning solves much more effectively than "traditional regression approaches." Here, they use the fanciful but unfortunately primitivist language of "rain dance problems," where we need to know if a particular intervention will result in the desired outcome, versus "umbrella problems," where we only need to know about the future state of the world in making our decisions.

In a later paper,[21] there is better language of "$\widehat{y}$ problems" and "$\widehat{\beta}$ problems," referring to the parts of a regression equation $\widehat{y} = \widehat{\beta}\mathbf{x}$, which is an estimate of the (hypothesized underlying relationship) $y = \beta\mathbf{x}+\varepsilon$. Here again they recognize a trade-off; with machine learning we can get $\widehat{y}$ ("predicted value" or "fitted value") that are closer to the actual values y, while not having $\widehat{\beta}$ close to the actual β (which is hopefully causal, and for which a causal interpretation is the goal of econometrics, but may also suffice as estimate a "pure relationship" as discussed above). In contrast, if we focus on getting $\widehat{\beta}$ as close to β as possible, we may sacrifice how close our $\widehat{y}$ is to y. Umbrella problems, or $\widehat{y}$ problems, are *prediction policy problems*: where we can apply whatever models give the best correlations, and don't care about if or how those correlations connect to underlying causal processes. Given the confusing nature of the technical sense of "prediction," these would be better (if less alliteratively) called *correlation policy problems*, although I will stay with the published term.

A persistent problem will be how to identify problems as prediction policy ones; they may be far less common than the mass deployment of machine learning would suggest in cases of decision-making about outcomes related to people. And, contrary to Kleinberg et al., there are perhaps insurmountable difficulties in identifying them. Alternatively, we might say that there are real cases of prediction policy problems, but the pendulum has swung too far in the direction of treating everything as a prediction policy problem, and what is needed is a return to the prevalence of the kind modeling on which Breiman critiqued the reliance of statisticians or, in fact, a turn to *qualitative* modeling.

While the framework of causality is effective at explaining how machine learning can fail, and what its limits are, it too is a limited frame of the world. Notably, taking Andrew Abbott's critique of the entire way in which statistical modeling sets up the world,[22] anything that requires things to be distinguished but associated with

merican0577-119.

[19] Note that "stability selection" tries to select a set of variables that are important across multiple subsets of data, as a way to try and get robustness. However, such a "stable set" of variables may not exist to select. Nicolai Meinshausen and Peter Bühlmann, "Stability Selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, no. 4 (2010): 417–473, doi:10.1111/j.1467-9868.2010.00740.x.

[20] Jon Kleinberg et al., "Prediction Policy Problems," *American Economic Review* 105, no. 5 (2015): 491–95, doi:10.1257/aer.p20151023.

[21] Sendhil Mullainathan and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106, doi:10.1257/jep.31.2.87.

[22] Andrew Abbott, "Transcending General Linear Reality," *Sociological Theory* 6, no. 2 (1988): 169–186, doi:10.2307/202114.

each other as units of a population with fixed and independent attributes, and that relies on a central tendency to characterize the population,[23] is a limited view of the world and one that excludes and is used to invalidate other ways of knowing, such as lived experience.[24] More directly, how might we fit structural racism into the language of certain variables causing others? Eugene Richardson is developing a critique about how such diffuse, long-range effects are difficult to fit into a quantitative framework, leading to them being ignored.[25] Even before questions of causality or correlations, or whether we put in effort to quantify the uncertainty of out models[26] instead of ignoring it as machine learning usually does, we should consider the consequences of standardization, bureaucratization, formalization, quantification, datafication, and mass systems overall and possible alternatives—such as qualitative analysis, participatory action research that recognizes people (rather than quantitative measures) as the experts of their own experience, or prison abolition that rejects calls for greater efficiency and/or justice in carceral systems and instead recognizes those systems themselves as inherently unjust and deserving of dissolution in favor of alternatives.

Interpreting the wrong thing: A decision tree on the *Titanic* data

I now turn to my central example, presenting what is unquestionably an 'interpretable' model, a decision tree, going through some of what interpreting it might mean and how those interpretations are insufficient for validation, anticipating breakdowns, or fine-tuning. Using the same data as used by Rudin and colleagues, survival aboard the *Titanic*, with the same training/test split in the data as used in the Datacamp tutorial using these data.

I construct a decision tree in the R package `rpart` with default parameters (fig. 2).[27] From seeing the variables that the tree splits upon, and at what values and in what order, give an overall sense of what the model is finding important in the system (we can also quantify this through "feature importance," but that is one level of interpretation removed from the way the model will run, and is not necessary for a single, relatively shallow tree, unlike a forest of shallow trees or a single tree with hundreds of branchings). We could also convert this to a set of rules in terms of if/then statements, but I will stay with the tree representation.

Face validity. Consider the split on fare at node 6. According to this, having a higher fare makes female third-class passengers *more* likely to perish. This holds on the test data as well: among 3rd class female passengers with fares below 23.35, 13 perished and 16 survived. Among 3rd class female passengers with fares above 23.35, 12 perished and 9 survived.

The tree has uncovered a counterintuitive pattern, but one that is robust (at least across this training/test division—which does not say anything about whether this might generalize to disasters in general, or only boat disasters, or only boat disasters in the 19th century, or only boat disasters in the 19th century with primarily western passengers, etc.). Of course, outside of this subset of data (female, 3rd-class passengers), having a lower fare correlates strongly with perishing, but much of this effect is "soaked up" by splitting passengers by class (which itself correlates strongly with fare, i.e. is strongly collinear). The tree does not reveal these collinearities, so any reasoning we do about this split making sense after all is speculation, or is drawing on additional analysis.

Anticipating breakdowns. As we saw above, a counterintuitive pattern held not only in training data but in test data as well. Then, a split that violates intuition may nevertheless be robust. Can a split that violates intuition fail to be robust? And, can a split that agrees with intuition fail to be robust? To look for possible examples, we can see how the test data percolates through the same decision tree (fig. 3). Node 5 gives us an example of an intuitive pattern that fails to hold in the test data: more males younger than 6.5 died than lived. It also gives us

[23] Todd Rose, *The End of Average: How We Succeed in a World that Values Sameness* (New York: HarperCollins Publishers, 2015).

[24] Candice Lanius, *Fact Check: Your Demand for Statistical Proof is Racist*, *Cyborgology* [blog], January 2015, https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/.

[25] Eugene Richardson, forthcoming, "Not-so-big Data and Ebola Virus Disease."

[26] D. R. Cox, "Role of Models in Statistical Analysis," *Statistical Science* 5, no. 2 (1990): 169–174, doi:10.1214/ss/1177012165.

[27] The default parameters for other decision tree-fitting packages, namely `tree` and `party`, produce different trees. I originally developed my critique around the result of the default parameters for `tree`, which make the argument more easily, but I decided to switch to the "standard" package for decision trees. Decision trees are unstable (i.e., slightly different subsets of data, and different tuning parameters, often result in very different trees although with similar performance), so a given tree might not have these objectionable qualities, but many trees will have something that contradicts intuition or domain knowledge.

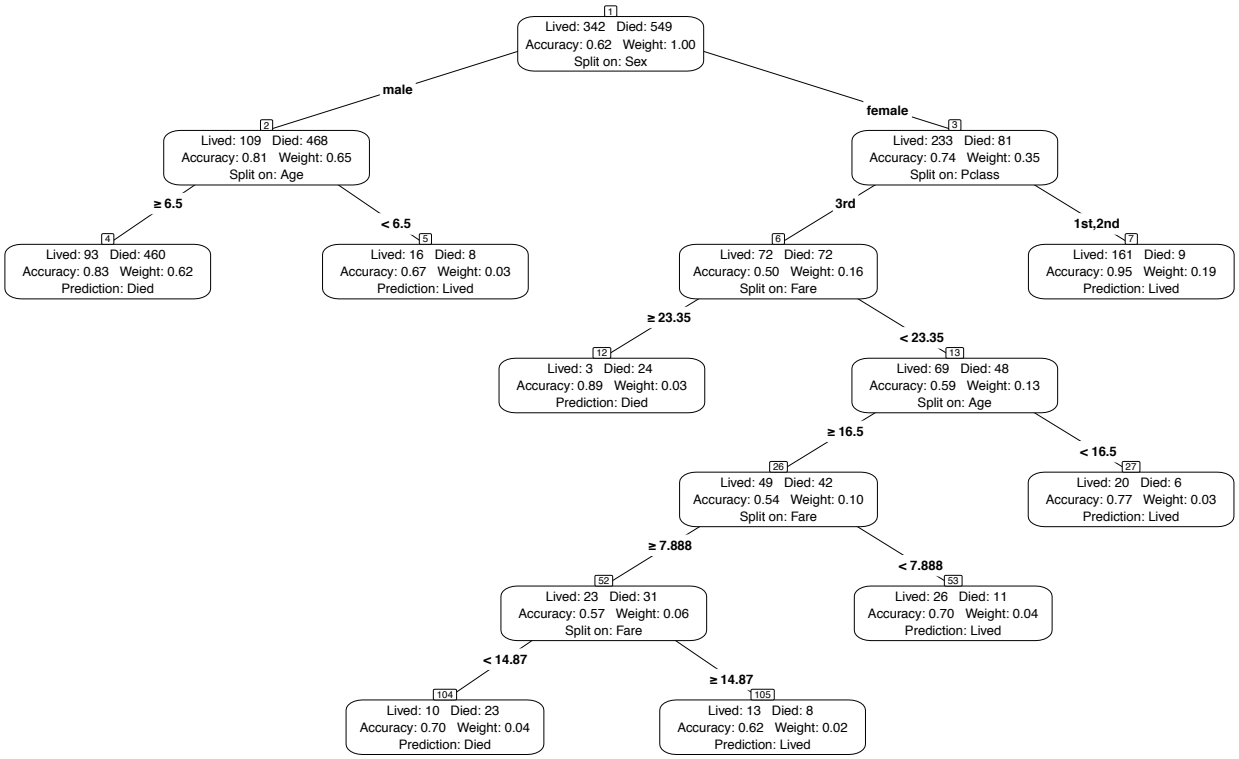


Figure 2: A decision tree for classifying survival aboard the *Titanic* based on age, sex, passenger class, and fare, fitted with the R library `rpart` under default parameters. Each node is numbered, above. Within each node is number of observations in each class (lived/died), the accuracy in that node if we were to predict the majority class, and the weight out of the overall dataset present in that node (i.e., the fraction of the total observations in the dataset, such that the weighted accuracy of the terminal nodes is the overall accuracy), and the variable on which the observations are split. If a terminal node, the final prediction is given (note that the "predictions" are in past tense, as they are about something that already happened). The edges give the values of the split variable at which the split happens.

an example of a split that violates intuition not holding: node 53 (split on the fare being less than 7.888) predicts that the passenger lived, but the majority of test observations falling into this node died. Here, intuition would indeed guide us in the right direction (although only by improving accuracy for 7% of the test data from 45% to 55%).

Fine-tuning. If the logic of our interpretations was a valid way to understand models, then we should be able to 'fine-tune' decision trees using our domain knowledge. A specific example of a place where we might do this is around age. An econometrics paper, published in several venues, modeled this same dataset not in terms of best "predicting" survival, but of studying social norms (such as "women and children first,"[28] or "noblesse oblige"[29]). Meredith Broussard's discussion of the *Titanic* dataset also notes this norm being present as early as 1852,[30] and that the *Titanic* captain explicitly made an order to put women and children in lifeboats when evacuating the ship.[31] Particularly for studying the question of women and children first, the econometrics work needed a threshold for childhood: they decided on using the (contemporary) United Nations definition of children as being 15 and younger.[32] While we can question the accuracy of using a contemporary definition, rather than

[28] Bruno S. Frey, David A. Savage, and Benno Torgler, "Behavior under Extreme Conditions: The *Titanic* Disaster," *Journal of Economic Perspectives* 25, no. 1 (2011): pp. 36, doi:10.1257/jep.25.1.209.

[29] Bruno S. Frey, David A. Savage, and Benno Torgler, "Noblesse Oblige? Determinants of Survival in a Life-or-Death Situation," *Journal of Economic Behavior & Organization* 74, nos. 1–2 (2010): 1–11, doi:10.1016/j.jebo.2010.02.005.

[30] Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2018), pp. 101.

[31] Ibid., pp. 116.

[32] Bruno S. Frey, David A. Savage, and Benno Torgler, *Surviving the Titanic Disaster: Economic, Natural and Social Determinants*, technical report 2009-03 (CREMA Gellertstrasse 18 CH - 4052 Basel: Center for Research in Economics, Management and the Arts, 2009), pp. 13, http://www.crema-research.ch/papers/2009-03.pdf.

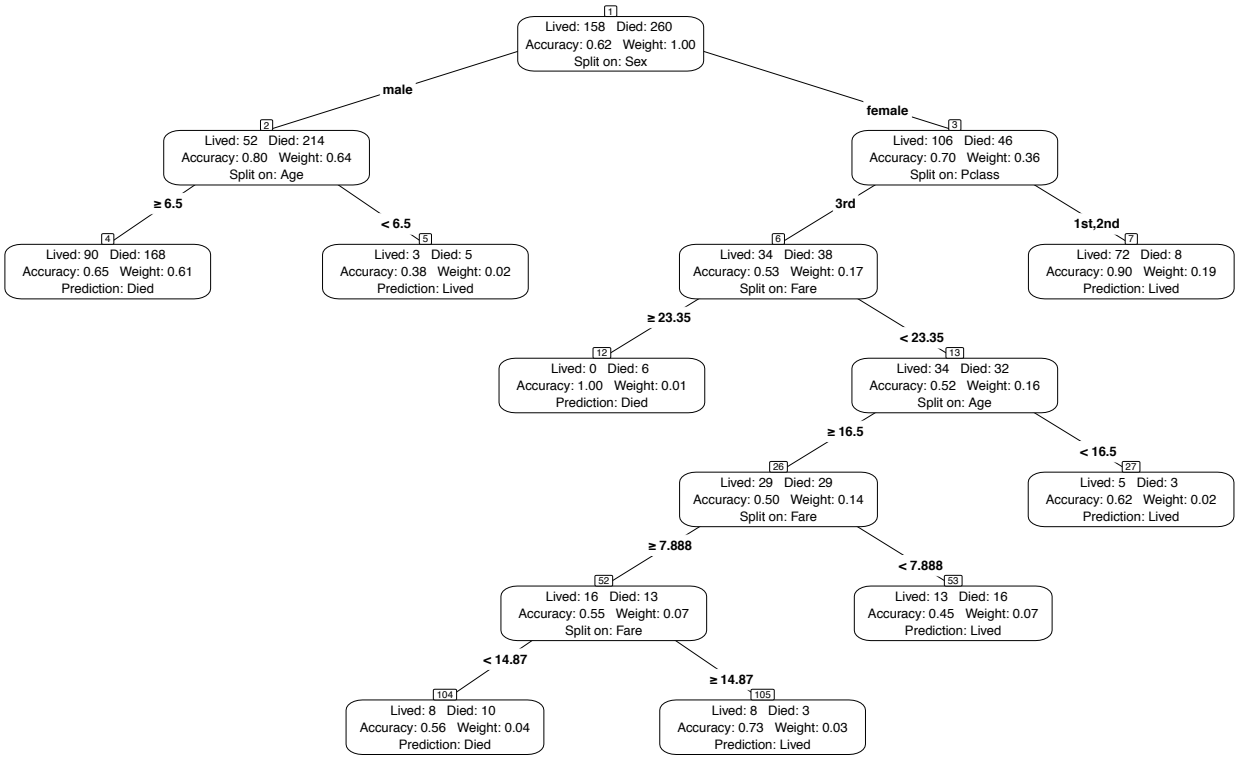


Figure 3: The breakdown of how the test data goes through the fitted decision tree. I manually traced the test data through the fitted decision tree, and manually changed the text of the decision tree to reflect this tracing.

historically investigating what the category of 'children' might have meant at that time, it is reasonable to have an *a priori* definition for the modeling rather than one discovered from data, as that would be tautological (using survival rates as a proxy for defining children, then using the label of a child to predict survival) or, in statistical terms, re-using data which harms generalizability.[33]

For the sake of argument, then, let's say we took the definition of 'children' as being 15 and younger. If we combined this with the domain knowledge of "women and children first," we would get the suggestion that the various splits on age *should* be at <16 and ≥ 16. One of the splits, at node 13, is near this; however, it includes 16-year-olds, and furthermore, the split on age at node 2 is at 6.5. What would be the effect if we used domain knowledge to dispute the value of the splits, and suggest fine-tuning to improve the tree?

This, too, degrades the empirical effectiveness of the tree. Splitting at node 2 at age 16 and above results in the left leaf node having 88 who lived and 449 who died, for an accuracy of 0.84 on 0.60 of the data. The right leaf node now has 21 who lived and 19 who died, for an accuracy of 0.52 on 0.04 of the data. The weighted accuracy of the "male" branch of the decision tree is then $0.84 \times 0.60 + 0.52 \times 0.04 = 0.52$, versus the original tree, $0.83 \times 0.62 + 0.67 \times 0.03 = 0.53$. After all, if splitting at age 16 were more empirically effective, the tree would have selected this. On the test data, the original test accuracy for the "male" branch is $0.81 \times 0.62 + 0.62 \times 0.02 = .5146$, versus the revised branching having accuracy $0.82 \times 0.59 + 0.58 \times 0.05 = .5128$. This is even less of a difference than on the training data (indeed, I had to include more significant figures), and may not be "significant" if we were to do a statistical test of the difference in accuracy (e.g., with a bootstrap), but it is a difference nonetheless in the same direction. The current state of published, quantitative domain knowledge, applied to this tree, would hurt predictive performance.

[33] Bradley Efron et al., "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation [with Comments, Rejoinder]," *Journal of the American Statistical Association* 99, no. 467 (2004): 619–642, doi:10.1198/016214504000000692.

Interpreting the wrong thing in general

I have gone through hypothetical ways in which interpretation could go awry, but in Breiman's article itself, I would argue that there is a real-world example:

> "A project I worked on in the late 1970s was the analysis of delay in criminal cases in state court systems... The dependent variable for each criminal case was the time from arraignment to the time of sentencing. All of the other information in the trial history were the predictor variables. A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, 'I knew those guys in District N were dragging their feet."'

I would say that this, and this overall project (the *sources* of delay), is an explanation/$\widehat{\beta}$ problem wrongly treated as a predictive/$\widehat{y}$ one. Was District N actually dragging its feet, i.e. were faster criminal trials within their ability but simply not done? To properly answer this question would be to provide an estimate of the effect of being in District N, after controlling for other factors. But a decision tree might use an indicator for being in District N as a proxy for the actual causal factors. The judge interpreting the decision tree, thus, arrived at an unjustified conclusion. That is, maybe the conclusion was actually accurate, but using the decision tree was not a valid way to make that determination. It is hard to read Breiman's stance on this anecdote: he notes that he was careful not to comment, but it is unclear what he is saying about the judge's interpretation. The next statement is about decision trees rating A+ on interpretability but a B on prediction, which I could be suggesting that a model that predicted better could have been interpreted causally, or perhaps that a model that predicted better but was less interpretable would not be subject to such erroneous interpretations.

This is an explicit example of misinterpretation by policymakers, but multiple other authors allude to the difficulty of client audiences confusing interpretability and causality. Doshi-Velez and Kim write, "one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern." In a famous example of asthma being robustly associated with a lessened risk of dying from pneumonia (a spurious correlation, with the underlying cause that asthmatic patients received more attention), Rich Caruana et al. write, "Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation."[34] Then, in an early version of "The Mythos of Model Interpretability", Zachary Lipton writes, "Another problem is that such an interpretation might explain the behavior of the model but not give deep insight into the causal associations in the underlying data. That's because linear models are subject to covariate effects through the process of feature selection. This can be problematic if you expect to understand anything about the underlying reality simply by a model's weights.".[35] In a 2018 update, the equivalent statement reads,

> "While the discussed desiderata, or objectives of interpretability, are diverse, they typically speak to situations where standard ML problem formulations, e.g. maximizing accuracy on a set of hold-out data for which the training data is perfectly representative, are imperfectly matched to the complex real-life tasks they are meant to solve. Consider medical research with longitudinal data. The real goal may be to discover potentially causal associations that can guide interventions, as with smoking and cancer. The optimization objective for most supervised learning models, however, is simply to minimize error, a feat that might be achieved in a purely correlative fashion."[36]

Are there really prediction policy problems?

The just use of machine learning, then, relies on the question of whether there actually are prediction policy problems, entirely apart from questions of interpretability. I would argue that such problems are frequently present around physical systems; Breiman's example of detecting chemicals on the water is an excellent one.

[34] Rich Caruana et al., "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15 (2015), 1721–1730, doi:10.1145/2783258.2788613.

[35] Zachary C. Lipton, "The Myth of Model Interpretability," *KDnuggets* 15, no. 13 (April 2015), https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html.

[36] Lipton, "The Mythos of Model Interpretability."

If we can design a model that uses correlations between the numbers of a cheap and/or fast test as a proxy for a more expensive and/or slower test that more directly measures the chemical, it is indeed a prediction[-only] problem.

On the other hand, if the 'water' in question were blood inside the human body, and the chemical was the presence of a steroid, then it would not be a prediction policy problem because the humans could react to the proxy, e.g. by finding chemicals that could trick the proxy (or even trick a gold standard test).[37]

The possibility of systems being gamed is ever-present in human systems. For example, a major argument against making public the methodology behind credit scoring is that it would make it easy to game the system. The "quantitative fallacy" or "McNamara fallacy" also relates to gaming systems; this is named after Robert McNamara's extreme reliance on unverified counts of enemy deaths as a measure of success during the Vietnam War, and the incentives created for those relaying the numbers to lie to him. The possibility of gaming, along with intervening on a covariate that influences predicted outcomes, are ways in which problems are causal and not prediction policy ones.

I have already argued that the example of delays in court cases provided by Breiman is not a prediction policy problem, based on the actual goal being finding the causal sources of delays. Consider, then, the four examples given by Kleinberg et al. of prediction policy problems:

> "Other illustrative examples include: (i) in education, predicting which teacher will have the greatest value added (Rockoff et al. 2011); (ii) in labor market policy, predicting unemployment spell length to help workers decide on savings rates and job search strategies; (iii) in regulation, targeting health inspections (Kang et al. 2013); (iv) in social policy, predicting highest risk youth for targeting interventions (Chandler, Levitt, and List 2011); and (v) in the finance sector, lenders identifying the underlying credit-worthiness of potential borrowers."

I would argue that none of these are prediction policy problems.

(i) The phrasing of value-added (critiques of the value-added model[38] aside) obscures that it is only a proxy for the real underlying goal: if a particular teacher will *cause* a student or a group of students to, on the whole, learn better.

(ii) Saving rates and job search strategies will influence unemployment spell length. Indeed, the goal of decreasing unemployment spell length is a causal one.

(iii) First note that the cited paper is subject of a recent critique and re-analysis of the data,[39] suggesting that the results were not accurate and that online reviews are not sufficient for predicting health violations. Second, note that there is a game-theoretic aspect to inspections: the goal of health inspections is not to catch violators, but to have establishments have high health standards, for which inspections and penalties are a tool. It is likely that only a fraction of violators will ever be caught, but if enough establishments calculate that the costs of penalties, combined with the chance of being caught, is too high to risk committing violations, then the overall goal is achieved.[40] But assuming that online reviews were a valid data source, and predicting violations a valid objective, there would still be the problem of relying on a proxy that could be gamed. In this case, confusing the signal by leaving fake reviews on competitors' pages, and drowning out negative reviews on an establishment's own page, would push the signal into the cat-and-mouse game of detecting fake reviews.

(iv) Interventions will affect the variables that go into the model. And, the purpose is ultimately to reduce the frequency of adverse outcomes, for which we want to know not who has the highest risk, but the *highest risk that an intervention will lower*.

(v) The *underlying* creditworthiness of potential borrowers is a causal question. The use of proxies, and the predictive modeling for determining credit scores that are acted on as proxies for creditworthiness are *a*

[37] And insofar as any systems with which humans interact are never purely natural ones, as in the case of water contamination that might be caused by industrial waste disposal, natural resource extraction, or infrastructural resource deprivation.
[38] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York, NY: Crown, 2016).
[39] Daniel E. Ho and Kristen M. Altenburger, "Is Yelp Actually Cleaning Up the Restaurant Industry? A Re-Analysis on the Relative Usefulness of Consumer Reviews," in *Proceedings of the 2019 Web Conference* (2019).
[40] This insight is due to Rayid Ghani.

major source of injustice in the world today.[41] This is perhaps the best example of a causal question, treated as a prediction policy problem, leading to injustice.

Kleinberg et al. also refer to then-unpublished work of theirs, subsequently published in 2017,[42] arguing that bail decisions are a prediction policy problem. Quite a bit of machinery is used in making this argument: specifically, I see the paper as cleverly using judge leniency as an instrumental variable to try and get an unbiased estimate of the generalization error from the machine learning models—after all, test error is an estimator that can be biased by data missing not at random (in this case, the counterfactuals are the missing data), and hence causal estimation techniques can be used to try to get unbiased estimates. There is a certain irony in using causal estimation of one quantity in an argument that causal estimation of another is unnecessary, but I think that deciding if something is a prediction policy problem deserves just as much care as goes into causal estimation and so this approach should be more widely adopted.

This is perhaps the most solid theoretical case for something being a prediction policy problem. In theory, the system cannot be gamed: after all, if gaming the system equates to not getting arrested or charged, then the underlying goal is met… assuming arrests and charges are fair in practice. But therein lies the rub. The connection between crime and getting arrested and/or getting charged depends on whether the accused are members of overpoliced populations or underpoliced populations. The discrepancy between actual *crime* (not even going into how acts are constructed as crime) and heavy biases in available proxies, either arrest data or crime report data, is a perennial problem of criminology and certainly of anything predictive.[43]

There is also the problem that finding good instrumental variables is a matter of cleverness and luck,[44] and so the ability to rigorously establish a problem as a prediction policy one might, in general, be unavailable.

Conclusion

I have tried to focus on the issue of model interpretability, and not of bias in data. But, as shown in the final example, the role of data in just usage might outweigh all other factors, and so perhaps should remain the primary consideration. Furthermore, while focused on the specific topic of interpretability, I sought to open up the focus to what I believe is the core issue: that of the possibility of prediction policy problems. Insofar as interpretability is not causality, the logic of our interpretation does not correspond with the logic of modeling, failing to give us insight into why, when, and how a model works or doesn't work and what we might do about it. The logic of modeling, wherein a model can "predict" well without reflecting causal processes (with prediction defined in a particular way), is something we have to contend with through the question of prediction policy problems.

I do want to acknowledge that it is quite easy to use machine learning in just ways in pubic policy, so long as those uses relate to physical systems that affect people rather than people's behavior, like Breiman's example of water contamination. I draw on several examples from the Data Science for Social Good program, in which I was a 2017 fellow. In some cases, potential feedback loops might be so large, and second-order effects from gaming the system so distant, that machine learning would be justified in practice if not in theory. For example, water main breaks[45] are not just a function of nature, but of decisions that go into the built environment and urban infrastructure. But, unless machine learning were deployed on a large scale in ways that those who have control over infrastructure might try to game, it is unlikely that causal factors would be a critical consideration. Similarly, the decisions that went into using lead paint are in the past, and so only the cleanup of lead paint is an active system today. Again, which houses' lead paint is ignored depends on structural racism and economic

[41] Martha Poon, "Scorecards as Devices for Consumer Credit: The Case of Fair, Isaac & Company Incorporated," *The Sociological Review* 55, no. 2 supplement (2007): 284–306, doi:10.1111/j.1467-954X.2007.00740.x; Josh Lauer, *Creditworthy: A History of Consumer Surveillance and Financial Identity in America* (New York: Columbia University Press, 2017); Marion Fourcade and Kieran Healy, "Classification Situations: Life-Chances in the Neoliberal Era," *Accounting, Organizations and Society* 38, no. 8 (2013): 559–572, doi:10.1016/j.aos.2013.11.002.

[42] Jon Kleinberg et al., "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics* 133, no. 1 (August 2017): 237–293, doi:10.1093/qje/qjx032.

[43] Kristian Lum and Patrick Ball, *Estimating Undocumented Homicides with Two Lists and List Dependence*, technical report (Human Rights Data Analysis Group, 2015), https://hrdag.org/publications/estimating-undocumented-homicides-with-two-lists-and-list-dependence/; Kristian Lum and William Isaac, "To predict and serve?," *Significance* 13, no. 5 (2016): 14–19, doi:10.1111/j.1740-9713.2016.00960.x.

[44] Andrew Gelman, "A Statistician's Perspective on *Mostly Harmless Econometrics: An Empiricist's Companion*, by Joshua D. Angrist and Jörn-Steffen Pischke," *Stata Journal* 9 (2 2009): 315–320, http://www.stata-journal.com/article.html?article=gn0046.

[45] Avishek Kumar et al., "Using Machine Learning to Assess the Risk of and Prevent Water Main Breaks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18 (2018), 472–480, doi:10.1145/3219819.3219835.

disparities, but machine learning may be appropriate for short-term action.[46] Then, treating a causal question as a prediction policy problems in ways that apply any potential injustices to those in positions of power, such as predicting police misconduct and "adverse interactions,"[47] or mapping out resource distribution within US congressional actions,[48] is far less worrisome.

If we determine that we have a prediction policy problem and that the use of machine learning is therefore just, interpretability may still be an important rhetorical tool for convincing people to adopt the tool. However, I fear this does a disservice to domain practitioners deciding whether or not to use a predictive model; the case for using the model should be made on the basis of the existence and nature of prediction policy problems, of the current case being one, of the data being acceptable, and of due diligence being done in careful input/output robustness checks for edge cases and unit tests for implementation correctness. The case should not be made on the basis of what I have demonstrated is ultimately an auxiliary concern, interpretability.

References

Abbott, Andrew. "Transcending General Linear Reality." *Sociological Theory* 6, no. 2 (1988): 169–186. doi:10.2307/202114.

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media." In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 492–499. WI-IAT '10. 2010. doi:10.1109/WI-IAT.2010.63.

Breiman, Leo. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3 (2001): 199–231. doi:10.1214/ss/1009213726.

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. KDD '15. 2015. doi:10.1145/2783258.2788613.

Cox, D. R. "Role of Models in Statistical Analysis." *Statistical Science* 5, no. 2 (1990): 169–174. doi:10.1214/ss/1177012165.

Doshi-Velez, Finale, and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608. https://arxiv.org/abs/1702.08608.

Efron, Bradley, Prabir Burman, L. Denby, J. M. Landwehr, C. L. Mallows, Xiaotong Shen, Hsin-Cheng Huang, Jianming Ye, Jimmy Ye, and Chunming Zhang. "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation [with Comments, Rejoinder]." *Journal of the American Statistical Association* 99, no. 467 (2004): 619–642. doi:10.1198/016214504000000692.

Efron, Bradley, and Carl Morris. "Stein's Paradox in Statistics." *Scientific American* 236, no. 5 (1977): 119–127. doi:10.1038/scientificamerican0577-119.

Fourcade, Marion, and Kieran Healy. "Classification Situations: Life-Chances in the Neoliberal Era." *Accounting, Organizations and Society* 38, no. 8 (2013): 559–572. doi:10.1016/j.aos.2013.11.002.

Frey, Bruno S., David A. Savage, and Benno Torgler. "Behavior under Extreme Conditions: The *Titanic* Disaster." *Journal of Economic Perspectives* 25, no. 1 (2011): 209–222. doi:10.1257/jep.25.1.209.

———. "Noblesse Oblige? Determinants of Survival in a Life-or-Death Situation." *Journal of Economic Behavior & Organization* 74, nos. 1–2 (2010): 1–11. doi:10.1016/j.jebo.2010.02.005.

[46] Eric Potash et al., "Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15 (2015), 2039–2047, doi:10.1145/2783258.2788629.

[47] Jennifer Helsby et al., "Early Intervention Systems: Predicting Adverse Interactions Between Police and the Public," *Criminal Justice Policy Review* 29, no. 2 (2018): 190–209, doi:10.1177/0887403417695380.

[48] Ellery Wulczyn et al., "Identifying Earmarks in Congressional Bills," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (2016), 303–311, doi:10.1145/2939672.2939711.

Frey, Bruno S., David A. Savage, and Benno Torgler. *Surviving the Titanic Disaster: Economic, Natural and Social Determinants*. Technical report 2009-03. CREMA Gellertstrasse 18 CH - 4052 Basel: Center for Research in Economics, Management and the Arts, 2009. http://www.crema-research.ch/papers/2009-03.pdf.

Friedman, Milton. *Essays in Positive Economics*. University of Chicago Press, 1953.

Gayo-Avello, Daniel. *'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper' — A Balanced Survey on Election Prediction using Twitter Data*. arXiv:1204.6441, May 2012. http://arxiv.org/abs/1204.6441v1.

———. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16, no. 6 (2012): 91–94. doi:10.1109/MIC.2012.137.

Gelman, Andrew. "A Statistician's Perspective on *Mostly Harmless Econometrics: An Empiricist's Companion*, by Joshua D. Angrist and Jörn-Steffen Pischke." *Stata Journal* 9 (2 2009): 315–320. http://www.stata-journal.com/article.html?article=gn0046.

Helsby, Jennifer, Samuel Carton, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Andrea Navarrete, Klaus Ackermann, et al. "Early Intervention Systems: Predicting Adverse Interactions Between Police and the Public." *Criminal Justice Policy Review* 29, no. 2 (2018): 190–209. doi:10.1177/0887403417695380.

Ho, Daniel E., and Kristen M. Altenburger. "Is Yelp Actually Cleaning Up the Restaurant Industry? A Re-Analysis on the Relative Usefulness of Consumer Reviews." In *Proceedings of the 2019 Web Conference*. 2019.

Jones, Matthew L. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673–684. doi:10.1525/hsns.2018.48.5.673.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133, no. 1 (August 2017): 237–293. doi:10.1093/qje/qjx032.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. "Prediction Policy Problems." *American Economic Review* 105, no. 5 (2015): 491–95. doi:10.1257/aer.p20151023.

Kumar, Avishek, Syed Ali Asad Rizvi, Benjamin Brooks, R. Ali Vanderveld, Kevin H. Wilson, Chad Kenney, Sam Edelstein, et al. "Using Machine Learning to Assess the Risk of and Prevent Water Main Breaks." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 472–480. KDD '18. 2018. doi:10.1145/3219819.3219835.

Lanius, Candice. *Fact Check: Your Demand for Statistical Proof is Racist*. *Cyborgology* [blog], January 2015. https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/.

Lauer, Josh. *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*. New York: Columbia University Press, 2017.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343, no. 6176 (2014): 1203–1205. doi:10.1126/science.1248506.

Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model." *The Annals of Applied Statistics* 9, no. 3 (September 2015): 1350–1371. doi:10.1214/15-AOAS848.

Lipton, Zachary C. "The Myth of Model Interpretability." *KDnuggets* 15, no. 13 (April 2015). https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html.

———. "The Mythos of Model Interpretability." *ACM Queue* 16, no. 3 (June 2018): 31–57. doi:10.1145/3236386.3241340.

Lum, Kristian, and Patrick Ball. *Estimating Undocumented Homicides with Two Lists and List Dependence*. Technical report. Human Rights Data Analysis Group, 2015. https://hrdag.org/publications/estimating-undocumented-homicides-with-two-lists-and-list-dependence/.

Lum, Kristian, and William Isaac. "To predict and serve?" *Significance* 13, no. 5 (2016): 14–19. doi:10.1111/j.1740-9713.2016.00960.x.

Meinshausen, Nicolai, and Peter Bühlmann. "Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, no. 4 (2010): 417–473. doi:10.1111/j.1467-9868.2010.00740.x.

Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106. doi:10.1257/jep.31.2.87.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown, 2016.

Poon, Martha. "Scorecards as Devices for Consumer Credit: The Case of Fair, Isaac & Company Incorporated." *The Sociological Review* 55, no. 2 supplement (2007): 284–306. doi:10.1111/j.1467-954X.2007.00740.x.

Potash, Eric, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. "Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2039–2047. KDD '15. 2015. doi:10.1145/2783258.2788629.

Rose, Todd. *The End of Average: How We Succeed in a World that Values Sameness*. New York: HarperCollins Publishers, 2015.

Rudin, Cynthia. "Please Stop Explaining Black Box Models for High Stakes Decisions." In *The NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning*. 2018 Conference on Neural Information Processing Systems (NeurIPS 2018). 2018. https://arxiv.org/abs/1811.10154.

Shmueli, Galit. "To Explain or to Predict?" *Statistical Science* 25, no. 3 (2010): 289–310. doi:10.1214/10-STS330.

Spirtes, Peter, and Kun Zhang. "Causal Discovery and Inference: Concepts and Recent Methodological Advances." *Applied Informatics* 3, no. 3 (2016): 1–28. doi:10.1186/s40535-016-0018-x.

Swendsen, Robert H. *An Introduction to Statistical Mechanics and Thermodynamics*. Oxford University Press, 2012. doi:10.1093/acprof:oso/9780199646944.001.0001.

Weisberg, Sanford. *Applied Linear Regression*. 3rd ed. Wiley, 2005.

Wu, Shaohua, T. J. Harris, and K. B. McAuley. "The Use of Simplified or Misspecified Models: Linear Case." *The Canadian Journal of Chemical Engineering* 85, no. 4 (2007): 386–398. doi:10.1002/cjce.5450850401.

Wulczyn, Ellery, Madian Khabsa, Vrushank Vora, Matthew Heston, Joe Walsh, Christopher Berry, and Rayid Ghani. "Identifying Earmarks in Congressional Bills." In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 303–311. KDD '16. 2016. doi:10.1145/2939672.2939711.