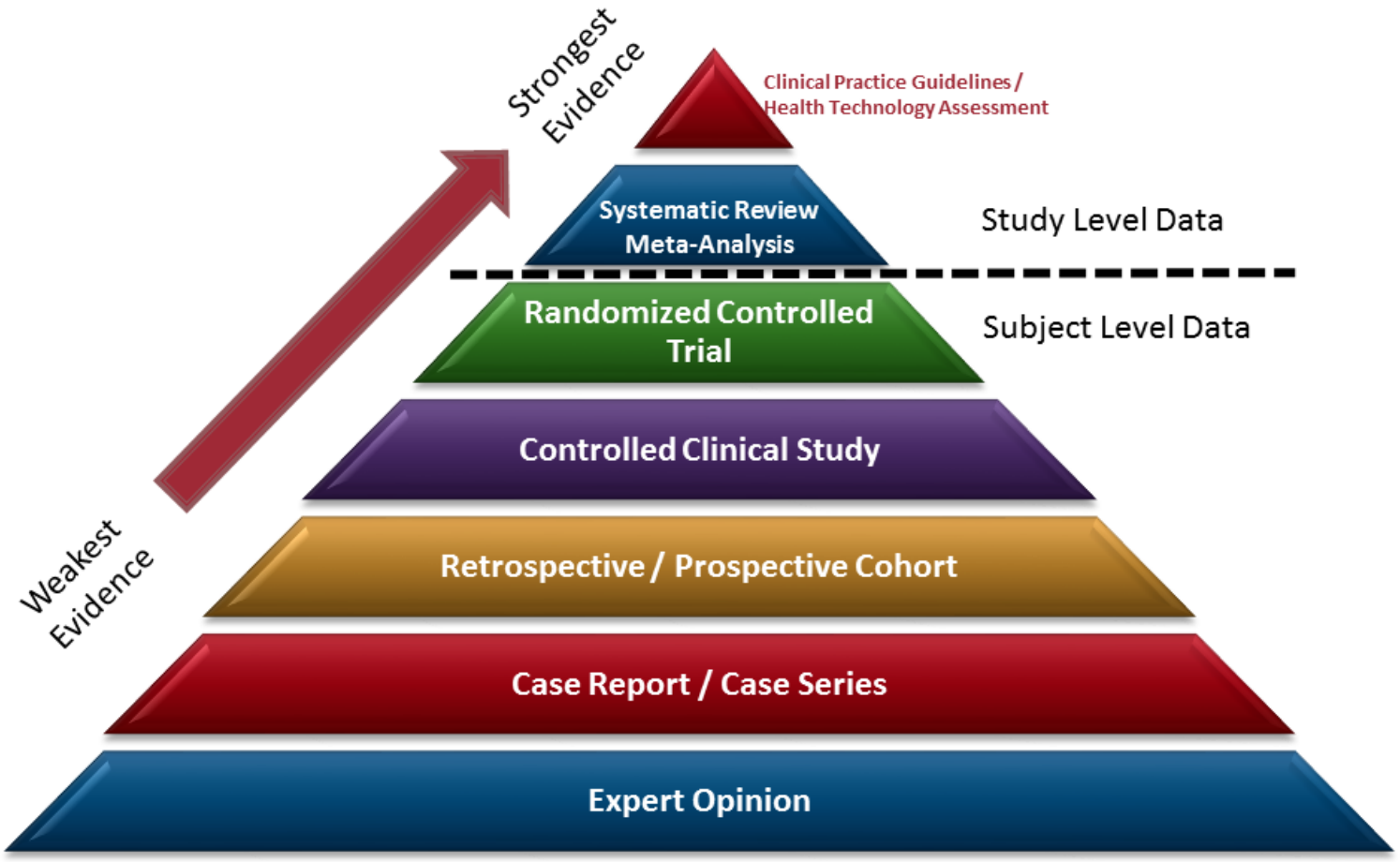
Evidence-based decision making

Rui Mata, FS 2022



https://en.wikipedia.org/wiki/Levels_of_evidence

Policy capturing in practice

https://evidencebaseddm.formr.org

1

Perspectives on Psychological Science
2017, Vol. 12(6) 1100–1122

DOI: 10.1177/1745691617693393
www.psychologicalscience.org/PPS
SAGE

Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning

Tal Yarkoni and Jacob Westfall
University of Texas at Austin

Abstract

Psychology has historically been concerned, first and foremost, with explaining the causal mechanisms that give rise to behavior. Randomized, tightly controlled experiments are enshrined as the gold standard of psychological research, and there are endless investigations of the various mediating and moderating variables that govern various behaviors. We argue that psychology's near-total focus on explaining the causes of behavior has led much of the field to be populated by research programs that provide intricate theories of psychological mechanism but that have little (or unknown) ability to predict future behaviors with any appreciable accuracy. We propose that principles and techniques from the field of machine learning can help psychology become a more predictive science. We review some of the fundamental concepts and tools of machine learning and point out examples where these concepts have been used to conduct interesting and important psychological research that focuses on predictive research questions. We suggest that an increased focus on prediction, rather than explanation, can ultimately lead us to greater understanding of behavior.

Yarkoni & Westfall

"research programs that provide intricate theories of psychological mechanism but that have little (or unknown) ability to predict future behaviors"

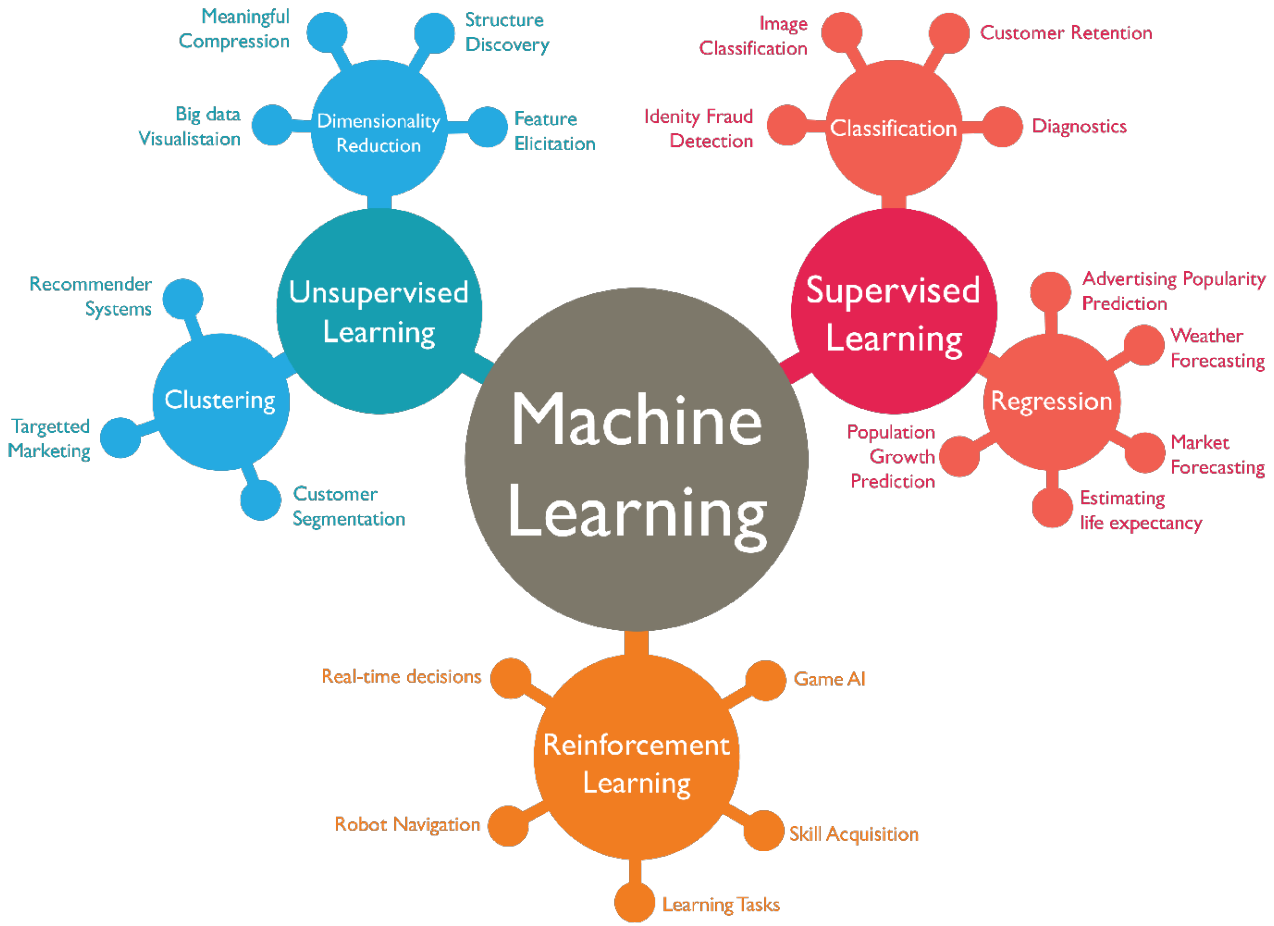
2

O'Neil

".. algorithms are opinions embedded in code"!

Actuarial judgment today = machine learning…

"Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?"



Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. http://doi.org/10.1126/science.aaa8415

Limitations of actuarial judgment: Overfitting

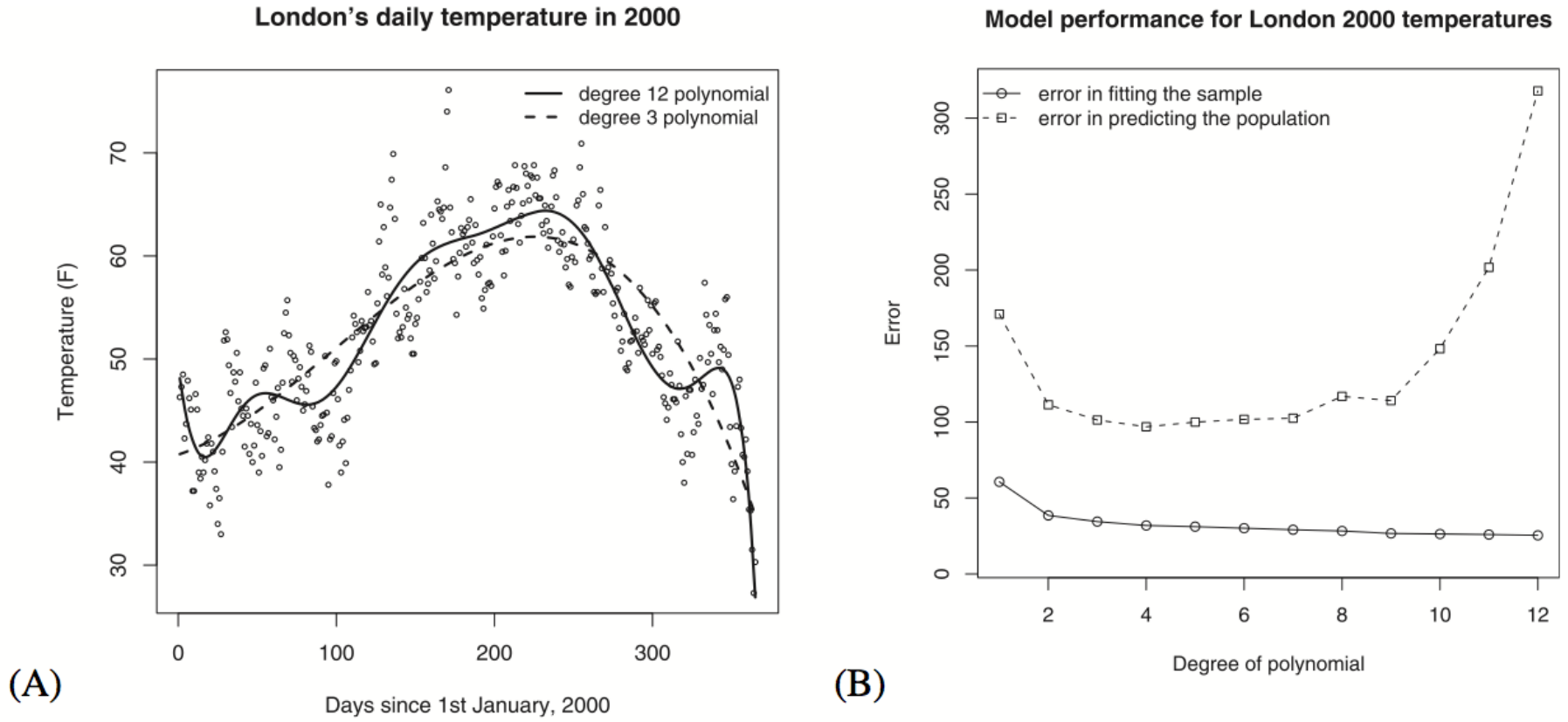


Fig. 3. Plot (A) shows London's mean daily temperature in 2000, along with two polynomial models fitted with using the least squares method. The first is a degree-3 polynomial, and the second is a degree-12 polynomial. Plot (B) shows the mean error in fitting samples of 30 observations and the mean prediction error of the same models, both as a function of degree of polynomial.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1(1)*, 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Bias-variance dilemma

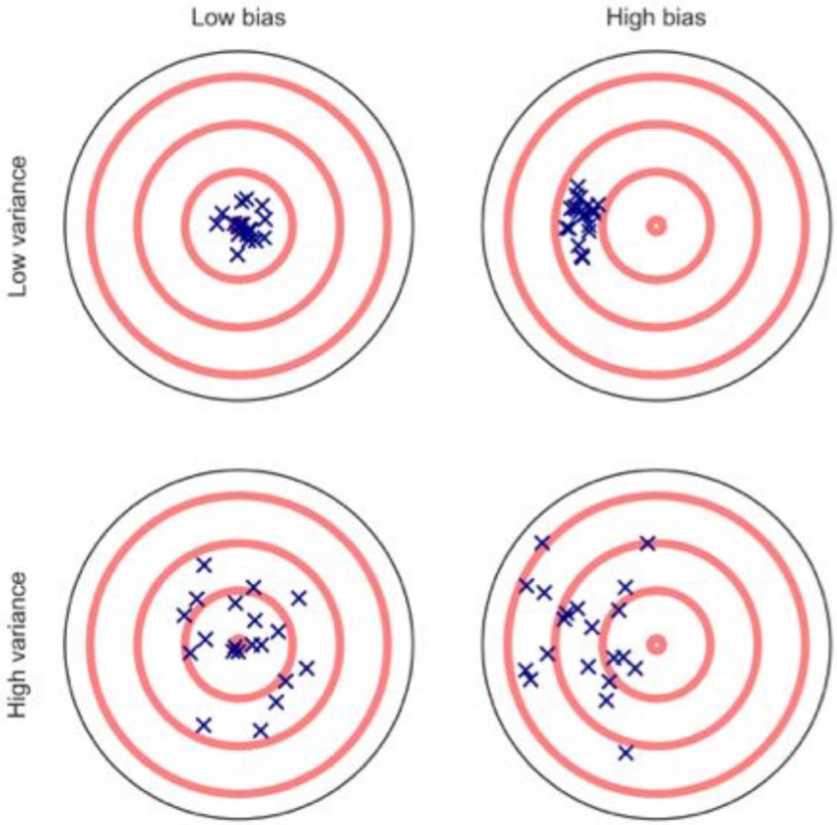


Figure 2. An estimator's predictions can deviate from the desired outcome (or true scores) in two ways. First, the predictions may display a systematic tendency (or *bias*) to deviate from the central tendency of the true scores (compare right panels with left panels). Second, the predictions may show a high degree of *variance*, or imprecision (compare bottom panels with top panels).

A dilemma exists because bias and variance are not independent: Methods for reducing variance tend to increase bias, and methods for reducing bias tend to increase variance (Brighton & Gigerenzer, 2015)

Yarkoni, T. & Westafall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research, 68*(8), 1772–1784. http://doi.org/10.1016/j.jbusres.2015.01.061

How can one avoid overfitting?

- **Regularization**: Use regularisation in the estimation of model parameters (e.g. ridge or lasso regression)

$$\textit{Regularized loss} = \sum_{i}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j}^{p} f(\beta_j))$$

lasso regression: |β| (βs are reduced in size, resulting in automatic feature selection, with some βs becoming zero)

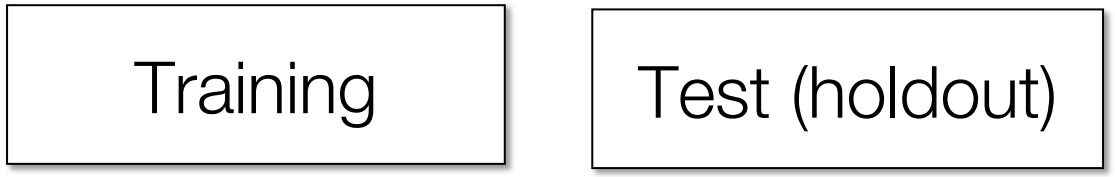
ridge regression: β^2 (squaring reduces the size of extreme βs).

elastic net: |β| + β^2 (the best of both worlds)

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. ***Perspectives on Psychological Science, 12***(6), 1100-1122.

How can one avoid overfitting?

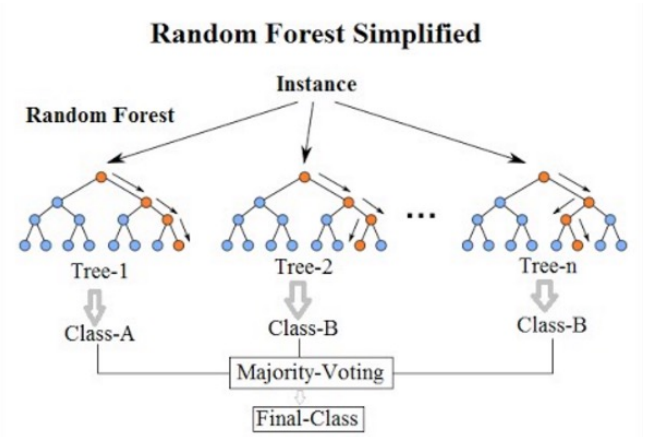
- **Cross-validation**: Compare models in how well they predict out-of-sample (cross-validation/prediction)



Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. ***Perspectives on Psychological Science, 12***(6), 1100-1122.

How can one avoid overfitting?

- **Averaging**: Use modeling approaches that integrate averaging (e.g., random forest) or use different models and combine their predictions



For example, one may have three predictive models, one based on a random forest, leading to predictions $\hat{Y}_i^{\text{RF}}$; one based on a neural net, with predictions $\hat{Y}_i^{\text{NN}}$; and one based on a linear model estimated by LASSO, leading to $\hat{Y}_i^{\text{LASSO}}$. Then, using a test sample, one can choose weights p^{rf}, p^{nn}, and p^{lasso} by minimizing the sum of squared residuals in the test sample:

$$(\hat{p}^{\text{rf}}, \hat{p}^{\text{nn}}, \hat{p}^{\text{lasso}}) = \arg \min_{p^{\text{rf}}, p^{\text{nn}}, p^{\text{lasso}}} \sum_{i=1}^{N^{\text{test}}} \left(Y_i - p^{\text{rf}} \hat{Y}_i^{\text{RF}} - p^{\text{nn}} \hat{Y}_i^{\text{NN}} - p^{\text{lasso}} \hat{Y}_i^{\text{LASSO}} \right)^2,$$

$$\text{subject to } p^{\text{rf}} + p^{\text{nn}} + p^{\text{lasso}} = 1 \quad \text{and} \quad p^{\text{rf}}, p^{\text{nn}}, p^{\text{lasso}} \geq 0.$$

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. Annual Review of Economics, 11(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Complex algorithms may lead to only small increases in performance!

Table 1

Performance of Different Algorithms in Predicting House Values

	Prediction performance (R^2)		*Relative improvement over ordinary least squares by quintile of house value*				
Method	*Training sample*	*Hold-out sample*	1st	2nd	3rd	4th	5th
Ordinary least squares	47.3%	41.7% [39.7%, 43.7%]	–	–	–	–	–
Regression tree tuned by depth	39.6%	34.5% [32.6%, 36.5%]	–11.5%	10.8%	6.4%	–14.6%	–31.8%
LASSO	46.0%	43.3% [41.5%, 45.2%]	1.3%	11.9%	13.1%	10.1%	–1.9%
Random forest	85.1%	45.5% [43.6%, 47.5%]	3.5%	23.6%	27.0%	17.8%	–0.5%
Ensemble	80.4%	45.9% [44.0%, 47.9%]	4.5%	16.0%	17.9%	14.2%	7.6%

Note: The dependent variable is the log-dollar house value of owner-occupied units in the 2011 American Housing Survey from 150 covariates including unit characteristics and quality measures. All algorithms are fitted on the same, randomly drawn training sample of 10,000 units and evaluated on the 41,808 remaining held-out units. The numbers in brackets in the hold-out sample column are 95 percent bootstrap confidence intervals for hold-out prediction performance, and represent measurement variation for a fixed prediction function. For this illustration, we do not use sampling weights. Details are provided in the online Appendix at http://e-jep.org.

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

Complex algorithms do not guarantee predictability!

"How predictable are life trajectories? We investigated this question with a scientific mass collaboration using the common task method; 160 teams built predictive models for six life outcomes using data from the Fragile Families and Child Wellbeing Study, a high-quality birth cohort study. Despite using a rich dataset and applying machine-learning methods optimized for prediction, the best predictions were not very accurate and were only slightly better than those from a simple benchmark model. Within each outcome, prediction error was strongly associated with the family being predicted and weakly associated with the technique used to generate the prediction. Overall, these results suggest practical limits to the predictability of life outcomes in some settings and illustrate the value of mass collaborations in the social sciences."

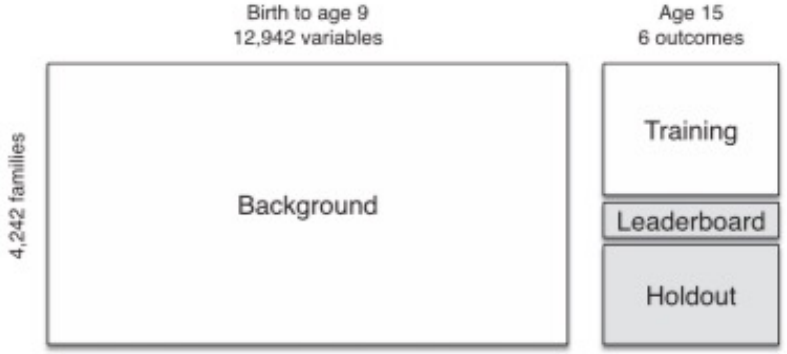
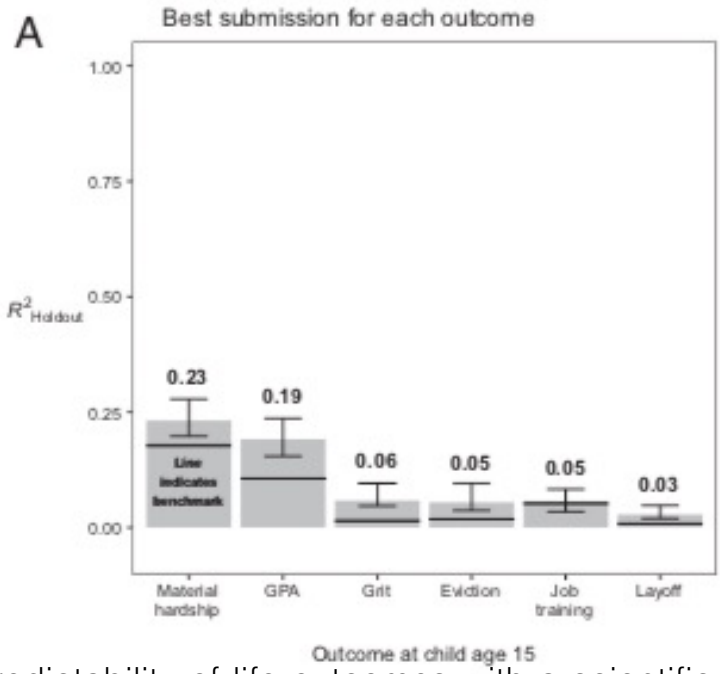


Fig. 2. Datasets in the Fragile Families Challenge. During the Fragile Families Challenge, participants used the background data (measured from child's birth to age 9 y) and the training data (measured at child age 15 y) to predict the holdout data as accurately as possible. While the Fragile Families Challenge was underway, participants could assess the accuracy of their predictions in the leaderboard data. At the end of the Fragile Families Challenge, we assessed the accuracy of the predictions in the holdout data.



Salganik, M. J., et al. 2020. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. PNAS, http://doi.org/10.7910/DVN/CXSECU

1

Perspectives on Psychological Science
2017, Vol. 12(6) 1100–1122

DOI: 10.1177/1745691617693393
www.psychologicalscience.org/PPS
SAGE

Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning

Tal Yarkoni and Jacob Westfall
University of Texas at Austin

Abstract
Psychology has historically been concerned, first and foremost, with explaining the causal mechanisms that give rise to behavior. Randomized, tightly controlled experiments are enshrined as the gold standard of psychological research, and there are endless investigations of the various mediating and moderating variables that govern various behaviors. We argue that psychology's near-total focus on explaining the causes of behavior has led much of the field to be populated by research programs that provide intricate theories of psychological mechanism but that have little (or unknown) ability to predict future behaviors with any appreciable accuracy. We propose that principles and techniques from the field of machine learning can help psychology become a more predictive science. We review some of the fundamental concepts and tools of machine learning and point out examples where these concepts have been used to conduct interesting and important psychological research that focuses on predictive research questions. We suggest that an increased focus on prediction, rather than explanation, can ultimately lead us to greater understanding of behavior.

Yarkoni & Westfall

"research programs that provide intricate theories of psychological mechanism but that have little (or unknown) ability to predict future behaviors"

2

O'Neil

".. algorithms are opinions embedded in code"!

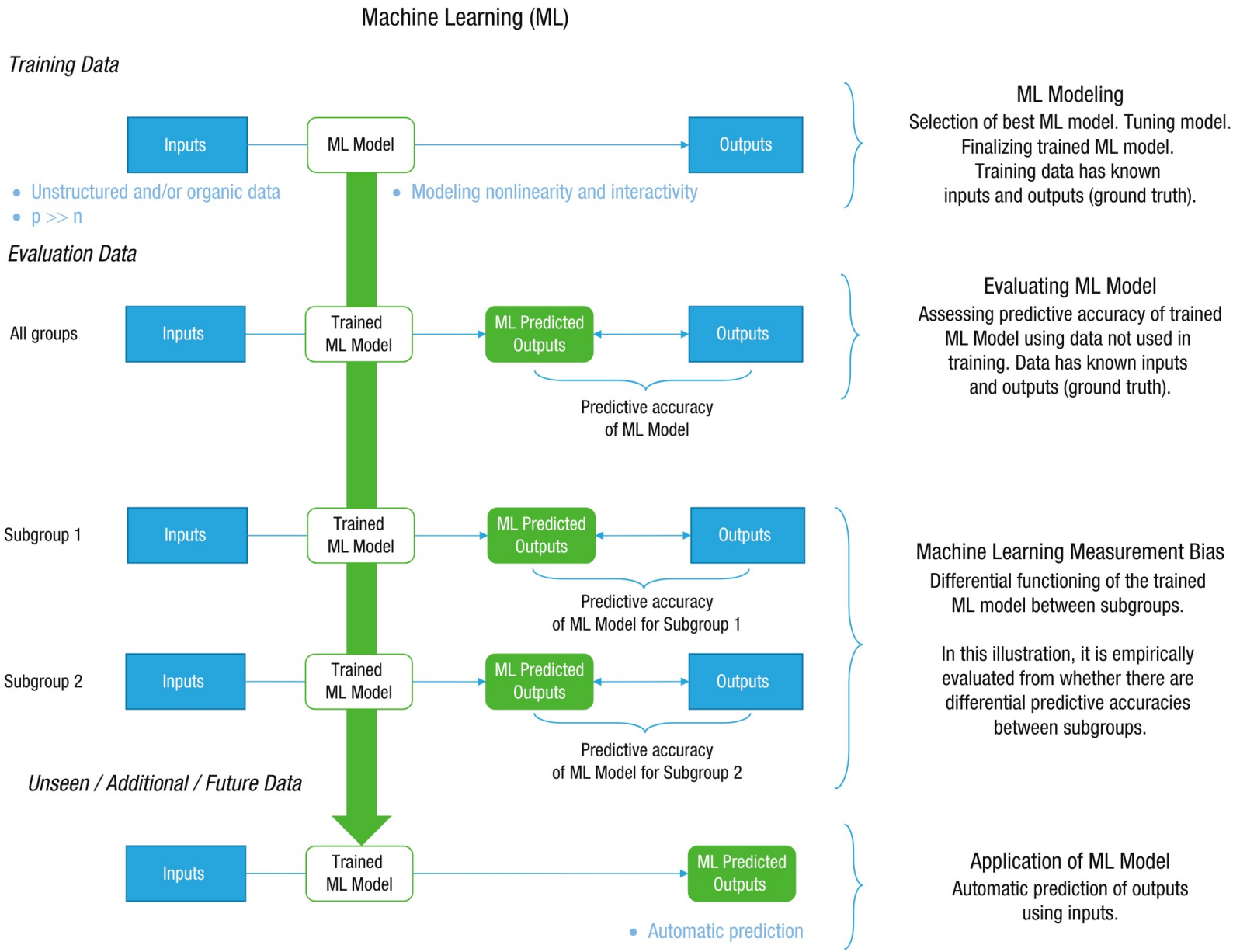


Fig. 1. Simplified process of machine-learning modeling.

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science, 5* (1),1–30.

Table 2. Comparison Between Traditional Measurement Bias and Machine-Learning Measurement Bias

Key issues	Measurement bias	Machine-learning measurement bias
Types of scores that are relevant	Predicted observed scores typically derived from CFA or IRT models of psychological assessments	ML-model-predicted scores that are predictions produced by the ML model
	Latent scores typically derived from CFA or IRT models of psychological assessments	Ground-truth scores typically in the form of observed scores from psychological assessments
Defining bias	Defined as a differential relationship between the latent score and the predicted observed score or differential functioning of the measurement tool across subgroups	Defined as differential functioning of the trained ML model between subgroups
	One empirical manifestation is that the measurement model produces different scores for individuals belonging to different subgroups despite the same latent-score level.	One empirical manifestation is when a trained ML model produces different predicted score levels for individuals belonging to different subgroups despite them having the same ground-truth level for the underlying construct of interest.
	Another empirical manifestation is that the same measurement model does not fit subgroups equally well.	Another empirical manifestation is that the ML model yields differential predictive accuracies across the subgroups.
Empirical manifestation of bias	Most typically assessed via differences in model-data fit: (a) differences in CFA fit between subgroups and (b) item-level subgroup differences in IRT fit	Ground-truth score level: different ML-predicted score levels between subgroups when subgroups have the same ground-truth score level
	Can also be assessed based on different model-predicted scores for the same latent-trait level	Ground-truth distribution level: different ML-predicted score distributions (e.g., means, variances) between subgroups for equivalent subgroup ground-truth distributions or the discrepancy between ML-predicted subgroup score distributions and ground-truth subgroup score distributions
		Predictive accuracy: different ML-model prediction accuracies (i.e., nonequivalent convergence of predicted scores and ground-truth scores) between subgroups
		Modeling ground-truth score and ML-predicted scores: applying (regression) models between ground-truth scores and ML-predicted scores and finding that significantly different models are needed between subgroups

Note: CFA = confirmatory factor analysis; IRT = item response theory; ML = machine learning.

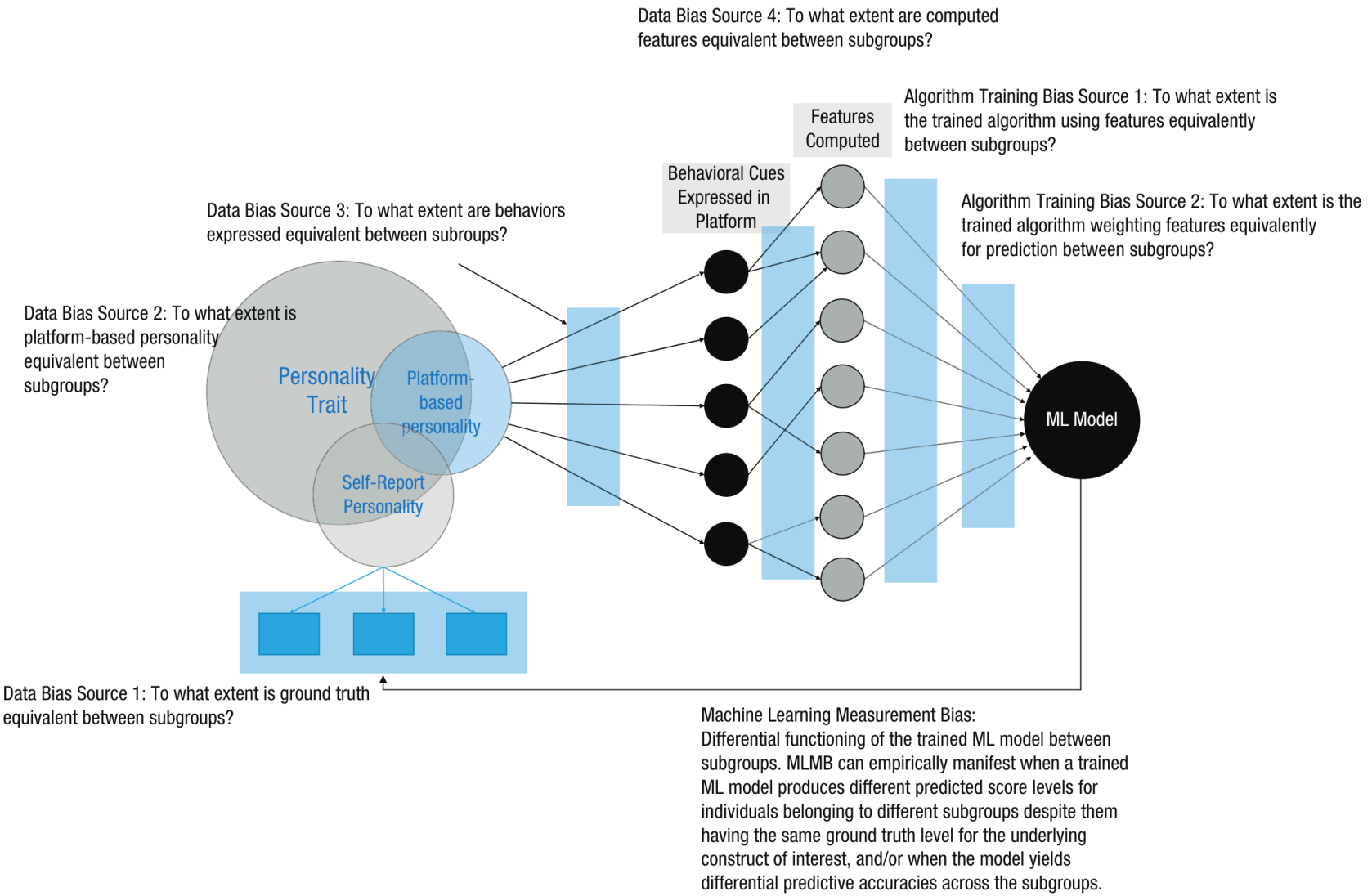
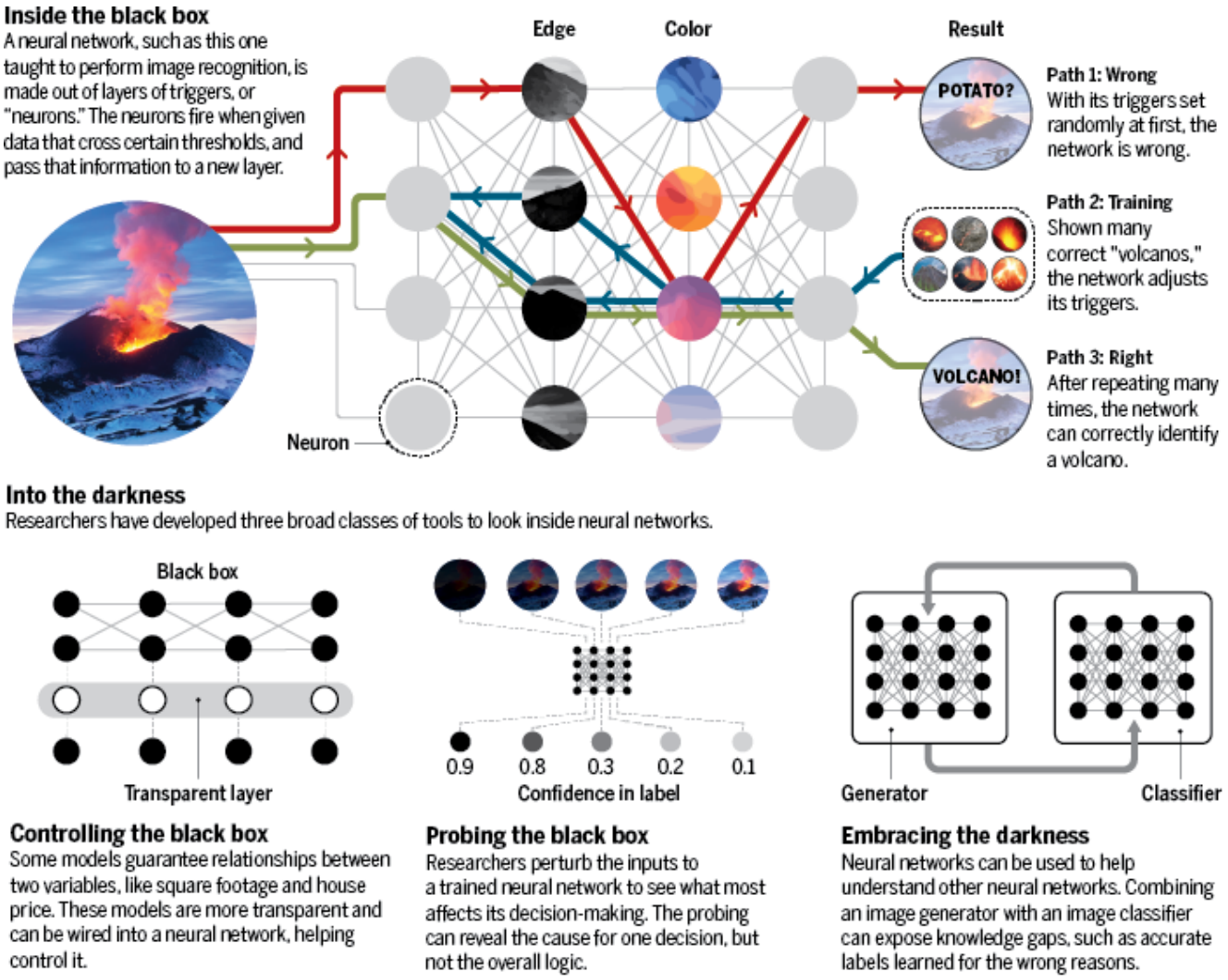


Fig. 4. Expanding the Brunswik lens model to identify the sources of machine-learning measurement bias: an illustration using personality as the focal construct. Areas highlighted in blue represent possible sources of machine-learning measurement bias; "platform-based personality": the personality construct measured by input data (e.g., online personality assessed by social media data) used in machine-learning models to predict self-report personality.

The need for Explainable AI



Voosen (2017) How AI detectives are cracking open the black box of deep learning, *Science*.

see also https://distill.pub/2018/building-blocks/

Algorithm aversion vs. Algorithm Appreciation

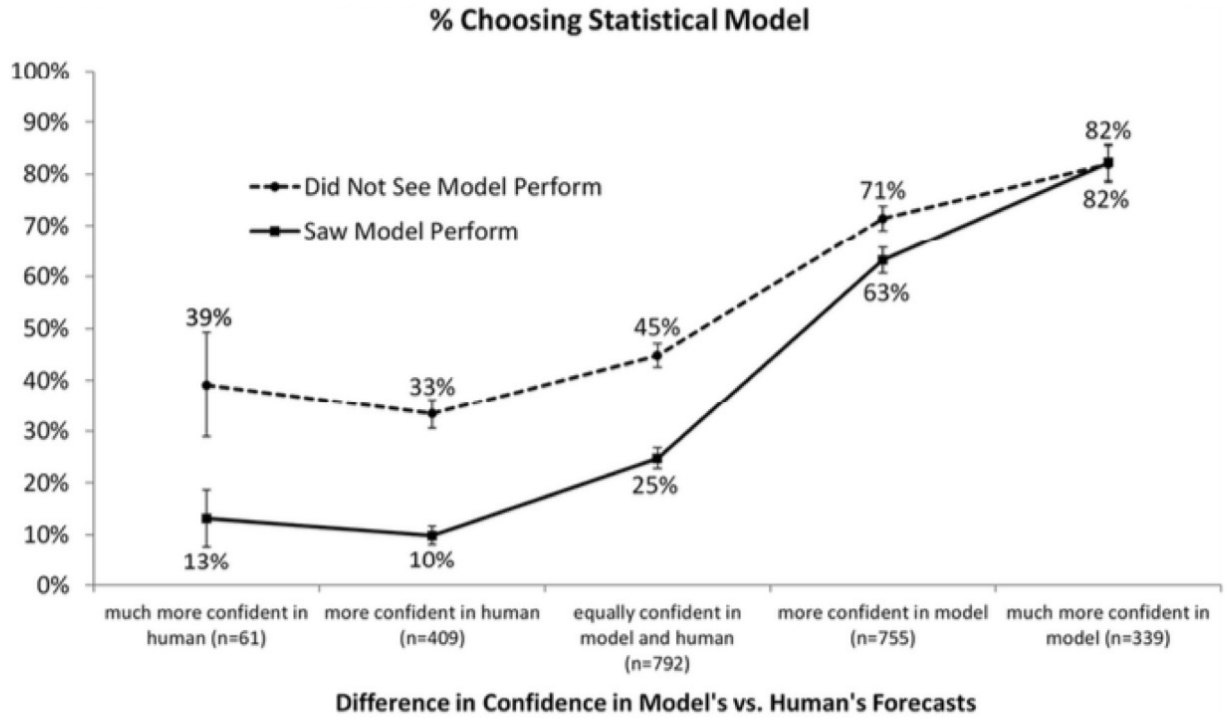


Figure 4. Most people do not choose the statistical model unless they are more confident in the model's forecasts than in the human's forecasts. Errors bars indicate ±1 standard error. The "Did Not See Model Perform" line represents results from participants in the control and human conditions. The "Saw Model Perform" line represents results from participants in the model and model-and-human conditions. Differences in confidence between the model's and human's forecasts were computed by subtracting participants' ratings of confidence in the human forecasts from their ratings of confidence in the model's forecasts (i.e., by subtracting one 5-point scale from the other). From left to right, the five *x*-axis categories reflect difference scores of: <–1, −1, 0, +1, and >1. The figure includes results from all five studies.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. http://doi.org/10.1037/xge0000033

Algorithm aversion vs. Algorithm Appreciation

A systematic review of algorithm aversion in augmented decision making

Jason W. Burton[1] | Mari-Klara Stein[2] | Tina Blegind Jensen[2]

[1]Department of Psychological Sciences, Birkbeck, University of London, London, UK

[2]Department of Digitalization, Copenhagen Business School, Frederiksberg, Denmark

Correspondence
Jason W. Burton, Department of Psychological Sciences, Birkbeck, University of London, Male Street, London, WC1E 7HX, UK.
Email: jasonwilliamburton@gmail.com

Abstract

Despite abundant literature theorizing societal implications of algorithmic decision making, relatively little is known about the conditions that lead to the acceptance or rejection of algorithmically generated insights by individual users of decision aids. More specifically, recent findings of algorithm aversion—the reluctance of human forecasters to use superior but imperfect algorithms—raise questions about whether joint human-algorithm decision making is feasible in practice. In this paper, we systematically review the topic of algorithm aversion as it appears in 61 peer-reviewed articles between 1950 and 2018 and follow its conceptual trail across disciplines. We categorize and report on the proposed causes and solutions of algorithm aversion in five themes: expectations and expertise, decision autonomy, incentivization, cognitive compatibility, and divergent rationalities. Although each of the presented themes addresses distinct features of an algorithmic decision aid, human users of the decision aid, and/or the decision making environment, apparent interdependencies are highlighted. We conclude that resolving algorithm aversion requires an updated research program with an emphasis on theory integration. We provide a number of empirical questions that can be immediately carried forth by the behavioral decision making community.

KEYWORDS
algorithm aversion, augmented decision making, human-algorithm interaction, systematic review

Burton, J. W., Stein, M. K., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. ***Journal of Behavioral Decision Making***, ***27***(11), 1309–20. http://doi.org/10.1002/bdm.2155

Algorithm aversion vs. Algorithm Appreciation

False expectations -> decision makers may have incorrect/unrealistic beliefs about the performance of algorithms

Lack of decision control -> decision makers may strive for autonomy

Lack of incentivization -> incentives for algorithmic use may be unclear or misaligned (effort vs. performance)

Combatting intuition -> decision makers may have incorrect (overconfident) beliefs about own intuition

Conflicting concepts of rationality -> lack of a match between algorithm's knowledge and those of the individual (risk vs. uncertainty)

Burton, J. W., Stein, M. K., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. ***Journal of Behavioral Decision Making***, ***27***(11), 1309–20. http://doi.org/10.1002/bdm.2155

Summary

- Potential limitations of actuarial methods: overfitting, complexity, lack of interpretability; bias
- Potential remedies: regularization, cross-validation, transparency/interpretable AI, checking for bias
- Algorithmic aversion vs. algorithmic appreciation: academic debate has centered around reasons for lack of adoption of algorithms in professional/forecasting settings; multifactorial issue (role of training, beliefs in intuition, incentives, etc.). More interdisciplinary work is needed…