ESSAY

Prediction and explanation in social systems

Jake M. Hofman,* Amit Sharma,* Duncan J. Watts*

Historically, social scientists have sought out explanations of human and social phenomena that provide interpretable causal mechanisms, while often ignoring their predictive accuracy. We argue that the increasingly computational nature of social science is beginning to reverse this traditional bias against prediction; however, it has also highlighted three important issues that require resolution. First, current practices for evaluating predictions must be better standardized. Second, theoretical limits to predictive accuracy in complex social systems must be better characterized, thereby setting expectations for what can be predicted or explained. Third, predictive accuracy and interpretability must be recognized as complements, not substitutes, when evaluating explanations. Resolving these three issues will lead to better, more replicable, and more useful social science.

For centuries, prediction has been considered an indispensable element of the scientific method. Theories are evaluated on the basis of their ability to make falsifiable predictions about future observations—observations that come either from the world at large or from experiments designed specifically to test the theory. Historically, this process of prediction-driven explanation has proven uncontroversial in the physical sciences, especially in cases where theories make relatively unambiguous predictions and data are plentiful. Social scientists, in contrast, have generally deemphasized the importance of prediction relative to explanation, which is often understood to mean the identification of interpretable causal mechanisms. In part, this emphasis may reflect the intrinsic complexity of human social systems and the relative paucity of available data. But it also partly reflects the widespread adoption within the social and behavioral sciences of a particular style of thinking that emphasizes unbiased estimation of model parameters over predictive accuracy (*1*). Rather than asking whether a given theory can predict some outcome of interest, the accepted practice in social science instead asks whether a particular coefficient in an idealized model is statistically significant and in the direction predicted by the theory.

Recently, this practice has come under increasing criticism, in large part out of concern that an unthinking "search for statistical significance" (*2*) has resulted in the proliferation of nonreplicable findings (*3*, *4*). Concurrently, growing interest among computational scientists in traditionally social scientific topics, such as the evolution of social networks (*5*), the diffusion of information (*6*, *7*), and the generation of inequality (*8*), along with massive increases in the volume and type of social data available to researchers (*9*), has raised awareness of methods from machine learning that evaluate performance largely in terms of predictive accuracy. We believe that the confluence of these two trends presents an opportune moment to revisit the historical separation of explanation and prediction in the social sciences, with productive lessons for both points of view. On the one hand, social scientists could benefit by paying more attention to predictive accuracy as a measure of explanatory power; on the other hand, computer scientists could benefit by paying more attention to the substantive relevance of their predictions, rather than to predictive accuracy alone.

Microsoft Research, 641 Avenue of the Americas, 7th Floor, New York, NY 10003, USA.
*Corresponding author. Email: jmh@microsoft.com (J.M.H.); amshar@microsoft.com (A.S.); duncan@microsoft.com (D.J.W.)

Standards for prediction

Predictive modeling has generated enormous progress in artificial intelligence (AI) applications (e.g., speech recognition, language translation, and driverless vehicles), in part because AI researchers have converged on simple-to-understand quantitative metrics that can be compared meaningfully across studies and over time. In light of this history, it is perhaps surprising that applications of similar methods in the social sciences often fail to adhere to common reporting and evaluation standards, making progress impossible to assess. The reason for this incoherence is that prediction results depend on many of the same "researcher degrees of freedom" that lead to false positives in traditional hypothesis testing (*3*). For example, consider the question of predicting the size of online diffusion "cascades" to understand how information spreads through social networks, a topic of considerable recent interest (*6*, *7*, *10*, *11*). Although seemingly unambiguous, this question can be answered only after it has first been translated into a specific computational task, which in turn requires the researcher to make a series of subjective choices, including the selection of the task, data set, model, and performance metric. Depending on which specific set of choices the researcher makes, what appear to be very different answers can be obtained.

To illustrate how seemingly innocuous design choices can affect stated results, we reanalyzed data from (*11*) comprising all posts made to Twitter during the month of February 2015 that contained links to the top 100 most popular websites, as measured by unique visitors. In addition to holding the data set fixed, for simplicity, we also restricted our analysis to a single choice of model, reported in (*11*), that predicts cascade size as a linear function of the average past performance of the "seed" individual (i.e., the one who initiated the cascade). Even with the data source and model held fixed, Fig. 1 (top) shows that many potential research designs remain: Each node represents a decision that a researcher must make, and each distinct path from the root of the tree to a terminal leaf node represents a potential study (*12*). We emphasize that none of these designs is intrinsically wrong. Nevertheless, Fig. 1 (bottom) shows that different researchers—each making individually defensible choices—can arrive at qualitatively different answers to the same question. For example, a researcher who chose to measure the AUC [the area under the receiver operating characteristic (ROC) curve] on a subset of the data could easily reach the conclusion that their predictions were "extremely accurate" [e.g., (*10*)], whereas a different researcher who decided to measure the coefficient of determination (R^2) on the whole data set would conclude that 60% of variance could not be explained [e.g., (*6*)].

Reality is even more complicated than our simple example would suggest, for at least three reasons. First, researchers typically start with different data sets and choose among potentially many different model classes; thus, the schematic in Fig. 1 is only a portion of the full design space. Second, researchers often reuse the same data set to assess the out-of-sample performance of many candidate models before choosing one. The resulting process, sometimes called "human-in-the-loop overfitting," can produce gross overestimates of predictive performance that fail to generalize to new data sets. Third, in addition to arriving at different answers to the same question, researchers may choose similar-sounding prediction tasks that correspond to different substantive questions. For example, a popular variant of the task described above is to observe the progression of a cascade for some time before making a prediction about its eventual size (*7*). "Peeking" strategies of this sort generally yield much better predictive performance than ex ante predictions, which use only information available before a given cascade. Importantly, however, they achieve this gain by, in effect, changing the objective from explanation (i.e., which features account for success?) to early detection (i.e., which cascades will continue to spread?). Using the same language ("predicting cascades") to describe both exercises therefore creates confusion about what has been accomplished, as well as how to compare results across studies.

Resolving these issues is nontrivial; nevertheless, some useful lessons can be learned from the past three decades of progress in the AI applications of machine learning, as well as from recent efforts to improve the replicability of scientific claims in behavioral science (*3*, *4*, *12*). First, comparability of results would be improved by

establishing consensus on the substantive problems that are to be solved. If early detection of popular content is the goal, for example, then peeking strategies are admissible, but if explanation is the goal, then they are not. Likewise, AUC is an appropriate metric when balanced classification (i.e., between classes of equal size) is a meaningful objective, whereas R^2 or root mean square error (RMSE) may be more appropriate when the actual cascade size is of interest. Second, where specific problems can be agreed upon, claims about prediction can be evaluated using the "common task framework" (e.g., the Netflix prize), in which competing algorithms are evaluated by independent third parties on standardized, publicly available data sets, agreed-upon performance metrics, and high-quality baselines (*13*). Third, in the absence of common tasks and data, researchers should transparently distinguish exploratory from confirmatory research. In exploratory analyses, researchers are free to study different tasks, fit multiple models, try various exclusion rules, and test on multiple performance metrics. When reporting their findings, however, they should transparently declare their full sequence of design choices to avoid creating a false impression of having confirmed a hypothesis rather than simply having generated one (*3*). Relatedly, they should report performance in terms of multiple metrics to avoid creating a false appearance of accuracy. In cases where data are abundant, moreover, researchers can increase the validity of exploratory research by using a three-way split of their data into a training set used to fit models, a validation set used to select any free parameters that control model capacity and to compare different models, and a test set that is used only once to quote final performance. Last, having generated a firm hypothesis through exploratory research, researchers may then choose to engage in confirmatory research, which allows them to make stronger claims. To qualify research as confirmatory, however, researchers should be required to preregister their research designs, including data preprocessing choices, model specifications, evaluation metrics, and out-of-sample predictions, in a public forum such as the Open Science Framework (https://osf.io). Although strict adherence to these guidelines may not always be possible, following them would dramatically improve the reliability and robustness of results, as well as facilitating comparisons across studies.

Limits to prediction

How predictable is human behavior? There is no single answer to this question because human behavior spans the gamut from highly regular to wildly unpredictable. At one extreme, a study of 50,000 mobile phone users (*14*) found that in any given hour, users were in their most-visited location 70% of the time; thus, one could achieve 70% accuracy on average with the simple heuristic "Jane will be at her usual spot today." At the other extreme, so-called "black swan" events (e.g., the impact of the Web or the 2008 financial crisis) are thought to be intrinsically impossible to predict in any meaningful sense (*15*). Last, for outcomes of intermediate predictability, such as presidential elections, stock market movements, and feature films revenues, the difficulty of prediction can vary tremendously with the details of the task (e.g., predicting box office revenues a week versus a year in advance). To evaluate the accuracy of any particular predictive model, therefore, we require not only the relevant baseline comparison—that is, the best known performance—but also an understanding of the best possible performance. The latter is important because when predictions are imperfect, the reason could be insufficient data and/or modeling sophistication, but it could also be that the phenomenon itself is unpredictable, and hence that predictive accuracy is subject to some fundamental limit. In other words, to the extent that outcomes in complex social systems resemble the outcome of a die roll more than the return of Halley's Comet, the potential for accurate predictions will be correspondingly constrained.

To illustrate the potential for predictive limits, consider again the problem of predicting diffusion cascades. As with "success" in many domains [e.g., in cultural markets (*8*)], the distribution of outcomes resembles Fig. 2 (top) in two important respects: First, both the average and modal success is low (i.e., most tweets, books, songs, or people experience modest success), and second, the right tail is highly skewed, consistent with the observation that a small fraction of items ("viral" tweets, best-selling books, hit songs, or celebrities) are orders of magnitude more successful than average. The key question posed by this picture, both for prediction and for explanation, is what determines the position of a given item in this highly unequal distribution. One extreme stylized explanation, which we label "skill world" (Fig. 2, bottom left), holds that success is almost entirely explained by some property that is intrinsic, albeit possibly hard to measure, which can be interpreted loosely as skill, quality, or fitness. At the opposite extreme, what we call "luck world" (Fig. 2, bottom right) contends that skill has very little impact on eventual success, which is instead driven almost entirely by other factors, such as luck, that are external to the item in question and effectively random in nature. Where exactly the real world lies in between these two extremes has important consequences for prediction. In skill world, for example, if one could hypothetically measure skill, then in principle it would be

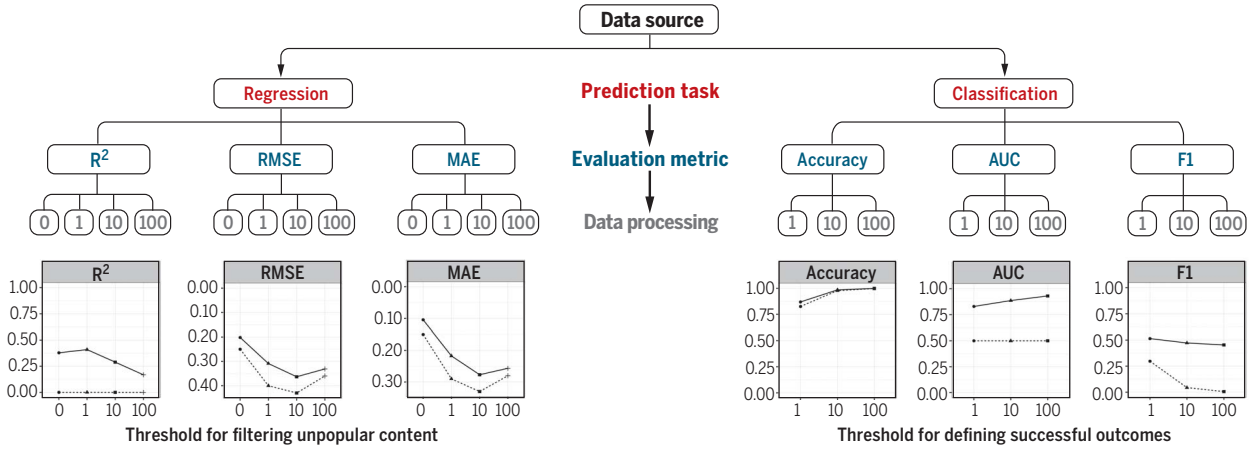

Fig. 1. A single question may correspond to many research designs, each yielding different answers. (**Top**) A depiction of the many choices involved in translating the problem of understanding diffusion cascades into a concrete prediction task, including the choice of data source, task, evaluation metric, and data preprocessing. The preprocessing choices shown at the terminal nodes refer to the threshold used to filter observations for regression or define successful outcomes for classification. Cascade sizes were log-transformed for all of the regression tasks. (**Bottom**) The results of each prediction task, for each metric, as a function of the threshold used in each task. The lower limit of each vertical axis gives the worst possible performance on each metric, and the top gives the best. Dashed lines represent the performance of a naive predictor (always forecasting the global mean for regression or the positive class for classification), and solid lines show the performance of the fitted model. R^2, coefficient of determination; AUC, area under the ROC curve; RMSE, root mean squared error; MAE, mean absolute error; F1 score, the harmonic mean of precision and recall.

possible to predict success with almost perfect precision. In luck world, in contrast, even a "perfect" predictor would yield mediocre performance, no better than predicting that all items will experience the same (i.e., average) level of success (*11*). It follows, therefore, that the more that outcomes are determined by extrinsic random factors, the lower the theoretical best performance that can be attained by any model.

Aside from some special cases (*14*), the problem of specifying a theoretical limit to predictive accuracy for any given complex social system remains open, but it ought to be of interest both to social scientists and computer scientists. For computer scientists, if the best-known performance is well below what is theoretically possible, efforts to find better model classes, construct more informative features, or collect more or better data might be justified. If, however, the best-known model is already close to the theoretical limit, scientific effort might be better allocated to other tasks, such as devising interventions that do not rely on accurate predictions (*16*). For social scientists, benchmarking of this sort could also be used to evaluate causal explanations. For example, to the extent that a hypothesized mechanism accounts for less observed variance than the theoretical limit, it is likely that other mechanisms remain to be identified. Conversely, where the theoretical limit is low (i.e., where outcomes are intrinsically unpredictable), expectations for what can be explained should be reduced accordingly. For example, although success is likely determined to some extent by intrinsic factors such as quality or skill, it also likely depends to some (potentially large) extent on extrinsic factors such as luck and cumulative advantage (*8*). Depending on the balance between these two sets of factors, any explanation for why a particular person, product, or idea succeeded when other similar entities did not will be limited, not because we lack the appropriate model of success, but rather because success itself is in part random (*17*).

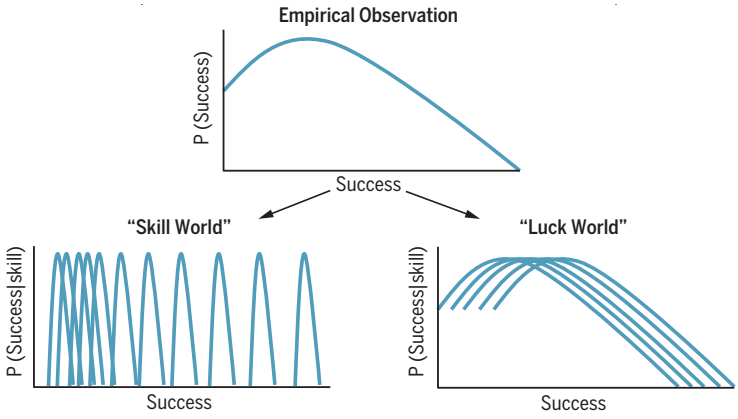

Fig. 2. Schematic illustration of two stylized explanations for an empirically observed distribution of success. In the observed world (**top**), the distribution of success is right-skewed and heavy-tailed, implying that most items experience relatively little success, whereas a tiny minority experience extraordinary success. In "skill world" (**bottom left**), the observed distribution is revealed to comprise many item-specific distributions sharply peaked around the expected value of some (possibly unobservable) measure of skill; thus, conditioning correctly on skill accounts for almost all observed variance. In contrast, in "luck world" (**bottom right**), almost all the observed variance is attributable to extrinsic random factors; thus, conditioning on even a hypothetically perfect measure of skill would explain very little variance. [Adapted from (*11*)]

Prediction versus interpretation

Conversations about the place of prediction in social science almost always elicit the objection that an emphasis on predictive accuracy leads to complex, uninterpretable models that generalize poorly and offer little insight. There is merit to this objection: The best-performing models are often complex, and, as we have already emphasized, an unthinking focus on predictive accuracy can lead to spurious claims. However, it does not follow that predictive accuracy is necessarily at odds with insight into causal mechanisms, for three reasons. First, simple models do not necessarily generalize better than complex models (*1*, *18*). Rather, generalization error is a property of the entire modeling process, including researcher degrees of freedom (*3*) and algorithmic constraints on the model search (*18*). Generalization error should therefore be minimized directly, as illustrated by ensemble methods such as bagging and boosting (*19*), which often succeed in lowering generalization error despite increasing model complexity. Second, there is increasing evidence from the machine learning literature that the trade-off between predictive accuracy and interpretability may be less severe than once thought. Specifically, by optimizing first for generalization error and then searching for simpler and more interpretable versions of the resulting model, it may be possible to achieve close to optimal prediction (subject to the limits discussed above) while also gaining insight into the relevant mechanisms (*20*). Third, it is important to clarify that "understanding" is often used to refer both to the subjective feeling of having made sense of something (i.e., interpreted it) and also to having successfully accounted for observed empirical regularities (i.e., predicted it). Although these two notions of understanding are frequently conflated, neither one necessarily implies the other: It is both possible to make sense of something ex post that cannot be predicted ex ante and to make successful predictions that are not interpretable (*17*). Moreover, although subjective preferences may differ, there is no scientific basis for privileging either form of understanding over the other (*18*).

None of this is to suggest that complex predictive modeling should supplant traditional approaches to social science. Rather, we advocate a hybrid approach in which researchers start with a question of substantive interest and design the prediction exercise to address that question, clearly stating and justifying the specific choices made during the modeling process. These requirements do not preclude exploratory studies, which remain both necessary and desirable for a variety of reasons—for example, to deepen understanding of the data, to clarify conceptual disagreements or ambiguities, or to generate hypotheses. When evaluating claims about predictive accuracy, however, preference should be given to studies that use standardized benchmarks that have been agreed upon by the field or, alternatively, to confirmatory studies that preregister their predictions. Mechanisms revealed in this manner are more likely to be replicable, and hence to qualify as "true," than mechanisms that are proposed solely on the basis of exploratory analysis and interpretive plausibility. Properly understood, in other words, prediction and explanation should be viewed as complements, not substitutes, in the pursuit of social scientific knowledge.

REFERENCES AND NOTES

1. L. Breiman, *Stat. Sci.* **16**, 199–231 (2001).
2. G. Gigerenzer, *J. Socio-Econ.* **33**, 587–606 (2004).
3. J. P. Simmons, L. D. Nelson, U. Simonsohn, *Psychol. Sci.* **22**, 1359–1366 (2011).
4. Open Science Collaboration, *Science* **349**, aac4716 (2015).
5. D. Liben-Nowell, J. Kleinberg, *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019–1031 (2007).
6. E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, "Everyone's an influencer: Quantifying influence on Twitter," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* [ACM (Association for Computing Machinery), 2011], pp. 65–74.
7. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd International Conference on World Wide Web* (ACM, 2014), pp. 925–936.
8. M. J. Salganik, P. S. Dodds, D. J. Watts, *Science* **311**, 854–856 (2006).
9. D. Lazer *et al.*, *Science* **323**, 721–723 (2009).
10. M. Jenders, G. Kasneci, F. Naumann, "Analyzing and predicting viral tweets," in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, 2013), pp. 657–664.
11. T. Martin, J. M. Hofman, A. Sharma, A. Anderson, D. J. Watts, "Exploring limits to prediction in complex social systems," in *Proceedings of the 25th International Conference on World Wide Web* (International World Wide Web Conference Committee, 2016), pp. 683–694.
12. A. Gelman, E. Loken, *Am. Sci.* **102**, 460 (2014).
13. M. Liberman, *Comput. Linguist.* **36**, 595–599 (2010).
14. C. Song, Z. Qu, N. Blumm, A.-L. Barabási, *Science* **327**, 1018–1021 (2010).
15. N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable* (Random House, 2007).
16. D. J. Watts, *Everything is Obvious*: *Once You Know the Answer* (Crown Business, 2011).
17. D. J. Watts, *Am. J. Sociol.* **120**, 313–351 (2014).
18. P. Domingos, *Data Min. Knowl. Discov.* **3**, 409–425 (1999).
19. R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear Estimation and Classification* (Springer, 2003), pp. 149–171.
20. M. T. Ribeiro, S. Singh, C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016).

10.1126/science.aal3856

Prediction and explanation in social systems

Jake M. Hofman, Amit Sharma and Duncan J. Watts

Science **355** (6324), 486-488.
DOI: 10.1126/science.aal3856