Explanation, prediction, and causality: Three sides of the same coin?

Emorie Beck (Washington University - St. louis), Elisa Jayne Bienenstock (Arizona State University), Jake Bowers (University of Illinois at Urbana–Champaign), Aaron B. Frank (RAND Corporation), Tony Grubesic (Arizona State University), Jake Hofman (Microsoft), Julia Rohrer (Leipzig University), Matt Salganik (Princeton), Duncan J. Watts (Microsoft)

Abstract

In this essay we make four interrelated points. First, we reiterate previous arguments (Kleinberg et al 2015) that forecasting problems are more common in social science than is often appreciated. From this observation it follows that social scientists should care about predictive accuracy in addition to unbiased or consistent estimation of causal relationships. Second, we argue that social scientists should be interested in prediction *even if they have no interest in forecasting per se*. Whether they do so explicitly or not, that is, causal claims necessarily make predictions; thus it is both fair and arguably useful to hold them accountable for the accuracy of the predictions they make. Third, we argue that prediction, used in either of the above two senses, is a useful metric for quantifying progress. Important differences between social science explanations and machine learning algorithms notwithstanding, social scientists can still learn from approaches like the Common Task Framework (CTF) which have successfully driven progress in certain fields of AI over the past 30 years (Donoho, 2015). Finally, we anticipate that as the predictive performance of forecasting models and explanations alike receives more attention, it will become clear that it is subject to some upper limit which lies well below deterministic accuracy for many applications of interest (Martin et al 2016). Characterizing the properties of complex social systems that lead to higher or lower predictive limits therefore poses an interesting challenge for computational social science.

"In general, we look for a new law by the following process. First, we guess it ... Then we compute the consequences of the guess, … to see what it would imply and then we compare the computation results to ... experiment or experience [or] with observations to see if it works. If it disagrees with experiment, it's wrong. In that simple statement is the key to science. It doesn't make any difference how beautiful your guess is, it doesn't make any difference how smart you are who made the guess, or what his name is… If it disagrees with experiment, it's wrong. That's all there is to it."

Richard Feynman, 1964

Prediction vs. explanation: A false dichotomy

Social science traditionally distinguishes between prediction and explanation. Explanation is viewed as being primarily concerned with the identification of a causal mechanism, with the purpose of developing understanding and potentially designing interventions to change outcomes. Prediction, on the other hand, is viewed as simply forecasting future events and outcomes without changing them, often for the purpose of incorporating that information into some other plan or decision making process. For example, a highly accurate forecast of rain would enable people to carry an umbrella only when it would be needed. Less trivially, the ability to accurately predict the likelihood of disruptive events such as wars, sectarian conflicts, investment surges or mass migrations may be critical to strategic planning processes (Mitchell 2009; Guston 2014; Selin and VanDeveer 2007). By contrast, if the purpose is to avert or otherwise impact these outcomes, an understanding of cause and effect may be essential.

Kleinberg et al (2015) formalize the distinction between prediction/forecasting and explanation/causation as follows. If Y is some outcome of interest (the weather, in their example) which depends on some intervention X and $\pi(X, Y)$ is the associated payoff function for some actor, then we can quantify the effect of the intervention through the following derivative:

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0} \underbrace{(Y)}_{\text{prediction}} + \frac{\partial \pi}{\partial Y} \underbrace{\frac{\partial Y}{\partial X_0}}_{\text{causation}} .$$

These two terms represent the two different pathways through which an intervention might affect the payoff (see Fig. 1 for examples from Kleinberg et al 2015). The first corresponds to a mechanism that directly changes the payoff, but not the outcome (e.g., superstitions aside, bringing an umbrella doesn't affect the chance of rain, but it can prevent you from getting wet). This is what we might call a "pure forecasting" task, because it doesn't require understanding the process that causes rain, but only a reliable forecast of whether it will rain or not. The second term corresponds to a different pathway, where the intervention changes the outcome, which in turn alters the payoff (e.g., performing a rain dance is presumed to save dying crops only if it actually causes rain). This is what we might call a "pure causation" task, where a proper understanding of the underlying mechanism is crucial.

Clearly, this forecasting-causation dichotomy is an oversimplification. As Athey (2017) points out, many applied problems in social science are mixtures of the two. For example, predictive algorithms used by cities to assign fire and health inspectors should take into account not only the factors that predict that a particular establishment will be in violation of fire and/or health codes (a forecasting problem) but also the causal effect on an establishment's behavior of receiving an inspection or not (a causation problem). Just as with targeted advertising schemes which rank potential customers by probability of future purchase, the most efficient allocation of resources is the one that makes the highest marginal impact on behavior, a calculation that requires accurately predicting risk and also the causal effect of the intervention.

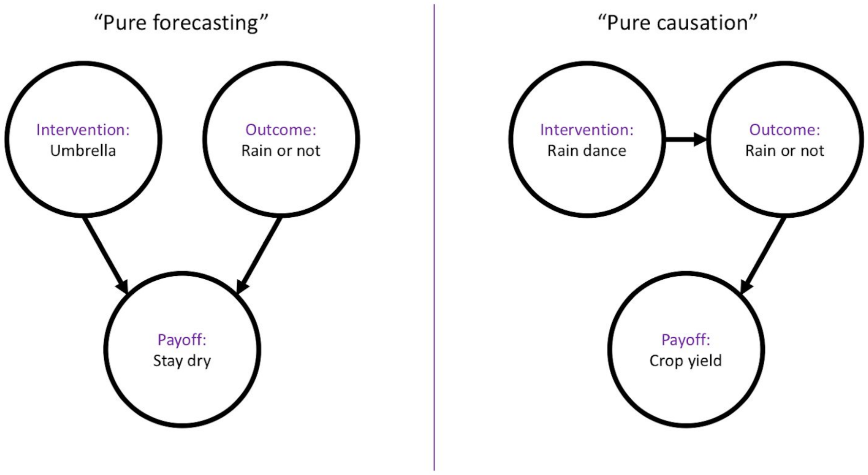

Fig. 1. Alternate causal pathways for "forecasting" vs. "causal" intervention, following Kleinberg et al. 2015.

Nevertheless, we concur with Kleinberg et al. that forecasting problems are important in their own right, but have historically been underappreciated in the social sciences. Mullainathan and Spiess (2017) provide several such examples, including the value of predictive algorithms in forecasting bail decisions (Kleinberg et al. 2017), predicting teacher quality (Chalfin et al. 2017), identifying high-risk youth behavior (Chandler et. al 2011), and improving poverty targeting (McBridge et. al 2016). Each of these highlights cases in which increases in predictive performance have direct policy benefits, and a broader embrace of predictive modeling across the social sciences will surely reveal many others.

A second, and less obvious, claim is that prediction remains relevant even in cases where one's interest lies exclusively in the second term $\left(\frac{\partial Y}{\partial X}\right)$; that is, with understanding causal effects of X on Y. Very often social scientists do espouse a lack of interest in prediction on the grounds that they do not care about forecasting problems (Lieberson and Lynn 2002; Hedström and Ylikoski 2010).[1] Yet, as Ward et al (2010) argue, explanations that invoke causal mechanisms **always** make predictions---specifically predictions about what will happen under an intervention---whether these predictions are articulated explicitly or not. Any explanation that identifies an effect of some factor X on some outcome Y--for example, that economic inequality increases the risk of civil war, or that psychological safety improves team performance, or that

[1] We find this tendency especially confusing since it runs directly against current norms of understanding causal effects in terms of comparisons of counter-factual potential outcomes: i.e. Y|X=x_1 vs Y|X=x_2 (Holland 1986) or Pearl's *do*-calculus (cite), which also formalized causal relations in terms of Y.

neighborhood characteristics impact life-course outcomes--in effect makes a prediction about how Y should change under different values of X. Importantly, these “implied predictions” are not incidental to the purpose of such explanations. Indeed, assertions of the form “Y would have been different if X had been different” owe much of their theoretical interest (and their policy relevance) to their implication that “Y *will be* different if X is changed” in cases other than those in the analysis, including in the future (Watts 2014). Even when explanations are not explicitly asserted to be causal, the language used very often insinuates causality, as when the reader is encouraged to treat a series of events as the “reason” for some outcome, or to consider some “what if” counterfactual outcome, or to generalize beyond the specific situation under analysis. Again, this insinuation of causality is not incidental to the purpose of the explanation being offered, and very often is central to its appeal.

Regardless of the details, whenever it is asserted or implied that something (Y) happened or will happen or might happen because of something else (X), it is important to realize that a causal claim is being made. From this observation it necessarily follows that a prediction is also being made, about how other events similar to Y, whether in the past or in the future, should be expected to turn out, or to have have turned out, depending on the presence or absence of X. To be clear, specifying the claim in a form that admits a quantifiable prediction may be nontrivial. Likewise, it may be prohibitively difficult to gather adequate data to test such predictions, even if they can be quantified, as could be the case for predictions about the probability of rare or unique events such as wars or presidential elections. But none of these practical difficulties associated with evaluating the predictive consequences of a causal explanation should be taken to mean that all such explanations do not have predictive consequences. If one is unwilling or unable to articulate these consequences and to have one’s explanation evaluated in terms of its ability to account for them then it is the explanation not the evaluation procedure that should be viewed with skepticism. In other words, just because Feynman’s prescription is oversimplified and difficult to implement in practice does not reduce its relevance to social scientific explanations in principle.

Nevertheless, the idea that causal explanations are distinct from prediction, with the former ranking as more important than the latter, continues to have traction in various communities (Mitchell 2004). In part these objections may arise from the aforementioned conflation of “prediction” with what we have called “forecasting.” By interpreting calls to emphasize prediction as merely about improving forecasts, researchers can safely disassociate the business of producing explanations from that of evaluating their implied predictions. But in part the objection may stem from the standard practice of “hypothesis testing” in quantitative social science, wherein researchers derive from theory some prediction about the sign and significance of a particular regression coefficient. Because this practice follows the form of the scientific method--start with a theory, derive one or more hypotheses which make predictions, test these predictions on data, etc.--practitioners may believe that they are already evaluating the predictions of their theories. However, hypothesis testing as it is almost universally practiced does not directly test hypotheses of interest at all. Rather, one instead proposes a “null hypothesis” of the form “the effect predicted by my theory does not exist” and then attempts to reject this null (or “nil”) hypothesis. Although this procedure can be appropriate in instances where the theory itself does not make sharp predictions (and thus the rejection of “no effect” rules out one possible alternative), for the same reason it does not directly validate the theory itself, either in isolation or relative to alternative theories, a point that has been made for decades by statisticians, social scientists, and life scientists alike (Gill 1999; Johnson 1999; Gigerenzer 2004; Gelman and Carlin 2017; Lakens et al. 2018). Likewise, null hypothesis testing offers no direct way of comparing two competing theories unless one is taken to be the null and the other the alternative, a procedure which is rarely adopted in practice. In other words, null hypothesis testing gives no way of identifying whether one explanation is better than another, and if so, by how much.

Assuming that the goal of scientific research is (a) to generate, test, and compare competing explanations of observed phenomena, and (b) to provide useful advice to policy makers and other practitioners, predictive validation is essential for evaluating the progress of research programs. Adopting such a framework for the social sciences would confer three main benefits. First, it would highlight that some problems of interest can be addressed primarily with accurate forecasts, irrespective of the underlying causal mechanisms. Second, it would explicitly acknowledge that causal explanations necessarily generate predictions; testing those predictions would provide a robust and generalizable way of assessing the quality of these explanations. And third, it would facilitate direct comparison of competing theories and models to provide a clear sense of how our collective understanding of social science phenomena has progressed over time.

Using prediction to evaluate explanations

We now turn to the question of how, precisely, prediction should be used to evaluate explanations. There are many ways of measuring predictive performance, the most common being R^2, or the fraction of the variance “explained” by a model. Others include mean average error (MAE), root mean squared error (RMSE), area under the “receiver operating curve” (AUC), accuracy (the fraction of correct labels), precision (the fraction of positive labels that are true positives), recall (the fraction of true positives that are positively labeled), and F1 score (the harmonic mean of precision and recall).

For the special case of analyzing a dichotomous treatment, a simple option is to focus on effect sizes in place of p-values. For example, measures such as Cohen's d (Cohen 1988) are increasingly used in medicine, psychology, and other domains (Coe 2002; Sullivan and Feinn 2012; Wasserstein, et. al 2016). This shift to effect sizes provides several benefits, enabling meta-analyses that combine results from different studies on the same topic and allowing for comparisons between competing factors and theories. Although perhaps not obvious, effect sizes are useful precisely because they are themselves a measure of predictive performance. Indeed, there is a one-to-one mapping between Cohen's d for dichotomous treatments and predictive measures such as R^2 (Bruce 1998). Placing emphasis on either measure stands in contrast with many published studies that present tables of estimated model coefficients and p-values but pay little or no attention to the corresponding R^2 values to assess how much of the variance in outcomes these models explain. This is akin to ignoring effect sizes and looking only at which effects are non-zero.

The plethora of performance metrics is both a benefit and a liability. On the one hand, different prediction tasks are intrinsically suited to different metrics. For example, accuracy is appropriate for balanced classification tasks (where the number of positive and negative cases is equal), but can be highly misleading for imbalanced tasks. MAE or RMSE may be more informative than R^2 for when the magnitude of error is important but R^2 may be more appropriate when comparing accuracy across tasks that involve different scales. Given the vast range of questions posed by the social sciences, it is therefore helpful to have many measures of predictive performance from which to choose.

On the other hand, the flexibility to choose different metrics for the same prediction task can yield apparently different performances even in the absence of any underlying difference in real accuracy. Indeed it is simple to show that by switching between regression and classification and by choosing different metrics for each, one can reach qualitatively different conclusions about predictive performance even while using the same underlying model on the same data (Hofman et al. 2017). This is especially problematic given issues of publication bias, wherein outsized importance is placed on the discovery of positive results while null findings are largely ignored. As a result, myopically focusing on prediction without a grounding in explanation can lead researchers to put the cart before the horse, searching for a prediction problem to which they can declare “success” instead of using prediction in service of furthering our understanding of how the world works. Absent a causal understanding of the world, predictive models may conclude that playing basketball makes people taller, or ashtrays cause cancer (Shapiro 2005, p. 16; Cartwright and Hardie 2012, p. 50.). For predictive performance to be a useful metric for comparing theories or assessing progress, therefore, it is first necessary to reach agreement on the specific prediction problems and evaluation metrics that are appropriate for a given research question.

A final consideration is that when studies do discuss predictive measures, they often do so only for the data on which the model was fit. This amounts to postdiction, or “explaining after the fact”, which often generates explanations that work only for the particular dataset under study but fail to generalize to future observations. Indeed, it is possible to increase the complexity of a model until its predictions perfectly match already observed outcomes, but doing so does not imply we have achieved a perfect model of the phenomenon being studied; it merely means that we have found another way to represent the data we have collected. Thankfully, this problem is easily dealt with by adhering to Feynman’s recipe: not only should we make a guess (or prediction), but we should check this guess against new data (that we had not seen prior to making the guess). In machine learning, this is referred to as focusing on *generalization,* or *out-of-sample,* error to avoid *overfitting*. Ideally this is done by making predictions for future data, but can be approximated using only historical data that has already been collected. The procedure, known as cross-validation, involves splitting a dataset into two parts: one that is used only to fit the model (referred to as the training set) and another that is used only to assess its performance (the test set). Selecting the model that minimizes error on the test set (instead of the training set) avoids the problem of overfitting, and reporting the test set error gives a reliable approximation for how well we expect the model to perform in the future, provided the underlying data generating process remains the same (Janeksela 1982).

Limits to prediction

Assuming social scientists choose to evaluate explanations in terms of out-of-sample predictive performance, two related questions naturally arise.

1. In any given context and for any given research question, how accurate can we expect our predictions to be? Setting aside practical difficulties such as missing or biased data, measurement error, or ignorance of the underlying model, that is, how accurately can some outcome of interest be predicted *in principle*, given constraints around the data available for the task? The definitive upper boundary of predictive accuracy is perfect prediction ($R^2 = 1$, MSE = 0); however, in complex systems in general, and in the social world in particular, empirical predictions are typically well below this limit (Goel et al. 2010; Bakshy et al. 2011; Glaeser, Sacerdote, and Scheinkman 1996; Martin et al. 2016; Kleinberg, Liang, and Mullainathan 2016). To what extent can we can expect that predictions will continually improve, given time, effort, and resources? At what point does the marginal return on additional data/model complexity/effort become zero?
2. How does this upper limit to predictiability, assuming one exists, vary across contexts and research questions? For example, it is presumably an easier task to predict the lifetime box office revenue for a feature film the week after the opening weekend than it is a year in advance; correspondingly, one might expect that the best possible prediction also depends on how long in advance the prediction is made. Likewise, in some cases it is easier to make predictions about average outcomes than it is about individual events (e.g. predicting average wealth of an entire cohort of individuals vs. individual life outcomes, or predicting a student's average GPA in high school compared to their GPA in only 9th grade) and in some cases the reverse applies (e.g. predicting which songs an individual will like based on their history of song listens vs. predicting hits songs for an entire market).

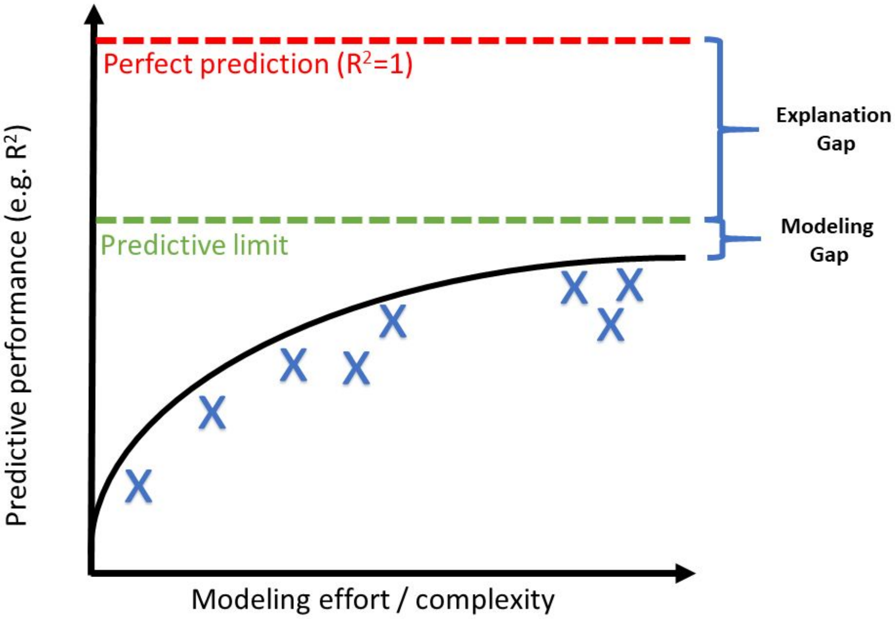
Perfect prediction (R^2=1)
Explanation Gap
Predictive limit
Modeling Gap
Predictive performance (e.g. R^2)
Modeling effort / complexity
X

Fig. 2. Hypothetical limit to predictive performance. Empirical models (X) show diminishing returns to modeling complexity and/or effort, implying a theoretical upper limit to predictive accuracy that is bounded away from deterministic accuracy (R^2=1). The gap between the best-available empirical model and the theoretical limit is the “modeling gap,” whereas the gap between the theoretical limit and R^2=1 is the “explanation gap,” where the latter implies that some variance is intrinsic and cannot be explained.

A clearer understanding of the limits of prediction could facilitate and/or guide progress when researchers are most concerned with their absolute success in prediction improvement. If the research team is close to the known limit, there may be a limited return on invested effort to improve prediction. As a result, this would suggest that researchers should focus other aspects of the problem at hand, such as consolidating high-performing predictive models into new theory, or testing the suitability of predictive models for prescriptive action and control. Conversely, if the research team is far from the known or expected limit, efforts to improve prediction are justified. Likewise, policy makers could decide whether or not to invest limited resources into efforts for improving predictions, or choose to focus on developing strategies for interventions that are robust to uncertainties (Raynor 2007; Watts 2011; Mintzberg 1994; Lempert, Popper, and Bankes, 2003) This would include the development of flexible plans to adapt to uncertainties that cannot be reduced further.

Establishing a predictive limit for a given domain

Given a prediction task with its underlying social system, is there an inherent limit to how well we can predict, and if so what is it? To our knowledge little is known that can answer this question in any general sense. Perhaps the domain in which this question has been broached most explicitly is in human mobility studies that leverage a combination of smart-phone data and social media traces (Cuttone, Lehmann, and González 2018; Song et al. 2010). Another area in which predictive performance across studies has been a focus of concern is with respect to predicting retweet cascade size on Twitter (Bakshy et al. 2011; Martin et al. 2016). Although both sets of studies suggest that some fundamental limit exists, neither provides a clear answer to what it might be even in the specific context under consideration.

We propose three approaches for establishing upper limits to predictive performance:

1. **Reviews of the existing literature**. Take R^2s (or equivalent) from the literature and plot them against various characteristics of the model and data, such as the number of included features, sample size, statistical approach, etc. A major impediment to this approach is that in the social sciences, many studies do not report model performance, thus is can be impossible to assess without reproducing the analysis. Another is that even in fields like computer science where it is common to report model performance, different authors may choose to report different performance metrics or to choose different prediction tasks (e.g. regression vs. classification), both of which also complicate efforts to make comparisons across studies (Hofman, Sharma, and Watts 2017). Looking forward, both problems could be addressed by encouraging researchers to a) adopt standardized reporting procedures, which could potentially comprise a "basket" of measures and b) make their data and code available for other researchers to verify and extend.

2. **Mass collaborations**. Social scientists could adapt the "Common Task Framework" that is popular in AI and machine learning communities (Donoho 2017). An example of the CTF applied to a social science problem is the ("Fragile Families Challenge" n.d.)(FFC), a Netflix-style "prize contest" was based on a large, longitudinal data set (the Fragile Families and Child Wellbeing Study) that has followed a cohort of nearly 5,000 children from birth through adolescence (Kindel et al. 2018). The "challenge" part comprised an open solicitation to researchers of any discipline and seniority to submit predictive models of individual-level outcomes (e.g. grit, GPA, material hardship, etc.) from the most recent wave of data, which at the time had been collected but not yet made public. Over 400 participants, ranging from undergraduate machine learning students to full professors of sociology, economics, and political science submitted entries, either as individuals or in teams, over the course of several months. Although in principle this model of mass collaboration could be applied to any longitudinal social dataset, the nature of social data encounters complications such as protection of data-subject privacy and autonomy (Lundberg et al. 2018) and absence of machine readable metadata that are less problematic for more traditional AI applications (Kindel et al. 2018). Another challenge is how to incentive large numbers of researchers to collaborate on a single

problem, although recent large-scale collaborations in psychology (https://osf.io/wx7ck/, https://psysciacc.org/) are encouraging.

3. **Derive theoretical limits from models**. Another possible approach to establishing an upper limit to predictive accuracy is to show that such a limit exists in a theoretical model of the system. For example, Song et al (2010) proposed that when an individual divides his or her time among N locations, the predictive performance for any algorithm is bounded above by Fano's inequality, which in turn depends on the entropy S of the individual's location: individuals with high entropy divide their time relatively evenly among N locations and are relatively unpredictable whereas individuals with low entropy are the opposite. In the case analyzed by Song et al, individuals spent an average of 70% of their time at their most visited location for each hour of the day (at night the number was 90%); thus had relatively low entropy and high maximum predictability (over 90%). In the context of Twitter cascadest, Martin et al. (2016) used simulations to show that under certain idealized assumptions, the maximum R^2 for any model predicting cascade size would depend sensitively on (a) the heterogeneity of the underlying content and (d) measurement error. Although promising, we note that neither of these approaches applies generally to complex systems; thus the existence of some theoretical limit to predictive accuracy for any particular system remains, to our knowledge, an open question.

Factors that affect predictive limits across domains

Assuming that some general theoretical framework for describing the limits to prediction for any given system could be developed, what are the key elements that any such framework would have to incorporate? Although we regard this exercise as a major open research suggest three sets of features that may determine predictive performance in a given setting: 1) prediction type, 2) available data, 3) features inherent to the system of interest.

First, the "type" of prediction may vary along at least three dimensions, each of which might limit predictive accuracy.

1. Time lag between prediction and outcome. Predicting the outcome of a presidential election or the box office revenue of a feature film is clearly more difficult a year in advance of the event in question (election day or opening day respectively) than it is a day in advance. Conversely, predictions of school performance for children based on prior classroom behavior may be less noisy over long intervals than over short ones (Alexander, Entwisle, and Dauber 1993). Regardless, the time interval between when a prediction is made and the outcome in question is likely an important factor in determining the upper limit of predictive performance.
2. Routine vs. rare events. In some domains (e.g. credit card defaults, click-through rates on display ads) predictive models can benefit from large amounts of comparable data, whereas at in other domains (e.g. the next financial crisis) the events in question are arguably unique.

3. *Individual vs. aggregate outcomes*. Outcomes for individual people may be highly stochastic, and hence even best-possible predictions may be of limited accuracy. By contrast, predictions about whole categories or populations of people (e.g. in cross country comparisons) may benefit from averaging of individual-level errors and hence allow for much higher accuracy in principle. Likewise, predictions about changes in distributions (e.g. inequality as a function of increasing social influence (Salganik, Dodds, and Watts 2006)) may be subject to weaker limits than predictions about the position of individual entities in those distributions.

Second, the volume, quality, and relevance of data that is available can limit predictive accuracy. Many predictions are difficult because available data are too difficult or expensive to sample (historically this has applied to social network data). Others are difficult because the data, even if obtainable, is too noisy, biased, or indicates some proxy of the desired outcome rather than the outcome itself (e.g. communication frequency as a proxy for tie strength, gregariousness as a proxy of interpersonal influence).

Third, inherent features of the underlying system can impact predictability. Factors such as nonlinearity, stochasticity, interdependency of units, and multiplicity of scales, types and interactions all contribute to system complexity, which may in turn limit predictability (Weaver 1948; Watts 2011) Finally, a factor that is unique to social systems is “performativity” (MacKenzie and Millo 2003): that systems comprising people may react to predictions about systems in ways that either reinforce the prediction (Merton 1948) or negate it (cf. the so-called Lucas critique (Lucas 1976)).

Conclusion

Throughout this paper we have argued that prediction and explanation should be viewed as a complements, not substitutes, when studying social scientific phenomena. In particular, prediction has traditionally been undervalued in the social sciences, and the field would benefit from adopting tools and techniques from predictive modeling. This holds for both "pure forecasting" problems and for "pure causation” problems alike. As others have pointed out (Kleinberg et al. 2015; Mullainathan and Spiess 2017), prediction is beneficial in the "pure forecasting" sense simply because high-quality forecasts allow one to plan for future events, provided that the assumptions and scope of the model(s) used remain valid for the considered circumstances. But prediction is also useful in "pure causation" contexts because all explanations make predictions, and explicitly testing these predictions provides a straightforward way of evaluating the corresponding explanations. Predictive validation of social science models, such as focusing on effect sizes and out-of-sample predictive performance, provides a better framework for assessing and comparing models than the standard practice of null hypothesis testing.

That said, predictive validation is not a panacea. For predictive performance to be a meaningful way of comparing theories or assessing progress, it is first necessary to reach agreement on how to operationalize a prediction task corresponding to a given research question. This includes a clear articulation of the underlying task, including domain, scope, and evaluation criteria. Without this, it becomes difficult to compare the results of one study with the next, defeating the purpose of the framework. We make several recommendations to avoid this issue and to maximize the value of predictive validation for furthering our understanding of the social scientific phenomena. The first is systematic reviews of the literature to compare the predictive performance of existing models where possible, with the adoption of standardized reporting and open data and code to enable such investigations going forward (e.g., see Ward et al. 2010). For settings in which such comparisons cannot be made, we recommend the common task framework for mass collaborations on problems of mutual interest, as was recently employed by the Fragile Families Challenge to enable over a hundred teams to work together to better predict and understand future outcomes for disadvantaged children.

Finally, we note that adopting these practices would likely reveal certain limits in our ability to predict social events and outcomes. Studying these limits is itself an important topic, as a proper understanding of predictive limits would inform the investment of scientific attention and resources. We provide several dimensions on which to characterize predictive limits, including the type of prediction being made, features of the data available for the problem, and inherent aspects of the system being studied. Characterizing how and why these properties lead to higher or lower predictive limits poses an interesting challenge for the field of computational social science.

Taken together, we believe that following these recommendations for increased attention to prediction in the social sciences would improve the utility of resulting explanations, enabling new capabilities in planning and policy and furthering our understanding of how people behave and interact.

Bibliography

Alexander, Karl L., Doris R. Entwisle, and Susan L. Dauber. 1993. “First-Grade Classroom Behavior: Its Short-and Long-Term Consequences for School Performance.” *Child Development* 64 (3): 801–14.

Athey, Susan. 2017. “Beyond Prediction: Using Big Data for Policy Problems.” *Science* 355 (6324): 483–85.

Bakshy, Eytan, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. “Everyone’s an Influencer: Quantifying Influence on Twitter.” Proceedings of the Fourth ACM International Conference on Web Search and Data Mining 978-1-4503-0493-1. Hong Kong, China: ACM.

Braha, Dan. 2012. “Global Civil Unrest: Contagion, Self-Organization, and Prediction.” *PloS One* 7 (10): e48596.

Brunner, Karl, and Alan Meltzer. "Econometric policy evaluation. A critique." In Theory, Policy, Institutions: Papers from the Carnegie-Rochester Conferences on Public Policy, vol. 1, p. 257. North Holland, 1983.

Cartwright, Nancy, Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford University Press.

Cuttone, Andrea, Sune Lehmann, and Marta C. González. 2018. “Understanding Predictability and Exploration in Human Mobility.” *EPJ Data Science* 7 (1). https://doi.org/10.1140/epjds/s13688-017-0129-1.

Donoho, David. 2017. “50 Years of Data Science.” *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 26 (4): 745–66.

“Fragile Families Challenge.” n.d. Accessed October 15, 2018. http://www.fragilefamilieschallenge.org.

Gelman, Andrew, and John Carlin. "Some natural solutions to the p-value communication problem—and why they won’t work." Journal of the American Statistical Association 112, no. 519 (2017): 899-901.

Gigerenzer, Gerd. 2004. “Mindless Statistics.” *The Journal of Socio-Economics* 33 (5): 587–606.

Gill, Jeff. 1999. “The Insignificance of Null Hypothesis Significance Testing.” *Political Research Quarterly* 52 (3): 647–74.

Glaeser, Edward L., Bruce Sacerdote, and Jose A. Scheinkman. 1996. “Crime and Social Interaction.” *The Quarterly Journal of Economics* 111 (2): 507–48.

Goel, Sharad, Sebastien Lahaie, Jake Hofman, David M. Pennock, and Duncan J. Watts. 2010. *What Can Search Predict?* 19th International World Wide Web Conference. Raleigh, North Carolina.

Guston, David H. 2014. “Understanding ‘Anticipatory Governance.’” *Social Studies of Science* 44 (2): 218–42.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. “Prediction and Explanation in Social Systems.” *Science* 355 (6324): 486–88.

Janeksela, Galan M. 1982. “INCREASING CONFIDENCE IN SOCIAL SCIENCE RESEARCH FINDINGS VIA CROSS-VALIDATION.” *International Review of Modern Sociology* 12 (1): 67–75.

Johnson, D. H. 1999. “The Insignificance of Statistical Significance Testing.” *The Journal of Wildlife Management*. https://www.jstor.org/stable/3802789?casa_token=a4aBrtRoZocAAAAA:ZzcwxVzlbIe8xWxj84vwjYzA4iNaIW0li2cdkII50zKdz3hegIj45h2V3st7D3h54MMJq7I8eou49RbOlNq2LjkrEnCAxoUmgMC3G4-543ybVctBea8.

Kindel, Alexander, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, et al. 2018. “Improving Metadata Infrastructure for Complex Surveys: Insights from the Fragile Families Challenge.” SocArXiv. September 19. doi:10.31235/osf.io/u8spj.

Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan. 2016. “The Theory Is Predictive, but Is It Complete? An Application to Human Perception of Randomness.” *Working Paper*.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. “Equivalence Testing for Psychological Research: A Tutorial.” *Advances in Methods and Practices in Psychological Science* 1, no. 2 (June 2018): 259–69. doi:10.1177/2515245918770963.

Langley, Pat, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. 1987. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press.

Lempert, Robert J., Steven W. Popper, Steven C. Bankes. 2003. *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. RAND Corporation.

Lucas, Robert. 1976. “Econometric Policy Evaluation. A Critique.” In *The Phillips Curve and Labor Markets. Carnegie-Rochester Conference Series on Public Policy. Vol 1*, edited by Karl Brunner and Alan Meltzer, 19–46.

Lundberg, Ian, Arvind Narayanan, Karen Levy, and Matthew J. Salganik. 2018. “Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge.” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1809.00103.

MacKenzie, Donald, and Yuval Millo. 2003. “Constructing a Market, Performing Theory: The Historical Sociology of a Financial Derivatives Exchange.” *The American Journal of Sociology* 109 (1): 107–45.

Martin, Travis, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. 2016. “Exploring Limits to Prediction in Complex Social Systems.” Proceedings of the 25th International Conference on World Wide Web 978-1-4503-4143-1. Montreal, Quebec, Canada: International World Wide Web Conferences Steering Committee.

Merton, Robert K. 1948. “The Self-Fulfilling Prophecy.” *The Antioch Review* 8: 193.

Mintzberg, Henry. 1994. *Rise and Fall of Strategic Planning*. Simon and Schuster.

Mitchell, Sandra D. 2009. *Unsimple Truths: Science, Complexity, and Policy*. University of Chicago Press.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *The Journal of Economic Perspectives: A Journal of the American Economic Association* 31 (2): 87–106.

Raynor, Michael E. 2007. *The Strategy Paradox: Why Committing to Success Leads to Failure, and What to Do about It*. Crown Business.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market.” *Science* 311 (5762): 854–56.

Selin, Henrik, and Stacy D. VanDeveer. 2007. “Political Science and Prediction: What’s Next for U.S. Climate Change Policy?” *The Review of Policy Research* 24 (1): 1–27.

Shapiro, Ian, 2005. *The Flight from Reality in the Human Sciences*. Princeton University Press.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. “Limits of Predictability in Human Mobility.” *Science* 327 (5968): 1018–21.

Watts, Duncan J. 2011. *Everything Is Obvious: How Common Sense Fails Us*. Crown Pub.

Weaver, W. 1948. “Science and Complexity.” *American Scientist* 36 (4): 536–44.

Yarkoni, Tal, and Jacob Westfall. 2017. “Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning.” *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 12 (6): 1100–1122.