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

To Explain or to Predict?

Galit Shmueli

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

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

1. INTRODUCTION

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

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

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

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

tinct scientific goals that they are aimed at: causal explanation and empirical prediction, respectively. *Explanatory modeling* and *predictive modeling* reflect the process of using data and statistical (or data mining) methods for explaining or predicting, respectively. The term *modeling* is intentionally chosen over *models* to highlight the entire process involved, from goal definition, study design, and data collection to scientific use.

1.1 Explanatory Modeling

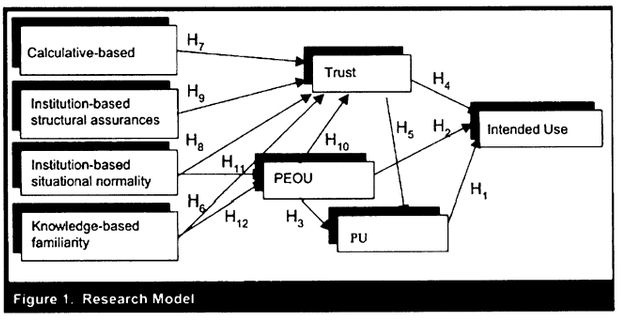
In many scientific fields, and especially the social sciences, statistical methods are used nearly exclusively for testing causal theory. Given a causal theoretical model, statistical models are applied to data in order to test causal hypotheses. In such models, a set of underlying factors that are measured by variables X are assumed to cause an underlying effect, measured by variable Y. Based on collaborative work with social scientists and economists, on an examination of some of their literature, and on conversations with a diverse group of researchers, I conjecture that, whether statisticians like it or not, the type of statistical models used for testing causal hypotheses in the social sciences are almost always association-based models applied to observational data. Regression models are the most common example. The justification for this practice is that the theory itself provides the causality. In other words, the role of the theory is very strong and the reliance on data and statistical modeling are strictly through the lens of the theoretical model. The theory–data relationship varies in different fields. While the social sciences are very theory-heavy, in areas such as bioinformatics and natural language processing the emphasis on a causal theory is much weaker. Hence, given this reality, I define *explaining* as causal explanation and *explanatory modeling* as the use of statistical models for testing causal explanations.

To illustrate how explanatory modeling is typically done, I describe the structure of a typical article in a highly regarded journal in the field of Information Systems (IS). Researchers in the field of IS usually have training in economics and/or the behavioral sciences. The structure of articles reflects the way empirical research is conducted in IS and related fields.

The example used is an article by Gefen, Karahanna and Straub (2003), which studies technology acceptance. The article starts with a presentation of the prevailing relevant theory(ies):

> Online purchase intensions should be explained in part by the technology acceptance model (TAM). This theoretical model is at present a preeminent theory of technology acceptance in IS.

The authors then proceed to state multiple causal hypotheses (denoted H_1, H_2, ... in Figure 1, right panel), justifying the merits for each hypothesis and grounding it in theory. The research hypotheses are given in terms of theoretical *constructs* rather than measurable variables. Unlike measurable variables, constructs are abstractions that "describe a phenomenon of theoretical interest" (Edwards and Bagozzi, 2000) and can be observable or unobservable. Examples of constructs in this article are trust, perceived usefulness (PU), and perceived ease of use (PEOU). Examples of constructs used in other fields include anger, poverty, well-being, and odor. The hypotheses section will often include a causal diagram illustrating the hypothesized causal relationship between the constructs (see Figure 1, left panel). The next step is *construct operationalization*, where a bridge is built between theoretical constructs and observable measurements, using previous literature and theoretical justification. Only after the theoretical component is completed, and measurements are justified and defined, do researchers proceed to the next



H_1: *PU will positively affect intended use of a business-to-consumer (B2C) Web site.*

H_2: *PEOU will positively affect intended use of a business-to-consumer (B2C) Web site.*

H_3: *PEOU will positively affect PU of a business-to-consumer (B2C) Web site.*

FIG. 1. *Causal diagram* (left) *and partial list of stated hypotheses* (right) *from Gefen, Karahanna and Straub* (2003).

step where data and statistical modeling are introduced alongside the statistical hypotheses, which are operationalized from the research hypotheses. Statistical inference will lead to "statistical conclusions" in terms of effect sizes and statistical significance in relation to the causal hypotheses. Finally, the statistical conclusions are converted into research conclusions, often accompanied by policy recommendations.

In summary, *explanatory modeling* refers here to the application of statistical models to data for testing causal hypotheses about theoretical constructs. Whereas "proper" statistical methodology for testing causality exists, such as designed experiments or specialized causal inference methods for observational data [e.g., causal diagrams (Pearl, 1995), discovery algorithms (Spirtes, Glymour and Scheines, 2000), probability trees (Shafer, 1996), and propensity scores (Rosenbaum and Rubin, 1983; Rubin, 1997)], in practice association-based statistical models, applied to observational data, are most commonly used for that purpose.

1.2 Predictive Modeling

I define *predictive modeling* as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations. In particular, I focus on nonstochastic prediction (Geisser, 1993, page 31), where the goal is to predict the output value (Y) for new observations given their input values (X). This definition also includes temporal forecasting, where observations until time t (the input) are used to forecast future values at time $t + k, k > 0$ (the output). *Predictions* include point or interval predictions, prediction regions, predictive distributions, or rankings of new observations. *Predictive model* is any method that produces predictions, regardless of its underlying approach: Bayesian or frequentist, parametric or nonparametric, data mining algorithm or statistical model, etc.

1.3 Descriptive Modeling

Although not the focus of this article, a third type of modeling, which is the most commonly used and developed by statisticians, is descriptive modeling. This type of modeling is aimed at summarizing or representing the data structure in a compact manner. Unlike explanatory modeling, in descriptive modeling the reliance on an underlying causal theory is absent or incorporated in a less formal way. Also, the focus is at the measurable level rather than at the construct level. Unlike predictive modeling, descriptive modeling is not aimed at prediction. Fitting a regression model can be descriptive if it is used for capturing the association between the dependent and independent variables rather than for causal inference or for prediction. We mention this type of modeling to avoid confusion with causal-explanatory and predictive modeling, and also to highlight the different approaches of statisticians and nonstatisticians.

1.4 The Scientific Value of Predictive Modeling

Although explanatory modeling is commonly used for theory building and testing, predictive modeling is nearly absent in many scientific fields as a tool for developing theory. One possible reason is the statistical training of nonstatistician researchers. A look at many introductory statistics textbooks reveals very little in the way of prediction. Another reason is that prediction is often considered unscientific. Berk (2008) wrote, "In the social sciences, for example, one either did causal modeling econometric style or largely gave up quantitative work." From conversations with colleagues in various disciplines it appears that predictive modeling is often valued for its applied utility, yet is discarded for scientific purposes such as theory building or testing. Shmueli and Koppius (2010) illustrated the lack of predictive modeling in the field of IS. Searching the 1072 papers published in the two top-rated journals *Information Systems Research* and *MIS Quarterly* between 1990 and 2006, they found only 52 empirical papers with predictive claims, of which only seven carried out proper predictive modeling or testing.

Even among academic statisticians, there appears to be a divide between those who value prediction as the main purpose of statistical modeling and those who see it as unacademic. Examples of statisticians who emphasize predictive methodology include Akaike ("The predictive point of view is a prototypical point of view to explain the basic activity of statistical analysis" in Findley and Parzen, 1998), Deming ("The only useful function of a statistician is to make predictions" in Wallis, 1980), Geisser ("The prediction of observables or potential observables is of much greater relevance than the estimate of what are often artificial constructs-parameters," Geisser, 1975), Aitchison and Dunsmore ("prediction analysis... is surely at the heart of many statistical applications," Aitchison and Dunsmore, 1975) and Friedman ("One of the most common and important uses for data is prediction," Friedman, 1997). Examples of those who see it as unacademic are Kendall and Stuart ("The Science of Statistics deals with the properties of populations. In considering

a population of men we are not interested, statistically speaking, in whether some particular individual has brown eyes or is a forger, but rather in how many of the individuals have brown eyes or are forgers," Kendall and Stuart, 1977) and more recently Parzen ("The two goals in analyzing data... I prefer to describe as "management" and "science." Management seeks profit... Science seeks truth," Parzen, 2001). In economics there is a similar disagreement regarding "whether prediction per se is a legitimate objective of economic science, and also whether observed data should be used only to shed light on existing theories or also for the purpose of hypothesis seeking in order to develop new theories" (Feelders, 2002).

Before proceeding with the discrimination between explanatory and predictive modeling, it is important to establish prediction as a necessary scientific endeavor beyond utility, for the purpose of developing and testing theories. Predictive modeling and predictive testing serve several necessary scientific functions:

1. Newly available large and rich datasets often contain complex relationships and patterns that are hard to hypothesize, especially given theories that exclude newly measurable concepts. Using predictive modeling in such contexts can help uncover potential new causal mechanisms and lead to the generation of new hypotheses. See, for example, the discussion between Gurbaxani and Mendelson (1990, 1994) and Collopy, Adya and Armstrong (1994).
2. The development of new theory often goes hand in hand with the development of new measures (Van Maanen, Sorensen and Mitchell, 2007). Predictive modeling can be used to discover new measures as well as to compare different operationalizations of constructs and different measurement instruments.
3. By capturing underlying complex patterns and relationships, predictive modeling can suggest improvements to existing explanatory models.
4. Scientific development requires empirically rigorous and relevant research. Predictive modeling enables assessing the distance between theory and practice, thereby serving as a "reality check" to the relevance of theories.[1] While explanatory power provides information about the strength of an underlying causal relationship, it does not imply its predictive power.
5. Predictive power assessment offers a straightforward way to compare competing theories by examining the predictive power of their respective explanatory models.
6. Predictive modeling plays an important role in quantifying the level of predictability of measurable phenomena by creating benchmarks of predictive accuracy (Ehrenberg and Bound, 1993). Knowledge of un-predictability is a fundamental component of scientific knowledge (see, e.g., Taleb, 2007). Because predictive models tend to have higher predictive accuracy than explanatory statistical models, they can give an indication of the potential level of predictability. A very low predictability level can lead to the development of new measures, new collected data, and new empirical approaches. An explanatory model that is close to the predictive benchmark may suggest that our understanding of that phenomenon can only be increased marginally. On the other hand, an explanatory model that is very far from the predictive benchmark would imply that there are substantial practical and theoretical gains to be had from further scientific development.

For a related, more detailed discussion of the value of prediction to scientific theory development see the work of Shmueli and Koppius (2010).

1.5 Explaining and Predicting Are Different

In the philosophy of science, it has long been debated whether explaining and predicting are one or distinct. The conflation of explanation and prediction has its roots in philosophy of science literature, particularly the influential hypothetico-deductive model (Hempel and Oppenheim, 1948), which explicitly equated prediction and explanation. However, as later became clear, the type of uncertainty associated with explanation is of a different nature than that associated with prediction (Helmer and Rescher, 1959). This difference highlighted the need for developing models geared specifically toward dealing with predicting future events and trends such as the Delphi method (Dalkey and Helmer, 1963). The distinction between the two concepts has been further elaborated (Forster and Sober, 1994; Forster, 2002; Sober, 2002; Hitchcock and Sober, 2004; Dowe, Gardner and Oppy, 2007). In his book *Theory Building*, Dubin (1969, page 9) wrote:

> Theories of social and human behavior address themselves to two distinct goals of

[1] Predictive models are advantageous in terms of negative empiricism: a model either predicts accurately or it does not, and this can be observed. In contrast, explanatory models can never be confirmed and are harder to contradict.

> science: (1) prediction and (2) understanding. It will be argued that these are separate goals [...] I will not, however, conclude that they are either inconsistent or incompatible.

Herbert Simon distinguished between "basic science" and "applied science" (Simon, 2001), a distinction similar to explaining versus predicting. According to Simon, basic science is aimed at knowing ("to describe the world") and understanding ("to provide explanations of these phenomena"). In contrast, in applied science, "Laws connecting sets of variables allow inferences or predictions to be made from known values of some of the variables to unknown values of other variables."

Why should there be a difference between explaining and predicting? The answer lies in the fact that measurable data are not accurate representations of their underlying constructs. The operationalization of theories and constructs into statistical models and measurable data creates a disparity between the ability to explain phenomena at the conceptual level and the ability to generate predictions at the measurable level.

To convey this disparity more formally, consider a theory postulating that construct $\mathcal{X}$ causes construct $\mathcal{Y}$, via the function $\mathcal{F}$, such that $\mathcal{Y} = \mathcal{F}(\mathcal{X})$. $\mathcal{F}$ is often represented by a path model, a set of qualitative statements, a plot (e.g., a supply and demand plot), or mathematical formulas. Measurable variables $\mathbf{X}$ and Y are operationalizations of $\mathcal{X}$ and $\mathcal{Y}$, respectively. The operationalization of $\mathcal{F}$ into a statistical model f, such as $E(Y) = f(\mathbf{X})$, is done by considering $\mathcal{F}$ in light of the study design (e.g., numerical or categorical Y; hierarchical or flat design; time series or cross-sectional; complete or censored data) and practical considerations such as standards in the discipline. Because $\mathcal{F}$ is usually not sufficiently detailed to lead to a single f, often a set of f models is considered. Feelders (2002) described this process in the field of economics. In the predictive context, we consider only $\mathbf{X}$, Y and f.

The disparity arises because the goal in explanatory modeling is to match f and $\mathcal{F}$ as closely as possible for the statistical inference to apply to the theoretical hypotheses. The data $\mathbf{X}$, Y are tools for estimating f, which in turn is used for testing the causal hypotheses. In contrast, in predictive modeling the entities of interest are $\mathbf{X}$ and Y, and the function f is used as a tool for generating good predictions of new Y values. In fact, we will see that even if the underlying causal relationship is indeed $\mathcal{Y} = \mathcal{F}(\mathcal{X})$, a function other than $\hat{f}(\mathbf{X})$ and data other than $\mathbf{X}$ might be preferable for prediction.

The disparity manifests itself in different ways. Four major aspects are:

Causation–Association: In explanatory modeling f represents an underlying causal function, and X is assumed to cause Y. In predictive modeling f captures the association between $\mathcal{X}$ and $\mathcal{Y}$.

Theory–Data: In explanatory modeling, f is carefully constructed based on $\mathcal{F}$ in a fashion that supports interpreting the estimated relationship between X and Y and testing the causal hypotheses. In predictive modeling, f is often constructed from the data. Direct interpretability in terms of the relationship between X and Y is not required, although sometimes transparency of f is desirable.

Retrospective–Prospective: Predictive modeling is forward-looking, in that f is constructed for predicting new observations. In contrast, explanatory modeling is retrospective, in that f is used to test an already existing set of hypotheses.

Bias–Variance: The expected prediction error for a new observation with value x, using a quadratic loss function,[2] is given by Hastie, Tibshirani and Friedman (2009, page 223)

$$\begin{aligned} \text{EPE} &= E\{Y - \hat{f}(x)\}^2 \\ &= E\{Y - f(x)\}^2 + \{E(\hat{f}(x)) - f(x)\}^2 \\ &\quad + E\{\hat{f}(x) - E(\hat{f}(x))\}^2 \\ &= \text{Var}(Y) + \text{Bias}^2 + \text{Var}(\hat{f}(x)). \end{aligned} \tag{1}$$

Bias is the result of misspecifying the statistical model f. Estimation variance (the third term) is the result of using a sample to estimate f. The first term is the error that results even if the model is correctly specified and accurately estimated. The above decomposition reveals a source of the difference between explanatory and predictive modeling: In explanatory modeling the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision. This point is illustrated in the Appendix, showing that the "wrong" model can sometimes predict better than the correct one.

[2]For a binary Y, various 0–1 loss functions have been suggested in place of the quadratic loss function (Domingos, 2000).

The four aspects impact every step of the modeling process, such that the resulting f is markedly different in the explanatory and predictive contexts, as will be shown in Section 2.

1.6 A Void in the Statistics Literature

The philosophical explaining/predicting debate has not been directly translated into statistical language in terms of the practical aspects of the *entire* statistical modeling process.

A search of the statistics literature for discussion of explaining versus predicting reveals a lively discussion in the context of *model selection*, and in particular, the derivation and evaluation of model selection criteria. In this context, Konishi and Kitagawa (2007) wrote:

> There may be no significant difference between the point of view of inferring the true structure and that of making a prediction if an infinitely large quantity of data is available or if the data are noiseless. However, in modeling based on a finite quantity of real data, there is a significant gap between these two points of view, because an optimal model for prediction purposes may be different from one obtained by estimating the 'true model.'

The literature on this topic is vast, and we do not intend to cover it here, although we discuss the major points in Section 2.6.

The focus on prediction in the field of machine learning and by statisticians such as Geisser, Aitchison and Dunsmore, Breiman and Friedman, has highlighted aspects of predictive modeling that are relevant to the explanatory/prediction distinction, although they do not directly contrast explanatory and predictive modeling.[3] The prediction literature raises the importance of evaluating predictive power using holdout data, and the usefulness of algorithmic methods (Breiman, 2001b). The predictive focus has also led to the development of inference tools that generate predictive distributions. Geisser (1993) introduced "predictive inference" and developed it mainly in a Bayesian context. "Predictive likelihood" (see Bjornstad, 1990) is a likelihood-based approach to predictive inference, and Dawid's prequential theory (Dawid, 1984) investigates inference concepts in terms of predictability. Finally, the bias–variance aspect has been pivotal in data mining for understanding the predictive performance of different algorithms and for designing new ones.

Another area in statistics and econometrics that focuses on prediction is time series. Methods have been developed specifically for testing the predictability of a series [e.g., random walk tests or the concept of Granger causality (Granger, 1969)], and evaluating predictability by examining performance on holdout data. The time series literature in statistics is dominated by extrapolation models such as ARIMA-type models and exponential smoothing methods, which are suitable for prediction and description, but not for causal explanation. Causal models for time series are common in econometrics (e.g., Song and Witt, 2000), where an underlying causal theory links constructs, which lead to operationalized variables, as in the cross-sectional case. Yet, to the best of my knowledge, there is no discussion in the statistics time series literature regarding the distinction between predictive and explanatory modeling, aside from the debate in economics regarding the scientific value of prediction.

To conclude, the explanatory/predictive modeling distinction has been discussed directly in the model selection context, but not in the larger context. Areas that focus on developing predictive modeling such as machine learning and statistical time series, and "predictivists" such as Geisser, have considered prediction as a separate issue, and have not discussed its principal and practical distinction from causal explanation in terms of developing and testing theory. The goal of this article is therefore to examine the explanatory versus predictive debate from a statistical perspective, considering how modeling is used by nonstatistician scientists for theory development.

The remainder of the article is organized as follows. In Section 2, I consider each step in the modeling process in terms of the four aspects of the predictive/explanatory modeling distinction: *causation–association, theory–data, retrospective–prospective* and *bias–variance*. Section 3 illustrates some of these differences via two examples. A discussion of the implications of the predict/explain conflation, conclusions, and recommendations are given in Section 4.

2. TWO MODELING PATHS

In the following I examine the process of statistical modeling through the explain/predict lens, from goal definition to model use and reporting. For clarity, I broke down the process into a generic set of steps,

[3] Geisser distinguished between "[statistical] parameters" and "observables" in terms of the objects of interest. His distinction is closely related, but somewhat different from our distinction between theoretical constructs and measurements.

FIG. 2. *Steps in the statistical modeling process.*

as depicted in Figure 2. In each step I point out differences in the choice of methods, criteria, data, and information to consider when the goal is predictive versus explanatory. I also briefly describe the related statistics literature. The conceptual and practical differences invariably lead to a difference between a final explanatory model and a predictive one, even though they may use the same initial data. Thus, a priori determination of the main study goal as either explanatory or predictive[4] is essential to conducting adequate modeling. The discussion in this section assumes that the main research goal has been determined as either explanatory or predictive.

2.1 Study Design and Data Collection

Even at the early stages of study design and data collection, issues of what and how much data to collect, according to what design, and which collection instrument to use are considered differently for prediction versus explanation. Consider sample size. In explanatory modeling, where the goal is to estimate the theory-based f with adequate precision and to use it for inference, statistical power is the main consideration. Reducing bias also requires sufficient data for model specification testing. Beyond a certain amount of data, however, extra precision is negligible for purposes of inference. In contrast, in predictive modeling, f itself is often determined from the data, thereby requiring a larger sample for achieving lower bias and variance. In addition, more data are needed for creating holdout datasets (see Section 2.2). Finally, predicting new individual observations accurately, in a prospective manner, requires more data than retrospective inference regarding population-level parameters, due to the extra uncertainty.

A second design issue is sampling scheme. For instance, in the context of hierarchical data (e.g., sampling students within schools) Afshartous and de Leeuw (2005) noted, "Although there exists an extensive literature on estimation issues in multilevel models, the same cannot be said with respect to prediction." Examining issues of sample size, sample allocation, and multilevel modeling for the purpose of "predicting a future observable y_{*j} in the Jth group of a hierarchial dataset," they found that allocation for estimation versus prediction should be different: "an increase in group size n is often more beneficial with respect to prediction than an increase in the number of groups J... [whereas] estimation is more improved by increasing the number of groups J instead of the group size n." This relates directly to the bias–variance aspect. A related issue is the choice of f in relation to sampling scheme. Afshartous and de Leeuw (2005) found that for their hierarchical data, a hierarchical f, which is more appropriate theoretically, had poorer predictive performance than a nonhierarchical f.

A third design consideration is the choice between experimental and observational settings. Whereas for causal explanation experimental data are greatly preferred, subject to availability and resource constraints, in prediction sometimes observational data are preferable to "overly clean" experimental data, if they better represent the realistic context of prediction in terms of the uncontrolled factors, the noise, the measured response, etc. This difference arises from the theory–data and prospective–retrospective aspects. Similarly, when choosing between primary data (data collected for the purpose of the study) and secondary data (data collected for other purposes), the classic criteria of data recency, relevance, and accuracy (Patzer, 1995) are considered from a different angle. For example, a predictive model requires the secondary data to include the exact $\mathbf{X}, Y$ variables to be used at the time of prediction, whereas for causal explanation different operationalizations of the constructs $\mathcal{X}, \mathcal{Y}$ may be acceptable.

In terms of the data collection instrument, whereas in explanatory modeling the goal is to obtain a reliable and valid instrument such that the data obtained represent the underlying construct adequately (e.g., item response theory in psychometrics), for predictive purposes it is more important to focus on the measurement quality and its meaning in terms of the variable to be predicted.

[4]The main study goal can also be descriptive.

Finally, consider the field of design of experiments: two major experimental designs are factorial designs and response surface methodology (RSM) designs. The former is focused on causal explanation in terms of finding the factors that affect the response. The latter is aimed at prediction—finding the combination of predictors that optimizes Y. Factorial designs employ a linear f for interpretability, whereas RSM designs use optimization techniques and estimate a nonlinear f from the data, which is less interpretable but more predictively accurate.[5]

2.2 Data Preparation

We consider two common data preparation operations: handling missing values and data partitioning.

2.2.1 *Handling missing values.* Most real datasets consist of missing values, thereby requiring one to identify the missing values, to determine the extent and type of missingness, and to choose a course of action accordingly. Although a rich literature exists on data imputation, it is monopolized by an explanatory context. In predictive modeling, the solution strongly depends on whether the missing values are in the training data and/or the data to be predicted. For example, Sarle (1998) noted:

> If you have only a small proportion of cases with missing data, you can simply throw out those cases for purposes of estimation; if you want to make predictions for cases with missing inputs, you don't have the option of throwing those cases out.

Sarle further listed imputation methods that are useful for explanatory purposes but not for predictive purposes and vice versa. One example is using regression models with dummy variables that indicate missingness, which is considered unsatisfactory in explanatory modeling, but can produce excellent predictions. The usefulness of creating missingness dummy variables was also shown by Ding and Simonoff (2010). In particular, whereas the classic explanatory approach is based on the Missing-At-Random, Missing-Completely-At-Random or Not-Missing-At-Random classification (Little and Rubin, 2002), Ding and Simonoff (2010) showed that for predictive purposes the important distinction is whether the missingness depends on Y or not. They concluded:

> In the context of classification trees, the relationship between the missingness and the dependent variable, rather than the standard missingness classification approach of Little and Rubin (2002)... is the most helpful criterion to distinguish different missing data methods.

Moreover, missingness can be a blessing in a predictive context, if it is sufficiently informative of Y (e.g., missingness in financial statements when the goal is to predict fraudulent reporting).

Finally, a completely different approach for handling missing data for prediction, mentioned by Sarle (1998) and further developed by Saar-Tsechansky and Provost (2007), considers the case where to-be-predicted observations are missing some predictor information, such that the missing information can vary across different observations. The proposed solution is to estimate multiple "reduced" models, each excluding some predictors. When predicting an observation with missingness on a certain set of predictors, the model that excludes those predictors is used. This approach means that different reduced models are created for different observations. Although useful for prediction, it is clearly inappropriate for causal explanation.

2.2.2 *Data partitioning.* A popular solution for avoiding overoptimistic predictive accuracy is to evaluate performance not on the training set, that is, the data used to build the model, but rather on a holdout sample which the model "did not see." The creation of a holdout sample can be achieved in various ways, the most commonly used being a random partition of the sample into training and holdout sets. A popular alternative, especially with scarce data, is cross-validation. Other alternatives are resampling methods, such as bootstrap, which can be computationally intensive but avoid "bad partitions" and enable predictive modeling with small datasets.

Data partitioning is aimed at minimizing the combined bias and variance by sacrificing some bias in return for a reduction in sampling variance. A smaller sample is associated with higher bias when f is estimated from the data, which is common in predictive modeling but not in explanatory modeling. Hence, data partitioning is useful for predictive modeling but less so for explanatory modeling. With today's abundance of large datasets, where the bias sacrifice is practically small, data partitioning has become a standard preprocessing step in predictive modeling.

[5] I thank Douglas Montgomery for this insight.

In explanatory modeling, data partitioning is less common because of the reduction in statistical power. When used, it is usually done for the retrospective purpose of assessing the robustness of $\hat{f}$. A rarer yet important use of data partitioning in explanatory modeling is for strengthening model validity, by demonstrating some predictive power. Although one would not expect an explanatory model to be optimal in terms of predictive power, it should show some degree of accuracy (see discussion in Section 4.2).

2.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is a key initial step in both explanatory and predictive modeling. It consists of summarizing the data numerically and graphically, reducing their dimension, and "preparing" for the more formal modeling step. Although the same set of tools can be used in both cases, they are used in a different fashion. In explanatory modeling, exploration is channeled toward the theoretically specified causal relationships, whereas in predictive modeling EDA is used in a more free-form fashion, supporting the purpose of capturing relationships that are perhaps unknown or at least less formally formulated.

One example is how data visualization is carried out. Fayyad, Grinstein and Wierse (2002, page 22) contrasted "exploratory visualization" with "confirmatory visualization":

> Visualizations can be used to explore data, to confirm a hypothesis, or to manipulate a viewer... In exploratory visualization the user does not necessarily know what he is looking for. This creates a dynamic scenario in which interaction is critical... In a confirmatory visualization, the user has a hypothesis that needs to be tested. This scenario is more stable and predictable. System parameters are often predetermined.

Hence, *interactivity*, which supports exploration across a wide and sometimes unknown terrain, is very useful for learning about measurement quality and associations that are at the core of predictive modeling, but much less so in explanatory modeling, where the data are visualized through the theoretical lens.

A second example is numerical summaries. In a predictive context, one might explore a wide range of numerical summaries for all variables of interest, whereas in an explanatory model, the numerical summaries would focus on the theoretical relationships. For example, in order to assess the role of a certain variable as a mediator, its correlation with the response variable and with other covariates is examined by generating specific correlation tables.

A third example is the use of EDA for assessing assumptions of potential models (e.g., normality or multicollinearity) and exploring possible variable transformations. Here, too, an explanatory context would be more restrictive in terms of the space explored.

Finally, dimension reduction is viewed and used differently. In predictive modeling, a reduction in the number of predictors can help reduce sampling variance. Hence, methods such as principal components analysis (PCA) or other data compression methods that are even less interpretable (e.g., singular value decomposition) are often carried out initially. They may later lead to the use of compressed variables (such as the first few components) as predictors, even if those are not easily interpretable. PCA is also used in explanatory modeling, but for a different purpose. For questionnaire data, PCA and exploratory factor analysis are used to determine the validity of the survey instrument. The resulting factors are expected to correspond to the underlying constructs. In fact, the rotation step in factor analysis is specifically aimed at making the factors more interpretable. Similarly, correlations are used for assessing the reliability of the survey instrument.

2.4 Choice of Variables

The criteria for choosing variables differ markedly in explanatory versus predictive contexts.

In explanatory modeling, where variables are seen as operationalized constructs, variable choice is based on the role of the construct in the theoretical causal structure and on the operationalization itself. A broad terminology related to different variable roles exists in various fields: in the social sciences—*antecedent, consequent, mediator* and *moderator*[6] variables; in pharmacology and medical sciences—*treatment* and *control* variables; and in epidemiology—*exposure* and *confounding* variables. Carte and Craig (2003) mentioned that explaining moderating effects has become an important scientific endeavor in the field of Management Information Systems. Another important term common in economics is *endogeneity* or "reverse causation," which results in biased parameter estimates. Endogeneity can occur due to different reasons. One

[6] "A moderator variable is one that influences the strength of a relationship between two other variables, and a mediator variable is one that explains the relationship between the two other variables" (from http://psych.wisc.edu/henriques/mediator.html).

reason is incorrectly omitting an input variable, say Z, from f when the causal construct $\mathcal{Z}$ is assumed to cause $\mathcal{X}$ and $\mathcal{Y}$. In a regression model of Y on X, the omission of Z results in X being correlated with the error term. Winkelmann (2008) gave the example of a hypothesis that health insurance ($\mathcal{X}$) affects the demand for health services $\mathcal{Y}$. The operationalized variables are "health insurance status" (X) and "number of doctor consultations" (Y). Omitting an input measurement Z for "true health status" ($\mathcal{Z}$) from the regression model f causes endogeneity because X can be determined by Y (i.e., reverse causation), which manifests as X being correlated with the error term in f. Endogeneity can arise due to other reasons such as measurement error in X. Because of the focus in explanatory modeling on causality and on bias, there is a vast literature on detecting endogeneity and on solutions such as constructing instrumental variables and using models such as two-stage-least-squares (2SLS). Another related term is *simultaneous causality*, which gives rise to special models such as Seemingly Unrelated Regression (SUR) (Zellner, 1962). In terms of chronology, a causal explanatory model can include only "control" variables that take place before the causal variable (Gelman et al., 2003). And finally, for reasons of model identifiability (i.e., given the statistical model, each causal effect can be identified), one is required to include main effects in a model that contains an interaction term between those effects. We note this practice because it is not necessary or useful in the predictive context, due to the acceptability of uninterpretable models and the potential reduction in sampling variance when dropping predictors (see, e.g., the Appendix).

In predictive modeling, the focus on association rather than causation, the lack of $\mathcal{F}$, and the prospective context, mean that there is no need to delve into the exact role of each variable in terms of an underlying causal structure. Instead, criteria for choosing predictors are quality of the association between the predictors and the response, data quality, and availability of the predictors at the time of prediction, known as ex-ante availability. In terms of ex-ante availability, whereas chronological precedence of X to Y is necessary in causal models, in predictive models not only must X precede Y, but X must be available at the time of prediction. For instance, explaining wine quality retrospectively would dictate including barrel characteristics as a causal factor. The inclusion of barrel characteristics in a predictive model of future wine quality would be impossible if at the time of prediction the grapes are still on the vine. See the eBay example in Section 3.2 for another example.

2.5 Choice of Methods

Considering the four aspects of causation–association, theory–data, retrospective–prospective and bias–variance leads to different choices of plausible methods, with a much larger array of methods useful for prediction. Explanatory modeling requires interpretable statistical models f that are easily linked to the underlying theoretical model $\mathcal{F}$. Hence the popularity of statistical models, and especially regression-type methods, in many disciplines. Algorithmic methods such as neural networks or k-nearest-neighbors, and uninterpretable nonparametric models, are considered ill-suited for explanatory modeling.

In predictive modeling, where the top priority is generating accurate predictions of new observations and f is often unknown, the range of plausible methods includes not only statistical models (interpretable and uninterpretable) but also data mining algorithms. A neural network algorithm might not shed light on an underlying causal mechanism $\mathcal{F}$ or even on f, but it can capture complicated associations, thereby leading to accurate predictions. Although model transparency might be important in some cases, it is of secondary importance: "Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why" (Breiman, 2001b).

Breiman (2001b) accused the statistical community of ignoring algorithmic modeling:

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models.

From the explanatory/predictive view, algorithmic modeling is indeed very suitable for predictive (and descriptive) modeling, but not for explanatory modeling.

Some methods are not suitable for prediction from the retrospective–prospective aspect, especially in time series forecasting. Models with coincident indicators, which are measured simultaneously, are such a class. An example is the model $Airfare_t = f(OilPrice_t)$, which might be useful for explaining the effect of oil price on airfare based on a causal theory, but not for predicting future airfare because the oil price at the

time of prediction is unknown. For prediction, alternative models must be considered, such as using a lagged *OilPrice* variable, or creating a separate model for forecasting oil prices and plugging its forecast into the airfare model. Another example is the centered moving average, which requires the availability of data during a time window before and after a period of interest, and is therefore not useful for prediction.

Lastly, the bias–variance aspect raises two classes of methods that are very useful for prediction, but not for explanation. The first is shrinkage methods such as ridge regression, principal components regression, and partial least squares regression, which "shrink" predictor coefficients or even eliminate them, thereby introducing bias into f, for the purpose of reducing estimation variance. The second class of methods, which "have been called the most influential development in Data Mining and Machine Learning in the past decade" (Seni and Elder, 2010, page vi), are ensemble methods such as bagging (Breiman, 1996), random forests (Breiman, 2001a), boosting[7] (Schapire, 1999), variations of those methods, and Bayesian alternatives (e.g., Brown, Vannucci and Fearn, 2002). Ensembles combine multiple models to produce more precise predictions by averaging predictions from different models, and have proven useful in numerous applications (see the Netflix Prize example in Section 3.1).

2.6 Validation, Model Evaluation and Model Selection

Choosing the final model among a set of models, validating it, and evaluating its performance, differ markedly in explanatory and predictive modeling. Although the process is iterative, I separate it into three components for ease of exposition.

2.6.1 *Validation.* In explanatory modeling, validation consists of two parts: *model validation* validates that f adequately represents $\mathcal{F}$, and *model fit* validates that $\hat{f}$ fits the data $\{X, Y\}$. In contrast, validation in predictive modeling is focused on *generalization*, which is the ability of $\hat{f}$ to predict new data $\{X_{\text{new}}, Y_{\text{new}}\}$.

Methods used in explanatory modeling for model validation include model specification tests such as the popular Hausman specification test in econometrics (Hausman, 1978), and construct validation techniques such as reliability and validity measures of survey questions and factor analysis. Inference for individual coefficients is also used for detecting over- or underspecification. Validating model fit involves goodness-of-fit tests (e.g., normality tests) and model diagnostics such as residual analysis. Although indications of lack of fit might lead researchers to modify f, modifications are made carefully in light of the relationship with $\mathcal{F}$ and the constructs $\mathcal{X}, Y$.

In predictive modeling, the biggest danger to generalization is overfitting the training data. Hence validation consists of evaluating the degree of overfitting, by comparing the performance of $\hat{f}$ on the training and holdout sets. If performance is significantly better on the training set, overfitting is implied.

Not only is the large context of validation markedly different in explanatory and predictive modeling, but so are the details. For example, checking for multicollinearity is a standard operation in assessing model fit. This practice is relevant in explanatory modeling, where multicollinearity can lead to inflated standard errors, which interferes with inference. Therefore, a vast literature exists on strategies for identifying and reducing multicollinearity, variable selection being one strategy. In contrast, for predictive purposes "multicollinearity is not quite as damning" (Vaughan and Berry, 2005). Makridakis, Wheelwright and Hyndman (1998, page 288) distinguished between the role of multicollinearity in explaining versus its role in predicting:

> Multicollinearity is not a problem unless either (i) the individual regression coefficients are of interest, or (ii) attempts are made to isolate the contribution of one explanatory variable to Y, *without* the influence of the other explanatory variables. Multicollinearity will not affect the ability of the model to predict.

Another example is the detection of influential observations. While classic methods are aimed at detecting observations that are influential in terms of model estimation, Johnson and Geisser (1983) proposed a method for detecting influential observations in terms of their effect on the predictive distribution.

2.6.2 *Model evaluation.* Consider two performance aspects of a model: explanatory power and predictive power. The top priority in terms of model performance in explanatory modeling is assessing *explanatory power*, which measures the *strength of relationship* indicated by $\hat{f}$. Researchers report R^2-type values

[7] Although boosting algorithms were developed as ensemble methods, "[they can] be seen as an interesting regularization scheme for estimating a model" (Bohlmann and Hothorn, 2007).

and statistical significance of overall F-type statistics to indicate the level of explanatory power.

In contrast, in predictive modeling, the focus is on *predictive accuracy* or *predictive power*, which refer to the performance of $\hat{f}$ on new data. Measures of predictive power are typically out-of-sample metrics or their in-sample approximations, which depend on the type of required prediction. For example, predictions of a binary Y could be binary classifications ($\hat{Y} = 0, 1$), predicted probabilities of a certain class [$\hat{P}(Y = 1)$], or rankings of those probabilities. The latter are common in marketing and personnel psychology. These three different types of predictions would warrant different performance metrics. For example, a model can perform poorly in producing binary classifications but adequately in producing rankings. Moreover, in the context of asymmetric costs, where costs are heftier for some types of prediction errors than others, alternative performance metrics are used, such as the "average cost per predicted observation."

A common misconception in various scientific fields is that predictive power can be inferred from explanatory power. However, the two are different and should be assessed separately. While predictive power can be assessed for both explanatory and predictive models, explanatory power is not typically possible to assess for predictive models because of the lack of $\mathcal{F}$ and an underlying causal structure. Measures such as R^2 and F would indicate the level of association, but not causation.

Predictive power is assessed using metrics computed from a holdout set or using cross-validation (Stone, 1974; Geisser, 1975). Thus, a major difference between explanatory and predictive performance metrics is *the data from which they are computed*. In general, measures computed from the data to which the model was fitted tend to be overoptimistic in terms of predictive accuracy: "Testing the procedure on the data that gave it birth is almost certain to overestimate performance" (Mosteller and Tukey, 1977). Thus, the holdout set serves as a more realistic context for evaluating predictive power.

2.6.3 *Model selection.* Once a set of models f_1, f_2, ... has been estimated and validated, model selection pertains to choosing among them. Two main differentiating aspects are the data–theory and bias–variance considerations. In explanatory modeling, the models are compared in terms of explanatory power, and hence the popularity of nested models, which are easily compared. Stepwise-type methods, which use overall F statistics to include and/or exclude variables, might appear suitable for achieving high explanatory power. However, optimizing explanatory power in this fashion conceptually contradicts the validation step, where variable inclusion/exclusion and the structure of the statistical model are carefully designed to represent the theoretical model. Hence, proper explanatory model selection is performed in a constrained manner. In the words of Jaccard (2001):

> Trimming potentially theoretically meaningful variables is not advisable unless one is quite certain that the coefficient for the variable is near zero, that the variable is inconsequential, and that trimming will not introduce misspecification error.

A researcher might choose to retain a causal covariate which has a strong theoretical justification *even if is statistically insignificant*. For example, in medical research, a covariate that denotes whether a person smokes or not is often present in models for health conditions, whether it is statistically significant or not.[8] In contrast to explanatory power, statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, *even if they are statistically significant*, results in improved prediction accuracy (Greenberg and Parks, 1997; Wu, Harris and McAuley, 2007, and see the Appendix). Stepwise-type algorithms are very useful in predictive modeling as long as the selection criteria rely on predictive power rather than explanatory power.

As mentioned in Section 1.6, the statistics literature on model selection includes a rich discussion on the difference between finding the "true" model and finding the best predictive model, and on criteria for explanatory model selection versus predictive model selection. A popular predictive metric is the in-sample Akaike Information Criterion (AIC). Akaike derived the AIC from a predictive viewpoint, where the model is not intended to accurately infer the "true distribution," but rather to predict future data as accurately as possible (see, e.g., Berk, 2008; Konishi and Kitagawa, 2007). Some researchers distinguish between AIC and the Bayesian information criterion (BIC) on this ground. Sober (2002) concluded that AIC measures predictive accuracy while BIC measures goodness of fit:

[8] I thank Ayala Cohen for this example.

> In a sense, the AIC and the BIC provide estimates of different things; yet, they almost always are thought to be in competition. If the question of which estimator is better is to make sense, we must decide whether the average likelihood of a family [=BIC] or its predictive accuracy [=AIC] is what we want to estimate.

Similarly, Dowe, Gardner and Oppy (2007) contrasted the two Bayesian model selection criteria Minimum Message Length (MML) and Minimum Expected Kullback–Leibler Distance (MEKLD). They concluded,

> If you want to maximise predictive accuracy, you should minimise the expected KL distance (MEKLD); if you want the best inference, you should use MML.

Kadane and Lazar (2004) examined a variety of model selection criteria from a Bayesian decision–theoretic point of view, comparing prediction with explanation goals.

Even when using predictive metrics, the fashion in which they are used within a model selection process can deteriorate their adequacy, yielding overoptimistic predictive performance. Berk (2008) described the case where

> statistical learning procedures are often applied several times to the data with one or more tuning parameters varied. The AIC may be computed for each. But each AIC is ignorant about the information obtained from prior fitting attempts and how many degrees of freedom were expended in the process. Matters are even more complicated if some of the variables are transformed or recoded... Some unjustified optimism remains.

2.7 Model Use and Reporting

Given all the differences that arise in the modeling process, the resulting predictive model would obviously be very different from a resulting explanatory model in terms of the data used ($\{X, Y\}$), the estimated model $\hat{f}$, and explanatory power and predictive power. The use of $\hat{f}$ would also greatly differ.

As illustrated in Section 1.1, explanatory models in the context of scientific research are used to derive "statistical conclusions" using inference, which in turn are translated into scientific conclusions regarding $\mathcal{F}, X, Y$ and the causal hypotheses. With a focus on theory, causality, bias and retrospective analysis, explanatory studies are aimed at testing or comparing existing causal theories. Accordingly the statistical section of explanatory scientific papers is dominated by statistical inference.

In predictive modeling $\hat{f}$ is used to generate predictions for new data. We note that generating predictions from $\hat{f}$ can range in the level of difficulty, depending on the complexity of $\hat{f}$ and on the type of prediction generated. For example, generating a complete predictive distribution is easier using a Bayesian approach than the predictive likelihood approach.

In practical applications, the predictions might be the final goal. However, the focus here is on predictive modeling for supporting scientific research, as was discussed in Section 1.2. Scientific predictive studies and articles therefore emphasize data, association, bias–variance considerations, and prospective aspects of the study. Conclusions pertain to theory-building aspects such as new hypothesis generation, practical relevance, and predictability level. Whereas explanatory articles focus on theoretical constructs and unobservable parameters and their statistical section is dominated by inference, predictive articles concentrate on the observable level, with predictive power and its comparison across models being the core.

3. TWO EXAMPLES

Two examples are used to broadly illustrate the differences that arise in predictive and explanatory studies. In the first I consider a predictive goal and discuss what would be involved in "converting" it to an explanatory study. In the second example I consider an explanatory study and what would be different in a predictive context. See the work of Shmueli and Koppius (2010) for a detailed example "converting" the explanatory study of Gefen, Karahanna and Straub (2003) from Section 1 into a predictive one.

3.1 Netflix Prize

Netflix is the largest online DVD rental service in the United States. In an effort to improve their movie recommendation system, in 2006 Netflix announced a contest (http://netflixprize.com), making public a huge dataset of user movie ratings. Each observation consisted of a user ID, a movie title, and the rating that the user gave this movie. The task was to accurately predict the ratings of movie-user pairs for a test set such

that the predictive accuracy improved upon Netflix's recommendation engine by at least 10%. The grand prize was set at $ 1,000,000. The 2009 winner was a composite of three teams, one of them from the AT&T research lab (see Bell, Koren and Volinsky, 2010). In their 2008 report, the AT&T team, who also won the 2007 and 2008 progress prizes, described their modeling approach (Bell, Koren and Volinsky, 2008).

Let me point out several operations and choices described by Bell, Koren and Volinsky (2008) that highlight the distinctive predictive context. Starting with sample size, the very large sample released by Netflix was aimed at allowing the estimation of f from the data, reflecting the absence of a strong theory. In the data preparation step, with relation to missingness that is predictively informative, the team found that "the information on which movies each user chose to rate, regardless of specific rating value" turned out to be useful. At the data exploration and reduction step, many teams including the winners found that the noninterpretable Singular Value Decomposition (SVD) data reduction method was key in producing accurate predictions: "It seems that models based on matrix-factorization were found to be most accurate." As for choice of variables, supplementing the Netflix data with information about the movie (such as actors, director) actually decreased accuracy: "We should mention that not all data features were found to be useful. For example, we tried to benefit from an extensive set of attributes describing each of the movies in the dataset. Those attributes certainly carry a significant signal and can explain some of the user behavior. However, we concluded that they could not help at all for improving the accuracy of well tuned collaborative filtering models." In terms of choice of methods, their solution was an ensemble of methods that included nearest-neighbor algorithms, regression models, and shrinkage methods. In particular, they found that "using increasingly complex models is only one way of improving accuracy. An apparently easier way to achieve better accuracy is by blending multiple simpler models." And indeed, more accurate predictions were achieved by collaborations between competing teams who combined predictions from their individual models, such as the winners' combined team. All these choices and discoveries are very relevant for prediction, but not for causal explanation. Although the Netflix contest is not aimed at scientific advancement, there is clearly scientific value in the predictive models developed. They tell us about the level of predictability of online user ratings of movies, and the implicated usefulness of the rating scale employed by Netflix. The research also highlights the importance of knowing which movies a user does not rate. And importantly, it sets the stage for explanatory research.

Let us consider a hypothetical goal of *explaining* movie preferences. After stating causal hypotheses, we would define constructs that link user behavior and movie features $\mathcal{X}$ to user preference $\mathcal{Y}$, with a careful choice of $\mathcal{F}$. An operationalization step would link the constructs to measurable data, and the role of each variable in the causality structure would be defined. Even if using the Netflix dataset, supplemental covariates that capture movie features and user characteristics would be absolutely necessary. In other words, the data collected and the variables included in the model would be different from the predictive context. As to methods and models, data compression methods such as SVD, heuristic-based predictive algorithms which learn f from the data, and the combination of multiple models would be considered inappropriate, as they lack interpretability with respect to $\mathcal{F}$ and the hypotheses. The choice of f would be restricted to statistical models that can be used for inference, and would directly model issues such as the dependence between records for the same customer and for the same movie. Finally, the model would be validated and evaluated in terms of its explanatory power, and used to conclude about the strength of the causal relationship between various user and movie characteristics and movie preferences. Hence, the explanatory context leads to a completely different modeling path and final result than the predictive context.

It is interesting to note that most competing teams had a background in computer science rather than statistics. Yet, the winning team combines the two disciplines. Statisticians who see the uniqueness and importance of predictive modeling alongside explanatory modeling have the capability of contributing to scientific advancement as well as achieving meaningful practical results (and large monetary awards).

3.2 Online Auction Research

The following example highlights the differences between explanatory and predictive research in online auctions. The predictive approach also illustrates the utility in creating new theory in an area dominated by explanatory modeling.

Online auctions have become a major player in providing electronic commerce services. eBay (www.eBay.com), the largest consumer-to-consumer auction website, enables a global community of buyers and

sellers to easily interact and trade. Empirical research of online auctions has grown dramatically in recent years. Studies using publicly available bid data from websites such as eBay have found many divergences of bidding behavior and auction outcomes compared to ordinary offline auctions and classical auction theory. For instance, according to classical auction theory (e.g., Krishna, 2002), the final price of an auction is determined by a priori information about the number of bidders, their valuation, and the auction format. However, final price determination in online auctions is quite different. Online auctions differ from offline auctions in various ways such as longer duration, anonymity of bidders and sellers, and low barriers of entry. These and other factors lead to new bidding behaviors that are not explained by auction theory. Another important difference is that the total number of bidders in most online auctions is unknown until the auction closes.

Empirical research in online auctions has concentrated in the fields of economics, information systems and marketing. Explanatory modeling has been employed to learn about different aspects of bidder behavior in auctions. A survey of empirical explanatory research on auctions was given by Bajari and Hortacsu (2004). A typical explanatory study relies on game theory to construct $\mathcal{F}$, which can be done in different ways. One approach is to construct a "structural model," which is a mathematical model linking the various constructs. The major construct is "bidder valuation," which is the amount a bidder is willing to pay, and is typically operationalized using his observed placed bids. The structural model and operationalized constructs are then translated into a regression-type model [see, e.g., Sections 5 and 6 in Bajari and Hortacsu (2003)]. To illustrate the use of a statistical model in explanatory auction research, consider the study by Lucking-Reiley et al. (2007) who used a dataset of 461 eBay coin auctions to determine the factors affecting the final auction price. They estimated a set of linear regression models where $Y = \log(Price)$ and X included auction characteristics (the opening bid, the auction duration, and whether a secret reserve price was used), seller characteristics (the number of positive and negative ratings), and a control variable (book value of the coin). One of their four reported models was of the form

$$\begin{aligned}\log(Price) = {} & \beta_0 + \beta_1 \log(BookValue) \\ & + \beta_2 \log(MinBid) + \beta_3 Reserve \\ & + \beta_4 NumDays + \beta_5 PosRating \\ & + \beta_6 NegRating + \varepsilon.\end{aligned}$$

The other three models, or "model specifications," included a modified set of predictors, with some interaction terms and an alternate auction duration measurement. The authors used a censored-Normal regression for model estimation, because some auctions did not receive any bids and therefore the price was truncated at the minimum bid. Typical explanatory aspects of the modeling are:

Choice of variables: Several issues arise from the causal-theoretical context. First is the exclusion of the number of bidders (or bids) as a determinant due to endogeneity considerations, where although it is likely to affect the final price, "it is endogenously determined by the bidders' choices." To verify endogeneity the authors report fitting a separate regression of $Y = Number\ of\ bids$ on all the determinants. Second, the authors discuss operationalization challenges that might result in bias due to omitted variables. In particular, the authors discuss the construct of "auction attractiveness" ($\mathcal{X}$) and their inability to judge measures such as photos and verbal descriptions to operationalize attractiveness.

Model validation: The four model specifications are used for testing the robustness of the hypothesized effect of the construct "auction length" across different operationalized variables such as the continuous number of days and a categorical alternative.

Model evaluation: For each model, its in-sample R^2 is used for determining explanatory power.

Model selection: The authors report the four fitted regression models, including both significant and insignificant coefficients. Retaining the insignificant covariates in the model is for matching f with $\mathcal{F}$.

Model use and reporting: The main focus is on inference for the β's, and the final conclusions are given in causal terms. ("A seller's feedback ratings... have a measurable effect on her auction prices... when a seller chooses to have her auction last for a longer period of days [sic], this significantly increases the auction price on average.")

Although online auction research is dominated by explanatory studies, there have been a few predictive studies developing forecasting models for an auction's final price (e.g., Jank, Shmueli and Wang, 2008; Jap and Naik, 2008; Ghani and Simmons, 2004; Wang, Jank and Shmueli, 2008; Zhang, Jank and Shmueli, 2010). For a brief survey of online auction forecasting research see the work of Jank and Shmueli (2010, Chapter 5). From my involvement in several of these predictive studies, let me highlight the purely predictive aspects that appear in this literature:

Choice of variables: If prediction takes place before or at the start of the auction, then obviously the total number of bids or bidders cannot be included as a predictor. While this variable was also omitted in the explanatory study, the omission was due to a different reason, that is, endogeneity. However, if prediction takes place at time t during an ongoing auction, then the number of bidders/bids present at time t is available and useful for predicting the final price. Even more useful is the time series of the number of bidders from the start of the auction until time t as well as the price curve until time t (Bapna, Jank and Shmueli, 2008).

Choice of methods: Predictive studies in online auctions tend to learn f from the data, using flexible models and algorithmic methods (e.g., CART, k-nearest neighbors, neural networks, functional methods and related nonparametric smoothing-based methods, Kalman filters and boosting (see, e.g., Chapter 5 in Jank and Shmueli, 2010). Many of these are not interpretable, yet have proven to provide high predictive accuracy.

Model evaluation: Auction forecasting studies evaluate predictive power on holdout data. They report performance in terms of out-of-sample metrics such as *MAPE* and *RMSE*, and are compared against other predictive models and benchmarks.

Predictive models for auction price cannot provide direct causal explanations. However, by producing high-accuracy price predictions they shed light on new potential variables that are related to price and on the types of relationships that can be further investigated in terms of causality. For instance, a construct that is not directly measurable but that some predictive models are apparently capturing is competition between bidders.

4. IMPLICATIONS, CONCLUSIONS AND SUGGESTIONS

4.1 The Cost of Indiscrimination to Scientific Research

Currently, in many fields, statistical modeling is used nearly exclusively for causal explanation. The consequence of neglecting to include predictive modeling and testing alongside explanatory modeling is losing the ability to test the relevance of existing theories and to discover new causal mechanisms. Feelders (2002) commented on the field of economics: "The pure hypothesis testing framework of economic data analysis should be put aside to give more scope to learning from the data. This closes the empirical cycle from observation to theory to the testing of theories on new data." The current accelerated rate of social, environmental, and technological changes creates a burning need for new theories and for the examination of old theories in light of the new realities.

A common practice due to the indiscrimination of explanation and prediction is to erroneously infer predictive power from explanatory power, which can lead to incorrect scientific and practical conclusions. Colleagues from various fields confirmed this fact, and a cursory search of their scientific literature brings up many examples. For instance, in ecology an article intending to predict forest beetle assemblages infers predictive power from explanatory power ["To study... predictive power, ... we calculated the R^2"; "We expect predictabilities with R^2 of up to 0.6" (Muller and Brandl, 2009)]. In economics, an article entitled "The predictive power of zero intelligence in financial markets" (Farmer, Patelli and Zovko, 2005) infers predictive power from a high R^2 value of a linear regression model. In epidemiology, many studies rely on in-sample hazard ratios estimated from Cox regression models to infer predictive power, reflecting an indiscrimination between description and prediction. For instance, Nabi et al. (2010) used hazard ratio estimates and statistical significance "to compare the predictive power of depression for coronary heart disease with that of cerebrovascular disease." In information systems, an article on "Understanding and predicting electronic commerce adoption" (Pavlou and Fygenson, 2006) incorrectly compared the predictive power of different models using in-sample measures ("To examine the predictive power of the proposed model, we compare it to four models in terms of R^2 adjusted"). These examples are not singular, but rather they reflect the common misunderstanding of predictive power in these and other fields.

Finally, a consequence of omitting predictive modeling from scientific research is also a gap between research and practice. In an age where empirical research has become feasible in many fields, the opportunity to bridge the gap between methodological development and practical application can be easier to achieve through the combination of explanatory and predictive modeling.

Finance is an example where practice is concerned with prediction whereas academic research is focused on explaining. In particular, there has been a reliance on a limited number of models that are considered

pillars of research, yet have proven to perform very poorly in practice. For instance, the CAPM model and more recently the Fama–French model are regression models that have been used for explaining market behavior for the purpose of portfolio management, and have been evaluated in terms of explanatory power (in-sample R^2 and residual analysis) and not predictive accuracy.[9] More recently, researchers have begun recognizing the distinction between in-sample explanatory power and out-of-sample predictive power (Goyal and Welch, 2007), which has led to a discussion of predictability magnitude and a search for predictively accurate explanatory variables (Campbell and Thompson, 2005). In terms of predictive modeling, the Chief Actuary of the Financial Supervisory Authority of Sweden commented in 1999: "there is a need for models with predictive power for at least a very near future... Given sufficient and relevant data this is an area for statistical analysis, including cluster analysis and various kind of structure-finding methods" (Palmgren, 1999). While there has been some predictive modeling using genetic algorithms (Chen, 2002) and neural networks (Chakraborty and Sharma, 2007), it has been performed by practitioners and nonfinance academic researchers and outside of the top academic journals.

In summary, the omission of predictive modeling for theory development results not only in academic work becoming irrelevant to practice, but also in creating a barrier to achieving significant scientific progress, which is especially unfortunate as data become easier to collect, store and access.

In the opposite direction, in fields that focus on predictive modeling, the reason for omitting explanatory modeling must be sought. A scientific field is usually defined by a cohesive body of theoretical knowledge, which can be tested. Hence, some form of testing, whether empirical or not, must be a component of the field. In areas such as bioinformatics, where there is little theory and an abundance of data, predictive models are pivotal in generating avenues for causal theory.

4.2 Explanatory and Predictive Power: Two Dimensions

I have polarized explaining and predicting in this article in an effort to highlight their fundamental differences. However, rather than considering them as extremes on some continuum, I consider them as two dimensions.[10,11] Explanatory power and predictive accuracy are different qualities; a model will possess some level of each.

A related controversial question arises: must an explanatory model have some level of predictive power to be considered scientifically useful? And equally, must a predictive model have sufficient explanatory power to be scientifically useful? For instance, some explanatory models that cannot be tested for predictive accuracy yet constitute scientific advances are Darwinian evolution theory and string theory in physics. The latter produces currently untestable predictions (Woit, 2006, pages x–xii). Conversely, there exist predictive models that do not properly "explain" yet are scientifically valuable. Galileo, in his book *Two New Sciences*, proposed a demonstration to determine whether light was instantaneous. According to Mackay and Oldford (2000), Descartes gave the book a scathing review:

> The substantive criticisms are generally directed at Galileo's not having identified the causes of the phenomena he investigated. For most scientists at this time, and particularly for Descartes, that is the whole point of science.

Similarly, consider predictive models that are based on a *wrong* explanation yet scientifically and practically they are considered valuable. One well-known example is Ptolemaic astronomy, which until recently was used for nautical navigation but is based on a theory proven to be wrong long ago. While such examples are extreme, in most cases models are likely to possess some level of both explanatory and predictive power.

Considering predictive accuracy and explanatory power as two axes on a two-dimensional plot would place different models (f), aimed either at explanation or at prediction, on different areas of the plot. The bi-dimensional approach implies that: (1) In terms of modeling, the goal of a scientific study must be specified a priori in order to optimize the criterion of interest; and (2) In terms of model evaluation and scientific reporting, researchers should report *both the explanatory and predictive qualities* of their models. Even if prediction is not the goal, the predictive qualities of a model should be reported alongside its explanatory

[9] Although in their paper Fama and French (1993) did split the sample into two parts, they did so for purposes of testing the sensitivity of model estimates rather than for assessing predictive accuracy.

[10] Similarly, descriptive models can be considered as a third dimension, where yet different criteria are used for assessing the strength of the descriptive model.

[11] I thank Bill Langford for the two-dimensional insight.

power so that it can be fairly evaluated in terms of its capabilities and compared to other models. Similarly, a predictive model might not require causal explanation in order to be scientifically useful; however, reporting its relation to causal theory is important for purposes of theory building. The availability of information on a variety of predictive and explanatory models along these two axes can shed light on both predictive and causal aspects of scientific phenomena. The statistical modeling process, as depicted in Figure 2, should include "overall model performance" in terms of both predictive and explanatory qualities.

4.3 The Cost of Indiscrimination to the Field of Statistics

Dissolving the ambiguity surrounding explanatory versus predictive modeling is important for advancing our field itself. Recognizing that statistical methodology has focused mainly on inference indicates an important gap to be filled. While our literature contains predictive methodology for model selection and predictive inference, there is scarce statistical predictive methodology for other modeling steps, such as study design, data collection, data preparation and EDA, which present opportunities for new research. Currently, the predictive void has been taken up the field of machine learning and data mining. In fact, the differences, and some would say rivalry, between the fields of statistics and data mining can be attributed to their different goals of explaining versus predicting even more than to factors such as data size. While statistical theory has focused on model estimation, inference, and fit, machine learning and data mining have concentrated on developing computationally efficient predictive algorithms and tackling the bias–variance trade-off in order to achieve high predictive accuracy.

Sharpening the distinction between explanatory and predictive modeling can raise a new awareness of the strengths and limitations of existing methods and practices, and might shed light on current controversies within our field. One example is the disagreement in survey methodology regarding the use of sampling weights in the analysis of survey data (Little, 2007). Whereas some researchers advocate using weights to reduce bias at the expense of increased variance, and others disagree, might not the answer be related to the final goal?

Another ambiguity that can benefit from an explanatory/predictive distinction is the definition of parsimony. Some claim that predictive models should be simpler than explanatory models: "Simplicity is relevant because complex families often do a bad job of predicting new data, though they can be made to fit the old data quite well" (Sober, 2002). The same argument was given by Hastie, Tibshirani and Friedman (2009): "Typically the more complex we make the model, the lower the bias but the higher the variance." In contrast, some predictive models in practice are very complex,[12] and indeed Breiman (2001b) commented: "in some cases predictive models are more complex in order to capture small nuances that improve predictive accuracy." Zellner (2001) used the term "sophisticatedly simple" to define the quality of a "good" model. I would suggest that the definitions of parsimony and complexity are task-dependent: predictive or explanatory. For example, an "overly complicated" model in explanatory terms might prove "sophisticatedly simple" for predictive purposes.

4.4 Closing Remarks and Suggestions

The consequences from the explanatory/predictive distinction lead to two proposed actions:

1. It is our responsibility to be aware of how statistical models are used in research outside of statistics, why they are used in that fashion, and in response to develop methods that support sound scientific research. Such knowledge can be gained within our field by inviting scientists from different disciplines to give talks at statistics conferences and seminars, and to require graduate students in statistics to read and present research papers from other disciplines.
2. As a discipline, we must acknowledge the difference between explanatory, predictive and descriptive modeling, and integrate it into statistics education of statisticians and nonstatisticians, as early as possible but most importantly in "research methods" courses. This requires creating written materials that are easily accessible and understandable by nonstatisticians. We should advocate both explanatory and predictive modeling, clarify their differences and distinctive scientific and practical uses, and disseminate tools and knowledge for implementing both. One particular aspect to consider is advocating a more careful use of terms such as "predictors," "predictions" and "predictive power," to reduce the effects of terminology on incorrect scientific conclusions.

[12] I thank Foster Provost from NYU for this observation.

Awareness of the distinction between explanatory and predictive modeling, and of the different scientific functions that each serve, is essential for the progress of scientific knowledge.

APPENDIX: IS THE "TRUE" MODEL THE BEST PREDICTIVE MODEL? A LINEAR REGRESSION EXAMPLE

Consider $\mathcal{F}$ to be the true function relating constructs $\mathcal{X}$ and $\mathcal{Y}$ and let us assume that f is a valid operationalization of $\mathcal{F}$. Choosing an intentionally biased function f^* in place of f is clearly undesirable from a theoretical–explanatory point of view. However, we will show that f^* can be preferable to f from a predictive standpoint.

To illustrate this, consider the statistical model $f(x) = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ which is assumed to be correctly specified with respect to $\mathcal{F}$. Using data, we obtain the estimated model $\hat{f}$, which has the properties

$$\text{Bias} = 0, \tag{2}$$

$$\begin{aligned}\operatorname{Var}(\hat{f}(x)) &= \operatorname{Var}(x_1\hat{\beta}_1 + x_2\hat{\beta}_2)\\ &= \sigma^2 x'(X'X)^{-1}x,\end{aligned} \tag{3}$$

where x is the vector $x = [x_1, x_2]'$, and X is the design matrix based on both predictors. Combining the squared bias with the variance gives

$$\begin{aligned}\text{EPE} &= E\big(Y - \hat{f}(x)\big)^2\\ &= \sigma^2 + 0 + \sigma^2 x'(X'X)^{-1}x\\ &= \sigma^2\big(1 + x'(X'X)^{-1}x\big).\end{aligned} \tag{4}$$

In comparison, consider the estimated underspecified form $\hat{f}^*(x) = \hat{\gamma}_1 x_1$. The bias and variance here are given by Montgomery, Peck and Vining (2001, pages 292–296):

$$\begin{aligned}\text{Bias} &= x_1\gamma_1 - (x_1\beta_1 + x_2\beta_2)\\ &= x_1(x_1'x_1)^{-1}x_1'(x_1\beta_1 + x_2\beta_2)\\ &\quad - (x_1\beta_1 + x_2\beta_2),\\ \operatorname{Var}(\hat{f}^*(x)) &= x_1 \operatorname{Var}(\hat{\gamma}_1)x_1 = \sigma^2 x_1(x_1'x_1)^{-1}x_1.\end{aligned}$$

Combining the squared bias with the variance gives

$$\begin{aligned}\text{EPE} &= \big(x_1(x_1'x_1)^{-1}x_1'x_2\beta_2 - x_2\beta_2\big)^2\\ &\quad + \sigma^2\big(1 + x_1(x_1'x_1)^{-1}x_1'\big).\end{aligned} \tag{5}$$

Although the bias of the underspecified model $f^*(x)$ is larger than that of $f(x)$, its variance can be smaller, and in some cases so small that the overall EPE will be lower for the underspecified model. Wu, Harris and McAuley (2007) showed the general result for an underspecified linear regression model with multiple predictors. In particular, they showed that the underspecified model that leaves out q predictors has a lower EPE when the following inequality holds:

$$q\sigma^2 > \beta_2' X_2'(I - H_1)X_2\beta_2. \tag{6}$$

This means that the underspecified model produces more accurate predictions, in terms of lower EPE, in the following situations:

- when the data are very noisy (large σ);
- when the true absolute values of the left-out parameters (in our example β_2) are small;
- when the predictors are highly correlated; and
- when the sample size is small or the range of left-out variables is small.

The bottom line is nicely summarized by Hagerty and Srinivasan (1991): "We note that the practice in applied research of concluding that a model with a higher predictive validity is "truer," is not a valid inference. This paper shows that a parsimonious but less true model can have a higher predictive validity than a truer but less parsimonious model."

ACKNOWLEDGMENTS

I thank two anonymous reviewers, the associate editor, and editor for their suggestions and comments which improved this manuscript. I express my gratitude to many colleagues for invaluable feedback and fruitful discussion that have helped me develop the explanatory/predictive argument presented in this article. I am grateful to Otto Koppius (Erasmus) and Ravi Bapna (U Minnesota) for familiarizing me with explanatory modeling in Information Systems, for collaboratively pursuing prediction in this field, and for tireless discussion of this work. I thank Ayala Cohen (Technion), Ralph Snyder (Monash), Rob Hyndman (Monash) and Bill Langford (RMIT) for detailed feedback on earlier drafts of this article. Special thanks to Boaz Shmueli and Raquelle Azran for their meticulous reading and discussions of the manuscript. And special thanks for invaluable comments and suggestions go to Murray Aitkin (U Melbourne), Yoav Benjamini (Tel Aviv U), Smarajit Bose (ISI), Saibal Chattopadhyay (IIMC), Ram Chellapah (Emory), Etti Doveh (Technion), Paul Feigin (Technion), Paulo Goes (U Arizona), Avi Goldfarb (Toronto U), Norma Hubele (ASU), Ron Kenett (KPA Inc.), Paul Lajbcygier (Monash),

Thomas Lumley (U Washington), David Madigan (Columbia U), Isaac Meilejson (Tel Aviv U), Douglas Montgomery (ASU), Amita Pal (ISI), Don Poskitt (Monash), Foster Provost (NYU), Saharon Rosset (Tel Aviv U), Jeffrey Simonoff (NYU) and David Steinberg (Tel Aviv U).

REFERENCES

AFSHARTOUS, D. and DE LEEUW, J. (2005). Prediction in multilevel models. *J. Educ. Behav. Statist.* **30** 109–139.

AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge Univ. Press. MR0408097

BAJARI, P. and HORTACSU, A. (2003). The winner's curse, reserve prices and endogenous entry: Empirical insights from ebay auctions. *Rand J. Econ.* **3** 329–355.

BAJARI, P. and HORTACSU, A. (2004). Economic insights from internet auctions. *J. Econ. Liter.* **42** 457–486.

BAPNA, R., JANK, W. and SHMUELI, G. (2008). Price formation and its dynamics in online auctions. *Decision Support Systems* **44** 641–656.

BELL, R. M., KOREN, Y. and VOLINSKY, C. (2008). The BellKor 2008 solution to the Netflix Prize.

BELL, R. M., KOREN, Y. and VOLINSKY, C. (2010). All together now: A perspective on the netflix prize. *Chance* **23** 24.

BERK, R. A. (2008). *Statistical Learning from a Regression Perspective*. Springer, New York.

BJORNSTAD, J. F. (1990). Predictive likelihood: A review. *Statist. Sci.* **5** 242–265. MR1062578

BOHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140. MR1425957

BREIMAN, L. (2001a). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L. (2001b). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–215. MR1874152

BROWN, P. J., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 519–536. MR1924304

CAMPBELL, J. Y. and THOMPSON, S. B. (2005). Predicting excess stock returns out of sample: Can anything beat the historical average? Harvard Institute of Economic Research Working Paper 2084.

CARTE, T. A. and CRAIG, J. R. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quart.* **27** 479–501.

CHAKRABORTY, S. and SHARMA, S. K. (2007). Prediction of corporate financial health by artificial neural network. *Int. J. Electron. Fin.* **1** 442–459.

CHEN, S.-H., ED. (2002). *Genetic Algorithms and Genetic Programming in Computational Finance*. Kluwer, Dordrecht.

COLLOPY, F., ADYA, M. and ARMSTRONG, J. (1994). Principles for examining predictive–validity—the case of information-systems spending forecasts. *Inform. Syst. Res.* **5** 170–179.

DALKEY, N. and HELMER, O. (1963). An experimental application of the delphi method to the use of experts. *Manag. Sci.* **9** 458–467.

DAWID, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811

DING, Y. and SIMONOFF, J. (2010). An investigation of missing data methods for classification trees applied to binary response data. *J. Mach. Learn. Res.* **11** 131–170.

DOMINGOS, P. (2000). A unified bias–variance decomposition for zero–one and squared loss. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence* 564–569. AAAI Press, Austin, TX.

DOWE, D. L., GARDNER, S. and OPPY, G. R. (2007). Bayes not bust! Why simplicity is no problem for Bayesians. *Br. J. Philos. Sci.* **58** 709–754. MR2375767

DUBIN, R. (1969). *Theory Building*. The Free Press, New York.

EDWARDS, J. R. and BAGOZZI, R. P. (2000). On the nature and direction of relationships between constructs. *Psychological Methods* **5 2** 155–174.

EHRENBERG, A. and BOUND, J. (1993). Predictability and prediction. *J. Roy. Statist. Soc. Ser. A* **156** 167–206.

FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in stock and bond returns. *J. Fin. Econ.* **33** 3–56.

FARMER, J. D., PATELLI, P. and ZOVKO, I. I. A. A. (2005). The predictive power of zero intelligence in financial markets. *Proc. Natl. Acad. Sci. USA* **102** 2254–2259.

FAYYAD, U. M., GRINSTEIN, G. G. and WIERSE, A. (2002). *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, San Francisco, CA.

FEELDERS, A. (2002). Data mining in economic science. In *Dealing with the Data Flood* 166–175. STT/Beweton, Den Haag, The Netherlands.

FINDLEY, D. Y. and PARZEN, E. (1998). A conversation with Hirotsugo Akaike. In *Selected Papers of Hirotugu Akaike* 3–16. Springer, New York. MR1486823

FORSTER, M. (2002). Predictive accuracy as an achievable goal of science. *Philos. Sci.* **69** S124–S134.

FORSTER, M. and SOBER, E. (1994). How to tell when simpler, more unified, or less ad-hoc theories will provide more accurate predictions. *Br. J. Philos. Sci.* **45** 1–35. MR1277464

FRIEDMAN, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1** 55–77.

GEFEN, D., KARAHANNA, E. and STRAUB, D. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quart.* **27** 51–90.

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

GEISSER, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall, London. MR1252174

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC New York/Boca Raton, FL. MR1385925

GHANI, R. and SIMMONS, H. (2004). Predicting the end-price of online auctions. In *International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management*, Pisa, Italy.

GOYAL, A. and WELCH, I. (2007). A comprehensive look at the empirical performance of equity premium prediction. *Rev. Fin. Stud.* **21** 1455–1508.

GRANGER, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–438.

GREENBERG, E. and PARKS, R. P. (1997). A predictive approach to model selection and multicollinearity. *J. Appl. Econom.* **12** 67–75.

GURBAXANI, V. and MENDELSON, H. (1990). An integrative model of information systems spending growth. *Inform. Syst. Res.* **1** 23–46.

GURBAXANI, V. and MENDELSON, H. (1994). Modeling vs. forecasting—the case of information-systems spending. *Inform. Syst. Res.* **5** 180–190.

HAGERTY, M. R. and SRINIVASAN, S. (1991). Comparing the predictive powers of alternative multiple regression models. *Psychometrika* **56** 77–85. MR1115296

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR1851606

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271. MR0513692

HELMER, O. and RESCHER, N. (1959). On the epistemology of the inexact sciences. *Manag. Sci.* **5** 25–52.

HEMPEL, C. and OPPENHEIM, P. (1948). Studies in the logic of explanation. *Philos. Sci.* **15** 135–175.

HITCHCOCK, C. and SOBER, E. (2004). Prediction versus accommodation and the risk of overfitting. *Br. J. Philos. Sci.* **55** 1–34.

JACCARD, J. (2001). *Interaction Effects in Logistic Regression*. SAGE Publications, Thousand Oaks, CA.

JANK, W. and SHMUELI, G. (2010). *Modeling Online Auctions*. Wiley, New York.

JANK, W., SHMUELI, G. and WANG, S. (2008). Modeling price dynamics in online auctions via regression trees. In *Statistical Methods in eCommerce Research*. Wiley, New York. MR2414052

JAP, S. and NAIK, P. (2008). Bidanalyzer: A method for estimation and selection of dynamic bidding models. *Marketing Sci.* **27** 949–960.

JOHNSON, W. and GEISSER, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Assoc.* **78** 137–144. MR0696858

KADANE, J. B. and LAZAR, N. A. (2004). Methods and criteria for model selection. *J. Amer. Statist. Soc.* **99** 279–290. MR2061890

KENDALL, M. and STUART, A. (1977). *The Advanced Theory of Statistics* **1**, 4th ed. Griffin, London.

KONISHI, S. and KITAGAWA, G. (2007). *Information Criteria and Statistical Modeling*. Springer, New York. MR2367855

KRISHNA, V. (2002). *Auction Theory*. Academic Press, San Diego, CA.

LITTLE, R. J. A. (2007). Should we use the survey weights to weight? JPSM Distinguished Lecture, Univ. Maryland.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York. MR1925014

LUCKING-REILEY, D., BRYAN, D., PRASAD, N. and REEVES, D. (2007). Pennies from ebay: The determinants of price in online auctions. *J. Indust. Econ.* **55** 223–233.

MACKAY, R. J. and OLDFORD, R. W. (2000). Scientific method, statistical method, and the speed of light. Working Paper 2000-02, Dept. Statistics and Actuarial Science, Univ. Waterloo. MR1847825

MAKRIDAKIS, S. G., WHEELWRIGHT, S. C. and HYNDMAN, R. J. (1998). *Forecasting: Methods and Applications*, 3rd ed. Wiley, New York.

MONTGOMERY, D., PECK, E. A. and VINING, G. G. (2001). *Introduction to Linear Regression Analysis*. Wiley, New York. MR1820113

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.

MULLER, J. and BRANDL, R. (2009). Assessing biodiversity by remote sensing in mountainous terrain: The potential of lidar to predict forest beetle assemblages. *J. Appl. Ecol.* **46** 897–905.

NABI, J., KIVIMÄKI, M., SUOMINEN, S., KOSKENVUO, M. and VAHTERA, J. (2010). Does depression predict coronary heart diseaseand cerebrovascular disease equally well? The health and social support prospective cohort study. *Int. J. Epidemiol.* **39** 1016–1024.

PALMGREN, B. (1999). The need for financial models. *ERCIM News* **38** 8–9.

PARZEN, E. (2001). Comment on statistical modeling: The two cultures. *Statist. Sci.* **16** 224–226. MR1874152

PATZER, G. L. (1995). *Using Secondary Data in Marketing Research: United States and Worldwide*. Greenwood Publishing, Westport, CT.

PAVLOU, P. and FYGENSON, M. (2006). Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *Mis Quart.* **30** 115–143.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–709. MR1380809

ROSENBAUM, P. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

RUBIN, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* **127** 757–763.

SAAR-TSECHANSKY, M. and PROVOST, F. (2007). Handling missing features when applying classification models. *J. Mach. Learn. Res.* **8** 1625–1657.

SARLE, W. S. (1998). Prediction with missing inputs. In *JCIS 98 Proceedings* (P. Wang, ed.) **II** 399–402. Research Triangle Park, Durham, NC.

SENI, G. and ELDER, J. F. (2010). *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions (Synthesis Lectures on Data Mining and Knowledge Discovery)*. Morgan and Claypool, San Rafael, CA.

SHAFER, G. (1996). *The Art of Causal Conjecture*. MIT Press, Cambridge, MA.

SCHAPIRE, R. E. (1999). A brief introduction to boosting. In *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* 1401–1406. Stockholm, Sweden.

SHMUELI, G. and KOPPIUS, O. R. (2010). Predictive analytics in information systems research. *MIS Quart.* To appear.

SIMON, H. A. (2001). Science seeks parsimony, not simplicity: Searching for pattern in phenomena. In *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple* 32–72. Cambridge Univ. Press. MR1932928

SOBER, E. (2002). Instrumentalism, parsimony, and the Akaike framework. *Philos. Sci.* **69** S112–S123.

SONG, H. and WITT, S. F. (2000). *Tourism Demand Modelling and Forecasting: Modern Econometric Approaches*. Pergamon Press, Oxford.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

STONE, M. (1974). Cross-validatory choice and assesment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 111–147. MR0356377

TALEB, N. (2007). *The Black Swan*. Penguin Books, London.

VAN MAANEN, J., SORENSEN, J. and MITCHELL, T. (2007). The interplay between theory and method. *Acad. Manag. Rev.* **32** 1145–1154.

VAUGHAN, T. S. and BERRY, K. E. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *J. Statist. Educ.* **13** online.

WALLIS, W. A. (1980). The statistical research group, 1942–1945. *J. Amer. Statist. Assoc.* **75** 320–330. MR0577363

WANG, S., JANK, W. and SHMUELI, G. (2008). Explaining and forecasting online auction prices and their dynamics using functional data analysis. *J. Business Econ. Statist.* **26** 144–160. MR2420144

WINKELMANN, R. (2008). *Econometric Analysis of Count Data*, 5th ed. Springer, New York. MR2148271

WOIT, P. (2006). *Not Even Wrong: The Failure of String Theory and the Search for Unity in Physical Law*. Jonathan Cope, London. MR2245858

WU, S., HARRIS, T. and MCAULEY, K. (2007). The use of simplified or misspecified models: Linear case. *Canad. J. Chem. Eng.* **85** 386–398.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235

ZELLNER, A. (2001). Keep it sophisticatedly simple. In *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple* 242–261. Cambridge Univ. Press. MR1932939

ZHANG, S., JANK, W. and SHMUELI, G. (2010). Real-time forecasting of online auctions via functional k-nearest neighbors. *Int. J. Forecast.* **26** 666–683.