Journal of Business Research 68 (2015) 1772–1784

Contents lists available at ScienceDirect

Journal of Business Research

The bias bias

Henry Brighton *, Gerd Gigerenzer

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

ARTICLE INFO

Article history:
Received 1 March 2014
Received in revised form 1 October 2014
Accepted 1 January 2015
Available online 4 April 2015

Keywords:
Simple heuristics
Uncertainty
Bias–variance dilemma
Occam's razor
Out-of-sample prediction
Bias bias

ABSTRACT

In marketing and finance, surprisingly simple models sometimes predict more accurately than more complex, sophisticated models. Here, we address the question of when and why simple models succeed — or fail — by framing the forecasting problem in terms of the bias–variance dilemma. Controllable error in forecasting consists of two components, the "bias" and the "variance". We argue that the benefits of simplicity are often overlooked because of a pervasive "bias bias": the importance of the bias component of prediction error is inflated, and the variance component of prediction error, which reflects an oversensitivity of a model to different samples from the same population, is neglected. Using the study of cognitive heuristics, we discuss how to reduce variance by ignoring weights, attributes, and dependencies between attributes, and thus make better decisions. Bias and variance, we argue, offer a more insightful perspective on the benefits of simplicity than Occam"s razor.

1. Introduction

Do complex problems require complex solutions? Consider managers in the retail industry who often need to distinguish between active and inactive customers. One strategy is to use observations of past customer activity to estimate the parameters of a sophisticated probabilistic model. For example, the Pareto/NBD model estimates the parameters of a Poisson process modeling customer purchasing behavior and the parameters of exponential distribution modeling customer dropout rates (Schmittlein, Morrison, & Colombo, 1987).

Combined with further probabilistic assumptions about the heterogeneity of customers within the population, categorization decisions are then made using a computationally demanding maximum likelihood calculation (Fader, Hardie, & Lee, 2005). An alternative strategy is to use a simple hiatus rule where customers who have not made a purchase within a hiatus period of, say, 9 months are classified as inactive, and all other customers are categorized as active. Researchers suffering from what we term the "bias bias" place their faith in complex models like the Pareto/NBD model and expect that the simple hiatus rule will perform poorly in comparison.

Putting this intuition to the test, Wübben and Wangenheim (2008) compared the Pareto/NBD model and simple hiatus rules using transaction data from the apparel, airline, and music industries. First, they used 40 weeks of customer transaction data to estimate the parameters of the Pareto/NBD model. Using transaction data for the subsequent 40 weeks, they then estimated how accurately each model predicted future customer activity. For the apparel, airline, and music customers, the Pareto/NBD model achieved predictive accuracies of 75%, 74%, and 77%. Hiatus rules with cutoff periods recommended by experienced managers, on the other hand, predicted customer activity with accuracies of 83%, 77%, and 77%. Here, a simple hiatus rule either matched or outperformed the Pareto/NBD model. This example illustrates the potential of the bias bias to mislead. But what exactly is the bias bias?

Our use of the term bias makes reference to the bias–variance decomposition in understanding the prediction error incurred by statistical models (Bishop, 2006; Geman, Bienenstock, & Doursat, 1992; Hastie, Tibshirani, & Friedman, 2001; O'Sullivan, 1986). The bias component of prediction error reflects the inability of a model to represent the systematic patterns that govern the observations. The variance component of prediction error reflects the sensitivity of the model's predictions to different observations of the same problem, such as a different sample from the same population. Together, bias and variance additively contribute to the total prediction error:

$$\text{Total error} = (\text{bias})^2 + \text{variance} + \text{noise}. \quad (1)$$

The "bias bias" refers to a cluster of commonly held statistical intuitions that consider bias but pay little attention to variance:

> *The bias bias*: To suffer from the bias bias is to develop, deploy, or prefer models that are likely to achieve low bias, while simultaneously paying little or no attention to models with low variance.

For example, the Pareto/NBD model aims to accurately represent the probabilistic structure of the problem. This modeling strategy is likely to

* Corresponding author.
E-mail addresses: hbrighton@mpib-berlin.mpg.de (H. Brighton), gigerenzer@mpib-berlin.mpg.de (G. Gigerenzer).

http://dx.doi.org/10.1016/j.jbusres.2015.01.061

incur low bias, whereas a simple 9-month hiatus rule is likely to incur high bias. However, because a hiatus rule has only one parameter to estimate (the length of the hiatus), it will incur low variance. The ability to incur low variance explains why simple hiatus rules, and other simple models, can outperform models deemed more "accurate" and "sophisticated" (e.g., Wright and Stern, 2015-in this issue). This is why a researcher suffering from the bias bias is likely to assume, incorrectly, that a hiatus rule will achieve lower predictive accuracy than the Pareto/NBD model.

The importance of model simplicity and robustness has long been recognized in management science (Little, 1970). Our concern here is that, even among those aware of the benefits of simplicity, the role of variance reduction is frequently overlooked during the search for predictive models. Table 1 summarizes five common symptoms of this bias bias: (1) relying solely on goodness of fit to evaluate models; (2) equating model complexity, that is, the ability of a model to fit diverse patterns of data, with a model's parametric complexity, that is, the number of parameters estimated by the model; (3) drawing conclusions from a single model; (4) seeking unbiased models without assessing the predictive accuracy of biased models; and (5) assuming the existence of an accuracy–effort tradeoff. To take point 2 as an example, parametric complexity provides only a limited view onto the benefits of simple, low-variance models. When focusing on factors such as these, subtle trade-offs exist that are not easily understood, or discovered. For instance, in marketing, portfolio management, and problems of financial regulation, implementing "wrong" constraints can often increase the predictive accuracy of a model (e.g., Haldane & Madouros, 2012; Jagannathan & Ma, 2003). Drawing on the study of simple heuristics, we show how factors such as limited search provide another means to limit variance. Put simply, our goal is to illustrate how simplicity can guide the search for predictive models in many ways, but it is often more insightful to view simplicity as addressing a single, more fundamental problem: variance reduction.

2. Simplicity and complexity in forecasting

Uncertainty exists when we have limited observations and knowledge of the causal processes of interest. Under uncertainty, problem solving is a process of search guided by heuristics of discovery. When developing forecasting models in areas such as finance, management, marketing, consumer research, and healthcare, a preference for simple models, commonly referred to as Occam's razor, can steer the discovery process away from overly complex models prone to overfitting. Models that overfit excel at describing the past, but offer poor predictors of the future. Why, though, believe that simple models are more likely to result in accurate predictions? Using predictive accuracy as a criterion of success, statistical measures of model complexity provide the most direct formal relationship to forecasting accuracy (e.g., Solomonov, 1964; Pearl, 1978; MacKay, 1992; Rissanen, 2007; Li and Vitányi, 1997; Grünwald, 2005; Myung et al., 2000).

Generally speaking, statistical notions of complexity consider "the flexibility inherent in a model that enables it to fit diverse patterns of data" (Pitt, Myung, & Zhang, 2002, p. 473). Fig. 1 provides a basic illustration of the role of model complexity in prediction. Readers familiar with this relationship can skip the remainder of this section. First of all, Fig. 1(a) plots two polynomial models fitted to London's mean daily temperature on each day of the year 2000. The first model is a degree-3 polynomial (a cubic equation with 4 parameters), and the second is a degree-12 polynomial (which has 13 parameters). Comparing these two models, we see that the degree-12 polynomial captures monthly fluctuations in temperature while the degree-3 polynomial captures a simpler pattern charting a rise in temperature that peaks in the summer, followed by a slightly sharper fall. Which of these two models best captures the process governing London's daily temperature?

Table 1
Five common symptoms of the bias bias.

	Symptom	Relationship to the bias bias
1.	Relying solely on goodness of fit to evaluate models	Designing, deploying, and evaluating models by considering their ability to fit rather than predict observations is one symptom of the bias bias. Variance is irrelevant to achieving a good fit, but critical to prediction. For this reason, the majority of researchers in education, sociology, and many other branches of the social and behavioral sciences suffer from the bias bias. Roberts and Pashler (2000), for example, estimated that in psychology alone, the number of articles relying on a good fit as the only indication of a good model runs into the thousands. See Section 2.
2.	Equating model complexity with parametric complexity	Penalizing models that achieve a good fit by considering only parametric complexity (for example, by using AIC and BIC, discussed in the main text) is the second symptom of the bias bias. Although the researcher attempts to assess the generalizability of a model, the policy of only considering parametric complexity will mask other factors contributing to model complexity, such as the functional form of the model and the range of possible values that the model parameters can be assigned. See Sections 2, 3 and 5.
3.	Drawing conclusions from a single model	Testing a single model, such as linear or logistic regression, as opposed to competitive testing of several models is the third symptom of the bias bias. Without comparing diverse models, the ability of a particular model to strike a good balance between reducing bias and variance is impossible to estimate. See Sections 4 and 6.
4.	Seeking unbiased models	Designing and deploying models capable of achieving low or zero bias by placing little or no restrictions on the class of functions that the model is capable of approximating is the fourth symptom of the bias bias. This policy assumes that models need to closely approximate the underlying data-generating process in order to achieve high predictive accuracy. However, by virtue of reducing more variance than they add bias, biased models can result in higher predictive accuracy. A practitioner suffering from the bias bias is likely to overlook such models and implicitly assume that sophisticated "accurate" models adequately reduce variance. See Sections 4 and 6.
5.	Assuming an accuracy–effort tradeoff	Designing and deploying models under the assumption of an accuracy–effort tradeoff is the fifth symptom of the bias bias. Believing in the general existence of an accuracy–effort tradeoff is to assume that more information and computation will always result in more accurate predictions. Effort can refer to the computational resources expended, the number of variables considered, or the complexity of the assumed relationships between variables. This widely held belief (e.g., Shah & Oppenheimer, 2008) is another example of the bias bias.

2.1. Goodness of fit and model complexity

Goodness of fit, which measures the discrepancy between the model and the observations, is one criterion for judging models. To understand the relationship between goodness of fit and model complexity, we will consider the problem of selecting a model after observing a sample of 50 observations of London's daily temperature, drawn at random. Using least squares, these 50 observations are used to estimate the parameters

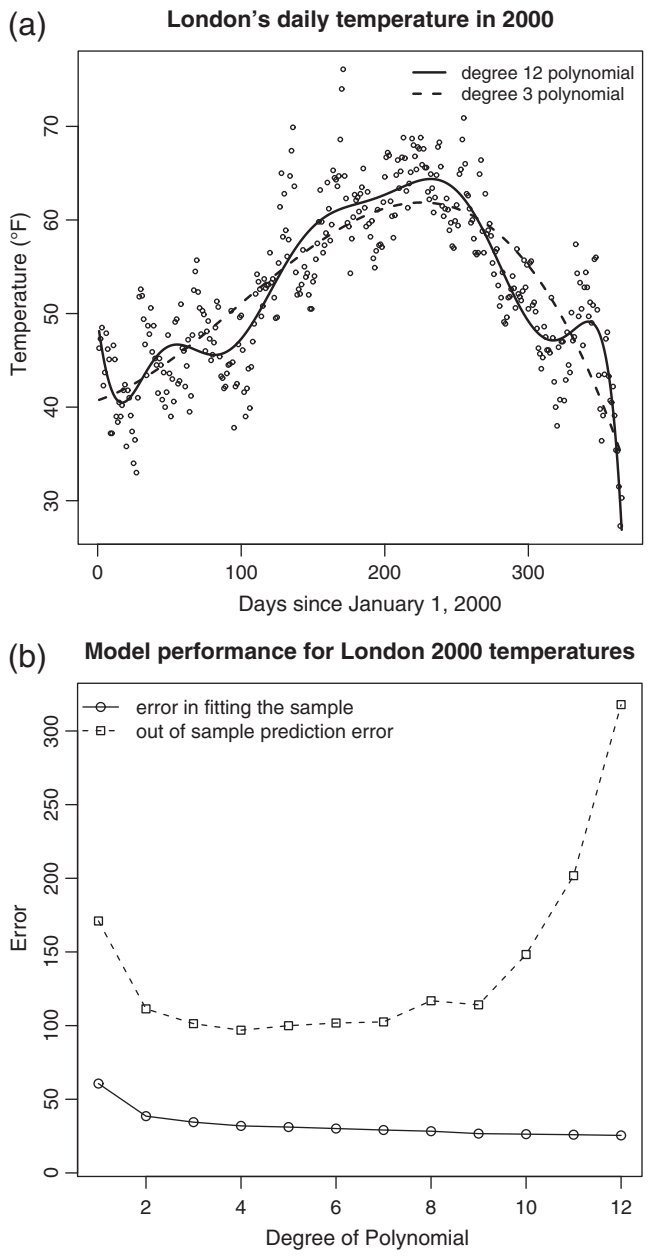

Fig. 1. Modeling London's mean daily temperature in 2000: Here we examine the relationship between polynomial models of increasing degree (and therefore increasing parametric complexity), their error in fitting the observations, and their error in predicting temperatures on unobserved days. Plot (a) shows London's mean daily temperature in 2000, along with two polynomials fitted to these observations. As a function of polynomial degree, Plot (b) shows both the mean error incurred when fitting 50 randomly drawn observations, and the mean error for the same polynomials when predicting the mean temperature on unobserved days. For the task of data fitting, greater parametric complexity ensures lower error. For the task of predicting temperatures on unobserved days, a U-shaped relationship exists, where low to intermediate parametric complexity yields the lowest prediction error.

of a series of polynomial models increasing in complexity from degree-1 to degree-12. Specifically, we measure goodness of fit as the sum squared difference between the 50 observed temperatures and the temperature estimates of the model on the same 50 days. A lower score indicates a better fit. The solid line in Fig. 1(b) plots the mean goodness of fit, relative to 5000 samples of size 50, achieved by each model as a function of its degree. As the polynomial degree increases, the error decreases, revealing that models with more parameters achieve lower error by virtue of their increased flexibility, a property that enables them to fit the observations accurately.

2.2. Predictive accuracy and model complexity

Predictive accuracy, the issue driving this discussion, is another criterion for judging models. Continuing the temperature example, predictive accuracy is measured by taking the fitted polynomials models used above but instead of measuring their ability to fit the observations we measure their ability to predict the temperature on those days we did not observe. Plotting predictive accuracy as a function of polynomial degree, the broken line in Fig. 1(b) reveals a different, U-shaped relationship between complexity and prediction error. Unlike goodness of fit, prediction error first decreases, and then increases as a function of model complexity. Models that result in poor predictive accuracy due to excess complexity are said to overfit. This trade-off between model complexity and predictive accuracy is a basic, yet often overlooked problem in statistical modeling and scientific inquiry more generally (Einhorn, 1972; Hitchcock & Sober, 2004; Pitt et al., 2002; Roberts & Pashler, 2000). To understand the merits of simplicity more generally and the principles needed to explain when and why simple models perform well, we will first distinguish different aspects of simplicity.

2.2.1. Model complexity and Occam's razor

Model complexity is typically used to guide the selection of competing models via the model selection criteria that penalize the goodness of fit achieved by each model by a complexity term; the greater the complexity of the model, the less weight should be placed on its ability to fit the observations. Model complexity is determined by two properties:

1. The number of parameters needed to specify the model. For example a line in a two dimensional space $y = ax + b$ is defined by 2 parameters (a and b), and a polynomial of degree 3, $y = ax^3 + bx^2 + cx + d$ is defined by 4 parameters (a, b, c and d).
2. The functional form of the model. Functional form refers to how the dependent variable(s) are related to the response variable. For example, the two models $y = b \cdot \ln(x + a)$ and $y = b \cdot x^a$ both have 2 parameters, but differ in functional form. Functional form plays a critical role in defining which patterns the model is capable of fitting (e.g., Pitt et al., 2002; Myung et al., 2000).

Popular model selection criteria include Akaike's information criterion (AIC, Akaike, 1973) and the Bayesian information criterion (BIC, Schwarz, 1978), both of which consider only the number of model parameters. More sophisticated criteria such as the minimum description length principle (MDL) take into account the contribution of the number of parameters, the functional form of the model, and other factors such as the range of values that each parameter can potentially be assigned (Grünwald, 2005; Pitt et al., 2002; Rissanen, 1996). In short, there exist several formal measures of complexity, with the application of one measure rather than another often depending on practical considerations.

3. A closer look at simple heuristics

Informal use of the terms “simple” and “complex” tend not to refer to the statistical measures of complexity mentioned above. The terms are more often used to refer to how easy it is to define a model or understand it, or what computational costs are incurred when applying the model, factors that are typically regarded as unrelated to the statistical measures of complexity used in the formal application of Occam's razor (e.g., MacKay, 1992; Domingos, 1999). In practice, however, computational costs and statistical complexity often go hand-in-hand. Indeed, we will use the study of simple heuristics to examine this relationship further and to shed light on the role of computational simplicity as a means to discover predictive models. Table 2 summarizes the key concepts that we will use to examine the relationship.

The study of simple heuristics examines the hypothesis that cognitive systems of humans and other animals often rely on

surprisingly simple strategies to make accurate inferences in uncertain environments (Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011; Gigerenzer, Todd, & The ABC Research Group, 1999). Use of the term "simple" in the study of heuristics refers to a range of strategies for ignoring information, such as ignoring dependencies between cues, making a prediction using a single cue, or foregoing the computation of cue weights. Use of the term "surprising" refers to the benefits of ignoring information, relative to commonly assumed models of information processing that attempt to integrate all information, calculate weights, and model potential dependencies between cues (Chater, Oaksford, Nakisa, & Redington, 2003; Gigerenzer & Brighton, 2009).

Strategies for ignoring information can result in reduced computational complexity, reduced model complexity, or both. Absolutely key to this approach is the hypothesis that these simplifications can increase predictive accuracy. This is in contrast to the commonly held belief that heuristics reduce effort at the expense of accuracy (Shah & Oppenheimer, 2008). This belief can be seen as another example of the bias bias (see Table 1), given that it appeals to the intuition that accurate inferences require accurate, usually complex representations of the problem, which in turn require complex calculations to apply. In contrast, our "simplicity first" approach views computational simplifications as a key consideration in the search for predictive models. Adopting this approach in no way assumes that simple models will necessarily outperform more complex models. Rather, the search for predictive models should be guided by an analysis of both simple and complex models. The bias bias describes a tendency among many researchers to focus on complex models at the expense of simple models.

Table 2
A glossary of key terms. Note that, in principle, the three categories of simplicity/complexity defined above are orthogonal. In practice, however, they tend not to be: Computationally demanding models tend to be more flexible and harder to interpret.

Term	Description
Categories of	simplicity/complexity
1. Computational complexity	A measure of the time and space resources used when estimating the parameters of a model and when using the model to make predictions. For example, a 9-month hiatus has low computational complexity because it has no parameters to estimate, whereas the Pareto/NBD model has higher computational complexity as several parameters must be estimated using a maximum likelihood procedure.
2. Model complexity	A measure of the ability of the model to fit diverse patterns of data. As we discussed in the main text, the two key contributors to model complexity are the number of parameters and the functional form of the model.
3. Model comprehensibility	How easy the model is to use and explain, or how transparent the relationship is between the assumptions made by the model and its predictions. For example, a 9-month hiatus rule is trivial to explain and use, while the Pareto/NBD model is far harder to use and interpret
Occam's razor	Usually defined as the following maxim: Among competing explanations consistent with the observed phenomena, the simplest should be preferred.
Bias–variance decomposition	For a given problem, a statistical decomposition of a model's expected prediction error into three components: bias, variance, and noise (see Eq. (1)).
Bias	For all k possible samples of a given size, we fit k models. Bias is the difference between the "mean" response of these k models and the "true" model. Bias usually reflects the inability of the model to represent the predictive regularities governing the observations.
Variance	A measure of the degree to which the k models described above vary about their mean. The variance component of prediction error reflects an oversensitivity of the parameter estimates to different observations of the same problem.
Bias–variance dilemma	A dilemma exists because bias and variance are not independent: Methods for reducing variance tend to increase bias, and methods for reducing bias tend to increase variance.
Bias bias	To suffer from the bias bias is to develop, deploy, or prefer models that are likely to achieve low bias, while simultaneously paying little or no attention to models with low variance.

3.1. Simple models in marketing and finance

Simple heuristics have been applied to problems of finance, management, marketing, and consumer research. Consider the problem of investing money into N funds. Harry Markowitz received the Nobel prize in economics for finding the optimal solution, the mean–variance portfolio (Markowitz, 1959). Taking seven investment problems, DeMiguel, Garlappi, and Uppal (2009) compared a simple 1/N heuristic, which allocates money equally to N funds, with 14 optimizing models, including the mean–variance portfolio and Bayesian and non-Bayesian models. These optimizing strategies had 10 years of stock data for estimating their parameters and on that basis had to predict the next month's performance; after this, the 10-year window was moved 1 month ahead, and the next month had to be predicted, and so on until the data ran out. 1/N, in contrast, did not need any past information because it has no parameters to estimate. In spite (or because) of this, 1/N ranked first (out of 15) on certainty equivalent returns, second on turnover, and fifth on the Sharpe ratio, respectively. Even with their complex estimations and computations, none of the optimization methods could consistently earn better returns than this simple 1/N heuristic.

More broadly, simple heuristics have been used to describe how consumers narrow down their consideration sets (Dzyabura & Hauser, 2011; Hauser, Toubia, Evgeniou, Befurt, & Dzyabura, 2010), how companies develop strategies to cope with fast-moving and uncertain markets (Bingham & Eisenhardt, 2011; Eisenhardt & Sull, 2001), and why early-stage ventures succeed or fail (Åstebro & Elhedhli, 2006; Furubotn, 2009; Guercini, 2012).

Our focus here is to understand how the analysis of simple cognitive heuristics — specifically, uncovering the statistical basis for their success and failure — can contribute to the broader question of simplicity in forecasting (Goldstein & Gigerenzer, 2009; Makridakis & Hibon, 2000; Makridakis, Hibon, & Moser, 1979). Why might hiatus rules used in marketing and the 1/N heuristic used in finance, for instance, result in higher predictive accuracy than the more established, sophisticated methods? To explore the benefits of simplicity further, we will consider perhaps the most intensively studied simple heuristic, take-the-best (Gigerenzer & Goldstein, 1996).

3.2. Take-the-best

Take-the-best models how decision makers infer which of two objects, say houses, scores highest on some criteria of interest, such as price. The decision maker is assumed to base this inference on a set of m binary cues $\{c_1, c_2, \ldots, c_m\}$ describing the task environment. Each cue describes a feature of the objects in the environment, such as the presence of swimming pool, garage, or whether the house has three or more bedrooms. Each object in the environment also has an associated criterion value – the dependent variable – such as the price of the house.

When deciding which of two objects has a greater criterion value, take-the-best operates as follows:

1. Search rule: Search through cues in order of their validity.
2. Stopping rule: Stop on finding the first cue that discriminates between the objects (i.e., cue values are 1 and 0).
3. Decision rule: Infer that the object with a cue value (0 or 1) that correlates positively with the criterion has the higher criterion value.

If none of the cues discriminate between the two objects in question, take-the-best guesses as to which object has a larger criterion value.

Take-the-best simplifies decision-making by both stopping after finding the first discriminating cue and ordering cues unconditionally by validity, which for the *i*th cue, c_i, is given by

$$v(c_i) = \frac{\text{number of correct inferences using } c_i}{\text{number of possible inferences using } c_i}. \quad (2)$$

Viewed as a statistical model, the functional form of take-the-best is shown in Fig. 2(a): Take-the-best can be viewed as a decision tree, where the boxes contain comparisons between objects using a given cue, and the cue value (1 or 0) that determines the choice of which object is inferred as having the greater criterion value is contained in each circle. The process of parameter estimation fills in the boxes and circles, as shown in Fig. 2(b).

The validity cue order determines which cues are placed in which boxes, and the values in the circles are determined by whichever cue value correlates positively with the criterion. This tree structure implements a form of noncompensatory processing: the search stops when the first discriminating cue is found, and all subsequent cues are ignored irrespective of their values. Noncompensatory processing is a form of simplicity studied extensively in marketing research, where it has proven highly predictive of how consumers form small consideration sets when faced with a large number of potential products (Hauser et al., 2010; Yee, Dahan, Hauser, & Orlin, 2007).

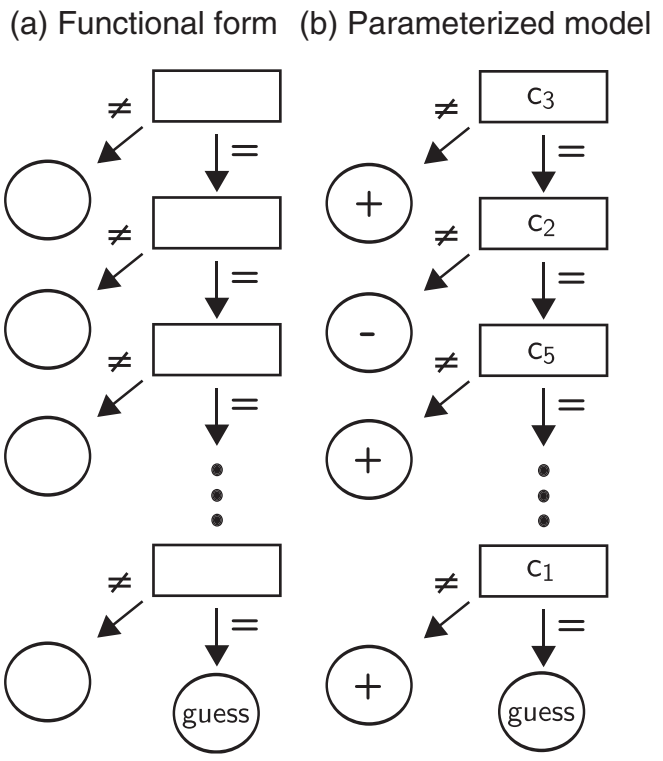

Fig. 2. Which parameters does take-the-best estimate? Here, we illustrate the structure of the models induced by take-the-best. These models are formally equivalent to a decision tree. In (a), the functional form of the unparameterized decision tree used by take-the-best is shown, where boxes denote comparisons between the two objects using a single cue, and circles denote decisions made on the basis of these comparisons. If none of the cues discriminate between the two objects, take-the-best guesses. In (b), a parameterized decision tree illustrates the ordering of the *m* cues ($c_1, \ldots, c_m$), and the "direction" of each cue. The cue directions define which cue value is used to infer which of the two object scores higher on the criterion. Here, cue directions are denoted by one of {+,−}, where "+" indicates that a positive cue value indicates a larger score on the criterion, and indicates a lower score.

3.3. Six alternative models

Whereas early studies compared take-the-best with decision-making models such as linear regression and unit-weighted linear models (Czerlinski, Gigerenzer, & Goldstein, 1999), we will conduct a comparison focusing on tried-and-tested models used in cognitive modeling, data mining, and machine learning, models that "should" outperform take-the-best (Chater et al., 2003). In particular, we will compare take-the-best with six models that are summarized in Table 3: Logistic regression (Hosmer & Lemeshow, 2000), classification and regression trees (CART, Breiman, Friedman, Olshen, & Stone, 1994), a single-layer neural network (Bishop, 1995; Rosenblatt, 1959), the nearest neighbor classifier (Cover & Hart, 1967; Fix & Hodges, 1951), a support vector machine (Schölkopf & Smola, 2002), and a variant of take-the-best that differs only in how cues are searched (Martignon & Hoffrage, 2002; Schmitt & Martignon, 2006). These models provide a broad span of widely used and fairly sophisticated modeling strategies.

The variant of take-the-best, referred to as greedy take-the-best, will serve as a "control condition" because it differs in one important respect: The validity of a cue is computed conditionally on the decisions made by cues appearing earlier in the cue order. For example, the validity of the final cue in the cue order is, in practice, likely to change

Table 3
A description of the inference algorithms used in the model comparison appearing in Fig. 3.

Model	Description
Logistic regression	A linear model for binary classification similar to multiple linear regression. Instead of estimating a numeric independent variable, logistic regression provides a probability estimate of an observation belonging to a given class (e.g., Bishop, 2006; Hosmer & Lemeshow, 2000).
Classification and regression trees (CART)	Decision trees guide the decision maker through a series of decision nodes until a leaf node is reached that specifies which class to predict. The path through the tree is determined by the values of the dependent variables. CART provides a set of techniques for constructing decision trees by recursively partitioning the observations into subsets containing observations of approximately the same class (Breiman et al., 1994).
Single-layer neural network	Neural networks model a collection of artificial neurons connected with varying degrees of strength. By feeding the values of the dependent variables into the input nodes, the network propagates the values to an output node that codes the prediction. Training a neural network involves adjusting the connection strengths between neurons to minimize the errors made in predicting the independent variable (Bishop, 1995; Rosenblatt, 1959).
Nearest neighbor classifier	Store all observations, and when called to assign a class to a novel observation, predict the class of the most similar stored observation (Cover & Hart, 1967; Fix & Hodges, 1951).
Support vector machine	Observations can be seen as points in a multidimensional feature space. Support vector machines transform the observations into a space where a linear decision boundary separates the two classes by some (preferably large) margin. By transforming the observations into a higher dimensional space, support vector machines can treat complex nonlinear problems as linear problems (Schölkopf & Smola, 2002).
Take-the-best	Take-the-best decides which object has a higher criterion value by searching the cues in sequence of validity order and using the first discriminating cue to make the prediction (Gigerenzer and Goldstein (1996); see main text).
Greedy take-the-best	Identical to take-the-best except that cues are searched in conditional validity order rather than validity order (Martignon and Hoffrage (2002); Schmitt and Martignon (2006); see main text).

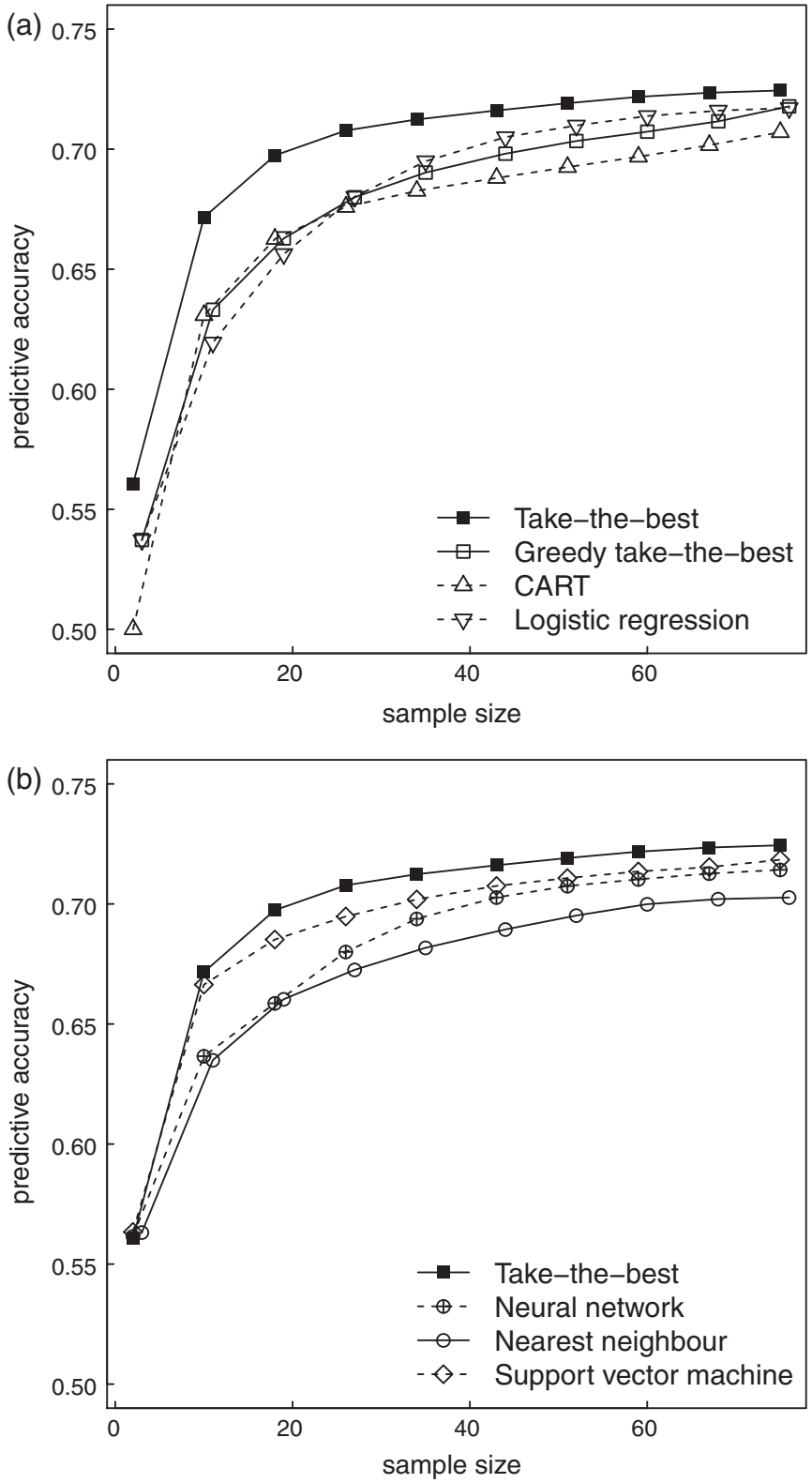

Fig. 3. An illustration of take-the-best achieving greater predictive accuracy than a range of widely used, relatively complex learning algorithms. For the problem of inferring which of two German cities has a greater population, (a) plots the predictive accuracy of take-the-best, its greedy variant, classification and regression trees, and logistic regression as a function of the sample size r; (b) compares take-the-best with a single-layer neural network, the nearest neighbor classifier, and a support vector machine. Take-the-best outperforms all these methods across nearly all sample sizes.

because some of the comparisons contributing to its validity will have been made by cues appearing earlier in the order. Greedy take-the-best addresses this issue by estimating validities sequentially, recomputing them at each stage in the construction of the tree.

Specifically, we will compare the predictive accuracy achieved by each model in a task environment detailing German cities with a population greater than 100,000 (83 cities). Here, the task is to infer as to which of two cities has a greater population. Each city is described

by nine binary cues indicating the properties of the cities, such as the presence of an airport, a top-flight soccer team, or a university. For a given sample size, say 10 cities, we first generate a training set of paired comparisons between the cities and allow each model to estimate its parameters using these observations. The predictive accuracy of each model is then estimated by measuring how accurately they predict which city is largest among all possible comparisons between the cities in the test set, which contains those cities not appearing in the training set. For a range of sample sizes, we report mean predictive accuracy with respect to 5000 random partitions of the data into training and testing sets.

Fig. 3(a–b) shows the result of the model comparison. On the x-axis is the sample size, the number of cities used to generate the training sets used to estimate the model parameters. The y-axis represents the out-of-sample predictive accuracy. The result is clear: Simplicity, again, wins out. Take-the-best outperforms all the alternative models over nearly all sample sizes.

3.4. Can Occam's razor explain the success of take-the-best?

The preceding analysis of take-the-best poses a conundrum. Occam's razor – the idea that performance differences arise due to factors related to model complexity – should explain the success of simple heuristics like take-the-best, but a close inspection of Fig. 3(a), in particular the relative predictive accuracies achieved by take-the-best and its greedy counterpart, rule out an explanation based on model complexity.

These two algorithms achieve very different degrees of performance, yet induce models drawn from the same class, with the same number of parameters and identical functional form. The only difference between the two models is the criterion used to order the cues (that is, the search rule). In addition, the greedy variant of take-the-best provably achieves a better fit to the observations than take-the-best (Schmitt & Martignon, 2006). So, why does take-the-best achieve a higher predictive accuracy than its greedy counterpart here, and for many other problems? To answer this question, a more in-depth treatment of what causes a model to "overfit" is required.

4. The analysis of bias and variance

Informally stated, a "no free lunch" theorem holds in statistical pattern recognition: Without restricting the range assumed problem characteristics, any two forecasting models will have precisely equal predictive accuracy when averaged over all possible problems (Duda, Hart, & Stork, 2001; Wolpert, 1996). Thus, no single predictive model is *inherently* superior to any other; the assumptions implicit in a model must to one degree or another match the characteristics of the problem at hand in order to yield accurate inferences. One way of understanding the match between a model and problem, and the properties of the model responsible for this match, is to decompose the prediction error of the model into bias, variance and noise:

$$\text{Total error} = (\text{bias})^2 + \text{variance} + \text{noise}.$$

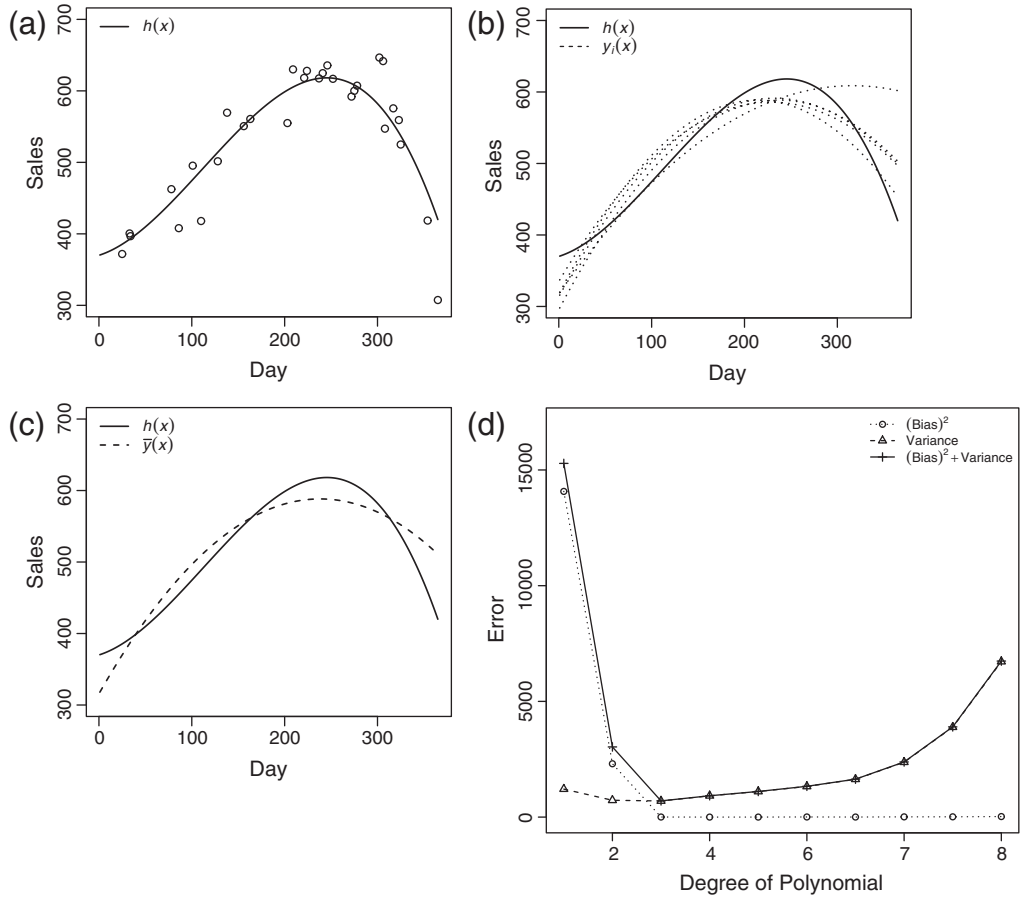

Fig. 4. An illustration of the role of bias and variance. Plot (a) shows the function $h(x)$, which determines units sales of some product for each day of the year, along with 40 noisy observations of $h(x)$ on randomly selected days. For four such sets of 40 observations, Plot (b) shows four induced polynomials. Bias measures the difference between $h(x)$ and $\bar{y}(x)$ — where $\bar{y}(x)$ is the mean response of the four polynomials shown in (b) — which are plotted together for comparison in (c). In (d), we plot prediction error as a function of polynomial degree, decomposed into bias and variance. Notice how, in (d), the solid line plotting $(bias)^2 + variance$ follows the same U-shaped relationship between polynomial degree and error found in Fig. 1(b). In (d), note that for polynomials of degree 4 and higher, the negative consequences of the bias bias become evident.

The noise component of error cannot be reduced, however we design the model. Bias and variance, in contrast, arise from the interaction between properties of the model, properties of the problem, and the sample size. Decomposing error into these three components, and the analysis of their relative contribution, provides a great deal of insight on the art of forecasting (Bishop, 2006; Geman et al., 1992; Hastie et al., 2001; O'Sullivan, 1986; Van Der Putten & Van Someren, 2004).

4.1. An illustration of bias and variance

Consider the problem of modeling product sales throughout the year. In Fig. 4(a) we have plotted a function $h(x)$ governing unit sales of some product over the course of a year. We have also plotted 40 observations of $h(x)$ on a random sample of days, each subject to some measurement error. A modeler who has observed these sales levels, but does not know $h(x)$ uses the observations to construct a model for the entire year. For five such samples of 40 observations, Fig. 4(b) plots the five models, $y_1(x), \ldots, y_5(x)$, induced by the forecaster for these different "replays" of the tape of experience. In addition to these five models, we have plotted $\overline{y}(x)$, which is the mean prediction of these five models. Bias, when used to analyze the prediction error of a statistical model, is the difference between $h(x)$ and $\overline{y}(x)$, both depicted in Fig. 4(c). Variance measures how much the five individual models, $y_1(x), \ldots, y_5(x)$, vary about their mean, $\overline{y}(x)$.

Fig. 4(d), much like Fig. 1(b), plots the mean prediction error as a function of polynomial degree. Here, though, we have also plotted the prediction error and its components, bias and variance. For polynomials of degree 10, for example, bias is zero and variance is high. Notice how the plot of "$(\text{bias})^2 + \text{variance}$" mirrors the relationship between model complexity and prediction error discussed earlier, and depicted by Fig. 1(b). When broken down into bias, variance, and noise, models with too many free parameters suffer from excess variance and models with too few parameters suffer from excess bias.

A tempting and commonly drawn conclusion is that considerations of bias and variance simply rephrase the problem of balancing goodness of fit and parametric, or some other measure of complexity. The comparison of take-the-best and its greedy variant in Fig. 3 illustrates that the picture is not quite so simple, and a more accurate interpretation reverses this relationship: Model complexity provides a window into more general considerations of bias and variance.

4.2. The bias/variance dilemma

Bias and variance highlight a fundamental problem in inductive inference known as the bias/variance dilemma (Geman et al., 1992). At one extreme, a statistical model could express a wild guess by ignoring the observations altogether and always selecting the same parameter values. For example, the marketing executive who uses the same hiatus rule for distinguishing active from inactive customers adopts this policy. This approach guarantees zero variance, but can lead to high bias unless the guess turns out to be correct or close to correct. At the other extreme, the statistical model could hedge its bets, let the observations speak for themselves, and select from a highly flexible model space capable of approximating any function. Given enough observations, this policy could in principle guarantee zero bias, but usually at the expense of high variance, since the flexibility of the model space is likely to lead to an oversensitivity to the vagaries of particular samples.

The bias/variance dilemma arises because methods for minimizing variance tend to increase bias and methods for minimizing bias tend to increase variance. The two need to be balanced, a process that should be guided by knowledge of the task at hand. As pointed out above, high model complexity can lead to excess variance and overfitting but it is not the only cause. Similarly, reducing the complexity of a model is not the only way of reducing variance.

4.2.1. Bias and variance in practice

Recall that a theorist suffering from the bias bias views problems through the lens of bias, and places little or no weight on considerations of variance. To avoid the bias bias during the search for predictive models, and sidestep the pathologies summarized in Table 1, different points in the trade-off between bias and variance need to be assessed. This is an exploratory task that ideally involves a comparison between models with varying complexity. A critic, however, may wonder how relevant the above "toy" prediction example is to real-world forecasting, or how critical variance will be when large samples of observations are available.

In practice, variance is always critical. Van Der Putten and Van Someren (2004), for example, conducted a bias–variance analysis of 8 classes of model submitted to the Computational Intelligence and Learning Cluster (CoIL) challenge, a competition to predict insurance purchases. They found that differences between submitted models were overwhelmingly due to the variance component of error. The best performing competitor was the naïve Bayes classifier, which, like take-the-best, makes the typically "false" assumption that cues are conditionally independent (an issue we discuss further below). Findings such as these are easily explained by appealing to bias and variance: the naïve Bayes classifier tends to have high bias in practice but achieves low prediction error due to incurring low variance (Domingos & Pazzani, 1997; Hand & Yu, 2001; Hastie et al., 2001).

Questioning the increasingly complex measures imposed by regulators to avoid financial crises, Haldane and Madouros (2012) used a bias–variance decomposition to assess a range of models for forecasting return volatility and concluded that misspecified, and hence biased, low-variance models achieved higher out-of-sample predictive accuracy. Their conclusion that "simple does not just defeat complex; it trumps the truth" (Haldane & Madouros, 2012, p. 17) has a natural interpretation when considering bias and variance. In short, the bias bias describes a general reticence in trying and trusting simple models. In the case of predicting insurance purchases, Van Der Putten and Van Someren (2004) reported that many competitors abandoned simple models in favor of fine-tuned, better fitting models prior to entering their final model into the competition. These competitors suffered from the bias bias.

5. A bias bias in the analysis of simple heuristics

The question of when and why simple heuristics like take-the-best achieve high predictive accuracy has been the topic of sustained research (e.g., Hogarth and Karelaia, 2006, 2007; Katsikopoulos and Martignon, 2006; Martignon and Schmitt, 1999; Schmitt & Martignon, 2006). Previous analyses can be seen as focusing on bias and asking the question of when heuristics like take-the-best make accurate inferences when cue validities are known rather than estimated from a sample. This approach does not consider the role of variance. Instead, we will conduct a bias–variance decomposition of take-the-best's performance in a more realistic setting where cue validities are uncertain and must be estimated from observations. By considering the impact of variance, we will show that the conditions favoring take-the-best differ significantly from those identified when focusing on bias. Specifically, we will analyze the performance of take-the-best and its greedy variant in two artificial environments. Working under these restricted "laboratory conditions" will allow us to understand more clearly when and why two models with equivalent complexity perform so differently.

5.1. Noncompensatory models and environments

Recall that a noncompensatory model, such as take-the-best, searches cues sequentially, selects one cue to make a decision, and ignores all subsequent cues. A foundational result in the study of simple heuristics seeks to establish a correspondence between

noncompensatory models and noncompensatory environments. The general idea is that "Take The Best 'bets' that the cues in the environment are noncompensatory" (Martignon & Hoffrage, 2002, p. 47). A noncompensatory environment is one where the validities or weights of the cues under consideration are highly skewed, such that when ordered by their validity, a decision made by a given cue cannot be overturned by any combination of cues that appear subsequently in the cue order.

The two environments used to examine this correspondence include, first, an instance of the class of *binary* environments, which have a noncompensatory structure such that the weights of the cues decay rapidly, such as $1, \frac{1}{2}, \frac{1}{4}, \ldots$ The second environment is a member of the class of *Guttman* environments, where all cues have a validity of 1.0, and are therefore "compensatory." Appendix A provides a detailed definition of these classes of environments, and a discussion of this analysis can be found in Gigerenzer and Brighton (2009). The important point to note here is that the structure of take-the-best is "mirrored" by binary environments, but not Guttman environments.

5.2. Performance at environmental extremes

On the left-hand side of Fig. 5, we compare the predictive accuracy of take-the-best and its greedy counterpart in a (a) binary environment with 6 cues and 32 objects, and (b) a Guttman environment with 32 cues and 33 objects. For both environments, we have also decomposed the error of both models into bias and variance, shown by the subplots on the right-hand side. First, notice that take-the-best performs poorly in comparison to its greedy variant in the binary environment. This is not what we should expect to see, given the preceding discussion. Second, in the Guttman environment, take-the-best outperforms its greedy variant by a significant margin. Again, this finding runs counter

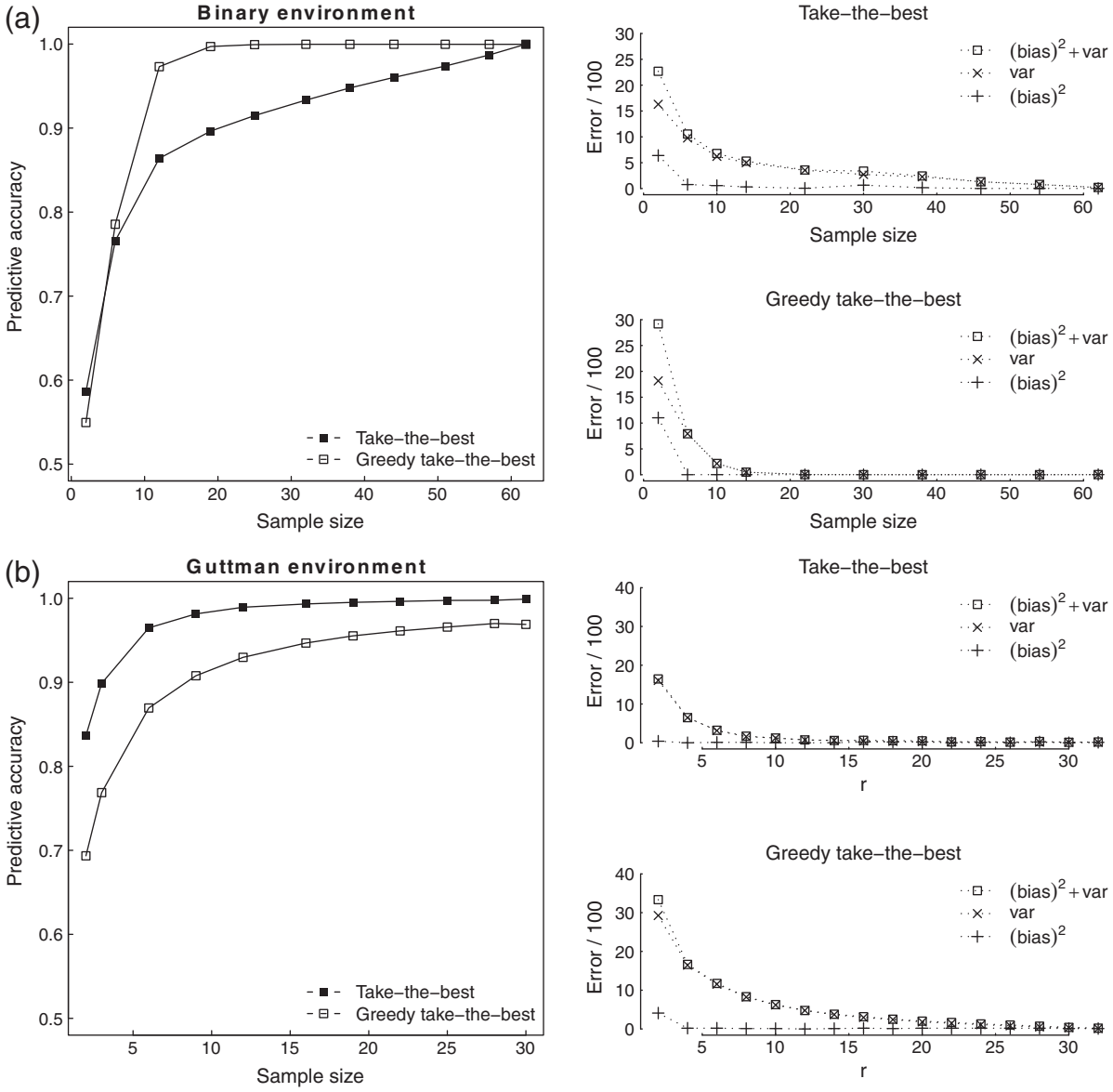

Fig. 5. When does take-the-best excel, and what role do bias and variance play in determining performance? Here, we compare take-the-best and its greedy counterpart in two artificial environments. In the main plots shown on the left, (a) shows an analysis of take-the-best and its greedy variant in the binary environment $h_{binary}(6)$. Despite this environment having noncompensatory cue validities, take-the-best performs poorly. Plot (b) shows the same analysis for the environment $h_{Guttman}(32)$, where cue validities have a compensatory structure, and take-the-best outperforms its greedy variant. In the sub-plots shown on the right, in both environments, and for both models, bias and the variance are plotted as a function of the sample size. Variance proves to be the key determinant of the observed performance differences.

to the idea of a general correspondence between noncompensatory models and noncompensatory environments.

This mismatch is easily explained by appealing to bias and variance. Fig. 5(a) shows that given enough observations, take-the-best achieves maximum predictive accuracy in the binary environment. This tells us that take-the-best is unbiased. Crucially though, no previous analysis considered take-the-best's variance, which for this environment is high in comparison to practically all other models. In the Guttman environment, we see the opposite: Take-the-best has low variance, and will outperform most models. Critically, in both environments, bias is largely irrelevant, with performance differences arising due to variance. Put simply, early work on the question of when simple heuristics like take-the-best succeed suffers from the bias bias: Previous analyses considered conditions for achieving low bias, but not low variance.

5.2.1. The "trick" behind take-the-best

Recall that familiar measures of model complexity failed to explain the performance differences between take-the-best and its greedy variant. The key difference between these two models, the property that leads to the extreme differences in performance, is the method used to infer the cue order from observations. Take-the-best relies on the trick of using the naïve measure of cue validity to order cues and construct a decision tree, rather than conducting the extra computations required when estimating potential dependencies between cues. Unlike the greedy variant of take-the-best, no attempt is made to find an ordering of cues that maximize goodness of fit. Relative to its greedy counterpart, this is a computational shortcut that often works very well in natural environments. As shown in Fig. 5(a–b), the shortcut "works" or "fails" to the degree that it reduces variance.

A number of studies have attempted to explain when and why take-the-best succeeds. Almost without exception, the noncompensatory nature of take-the-best has been viewed as the key property of interest (which is justified when the ecological validity of cues is known, rather than being estimated from a sample). Here, and elsewhere, we have shown that the high predictive accuracy of take-the-best relative to many other models can, informally speaking, be switched "on and off" (Brighton & Gigerenzer, 2007; Gigerenzer & Brighton, 2009). Rather than a property of how cues are processed once a cue order has been selected, we have shown that the performance differences are a property of the method for selecting the cue order. The noncompensatory nature of take-the-best contributes little or nothing to the problem of variance reduction. The issue of noncompensatory processing is, however, relevant to the bias of take-the-best. As we have shown, however, bias is irrelevant to the performance differences we observe.

5.3. Questioning Occam's razor

These findings serve as a cautionary tale, with broader implications. As a basic consequence of the design of a model, the method for estimating model parameters will have an impact on the stability of the model's predictions with respect to perturbations to the observations. Instabilities tend to cause variance and need to be controlled. Clearly, the stability properties of a model will depend on the statistical properties of the problem. This is what we observe in Fig. 5, where take-the-best's achieves low variance in the binary environment, but not in the Guttman environment. Thus, rather than a characteristic of the class of models being selected from, stability is a property of the *interaction* between the model and the environment (Poggio, Rifkin, Mukherjee, & Niyogi, 2004). In some cases, model complexity will provide an insightful perspective on this phenomenon: The more flexible the choice of model is, the greater the opportunity for instability to result in a given environment.

The idea that model complexity, a formal measure of a model's inherent flexibility, fails to provide a sufficiently fundamental concept to fully explain performance in general is not new. Domingos (1999) discusses a range of counterexamples to Occam's razor in the machine learning and data mining literature (see also Webb, 1996). The use of ensemble methods, which can improve predictive accuracy by combining the predictions of multiple models, is another class of counterexamples commonly explained by appealing to problems of variance reduction (e.g., Breiman, 1996, 2001a; Seni & Elder, 2010). Thus, considerations of bias and variance do not imply a straightforward relationship between simple models and predictive accuracy; the problem of variance reduction can also be addressed by adding complexity through combining multiple forecasts (Armstrong, 2001, 2005; Graefe, Armstrong, Cuzan, & Jones, forthcoming).

Of particular relevance to this discussion is the proposal that overfitting is often a pathology associated with failing to control for different sample sizes when conducting and comparing multiple statistical tests during the parameter estimation process (Jensen & Cohen, 2000; Zahálka & Železný, 2011). Greedy take-the-best, for example, calculates cue validities with respect to reference classes of observations which decrease significantly as the decision tree is built one level at a time. No attempt is made to control for these size differences. Take-the-best, on the other hand, computes unconditional estimates of cue validity relative to a single reference class containing all the observations. An interesting avenue for future research would be to investigate the relationship between strategies like greedy take-the-best that "oversearch" and variance. Here, the relevant notion of complexity appears to be computational rather than model complexity.

6. General discussion

In an ideal world, the problem of forecasting would be reduced to the problem of formalizing accurate probabilistic representations of the causal processes determining future event of interests, and then applying the laws of probability theory to yield optimal predictions. In reality, forecasters face degrees of uncertainty that make optimality an unobtainable goal: Underlying causal processes tend to be latent, complex, interacting, and often nonstationary. Moreover, observations tend to be sparse, rendering reliable parameter estimation problematic. Operating outside the idealized world of mathematical statistics, the practice of forecasting is more accurately seen as a process of exploratory data analysis, an incremental search for models that reduce but do not resolve uncertainty (Breiman, 2001b; Tukey, 1962, 1977). Our goal has been to clarify the role of simplicity in guiding this search, and to understand why it appears to be such a powerful heuristic of discovery.

6.1. What is the bias bias?

To recap, a theorist suffering from the bias bias views problems through the lens of bias, and places little or no weight on considerations of variance. Underestimating the variance component of error, the bias bias masks the benefits of simplicity, obscures the search for predictive models, and reinforces an overly simplistic view on the problem of statistical inference. Recall that the bias component of error reflects an inherent inability of the model to consistently recover predictive regularities. The variance component reflects the sensitivity of the method used to recover these regularities from different samples of the same underlying process.

Contemporary use of the bias–variance perspective owes a great deal to a landmark paper by Geman et al. (1992), who used the analysis of bias and variance to clarify why the ability of neural network models to provide a universal framework for modeling learning systems does not, by itself, solve the learning problem. Models relying on complex systems of representation need constraints in order to learn under conditions of uncertainty, constraints that keep variance within acceptable limits. The bias bias, which is manifest to varying degrees across most disciplines, exacerbates this problem. Put simply, models capable of representing a problem cannot be assumed to address the problem adequately, because issues of representation fail to offer a complete

view on the issue of variance reduction. It should be stressed that our claim is not that sophisticated models relying on rich representations result in inaccurate predictions. Our claim is that under uncertainty, the search for predictive models is a matter of comparison rather than confirmation, and that an informative comparison requires diverse competitors. Simplicity, in its various guises, provides dimensions of variation for introducing diversity.

6.2. Overcoming the bias bias

How can practitioners overcome the bias bias? The first step is to recognize the importance of variance and its causes. The second step is to investigate strategies for reducing variance. Although there is no set recipe for reducing variance, such as using fewer parameters, there are a number of techniques to consider. Before summarizing these techniques, it should be stressed that simplicity also plays a role in the processing and preparation of observations; discretizing continuous cues is one heuristic trick that can improve predictive accuracy (Liu, Hussain, Tan, & Dash, 2002), as is ignoring a large proportion of observations altogether (Brighton & Mellish, 2002). As we mentioned above, it should also be stressed that variance can be reduced by adding complexity, for example, by combining the forecasts of multiple models (Armstrong, 2001, 2005; Breiman, 1996, 2001a; Graefe et al., forthcoming; Seni & Elder, 2010).

6.2.1. Predict using a single cue

Predictions based on a single cue are perhaps the most simple of all strategies: A single cue is vanishingly unlikely to provide an unbiased model of system being predicted. For instance, taking a "simplicity first" approach, Holte (1993) compared how a single cue model, 1R, and more familiar statistical models, which tend to use all available cues, performed on 16 prediction tasks. Holte found that 1R's performance often matched or approached the performance of more sophisticated methods. In cases like these, single-cue models provide a useful benchmark for assessing the potential benefits of using more cues or more sophisticated models, the advantages of which are sometimes inflated (Hand, 2006). In other cases, single cue models have been shown to outperform more sophisticated models relying on more cues. The hiatus rule for distinguishing active from inactive customers, discussed earlier, is one example. Forecasting elections, sporting events, and the stock market by measuring lay people's recognition of the alternative candidates, players, or stocks can also prove remarkably predictive (Borges, Goldstein, Ortmann, & Gigerenzer, 1999; Gaissmaier & Marewski, 2011; Goldstein & Gigerenzer, 2009; Scheibehenne & Bröder, 2007; Serwe & Frings, 2006).

6.2.2. Ignore or restrict cue weights

Multiple linear regression is the most widely used and trusted statistical model in use today. In the 1970s, the discovery that unit weights (either −1 or 1) in a linear regression can increase in predictive accuracy was a key discovery (Dawes, 1979; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975; Schmidt, 1971; Wainer, 1976). When Robyn Dawes presented the results at professional conferences, distinguished attendees told him that they were impossible. This reaction illustrates the negative impact of the bias bias: Dawes' paper with Corrigan was first rejected and deemed premature, and a sample of recent textbooks in econometrics revealed that none referred to their findings (Hogarth, 2012). These examples are an extreme case of shrinkage, a statistical technique for reducing variance by imposing restrictions on estimated parameter values (Hastie et al., 2001; Hoerl & Kennard, 2000). Portfolio management is one application of this technique, where by imposing "wrong" constraints — assumptions that are known not to hold in practice — practitioners can increase bias in order to achieve a greater reduction in variance, and lower overall error (Jagannathan & Ma, 2003).

6.2.3. Make the naïvety assumption

The naïvety assumption, also termed the assumption of conditional independence, concerns the relationship between the available cues and the independent variable being predicted. For example, if urban customers tend to purchase more than rural customers, and customers who own cars also tend to purchase more than those without, then making the conditional independence assumption means that when a given urban customer purchases more than a given rural customer, this information tells us nothing about whether the urban customer also owns a car. As we saw, take-the-best implicitly makes this assumption by relying on cue validities to make inferences: Validities are calculated independently from the contribution of other cues. Linear regression and logistic regression models, on the other hand, attempt to estimate these dependencies. The most widely used naïve method, the naïve Bayes classifier, is closely related to logistic regression (Ng & Jordan, 2002) but its assumption of conditional independence has been found, in many contexts, to yield higher predictive accuracy. Indeed, the naïve Bayes classifier is ranked in the top 10 models in data mining (Wu et al., 2007), and like take-the-best, its low variance often results in improved predictive accuracy over more sophisticated methods, even when its assumption of conditional independence is known to be incorrect (Domingos & Pazzani, 1997; Hand & Yu, 2001; Van Der Putten & Van Someren, 2004).

7. Conclusion

What is the right level of simplicity for forecasting methods? We have approached this question by framing the forecasting problem as one of exploratory data analysis and specifically, an incremental search for improved responses to the bias/variance dilemma. This changes the question to the following: What role does simplicity play in the search for models with low variance? Simplicity provides a powerful heuristic of discovery, but the term "simple" refers to a broad range of techniques for — and perspectives on — limiting variance. Thus, without considering and comparing simplifications that differ in both degree and kind, the benefits of simplicity cannot be judged. What we term the bias bias obscures the often subtle and surprising benefits of simplicity in this search for predictive, low-variance models. Moreover, the bias/variance perspective also clarifies why *adding* complexity, such as combining multiple forecasts, can also reduce variance and increase forecasting accuracy.

When is simplicity likely to confer a performance advantage? In broad terms, we have stressed the role of simplicity under conditions of high uncertainty, where observations are sparse and little is known about the causal processes that determine future events of interest. This is when variance is most problematic. Indeed, we have shown how the bias bias can obscure the analysis of conditions under which simple models perform well relative to more familiar, sophisticated models. The benefits of simple heuristics like take-the-best, 1/N, and hiatus rules center on their ability to reduce variance, and the conditions under which a model will achieve low variance are much harder to

Table 4
An example of a binary environment, h_binary (3). This environment codes 8 objects (labeled A–H) using the 3 cues c_1, …, c_3. Cue validities follow a noncompensatory pattern, and range from 1 to .5.

Object	Cues			Criterion
	C1	C2	C3	
A	0	0	0	1
B	0	0	1	2
C	0	1	0	3
D	0	1	1	4
E	1	0	0	5
F	1	0	1	6
G	1	1	0	7
H	1	1	1	8

Table 5
An example of a Guttman environment, h_Guttman (7). This environment codes 8 objects (labeled A–H) using the 8 cues c_1, …, c_7. Cue validities are all 1.0 because for all pairs of objects, the cues that discriminate between the objects always points to the one with the larger criterion value.

Object	Cues							Criterion
	C1	C2	C3	C4	C5	C6	C7	
A	0	0	0	0	0	0	0	1
B	1	0	0	0	0	0	0	2
C	1	1	0	0	0	0	0	3
D	1	1	1	0	0	0	0	4
E	1	1	1	1	0	0	0	5
F	1	1	1	1	1	0	0	6
G	1	1	1	1	1	1	0	7
H	1	1	1	1	1	1	1	8

determine than when a model will achieve low bias. Thus, overcoming the bias bias is critical to discovering new models, but also to understanding when and why they work.

Appendix A. Binary and Guttman environments

The class of binary environments has the noncompensatory property discussed in the main text. Parameterized by the number of cues m, binary environments are defined as follows:

$$h_{binary}(m) = \{\langle B_m(i), i+1\rangle : 0 \le i \le 2^m - 1\}. \quad (A.1)$$

This expression defines a set of 2^m objects, each one mapping m binary cues onto an integer-valued criterion. For illustrative purposes, Table 4 lists the 8 objects in the environment $h_{binary}(3)$ where, for each object, the cue values are given as a function of the criterion value by $B_m : Z \mapsto (c_1, \ldots, c_m)$, which maps integers onto their m-bit binary encodings [e.g., $B_3(2) = (0, 1, 0)$]. Each bit of the binary encoding represents a cue value. Environments in this class always have cue validities (and β-weights) with the noncompensatory property, and use the minimum number of cues required to unambiguously code N objects, which is $\log_2(N)$ cues.

The class of Guttman environments, also parameterized by the number cues m, is defined by

$$h_{Guttman}(m) = \left\{ \left\langle B_m\left(\sum_{i=0}^{j-1} 2^i\right), j \right\rangle : 1 \le j \le m-1 \right\}. \quad (A.2)$$

This expression defines a set of $m + 1$ objects. Table 5 lists the 8 objects in the environment $h_{Guttman}(7)$. Guttman environments are inspired by the Guttman (1944) scale: The ith object in the environment has the first $i - 1$ cue values set to 1, and all others set to 0. Guttman environments represent an opposing extreme to noncompensatory environments: cues have a validity of 1. Guttman environments are therefore "compensatory," and provide an inefficient coding of objects in comparison to binary environments because they require $N - 1$ cues to code N objects.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle [Second International Symposium on Information Theory]. In B. N. Petrox, & F. Caski (Eds.), Budapest: Akademiai Kiado.

Armstrong, J.S. (2001). Combining forecasts. In J.S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J.S. (2005). The forecasting canon: Nine generalizations to improve forecast accuracy. *Foresight: The International Journal of Applied Forecasting*, *1*(1), 29–35.

Åstebro, T., & Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, *52*(3), 395–409.

Bingham, C.B., & Eisenhardt, K.M. (2011). Rational heuristics: The simple rules that strategists learn from process experience. *Strategic Management Journal*, *32*, 1437–1464.

Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Borges, B., Goldstein, D.G., Ortmann, A., & Gigerenzer, G. (1999). Can ignorance beat the stock market? In G. Gigerenzer, P.M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 59–72). New York: Oxford University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199–231.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, P.J. (1994). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Brighton, H., & Gigerenzer, G. (2007). Bayesian brains and cognitive mechanisms: Harmony or dissonance? In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 189–208). Cambridge: Cambridge University Press.

Brighton, H., & Mellish, C. (2002). Advances in instance selection for instance-based learning algorithms. *Data Mining and Knowledge Discovery*, *6*, 153–172.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes*, *90*, 63–86.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Czerlinski, J., Gigerenzer, G., & Goldstein, D.G. (1999). How good are simple heuristics? In G. Gigerenzer, P.M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 119–140). Oxford: Oxford University Press.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Dawes, R.M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, *22*, 1915–1953.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, *3*, 409–425.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Duda, R.O., Hart, P.E., & Stork, D.G. (2001). *Pattern classification*. New York: Wiley.

Dzyabura, D., & Hauser, J.R. (2011). Active machine learning for consideration heuristics. *Marketing Science*, *30*, 801–819.

Einhorn, H. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly*, *36*, 367–378.

Einhorn, H.J., & Hogarth, R.M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, *13*, 171–192.

Eisenhardt, K.M., & Sull, D.N. (2001). Strategy as simple rules. *Harvard Business Review*, *79*(1), 106–116.

Fader, P.S., Hardie, B.G.S., & Lee, K.L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, *24*, 275–284.

Fix, E., & Hodges, J. (1951). Discriminatory analysis — Nonparametric discrimination: Consistency properties. *Project 21-49-004, report no. 4* (pp. 261–279). Randolph Field, TX, USA: USAF School of Aviation Medicine.

Furubotn, E.G. (2009). Heuristics, the non-maximizing firm and efficient allocation. *Metroeconomica*, *60*(1), 1–23.

Gaissmaier, W., & Marewski, J.N. (2011). Forecasting elections with mere recognition from small, lousy samples: A comparison of collective recognition, wisdom of crowds, and representative polls. *Judgment and Decision Making*, *6*, 73–88.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., Todd, P.M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Goldstein, D.G., & Gigerenzer, G. (2009). Fast and frugal forecasting. *International Journal of Forecasting*, *25*, 760–772.

Graefe, A., Armstrong, J. S., Jones, R. J., Jr., & Cuzan, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, *30*(1), 43–54.

Grünwald, P. (2005). Minimum description length tutorial. In P. Grünwald, I.J. Myung, & M.A. Pitt (Eds.), *Advances in minimum description length* (pp. 23–79). Cambridge: MIT Press.

Guercini, S. (2012). New approaches to heuristic processes and entrepreneurial cognition of the market. *Journal of Research in Marketing and Entrepreneurship*, *14*(2), 199–213.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Haldane, A., & Madouros, V. (2012). The dog and the frisbee. *Speech given at the Federal Reserve Bank of Kansas City's 36th Economic Policy Symposium "The Changing Policy Landscape", Jackson Hole, Wyoming, USA.*

Hand, D.J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*, 1–14.

Hand, D.J., & Yu, K. (2001). Idiot's Bayes: Not so stupid after all? *International Statistical Review*, *69*, 385–398.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hauser, J.R., Toubia, O., Evgeniou, T., Befurt, R., & Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, *47*(3), 485–496.
Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal of the Philosophy of Science*, *55*, 1–34.
Hoerl, A.E., & Kennard, R.W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*, 80–86.
Hogarth, R.M. (2012). When simple is hard to accept. In P.M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 61–79). Oxford: Oxford University Press.
Hogarth, R. M., & Karelaia, N. (2006). "Take-the-best"' and other simple strategies: Why and when the work "well" with binary cues. *Theory and Decision*, *61*, 205–249.
Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, *114*, 733–758.
Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63–91.
Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley and Sons.
Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, *58*, 1651–1684.
Jensen, D.D., & Cohen, P.R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, *38*, 309–338.
Katsikopoulos, K. V., & Martignon, L. (2006). Naïve heuristics for paired comparison: Some results on their relative accuracy. *Journal of Mathematical Psychology*, *50*, 488–494.
Li, M., & Vitányi, P. M. B. (1997). *An introduction to Kolmogorov complexity and its applications.* Berlin: Springer Verlag.
Little, J.D.C. (1970). Models and managers: The concept of a decision calculus. *Management Science*, *16*(8), B466–B485.
Liu, H., Hussain, F., Tan, C., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, *6*, 393–423.
MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*, 415–447.
Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.
Makridakis, S., Hibon, M., & Moser, C. (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society, Series A*, *142*, 97–145.
Markowitz, H. (1959). *Portfolio selection: Efficient diversification of investments.* New York: John Wiley & Sons.
Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparisons. *Theory and Decision*, *52*, 29–71.
Martignon, L., & Schmitt, M. (1999). Simplicity and robustness of fast and frugal heuristics. *Minds and Machines*, *9*, 565–593.
Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11170–11175.
Ng, A.Y., & Jordan, M.I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (pp. 14).
O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, *1*, 502–518.
Pearl, J. (1978). On the connection between the complexity and credibility of inferred models. *International Journal of General Systems*, *4*, 255–264.
Pitt, M.A., Myung, I.J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.
Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, *428*, 419–422.
Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *IT-42*, 40–47.
Rissanen, J. (2007). *Information and complexity in statistical modeling.* Berlin: Springer Verlag.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.
Rosenblatt, F. (1959). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 368–408.
Scheibehenne, B., & Bröder, A. (2007). Predicting Wimbledon 2005 tennis results by mere player name recognition. *International Journal of Forecasting*, *23*, 415–426.
Schmidt, F.L. (1971). The relative efficiency of regression and simple unit weighting predictor weights in applied differential psychology. *Educational and Psychological Measurement*, *31*, 699–714.
Schmitt, M., & Martignon, L. (2006). On the complexity of learning lexicographic strategies. *Journal of Machine Learning Research*, *7*, 55–83.
Schmittlein, D.C., Morrison, D.G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science*, *33*, 1–24.
Schwarz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics*, *6*, 461–464.
Schölkopf, B., & Smola, A. (2002). *Learning with kernels.* Cambridge, MA: MIT Press.
Seni, G., & Elder, J.F. (2010). *Ensemble methods in data mining: Improving accuracy through combining predictions.* San Francisco: Morgan & Claypool.
Serwe, S., & Frings, C. (2006). Who will win Wimbledon? The recognition heuristic in predicting sports events. *Journal of Behavioral Decision Making*, *19*, 321–332.
Shah, A.K., & Oppenheimer, D.M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *137*, 207–222.
Solomonov, R. J. (1964). A formal theory of inductive inference. *Part I. Information and Control*, *7*(1), 1–22.
Tukey, J.W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*, 1–67.
Tukey, J.W. (1977). *Exploratory data analysis.* Menlo Park, CA: Addison-Wesley.
Van Der Putten, P., & Van Someren, M. (2004). A bias–variance analysis of a real world learning problem: The CoIL challenge 2000. *Machine Learning*, *57*, 177–195.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213–217.
Webb, G. I. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, *4*, 397–417.
Wolpert, D.H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, 1341–1390.
Wright, M. W., & Stern, P. (2015). Forecasting new product trial with analogous series. *Journal of Business Research*, *68*(8), 1732–1738.
Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., et al. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*, 1–37.
Wübben, M., & Wangenheim, F. v. (2008). Instant customer base analysis: Managerial heuristics often get it right. *Journal of Marketing*, *72*, 82–93.
Yee, M., Dahan, E., Hauser, J.R., & Orlin, J.B. (2007). Greedoid-based noncompensatory inference. *Marketing Science*, *26*(4), 532–549.
Zahálka, J., & Železný, F. (2011). An experimental test of Occam's razor in classification. *Machine Learning*, *82*, 475–481.