Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Causal Interpretations of Black-Box Models

Qingyuan Zhao & Trevor Hastie

To cite this article: Qingyuan Zhao & Trevor Hastie (2019): Causal Interpretations of Black-Box Models, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2019.1624293

To link to this article: https://doi.org/10.1080/07350015.2019.1624293

Causal Interpretations of Black-Box Models

Qingyuan ZHAO
Department of Statistics, University of Pennsylvania, 400 Huntsman Hall, 3730 Walnut St, Philadelphia, PA 19104
(*qyzhao@wharton.upenn.edu*)

Trevor HASTIE
Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford CA 94305
(*hastie@stanford.edu*)

The fields of machine learning and causal inference have developed many concepts, tools, and theory that are potentially useful for each other. Through exploring the possibility of extracting causal interpretations from black-box machine-trained models, we briefly review the languages and concepts in causal inference that may be interesting to machine learning researchers. We start with the curious observation that Friedman's partial dependence plot has exactly the same formula as Pearl's back-door adjustment and discuss three requirements to make causal interpretations: a model with good predictive performance, some domain knowledge in the form of a causal diagram and suitable visualization tools. We provide several illustrative examples and find some interesting and potentially causal relations using visualization tools for black-box models.

KEY WORDS: Back-door adjustment; Data visualization; Machine learning; Mediation analysis; Partial dependence plot.

1. INTRODUCTION

A central task of statistics and machine learning is to study the relationship between *independent* or *predictor variables*, commonly denoted by X, and *dependent* or *response variables*, Y. Linear regression (and its generalizations such as logistic regression) persists as the main workhorse for this purpose and is being routinely taught at all levels of statistics courses. However, the legitimacy of linear regression (as a universal tool) has been seriously challenged from at least two angles:

1. When the goal is to *predict* Y using X, the predictive accuracy of linear regression is often far worse than other alternatives.

2. When the goal is to infer the *structural relationship*, coefficients of X in the linear regression may not have any causal interpretation.

Fundamentally, the problem is that linearity may be too simplistic in describing associational and structural relationships in real data.

Many researchers took these challenges in the past decades. Once considered minorities, two subjects—*machine learning* and *causal inference*—eventually grew out of these challenges and the developments are now largely embraced by the statistics community. In machine learning (in particular supervised learning), researchers have devised sophisticated or black-box algorithms such as random forests and neural networks to greatly improve the predictive accuracy of linear regression. Some statistical theory has also been developed to understand the behavior of these black-box algorithms. In causal inference, we now understand the basic assumptions that are necessary to identify the causal effect of X on Y using causal graphical models and/or counterfactual languages. In other words, to make causal inference we no longer require a linear model that correctly specifies the structural relationship.

Although machine learning and causal inference may both trace their origins to dissatisfaction with linear regression, the two subjects were developed mostly in parallel and the two communities had few shared research interests. But more recently, researchers in both fields start to realize that the theory and methods developed in the other field may help in fundamental challenges that have emerged in their own field. For example, a fundamental challenge in causal inference is the estimation of nuisance functions, for which machine learning may provide many useful and flexible tools (van der Laan and Rose 2011; Chernozhukov et al. 2018). Machine learning algorithms may also help us to discover heterogeneity in causal relations and optimize treatment decision (Hill 2011; Shortreed et al. 2011; Green and Kern 2012; Zhao et al. 2012; Wager and Athey 2018). This literature is actually exploding and only a short list of references is provided here to show the variety of problems and techniques that are being considered. The Atlantic Causal Inference Conference has even held a machine-learning-style competition to compare the various causal inference methods every year since 2016 (Dorie et al. 2019).

On the other side, the good performance of black-box machine learning systems may fail to generalize to the environments that are different from the training dataset. For example, Caruana et al. (2015) built several machine learning models to predict the risk of death among those who develop pneumonia. A rule-based learning algorithm learned a counterintuitive

Journal of Business & Economic Statistics
Month, 2019, Vol. 00, No. 0
DOI: 10.1080/07350015.2019.1624293
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

phenomenon that patients with asthma are less likely to die from pneumonia. However, this was due to an existing policy that asthmatics with pneumonia should be directly admitted to the intensive care unit and thus received better care. Thus, this model may be dangerous to deploy in practice because asthmatics may actually have much higher risk if not hospitalized. To resolve this problem, more intelligible or even causal models must be built (Caruana et al. 2015; Schulam and Saria 2017). Fairness is another crucial concern for deploying black-box models in practice. The definition of fairness and the solution to this problem may be closely related to causality (Kilbertus et al. 2017; Kusner et al. 2017). Causal interpretation of black-box models, the main topic of this article, is another good example for the usefulness of causal language and theory for machine learning researchers. Pearl (2018) postulated fundamental impediments to today's machine learning algorithms and summarized "sparks" from "The Causal Revolution" that may help us circumvent them.

There are a large number of excellent books, tutorials, and online resources for machine learning, some even directed to applied researchers working on causal questions (e.g., Mullainathan and Spiess 2017; Mooney and Pejaver 2018; Athey and Imbens 2019). In comparison, despite some good efforts to introduce causal inference concepts to the machine learning community (Spirtes 2010; Peters, Janzing, and Schölkopf 2017; Pearl 2018), the learning curve remains steep for researchers who are used to building black-box predictive models. We, the authors of this article, frequently encounter machine learning researchers and statisticians who find the language used in causal inference obscure, and we also deliberate upon causal concepts in our own research when the goal seemed to be prediction.

By exploring the possibilities of making causal interpretations of black-box machine-learned models, this article is aimed at introducing language, concepts, and problems in causal inference to researchers who are not trained to grasp such subtleties. We will assume no prior knowledge about causal inference and use a language that we believe will be most accessible to machine learning researchers. We will discuss a popular visualization tool for black-box predictive models, the *partial dependence plot* (Friedman 2001), and its generalization, *individual conditional expectation* (Goldstein et al. 2015), and use several examples from the UCI machine learning repository, a widely used public database for machine learning research. Our hope is that this will arouse broader interest in accomplishments and ongoing research in causal inference. More resources that we find helpful to learn about causal inference can be found at the end of this article.

2. INTERPRETATIONS OF BLACK-BOX MODELS

Interpretation is an ambiguous term and we will start with discussing possible interpretations of black-box models. Many if not most of the statistical analyses implicitly hold a *determinism* view regarding this relationship: the input variables X go into one side of a black box and the response variables Y come out from the other side. Pictorially, this process can be described by

A common mathematical interpretation of this picture is

$$Y = f(X, \epsilon), \tag{1}$$

where f is the law of nature and ϵ is some random noise. Having observed data that is likely generated from (1), there are two goals in the data analysis:

Science Extract information about the law of nature—the function f.

Prediction Predict what the response variables Y are going to be with the predictor variables X revealed to us.

In an eminent article, Breiman (2001b) contrasted two cultures of statistical analysis that emphasize on different goals. The "data modeling culture" assumes a parametric form for f (e.g., generalized linear model). The parameters are often easy to interpret. They are estimated from the data and then used for science and/or prediction. The "algorithmic modeling culture," more commonly known as machine learning, trains complex models (e.g., random forest, neural nets) that approximates f to maximize predictive accuracy. These black-box models often perform significantly better than the parametric models (in terms of prediction) and have achieved tremendous success in applications across many fields (see, e.g., Hastie, Tibshirani, and Friedman 2009).

However, the results of the black-box models are notoriously difficult to interpret. The machine learning algorithms usually generate a high-dimensional and highly nonlinear function $g(x)$ as an approximation to $f(x)$ with many interactions, making the visualization very difficult. Yet this is only a technical challenge. The real challenge is perhaps a conceptual one. For example, one of the most commonly asked question is the importance of a component of X. Jiang and Owen (2002) noticed that there are at least three notions of variable importance:

1. The first notion is to take the black-box function $g(x)$ at its face value and ask which variable x_j has a big impact on $g(x)$. For example, if $g(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ is a linear model, then β_j can be used to measure the importance of x_j given it is properly normalized. For more general $g(x)$, we may want to obtain a functional analysis of variance (ANOVA). See Jiang and Owen (2002) and Hooker (2007) for methods of this kind.

2. The second notion is to measure the importance of a variable X_j by its contribution to predictive accuracy. For decision trees, Breiman et al. (1984) used the total decrease of node impurity (at nodes split by X_j) as an importance measure of X_j. This criterion can be easily generalized to additive trees such as boosting (Freund and Schapire 1996; Friedman, Hastie, and Tibshirani 2000) and random forests (Breiman 2001a). Breiman (2001a) proposed to permute the values of X_j and use the degradation of predictive accuracy as a measure of variable importance.

3. The third notion is causality. If we are able to make an intervention on X_j (change the value of X_j from a to b with the other variables fixed), how much will the value of Y change?

Among the three notions above, only the third is about the science instead of prediction. Lipton (2018) discussed several other notions of model interpretability and acknowledges the difficulty of making causal interpretations. Next we will examine whether certain causal interpretations can indeed be made if we use the right visualization tool and are willing to make additional assumptions.

3. PARTIAL DEPENDENCE PLOTS

Our discussion starts with a curious coincidence. One of the most used visualization tools of black-box models is the partial dependence plot (PDP) proposed in Friedman (2001). Given the output $g(x)$ of a machine learning algorithm, the partial dependence of g on a subset of variables $X_\mathcal{S}$ is defined as (let $\mathcal{C}$ be the complement set of $\mathcal{S}$)

$$g_\mathcal{S}(x_\mathcal{S}) = \mathrm{E}_{X_\mathcal{C}}[g(x_\mathcal{S}, X_\mathcal{C})] = \int g(x_\mathcal{S}, x_\mathcal{C})\, \mathrm{d}P(x_\mathcal{C}). \quad (2)$$

That is, the PDP $g_\mathcal{S}$ is the expectation of g over the *marginal* distribution of all variables other than $X_\mathcal{S}$. This is different from the conditional expectation

$$\begin{aligned}\mathrm{E}[g(X_\mathcal{S}, X_\mathcal{C})|X_\mathcal{S} = x_\mathcal{S}] =& \mathrm{E}_{X_\mathcal{C}}[g(x_\mathcal{S}, X_\mathcal{C})] \\ =& \int g(x_\mathcal{S}, x_\mathcal{C})\, \mathrm{d}P(x_\mathcal{C}|X_\mathcal{S} = x_\mathcal{S}),\end{aligned}$$

where the expectation is taken over the conditional distribution of $X_\mathcal{C}$ given $X_\mathcal{S} = x_\mathcal{S}$. In practice, PDP is simply estimated by averaging over the training data $\{X_i,\ i = 1, \ldots, n\}$ with fixed $x_\mathcal{S}$

$$\bar{g}_\mathcal{S}(x_\mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} g(x_\mathcal{S}, X_{i\mathcal{C}}).$$

The consideration of *partial effect* of some independent variables $X_\mathcal{S}$ on a dependent variable Y is common in social science when model parameters are not immediately interpretable (e.g., King, Tomz, and Wittenberg 2000; Imai, Keele, and Yamamoto 2010; Wooldridge 2015). What is under the spotlight here is the distribution of $X_\mathcal{S}$ where the partial effect should be averaged over. An appealing property that motivated the proposal of PDP is that it recovers the corresponding individual components if g is additive. For example, if $g(x) = h_\mathcal{S}(x_\mathcal{S}) + h_\mathcal{C}(x_\mathcal{C})$, then the PDP $g_\mathcal{S}$ is equal to $h_\mathcal{S}(x_\mathcal{S})$ up to an additive constant. Furthermore, if g is multiplicative $g(x) = h_\mathcal{S}(x_\mathcal{S}) \cdot h_\mathcal{C}(x_\mathcal{C})$, then the PDP $g_\mathcal{S}$ is equal to $h_\mathcal{S}(x_\mathcal{S})$ up to a multiplicative constant. These two properties do not hold for conditional expectation.

Interestingly, Equation (2) that defines PDP is exactly the same as the famous back-door adjustment formula of Pearl (1993) to identify causal effect of $X_\mathcal{S}$ on Y from observational data. To be more precise, Pearl (1993) showed that if the causal relationship of the variables in (X, Y) can be represented by a graph and $X_\mathcal{C}$ satisfies a graphical back-door criterion (to be defined in Section 4.2) with respect to $X_\mathcal{S}$ and Y, then the *causal effect* of $X_\mathcal{S}$ on Y is identifiable and is given by

$$\mathrm{P}(Y|do(X_\mathcal{S} = x_\mathcal{S})) = \int \mathrm{P}(Y|X_\mathcal{S} = x_\mathcal{S}, X_\mathcal{C} = x_\mathcal{C})\, \mathrm{d}P(x_\mathcal{C}). \quad (3)$$

Here $\mathrm{P}(Y|do(X_\mathcal{S} = x_\mathcal{S}))$ stands for the distribution of Y after we make an intervention on $X_\mathcal{S}$ that sets it equal to $x_\mathcal{S}$ (Pearl 2009). We can take expectation on both sides of (3) and obtain

$$\mathrm{E}[Y|do(X_\mathcal{S} = x_\mathcal{S})] = \int \mathrm{E}[Y|X_\mathcal{S} = x_\mathcal{S}, X_\mathcal{C} = x_\mathcal{C}]\, \mathrm{d}P(x_\mathcal{C}). \quad (4)$$

Typically, the black-box function g is the expectation of the response variable Y. Therefore, the definition of PDP (2) appears to be the same as the back-door adjustment formula (4), if the conditioning set $\mathcal{C}$ is the complement of $\mathcal{S}$.

Readers who are more familiar with the potential-outcome notations may interpret $\mathrm{E}[Y|do(X_\mathcal{S} = x_\mathcal{S})]$ as $\mathrm{E}[Y(x_\mathcal{S})]$, where $Y(x_\mathcal{S})$ is the potential outcome that would be realized if treatment $x_\mathcal{S}$ is received. When $X_\mathcal{S}$ is a single binary variable (0 or 1), the difference $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ is commonly known as the average treatment effect (ATE) in the literature. We refer the reader to Holland (1986) for some introduction to the Neyman–Rubin potential outcome framework and the Ph.D. thesis of Zhao (2016) for an overview of the different frameworks of causality.

Next, we shall use several illustrative examples to discuss under what circumstances we can make causal interpretations by PDP and other visualization tools for machine learning algorithms.

4. CAUSAL MODEL

4.1. Structural Equation Model

First of all, we need a causal model to talk about causality. In this article, we will use the non-parametric structural equation model (NPSEM) of Pearl (2009, chap. 5). In the NPSEM framework, each random variable is represented by a node in a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set (in our case $\mathcal{V} = \{X_1, X_2, \ldots, X_p, Y\}$) and $\mathcal{E}$ is the edge set. A NPSEM assumes that the observed variables are generated by a system of nonlinear equations with random noise. In our case, the causal model is

$$Y = f(\mathrm{pa}(Y), \epsilon_Y), \quad (5)$$

$$X_j = f_j(\mathrm{pa}(X_j), \epsilon_j), \quad (6)$$

where $\mathrm{pa}(Y)$ is the parent set of Y in the graph $\mathcal{G}$ and the same for $\mathrm{pa}(X_j)$.

Notice that (5) and (6) are different from regression models in the sense that they are *structural* (the law of nature). To make this difference clear, consider the following hypothetical example

Example 1. Suppose a student's grade is *determined* by the hours she studied via

$$\text{Grade} = \alpha + \beta \cdot (\text{Hours studied}) + \epsilon, \quad (7)$$

where the noise variable ϵ is independent of "Hours studied." This corresponds to the following causal diagram

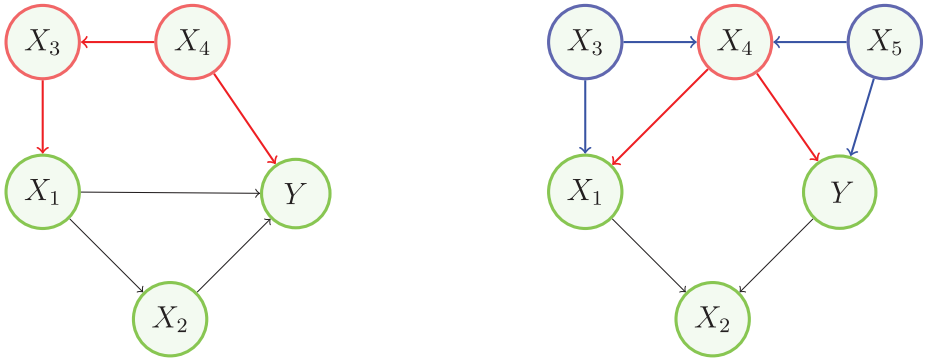


Figure 1. Two examples: the thick edges are back-door paths from X_1 to Y. $\{X_4\}$ blocks all the back-door paths in the left panel but not the right panel (because X_4 is a collider in the path $X_1 \leftarrow X_3 \rightarrow X_4 \leftarrow X_5 \rightarrow Y$).

If we are given the grades of many students and wish to estimate how many hours they studied, we can invert (7) and run a linear regression

$$\text{Hours studied} = \alpha' + \beta' \cdot \text{Grade} + \epsilon'. \tag{8}$$

Equation (7) is structural but Equation (8) is not. To see this, (7) means that if a student can study one more hour (either voluntarily or asked by her parents), her grade will increase by β on average. However, we cannot make such interpretation for (8). The linear regression (8) may be useful for the teacher to estimate how many hours a student spent on studying, but that time will not change if the teacher gives the student a few more points since "hours studied" is not an effect of "grade" in this causal model. Equation (8) is not structural because it does not have any predictive power in the interventional setting. For more discussion on the differences between a structural model and a regression model, we refer the reader to Freedman (2009) and Bollen and Pearl (2013).

Notice that it is not necessary to assume a structural equation model to derive the back-door adjustment formula (3). Here we use NPSEM mainly because it is easy to explain and is close to what a black-box model tries to capture.

4.2. The Back-Door Criterion

Pearl (1993) showed that the adjustment formula (3) is valid if the variables X_C satisfy the following back-door criterion (with respect to X_S and Y) in the DAG $\mathcal{G}$:

1. No node in X_C is a descendant of X_S; and
2. X_C blocks every "back-door" path between X_S and Y. (A path is any consecutive sequence of edges, ignoring the direction. A back-door path is a path that contains an arrow into X_S. A set of variables block or d-separates a path if the path contains a chain $X_i \rightarrow X_m \rightarrow X_j$ or a fork $X_i \leftarrow X_m \rightarrow X_j$ such that the middle node X_m is in the set, or the path contains a collider $X_i \rightarrow X_m \leftarrow X_j$ such that X_m nor its descendant is in the set.)

More details about the back-door criterion can be found in Pearl (2009, sec. 3.3). Heuristically, each back-door path corresponds to a common cause of X_S and Y. To compute the causal effect of X_S on Y from observational data, one needs to adjust for all back-door paths including those with hidden variables (often called unmeasured confounders).

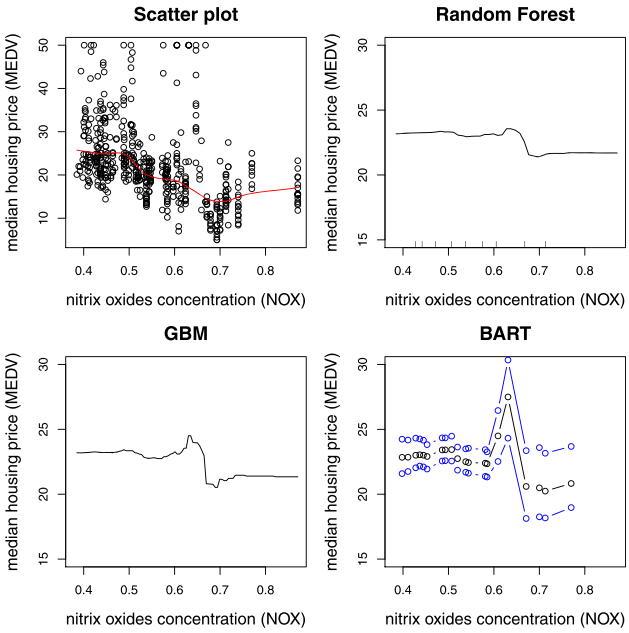
Figure 1 gives two examples where we are interested in the causal effect of X_1 on Y. In the left panel, $X_1 \leftarrow X_3 \leftarrow X_4 \rightarrow Y$ (in red color) is a back-door path but $X_1 \rightarrow X_2 \rightarrow Y$ is not. The set X_C to adjust can be $\{X_3\}$ or $\{X_4\}$. In the right panel $X_1 \leftarrow X_4 \rightarrow Y$ and $X_1 \leftarrow X_3 \rightarrow X_4 \leftarrow X_5 \rightarrow Y$ are back-door paths, but $X_1 \rightarrow X_2 \leftarrow Y$ is not. In this case, applying the adjustment formula (3) with $X_C = \{X_4\}$ is not enough because X_4 is a collider in the second back-door path.

Thus the PDP of black-box models estimates the causal effect of X_S on Y, given that the complement set C satisfies the back-door criterion. This is indeed a fairly strong requirement as no variables in X_C can be a causal descendant of X_S. Alternatively if C does not satisfy the back-door criterion, PDP does not have a clear causal interpretation and domain knowledge is required to select the appropriate set C.

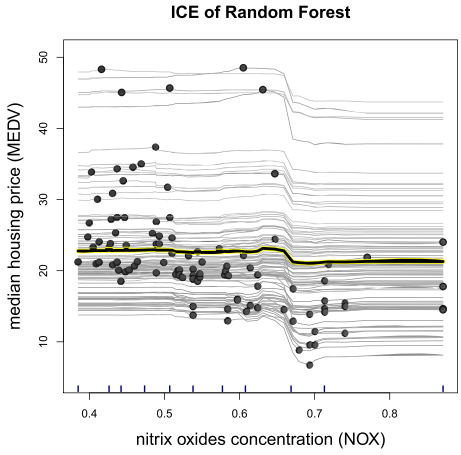
Example 2 (Boston housing data[1]). We next apply PDP for three machine learning algorithms in our first real data example. In an attempt to quantify people's willingness to pay for clean air, Harrison and Rubinfeld (1978) gathered the housing price and other attributes of 506 suburb areas of Boston. The primary variable of interest X_S is the nitrix oxides concentration (NOX, in parts per 10 million) of the areas, and the response variable Y is the median value of owner-occupied homes (MEDV, in \$1000). The other measured variables include the crime rate, proportion of residential/industrial zones, average number of rooms per dwelling, age of the houses, distance to the city center and highways, pupil-teacher ratio, the percentage of blacks and the percentage of lower class.

To obtain causal interpretations from the PDP, we shall assume that NOX is not a cause of any other predictor variables.[2] This assumption is quite reasonable as air pollution is most likely a causal descendant of the other variables in the dataset. If we further assume these predictors block all the back-

[1]Taken from *https://archive.ics.uci.edu/ml/datasets/Housing*.
[2]This statement, together with all other structural assumptions in the real data examples of this article, are only based on the authors' subjective judgment.



(A) Scatter plot and partial dependence plots using different black-box algorithms. The blue curves in the BART plot are Bayesian credible intervals of the PDP.



(B) ICE plot. The thick curve with yellow shading is the average of all the individual curves, i.e. the PDP. The dots indicate the actual NOX for each curve.

Figure 2. Boston housing data: impact of the nitrix oxides concentration (NOX) on the median value of owner-occupied homes (MEDV). The PDPs suggest that the housing price could be (causally) insensitive to air quality until it reaches certain pollution level. The ICE plot indicates that the effect of NOX is roughly additive.

door paths, PDP indeed estimates the causal effect of air quality on housing price.

Three predictive models for the housing price are trained using random forest (Liaw and Wiener 2002, R package `randomForest`), gradient boosting machine (Ridgeway 2015, R package `gbm`), and Bayesian additive regression trees (Chipman and McCulloch 2016, R package `BayesTree`). Figure 2(a) shows the smoothed scatterplot (top left panel) and the PDPs. The PDPs suggest that the housing price seem to be insensitive to air quality until it reaches certain pollution level around 0.67. The PDP of BART has some abnormal behaviors when NOX is between 0.6 and 0.7. These observations do not support the presumption in the theoretical development in Harrison and Rubinfeld (1978) that the utility of a house is a smooth function of air quality. Whether the drop around 0.67 is actually causal or due to residual confounding requires further investigation.

5. FINER VISUALIZATION

The lesson so far is that we should average the black-box function over the marginal distribution of some appropriate variables $X_{\mathcal{C}}$. A natural question is: if the causal diagram is unavailable and hence the confounder set $\mathcal{C}$ is hard to determine, can we still peek into the black box and give some causal interpretations? Of course this is not always possible, but next we shall see that a finer visualization tool may help us generate another kind of causal hypothesis, namely which variables $X_{\mathcal{M}}$ mediates the causal effect of $X_{\mathcal{S}}$ on Y.

5.1. Individual Curves

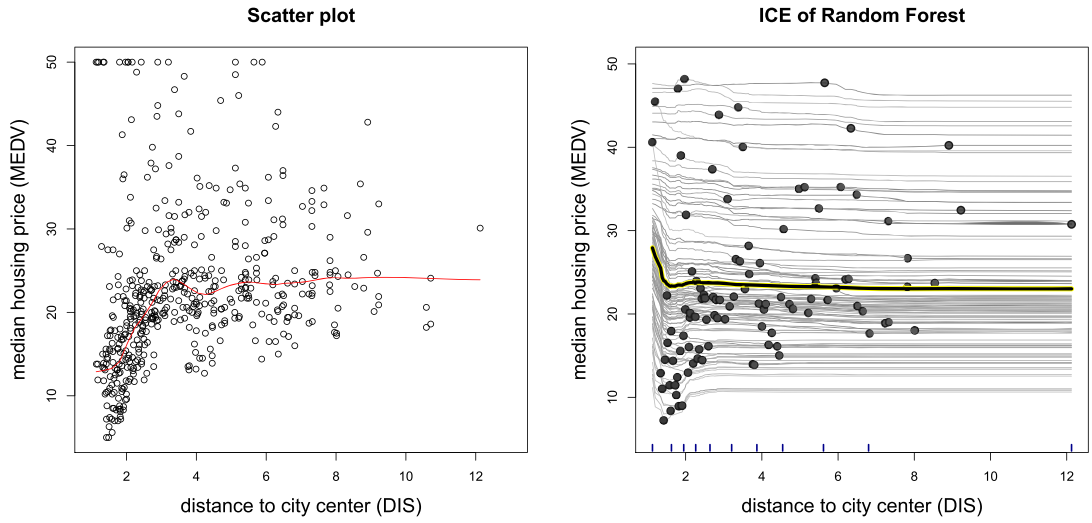
The *individual conditional expectation* (ICE) of Goldstein et al. (2015) is an extension to PDP and can help us to extract more information about the nature f. Instead of averaging the black-box function $g(x)$ over the marginal distribution of $X_{\mathcal{C}}$, ICE plots the curves $g(x_{\mathcal{S}}, X_{i\mathcal{C}})$ for each $i = 1, \ldots, n$, so PDP is simply the average of all the individual curves. ICE is first introduced to discover interaction between the predictor variables and visually test if the function g is additive (i.e., $g(x) = g_{\mathcal{S}}(x_{\mathcal{S}}) + g_{\mathcal{C}}(x_{\mathcal{C}})$).

Example 3 (Boston housing data, continued). Figure 2(b) shows the ICE of the black-box model trained by random forest for the Boston housing data. The thick curve in the middle (with yellow shading) is the average of all the individual curves, that is, the PDP. The solid black dots represent the actual value of $X_{i\mathcal{S}}$, so each curve shows what "might happen" if $X_{i\mathcal{S}}$ is changed to a different value based on the predictive model. All the individual curves drop sharply around NOX $= 0.67$ and are quite similar throughout the entire region. This indicates that NOX might have (or might be a proxy for another variable that has) an additive and non-smooth causal impact on housing value.

As a remark, the name "individual conditional expectation" given by Goldstein et al. (2015) can be misleading. If the response Y is truly generated by g (i.e., $g = f$), the ICE curve $g(x_{\mathcal{S}}, X_{i\mathcal{C}})$ is the conditional expectation of Y only if none of $X_{\mathcal{C}}$ is a causal descendant of $X_{\mathcal{S}}$ (the first criterion in the back-door condition).

5.2. Mediation Analysis

In many problems, we already know some variables in the complement set $\mathcal{C}$ are causal descendants of $X_{\mathcal{S}}$, so the



(A) Scatter plot.

(B) ICE plot. The thick curve (with yellow shading) in the middle is the average of all the individual curves, i.e. the PDP.

Figure 3. Boston housing data: impact of weighted distance to the five Boston employment centers (DIS) on median value of owner-occupied homes (MEDV). The ICE plot shows that longer distance to the city center has a negative causal effect on housing price. This is opposite to the trend in the marginal scatterplot.

back-door criterion in Section 4.2 is not satisfied. If this is the case, quite often we are interested in learning how the causal impact of X_S on Y is *mediated* through these descendants. For example, in the left panel of Figure 1, we may be interested in how much X_1 directly impacts Y and how much X_1 indirectly impacts Y through X_2.

Formally, we can define these causal targets through the NPSEM (Pearl 2014; VanderWeele 2015). Let X_C be some variables that satisfy the back-door criterion and X_M be the mediation variables. Suppose X_M is determined by the structural equation $X_M = h(X_S, X_C, \epsilon_M)$ and Y is determined by $Y = f(X_S, X_M, X_C, \epsilon)$. In this article, we are interested in comparing the following two quantities (x_S and x'_S are fixed values):

Total effect $\text{TE} = \text{E}[f(x_S, h(x_S, X_C, \epsilon_M), X_C, \epsilon)] - \text{E}[f(x'_S, h(x'_S, X_C, \epsilon_M), X_C, \epsilon)]$. The expectations are taken over X_C, ϵ_M and ϵ. This is how much X_S causally impacts Y in total.

Controlled direct effect $\text{CDE}(x_M) = \text{E}[f(x_S, x_M, X_C, \epsilon)] - \text{E}[f(x'_S, x_M, X_C, \epsilon)]$. The expectations are taken over X_C and ϵ. This is how much X_S causally impacts Y when X_M is fixed at x_M.

In general, these two quantities can be quite different. When a set C (not necessarily the complement of S) satisfying the back-door condition is available, we can visualize the total effect by the PDP. For the controlled direct effect, the ICE is more useful since it essentially plots $\text{CDE}(x_M)$ at many different levels of x_M. When the effect of X_S is additive, that is, $f(X_S, X_M, X_C, \epsilon) = f_S(X_S) + f_{M,C}(X_M, X_C, \epsilon)$, the controlled direct effect does not depend on the mediators: $\text{CDE}(x_M) \equiv f_S(x_S) - f_S(x'_S)$. The causal interpretation is especially simple in this case.

Example 4 (Boston housing data, continued). Here we consider the causal impact of the weighted distance to five Boston employment centers (DIS) on housing value. Since the geographical location is unlikely a causal descendant of any other variables, the total effect of DIS can be estimated by the conditional distribution of housing price. From the scatterplot in Figure 3(a), we can see that the suburban houses are preferred over the houses close to city center. However, this effect is probably indirect (e.g., urban districts may have higher criminal rate, which lowers the housing value). The ICE plot for DIS in Figure 3(b) shows that the direct effect of DIS has an opposite trend. This suggests that when two districts have the same other attributes, people are indeed willing to pay more for the house closer to city center. However, this effect is substantial only when the house is very close to the city (DIS < 2), as indicated by Figure 3(b).

6. MORE EXAMPLES

Finally, we provide two more examples to illustrate how causal interpretations may be obtained after fitting black-box models.

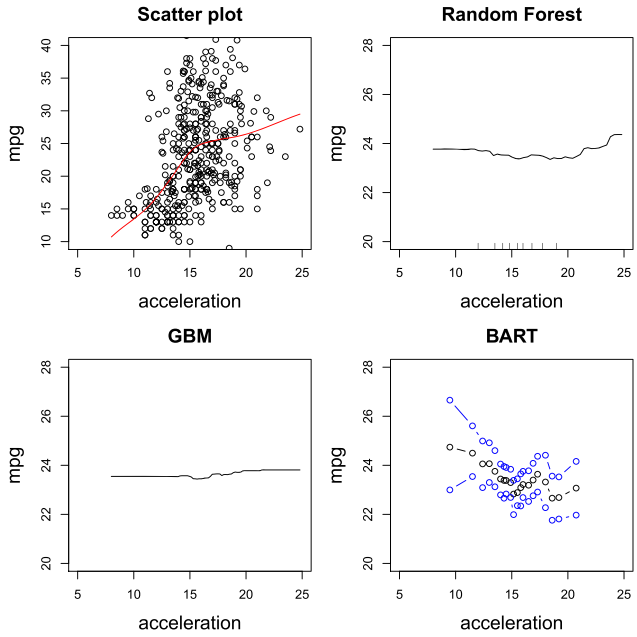
Example 5 (Auto MPG data[3]). Quinlan (1993) used a dataset of 398 car models from 1970 to 1982 to predict the miles per gallon (MPG) of a car from its number of cylinders, displacement, horsepower, weight, acceleration, model year and

[3]Taken from *https://archive.ics.uci.edu/ml/datasets/Auto+MPG*.

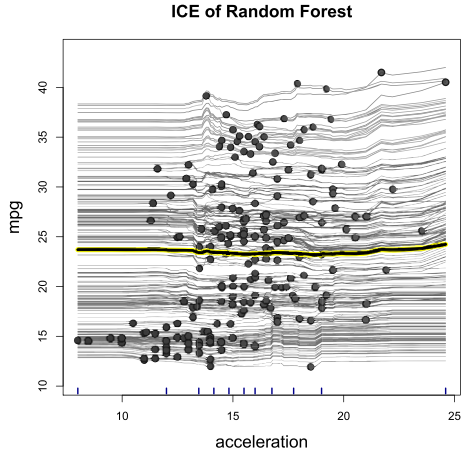
origin. Here we investigate the causal impact of acceleration and origin.

First, acceleration (measured by the number of seconds to run 400 m) is a causal descendant of the other variables, so we can use PDP to visualize its causal effect. The top left panel of Figure 4(a) shows that acceleration is strongly correlated with MPG. However, this correlation can be largely explained by the other variables. The other three panels of Figure 4(a) suggest that the causal effect of acceleration on MPG is quite small. However, different black-box algorithms disagree on the trend of this effect. The ICE plot in Figure 4(b) shows that the effect acceleration perhaps has some interaction with the other variables (some curves decrease from 15 to 20 while some other curves increase).

Next, origin (US for American, EU for European and JP for Japanese) are causal ancestors of all other variables, so its total effect can be inferred from the boxplot in Figure 5(a). It is apparent from this plot that Japanese cars have the highest MPG, followed by European cars. However, this does not necessarily mean Japanese manufacturers have the techno-

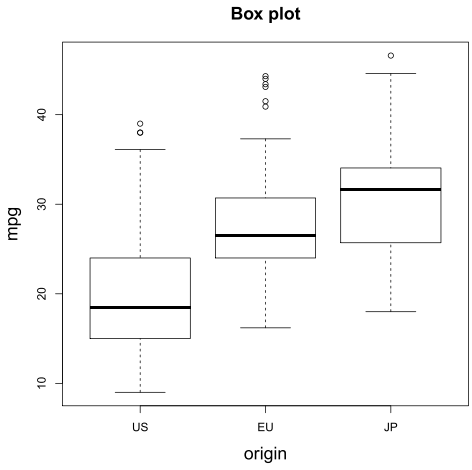


(A) Scatter plot and partial dependence plots using different black-box algorithms.

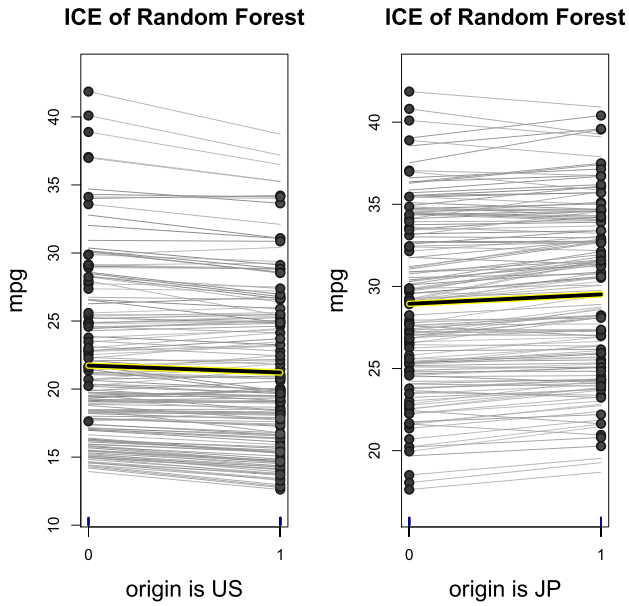


(B) ICE plot. The thick curve (with yellow shading) in the middle is the average of all the individual curves, i.e. the PDP.

Figure 4. Auto MPG data: impact of acceleration (in number of seconds to run 400 m) on MPG. The PDPs show that the causal effect of acceleration is smaller than what the scatterplot may suggest. The ICE plot shows that there are some interactions between acceleration and other variables.



(A) Box plot.



(B) ICE plots. The baseline (level 0) in both plots is origin being EU. For clarity, only 50% of the ICE curves in the left panel are shown.

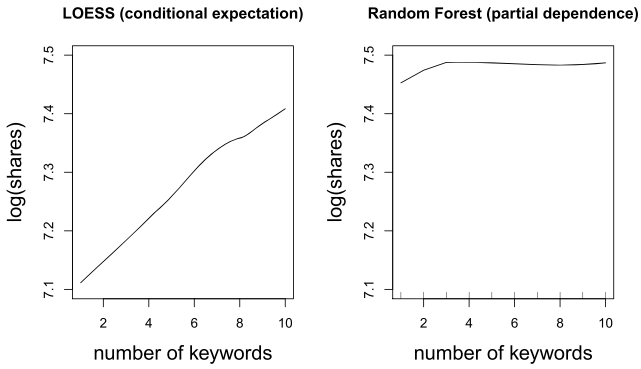
Figure 5. Auto MPG data: impact of origin on MPG. Marginally, Japanese cars have much higher MPG than American cars. This trend is maintained in the ICE plots but the difference is much smaller.

logical advantage of saving fuel. For example, the average displacement (the total volume of all the cylinders in an engine) of American cars in this dataset is 245.9 cubic centimeters, but this number is only 109.1 and 102.7 for European and Japanese cars. In other words, American cars usually have larger engines or more cylinders. To single out the direct effect of manufacturer origin, we can use the ICE plots of a random forest model, shown in Figure 5(b). From these plots, one can see Japanese cars seem to be slightly more fuel-efficient (as the ICE curves are mostly increasing) and American cars seem to be slightly less fuel-efficient than European cars even after considering the indirect effects of displacement and other variables.

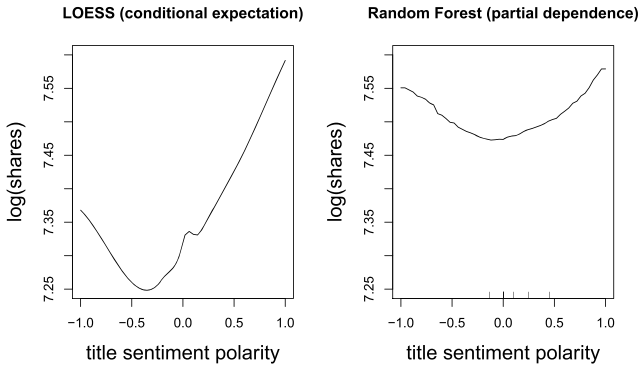
Example 6 (Online news popularity dataset[4]). Fernandes, Vinagre, and Cortez (2015) gathered 39,797 news article published by Mashable and used 58 predictor variables to predict the number of shares in social networks. For a complete list of the variables, we refer the reader to their dataset page on the UCI machine learning repository. In this example, we study the causal impact of the number of keywords and title sentiment polarity. Since both of them are usually decided near the end of the publication process, we treat all other variables as potential confounders and use the partial dependence plots to estimate the causal effect.

The results are plotted in Figure 6. For the number of keywords, the left panel of Figure 6(a) shows that it has a positive marginal effect on the number of shares. The PDP in the right panel shows that the actual causal effect might be much smaller and only occur when the number of keywords is less than 4.

For the title sentiment polarity, both the LOESS plot of conditional expectation and the PDP suggest that articles with more extreme titles get more shares, although the inflection points are different. Interestingly, sentimentally positive titles attract more reshares than negative titles on average. The PDP shows that the causal effect of title sentiment polarity (no more than 10%) is much smaller than the marginal effect (up to 30%) and the effect seems to be symmetric around 0 (neutral title).



(A) Impact of number of keywords on log of shares. The PDP shows that the actual causal effect might be much smaller than the marginal effect and only occur when the number of keywords is less than 4.



(B) Impact of title sentiment polarity on log of shares. Both plots suggest that extreme titles get more shares. The PDP shows that the causal effect might be much smaller than the marginal effect.

Figure 6. Results of online news popularity dataset.

7. CONCLUSION

We have demonstrated that it is possible to extract causal information from these models using the partial dependence plots (PDP) and the individual conditional expectation (ICE) plots, but this does not come for free. In summary, a successful attempt of causal interpretation requires at least three elements:

1. A good predictive model, so the estimated black-box function g is (hopefully) close to the law of nature f.
2. Some domain knowledge about the causal structure to assure the back-door condition is satisfied.
3. Visualization tools such as the PDP and its extension ICE.

For these reasons, we want to emphasize that PDP and ICE, although useful to visualize and possibly make causal interpretations about the black-box models, should not replace a randomized controlled experiment or a carefully designed observational study to establish causal relationships. Verifying the back-door condition often requires considerable domain knowledge and deliberation, which is usually neglected when collecting data for a predictive task. PDPs can suggest causal hypotheses which should be verified by a more carefully designed study. When a PDP behaves unexpectedly (such as the PDP of BART in Figure 2(a)), it is important to dig into the data and look for the root of spurious association such as unmeasured confounding or conditioning on a causal descendant of the response. Structural learning tools developed in causal inference may be helpful for this purpose, see Spirtes et al. (2000, chap. 8) for some examples.

Our hope is that this article can encourage more machine learning practitioners to peek into their black-box models and look for causal interpretations. This article only reviews the minimal language and concepts in causal inference needed

[4]Taken from *https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity*.

to discuss possible causal interpretations of PDP and ICE. There are many additional resources that an intrigued reader may find useful. Lauritzen (2001) is an excellent review of probabilistic graphical models for causal inference; a more thorough treatment can be found in Spirtes et al. (2000). Pearl (2009) contains many philosophical considerations about statistics and causal inference and also gives a good coverage of nonparametric structural equation model that is used in this article. Rosenbaum (2002); Imbens and Rubin (2015) focused on statistical inference for the causal effect of a single treatment variable X and another dependent variable Y. Morgan and Winship (2015) is a good introduction to causal inference from a social science perspective, and a book by Hernan and Robins (2019) gives a cohesive presentation of concepts and methods in causal inference to readers of a broader background.

ACKNOWLEDGMENTS

The authors would like to thank Dylan Small and two anonymous reviewers for their helpful comments.

FUNDING

Trevor Hastie was partially supported by grant DMS-1407548 from the National Science Foundation, and grant 5R01 EB 001988-21 from the National Institutes of Health.

[Received October 2018. Revised April 2019.]

REFERENCES

Athey, S., and Imbens, G. (2019), "Machine Learning Methods Economists Should Know About," arXiv no. 1903.10075. [2]

Bollen, K. A., and Pearl, J. (2013), "Eight Myths About Causality and Structural Equation Models," in *Handbook of Causal Analysis for Social Research*, ed. S. Morgan, Dordrecht: Springer, pp. 301–328. [4]

Breiman, L. (2001a), "Random Forests," *Machine Learning*, 45, 5–32. [2]

——— (2001b), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231. [2]

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Boca Raton, FL: CRC Press. [2]

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015), "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1721–1730. [1,2]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [1]

Chipman, H., and McCulloch, R. (2016), "BayesTree: Bayesian Additive Regression Trees," R Package Version 0.3-1.3, available at *https://CRAN.R-project.org/package=BayesTree*. [5]

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019), "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned From a Data Analysis Competition," *Statistical Science*, 34, 43–68. [1]

Fernandes, K., Vinagre, P., and Cortez, P. (2015), "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News," in *Portuguese Conference on Artificial Intelligence*, eds. F. Pereira, P. Machado, E. Costa, and A. Cardoso, Cham: Springer, pp. 535–546. [8]

Freedman, D. A. (2009), *Statistical Models: Theory and Practice*, New York: Cambridge University Press. [4]

Freund, Y., and Schapire, R. E. (1996), "Experiments With a New Boosting Algorithm," in *Proceedings of the 13th International Conference of Machine Learning*, pp. 148–156. [2]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 29, 1189–1232. [2,3]

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting," *The Annals of Statistics*, 28, 337–407. [2]

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015), "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics*, 24, 44–65. [2,5]

Green, D. P., and Kern, H. L. (2012), "Modeling Heterogeneous Treatment Effects in Survey Experiments With Bayesian Additive Regression Trees," *Public Opinion Quarterly*, 76, 491–511. [1]

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102. [4,5]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *Elements of Statistical Learning*, New York: Springer. [2]

Hernan, M. A., and Robins, J. M. (2019), *Causal Inference*, Boca Raton, FL: Chapman & Hall/CRC (forthcoming). [9]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [1]

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960. [3]

Hooker, G. (2007), "Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables," *Journal of Computational and Graphical Statistics*, 16, 709–732. [2]

Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71. [3]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, New York: Cambridge University Press. [9]

Jiang, T., and Owen, A. B. (2002), "Quasi-Regression for Visualization and Interpretation of Black Box Functions," Technical Report, Stanford University, Stanford. [2]

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017), "Avoiding Discrimination Through Causal Reasoning," in *Advances in Neural Information Processing Systems*, pp. 656–666. [2]

King, G., Tomz, M., and Wittenberg, J. (2000), "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science*, 44, 341–355. [3]

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017), "Counterfactual Fairness," in *Advances in Neural Information Processing Systems*, pp. 4066–4076. [2]

Lauritzen, S. L. (2001), "Causal Inference From Graphical Models" (Chapter 2), in *Complex Stochastic Systems*, eds. O. E. Barndorff-Nielsen and C. Kluppelberg, London: Chapman and Hall, pp. 63–107. [9]

Liaw, A., and Wiener, M. (2002), "Classification and Regression by Random Forest," *R News*, 2, 18–22, available at *http://CRAN.R-project.org/doc/Rnews/*. [5]

Lipton, Z. C. (2018), "The Mythos of Model Interpretability," *Queue*, 16, 30:31–30:57, DOI: 10.1145/3236386.3241340, ISSN 1542-7730. [3]

Mooney, S. J., and Pejaver, V. (2018), "Big Data in Public Health: Terminology, Machine Learning, and Privacy," *Annual Review of Public Health*, 39, 95–112. [2]

Morgan, S. L., and Winship, C. (2015), *Counterfactuals and Causal Inference*, New York: Cambridge University Press. [9]

Mullainathan, S., and Spiess, J. (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106. [2]

Pearl, J. (1993), "Comment: Graphical Models, Causality and Intervention," *Statistical Science*, 8, 266–269. [3,4]

——— (2009), *Causality*, New York: Cambridge University Press. [3,4,9]

——— (2014), "Interpretation and Identification of Causal Mediation," *Psychological Methods*, 19, 459. [6]

——— (2018), "Theoretical Impediments to Machine Learning With Seven Sparks From the Causal Revolution," arXiv no. 1801.04016. [2]

Peters, J., Janzing, D., and Schölkopf, B. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*, Cambridge, MA: MIT Press. [2]

Quinlan, J. R. (1993), "Combining Instance-Based and Model-Based Learning," in *Proceedings of the Tenth International Conference on Machine Learning*, pp. 236–243. [6]

Ridgeway, G. (2015), "gbm: Generalized Boosted Regression Models," R Package Version 2.1.1, available at *https://CRAN.R-project.org/package=gbm*. [5]

Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer-Verlag. [9]

Schulam, P., and Saria, S. (2017), "Reliable Decision Support Using Counterfactual Models," in *Advances in Neural Information Processing Systems*, pp. 1697–1708. [2]

Shortreed, S. M., Laber, E., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. (2011), "Informing Sequential Clinical Decision-Making Through Reinforcement Learning: An Empirical Study," *Machine Learning*, 84, 109–136. [1]

Spirtes, P. (2010), "Introduction to Causal Inference," *Journal of Machine Learning Research*, 11, 1643–1662. [2]

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000), *Causation, Prediction, and Search*, Cambridge, MA: MIT Press. [8,9]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning*, New York: Springer. [1]

VanderWeele, T. (2015), *Explanation in Causal Inference: Methods for Mediation and Interaction*, New York: Oxford University Press. [6]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [1]

Wooldridge, J. M. (2015), *Introductory Econometrics: A Modern Approach*, New York: Nelson Education. [3]

Zhao, Q. (2016), "Topics in Causal and High Dimensional Inference," PhD thesis, Stanford University. [3]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [1]