The Leadership Quarterly xxx (xxxx) xxxx

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Determining causal relationships in leadership research using Machine Learning: The powerful synergy of experiments and data science

Allan Lee[a,1], Ilke Inceoglu[a,*,1], Oliver Hauser[a], Michael Greene[b]

[a] University of Exeter Business School, University of Exeter, United Kingdom
[b] Deloitte Consulting, Boston, USA

ARTICLE INFO

Keywords:
Leadership effectiveness
Leadership processes
Machine Learning
Artificial intelligence
Causality
Experiments
Big Data
Heterogeneous treatment effects

ABSTRACT

Machine Learning (ML) techniques offer exciting new avenues for leadership research. In this paper we discuss how ML techniques can be used to inform predictive and causal models of leadership effects and clarify why both types of model are important for leadership research. We propose combining ML and experimental designs to draw causal inferences by introducing a recently developed technique to isolate "heterogeneous treatment effects." We provide a step-by-step guide on how to design studies that combine field experiments with the application of ML to establish causal relationships with maximal predictive power. Drawing on examples in the leadership literature, we illustrate how the suggested approach can be applied to examine the impact of, for example, leadership behavior on follower outcomes. We also discuss how ML can be used to advance leadership research from theoretical, methodological and practical perspectives and consider limitations.

Introduction

The ability to apply Machine Learning (ML) techniques to data collected in organizations holds great promise for application in leadership and management research more generally (e.g., Chaffin et al., 2017; Wenzel & Van Quaquebeke, 2018). ML is a subfield of computer science, referring to algorithms with ability to learn from patterns in data to make predictions of outcomes without constant supervision and reprogramming by a human (e.g., Mikalef, Pappas, Krogstie, & Giannakos, 2018). This type of approach can be highly effective when large amounts of data are available to analyze and is often – especially in disciplines outside of Computer Science and in practice – referred to as "Big Data Analytics" (e.g., Mikalef et al., 2018; Oswald, Behrend, Putka, & Sinar, 2020). Enormous amounts of data can be fed into ML models, which, if trained accordingly, can result in prediction models, that are produced quickly and free from some types of human bias (e.g., George, Osinga, Lavie, & Scott, 2016).

The use of ML in organizations has received increased interest (in research and practice) which can be partly attributed to the fact that "sophisticated technologies for collecting and storing data allow for an exponential increase in organizational data that can be collected" (Oswald et al., 2020, p. 506). Data gathered from these sources require more powerful processing and offer opportunities for exploration of research questions that in the past were not easily accessible. For instance, ML techniques have been widely applied to disease prognosis and prediction, such as predicting cancer susceptibility, recurrence and survival (e.g., Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015). Such predictive models can be transferred to an organizational setting where ML has been used to predict, for example, employee turnover (e.g., de Oliveira, Zylka, Gloor, & Joshi, 2019), return to work after sick leave (e.g., Na & Kim, 2019), physiological markers of stress (e.g., Bacciu, Colombo, Morelli, & Plans, 2018; Reddy, Thota, & Dharun, 2018), and employee performance (Kirimi & Moturi, 2016). For leadership scholars who are interested in understanding relationships between leadership and follower, team, or organizational outcomes, the application of ML to data collected in the field represents an opportunity to examine exactly such relationships and build powerful leadership models.

Leadership researchers primarily seek to understand leadership phenomena and processes to be able to *predict* the occurrence of leader and follower behaviors and/or *explain* the effects that leaders have and describe its causal underpinnings. Thus, leadership scholars typically attempt to answer two key questions that are of high theoretical and practical relevance: first, how can we best predict the behaviors, decisions, and performance of individuals, groups, or firms based on leader characteristics or behaviors? Second, identify the causes of those outcomes – for example, what interventions in leadership behavior would cause positive improvements in team performance?

* Corresponding author at: University of Exeter Business School, University of Exeter, United Kingdom.
E-mail address: i.inceoglu@exeter.ac.uk (I. Inceoglu).
[1] Joint first authors.

https://doi.org/10.1016/j.leaqua.2020.101426
Received 2 November 2019; Received in revised form 16 March 2020; Accepted 28 May 2020

The application of ML can help to address both questions related to prediction and causality in leadership research. However, while advanced analytical techniques such as ML have developed at pace to deal effectively with large datasets (e.g., Oswald & Putka, 2015), they are usually designed to solve prediction problems (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015). They do so by learning from large quantities of data how to make predictions on "out of sample" data with high accuracy. Moreover, while ML approaches can do well in addressing prediction problems, they are not often applied (correctly) to answer causal questions.

This paper will discuss the potential of using ML in research (and practice) to inform and extend leadership research and theory. We will discuss the application of ML in leadership research questions, addressing both prediction and causality questions through different research designs. As opportunities for using ML in organizations often go in tandem with the availability of Big Data (e.g., Oswald et al., 2020) our discussion will pay specific attention to the application of ML to Big Data in organizations. ML can, however, be applied to data that do not fall within common definitions of "Big Data" (see discussion of definition further below).

Researchers have recently begun to introduce ML tools designed to address questions of causal inference (Athey & Imbens, 2016; Wager & Athey, 2018). With these methods, it is possible to delve deeper into causal inference by combining ML techniques with experimental methods. Thus, a key objective of this paper is to propose steps to establish causal inference between variables in the realm of leadership research using randomized controlled experimental designs - the gold standard for causal inference (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Hauser, Linos, & Rogers, 2017) - and incorporate novel statistical techniques (Pearl & Mackenzie, 2018) to gain a deeper understanding of the inferences that can be drawn from the application and outcomes of experimental methods. We propose that combining the application of predictive algorithms and experimental designs to draw causal inferences through a recently developed technique to isolate heterogeneous treatment effects (Athey & Imbens, 2016; Wager & Athey, 2018) can help advance leadership theory, methodological approaches and create research with strong policy and practical implications (Antonakis, 2017).

By reviewing the application of ML (specifically, but not exclusively applied to Big Data) in the leadership domain, focusing on issues of prediction and causality, we seek to make three contributions to the literature. First, we disentangle how ML techniques can be used to inform predictive and causal models of leadership effects, respectively, and clarify why both types of model are important for leadership research. In discussing both predictive and causal models we provide guidance on how ML techniques can be used to extend our understanding of leadership effects on outcome variables such as follower performance and affect. Second, whereas previous research in management has focused on the use of ML (often in conjunction with Big Data) to inform predictive models, we highlight the potential for future research to extend these approaches to causal models and discuss the state-of-the-art techniques to apply ML to experimental designs. With these new methods, ML techniques can be used to delve deeper into causal mechanisms when combined with experimental methods. Thus, in addition to understanding if a leadership intervention will improve, for example, team performance or employee well-being, we highlight how ML tools can identify for which leaders and team members the intervention is likely to work particularly well (or badly). Third, in addition to discussing the theoretical implications of using ML to understand causal models, we provide practical guidance for how this can be done. In doing so, we seek to provide impetus for future research in this area and encourage researchers to consider these techniques to answer leadership questions for the range of study designs being used.

The remainder of this review is organized as follows. We begin by introducing concepts of ML and Big Data and discussing their application in management and leadership research. We then discuss how ML techniques can be used to explore leadership processes in both predictive and causal models. Finally, we identify key areas for future research that can help to produce a reliable and systematic body of evidence to serve as a platform for leadership theory development and trustworthy policy recommendations.

Machine Learning and Big Data

Both ML and Big Data are terms that tend to be used rather casually with differing definitions across academic disciplines and the business community. In order to understand their definitions (and perhaps misconceptualizations) and current applications in research and practice better, we briefly review the origins of these terms below.

ML has become a catch-all term for algorithms that predict anything. Often credited to Arthur Samuel (Samuel, 1959), the term ML first appeared much later (Koza, Bennett, Andre, & Keane, 1996) and refers to a computer program that learns patterns from data without being explicitly programmed. In his 1959 paper, Samuel suggested that a computer can be programmed so that it will learn to play a better game of checkers better than can be played by the person who wrote the program (Samuel, 1959). Today, ML algorithms sift through vast numbers of variables, looking for combinations that reliably predict outcomes. In some ways, this process resembles using the use of traditional regression models,[2] but ML is adept at handling enormous numbers of predictors and combining them in nonlinear and highly interactive and complex ways (Obermeyer & Emanuel, 2016). This functionality allows the application to new kinds of data, whose sheer volume or complexity would previously have made analyzing them difficult or impossible. In this paper we use the term ML synonymously with "Predictive Modeling" as defined by Donoho (2015, 2017): the application of algorithms trained on data with a focus on accurately predicting a specific outcome. The concept harkens back to the statistician Leo Breiman's call for more focus on predictive algorithms in addition to the traditional focus among statisticians on generative algorithms, where the attention instead lies in identifying the proper stochastic model for a process and examining the parameters resulting from fitting that model with the data (Breiman, 2001; Donoho, 2015, 2017). Donoho (2017) credits to Breiman (2001) that "the relatively recent discipline of ML, often sitting within computer science departments… [is] the epicenter of the Predictive Modeling culture." (p. 751). Popular use of the term ML is in this vein, a class of algorithms focused on estimating an outcome, with less focus on interpretation of the parameterization of the model itself.

ML algorithms are often grouped into categories such as supervised algorithms, unsupervised algorithms, and reinforcement learning algorithms (e.g., Hastie, Tibshirani, & Friedman, 2009). Simply put, supervised algorithms are those with a clearly defined dependent variable (i.e., an outcome, or target variable) which is described as a function of independent variables (i.e., covariates, predictive variables, or features) given a specified probability and loss functions (Hastie et al., 2009). Examples of supervised models include the linear regression model, classification tree, and artificial neural network. Unsupervised models, by contrast, lack a unique target variable on which to focus the optimization of the model training procedure. Instead, unsupervised models seek to describe all variables in a dataset by uncovering multivariate relationships. Examples of unsupervised models include Principal Components Analysis, the association rules algorithms (e.g., the a-priori algorithm), and clustering algorithms including k-means, hierarchical clustering, and Gaussian mixture models. Reinforcement learning is often considered a separate branch of ML, where an agent learns the behavior of a system through trial and error (Kaelbling, Littman, &

[2] In fact, standard regression techniques, such as ordinary least squares (OLS), can be viewed more appropriately as a special case of a more general set of ML techniques (Kleinberg et al., 2015).

Moore, 1996). Compared to supervised and unsupervised algorithms which work on static datasets, reinforcement learning relies on the dynamic ability to posit a new case and retrieve feedback, learning the outcome associated with the case. Reinforcement learning is an area of artificial intelligence at the forefront of current research (Dabney et al., 2020).

In this paper, our discussion of ML focuses on supervised ML for two reasons. First, while the literature on unsupervised ML and causality has seen some recent advancements (see An, Xiao, Yuan, Yang, & Alterovitz, 2019), uncovering causality in supervised ML has garnered more attention and, as a result, a more comprehensive toolkit is readily available (which we will be introducing in this paper). Second, the field of leadership research has traditionally studied well-defined outcome variables (such as specific leaders' or followers' behaviors or traits), which is the focus of supervised ML.

Like ML, Big Data has been defined in various ways but, in contrast to ML, there is no single well-accepted definition. The extensive use of digital technologies and the wide range of data-reliant applications have made the term "Big Data" pervasive across a range of disciplines including sociology, medicine, biology, economics, management, and information science (De Mauro, Greco, & Grimaldi, 2016). However, the popularity of this phenomenon has not been accompanied by the development of an accepted definition. Donoho's (2015, 2017) historical charting of the term helps understand the development of various conceptualizations of Big Data. Most studies that discuss Big Data treat the term as a "catch-all, amorphous phrase that assumes that all Big Data share a set of general traits" (Kitchin & McArdle, 2016, pg., 2016). For instance, it is common for researchers to suggest that Big Data possesses three traits (Laney, 2001): volume (consisting of enormous quantities of data), velocity (created in real-time), and variety (being structured, semi-structured and unstructured) – the "three Vs." However, analysis of 26 datasets revealed that Big Data do not all share the same characteristics and that there are multiple forms of Big Data (Kitchin & McArdle, 2016). Many authors use a wide range of defining characteristics which exceed the much cited three Vs (e.g., Shaffer, 2017). Kitchin and McArdle (2016) argue that to be considered "Big Data" datasets should possess the majority of the seven traits set out in Kitchin's (2013, 2014) typology of Big Data (volume, velocity, variety, exhaustivity, resolution and indexicality, relationality, extensionality and scalability), of which velocity and exhaustivity are the most important. For instance, rather than data being occasionally sampled (either on a one-off basis or with a temporal gap between samples), Big Data are typically produced much more continually. Exhaustive data sets refer to those where an entire system (such as an organization) is captured, rather than being sampled (Mayer-Schonberger & Cukier, 2013).

Within the field of organizational behavior and industrial/organizational psychology, new technologies have greatly expanded the type and amount of data that is accessible to researchers and practitioners (Guzzo, Fink, Tonidandel, King, & Landis, 2015; Oswald et al., 2020). For instance, the use of wearable devices to capture, simultaneously and rapidly, large quantities of data can be used to examine research questions that were previously difficult to address in the field. As highlighted by Wenzel and Van Quaquebeke (2018) in a review of Big Data in organizational and management research: "the act of gathering, analyzing, and interpreting Big Data is, by and large, unfamiliar territory" (p. 201) for management and leadership researchers. In these fields the lack of a clear and widely applicable definition is often acknowledged and has resulted in rather pragmatic approaches (with many authors adopting definitions encompassing three or four characteristic traits). Wenzel and Van Quaquebeke (2018), for example, define Big Data as "observational records that may be exceptionally numerous, highly heterogeneous, and/or generated at high rate and systematically captured, aggregated, and analyzed to useful ends" (p. 550) and also point to key drivers of Big Data: instrumentation (technological instruments that emit a range of data modalities), interaction (temporal interactions resulting in ordered records), and interconnection (communication, collaboration and creation of content). In their focal article on "Big Data recommendations for Industrial and Organizational Psychology," Guzzo et al. (2015) define Big Data "by more than just the volume, variety, and velocity of electronic records, however. It also encompasses new sets of tools and techniques for statistical analysis" (p. 493).

Acknowledging this definitional challenge, Oswald et al. (2020) propose to "remain practical and problem-focused as a way to accumulate practical intelligence on Big Data questions from a more bottom-up approach" (p. 506). In the realm of strategic Human Resource Management, Minbaeva (2017) advocates that evidence based, strategic decision-making in organizations requires "smart data" rather than Big Data, with smart data referring to "data that is organized, structured, and continuously updated" (p. 112). Notwithstanding these definitional issues, we will discuss various types of data that organizations collect as examples and would like to emphasize that for the purpose of our paper, it is the application of "learning" algorithms (i.e., ML) in leadership research — not Big Data by itself - that is fundamental.

Examples of continuously collected data in organizations include selection and assessment data (e.g., psychometric testing data, selection interview records), salary, performance data, promotion records, absence data, employee turnover data, or employees' and managers' free-text responses in annual performance appraisals. Indeed, a "defining characteristic of contemporary organizations is the rapid pace at which massive amounts of information are collected and stored" (McAbee, Landis, & Burke, 2017, p. 278). With increased application of sophisticated technology in Human Resource Management processes, data that are being collected continuously have grown dramatically in terms of volume and complexity, which explains why such data are typically labelled Big Data. For example, electronic performance monitoring - i.e., the use of technological means to observe, record, and analyze information that directly or indirectly relates to employee job performance (Bhave, 2014) - is now common within many organizations (Ravid, Tomczak, White, & Behrend, 2020). Examples include call monitoring, wearable sensors, e-mail and internet usage monitoring. Such information can be analyzed by ML tools to predict, for instance, employee personality, job attitudes, health, and performance (e.g., Kosinski, Bachrach, Kohli, Stillwell, & Graepel, 2014; Kozlowski, Chao, Chang, & Fernandez, 2020). Furthermore, Hauser and Luca (2015) argue that "data audits" in organizational research should be accompanied by increasingly readily available external data sources. Examples include collecting information from social networking websites (e.g., Facebook, Twitter) to evaluate/screen applicants during the selection process for employee recruitment (e.g., Brown & Vaughn, 2011; Roulin & Bangerter, 2013) or for research purposes to predict personality profiles (Kosinski, Stillwell, & Graepel, 2013). For leadership scholars who are interested in understanding relationships between, for example, leader and follower variables, the application of ML to Big Data has created major opportunities to examine exactly such relationships in organizations in the field. In the section below we will discuss how ML, especially applied to Big Data, can inform the development of predictive leadership models.

Machine Learning, Big Data, and predictive models in leadership research

The ability to apply ML in field research, coupled with the availability of "Big Data" in organizations, has the potential to expand our understanding of leadership processes and models. The reason for this is simple: as leadership researchers we are often interested in prediction; we want to know, for example, which leadership characteristics or behaviors will predict future, for instance, employee performance (e.g., Cavazotte, Moreno, & Hickmann, 2012) or well-being (Inceoglu, Thomas, Chu, Plans, & Gerbasi, 2018). We are interested in predicting who is likely to be promoted to a leadership position and become an

effective leader (e.g., Reichard et al., 2011). The application of ML is a powerful tool that can aid leadership scholars with optimizing such prediction models. For example, we know that certain leadership styles correlate with outcomes such as followers' job satisfaction and performance (e.g., Lee, Lyubovnikova, Tian, & Knight, 2020; Piccolo et al., 2012). ML can contribute to leadership research by identifying more, disparate, sets of variables that are predictive of effective leadership and have a positive effect on followers. For instance, a study by Spisak, van der Laken, and Doornenbal (2019) used ML to examine multiple personality and contextual predictors of leader effectiveness across a range of analytical methods, testing competing leadership theories at a level of complexity that was previously not conceived as being possible. Such tests allow for "naïve" data exploration without needing to specify a theoretical model beforehand. In fact, many contemporary applications of ML do not have a priori expectations, theories, or hypotheses about the underlying relations, but rather find patterns in the data to build predictive models (McAbee et al., 2017). This data-driven approach has similarities to inductive reasoning – representing a departure from typical approaches in leadership research which are theory-led and deductive, using quantitative data, or emphasize theory building and are inductive, using qualitative data. Observation-driven, exploratory approaches may identify patterns and highlight boundary conditions, thereby generating novel research questions (Woo, O'Boyle, & Spector, 2017). Such approaches can also be abductive: starting from an initial idea or "hunch" which is used to interpret empirical findings and generate plausible explanations (e.g., Van Maanen, Sørensen, & Mitchell, 2007). These insights (inductively or abductively derived) can provide new theoretical directions for future research in leadership, for example, by considering specific contextual variables which are receiving increasing attention and can be complex to model (e.g., Oc, 2018).

While ML algorithms excel at identifying patterns in data, they can suffer from "overfitting" – learning patterns within a given dataset so well that the algorithm will not make accurate predictions on another dataset stemming from the same process (Cawley & Talbot, 2010). To guard against overfitting, ML models are often iteratively trained and tested on separate datasets using methods such as bootstrapping, cross validation, or the jackknife method (Efron & Gong, 1983). These iterative steps aim to ensure that the resulting ML model is robust and, as much as possible, generalizable to new data which will be fed into the model in the future (Hastie et al., 2009). This iterative process makes ML techniques better equipped to handle and predict outcome(s) on new "unseen" data, allowing them to be employed across different contexts, datasets, and decision problems. While many ML techniques do require human-led decisions, one critical insight from the ML literature is that these models use an empirically automated way to minimize out-of-sample error, using the data itself to create a model and test it iteratively (Kleinberg et al., 2015). That is, while bootstrapping, cross validation, and jackknife work differently, the goal is the same: to estimate the variability of the predictions from the ML models on "out-of-sample" data. The bootstrap resamples observations with replacement, the jackknife holds out a single observation from training, and cross validation breaks the data into "folds" (i.e., non-overlapping subsets of the data), where the model is trained on some folds and predictions are made on the remaining folds. Although different mechanically, these methods ensure that the model is not overfitted on just one dataset while performing poorly when new data are added, but quite the opposite: the "train–test" processes ensure that ML models are optimized to do well in "out of sample" data (Hastie et al., 2009; Kleinberg et al., 2015).

Clearly ML can extend our understanding of leadership processes, for example by allowing us to test variables that predict leadership emergence, or which leader characteristics or behaviors predict follower outcomes. However, as with many analytical approaches, the use of ML is not without limitations, posing, for instance, risks of biased sampling or deceiving data quality (see review by Wenzel & Van Quaquebeke, 2018). Further, the application of ML to Big Data does not solve the fundamental problem of drawing causal inferences from observational data sets. This is, of course, an old issue in the social sciences, from psychology to economics to management: the ability "to *explain* behavior - that is, to accurately describe its causal underpinnings - and to *predict* behavior - that is, to accurately forecast behaviors that have not yet been observed" (Yarkoni & Westfall, 2017, p. 1).

While the goal of some leadership studies is prediction, for many it is causation. For instance, a common leadership study involves the examination a leadership variable (e.g., a leader's behavior of style), a distal outcome (e.g., team performance) and a more proximate mediating construct, such as follower motivation (Fischer, Dietz, & Antonakis, 2017). This type of study design attempts to explain the link between the leadership variable and outcome through some underlying mechanism that explains the causal relationship, following an input-process-output logic (see Fischer, Dietz, & Antonakis, 2017). One critical problem remains in a world of data science, ML and Big Data: the research designs used in the majority of such studies typically suffer from endogeneity issues (i.e., the predictor variable is correlated with the error term of the outcome, and/or mediator variable), and do not allow us to determine the causal relationships of the variables of interest, or whether the causal effects even exist at all (e.g., Antonakis et al., 2010; Hughes, Lee, Tian, Newman, & Legood, 2018). An endogenous predictor is related to the measured outcome/mediator in multiple ways, as a meaningful antecedent, but also in some unanticipated way(s) (e.g., common method bias, reciprocal effects, relationship with a common cause). Despite ever more data (access) and sophisticated statistical techniques, endogeneity problems are not solved with the application of ML. ML may be good at *predicting* outcomes, but the identified predictors are not typically *causes* (and even in those cases where they are, standard "out of the box" ML methods are not apt to identify them as such, requiring other ways to demonstrate causality). Thus, the usual caveats about not confusing correlation with causation still apply; in fact, they become even more important as researchers begin including potentially thousands or even millions of variables in statistical models (Obermeyer & Emanuel, 2016). Big Data may involve ongoing observations of discrete events with temporal ordering, which can facilitate more nuanced examinations of direction, magnitude, frequency, speed, and points of change associated with a phenomenon (Wenzel & Van Quaquebeke, 2018). However, although time series data can help support causal claims, which require that X precedes Y temporally and that X and Y are correlated beyond chance, many temporally ordered observations do not inherently demonstrate causality (Antonakis et al., 2010). This is because the third condition that is required to demonstrate causality is often not met: ruling out any other causes that could explain the relationship between X and Y (Kenny, 1979).

As with all organizational and management research, it is therefore imperative to clarify the objectives of one's research (what is the problem we want to address?) and to explicitly distinguish between issues of prediction and causality before deciding whether ML is an appropriate tool for examining this research question. The ability to predict an outcome or phenomenon using ML (without establishing causal relationships between variables) can of course be a valuable objective. As an illustrative example, consider team performance as an outcome variable of interest. Viewed as a prediction problem alone, an organization might ask – do we need to set aside a big pot of performance bonuses this year? We do not need to focus on the causes of good team performance in this case; instead we want to know what the outcome will be in a year's time and how much money to set aside. With enough historical data to train on, ML can help us get an accurate answer. However, viewed as a causal problem, one could ask: does a certain leadership behavior make it more likely that the team will perform well? This is a causal question – for example, whether a certain leadership behavior causes team performance to change. If so, identifying

these behaviors can help inform recruitment and leadership development processes in the organization (therefore addressing a different issue).

More recently, advances in econometrics and statistics have introduced the use of ML tools for estimating causal effects within subsets of the data (Athey & Imbens, 2016; Athey, Tibshirani, & Wager, 2019; Wager & Athey, 2018). With these new methods, ML techniques can be used to delve deeper into causal inference by combining them with experimental methods. So, in addition to understanding if a leadership intervention will improve team performance, these causal ML tools can identify for which leader and team characteristics the intervention works particularly effectively.

Machine Learning, Big Data and causal models in leadership research

Randomized controlled experiments are the gold standard for establishing causal inference (Hauser et al., 2017; Lonati, Quiroga, Zehnder, & Antonakis, 2018). Rigorously designed experiments, with randomly assigned treatment and control groups can establish that a treatment has a causal impact on an outcome through the direct manipulation of that treatment; for example, finding that a leader's charismatic speech caused an increase in workers' task output by about 17%, compared to workers who listened to a standard motivational speech (Antonakis, d'Adda, Weber, & Zehnder, 2014). However, while ML tools are often used to inform predictive models based on observational (a.k.a., non-experimental) data, until recently they have been seldom used to improve the ability of lab or field experiments to understand causal effects and draw inferences. Here we first discuss the foundation of experimental design before diving into the potential that ML offers experimentalists interested in examining causal relationships.

Causal inference has a simple premise – to understand whether a variable X has a causal effect on variable Y. We typically consider the role of an "intervention" (or "treatment") that manipulates X and we are interested in any change in Y as a result of manipulating X exogenously (for detailed information and background on experiments in management and leadership research, see, for example, Antonakis et al., 2010, Hauser et al., 2017; Hughes et al., 2018; here we summarize some key points). Rubin and collaborators first postulated and explored the "potential outcomes" framework (Rubin, 1974; Rosenbaum, 1984a, 1984b; Rosenbaum & Rubin, 1983a, 1983b; Holland & Rubin, 1983, 1987), providing a mathematical foundation for causal inference. Rubin noted that different outcomes exist for an individual observation – one outcome if the intervention is applied and one if withheld – however, in reality, only one of these potential outcomes can ever be observed; either the outcome for the observation when it was treated, or the outcome when it was not treated. While it will never be possible to observe both potential outcomes for the same individual observation (nobody can be "treated" and "not treated" at the same time), we can compare *groups* of individuals where one did receive the intervention (the "treatment" group) and one did not ("control" group). If the groups are similar enough before the treatment is applied, any difference in the outcome can then be attributed to the treatment. One intuitive method for creating similar groups is to randomly assign the treatment – i.e., individuals do not choose whether to receive the intervention. As a result of randomizing, the control and treatment groups are, in expectation, comparable (or exchangeable) in all possible dimensions (at least in expectation across both observable characteristics, such as gender, race or age, and unobservable factors, such as attitudes, behaviors or motivations). Thus, after the experiment is completed, it is then possible to compare the outcome of interest between the two groups and attribute the difference in outcome to the treatment, thereby establishing causality. The difference observed in the outcome between the groups is known as the average treatment effect (Angrist & Pischke, 2008; Glennerster & Takavarasha, 2013; Rubin, 1974). Today, this causal framework underpins much scientific discovery where randomized control trials and propensity score methods are deployed, having built on and displaced previous methods attributed to Fisher (1935), Kempthorne (1952), Cochran and Chambers (1965), and Cox (1958; see Holland, 1986). This rigorous definition of causality supplants previous methods such as those suggested by Granger (1969) that temporal considerations can indicate causality. While some approaches to understanding causality can use temporal separation, Granger's approach is not appropriate in all situations (Holland, 1988).

When randomization is not possible, other frameworks have been proposed for drawing causal inferences (for an overview, see Angrist & Pischke, 2008), such as regression discontinuity, instrumental variables, special instances of time series, leveraging graphical models (see Pearl, 2019), and propensity score matching (Imbens & Rubin, 2015; Rubin, 1974). These methods often rely on data that are akin to randomization, or where quasi-randomization can be inferred, even if the setting was not designed or implemented as a randomized experiment (see Antonakis et al., 2010 for an overview). For example, if a training course were offered to potential future leaders based on performance ratings, these approaches might focus on leaders just above and below the cut-off for admittance to the course. For simplicity, in this paper we focus on the experimental results of randomized control trials but most of our discussion of the methods proposed below can be extended to quasi-experimental methods as well.

Even when randomized control trials are used, leadership researchers are left with a conundrum. Experimental methods focus on the average treatment effect – that is, the impact of a treatment *on average* in a study population. Of further interest, researchers might want to know for whom the treatment works particularly well (or badly) – since there may be a distribution of smaller and larger effects around the average treatment effect in a population. For example, say a leadership researcher wants to understand the impact of bonuses on performance. Consider that a hypothetical study finds that the introduction of a one-time annual bonus improves performance by 10% on average in the treatment group, relative to the control group (or using another method mentioned above). A researcher may then also explore "heterogeneous treatment effects": treatment effects within specific groups of interest. For example, do bonuses improve performance more for women or men? Is performance boosted by bonuses more for experienced employees compared with less experienced employees? Traditionally, researchers would need to hypothesize in advance that an intervention is going to be useful for a specific subpopulation (guided by theory or prior empirical literature) to test for heterogeneous treatment effects. Alternatively, some do not pre-specify the analysis and engage in "data mining", a process that is frowned upon as the researcher could exploit "researcher's degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) – searching the data for any statistically significant results, which may result in spurious and non-replicable findings (Simmons et al., 2011). Even with prior theorizing and pre-specification, it is plausible that researchers may not hypothesize all relevant subgroups that would benefit from the population – perhaps in part because some groups are very specific (e.g. the bonuses may especially boost the performance of female hires with an engineering background) and would be missed, even though they may be practically and/or theoretically relevant.

Athey and Imbens (2016) and Wager and Athey (2018) proposed a novel solution to this problem using ML (which we collectively refer to as the Causal Forests method henceforth). Because ML methods excel at finding patterns by searching through data, guarded against overfitting by the methods described above (such as cross validation, bootstrapping or the jackknife), they are appropriate for empirically guided data exploration – and they enable researchers to do so while not exploiting researchers' degrees of freedom (Simmons et al., 2011), as much of the model testing and validation process is automated. In short, this method estimates heterogeneous treatment effects by casting a wide net of potentially relevant predictor variables to locate subpopulations that differ in the extent to which they respond to the

treatment. This enables new kinds of causal insights: for example, what would have been the causal effect of the intervention for an individual (based on certain covariates) had they been in the treatment group, not in the control group? For which individuals (based on certain covariates) was there no effect, or even a negative effect? Answering these questions can help in the application of future treatments into the field, deploying interventions where they will be most effective and least detrimental.

The Causal Forests method offers an empirical ML-enabled way to answer these questions. This approach usually starts with the application of a randomized experiment and then uses ML to discover and estimate treatment effect heterogeneity within relevant subpopulations. Specifically, an experiment would randomly assign an intervention to a treatment group while the control would group not receive the intervention. For example, a leadership development training could be tested as an intervention for managers to improve employee-manager relationships (we will discuss this example below further). Once the experiment has concluded, a ML method referred to as "Causal Random Forests" (Athey et al., 2019; Wager & Athey, 2018) can then be applied to identify those for whom the treatment worked most effectively (or least). This technique works by identifying comparable groups of individuals in both the control and treatment groups. The algorithm splits the data into partitions based on covariates, aiming to maximize the *causal* treatment effect between the treatment and control groups within a partition (i.e., the difference in outcomes between employees similar on covariates but who happen to be in either the control or treatment group). Intuitively, this method applies the same iterative process described above to answer the question: which variables (i.e., covariates) are most indicative of large treatment effects in the population? The result of applying the algorithm is that every individual in the study population receives an estimated "individualized" treatment effect – a measure of how large (or small) the treatment effect would have been had that individual been treated. This method allows researchers to identify heterogeneity in treatment effects, answering the question: for whom - based on ML - identified characteristics — did the treatment work particularly well (or badly)?

While this paper focuses on the Causal Forests method, it is worth noting that determining causality through ML is a growing area of interest, with other approaches having been studied as well. For example, the Transformed Outcome Trees (Beygelzimer & Langford, 2009; Sigovitch, 2007; Weisberg & Pontes, 2015), Fit Based Trees (Zeileis, Hothorn, & Hornik, 2008), and Squared t-statistic trees (Su, Tsai, Wang, Nickerson, & Li, 2009) all share the goal to make causal inferences but with differing technical approaches.

One reason we focus on the Causal Forests method is because it is a natural extension of Rubin's original causal model which offers desirable characteristics for causality, which we reviewed above. However, two other well-known other approaches that have found widespread application include Granger Causality (Granger, 1969) and probabilistic graphical models (Koller & Friedman, 2009). Granger Causality is a specific approach used with time-series data that tests for similarity in time-lagged variation. In this context, causality means that a change in a variable temporally precedes the change in another variable. While still widely applied as a causal method, Granger's formulation does not exclude the possibility of a third confounding variable that may cause both variables to change (Antonakis et al., 2010), making it less stringent than the Rubin causal model. Nonetheless, Granger Causality has found widespread appeal, and is sometimes referred to as a "predictive causality" (Diebold, 1998) because this time-series technique can be applied with great success to problems where time-lagged predictable variation is common, such as in neuroscience (Chockanathan, DSouza, Abidin, Schifitto, & Wismüller, 2019). Probabilistic graphical models, on the other hand, operate on a statistical basis; learning and drawing inferences from observational (i.e., non-experimental) datasets in which relationship between variables are statistically likely or unlikely to be causal linkages (Dawid, 2010). The Graphical Causal Bayesian Model, for example, uses a Bayesian score (Sucar, 2015, p. 243) to calculate the reliability of a causal relation between two variables. However, these models do not proceed from the potential outcomes framework and do not rely on randomized assignment to infer causality; instead, they offer a probabilistic pathway to causality (for a broader discussion of probabilistic causality and related concepts, see Hausman, 2010). While these approaches are useful, both Granger Causality and probabilistic graphical models suffer from a shortcoming common to other discussions of causality in statistics – outside of ML – that an unobserved confounding variable may be the ultimate cause for the observed change. Or, that these models are not counterfactual – where two groups are compared directly but for the applied intervention. This issue cannot be resolved without a proper, randomized control group – i.e., using the counterfactual logic of what would have happened in the absence of treatment, which prompted Pearl (2000) to introduce the language of the "do" operator. The "do" operator is a formalization in Pearl's Causal Calculus framework to explicitly convey the random (exogenous) variation of a variable of interest (in contrast to an endogenous variation often found in non-experimental data). Therefore, while the above approaches are often useful, we believe that Rubin's causal framework offers the most compelling basis to extend into machine learning – which is the proposition of the Causal Forests method.

Another reason to focus on the Causal Forests method is its effectiveness in determining causal relationships. Athey and Imbens (2016) quantitatively evaluated many well-known algorithms that aim to uncover causal effects using simulated datasets. The Causal Tree (Athey & Imbens, 2016) performs best at recovering the heterogeneous causal effects in point estimate and coverage. Wager and Athey (2018) further describe other research in the area include applying transformations to the outcome variable and applying the LASSO algorithm (Tian, Alizadeh, Gentles, & Tibshirani, 2014; Tibshirani, 1996, 2011), using the Random Forest algorithm to separately model outcomes for treated and control groups (Foster, Taylor, & Ruberg, 2011), using LASSO for estimating interaction effects (Imai & Ratkovic, 2013), with Bayesian additive regression trees (Green & Kern, 2012), and exploring linear outcomes under interactions with the treatment (Taddy, Gardner, Chen, & Draper, 2016). Other methods for exploring causality with ML include using targeted learning (Van der Laan & Rose, 2011), explicitly using experimental design principles (Rosenblum & van der Laan, 2011), adjusting confidence intervals to account for adaptive estimation (Wager & Walther, 2015), and brute force methods such as exhaustive search (e.g., Chisholm & Tadepalli, 2002). When directly compared, the Causal Forests method outperforms other methods in simulation comparisons (Athey et al., 2019; Athey, Imbens, Pham, & Wager, 2017; Athey, Imbens, & Wager, 2016; Wager & Athey, 2018).

Step-by-step guides

In the following, we offer a short and practical guide on to how the Causal Forests can be used. The focus of this guide is to introduce this method for practical application in leadership research. However, for completeness, we also cover the basic principles of a typical "ML prediction" approach and a typical "experimental" approach before explaining how the two can be combined in the Causal Forests method.

1. Machine learning approach

The traditional supervised ML approach aims to predict outcomes – to understand which variables *predict* an outcome variable. We will briefly outline how a researcher might go about testing and validating a ML algorithm for this prediction exercise. We accompany each step with an illustrative (hypothetical) example that is relevant to leadership scholars. For example, a leadership researcher may want to know what variables predict employees' well-being.

1. Define the problem. Commonly leadership researchers are interested in predicting individual, team or organizational outcome variables; such as employee, or team performance, well-being, attitudes or turnover. In addition to conventional measures of these outcomes (e.g., observational data in the form of questionnaires), technological devices enable the use of Big Data such as physiological indicators of well-being (Henning & van de Ven, 2017) or geospatial and verbal tracking data (Pentland, 2012).
2. Define a study population, of those individuals on whom we want to make predictions. For example, we focus on employees in an organization.[3]
3. Define the outcome variable. In our example, we focus on the employees' well-being, which we assume is measured on a scale from 1 to 5 through an employment engagement survey on an annual basis.[4]
4. Choose ML algorithms. The choice of algorithm depends on multiple factors such as the goal of prediction, the volume and nature of the data, and the outcome variable. Further, multiple algorithms can be compared for performance. For example, our outcome variable (employee well-being) could be interpreted as a continuous variable or an ordinal categorical variable. Appropriate models should be selected for comparison given the goals of the analysis and outcome variable, such as Ordinary Least Squares (OLS) for continuous or ordered probit in case of ordinal categorical. Further, if we operate under the assumption that we have a large number of potentially relevant predictors (i.e., covariates) in our dataset, an algorithm with coefficient-shrinkage, such as LASSO (Tibshirani, 1996, 2011) or Elastic Net (Zou & Hastie, 2005), may be an appropriate choice. Random Forests (Breiman, 2001), on the other hand, may be preferred if the number of predictors is particularly large, or we want an algorithm that requires little tuning and works well out-of-the-box, or the outcome variable is categorical in nature. While some algorithms such as artificial neural networks or "deep learning" networks work best with large datasets, others such as Gaussian processes or some types of Bayesian analyses are challenged computationally in the face of large amounts of data, working best with smaller datasets. As with any statistical model, choose an algorithm that best fits the problem. For an excellent introduction to a variety of models and their application in the social sciences, see Athey and Imbens (2019). Today, there exist many implementations of ML algorithms in for all popular statistical software programs; for first-time users, the "caret" package in R, the Python-based libraries such as scikit-learn or Keras (https://keras.io), are a good place to start, bringing together a large variety of ML algorithms with ample of documentation, tutorials and online help available.
5. Divide the data into "folds" (or use another "train–test" or validation technique). In cross validation — a popular validation technique — folds are k non-overlapping, randomly partitioned subsets of the data (in our example, employees with all available covariates). A common choice is $k = 10$, so that the data are split into 10 equally sized folds. To prepare the data for the training and testing procedure, the dataset must contain a column for the outcome variable (e.g., well-being) and J additional columns (covariates, or "features" in ML) that could be potential predictors of the outcome variable.[5] There can be a large number of J predictors, possibly even more than there are M rows or observations.[6] Some software packages, such as Python scikit-learn, will perform this cross-validation procedure for the user automatically.
6. Train and test the algorithm on the data. Each of the k folds is then used, in turn, for testing purposes with the others to train the model, which reduces the risk that the model is overfit. First, model parameters are estimated where the kth fold (for example, the 1st fold) is held as the test data. The model is trained on all other folds (in our example, folds 2 through 10) combined together. Next, the model is tested on the kth fold (which was held for test purposes) to see how well it does "out of sample." Then the algorithm moves on to the next fold (e.g., 2nd fold), which is held as the test data, where parameters are again estimated using the remaining k–1 folds (e.g., 1, 3, 4, 5, …, 10), and so forth.
7. Define a measure of "success". Unlike typical regression models based on frequentist statistics, ML algorithms do not usually evaluate success by returning p-values. This is in part because, given a large enough dataset, almost anything might be statistically significant (as defined by frequentist statistics) and no valuable insight would be gained. Instead, some ML algorithms, such as Random Forests, return lists of "variable importance" (Archer & Kimes, 2008) that aim to give the researcher an insight into which predictors play a particularly important role in predicting the outcome variable.

Once the algorithm has been tested the researcher is left with a list of "important" variables that are predictive of the outcome variable. For example, the researcher might find that a good employee-manager relationship, regular working hours and level of seniority are predictive of employee well-being at the firm.

2. Experimental approach

The experimental approach aims to test the *causal* effect of an intervention – to explain what *changes* an outcome. Here we briefly describe how a researcher might go about testing an intervention using an experiment. We continue with the example of employee well-being introduced above. Based on the (hypothetical) finding above that a good employee-manager relationship is an important predictor of employee well-being, a researcher might hypothesize that developing certain leadership qualities could improve the relationship between employee and manager and thereby improve employee well-being A leadership scholar might ask, for example, whether a novel leadership development training can improve a manager's ability to connect with their employees and, as a result, increase employee well-being. (This guide is a shortened version of Hauser and Luca (2015) and Hauser et al. (2017), which discuss each step in more detail, especially within a field organizational context.)

1. Define a study population, among whom the intervention will be tested. For example, we focus on managers (i.e., leaders) and employees (i.e., followers) in an organization. Note that the study population here is different from the ML example above, since the intervention will be delivered to managers, but the outcome measured at the employee level. For simplicity, we assume that every follower in our sample has exactly one leader and leaders supervise

[3] As with all empirical research, the study population is usually a sample from a much larger set. For example, even if we include all managers in one organization in our study, this is not the same as the population of all managers. Generalizability issues continue to exist and are not eliminated with ML or experiments.

[4] For simplicity, we focus on one time period only. Repeated measures as well as lagged variables from prior periods can be integrated but this goes beyond our simple guide; for more information and recommendations, see Ahmed, Atiya, Gayar, and El-Shishiny (2010) and Karch (2016).

[5] Unlike in traditional analyses where theoretical frameworks guide variable selection, there is no need to prune or select which predictors may be relevant. (*footnote continued*) One critical achievement of ML algorithms is that they use empirical methods to select and retain variables based on their predictive power.

[6] While the possibility of the number of predictors exceeding the number of observations may seem odd at first, a number of applications suffer from this unfortunate condition such as genomics or image and video processing. Unlike the classic Ordinary Least Squares (OLS) regression, a number of ML algorithms are able to operate when there are more covariates than observations, such as LASSO regression and Decision Trees.

exactly one follower, not an entire team.[7]
2. Define the outcome variable. In our example, we are interested in employees' (i.e., followers') well-being (e.g., measured through a regular employee engagement survey on a scale from 1 to 5).
3. Design an intervention. The intervention (i.e., treatment) is typically of primary interest to scholars, and the aim of the study is to identify the causal effect of the intervention on an outcome variable of interest. In our example, we assume that the intervention is a novel leadership development program that aims to make managers more empathetic to employees' concerns (for an overview of theories of behavior change using experimental treatments, see also Hauser, Gino, & Norton, 2018 and Rogers & Frey, 2014).
4. Divide the study population in two, *randomly* assigning half of the participants (in our example, randomly selected managers) into the treatment group and the other half into the control group. The treatment group receives the intervention, the control group does not.
5. Define the hypothesis to be tested. We might hypothesize that the leadership development training (our intervention) among leaders has a causal effect on followers' well-being (our outcome variable).
6. Define a measure of "success". In frequentist statistics, which are still predominant in the social sciences, if the *p*-value of the coefficient in front of the intervention dummy (treatment = 1, control = 0) in a standard OLS regression is below 5%, researchers typically declare the result to be statistically significant and the intervention to have been successful.

Thus, once the experiment has been run, we would simply run an ANOVA analysis or an ordinary least square (OLS) regression predicting employee well-being based on whether the manager has randomly been assigned to receive the intervention (leadership development training). The coefficient associated with the intervention dummy in the OLS regression is our measure of success. For example, the researcher might find that the *p*-value is below 5% (and the direction of the coefficient is in the predicted direction) and therefore conclude that the treatment *on average* had a causal effect on employee well-being (i.e., managers attending the training display more empathic behavior towards their employees, which has a positive impact on the employees' well-being).

3. The combined experimental-Machine Learning approach

The Causal Forests method is a combination of the aforementioned ML and experimental approach. Where ML focuses on prediction and the experimental approach isolates the causal pathway, the combination of both enables researchers to answer additional questions. For example, for whom (based on available covariates) did the treatment work particularly well? Assuming highly granular data on participants, the interpretation of the estimates from this method might no longer be described as an "average treatment effects" (at the study population level). Instead, this method edges closer to what one might describe as "individual treatment effects" (which has been described in the medical literature as "personalized medicine" or "precision medicine": Ghahramani, 2015; Mesko, 2017) and in marketing as "personalized advertising" (Matz, Kosinski, Nave, & Stillwell, 2017; Matz, Segalin, Stillwell, Müller, & Bos, 2019). Real "individualization" is, of course, not technically possible: as Rubin (1974) observed, no individual can ever be in two states at the same time (receiving the treatment and not receiving the treatment). However, given enough information (covariates) about every individual and a large enough sample size, the *causal effect* of a treatment on *similar* individuals can be estimated.

[7] These simplifying assumptions can be relaxed but doing so requires a hierarchical statistical model with clustered standard errors. While it is possible to introduce these modifications, we chose to keep it simple to illustrate the main concept without complications.

The general procedure of this method is as follows. First, the researcher runs an experiment with randomized control and treatment groups, as described above. As before, the experiment returns a (causal) average treatment effect of the intervention on an outcome variable. Then a similar ML algorithm is applied, using many potential predictors – but not to predict the outcome variable as before but to predict the *response to the treatment* (the difference between the control and treatment groups' outcome variable). For each individual, an estimate can be calculated – the extent to which an individual (identified by a large number of covariates) would respond to the treatment, compared to other individuals with similar covariates in the control group. The data can thus be partitioned into individuals who have (or would have, had they been in the treatment group) experienced large treatment effects, and those who have (or would have) experienced small, no or even negative treatment effects.

We explain this with the below step-by-step guide, continuing the example of employee well-being as the outcome of interest. Using the Causal Forests method, we are not only interested in understanding the average treatment effect of a leadership development training intervention but also the heterogenous treatment effects (based on leaders' and/or employees' covariates) for whom the intervention had the largest causal effects and which covariates were important in differentiating the heterogeneous treatment effects.

1. Define a study population, among whom the intervention will be tested. We again focus on managers (i.e., leaders) and employees (i.e., followers) in an organization.
2. Define the outcome variable. We continue with the focus on employees' (i.e., followers') well-being, measured on a continuous scale from 1 to 5.
3. Design an intervention. We assume the same intervention as before: a novel leadership development program that aims to make managers more empathetic to employees' concerns.
4. Divide the study population in two, *randomly* assigning half of the participants (here, managers) into the treatment group which receives the intervention and the other half into the control group which does not receive the intervention.
5. Run the experiment. At the end of the experiment, each employee-manager pair is associated with a column that contains the treatment status of the manager (1 = treatment, 0 = group) and the outcome variable of the employee (a continuous measure between 1 and 5), assessed at the end of the experiment.
6. Apply the Causal Forests algorithm. In principle, this algorithm follows a similar procedure as discussed above. Instead of predicting well-being directly, however, it estimates the *difference* in the potential outcomes were the individual assigned to the control and treatment groups. Based on the Random Forest algorithm (Breiman, 2001), the Causal Forest algorithm searches through the covariates, looking for the variables and splits that maximize the difference in the outcome between the treated and control groups. The process is iterated many times using different subsets of the data, resulting in a forest of decision trees. The algorithm is available for free in the R package "grf" (HYPERLINK "https://grf-labs.github.io/grf/" https://grf-labs.github.io/grf/; https://CRAN.R-project.org/package=grf; Athey et al., 2019). In addition to estimating the treatment effect for each individual, the algorithm estimates the variance of the treatment effect, allowing the researcher to evaluate if each individual treatment effect is non-zero. For our example, while the training might lead to higher *average* employee well-being, the Causal Forests method can identify a subset of the employees where the intervention substantially increased well-being (for example, due to low a priori well-being), and a subset where the well-being is not impacted (say, those who already report the highest well-being before the treatment was applied). In addition to estimating the individual treatment effects, the Causal Forest algorithm provides a variable importance measure – a numerical value for

each covariate, representing how important each variable is in differentiating the treatment effects. For example, the Causal Forest might indicate that the most important variables in determining treatment effects are due to manager experience, and personality traits of the employees such as introversion. These importance metrics can provide leadership researchers with insights into future areas of research.

7. Define a measure of "success". The outcome of this method is a quantifiable measure of treatment heterogeneity at an individual level – or, put differently, the extent to which an individual would respond to the treatment, as estimated by their covariates. In this setting, there is no well-defined concept of "success" as the information is the *heterogeneity* of the treatment success *across* individuals. (Although, arguably, general measures of model fit should be considered in evaluating whether the covariates are important and practically useful in the exercise.) The interpretation of these estimates is of interest, however. One way to look at the results is to split the data into partitions, by the magnitude to which the treatment would have improved (or worsened) the outcome for different individuals, relative to control.

The result of the Causal Forests method is an insight into the *heterogeneity* of causal treatment effects. While heterogeneity analyses have been done many times for specific subsamples in the past (for example, for men or women only; for low-income households; or for experienced or inexperienced managers), this approach gives the researcher a comprehensive and systematic approach to identify subsets of data that she would likely not be able to access otherwise or test for (or if she did, one might fear that she was "data mining" or "p-hacking" the data, which is not an issue when using the Causal Forests approach).

Furthermore, this approach not only provides the researcher with an empirical overview at the causal relationship of the treatment for specific subsets of the data, but also with a clear prediction for whom this causal relationship might hold in the future, bringing causal explanations together with prediction. For example, the researcher would not only be able to conclude that the intervention — the leadership development training — has on average had an effect on improve followers' well-being but she might also be able to show that it works best for inexperienced leaders who lead teams with introverts (an area of the organization where this intervention could be rolled out to with likely success), whereas the intervention does little to help experienced leaders or a team with many extroverts. Finally, this approach also enables the researcher to find pockets of the organization where the intervention could backfire – for example, if the team is composed of all male followers led by a female manager – providing both practically and theoretically relevant insights.

Discussion and future research

The aim of our paper was to highlight the application of ML in leadership research by focusing on questions of prediction and causal inference. We discussed and proposed practical steps for combining the application of predictive algorithms and experimental designs to examine causal relationships using a recently developed technique to isolate heterogeneous treatment effects (Athey & Imbens, 2016; Wager & Athey, 2018). We paid specific attention to the application of ML to Big Data as interest in this area by researchers and practitioners in management and organizational/industrial psychology has grown enormously with the advancement of technological devices and systems in organizations to collect, store and retrieve data.

We have also drawn attention to the enduring issue of confounding prediction and causation models in leadership and social science research more generally (e.g., Obermeyer & Emanuel, 2016; Yarkoni & Westfall, 2017) which is not resolved by merely applying ML. We observed that, for the most part, the use of ML (often applied to Big Data) within Organizational/Industrial Psychology and Management Research has been discussed in relation to predictive models based on correlational data. While such data can be useful for building predictive models, these models typically suffer from issues of endogeneity, not permitting us to use ML to make (strong) causal claims (Antonakis et al., 2010). To help clarify the distinction between prediction and causation models in the application of ML, our paper provided an overview of some key approaches available to examine prediction and causality using ML and considered limitations of these approaches.

We argue that the application of ML techniques combined with randomized experimental designs in order to isolate heterogeneous treatment effects can move the field in leadership research forward from both theoretical, practical and methodological perspectives. We have provided a step-by-step guide on how to design studies that combine experiments with the application of ML to establish causal relationships with maximal predictive power. By doing so, we aim to introduce these ideas that have been established in other disciplines (such as computer science) to the application in leadership research. Below we discuss the implications of our review for future leadership research.

1. Theoretical implications

Theory testing and refinement. Our suggested step-by-step guides provide the basis for leadership research to apply ML to Big Data to address questions related to both prediction and causality. Such techniques offer the opportunity for stringent theory testing of causal relationships in a deductive way, which can be followed by abductive or inductive approaches for subsequent theory refinement or new theory building. Specifically, causal models can be tested in a deductive way by first examining a treatment effect in a field experiment (for example, does training leaders to become more charismatic result in higher follower satisfaction with leadership; does training leaders in job design for increasing employee health/well-being improve the well-being of their employees). As discussed, isolating heterogeneous treatment effects can, for example, help to better understand which leaders or future leaders will benefit the most from a specific leadership training. This can be done using abductive or inductive approaches. Abductive approaches involve that there are personality differences in how people react to charisma or stressors at work but we do not know which combination of traits, so the enquiry could focus on follower personality profiles, for example we have a hunch. Inductive approaches involve using all person-related data available to test if any profiles emerge.

Theory building through exploratory approaches. New types of data and the ability to combine diverse (and numerous) data sources can help inform theory building, using inductive and abductive approaches. This can be done by drawing on quantitative data and qualitative data (e.g., text analysis of written documents, speech, email data, or image analysis of facial expression). Such approaches have started to emerge in leadership research, for example, by examining an exploratory question guided by theory which reflects an abductive approach (e.g., Spisak et al., 2019), or combining deductive approaches with inductive (e.g., Doldor, Wyatt, & Silvester, 2019). Purely exploratory approaches (e.g., letting patterns emerge from the data in predicting leadership effectiveness) can also be the starting point for new theory, or theory refinement - these can then be followed by a deductive approach, testing theory and causal models as suggested above. It appears that such approaches are being applied in practice in organizations (e.g., aiming to predict leadership performance in recruitment and selection processes using all available data) but a more systematic, scientific overall approach is needed to inform theory and practice, and transparency by publishing findings. We propose a more conscious choice of approach and systematic research design and agenda, where the research enquiry and research findings build on each other.

Modeling time. Time is integral to leadership development and

performance and leadership phenomena reflect dynamic processes (Fischer et al., 2017; Shamir, 2011). ML techniques can be used to explore leadership processes in more complex ways than incorporate longitudinal data (e.g., Gruda & Hasan, 2019). New technologies (e.g., sensors) allow for more continuous measurement of data, drawing on a range of data sources (e.g., physiological data measuring stress response; behavioral data reflected in email patterns, performance), which can help examine leadership processes and phenomena better through consideration of time such as the unfolding of leadership development and performance (e.g., Fischer et al., 2017) or leadership behavior and employee well-being (Arnold, 2017; Inceoglu et al., 2018). Other areas in management research have begun to examine the role of time in more depth conceptually and empirically (e.g., Organizational Citizenship Behavior: Methot, Lepak, Shipp, & Boswell, 2017; newcomer identification: Zhu, Tatachari, & Chattopadhyay, 2017; motivation and performance: Roe, 2014) but research on leadership processes and phenomena is still emergent (e.g., Castillo & Trinh, 2018; Lang, Bliese, & de Voogt, 2018; Shamir, 2011). Complex question relating to how leadership processes emerge and have effects over time can be answered with a combination of ML and Big Data.

Modeling context. Attempts to specify situational/context variables in leadership in theoretical models (e.g., Fiedler, 1967) have only been moderately successful with empirical evidence not (fully) supporting these models (e.g., Peters, Hartke, & Pohlmann, 1985). As a result, situational approaches of leadership have somewhat faded from the literature, most likely because of the inherent complexity of developing models that capture key situational factors that are relevant in predicting, for example, leadership effectiveness. While many studies consider boundary conditions in the form of moderator variables, we do not have robust, empirically supported theoretical models for incorporating context into leadership theory. Contextual approaches to leadership have started to develop again, by analyzing contextual factors that may influence the leadership process in more detail as proposed by the integrative framework linking context to leadership by Oc (2018). Applying ML to leadership research can help model such complex contextual variables and contingencies, also in combination with time (context change over time) especially when developing and testing causal models.

2. Methodological implications

Identifying instrumental variables. A criticism that is often levelled at the leadership literature is the failure to address issues of endogeneity between independent and dependent variables (e.g., Antonakis et al., 2010; Hughes et al., 2018). Researchers typically model survey data using random effects and multilevel models, which assume that the random effects are uncorrelated with the regressors. Violation of this assumption creates an endogeneity problem (e.g., Antonakis et al., 2010). While the ideal solution to address this issue is the use of randomized controlled experiments, for many research questions, the use of experimental designs may not be practical, ethical, or suitable. Such methods can also suffer from a lack of ecological validity if not well designed. An alternative approach to addressing endogeneity issues in correlational data is to adopt the instrumental-variable estimation otherwise known as the two-stage least squares (2SLS) procedure. 2SLS uses instrumental variables – exogenous predictors of the endogenous predictor (X) which will be associated with the outcome (Y) but only through the endogenous predictor (X) (Antonakis et al., 2010). Instrumental variables are widely used in other disciplines, such as economics, but have had limited use within organizational research, including in the field of leadership. This is perhaps partly because identifying appropriate instrumental variables can be difficult, and poorly chosen instruments (especially if not strongly related to the endogenous predictor) can bias findings (cf. Larcker & Rusticus, 2010). Big Data has the potential to help with this issue by providing a range of potential instruments that can be incorporated into leadership research. Furthermore, ML techniques can be useful in identifying such instrumental variables. With careful consideration it is likely that many instrumental variables could be extracted from Big Data and be incorporated within a ML predictive model to help to address issues of endogeneity. For example, combining Big Data and research from genetics, it is very possible that we will be able to estimate the causal effects of genetic markers of, for example, intelligence on outcomes such as leadership effectiveness (e.g., see DiPrete, Burik, & Koellinger, 2018; Von Hinke, Smith, Lawlor, Propper, & Windmeijer, 2016).

Making use of new/additional data sources. Big Data and ML provide an opportunity to explore a wider range of variables and allow for more complex testing of different variable profiles and situational characteristics. For instance, the leader trait paradigm can be further explored using a range of ML methods allowing optimal sets of potential leader trait covariates to be identified, examining the effects of multiple interactions simultaneously, and incorporating non-linear relationships (e.g., Spisak et al., 2019). Big Data may also allow for leadership researchers to make greater use of a variety of objective data to reduce common method bias. To date, the leadership literature has largely relied on survey data (e.g., Hughes et al., 2018), not making use of the data that many organizations collect with regularity that could help us build predictive models of leadership. For example, using ML expansions in computer science, such as social sensing (e.g., analysis of emotional and behavioral cues) and natural language processing, have grown in popularity (LeCun, Bengio, & Hinton, 2015). For instance, research has explored how algorithms can give managers awareness of how their employees are feeling (Waddell, 2016) and research has examined how automatically sensed behavior can predict job performance (Schmid Mast, Gatica-Perez, Frauendorfer, Nguyen, & Choudhury, 2015). Incorporating more diverse data, like this, into our leadership models may help build new theory and better understand our current theories.

Data quality and interpretation of results. Throughout this paper, we have posited potential benefits of applied ML techniques to Big Data. It is important to caveat our suggestions and highlight that more data does not necessarily mean good data (Tonidandel, King, & Cortina, 2018; see also Minbaeva, 2017: “smart data”). Furthermore, it is important to highlight the fact that many leadership researchers will be unfamiliar with ML techniques. Previous reviews have been quick to highlight both the benefits and drawbacks of using ML and Big Data (e.g., Tonidandel et al., 2018; Wenzel & Van Quaquebeke, 2018). For instance, the inclusion of multi-faceted measures (i.e., physiological data, location monitoring, non-linguistic social signals) can create issues surrounding data integrity and difficulties combining data of different types. An additional concern with some of the measures used in Big Data studies is construct validity (see Tonidandel et al., 2018). If not carefully managed, Big Data has the potential to exacerbate the “Garbage-In, Garbage-Out” phenomenon.

Technical skills required for the application of ML. The effective use of ML to inform leadership research requires training in certain areas with software with which many of us are unfamiliar (e.g., Hadoop, MapReduce, Python) and statistical techniques we have likely had no previous exposure to (Tonidandel et al., 2018). These issues make the use of multidisciplinary teams even more essential in the future.

3. Practical implications

Developing interventions. Determining causal factors through experimental research and isolating heterogenous treatment effects can help develop more targeted evidence-based interventions in organizations. For example, subpopulations can be dynamically targeted using ML algorithms, in order to make policies or organizational interventions more effective and efficient (see for example, Einav, Finkelstein, Mullainathan, & Obermeyer, 2018; Hauser, Greene, DeCelles, Norton, &

Gino, 2018). In leadership research, there exist many potential applications: leadership training to increase team effectiveness could be more focused on aspects of leadership behavior (e.g. charisma) and the leadership process (e.g. specific types of interactions with the team, the quality of the leader-follower relationship) that have been shown to have a positive effect on teams in (a specific) organization. We would also know which leaders would benefit most from a specific type of training (e.g. extroverts) and could tailor the training accordingly. Such an evidence-based approach would be more effective compared to one-size-fits all approaches and save costs by focusing trainings on desired outcomes. Additional positive "side effects" could include higher training motivation of leaders and employees which in itself is likely to increase training success.

Collaborations between organizations and researchers. Organizations have started using ML approaches to address a wide range of business issues, for example in Human Resource Management (e.g. Minbaeva, 2018). Leadership researchers need to be more closely involved in developing research with organizations in that area to help with the choice of research approach (e.g. data driven, inductive or starting with initial ideas using an abductive approach), data and research design (e.g. designing field experiments). Closer collaboration requires a dialogue with organizations in which researchers also need to learn about available data and contextual variables, help organizations to understand the benefits of differentiating between prediction and causation models and, quite crucially, to identify the research questions at hand (What issue are we trying to solve? How can existing or new data help address a specific issue?). More systematic, evidence-based theory guided approaches will help organizations move from operational data reporting to using data for strategic decision-making (Minbaeva, 2017). The need for closer collaboration between organizations and researchers to improve evidence-based interventions and applied science is not new, but with the rapid development ML applications and use of Big Data we argue that such close collaborations will become even more important.

Training in organizations to use ML effectively and understand Big Data better.

Organizations have started to employ data scientists to apply ML to Big Data (e.g. in HRM) to address business issues and make work processes more efficient. One common challenge observed in practice is the interpretation of data outputs by other business functions, by leaders and employees without a technical background (e.g. Sinar, Ray, & Canwell, 2018). Using ML to address business issues and move leadership research forward, requires training of researchers but also practitioners, leaders and stakeholders in organizations to be able to interpret and communicate findings for decision-making. This involves, for example, understanding basic principles of ML, and forms of data used for ML (including limitations and pitfalls), the distinction between prediction and causation and the application of Big Data and ML for examining prediction and causal relationships.

4. Limitations of approaches and issues to consider

Spurious findings. ML methods applied to very large data sets can lead researchers to falsely reject the null hypothesis and may uncover many significant relationships that are spurious (Fan, Han, & Liu, 2014), although many ML techniques are now able to guard against spurious findings with appropriate penalization (for more details, see Tibshirani, 1996, 2011). Even when statistically significant results are not spurious, researchers have to examine critically whether a result also has practical (i.e. economic) significance. As Yarkoni and Westfall (2017) put it, despite the hyped-up Big Data revolution and the use of increasingly sophisticated methods, "explanatory utility of a massively complex model fitted to enormous amounts of data remains somewhat unclear" (p. 9).

Issues of (cross-) validation. Despite being praised for their ability to do well in "out of sample" prediction, ML algorithms are not universally applicable to all datasets and all problems after conducting an initial training and validation exercise. "Out of sample" does have its limits. Within the realm of ML, "out of sample" narrowly refers to the ability to predict behaviors in a dataset that is similarly structured, similarly sourced, and has similar properties (such as patterns of behaviors that have been identified). This, conversely, means that data that has different predictors or data structures, comes from a different source, or shows different patterns of behavior altogether. For example, an algorithm trained and tested on human resource data in a large financial firm to predict employee turnover does not necessarily (and not likely) do well in predicting employee turnover in a dataset from a manufacturing company or marketing agency. Furthermore, similar predictors may not be available or comparable across industries – and even if comparable predictors can be found, questions related to specific predictions and the size of coefficients arise. For example, should the weight on the coefficient of a trader's stock performance, be the same as a plant manager's processing efficiency or a marketing agent's creative portfolio? Furthermore, the patterns (such as a list of important coefficients that predict employee turnover) that the ML algorithm has learned in the financial firm dataset will likely not hold in other industries. Finally, while it seems plausible that an algorithm trained and tested on one financial firm's data may be relevant for another financial firm's data, it is still advisable to treat them as separate to examine distinct problems/questions – the availability of predictors may vary, internal politics and interpersonal dynamics within a firm may play more or less a role, or different financial incentives might alter behavior in unexpected ways. Thus, while it is technically possible to use the method with data that comes from multiple firms, whether doing so would be useful depends on how similar or different the firms are. ML is both powerful and limited in this context: powerful because it allows you to predict future events but limited that it only does so within the confines of the historic, representative data. That is not to say that experience with an algorithm in one industry or firm does not help inform practices in another. From practical experience, we can attest to the fact that many similar algorithms can be applied to disparate problems, although they require model selection and parameter tuning as if it were a completely new problem.

Ethical issues. Accessing certain types of data for research purposes and practical application in organizations might be sensitive (e.g. email, using social network data for selection processes in non-standardized processes) (see Tonidandel et al., 2018). There is variation in legislation across countries with some of these being more stringent in terms of boundaries and protection of employee data (e.g. the General Data Protection Regulation in European Union countries) than in others.

Moreover, the application of ML to Big Data in leadership research can bring with it new questions that might not be covered by clear legislative guidance. For example, if we establish causal relationships between a specific type of leadership training to increase leadership effectiveness (e.g. by increasing charisma) and, by isolating heterogenous treatment effects, know which leaders or potential/aspiring leaders (e.g. with a specific personality profile) benefit from this type of training most or least, do we only offer this training to some leaders (those with a specific personality profile)? Or to everyone but communicate clearly that not everyone might benefit equally from the training? Another example is the examination of relationships between leadership behavior and employee stress-responses using physiological data: if we know that a certain type of leadership behavior is more likely to have negative effects on an employees' health and which employees are most likely to show a higher stress response, what is the responsibility of the organization to intervene and the leader to take action?

Furthermore, combining Big Data and research from genetics to predict, for example, leadership effectiveness (e.g., see DiPrete et al., 2018; Von Hinke et al., 2016) would certainly be powerful research, but raises ethical questions of whether we should, for example, select

leaders based on genetically encoded information.

Conclusion

The application of ML has opened up exciting new avenues for leadership research, especially applied to Big Data. Yet, while ML offers new tools to solve new problems, the fundamental question of causality remains as important as ever in the social sciences. Randomized controlled experiments are the gold standard to address causality. In this paper, we hope to have equipped researchers with the practical toolkits they need to study different questions – questions of either prediction (for which we refer them to our step-by-step ML guide) or causality (using the step-by-step experiment guide) or both (using the joint step-by-step ML-experiment guide). Taken together, these methods will move the field of leadership research into new directions, providing a more comprehensive and nuanced understanding of what, why and who will benefit from different kinds of leadership.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.leaqua.2020.101426.

References

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews, 29*(5–6), 594–621. https://doi.org/10.1080/07474938.2010.481556.

An, N., Xiao, Y., Yuan, J., Yang, J., & Alterovitz, G. (2019). Extracting causal relations from the literature with word vector mapping. *Computers in Biology and Medicine, 115*, Article 103524.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton: Princeton University Press.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly, 28*, 5–21. https://doi.org/10.1016/j.leaqua.2017.01.006.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1082–1120. https://doi.org/10.1016/j.leaqua.2010.10.010.

Antonakis, J., d'Adda, G., Weber, R. A., & Zehnder, C. (2014). *Just words, just speeches? On the economic value of charismatic leadership.* Working PaperUniversity of Zurich.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*, 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

Arnold, K. A. (2017). Transformational leadership and employee psychological well-being: A review and directions for future research. *Journal of Occupational Health Psychology, 22*(3), 381–393. https://doi.org/10.1037/ocp0000062.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*, 7353–7360. https://doi.org/10.1073/pnas.1510489113.

Athey, S., Imbens, G., Pham, T., & Wager, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review, 107*(5), 278–281. https://doi.org/10.1257/aer.p201 71042.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*. https://doi.org/10.1146/annurev-economics-080217-053433.

Athey, S., Imbens, G. W., & Wager, S. (2016). *Efficient inference of average treatment effects in high dimensions via approximate residual balancing.* Preprint arXiv:1604.07125Stanford: Stanford University.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*, 1148–1178. https://doi.org/10.1214/18-AOS1709https://CRAN.R-project.org/package=grfhttps://grf-labs.github.io/grf/.

Bacciu, D., Colombo, M., Morelli, D., & Plans, D. (2018). Randomized neural networks for preference learning with physiological data. *Neurocomputing, 298*, 9–20. https://doi.org/10.1016/j.neucom.2017.11.070.

Beygelzimer, A., & Langford, J. (2009, June). The offset tree for learning with partial labels. *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 129–138). .

Bhave, D. P. (2014). The invisible eye? Electronic performance monitoring and employee job performance. *Personnel Psychology, 67*, 605–635. https://doi.org/10.1111/peps.12046.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*, 199–231.

Brown, V. R., & Vaughn, E. D. (2011). The writing on the (Facebook) wall: The use of social networking sites in hiring decisions. *Journal of Business and Psychology, 26*, 219–225. https://doi.org/10.1007/s10869-011-9221-x.

Castillo, E. A., & Trinh, M. P. (2018). In search of missing time: A review of the study of time in leadership research. *The Leadership Quarterly, 29*, 165–178. https://doi.org/10.1016/j.leaqua.2017.12.001.

Cavazotte, F., Moreno, V., & Hickmann, M. (2012). Effects of leader intelligence, personality and emotional intelligence on transformational leadership and managerial performance. *The Leadership Quarterly, 23*, 443–455. https://doi.org/10.1016/j.leaqua.2011.10.003.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods, 20*, 3–31. https://doi.org/10.1177/2F1094428115617004.

Chisholm, M., & Tadepalli, P. (2002). Learning decision rules by randomized iterative local search. *Nineteenth international conference on machine learning* (pp. 75–82). Morgan Kaufmann.

Chockanathan, U., DSouza, A. M., Abidin, A. Z., Schifitto, G., & Wismüller, A. (2019). Automated diagnosis of HIV-associated neurocognitive disorders using large-scale Granger causality analysis of resting-state functional MRI. *Computers in Biology and Medicine, 106*, 24–30. https://doi.org/10.1016/j.compbiomed.2019.01.006.

Cochran, W. G., & Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General), 128*, 234–266. https://doi.org/10.2307/2344179.

Cox, D. R. (1958). *The planning of experiments.* New York: Wiley.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature, 577*, 671–675. https://doi.org/10.1038/s41586-019-1924-6.

Dawid, A. P. (2010). Beware of the DAG!. In: *Journal of Machine Learning Research Workshop Conf. Proc. 6*, 59–86In: . http://tinyurl.com/33va7tm.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review, 65*, 122–135. https://doi.org/10.1108/LR-06-2015-0061.

de Oliveira, J. M., Zylka, M. P., Gloor, P. A., & Joshi, T. (2019). Mirror, Mirror on the wall, who is leaving of them all: Predictions for employee turnover with gated recurrent neural networks. *Collaborative innovation networks* (pp. 43–59). Cham: Springer.

Diebold, F. X. (1998). *Elements of forecasting.* South-Western College Pub.

DiPrete, T. A., Burik, C. A., & Koellinger, P. D. (2018). Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences, 201707388*. https://doi.org/10.1073/pnas.1707388115.

Doldor, E., Wyatt, M., & Silvester, J. (2019). Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. *The Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2019.101308.

Donoho, D. (2015). 50 years of data science. *Tukey centennial workshop* (pp. 1–41). September. (NJ: Princeton).

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*, 745–766. https://doi.org/10.1080/10618600.2017.1384734.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36–48. https://doi.org/10.1080/00031305.1983.10483087.

Einav, L., Finkelstein, A., Mullainathan, S., & Obermeyer, Z. (2018). Predictive modeling of US health care spending in late life. *Science, 360*(6396), 1462–1465.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*, 293–314. https://doi.org/10.1093/nsr/nwt032.

Fiedler, F. E. (1967). *A theory of leadership effectiveness.* New York, NY: McGraw-Hill.

Fischer, T., Dietz, J., & Antonakis, J. (2017). Leadership process models: A review and synthesis. *Journal of Management, 43*(6), 1726–1753. https://doi.org/10.1177/0149206316682830.

Fisher, R. A. (1935). *The Design of Experiments.* First: Oliver and. Boyd.

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine, 30*, 2867–2880. https://doi.org/10.1002/sim.4322.

George, G., Osinga, C. E., Lavie, D., & Scott, B. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59*, 1493–1507. https://doi.org/10.5465/amj.2016.4005.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452. https://doi.org/10.1038/nature14541.

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide.* Princeton University Press.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society,* 424–438. https://doi.org/10.2307/1912791.

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly, 76*, 491–511. https://doi.org/10.1093/poq/nfs036.

Gruda, D., & Hasan, S. (2019). Feeling anxious? Perceiving anxiety in tweets using machine learning. *Computers in Human Behavior, 98*, 245–255. https://doi.org/10.1016/j.chb.2019.04.020.

Guzzo, R. A., Fink, A. A., Tonidandel, S., King, E., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology, 8*(4), 491–508. https://doi.org/10.1017/iop.2015.40.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning.* New York: Springer.

Hauser, O., Greene, M., DeCelles, K., Norton, M., & Gino, F. (2018). Minority report: A big data approach to organizational attempts at deterring unethical behavior. *Academy of Management Global Proceedings, 2018*, 125.

Hauser, O., & Luca, M. (2015). *How to design (and analyze) a business experiment.* Harvard

Business Review.
Hauser, O. P., Gino, F., & Norton, M. I. (2018). Budging beliefs, nudging behaviour. *Mind & Society, 17*(1–2), 15–26.
Hauser, O. P., Linos, E., & Rogers, T. (2017). Innovation with field experiments: Studying organizational behaviors in actual organizations. *Research in Organizational Behavior, 37*, 185–198. https://doi.org/10.1016/j.riob.2017.10.004.
Hausman, D. M. (2010). Probabilistic causality and causal generalizations. *The place of probability in science* (pp. 47–63). Dordrecht: Springer.
Henning, A., & van de Ven, K. (2017). "Counting your steps": The use of wearable technology to promote employees' health and wellbeing. *Performance Enhancement & Health, 5*, 123–124. https://doi.org/10.1016/j.peh.2017.11.002.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960. https://doi.org/10.1080/01621459.1986.10478354.
Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. In C. C. Clogg (Ed.). *Sociological methodology* (pp. 449–484). Washington, DC: American Sociological Association.
Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer, & S. Messick (Eds.). *Principals of modern psychological measurement* (pp. 3–35). Hillsdale, NJ: Erlbaum.
Holland, P. W., & Rubin, D. B. (1987). Causal inference in retrospective studies. *ETS Research Report Series, 1987*(1), 203–231. https://doi.org/10.1002/j.2330-8516.1987.tb00211.x.
Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly, 29*, 549–569. https://doi.org/10.1016/j.leaqua.2018.03.001.
Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics, 7*, 443–470. https://doi.org/10.1214/12-AOAS593.
Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.
Inceoglu, I., Thomas, G., Chu, C., Plans, D., & Gerbasi, A. (2018). Leadership behavior and employee well-being: An integrated review and a future research agenda. *Leadership Quarterly, 29*, 179–202. https://doi.org/10.1016/j.leaqua.2017.12.006.
Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research, 4*, 237–285.
Karch (2016). *A machine learning perspective on repeated measures: Gaussian process panel and person-specific EEG modeling.* Unpublished Doctoral DissertationHumboldt University Berlinhttps://doi.org/10.18452/17641.
Kempthorne, O. (1952). *The design and analysis of experiments.* Malabar, Florida: Robert Krieger Publishing Company.
Kenny, D. A. (1979). *Correlation and causality.* New York, NY: Wiley.
Kirimi, J. M., & Moturi, C. A. (2016). Application of data mining classification in employee performance prediction. *International Journal of Computer Applications, 146*, 28–35.
Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography, 3*, 262–267. https://doi.org/10.1177/2F2043820613513388.
Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences.* London: Sage.
Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1), 1–10. https://doi.org/10.1177/2F2053951716631130.
Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*, 491–495. https://doi.org/10.1257/aer.p20151023.
Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques.* MIT press.
Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning, 95*, 357–380. https://doi.org/10.1007/s10994-013-5415-y.
Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110.
Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated design of both the topology and sizing of analog electrical circuits using genetic programming. *Artificial Intelligence in Design '96* (pp. 151–170). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-0279-4_9.
Kozlowski, S. W. J., Chao, G. T., Chang, C.-H., & Fernandez, R. (2020). Team dynamics: Using "big data" to advance the science of team effectiveness. In S. Tonidandel, E. King, & J. Cortina (Eds.). *Big data at work: The data science revolution and organizational psychology*New York, NY: Routledge Academic (in press).
Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. February 6 META Group Research Notehttp://goo.gl/Bo3GS.
Lang, J. W. B., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence using longitudinal multilevel models. *Personnel Psychology, 71*, 255–281. https://doi.org/10.1111/peps.12260.
Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics, 49*, 186–205. https://doi.org/10.1016/j.jacceco.2009.11.004.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539.
Lee, A., Lyubovnikova, J., Tian, A. W., & Knight, C. (2020). Servant leadership: A meta-analytic examination of incremental contribution, moderation, and mediation. *Journal of Occupational and Organizational Psychology, 93*, 1–44. https://doi.org/10.1111/joop.12265.
Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management, 64*, 19–40. https://doi.org/10.1016/j.jom.2018.10.003.
Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences, 114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114.
Matz, S. C., Segalin, C., Stillwell, D., Müller, S. R., & Bos, M. W. (2019). Predicting the personal appeal of marketing images using computational methods. *Journal of Consumer Psychology, 29*(3), 370–390. https://doi.org/10.1002/jcpy.1092.
Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will change how we live, work and think.* London: John Murray.
McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review, 27*, 277–290. https://doi.org/10.1016/j.hrmr.2016.08.005.
Mesko, B. (2017). The role of artificial intelligence in precision medicine. *Expert Review of Precision Medicine and Drug Development, 2*(5), 239–241. https://doi.org/10.1080/23808993.2017.1380516.
Methot, J. R., Lepak, D., Shipp, A. J., & Boswell, W. R. (2017). Good citizen interrupted: Calibrating a temporal theory of citizenship behavior. *Academy of Management Review, 42*(1), 10–31. https://doi.org/10.5465/amr.2014.0415.
Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: A systematic literature review and research agenda. *Information Systems and e-Business Management, 16*, 547–578. https://doi.org/10.1007/s10257-017-0362-y.
Minbaeva, D. (2017). Human capital analytics: Why aren't we there? Introduction to the special issue. *Journal of Organizational Effectiveness, 4*(2), 110–118. https://doi.org/10.1108/JOEPP-04-2017-0035.
Minbaeva, D. B. (2018). Building credible human capital analytics for organizational competitive advantage. *Human Resource Management, 57*(3), 701–713. https://doi.org/10.1002/hrm.21848.
Na, K. S., & Kim, E. (2019). A machine learning-based predictive model of return to work after sick leave. *Journal of Occupational and Environmental Medicine, 61*, e191–e199. https://doi.org/10.1097/JOM.0000000000001567.
Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. *The New England Journal of Medicine, 375*, 1216. https://doi.org/10.1056/NEJMp1606181.
Oc, B. (2018). Contextual leadership: A systematic review of how contextual factors shape leadership and its outcomes. *The Leadership Quarterly, 29*, 218–235. https://doi.org/10.1016/j.leaqua.2017.12.004.
Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior, 7*(1), 505–533.
Oswald, F. L., & Putka, D. J. (2015). Statistical methods for big data. In S. Tonidandel, E. King, & J. Cortina (Eds.). *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge.
Pearl, J. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association, 95*, 428–431. https://doi.org/10.2307/2669380.
Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*, 54–60. https://doi.org/10.1145/3241036.
Pearl, J., & Mackenzie, D. (2018). AI can't reason why. *Computer, 11*, 30.
Pentland, S. (2012). The new science of building great teams. *Harvard Business Review, 90*, 60–69. Retrieved from https://hbr.org/2012/04/the-new-science-of-building-great-teams/ar/1.
Peters, L. H., Hartke, D. D., & Pohlmann, J. T. (1985). Fiedler's contingency theory of leadership: An application of the meta-analysis procedures of Schmidt and hunter. *Psychological Bulletin, 97*, 274–285. https://doi.org/10.1037/0033-2909.97.2.274.
Piccolo, R. F., Bono, J. E., Heinitz, K., Rowold, J., Duehr, E., & Judge, T. A. (2012). The relative impact of complementary leader behaviors: Which matter most? *The Leadership Quarterly, 23*, 567–581. https://doi.org/10.1016/j.leaqua.2011.12.008.
Ravid, D. M., Tomczak, D. L., White, J. C., & Behrend, T. S. (2020). EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. *Journal of Management, 46*, 100–126. https://doi.org/10.1177/2F0149206319869435.
Reddy, U. S., Thota, A. V., & Dharun, A. (2018). Machine learning techniques for stress prediction in working employees. *2018 IEEE international conference on computational intelligence and computing research (ICCIC)* (pp. 1–4). IEEE.
Reichard, R. J., Riggio, R. E., Guerin, D. W., Oliver, P. H., Gottfried, A. W., & Gottfried, A. E. (2011). A longitudinal analysis of relationships between adolescent personality and intelligence with adult leader emergence and transformational leadership. *The Leadership Quarterly, 22*, 471–481. https://doi.org/10.1016/j.leaqua.2011.04.005.
Roe, R. A. (2014). Time, performance and motivation. In A. J. Shipp, & Y. Fried (Eds.). *Time and work: How time impacts individuals* (pp. 63–110). New York: Psychology Press.
Rogers, T., & Frey, E. (2014). *Changing behavior beyond the here and now. HKS Working Paper No. RWP14-014. Available at SSRN*https://ssrn.com/abstract=2410465. orhttps://doi.org/10.2139/ssrn.2410465.
Rosenbaum, P. R. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*, 41–48. https://doi.org/10.1080/01621459.1984.10477060.
Rosenbaum, P. R. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General), 147*, 656–666. https://doi.org/10.2307/2981697.
Rosenbaum, P. R., & Rubin, D. B. (1983a). The central role of the propensity score in

observational studies for causal effects. *Biometrika, 70*, 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rosenbaum, P. R., & Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological), 45*, 212–218. https://doi.org/10.1111/j.2517-6161.1983.tb01242.x.

Rosenblum, M., & van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika, 98*, 845–860. https://doi.org/10.1093/biomet/asr055.

Roulin, N., & Bangerter, A. (2013). Social networking websites in personnel selection. *Journal of Personnel Psychology, 12*, 143–151. https://doi.org/10.1027/1866-5888/a000094.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66*, 688–701. https://psycnet.apa.org/doi/10.1037/h0037350.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229.

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science, 24*, 154–160. https://doi.org/10.1177/2F0963721414560811.

Shaffer, T. (2017). The 42 V's of big data and data science. KDnuggets. April https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html.

Shamir, B. (2011). Leadership takes time: Some implications of (not) taking time seriously in leadership research. *The Leadership Quarterly, 22*, 307–315. https://doi.org/10.1016/j.leaqua.2011.02.006.

Sigovitch, J. (2007). *Identifying informative biological markers in high-dimensional genomic data and clinical trials.* PhD thesisCambridge, MA: Harvard Univ.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/2F0956797611417632.

Sinar, E., Ray, R. L., & Canwell, A. L. (2018). HR leaders need stronger data skills. *Harvard Business Review,* 2–5. Retrieved from https://hbr.org/2018/10/hr-leaders-need-stronger-data-skills.

Spisak, B. R., van der Laken, P. A., & Doornenbal, B. M. (2019). Finding the right fuel for the analytical engine: Expanding the leader trait paradigm through machine learning? *The Leadership Quarterly, 30*, 417–426. https://doi.org/10.1016/j.leaqua.2019.05.005.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research, 10*, 141–158.

Sucar, L. E. (2015). Graphical causal models. *Probabilistic graphical models* (pp. 237–246). London: Springer.

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation. *Journal of Business & Economic Statistics, 34*, 661–672. https://doi.org/10.1080/07350015.2016.1172013.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association, 109*, 1517–1532. https://doi.org/10.1080/01621459.2014.951443.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*, 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods, 21*(3), 525–547. https://doi.org/10.1177/2F1094428116677299.

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data.* New York: Springer Science & Business Media.

Van Maanen, J., Sørensen, J. B., & Mitchell, T. R. (2007). The interplay between theory and method. *Academy of Management Review, 32*, 1145–1154. https://doi.org/10.5465/amr.2007.26586080.

Von Hinke, S., Smith, G. D., Lawlor, D. A., Propper, C., & Windmeijer, F. (2016). Genetic markers as instrumental variables. *Journal of Health Economics, 45*, 131–148. https://doi.org/10.1016/j.jhealeco.2015.10.007.

Waddell, K. (2016). The algorithms that tell bosses how employees are feeling. *The Atlantic,* 29.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*, 1228–1242. https://doi.org/10.1080/01621459.2017.1319839.

Wager, S., & Walther, G. (2015). *Adaptive concentration of regression trees, with application to random forests.* arXiv:1503.06388.

Weisberg, H. I., & Pontes, V. P. (2015). Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials, 12*, 357–364. https://doi.org/10.1177/2F1740774515588096.

Wenzel, R., & Van Quaquebeke, N. (2018). The double-edged sword of big data in organizational and management research: A review of opportunities and risks. *Organizational Research Methods, 21*, 548–591. https://doi.org/10.1177/1094428117718627.

Woo, S. E., O'Boyle, E. H., & Spector, P. E. (2017). Best practices in developing, conducting, and evaluating inductive research. *Human Resource Management Review, 27*, 255–264. https://doi.org/10.1016/j.hrmr.2016.08.004.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*, 1100–1122. https://doi.org/10.1177/2F1745691617693393.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*, 492–514. https://doi.org/10.1198/106186008X319331.

Zhu, J., Tatachari, S., & Chattopadhyay, P. (2017). Newcomer identification: Trends, antecedents, moderators, and consequences. *Academy of Management Journal, 60*, 855–879. https://doi.org/10.5465/amj.2015.0466.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology), 67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.