INSEAD
The Business School for the World®

Working Paper Series
2018/11/STR

Algorithmic Induction through Machine Learning: Opportunities for Management and Organization Research

Phanish Puranam
INSEAD, phanish.puranam@insead.edu

Yash Raj Shrestha
ETH Zurich, yshrestha@ethz.ch

Vivianna Fang He
ETH Zurich, fhe@ethz.ch

Georg von Krogh
ETH Zurich, gvkrogh@ethz.ch

This draft: March 14, 2018

Machine learning (ML) algorithms are rapidly advancing research across many fields of social science, including economics, marketing, and management information systems. Management and organization studies are yet to (fully) leverage these methods. We argue that ML algorithms can benefit both qualitative researchers engaged in a small number of cases and quantitative researchers faced with a large number of observations. Such benefits arise from the ability of MLtechniques to facilitate "algorithmic induction"—a form of inductive inference that yields identical (or highly similar) conclusions when applied by different observers to the same data. Algorithmic induction is valuable for researchers interested in theorizing through interpretative and comparative case analysis as well as generating hypotheses from large sets of quantitative data (followed by traditional testing approaches). We introduce variants of ML algorithms to management and organization researchers, develop the concept of algorithmic induction, and discuss its general potential for inductive theorizing in the field.

Keywords: Machine Learning; Algorithmic Induction; Theory Building

Algorithmic induction through machine learning: Opportunities for management and organization research

Phanish Puranam (INSEAD)

Yash Raj Shrestha, Vivianna Fang He, & Georg von Krogh (ETH Zurich)

This draft: March 14, 2018

ABSTRACT

Machine learning (ML) algorithms are rapidly advancing research across many fields of social science, including economics, marketing, and management information systems. Management and organization studies are yet to (fully) leverage these methods. We argue that ML algorithms can benefit both qualitative researchers engaged in a small number of cases and quantitative researchers faced with a large number of observations. Such benefits arise from the ability of ML techniques to facilitate "algorithmic induction"—a form of inductive inference that yields identical (or highly similar) conclusions when applied by different observers to the same data. Algorithmic induction is valuable for researchers interested in theorizing through interpretative and comparative case analysis as well as generating hypotheses from large sets of quantitative data (followed by traditional testing approaches). We introduce variants of ML algorithms to management and organization researchers, develop the concept of algorithmic induction, and discuss its general potential for inductive theorizing in the field.

INTRODUCTION

Inductive inferences occupy a prominent role in empirical management and organization research. Traditionally, theory induction appears to be reserved for researchers working with a "small N" rather than for those working with a "large N". In this paper, we argue that machine learning (ML) represents a useful new methodology to facilitate inductive inferences for management researchers, be they deeply engaged in a small number of cases or faced with a large number of observations. This versatile utility arises from the ability of ML techniques to facilitate pattern detection and prediction in a flexible and robust manner. This allows for algorithmic induction, which we define as a form of inductive inference that yields identical (or highly similar) conclusions when applied by different observers to the same data. The applications of algorithmic induction span the entire empirical research process, from data coding and data reduction to stylized fact generation. In its ideal form, algorithmic induction should be less prone to biases arising from the limits of human judgment and the danger of over-fitting (i.e., results that are highly idiosyncratic to the observed sample).

In the current stage of development, ML techniques do not represent a substitute for researchers' human judgement. What ML can do in a powerful manner is establishing robust patterns in data. However, other fundamental steps in the research process that precede and follow the establishment of patterns, such as conceptualization (defining the constructs of interest), measurement (selecting or developing measures for the constructs), and explanation (theorizing via inductive and abductive reasoning, Sheperd & Sutcliffe, 2011) of the observed patterns, remain largely human prerogatives. Thus, we believe that ML techniques complement the current inductive approaches used by organization and management scholars, although they do require us to alter how we think about theorization, explanation, and prediction.

ML does not equate to, or always require, "Big Data". While ML is a powerful way to analyze and make sense of large volumes of data, it can also be used to analyze datasets that are fairly small (with the obvious need to take extra precautions and be more skeptical about the conclusions). ML is fundamentally a tool for prediction, not statistical inferences (Mullainathan and Spiess, 2017). As a consequence, no discontinuity in sample size occurs beyond which the central limit theorem and its valuable properties become applicable. Instead, a primary concern in applying ML to small samples is over-fitting, which increases as the ratio of the number of observations to the number of variables decreases (Abu-Mostafa, Magdon-Ismail & Lin, 2012). This risk is particularly relevant to researchers who are interested in inductive work with relatively small samples. However, as long as the risks of over-fitting can be mitigated (by keeping the model complexity low and sampling error low) or at least explicitly acknowledged, the advantages of ML can apply to studies with smaller sample sizes.

Very few papers published in the leading organization and management journals have employed ML techniques thus far, yet our colleagues in adjacent disciplines have advanced more rapidly in applying these methods. In marketing, recent contributions draw on ML principles to improve model selection in order to enhance the model fit to the data while maintaining generalizability (e.g., Schwartz, Bradlow & Fader, 2014). Chen, Iyengar & Iyengar (2016), for example, apply one such method to investigate consumer heterogeneity. In management information systems, recent years have seen a surge in the application of ML techniques to understand complex information sets and to design systems for processing them (e.g., Zheng & Padmanabhan, 2006). In economics, researchers have introduced ML techniques in conjunction with instrumental variable analysis (which requires prediction accuracy in stage-one models) to improve causal inferences (e.g., Belloni, Chernozhukov & Hansen, 2014). ML techniques can

also be used as an alternative to propensity score matching (Varian, 2016). Another application in economics pertains to estimating heterogeneity in causal effects (Athey & Imbens, 2016). Mullainathan and Spiess (2017) provide an overview of ML applications in economics (see also Varian, 2014; Kleinberg, Ludwig, Mullainathan & Obermeyer, 2015).

In this paper, our intent is to contribute to the vigorous debate on the methods of theory building in the Academy of Management Review (e.g., Eisenhardt, 1989; Eisenhardt, 1991; Dyer & Wilkins, 1991; Shepherd & Sutcliffe, 2011; Lewis & Grimes, 1999). Thus, we neither review all the emerging literature in adjacent disciplines nor provide a comprehensive discussion of ML techniques per se. Instead, we aim to provide an accessible introduction to the core logic of ML techniques with a unique epistemic focus on employing algorithmic induction to support theory building, whether in large or small samples. To explain the technical aspects of ML, we draw on analogies to regression methods and psychological learning principles with which management researchers are already familiar. Our aim is motivated by the observation that no prior attempts (to our knowledge) have been made to develop a systematic approach to use ML as a tool for inductive theory building in a manner that spans small and large sample analysis.

Organization and management researchers, as with most social scientists, typically aspire to be sophisticated users rather than producers of statistical methodology. Identical to the adoption of techniques developed by statisticians, the adoption of ML techniques in organization and management research requires not only a solid conceptual understanding of what the algorithms do and what they assume but also familiarity and access to software that embeds these procedures. We see opportunities for improvement concerning both conditions in our field.

This paper is organized as follows. First, we offer a concise and accessible introduction to ML principles. While we do not recommend doing so, a reader who has no interest in the technical details and is focused only on possible applications can skip directly to the summary at the end of this first section. Second, we show how ML's core analytical property—robust pattern detection—allows management and organization researchers to use ML for algorithmic induction. Third, we conclude with some thoughts about how ML principles can be more widely diffused in our field, and we suggest an agenda for future research.

WHAT IS MACHINE LEARNING?

ML is a subdomain within the field of Artificial Intelligence (AI). ML provides computers with "the ability to learn without being explicitly programmed" (Samuel, 1959: 120). ML has found extensive applications in various domains. In natural science, ML as a tool has been applied across various fields, from astronomy (Sokol, 2017) to biology (Shipp et al., 2002). In everyday life, ML applications appear in many forms, such as email spam detectors, software games, personal assistants, search engines and automatic translation.

The following definition by Mitchell (1997) provides a helpful base on which to ground our discussion on the components of a learning problem: *"A computer program is said to learn* [effectively; author's note] *from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E"*. Suppose that task T is that we want to predict the profits of a firm in the current year and future years. We want to do this based on the data from past years (experience E). Experience may comprise data on past profits with a set of predictors such as price and sales changes, CEO characteristics, planned expenditures, demand trends, liquidity position, or any of

the factors that researchers believe are associated with firm profitability. The performance of the algorithm (P) is considered satisfactory if it can predict a firm's profits accurately. Hence, with increasing data from experience, the algorithm should behave in such a way that the difference between the actual and predicted profit is minimized. In the following text, we briefly introduce the three main ML paradigms that have been developed to address the learning problem.

Supervised Learning

In this setting, the dataset (E) contains explicit examples of what the correct output should be for a given (set of) inputs, i.e., both the set of predictors (X) and the outcome (Y). In the learning problem discussed above, the goal of a supervised learning algorithm is to make good predictions about future profits (future Ys) based on knowing only the future Xs. This is accomplished by fitting a model that does a good job of "predicting" past Ys based on past Xs and then assuming that the same relationship between Xs and Ys will hold in the future.

We could obtain a similar prediction using the familiar ordinary least squares (OLS) regression. Through techniques such as step-wise regression, we can also find the best fitting model (i.e., linear weighted combination of Xs) that predicts Y in the past data and then use this model to make predictions about future Ys based on future Xs. In one critical way, ML algorithms are exactly like these familiar models in that the model fitting occurs by minimizing a "loss function". A loss function penalizes the discrepancy between the predicted outcome and the actual outcome in past data. There are several types of possible loss functions (see Table 1 for a list of the most popular ones). OLS regression users will be aware that the loss function in OLS is the sum of squared error. ML algorithms use a broader range of loss functions. A loss function

can be minimized by rules (e.g., setting the derivative equal to zero and solving) or by searches (e.g., gradient descent, Newton–Raphson) when the former is not feasible.

However, ML algorithms also offer improvements over statistical models such as OLS in two major directions: sophistication in functional form and protection against over-fitting. First, compared with most regression models commonly used in management research (e.g., OLS, Logistic, Poisson), ML algorithms can model a much more complex relationship between the Xs and Ys in existing data. Higher-order polynomials and interactions can be incorporated without having to be specified in detail in advance. The resulting models can achieve higher levels of fit in the data, which hopefully leads to better predictions in the future. This sophisticated functional form may make the models harder to interpret (e.g., what do we make of a result that the fourth power of CEO charisma interacts with sales volatility to predict profit?). However, this is not necessarily a concern if the goal is prediction rather than explanation (i.e., hypothesis testing).

Second, ML algorithms use more advanced procedures than OLS (or other commonly used regression models) to guard against over-fitting. If we were to build a well-fitting OLS model by selectively adding or dropping variables, we would run two related risks of over-fitting: a) Excessive model complexity: The realized R-squared may be high simply because we have too many parameters in the model. As users of OLS regression techniques know, we could fit a model with a perfect R-squared if we added as many variables as cases to the regression, b) Excessive sample dependence: Including cherry-picked variables can produce a model that may fit the particular sample of data but may not be generalized beyond the data at hand.

More generally, statistical learning theory indicates that the complexity of a model selected to fit data has a U-shaped relationship with prediction error (see Figure 1). Put

differently, the prediction error initially decreases upon increasing the model complexity and then increases afterwards (Abu-Mostafa et al., 2012). The model is under-fitting in the region before the inflection point and over-fitting in the region after the inflection point. Under-fitting produces prediction errors that are systematically biased (because they represent a systematic deviation from true model), whereas over-fitting produces more variance (because the deviation is not systematic). The goal is to find the point where the model is sufficiently complex to accomplish the lowest prediction error. Excessive model complexity and excessive sample dependence both produce high prediction errors through over-fitting.

------INSERT FIGURE 1 ABOUT HERE------

OLS addresses the problem of excessive model complexity using adjusted R-squared, but it offers no standard solution to the second problem (excessive sample dependence). The second problem is also closely related to the issue of hypothesizing after knowing the results ("HARKing," Kerr, 1998). When mining a sample for "good" predictors (i.e., those that have large effect sizes and/or statistical significance), we will likely find at least a few if we search hard enough, but we face a significant risk that these associations do not generalize beyond the current sample. This is why pretending that the results reported in a paper were based on tests of prior hypotheses when they actually resulted from data mining is not only unethical in misrepresenting inductive work as deductive hypothesis testing; it can also seriously impede the progress of a field by accumulating non-replicable results that are too sample specific (Ioannidis, 2005; Gelman & Loken, 2014).

ML algorithms solve the first problem (i.e., fitting overly complex models) through a procedure known as **regularization** and the second problem (i.e., excessive sample dependence)

through procedures known as **cross-validation**. Regularization penalizes model fit for complexity. The intuition is similar to the use of adjusted R-squared in OLS, though a wider variety of constraints on complexity can be adopted.

For instance, the least absolute shrinkage and selection operator (LASSO), a popular ML algorithm, performs what is known as "L1 regularization". This adds a penalty proportional to the absolute sum of the standardized coefficients in a linear regression model. This is comparable to minimizing the sum of squares with the additional constraint that the absolute sum of the standard coefficients should be less than or equal to a constant (e.g., 1). This type of regularization can result in sparse models with few coefficients. Coefficients of some variables with small effects can become zero and be eliminated from the model.

Cross-validation, which is used to solve the problem of excessive sample dependence, is closely related to the idea of a hold-out sample. In this method, we split the available data on Xs and Ys into random sub-samples. Some of these sub-samples are used to fit the model (or "train" it), whereas others are used to evaluate or "test" the fitted model for its predictive accuracy. Models that fit the training datasets well while also achieving good predictive accuracy in the test sets can be found by repeating this procedure a large number of times.

With these concepts, researchers can comprehend a large class of supervised ML models in terms of functional form complexity, loss functions, regularization strategies and cross-validation techniques (see Table 1). Given the variety of functional forms and the models resulting from them, the task of **model selection** in supervised learning is also carried out by cross-validation. In traditional OLS regression, the user selects the best model by hand picking a set of models and adopting a criterion such as R-squared, adjusted R-squared, or Akaike's

information criterion. Supervised learning typically applies cross-validation to select from a set of automatically generated models based on different levels of regularization. One can also rely on an "ensemble" of models, averaging across many different types.

------INSERT TABLE 1 ABOUT HERE------

In essence, supervised learning involves learning from the past what correlates with a given outcome. It is similar to observing and learning from one's own or others' experiences that demonstrate relationships between actions and their outcomes (Bandura & Walters, 1963; Bandura, 1962). The key feature is that the learning involves studying patterns of Xs and Ys that have already occurred and finding the correlates of a particular "target" outcome. One can also think of this in terms of classical Pavlovian (or respondent) conditioning, which is well known to psychologists (Rescorla, 1967, 1988; Pearce, 1987).

Unsupervised Learning

Unsupervised learning algorithms, as the name suggests, operate in the absence of a "supervisor" variable. The data (E) lack any specific target outputs (i.e., Y) associated with each input. These algorithms are generally tasked with detecting patterns of correlations between groups of X variables, without any particular variable being selected as the dependent variable. Unsupervised learning also has psychological analogies to classical conditioning in that the agent is learning the relationships between events that have already occurred, but in this case, none of the variables necessarily has to be distinguished as a "target".

Clustering is a canonical example of unsupervised learning with which most management scholars are already familiar. Its purpose is to partition cases into sub-sets such that similar cases are in the same cluster and dissimilar cases are in different clusters. In our profit prediction

example, unsupervised learning can help users find a cluster of firms that are similar to one another on observed dimensions such as CEO attributes, demand trends and profitability. By observing the other attributes of firms that fall into the clusters with high and low profitability, we can infer the correlates of profitability. Once clusters are known, we can use partial information (some of the Xs) to guess what the remaining Xs will be by assuming that the clustering structure stays the same in the future. Alternatively, we may be satisfied with identifying sub-groups of firms that seem very similar (and therefore, we believe, will react similarly to future events). Strategic group analysis (e.g., Harrigan, 1985), is a well-established technique based on cluster analysis.

Unsupervised ML techniques for clustering use the same basic logic but provide much more flexibility in terms of choosing different types of algorithms. The algorithms include K-means, hierarchical and spectral clustering, all of which largely share a similar intuition. These algorithms iteratively partition the data into sub-sets that show high similarity of cases within and low similarity between sub-sets. Users may also cluster variables rather than cases. Principal component analysis (PCA) is another multivariate technique that is familiar to most management researchers. It involves reducing the number of variables by grouping together variables that contain essentially the same information. For instance, height and weight may contain much shared information that together they form a body index. ML algorithms embed PCA and other techniques (e.g., singular value decomposition) to obtain exactly these kinds of results. As with supervised learning, concerns around over-fitting and model selection can be addressed through regularization and cross-validation.

Reinforcement Learning

Reinforcement learning is a method where the algorithm determines the actions that will help it attain its goals by interacting with its environment (for an accessible introduction, c.f. Sutton & Barto, 1998). This learning involves producing actions and discovering errors and rewards, which assists the algorithm in automatically determining the ideal actions within a specific context that maximizes the total reward (Alpaydin, 2004).

The process of reinforcement learning is known under the labels of trial-and-error learning, experiential learning, operant or instrumental conditioning, and "win-stay-lose-shift" rules in the relevant literature in psychology, computer science, organization theory and evolutionary biology (Thorndike, 1911; March, 1991; Domjan 2010; Nowak and Sigmund, 1993; Sutton and Barto, 1998). All instantiate Thorndike's law of effect from psychology, which holds that *responses that produce a satisfying effect in a particular situation become more likely to occur again in that situation, and responses that produce a discomforting effect become less likely to occur again in that situation.* This implies that favorable feedback (i.e., bringing desired outcomes) on selecting an alternative tends to positively reinforce the belief about an alternative and thereby make it more likely to be selected. Whereas agents cannot influence the events being observed in supervised and unsupervised learning, they can do so in reinforcement learning. It is thus a form of *online* learning (Levinthal, 1997; Gavetti and Levinthal, 2000), as opposed to supervised and unsupervised learning, which are both *offline*.

Software installed in self-driving cars, autonomous robots and drones applies reinforcement learning to dynamically update and determine their trajectory based on input from the environment. Recently, such algorithms have also been successful in outperforming humans in interactive games such as chess and Go. One could, in theory, imagine a reinforcement learning algorithm that takes the place of a CEO in making various policy changes, observing the

resulting profits and adjusting actions accordingly. This is the spirit with which many modelers of organizations as adaptive systems have used reinforcement learning principles (c.f. Puranam, Stieglitz, Osman & Pillutla, 2015 for a review), though they have done so in a metaphorical rather than literal sense. There are some early indications of more literal approaches in which algorithms learn to make managerial decisions in, for example, foreign exchange trading (Austin, Bates, Dempster, Leemans & Williams, 2004) or picking ventures to invest in (Nevmyvaka, Feng & Kearns, 2006); however, we are not aware of any application to the problems of general management more broadly defined. Nonetheless, a suite of algorithms that are each specialized to a particular task (e.g., project selection, employee recruitment and retention), each of which exists at least in rudimentary form today, may well come together in the future.

Summary

It is crude but accurate to think of all ML algorithm families as producing **robust associations**, i.e., associations between variables that are unlikely to be the result of sample idiosyncrasy. The "robustness" in associations produced by ML algorithms results from procedures that allow complex models to fit the data (reducing bias in prediction) while also mitigating against over-fitting (reducing variance in predictions). The strength of these procedures lies in their ability to hunt for the best fitting functional forms with considerable flexibility while simultaneously guarding against over-fitting through the use of regularization, cross-validation and ensemble methods. However, three caveats must be borne in mind about these procedures:

1. They are **not** a substitute for randomization to obtain a causal inference. All ML methods are meant to be correlational. As is well known, it is not possible to go from correlation to causation without additional assumptions (Shadish, Cook & Campbell, 2002).

2. All assume that the future **can** be predicted from the past. (More technically, probability distributions for relevant variables remain stationary.)

3. In addition, **none** are primarily geared towards testing an explanation through inferences about the relationships between variables; instead, they focus on prediction ("y hat", not "beta", to use econometric terminology, as noted by Mullainathan & Spiess, 2017).

ALGORITHMIC INDUCTION THROUGH MACHINE LEARNING

At least in its idealized form, the scientific method begins with the observation of an empirical pattern, which then becomes the target of theorizing to be tested in additional data (e.g., Bernard 2012, Lave & March, 1993). The preliminary step of constructing the pattern to be explained is essentially inductive, as it involves the observation of a pattern in the data. We would ideally like this pattern to be a property of the population, as well. While nothing but a census can guarantee such generalization, random sampling can help make this more likely.

However, even within a randomly selected sample, sampling error will create random variation. This poses a key danger to inductive inference: over-fitting. We might mistakenly assume that a pattern (causal or not) we see in our sample will also occur in the population. Some philosophers of science have long argued that induction is not a logically defensible procedure because it involves the fallacy of assuming that samples are identical to the population (Popper, 1962). Nonetheless, it is routinely used by scientists in both a statistical and intuitive sense, and it proves a very fertile basis to obtain insights about a phenomenon and for theory generation. Thereafter, we may gain confidence in a theory if it escapes repeated attempts at falsification across a range of contexts and formulations of the test (i.e. the Duheme-Quine synthesis). If the central problem in deductive theory testing is spuriousness (i.e., omitted

variables that provide alternative explanations), the central problem in inductive theory building is over-fitting. However, it has not been given nearly as much attention as the problem of spuriousness, for which we have available a suite of statistical techniques (e.g., instrumental variables, matching, regression discontinuity designs, and ideally, of course, randomization). Note that expanding the sample size will solve the problem of over-fitting but not the spuriousness of associations, whereas expanding the number of variables measured can help resolve spuriousness but not over-fitting.

Algorithmic induction based on ML offers three potential advantages for inductive theorizing. First, being algorithmic in nature, it has high inter-subject reliability. An algorithm used by different individuals will still yield the same results. This is not necessarily an advantage if the goal is to enhance creative variation of interpretation, but it is an advantage if we want to enhance the reproducibility of an inductive inference. Second, the algorithms do not suffer from human "comprehension constraints". The functional forms we use in our research to test hypotheses are rarely the result of a theoretical commitment to their shapes on our part; more often, they simply represent what we can easily comprehend and interpret. For instance, we doubt that any management theorist would hold an entrenched view about the key relationships in their models being linear, though linear regression is our workhorse for theory testing. Again, relaxing this constraint is not necessarily an advantage if our goal is to build theory that is comprehensible to fellow humans. However, when we want to reliably code a large dataset in ways that closely approximate how humans would do it, complexity need not be shunned. Third, the algorithms offer protection against results that are highly idiosyncratic to what is observed (over-fitting). This is an advantage if we wish to build generalizable theory from our inductive

efforts. It is irrelevant if we do not, as is the stance of some anthropologists and ethnographers who may favor descriptions of specific cases over generalized explanation.

Next, we describe how a researcher may exploit these advantages across different kinds of inductive work. We illustrate how ML algorithms facilitate inductive theorizing in interpretative or comparative case methods and large-sample quantitative analysis. The discussion below is not exhaustive, and we believe the researcher's creativity is the only real constraint to finding new applications.

Machine Learning and Interpretative Case Analysis

The interpretivist tradition of case analysis in organization research is rooted in work in anthropology and sociology. The emphasis is on an immersive, ethnographic observation of the empirical context (e.g., Bechky, 2003; Hargadon & Bechky, 2006; Kellogg et al. 2006). The basic procedures involved in systematically making sense of experience is well captured in Glaser and Strauss's account of grounded theory (1967). "The basic idea of the grounded theory approach is to read (and re-read) a textual database (such as a corpus of field notes) and "discover" or label variables (categories, concepts and properties) and their interrelationships.

The ability to perceive variables and relationships is termed "theoretical sensitivity" and is affected by a number of factors including comprehension of the literature and sensitivity of the phenomena (Glaser and Strauss, 1967: 46). The main steps involve 1) open coding, which is concerned with identifying, naming, categorizing and describing phenomena found in a text, 2) the process of relating codes (categories and properties) to one another via a combination of inductive and deductive thinking and 3) selective coding and memos (Glaser & Strauss, 1967). Selective coding isolates those codes that are particularly salient for explaining the phenomena within emerging theory. Such coding makes for frugal and fast interpretations but may expose

theorizing to various biases in individual judgement and over-fitting. The latter is particularly challenging when selective coding is used to direct further data gathering.

While interpretative case analysis has the virtue of generating a deep understanding of phenomena and the causal mechanisms that may underlie them, it is prone to the dangers of over-fitting since the number of units being observed is typically equal to one (though there may be significant volumes of data capturing variations in sub-units within this unit). Some proponents of the interpretative method argue that they have no intention of generalizing beyond their sample, so over-fitting is irrelevant at the early stage of theorizing. For them, immersion into the field offers a wide variety of contextual data, allowing researchers to strive for contextual certainty; "One attains contextual certainty when there is a great deal of positive evidence supporting a conclusion and no contradictory evidence" (Locke 2007: 885).

Other researchers, however, seek the breadth of insight generated by grounded theory building. They emphasize the importance of relying on standardized procedures for converting experience to data, thereby moving the inductive process more towards the algorithmic (Gioia & Chittipeddi, 1991; Corley & Gioia, 2004; Gioia, Corley & Hamilton, 2013). A comprehensive data structure is paramount to such procedures, demonstrating how the researcher makes interpretative "moves" from the raw data, via first-order categories and second-order theoretical themes, to aggregate dimensions. These dimensions are next incorporated (or, as Glaser & Strauss would say, "integrated") into an overall explanatory theoretical framework using inductive and abductive reasoning. An advantage of this method that is relevant to the current argument is the high level of transparency in the interpretative process.

Functionality 1: Coding data. We realize that the purpose of grounded theory and related methods is neither primarily the accurate presentation of raw data nor the "routine

application of formulaic techniques" to make sense of those data (Suddaby, 2006). Instead, grounded theory aims to understand social phenomena using a methodology that is attentive to issues of interpretation and processes and that does not bind one too closely to long-standing assumptions (Suddaby, 2006). To this end, ML algorithms can be useful in the interpretivist tradition when the quantity of data may be large, even though the number of units being studied is small or even one. It is not unusual for a qualitative researcher to gather large volumes of text, audio, video and image data. Current approaches to coding these data rely heavily on human judgement. While human judgement is critical to the construction of de novo categories and the interpretation of phenomena from the participant-observer's viewpoint, the coding of (large volumes) of data along the lines of the schema generated by the researcher is more reliably done by algorithms than humans.

Put differently, by using a small amount of hand-coded data as the training set, ML algorithms can learn the patterns implicit in this coding and "predict" the coding for a much larger dataset. For such an application, the interpretability of the models is less important; predictive accuracy is key. For the purpose of expanding the human-generated coding to a larger sample, semi-automatic annotation/coding techniques can be used for coding the dataset, where a limited amount of data is coded by humans, which is then used as a "seed" by an algorithm to "code" the rest of the dataset (c.f. Medlock & Briscoe, 2007). The challenge of low inter-rater reliability is mitigated (it only needs to be established in the seed or training data), while the scale of data that one can analyze within a single unit of observation can be almost limitless.

A large set of natural language-processing algorithms are currently available that can assist researchers in data coding when dealing with large volumes of text data. Researchers can use automatic summary generation algorithms (c.f Dohare, Karnick & Gupta, 2017), which input

large volumes of text and generate a concise summary out of them. The summary can then be human coded or used to select documents to be human coded from a larger set. A set of sentiment analysis algorithms is currently available that facilitate the quantification of the sentiments embedded in texts. For example, in a recent study of leadership in online communities, Johnshon, Safadi & Faraj (2015) applied sentiment analysis to distinguish the language use patterns of leaders from those of non-leaders in these communities. Topic modeling approaches such as Latent Dirichlet Allocation (LDA) can be used to discover themes and trends in a collection of documents (Blei, 2012). A large number of studies on strategic alliances revolve around deep analysis of strategic alliance contracts (c.f. Reuer & Arino, 2007). In such a study, one could use LDA to automatically generate representative and meaningful topics from a large set of strategic alliance contracts.

The neighboring fields of marketing and economics have already adopted ML algorithms to manage large volumes of data in this manner. For example, in a recent study, Cui, Wong & Lui (2006) adopted artificial neural networks to gain insight into consumer behavior from large volumes of consumer datasets. The nature of these datasets varies from images, audio, and videos to human language. Due to their nature, these demand novel tools and techniques, which are available with ML. Puranam, Narayan & Kadiyali (2017) adopted flexible LDA models to investigate consumer opinions from online reviews of restaurants. Bajari, Nekipelov, Ryan & Yang (2015) applied ML methods based on regression trees to analyze and estimate consumer demand from scanner panel data related to sales from a grocery store. Kaminski, Jiang, Piller & Hopp (2017) used ML on video pitches to identify “lead user” entrepreneurs. Similarly, Zhang, Lee, Singh & Srinivasan (2016) used ML techniques to analyze the aesthetic quality of images on the Airbnb marketplace to investigate how the quality of images impacts room demand.

We illustrate how ML algorithms can be applied in an interpretative case for data coding. Ben-Menahem, von Krogh, Erden & Schneider (2016) gathered and analyzed qualitative data to inductively generate a model that stipulates interactions between formal and informal coordination in such projects. An open question from this study is how drug discovery projects originate in specific domain information (e.g., incorporating sources of detailed information on diseases or drug molecules) and whether and how such information potentially impacts their subsequent coordination. Patent analysis is a rapidly emerging application of ML that can shed light on this question (Trappey, Hsu, Trappey & Lin, 2006). More specifically, algorithmic induction could help researchers specify a grounded categorization and visualization of a firm's knowledge domains using an algorithm that learns to classify the firm's detailed patent information. Such categorization could form a basis for subsequent coding. The coded categories could then support within-case project sampling according to the novelty of the target proteins (disease) relative to the firm's knowledge base. A within-case sample constructed accordingly may reveal whether and how the novelty of the target-domain relation shapes project coordination in the firm. Without the assistance of ML algorithms, it would be impractical to couple the analysis of detailed patent information (a large pharmaceutical firm may hold in excess of 50,000 patents) with extensive qualitative data gathering within one and the same field study.

The use of algorithms for data coding simply expands the scale of data that can be coded and the reliability of coding in the interpretivist case analysis; the critical steps of gathering data, identifying patterns and building an abductive theoretical explanation of the pattern remain within the remit of the researcher. The inductive algorithms of ML are relegated to the role of a researcher's data assistant, but the researcher remains in control of the induction needed to

observe patterns and the abductive reasoning needed to explain them. However, as Gerring (1994) noted, the analysis of variance in data is a feature of both interpretative and comparative case methodologies. In the former, the variation studied is within a unit of observation; in the latter, it occurs across units of observation. The interpretivist researcher may thus also find some value in the functionality of the algorithms that we describe below in terms of their use for comparative case analysis.

Machine Learning and Comparative Case Analysis

In contrast to the interpretative case analysis that emphasizes the insights derived from a single unit of observation (e.g., a community, an organization), the comparative method focuses on insights derived from comparison across cases. The interpretative analysis is stronger in terms of deep understanding and meaning construction within a unit. In contrast, the use of multiple units of observation in the comparative method offers an advantage over the interpretative method in terms of avoiding over-fitting.

For instance, in the method described by Eisenhardt (1989), a single case is first used to tentatively establish a pattern. This is followed by a logic of replication, where extension to multiple cases is used to make the generated theory more likely to be "parsimonious, accurate and generalizable" (Eisenhardt 1989: 542). This method combines the logic (but not the statistical-algorithmic analysis) of quantitative induction in case control designs with a qualitative approach to within-case pattern discovery and categorization. Multiple cases (usually 4-10) are first selected such that the cases sharply differ on one (or a few) key dimension(s) (e.g., performance) while remaining similar on others. Interview (and other qualitative) data are collected from diverse informants within the cases. The task of the researcher is next to identify

elements that distinguish high- and low-performing cases, building on cross-case comparison (Brown & Eisenhardt, 1997, Martin & Eisenhardt, 2010, Vuori & Huy, 2016).

The comparative case analysis method draws its inspiration from Yin (2009) and Miles, Huberman & Saldana (1984). Compared with work in the interpretative tradition, comparative case analysis also tends to draw earlier and more extensively on prior theory (unfolding literature) in order to strengthen internal validity, sharpen construct definitions, and eventually increase theoretical sophistication and generalization.

As Eisenhardt (1989: 545) notes, induction is a fairly subjective process in comparative case analysis. Despite the best attempts of researchers to ensure inter-subjective reliability and replication across cases, they could fall prey to motivated reasoning (where they may be unconsciously motivated to selectively pay more attention to evidence confirming an emerging pattern and ignore disconfirming evidence). The induced pattern is also limited in complexity to the functional forms with which our minds can work. Further, the risk of over-fitting, while perhaps lower than in the interpretivist case analysis, is still significant. What may be true in a sample of 8-10 cases may not be true in the population, although steps may be taken relate within- and across case patterns to what is known about the overall population (e.g., industry; see Ozcan & Eisenhardt, 2009). These weaknesses are well known to proponents of the method. They engage in a series of procedures such as theoretical sampling (although used differently than in the grounded theory method discussed earlier), cross-case analysis conducted after data collection has been completed, conceptual replication, iteration between data and theory, and the establishment of subjective convergence across researchers to mitigate against these weaknesses as best as they can. In some sense, these challenges are the price paid for the depth of insight obtained from the qualitative induction within cases.

Another approach to comparative cases that is popular in organization research is the qualitative comparative analysis method of Charles Ragin (Ragin, 1987, 2000). In this method, a case is coded in terms of its (fuzzy) membership in different sets, instead of variables. For instance, a firm may belong (to some extent) to the set "Has high profits" and to the sets "Has charismatic CEO", "Has employee engagement program" and "Is in industry with low entry barriers". This set of theoretic representations then allows the application of algorithms such as DeMorgan's law in the crisp membership case and Quine–McCluskey reduction in the fuzzy set case (refer to Ragin, 2000 for details on algorithms) to identify the combinations of conditions associated with a given outcome (e.g., membership in the set "Has high profits"). In contrast to traditional regression methods, this approach produces combinations of factors that associate with the presence or absence an outcome rather than the marginal effect of each factor. One can see qualitative comparative analysis (QCA) results in terms of a regression-like equation that fits many-way interactions. Consequently, it can also produce results that are hard to interpret.

QCA typically works with many more cases than the Eisenhardt/Yin method. It is thus less prone to over-fitting (which necessarily decreases with sample size) but is also less geared towards the generation of qualitative insight within cases. Further, the induction itself is algorithmic, making it less susceptible to the twin problems of motivated reasoning and complexity constraints to human cognition. However, in absolute terms, the risk of over-fitting remains in the multi-case method in the same way that data mining, even with a large sample, is prone to over-fitting.

ML can contribute to strengthening both of the approaches to comparative case analysis discussed above. Exactly as with the interpretive case approach, ML can simply help with data coding at a large scale and in a reliable fashion. For example, one could use existing sentiment

analysis tools to quantify sentiments in written texts or speeches (Feldman, 2013) for an input to QCA. However, an additional functionality of ML is data simplification, which is a powerful complement to comparative case analysis with QCA.

Functionality 2: Simplifying data. A technical challenge for researchers using QCA is that the number of cases may fail to accommodate all possible combinations of variables or conditions, as the number of possible combinations increases exponentially with the number of conditions. Having too many conditions means that no case falls into a particular configuration or that a configuration cannot be assessed by empirical examples, causing limited diversity (Ragin, 2008; Schneider & Wagemann, 2012). We illustrate how ML algorithms can be used to manage such situations in a hypothetical comparative case study of new ventures. Assume the researchers collect rich data on a cohort of 20 new ventures over a course of five years. At the end of the observation period, 10 ventures went bankrupt, and the other 10 survived. Within the study sample, the researchers have identified 12 variables (i.e., lead founder's social capital, founding team size, etc.) that might influence the venture's survival outcome, and they would like to use QCA to further identify the configuration of conditions and make causal inferences. However, given the sample size of 20, they face the problem of limited diversity. Assuming that the researchers do not have sufficient theoretical grounding to guide their selection of variables for the subsequent QCA, ML algorithms can prove highly useful for variable reduction. The researchers could use ML algorithms to identify a smaller set of variables (those that are robustly associated with venture survival outcomes), which can then form the basis of QCA.

There are two forms of variable reduction with which ML algorithms can help researchers. First, there is the familiar summarization of multiple variables that contain redundant information into small set of factors through PCA. Second, variables can be discarded

based on their low association strength with outcomes of interest. Using the principles of regularization and cross-validation, the variables that reliably have limited effects on a target variable can be dropped from further consideration, without incurring the risk that these variables in fact mattered. This too can lead to over-fitting if not validated; the model that drops variables because they did not fit in the current sample (but do in fact matter in the population) is also an over-fitted model. Algorithms such as LASSO can deliver this form of data reduction while mitigating the risks of over-fitting.

Data simplification methods are very useful when the ratio of case to variable is small (c.f. Wasikowski & Chen, 2010, for comparison of variable reduction methods), which is a common situation in comparative case analysis. A preliminary reduction step can help reduce the number of variables to a point where comparison across cases becomes more interpretable. Other methods for small datasets include aggregation of regularized classifiers (Lu, Eng, Guan, Plataniotis & Venetsanopoulos, 2010), robust sparse representation (Haq, Tao, Sun & Yang, 2012) and discriminant analysis (Chen, Liao, Ko, Lin & Yu, 2000).

Functionality 3: Constructing stylized facts. Ultimately, the comparative case method aims to build "stylized facts"—patterns of associations between a few key variables that must then be explained through inductive and abductive theorizing. This is why comparative case researchers espouse the logic of replication, where the same patterns of co-occurrence of variables are documented across cases (Eisenhardt, 1989). Cross-validation procedures in ML offer a ready-made set of tools to help the researcher produce reliable associations that replicate across sub-samples of data. Indeed, the result of the cross-validation exercise may reveal few robust patterns—a finding that is valuable in and of itself. Mullainathan and Speiss (2017) demonstrated that it was possible to build comparably predictive models of house prices across

sub-samples of data, but the predictors used in each sample differed substantially. In our view, this points a) to the need to focus on models that work reasonably well across a range of data segments even though their predictive power may be lower and b) to the acceptance that sometimes, models that work across all data simply may not exist. If the latter is true, then the algorithms have saved the researcher from making an egregious error of over-fitting in their inductive theorizing, though the result in terms of publishability may not be as uplifting.

At this point, a natural question arises concerning the sample size required to be able to apply ML algorithms. In principle, it is possible to build robust induction of patterns through ML with relatively small sample sizes. In contrast to hypothesis testing, the key concern with induction through ML is over-fitting, not statistical inference. For instance, one of the most widely known datasets for teaching ML, known as "iris", contains only 150 observations of data with 5 columns variables. This is a dataset for three species of iris flowers and has been used extensively to test and validate diverse ML algorithms and models in more than 100 academic papers (Iris Data Set). In medicine, where generating cases is costly, ML techniques to work with small dataset has been developed (see, e.g., Shaikhina & Khovanova, 2017 where N=56).

One can imagine using algorithms sequentially in a "cascade of comprehensibility" in which a small dataset is initially analyzed using an ML algorithm with very high flexibility of functional form (e.g., support vector machine or tree induction). Such an algorithm will typically suffer from low interpretability. However, if the model has good predictive accuracy, it can be used to create a much larger predicted sample of data, which can then be analyzed using ML algorithms with much greater interpretability (e.g., LASSO, or step-wise logit). This is basically a way to amplify the signal-to-noise ratio in the small sample. A recent study by Jiang, Li &

Zhou (2009) is an instance of such an approach; they proposed clever modifications in existing ML algorithms for learning from samples as small as **24** cases (see also Zhou & Jiang, 2003).

In sum, algorithmic induction can be a powerful tool for researchers who perform comparative case analysis. It can accommodate a larger number of cases than is traditionally used in the Eisenhardt method, but well within the range of the sample size common to QCA, and larger samples. It helps to separate the inductive process needed to discover patterns in data (to determine which algorithm is helpful) from the abductive creative leaps needed to explain them (for which the researcher is still responsible). This separation itself may eliminate some of the biases arising from motivated reasoning to which researchers, as any humans, could potentially fall prey. To be clear, when ML is combined with these traditional comparative case analysis methods, it helps make the induction process underlying the construction of stylized patterns more robust because its algorithmic nature and measures help avoid over-fitting. It does not substitute for the inductive and abductive reasoning that researchers routinely perform when faced with the stylized patterns and the need for an interpretable theory (i.e., a proposed set of causal mechanisms) that accounts for these patterns.

Machine Learning and Quantitative Induction

It becomes clear that the first two functionalities of ML (i.e., data coding and data simplification) are as (if not even more) useful with large-sample quantitative analysis as they are in interpretive or comparative case methods. However, we focus here on the crucial importance of the third functionality (i.e., stylized fact generation) to quantitative researchers.

In management and organization research, explicit quantitative inductive inferences, i.e., the use of data primarily to describe a pattern rather than test a hypothesis, is rare. This may be partly the result of an incorrect (but, in our experience, widely held) premise that induction is

necessarily restricted to qualitative data (c.f. Shah & Corley, 2006; see also Locke, 2015). As also noted by Glaser (2008), who helped lay the foundations for qualitative induction along with Strauss (Glaser & Strauss, 1967), quantitative induction can serve as a powerful stimulus to theory building. Case control designs, which are popular in medical research, represent quantitative induction. In this method, a sample of cases that vary in their outcome of interest are statistically (i.e., algorithmically) analyzed to detect the correlates of the outcome in the data (Shadish, Cook & Campbell: 128).

At the same time, many scholars suspect that at least some quantitative studies that are reported as purely deductive may in fact have some disguised inductive elements. The "inductiveness" in these studies lies in the moments when the quantitative researchers look for patterns in the data before finalizing their hypothesis. The inductive elements are disguised because the patterns are then stated as if they are a priori expectations, which the data are then argued to confirm (Kerr, 1998). Setting aside the ethical issues this involves, from a purely statistical sense, the danger of over-fitting in such disguised inductive practice is eminent.

ML algorithms, as we have discussed, come equipped with two powerful techniques (i.e., regularization and cross-validation) to avoid over-fitting. These techniques can help us find patterns inductively that are less likely to be sample specific than if we did not use these techniques. Consider the following two approaches (shown in Table 2) that quantitative researchers may follow when they acquire a new and interesting dataset. We fear the description of Procedure A is less of a caricature of current ways of working than we hope it is.

------INSERT TABLE 2 ABOUT HERE------

Now consider Procedure B. The key point is that Sample I is reserved for inductivework, and Sample II is the deductive or "hold-out" sample. We use the inductive samples to search for

interesting patterns in the data (Locke, 2015)—this is honest and sophisticated data mining—and the hold-out sample to test the hypotheses so generated. The smaller sample used in Sample II may imply lower power unless the initial unpartitioned sample was large enough. However, recall that a) Samples IA, IB and II need not be of identical size, and b) low power has asymmetric effects—if no effect is detected, it may still exist in the population. Therefore, if an effect is detected, it is very likely present in the population.

To us, it seems, Procedure B dominates Procedure A in terms of transparency, and it is particularly useful when we do not start with strong theoretical priors. It has the virtue of allowing for induction and hypothesis in the same study (but in different sub-samples), and it delivers a result that we can feel confident is unlikely to be the result of over-fitting. As the reader would have noticed, almost everything we have said regarding the application of algorithms to the comparative case method stays true for quantitative induction. Indeed, quantitative induction is an instance of comparative case analysis, one that yields considerably fewer within-case findings but still follows the logic of replication. As with comparative case analysis, the abductive reasoning needed to formulate theories in quantitative induction remains the researcher's forte, whereas the inductive establishment of a stylized pattern can now be made algorithmic. The appropriate choice of algorithm can maintain comprehensibility while improving the protection against biases arising from subjectivity and over-fitting.

CONCLUSION: TRAINING PAVLOV'S DOG

ML algorithms such as supervised and unsupervised learning can be considered the descendants of Pavlov's dogs: they are "trained" to develop associations between variables (e.g., establish the co-presence of bell ringing and food) and then tested in their ability to predict the rest when presented with only some of the variables (e.g., will the bell ringing predict the presence of

food?). Reinforcement learning, in turn, is the direct descendant of Thorndike's cat, who learned through reinforcement (i.e., reward on success and punishment on failure) how to escape a cage.

However, these rudimentary learning mechanisms found (even) in our pets have enormous power and are at the heart of the current explosion of interest in ML. Robust pattern detection, beyond human comprehension constraints and free of judgement biases, is the key functionality on offer here. The adjacent fields of marketing, management information systems, and economics have already initiated important attempts to integrate this functionality into their traditional concerns such as model fitting, similarity detection, and causal inference. Inductive theorizing occupies an important place in management and organization research and, to us, seems to be a useful place to consider the value provided by ML algorithms. Once harnessed, algorithmic induction may constitute a domain in which organization researchers can contribute to the broader social science community.

We have outlined three broad applications of ML algorithms: data coding, data simplification and stylized fact generation. In interpretative case analysis, data coding—and to some extent data simplification—are the primary application. In comparative case and large-sample analysis, data simplification and stylized fact generation apply, in addition to data coding (see Table 3). While interpretability may not matter as much for data coding, it will be an important factor in data simplification and stylized fact generation. Some ML algorithms, such as LASSO, use a simpler functional form that does not violate the comprehension constraint.

------INSERT TABLE 3 ABOUT HERE------

In our view, the most promising application of ML is stylized fact generation. Robust (i.e., replicable) stylized facts form a basis on which new theoretical progress is founded, to which existing theories compete to offer useful explanations, and with which additional data is

used for testing theories. In management and organization research, there are very few such robust stylized facts, and these have been built across many individual studies (e.g., the variance decomposition studies in strategy). ML techniques can help researchers generate robust stylized patterns even within single studies. This insight may have long-term consequences for the way we craft theories in our field; in the future, we may come to think of a pronounced search for stylized facts as a sign of high-quality work. At the very least, ML algorithms are likely to lend increasing prominence to theorizing from single and multiple cases, as a manner of knowledge creation in our field.

As we have stressed at multiple points in the paper, ML algorithms can support but cannot replace human judgement in inductive research. Choosing what to measure (constructing categories), how to measure (developing a coding protocol) and what explanation to offer for the observed pattern among variables (theorizing via inductive or abductive reasoning) remain human prerogatives, at least at the current state of development in the field of AI. We can choose to use algorithms simply as robotic coders that never tire or make errors and can be taught by example rather than instruction (e.g., the application to interpretivist case work). Alternatively, we can use them to generate the raw material every theorist craves: robust stylized facts. With these robust patterns, we take joy in developing different, internally consistent explanations, which are then submitted to the challenge of critical hypothesis testing. We must also be careful to remember that no robust pattern may emerge in some cases and further that the result of this inductively derived exercise awaits hypothesis testing. Falsification of a hypothesis constructed most carefully with algorithm-assisted induction can (of course) occur; this indicates that despite our best efforts, we stand defeated by sampling error or undetected flaws in measuring the data used for induction. That is learning too, of course.

The adoption of algorithmic techniques for induction requires management and organization researchers to be willing to try them out with their own data. The barriers to such a trial, in our view, are quite low. Excellent introductions to the various algorithms used are readily available in online courses, often pitched at the same or even lower levels of technical difficulty as that of most graduate-level statistics or econometrics courses that our PhD students take. We can, for example, recommend ML courses available in Coursera (ML–Stanford University, Practical ML–Johns Hopkins University) and OCW MIT (Introduction to Neural Networks). Analysis packages available in R and Python come equipped with modules that make it fairly easy, given some basic programming skill, to apply. This skill, in our view, is critical for management and organization researchers to acquire irrespective of whether they will ever use algorithm-assisted induction.

Future work on algorithmic induction in management and organizations should focus on the match between various types of data (video, image, text, speech) gathered by the researcher and various ML techniques (supervised, unsupervised, reinforcement). Overall, future inductive theorizing would benefit from detailed reporting on the procedures undertaken to integrate ML algorithms into the various steps in the research process. For journal editors, this implies that papers utilizing ML methods should be submitted with a detailed Appendix specifying the algorithms used. Finally, an exciting avenue for future research in the field of management and organization would be to "replicate" prior hypothetico-deductive studies with publicly accessible data using algorithmic induction. Scholars may be able to advance theory rapidly by comparing, rejecting or compounding a corpus of established theory confirmed in prior work, with conclusions obtained from automated induction processes applied to the same dataset.

REFERENCES

Abu-Mostafa, Y., Magdon-Ismail, M., & Lin, H.-T. (2012). ***Learning from data*** (Vol. 4). New York, NY: AML Book.

Alpaydin, E. (2004). ***Introduction to machine learning.*** Cambridge, MA: The MIT Press.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. ***Proceedings of the National Academy of Sciences,*** 113(27): 7353–7360.

Austin, M. P., Bates, G., Dempster, M. A. H., Leemans, V., & Williams, S. N. (2004). Adaptive systems for foreign exchange trading. ***Quantitative Finance,*** 4(4): 37–45.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. ***American Economic Review,*** 105(5): 481–485.

Bandura, A. (1962). Social learning through imitation. In M. R. Jones (Ed.), ***Nebraska Symposium on Motivation:*** 211–274. Oxford, England: University Nebraska Press.

Bandura, A., & Walters, R. H. (1963). **Social learning and personality development.** New York, NY: Holt, Rinehart and Winston.

Bechky, B. A. (2003). Sharing meaning across occupational communities: The transformation of understanding on a production floor. ***Organization Science,*** 14(3): 312–330.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. ***The Review of Economic Studies,*** 81(2): 608–650.

Ben-Menahem, S. M., von Krogh, G., Erden, Z., & Schneider, A. (2016). Coordinating

knowledge creation in multidisciplinary teams: Evidence from early-stage drug discovery. ***Academy of Management Journal,*** 59(4): 1308–1338.

Bernard, H. R. (2013). ***Social research methods: Qualitative and quantitative approaches*** (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Bilmes, J. A. (1998). ***A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models.*** International Computer Science Institute, 4(510): 126.

Blei, D. M. (2012). Probabilistic topic models. ***Communications of the ACM,*** 55(4): 77–84.

Brown, S. L., & Eisenhardt, K. M. (1997). The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations. ***Administrative Science Quarterly,*** 42(1): 1–34.

Chen, L.-F., Liao, H.-Y. M., Ko, M.-T., Lin, J.-C., & Yu, G.-J. (2000). A new LDA-based face recognition system which can solve the small sample size problem. ***Pattern Recognition,*** 33(10): 1713–1726.

Chen, Y., Iyengar, R., & Iyengar, G. (2016). Modeling multimodal continuous heterogeneity in conjoint analysis - A sparse learning approach. ***Marketing Science,*** 36(1): 140–156.

Corley, K. G., & Gioia, D. A. (2004). Identity ambiguity and change in the wake of a corporate spin-off. ***Administrative Science Quarterly,*** 49(2): 173–208.

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. ***Management Science,*** 52(4): 597–612.

Dohare, S., Karnick, H., & Gupta, V. (2017). ***Text summarization using Abstract Meaning Representation.*** https://arxiv.org/abs/1706.01678, first accessed June 2017.

Domjan, M. (2010***). The principles of learning and behavior*** (6th ed.). Belmont, CA: Wadsworth Publishing.

Dyer, W. G., & Wilkins, A. L. (1991). Better stories, not better constructs, to generate better theory: A rejoinder to Eisenhardt***. Academy of Management Review,*** 16(3): 613–619.

Eisenhardt, K. M. (1989). Building theories from case study research. ***Academy of Management Review,*** 14(4): 532–550.

Eisenhardt, K. M. (1991). Better stories and better constructs: The case for rigor and comparative logic. ***Academy of Management Review,*** 16(3): 620–627.

Feldman, R. (2013). Techniques and applications for sentiment analysis. ***Communications of the ACM,*** 56(4): 82–89.

Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. ***Administrative Science Quarterly,*** 45(1): 113–137.

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis - a "garden of forking paths" - explains why many statistically significant comparisons don't hold up. ***American Scientist,*** 102(6): 460–465.

Gerring, J. (1994). ***Social science methodology: A criterial framework.*** Cambridge, England: Cambridge University Press.

Gioia, D. A., & Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change

initiation. ***Strategic Management Journal,*** 12(6): 433–448.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. ***Organizational Research Methods,*** 16(1): 15–31.

Glaser, B. G. (2008). ***Doing quantitative grounded theory.*** Mill Valley, CA: Sociology Press.

Glaser, B., & Strauss, A. L. (1967). ***The discovery of grounded theory.*** Chicago, IL: Aldine Transaction.

Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesus, O. (2014). ***Neural network design*** (2nd ed.). Boston, MA: PWS Publishing Company.

Haq, Q. S. ul, Tao, L., Sun, F., & Yang, S. (2012). A fast and robust sparse approach for hyperspectral data classification using a few labeled samples***. IEEE Transactions on Geoscience and Remote Sensing,*** 50(6): 2287–2302.

Hargadon, A. B., & Bechky, B. A. (2006). When collections of creatives become creative collectives: A field study of problem solving at work. ***Organization Science,*** 17(4): 484–500.

Harrigan, K. R. (1985). An application of clustering for strategic group analysis. ***Strategic Management Journal,*** 6(1): 55–73.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. ***Journal of the Royal Statistical Society,*** Series C (Applied Statistics), 28(1): 100–108.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal

problems. ***Technometrics,*** 12(1): 55–67.

Ioannidis, J. P. (2005). Why most published research findings are false. ***PLoS Medicine,*** 2(8): e124.

Iris Data Set. Available at https://archive.ics.uci.edu/ml/datasets/iris, first accessed June 2017.

Jiang, Y., Li, M., & Zhou, Z.-H. (2009). Mining extremely small data sets with application to software reuse. ***Software: Practice and Experience,*** 39(4): 423–440.

Johnson, S. L., Safadi, H., & Faraj, S. (2015). The emergence of online community leadership. ***Information Systems Research,*** 26(1): 165-187.

Kaminski, J., Jiang, Y., Piller, F., & Hopp, C. (2017). Do user entrepreneurs speak different?: Applying natural language processing to crowdfunding videos. In ***Proceedings of the 2017 CHI Conference*** Extended Abstracts on Human Factors in Computing Systems: 2683–2689. New York, NY: ACM.

Kellogg, K. C., Orlikowski, W. J., & Yates, J. A. (2006). Life in the trading zone: Structuring coordination across boundaries in postbureaucratic organizations. ***Organization Science,*** 17(1): 22–44.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. ***Personality and Social Psychology Review,*** 2(3): 196–217.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems***. American Economic Review,*** 105(5): 491–495.

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In

Advances in Neural Information Processing Systems, 4: 950–957. San Mateo, CA: Morgan Kaufmann Publishers.

Lave, C. A., & March, J. G. (1993). ***An introduction to models in the social sciences.*** Lanham, ML: Press of America.

Levinthal, D. A. (1997). Adaptation on rugged landscapes. ***Management Science,*** 43(7): 934–950.

Lewis, M. W., & Grimes, A. I. (1999). Metatriangulation: Building theory from multiple paradigms. ***Academy of Management Review,*** 24(4): 672–690.

Locke, E. A. (2007). The case for inductive theory building. ***Journal of Management,*** 33(6): 867–890.

Locke, K. (2015). Pragmatic reflections on a conversation about grounded theory in management and organization studies. ***Organizational Research Methods,*** 18(4): 612–619.

Lu, H., Eng, H.-L., Guan, C., Plataniotis, K. N., & Venetsanopoulos, A. N. (2010). Regularized common spatial pattern with aggregation for EEG classification in small-sample setting. ***IEEE Transactions on Bio-Medical Engineering,*** 57(12): 2936–2946.

March, J. G. (1991). Exploration and exploitation in organizational learning. ***Organization Science,*** 2(1): 71–87.

Martin, J. A., & Eisenhardt, K. M. (2010). Rewiring: Cross-business-unit collaborations in multibusiness organizations. ***Academy of Management Journal,*** 53(2): 265–301.

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in

scientific literature. In ***Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics:*** 992–999.

Menard, S. (2002). ***Applied logistic regression analysis*** (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Miles, M., Huberman, A., & Saldana, J. (1984). ***Qualitative data analysis: A methods sourcebook.*** Thousand Oaks, CA: SAGE Publications.

Mitchell, T. (1997). ***Machine learning.*** Burr Ridge, IL: McGraw Hill.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. ***Journal of Economic Perspectives,*** 31(2): 87–106.

Nevmyvaka, Y., Feng, Y., & Kearns, M. (2006). Reinforcement learning for optimized trade execution. In ***Proceedings of the 23rd International Conference on Machine Learning:*** 673–680. New York, NY: ACM.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. ***Nature,*** 364(6432): 56–58.

Ozcan, P., & Eisenhardt, K. M. (2009). Origin of alliance portfolios: Entrepreneurs, network strategies, and firm performance. ***Academy of Management Journal,*** 52(2): 246–279.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. ***Psychological Review,*** 94(1): 61–73.

Popper, K. R. (1962). On the sources of knowledge and of ignorance. ***Philosophy and Phenomenological Research,*** 23 (2):292-293.

Puranam, D., Narayan, V., & Kadiyali, V. (2017). The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors. ***Marketing Science,*** 36(5): 726–746.

Puranam, P., Stieglitz, N., Osman, M., & Pillutla, M. M. (2015). Modelling bounded rationality in organizations: Progress and prospects. ***The Academy of Management Annals,*** 9(1): 337–392.

Ragin, C. C. (1987). ***The comparative method: Moving beyond qualitative and quantitative strategies.*** Oakland, CA: University of California Press.

Ragin, C. C. (2000). ***Fuzzy-set social science.*** Chicago, IL: University of Chicago Press.x

Ragin, C. C. (2008). ***Redesigning social inquiry: Fuzzy sets and beyond***. Chicago, IL: University Press Chicago.

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. ***Psychological Review,*** 74: 71–80.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. ***American Psychologist,*** 43(3): 151.

Reuer, J. J., & Arino, A. (2007). Strategic alliance contracts: Dimensions and determinants of contractual complexity. ***Strategic Management Journal,*** 28(3): 313–330.

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

Schneider, C. Q., & Wagemann, C. (2012). ***Set-theoretic methods for the social sciences: A***

guide to qualitative comparative analysis. Cambridge, UK: Cambridge University Press.

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. ***Marketing Science,*** 33(2), 188–205.

Seber, G. A. F., & Lee, A. J. (2012). ***Linear regression analysis*** (Vol. 936). Chichester, UK: John Wiley & Sons Inc.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). ***Experimental and quasi-experimental designs for generalized causal inference.*** Belmont, CA: Wadsworth Publishing Company.

Shah, S. K., & Corley, K. G. (2006). Building better theory by bridging the quantitative–qualitative divide. ***Journal of Management Studies,*** 43(8): 1821–1835.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2017). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. ***Biomedical Signal Processing and Control.***

Shepherd, D. A., & Sutcliffe, K. M. (2011). Inductive top-down theorizing: A source of new theories of organization. ***Academy of Management Review,*** 36(2): 361–380.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. ***Nature Medicine,*** 8(1): 68–74.

Sokol, J. (2017). AI in action: Machines that make sense of the sky. ***Science,*** 357(6346): 26–26.

Specht, D. F. (1991). A general regression neural network***. IEEE Transactions on Neural Networks,*** 2(6): 568–576.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting***. Journal of Machine Learning Research,*** 15: 1929–1958.

Suddaby, R. (2006). From the editors: What grounded theory is not. ***Academy of Management Journal,*** 49(4): 633–642.

Sutton, R. S., & Barto, A. G. (1998). ***Reinforcement learning: An introduction*** (1st ed.). Cambridge, MA: MIT Press.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. ***Neural Processing Letters,*** 9(3): 293–300.

Thorndike, E. L. (1911). ***Animal intelligence: Experimental studies.*** New York, NY: The Macmillan Company.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. ***Journal of the Royal Statistical Society.*** Series B (Methodological), 58(1): 267 – 288.

Trappey, A. J., Hsu, F. C., Trappey, C. V., & Lin, C. I. (2006). Development of a patent document classification and search platform using a back-propagation network. ***Expert Systems with Applications,*** 31(4): 755–765.

Varian, H. R. (2014). Big Data: New tricks for econometrics. ***Journal of Economic Perspectives,*** 28(2): 3–28.

Varian, H. R. (2016). How to build an economic model in your spare time. ***The American Economist,*** 61(1): 81–90.

Vuori, T. O., & Huy, Q. N. (2016). Distributed attention and shared emotions in the innovation process. ***Administrative Science Quarterly,*** 61(1): 9–51.

Wasikowski, M., & Chen, X. (2010). Combating the small sample class imbalance problem using feature selection. ***IEEE Transactions on Knowledge and Data Engineering,*** 22(10): 1388–1400.

Yin, R. (2009). ***Case study research: Design and methods.*** Thousand Oaks, CA: Sage Publications Inc.

Zhang, S., Lee, D., Singh, P., & Srinivasan, K. (2016). How much is an image worth? An empirical analysis of property's image aesthetic quality on demand at AirBNB. In ***Proceedings of the International Conference on Information Systems:*** 1-20.

Zheng, Z., & Padmanabhan, B. (2006). Selectively acquiring customer information: A new data acquisition problem and an active learning-based solution. ***Management Science,*** 52(5): 697–712.

Zhou, Z.-H., & Jiang, Y. (2003). Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble. IEEE Transactions on Information Technology in Biomedicine, 7(1), 37–42.

TABLE 1Common Functional Forms, Loss Functions and Regularization Techniques in Machine Learning.

Type	Model Name	Functional Form	Loss Function	Regularization	Loss Minimization Technique
Regression	Linear regression (Seber & Lee, 2012)	$f(x;\beta)=\beta_0+\beta_1\cdot x_1+\cdots+\beta_p\cdot x_p$ ($x_1,\cdots,x_p$ are p independent variables, and $\beta_0,\cdots,\beta_p$ are model parameters.)	Squared loss $l(y,f(x;\beta))=(y-f(x;\beta))^2$ (y is the dependent variable/target)	-	First order optimality
	LASSO regression (Tibshirani, 1996)			L1-norm: $\lambda\sum_{j=0}^{p}\lvert\beta_j\rvert$ (λ controls the extent of regularization)	Sub-gradient descent, proximal gradient descent
	Ridge regression (Hoerl & Kennard, 1970)			L2-norm: $\lambda\sum_{j=0}^{p}\beta_j^2$ (λ controls the extent of regularization)	First order optimality
	Neural network (Specht,	$g(x)=f(\alpha+\sum_{h=1}^{q}w_h\cdot\phi(\tilde{\alpha}_h+\sum_{j=1}^{p}\tilde{w}_{jh}\cdot x_j))$	Squared loss	Weight decay (Krogh & Hertz, 1992), L1-norm,	Back-propagation (gradient descent)

	1991)	($x_1, \cdots, x_p$ are p independent variables, and w, α are model parameters.)	$l(y, g(x)) = (y - g(x))^2$ (y is the dependent variable/target)	dropout (Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov, 2014)	
Classification	Support Vector Machine (SVM) (Suykens & Vandewalle, 1999)	$y(\vec{w} \cdot \vec{x} - b) \geq 1$ (x is a vector of independent variables and y is the predicted class. w and b are model parameters)	Hinge loss $l(\vec{x}, \vec{w}, b, y') = max(0, 1 - (\vec{w} \cdot \vec{x} -$ (y' is the target class)	L1-norm, L2-norm	Sub-gradient descent
	Logistic regression (Menard, 2002)	$y = \ln(\frac{P(Y=1 \mid x)}{1 - P(Y=1 \mid x)}) = w_0 + \sum_{i=1}^{p} w_i \cdot x_i$ (x is a vector of independent variables of dimension p and y is the logit (log odds). $w_0, \cdots, w_p$ are model parameters)	Logistic loss $l(y, y') = log(1 + \exp(-y \cdot y'))$ (y' is the target class)	-	Gradient descent
	Neural network Hagan, Demuth,	$g_k(x) = f_0(\alpha_k + \sum_{h=1}^{q} w_{hk} \cdot \phi(\tilde{\alpha}_h + \sum_{j=1}^{p} \tilde{w}_{jh} \cdot x_j$ (1 hidden layer, $x_1, \cdots, x_p$ are p	Cross entropy loss $l(y, y') = -y \ln(y') - (1 - y) \ln(1 -$	Weight decay, L1-norm, dropout	Back-propagation (gradient descent)

	Beale & De Jesus, 2014)	independent variables, and w, α are model parameters.) $C(x) = \arg\max_j g_j(x)$ (C(x) is the predicted class)	($y = C(x)$ is the predicted class and y' is the target class)		
Clustering	Gaussian mixture model (Bilmes, 1998)	$P(x \mid \mu_k, \Sigma_k) = N(x \mid \mu_k, \Sigma_k)$ $P(x) = \sum_{k=1}^{K} \Phi_k P(x \mid \mu_k, \Sigma_k), \sum_{k=1}^{K} \Phi_k = 1$ ($N(x \mid \mu_k, \Sigma_k)$ is the multivariate normal distribution, x is a vector of independent variables and μ, Σ, Φ are model parameters)	Negative log-likelihood $l(x, \mu, \Sigma, \Phi) = -\log(\sum_{k=1}^{K} \Phi_k \cdot P(x$	-	EM algorithm
	K-means (Hartigan & Wong, 1979)	$C(x) = \arg\min_k \|x - \mu_k\|_2^2$ (x is the vector of independent variables, $\mu_0, \cdots, \mu_{K-1}$ are model parameters, C(x) is the predicted class)	L2-distance $l(x, u) = \|x - \mu_{C(x)}\|_2^2$	-	EM algorithm

TABLE 2 Two Approaches to Large Sample Analysis

Procedure A	Procedure B
1. Examine the variables and consider how they link to prior theory as well as what relationships one should expect between them. 2. Test for these relationships. 3. If no "statistically significant" patterns are found, revise the thinking about which variables (and their transformations) and relationships to focus on. 4. Iterate between steps 2 & 3 until some interesting (i.e., publishable) results are found. 5. Write up the results in the format of the results of a hypothesis testing exercise, in which the hypothesis (or at least its specific wording) is consistent with the results.	1. Examine the variables and consider what relationships one should expect between them. 2. Partition the data randomly into two sub-samples, Sample I, and Sample II. Put Sample II away in the "vault" and focus for now on Sample I. 3. Split Sample I further into Samples IA and IB. 4. Use ML algorithms (including cross-validation, within Sample IA) to generate a robust pattern of associations that seems theoretically interesting. However, choose models carefully since interpretability remains important (e.g., LASSO rather than neural nets). 5. Derive additional implications of the theories we construct that explain the observed patterns in Sample IA. 6. Test for the replication of the observed pattern as well as the additional theoretical implications derived from Sample IA in Sample IB.

	7. Iterate between steps 3 and 6 until the results in Sample IB confirm the observed patterns and predictions derived from Sample IA. 8. Test for the replication of the observed pattern as well as the additional theoretical implications in Sample II. Report the results of Sample II alone as the result of the hypothesis tests and the analysis of Sample I as a purely inductive process.

TABLE 3. Machine Learning Functionality and Application across Research Approaches

	Data coding	Data simplification	Stylized fact construction
Interpretive case analysis	Yes	Yes	--
Comparative case analysis	Yes	Yes	Yes
Large sample analysis	Yes	Yes	Yes

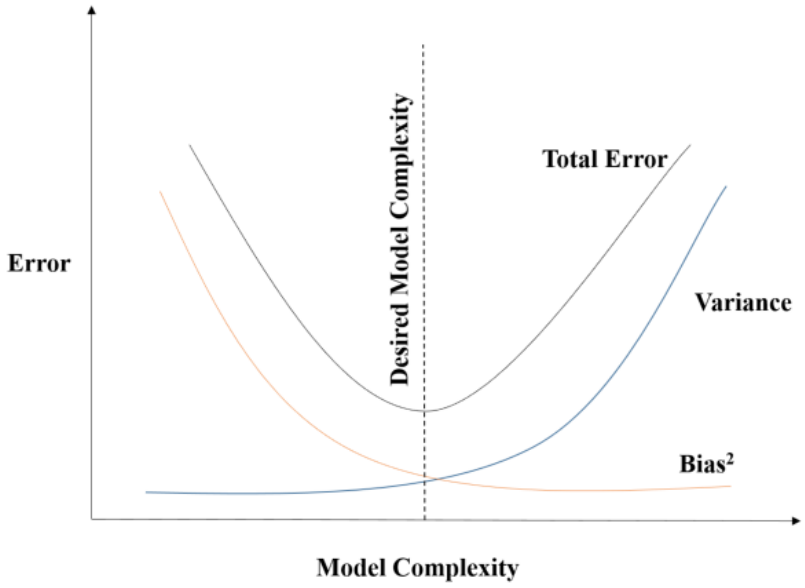

FIGURE 1 Relationship between Model Complexity and Prediction Error