Making the Most of AI and Machine Learning in Organizations and Strategy Research: Supervised Machine Learning, Causal Inference, and Matching Models

Jason Rathje
jason.rathje@gmail.com
Department of Defense

Riitta Katila
rkatila@stanford.edu (corresponding author)
Stanford University

Philipp Reineke
preineke@stanford.edu
Stanford University

Strategic Management Journal, in press

Making the Most of AI and Machine Learning in Organizations and Strategy Research: Supervised Machine Learning, Causal Inference, and Matching Models

ABSTRACT

Research summary. We spotlight the use of machine learning in two-stage matching models to deal with sample selection bias. Recent advances in machine learning have unlocked new empirical possibilities for inductive theorizing. In contrast, the opportunities to use machine learning in regression studies involving large-scale data with many covariates and a causal claim are still less well understood. Our core contribution is to guide researchers in the use of machine learning approaches to choosing matching variables for enhanced causal inference in propensity score matching (PSM) models. We use an analysis of real-world technology invention data of public-private relationships to demonstrate the method and find that machine learning can provide an alternative approach to ad-hoc matching. However, as with any method, it is also important to understand its limitations.

Managerial summary. This paper explores the use of machine learning to enhance decision-making, particularly in addressing sample selection bias in large-scale datasets. The rapid development of AI and machine learning offers new, powerful tools especially for digital ecosystems where complex data and causal relationships are complex to analyze. We offer managers and stakeholders insight into the effective integration of machine learning for selecting critical variables in propensity score matching models. Through a detailed examination of real-world data on technology inventions within public-private relationships, we demonstrate the effectiveness of machine learning as a robust alternative to traditional matching methods.

Keywords:
AI and machine learning, Propensity-score matching, Technological Change and Types of Innovation, Patents and R&D, Government Regulation and Public-Private Relationship

1| INTRODUCTION

Machine learning (ML) offers new empirical tools for strategy research, but how do these methods fit with what we know and need in regression studies? In particular, (how) do predictions of an outcome variable *y* from *x* that are at the core of supervised machine learning fit in with regression methods used by strategy scholars? Our contribution is to highlight the use of supervised machine learning in two-stage matching models to deal with sample selection bias (Heckman & Todd, 2009). In particular, we demonstrate the use of ML to determine variables on which to match in the first-stage model.[1]

Discussion on how to use machine learning in the matching process is increasingly relevant for several reasons. Causal inference using two-stage matching methods—in which the first stage is effectively a prediction and thus particularly fit for supervised ML methods—is now used frequently in strategy and organizations research. As the bar for causal inference has increased in our field (Bettis, Gambardella, Helfat, & Mitchell, 2014), this trend is only likely to intensify. Yet, despite the popularity of such matching techniques, there is a need for scholars to more closely justify the first stage selection of variables, and ML methods can play an important role.

Second, increasing availability of "big data" through online platforms (e.g., job postings or employee reviews; Reineke, Katila, & Eisenhardt, 2024) and open-source data collection efforts (e.g., crowdsourced repositories such as Wikipedia and Mobygames; Katila, Piezunka, Reineke, & Eisenhardt, 2022) increase opportunities to create many new, often interdependent variables. Again, the ability of ML to help evaluate a comprehensive set of variables and their interactions in first-stage matching, and then prune, is likely to be increasingly valuable.

More specifically, in matching models, machine learning can play a role in the pretreatment

[1] Other studies such as Choudhury, Allen, and Endres (2021) focus on supervised machine learning and Choudhury, Wang, Carlson, and Khanna (2019) on unsupervised and supervised machine learning.

covariate selection process. It can be used in place of or to augment, *ad-hoc* selection of first-stage variables, including pretreatment confounding covariates. By improving the first stage, second-stage estimates of causal effects can potentially be made more precise and more reproducible (Angrist & Frandsen, 2022). Machine learning can thus improve robustness of second-stage parameter estimates by acting as a data pre-processing method (Goller, Lechner, Moczall, & Wolff, 2020; Rathje & Katila, 2021).

Adding ML is also potentially beneficial because with complex data, researcher reasoning used to pick and prioritize pretreatment confounders is likely to reach its limits. Machine learning can help by pointing to which confounders might matter more than others from the list originally generated by the researcher. Further, by examining highly correlated covariates—of which ML typically picks a subset—the method pushes researchers to think deeper about their own underlying assumptions of which confounders matter.

Another possible benefit is a more thorough covariate selection process. Manually including all pairwise interactions and complex functional forms is often infeasible and computationally not possible, so researchers pick a subset. In contrast, supervised machine learning sorts through these interactions and functional forms automatically, and "allows us to let the data explicitly pick effective specifications" (Mullainathan & Spiess, 2017, p. 101), potentially resulting in a more comprehensive approach to estimate the first stage. A related example is that in a regression context with many possible controls, machine learning can be used to "automatically" prune the control or variable set (Belloni, Chernozhukov, & Hansen, 2014; Vanneste & Gulati, 2022; Kumar, Liu, & Zaheer, 2023) and possibly even defend against data mining.

While all machine learning methods are associative, none of our other typical methods in

strategy (ML or not) can entirely deal with the challenge of unobservables like a well-designed randomized controlled trial could. The unique aspect of ML matching here is that the techniques can include a fair number of observables, and, thus potentially reduce unobserved heterogeneity because so many more observables and their functional forms can be included and tested. Altogether, given its automated nature, machine learning matching is likely a useful addition to strategy scholar's toolbox.

Our review of the strategy literature suggests that recent advances in AI and machine learning methods have opened up new empirical possibilities for comparing human strategies against algorithmic ones (Bahceci, Katila, & Miikkulainen, 2015), for creating comprehensive strategies for search and innovation discovery (Bahceci, Katila, & Miikkulainen, 2023), and for inductive theorizing, such as variable creation based on swaths of unstructured text data (Helfat et al., 2023; Miric, Jia, & Huang, 2022). At the same time, other sister disciplines including economics (e.g., Angrist & Frandsen, 2022; Goller et al., 2020), medicine, and biosciences (D'Agostino, 1998; McCaffrey, Ridgeway, & Morral, 2004; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Cannas & Arpino, 2019; Lee, Lessler, & Stuart, 2010; Blakely, Lynch, Simons, Bentley, & Rose, 2020) have a recent but already robust stream of research applying machine learning to matching designs. Although studies in other disciplines are informative, it is also important to understand which approaches may, or may not, apply to strategy problems.

In this paper, we first review prior studies to identify the "sweet spot" for machine learning for matching studies in strategy—i.e., large-scale data with many interrelated covariates involving a causal claim. In particular, we use the propensity score method to illustrate one application that involves matching in strategy, as Bettis and Blettner (2020) suggest.

To be as close as possible to a real situation that empirical strategy researchers might face, we then use a rich technology invention dataset to show the technique's application to research in technology strategy that evaluates the impact of public funding. It is important to note that we have chosen these data simply as an example to *demonstrate* the ML-matching (ML-PSM) technique. Although we report the second stage results—and these results are consistent across the matching methods, which increases our confidence in the method as an additional tool that researchers can use for robustness testing—our focus is on demonstration of the ML-PSM technique. Our hope is that by looking "under the hood" of a real-world strategy application, we can encourage more researchers to consider the potential of the ML method to deal with sample selection bias.

As with any method, it is important to note its boundaries. First, supervised machine learning methods do not provide standard errors, so no coefficient estimates can be readily interpreted (Mullainathan & Spiess, 2017, p. 96).[2] A second boundary condition of machine-learning methods is that we cannot relinquish our responsibility as researchers to a "machine". Human interpretation still plays a crucial role: For example, the raw material from which the process prunes—such as the set of confounding variables that are originally included—needs to be selected by the researcher, pointing to a boundary condition for the method. Another boundary condition is that the confounders automatically selected for prediction in the first stage do not always match researcher reasoning because code picks "randomly" of highly multicollinear variables (although the researcher can require a particular variable to be included). So, researcher expertise is still very much needed if coefficients are interpreted. Fourth, machine learning approaches carry the risk of introducing more uncertainty than reducing it. Although one can be more precise about what to optimize for, as we describe in more detail below, how an algorithm

[2] Even when machine learning algorithms produce regression coefficients, the estimates are rarely consistent. Further, the tools do not yet provide consistent standard errors (Mullainathan & Spiess, 2017, p. 96).

selects variables can remain somewhat of a black box which could produce unexpected results. Such results naturally require more human interpretation of the type that a researcher might be trying to avoid in the first place. Finally, supervised ML shines when the regression datasets are large with many, complex confounders. In small datasets with few confounders, an analytic (e.g. "vanilla" logistic) regression is likely to be an equally effective alternative.

2| PREDICTION TASKS IN STRATEGY: PRIOR WORK

In many classic strategy problems, prediction is an essential part. There are two broad categories of research that are particularly relevant. The first category is inductive, discovery tasks using complex data that includes a prediction task. Here, machine learning can help researchers discover data patterns. One characteristic of these data is that an "almost unlimited number of variables can be created" (Einav & Levin, 2014) and analyzed, making machine learning methods attractive. Examples include predictions of consumer behavior, or predictions from more recently available data, such as whether pixels in a satellite picture represents a geographical area with economic activity (Mullainathan & Spiess, 2017). These data come both directly from companies and from publicly available open-source data (Nagaraj, 2022). Strategy scholars more and more frequently use these types of complex data that often generates variables for follow-on analysis.

The second category—and our focus in this paper—is causal inference using matching models which inherently includes a prediction task in the first-stage regression. Common examples in strategy scholarship include estimating a causal effect of a treatment which often includes a first stage pre-processing step of flexibly controlling for many confounders to deal with sample selection bias, i.e., the first stage is effectively a prediction. In these models, the goal is to evaluate consequences of decisions such as which growth strategy (acquisition vs organic), policy (antitrust vs not), or technology (renewable vs not) to pick. The challenge is that "treated" observations often

can differ in many covariates from those in the counterfactual control group (e.g., in age, industry, experience, location, etc.).[3] In such cases, matching becomes an important component of the policy evaluation (Imbens, 2004). Without the first-stage prediction, evaluation of strategic decisions often suffers from selection bias, potentially causing causal identification to fail and rendering the results from the study invalid. Given the complexity of these real-world strategy decisions, covariates that underlie the first-stage prediction can be not only large in number, but also interdependent with each other, complicating the ability of strategy researchers to make the evaluation, and suggesting benefits to considering machine learning approaches. Consistent with this view, Leiblein, Reuer, & Zenger (2018) note that in strategic decisions, "decision interdependencies, including higher order interactions… call [traditional] regression approaches into question." So while supervised machine learning is not designed to directly assist parameter estimates in the second-stage regression, it can have a central role when the first stage involves a prediction. We focus on demonstrating such matching designs.

A practical problem that strategy researchers encounter is deciding *which* variables to include in the first stage. "…intuition will suggest a set of variables that might be important to control for but will not identify exactly which variables are important or the functional form with which variables should enter the model" (Belloni, Chernozhukov, Fernández-Val, & Hansen, 2017). Given the complexity, Bettis and Blettner (2020, p. 82) observe that "The nature of complexity … strongly suggests the expanded development and use of … appropriate machine learning algorithms." The causal inference method that we focus on is propensity score matching

[3] Some statisticians have called for the extreme measure to limit research to only those empirical contexts which allow for randomized experiments (Grushka-Cockayne, Jose, & Lictendhal, 2016), but we are more hopeful.

(PSM).[4] Below, we first discuss how propensity scores are used in current work in strategy, what could be added, and what role supervised ML could play in these causal inference tasks.

Example of prediction: Propensity score matching in two-stage designs

We chose propensity score matching (PSM) as an example to demonstrate machine learning use for several reasons. First, propensity score matching is used when data are relatively complex (large dataset, many predictors), and it involves a first-stage prediction step[5], both consistent with settings in which supervised machine learning methods are useful. Second, PSM has become increasingly common in strategy. Review of strategy research (Strategic Management Journal articles published since 2010) points to an increase in the use of the method to address endogeneity concerns.

Prior propensity score matching papers in strategy research. Because much research in strategy centers around the impacts of "treatments" or "policies," propensity scores is a common method. We catalogued over 50 SMJ papers using the propensity score method in the past few years (details available from the authors). Our focus was to identify (1) the kinds of strategy problems where propensity scores have been used to enhance causal inference, (2) the variables and procedures used in the first-stage analyses to calculate the propensity score for the estimation, and (3) the trends in research using propensity scores over time.

Strategy research we reviewed used propensity scores to examine growth strategy interventions such as alliances (Asgari, Singh, & Mitchell, 2017), joint ventures (Chang, Chung, & Moon, 2013), corporate diversification (Rawley, 2010; Chang, Kogut, & Yang, 2016),

[4] Propensity score matching is often used for large-scale samples, and is increasingly common in strategy research. A closely related variant is inverse probability of treatment weights (IPTWs) which many strategy studies also use to strengthen causal inference and for which supervised ML can also be applied (c.f. Katila, Thatchenkery, Christensen, & Zenios, 2017; Blakely et al., 2020).

[5] Propensity score matching uses prediction to match observations based on their propensity to be treated (Rosenbaum & Rubin, 1983).

competitive positioning (Thatchenkery & Katila, 2023) and refocusing moves (de Figueiredo, Feldman, & Rawley, 2019), using propensity scores to identify comparable control groups that were not exposed to the intervention. Other research that similarly employed propensity scores to enhance causal inference examined the performance impact of affiliations with high-status actors (Schuler, Shi, Hoskisson, & Chen, 2017), particular types of CEOs and top executives (Cummings & Knott, 2018; Patel & Cooper, 2014; Mata & Alves, 2018) or investors (Hasan, Kobeissi, & Wang, 2011; DesJardine & Durand, 2020; Oehmichen, Firk, Wolff, & Maybuechen, 2021). More recent work examined the implications of business strategy decisions such as adoption of sustainable business practices (Ortiz-de-Mandojana & Bansal, 2016; Durand, Paugam, & Stolowy, 2019). Overall, PSM was used to maintain a *comparable* control group that was not treated.

Only a handful of the papers we reviewed (~3-4%) however tested for different functional forms of the variables in the first stage, and just about 30% of the papers were sensitive to explaining why particular variables were included in the first stage. In addition, typical justifications were rather general such as "based on prior work". Therefore, adding machine learning in matching, together with researcher's reasoning and justification, can provide a major upside to causal inference in future studies. With machine learning's added help, researchers could discover complex structures that were not specified in advance, such as high-level interactions, and thus potentially suggest new theoretical criteria to be included in matching (that need to be interpreted with scholar's insight). At the same time, adding the supervised method might lower concerns regarding fitting the model to a particular circumstance because machine learning uses quantifiable criteria to prune the model, as discussed in more detail below. The conclusion of our review of the prior propensity score papers in strategy is that supervised machine learning could meaningfully expand what is currently at least partly an ad-hoc process of determining the first

stage variables and thus suggests value to adding ML matching in the strategy scholar's causal-inference methods toolbox.

How machine learning can help. Given the common use of matching in the first stage, what alternatives does machine learning offer for matching models to deal with sample selection bias? We discuss several below, such as employing additional tools to prevent underfitting and overfitting, along with quantifiable metrics to effectively measure how well selection bias is accounted for.

First, with complex data, as noted above, researcher reasoning used to pick *confounders* is likely to reach its limits. Machine learning helps by pointing to which confounders might matter, thus reducing subjectivity in the covariate selection process. This way, machine learning is also likely to motivate researchers to think deeper about their own underlying assumptions of which confounders matter. By examining highly correlated covariates—of which ML typically picks a subset (which may or may not overlap with the researcher's set as Angrist and Frandsen (2022) note) —the researcher is pushed to reason why their preferred confounder should be on the list.

Second, supervised machine learning helps determine precise *functional forms*. As Mullainathan and Spiess (2017, p. 101) note, "including all pairwise interactions would be infeasible as it produces more regressors than data points (especially considering that some variables are categorical)." In contrast, supervised machine learning searches for these interactions automatically, and "allows us to let the data explicitly pick effective specifications, and thus allows us to recover more of the variation and construct stronger" instruments and predictions (Mullainathan & Spiess, 2017, p. 101). Benefit is a more thorough covariate selection process.

Third, even if scholars have the foresight and patience to hand-curate and test a large number of variables and their interactions, the expansion of dimensionality may lead to *overfitting*. Overfitting occurs when additional dimensions[6] do not increase the ability of the propensity score model to predict treatment, but rather explain some noisy covariance. The resulting propensity scores will then deviate from the true propensity, which risks making the second stage estimations invalid. Simulations have shown that relying on overfit propensity scores results in inflated standard errors (i.e., lack of precision) (Schuster, Lowe, & Platt, 2016), and over-estimated or under-estimated second-stage effects (Cepeda, Boston, Farrar, & Strom, 2003). Thus, overfitting can severely limit the interpretability of the second stage's estimates of treatment's performance effects, and as such, points to the ability of machine learning to help with matching in the first stage. Traditional propensity score models for instance rely on measurements of model fit (pseudo-R^2) to judge confidence. The higher the R^2 metric, the better the propensity score model fits the available data, and scholars have greater confidence that their model is functioning correctly. However, a higher R^2 is not necessarily related with predictive performance, only with how well the model fits the currently available data (UCLA, 2012; Ahrens, Hansen, & Schaffer, 2020). Therefore, in instances of overfitting, the model fit will either stay the same or increase, giving researchers false confidence. While traditional first-stage models (or any model with a binary outcome) have no *objective measurement* to assess the severity of the problem, with supervised ML, additional metrics including AUC (area-under-the-curve) are available.

Overall, machine learning methods can possibly assist strategy researchers in first-stage prediction. Two core features of supervised machine learning - regularization and cross-validation

[6] In ML terms, a covariate is defined as a group of one or more "dimensions" (Mullainathan & Spiess, 2017). This distinction is important for discrete variables, such as technology class. If there are 473 different technology classes, for example, it means there are 473 potentially confounding technology-class dimensions. On the other hand, continuous variables are each considered as one-dimensional covariates.

- are particularly relevant. In the next section, we demonstrate how to apply these techniques to a two-stage matching design using technology strategy data.

3| EMPIRICAL DEMONSTRATION: THE CASE OF TWO-STAGE MATCHING MODELS AND MACHINE LEARNING

Following prior work's suggestion that new applications of machine learning in strategy research should start with an interesting research question and an empirical data source, we investigate the relationship between public-private sector R&D relationships and their potential to produce invention disruption. In particular, we ask whether partnering with the federal government increases how disruptive the resulting inventions are. Per Funk and Owen-Smith (2017), we define disruptiveness by how much a patented technology disrupts existing patent citation patterns, i.e., how much it challenges the existing order and decreases the use of incumbent technologies.

The question about the federal government's role in disrupting status quo has a long-standing tradition in the strategy field and we continue this line of research (e.g., Trajtenberg, Henderson, & Jaffe, 1997; Pahnke, Katila, & Eisenhardt, 2015; Bruce, de Figueiredo, & Silverman, 2019; Rathje, 2019; David, Hall, & Toole, 2000).[7] Much research points in the direction that partnering with the federal government (rather than with private firms, for example) has a positive relation with disruptiveness of the resulting invention. Katila and Shane (2005: 826) for example suggest that private sources of capital such as VC may "lag behind rather than lead innovation" possibly because they favor more local, less riskier innovation search paths (Katila & Ahuja, 2002), and that instead, "contract research, and government programs, … may be more important sources of capital for firms at early stages of technological development" (Katila and

[7] Prior research has for example studied the effect that different public funding mechanisms (e.g., grants, contracts, cooperative agreements) have on technological innovation (Bruce et al., 2019; Rathje, 2019). As it is nearly impossible to run policy "experiments" that randomly treat technologies with public funding, yet public funding's impact is critical to understand, the setting is relevant to illustrate the use of supervised machine learning for matching.

Shane, 2005: 862). Similarly, Funk and Owen-Smith (2017) find that when universities receive more federal funding for academic research, they tend to produce more disruptive inventions. Scholarship also suggests that federal funding of private R&D likely results in more "radical" innovations than private research alone (see e.g., Corredoira, Goldfarb, & Shi, 2018).

Our data specifically examines R&D relations where a public funding partner and a private firm jointly solve a complex technical problem (David et al., 2000) and the private firm typically retains exclusive rights to the resulting invention. In the U.S. (the setting for our study), legislation, i.e., the Federal Grant and Cooperative Agreement Act of 1977, defines the types of R&D relations that these public-private relationships can take. The three types are cooperative agreements, contracts, and grants, and we examine their disruptive invention impact relative to inventions developed by the private sector firm alone.

In contrasting inventions developed with public sector partners vs without them, a key challenge is selection bias. The ideal solution is a randomized experiment, but this is impractical. Instead, we use matching on the first stage to find comparable treated and control groups, and, using the matched sample, examine the invention impact in the second stage model. In particular, we compare the "treated" group of technologies developed by private firms together with the public sector partner with the "control" group of technologies developed without the public sector partner (i.e., by the private firm alone). In other words, some technologies were "treated" with public partners, some not, and we are interested in effects of treatment on disruptiveness. Empirically, we examine what machine learning can add to the strategist's toolkit in first-stage matching to determine variables on which to match.

Data sources and variables. Our dataset is the population of patented technologies (inventions) granted to private firms in the U.S. over a 31-year period between 1982-2012. The

data include 3,337,229 patented technologies, and these data were collected from United States Patent and Trademark Office's (USPTO) PatentsView and triangulated with data from Google's Patent Search and USPTO main data. (Less than 1% of the PatentsView records had missing dates. We used Google's Patent Search to rectify the missing data.). Because the Bayh-Dole Act of 1980 required patent applicants to disclose federal support (Rathje, 2019), U.S. patents funded by the federal government are required to file a "government interest statement," detailing the government's involvement. We used the government interest statement data compiled by PatentsView to track public sector partners.

Of the patented technologies, 58,082 had public sector partners (i.e. were treated), and we used machine learning for the first-stage prediction task to create propensity scores to identify the treated and control groups. Consistent with our research question on public partner's potential to produce disruption, the second-stage dependent variable is *invention disruption.* It tracks how future inventions make use (or not) of the technological predecessors of the focal invention. Invention disruption was measured by how much a patented technology disrupts patent citation patterns using the ratio of forward citations that uniquely cite the originating patent and not its predecessors (Funk & Owen-Smith, 2017). Larger values of invention disruption indicate more radical break from prior work: Disruptive technologies, i.e., patents uniquely cited by future patents receive a score of one, and consolidating technologies, i.e., patents that contribute to reinforcing current citation patterns receive a score of minus one. Per Funk and Owen-Smith (2017), we calculated invention disruption five years after the focal patent's issue date.[8]

[8] We operationalize invention disruption with the formula $\frac{1}{n_t}\sum_{i=1}^{n}\frac{-2f_{it}b_{it}+f_{it}}{1}, w_{it} > 0$ (Funk & OwenSmith, 2017) where n_t is the number of forward citations of both the focal patent and its predecessors, f_{it} equals 1 if a new patent cites the focal patent and is 0 otherwise, b_{it} equals 1 if a new patent cites the focal patent's predecessors and is 0 otherwise.

We also include several independent and control variables. As noted above, all U.S. patents must include a government interest statement. We used the government interest statements to code *public-private relationships*, of which there are three types, *contracts*, *grants* and *cooperative agreements*, per legislation (Federal Grant and Cooperative Agreement Act of 1977, U. S. Congress, 1977). Other key variables in the data include *patent originality*, measured by the breadth of patent's technical fields in backward citations (Trajtenberg et al., 1997). Three-digit U.S. patent classes were used to categorize technical fields. Breadth is calculated using a Herfindahl index ranging from zero to one. If a patent cites previous patents in a narrow set of fields the originality score will be low whereas wide range yields a high score. *Previous relationships* was measured by the natural logarithm of the count of prior times the private firm had engaged in public-private relationships that resulted in a patent. *Patent age* was measured by the number of years since a patent was granted, *time to grant* by the number of years from filing to grant of a patent, and *number of inventors* was measured by the count of total number of inventors listed on the patent document. We also included fixed effects for invention *location* using categorical variables *state* and *country* at the time of invention, for patent *technology class* using dummy variables for the main U.S. patent classes and for patent *application year* using calendar years.

We chose the empirical setting because it is a particularly relevant context to illustrate how machine learning can be used. The sample size is reflective of larger samples and many covariates now more and more commonly analyzed by strategy scholars. Further, first-stage confounding covariates are complex, likely interdependent, and not readily and fully understood in this setting.

Looking Back: Traditional Approaches

To test the public sector treatment effect without machine learning, technology strategy scholars typically use three steps. The first step is to use prior literature, data and reasoning to identify the potentially confounding covariates that would limit the ability to interpret (causal) treatment effects. In the case of public-private R&D relationships, the norm is to use calendar time and technology (Agrawal, Cockburn, & Rosell, 2009; Belenzon & Schankerman, 2013; Trajtenberg et al., 1997). In practice, patented technologies are typically matched on age (calendar years that proxy for factors such as variations in macroeconomic and technical climate at the time of patent filing and granting), and patent technology class (e.g., electronics, computers, or manufacturing with different underlying technical requirements and ecosystems).[9] An important question is whether these 2-3 covariates are enough. For example, will it be enough to take into account broad technology (e.g., information technology) but not technology specialization (e.g., AI)? Using traditional methods, we take it for granted that the assumption to use the traditional covariates is "correct" and proceed with the analysis such as matching.

Possible shortfalls of traditional approaches. In addition to the standard covariates noted above, others may also be important. In pioneering work, Alcácer and Gittelman (2006) found that patent examiners are strongly associated with both (1) the specific scientific or technical specialization of the underlying technology, and (2) the number of forward citations a patent receives (i.e., patent quality) (see also Katila, 2000). It is reasonable to believe then, that patent examiners, as a reasonable proxy for the underlying technology specialization, could be correlated

[9] Prior research identified technology class, application year, and grant year as important (Pavitt, 1982) because it is reasonable to believe that if public-private relationship patents come from a different population of technology classes than corporate (private only) patents, the success of public-private relationship patents may in fact be derived from differences in the technology class (e.g., some technology classes have more impact than others). Similarly, if public-sector funding for R&D in the United States is more common during a Democratic administration than a Republican administration, then public-sector patent performance may be influenced by application or grant year (e.g. when public organizations have more additional funding to spend on research or patenting in general).

with both the treatment (public sector relationship) and the outcome (such as a patent's disruptive quality). Similarly, prior work associates the size of the R&D team (number of inventors on a patent) with treatment (public sector relationship) and subsequent impact of the patented technology (patent quality) (Pahnke et al., 2015). Public-interest inventions may for example require larger teams and also have more disruptive potential. Similarly, team's geographic location may be correlated with both likelihood of public sector relationships (treatment) and with patent quality. In combination, patent examiners, number of inventors, and geographic location represent missing but potentially relevant confounding covariates.

Why is it, then, that despite our theoretical understanding of their importance, patent examiners for example are not included in current (typical) matching criteria? A key reason is that current methods do not provide an objective way to determine which confounders to include when the number of confounders increases beyond what can be included in the models. For instance, there are 25,439 unique patent examiners (i.e., examiner names) in the patent population that we study (1982-2012), so it would be practically impossible to match exactly on each examiner, or to include additional functional forms. As the number of covariates increases, dimensionality grows dramatically. Even if each covariate is binary, thus containing only two discrete dimensions, as the number of covariates grows by P, the number of dimensional combinations grows by 2^P (Rosenbaum & Rubin, 1983). Very quickly, this value grows beyond the total number of observations. Thus, the increase in observations severely limits the effectiveness of traditional techniques to capture the additional confounders.

Additionally, a limitation is that it is often not possible to know which confounders to add. What scholars with traditional methods cannot provide is a numerical evaluation of this step. As we describe in more detail below, machine learning offers a quantifiable step in this direction.

Moving Forward: Steps of Including Machine Learning

How to apply machine learning. We start by including the standard covariate set of technology class, application year, and grant year. We then consider additional covariates which have not been used in matching before but are also likely to be confounding covariates per prior theory and research in the area, including number of patent inventors, geographic location (state, country) (Pahnke et al., 2015), and patent examiners (Alcácer & Gittelman, 2006), as described above. In total, this results in 26,269 potentially confounding dimensions. Note that while exact matching could be used, it would not be able to (and could not) identify exact matches for every possible combination.

(1) Split the data. Machine learning approach starts by splitting the data into subsets: training, validation, and test (Chernozhukov et al., 2018).[10] This step allows us to quantitatively test how well the results generalize. There are no exact rules of how to split the data. However, it is preferable to increase training efficiency by maximizing the size of the training set when possible (Jain & Chandrasekaran, 1982; Raudys & Jain, 1991). Given the large sample size of our patent data, we need a relatively small share of observations from which to test our predictions, and therefore chose a 98-1-1 data split (Picard & Berk, 1990; Reitermanová, 2010). This means that 98% of the data were used for the training set, while one percent was used for validation and test sets, respectively.[11] A reasonable general rule of thumb is ensuring that the test data set has at least as many observations as the total number of dimensions in the data.[12] If

[10] The validation (aka development) set is an intermediary between the training and test sets. The validation set is used during training to fine-tune the model parameters such as regularization and to assess model performance. In contrast, the test set is used only at the end.

[11] We conducted preliminary experiments with 80-10-10 or 90-5-5 and found consistent results. In smaller datasets, a typical split is 60-20-20.

[12] In ML terms, a covariate is defined as a group of one or more "dimensions" (Mullainathan & Spiess, 2017). This distinction is important for discrete variables, such as technology class. In Trajtenberg et al. (1997), there are 496

there are less, then k-fold cross validation (k times with different splits) could be used, although some researchers have reservations about the method (Ahrens et al., 2020). A good rule of thumb is that if the researcher is interested in inference, training set should be large, and if the researcher is interested in prediction, the test set should be large (Ahrens et al., 2020). The key is to split the sample randomly: We train the prediction on a training set which is separate and distinct from the test set. As Varian (2014, p. 7) notes: "For many years, economists have reported in-sample goodness-of-fit measures using the excuse that we had small datasets. But now that larger datasets have become available, there is no reason not to use separate training and testing sets."

(2) Select a machine learning model, add regularization, and cross-validate. Regularization and cross-validation are employed in tandem in the next phase (for readers not familiar with these methods, we refer to Rathje (2019) and Choudhury and colleagues (2021)). First, we select a supervised learning method for our predictive model, H_θ. Because we use propensity score matching (PSM), we use a logistic regression model.[13] Logistic regression is useful because it generates a continuous probability of an observation being treated by regressing the observation's set of covariates (X_i) on its observed treatment (Y_i). As a result, logistic regression model (as opposed to other machine learning techniques, such as support vector machine (SVM)) allows us to generate continuous estimates of propensity scores.

Regularization aims to decrease subjectivity in the covariate selection process and prevent overfitting. While traditional methods require ample justification that some observable covariates

different technology classes, which means there are 496 potentially confounding technology-class dimensions. On the other hand, since time is a continuous variable, application year and grant year are each considered as one-dimensional covariates. In total, Trajtenberg et al. (1997) then included 498 potentially confounding dimensions.

[13] A few other regression models have been suggested and discussed in the statistical literature, regression trees in particular. See Lee et al. (2010) for an interesting comparative analysis of alternative models.

are more important than others (e.g., see Stuart & Rubin, 2008 re matching), through regularization, machine learning "suggests" covariates important for prediction. Regularization thus starts with a list of variables that the researcher has compiled and then prunes down the list, limiting model complexity. A standard way to regularize is to penalize complex models in favor of simpler models. For instance, like ordinary least squares regression (OLS), regularized linear regression minimizes the sum of squared deviations between observed and model-predicted values but imposes a regularization penalty aimed at limiting model complexity (see Rathje, 2019 for more details). The most common forms of regularization applicable for regression models are Lasso (L1), Ridge (L2) and Elastic Net, implemented for example in Stata (Ahrens et al., 2020).

Cross-validation allows scholars to validate the quality of the prediction by quantitatively testing how well the results generalize to yet unseen test data. In cross-validation, machine learning again adds a quantitative performance evaluation metric as researcher reasoning is augmented with a quantifiable, "automated" approach.

With the final model (i.e., after regularization and cross-validation is complete), we can investigate the coefficient weights (sizes of θ) to determine which covariates contributed to treatment selection bias.[14] Those covariates with the largest coefficient weights in magnitude contribute the most to treatment selection bias, while those with smaller weights have less of an effect. As a result, higher weights indicate the need to include the corresponding covariates in the matching model.

[14] "Although not as rich as partial dependence plots, examining variable importance (when possible) is a reasonable first step for interpreting ML models" (Choudhury et al., 2021, p. 49). In its simplest form, supervised machine learning estimates a predictive model's coefficient weights with a maximum likelihood estimator. Regularization term drives the weights of coefficients which do not contribute to selection to zero and/or normalizes coefficient weights to ensure that no singular dimension dominates the model. More details about maximum likelihood estimators and regularization are provided in the appendix.

In practice, to get more insight on covariates, researchers often plot a covariate-weight graph as a visual tool to investigate coefficient weights as we have done in figure 1. Covariate-weight graph is a bar chart representing the range of dimension weights, grouped by covariates (e.g., in figure 1 each examiner name is a dimension, examiners is a covariate) with the goal to get intuitive understanding of the first stage. It is important to note that variables that are highly correlated with one another contribute redundant information, so the model may prefer one variable over another in a seemingly random fashion (Mullainathan & Spiess, 2017). That is, the graph is not a good tool to compare variables in terms of causal explanation, and inspecting the correlations is an important first step (which we do in table 1).

As an example, figure 1 represents the covariate-weight graph for the top five most heavily weighted covariates in our data. Interestingly, patent examiners are the most heavily weighted covariate, followed by technology class and then country. These top three covariates indicate a need to include a partly different set of confounding covariates in estimating treatment effects than the *traditional* covariate set (i.e., the traditional set is technology class, application year, grant year). Indeed, 2 of the top 3 covariates are different from the traditional set, and only technology class makes it to the top set in figure 1.

INSERT TABLE 1 AND FIGURE 1 ABOUT HERE

To dive deeper into the examiner covariate and to investigate specific examiners, table 2 represents the coefficient weights of the top three highly weighted patent examiners (both positive and negative) along with three patent examiners who were relatively unimportant and driven to zero by regularization. Weights correspond to the probability of treatment. As an example, patent examiners Zarfas, Ansher, and Douglas have large and negative weights, indicating that they are more heavily assigned to examine technologies in the control group, i.e. patents granted to private

firms alone, without public partners. Patent examiners Goddard, Grarsay, and Ryam have large and positive weights, indicating that they, in contrast, are more heavily assigned to technologies in the treated group, i.e. they more likely examine patents from public-private relationships rather than those where private firms worked alone. (In contrast, Knife, Renner, and Chan likely examine both equally, and so do not help to predict propensity scores. Therefore, regularization drove their weights to zero.) Said differently, if patent examiners indeed proxy technology specialty, negative and positive weights indicate that treated technologies likely differ from control ones in technology specialty and therefore may contribute to selection bias unless accounted for.

INSERT TABLE 2 ABOUT HERE

(3) Repeat by adding interactions. Next, the goal is to test the model by iteratively adding more covariates. As is typical in machine learning, and to illustrate the process step-by-step, we repeat the previous steps by first adding all quadratic interactions, and then adding all quadratic and tertiary interactions as potentially confounding covariates. For example, we include year x technology class, year x examiner name, and year x technology class x examiner name, because it is reasonable to assume that in certain years, for example, particular technology classes were more likely to be "treated" with public funding. To protect against overfitting, we evaluate the performance of each model independently.

(4) Evaluate and select the ML-PSM model. Once all the models are created, the next step is to compare their performance. When evaluating performance, an effective combination is to measure predictive performance using area under the curve (AUC) and the model fit using R^2. As noted above, because AUC is the measurement of predictiveness, supervised machine learning prioritizes it as the principal measurement for model selection. Simply put, AUC is a ratio of true positive rate (TPR = (True Positives / (True Positives + False Negatives)) and false

positive rate (FPR = (False Positives / (False Positives + True Negatives)) (Narkhede, 2018).[15] R^2, additionally, allows us to interpret the impact of increasing model complexity on fit.[16] As model complexity grows (e.g., adding all interactions), if AUC begins to decrease while R^2 increases, the model is likely overfitting (Ahrens et al., 2020). Other classification metrics such as precision, recall, and accuracy can also be considered.

To illustrate, we evaluate five different models by comparing changes in AUC (predictive performance) and R^2 (fit): (1) machine learning-propensity score matching method using three traditional confounders, (2) machine learning-propensity score matching with the addition of four other potential confounders suggested by theory (patent inventors, geographic location, originality, patent examiners), (3) machine learning-propensity score matching adding quadratic (two-way) interactions, (4) repeating step #3 with regularization, and (5) adding cubic (three-way) interactions.

INSERT TABLE 3 ABOUT HERE

In table 3, the baseline (model 1) has the lowest predictive performance (AUC) and fit (R-squared). Adding 4 additional covariates in model 2, and quadratic interactions in model 3 further increase performance. As expected, the AUC scores in the training data are higher than in the test data, indicating that increasingly complex models that fit the training data well may not generalize to the test data as well as simpler ones. Adding quadratic interactions in model 3 again increases predictive performance but exacerbates over-fitting to the training data. Adding regularization

[15] AUC ranges in value from 0 to 1, with 1 being more predictive. Simply put, AUC is the probability that the model ranks true positives more highly than false positives. In technical terms, AUC stands for "Area under the ROC Curve." That is, AUC measures the two-dimensional area underneath the ROC (receiver operating characteristic) curve where ROC curve plots TPR vs. FPR at different thresholds, and top-left side curves are better. It should be noted that in use cases where the objective is to minimize false positives (rather than maximize true positives), AUC is less useful.

[16] Since logistic regression models cannot generate accurate R^2, a pseudo-R^2 is applied. Adjusted count is a simple pseudo-R^2 commonly used in logistic regressions, and the one used in our approach (UCLA, 2012).

(variable pruning) in model 4 illustrates this effect. Unsurprisingly, by simplifying the model through regularization, model performance using the *training* data *decreases*. However, simplifying through regularization *increases* the predictive performance (AUC) in the *test* data. These differences provide evidence that regularization and cross-validation can help against overfitting and help with generalizability. Finally, the addition of cubic interactions in model 5 provides strong evidence of overfitting. The model simply has too many terms now and explains random noise on top of the trend. Overfitting is particularly apparent in the test data, where R^2 increases while AUC decreases. Such a shift is a clear example of limits of elastic net regularization. Importantly, combination of AUC and R-squared scores provides a useful set of metrics to indicate what the preferred first-stage model is (in this case model 4, i.e., the regularized ML-PSM model including quadratic interactions).

(5) Match on propensity scores and plot the results. Next, we can evaluate how our preferred model from step 4 does in terms of generating a balanced sample. Using the standard steps of propensity score matching, we first generate propensity scores for each observation (i.e. predicted likelihood of public sector relationship). Next, we match treatment and control propensity scores one-to-one using a global-optimum matching algorithm. There are a wide variety of matching algorithms, but in cases where the control group is much larger than the treatment group, a global optimum strategy is preferred (Stuart & Rubin, 2008). Lastly, we remove the unmatched sample, leaving us with balanced treatment and control groups.

To evaluate balance, we revisit the balance plots from step one using the matched sample. First, balanced support in propensity scores is assessed visually (figure 2), that is, evaluating whether treatment and control groups are balanced across a composite of all covariates (Dehejia & Wahba, 2002). Second, balanced support between the covariates is assessed (figure 3, right-hand

side). Ideally, the distributions should be identical across the treatment and control groups, as shown in figure 3. Other methods to evaluate balance are discussed for example in Thatchenkery and Katila (2021).

INSERT FIGURES 2 and 3 ABOUT HERE

(6) **Inspect and compare second-stage regressions.** After validating that the treatment and control groups are balanced across confounding covariates, we finish executing the ML-PSM method by assessing the treatment effects. Comparing treatment effects across different inference strategies is a common way to assess whether or not the treatment is subject to selection bias (Imbens, 2004). If the treatment effect stays consistent across all strategies, then matching is likely not required. However, if the coefficients are inconsistent (e.g., changes in sign), then more robust matching approaches are advised (Imbens, 2004; Stuart & Rubin, 2008). In the empirical example used in this paper, the power of machine learning method to make treatment and control groups comparable can be assessed by comparing the second stage treatment effects using ML-PSM vs. the second stage treatment effects using other inference strategies. By comparing multiple strategies, we can determine when machine learning-matching brings in benefits. Although assessing the treatment effects is not at the core of our analysis, there are several interesting observations about these second stage results that can help compare ML with the other approaches.

Table 4 compares matching on some typical variables used in similar prior studies with what ML produces:[17] pre-match (unmatched) full population (models 1a and b), exact matching using the typical criteria of technology class, application year, grant year per Trajtenberg and colleagues (1997) (models 2a and b), PSM matching using the typical criteria of technology class, application year, grant year (models 3a and b) and ML-PSM (models 4a and b). In all models, the

[17] This is feasible as no patent can be treated with multiple funding types. It, therefore, adheres to the stable unit treatment value assumption (SUTVA) allowing for sub-classification in second stage regressions (Rubin, 1980).

second stage outcome is invention disruption, measured by how much a patented technology disrupts patent citation patterns (Funk & Owen-Smith, 2017). As noted above, our "treatment" condition is those patented technologies by private firms which were developed together with public partners (see 1a; 2a; 3a; 4a). Per Stuart and Rubin (2008) to better assess when matching methods are useful, we decompose the treatment in three sub-classifications discussed above - contracts, grants, and cooperative agreements (see models 1b; 2b; 3b; 4b). To calculate the treatment effect, we run an OLS regression to predict invention disruption, given the treatment funding mechanism.

There are several results. First, matching matters for second stage results in table 4: several key coefficients in the unmatched regressions (models 1a and b) differ from the matching models (models 2a-4a and 2b-4b). Comparing treatment effects across different inference strategies (Imbens, 2004) in table 4, the coefficient for public-private relationship ranges from negative for unmatched (β=-0.01; model 1a) to positive for ML-matching (β=0.02; model 4a), pointing to importance of matching. Similarly, decomposing the public-private relationships into subcategories (Imbens, 2004), the coefficients of contracts and grants on invention disruption switch from null and negative respectively (models 1b) to positive (models 4b) when matching is added whereas coefficient values of cooperative agreements remain consistently negative. We return to this interesting difference below.

Examination of matching models also indicates that the estimated coefficients for the public-private relationship, contract, grant and agreement variables are consistent across matching models (with controls) although the coefficient on the grant variable is estimated with more precision in the ML-PSM model. In other words, the matching results are consistent across approaches (models 2a-4a and 2b-4b), and only slightly different with the use of ML (models 4a

and b). Overall, given that the original (i.e. unmatched) estimates of contracts and grants appear to be biased, it seems likely that matching is a useful method to control for selection bias for grants and contracts.

A qualitative examination of the selection (i.e., solicitation) process for grants and contracts vs cooperative agreements is helpful to explain why grants and contracts are more likely exposed to treatment selection bias while cooperative agreements are less likely (see table 4). As noted above, in our setting, patents are "treated" with public funding when a public funding agency funds a corporate R&D project. However, these funding approaches differ for grants and contracts vs cooperative agreements. Grants and contracts are initiated when a public funding agency publicly posts a solicitation, defined as "a document which provides the requirements and instructions for the submission by eligible applicants" (U.S. Department of Labor, 2010). In the solicitation of grants and contracts, the public funding agency determines the parameters of the R&D project before beginning a project with a company. For example, the Office of Basic Energy Sciences (BES), out of the U.S. Department of Energy, recently released a solicitation calling for Advanced Fossil Energy Technology Research, which would fund firms up to nearly two million dollars to work on: "the development of innovative, cost-effective technologies for improving the efficiency and environmental performance of advanced industrial and utility fossil energy power generation and natural gas recovery systems (U.S. Department of Energy, 2019)." Similarly, the U.S. Department of the Navy released a solicitation calling for a High-Power Compact Fuel Cell System which could fund firms up to three million dollars to, "develop a compact fuel cell system (e.g., stackable fuel cells, hydrogen and oxygen fuel sources, all balance-of-plant equipment including by-product management components) capable of producing, at a minimum, 500 kW of power." (U. S. Navy SBIR/STTR, 2019) In both instances, grant and contract R&D projects, respectively,

are initiated by the public funding agency. After reading the solicitation, the company can then choose, or choose not, to apply. Therefore, with grants and contracts, the public R&D organization solicits the private company, and the company must select into the funding relationship.

Cooperative agreements have a different selection process, which may explain why they may be less likely to be exposed to treatment selection bias. With cooperative agreements, public-funding agency solicitations are not required. Instead, the parameters of the research project are often *co-developed* by the public funding agency and the company (Ham & Mowery, 1998). As opposed to grants or contracts, cooperative agreements are used when "substantial involvement is anticipated between the executive agency (public funding agency)…[and the] other recipient (company) during performance of the contemplated activity." (U. S. Congress, 1977) Given this social participatory requirement, policymakers have allowed companies and public funding agencies to *co-develop* the parameters of the R&D project before selecting into the project. A typical result is that the company, not the public funding agency, initiates the cooperative agreement project. In turn, prior work has shown that the public funding agency agrees to the cooperative agreement only if its interests align with the company (Bruce et al., 2019). This solicitation process stands in stark comparison to grants and contracts. Instead of a public funding agency soliciting the company, it begins with the company soliciting the government which likely explains the differences in treatment selection.

Legally, the difference in solicitation processes draws from the fact that grants and contracts must be awarded through open competition[18] (Office of Acquisition and Property Management, 2011). For cooperative agreements, this is strongly encouraged, but not necessary. For example, the National Institutes of Health notes that "fair access to CRADAs [i.e., cooperative

[18] There are exceptions when grants and contracts can be "sole-sourced" to specific companies, but those events are rare in comparison to cooperative agreements.

agreements] is not to be considered as synonymous with the term 'open competition,' as defined for contracts and small purchases." (NIH, 2023) This legal justification allows for a difference in who approaches whom when entering into an R&D project.

Figure 4 provides additional evidence. It illustrates the propensity scores for contracts, grants, and cooperative agreements in comparison with the propensity score distribution for private firm-only patents. While grant and contract distributions are different from that of private firm-only patents, cooperative agreement distribution aligns closely with private firm-only patents. Again, this additional evidence provides further support for the argument that cooperative agreement treated patents are sampled from a similar population to the private firm-only control group, whereas grants and contracts are likely sampled from a very different population necessitating careful matching to generate a balanced sample.

INSERT TABLE 4 AND FIGURE 4 ABOUT HERE

As an overview, the list of steps for applying machine learning is provided in table 5, and a comparison of traditional matching models vs ML-assisted matching models in table 6.

INSERT TABLES 5 AND 6 ABOUT HERE

4| DISCUSSION

This paper offered machine learning as an additional tool for two-stage matching models. Our particular focus was on the use of machine learning in the first stage to determine variables on which to match. Empirical researchers in strategy and organizations who tackle causal inference by including a prediction step in two-stage models (such as propensity score matching where the first stage is a prediction or IPTWs; e.g., Thatchenkery & Katila, 2023; Katila et al., 2017) can benefit from considering the method. The method can also contribute to strategy conversation around causal inference strategies in general given its partly "automated" nature.

There are several contributions. First, our literature review of existing work presented compelling evidence that traditional first-stage approaches (identifying potentially confounding variables by relying on somewhat ad-hoc researcher choice) could potentially be incomplete and could be enhanced by new additional matching methods. Our review of the machine learning methods useful for choosing matching variables can help address many of these limitations and thus help scholars in causal inference.

Second, we outlined a detailed step-by-step methodology that illustrated how to apply machine learning to propensity score matching and utilized a technology strategy data set – patent data on public-private relationships– to demonstrate how and when machine learning could be useful. By incrementally adding potentially confounding covariates and assessing treatment effects enabled us to illustrate the core details of the method. We found, per Imbens (2004) that matching methods were advised due to coefficients changing in the second stage models. We used a sequential, data-driven approach to adding confounding covariates and their complex functional forms. This provided a systematic approach to not only enhance robustness of model prediction in the first stage, but also potentially de-bias estimates in the second stage.

It is important to note that we chose these data simply as an example to demonstrate the ML-PSM matching technique. Although our intent is not second stage policy evaluation, we find our second stage results to be consistent with those of other matching methods, and in one case of public sector partners (i.e. grants), the machine learning estimates to be relatively more precise. These results increase our confidence in the ML method as an additional tool that researchers can use in regression studies.

As with any method, there are of course boundary conditions. First, we cannot relinquish our responsibility as researchers to automation or to a "machine". Human interpretation still plays

a crucial role: Researcher expertise is still very much needed, both in constructing and interpreting the models. In fact, without human reasoning, machine learning approaches carry the risk of black-boxing and thus introducing more uncertainty than reducing it.

Further, machine learning shines when the regression datasets are large with many, complex confounders (e.g., Katila, Piezunka, Reineke, & Eisenhardt, 2022). In small datasets with few confounders, an analytic regression is likely to be an equally effective alternative. A practical rule of thumb often suggested in the literature is that if there are many predictors relative to observations, it is worth considering machine learning approaches for variable selection. In the reverse case of a large number of observations relative to the number of predictors, running a logistic regression is often computationally feasible, and there is no need to run machine learning.

In conclusion, applying machine learning for causal inference is starting to surface as an increasingly useful method across various disciplines. Integrating machine learning concepts into existing practices can assist in a number of strategy problems examined today, encourage thoughtful and thorough covariate selection practices, and increase appreciation of 'automated' data practices.

ACKNOWLEDGEMENTS

We thank our editor Connie Helfat and our two anonymous reviewers as well as Raj Choudhury, Sharad Goel, Kenneth Huang, Rahul Kapoor, Ray Levitt, Risto Miikkulainen, Markus Pelger, Scott Stern, Ron Tidhar, Bart Vanneste, and Georg von Krogh for discussions, generous advice, and comments. We are thankful for the valuable comments from the audiences at the Academy of Management Big Data Conference, the Academy of Management Annual Meetings, the 2nd, 3rd and 4th AI and Strategy Consortia, LinkedIn Group Talk series on "The Organizations and AI", the Strategic Management Society Annual Meetings, and the Wharton Technology Conference. Our research project was supported by the Stanford Technology Ventures Program (STVP).

REFERENCES

Agrawal, A. K., Cockburn, I. M., & Rosell, C. (2009). Not invented here? Innovation in company towns. *NBER Working paper 15347*.

Ahrens, A., Hansen, C., & Schaffer, M. (2020). lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal,* 20(1), 176-235.
Alcácer, J., & Gittelman, M., (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economics and Statistics,* 88(4), 774–779.
Angrist, J., & Frandsen, B. (2022) Machine labor. *Journal of Labor Economics*, 40(S1), S97-S140.
Asgari, N., Singh, K., & Mitchell, W. (2017). Alliance portfolio reconfiguration following a technological discontinuity. *Strategic Management Journal,* 38(5), 1062-1081.
Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4), 597-623.
Bahceci, E., Katila R., Miikkulainen, R. 2015. Evolving strategies for social innovation games. *Proceedings of the Conference on Genetics and Evolutionary Computation (GECCO)*.
Bahceci, E., Katila, R., & Miikkulainen, R. 2023. Evolving strategies for competitive multi-agent search. arXiv:2306.10640
Belenzon, S., & Schankerman, M. (2013). Spreading the word: Geography, policy, and knowledge spillovers. *Review of Economics and Statistics*, 95(3), 884–903.
Belloni, A., Chernozhukov, V., & Hansen, C. (2014) High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29–50.
Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1), 233-298.
Bettis, R., & Blettner, D. (2020). Strategic reality today: Extraordinary past success, but difficult challenges loom. *Strategic Management Review*, 1(1): 75-101.
Bettis, R., Gambardella, A., Helfat, C. & Mitchell, W. 2014. Quantitative empirical analysis in strategic management. *Strategic Management Journal*, 35(7), 949-953.
Blakely, T., Lynch, J., Simons, K., Bentley, R., & Rose, S. (2020). Reflection on modern methods: When worlds collide - prediction, machine learning and causal inference. *International Journal of Epidemiology,* 49(6), 2058–2064.
Bruce, J. R., de Figueiredo, J. M., & Silverman, B. S. (2019). Public contracting for private innovation: Government capabilities, decision rights, and performance outcomes. *Strategic Management Journal,* 40(4), 533–555.
Cannas, M., & Arpino, B., (2019) A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal,* 61(4), 1049-1072.
Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology,* 158(3), 280–7.
Chang, S.-J., Chung, J., & Moon, J. (2013). When do wholly owned subsidiaries perform better than joint ventures? *Strategic Management Journal,* 34(3), 317-337.
Chang, S., Kogut, B., & Yang, J.-S. (2016). Global diversification discount and its discontents: A bit of self-selection makes a world of difference. *Strategic Management Journal,* 37(11), 2254-2274.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and causal parameters. *The Econometrics Journal*, 21(1): C1–C68.
Choudhury, P., Allen, R., & Endres, M. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal,* 42(1), 30-57.
Choudhury, P., Wang, D., Carlson, N., & Khanna, T. (2019) Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal,* 40(11), 1705-1732.
Corredoira, R. A., Goldfarb, B. D., & Shi, Y. (2018) Federal funding and the rate and direction of inventive activity. *Research Policy,* 47(9), 1777–1800.
Cummings, T., & Knott, A. M. (2018). Outside CEOs and innovation. *Strategic Management Journal,* 39(8), 2095-2119.

D'Agostino, R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19), 2265-2281.
David, P. A., Hall, B. H., & Toole, A. A. (2000). Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy,* 29(4), 497–529.
de Figueiredo, R., Feldman, E., & Rawley, E. (2019). The costs of refocusing: Evidence from hedge fund closures during the financial crisis. *Strategic Management Journal,* 40(8), 1268–1290.
Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics,* 84(1), 151–161.
DesJardine, M., & Durand, R., (2020). Disentangling the effects of hedge fund activism on firm financial and social performance. *Strategic Management Journal,* 41(6), 1054-1082.
Durand, R., Paugam, L., & Stolowy, H. (2019). Do investors actually value sustainability indices? Replication, development, and new evidence on CSR visibility. *Strategic Management Journal,* 40(9), 1471-1490.
Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210), 1243089.
Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science,* 63(3), 791–817.
Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2020). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, 65, 101855.
Grushka-Cockayne, Y., Jose, V. R. R., & Lictendhal, K. C. J. (2016). Ensembles of overfit and overconfident forecasts. *Management Science,* 63(4): 1110–1130.
Ham, R. M., & Mowery, D. C. (1998). Improving the effectiveness of public - private R&D collaboration: case studies at a US weapons laboratory. *Research Policy,* 26(6), 661–675.
Hasan, I., Kobeissi, N., & Wang, H. (2011). Global equity offerings, corporate valuation, and subsequent international diversification. *Strategic Management Journal,* 32(7), 787-796.
Heckman, J. J., & Todd, P. E. (2009). A note on adapting propensity score matching and selection models to choice based samples. *The Econometrics Journal*, 12(suppl_1), S230-S234.
Helfat, C.E., Kaul, A., Ketchen Jr, D.J., Barney, J.B., Chatain, O. and Singh, H., 2023. Renewing the resource-based view: New contexts, new concepts, and new methods. *Strategic Management Journal*, 44(6), 1357-1390.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics,* 86(1), 4-29.
Jain, A., & Chandrasekaran, B. (1982). Dimensionality and sample size consideration in pattern recognition practice. *Classification, Pattern Recognition and Reduction of Dimensionality. Handbook of Statistics,* 2, 835–856.
Katila, R. (2000). Using patent data to measure innovation performance. *International Journal of Business Performance*, 2(1/2/3), 180-193.
Katila, R., & Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. *Academy of Management Journal*, 45(6), 1183–1194.
Katila R., & Shane S. (2005). When does lack of resources make new firms innovative? *Academy of Management Journal*, 48: 814–829.
Katila, R., Thatchenkery, S., Christensen, M. Q., & Zenios, S. (2017). Is there a doctor in the house? Expert product users, organizational roles, and innovation. *Academy of Management Journal*, 60(6), 2415-2437.
Katila, R., Piezunka, H., Reineke, P., & Eisenhardt, K.M. (2022). Big fish versus big pond? Entrepreneurs, established firms, and antecedents of tie formation. *Academy of Management Journal*, 65(2), 427-452.
Kumar, P., Liu, X., & Zaheer, A. (2023) How much does the firm's alliance network matter? *Strategic Management Journal*, 43(8), 1433-1468.
Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistical Medicine*, 29(3), 337-346.
Leiblein, M., Reuer, J., & Zenger, T. (2018). What makes a decision strategic? *Strategy Science,* 3(4),

558-573.
Mata, J., & Alves, C. (2018). The survival of firms founded by immigrants: Institutional distance between home and host country, and experience in the host country. *Strategic Management Journal,* 39(11), 2965-2991.
McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4), 403-25.
Miric, M., Jia, N., & Huang, K. (2022). Using supervised machine learning for large-scale classification in management research: The case of identifying artificial intelligence patents. *Strategic Management Journal,* 44(2), 491-519.
Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives,* 31(2), 87–106.
Nagaraj, A. (2022). The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry. *Management Science,* 68(1), 564–582.
Narkhede, S. (2018). Understanding confusion matrix. Towards Data Science. Retrieved from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
National Institute of Health (NIH), (2023). CRADAs. Retrieved from: https://www.techtransfer.nih.gov/policy/cradas.
Oehmichen, J., Firk, S., Wolff, M., & Maybuechen, F. (2021). Standing out from the crowd: Dedicated institutional investors and strategy uniqueness. *Strategic Management Journal,* 42(6), 1083-1108.
Office of Acquisition and Property Management. (2011). Procurement contracts, grant and cooperative agreements. In *Federal Assistance Programs, Grant Administration*: 1–5.
Ortiz-de-Mandojana, N., & Bansal, P. (2016). The long-term benefits of organizational resilience through sustainable business practices. *Strategic Management Journal,* 37(8), 1615-1631.
Pahnke, E., Katila, R., & Eisenhardt, K. (2015). Who takes you to the dance? How partners' institutional logics influence innovation in young firms. *Administrative Science Quarterly*, 60(4), 596-633.
Patel, P., & Cooper, D., 2014. The harder they fall, the faster they rise: Approach and avoidance focus in narcissistic CEOs. *Strategic Management Journal,* 35(10), 1528-1540.
Pavitt, K. (1982). R&D, patenting and innovative activities. A statistical exploration. *Research Policy,* 11(1), 33–51.
Picard, R. R., & Berk, K. N. (1990). Data splitting. *The American Statistician,* 44(2), 140–147.
Rathje, J. (2019). Hybrid conflict and innovation. Ph.D. dissertation. Stanford University.
Rathje, J., & Katila, R. (2021). Enabling technologies and the role of private firms: A machine learning matching analysis. *Strategy Science*, 6(1), 5-21.
Reineke, P., Katila, R., & Eisenhardt, K.M. (2024). Decentralization in organizations: A revolution or a mirage? Stanford Technology Ventures Working Paper.
Raudys, S.J. & Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3), 252-264.
Rawley, E. (2010). Diversification, coordination costs, and organizational rigidity: Evidence from microdata. *Strategic Management Journal,* 31(8), 873-891.
Reitermanová, Z. (2010). Data splitting. *Week of Doctoral Students 2010 – Proceedings of Contributed Papers*: 31–36.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70(1), 41–55.
Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics* 36(2), 293–298.
Russell, S., & Norvig, P. (2010). *Artificial intelligence a modern approach.* Third Edition. Pearson.
Schuler, D., Shi, W., Hoskisson, R., & Chen, T. (2017). Windfalls of emperors' sojourns: Stock market reactions to Chinese firms hosting high-ranking government officials. *Strategic Management Journal,* 38(8), 1668-1687.
Schuster, T., Lowe, W. K., & Platt, R. W. (2016). Propensity score model overfitting led to inflated variance of estimated odds ratios. *Journal of Clinical Epidemiology,* 80, 97–106.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiological Drug Safety*, 17(6), 546-555.

Stuart, E. A, & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.) *Best Practices in Quantitative Social Science* (pp. 155-176).

Thatchenkery, S., & Katila, R. (2021). Seeing what others miss: A competition network lens on product innovation. *Organization Science*, 32(5), 1149-1390.

Thatchenkery, S., & Katila, R. (2023). Innovation and profitability following antitrust intervention against a dominant platform: The wild, wild west?. *Strategic Management Journal*, 44(4), 943-976.

Trajtenberg, M., Henderson, R., & Jaffe, A., (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology,* 5(1), 19–50.

U.S. Congress. (1977). *Federal grant and cooperative agreement act*. Public Law 95-224 Retrieved from: https://www.govinfo.gov/content/pkg/STATUTE-92/pdf/STATUTE-92-Pg3.pdf.

U.S. Department of Energy. (2019). *Advanced fossil energy technology research phase I*. Retrieved from: https://www.sbir.gov/node/1308631.

U.S. Department of Labor. (2010). *Veterans' employment & training service annual report to congress*. Retrieved from: https://www.dol.gov/sites/dolgov/files/VETS/legacy/files/USERRA-Annual-FY2010.pdf.

U.S. Navy Small Business Innovation Research (SBIR) / Small Business Technology Transfer (STTR). (2019). *High power compact fuel cell system navy SBIR NX191 - topic NX19-006*. Retrieved from: https://www.navysbir.com/nx19_1/nx19-006.htm.

UCLA. (2012). FAQ : What are pseudo R-squareds ? *Institute for Digital Research and Education*. Available at: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/.

Vanneste, B., & Gulati, R. (2022). Generalized trust, external sourcing, and firm performance in economic downturns. *Organization Science,* 33(4), 1251-1699.

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

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society,* 67(2), 301–320.

FIGURES

FIGURE 1 Covariate-weight graph (illustration tool)

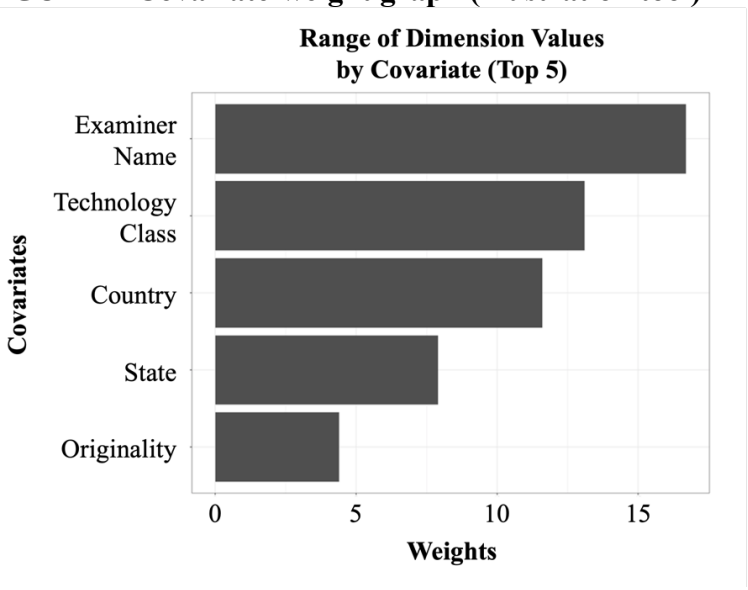


FIGURE 2 Machine learning-propensity score matching calculated propensity score balance plots, pre- vs post-match

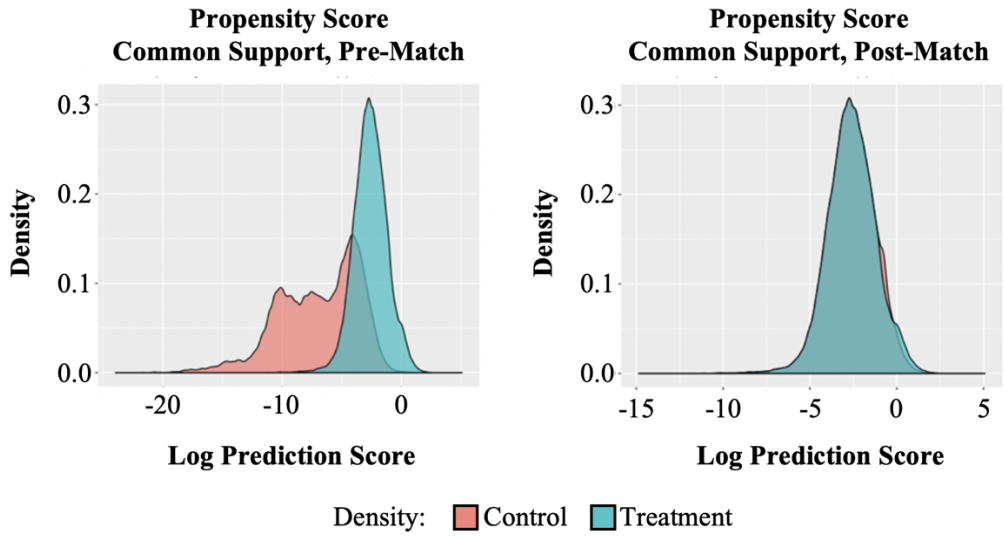


FIGURE 3 Balance plots, pre-match vs post-match

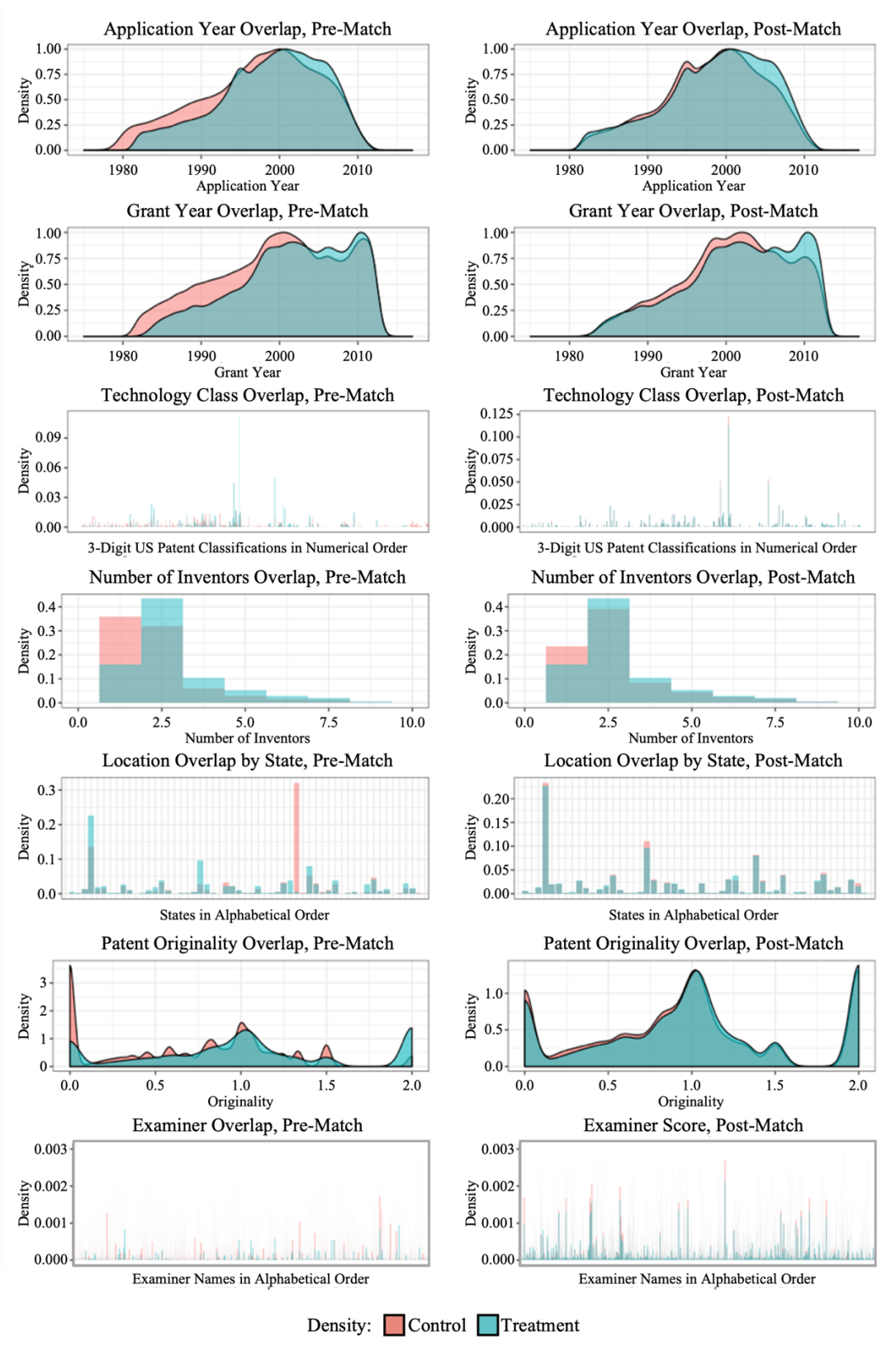
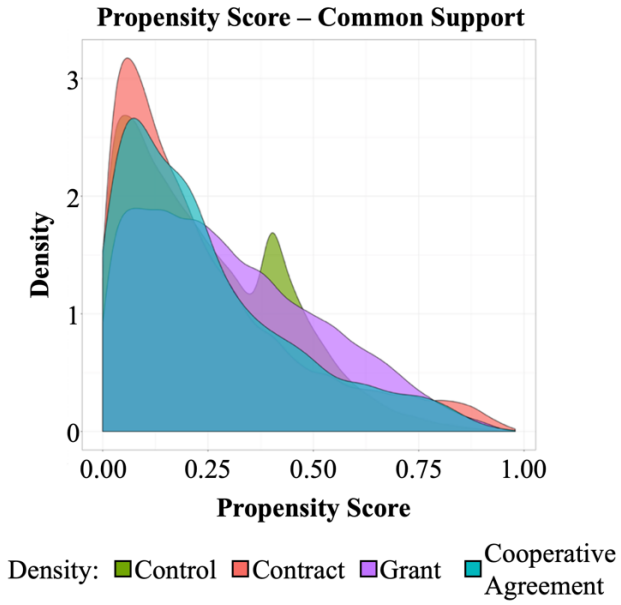


FIGURE 4 Propensity score balance across three treatments



TABLES

TABLE 1 Correlations

	Variables	Mean	S.D.	1	2	3	4
1	Invention disruption	0.36	0.78				
2	Public-private relationship	0.02	0.14	0.01			
3	Patent originality	0.65	0.39	0.03	0.02		
4	Application year	2000.5	11.02	-0.06	0.00	0.31	
5	Number of inventors	2.37	1.75	0.04	0.03	0.08	0.18

TABLE 2 Examiner name weights

Examiner ID	Examiner name	Covariate-dimension weight
A	Zarfas, S	-13.11
B	Ansher, B	-12.89
C	Douglas, P	-12.71
D	Knife, M	0
E	Renner, A	0
F	Chan, F	0
G	Goddard, B	6.3
H	Grarsay, T	6.45
I	Ryam, P	6.49

Note: Nine examiners who either were more likely to examine corporate-only patents (A-C), publicly-funded patents (G-I) or neither (D-F).

TABLE 3 Predictive performance (AUC) vs. fit (R^2) predicting public funding in the first stage

		Training data		Test data	
Model	**First stage method**	**AUC**	**R^2**	**AUC**	**R^2**
1	Propensity score matching - Machine learning with the traditional linear covariate set (technology class, application year, grant year)	0.768	0.000	0.768	0.000
2	Propensity score matching - Machine learning with the traditional covariate set from #1 and additional covariates (patent inventors, geographic location, originality, patent examiners)	0.896	0.003	0.886	0.003
3	Propensity score matching - Machine learning with the traditional and additional covariates from #2 and quadratic interactions, no regularization	0.939	0.058	0.894	0.038
4	Propensity score matching - Machine learning with the traditional and additional covariates from #2 and quadratic interactions, with regularization	0.912	0.058	**0.896**	0.038
5	Propensity score matching - Machine learning with the traditional and additional covariates from #2, quadratic and cubic interactions, with regularization	0.955	0.084	0.881	0.048

TABLE 4 OLS regression predicting invention disruption with different matching methods[1]

	Unmatched		Exact matching		PSM matching		ML-PSM matching	
	1a	**1b**	**2a**	**2b**	**3a**	**3b**	**4a**	**4b**
Public-private relationship	-0.01 (.01)		0.03 (.00)		0.03 (.00)		0.02 (.00)	
Contract		0.001 (.84)		0.02 (.00)		0.02 (.01)		0.02 (.01)
Grant		-0.02 (.00)		0.02 (.06)		0.01 (.12)		0.02 (.01)
Cooperative agreement		-0.10 (.00)		-0.08 (.00)		-0.08 (.00)		-0.07 (.00)
Controls								
Patent originality	0.12 (.00)	0.12 (.00)	0.12 (.00)	0.12 (.00)	0.13 (.00)	0.13 (.00)	0.13 (.00)	0.13 (.00)
Previous relationships (log)	0.02 (.00)	0.02 (.00)	0.01 (.00)	0.01 (.00)	0.01 (.00)	0.01 (.00)	0.01 (.00)	0.01 (.00)
Patent age	0.01 (.00)	0.01 (.00)	0.04 (.00)	0.04 (.00)	0.03 (.00)	0.03 (.00)	0.02 (.04)	0.02 (.04)
Time to grant	0.02 (.00)	0.02 (.00)	0.05 (.00)	0.05 (.00)	0.05 (.00)	0.05 (.00)	0.04 (.00)	0.04 (.00)
Number of inventors	-0.01 (.00)	-0.01 (.00)	-0.01 (.00)	-0.01 (.00)	-0.01 (.00)	-0.01 (.00)	-0.01 (.00)	-0.01 (.00)
R^2	0.10	0.10	0.09	0.09	0.09	0.09	0.06	0.06
Adjusted R^2	0.10	0.10	0.09	0.09	0.09	0.08	0.06	0.06
[1]First stage matching on			Technology class + application year + grant year		Technology class + application year + grant year		Technology class + application year + grant year + state + country + examiner + num inventors + all quadratic pairwise interactions	

Note: p=values in parentheses (two-tailed tests). N=3.2M (model 1), N=0.1M (models 2, 3, 4). Second stage fixed effects for application year, geographic location, technology class included in all models. First stage matching DV is public-private relationship for all models.

TABLE 5 Propensity score matching–machine learning roadmap

Step	Activities
1	Select covariates
2	Plot covariate distributions
3	Build a propensity score model
3.1	**Split the data into training, validation, and test sets**
3.2	**Select a propensity score matching-machine learning model (e.g., logistic regression)**
3.3	**Train & validate model using regularization methods**
3.4	**Evaluate model performance on test data**
3.5	**Repeat steps 3.2-3.4 with additional covariates/interactions**
3.6	**Select the best propensity score model**
4	Use propensity score model to generate propensity scores
5	Match treatment and control observations which have similar scores
6	Plot covariate distributions, post-match
7	Run second-stage regression

Note: The unique machine learning approach is bolded in steps 3.1-3.6. The other, un-bolded steps represent the conventional propensity score matching approach.

TABLE 6 Comparison of traditional vs ML-assisted regressions

	Causal inference models (2-stage regression models)	
Key elements	**Traditional 2-stage**	**Supervised ML-assisted 2-stage**
Objective	Robust causal inference	Robust causal inference
Typical application areas	# Variables << # observations	# Variables >> # observations; variable interdependencies
Defenses against overfitting (specialization to a particular circumstance)	Comparison across different datasets	Cross-validation, comparison of R^2 and AUC, regularization
Model selection	Researcher intuition	Combination of optimization and researcher intuition (determining the variables to be constructed, and forcing inclusion of certain variables in the model)
Variables and functional forms	Few, mostly linear	Many, with complex interdependencies

APPENDIX

Regularization and cross-validation details. In its simplest form, supervised learning estimates a predictive model's coefficient weights with a maximum likelihood estimator: $L \equiv \sum_i^n (Y_i - H_\theta(X_i))^2$ where $H_\theta(X_i)$ is the predictive model, Y_i is the outcome variable, X_i is the set of observed covariates, and θ is the set of coefficient weights. To build a model, the optimizing solver iteratively updates estimates of θ to drive likelihood (L) to zero. For reference, this "base" likelihood estimator is equivalent to traditional econometric approaches (Rosenbaum & Rubin, 1983). The difference is the regularization term. This term prevents overfitting by penalizing the estimator if the estimator tries to weight particular coefficients too heavily.

The two most common forms of regularization applicable for propensity scores are Lasso (L1) and Ridge (L2). L1 regularization penalizes the estimator by driving the weights of dimension coefficients which do not contribute to selection to zero. Using L1 regularization, overfitting is controlled for by effectively deleting those observational factors which are *not confounding* (i.e., not predictive). The L1 penalized estimator takes the form of: $L \equiv \sum_{i=1}^n \left(Y_i - H_\theta(X_i)\right)^2 + \lambda_1 \sum_{i=1}^n |\theta_i|$ The L2 regularization normalizes coefficient weights to ensure that no singular dimension dominates the model. Thus, L2 regularization forces coefficient weights to be small, but keeps all information in the model. While it therefore minimizes the impact of any single dimension, it remains prone to including non-confounding dimensions. The L2 penalized estimator takes the form $L \equiv \sum_{i=1}^n \left(Y_i - H_\theta(X_i)\right)^2 + \lambda_2 \sum_{i=1}^n \theta_i^2$ Finally, the combination of both L1 and L2, i.e., *elastic net regularization*, is particularly useful for matching (Athey, Imbens, & Wagner, 2018; Zou & Hastie, 2005) computed as $L \equiv \sum_{i=1}^n \left(Y_i - H_\theta(X_i)\right)^2 + \lambda_1 \sum_{i=1}^n |\theta_i| + \lambda_2 \sum_{i=1}^n \theta_i^2$ Elastic net regularization not only handles overfitting but also decreases curse of dimensionality problems by algorithmically removing covariates and, therefore, limiting the number of dimensions. As the number of dimensions approaches, or surpasses, the number of observations, L1 regularization begins randomly deleting dimensions with equivalent covariance (Zou & Hastie, 2005). While this certainly minimizes overfitting, it is also likely to remove confounding dimensions, violating the exogeneity assumption. Elastic-net regularization combines L1 and L2 such that L2 can re-weight dimensions before L1 sets dimension coefficient weights to zero, minimizing the likelihood of equivalent covariance. As a second stage, then, L1 can more accurately remove non-confounding dimensions. In combination, elastic net regularization

quantitatively and algorithmically determines which covariates are confounding covariates and thus helps avoid overfitting and curse of dimensionality problems.

In this step, we include the regularization terms L1, $\lambda_1 \sum_{i=1}^{n} |\theta_i|$, and L2, $\lambda_2 \sum_{i=1}^{n} \theta_i^2$. These terms remove unconfounding covariates by penalizing their coefficients, θ, if the logistic model begins to overfit. Thus, our likelihood estimation function is: $L \equiv \sum_{i=1}^{n} \left(Y_i - H_\theta(X_i)\right)^2 + \lambda_1 \sum_{i=1}^{n} |\theta_i| + \lambda_2 \sum_{i=1}^{n} \theta_i^2$ where $H_\theta(X_i) = \frac{1}{1+e^{-(\theta_0 + \theta_1 X_i + \cdots + \theta_N X_N)}}$. With this function, we estimate the covariate coefficients (θ). First, λ_1 & λ_2 are initialized to zero, and the training data is used to generate predictions of θ. Next, the validation set is used to compare model performance. Iteratively, the optimizer first searches the set of θ that minimizes the training error (regularization), and then for the set λ_1 & λ_2 that minimize the validation error (cross-validation) (Zou & Hastie, 2005, p. 310). Once $\theta, \lambda_1, \&\ \lambda_2$ are determined, supervised learning is complete. In practice, there are a wide variety of available tools and statistical learning packages that make supervised learning relatively straight forward including Scikit-Learn (for python), Caret (for R), and Tensorflow. We used Vowpal Wabbit,[19] which takes advantage of AdaGRAD to minimize the loss function.

[19] Vowpal Wabbit is a fast, online learning code by Microsoft Research Group and (previously) Yahoo! Research. https://github.com/JohnLangford/vowpal_wabbit/wiki.