Testing Causal Theories with Learned Proxies*

Dean Knox† Christopher Lucas‡ Wendy K. Tam Cho§

January 19, 2022

Abstract

Social scientists commonly use computational models to estimate proxies of unobserved concepts, then incorporate these into subsequent tests of their theories. The consequences of this practice, which comprises over two-thirds of recent computational work in political science, are underappreciated. Imperfect proxies can reflect noise and contamination from other concepts, producing biased point estimates and standard errors. We demonstrate how analysts can use causal diagrams to articulate theoretical concepts and their relationships to learned proxies, then apply straightforward rules to assess which conclusions are rigorously supportable. We formalize and extend common heuristics for "signing the bias"—a technique for reasoning about unobserved confounding—to scenarios with imperfect proxies. Using these tools, we demonstrate how in often-encountered research settings, proxy-based analyses allow for valid tests for the existence and direction of theorized effects. We conclude with best-practice recommendations for the rapidly growing literature using learned proxies to test causal theories.

*For excellent research assistance, we thank Gechun Lin. For insightful comments, we thank Guilherme Duarte, Lucia Motolinia, Brandon Stewart, and Dustin Tingley.

†Dean Knox is Faculty Affiliate of Analytics at Wharton and Assistant Professor in the Operations, Information, and Decisions Department, the Wharton School of the University of Pennsylvania (dcknox@upenn.edu).

‡Christopher Lucas is an Assistant Professor in the Department of Political Science and a Faculty Affiliate with the Division of Computational and Data Sciences at Washington University in St. Louis (christopher.lucas@wustl.edu).

§Wendy K. Tam Cho is Professor in the Departments of Political Science, Statistics, Mathematics, Computer Science, and Asian American Studies, the College of Law, and the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign (wendycho@illinois.edu).

1 I Don't Know Y (and Other Challenges Arising from Imperfect Proxies in Social Science)

Social scientific theories often involve latent concepts that are not directly observed by researchers, such as "democracy" or "ideology." To empirically evaluate their theories, researchers must imperfectly measure these unobserved concepts. Classic examples include the use of expert panels to rate countries' political systems and factor analysis to construct weighted indices from survey responses, which respectively produce *proxies* of democracy and ideology. While various forms of measurement date to the advent of quantitative social science, the recent growth of machine learning and computation has led to an explosion of work that constructs learned proxies. Compared to classic approaches—which can require costly in-depth expert reading or derivation of case-specific measurement models—this new body of work increasingly uses rich, unstructured data and flexible, off-the-shelf statistical tools to measure concepts of theoretical importance. In this article, we review common approaches and key methodological considerations in this rapidly growing literature. In particular, we focus on best practices for incorporating imperfectly learned proxies into subsequent analyses, which pose underappreciated challenges for analysts seeking to rigorously test social scientific theories. Excellent references are available for measurement [Adcock and Collier, 2001] and statistical learning [Grimmer et al., 2021] more broadly. In contrast with these and similar review articles, we focus specifically on the use of learned proxies in causal tests, especially when the proxy is estimated from a computational model.

Regardless of how formally they are expressed, social scientific theories are precisely articulated, falsifiable statements about the causal structure of the world. In political science, the greatest impact of recent computational advances has been to improve the researchers' ability to test such theories. In a review of papers in the *American Journal of Political Science*, the *American Political Science Review*, and the *Journal of Politics* from 2018 to 2020, we identified 48 papers that employed statistical learning or other

computational methods in one fashion or another.[1] The vast majority of this work—over two-thirds—seeks to estimate a proxy for a concept in a causal theory that is not directly observable. Without this proxy, no empirical evaluation of the theory is possible.

While the use of proxies in social science is not new, our literature review highlights how computational methods have drastically increased their accessibility. For decades, the development of new proxies was a major effort, feasible only for well-funded research teams, that often attempted to produce a new measure shared across research teams. For example, an enormous literature theorizes the effects of democratic institutions on a host of outcomes ranging from economic development to life expectancy. However, because "democracy" is not observable directly, any empirical test of these theoretical predictions must rely on a proxy. Out of this necessity arose several costly efforts utilizing large groups of expert coders, which have seen close scrutiny and widespread use.[2] Similarly, to empirically test numerous theories about the origins and effects of legislator ideology, researchers commonly rely on a publicly available measure that was built from carefully derived statistical models based on application-specific functional form assumptions about ideology and voting [i.e., NOMINATE; Poole and Rosenthal, 1985]. With the exception of multidimensional scaling methods for survey data and votes [Poole, 2008], case-specific measurement models were, until recently, limited.

In a noteworthy paradigm shift, researchers now regularly estimate new proxies for individual studies, often from high-dimensional data for which traditional methods are inappropriate. At the same time, implementing computationally intensive parametric models has become considerably easier with the advent of languages like Stan [Carpenter et al., 2017] and the vast computational power now available to researchers. Advances in statistical learning now allow researchers to flexibly estimate proxies without application-specific knowledge, using increasingly rich data sources and generic statistical models that adapt to the data at hand.

Despite these technological advances, there remain a number of fundamental research-

[1] Appendix Section A describes our coding scheme employed, as well as the identified articles.

[2] For example, see Munck and Verkuilen [2002] for an evaluation of various measures of democracy including Polity [Gurr, 1974], Freedom House [2014], and others.

design considerations that receive little attention when conducting analyses with learned proxies. Throughout this paper, we use the term "proxy," as opposed to "measure," to emphasize that many such variables are substitutes that imperfectly measure the underlying theoretical concept. At a high level, this slippage can stem from three sources: (1) measures often fail to fully capture all aspects of the underlying concept, (2) they often contain some level of purely random noise, and (3) they are often systematically contaminated by other factors besides the concept of interest. While an extensive literature on measurement has focused on improving validity by eliminating these sources of error, resource-constrained researchers often do not have the luxury of perfecting their proxy variables, particularly when measurement modeling constitutes just one of many stages in the research process. Determining how to proceed in the face of these inevitable imperfections—the focus of this article—is therefore an important methodological question that confronts many applied researchers.

In the remainder of this paper, we explain how these issues can bias both treatment effect estimates and standard errors. We then illustrate how scholars can use causal diagrams to reason about various sources of error and their implications in terms of statistical biases. It is well known that these diagrams constitute an easy-to-use tool for conveying the essence of social scientific theories. What is less appreciated is that causal diagrams are also useful for concisely expressing the assumed quality of proxies used to approximate an underlying true concept, as well as for indicating potential sources of contamination. By writing down concrete assumptions in this easily digestible form, analysts can then apply well-established rules to determine which conclusions can be rigorously supported—all while avoiding implausible parametric assumptions about functional form and the distribution of random errors. Without such parametric assumptions, which are generally difficult to defend, analysts generally cannot recover accurate quantitative estimates of theorized effect sizes. However, we show that in many common research settings, analysts *can* reliably evaluate the qualitative existence of these effects and determine their direction. That is, despite measurement error and possible systematic contamination by other factors, analysts can nonetheless rigorously assess whether treatment variables causally

lead to the theorized increase or decrease in outcome variables. We provide numerous examples of research settings with proxied treatments, outcomes, and confounders in which such conclusions can be supported, along with straightforward procedures that analysts can use when confronted with more complex scenarios.

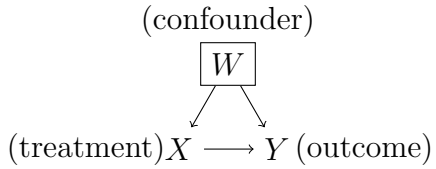
2 Integrating Machine-learning Techniques with Social-scientific Theory

Rigorous social-scientific theories are statements about the *causal structure* of the world [Pearl and Mackenzie, 2018]. That is, they assert that a dependent variable, Y, would or would not have unfolded differently if an independent variable, X, had been hypothetically modified. Such theories are distinct from empirical predictions that X will be *associated* with Y, in that they posit an explanation for *why* empirical associations appear: for example, because X has a direct effect on Y; because it has an indirect effect through some intermediate factors; or because X and Y are both influenced by some common cause that produces a spurious correlation.

Well-articulated theories are collections of statements about (1) the set of factors that are theoretically meaningful and (2) how these factors might influence one another. These statements can be concisely expressed in the form of a *causal diagram* depicting each factor, with arrows representing influence relationships; a generic example is given in Figure 1. We note that what causal diagrams do not convey is perhaps as important as what they do. Critically, causal diagrams do not make implausible claims about precisely *how* X affects Y. For example, they do not state that "the effect of increasing X by 1 unit is that Y will increase by an average of 2.5 units" or that "X_1 and X_2 have linear effects on Y and do not interact." In complex social-scientific settings, analysts rarely have enough knowledge to theorize such rigid and specific functional forms. Instead, these relationships must be flexibly estimated from data.

Causal diagrams have proven invaluable to the social sciences, guiding both qualitative process tracing [Waldner, 2015] and quantitative analyses [Keele et al., 2020] when

Figure 1: **Theorized causal structure.** A causal theory in which treatment X has an effect on outcome Y, but estimation is complicated by a common cause W that must be adjusted for (indicated with a rectangle) to recover $X \to Y$. Subsequent figures will consider scenarios in which X, Y, or Z cannot be directly observed and must instead be noisily measured.



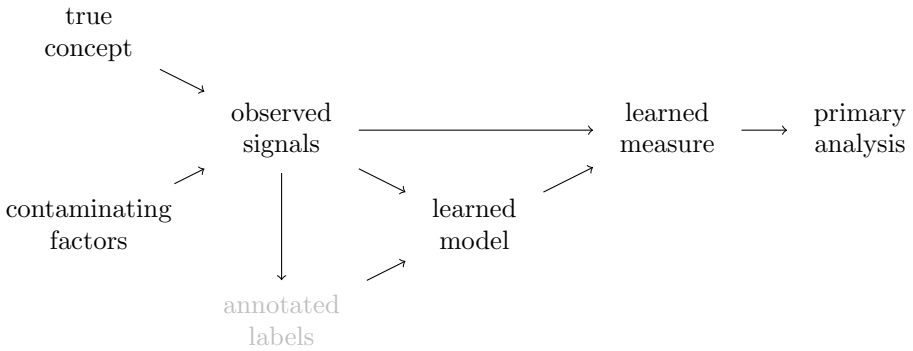
evaluating social-scientific theories. Classic references such as Pearl [2009] offer clear-cut guidelines for diagrammatically assessing alternative explanations that must be ruled out before analysts can draw firm conclusions. As a simple example, in the scenario of Figure 1, it can be seen that analysts must account for the common cause (or *confounder*, W), before estimating the theorized effect, $X \to Y$.[3] Without adjusting, analysts cannot rule out the possibility that observed associations between X and Y might be due to confounding and the theorized causal effect might be nonexistent. In this paper, we consider the complex issues that arise when analysts seek to evaluate their theories using indirect measures of key theoretical constructs—an increasingly common practice in social-scientific research that uses rich, newly available data to proxy for concepts that were previously difficult to operationalize. We outline the types of conclusions that can and cannot be rigorously supported when using learned proxies in a number of common research settings, as well as a set of rules to help guide analysts confronted with more complex scenarios.

The fundamental problem that proxy-based research seeks to address is that theoretical concepts in the social sciences are often abstract and lack precision [Weber, 2017]. Consider the ideological bias, or slant, of media outlets. A staggering volume of research examines the origins of media bias, as well as its effects on subsequent social phenomena [Puglisi and Snyder, 2015]. At the time of writing, searching for "media bias" on Google

[3] In potential outcomes notation, we have $X = X(W)$ and $Y = Y(W, X)$; the causal quantities of interest are taken to be various aggregations of or contrasts between the conditional average treatment effects, $\mathbb{E}[Y(x', w) - Y(x, w)|W = w]$.

Scholar yielded over 26,000 search results. Yet, media bias is not directly observable under any study design. Rather, this underlying *true concept* generates noisy and imperfect *observed signals* according to some generally unknown process. These imperfect signals may include (1) which politicians a given newspaper chooses to endorse [Ansolabehere et al., 2006]; (2) the textual similarity between language used by media outlets and members of congress [Gentzkow and Shapiro, 2010]; or (3) the way an outlet covers certain issues [Larcinese et al., 2011]. The absence of direct, high-quality data on the underlying concept greatly complicates the task of evaluating theories of media bias.

Figure 2: **Overview of computational measurement.** Each observation is associated with a specific value for the true concept of interest, which may be a confounder, a treatment, or an outcome. This attribute cannot in general be observed directly, but auxiliary information provides some signal about its value. However, these signals may be contaminated by additional factors; for example, if the attribute being measured is the treatment, the observed signals may contain not only treatment-related information, but also contamination from the confounder. The attribute of interest may be annotated for a subset of units (indicated with gray text), on the basis of observed signals. Annotations may contain errors or perfectly correspond to the true concept; if they contain errors, these errors may be independent or influenced by contaminating factors. After annotations are obtained (not obtained), supervised (unsupervised) machine-learning models are trained—either on the observed signals directly or, more commonly, on a reduced representation that may result in the loss of information. The learned model is applied to observed signals for all units. The resulting estimates constitute the learned measure, which is then incorporated into a primary analysis.



How do researchers use computational methods to address these challenges? Figure 2 graphically depicts the typical workflow. As already noted, analysts have access to *observed signals* that convey noisy information about the concept via some process that is generally unknown. These signals potentially capture not only the true concept, but also

other *contaminating factors* discussed in Section 2.2. To map these signals back to the true concept of interest, researchers typically convert them into a reduced format that is amenable to analysis, then apply an assumed *measurement model* to obtain a predicted value of the true concept. It is critical to recognize that the model used for measurement is, at best, a simplified representation of the unknown process by which the true concept manifests in observed signals. Moreover, contamination of the signals used to proxy the true concept can lead to systematic errors that must be carefully considered when seeking to draw conclusions. The task of constructing and validating measurement models, including with machine learning methods, has been the subject of much work [Adcock and Collier, 2001, Grimmer and Stewart, 2013]. We briefly review this extensive literature before turning to the question of how measures should be used in subsequent, theoretically motivated analyses—a key component of the social science workflow that has received far less attention.

2.1 Challenges with Computational Measures of Latent Variables

Measurement models are rich and varied, ranging from panels of human experts to keyword-based binary classification rules and trained neural networks. Here, we illustrate these and other choices confronting a researcher when developing a computational proxy, using Martin and McCrain [2019] as a running example. Martin and McCrain [2019] studies the effect of a sudden and widespread shift in media ownership, which we denote as X, in which the conservative Sinclair conglomerate acquired numerous media outlets in the United States. In this case, media consolidation was theorized to affect the unobserved concept of media slant, Y. Martin and McCrain [2019] uses the measurement model of Gentzkow and Shapiro [2010], proxying media bias based on the similarity between (1) the text of each media outlet's news and (2) the text of partisan speeches in the Congressional Record.[4] We will refer to the resulting predictions for each unit as the

[4]This model can variously be thought of as (1) a weighted, rescaled dictionary or (2) as an instance of a supervised model trained on the Congressional Record and transferred to the domain of news.

measure, $\hat{Y}$. We emphasize that the observed signal (which is often a rich information source, such as a television station's audiovisual stream) is conceptually distinct from the inputs to the measurement model (which can be lossy reductions, such as counts of various words obtained by a imperfect transcription).[5]

How might this proxy—textual similarity with the text of partisan speeches in the Congressional Record—differ from media slant, the unobserved latent concept of interest? For example, imagine that a researcher hopes to measure the slant of news articles covering *local* policy, which is generally not discussed in congressional speeches but still plausibly contains partisan bias. In this case, the similarity of these articles to the Congressional Record is not necessarily a good measurement model; its use requires researchers to assume that a model based on partisan speeches extrapolates well to topics not discussed in those speeches (local policy). If this assumption fails, relying on textual similarity between local news and Congressional speech will yield unreliable results.

Martin and McCrain [2019] address this concern by only applying the textual similarity measure to news segments covering national issues. However, if a researcher was interested in the ideological slant of local news coverage, they could alternatively rely on human annotators to inspect the observed signals (e.g., the text of news articles) and label the ideological slant of each document. In the case of labeling documents on a continuous spectrum, like ideological slant, obtaining labels with pairwise comparisons simplifies the task for coders [Carlson and Montgomery, 2017].

A benefit to human annotators is that they often have tremendous contextual knowledge, understand ambiguous instructions, and can learn a large and flexible set of measurement models. Human annotators can also be given direct access to unstructured signals, such as audiovisual recordings of a television news broadcast, in their entirety. However, a limitation is that human annotators are expensive, so it may not be feasible to annotate every observation in the data. Moreover, even when annotations are available, humans are well known to exhibit prejudices and cognitive limits, meaning that annotated

[5]The practice of manually specifying informative inputs is referred to as *feature engineering* and can include stemming/lemmatizing of words, extraction of n-grams, and computation of interactions or other higher-order terms.

labels do not always reflect the underlying true concept.[6] In some cases, key concepts may be difficult to precisely quantify even for experienced subject-matter experts, let alone the low-cost annotators that are often used for this task. *Annotation errors*, or deviations between truth and human labels, contribute to *measurement error*—a broader concept that refers to any deviation between the true concept and a proxy (including machine predictions that may rely in part on human labels). Importantly, these errors exist at a conceptual level even when the underlying truth is unknown for any observation: as long as the true construct exists as theorized, then proxies must either deviate or not deviate from the underlying, unknown value, even though this deviation is not directly calculable. These deviations may either be purely random (e.g., accidental mislabeling) or systematic (e.g., higher slant scores for articles that are in ideological disagreement with the annotator). We return to this issue of proxy quality issue later in this article, as it can introduce confounding and other statistical biases in subsequent analyses.

Setting aside challenges inherent in human coding, at a high level, *supervised learning* refers to the general approach of obtaining small to moderate amounts of annotation, then training a model that attempts to reconstruct the resulting labels on the basis of some reduced feature set, such as word frequencies [Grimmer and Stewart, 2013]. The resulting learned model can then be cheaply applied to millions of unlabeled articles to obtain learned measures. Here, annotation errors can lead the model astray, but they are not the only problem: small training sample sizes or incomplete feature sets represent other sources of measurement error. However, annotation is not always needed, as indicated by the gray coloring of this step in the research workflow depicted by Figure 2. In contrast, *unsupervised learning* approaches attempt to identify latent clusters or dimensions that explain patterns in the observed signal without the need for human review. For instance, Poole and Rosenthal [1985] scales legislators according to voting patterns. Similarly, Slapin and Proksch [2008] scale documents according to word frequencies, based solely on co-occurrence patterns; these measurement models do not use human annotations

[6]For example, a media-slant researcher might worry that a human annotators cannot reliably score bias for individual news articles on a numeric scale without additional points of reference. In this case, researchers could ask annotators to conduct pairwise comparisons, then apply an appropriate machine-learning measurement model to obtain numeric scaling estimates [Carlson and Montgomery, 2017].

to guide the process. Numerous variations (e.g., active learning, transfer learning) and hybrid approaches (e.g., semi-supervised learning, zero-shot learning) exist.

These computational methods represent powerful tools for mapping imperfect, messy, and high-dimensional signals about an unobserved theoretical concept to low-dimensional measures that can be used in statistical analyses. The tradeoffs are well documented: among other issues, such methods typically require moderate to large quantities of data to learn patterns without contextual knowledge; can overfit to limited data and memorize noise rather than learning generalizable patterns; and can learn only from the reduced space of features provided by analysts, which are typically more limited than the information available to human annotators. A number of approaches have been developed to help address these obstacles to supervised learning, including cross-validation, transfer learning, and novel architectures that can ingest complex data. A full examination of these techniques is beyond the scope of this paper. For a thorough review, including machine-learning applications beyond those considered here, see Grimmer et al. [2021].

2.2 Using Computational Measures in Subsequent Analyses Will Bias Causal Estimates, but Not All Is Lost

Much of the prior literature on measurement focuses on improving the validity of the measure itself—that is, eliminating measurement error, especially from systematic sources [Adcock and Collier, 2001]. Yet while measuring concepts has intrinsic value, we find that a far larger body of work is devoted to the next step of the scientific process: analyzing the origins and effects of the measured concept to improve our understanding of its broader social context. In our review of machine learning applications in *APSR*, *AJPS*, and *JOP*, we found 48 papers that employed machine learning in one fashion or another. Within this set, we identified two types of papers that estimate a proxy for use in an empirical test of a causal theory. The first of these two types (26 papers) makes a primarily substantive contribution by developing a causal theory, then estimates a proxy variable in order to empirically test the theory. The second set (7 papers) makes primarily a methodological contribution, focusing directly on the estimation and validation of a novel proxy variable

for use in empirical tests of a range of causal theories.

These papers all confront a shared obstacle: how do we test theories that involve variables that we cannot directly observe? To do so, analysts employ a measurement model to create the learned proxy, which is then incorporated into a *primary analysis*. For example, Martin and McCrain [2019] conducts a regression of a proxy media-bias outcome, $\hat{Y}$, on the theorized cause of media consolidation, X, as well as other confounders, W. More generally, any theory that includes variables which are not directly observable is untestable without a learned proxy. In every other sense, proxy-dependent analyses are unremarkable, often employing common research designs intended to address classic threats to causal inference like unobserved confounding. For example, Martin and McCrain [2019] leverage a differences-in-differences design that compares acquired stations to other stations in the same market.

But while proxy-dependent designs often take seriously inferential threats like confounding, they commonly ignore challenges that arise from the use of a proxy variable. For example, random and systematic measurement error in a proxy can induce additional statistical biases in this primary analyses, with consequences that vary substantially depending on the quantity proxied and the precise nature of the error. But researchers often do not have the luxury of perfecting the measurement process, which can demand considerable time, personnel, and research funds. In many cases, noisy or contaminated observable signals can make it entirely impossible to obtain ideal measures of key theorized concepts. How can research proceed in the face of this challenge? We now illustrate how to reason about limitations of learned measures and how these limitations relate to the theorized causal structure. In Section 3, we then examine a number of common research settings and show that despite the statistical biases induced by imperfectly learned proxies, it nonetheless remains possible to draw meaningful conclusions about the theorized causal process.

Specifically, we focus on lesser-known implications of measurement error and what researchers can credibly conclude in the presence of bias that results from this error. Specifically, we now review how social scientists can draw conclusions about the existence

and direction of a causal effect, even when the point estimate of that effect is almost certainly biased. This result should be encouraging, because social scientists are generally most interested in demonstrating the existence of theorized effects rather than precisely quantifying the exact effect size. This goal is a particular form of *causal discovery*, a branch of causal inference which attempts to learn causal diagrams; in this context, researchers focus specifically on "discovering" a single theorized $X \rightarrow Y$ relationship, rather than considering all possible influence relationships. To the extent that a theorized effect is discovered, researchers may then seek to evaluate whether its direction, or sign, accords with theoretical expectations. (In Section 3, we formalize terminology for various senses in which effects can be described as "positive" or "negative.") This research objective is distinct from the goal of *causal estimation*—which seeks to make precise quantitative statements about the magnitude of effects—which appears in literatures such as voter turnout and incumbency advantage, where the presence of an effect is already established with high confidence. As we discuss in this paper, causal estimation is difficult when using learned proxies to approximate key unobserved steps in the theorized causal process.[7] In contrast, discovery and signing of an $X \rightarrow Y$ effect can be conducted under generally weaker assumptions about the types of contamination affecting a proxy.[8]

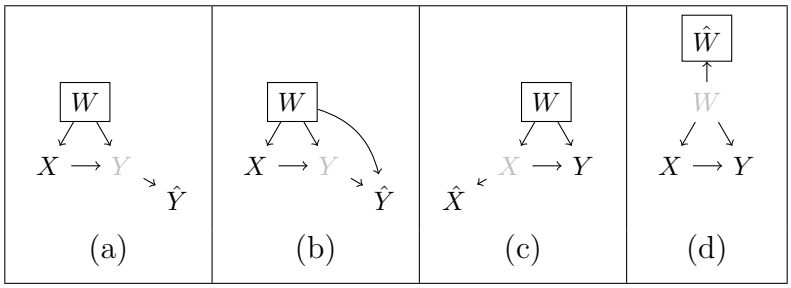
Figure 3(a) depicts one possible causal structure representing not only the theorized concepts, but also a measurement process (in this case for the outcome Y). The $Y \rightarrow \hat{Y}$ arrow indicates a causal process by which the true outcome Y leads to the proxy $\hat{Y}$, compactly summarizing the entirety of the measurement workflow: (1) the generation of observed signals; (2) annotation of labeled units, if any; (3) training of the model; and (4) prediction of learned measures. It does so without expressing untenable parametric assumptions. As a case in point, the $Y \rightarrow \hat{Y}$ arrow does not state that measures are centered on the true concept, i.e. satisfy $\mathbb{E}[\hat{Y} - Y] = 0$. As the media slant illustration makes clear, such assumptions are often facially implausible. However, Figure 3(a) does

[7] Except in very specific cases that can be sensitive to violations of unverifiable assumptions about, for example, the functional form of the outcome.

[8] We note that causal discovery can be regarded as a precursor to causal estimation: effects discovered with noisy proxies can highlight areas where improved measurement is necessary to obtain precise estimates.

encode structural assumptions in the *absence* of arrows from W or X to $\hat{Y}$, which state that the learned measure is *uncontaminated*—that is, free of influence from these factors, meaning that $\mathbb{E}[\hat{Y} - Y|W = w, X = x]$ is constant across all w and x.

Figure 3: **Causal structures of theory and measurement.** Data environments corresponding to the theory of Figure 1. Panels (a) and (b) illustrate settings in which analysts are unable to directly observe the outcome Y, and thus must resort to a learned measures $\hat{Y}$ that is either uncontaminated (a) or contaminated by a confounder (b). Panels (c) and (d) respectively depict cases in which the treatment X or the confounder W cannot be observed, so that analysts can only adjust for learned proxies ($\hat{X}$ or $\hat{W}$).



This is a difficult requirement to satisfy; in many settings, analysts will be unable to defend the assumption that a proxy is uncontaminated. The main way to ensure it holds is to verify that the observed signal does not convey additional information about factors other than the true concept of interest. For example, in the media bias setting, contamination would occur if nonpartisan issues of local interest tend to be discussed both in rural news and by legislators representing rural districts. In this case, the confounder of rural-urban status would contaminate the media bias measure. In other words, rural-urban status (W) might distort the measure of media bias ($\hat{Y}$), above and beyond any influence that it might have on the true concept (Y). If this were true, Figure 3(a) would not be an accurate representation of the theory and measurement structure; Figure 3(b), in which W has a direct arrow to $\hat{Y}$, would be the correct representation. In other contexts, researchers may encounter scenarios where learned proxies must be used for the treatment of interest (X) or for key confounders (W); Figure 3(c–d) depict these in turn.

When the observed signal is rich and unstructured (for example, when they contain text, audio, or images) it can be challenging to verify that they are free of contamination. It is therefore extraordinarily difficult to guarantee that unsupervised machine-learning

methods applied to such datasets will produce uncontaminated proxies of the true concept of interest. In the supervised setting, it is in theory possible to obtain uncontaminated measures from contaminated features. When constructing a training dataset, human annotators can be instructed to set aside their cognitive biases and label each unit according to objective scoring rubrics. For example, in the media bias case, annotators could be instructed that agriculture-related news should not be used as a signal of a news outlet's Republican leanings. But even if annotators perfectly adhere to these guidelines, a measurement model that is regularized or incorrectly specified can often learn inappropriate shortcuts that reintroduce contamination, despite training on uncontaminated labels. We therefore recommend that analysts err on the side of caution. Measures should be thoroughly validated and probed for signs of contamination, e.g. by examining the predictive features used by the measurement model or by assessing whether agricultural keyword proportions continue to correlate with the media bias measure even after adjusting for obvious political keywords. However, definitive tests for contamination are often infeasible—in the above example, requiring countless model specifications and extensive keyword lists that range from "soybeans" to "pesticide." We therefore recommend that when writing down assumptions in the form of a causal diagram, scholars should err on the conservative side by drawing arrows from all possible contaminating factors to the learned measure.

Having reviewed several challenges that arise when using computational measure of a latent variable, we now explain precisely how researchers can make credible, correct claims in the presence of bias that results from these challenges. In the next section, we cover three main contexts in which a computational measure is used: when the measure is the treatment, the outcome, or a confounder. We formalize and extend the common practice colloquially referred to as "signing the bias," then use similar logic to show how valid inferences can be made about the presence of an effect even when point estimates are not point identified. By applying the rules in the subsequent section, analysts can then determine how various forms of contamination impact their ability to draw conclusions from available data.

3 Articulating Assumptions and Structure

As noted in the previous section, researchers often use a measurement model to generate learned proxies when a true theoretical concept cannot be directly observed, as with media slant. In this section, we discuss the various research-design complications that arise when testing causal theories with computational measures, depending on the theorized role of the proxied variable. We consider three cases in turn: where the learned measure proxies a treatment (Section 3.1), an outcome (Section 3.2), and finally a confounder (Section 3.3). We also consider certain settings where multiple factors are proxied simultaneously. Prominent examples of each are discussed, drawing on recent research in international relations [Carroll and Kenkel, 2019], comparative politics [Motolinia, 2021], and American politics [Nyhan et al., 2012].

3.1 Learned treatments

We first examine the case in which treatment, X, is approximated with a noisily learned proxy $\hat{X}$ (i.e., the error, $X - \hat{X}$, is nonzero). In this section, while discussing proxied treatments, we will assume that the outcome (Y) and confounders (W, representing common causes of treatment and outcome) are perfectly observed. In subsequent discussion of learned outcomes and confounders, except where otherwise noted, we will consider cases in which only one variable is proxied and that all other variables in the causal structure are observed without error. Finally, we assume that analysts either use a model which does not make rigid functional form assumptions (or, less plausibly, that analysts know the exact functional form for the primary regression).

To illustrate the task of estimating causal effects of a treatment for which only a proxy is available, we point readers to Carroll and Kenkel [2019], a recent article drawn from our review of machine learning applications. Carroll and Kenkel [2019] reexamines existing findings on the role of state power in conflict. As background, numerous theories predict that changes in a state's power will causally affect the chances of international conflict. In this study, the true treatment of interest, X, is state power—which is never directly

observed. Carroll and Kenkel [2019] use machine learning to build a proxy measure of military power that improves on prior work. Drawing on data about the capabilities and outcomes of states involved in global military disputes, Carroll and Kenkel [2019] train a measurement model based on the material capabilities of the involved states and the outcomes of conflict, then demonstrate how their approach improves over existing measures. Ultimately, the learned measure is used in the study's main objective: to revisit the findings of Reed et al. [2008] with this improved data. In this primary analysis, which uses a selection-on-observables design, Carroll and Kenkel [2019] find—contrary to existing work—that conflict is most likely when the state with the least benefits of war has a preponderance of power.

With this example in mind, we now consider the general problem of drawing conclusions from proxied treatments and review related methodological work. First, it is well-established that even when the measurement error $X - \hat{X}$ is independent noise, using a proxy $\hat{X}$ in place of X in a linear regression will result in attenuation bias [Wooldridge, 2015, Chapter 9.4]. A number of theoretical results are available for linear and other parametric errors-in-variables models; we refer interested readers to Cheng and Van Ness [1999]. And though scholars have known about bias induced by measurement error for decades, we find little mention of it in social-science applications employing proxies. (In general, the use of imperfect proxies results in skewed estimates, with certain exceptions that we discuss below.) Moreover, further complications can arise if $\hat{X}$ is contaminated by additional factors, or if errors depend on the value of X itself, as commonly occurs.

When the true values of treatment X are available for a subset of the data (e.g., when learning the proxy $\hat{X}$ with a supervised model), Fong and Tyler [2018] offer a solution in contexts where the regression is known to follow a linear functional form. Intuitively, their approach uses $\hat{X}$ as an instrument for X. Specifically, in the first stage regression, they relate X to $\hat{X}$ using the labeled data. In the second stage, where $\hat{X}$ is available but X is not, they use the full data to regress Y on $\hat{X}$. We discuss this procedure further in Section 3.3. However, a common concern is that linearity is unlikely to hold in complex social-scientific settings.

We now review how analysts can still draw principled *partial* conclusions in the face of the aforementioned issues, with weaker assumptions than those introduced by Fong and Tyler [2018]. Specifically, after accounting for confounders W, analysts can conduct *falsification tests*—a test where the null hypothesis is the absence of an effect—about the causal effect of X on Y in Figures 4(a–c). This is true even in the presence of measurement error, because under the null hypothesis (i.e., in the absence of the $X \rightarrow Y$ arrow in the causal structure), there should be no association between $\hat{X}$ and Y (after adjusting for W).

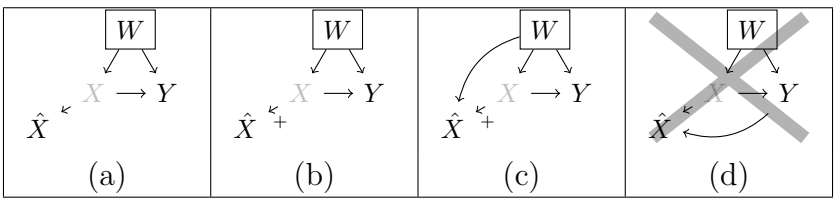
Social scientists often seek to characterize the direction of causal effects, rather than simply testing null hypotheses about their nonexistence. By building this into the statistical test, we are able to make conclusions that would otherwise not be possible. To do so, we introduce the notion of signed causal diagrams [VanderWeele and Robins, 2010], which allow formal statements about the theorized direction of the effect, rather than simply the presence of one. These signed diagrams build on the causal diagrams introduced in Section 2. In signed causal diagrams, researchers specify not just the presence of effects in their theory, but the direction of the effect. This practice is closely related to the common practice of "signing the bias," in which the researcher informally reasons through how unobserved confounding might positively or negatively skew their results.

3.1.1 An Introduction to Signed Causal Diagrams

Researchers often informally state that one variable should have a "positive" or "negative" effect on another, but the precise meaning of these directions can be ambiguous. In this paper, we focus on two possible assumptions about the direction of an effect, beginning with the assumption of *average monotonicity*. Informally, for two variables A and B, positive (negative) average monotonicity simply assumes that on average, as A increases, B either increases (decreases) or stays the same. Formally, the assumption of positive average monotonicity states that $\mathbb{E}[B(a') - B(a)] \geq 0$ for all $a' > a$. In a signed causal structure, we simply modify the arrows that we introduced in Section 2 to indicate whether the theorized effect is positive or negative. If A and B have a causal relation

satisfying positive average monotonicity, we label the corresponding arrow as $A \xrightarrow[+]{} B$, indicating that A has a nonnegative effect on the average value of B.[9] Later, we will discuss how a stronger condition, *distributional monotonicity*, is sometimes needed to draw conclusions about the direction of an effect. If positive distributional monotonicity holds, we write $A \xrightarrow[++]{} B$, indicating that increasing A will increase every quantile of B (including, e.g., the median of B). Formally, this is a statement about first-order stochastic dominance, requiring that $\Pr[B(a') \leq c] \leq \Pr[B(a) \leq c]$ for all $a' > a$ and all c.[10] Some readers may be familiar with yet another type of signed effect, *unit-level monotonicity*, an even stronger assumption that we will not use in this paper. This assumption states that if A is increased *for any unit*, B will also increase or stay the same *for that unit*.[11] These conditions are nested within one another: unit-level monotonicity implies distributional monotonicity, which in turn implies on-average monotonicity. Figure 4 presents several possible causal structures describing theory and measurement using signed diagrams indicating on-average monotonicity, the weakest and most plausible of the above monotonicity assumptions; we primarily focus on results involving this assumption.

Figure 4: **Learned treatments.** Settings in which the true treatment X is unknown but $\hat{X}$ can be estimated from auxiliary information. In all cases, both confounders W and outcome Y are known, and the analyst adjusts for W. Panel (a) describes a simple case in which $\hat{X}$ is an uncontaminated proxy for X; this setting permits a falsification test for the existence of an $X \to Y$ effect. Panel (b) adds the generally plausible assumption that X has an on-average monotonic effect on $\hat{X}$, in which case the sign of the $X \to Y$ effect can also be identified. Even when $\hat{X}$ is contaminated by W, as in panel (c), these results hold as long as W is adjusted for in a subsequent regression. However, if the proxy is contaminated by the outcome itself, as in panel (d), association between $\hat{X}$ and Y cannot be interpreted as evidence for the theorized $X \to Y$ effect.



[9] Note that if other parents of B exist, this must hold conditional on all possible values of these parents.

[10] Note that distributional monotonicity implies average monotonicity, but not vice versa.

[11] Readers may be familiar with strong monotonicity from the "no defiers" assumption of Angrist et al. [1996] in the instrumental-variables setting. Formally, unit-level monotonicity $A \xrightarrow[+++]{} B$ states that $B_i(a') \geq B_i(a)$ for all $a' \geq a$ and all units i.

3.1.2 Using Signed Causal Diagrams for Proxied Treatments

Usefully, when learning $\hat{X}$ from data that is informative about X, it is plausible to assume that $X \rightarrow \hat{X}$ satisfies positive average monotonicity. This assumption states that $\hat{X}$ be informative about X in the sense that when X is larger, the estimated $\hat{X}$ will also tend to be larger, on average. In the context of learned proxies, we consider this to be a weak assumption; it is generally satisfied when well-calibrated machine-learning models are used. This assumption can also be empirically assessed whenever the proxy is learned from labeled data. To do so, the researcher can simply train the model on a fraction of the labeled data (a training set) and inspect the accuracy of predictions in the remaining labeled data (a test set) by generating predicted values for the test set from the model learned in the training set. If average monotonicity holds, then predictions should correlate with the labeled values (which are known in the test set).

When average monotonicity between X and $\hat{X}$ is satisfied and W is correctly adjusted for, as in Figure 4(b), any positive association between $\hat{X}$ and Y implies a positive $X \rightarrow Y$ effect [VanderWeele and Hernán, 2012]. By the same logic, this is true for negative associations, which imply negative effects. Perhaps surprisingly, this is also also holds in the setting of Figure 4(c), in which the learned treatment $\hat{X}$ is contaminated by confounders W. At first glance, this contamination may appear to be problematic, as the measurement error will generally be associated with the outcome (as both $\hat{X}$ and Y are affected by the confounder W). However, because analysts adjust for W, this concern is in fact unwarranted. Conditioning on W controls for the non-causal relationship between $\hat{X}$ and Y that results from contamination from W. Specifically, controlling for W blocks two non-causal alternative explanations—termed "backdoor" paths by Pearl [1995]—from $\hat{X}$ to Y.[12] The first alternative explanation is that X does not have an effect on Y, but X is spuriously associated with Y due to confounding by W; because $\hat{X}$ is influenced by X, this then also manifests in a spurious association between $\hat{X}$ and Y. This possibility, which is present in Figure 4(a–c), can be concisely expressed as $\hat{X} \leftarrow X \leftarrow W \rightarrow Y$.

[12]This adjustment is straightforward when W is discrete (so that the association can be tested within levels of W) or when W's contribution to Y is additively separable from X's contribution (i.e., when $E[Y|W, X] = f(W) + g(X)$); it may be difficult when W is continuous and interacts with X.

The second alternative explanation is that the measurement $\hat{X}$ is directly contaminated by the confounder W (i.e. that $\hat{X} - X$ is influenced by W), denoted $\hat{X} \leftarrow W \rightarrow Y$; this appears only in Figure 4(c). Both backdoor paths can be eliminated by adjusting for W, thereby breaking the chain of association. We refer interested readers to Pearl [2009] for a more comprehensive introduction to these concepts. We caution that if $\hat{X}$ is contaminated by Y itself, as in Figure 4(d), then association between the two clearly cannot be interpreted as evidence of a causal effect of X on Y.

We emphasize that the reverse is not true. Failure to find an association between $\hat{X}$ and Y does not necessarily indicate that no $X \rightarrow Y$ effect exists. In addition to standard issues of power in null hypothesis testing, there is the added issue that a poorly learned $\hat{X}$ may have no or vanishingly little association with X; this problem compounds with any power limitations that would arise in a non-proxied primary analysis. In other words, a lack of detectable association may be due to $X \rightarrow \hat{X}$ as well as $X \rightarrow Y$ path. We will return to additional issues around uncertainty in Section 4.

3.2 Learned outcomes

We next turn to the case when the true outcome Y is unobserved and analysts seek to draw causal inferences from a noisily learned proxy $\hat{Y}$. There are now countless examples of applications in which machine learning was used to learn the outcome. These include every application of text analysis using topic proportions—an unsupervised measure based on observed term frequencies—as an outcome measure.[13] Several possible causal structures depicting theory and measurement are given in Figure 5.

Here, we highlight one prominent recent example to illustrate the concept. Motolinia [2021] studies the effect of allowing reelection on legislator provision of particularistic legislation. The theory states that for a legislator that is seeking votes and deciding where to allocate their effort, providing particularistic legislation will yield the most votes due to its targeted focus on constituent services. To identify the effect of term

[13]For example, approaches that explicitly couple the measurement and inferential processes like the structural topic model [Roberts et al., 2013, 2014, 2016a], as well those that separately learn a topic model with, for example, latent dirichlet allocation [Blei et al., 2003].

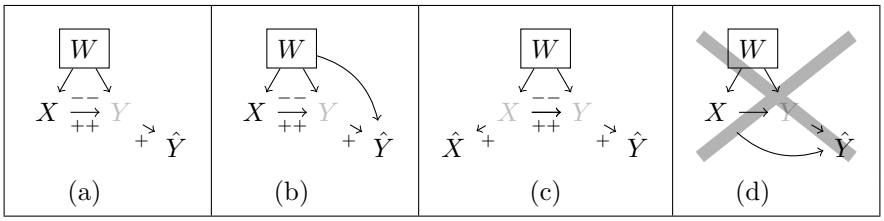
limits, Motolinia [2021] use a difference-in-difference design leveraging a staggered reform to elections in Mexico, which lifted a ban on reelection. Here, X is the ability of a politician to run for reelection, which is perfectly observed. In this case, confounders W are accounted for with state and month-year fixed effects. However, Y, the amount of particularistic legislation proposed, is not directly observed. To estimate the effect of the institutional transition, Motolinia [2021] must generate a measure $\hat{Y}$ of the outcome. To do so, Motolinia [2021] fits a correlated topic model [Blei et al., 2007] to legislative session transcripts, then classifies the resulting topics according to the legislation type. First, topics are grouped according to whether the legislation is procedural (e.g., protocol, voting rules), general (benefits to all constituents), or particularistic (benefits a fraction of constituents). Motolinia [2021] validates this measure with extensive qualitative inspection and by confirming that the measure varies predictably in contexts where theory suggests it ought to (a test of face validity). The core outcome of interest Y is the proportion of particularistic legislation, and the proxy $\hat{Y}$ used in the regression is the *estimated* proportion according to this procedure.

We now demonstrate how estimated effects depending on a learned proxy, like that in Motolinia [2021], estimate the correct sign of the effect on the latent, unobserved variable that is proxied with the measurement model. To do so, we turn to the more general problem of how researchers can draw partial conclusions when using a learned proxy for the outcome. It is well-known that when the learned proxy is correct on average (i.e., when the error $\hat{Y} - Y$ has zero conditional mean), there is no bias in the point estimate [Wooldridge, 2015, Chapter 9.4]. That is, using an unbiased $\hat{Y}$ in place of Y in a regression on W and X is equivalent to simply adding noise to the outcome. However, such perfect conditions rarely hold in practice; for example, when the proxied outcome is binary, the zero conditional mean assumption is violated if the learned model is more likely to misclassify a "true zero" as a "predicted one" than a "true one" as a "predicted zero" (i.e., if misclassification is asymmetrical). This is commonly the case, especially when one value of the outcome is more common than the other.[14] In the more general

[14]Even more implausibly, this perfect symmetry of misclassification must hold within all levels of W.

case, if the measurement error $\hat{Y} - Y$ depends on treatment X, standard Gauss-Markov assumptions are violated and estimates will be biased.

Figure 5: **Learned outcomes.** Settings in which the true outcome Y is unknown but $\hat{Y}$ can be estimated from auxiliary information. In all cases, both confounders W and treatment X are known, and the analyst adjusts for W. Panel (a) describes a simple case in which Y has an on-average monotonic effect on an uncontaminated proxy $\hat{Y}$, and the $X \to Y$ effect is known to be distributionally monotonic. In this case, positive (negative) distributional monotonicity in $X \to Y$ is guaranteed to produce weakly positive (negative) $\text{Cov}(\hat{X}, \hat{Y}|W)$, and the sign of the $X \to Y$ effect can be identified. This result holds even when $\hat{Y}$ is contaminated by W, as in panel (b), as long as these contextual factors are adjusted for in a subsequent regression; it also holds when X is also imperfectly but on-average monotonically learned, as in panel (c). However, if the proxy is contaminated by the treatment itself, as in panel (d), association between X and $\hat{Y}$ cannot be interpreted as evidence for the theorized $X \to Y$ effect.



To describe the conditions under which analysts can draw partial conclusions about the sign of $X \to Y$, even when Y is imperfectly observed, we begin with the structure of Figure 5(a). As before, we find the assumption of average monotonicity on $Y \underset{+}{\longrightarrow} \hat{Y}$ to be generally plausible and empirically verifiable. Unfortunately, this assumption alone is not generally sufficient to assess whether $X \to Y$ is "on-average positive" or "on-average negative." Because it is possible to construct examples where $X \underset{-}{\longrightarrow} Y \underset{+}{\longrightarrow} \hat{Y}$ and $X \underset{+}{\longrightarrow} Y \underset{+}{\longrightarrow} \hat{Y}$ both lead to positive correlations between X and $\hat{Y}$, analysts cannot conclude that X has an on-average positive effect on Y simply from observing that $\text{Cov}(X, \hat{Y}) > 0$. For examples and detailed explanations, we direct interested readers to VanderWeele et al. [2008] and VanderWeele and Hernán [2012].

Instead, a stronger *distributional monotonicity* condition is required. It has been shown that $X \underset{++}{\longrightarrow} Y \underset{+}{\longrightarrow} \hat{Y}$ always leads to a positive correlation between X and $\hat{Y}$, and similarly that $X \underset{--}{\longrightarrow} Y \underset{+}{\longrightarrow} \hat{Y}$ always produces a negative correlation. Therefore, if the $X \to Y$ effect is assumped to be distributionally monotonic, the sign of that effect can be inferred. We caution that when X and Y are continuous, distributional monotonicity in

$X \xrightarrow[++]{} Y$ is a strong assumption that must be justified with domain expertise. However, an important special case is when both treatment and outcomes are binary, in which case average and distributional monotonicity are equivalent. In this case, this assumption is considerably simpler and analysts can safely infer the sign of $X \rightarrow Y$ using the proxy $\hat{Y}$.

Next, we highlight two more complex cases in which analysts can nonetheless draw partial causal inferences from imperfect proxies. The first case is when W contaminates $\hat{Y}$, as in Figure 5(b); like in Section 3.1, this is less of a problem because spurious association from W can be adjusted for in the subsequent primary regression.[15] The second case is when learned versions of both $\hat{X}$ and $\hat{Y}$ are used in place of the true treatment and outcome, as in Figure 5(c). In this case, results generalize straightforwardly: due to a technical result from VanderWeele et al. [2008], if $X \rightarrow Y$ is known to be distributionally monotonic, then the conditional effect must share the sign of $\text{Cov}(\hat{X}, \hat{Y}|W)$.[16]

3.3 Learned confounders

Finally, we consider the difficult task of estimating causal effects by adjusting for imperfectly learned confounders, $\hat{W}$, instead of the true concept, W. Again, illustrations of proxied confounders are plentiful in the social sciences. An especially prominent example is legislator ideal points [Poole and Rosenthal, 1985], which researchers often wish to control for when explaining legislator behavior. Because ideology cannot be directly observed, political scientists construct proxies for ideal points with unsupervised scaling methods. There are at least two reasons for this. First, it would be very difficult to reliably hand-label each legislator on a continuous scale. Second, because all legislators routinely vote on the same bills, latent trait models are a reasonable way to project these votes down to one or two dimensions. Much research is devoted to the measurement of this variable, and we direct interested researchers to Clinton [2012] for further discussion.

[15] As noted in Section 3.1, the issue is fully resolved when W is discrete, or alternatively when $W \rightarrow Y$ and $X \rightarrow Y$ are additively separable.

[16] This is because the sign of the correlation induced by a path can be inferred by multiplying the signs of edges along that path when either (1) all edges are distributionally monotonic or (2) intermediate edges are distributionally monotonic and final edges are on-average monotonic.

Nyhan et al. [2012] demonstrates the importance of adjusting for ideological position when studying legislative voting. Specifically, they examine the effect of controversial roll-call votes, X—specifically, high-profile votes against the Republican-led healthcare reform in 2010—on subsequent electoral performance, Y. To estimate this effect, Nyhan et al. [2012] uses a selection-on-observables design and note clear confounding by legislator ideal point W, which shapes both legislative positions and voter evaluation. After conditioning on estimated ideal points $\hat{W}$ and other confounders, Nyhan et al. [2012] estimate that votes against healthcare reform may have cost the Democrats the majority in subsequent midterm elections.

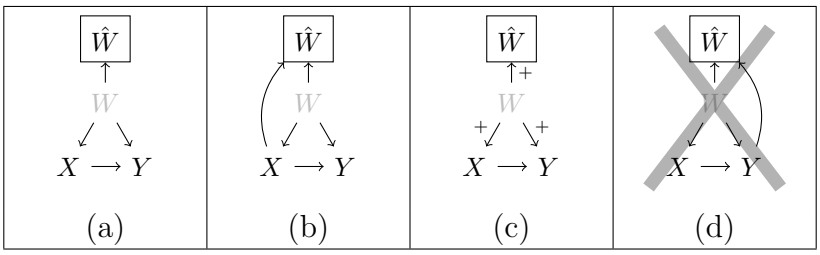
We now consider the general problem of drawing conclusions with proxied confounders. Intuitively, it is generally insufficient to simply treat $\hat{W}$ as if it were W, because the remaining error $W - \hat{W}$ also contributes to confounding. Strategies nonetheless exist for recovering causal effects in certain settings, though we caution that available solutions are fragile in various ways elaborated below. We further emphasize that common practice deviates substantially from these solutions for estimating causal effects in the presence of proxied confounding.

We begin by examining the simple setting of Figure 6. Greenland and Lash [2008] and Kuroki and Pearl [2014] establish that when W and $\hat{W}$ are discrete, causal effects are nonparametrically identified if analysts know the *error mechanism*, the distribution $p(\hat{w}|W = w)$—in other words, the pattern of correct and incorrect proxy values that arise from each possible true value. The basic idea is that when this error mechanism is known, it can be used in combination with the observed distribution of proxy values to back out the unobserved distribution of underlying true values.[17] This procedure can then be applied within each level of X and Y.

Approaches that rely on quantifying the error distribution are particularly attractive in supervised learning, where analysts following best practices already evaluate models in held-out validation sets. This validation set offers a way to unbiasedly estimate the

[17]This requires that $\hat{W}$ must be sufficiently informative about W. In particular, one concern is that two confounder values, w and w', may produce the same proxy distribution, $p(\hat{w}|W = w) = p(\hat{w}|W = w')$. This could occur if, for example, limited signals are incapable of distinguishing between two classes, or if there is a "ceiling effect" beyond which increasing W no longer affects $\hat{W}$.

Figure 6: **Learned confounders.** Settings in which the true confounder W is unknown but $\hat{W}$ can be estimated from auxiliary information. In all cases, both treatment X and outcome Y are known, but the analyst is only able to adjust for W. Panel (a) describes a simple case in which $\hat{W}$ is an uncontaminated proxy for X; panel (b) describes a generalization in which $\hat{W}$ may be contaminated by X. Merely controlling for $\hat{W}$ is insufficient to unbiasedly estimate $X \to Y$ in these cases. However, the methods described in Kuroki and Pearl [2014] and Miao et al. [2018] can in principle recover these effects, if certain conditions are satisfied. Panel (c) depicts a true confounder W that has a positive, on-average monotonic effect on both X and Y. In this case, an observed negative association between X and Y implies that a negative $X \to Y$ effect exists and is sufficiently strong to overpower the positive association induced by confounding; this remains true whether or not analysts adjust for $\hat{W}$ (the reverse holds for negative confounding and positive association between X and Y). If the proxy is contaminated by the outcome itself, as in panel (d), causal effects are difficult to recover.



required information essentially for free, allowing analysts to recover the causal effects of interest. Indeed, in binary classification, widely used evaluation metrics—true and false positive rates—correspond exactly to $p(\hat{w}|W = w)$. Fong and Tyler [2018] build on this intuition in the linear case, developing a general method of moments estimator that simultaneously estimates the error distribution and the primary regression.

Results on proxied confounding also hold for the case when the learned confounder is contaminated by the treatment, as in Figure 6(b). However, if $\hat{W}$ is contaminated by Y, then the $X \to Y$ effect cannot be recovered. For this reason, Fong and Tyler [2018] recommend explicitly excluding Y from features used to train machine-learning models. Even so, contamination can creep into the learned measure through numerous channels, including (1) learned models that inappropriately leverage correlates of Y, (2) model misspecification as discussed in Section 2, or (3) contamination of observed signals that influence human annotations.

Kuroki and Pearl [2014] and Miao et al. [2018] extend these results to the challenging case where the true confounder is completely unobserved, as in unsupervised learning. A

review of these techniques is beyond the scope of this paper; for an overview of causal inference with proxy confounders, see Tchetgen Tchetgen et al. [2020]. However, we emphasize that these methods are substantially more involved than the common two-stage practice of fitting an unsupervised measurement model and then controlling for the result in the primary regression—a procedure that does not, in general, consistently recover causal estimates. Broadly speaking, consistent estimation of $X \rightarrow Y$ in the presence of imperfectly measured confounding is an extremely difficult task. Kuroki and Pearl [2014] note poor finite sample performance of these methods; importantly, in the settings we examine, asymptotics depend critically on the size of the validation set; sampling errors in $\hat{p}(\hat{w}|W = w)$ can lead to substantial errors even when the primary analysis is based on infinite data. Tchetgen Tchetgen et al. [2020] also note issues with numerical instability, though these can be partially addressed with additional parametric modeling assumptions.

4 Accurately Reporting Uncertainty

In the preceding section, we describe how imperfect proxies lead to biased point estimates. This raises an obvious question: do imperfect proxies also lead to biased statements about uncertainty? In short, the answer is "yes." We now explain how analysts can nevertheless draw principled conclusions despite these challenges.

In general, the common practice of using learned proxies *as if* they directly reflect the underlying true concept will bias standard errors downward, leading to overconfident conclusions that may fail to replicate. This is because standard procedures only account for uncertainty due to sampling variability in the primary analysis (the second stage of a proxy-based research workflow, which occurs after fitting the measurement model and estimating a proxy). In doing so, researchers do not account for the fact that the learned measurement model (the first stage) is *also* estimated with a sample of data, introducing variability into the resulting proxy and therefore also contributing to overall uncertainty.

In other forms of research, such as analyses with missing data, it is well known that

ignoring uncertainty from earlier stages (i.e., multiple imputation) leads to unreliable standard errors in subsequent regressions [Blackwell et al., 2017]. Despite widespread awareness of this issue in related contexts, our review of published proxy-based work suggests that researchers rarely attempt to correct their standard errors. Among papers using computational methods in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*, the vast majority of papers analyzing learned proxies ignore the fact that these proxies are estimated with uncertainty.[18] The only exceptions were applications of the Structural Topic Model [STM, Roberts et al., 2013, 2014, 2016a]. Interestingly, while substantive papers conducting a proxy-dependent empirical test generally ignored uncertainty in the learned measure, methodological papers proposing a novel proxy often included a method for measuring uncertainty in the learned measure. For instance, Caughey et al. [2019] develop a proxy for mass policy ideology in Europe with a Bayesian dynamic group-level IRT model, from which uncertainty is easily extracted from the posterior estimates. But while this is common across Bayesian models, none of the papers that we identified incorporated this uncertainty into subsequent empirical tests.

Before describing solutions to this issue, including the approach used by STM, we first provide a more in-depth review of sources of uncertainty that are often unaccounted for when using learned proxies.

4.1 (Mostly) Ignored Sources of Uncertainty When Using Learned Proxies

Why do learned proxies lead to overconfident conclusions? Here, we briefly enumerate sources of uncertainty that, when ignored, lead to inappropriately small standard errors for the causal estimate of theoretical interest.

We begin by considering a supervised analysis in which the learned proxy is estimated from a training set, which is a sample from a population of possible training units. (Note that the same logic holds for unsupervised measurement models.) Our first source of

[18] For details, see Appendix Section A.

uncertainty is the sampling variability that results from drawing one of many possible training sets. For simplicity of exposition, we will assume that (1) annotators correctly label the underlying ground truth for each unit and (2) analysts recover a global maximum likelihood estimate for the measurement model, rather than a "local mode" that depends on randomly selected starting values [Roberts et al., 2016b]. However, we note that in reality, these and other sources of nuisance variation can also undermine replicability of empirical conclusions.

Under these simplifying assumptions, given a particular training sample, applying a measurement model to this training set will lead deterministically to an estimate for the measurement model parameters. However, if a different training sample had been drawn, then the measurement model would have learned a different mapping from the observed signal to the concept of interest. This leads to a sampling distribution over learned measurement models.

These model parameters are in turn used to generate learned measures—whether $\hat{W}$, $\hat{X}$, or $\hat{Y}$—for each unlabeled unit in the primary analysis. Here, it is important to note that a slight change in the learned model (including changes due to a slightly different sample of training observations) will alter the generated proxy values for many units simultaneously. Put another way, over repeated sampling of the training set, learned measures in the primary analysis set are correlated across units. Moreover, units that have similar observed information will tend to shift similarly.

Our final source of uncertainty arises when learned measures are used in a primary analysis. The primary-analysis dataset is also a sample from a broader population, producing another source of random variation. In our review, with the exception of STM applications, every paper conducting a direct test with a learned proxy neglected training uncertainty and reported only uncertainty from the primary regression.

Thus, a widespread methodological issue in existing work is the failure to adequately report uncertainty from the training process. There are numerous reasons why this issue has persisted. When analysts obtain pretrained machine-learning models from third parties—e.g., commercial sources or other researchers—they may not know precisely how

this sampling was done, and uncertainty may not be adequately reported. For example, estimated sampling variances of model parameters might be reported, but not covariances. Similarly, if unit-level features are supplied to a cloud service, the service might respond with predictions and associated uncertainty for the unit-level learned measure, but cross-unit covariance is rarely reported by currently available services.

A simple *reductio ad absurdum* argument further illustrates the importance of training uncertainty for analyses based on $\hat{X}$ and $\hat{W}$ as well. If training uncertainty could in fact be safely ignored, in the limit, it would imply that binary classifiers could be trained on only two randomly sampled observations—one positive case and one negative case. The resulting model could then be used to learn measures for an infinite number of units. A primary analysis in this group would contribute no additional sampling uncertainty, due to its size. As a result, an analyst ignoring the training stage would claim perfect certainty in the results of their primary regression—a facially absurd claim, given that the entire analytic workflow hinges on a miniscule sample of two units. This illustration reveals that when properly accounted for, uncertainty vanishes only as both the measurement-model (first-stage) and primary-analysis (second-stage) datasets grow large. For this reason, we strongly discourage the widespread practice of ignoring training uncertainty (or, equivalently, reporting results "conditional on" pretrained models or learned measures based on their predictions). Because the causal theories being analyzed are ultimately about X, Y, and Z—not $\hat{X}$, $\hat{Y}$, and $\hat{Z}$, which are merely proxies with no intrinsic causal role in the theory—analysts must take the underlying true concepts seriously.

As a final illustration, consider the use of a learned proxy, $\hat{Y}$. Here, correlation in $\hat{Y} - Y$ across units is functionally identical to correlation in the error term of a regression (e.g., as can occur in cluster randomized trials, where units within a cluster may be simultaneously influenced by unobserved factors). It is easy to see that failure to account for correlated errors in the primary regression will typically lead to underestimates of uncertainty in the resulting estimates, much like failure to use clustered standard errors in a clustered design.

Given this challenge, how should researchers correct uncertainty estimates when using

learned proxies? Next, we make recommendations which apply broadly across a range of designs.

4.2 Correcting Errors in Estimated Uncertainty

To represent uncertainty in the initial measurement stage, researchers can employ a range of common methods. Specifically, this uncertainty may be represented (1) with draws from a multivariate normal distribution, using point estimates and an estimated covariance matrix for the measurement-model parameters; (2) with draws from the joint posterior of parameters in a Bayesian analysis; or (3) with bootstrap draws of learned parameters, obtained by resampling of the training set and rerunning of the measurement model. Regardless of how it is obtained, each draw represents one possible measurement model that could have been learned; together, they approximate the spread of learned models that are plausible, given the finite training sample.

One improved and easy-to-implement method for reporting uncertainty follows the procedure of Treier and Jackman [2008]. Take the first draw, $t = 1$, corresponding to one of the T trained measurement models drawn as described above. Compute the proxy, e.g. $\hat{X}^{(t=1)}$, under this measurement model. Next, conduct the primary analysis using this proxy and extract the biased estimate of the quantity of interest, e.g. the $\hat{X}^{(t=1)}$ coefficient in a regression of Y on $\hat{X}$ and W. Uncertainty on this quantity of primary interest can then be accounted for by taking P draws as above—i.e., by drawing from a multivariate normal approximation, drawing from a Bayesian posterior, or taking bootstrap draws. These draws approximate the uncertainty in the primary analysis *taking the* $t = 1$ *proxy as given*. The current standard practice essentially stops at this point and, as a result, only accounts for primary-analysis uncertainty. In contrast, we recommend repeating the process T times, producing a total of $T \times P$ samples for the quantity of interest. Taking the 2.5th and 97.5th percentiles of the resulting distribution will produce an interval that reflects both uncertainty from both measurement and primary analysis.

We caution that this procedure lacks many properties possessed by traditional confidence intervals. In particular, due to the bias in point estimates that we discuss exten-

sively above, it will not generally contain the true causal estimand in 95% of repeated samples. Despite this issue, it can be used in conjunction with the null-hypothesis-testing and effect-signing techniques developed above, while accounting for sampling variability in both stages of the analytic workflow.

The case of STM [STM, Roberts et al., 2013, 2014, 2016a] illustrates an alternative, more complex approach for obtaining principled uncertainty estimates. Specifically, STM estimates a single model that encompasses both the initial measurement stage and the subsequent primary-analysis stage. This allows information to be passed back and forth between stages—e.g., using patterns from in the primary analysis to refine proxy predictions from the measurement model—leading to greater statistical power. In general, joint modeling is possible for a wide variety of modeling approaches (e.g., Knox and Lucas [2021] develops a joint model for a very different application — speech audio — than that addressed by STM). But relative to the sampling-based procedures describe above, the tradeoff is that the joint modeling approach requires somewhat more technical familiarity and case-specific coding to implement. However, joint modeling is increasingly feasible to implement in languages such as Stan.

5 Recommendations and Concluding Thoughts for Credible Estimates with Learned Proxies

As our review demonstrates, it is now commonplace to learn proxies with computational methods as a first step in testing a causal theory. Though this approach has opened the door to numerous new and innovative studies, the use of imperfect proxies also presents challenges. To address these challenges, we now outline a series of best-practice recommendations for drawing principled conclusions from analyses using computational proxies of theorized concepts. We conclude by noting that while computational methods are transforming the social sciences, the underlying statistical issues resemble those seen decades ago—highlighting the need for careful research design and methodological caution in social-scientific research.

5.1 Recommendations

Explicitly state your causal theory. As this article makes clear, formally specifying the theorized causal diagram has numerous benefits. These diagrams are concise and easy-to-use tools for communicating concepts to readers and clarifying the assumptions that underlie an analysis. Importantly, a well-specified causal diagram includes not only the theorized process, but also a discussion of possible contamination sources and measurement-quality assumptions for proxies of unobserved variables. As we discuss above, clearly specified causal diagrams also help in reasoning about sources of error and assessing what conclusions can be supported with a particular research design. Most notably, they reveal when null hypotheses and effect signs can be reliably tested.

Avoid overclaiming based on biased point estimates. We show that primary analyses based on imperfectly learned proxies are almost always biased. Given this, researchers should be conservative in their interpretation by simply characterizing the sign of an effect, rather than making unsupportable claims about effect magnitude. If a researcher wishes to draw inferences about precise effect sizes, methods such as Duarte et al. [2021] offer a way to obtain bounds on possible effect sizes that account for the issues discussed here. Incorporating and expanding on this cautious approach to causal inference—whether by focusing on effect sign or through effect bounding—is an important avenue for future work by applied reearchers and methodologists.

Test your assumptions. Researchers making assumptions (e.g., about the monotonicity of an effect) ought to assess their plausibility by drawing on past work, domain expertise, or empirical evaluation where possible. In particular, assumptions about on-average monotonicity in measurement, such as $X \underset{+}{\longrightarrow} \hat{X}$, are straightforward to evaluate with procedures described in this paper.

Always assess and report measurement performance. The performance of the measurement model undergirds any research design employing learned proxies. Proxies that are noisier or more skewed will tend to exacerbate the issues that we describe above. Among other issues, they can lead to "false negative" results: failure to find evidence in support of a theory, even when that theory is true. In general, researchers

should not trust results from any machine learning model until the performance of that model is suitably demonstrated. To demonstrate satisfactory performance, all applications should include a confusion table (i.e., cross-tabulation of true and predicted values) and other performance metrics obtained from a held-out validation set that was not used for training or parameter tuning. Finally, researchers should include measures of inter-coder reliability, especially when annotating ambiguous labels. With limited resources, it may be inefficient to annotate each example in the labbeled set repeatedly; instead, we encourage re-labeling a sufficient number of cases to assess reliability. For example, with 2,000 training examples, it may be sufficient to hire a second coder to label only 100 for comparison.

Correct your standard errors. As we note above, reported uncertainty for proxy-based analyses are almost certainly biased, generally in a downward (anti-conservative) direction. This is intuitive given that analysts typically only report uncertainty for the second-stage model (the primary analysis, targeting the causal effect of interest) and neglect uncertainty and bias in the first stage (learning proxies). Unfortunately, without access to the learned model, it can be extraordinarily difficult to characterize how measurement error covaries between units. In the previous section, we describe methods for correcting this downward bias. Regardless which approach is used, analysts should seek to accurately report uncertainty from all stages of the model.

Compare estimates in the full data to estimates using only the labeled observations. We echo the observation by Fong and Tyler [2018] that, in the supervised case, a simple and consistent estimator exists: fitting a model using only the labeled data. This estimator is desirable for several reasons. First, and most obviously, it does not use proxies and therefore does not suffer from any of the sources of bias that we discuss in this article—at least, as long as the researcher can ensure that human labels reflect true, gold-standard values for the underlying concept of interest. Second, substantial differences between the smaller-N unproxied analysis and the larger-N proxied analysis may indicate deeper issues that warrant further investigation. For instance, these estimates may diverge if the labeled data is not representative of the full data or if there are

systematic errors in the classifier.

5.2 Conclusion

Our review highlights that advances in computational statistics are transforming research in the social sciences, primarily by allowing researchers to measure theoretized concepts and use the resulting proxies in subsequent causal analyses. Yet despite the increasing prevalence of this research strategy, little methodological guidance is available for applied scholars. This is troubling because, as we note, the common practice of conflating proxies with the underlying true concept leads to biased point estimates and standard errors, undermining the conclusions drawn from this work.

Our analysis reveals that in spite of the recent computational revolution, core statistical obstacles faced by the discipline remain largely unchanged. In fact, our emphasis on precisely articulating theory and assumptions highlights that, ultimately, credible causal inference is about research design—same as it ever was. While new models may improve our ability to approximate previously unobservable concepts, no amount of computation can evaluate the plausibility of assumptions or prevent researchers from drawing unsupportable conclusions. It is thus unsurprising that new research using new computational methods suffers from issues similar to those that ailed proxy-based studies decades ago.

But more optimistically, our review demonstrates how recent advances in causal inference can augment concurrent computational developments. By writing down their assessments of proxy contamination and measurement quality in the form of simple causal diagrams, analysts can now easily assess if a causal claim—whether about the existence, direction, or magnitude of an effect—is defensible. However, methodology in this area is far from complete. As computational social science continues to grow, much more work is needed to ensure that this rapidly expanding research area produces reliable scientific knowledge.

References

Adcock, R. and Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Ansolabehere, S., Lessem, R., and Snyder Jr, J. M. (2006). The orientation of newspaper endorsements in us elections, 1940–2002. *Quarterly Journal of political science*, 1(4):393.

Blackwell, M., Honaker, J., and King, G. (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, 46(3):303–341.

Blei, D. M., Lafferty, J. D., et al. (2007). A correlated topic model of science. *The annals of applied statistics*, 1(1):17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Carlson, D. and Montgomery, J. M. (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review*, 111(4):835–843.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32.

Carroll, R. J. and Kenkel, B. (2019). Prediction, proxies, and power. *American Journal of Political Science*, 63(3):577–593.

Caughey, D., O'GRADY, T., and Warshaw, C. (2019). Policy ideology in european mass publics, 1981–2016. *American Political Science Review*, 113(3):674–693.

Cheng, L. and Van Ness, J. W. (1999). *Statistical regression with measurement error*. Arnold; New York: Oxford University Press,.

Clinton, J. D. (2012). Using roll call estimates to test models of politics. *Annual Review of Political Science*, 15:79–99.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. (2021). An automated approach to causal inference in discrete settings. https://arxiv.org/pdf/2109.1347.pdf.

Fong, C. and Tyler, M. (2018). Machine learning predictions as regression covariates. *Political Analysis*, pages 1–18.

Freedom House (2014). *Freedom in the world 2014: The annual survey of political rights and civil liberties*. Rowman & Littlefield.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Greenland, S. and Lash, T. L. (2008). Bias analysis. In Rothman, K. J., Greenland, S., and Lash, T. L., editors, *Modern Epidemiology*. Lippincott Williams & Wilkins.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Gurr, T. R. (1974). Persistence and change in political systems, 1800–1971. *American political science review*, 68(4):1482–1504.

Keele, L., Stevenson, R. T., and Elwert, F. (2020). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, 8(1):1–13.

Knox, D. and Lucas, C. (2021). A dynamic model of speech for the social sciences. *American Political Science Review*, 115(2):649–666.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Larcinese, V., Puglisi, R., and Snyder Jr, J. M. (2011). Partisan bias in economic news: Evidence on the agenda-setting behavior of us newspapers. *Journal of public Economics*, 95(9-10):1178–1189.

Martin, G. J. and McCrain, J. (2019). Local news and national politics. *American Political Science Review*, 113(2):372–384.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Motolinia, L. (2021). Electoral accountability and particularistic legislation: Evidence from an electoral reform in mexico. *American Political Science Review*, 115(1):97–113.

Munck, G. L. and Verkuilen, J. (2002). Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative political studies*, 35(1):5–34.

Nyhan, B., McGhee, E., Sides, J., Masket, S., and Greene, S. (2012). One vote out of step? the effects of salient roll call votes in the 2010 election. *American Politics Research*, 40(5):844–879.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic books.

Poole, K. T. (2008). The evolving influence of psychometrics in political science. In Box-Steffensmeier, J. M., Brady, H. E., and Collier, D., editors, *The Oxford Handbook of Political Methodology*, volume 10. Oxford Univerity Press.

Poole, K. T. and Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

Puglisi, R. and Snyder, J. M. (2015). Empirical studies of media bias. In *Handbook of media economics*, volume 1, pages 647–667. Elsevier.

Reed, W., Clark, D. H., Nordstrom, T., and Hwang, W. (2008). War, power, and bargaining. *The Journal of Politics*, 70(4):1203–1216.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016a). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2016b). Navigating the local modes of big data. *Computational social science*, 51.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Treier, S. and Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, 52(1):201–217.

VanderWeele, T. J. and Hernán, M. A. (2012). Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology*, 175(12):1303–1310.

VanderWeele, T. J., Hernán, M. A., and Robins, J. M. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology*, 19(5):720.

VanderWeele, T. J. and Robins, J. M. (2010). Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):111–127.

Waldner, D. (2015). Process tracing and qualitative causal inference. *Security Studies*, 24(2):239–250.

Weber, M. (2017). *Methodology of social sciences*. Routledge.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

A Literature Review

To establish the prevalence of machine learning in modern political science, we survey all papers published in the *American Journal of Political Science*, *American Political Science Review*, and *Journal of Politics* between January 2018 and January 2021. To determine the number of articles employing machine learning in these outlets over this time period, we employ the following procedure:

1. Read the abstract to see if it contains enough information about what method or model is used. If the abstract states that the paper used machine learning, topic model, etc, classify accordingly.

2. Search for key words based on the following list. If a keyword is found, read the surrounding text to confirm that the keyword correctly identifies a machine learning application.

    - Machine learning
    - Supervised learning
    - Unsupervised learning
    - Semi-supervised learning
    - Deep learning
    - Classifier/Classification
    - Text analysis
    - Image analysis
    - High dimension/dimensional/dimensionality
    - Topic model
    - NLP (natural language processing)
    - Network
    - Computational
    - Artificial
    - Automated

Then, within all articles classified employing machine learning, we coded the following fields:

Journal: Journal in which the article was published.

Year: Year in which the article was published.

Volume: Volume in which the article was published.

Issue: Issue in which the article was published.

Author: Authors of the article.

Title: Title of the article.

Abstract: Abstract of the article.

Data URL: URL at which replication data can be found.

imated Variable: Description of proxy.

Used As: Treatment, Outcome, Confounder (control), or NA.

Features: Broad class of features used to predict the missing value.

Model: Model employed by the study.

d-stage model...: Was a learned proxy estimated then used in a second-stage model to test a theory?

d stage model...: If a learned proxy wasn't used in a second-stage model, was a learned proxy developed and proposed for use in second-stage models?

lles Uncertainty: If a proxy was used in a second-stage model, did the second-stage analysis account for uncertainty in the proxy? If a proxy was proposed for use in a second-stage model, does the model provide a method for proxy uncertainty?

If NA, why?: If NA in some columns, explain here.

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
1	AJPS	2018	62	1	Connor Huff, Joshua D. Kertzer	How the Public Defines Terrorism	Every time a major violent act takes place in the United States, a public debate erupts as to whether it should be considered terrorism. Political scientists have offered a variety of conceptual frameworks, but have neglected to explore how ordinary citizens understand terrorism, despite the central role the public plays in our understanding of the relationship between terrorism and government action in the wake of violence. We synthesize components of both scholarly definitions and public debates to formulate predictions for how various attributes of incidents affect the likelihood they are perceived as terrorism. Combining a conjoint experiment with machine learning techniques and automated content analysis of media coverage, we show the importance not only of the type and severity of violence, but also the attributed motivation for the incident and social categorization of the actor. The findings demonstrate how the language used to describe violent incidents, for which the media has considerable latitude, affects the likelihood the public classifies incidents as terrorism.	Likelihood of an incident being understood as terrorism	descriptive	Descriptors of a violect act	SVM	No		No	
2	AJPS	2018	62	2	Lucy Barnes, Timothy Hicks	Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy	What explains variation in individual attitudes toward government deficits? Although macroeconomic stance is of paramount importance for contemporary governments, our understanding of its popular politics is limited. We argue that popular attitudes regarding austerity are influenced by media (and wider elite) framing. Information necessary to form preferences on the deficit is not provided neutrally, and its provision shapes how voters understand their interests. A wide range of evidence from Britain between 2010 and 2015 supports this claim. In the British Election Study, deficit attitudes vary systematically with the source of news consumption, even controlling for party identification. A structural topic model of two major newspapers' reporting shows that content varies systematically with respect to coverage of public borrowing—in ways that intuitively accord with the attitudes of their readership. Finally, a survey experiment suggests causation from media to attitudes: deficit preferences change based on the presentation of deficit information.	Topics about fiscal policy	descriptive	text	STM(structual topic model)	No		No	
3	AJPS	2018	62	3	Zachary M. Jones, Yonatan Lupu	Is There More Violence in the Middle?	Is there more violence in the middle? Over 100 studies have analyzed whether violent outcomes such as civil war, terrorism, and repression are more common in regimes that are neither full autocracies nor full democracies, yet findings are inconclusive. While this hypothesis is ultimately about functional form, existing work uses models in which a particular functional form is assumed. Existing work also uses arbitrary operationalizations of "the middle." This article aims to resolve the empirical uncertainty about this relationship by using a research design that overcomes the limitations of existing work. We use a random forest-like ensemble of multivariate regression and classification trees to predict multiple forms of conflict. Our results indicate the specific conditions under which there is or is not more violence in the middle. We find the most consistent support for the hypothesis with respect to minor civil conflict and no support with respect to repression.	NA	NA	NA	CART (random forest-like ensemble of multivariate regression and classification trees)	NA		NA	Uses ML for functional flexibility, not measurement
4	AJPS	2018	62	3	Jacob M. Montgomery, Santiago Olivella	Tree-Based Models for Political Science Data	Political scientists often find themselves analyzing data sets with a large number of observations, a large number of variables, or both. Yet, traditional statistical techniques fail to take full advantage of the opportunities inherent in "big data," as they are too rigid to recover nonlinearities and do not facilitate the easy exploration of interactions in high-dimensional data sets. In this article, we introduce a family of tree-based nonparametric techniques that may, in some circumstances, be more appropriate than traditional methods for confronting these data challenges. In particular, tree models are very effective for detecting nonlinearities and interactions, even in data sets with many (potentially irrelevant) covariates. We introduce the basic logic of tree-based models, provide an overview of the most prominent methods in the literature, and conduct three analyses that illustrate how the methods can be implemented while highlighting both their advantages and limitations.	simulated outcomes	NA	number	CART, RF, GBM, and BART	NA		NA	Methods article without measurement component
								whether the advertising gone negative in a given week	NA	advertisement	GBM, BART	NA		NA	Methods article without measurement component
								turnout and vote for McCain	NA	vote	Poststratified BART	NA		NA	Methods article without measurement component
5	AJPS	2018	62	4	Adam Bonica	Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning	This article develops a generalized supervised learning methodology for inferring roll-call scores from campaign contribution data. Rather than use unsupervised methods to recover a latent dimension that best explains patterns in giving, donation patterns are instead mapped onto a target measure of legislative voting behavior. Supervised models significantly outperform alternative measures of ideology in predicting legislative voting behavior. Fundraising prior to entering office provides a highly informative signal about future voting behavior. Impressively, forecasts based on fundraising as a non-incumbent predict future voting behavior as accurately as in-sample forecasts based on votes cast during a legislator's first 2 years in Congress. The combined results demonstrate campaign contributions are powerful predictors of roll-call voting behavior and resolve an ongoing debate as to whether contribution data successfully distinguish between members of the same party.	ideology	descriptive	campaign contribution data	SVM; random forest	No	Yes	No	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
6	AJPS	2019	63	2	Kenneth Benoit, Kevin Munger, Arthur Spirling	Measuring and Explaining Political Sophistication through Textual Complexity	Political scientists lack domain-specific measures for the purpose of measuring the sophistication of political communication. We systematically review the shortcomings of existing approaches, before developing a new and better method along with software tools to apply it. We use crowdsourcing to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of sophistication. This includes previously excluded features such as parts of speech and a measure of word rarity derived from dynamic term frequencies in the Google Books data set. Our technique not only shows which features are appropriate to the political domain and how, but also provides a measure easily applied and rescaled to political texts in a way that facilitates probabilistic comparisons. We reanalyze the State of the Union corpus to demonstrate how conclusions differ when using our improved approach, including the ability to compare complexity as a function of covariates.	textual sophistication	descriptive	text summaries	random forest	No	Yes	No	
7	AJPS	2019	63	3	Robert J. Carroll, Brenton Kenkel	Prediction, Proxies, and Power	Many enduring questions in international relations theory focus on power relations, so it is important that scholars have a good measure of relative power. The standard measure of relative military power, the capability ratio, is barely better than random guessing at predicting militarized dispute outcomes. We use machine learning to build a superior proxy, the Dispute Outcome Expectations (DOE) score, from the same underlying data. Our measure is an order of magnitude better than the capability ratio at predicting dispute outcomes. We replicate Reed et al. (2008) and find, contrary to the original conclusions, that the probability of conflict is always highest when the state with the least benefits has a preponderance of power. In replications of 18 other dyadic analyses that use power as a control, we find that replacing the standard measure with DOE scores usually improves both in-sample and out-of-sample goodness of fit.	military power	treatment	material capabilities	super learner	Yes		No	
8	AJPS	2019	63	4	Peter K. Hatemi, Charles Crabtree, Kevin B. Smith	Ideology Justifies Morality: Political Beliefs Predict Moral Foundations	Moral Foundations Theory (MFT) is employed as a causal explanation of ideology that posits political attitudes are products of moral intuitions. Prior theoretical models, however, suggest the opposite causal path, that is, that moral judgments are driven by political beliefs. In both instances, however, extant research has assumed rather than explicitly tested for causality. So do moral intuitions drive political beliefs or do political beliefs drive moral intuitions? We empirically address this question using data from two panel studies and one nationally representative study, and find consistent evidence supporting the hypothesis that ideology predicts moral intuitions. The findings have significant implications for MFT as a theory of ideology, and also about the consequences of political beliefs for shaping how individuals rationalize what is right and what is wrong.	NA	NA	NA	random forest	NA		NA	Uses ML for functional flexibility, not measurement
9	AJPS	2020	64	1	Andreu Casas, Matthew J. Denny, John Wilkerson	More Effective Than We Thought: Accounting for Legislative Hitchhikers Reveals a More Inclusive and Productive Lawmaking Process	For more than half a century, scholars have been studying legislative effectiveness using a single metric—whether the bills a member sponsors progress through the legislative process. We investigate a less orthodox form of effectiveness—bill proposals that become law as provisions of other bills. Counting these "hitchhiker" bills as additional cases of bill sponsorship success reveals a more productive, less hierarchical, and less partisan lawmaking process. We argue that agenda and procedural constraints are central to understanding why lawmakers pursue hitchhiker strategies. We also investigate the legislative vehicles that attract hitchhikers and find, among other things, that more Senate bills are enacted as hitchhikers on House laws than become law on their own.	textual similarity (measuring hitchhiker bills)	outcome	text	A New Sequence-Based Algorithm for Characterizing Document Similarity	Yes		No	
10	AJPS	2020	64	1	Richard A. Nielsen	Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers	How do women gain authority in the public sphere, especially in contexts where patriarchal norms are prevalent? I argue that the leaders of patriarchal social movements face pragmatic incentives to expand women's authority roles when seeking new movement members. Women authorities help patriarchal movements by making persuasive, identity-based arguments in favor of patriarchy that men cannot, and by reaching new audiences that men cannot. I support this argument by examining the rise of online female preachers in the Islamist Salafi movement, using interviews, Twitter analysis, and automated text analysis of 21,000 texts by 172 men and 43 women on the Salafi-oriented website saaid.net. To show the theory's generality, I also apply it to the contemporary white nationalist movement in the United States. The findings illustrate how movements that aggressively enforce traditional gender roles for participants can nevertheless increase female authority for pragmatic political reasons.	topics in religious texts	outcome	text	STM	Yes		Yes	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
11	AJPS	2020	64	1	In Song Kim, Steven Liao, Kosuke Imai	Measuring Trade Profile with Granular Product-Level Data	The product composition of bilateral trade encapsulates complex relationships about comparative advantage, global production networks, and domestic politics. Despite the availability of product-level trade data, most researchers rely on either the total volume of trade or certain sets of aggregated products. In this article, we develop a new dynamic clustering method to effectively summarize this massive amount of product-level information. The proposed method classifies a set of dyads into several clusters based on their similarities in trade profile—the product composition of imports and exports—and captures the evolution of the resulting clusters over time. We apply this method to two billion observations of product-level annual trade flows. We show how typical dyadic trade relationships evolve from sparse trade to interindustry trade and then to intra-industry trade. Finally, we illustrate the critical roles of our trade profile measure in international relations research on trade competition.	trade profile (assign a cluster membership to each dyad so that a set of dyads with similar trade profiles (i.e., product compositions of exports and imports) are grouped together.)	descriptive	product-level trade data	a new dynamic clustering method	No	Yes	No	
12	AJPS	2020	64	3	Anita R. Gohdes	Repression Technology: Internet Accessibility and State Violence	This article offers a first subnational analysis of the relationship between states' dynamic control of Internet access and their use of violent repression. I argue that where governments provide Internet access, surveillance of digital information exchange can provide intelligence that enables the use of more targeted forms of repression, in particular in areas not fully controlled by the regime. Increasing restrictions on Internet accessibility can impede opposition organization, but they limit access to information on precise targets, resulting in an increase in untargeted repression. I present new data on killings in the Syrian conflict that distinguish between targeted and untargeted events, using supervised text classification. I find that higher levels of Internet accessibility are associated with increases in targeted repression, whereas areas with limited access experience more indiscriminate campaigns of violence. The results offer important implications on how governments incorporate the selective access to communication technology into their strategies of coercion.	type of killing (targeted or untargeted)	outcome	text	extreme gradient booster	Yes		No	
13	AJPS	2020	64	4	Margaret E. Roberts, Brandon M. Stewart, Richard A. Nielsen	Adjusting for Confounding with Text Matching	We identify situations in which conditioning on text can address confounding in observational studies. We argue that a matching approach is particularly well-suited to this task, but existing matching methods are ill-equipped to handle high-dimensional text data. Our proposed solution is to estimate a low-dimensional summary of the text and condition on this summary via matching. We propose a method of text matching, topical inverse regression matching, that allows the analyst to match both on the topical content of confounding documents and the probability that each of these documents is treated. We validate our approach and illustrate the importance of conditioning on text to address confounding with two applications: the effect of perceptions of author gender on citation counts in the international relations literature and the effects of censorship on Chinese social media users.	topic proportions	confounder	text	topical inverse regression matching	Yes		No	
								topic proportions	confounder	text		Yes		No	
14	AJPS	2020	64	4	Anselm Hager, Hanno Hilbig	Does Public Opinion Affect Political Speech?	Does public opinion affect political speech? Of particular interest is whether public opinion affects (i) what topics politicians address and (ii) what positions they endorse. We present evidence from Germany where the government was recently forced to declassify its public opinion research, allowing us to link the content of the research to subsequent speeches. Our causal identification strategy exploits the exogenous timing of the research's dissemination to cabinet members within a window of a few days. We find that exposure to public opinion research leads politicians to markedly change their speech. First, we show that linguistic similarity between political speech and public opinion research increases significantly after reports are passed on to the cabinet, suggesting that politicians change the topics they address. Second, we demonstrate that exposure to public opinion research alters politicians' substantive positions in the direction of majority opinion.	topic of speech	outcome	text	support vector machine that takes the tf-idf document-term matrix	Yes		No	
15	APSR	2018	112	1	Jonathan B. Slapin, Justin H. Kirkland, Joseph A. Lazzaro, Patrick A. Leslie, Tom O'Grady	Ideology, Grandstanding, and Strategic Party Disloyalty in the British Parliament	Strong party discipline is a core feature of Westminster parliamentary systems. Parties typically compel members of Parliament (MPs) to support the party regardless of MPs' individual preferences. Rebellion, however, does occur. Using an original dataset of MP votes and speeches in the British House of Commons from 1992 to 2015, coupled with new estimations of MPs' ideological positions within their party, we find evidence that MPs use rebellion strategically to differentiate themselves from their party. The strategy that MPs employ is contingent upon an interaction of ideological extremity with party control of government. Extremists are loyal when their party is in the opposition, but these same extremists become more likely to rebel when their party controls government. Additionally, they emphasize their rebellion through speeches. Existing models of rebellion and party discipline do not account for government agenda control and do not explain these patterns.	extremism score	treatment	text	wordscores	Yes		No	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
16	APSR	2018			Amy Catalinac	Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos	We study a core question of interest in political science: Do candidates position themselves differently under different electoral systems and is their positioning in line with the expectations of spatial theories? We use validated estimates of candidate ideological positions derived from quantitative scaling of 7,497 Japanese-language election manifestos written by the near universe of candidates who competed in the eight House of Representatives elections held on either side of Japan's 1994 electoral reform. Leveraging variation before and after Japan's electoral reform, as well as within each electoral system, we find that candidates converge in single-member districts and diverge in multimember districts, and converge on copartisans when not faced with intraparty competition and diverge when they do. Our study helps to clarify debates about the effects of electoral systems on ideological polarization and party cohesion in Japan and more generally.	ideological position	outcome	text	wordfish	Yes		No	
17	APSR	2018	112	2	Hannes Mueller, Christopher Rauh	Reading Between the Lines: Prediction of Political Violence Using Newspaper Text	This article provides a new methodology to predict armed conflict by using newspaper text. Through machine learning, vast quantities of newspaper text are reduced to interpretable topics. These topics are then used in panel regressions to predict the onset of conflict. We propose the use of the within-country variation of these topics to predict the timing of conflict. This allows us to avoid the tendency of predicting conflict only in countries where it occurred before. We show that the within-country variation of topics is a good predictor of conflict and becomes particularly useful when risk in previously peaceful countries arises. Two aspects seem to be responsible for these features. Topics provide depth because they consist of changing, long lists of terms that make them able to capture the changing context of conflict. At the same time, topics provide width because they are summaries of the full text, including stabilizing factors.	topic of news article	NA	text	LDA topic model to estimate topic	NA		No	Interested only in forecasting, no causal claims
18	APSR	2018	112	3	Jennifer Pan, Kaiping Chen	Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances	A prerequisite for the durability of authoritarian regimes as well as their effective governance is the regime's ability to gather reliable information about the actions of lower-tier officials. Allowing public participation in the form of online complaints is one approach authoritarian regimes have taken to improve monitoring of lower-tier officials. In this paper, we gain rare access to internal communications between a monitoring agency and upper-level officials in China. We show that citizen grievances posted publicly online that contain complaints of corruption are systematically concealed from upper-level authorities when they implicate lower-tier officials or associates connected to lower-tier officials through patronage ties. Information manipulation occurs primarily through omission of wrongdoing rather than censorship or falsification, suggesting that even in the digital age, in a highly determined and capable regime where reports of corruption are actively and publicly voiced, monitoring the behavior of regime agents remains a challenge.	topic proportions	outcome	text	STM	Yes		Yes	
19	APSR	2018	112	4	Kenneth Lowande	Who Polices the Administrative State?	Scholarship on oversight of the bureaucracy typically conceives of legislatures as unitary actors. But most oversight is conducted by individual legislators who contact agencies directly. I acquire the correspondence logs of 16 bureaucratic agencies and re-evaluate the conventional proposition that ideological disagreement drives oversight. I identify the effect of this disagreement by exploiting the transition from George Bush to Barack Obama, which shifted the ideological orientation of agencies through turnover in agency personnel. Contrary to existing research, I find ideological conflict has a negligible effect on oversight, whereas committee roles and narrow district interests are primary drivers. The findings may indicate that absent incentives induced by public auditing, legislator behavior is driven by policy valence concerns rather than ideology. The results further suggest collective action in Congress may pose greater obstacles to bureaucratic oversight than previously thought.	type of contact (casework or policy)	outcome	text	supervised classifier	Yes		No	
20	APSR	2018	112	4	Sung Eun Kim	Media Bias against Foreign Firms as a Veiled Trade Barrier: Evidence from Chinese Newspapers	While the rules of international trade regimes prevent governments from employing protectionist instruments, governments continue to seek out veiled means of supporting their national industries. This article argues that the news media can serve as one channel for governments to favor domestic industries. Focusing on media coverage of auto recalls in China, I reveal a systematic bias against foreign automakers in those newspapers under strict government control. I further analyze subnational reporting patterns, exploiting variation in the level of regional government interest in the automobile industry. The analysis suggests that the media's home bias is driven by the government's protectionist interests but rules out the alternative hypothesis that home bias simply reflects the nationalist sentiment of readers. I show that this home bias in news coverage has meaningful impact on actual consumer behavior, combining automobile sales data and information on recall-related web searches.	topic proportion	outcome	text	STM	Yes		Yes	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
21	APSR	2019	113	1	Azusa Katagiri, Eric Min	The Credibility of Public and Private Signals: A Document-Based Approach	Crisis bargaining literature has predominantly used formal and qualitative methods to debate the relative efficacy of actions, public words, and private words. These approaches have overlooked the reality that policymakers are bombarded with information and struggle to adduce actual signals from endless noise. Material actions are therefore more effective than any diplomatic communication in shaping elites' perceptions. Moreover, while ostensibly "costless," private messages provide a more precise communication channel than public and "costly" pronouncements. Over 18,000 declassified documents from the Berlin Crisis of 1958–63 reflecting private statements, public statements, and White House evaluations of Soviet resolve are digitized and processed using statistical learning techniques to assess these claims. The results indicate that material actions have greater influence on the White House than either public or private statements; that public statements are noisier than private statements; and that private statements have a larger effect on evaluations of resolve than public statements.	beliefs about resolve	outcome	text	random forest	Yes		No	
22	APSR	2019	113	1	Tamar Mitts	From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West	What explains online radicalization and support for ISIS in the West? Over the past few years, thousands of individuals have radicalized by consuming extremist content online, many of whom eventually traveled overseas to join the Islamic State. This study examines whether anti-Muslim hostility might drive pro-ISIS radicalization in Western Europe. Using new geo-referenced data on the online behavior of thousands of Islamic State sympathizers in France, the United Kingdom, Germany, and Belgium, I study whether the intensity of anti-Muslim hostility at the local level is linked to pro-ISIS radicalization on Twitter. The results show that local-level measures of anti-Muslim animosity correlate significantly and substantively with indicators of online radicalization, including posting tweets sympathizing with ISIS, describing life in ISIS-controlled territories, and discussing foreign fighters. High-frequency data surrounding events that stir support for ISIS—terrorist attacks, propaganda releases, and anti-Muslim protests—show the same pattern.	topic of tweets (whether is pro-ISIS)	outcome	text	penalized logit	Yes		No	
23	APSR	2019	113	1	William Hobbs, Nazita Lajevardi	Effects of Divisive Political Campaigns on the Day-to-Day Segregation of Arab and Muslim Americans	How have Donald Trump's rhetoric and policies affected Arab and Muslim American behavior? We provide evidence that the de facto effects of President Trump's campaign rhetoric and vague policy positions extended beyond the direct effects of his executive orders. We present findings from three data sources—television news coverage, social media activity, and a survey—to evaluate whether Arab and Muslim Americans reduced their online visibility and retreated from public life. Our results provide evidence that they withdrew from public view: (1) Shared locations on Twitter dropped approximately 10 to 20% among users with Arabic-sounding names after major campaign and election events and (2) Muslim survey respondents reported increased public space avoidance.	dimensions of TV news coverage of Muslims	descriptive	TV news transcript	text scaling model (Hobbs 2017)	No		No	
24	APSR	2019	113	2	Ted Enamorado, Benjamin Fifield, Kosuke Imai	Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records	Since most social science research relies on multiple data sources, merging data sets is an essential part of researchers' workflow. Unfortunately, a unique identifier that unambiguously links records is often unavailable, and data may contain missing and inaccurate information. These problems are severe especially when merging large-scale administrative records. We develop a fast and scalable algorithm to implement a canonical model of probabilistic record linkage that has many advantages over deterministic methods frequently used by social scientists. The proposed methodology efficiently handles millions of observations while accounting for missing data and measurement error, incorporating auxiliary information, and adjusting for uncertainty about merging in post-merge analyses. We conduct comprehensive simulation studies to evaluate the performance of our algorithm in realistic scenarios. We also apply our methodology to merging campaign contribution records, survey data, and nationwide voter files. An open-source software package is available for implementing the proposed methodology.	identify which data observation are identical, simialr or different	descriptive	number or string	Canonical Model of Probabilistic Record Linkage	NA		Yes	Methods article without measurement component
25	APSR	2019	113	2	Gregory J. Martin, Joshua McCrain	Local News and National Politics	The level of journalistic resources dedicated to coverage of local politics is in a long-term decline in the US news media, with readership shifting to national outlets. We investigate whether this trend is demand- or supply-driven, exploiting a recent wave of local television station acquisitions by a conglomerate owner. Using extensive data on local news programming and viewership, we find that the ownership change led to (1) substantial increases in coverage of national politics at the expense of local politics, (2) a significant rightward shift in the ideological slant of coverage, and (3) a small decrease in viewership, all relative to the changes at other news programs airing in the same media markets. These results suggest a substantial supply-side role in the trends toward nationalization and polarization of politics news, with negative implications for accountability of local elected officials and mass polarization.	news topic proportion	outcome (used in a two-step process with LDA to measure national news reporting, then similarity to the Congressional Record for left-right scaling)	text	LDA topic model to estimate topic	Yes		No	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
26	APSR	2019	113	3	Ramya Parthasarathy, Vijayendra Rao, NethrA Palaniswamy	Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies	This paper opens the "black box" of real-world deliberation by using text-as-data methods on a corpus of transcripts from the constitutionally mandated gram sabhas, or village assemblies, of rural India. Drawing on normative theories of deliberation, we identify empirical standards for "good" deliberation based on one's ability both to speak and to be heard, and use natural language processing methods to generate these measures. We first show that, even in the rural Indian context, these assemblies are not mere "talking shops," but rather provide opportunities for citizens to challenge their elected officials, demand transparency, and provide information about local development needs. Second, we find that women are at a disadvantage relative to men; they are less likely to speak, set the agenda, and receive a relevant response from state officials. And finally, we show that quotas for women for village presidencies improve the likelihood that female citizens are heard.	topic proportion	outcome	text	STM	Yes		Yes	
27	APSR	2019	113	3	Devin Caughey, Tom O'Grady, Christopher Warshaw	Policy Ideology in European Mass Publics, 1981–2016	Using new scaling methods and a comprehensive public opinion dataset, we develop the first survey-based time-series–cross-sectional measures of policy ideology in European mass publics. Our dataset covers 27 countries and 36 years and contains nearly 2.7 million survey responses to 109 unique issue questions. Estimating an ordinal group-level IRT model in each of four issue domains, we obtain biennial estimates of the absolute economic conservatism, relative economic conservatism, social conservatism, and immigration conservatism of men and women in three age categories in each country. Aggregating the group-level estimates yields estimates of the average conservatism in national publics in each biennium between 1981–82 and 2015–16. The four measures exhibit contrasting cross-sectional cleavages and distinct temporal dynamics, illustrating the multidimensionality of mass ideology in Europe. Subjecting our measures to a series of validation tests, we show that the constructs they measure are distinct and substantively important and that they perform as well as or better than one-dimensional proxies for mass conservatism (left–right self-placement and median voter scores). We foresee many uses for these scores by scholars of public opinion, electoral behavior, representation, and policy feedback.	mass policy conservatism	descriptive	survey data	ordinal DGIRT model	No	Yes	Yes	
28	APSR	2019	113	3	Francisco Cantú	The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election	This paper investigates the opportunities for non-democratic regimes to rely on fraud by documenting the alteration of vote tallies during the 1988 presidential election in Mexico. In particular, I study how the alteration of vote returns came after an electoral reform that centralized the vote-counting process. Using an original image database of the vote-tally sheets for that election and applying Convolutional Neural Networks (CNN) to analyze the sheets, I find evidence of blatant alterations in about a third of the tallies in the country. This empirical analysis shows that altered tallies were more prevalent in polling stations where the opposition was not present and in states controlled by governors with grassroots experience of managing the electoral operation. This research has implications for understanding the ways in which autocrats control elections as well as for introducing a new methodology to audit the integrity of vote tallies.	whether the vote tally is altered	outcome	image	CNN: The inputs of the images consists of numerical arrays of 3 (RGB values) * 227 (height) * 227 (width) pixel values. The network contains six convoluted layers of 32, 32, 64, 64, 128, and 256 filters, respectively.	Yes		No	
29	APSR	2019	113	4	Pablo Barbera, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, Joshua A. Tucker	Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data	Are legislators responsive to the priorities of the public? Research demonstrates a strong correspondence between the issues about which the public cares and the issues addressed by politicians, but conclusive evidence about who leads whom in setting the political agenda has yet to be uncovered. We answer this question with fine-grained temporal analyses of Twitter messages by legislators and the public during the 113th US Congress. After employing an unsupervised method that classifies tweets sent by legislators and citizens into topics, we use vector autoregression models to explore whose priorities more strongly predict the relationship between citizens and politicians. We find that legislators are more likely to follow, than to lead, discussion of public issues, results that hold even after controlling for the agenda-setting effects of the media. We also find, however, that legislators are more likely to be responsive to their supporters than to the general public.	topic proprotion of tweets	outcome	text	LDA topic model to estimate topic	Yes		No	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
30	APSR	2019	113	4	James Bisbee	BARP: Improving Mister P Using Bayesian Additive Regression Trees	Multilevel regression and post-stratification (MRP) is the current gold standard for extrapolating opinion data from nationally representative surveys to smaller geographic units. However, innovations in nonparametric regularization methods can further improve the researcher's ability to extrapolate opinion data to a geographic unit of interest. I test an ensemble of regularization algorithms and find that there is room for substantial improvement on the multilevel model via more flexible methods of regularization. I propose a modified version of MRP that replaces the multilevel model with a nonparametric approach called Bayesian additive regression trees (BART or, when combined with post-stratification, BARP). I compare both methods across a number of data contexts, demonstrating the benefits of applying more powerful regularization methods to extrapolate opinion data to target geographical units. I provide an R package that implements the BARP method.	coefficients of covariates which predict opinon	descriptive	survey data	BARP	NA		NA	No mention of uncertainty but trivially easy to handle with proposed model
31	APSR	2020	114	1	Christopher Claassen	In the Mood for Democracy? Democratic Support as Thermostatic Opinion	Public support has long been thought crucial for the vitality and survival of democracy. Existing research has argued that democracy also creates its own demand: through early-years socialization and later-life learning, the presence of a democratic system coupled with the passage of time produces widespread public support for democracy. Using new panel measures of democratic mood varying over 135 countries and up to 30 years, this article finds little evidence for such a positive feedback effect of democracy on support. Instead, it demonstrates a negative thermostatic effect: increases in democracy depress democratic mood, while decreases cheer it. Moreover, it is increases in the liberal, counter-majoritarian aspects of democracy, not the majoritarian, electoral aspects that provoke this backlash from citizens. These novel results challenge existing research on support for democracy, but also reconcile this research with the literature on macro-opinion.	democratic mood	outcome	survey data	Bayesian dynmaic latent variable model	Yes		No	
32	APSR	2020	114	1	L. Jason Anastasopoulos, Anthony M. Bertelli	Understanding Delegation Through Machine Learning: A Method and Application to the European Union	Delegation of powers represents a grant of authority by politicians to one or more agents whose powers are determined by the conditions in enabling statutes. Extant empirical studies of this problem have relied on labor-intensive content analysis that ultimately restricts our knowledge of how delegation has responded to politics and institutional change in recent years. We present a machine learning approach to the empirical estimation of authority and constraint in European Union (EU) legislation, and demonstrate its ability to accurately generate the same discretionary measures used in an original study directly using all EU directives and regulations enacted between 1958–2017. We assess validity by training our classifier on a random sample of only 10% of hand-coded provisions and replicating an important substantive finding. While our principal interest lies in delegation, our method is extensible to any context in which human coding has been profitably produced.	a provision being classified as delegation or imposing constraint	descriptive	text	gradient-boosted tree (GBT) text classifiers	No	Yes	No	
33	APSR	2020	114	2	Kyle Peyton	Does Trust in Government Increase Support for Redistribution? Evidence from Randomized Survey Experiments	Why have decades of high and rising inequality in the United States not increased public support for redistribution? An established theory in political science holds that Americans' distrust of government decreases their support for redistribution, but empirical support draws primarily on regression analyses of national surveys. I discuss the untestable assumptions required for identification with regression modeling and propose an alternative design that uses randomized experiments about political corruption to identify the effect of trust in government on support for redistribution under weaker assumptions. I apply this to three survey experiments and estimate the effects that large, experimentally induced increases in political trust have on support for redistribution. Contrary to theoretical predictions, estimated effects are substantively negligible, statistically indistinguishable from zero, and comparable to estimates from two placebo experiments. I discuss implications for theory building about causes of support for redistribution in an era of rising inequality and eroding confidence in government.	NA	NA	experimental data	generalized random forests (GRF)	NA		NA	Methods article without measurement component
34	APSR	2020	114	3	Jane Esberg	Censorship as Reward: Evidence from Pop Culture Censorship in Chile	Censorship has traditionally been understood as a way for dictators to silence opposition. By contrast, this article develops and tests the theory that certain forms of censorship—in particular, prohibitions on popular culture—serve not only to limit political information but also to reward dictators' supporters. Using text analysis of all 8,000 films reviewed for distribution during Chile's dictatorship, I demonstrate that rather than focusing only on sensitive political topics, censors banned movies containing content considered immoral. Through a combination of qualitative and quantitative evidence, I show that these patterns cannot be explained by masked political content, distributor self-censorship, or censor preferences. Instead, they reflect the regime's use of censorship as a reward for supporters, particularly conservative Catholic groups. My findings suggest that even repressive measures can be used in part to maintain support for authoritarian regimes.	themes in film	treatment	text	supervised Indian Buffet Process (sIBP)	Yes		No	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
35	APSR	2020	114	3	Baekkwan Park, Kevin Greene, Michael Colaresi	Human Rights are (Increasingly) Plural: Learning the Changing Taxonomy of Human Rights from Large-scale Text Reveals Information Effects	This manuscript helps to resolve the ongoing debate concerning the effect of information communication technology on human rights monitoring. We reconceptualize human rights as a taxonomy of nested rights that are judged in textual reports and argue that the increasing density of available information should manifest in deeper taxonomies of human rights. With a new automated system, using supervised learning algorithms, we are able to extract the implicit taxonomies of rights that were judged in texts by the US State Department, Amnesty International, and Human Rights Watch over time. Our analysis provides new, clear evidence of change in the structure of these taxonomies as well as in the attention to specific rights and the sharpness of distinctions between rights. Our findings bridge the natural language processing and human rights communities and allow a deeper understanding of how changes in technology have affected the recording of human rights over time.	taxonomy of human rights	outcome	text	automated parser that detects the aspect and judgment phrases in human right reports (Parsing Unstructured Language into Sentiment–Aspect Representations (PULSAR))	Yes		No	
36	APSR	2020	114	4	Beatriz Magaloni, Luis Rodriguez	Institutionalized Police Brutality: Torture, the Militarization of Security, and the Reform of Inquisitorial Criminal Justice in Mexico	How can societies restrain their coercive institutions and transition to a more humane criminal justice system? We argue that two main factors explain why torture can persist as a generalized practice even in democratic societies: weak procedural protections and the militarization of policing, which introduces strategies, equipment, and mentality that treats criminal suspects as though they were enemies in wartime. Using a large survey of the Mexican prison population and leveraging the date and place of arrest, this paper provides causal evidence about how these two explanatory variables shape police brutality. Our paper offers a grim picture of the survival of authoritarian policing practices in democracies. It also provides novel evidence of the extent to which the abolition of inquisitorial criminal justice institutions—a remnant of colonial legacies and a common trend in the region—has worked to restrain police brutality.	topic proportions	outcome	text	STM	Yes		No	
37	APSR	2020	114	4	Jesse M. Crosson, Alexander C. Furnas, Geoffrey M. Lorenz	Polarized Pluralism: Organizational Preferences and Biases in the American Pressure System	For decades, critics of pluralism have argued that the American interest group system exhibits a significantly biased distribution of policy preferences. We evaluate this argument by measuring groups' revealed preferences directly, developing a set of ideal point estimates, IGscores, for over 2,600 interest groups and 950 members of Congress on a common scale. We generate the scores by jointly scaling a large dataset of interest groups' positions on congressional bills with roll-call votes on those same bills. Analyses of the scores uncover significant heterogeneity in the interest group system, with little conservative skew and notable inter-party differences in preference correspondence between legislators and ideologically similar groups. Conservative bias and homogeneity reappear, however, when weighting IGscores by groups' PAC contributions and lobbying expenditures. These findings suggest that bias among interest groups depends on the extent to which activities like PAC contributions and lobbying influence policymakers' perceptions about the preferences of organized interests.	ideal point	descriptive	position and votes	Bayesian IRT	No	Yes	Yes	
38	APSR	2020	114	4	Jesse Yoder	Does Property Ownership Lead to Participation in Local Politics? Evidence from Property Records and Meeting Minutes	Homeowners and renters have participated in politics at different rates throughout American history, but does becoming a property owner motivate an individual to participate in local politics? I combine deed-level property records in California and Texas with an original dataset on individual comments in local city council meetings to study the role of property ownership in shaping costly forms of political behavior, and I document large inequalities in who participates at city council meetings. I also link property records to individual-level contribution records and administrative voter files and find that becoming a property owner increases an individual's political activity. Over and above voting in local elections, property ownership motivates individuals to participate in local city council meetings and donate to candidates. These findings illustrate how the experience of homeownership leads property owners to become much more active in local politics.	topic proportion	outcome	text	STM	Yes		Yes	
39	APSR	2020	114	4	Jeff Carter, Charles E. Smith Jr.	A Framework for Measuring Leaders' Willingness to Use Force	Political leaders' willingness to use force is central to many explanations of foreign policy and interstate conflict. Unfortunately, existing indicators typically measure one aspect of this general concept, have limited coverage, and/or are not derived independently of leaders' participation in interstate conflicts. We develop a strategy for constructing measures of leaders' underlying willingness to use force with data on their background experiences, political orientations, and psychological traits in a Bayesian latent variable framework. Our approach produces measures of latent hawkishness for all national leaders between 1875 and 2004 that offer advantages over existing proxies along multiple dimensions, including construct validity, predictive validity, and measurement uncertainty. Importantly, our statistical framework allows scholars to build upon our measures by incorporating additional data and altering the assumptions underlying our models.	hawkishness of political leaders	descriptive	number	Bayesian latent variable	No	Yes	Yes	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
40	JOP	2018	80	2	Jack Blumenau, Benjamin E. Lauderdale	Never Let a Good Crisis Go to Waste: Agenda Setting and Legislative Voting in Response to the EU Crisis	The European Union's policy response to the recent global economic crisis transferred significant powers from the national to the European level. When exogenous shocks make status quo policies less attractive, legislators become more tolerant to proposed alternatives, and the policy discretion of legislative agenda setters increases. Given control of the EU agenda-setting process by pro-integration actors, we argue that this dynamic explains changes in voting patterns of the European Parliament during the crisis period. We observe voting coalitions increasingly dividing legislators along the pro-anti integration, rather than the left-right dimension of disagreement, but only in policy areas related to the crisis. In line with more qualitative assessments of the content of passed legislation, the implication is that pro-integration actors were able to shift policy further toward integration than they could have without the crisis.	crisis relevance	treatment	text	topic model with two stages OLS	Yes		No	
41	JOP	2018	80	4	Lisa Blaydes, Justin Grimmer, Alison McQueen	Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds	When did European modes of political thought diverge from those that existed in other world regions? We compare Muslim and Christian political advice texts from the medieval period using automated text analysis to identify four major and 60 granular themes common to Muslim and Christian polities, and examine how emphasis on these topics evolves over time. For Muslim texts, we identify an inflection point in political discourse between the eleventh to thirteenth centuries, a juncture that historians suggest is an ideational watershed brought about by the Turkic and Mongol invaders. For Christian texts, we identify a decline in the relevance of religious appeals from the Middle Ages to the Renaissance. Our findings also suggest that Machiavelli's Prince was less a turn away from religious discourse on statecraft than the culmination of centuries-long developments in European advice literature.	topic proportion of books	outcome	text	The model (1) estimates a set of specific themes, (2) estimates a set of broad themes, and (3) classifies each specific theme into a single broad theme. For each of the 46 books in the collection (i =1, ... , 46) the model (4) estimates how each book divides its attention over the 60 specific themes.	No		Yes	
42	JOP	2019	81	1	Kenneth Lowande	Politicization and Responsiveness in Executive Agencies	Scholarship on bureaucratic responsiveness to Congress typically focuses on delegation and formal oversight hearings. Overlooked are daily requests to executive agencies made by legislators that propose policies, communicate concerns, and request information or services. Analyzing over 24,000 of these requests made to 13 executive agencies between 2007 and 2014, I find agencies systematically prioritize the policy-related requests of majority party legislators—but that this effect can be counteracted when presidents politicize agencies through appointments. An increase in politicization produces a favorable agency bias toward presidential copartisans. This same politicization, however, has a net negative impact on agency responsiveness—agencies are less responsive to members of Congress, but even less responsive to legislators who are not presidential copartisans. Critically, this negative impact extends beyond policy-related requests to cases of constituency service. The results suggest that presidential appointees play an important, daily mediating role between Congress and the bureaucracy.	type of requests	outcome	text	supervised classifier	Yes		No	
43	JOP	2019	81	2	Jean-Christophe Boucher, Cameron G. Thies	"I Am a Tariff Man": The Power of Populist Foreign Policy Rhetoric under President Trump	This article contributes to the emerging literature on populist foreign policy by examining President Trump's ability to dominate and shape public discourse on trade. We develop an ideational approach to populism that focuses on the social network that emerges surrounding a populist leader's discourse. We hypothesize that populist leaders will generate a polarized social network along the elite-versus-people divide instead of the usual partisan boundary. Populist leaders like Trump are known to prefer direct, unmediated access to the people in order to spread their ideology. We therefore examine Trump's use of Twitter as he announced his steel and aluminum tariffs in March 2018 and its impact on the salience and content of debates around trade policy on the Twittersphere. Our findings highlight how Trump and his supporters use populist foreign policy themes to articulate their policy positions on social media.	narratives of tweets	descriptive	text	naive bayes	No		No	

	Journal	Year	Volume	Issue	Author	Title	Abstract	Estimated Variable	Used as	Features	Model	Learned proxy in second-stage model (or jointly fit measurement model and estimate of group differences)?	If no second stage model, is the contribution of the paper a proxy for use in other models?	Handles Uncertainty	If NA, why?
44	JOP	2019	81	3	Shea Streeter	Lethal Force in Black and White: Assessing Racial Disparities in the Circumstances of Police Killings	African Americans are nearly three times as likely to be killed by police as whites. This paper examines whether this racial disparity is due in part to racial differences in the circumstances of police killings. To assess whether and how these circumstances predict the race of a decedent, I use machine learning techniques and a novel data set of police killings containing over 120 descriptors. I find that decedent characteristics, criminal activity, threat levels, police actions, and the setting of the lethal interaction are not predictive of race, indicating that the police—given contact—are killing blacks and whites under largely similar circumstances. The findings suggest that the racial disparity in the rate of lethal force is most likely driven by higher rates of police contact among African Americans rather than racial differences in the circumstances of the interaction and officer bias in the application of lethal force.	NA	NA	number	random forest, lasso regression, SVM, neural network	NA		NA	Used for flexible functional form, not to learn a proxy
45	JOP	2019	81	4	Matthew J. Lacombe	The Political Weaponization of Gun Owners: The National Rifle Association's Cultivation, Dissemination, and Use of a Group Social Identity	There is substantial evidence indicating that the NRA's (National Rifle Association) political influence is closely tied to the deep political engagement of the minority of Americans who oppose strict gun control laws. This explanation of the NRA's influence, however, raises its own questions; namely, why are gun owners so devoted to their cause, and why is the NRA so effective at mobilizing them? I marshal a wide range of evidence covering nearly nine decades to demonstrate that an important cause of the political activity of gun owners is the NRA's long-term cultivation and dissemination of a distinct, politicized gun owner social identity, which the NRA uses to mobilize mass political action on its behalf. My findings shed new light on the ability of interest groups to develop and use resources that help them influence policy by altering the political behavior of members of the mass public.	NA	NA	text	STM: estimate topics of editorials; supervised classifier: code if use identity-language	NA		NA	STM used for sample selection
46	JOP	2019	81	4	Michael Horowitz, Brandon M. Stewart, Dustin Tingley, Michael Bishop, Laura Resnick Samotin, Margaret Roberts, Welton Chang, Barbara Mellers, Philip Tetlock	What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting	When do groups—be they countries, administrations, or other organizations—more or less accurately understand the world around them and assess political choices? Some argue that group decision-making processes often fail due to biases induced by groupthink. Others argue that groups, by aggregating knowledge, are better at analyzing the foreign policy world. To advance knowledge about the intersection of politics and group decision making, this paper draws on evidence from a multiyear geopolitical forecasting tournament with thousands of participants sponsored by the US government. We find that teams outperformed individuals in making accurate geopolitical predictions, with regression discontinuity analysis demonstrating specific teamwork effects. Moreover, structural topic models show that more cooperative teams outperformed less cooperative teams. These results demonstrate that information sharing through groups, cultivating reasoning to hedge against cognitive biases, and ensuring all perspectives are heard can lead to greater success for groups at forecasting and understanding politics.	topic proportion	outcome	text	STM	Yes		Yes	
47	JOP	2020	82	1	Junyan Jiang, Yu Zeng	Countering Capture: Elite Networks and Government Responsiveness in China's Land Market Reform	Government responsiveness is often viewed as a result of political pressure from the public, but why do politicians facing similar pressure sometimes differ in their responsiveness? This article considers the configurations of elite networks as a key mediating factor. We argue that access to external support networks helps improve politicians' responsiveness to ordinary citizens by reducing their dependence on vested interests, and we test this claim using China's land market reform as a case. Leveraging novel city-level measures of mass grievances and political networks, we demonstrate that the intensity of land-related grievances is on average positively associated with reform occurrence, but this association is only salient among a subset of city leaders who enjoy informal connections to the higher-level authority. We also show that connected leaders tend to implement policies less congruent with local bureaucratic and business interests. These findings underscore the importance of intra-elite dynamics in shaping mass-elite interactions.	topic of online petitions	treatment	text	topic model	Yes		No	
48	JOP	2020	82	2	Scott de Marchi, Michael Laver	Government Formation as Logrolling in High-Dimensional Issue Spaces	Analytical models of government formation typically assume low-dimensional real policy spaces. Behaviorally, however, politicians negotiate to form governments in high-dimensional discrete issue spaces. We model these negotiations, leveraging the fact that different politicians typically attach different importance to the same issue, allowing gains from trade when they negotiate agreed positions on large packages of issues. The set of issues in an agreed package is endogenous; politicians need not agree on every issue before they go into government together, "tabling" issues on which they agree to disagree. We exercise our model computationally, calibrating it to 91 real-world government formation settings, and mapping out the relative probability of Condorcet winning cabinets in different settings. This probability measures how hard it is for negotiators to find Condorcet winning cabinets in a giant high-dimensional state space. We test this claim empirically with a statistical model of the duration of negotiations after an election.	NA	NA	number	LASSO	NA		NA	LASSO used for variable selection

Appendix References

L. J. Anastasopoulos and A. M. Bertelli. Understanding delegation through machine learning: A method and application to the european union. *American Political Science Review*, 114(1):291–301, 2020.

P. Barberá, A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113 (4):883–901, 2019.

L. Barnes and T. Hicks. Making austerity popular: the media and mass attitudes toward fiscal policy. *American Journal of Political Science*, 62(2):340–354, 2018.

K. Benoit, K. Munger, and A. Spirling. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508, 2019.

J. Bisbee. Barp: Improving mister p using bayesian additive regression trees. *American Political Science Review*, 113(4):1060–1065, 2019.

L. Blaydes, J. Grimmer, and A. McQueen. Mirrors for princes and sultans: Advice on the art of governance in the medieval christian and islamic worlds. *The Journal of Politics*, 80(4):1150–1167, 2018.

J. Blumenau and B. E. Lauderdale. Never let a good crisis go to waste: Agenda setting and legislative voting in response to the eu crisis. *The Journal of Politics*, 80(2): 462–478, 2018.

A. Bonica. Inferring roll-call scores from campaign contributions using supervised machine learning. *American Journal of Political Science*, 62(4):830–848, 2018.

J.-C. Boucher and C. G. Thies. "i am a tariff man": the power of populist foreign policy rhetoric under president trump. *The Journal of Politics*, 81(2):712–722, 2019.

F. Cantú. The fingerprints of fraud: Evidence from mexico's 1988 presidential election. *American Political Science Review*, 113(3):710–726, 2019.

R. J. Carroll and B. Kenkel. Prediction, proxies, and power. *American Journal of Political Science*, 63(3):577–593, 2019.

J. Carter and C. E. Smith. A framework for measuring leaders' willingness to use force. *American Political Science Review*, 114(4):1352–1358, 2020.

A. Casas, M. J. Denny, and J. Wilkerson. More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process. *American Journal of Political Science*, 64(1):5–18, 2020.

A. Catalinac. Positioning under alternative electoral systems: evidence from japanese candidate election manifestos. *American Political Science Review*, 112(1):31–48, 2018.

D. Caughey, T. O'GRADY, and C. Warshaw. Policy ideology in european mass publics, 1981–2016. *American Political Science Review*, 113(3):674–693, 2019.

C. Claassen. In the mood for democracy? democratic support as thermostatic opinion. *American Political Science Review*, 114(1):36–53, 2020.

J. M. Crosson, A. C. Furnas, and G. M. Lorenz. Polarized pluralism: organizational preferences and biases in the american pressure system. *American Political Science Review*, 114(4):1117–1137, 2020.

S. de Marchi and M. Laver. Government formation as logrolling in high-dimensional issue spaces. *The Journal of Politics*, 82(2):543–558, 2020.

T. Enamorado, B. Fifield, and K. Imai. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2):353–371, 2019.

J. Esberg. Censorship as reward: Evidence from pop culture censorship in chile. *American Political Science Review*, 114(3):821–836, 2020.

A. R. Gohdes. Repression technology: Internet accessibility and state violence. *American Journal of Political Science*, 64(3):488–503, 2020.

A. Hager and H. Hilbig. Does public opinion affect political speech? *American Journal of Political Science*, 64(4):921–937, 2020.

P. K. Hatemi, C. Crabtree, and K. B. Smith. Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, 63(4):788–806, 2019.

W. Hobbs and N. Lajevardi. Effects of divisive political campaigns on the day-to-day segregation of arab and muslim americans. *American Political Science Review*, 113(1): 270–276, 2019.

M. Horowitz, B. M. Stewart, D. Tingley, M. Bishop, L. Resnick Samotin, M. Roberts, W. Chang, B. Mellers, and P. Tetlock. What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting. *The Journal of Politics*, 81(4):1388–1404, 2019.

C. Huff and J. D. Kertzer. How the public defines terrorism. *American Journal of Political Science*, 62(1):55–71, 2018.

J. Jiang and Y. Zeng. Countering capture: Elite networks and government responsiveness in china's land market reform. *The Journal of Politics*, 82(1):13–28, 2020.

Z. M. Jones and Y. Lupu. Is there more violence in the middle? *American Journal of Political Science*, 62(3):652–667, 2018.

A. Katagiri, E. Min, et al. The credibility of public and private signals: A document-based approach. *American Political Science Review*, 113(1):156–172, 2019.

I. S. Kim, S. Liao, and K. Imai. Measuring trade profile with granular product-level data. *American Journal of Political Science*, 64(1):102–117, 2020.

S. E. Kim et al. Media bias against foreign firms as a veiled trade barrier: Evidence from chinese newspapers. *American Political Science Review*, 112(4):954–970, 2018.

M. J. Lacombe. The political weaponization of gun owners: The national rifle association's cultivation, dissemination, and use of a group social identity. *The Journal of Politics*, 81(4):1342–1356, 2019.

K. Lowande. Politicization and responsiveness in executive agencies. *The Journal of Politics*, 81(1):33–48, 2019.

K. Lowande et al. Who polices the administrative state? *American Political Science Review*, 112(4):874–890, 2018.

B. Magaloni and L. Rodriguez. Institutionalized police brutality: Torture, the militarization of security, and the reform of inquisitorial criminal justice in mexico. *American Political Science Review*, 114(4):1013–1034, 2020.

G. J. Martin and J. McCrain. Local news and national politics. *American Political Science Review*, 113(2):372–384, 2019.

T. Mitts. From isolation to radicalization: anti-muslim hostility and support for isis in the west. *American Political Science Review*, 113(1):173–194, 2019.

J. M. Montgomery and S. Olivella. Tree-based models for political science data. *American Journal of Political Science*, 62(3):729–744, 2018.

H. MUELLER and C. RAUH. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2):358–375, 2018. doi: 10.1017/S0003055417000570.

R. A. Nielsen. Women's authority in patriarchal social movements: The case of female salafi preachers. *American Journal of Political Science*, 64(1):52–66, 2020.

J. Pan and K. Chen. Concealing corruption: How chinese officials distort upward reporting of online grievances. *The American Political Science Review*, 112(3):602–620, 2018.

B. Park, K. Greene, and M. Colaresi. Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects. *American Political Science Review*, 114(3):888–910, 2020.

R. Parthasarathy, V. Rao, and N. Palaniswamy. Deliberative democracy in an unequal world: A text-as-data study of south india's village assemblies. *American Political Science Review*, 113(3):623–640, 2019.

K. Peyton. Does trust in government increase support for redistribution? evidence from randomized survey experiments. *American Political Science Review*, 114(2):596–602, 2020.

M. E. Roberts, B. M. Stewart, and R. A. Nielsen. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903, 2020.

J. B. Slapin, J. H. Kirkland, J. A. Lazzaro, P. A. Leslie, and T. O'grady. Ideology, grandstanding, and strategic party disloyalty in the british parliament. *American Political Science Review*, 112(1):15–30, 2018.

S. Streeter. Lethal force in black and white: Assessing racial disparities in the circumstances of police killings. *The Journal of Politics*, 81(3):1124–1132, 2019.

J. Yoder. Does property ownership lead to participation in local politics? evidence from property records and meeting minutes. *American Political Science Review*, 114(4): 1213–1229, 2020.