

Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods

Justin Grimmer¹, Solomon Messing² and Sean J. Westwood³

¹ Associate Professor, Department of Political Science, University of Chicago, 5828 S. University Ave., Chicago, IL 60637, USA.
Email: jgrimmer@stanford.edu

² Director, Data Labs, Pew Research Center, 1615 L Street NW, Washington, DC, USA

³ Assistant Professor, Department of Government, Dartmouth College, USA

Abstract

Randomized experiments are increasingly used to study political phenomena because they can credibly estimate the average effect of a treatment on a population of interest. But political scientists are often interested in how effects vary across subpopulations—heterogeneous treatment effects—and how differences in the content of the treatment affects responses—the response to heterogeneous treatments. Several new methods have been introduced to estimate heterogeneous effects, but it is difficult to know if a method will perform well for a particular data set. Rather than using only one method, we show how an ensemble of methods—weighted averages of estimates from individual models increasingly used in machine learning—accurately measure heterogeneous effects. Building on a large literature on ensemble methods, we show how the weighting of methods can contribute to accurate estimation of heterogeneous treatment effects and demonstrate how pooling models lead to superior performance to individual methods across diverse problems. We apply the ensemble method to two experiments, illuminating how the ensemble method for heterogeneous treatment effects facilitates exploratory analysis of treatment effects.

1 Introduction

Experiments are increasingly used to test theories of politics and political conflict (Gerber and Green 2012). Experiments are used because they provide credible estimates of the effect of an intervention for a sample population. But underlying this average effect for a sample may be substantial variation in how particular respondents respond to treatments: there may be heterogeneous treatment effects (Athey and Imbens 2015). This variation may provide theoretical insights, revealing how the effect of interventions depends on participants' characteristics or how varying features of a treatment alter the effect of an intervention. The variation may also be practically useful, providing guidance on how to optimally administer treatments (Imai and Strauss 2011; Imai and Ratkovic 2013), or it may be useful for extrapolating the findings of an experiment to a broader population of interest (Hartman *et al.* 2012). Further, scholars are increasingly making use of experimental designs with many conditions, in order to examine how differences in treatment content affects response—the effect of heterogeneous treatments (Hainmueller and Hopkins 2015; Hainmueller, Hopkins, and Yamamoto 2014).

A growing literature has contributed new methods for estimating heterogeneous effects and the effects of heterogeneous treatments (Hastie, Tibshirani and Friedman 2001; Imai and Strauss 2011; Green and Kern 2012; Imai and Ratkovic 2013; Hainmueller and Hazlett 2013; Athey and Imbens 2015). Each of the methods provide new and important insights into how to reliably capture heterogeneity in treatment response or how individuals respond to high-dimensional treatments. For example, Ratkovic and Tingley (2017) provide a sparse estimation strategy for identifying

Political Analysis (2017)

DOI: 10.1017/pan.2017.15

Corresponding author
Justin Grimmer

Edited by
Dustin Tingley

© The Author(s) 2017. Published
by Cambridge University Press
on behalf of the Society for
Political Methodology.

Authors' note: Replication data available in Grimmer, Messing, and Westwood (2017).

heterogeneous treatment effects. To identify systematic variation in treatment response and to separate it from variation due to simple randomness each of the new methods combines information in the data with necessary and consequential assumptions about the data generating process (Hastie, Tibshirani and Friedman 2001). While the assumptions are often minimal and designed to maximize a method's flexibility, it is difficult to know before hand if a method's particular assumptions fit any one application well.

Rather than rely on a single method to estimate heterogeneous treatment effects, we show how a weighted average of methods for estimating heterogeneous effects—an ensemble—provides accurate estimates across diverse problems. We build on the ensemble method *super learning* (van der Laan, Polley, and Hubbard 2007), using a cross-validated measure of prediction performance to weight the contribution of methods to the final estimate of heterogeneous effects and show the close relationship of super learning to other ensemble methods (van der Laan, Polley, and Hubbard 2007; Hillard, Purpura and Wilkerson 2008; Montgomery, Hollenbach, and Ward 2012). Weighting based on out-of-sample performance is useful, we show, because methods that tend to perform well in cross validation prediction tasks also accurately estimate heterogeneous effects. Using Monte Carlo simulations we show that the ensemble outperforms constituent methods across diverse problems because the ensemble attaches greater weight to methods that have better estimates of the heterogeneous effects for the particular task at hand and that each method's performance varies across contexts.

We apply the ensemble method to two experiments that examine how constituents evaluate how legislators' claim credit for particularistic spending in the district and criticism of those credit-claiming efforts (Mayhew 1974; Grimmer, Messing, and Westwood 2012). In both examples we use the heterogeneous treatment effect method for explicitly exploratory purposes—to generate hypotheses useful for future rounds of experimentation. The exploratory purposes, however, are useful for other applications of heterogeneous treatment effect models: targeting particularly responsive subpopulations (Imai and Ratkovic 2013) and extrapolating treatment effects to new samples (Hartman *et al.* 2012). We also provide guidance on how to calculate standard errors for the heterogeneous treatment effects and provide a new visualization to reflect the data underlying the heterogeneous treatment effects. Our application of ensemble methods to the estimation of heterogeneous treatment effects builds on a growing literature in political science that uses super learning methods to improve inferences, including making more plausible comparisons in observational studies (Samii, Paler, and Daly 2017).

In our first application we show how ensembles can be used to estimate the response to heterogeneous treatments, facilitating an examination of how constituents evaluate credit claiming messages. Our results suggest that constituents focus on easily acquired information—such as the type of project the legislator claims credit for, rather than the amount of money allocated to the project. In our second application we show how ensembles can be used to model heterogeneity in response to a treatment, revealing ideological heterogeneity in response to criticism for government expenditures. In both examples the ensemble of heterogeneous treatment effect methods suggests new hypotheses to be tested in future experiments, where we can explicitly use a pre-analysis plan to test the variation.

Throughout the paper we explain that ensembles are useful because they are flexible and can be tuned to the particular problem at hand. They are also useful because they ensure that we make full use of impressive methodological innovations in the estimation of heterogeneous treatment effects. As we explain in the conclusion, ensemble methods are best conceived of as a companion to new constituent methods: better individual methods for estimating heterogeneous effects will lead to better ensemble estimates and the ensembles provide a new method for evaluating individual methods for estimating heterogeneous effects.

2 Experiments and Conditional Average Treatment Effects

We follow a large prior literature and formalize the estimation of heterogeneous effects using potential outcome notation (Holland 1986; Green and Kern 2012; Imai and Ratkovic 2013) and also borrow some notation from recent work on the analysis of conjoint experiments (Hainmueller, Hopkins, and Yamamoto 2014). Suppose that we have a sample of N , ($i = 1, \dots, N$) individuals from a population \mathcal{P} . We suppose that there are C ($c = 1, 2, \dots, C$) factors in the experiment and in each of the factors participants are randomly assigned to one of K_c conditions and that there is also a control condition. Participant i 's condition will be given by $\mathbf{T}_i = (T_{i,1}, T_{i,2}, \dots, T_{i,C})$ if assigned to a treatment condition and otherwise $T_i = 0$. We will denote respondent i 's response to condition \mathbf{T}_i with the potential outcome $Y_i(\mathbf{T}_i)$.¹ We will analyze dichotomous dependent variables, though the ensemble methods generalize easily to continuous or other dependent variables.

To measure the effect of an intervention for the entire population of interest, scholars commonly report an Average Treatment Effect (ATE) across two conditions. For simplicity, we will compare the effect of some treatment condition \mathbf{T}_i to the control condition $T_i = 0$, though comparisons of any two treatments are possible. We will write the ATE as,

$$\phi(\mathbf{T}) = E[Y(\mathbf{T}) - Y(0)]. \quad (1)$$

Randomly assigning participants to arms of the treatment ensures that the $\phi(\mathbf{T})$ is identified, which is commonly estimated with a difference in means across conditions.

The ATE measures the effect of the intervention over the entire population, but to measure how treatment effects vary across respondent characteristics we will estimate a *conditional* average treatment effect (CATE) (Imai and Strauss 2011; Green and Kern 2012; Imai and Ratkovic 2013). The CATE, measures the ATE for respondents who share a set of characteristics. To formalize this definition, suppose that for each respondent i we collect J covariates ($j = 1, 2, \dots, J$), $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,J})$, with values of the covariates collected in the set \mathcal{X} . We can then define the CATE for covariate profile $\mathbf{x} \in \mathcal{X}$, and treatment arm \mathbf{T} as $\phi(\mathbf{T}, \mathbf{x})$,

$$\phi(\mathbf{T}, \mathbf{x}) = E[Y(\mathbf{T}) - Y(0) | \mathbf{X} = \mathbf{x}]. \quad (2)$$

A treatment effect is heterogeneous if the value of Equation 2 varies as we consider different strata of participants. As before, random assignment to treatment conditions is sufficient to identify the CATE.

The CATE measures the effect for respondents who share identical values of all J covariates. We may be interested, however, in how responses to the treatment effect vary across a subset of covariates, a single covariate, or a single factor. Suppose that we are interested in estimating the marginal effect of a subset of covariates $\mathbf{X}_S = (X_{s,1}, X_{s,2}, \dots, X_{s,S})$. Define the *marginal* conditional average treatment effect (MCATE) for covariates \mathbf{X}_S and treatment conditions \mathbf{T} , $\phi(\mathbf{T}, \mathbf{x}_S)$,

$$\begin{aligned} \phi(\mathbf{T}, \mathbf{x}_S) &= \int \phi(\mathbf{T}, (X_1, X_2, \dots, \mathbf{X}_S = \mathbf{x}_S, \dots, X_J)) dF_{\mathbf{X}_{-S} | \mathbf{X}_S = \mathbf{x}_S} \\ &= \int E[Y(\mathbf{T}) - Y(0) | (X_1, X_2, \dots, \mathbf{X}_S = \mathbf{x}_S, \dots, X_J)] dF_{\mathbf{X}_{-S} | \mathbf{X}_S = \mathbf{x}_S}. \end{aligned} \quad (3)$$

And finally we define the marginal average treatment effect (MATE) for factor c as $\phi(T_c = k)$ as

$$\begin{aligned} \phi(T_c = k) &= \int \phi(\mathbf{T}) dF_{\mathbf{T}_{-c} | T_c = k} \\ &= \int E[Y(\mathbf{T}) - Y(0)] dF_{\mathbf{T}_{-c} | T_c = k} \end{aligned} \quad (4)$$

¹ We make the usual SUTVA assumptions, which are particularly likely to hold in our survey experiments.

where $F_{\mathbf{X}_{-S}|\mathbf{X}_S=\mathbf{x}_S}$ is the joint distribution of the covariates \mathbf{X}_{-S} , given that $\mathbf{X}_S = \mathbf{x}_S$ and $dF_{T_c|T_c=k}$. In words, Equation 3 shows that MCATEs are averages of CATEs, where the value of the covariate of interest is fixed and $F_{\mathbf{X}_{-S}|\mathbf{X}_S=\mathbf{x}_S}$ is used to weight other covariate profiles and Equation 4 shows that MATEs are averages of ATEs and $F_{T_c|T_c=k}$ is used to weight the treatment profiles.

This notation corresponds to well-known quantities of interest scholars commonly compute. For example, when applying a parametric bootstrap, such as Clarify, it is common to set variables other than the covariate of interest to the sample means to estimate marginal effects (King, Tomz and Wittenberg 2000). This is equivalent to $F_{\mathbf{X}_{-S}|\mathbf{X}_S=\mathbf{x}_S}$ to place all probability on the sample means and zero elsewhere. It may be possible to estimate the joint distribution, facilitating extrapolation from the sample population to some other sample or population (Hartman *et al.* 2012). And finally, we may average over all other possible covariate values—for discrete random variables—or a grid of values for continuous random variables. This is equivalent to setting $F_{\mathbf{X}_{-S}|\mathbf{X}_S=\mathbf{x}_S}$ to a uniform distribution.

3 Estimation of Heterogeneous Treatment Effects with a Weighted Ensemble of Methods

When there are a large number of observations in each condition and participants who share the same set of covariates, reliable estimation of ATEs, CATEs, MATEs, and MCATEs is straightforward. The random assignment of participants to treatments ensures that a difference in means across treatment arms will reliably estimate the ATE and a difference in means across arms among respondents with the same set of covariates provides an accurate estimate of CATEs and MCATEs. With a large number of participants, the differences computed with naïve differences in means will tend to reflect systematic differences (Gelman, Hill, and Yajima 2012).

But for more heterogeneous treatments with a large number of conditions, or covariates that have few observations who share the exact same covariates, a simple difference in means will be a less reliable estimate of the effect of treatments. When the sample size is relatively small, naïve differences will be likely to reflect random variation in the sample, rather than systematic differences in the underlying methods because there will be few observations who share the exact same characteristics. This renders ineffective the usual method for estimating heterogeneous treatment effects: computing a difference in means for observations with the same covariate value. It also makes simple comparisons of different levels of high-dimensional treatments highly problematic.

The goal in estimating heterogeneous effects is to separate the systematic responses from differences solely due to chance of the random assignment. Several new methods provide novel ways to identify the systematic effects. Each method m estimates the response surface for any treatment k and covariates \mathbf{x} ,

$$g_m(\mathbf{T}, \mathbf{x}) = E[Y|\mathbf{T}, \mathbf{x}] \quad (5)$$

and quantities of interest are computed by taking differences across the response surfaces.

To estimate Equation 5 each of the methods vary necessary and consequential assumptions about how treatment assignment and covariates alter the response surface. For example, one approach to estimating heterogeneous treatment effects is to use regression trees (Imai and Strauss 2011) and Bayesian Additive Regression Trees (BART) to estimate CATEs and MCATEs (Chipman, George, and McCulloch 2010; Green and Kern 2012). The trees subdivide the data repeatedly, developing decision rules to split the data to make more accurate predictions. An ensemble of the trees is then used to model the response surface and estimate the heterogeneous effects. Other methods start from a more familiar regression framework and then use data and assumptions to identify systematic differences. For example, LASSO methods use a penalty to

shrink some coefficients to zero (Hastie, Tibshirani and Friedman 2001; Athey and Imbens 2015). Imai and Ratkovic (2013) extend and generalize LASSO, introducing a model that has two different penalties—one for covariates and another for variation in treatment effects. A different, though related, approach is to impose a model that shrinks the coefficients to a common mean, allowing only the strongest coefficients to take on distinct values (Gelman *et al.* 2008). Hainmueller and Hazlett (2013) extend this idea further and include a much more flexible modeling approach with *Kernel Regularized Least Squares* (KRLS) and prove useful statistical properties of the algorithm.² And still other methods balance between the two types of smoothing. Elastic Net is a method that includes penalties that both shrink to zero and shrink to a common mean, with the weight attached to each penalty determined by a parameter α , $0 < \alpha < 1$ (Hastie, Tibshirani and Friedman 2001).

Each of the methods have been shown to perform well on important political science examples. Yet, knowing the ideal method to apply in any one experiment requires knowledge about the data generating process that assumes the heterogeneous effect sizes are known—exactly what we set out to estimate. For example, Imai and Ratkovic (2013) impose a sparseness assumption, using a method that identifies a set of treatments and covariates that have no effect on the response surface. In contrast, Hainmueller and Hazlett (2013) assume the estimates are more *dense*, smoothing many of the coefficient estimates to approximately the same value (Hastie, Tibshirani and Friedman 2001).

The appropriateness of those modeling assumptions will vary across substantive problems. New and diverse methods, then, are essential for estimating accurate heterogeneity in treatment effects. But relying on a single method will result in suboptimal performance across diverse problems. When the assumptions fit the data generation process, the model will perform well, but when the assumptions are a poor fit the method will perform poorly.

In place of using a single method to estimate heterogeneous effects, we argue for a weighted ensemble of estimators. As we show below, we use a weighted ensemble because it tends to attach the greatest weight to the methods that perform best at the task at hand (van der Laan, Polley, and Hubbard 2007). Asymptotically ensemble methods will select the best performing methods for a particular problem from the collection of methods (van der Laan, Polley, and Hubbard 2007). We show in simulations that in a finite sample ensembles lead to better estimates, as measured by root mean square error. Ensembles are also useful because they can estimate more complex functional forms than the underlying methods. And ensembles make estimates more robust—limiting the possibility that a coding error in any one method could lead to invalid conclusions (Dietterich 2000).

3.1 Constructing the ensemble via super learning

The ensemble estimator we utilize is a weighted average of heterogeneous treatment effect estimators, where the estimators out-of-sample performance will determine the weights. To construct this weighted average, we use the cross validation based methodology *super learning* introduced in van der Laan, Polley, and Hubbard (2007), a method that we show in the online Appendix is closely related to other ensemble methods (Raftery *et al.* 2005; Montgomery, Hollenbach, and Ward 2012). Ensembles of methods are increasingly used for diverse problems including supervised text classification (Hillard, Purpura and Wilkerson 2008) and prediction (Raftery *et al.* 2005; Montgomery, Hollenbach, and Ward 2012; van der Laan, Polley, and Hubbard 2007). Both classification and prediction tasks are closely related to the estimation of heterogeneous treatment effects. In classification and prediction, the goal is to estimate a function such as $g(\mathbf{T}, \mathbf{x})$, in order to make an out-of-sample estimate about a document or

² Both KRLS, Find It, and other methods discussed here have applications that extend well beyond measuring heterogeneous treatment effects.

future observation. Heterogeneous treatment effects share a similar goal, but take the difference between response surfaces to estimate the heterogeneous effects. As in classification and prediction, identifying features—covariates and treatment assignments—that systematically affect the response surface will improve our estimates of the quantities of interest.

To construct the ensemble, assume we include M models ($m = 1, 2, \dots, M$) for estimating heterogeneous effects. For each method m we will define its estimate of $E[Y(\mathbf{T})|\mathbf{T}] = g_m(\mathbf{T})$ and $E[Y(\mathbf{T})|\mathbf{T}, \mathbf{x}] = g_m(\mathbf{T}, \mathbf{x})$. Along with the M models, we will suppose that we have a set of weights attached to each of the models $\mathbf{w} = (w_1, w_2, \dots, w_M)$. We will assume that all weights are greater than or equal to zero ($w_m \geq 0$ for all m) and the weights sum to 1 ($\sum_{m=1}^M w_m = 1$).

With the weights and models, we define our ensemble estimate of the ATE for condition \mathbf{T} as $\hat{\phi}(\mathbf{T})$,

$$\hat{\phi}(\mathbf{T}) = \sum_{m=1}^M w_m g_m(\mathbf{T}) - \sum_{m=1}^M w_m g_m(0). \quad (6)$$

Analogously, define the ensemble estimate for the CATE for condition \mathbf{T} and covariates \mathbf{x} as $\hat{\phi}(\mathbf{T}, \mathbf{x})$,

$$\hat{\phi}(\mathbf{T}, \mathbf{x}) = \sum_{m=1}^M w_m g_m(\mathbf{T}, \mathbf{x}) - \sum_{m=1}^M w_m g_m(0, \mathbf{x}). \quad (7)$$

In words, Equations 6 and 7 show that the ensemble creates a final estimate of a heterogeneous treatment effect by weighting the estimates of the heterogeneous effect from the corresponding models. To estimate MCATEs we will take the appropriate averages over CATEs, using a joint distribution to weight the averages and similarly we will estimate MATEs by taking appropriate averages over the ATEs.

The ensembles that we use will weight the predictions from the component models. Following van der Laan, Polley, and Hubbard (2007) we determine the weight to attach to each method using the component methods' predictive performance as assessed using cross validation. Predictive performance (or classification) is used because methods that perform well at individual classification are also likely to perform well at estimating heterogeneous treatment effects. Intuitively, this occurs because a method that predicts individual responses well in a cross validation will also tend to identify systematic responses to treatments and systematic heterogeneity in response to treatments—the characteristics that lead to accurate estimation of heterogeneous effects. The result is that methods that separate systematic features that assist in prediction also identify systematic differences that represent heterogeneity.

Given the connection between performance in cross validation and accurate estimation of heterogeneous treatment effects, we create our ensemble in three broad steps (van der Laan, Polley, and Hubbard 2007). In order to weight methods by their cross validation performance, we first generate out-of-sample predictions for each observation using each of the component methods. To do this, we proceed as if we are performing D -fold cross validation: we randomly divide our data into D subsets ($d = 1, \dots, D$).³ For each subset d , we train all M component methods using the participants in the other $D - 1$ subsets, and often using cross validation within each fold to select tuning parameters. In each step, we set tuning parameters using whatever procedure we will use in the full data set. This might involve additional cross validation steps within each fold. Then, we generate predictions using the trained models and the participants' in subset d 's covariates and treatment assignment. The result of this procedure is an $N \times M$ matrix $\hat{\mathbf{Y}}$ where entry \hat{Y}_{im} contains the out-of-sample prediction for participant i from method m .

3 We use 10-fold cross validation below. If there are concerns about sample size, the number of folds can be increased.

Table 1. Super learnings's three steps to generating ensemble estimates for heterogeneous treatment effects.

- (1) Generate out-of-sample predictions for all N observations from all M component methods using D -fold cross validation.
- (2) Estimate weights for each method based on its out-of-sample performance using a constrained regression (or other procedure, see online appendix).
- (3) Fit each component model to entire sample and then weight estimates from model using weights estimated in step 2. To construct the final estimates we weight each stratum according to a set of weights from a target population.

Second, we estimate the weights using the predictions from the folds of the cross validation. For each participant i we regress the true response, Y_i , on the out-of-sample predictions, $(\hat{Y}_{i1}, \hat{Y}_{i2}, \dots, \hat{Y}_{im})$. Specifically, we fit the model,

$$Y_i = \sum_{m=1}^M w_m \hat{Y}_{im} + \epsilon_i \quad (8)$$

where ϵ_i is an error term. To ensure that each w_m are weights, we impose two constraints on \mathbf{w}_m : that the weights sum to 1 ($\sum_{m=1}^M w_m = 1$) and that the weights are greater than zero ($w_m \geq 0$). Fitting this regression is a straightforward quadratic programming problem, whose solution provides a set of weight estimates $\hat{\mathbf{w}}$ that we will use to produce our final ensemble. A method's weight will depend upon its accuracy in the cross validation folds and the distinctiveness of the predictions (van der Laan, Polley, and Hubbard 2007).

In the third and final step we use the weights and the component methods to generate an ensemble. We fit each of the component models to the entire sample, creating a prediction function for each method m , $\hat{g}_m(\cdot)$. To create final estimates of interest, we generate synthetic observations (Green and Kern 2012), with covariates \mathbf{x} and treatments \mathbf{T} . For discrete covariates we generate all unique covariate and treatment combinations. For continuous covariates we vary over the range of the covariate. We then use the component methods to generate estimates of the heterogeneous effects for each of the component methods. For each synthetic observation i we obtain $\hat{g}_m(\mathbf{T}_i, \mathbf{x}_i)$. We then use the estimated weights to create a weighted average for each observation, $\hat{\phi}(\mathbf{T}_i, \mathbf{x}_i) = \sum_{m=1}^M \hat{w}_m \hat{g}_m(\mathbf{T}_i, \mathbf{x}_i)$. Using the per-synthetic unit effects, we can then summarize the effect further, taking averages according to appropriate weights. We summarize the steps of the process in Table 1.

The weighting procedure ensures that we can include many different kinds of methods, even if those methods have closely related assumptions (van der Laan, Polley, and Hubbard 2007). The general advice is to include more methods in the ensemble. Additional, correlated predictions will be downweighted in the weighting step. Distinct methods will ensure that different data generating processes are included in the ensemble. In general, adding more models will improve the performance of the ensemble. Of course, we should not take the position too far. Additional methods may add little new information if correlated with already included methods and might slow down the estimation procedure.⁴ If scholars are in a position where they need to limit the number of included methods, we recommend including methods that make different fundamental assumptions about the underlying data generating process, rather than methods that have the same assumed underlying process but different tuning parameters.

⁴ Obviously the super learner function will depend on the number of included methods, but we can perform 10-fold cross validation on a data set of 5000 synthetic observations using an ensemble of LASSO, two elastic nets, Bayesian GLM, and an SVM in 266.4 seconds on a Mac Pro with a 3.5 GHz 6-Core Intel Xeon E5 processor.

3.2 Inference for heterogeneous treatment effect estimates

An important goal when making an inference about heterogeneity in treatment effect estimates or the effects of heterogeneous treatments is to have a measure of our uncertainty about our estimate. In this section we describe three different approaches to making inference and to convey to the reader the information underlying a particular inference. Before presenting these methods, we address a major concern with applying heterogeneous treatment effect methods: that they will facilitate fishing. Fishing occurs when researchers use a data set to find “significant” findings and then report only those findings and neglect to report the search used to obtain those findings (Humphreys, Sanchez de la Sierra, and van der Windt 2013). To avoid fishing, we suggest a sequential approach to using heterogeneous treatment effect methods. Our preferred approach is to use the heterogeneous treatment effect methods for exploration on an initial experiment and then conduct a follow up experiment explicitly targeting the heterogeneity in treatment effects of interest or assessing the effect in groups that are deemed to be particularly interesting.

This sequential approach, however, is not helpful if experiments are costly or researchers would like to report uncertainty estimates before designing the next round of the experiment. A similar sequential approach can be adopted in any one experiment by borrowing the notion of a training/test set split from machine learning and applying it to experiments (for similar applications of the training/test split applied to experiments, see Fong and Grimmer (2016), Wager and Athey (2017)).⁵ To create a training set, researchers can randomly select a subset of their data and use the methods described here to fit a model to the data and to explore the results, finding interesting heterogeneity. Using the remaining data as the test set, researchers would focus on the specific quantities of interest identified in the test set. If researchers know the particular heterogeneity in the data they are interested in before hand, there is no need to make the training/test set split.

Suppose that researchers use an ensemble of heterogeneous treatment effect methods in the training set and have identified either interesting CATEs or MCATEs. When turning to the test set to make final inferences, there are asymptotic and bootstrap approaches to uncertainty estimation. Once a set of strata are identified as the particular effects of interest, inference procedures from targeted maximum likelihood estimation (TMLE) can be used, where we “target” the CATEs and MCATEs. Specifically, an additional adjustment of the estimation of the effect is made using an additional equation that measures the propensity to be assigned to particular strata. Using an asymptotic argument Van der Laan and Rose (2011) provide a closed form formula for estimating the standard error of the target parameter and show that it is normally distributed, making it straightforward to create confidence intervals.

The closed form for estimating the standard error from Van der Laan and Rose (2011) is useful and provides a natural way to report standard errors, but there are reasons to be concerned about the use of the TMLE asymptotic argument when estimating standard errors for our estimators of MCATEs. The usual application of TMLE is to estimate the main effect of an intervention in a large experiment. This large sample makes the asymptotic argument more likely to hold. When estimating MCATEs, however, there are likely to be few observations that will contribute directly to the estimate of the quantity of interest. The result is that the estimate of the standard error may poorly approximate the true standard error.

A second approach to inference is to use a bootstrap to estimate uncertainty in the MCATEs (Efron and Tibshirani 1994). The benefit of using a bootstrap to calculate uncertainty is that

5 Of course, splitting the training and test in this way will undermine the power in both subsamples. When deciding about how to assign units to the training and test set, researchers will have to decide how to prioritize the discovery of heterogeneity in the training set relative to the precision of inferences in the test set. If they are primarily interested in discovering heterogeneity at the expense of more precise inferences, they should assign more units to the training set. If they are primarily interested in making precise inferences at the expense of discovering heterogeneity in the training set, they should assign more units to the test set.

it does not require asymptotic arguments in order to justify the uncertainty measures. There are, however, several concerns with applying bootstrapping to make inferences about MCATEs. First, the methods that comprise the ensemble make bias-variance tradeoffs. The result is the methods intentionally include bias, so that bootstrapped confidence intervals may not have the reported coverage rates. Second, several features of the ensemble method we presented here need to be modified to ensure that regularity conditions of the bootstrap are met. There are two separate concerns: if weights are estimated using a constrained regression—as we have presented above—then that estimation process violates the bootstrap regularity conditions. And several of the methods that comprise the ensemble, such as the LASSO, violate the same regularity conditions. These problems are not, however, without solution. We can use a Bayesian procedure to estimate the weights attached to the individual ensembles, such as the method presented in Montgomery, Hollenbach, and Ward (2012). We can also apply several of the recent advances in applying the bootstrap to machine learning methods to obtain valid confidence intervals (Chatterjee and Lahiri 2011; Wager and Athey 2017).

Given the limitations of both closed form and bootstrap methods, we report the information that comprises our MCATEs using a different method. When we visualize the effects, the size of each point will be proportional to the number of observations that are contributing to the inference in the sense that they are proportional to the number of observations in that stratum. We view this as a middle step between providing direct estimates of uncertainty—that are likely to be problematic—and failing to provide any guidance on the information used to make an inference. Readers are then aware what effects are based on substantial data and which effects depend primarily on model extrapolation.

4 Monte Carlo Simulations of Ensemble Based Methods

We use a series of Monte Carlo simulations to show the ensemble accurately estimates heterogeneous effects.⁶ The strong performance of the ensemble is not surprising, particularly given ensemble methods strong performance at classification and prediction tasks (van der Laan, Polley, and Hubbard 2007; Hillard, Purpura and Wilkerson 2008; Raftery *et al.* 2005; Montgomery, Hollenbach, and Ward 2012). Our simulations show that by evaluating methods' predictive performance, ensembles attach greater weight to methods that provide better estimates of the heterogeneous effects.⁷

We assess the performance of the ensemble across four distinct data generating processes, averaging the performance of the ensemble over 5 instances of the data generating process. The simulated data generating processes vary in the number of treatments that have systematic effects, the number of treatments that have heterogeneous effects with the included covariates, and the type of covariates that are included in the simulation. In two simulations, *Monte Carlo 1* and *Monte Carlo 2* are sparse data generating processes—with many of the simulated treatments specified to have an undetectable systematic effect and only a few specified to have heterogeneous effects. *Monte Carlo 3* and *Monte Carlo 4*, specify data generating processes that are more dense—with more treatments having systematic (and large) effects and heterogeneity across covariates. We provide specific details of each data generation process in the online appendix.

We apply our complete ensemble method to each of the simulated data set. In the Monte Carlo simulations—and in our applications, below—we form an ensemble with nine methods: (1) LASSO (Hastie, Tibshirani and Friedman 2001); (2) Elastic Net, with the mixing parameter set to 0.5 (Hastie, Tibshirani and Friedman 2001); (3) Elastic Net, with the mixing parameter set to 0.25 (Hastie,

⁶ Replication of our Monte Carlo simulations and all other analyses are found in our replication code Grimmer, Messing, and Westwood (2017).

⁷ This contributes to other simulation based evidence on the performance of ensembles on similar tasks (see, for example, van der Laan, Polley, and Hubbard (2007)).

Table 2. Across diverse problems, the ensemble estimator performs best (root mean square error of each method relative to root mean square error of the ensemble, reported).

Method	MC 1	MC 2	MC 3	MC 4	Average
LASSO	0.99	1.08	1.34	1.25	1.18
Elastic Net ($\alpha = 0.5$)	1.03	1.08	1.22	1.14	1.12
Elastic Net ($\alpha = 0.25$)	1.58	1.26	0.95	1.26	1.23
Find It	1.09	17.74	1.75	25.47	14.66
Bayesian GLM	1.87	1.02	1.1	0.99	1.12
BART	1.08	3.16	1.36	3.65	2.68
Random Forest	4.99	2.05	2.96	2.25	2.64
KRLS	1.53	2.76	1.11	3.31	2.42
SVM-SMO	2.61	2.1	1.77	2.22	2.12
Naive Average	1.38	1.86	1.06	2.52	1.83

Note: This table shows that the ensemble method outperforms other methods across diverse problems in recovering the heterogeneous treatment effects. This table presents the mean square error for each method averaged over 5 simulations of each data generating process, divided by the mean square error of the ensemble method average over 5 simulations of each data generating process. In individual simulations the ensemble method regularly outperforms the other methods. The right most column shows that, on average, the ensemble estimator has the lowest mean square error across different problems, reflecting its utility as a workhorse tool for estimating MCATEs.

Tibshirani and Friedman 2001); (4) Find It (Imai and Ratkovic 2013); (5) Bayesian GLM (Gelman *et al.* 2008); (6) Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2010; Green and Kern 2012); (7) Random Forest (Breiman 2001; Wager and Athey 2017); (8) KRLS (Hainmueller and Hazlett 2013); and a (9) Support Vector Machine (SVM) (Platt 1998; Keerthi *et al.* 2001). In the online appendix we provide details for each method’s estimation. After performing 10-fold cross validation to generate predictions, we use Equation 8 to determine each method’s weights. After estimating the weight we then apply the models to the entire sample and create our ensemble estimate of the treatment effects implied by the data generating process as a weighted average of the estimates from each of the component methods.

In addition to comparing the performance of our ensemble estimator to the component methods, we will compare its performance to a *naïve* ensemble—where all methods are assumed to contribute equally to the average. We measure the performance of the methods using the root mean square error of the estimated heterogeneous treatment effects (MCATEs), with a smaller root mean square error implying more accurate estimates of the heterogeneous effects.

The ensemble method outperforms the other methods across diverse data generating processes in our Monte Carlo simulations. For ease of interpretation, Table 2 presents the performance of each method in terms of the ratio of its root mean squared error to the root mean square error of the ensemble estimate (van der Laan, Polley, and Hubbard 2007). If this is greater than 1, then the ensemble has a smaller root mean square error, or performs better in estimating the heterogeneous effects. To calculate the entries in Table 2 we first averaged each method’s performance in five different iterations of the data generating processes and averaged the weighted ensemble’s performance. We then took the ratio of the average mean square errors. In the online appendix we provide the MSE for each iteration of the Monte Carlo simulations.

Consider the first two columns in Table 2, showing the results for the sparse data generation processes. Here, we see that methods that assume sparsity perform better—methods such as LASSO. The ensemble outperforms these component methods, however, and is able to more accurately estimate the heterogeneous effects. The third and fourth columns show that methods that assume a dense set of effects and interactions perform better—such as Elastic Net with

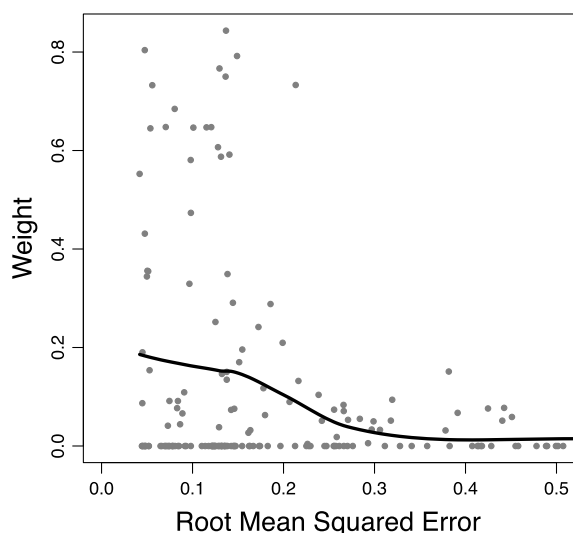


Figure 1. The ensemble tends to place greater weight on methods that more accurately measure heterogeneous treatment effects. (Note: This figure shows that the ensemble method places more weight on methods that more accurately measure the heterogeneous treatment effect. This occurs even though the method is weighting methods that are performing better at out-of-sample prediction. This occurs because of the close connection between estimating heterogeneous treatment effects and predicting out of sample.)

$\alpha = 0.25$, KRLS, and Bayesian GLM. Column 5 shows that, as a result, the ensemble estimate has the best average performance across data generating processes and iterations of the simulation. This exemplifies why ensembles are useful: because we never actually know the data generating process, we do not know how well a particular method's assumptions fit the underlying causal effects (Hastie, Tibshirani and Friedman 2001). Using ensembles ensures that we use methods that reliably capture the heterogeneous effects for a particular problem.

Figure 1 shows why the ensemble is able to outperform the constituent methods: the methods with a smaller RMSE in estimating MCATEs tend to receive more weight in the ensembles (van der Laan, Polley, and Hubbard 2007). Figure 1 presents the weight attached to the constituent methods (vertical axis) against the method's RMSE. This demonstrates a simple relationship: methods with a smaller RMSE in estimating the MCATEs receive greater weight in the final ensemble. Aggregated together, the methods with the best performance receive the greatest weight.⁸

With this strong performance of the methods in mind, we turn now to our applications. We use our methods to reveal how constituents reward legislators for securing money in the district and how constituents punish legislators for budget deficits.

5 Experiment 1: Rewarded For Type of Expenditure, Not Money

Our first experiment reexamines an experiment with heterogeneous treatments and how those treatments vary with respondent characteristics from Grimmer, Westwood, and Messing (2014). Grimmer, Westwood, and Messing (2014) argue that legislators' credit claiming statements—a message that legislators use to create the impression they are responsible for some government action—to receive credit for spending in the district and to cultivate a personal vote (Mayhew 1974; Grimmer, Westwood, and Messing 2014). The experiment from Grimmer, Westwood, and Messing (2014) analyzes how the content of a credit claiming message affects constituent credit allocation. In order to vary many of the features of the experiment, Grimmer, Westwood, and Messing (2014)

⁸ The method is also able to perform well at identifying particularly responsive subsets of observations. Following a simulation from Imai and Ratkovic (2013), we examined the method's ability to correctly discover the sign of the most responsive observations. Across different number of observations, the weighted ensemble method was able to correctly discover the sign of the most responsive and three most responsive subsets.

Table 3. Examining the effects of credit claiming statements on constituent credit allocation.

Advertising Condition
Headline: Representative (redacted) announces annual Congressional district art competition winner
Body: Representative (redacted) announced that 17-year-old Sara Fischer won 1st place in the annual Congressional district art competition. Sara’s winning art, “Medals” was created using colored pencils. Rep. (redacted) said Sara’s artwork will be displayed in the U.S. Capitol with other winning entries from districts nationwide.
Credit Claiming Condition
Headline: Representative (redacted) stageTitle moneyTitle typeTitle
Body: Representative (redacted), partyMain, alongMain stageMain moneyMain typeMain. Rep. (redacted) said “This money stageQuote typeQuote”
 stageTitle: [will request/requested/secured]
 moneyTitle: [\$50 thousand/\$20 million]
 typeTitle: [to purchase safety equipment for local firefighters/to purchase safety equipment for local police/to repave local roads, to beautify local parks/for medical equipment at the local planned parenthood/to help build a state of the art gun range]
 partyMain: [Democrat/Republican]
 alongMain: [(No text)/and Senator (redacted), a Democrat/ and Senator (redacted), a Republican]
 stageMain: [will request/requested/secured]
 moneyMain: [\$50 thousand/ \$20 million]
 typeMain: [to purchase safety equipment for local firefighters/to purchase safety equipment for local police/to repave local roads, to beautify local parks/for medical equipment at the local planned parenthood/to help build a state of the art gun range]
 stageQuote: [would help/would help/will help]
 typeQuote: [our brave firefighters stay safe as they protect our businesses and homes/our brave police officers stay safe as they protect our property from criminals/keep our roads in safe and working condition, ensuring that our local economy will continue to grow/create parks that add value to the community and provide our children a safe place to play/provide state of the art care for women in our community”/“provide local residents and local, state, and national law enforcement officials a place to sharpen their skills”]
Summary of Conditions
Funding Type: Planned Parenthood, Parks, Gun Range, Fire Department, Police, Roads
Money: \$50 thousand, \$20 million
Stage: Will Requested, Requested, Secured
Who: Alone, a Senate Democrat, a Senate Republican
Party: Democrat, Republican

created a template that allows them to vary several features of the message, while maintaining a coherent and realistic message from a legislator. The original authors report marginal effects from the experiment. Our goal in using an ensemble to reanalyze the experiment is to estimate the effects of heterogeneous treatments for exploratory purposes—to identify instances of large scale variation that future experiments could confirm.

Grimmer, Westwood, and Messing (2014) assign participants to a control condition (with a 10% chance) or the credit claiming condition (with a 90% chance). Participants in the control condition read a press release from a fictitious representative who “announced that 17-year old Sara Fisher won 1st place in the annual Congressional art competition.” This press release is an example of a common *advertising* press release—a message devoid of policy content intended to increase the legislators’ prominence (Mayhew 1974; Grimmer 2013). The full text of the condition is in Table 3.

Participants assigned to the credit claiming condition read a message about an expenditure in the district and the fictitious legislator’s role in securing that legislation. To assess how different

facets of the credit claiming process affect credit allocation, Grimmer, Westwood, and Messing (2014) vary five different components of the message: (1) type of expenditure, (2) amount of money, (3) stage in appropriations process, (4) collaboration with other legislators, and (5) representative's political party. Varying the features of message simultaneously allow us to identify the kind of information constituents use when evaluating credit claiming messages, in a way analogous to recent conjoint experiments (Hainmueller and Hopkins 2015; Hainmueller, Hopkins, and Yamamoto 2014). Table 3 summarizes the conditions and how the information was provided to participants in the study. In addition to the heterogeneous treatments, we also examine how the effects vary across two relevant respondent characteristics: respondent's *partisan identification*—respondents classify themselves as a *Democrat*, *Republican*, or *Independent/Other*—and *ideological orientation*—respondents classify themselves as *Conservative*, *Liberal*, or *Moderate*.

We examine the effect of legislators' credit claiming efforts on constituents' propensity to *Approve* of the representative's performance in office. Specifically, Grimmer, Westwood, and Messing (2014) asked participants if they "approve or disapprove" of the way the fictitious representative "is performing (his/her) job in Congress." Grimmer, Westwood, and Messing (2014) recruited 1,074 participants on Amazon.com's Mechanical Turk service, selecting only workers from the United States.⁹ After respondents were assigned to a treatment, they completed a brief postsurvey that included questions about the legislator and the respondent's own personal political preferences.

Figure 2 shows the (MATEs), averaging over the accompanying conditions and using the control condition for comparison. The vertical axis shows the information varied in the experiments—including the different types of expenditures, amount of spending, stage in the appropriations process, who announced the expenditure, and the fictitious legislator's party. The point is the MATE of that condition and the lines are 95 percent confidence intervals.

Figure 2 suggests that participants are able to make use of easily available information—such as the type of expenditure—but struggle to include other types of information in their evaluation of credit allocation. The type of expenditure matters considerably—gun ranges decrease approval for representatives (13.1 percentage point decrease, 95 percent confidence interval $[-0.25, -0.01]$), but other types of projects legislator increase approval over the control condition (37 percentage point increase, 95 percent confidence interval $[0.29, 0.46]$). Other information, however, has a smaller effect on constituent credit allocation. There is essentially no difference in the credit awarded legislators if they claim credit for \$50 thousand or \$20 million, who legislators announce with, or whether the legislator is a Democrat or Republican. Securing money does cause an increase in support, relative to stating that the legislator will request or requested the money, with securing causing an 8.1 percentage point great increase in support than requesting or stating an intent to request (95 percent confidence interval $[0.02, 0.15]$).

The effects in Figure 2 suggest that participants are evaluating legislators' credit claiming efforts by evaluating the type of expenditure, while struggling to use other information. If true, then how participants evaluate the type of expenditure will depend on their partisanship and ideology. This is clearest for two of the more polarizing types of expenditures Grimmer, Westwood, and Messing (2014) included: a gun range and funding for planned parenthood. Liberal elites and Democrats tend to vigorously defend planned parenthood, providing cues to like minded citizens that the organization provides valuable services. In contrast, conservatives and Republicans oppose planned parenthood, often working to strip the organization of money (for example, (Kasperowicz 2013)). Very different cues are available about gun ranges. Many Democrats—particularly liberal-urban Democrats—have argued for increased gun regulation. Republicans and conservatives have argued vigorously for constitutional protection of guns.

9 Grimmer, Westwood, and Messing (2014) included attention checks at the start and end of the survey to ensure that workers were not satisficing (Berinsky, Huber, and Lenz 2012).

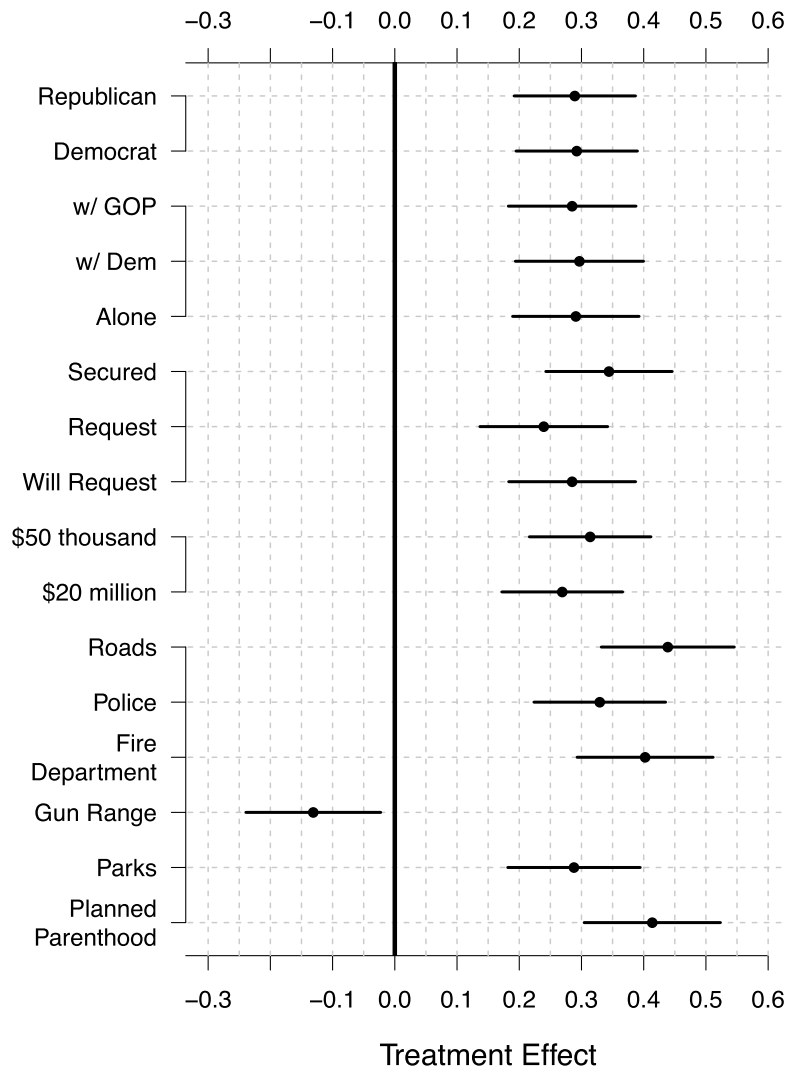


Figure 2. The marginal effects of the credit claiming experiment. (Note: This figure shows the marginal effects from the credit claiming experiment. Respondents appear to be evaluating the credit claiming statement based on the type of message, but struggle to include other types of information.)

To examine how the message content affects credit allocation across participants with different partisan identifications and ideological orientation we use our ensemble method.¹⁰ We apply the ensemble method using 10-fold cross validation, then use Equation 8 to estimate the weights attached to each method. The first column of Table 4 shows the weights attached to each method for the ensemble used to generate the effects for this experiment. Three methods receive nonzero weight: LASSO (0.62), KRLS (0.24), and Find It (0.14). In generating the final effects, we compare all treatment effects to the control advertising condition.

Figure 3 shows how the effect of the type of project claimed depends on the participant's partisan and ideological identification. On the right-hand side vertical axis we vary the type of expenditure announced and within each type of expenditure we vary participants ideological orientation and partisan identification. To ease interpretation and facilitate exploration, we draw lines to connect the heterogeneous responses to the same type of expenditure.

¹⁰ Together our experiment has $6 \times 2 \times 3 \times 2 \times 3 + 1$ conditions—a very heterogeneous treatment, along with the 9 unique partisan and ideological characteristics. Given our sample size limitations, we examine only pairwise interactions between our treatments in our method—reducing the number of potential conditions from 217 to 98 total conditions. This is a decision that we made balancing the desire to discover heterogeneity with the limits of our sample size.

Table 4. The weight attached to methods varies across experiments.

Method	Credit Allocation	Deficit Punishment
LASSO	0.62	0.00
Elastic Net ($\alpha = 0.5$)	0.00	0.00
Elastic Net ($\alpha = 0.25$)	0.00	0.00
Find It	0.14	0.10
Bayesian GLM	0.00	0.00
KRLS	0.24	0.81
SVM-SMO	0.00	0.09

Figure 3 reveals substantial heterogeneity in the effect of the *type* of project on constituent credit allocation, consistent with constituents evaluating the type of expenditure when evaluating legislators' credit claiming statements. Consider the response to money directed to planned parenthood. Liberal respondents, regardless of partisan affiliation, boost support for the representative who claims credit for securing money for planned parenthood. The fictitious legislator claiming credit for funds for planned parenthood caused a 30 percentage point increase in approval rating for liberals. Conservatives, however, have a more muted—and even negative—response to legislators who claim credit for planned parenthood spending. Indeed,

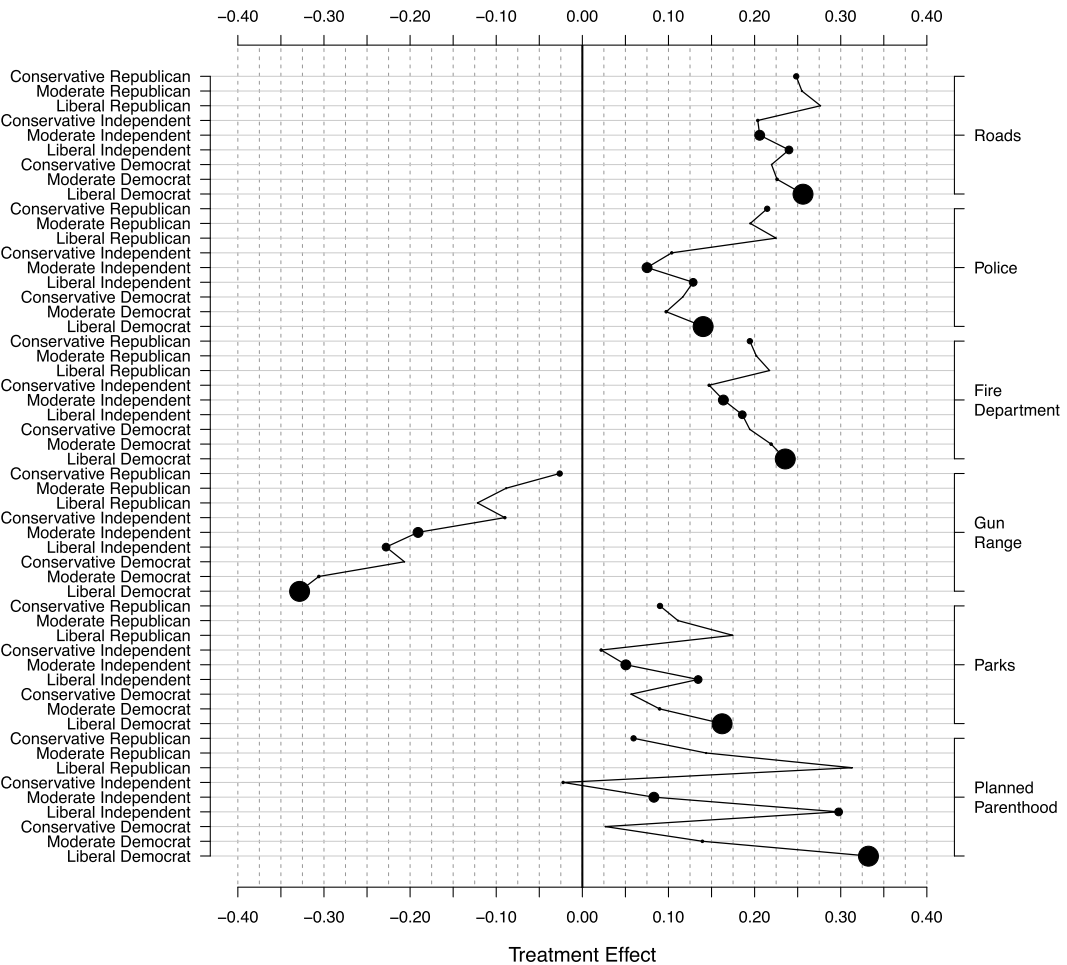


Figure 3. Constituent partisanship and ideology predicts differences in the effectiveness of credit allocation. (Note: This figure shows substantial heterogeneity in how participants respond to the different types of money legislators claim credit for securing.)

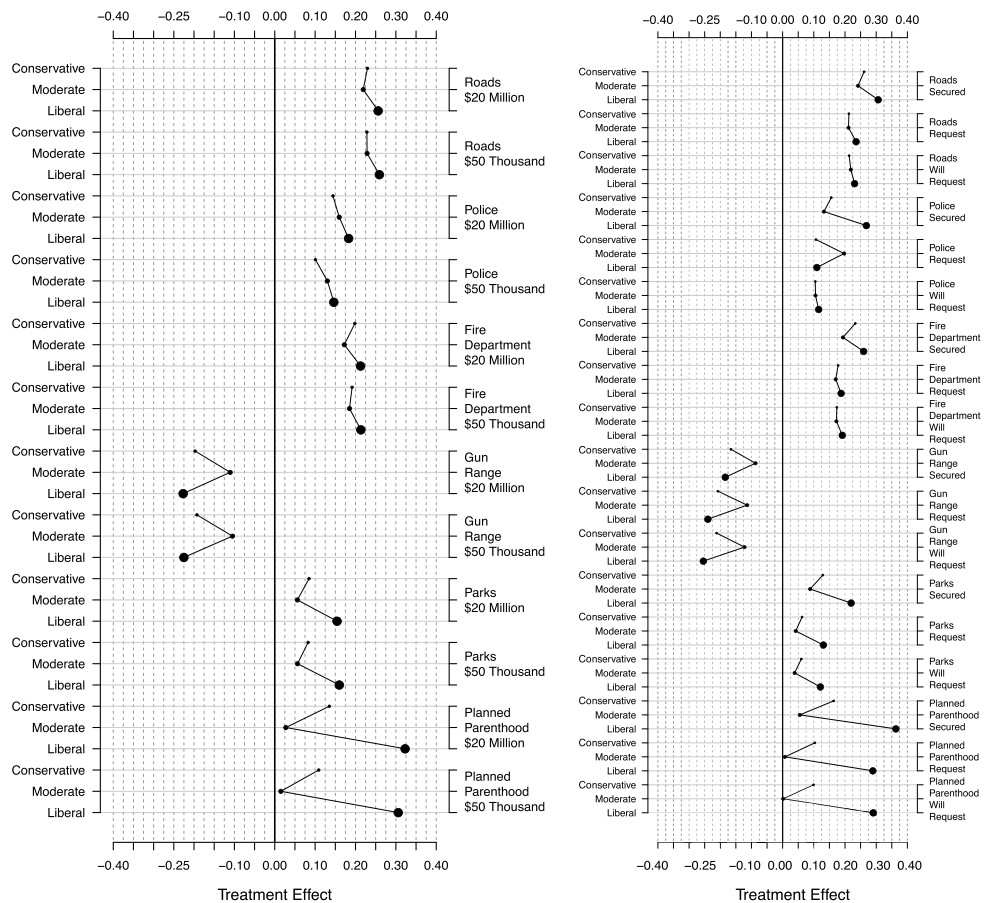


Figure 4. The money secured or the stage in the process appears to be less important. (Note: This figure shows that constituents struggle to include information about the amount of money legislators claim credit for securing (left-hand plot) or the stage in the process that the money is in (right-hand plot).)

claiming credit for money delivered to planned parenthood causes a decrease in approval ratings among independent conservatives and only a small increase among other conservatives. Of course, our visualization shows that there are few conservatives in our MTurk sample, leading us to have some caution in interpreting the findings for conservatives.

The effect of claiming credit for gun ranges also varies across participants. Constituents who are likely to have a negative attitude about guns have a strong negative reaction to legislators claiming credit for a new gun range. Claiming credit for a gun range causes an over 30 percentage point decline in a legislators' approval rating among liberal Democrats. But constituents who are likely to have a less negative view of guns have a more muted response to the gun range: claiming credit for a gun range causes only a 5 percentage point decline among conservative Republicans. But again, note the small sample of conservative Republicans in our MTurk sample. Conservatives and Republicans do not, however, simply have a smaller magnitude to all credit claiming efforts. Claiming credit for delivering money to police causes a 21 percentage point increase in approval among Republicans, while only an 11 percentage point increase among Democrats and independents.

The heterogeneity in response to the type of credit claiming message is consistent with constituents evaluating the type of expenditure when allocating credit for spending in the district. Figure 4 shows that constituents tend to be less responsive to other pieces of information in the credit claiming statements. The left-hand plot in Figure 4 shows how the effect of the type of project and the amount of money (right-hand vertical axis) for the project varies across

constituent ideology. This figure shows that constituents—across types of expenditure and ideological orientation—are largely unaffected by the amount claimed. Indeed the parallel lines are indicative of constituents who are largely unable to incorporate information about the size of the expenditure into their overall evaluations. Consistent with evidence from Grimmer, Messing, and Westwood (2012), the size of the expenditure appears to matter little when constituents are allocating credit to legislators. The right-hand plot in Figure 4 shows how the effect of credit claiming messages varies across type of money secured and stage in the appropriations process (right-hand vertical axis) and constituents' ideological orientation (left-hand vertical axis). While there is evidence that constituents do reward the fictitious legislator slightly more for *securing* rather than requesting money, the increase from having secured an expenditure appears to be much smaller than the heterogeneity across the type of expenditure.

This section uses the ensemble method for estimating heterogeneous treatment effects to explore the types of information constituents use when responding to legislators' credit claiming messages. This shows that constituents tend to evaluate easily available information—such as the type of money secured—but struggle to include information about the amount that legislators secured. Together, this shows how the ensemble of heterogeneous treatment effect methods can be used as an exploratory tool—to identify how the components of the messages vary across respondents. To validate this exploration, additional experiments could be run that explicitly test the heterogeneity discovered here along with a pre-analysis plan, ensuring that our exploration is not biasing the inferences we make (Humphreys, Sanchez de la Sierra, and van der Windt 2013).

5.1 Experiment 2: Punishment for budget deficits

Several studies argue that legislators claim credit for spending to cultivate a personal vote with constituents. Yet, in recent years a growing movement of conservatives, the *Tea Party*, has criticized expenditures as wasteful (Skocpol and Williamson 2011). Grimmer, Westwood, and Messing (2014) design an experiment to assess how the criticism of government spending affects how constituents allocate credit. Grimmer, Westwood, and Messing (2014) show that labeling an expenditure as contributing to the budget deficit undermines the credit legislators' receive. We apply the ensemble method to assess how the effect of the criticism varies across constituent characteristics—revealing that the effect of budget criticism varies substantially, depending on constituents' ideological orientation. While the previous example shows how the method can be used to estimate the effect of heterogeneous treatment, this example focuses on how the method can be used to estimate the heterogeneous effect of a relatively simple intervention.

Grimmer, Westwood, and Messing's (2014) experiment couples a legislator's claiming credit for an expenditure with criticism about the budget implications of the spending. For realism about the magnitude of the effects, this experiment utilizes participants' actual representatives, rather than fictitious representatives used in the first study. Table 5 contains the content of the experiment's three conditions. In the *credit claiming* condition participants are presented with their house member claiming credit for an \$84 million highway expenditure in their Congressional district. To customize the paragraph about each participant's legislator, Grimmer, Westwood, and Messing (2014) insert the representative's name at `|representativeName` and the participant's state at `|state` in the text.

Two *budget criticism* conditions vary the source of the information about how the spending will affect the federal deficit. The *CBO Budget Information* condition includes this same credit claiming about a highway expenditure, but pairs it with information about the budget consequences of the expenditure from the nonpartisan Congressional Budget Office (CBO). This condition has an official statement from the CBO, which includes the overall cost of the program and that expenditure would be deficit spending. In the *Partisan Information* condition, participants receive budget information from a political figure likely to criticize the participant's member of Congress:

Table 5. Content across conditions, experiment 2.

Headline: Representative representativeLastName Announces \$84 Million for Local Road Projects
Body: Representative representativeName (party - state) announced that the Department of Transportation Federal Highway Administration has released \$84 million for local road and highway projects. Representative representativeName said “I am pleased to announce that we will receive \$84 Million from the Federal Highway Administration. It is critical that we support our infrastructure to ensure that our roads are safe for travelers and the efficient flow of commerce.” This funding will add lanes to state highways.
CBO Budget Information: The nonpartisan Congressional Budget Office reported that the spending bill is wasteful and contributes to the growing federal deficit. “This bill contributes to federal spending without identifying a new source of revenue or off-setting budget cuts. Accounting for the total cost of this program across all Congressional districts, the bill costs taxpayers \$36.5 billion, all of which is added to the deficit and compounded with interest.”
Partisan Information: [Debbie Wasserman-Schultz, Chair of the Democratic National Committee/Reince Priebus, Chair of the Republican National Committee] said that the spending bill is wasteful and contributes to the growing federal deficit. “This bill contributes to federal spending without identifying a new source of revenue or off-setting budget cuts. Accounting for the total cost of this program across all Congressional districts, the bill costs taxpayers \$36.5 billion, all of which will be added to the deficit and compounded over time with interest.”
Key
representativeName: Representative’s name
party: Representative’s party
state: Representative’s state

the opposing party’s national chairperson. Participants with a Democratic representative see a statement from Reince Priebus—chair of the Republican national committee—and participants with a Republican member of Congress see a statement from Debbie Wasserman-Schultz—chair of the Democratic national committee.

Grimmer, Westwood, and Messing (2014) administered the experiment to 1,166 participants, using a census matched US sample from a Survey Sampling International (SSI) panel. Participants were assigned to conditions, the treatments were administered, and then a postsurvey was administered. Grimmer, Westwood, and Messing (2014) shows that the budget criticism have a strong and negative effect on legislators’ approval ratings. The budget information from the CBO causes an overall decrease of 8.2 percentage points (95 percent confidence interval [-0.16, -0.01]) and the partisan information causes a similar overall decrease in approval of 7.7 percentage points (95 percent confidence interval [-0.15, -0.00]).

To examine how the effects of the intervention vary across constituent and legislator characteristics, we apply the ensemble method to the experiment. Table 4 shows the diverse methods that receive weight for this experiment: including KRLS (0.81), Find It (0.10), and SVM (0.09). We then use the ensemble to compute MCATEs for combinations of respondent characteristics and particular treatments, with all effect sizes based on a comparison to the credit claiming message without criticism.

Following Imai and Strauss (2011) and Imai and Ratkovic (2013), we first use the ensemble method to assess the constituents who have the most negative and positive response to the criticism. Table 6 shows who has the most negative and positive response to the budget criticisms for Democratic and Republican representatives. The most negative responses for both Democratic and Republican representatives come from constituents who identify as conservative—strong conservatives for Republican representatives, conservatives for Democratic representatives. For Democratic legislators the most positive response comes from strong liberals who are also strong Democrats—the most positive response for Republican representatives is from moderate Republicans.

Table 6. Characteristics of constituents with the most negative response to budget criticism for democrat and republican representatives.

Democrat Representatives, Most Negative Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
−0.36	High School	\$50k–\$80k	50–63	Female	White	Cons.	Dem.
−0.36	High School	\$50k–\$80k	64+	Female	White	Cons.	Dem.
−0.36	Grad. Degree	\$50k–\$80k	50–63	Female	White	Cons.	Dem.
Democrat Representatives, Most Positive Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
0.28	College	<\$50k	36–49	Male	White	Strong Lib.	Strong Dem.
0.27	College	<\$50k	36–49	Female	White	Strong Lib.	Strong Dem.
0.27	College	<\$50k	<37	Male	White	Strong Lib.	Independent
Republican Representatives, Most Negative Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
−0.32	College	<\$50k	50–63	Male	Non-White	Strong Cons.	Strong Rep.
−0.32	College	\$50k–\$80k	50–63	Male	White	Strong Cons.	Strong Rep.
−0.31	College	<\$50k	50–63	Male	White	Strong Cons.	Strong Rep.
Republican Representatives, Most Positive Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
0.25	College	<\$50k	<37	Male	White	Moderate	Republican
0.24	College	<\$50k	<37	Male	White	Moderate	Republican
0.24	College	<\$50k	<37	Female	White	Moderate	Republican

Table 6 reveals who has the strongest response to the budget criticism treatment, suggesting that conservatives are particularly likely to punish legislators for deficit spending and that strong liberals are particularly likely to reject the criticism. To examine how the effect of the criticism varies across different types of respondents, Figure 5 shows how the effect of the budget criticism affects credit allocation for Democratic (left-hand plot) and Republican representatives (right-hand plot) and for constituents with varying ideological orientations (strong liberal, liberal, moderate, conservative, strong conservative) and partisan affiliations (strong Democrat, Democrat, Independent, Republican, strong Republican). This figure shows that strong liberals—particularly strong liberals with a Democratic representative—tend to have a *positive* response to the budget criticism. This aligns well with cues from political elites, who have suggested that deficit spending does less harm than members of the political right emphasize. Further, moderate Republicans appear unaffected by information about spending's budget implications. Conservatives, however, have a particularly negative response to learning about the budget implications of an expenditure. And as Table 6 illuminates, strong conservatives have a particularly negative response to the criticism.

The variation in response to the budget criticism aligns with intuition about how Tea Party rhetoric affects how constituents respond to budget criticism. For participants likely to be receptive to the rhetoric—conservatives—the budget criticism causes sharp punishment of the elected official. But for other constituents, the budget criticism is ineffective. Strong liberals—who have likely learned of the arguments for deficit spending—are unperturbed by the criticism and continue allocating credit to legislators for spending—in some cases greater expenditures. Together, the heterogeneity in response illuminates how critical rhetoric can make it costly for Republicans to claim credit for spending, because they depend on the support of strong conservatives in the primary. Democrats, however, do not have the same costs, because the

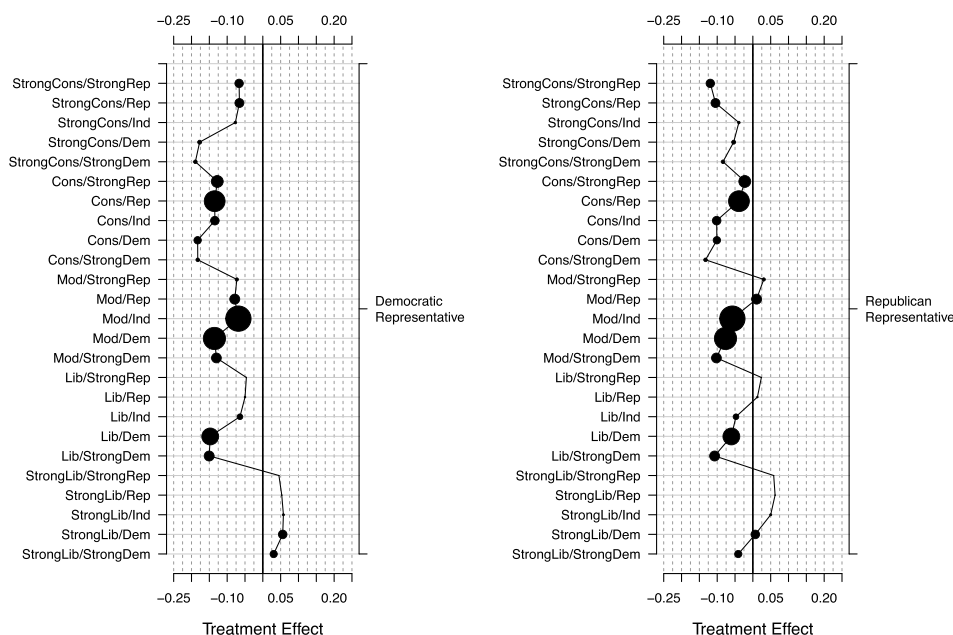


Figure 5. Strong liberals are unresponsive to budget criticisms. (Note: This figure shows how the response to budget criticism depends on constituents' characteristics (left-hand vertical axis). For both Democrats and Republicans (right-hand vertical axis), we see that Strong liberals are unresponsive to the budget criticism, or the criticism may cause an increase in legislators' approval ratings. Strong conservatives, however, a much more negative reaction to the credit claiming efforts. This is consistent with the rise of the Tea Party movement.)

most ideological members of their base—strong liberals, have a positive response to the budget criticism.

6 Conclusion

We have shown how weighted ensembles of methods provide reliable estimates of heterogeneous treatment effects. Across diverse problems the ensemble is able to provide accurate estimates, because it attaches more weight to methods that provide more accurate estimates for the particular task. Then applying the weighted ensemble to two experiments, we show how respondents evaluate easy to obtain information when accessing a legislators' credit claiming statement and the substantial ideological variation in how participants punish, or reward, legislators for deficit spending.

We have shown how ensembles provide accurate and reliable estimates of heterogeneous treatment effects across diverse problems. Ensembles do not, however, obviate the need for developing new methods for estimating heterogeneous effects. Quite to the contrary, the ensembles only work *because of the impressive innovations of the constituent methods*. New innovations in the estimation of heterogeneous effects, then, will improve the performance of the ensemble estimates.

Far from replacing individual methods, then, ensemble estimates of heterogeneous effects provide a way to make use of the impressive new innovations in the estimation of heterogeneous effects. By pooling together the methods, we make the most of new methods and the new experimental data.

Supplementary material

For supplementary material accompanying this paper, please visit <https://doi.org/10.1017/pan.2017.15>.

References

- Athey, Susan, and Guido Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. Preprint, [arXiv:1504.01132](https://arxiv.org/abs/1504.01132).
- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis* 20:351–368.
- Breiman, Leo. 2001. Random forests. *Journal of Machine Learning* 45(1):5–32.
- Chatterjee, Arindam, and Soumendra Nath Lahiri. 2011. Bootstrapping lasso estimators. *Journal of the American Statistical Association* 106(494):608–625.
- Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. BART: Bayesian additive regression trees. *Annals of Applied Statistics* 4(1):266–298.
- Dietterich, Thomas. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems. MCS 2000*. Lecture Notes in Computer Science, vol. 1857. Heidelberg: Springer-Verlag.
- Efron, Bradley, and Robert J. Tibshirani. 1994. *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.
- Fong, Christian, and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL 2016*, Berlin, Germany.
- Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4):1360–1383.
- Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons? *Journal of Research on Educational Effectiveness* 5(1):189–211.
- Gerber, Alan S., and Donald P. Green. 2012. *Field experiment: Design, analysis, and interpretation*. New York: W.W. Norton & Company.
- Green, Donald P., and Holger L. Kern. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* 76(3):491–511.
- Grimmer, Justin. 2013. *Representational style: What legislators say and why it matters*. Cambridge: Cambridge University Press.
- Grimmer, Justin, Sean J. Westwood, and Solomon Messing. 2014. *The impression of influence: Legislator communication, representation, and democratic accountability*. Princeton, NJ: Princeton University Press.
- Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2012. How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review* 106(4):703–719.
- Grimmer, Justin, Solomon Messing, and Sean Westwood. 2017. Replication data for estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. doi:[10.7910/DVN/BQMLQW](https://doi.org/10.7910/DVN/BQMLQW).
- Hainmueller, Jens, and Chad Hazlett. 2013. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis* 22(2):143–168.
- Hainmueller, Jens, Daniel Hopkins, and Teppei Yamamoto. 2014. Causal inference in conjoint analysis: Understanding multi-dimensional choices via stated preference experiments. *Political Analysis* 22(1):1–30.
- Hainmueller, Jens, and Daniel J. Hopkins. 2015. The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science* 59(3):529–548.
- Hartman, Erin, Richard Grieve, Roland Ramshai, and Jasjeet S. Sekhon. 2012. From SATE to PATT: Combining experimental with observational studies. University of California, Berkeley Mimeo.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning*. Springer.
- Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics* 4(4):31–46.
- Holland, Paul. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–960.
- Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* 21(1):1–20.
- Imai, Kosuke, and Aaron Strauss. 2011. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* 19(1):1–19.
- Imai, Kosuke, and Marc Ratkovic. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.
- Kasperowicz, Pete. 2013. GOP seeks planned parenthood study with hope to strip funding. [Politico.com](http://politico.com).
- Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation* 13(3):637–649.
- King, Gary, Michael Tomz, and Jason Wittenberg. 2000. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science* 44(2):347–361.
- Mayhew, David. 1974. *Congress: The electoral connection*. New Haven, CT: Yale University Press.
- Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2012. Improving predictions using ensemble Bayesian model averaging. *Political Analysis* 20(3):271–291.

- Platt, J. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, ed. B. Schoelkopf, C. Burges, and A. Smola. Cambridge, MA: MIT Press, <http://research.microsoft.com/jplatt/smo.html>.
- Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* 133:1155–1174.
- Ratkovic, Marc, and Dustin Tingley. 2017. Sparse estimation and uncertainty with application to subgroup analysis. *Political Analysis* 25(1):1–40.
- Samii, Cyrus, Laura Paler, and Sarah Daly. 2017. Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia. *Political Analysis* 24(4):434–456.
- Skocpol, Theda, and Vanessa Williamson. 2011. *The tea party and the remaking of republican conservatism*. Oxford: Oxford University Press.
- van der Laan, Mark, Eric Polley, and Alan Hubbard. 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology* 6(1):1–21.
- Van der Laan, Mark J, and Sherri Rose. 2011. *Targeted learning: Causal inference for observational and experimental data*. New York: Springer Science & Business Media.
- Wager, Stefan, and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, forthcoming.