Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 51–74

Political Campaigns and Big Data†

David W. Nickerson and Todd Rogers

The all-encompassing goal of political campaigns is to maximize the probability of victory. To that end, every facet of a campaign is evaluated by how many votes an activity will generate and at what cost. To perform this cost–benefit analysis, campaigns need accurate predictions about the preferences of voters, their expected behaviors, and their responses to campaign outreach. For instance, efforts to increase voter turnout are counterproductive if the campaign mobilizes people who support the opponent. Over the past six years, campaigns have become increasingly reliant on analyzing large and detailed datasets to create the necessary predictions. While the adoption of these new analytic methods has not radically transformed how campaigns operate, the improved efficiency gives data-savvy campaigns a competitive advantage. This has led the political parties to engage in an arms race to leverage ever-growing volumes of data to create votes. This paper describes the utility and evolution of data in political campaigns.

The techniques used as recently as a decade or two ago by political campaigns to predict the tendencies of citizens appear extremely rudimentary by current standards. At that time, citizens' likely support was gauged primarily by their party affiliations and the "performance" of the precincts in which they lived (that is, what

■ *David W. Nickerson is Associate Professor of Political Science, University of Notre Dame, South Bend, Indiana. He served as the "Director of Experiments" in the Analytics Department in the 2012 re-election campaign of President Barack Obama. Todd Rogers is Assistant Professor of Public Policy, Harvard Kennedy School of Government, Cambridge, Massachusetts. He co-founded the Analyst Institute, which uses field experiments and behavioral science insights to develop best practices in progressive political communications. Their email addresses are dnickers@nd.edu and Todd_Rogers@hks.harvard.edu.*

† To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.2.51 doi=10.1257/jep.28.2.51

percentage of the precinct had voted for a given party in the recent past). Whether a person was predicted to turn out and vote was often based on the past four general elections; for example, it was not uncommon to hear phrases like "2 of 4 voter" or "3 of 4 voter" used in campaign targeting plans. Past donors would be recontacted and asked for a flat amount of money (or perhaps asked for their highest previous contribution if that information was available) and prior volunteer captains would be recontacted, but intermittent volunteers were unlikely to appear on any lists. Back then, a "numbers-driven campaign" implied that candidates and their advisors paid close attention to poll numbers and adjusted policies in response to surveys. A memorable example of this dynamic is the story of President Clinton's advisor Dick Morris fielding a poll to choose Jackson Hole, Wyoming, as the vacation spot for the president (Kuhn 2004). Presidential campaigns targeted states based on historical notions of which states could see the vote swing either way, combined with the realities of the campaign budget.

In retrospect, the reliance of political campaigns on such rough—although often useful—heuristics is puzzling. Campaigns a decade ago already possessed considerable information on citizens' preferences based on what they had collected directly from volunteers, donors, and their own polling. Voter registration rolls were available at the state level from Secretaries of State. Detailed census information was available. Why did campaigns take so long to realize the value of information resources they already possessed?

Part of the answer is technological: adequate storage and computing power required large investments and were beyond the infrastructure of nearly all campaigns and state parties. Even if an entrepreneurial campaign made that investment, much of the available data would not have been as reliable as it is today. States were not required to keep electronic copies of which citizens voted in each past election until 2002 with the passage of the Help America Vote Act of 2002 (42 U.S.C. § 15483), so using the data on voting in federal elections would have been onerous in many regions.

But perhaps the biggest impediment to wider adoption of data-driven campaigning was simply that statistical thinking—and the human capital that produces it—had not yet taken root in the world of political consulting. Campaign consultants generate most of their business through social networks and are judged by win/loss records. Political candidates are typically trained in nonquantitative fields like law, education, and medicine and are more focused on fundraising and voter outreach than the nitty-gritty of managing a campaign. There were certainly consultants specializing in campaign data analytics, and the development of "predictive scores" for voters existed as a niche business, but most campaign decisions did not rely on these approaches. There were too few people with the skills required to make a noticeable impact on how campaigns operated and too few decisionmakers equipped to appreciate the effect that a fuller use of information could have. At that time, mail vendors were on the cutting edge of using consumer data for modeling purposes and at least a decade ahead of the political campaign learning curve (Malchow 2003).

These impediments to data-driven campaigning have changed in recent years. The costs of purchasing, storing, managing, and analyzing data have decreased exponentially. The supply of quantitatively oriented political operatives and campaign data analysts has increased as predictive analytics has gained footholds in other sectors of the economy like banking, consulting, marketing, and e-commerce. To reduce the need for individual campaigns to spend scarce funds purchasing citizen information from commercial vendors, the national parties have decided to construct, maintain, and regularly augment their own voter databases (McAuliffe with Ketten 2008, pp. 280–87).

These conditions have provided fertile ground for analytically minded consultants to apply statistical tools to campaign activities and campaign data. Contemporary political campaigns amass enormous databases on individual citizens and hire data analysts to create models predicting citizens' behaviors, dispositions, and responses to campaign contact. This data-driven campaigning gives candidates and their advisers powerful tools for plotting electoral strategy. A political campaign has limited financial resources. It can use this data-driven approach to shape decisions about who the campaign should target, with a sense of how much such contact will affect voter preferences, behaviors like fundraising, or turnout at the polls. This technology allows campaigns to target their outreach tactically at particular individuals and then also to aggregate these predictive estimates up to the jurisdiction level to inform large-scale strategic decisions.

Given that campaigns view their analytic techniques as secret weapons to be kept out of the hands of opponents, the public discourse on campaign data has been largely speculative and somewhat hypothetical, ranging from hyping the performance of the tools (Scherer 2012) to alarmist concerns about the personal privacy of voters (Duhigg 2012). This paper describes the state of contemporary campaign data analytics. We begin by explaining why campaigns need data and the "predictive scores" that they seek to calculate. We then describe where that data comes from and the techniques used to analyze political data. We conclude by noting several challenges facing campaigns as data analytics become more widely used and increasingly accurate. The analytics revolution has not radically transformed campaigns in the manner that television did in the 1960s, but in a close political contest, data-driven campaigning can have enough effect to make the difference between winning and losing.

Why Do Campaigns Need Data?

Contemporary campaigns use data in a number of creative ways, but the primary purpose of political data has been—and will be for the foreseeable future—providing a list of citizens to contact. Campaigns need accurate contact information on citizens, volunteers, and donors. Campaigns would like to record which citizens engage in specific campaign-supporting actions like donating money, volunteering, attending rallies, signing petitions, or expressing support for candidates or issues in tracking polls. Indeed, the Federal Election Commission requires campaigns and coordinated committees to disclose the identity of all individuals who contribute more than $200 during the calendar year. These disclosure requirements mean that campaigns have a legal requirement, as well as financial incentive, to maintain good lists of donors.

Campaigns also use data to construct predictive models to make targeting campaign communications more efficient and to support broader campaign strategies. These predictive models result in three categories of "predictive scores" for each citizen in the voter database: behavior scores, support scores, and responsiveness scores.

Behavior scores use past behavior and demographic information to calculate explicit probabilities that citizens will engage in particular forms of political activity. The primary outcomes campaigns are concerned with include voter turnout and donations, but other outcomes such as volunteering and rally attendance are also of interest.

Support scores predict the political preferences of citizens. In the ideal world of campaign advisers, campaigns would contact all citizens and ask them about their candidate and issue preferences. However, in the real world of budget constraints, campaigns contact a subset of citizens and use their responses as data to develop models that predict the preferences of the rest of the citizens who are registered to vote. These support scores typically range from 0 to 100 and generally are interpreted to mean "if you sample 100 citizens with a score of *X*, *X* percent would prefer the candidate/issue." A support score of "0" means that no one in a sample of 100 citizens would support the candidate/issue, "100" means that everyone in the sample would support the candidate/issue, and "50" means that half of the sample would support the candidate/issue. Support scores only predict the preferences at the aggregate level, not the individual level. That is, people assigned support scores of 50 are not necessarily undecided or ambivalent about the candidate/issue and, in fact, may have strong preferences. But when citizens have support scores of 50, it means that it is difficult to predict their political preferences.

Responsiveness scores predict how citizens will respond to campaign outreach. While there are theoretical rationales as to who might be most responsive to blandishments to vote (Arceneaux and Nickerson 2009) and attempts at persuasion (Hillygus and Shields 2008), in general predicting which types of individuals will be most and least responsive to particular direct communications in a given electoral context is difficult. Campaigns can use fully randomized field experiments to measure the average response to a campaign tactic (Gerber and Green 2000; Green and Gerber 2008; Nickerson and Rogers 2010; Arceneaux and Nickerson 2010; Nickerson 2005; Nickerson, Friedrichs, and King 2006; Bryan, Walton, Rogers, and Dweck 2011; Gerber and Rogers 2009; Bailey, Hopkins, and Rogers 2013; Rogers and Nickerson 2013). The results of these experiments can then be analyzed to detect and model heterogeneous treatment effects (in this case, predictive scores). The estimated model can then be used to predict treatment responsiveness for the entire target population and guide future targeting decisions (Issenberg 2012a, b, c).

Some of the results of these experiments can only be used to inform decisions in future elections: for example, the results of most voter turnout experiments necessarily come after Election Day. But other experiments can be conducted during the election cycle to improve efficiency in real time; for example, lessons from experiments evaluating the efficacy of treatments aimed at increasing observable behaviors like donations and volunteering can be put to immediate use. Similarly, the persuasiveness of campaign communications can be gauged through randomized experiments that measure voter preferences through post-treatment polling of the treatment and control groups. The types of citizens found to be especially responsive to the campaign treatment in these pilot experiments—as reflected in the responsiveness score—can be targeted during a larger rollout of the campaign treatment. Conversely, citizens who are unresponsive, or are predicted to respond negatively, can be avoided by the campaign.

Campaigns are primarily concerned with the practical question of how accurately predictive scores forecast the behaviors, preferences, and responses of individual citizens, not with testing an academic theory. As a result, the variables included in the construction of these scores often have thin theoretical justifications. That said, a variable in a dataset that is found to predict an outcome of interest but has no theoretical rationale for the relationship is more likely to prove to be spurious when validated against an "out-of-sample" dataset. For instance, the analyst may discover that people between the ages of 37 and 43 are more likely to support Republicans than older and younger age groups. However, there is no particular reason to suspect that this six-year cohort is especially conservative, suggesting that the finding could be a sample-specific fluke that would not generalize to the overall population. Successful predictive scores need not be based on theories or imply causal relationships, but campaign data analysts must still think critically and creatively about what variables sensibly relate to their outcomes of interest to generate predictive scores with the external validity required by campaigns.

Where Does Campaign Data Come From?

Procuring and maintaining large databases of citizens with up-to-date information from multiple sources may seem straightforward, but it is a nontrivial logistical hurdle and requires substantial financial commitment. After all, people frequently change residences and contact information (Nickerson 2006a). Campaigns also need to track their own behavior to limit awkward interactions with citizens who have been contacted multiple times previously.

In the recent past, campaigns struggled to manage and integrate the various sources of their data. The data collected by those working on digital communications rarely linked with the data collected by those working on field operations—meaning canvassing, phone calls, volunteer recruitment, and so on—or fundraising. One of the most heralded successes of the 2012 campaign to re-elect President Obama was the creation of *Narwhal*, a program that merged data collected from these digital, field, and financial sources into one database (Gallagher 2012; Madrigal 2012). As a result, the Obama re-election campaign began with a ten terabyte database (BigData-Startups 2013) that grew to be over 50 terabytes by the end of the election (Burt 2013).

The foundation of voter databases is the publicly available official voter files maintained by Secretaries of State, which ensure that only eligible citizens actually cast ballots and that no citizen votes more than once.[1] The official voter file contains a wide range of information. In addition to personal information such as date of birth and gender,[2] which are often valuable in developing predictive scores, voter files also contain contact information such as address and phone. More directly relevant to campaigns, certain details about past electoral participation are also recorded on official voter files. *Who* citizens vote for is secret, but *whether* citizens vote is reflected in official voter files—as is the method used to vote: for example, in person on Election Day or by use of absentee or another form of early voting. This information concerning past vote history unsurprisingly tends to be the most important data in the development of voter turnout behavior scores. The act of voting, of course, reveals higher propensity to vote.

The geographic location of citizens' residences can also provide valuable information, because campaigns can merge relevant Census and precinct data with the information on citizens in the voter database. Census data—such as average household income, average level of education, average number of children per household, and ethnic distribution—is useful for the development of a host of predictive scores. Campaign data analysts also append the aggregated vote totals cast for each office and issue in past elections in each citizen's precinct to individual voter records in the voter database. Even being mindful of ecological fallacy—that is, inferring someone's individual characteristics based on their membership in a larger group or cluster—this aggregate-level information in fact tends to increase predictive score accuracy.

Campaign data analysts also can append two types of data from consumer databases. First, and most essentially, they seek updated phone numbers. Phone calls are a critical feature of campaigns. While a volunteer knocking on doors will make successful contact with two to four people/hour, a volunteer making phone calls can reach 10–15 people/hour (Nickerson 2006b, 2007a). Using an automated dialer, the total can be even higher. While most official voter files contain phone numbers, they are often out of date and coverage is incomplete. Election officials only request a phone number from voters registering for the first time, and so if someone continues voting in the same jurisdiction over time, it's not uncommon to find phone numbers that are 20 years out of date. Because current phone numbers

[1] The exception to this rule is North Dakota, which does not have a voter registration system. Eligible voters simply show up and prove their eligibility by showing a valid ID, utility bill, or having a neighbor vouch for their residency.

[2] In states that were subject to the Voting Rights Act, the self-identified race of the registrants is included on official voter files, though this may change in light of the Supreme Court's June 25, 2013, ruling in *Shelby County v. Holder* 570 US ___ (2013).

are so important, campaigns find it worthwhile to purchase more accurate contact information available from consumer data firms.

Campaigns can also purchase a wide range of additional information from consumer data vendors relatively inexpensively, such as estimated years of education, home ownership status, and mortgage information. In contrast, information on magazine subscriptions, car purchases, and other consumer tastes are relatively expensive to purchase from vendors, and also tend to be available for very few individuals. Given this limited coverage, this data tends not to be useful in constructing predictive scores for the entire population—and so campaigns generally avoid or limit purchases of this kind of consumer data. The vast majority of these variables literally do nothing to increase the predictive power of models of mass behavior once prior behavior is accounted for (for example, any power of income or education measures to predict voter turnout are subsumed by controlling for prior voter turnout).

While campaigns do purchase some information, the vast majority of the useful information campaigns collect about individuals is provided by the individuals themselves. For example, those who have donated and volunteered in the past are high-value prospects for fundraising and volunteer-recruitment in the future. Moreover, the attributes of these individuals can be used to develop behavior scores to identify others who may be likely to donate or volunteer. Similarly, information about individuals who answered the phone or door in the past can be used to develop behavior scores for others who may be likely to be contactable moving forward. Data collected from online activities can be of particular value as well because such activities require a relatively low threshold for citizens to take action. For the small set of citizens who provide an email address to the campaign to receive campaign emails,[3] all of their activity concerning those emails—for example, sign up, opening emails, clicking links in emails, taking actions like signing petitions—can be tracked and used to predict levels of support for the candidate or focal issue, likelihood of taking action, and in many cases the policy areas of greatest interest (for example, imagine a voter who opens emails about taxes twice as often as any other topic). Thus, a state party or political organization can compile valuable information for developing predictive scores just by maintaining accurate records of its interactions with citizens over time.

In short, many of the claims about the information that campaigns purchase about individuals is overblown; little of the information that is most useful to campaigns is purchased. Official voter files are public records, census and precinct-level information are also freely available, and individual citizens themselves volunteer a wealth of data that can be used to develop scores that predict all citizens' behaviors and preferences. In fact, predictive scores can often allow campaigns to estimate some citizen preferences and behaviors more accurately

[3] In 2012, the Obama campaign had email addresses for 20 million supporters (Haberman 2013) compared with 13 million for the Obama campaign in 2008 and the three million addresses collected by the 2004 Kerry campaign (Vargas 2008).

than direct reports from citizens themselves (Rogers and Aida 2013; Ansolabehere and Hersh 2012). People may not be actively misrepresenting their intentions, but the desire to project a positive image of the self may lead voters to overestimate the degree to which they will participate in a given election. Again, the most important piece of information campaigns purchase tends to be phone numbers—and this is purchased with the intent of performing the old-fashioned task of calling citizens directly. Because the most useful information tends to be collected directly from citizens, one of the most valuable data acquisition activities in which campaigns engage is exchanging their information with that of other allied political organizations (when legal) to increase the breadth and scope of data that will be useful for the development of predictive scores.

An interesting result of the type of data that campaigns acquire directly from citizens is that campaigns are able to predict with greater accuracy which citizens will *support* their candidates and issues better than which citizens will *oppose* their candidates or issues. Information regarding citizens who donate, volunteer, and subscribe to email lists is available to campaigns and can be used to predict which other citizens will be similar. In contrast, citizens who do not perform such behaviors at all, or who perform similar behaviors for opposing campaigns, cannot be directly observed, so discriminating among the citizens who do not actively support a campaign is a much more challenging task. As a result the distribution of support scores typically have two to three times more voters with the highest scores (99 and 100) than the lowest (0 and 1). This imbalance does not imply that the opposition enjoys less passionate support or that the data analysts failed in their predictive task; it is a natural result of being able to observe the activity of only one campaign's supporters in an electoral competition. Similarly, because the foundations of voter databases are official voter files from states, campaigns tend to have much more information on citizens who have voted and are registered than citizens who have never voted and are not registered. Predictive models can still be constructed to predict fruitful geographies or people to target for registration drives, but the data available are much sparser and the models necessarily more coarse. This likely exacerbates the inequality in campaign communication and outreach between those who are already politically engaged and those who are not, and between voters and nonvoters (Rogers and Aida 2013).

How Do Campaigns Analyze Data to Develop Predictive Scores?

The predictive scores campaigns construct can be roughly divided into two types. The first predicts the behavior or attitudes of voters (that is, behavior scores or support scores). These models do not make any causal claim about why these individuals vote or donate or support the candidate; they merely predict the focal trait. As such, causation is not a major concern, and the goal of the analyst is primarily to avoid overfitting the data. The second type of score predicts how voters will respond to campaign outreach (that is, responsiveness scores). These responsiveness scores typically come from exploring heterogeneous reactions to campaign treatments in randomized field experiments. The causal effect of the campaign outreach is established by the experiment and these estimated effects are used as parameters for strategic decisionmaking. However, the moderators predicting strongly positive or weakly positive (or even negative) responsiveness to the treatment are not causal. In other words, the data may have been generated by an experiment, but the enterprise of modeling responsiveness to the treatment remains a matter of finding observed differences across types of subjects that predict large or small treatment effects. For instance, a campaign data analyst may discover that women are more responsive to a treatment than men, but since gender was not randomly manipulated by the campaign it is impossible to know that gender *caused* the differential response to treatment. The campaign data analyst only knows that gender is *correlated* with treatment responsiveness. Thus, even the search for moderators of the treatment effect in an experiment is essentially observational in nature.

Most of the analytic techniques employed by campaign data analysts are taught in standard undergraduate econometrics or statistics classes. Currently, the vast majority of the predictive scores used by campaigns are created by a campaign data analyst (or a team of them) using simple regression techniques: ordinary least squares for continuous outcomes; logistic regression for binary outcomes; and, rarely, tobit for truncated data like dollars donated or hours volunteered. The skills necessary for developing such models are widespread, and the models can easily be customized to specific political environments. For instance, party registration is not predictive of candidate preference for older citizens in many Southern states—because the South was historically solidly Democratic and remained so at the state level well after the civil rights movement transformed the national political environment—but campaign data analysts attuned to contextual facts like this can accommodate them in regression analyses.

There are two major downsides to using regression techniques for constructing campaign models. First, the utility of techniques that uncover correlations is highly dependent on the talent of the particular campaign data analyst employing them. A capable campaign data analyst who is familiar with the properties of the variables available in voter databases can generate highly accurate predictive scores for citizens. However, a slightly less-capable campaign data analyst might generate predictive scores that are only slightly better than the unsophisticated methods employed by earlier campaigns. As an example, consider the task of predicting a person's likelihood of voting in an election. Controlling for the whole set of turnout history available (often more than 50 elections) will typically predict around one-third more variance in individual turnout than the old "of 4" rule of thumb (that is, did the person vote in 0, 1, 2, 3, or 4 of the past elections). However, these variables all tap into a common latent propensity to vote and exhibit considerable collinearity. As a result, the coefficient for several of these variables will be negative and statistically significant. There is no theoretical rationale for why turnout in one election would decrease turnout in a future election, so observing negative coefficients would suggest that the analyst has overfitted the data and should pare back the number of variables used or model the propensity for turnout differently. Experienced analysts also construct relevant variables (for example, past turnout among people in the household) and insert theoretically informed interactions (for example, ethnicity of the voter by ethnicity of the candidate) to improve model fit. The marginal gains from these new variables are rarely as large as the initial gains from using a wide range of past turnout decisions, but that is to be expected—the gains from good predictive models are incremental. Since the people running campaigns rarely have experience or expertise in data analytics, the competence of the campaign data analysts they employ cannot be taken for granted.

The second drawback to using regression techniques in campaign models is that unique regression models typically need to be constructed for different regions, issues, and candidates, so the "modeling by hand" approach to analysis offers few economies of scale. While individual campaign data analysts likely become more efficient with each successive model they develop, constructing models for multiple races around the country requires either a small army of campaign data analysts, or else settling for very general national models that are not adapted for local contexts.

Thus, campaign data analysts have been seeking more systematic methods for selecting a preferred regression. The commercial marketing industry often uses a form of "machine learning" (for example, *k*-means clustering or *k*-nearest neighbor classifiers; see Gan, Ma, and Wu 2007) to divide consumers into categorical types like "blue collar, grilling, SUV owner." However, these statistical methods to group similar individuals or households are less useful for campaign data analysts because strategic cost–benefit decisions in campaign planning are based on individual-specific probabilities for particular outcomes, and knowing that a set of citizens are similar in many dimensions does not assist with targeting if those dimensions are not highly correlated with behaviors like voting, ideology, and propensity to donate. For this reason, *supervised learning* algorithms are typically more appropriate for the task of modeling political data.

Supervised machine learning includes methods such as classification and regression trees (Breiman, Friedman, Stone, and Olshen 1984). In a regression tree approach, the algorithm grows a "forest" by drawing a series of samples from existing data; it divides the sample based on where the parameters best discriminate on the outcome of interest; it then looks at how regressions based on those divisions would predict the rest of the sample and iterates to a preferred fit. The researcher chooses the number of "trees"—that is, how many times the data will be divided. In the particularly popular "random forests" algorithm for implementing a regression tree (Breiman 2001), the algorithm uses only a randomly drawn subset of variables in each tree to decide on the fit rather than the entire set of available variables. The payoff for this approach is that it generates estimates of what parameters are most important: that is, what parameters add the most predictive power when the group of other parameters is unchanged. Aside from its analytical advantages, "random trees" is a popular decision tree ensemble algorithm because it has very few tuning parameters and is available as an **R**-package, so that analysts with little formal education in statistics can develop the models. Bayesian Additive Regression

Trees have similar advantages (Chipman, George, and McCollough 2010; Green and Kern 2012).

Supervised machine learning presents three major advantages for campaign data analytics. First, these classes of estimators are typically nonlinear, so commonly known nonlinear relationships—such as the curvilinear relationship between age and turnout (older cohorts vote at higher rates than younger cohorts but this relationship peaks among group 60–70 years old and then reverses)—are easily accommodated by the algorithms. Second, the approach involves less discretion for the individual campaign data analyst, so the quality of the predictive scores generated is not as heavily dependent on the capabilities and integrity of analysts. People constructing the models still need to input the most diagnostic variables and set up rigorous out-of-sample tests to validate the models, but the algorithms are written in advance and run identically for every citizen in the voter database. Finally, these data-mining algorithms are relatively scalable. Some techniques may be computationally intensive and the variables included may need to be customized, but generally the marginal cost of constructing additional models is lower using these algorithms than having a campaign data analyst construct new models from similar databases by building a series of regressions from the ground up.

The major downside of these regression tree algorithms from the campaign's perspective is that their use is relatively new and not widespread, and it will take experience to see how to trim the regression trees and customize the tuning parameters in a way that satisfies political requirements. Campaign data analysts must also take great care to not overfit their models to their data (Dietterich 1995), in which case the results become less likely to apply outside the model. Typically, there will not be sufficient data from any single jurisdiction to create a unique model, so the data from several jurisdictions will need to be pooled to produce useful predictive scores. Most algorithms can be adapted to accommodate jurisdiction-specific political requirements, but only a small fraction of campaign data analysts today have the necessary skill set. In sum, as campaign data analytics becomes more common, sophisticated, and mature, it will likely move away from judgment-based regressions to regressions based on customized machine learning algorithms like regression trees.

How Are Predictive Scores Used?

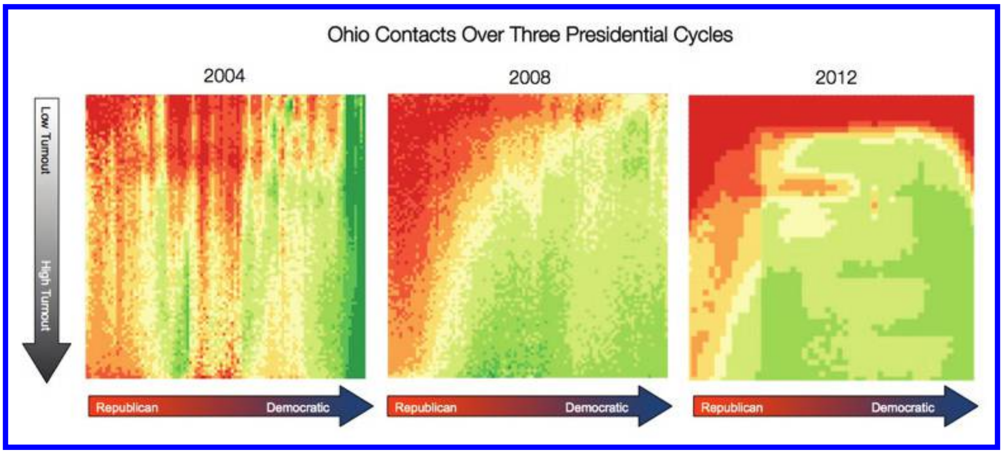
Campaigns use predictive scores to increase the efficiency of efforts to communicate with citizens. For example, professional fundraising phone banks typically charge $4 per completed call (often defined as reaching someone and getting through the entire script), regardless of how much is donated in the end. Suppose a campaign does not use predictive scores and finds that upon completion of the call 60 percent give nothing, 20 percent give $10, 10 percent give $20, and 10 percent give $60. This works out to an average of $10 per completed call. Now assume the campaign sampled a diverse pool of citizens for a wave of initial calls. It can then look through the voter database that includes all citizens it solicited for donations and all the donations it actually generated, along with other variables in the database such as past donation behavior, past volunteer activity, candidate support score, predicted household wealth, and Census-based neighborhood characteristics (Tam Cho and Gimpel 2007). It can then develop a fundraising behavior score that predicts the expected return for a call to a particular citizen. These scores are probabilistic, and of course it would be impossible to only call citizens who would donate $60, but large gains can quickly be realized. For instance, if a fundraising score eliminated half of the calls to citizens who would donate nothing, so that the resulting distribution would be 30 percent donate $0, 35 percent donate $10, 17.5 percent donate $20, and 17.5 percent donate $60, then the expected revenue from each call would increase from $10 to $17.50. Fundraising scores that increase the proportion of big donor prospects relative to small donor prospects would further improve on these efficiency gains.

The same logic can be applied to target expenditures for voter mobilization and persuasive communications. Targeting persuasive communications to citizens who are extremely unlikely to vote is inefficient. Even if the persuasive communication were effective at convincing these citizens to support the campaign's candidate or issue, the usual assumption among practitioners is that changing citizens' candidate or issue preferences does not meaningfully change their likelihood of voting. A similar logic could be applied to citizens who are already extremely likely to support a campaign's candidate or issue. If the support score predicts that a citizen is 98 percent likely to support a campaign's candidate or issue, and assuming the opposing campaign's activities will not meaningfully undermine this citizen's support likelihood, one might decide that persuasive communications would be better targeted to citizens who have a moderate or low likelihood of supporting the campaign's candidate or issue, along with a high likelihood of voting. Relying on turnout scores and support scores to target persuasion efforts in this manner represents an increase in efficiency, just as fundraising scores improve the cost effectiveness of fundraising calls.

The value of using predictive scores for targeting has become widely recognized by campaigns during the past five years. Sophisticated use of these predictive scores allows campaigns to simultaneously broaden the populations targeted while pruning away groups they believe will be cost ineffective.

Catalist, LLC, is a political data vendor that compiles and maintains nationwide registration, demographic, and other political data for progressive, civic, and nonprofit organizations such as labor unions, political candidates, and other advocacy groups. They build predictive scores using this data to help their clients analyze the electorate and target their activities more efficiently. The firm provided an aggregated data visualization for showing how its targeting of populations for its clients evolved over the last three presidential elections in Ohio (Ansolabehere and Hersh 2010). The discussion that follows references analyses of data aggregations that include the activities of independent groups as well as the activities of the Kerry campaign in 2004, the Obama campaign in 2008, and Ohio candidates in 2012 other than Obama. In each election, Catalist had several hundred clients across

Figure 1
Heatmap of Ohio Contacts over Three Presidential Cycles



Source: Derived from Catalist, LLC.
Notes: The x-axis is likelihood of supporting a Democratic candidate over a Republican candidate, ranging from 0 (left) to 100 (right). The y-axis is likelihood of voting ranging, from 100 (low) to 0 (high). Colors (or in grayscale, shade) of each cell indicate how many direct contacts were made by a particular campaign. In the grayscale version of the heatmap, darker means more contacts. In the color version, dark red represents the least contacts and dark green the most contacts. Readers can see the color heatmap in the online version of this paper.

the state of Ohio, for which data on contacts across all elections was aggregated. Catalist categorizes potential Ohio voters along two scales: whether or not they are likely to vote, and whether they are more likely to vote Democratic, Republican, or in-between. Divide each of these measures into a scale with 50 gradations, making a total of 2,500 different cells. You can then create a "heat map" of how often each one of those cells is contacted by campaigns allied with Catalist, including all modes of contact for all purposes across the election cycle, as in Figure 1. The heat maps used in the political campaigns are multicolored, but our print readers will see a grayscale version instead. Because of the centrality of Ohio in the past three presidential elections, the calculations represent tens of millions of voter contacts.

Although Catalist's client base differed across all three cycles, this graphical analysis of contacts for 2004, 2008, and 2012 show the increasing reliance on predictive scores for collective voter targeting efforts (see Figure 1). In 2004, when few clients relied on predictive scores for targeting, Catalist found that most contact was concentrated among people predicted to support Democratic candidates, regardless of their likelihoods of voting. This meant that campaign resources were probably inefficiently allocated, with a substantial share going to Democrats who were extremely unlikely to vote, or to Democrats who were extremely likely to vote and did not require either mobilization or persuasion. In 2008, Catalist clients appeared to have relied more on predictive scores for their targeting. The highest concentrations of direct contacts were observed among citizens who were predicted to support Democratic candidates but who had low likelihoods of voting—that is, those who might be reasonable targets for voter mobilization. They also targeted high-turnout citizens with middling partisanship scores, who might be reasonable targets for "persuasion." The reasonableness of targeting in these ways depends on the likelihood that voters can be moved to turn out, or be persuaded. As mentioned above, a current practice is to develop "responsiveness scores" based on pilot experiments to optimize targeting—particularly for persuasion outreach. As a result, the targeting in 2008 appears much closer to optimal than was observed in 2004. The heat map of contacts for 2012 looks much the same as that of 2008 except with smoother transitions and more consistency across the landscape, suggesting even wider adoption of predictive scores for targeting. One noticeable difference between the 2012 heat map and those of previous cycles is that Catalist clients appear to have avoided communicating with citizens with the lowest turnout probabilities. Catalist's clients may have chosen this strategy for a range of reasons, but regardless of their strategic reasons, apparently Catalist's Ohio clients in 2012 used predictive scores to manifest their strategic plans in ways that they had not in previous cycles.

What Are Predictive Scores Worth?

Campaign organizations have adopted predictive scores, which suggests that they are electorally useful. They use these scores to target nearly every aspect of campaign outreach: door-to-door canvassing; direct mail; phone calls; email; television ad placement; social media outreach (like Facebook and Twitter); and even web page display. Determining exactly how much using these scores affects electoral outcomes is difficult because the counterfactual is unclear. Is the appropriate comparison for assessing the value of campaign analytics to contrast the current uses of predictive scores for targeting with a complete absence of targeting? Or would it be to compare current uses to the basic heuristics that were used for targeting in the relatively recent past? Whatever the specific choice, it is possible to derive bounds as to how much campaign analytics could matter to campaigns.

Persuasive communications is a good place to begin because targeting is so diffuse. There are so many possible targets, including potentially all citizens, and so many strategies, from shoring up support to causing opposition supporters to defect. Thus, persuasive campaign outreach can be directed almost anywhere along the support score spectrum from hard-core supporters to hard-core opponents. Many campaigns use responsiveness scores as part of targeting their persuasive communications (Issenberg 2012a, b, c). Suppose a campaign's persuasive communications has an average treatment effect of 2 percentage points—a number on the high end of persuasion effects observed in high-expense campaigns: that is, if half of citizens who vote already planned to vote for the candidate, 52 percent would support the candidate after the persuasive communication. If a campaign indiscriminately attempted to persuade 8.5 million citizens—about the size of the Florida electorate—it would generate 170,000 votes under this scenario.

Table 1
Hypothetical Example of Persuasion Responsiveness Score's Value
(assuming average effect of campaign contact is 2 percentage points and electorate size is 8.5 million)

Quintile	*Effect multiplier*	*Votes created in quintile*	*Cumulative votes*	*Improvement over no targeting*
Top 20%	3	102,000	102,000	200%
60–80%	2	68,000	170,000	150%
Middle 20%	1	34,000	204,000	100%
20–40%	0	0	204,000	50%
Bottom 20%	−1	−34,000	170,000	20%

Notes: Imagine that a campaign has created a responsiveness score that predicts which citizens would be most responsive to its persuasive communications. Based on the responsiveness score, those in the top quintile are three times more responsive to the persuasive communications than the average citizen, the next quintile is twice as responsive, the middle quintile is no more responsive than average, the second quintile shows no average responsiveness to the persuasive communications, and the bottom quintile actually exhibited backlash to the persuasive communications equal to the overall average treatment effect.

Now imagine that the campaign has created a responsiveness score that predicts which citizens would be most responsive to its persuasive communications. Based on the responsiveness score, those in the top quintile are three times more responsive to the persuasive communications than the average citizen, the next quintile is twice as responsive, the middle quintile is no more responsive than average, the second quintile shows no average responsiveness to the persuasive communications, and the bottom quintile actually exhibited backlash to the persuasive communications equal to the overall average treatment effect. Table 1 illustrates these outcomes.[4] Actual campaign data analysts would construct a continuous responsiveness score, but this example involving quintiles suffices for illustration.

For campaigns with the resources to contact only 20 percent of the electorate, the responsiveness score allows them to create 102,000 votes ($1{,}700{,}000 \times 0.02 \times 3 = 102{,}000$). Without any form of targeting the campaign would generate only 34,000 votes ($1{,}700{,}000 \times 0.02 = 34{,}000$), so using predictive scores increases the number of votes by 200 percent (see Table 1, row 1). A better financed campaign that could contact 40 percent of the electorate and would target the two most promising quintiles of the population. This strategy would yield a total of 170,000 votes, which is a 150 percent increase over having no targeting ($3{,}400{,}000 \times 0.02 = 68{,}000$) (see Table 1, row 2). In this scenario, using predictive scores still improves the campaign's impact, but the gain is less than that of the more resource-constrained campaign. A campaign with the resources to push up against the zero bound where

[4] Backlash is not an uncommon observation among field experiments examining persuasive campaign effects (for example, Arceneaux and Kolodny 2009; Bailey, Hopkins, and Rogers 2013) and among other types of experiments (Nicholson 2012; Hersh and Shaffner 2013).

additional contacts begin to cost the campaign votes would see its efficiency improve by only 50 percent (see Table 1, row 4). This dynamic means that smaller campaigns will benefit most from targeting based on predictive scores, but they are also the campaigns that are least able to afford hiring campaign data analysts and voter databases. Well-financed campaigns benefit from targeting based on predictive scores, but yield smaller relative gains over not using predictive scores for targeting. In this sense, given that small campaigns tend to be less reliant on data analytics, it appears that smaller campaigns may be underinvesting in the development and use of predictive scores.

Again using a fairly generous multiplier regarding responsiveness scores and a baseline 2 percentage point average treatment effect, we can set an upper bound on how the use of such a score might affect campaign outcomes. If there are 8.5 million citizens who will vote in a state (roughly the number of votes cast in the 2012 presidential election in Florida), and a campaign can successfully administer the attempted direct persuasive communications to only half the targeted citizens because of inability to reach all citizens, then a campaign that does not use responsiveness scores would generate 85,000 votes while a campaign that uses responsiveness scores would generate 102,000 votes through direct persuasive communications. While the difference of 17,000 votes is notable, it constitutes only 0.2 percent of the overall vote in this jurisdiction. That said, it would have constituted 23 percent of the 74,309 vote margin of victory for the Obama campaign in Florida in 2012.

Campaigns do not want to mobilize citizens to vote who support their opponent, so one of the most important uses for support scores is to identify which citizens should be targeted during voter mobilization efforts. In an evenly divided electorate, indiscriminately mobilizing citizens would net zero votes—because as many opponents would be mobilized as supporters. In this setting, a naïve comparison of data-based campaigning to absolutely no targeting is not appropriate. Instead, consider a comparison with the following relatively basic targeting strategy that is still employed today in electoral settings that do not have access to predictive scores. Imagine that a campaign attempts to identify individual citizens who support their candidate or issue by directly contacting them in person or over the phone. Imagine that this campaign can successfully reach half of the population and accurately identify their candidate/issue preference. For the remaining half of the population for whom the campaign has not identified a preference, the campaign proceeds to sweep through neighborhoods where more than half of the population supports the campaign's candidate, on the assumption that this approach will lead to a net gain in votes. The only people not targeted in these sweeps are those individuals concretely identified as supporters of the opponent. We can therefore express the expected yield in votes from this targeting strategy as

$$0.5\beta N_j(\%Support_j) \text{ if } \%Support_j < 0.5$$

$$\beta N_j(\%Support) - 0.5\beta N_j(\%Oppose) \text{ if } \%Support_j > 0.5,$$

where β, is the mobilization effect from the campaign, $\%Support_j$ is the level of support for the candidate (a number between 0 and 1) in precinct j, and N_j is the number of registered voters in precinct j.

The first line points out that in precincts where support for the candidate is less than 50 percent, the only effect of this plan will be the direct contacts with supportive voters. However, by assumption the campaign only has the ability to identify half of these people. The second line points out that in areas where support for the candidate is more than 50 percent, the strategy will have two effects. The first is the benefit from mobilizing supporters in the precinct. Unfortunately, the sweep also mobilizes opponents in the proportion to which they are present ($\%Oppose$). However, the campaign managed to identify half of the people supporting the opposition and can choose to avoid these individuals, so the counterproductive mobilization can be cut in half.

We can now contrast this targeting strategy to an imagined predicted-support-score strategy. It would obviously be an unfair comparison to argue that the predicted-support-score strategy worked without error, so we assume that it includes both false positives (misidentifying opponents as supporters) and false negatives (misidentifying supporters as opponents). One can think of these errors as reflecting the political diversity of a given neighborhood. In precincts where the vote is split 50/50, the false positive and false negative error rates are both 15 percent, because these would be the precincts where it is most difficult to infer political beliefs. However, in this hypothetical example the error rate tapers linearly as the precinct becomes more informative of resident beliefs, so that if a precinct unanimously supports one candidate or another, the error rate would obviously be zero. The relationships below presents the formula used in this hypothetical model:

$$\beta N_j[\%Support_j(0.85) - \%Oppose_j(0.15)] \text{ if } \%Support_j = 0.50$$

$$\beta N_j\left[\%Support_j\left(1 - 0.15 \times \frac{\%Support_j}{0.5}\right) - \%Oppose_j \times 0.15\frac{\%Oppose_j}{0.5}\right]$$

$$\text{if } \%Support_j < 0.50$$

$$\beta N_j\left[\%Support_j\left(1 - 0.15 \times \left(1 - \frac{\%Support_j}{0.5}\right)\right) - \%Oppose_j \times 0.15 \times \left(\frac{1 - \%Oppose_j}{0.5}\right)\right]$$

$$\text{if } \%Support_j > 0.50.$$

The equations make clear one underappreciated aspect of predictive modeling; modeling can only increase the efficiency of mobilization efforts. If the outreach from the campaign is not effective (that is, $\beta = 0$), then no votes are generated. Big data analytics may receive media attention, but its effectiveness is entirely reliant on the strength of more traditional aspects of the campaign. If a campaign does not have effective outreach to voters, then predictive analytics cannot solve that problem.

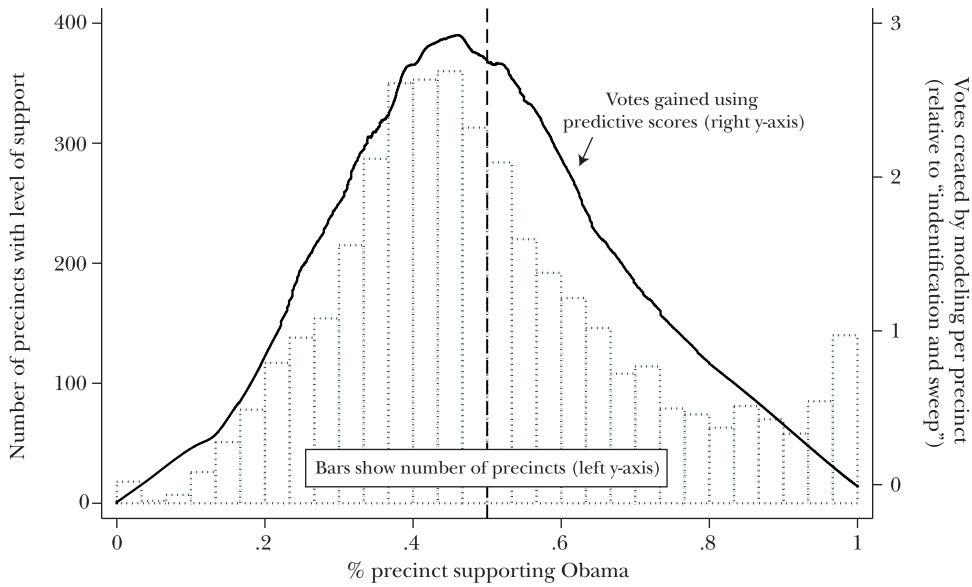
Comparing the traditional strategy of "identification and sweep" to the predictive model, two advantages of the predictive model become clear. First, predictive analytics allows the campaign to target likely supporters in otherwise unfriendly territory. Before accurate prediction was possible, campaigns would leave votes on the table by ignoring supporters living in opponent strongholds. Given the expense of actually identifying individual voter's preferences and the relatively low yield in terms of identifying supporters, avoiding these areas was not optimal tactically, but it was understandable. Second, precinct sweeps are inefficient because in evenly divided precincts many nonsupporters are also mobilized and thereby decrease the overall effectiveness of mobilization drives. Predictive scores (to the extent they are accurate) can reduce this inefficiency. As a result, conditional on precinct size, the biggest difference between the traditional "identification and sweep" tactic and modeled scores is found in the most evenly divided precincts.

Figure 2 shows the results of a thought experiment if these two tactics had been used in Florida across all 4,354 precincts during the 2012 election. The x-axis depicts the percent of votes cast in favor of President Obama in each precinct, and the left-hand y-axis shows in how many precincts President Obama received that share of the vote. Thus, President Obama received between 0 and 3 percent of the vote in about 20 precincts (the left-most bar) and received between 97 and 100 percent of the vote in 140 precincts (the right-most bar). Now imagine as a hypothetical example that the Obama campaign knows the distribution of its support across precincts before the election and is considering two possible strategies to increase its vote: the old-style "identification and sweep" combination of direct contact and precinct targeting, or the method using prediction scores. The solid line, measured on the right y-axis, shows the difference in the number of votes generated from these two approaches. The biggest difference between the two strategies takes place in the middle of the distribution where precincts are most evenly split.[5] The reason for this is clear when the tails are considered. In areas where support for Obama was low, there were not many Obama supporters to mobilize. In the areas where support for Obama was high, there were many supporters to mobilize, but both targeting strategies would target these citizens and neither would mistakenly mobilize those who support the opposing campaign's candidate. It is in areas where the precinct-level data is not predictive of which candidate the citizens support where predictive scores at the individual-level yield the greatest value—even given the inevitably higher number of false positives and false negatives in these precincts.

Using these assumptions, we can gain a rough sense of the impact of the Obama 2012 mobilization effort in Florida using the predictive scores for

[5] If the number of registered voters were held constant across precincts, then the point of maximum difference would be at 0.5. However, the precincts where Obama received 42–45 percent of the vote are larger than precincts with an even split, so there are more votes to be harvested just to the left of the 50/50 mark.

Figure 2
Difference between Predictive Scores and Older Campaign Targeting Heuristics



Notes: Figure 2 represents a thought experiment: In Florida during the 2012 presidential election campaign, how would use of predictive scores for targeting compare to a strategy of "identification and sweep." (See text for details.) The x-axis shows the percent of the two-party vote share for Barack Obama in a precinct in the 2012 general election. The height of dotted bars, read off the left y-axis, report the number of precincts with a given level of support for Obama. The height of the solid line, read off the right y-axis, reports the hypothesized difference between the use of predictive scores for targeting and the use of "identification and sweep." β is assumed to be 0.01. The distribution of precinct data comes from all 4,354 precincts in the 2012 presidential election in Florida.

targeting (which was the strategy the campaign reportedly employed) compared to a precinct-based targeting strategy. Assuming the campaign had a 1 percentage point effect on turnout among the half of the citizens that it targeted for mobilization and successfully contacted, we estimate that it would have generated 8,525 more votes in Florida targeting based on predictive scores relative to targeting based on precinct. This vote total would have been decisive in the 2000 election between George W. Bush and Al Gore, and still constitutes 11 percent of the 74,309 vote margin of victory Barack Obama enjoyed in that state in 2012. Combined with the persuasion analysis above, this thumbnail sketch makes an argument that the 2012 Obama re-election would have been closer in key states had it used the older and coarser targeting technologies, rather than the predictive scores produced by its campaign data analysts.

Conclusion: Some Thoughts on Coordination

Sophisticated campaigns develop and use voter databases that contain a range of detailed information on individual citizens. As a result, campaign data analysts occupy an increasingly important role in politics. They develop predictive models that produce individual-level scores that predict citizens' likelihoods of performing certain political behaviors, supporting candidates and issues, and responding to targeted interventions. The use of these scores has increased dramatically during the last few election cycles. Simulations suggest that these advances could yield sizable and electorally meaningful gains to campaigns that harness them.

Since predictive scores make campaigns more effective and efficient by increasing the cost effectiveness of communicating with citizens, a broad range of organizations do and will employ the technologies. To the extent that predictive scores are useful and reveal true unobserved characteristics about citizens, it means that multiple organizations will produce predictive scores that recommend targeting the same sets of citizens. For example, some citizens might find themselves contacted many times, while other citizens—like those with low turnout behavior scores in 2012—might be ignored by nearly every campaign. The marginal effect of the fifth or sixth contact from a campaign will be less than the marginal effect of the first contact from a campaign. Thus, concentrating attention on the same set of citizens due to widespread adoption of predictive scores may offset some of the gains reaped from developing predictive scores in the first place. In this way, developing and using predictive scores creates a coordination game in which allied organizations would prefer to partition the electorate and not to duplicate efforts.

Coordination could theoretically happen between partisan organizations, like state parties, candidate campaigns, and coordinated campaigns, and across nonpartisan activities, like civil rights groups, labor unions, and environmental groups. However, partisan and nonpartisan organizations are not allowed to coordinate their electoral activities. Since it is nearly impossible to observe whom campaigns target for direct communications—that is, direct mail, knock on doors, and making phone calls—this coordination game has incomplete information, which means that inefficiencies from overlapping contacts are inevitable.

Even when coordination is allowed by law, coalitions may have conflicting incentives. There is enough regional variation in ideology that it is possible for local candidates to appeal to citizens who oppose the national candidate. For instance, local Republicans mobilizing citizens in liberal districts could have hurt Mitt Romney, and local Democrats mobilizing citizens in conservative districts could have hurt Barack Obama in 2012. The same dynamic plays out among nonpartisan groups as well. While labor union members and environmentalists agree on many policies and values, it is likely that some members do not hold that same views on both labor and environmental issues. In states like West Virginia, where the local coal industry is considered "dirty" by environmentalists, the groups could be working at cross-purposes, both with regards to messaging and targeting. Thus, mobilizing a set of citizens for a labor-related ballot initiative might result in less support for an environmentally friendly candidate. This tension is endemic to the very nature of the federal system of representation and coalition politics. The tension has always been present, but now that groups can share very detailed targeting plans and support scores, the tension can and will bubble to the surface more often than in the past.

The improved capability to target individual voters offers campaigns an opportunity to concentrate their resources where they will be most effective. This power, however, has not radically transformed the nature of campaign work. One could argue that the growing impact of data analytics in campaigns has amplified the importance of traditional campaign work. Message polling (that is, polls designed to gauge voter reactions to different campaign messages) no longer solely dictates targeting, but the increased demand for information during the campaign has increased the amount of polling used to generate snapshots of the electorate. Professional phone interviews are still used for message development and tracking, but they are also essential for developing predictive scores of candidate support and measuring changes in voter preferences in randomized experiments. Similarly, better targeting has made grassroots campaign tactics more efficient and therefore more cost competitive with mass communication forms of outreach. Volunteers still need to persuade skeptical neighbors, but they are now better able to focus on persuadable neighbors and use messages more likely to resonate. This leads to higher-quality interactions and (potentially) a more pleasant volunteer experience. So while savvy campaigns will harness the power of predictive scores, the scores will only help the campaigns that were already effective.

References

Ansolabehere, Stephen, and Eitan Hersh. 2010. "The Quality of Voter Registration Records: A State-by-State Analysis." Working paper, Institute for Quantitative Social Science, Harvard University.

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 20(4): 437–59.

Arceneaux, Kevin, and Robin Kolodny. 2009. "Educating the Least Informed: Group Endorsements in a Grassroots Campaign." *American Journal of Political Science* 53(4): 755–70.

Arceneaux, Kevin, and David W. Nickerson. 2009. "Who is Mobilized to Vote? A Re-Analysis of 11 Randomized Field Experiments." *American Journal of Political Science* 53(1): 1–16.

Arceneaux, Kevin, and David W. Nickerson. 2010. "Comparing Negative and Positive Campaign Messages: Evidence from Two Field Experiments." *American Politics Research* 38(1): 54–83.

Bailey, Michael A., Daniel J. Hopkins, and Todd Rogers. 2013. "Unresponsive and Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts." Available at SSRN: http://ssrn.com/abstract=2307631.

BigData-Startups. 2013. "Big Data during the Obama Campaign." http://www.bigdata-startups.com/BigData-startup/big-data-obama-campaign/. Accessed February 18, 2014.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R. A. Olshen. 1984. *Classification and Regression Trees.* Boca Raton, FL: CRC Press.

Bryan, Christopher J., Gregory M. Walton, Todd Rogers, and Carol S. Dweck. 2011. "Motivating Voter Turnout by Invoking the Self." *PNAS* 108(31): 12653–156.

Burt, Gabriel. 2013. "My Time at 'Obama for America.'" *Tech in Motion,* August 27, 2013. http://www.techinmotionevents.com/blog/post/2013/08/27/my-time-at-obama-for-america.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1): 266–98.

Dietterich, Tom. 1995. "Overfitting and Undercomputing in Machine Learning." *ACM Computing Surveys* 27(3): 326–27.

Duhigg, Charles. 2012. "Campaigns Mine Personal Lives to Get Out Vote." *New York Times,* October 13, page A1.

Gallagher, Sean. 2012. "Built to Win: Deep inside Obama's Campaign Tech." Arstechnica.com, November 14, 2012. http://arstechnica.com/information-technology/2012/11/built-to-win-deep-inside-obamas-campaign-tech/.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. 2007. *Data Clustering: Theory, Algorithms, and Applications.* Alexandria, VA: SIAM, Society for Industrial and Applied Mathematics.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3): 653–63.

Gerber, Alan S., and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You." *Journal of Politics* 71(1): 1–14.

Green, Donald, and Alan S. Gerber. 2008. *Get Out the Vote! How to Increase Voter Turnout.* Washington, DC: Brookings Institution Press.

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.

Haberman, Maggie. 2013. "Obama '12 Vets to Share Some, Not All, Data." *Politico,* November 20, 2013. http://www.politico.com/story/2013/11/barack-obama-2012-campaign-data-100133.html.

Hersh, Eitan D., and Brian F. Schaffner. 2013. "Targeted Campaign Appeals and the Value of Ambiguity." *Journal of Politics* 75(2): 520–34.

Hillygus, D. Sunshine, and Todd G. Shields. 2008. *The Persuadable Voter: Wedge Issues in Presidential Campaigns.* Princeton, NJ: Princeton University Press.

Issenberg, Sasha. 2010. "Nudge the Vote: How Behavioral Science is Remaking Politics." *New York Times Sunday Magazine,* October 31, page MM28.

Issenberg, Sasha. 2012a. *The Victory Lab: The Secret Science of Winning Campaigns.* New York, NY: Crown Publishers.

Issenberg, Sasha. 2012b. "The Death of the Hunch." *Slate,* May 22. http://www.slate.com/articles/news_and_politics/victory_lab/2012/05/obama_campaign_ads_how_the_analyst_institute_is_helping_him_hone_his_message_.html.

Issenberg, Sasha. 2012c. "How President Obama's Campaign Used Big Data to Rally Individual Voters, Part 2." *MIT Technology Review,* December 17. http://www.technologyreview.com/featuredstory/508851/how-obama-wrangled-data-to-win-his-second-term/.

Kuhn, David Paul. 2004. "Vacation Politics." CBSNews.com, December 5, 2007. http://www.cbsnews.com/2100-250_162-607234.html.

Madrigal, Alexis C. 2012. "When the Nerds Go Marching In." TheAtlantic.com, November 16. http://www.theatlantic.com/technology/archive/2012/11/when-the-nerds-go-marching-in/265325.

Malchow, Hal. 2003. *The New Political Targeting.* Washington, DC: Campaigns and Elections Magazine.

McAuliffe, Terry, with Steve Kettman. 2007. *What A Party!: My Life Among Democrats: Presidents, Candidates, Donors, Activists, Alligators, and Other Wild Animals.* New York, NY: St. Martin's Press.

Nicholson, Steven P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1): 52–66.

Nickerson, David W. 2005. "Partisan Mobilization Using Volunteer Phone Banks and Door Hangers." *Annals of the American Academy of Political and Social Science* 601:10–27.

Nickerson, David W. 2006a. "Hunting the Elusive Young Voter." *Journal of Political Marketing* 5(3): 47–69.

Nickerson, David W. 2006b. "Volunteer Phone Calls Can Increase Turnout: Evidence from Eight Field Experiments." *American Politics Research* 34(3): 271–92.

Nickerson, David W. 2007a. "Quality is Job One: Professional and Volunteer Mobilization Calls." *American Journal of Political Science* 51(2): 269–82.

Nickerson, David W. 2007b. "Don't Talk to Strangers: Experimental Evidence of the Need for Targeting." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 12–15.

Nickerson, David W., Ryan F. Friedrichs, and David C. King. 2006. "Partisan Mobilization Campaigns in the Field: Results from a Statewide Turnout Experiment in Michigan." *Political Research Quarterly* 59(1): 85–97.

Nickerson, David, and Todd Rogers. 2010. "Do You Have a Voting Plan?: Implementation

Intentions, Voter Turnout, and Organic Plan Making." *Psychological Science* 21(2): 194–99.

Rogers, Todd, and Masa Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, And Is (Misleadingly) Unbiased." *American Politics Research*, published online before print, September 5.

Rogers, Todd, and David W. Nickerson. 2013. "Can Inaccurate Beliefs about Incumbents Be Changed? And Can Reframing Change Votes?" HKS Faculty Research Working Paper Series, RWP13-018.

Scherer, Michael. 2012. "Inside the Secret World of Quants and Data Crunchers Who Helped Obama Win." *Time*, November 19, p. 56–60.

Tam Cho, Wendy K., and James G. Gimpel. 2007. "Prospecting for (Campaign) Gold." *American Journal of Political Science* 51(2): 255–68.

Vargas, Jose Antonio. 2008. "Obama Raised Half a Billion Online." *Washingtonpost.com,* November 20. http://voices.washingtonpost.com/44/2008/11/20/obama_raised_half_a_billion_on.html.

This article has been cited by:

1. Hiba Mohammad Noor, Ozgur Turetken, Mehmet Akgul. 2024. Social Media, Sentiments and Political Discourse – An Exploratory Study of the 2021 Canadian Federal Election. *ACM Transactions on Social Computing* **7**:1-4, 1-23. [Crossref]
2. Natascha Löffler. 2024. Trusting tech firms' big data for political microtargeting? A qualitative analysis of parties' communication managers risk and trust perceptions. *Journal of Information Technology & Politics* **21**:4, 438-452. [Crossref]
3. Casey F. Breen, Dennis M. Feehan. 2024. New Data Sources for Demographic Research. *Population and Development Review* **59**. . [Crossref]
4. E. Dowling, C. Tolbert, N. Micatka, T. Donovan. 2024. Does ranked choice Voting Increase voter turnout and mobilization?. *Electoral Studies* **90**, 102816. [Crossref]
5. Juan Pablo Luna, Daniel Alcatruz, Cristian Pérez Muñoz, Fernando Rosenblatt, Sergio Toro Maureira, Sebastián Valenzuela. 2024. Local Government, Social Media and Management of COVID-19: The Case of Chilean Mayoral Communication. *Political Communication* **41**:4, 672-691. [Crossref]
6. Víctor Renobell Santaren. 2024. Opiniones políticas compartidas en grupos de WhatsApp: radicalización o diversidad. *European Public & Social Innovation Review* **9**, 1-19. [Crossref]
7. Scott Cameron, Babak Hamidzadeh. 2024. Preserving paradata for accountability of semi-autonomous AI agents in dynamic environments: An archival perspective. *Telematics and Informatics Reports* **14**, 100135. [Crossref]
8. Andrew Barclay, Katharine Dommett, Uta Russmann. 2024. Data Driven-Campaign Infrastructures in Europe: Evidence from Austria and the UK. *Journal of Political Marketing* 1-20. [Crossref]
9. Benjamin Farrer. 2024. Political Communication as a Tragedy of the Commons. *Political Studies* **72**:2, 701-718. [Crossref]
10. Jakob Thrane Mainz, Jørn Sønderholm, Rasmus Uhrenfeldt. 2024. Artificial intelligence and the secret ballot. *AI & SOCIETY* **39**:2, 515-522. [Crossref]
11. Laurent Bouton, Micael Castanheira, Allan Drazen. 2024. A Theory of Small Campaign Contributions. *The Economic Journal* **16**. . [Crossref]
12. Katharine Dommett, Sam Power, Andrew Barclay, Amber Macintyre. 2024. Understanding the Modern Election Campaign: Analysing Campaign Eras through Financial Transparency Disclosures at the 2019 UK General Election. *Government and Opposition* **13**, 1-27. [Crossref]
13. Michael J. Ritter, Joseph A. Coll, Caroline J. Tolbert. 2024. Can Election Administration Overcome the Effects of Restrictive State Voting Laws?. *Election Law Journal: Rules, Politics, and Policy* **23**:1, 19-38. [Crossref]
14. Katharine Dommett, Sam Power. 2024. Monitoring digital election campaigns: Assessing the transparency ecosystem in the United Kingdom. *Politics* **44**:1, 119-139. [Crossref]
15. Katharine Dommett, Andrew Barclay, Rachel Gibson. 2024. Just what is data-driven campaigning? A systematic review. *Information, Communication & Society* **27**:1, 1-22. [Crossref]
16. Samuel A. T. Johnston, Sinéad C. M. Harrington. 2024. Campaigning in the dark: theorising campaign strategies from the 2022 Seanad by-election. *Irish Political Studies* **39**:1, 1-23. [Crossref]
17. Heidi Haberl-Glantschnig, Katharina Janauschek. Zielsetzung 143-156. [Crossref]
18. Guocheng Feng, Kaihao Chen, Huaiyu Cai, Zhijian Li. A Hybrid Method of Sentiment Analysis and Machine Learning Algorithm for the U.S. Presidential Election Forecasting 1495-1500. [Crossref]
19. Martin J Riedl, Dariya Tsyrenzhapova, Jessica R Collier, Jacob Gursky, Katie Joseff, Samuel Woolley. 2023. The role of geolocation data in U.S. political campaigning: How digital political strategists

perceive it. *Convergence: The International Journal of Research into New Media Technologies* **29**:6, 1624-1640. [Crossref]

20. Rachel Gibson, Esmeralda Bon, Andrea Römmele. 2023. Operationalizing data-driven campaigning: designing a new tool for mapping and guiding regulatory intervention. *Policy Studies* **8**, 1-17. [Crossref]
21. McDonald Lewanika. 2023. ZANU-PF's bigwig rallies as performative politics during Zimbabwe's 2008 and 2013 elections. *Commonwealth & Comparative Politics* **111**, 1-20. [Crossref]
22. Yago de Souza Paiva. 2023. CAMPANHA POLÍTICA DIGITAL: DETERMINANTES DEMOGRÁFICOS E ECONÔMICOS DO IMPULSIONAMENTO DE CONTEÚDOS. *REVISTA FOCO* **16**:7, e2670. [Crossref]
23. Andreas Jungherr. 2023. Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society* **9**:3. . [Crossref]
24. David Dutwin, Patrick Coyle, Joshua Lerner, Ipek Bilgen, Ned English. 2023. Leveraging Predictive Modelling from Multiple Sources of Big Data to Improve Sample Efficiency and Reduce Survey Nonresponse Error. *Journal of Survey Statistics and Methodology* **3**. . [Crossref]
25. Michael Ritter. 2023. Assessing the Impact of the United States Postal System and Election Administration on Absentee and Mail Voting in the 2012 to 2020 U.S. Midterm and Presidential Elections. *Election Law Journal: Rules, Politics, and Policy* **22**:2, 166-184. [Crossref]
26. Tobias Mettler, Gianluca Miscione. 2023. Is altruism dead? A critical case study on the paradigm shift in Open Government Data. *Information Polity* **28**:2, 199-218. [Crossref]
27. Rachel Gibson. 2023. Data-Driven Campaigning as a Disruptive Force. *Political Communication* **40**:3, 351-355. [Crossref]
28. Glenn Kefford, Katharine Dommett, Jessica Baldwin-Philippi, Sara Bannerman, Tom Dobber, Simon Kruschinski, Sanne Kruikemeier, Erica Rzepecki. 2023. Data-driven campaigning and democratic disruption: Evidence from six advanced democracies. *Party Politics* **29**:3, 448-462. [Crossref]
29. Kelly LeRoux, Julie Langer, Samantha Plotner. 2023. Nonprofit Messaging and the 2020 Election: Findings from a Nonpartisan Get-Out-The-Vote (GOTV) Field Experiment. *Nonprofit Policy Forum* **14**:2, 157-183. [Crossref]
30. Peter Aagaard, Selma Marthedal. 2023. Political microtargeting: Towards a pragmatic approach. *Internet Policy Review* **12**:1. . [Crossref]
31. Tomás Baviera, Lorena Cano-Orón, Dafne Calvo. 2023. Tailored Messages in the Feed? Political Microtargeting on Facebook during the 2019 General Elections in Spain. *Journal of Political Marketing* **13**, 1-20. [Crossref]
32. Sonja Sperber, Susanne Täuber, Corinne Post, Cordula Barzantny. 2023. Gender Data Gap and its impact on management science — Reflections from a European perspective. *European Management Journal* **41**:1, 2-8. [Crossref]
33. Alexander V. Hirsch. 2023. Polarization and Campaign Spending in Elections. *The Journal of Politics* **85**:1, 240-249. [Crossref]
34. Quratulain Alvi, Syed Farooq Ali, Sheikh Bilal Ahmed, Nadeem Ahmad Khan, Mazhar Javed, Haitham Nobanee. 2023. On the frontiers of Twitter data and sentiment analysis in election prediction: a review. *PeerJ Computer Science* **9**, e1517. [Crossref]
35. Priya Sachdeva, Archan Mitra. Decoding the Twitter Sentiment Using Artificial Intelligence Tools: A Study on Tokyo Olympics 2020 539-554. [Crossref]
36. Samuel Baltz, Alexander Agadjanian, Declan Chin, John Curiel, Kevin DeLuca, James Dunham, Jennifer Miranda, Connor Halloran Phillips, Annabel Uhlman, Cameron Wimpy, Marcos Zárate, Charles Stewart. 2022. American election results at the precinct level. *Scientific Data* **9**:1. . [Crossref]

37. Magnus Høholt Kaspersen, Karl-Emil Kjær Bilstrup, Maarten Van Mechelen, Arthur Hjort, Niels Olof Bouvin, Marianne Graves Petersen. 2022. High school students exploring machine learning and its societal implications: Opportunities and challenges. *International Journal of Child-Computer Interaction* **34**, 100539. [Crossref]

38. Christopher Hare, Mikayla Kutsuris. 2022. Measuring Swing Voters with a Supervised Machine Learning Ensemble. *Political Analysis* **8**, 1-17. [Crossref]

39. Alberto López Ortega. 2022. Are microtargeted campaign messages more negative and diverse? An analysis of Facebook Ads in European election campaigns. *European Political Science* **21**:3, 335-358. [Crossref]

40. Pau Balart, Agustin Casas, Orestis Troumpounis. 2022. Technological change, campaign spending and polarization. *Journal of Public Economics* **211**, 104666. [Crossref]

41. Deana A. Rohlinger. 2022. Digital technologies, dysfunctional movement-party dynamics and the threat to democracy. *Information, Communication & Society* **25**:5, 591-597. [Crossref]

42. Enrico Cantoni, Vincent Pons. 2022. Does Context Outweigh Individual Characteristics in Driving Voting Behavior? Evidence from Relocations within the United States. *American Economic Review* **112**:4, 1226-1272. [Abstract] [View PDF article] [PDF with links]

43. Somya Mehta, Kristofer Erickson. 2022. Can online political targeting be rendered transparent? Prospects for campaign oversight using the Facebook Ad Library. *Internet Policy Review* **11**:1. . [Crossref]

44. Ismael Crespo Martínez, José Miguel Rojo Martínez, Claudia Mayordomo Zapata, Salvador Moreno Moreno. 2022. Conversación y difusión de información política en WhatsApp: un análisis de la "caja negra" desde las teorías de la interacción en redes sociales. *Revista de Comunicación* **21**:1, 117-136. [Crossref]

45. Mona Krewel. Wahlkampf 251-264. [Crossref]

46. Merve Schmitz-Vardar. Regieren nach Umfragen 695-709. [Crossref]

47. Jonathan Bonneau, Laurence Grondin-Robillard, Marc Ménard, André Mondoux. Fighting the "System": A Pilot Project on the Opacity of Algorithms in Political Communication 97-120. [Crossref]

48. Magnus Høholt Kaspersen, Karl-Emil Kjær Bilstrup, Maarten Van Mechelen, Arthur Hjort, Niels Olof Bouvin, Marianne Graves Petersen. 2022. Votestratesml: A High School Learning Tool for Exploring Machine Learning and its Societal Implications. *SSRN Electronic Journal* **52**. . [Crossref]

49. Lauren Maunder, Joshua Lyons, Lauren Anderson, Joe Harrison, Brian Timana-Gomez, Paul O'Donnell, Kiernan B. George, Alan J. Michaels. Identifying Corporate Political Trends Online 1-6. [Crossref]

50. Enrico Cantoni, Vincent Pons. 2021. Strict Id Laws Don't Stop Voters: Evidence from a U.S. Nationwide Panel, 2008–2018. *The Quarterly Journal of Economics* **136**:4, 2615-2660. [Crossref]

51. Ruben L. Bach, Christoph Kern, Ashley Amaya, Florian Keusch, Frauke Kreuter, Jan Hecht, Jonathan Heinemann. 2021. Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review* **39**:5, 862-883. [Crossref]

52. Melissa Maceyko. 2021. Political Campaigns, Voter Outreach, and American Democracy: Socializing Effective Participation and Citizen Agency in the United States. *American Anthropologist* **123**:3, 539-551. [Crossref]

53. Patrícia Rossini, Heloisa Sturm-Wikerson, Thomas J. Johnson. 2021. A wall of incivility? Public discourse and immigration in the 2016 U.S. Primaries. *Journal of Information Technology & Politics* **18**:3, 243-257. [Crossref]

54. Katharina Baum, Stefan Meissner, Hanna Krasnova. 2021. Partisan self-interest is an important driver for people's support for the regulation of targeted political advertising. *PLOS ONE* **16**:5, e0250506. [Crossref]

55. Doug Walker, Edward L. Nowlin. 2021. Data-Driven Precision and Selectiveness in Political Campaign Fundraising. *Journal of Political Marketing* **20**:2, 73-92. [Crossref]

56. Jørn Sønderholm, Jakob Mainz, Rasmus Uhrenfeldt. 2021. BIG DATA ANALYTICS AND HOW TO BUY AN ELECTION. *Public Affairs Quarterly* **35**:2, 119-139. [Crossref]

57. Ilgar SEYIDOV. 2021. "Büyük Verinin Gücü Adına": Siyasi Kampanyalarda Etkili Veri Kullanımı. *TRT Akademi* **6**:11, 32-49. [Crossref]

58. Glenn Kefford. Data and Analytics 45-67. [Crossref]

59. Glenn Kefford. Data-Driven Campaigning: A Case Study from the Ground 119-139. [Crossref]

60. Sabiha Mahmud Sumi. Automated Prediction of Voter's Party Affiliation Using AI 257-279. [Crossref]

61. Merve Schmitz-Vardar. Regieren nach Umfragen 1-15. [Crossref]

62. Natascha Löffler, Ulrike Röttger, Christian Wiencierz. Data-Based Strategic Communication as a Mediator of Trust: Recipients' Perception of an NPO's Automated Posts 115-134. [Crossref]

63. Christian Pentzold, Lena Fölsche. 2020. Data-driven campaigns in public sensemaking: Discursive positions, contextualization, and maneuvers in American, British, and German debates around computational politics. *Communications* **45**:s1, 535-559. [Crossref]

64. Barteld Braaksma, Kees Zeelenberg, Sofie De Broe. Big Data in Official Statistics 303-338. [Crossref]

65. Philipp Schmidt, Felix Biessmann, Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* **29**:4, 260-278. [Crossref]

66. Jonathan Bright, Scott Hale, Bharath Ganesh, Andrew Bulovsky, Helen Margetts, Phil Howard. 2020. Does Campaigning on Social Media Make a Difference? Evidence From Candidate Use of Twitter During the 2015 and 2017 U.K. Elections. *Communication Research* **47**:7, 988-1009. [Crossref]

67. David W. Nickerson, Todd Rogers. 2020. Campaigns influence election outcomes less than you think. *Science* **369**:6508, 1181-1182. [Crossref]

68. Joseph Gerald Hirschberg, Jeanette Ngaire Lye. 2020. GRADING JOURNALS IN ECONOMICS: THE ABCS OF THE ABDC. *Journal of Economic Surveys* **34**:4, 876-921. [Crossref]

69. Jessica Baldwin-Philippi. 2020. Data Ops, Objectivity, and Outsiders: Journalistic Coverage of Data Campaigning. *Political Communication* **37**:4, 468-487. [Crossref]

70. Uta Russmann. 2020. Voter Targeting Online in Comparative Perspectives: Political Party Websites in the 2008/2009 and 2013 Austrian and German Election Campaigns. *Journal of Political Marketing* **19**:3, 177-200. [Crossref]

71. Andreas Jungherr, Gonzalo Rivero, Daniel Gayo-Avello. Retooling Politics **3**, . [Crossref]

72. M. Wiesenberg, R. Tench. 2020. Deep strategic mediatization: Organizational leaders' knowledge and usage of social bots in an era of disinformation. *International Journal of Information Management* **51**, 102042. [Crossref]

73. Aitor Jiménez González. 2020. Tech power: a critical approach to digital corporations. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales* **17**:1, 77-85. [Crossref]

74. Michael Harker. 2020. Political advertising revisited: digital campaigning and protecting democratic discourse. *Legal Studies* **40**:1, 151-171. [Crossref]

75. Paola Mavriki, Maria Karyda. Big Data Analytics: From Threatening Privacy to Challenging Democracy 3-17. [Crossref]

76. Mona Krewel. Wahlkampf 1-14. [Crossref]

77. Emmanuelle Avril. Paradoxes of the Cyber Party: The Changing Organizational Design of the British Labour Party 609-625. [Crossref]
78. Ali A. Guenduez, Tobias Mettler, Kuno Schedler. 2020. Technological frames in public administration: What do public managers think of big data?. *Government Information Quarterly* **37**:1, 101406. [Crossref]
79. Thomas Ehrhard, Antoine Bambade, Samuel Colin. Digital Campaigning in France, a Wide Wild Web? 113-126. [Crossref]
80. David M. Steinberg, Eddie Aronovich. 2020. Thoughts on Data Science in business and industry. *Applied Stochastic Models in Business and Industry* **36**:1, 36-40. [Crossref]
81. Paola Mavriki, Maria Karyda. 2019. Automated data-driven profiling: threats for group privacy. *Information & Computer Security* **28**:2, 183-197. [Crossref]
82. Ka-Shing Cheung, Wai K Leung, Wai-Kay Seto. 2019. Application of Big Data analysis in gastrointestinal research. *World Journal of Gastroenterology* **25**:24, 2990-3008. [Crossref]
83. Tom Dobber, Damian Trilling, Natali Helberger, Claes de Vreese. 2019. Spiraling downward: The reciprocal relation between attitude toward political behavioral targeting and privacy concerns. *New Media & Society* **21**:6, 1212-1231. [Crossref]
84. Felix M. Simon. 2019. "We power democracy": Exploring the promises of the political data analytics industry. *The Information Society* **35**:3, 158-169. [Crossref]
85. Henry E. Brady. 2019. The Challenge of Big Data and Data Science. *Annual Review of Political Science* **22**:1, 297-323. [Crossref]
86. Nikolaos Dimisianos. Chapter 6 Political Campaigns, Social Media, and Analytics: The Case of the GDPR 73-88. [Crossref]
87. Pablo Olivera, Silvio Danese, Nicolas Jay, Gioacchino Natoli, Laurent Peyrin-Biroulet. 2019. Big data in IBD: a look into the future. *Nature Reviews Gastroenterology & Hepatology* **16**:5, 312-321. [Crossref]
88. Thomas Ehrhard, Antoine Bambade, Samuel Colin. 2019. À la conquête des élections 2.0. *Politiques de communication* **N° 12**:1, 97-133. [Crossref]
89. Biagio Aragona, Rosanna De Rosa. 2019. Big data in policy making. *Mathematical Population Studies* **26**:2, 107-113. [Crossref]
90. Gonzalo Rivero. 2019. Preaching to the choir: ideology and following behaviour in social media. *Contemporary Social Science* **14**:1, 54-70. [Crossref]
91. Javier Vidal-García, Marta Vidal, Rafael Hernández Barros. Computational Business Intelligence, Big Data, and Their Role in Business Decisions in the Age of the Internet of Things 1048-1067. [Crossref]
92. Javier Vidal-García, Marta Vidal, Rafael Hernández Barros. Business Applications of Big Data 1346-1367. [Crossref]
93. Ramona Sue McNeal, Lisa Dotterweich Bryan. Presidential Elections Web 2.0 244-253. [Crossref]
94. Bianica Pires, Ian Crandell, Madison Arnsbarger, Vicki Lancaster, Aaron Schroeder, Stephanie Shipp, Wendy Kang, Paula Robinson, Sallie Keller. 2018. Predicting postsecondary trajectories in Virginia high schools using publicly available data. *Statistical Journal of the IAOS* **34**:4, 553-565. [Crossref]
95. Ryan D. Enos, Anthony Fowler. 2018. Aggregate Effects of Large-Scale Campaigns on Voter Turnout. *Political Science Research and Methods* **6**:4, 733-751. [Crossref]
96. Sérgio Braga, Márcio Carlomagno. 2018. Eleições como de costume? Uma análise longitudinal das mudanças provocadas nas campanhas eleitorais brasileiras pelas tecnologias digitais (1998-2016). *Revista Brasileira de Ciência Política* **1**:26, 7-62. [Crossref]
97. Nick Anstead. 2018. Data and Election Campaigning. *Political Insight* **9**:2, 32-35. [Crossref]

98. Elizabeth Gibney. 2018. The scant science behind Cambridge Analytica's controversial marketing techniques. *Nature* **110**. . [Crossref]
99. Arthur Beckman. 2018. Political Marketing and Intellectual Autonomy. *Journal of Political Philosophy* **26**:1, 24-46. [Crossref]
100. Christine L. Exley, Ragan Petrie. 2018. The impact of a surprise donation ask. *Journal of Public Economics* **158**, 152-167. [Crossref]
101. Stephan D. Whitaker. 2018. Big Data versus a survey. *The Quarterly Review of Economics and Finance* **67**, 285-296. [Crossref]
102. Kyle Endres, Kristin J. Kelly. 2018. Does microtargeting matter? Campaign contact strategies and young voters. *Journal of Elections, Public Opinion and Parties* **28**:1, 1-18. [Crossref]
103. Mario Callegaro, Yongwei Yang. The Role of Surveys in the Era of "Big Data" 175-192. [Crossref]
104. Simon Kruschinski, André Haller. Back to the roots?! Der datengestützte Tür-zu-Tür-Wahlkampf in politischen Wahlkampagnen 289-317. [Crossref]
105. Lena Ulbricht, Sebastian Haunss, Jeanette Hofmann, Ulrike Klinger, Jan-Hendrik Passoth, Christian Pentzold, Ingrid Schneider, Holger Straßheim, Jan-Peter Voß. Dimensionen von Big Data: Eine politikwissenschaftliche Systematisierung 151-231. [Crossref]
106. Emmanuelle Avril. Paradoxes of the Cyber Party: The Changing Organizational Design of the British Labour Party 1-17. [Crossref]
107. Ramona Sue McNeal, Lisa Dotterweich Bryan. Presidential Elections Web 2.0 3612-3620. [Crossref]
108. Lena A. Hübner. Exploring Real-time Voter Targeting Strategies on Social Media 217-232. [Crossref]
109. Simon Kruschinski, André Haller. 2017. Restrictions on data-driven political micro-targeting in Germany. *Internet Policy Review* **6**:4. . [Crossref]
110. David Sathiaraj, William M. Cassidy, Eric Rohli. 2017. Improving Predictive Accuracy in Elections. *Big Data* **5**:4, 325-336. [Crossref]
111. Edward Elder, Justin B. Phillips. 2017. Appeals to the Hispanic Demographic: Targeting through Facebook Autoplay Videos by the Clinton Campaign during the 2015/2016 Presidential Primaries. *Journal of Political Marketing* **16**:3-4, 319-342. [Crossref]
112. Edward Elder, Justin B. Phillips. 2017. Appeals to the Hispanic demographic: Targeting through Facebook autoplay videos by the Clinton campaign during the 2015/2016 presidential primaries. *Journal of Political Marketing* **32**. . [Crossref]
113. Zachary C. Steinert-Threlkeld. 2017. Longitudinal Network Centrality Using Incomplete Data. *Political Analysis* **25**:3, 308-328. [Crossref]
114. Nick Anstead. 2017. Data-Driven Campaigning in the 2015 United Kingdom General Election. *The International Journal of Press/Politics* **22**:3, 294-313. [Crossref]
115. Todd Rogers, Donald P. Green, John Ternovski, Carolina Ferrerosa Young. 2017. Social pressure and voting: A field experiment conducted in a high-salience election. *Electoral Studies* **46**, 87-100. [Crossref]
116. Markus Wiesenberg, Ansgar Zerfass, Angeles Moreno. 2017. Big Data and Automation in Strategic Communication. *International Journal of Strategic Communication* **11**:2, 95-114. [Crossref]
117. Kim Fridkin, Patrick Kenney, Amanda Wintersieck, Jill Carle. 2017. The Upside of the Long Campaign. *American Politics Research* **45**:2, 186-223. [Crossref]
118. Javier Vidal-García, Marta Vidal, Rafael Hernandez Barros. Computational Business Intelligence, Big Data, and Their Role in Business Decisions in the Age of the Internet of Things 249-268. [Crossref]
119. Javier Vidal-García, Marta Vidal, Rafael Hernández Barros. Business Applications of Big Data 104-125. [Crossref]

120. Andrea Römmele, Dirk von Schneidmesser. 2016. Election campaigning enters a fourth phase: the mediatized campaign. *Zeitschrift für Politikwissenschaft* **26**:4, 425-442. [Crossref]
121. Kyle Endres. 2016. The Accuracy of Microtargeted Policy Positions. *PS: Political Science & Politics* **49**:04, 771-774. [Crossref]
122. Christopher Garcia. 2016. A nearest-neighbor algorithm for targeted interaction design in social outreach campaigns. *Kybernetes* **45**:8, 1243-1256. [Crossref]
123. Ho-Don Yan, Michael S. H. Heng. 2016. Political entrepreneurship and the Ko-P phenomenon in Taiwan's democratization *. *Asian Journal of Political Science* **24**:3, 285-304. [Crossref]
124. Lindsey Cormack. 2016. Extremity in Congress: Communications versus Votes. *Legislative Studies Quarterly* **41**:3, 575-603. [Crossref]
125. Andreas Jungherr. 2016. Four Functions of Digital Tools in Election Campaigns. *The International Journal of Press/Politics* **21**:3, 358-377. [Crossref]
126. Todd Rogers, Leanne ten Brinke, Dana R. Carney. 2016. Unacquainted callers can predict which citizens will vote over and above citizens' stated self-predictions. *Proceedings of the National Academy of Sciences* **113**:23, 6449-6453. [Crossref]
127. Saikat Banerjee, Bibek Ray Chaudhuri. 2016. Factors responsible behind political brand preference: an empirical study on Indian voters. *Marketing Intelligence & Planning* **34**:4, 559-582. [Crossref]
128. Markku Kaustia, Samuli Knüpfer, Sami Torstila. 2016. Stock Ownership and Political Behavior: Evidence from Demutualizations. *Management Science* **62**:4, 945-963. [Crossref]
129. Donald P Green, Michael Schwam-Baird. 2016. Mobilization, participation, and American democracy. *Party Politics* **22**:2, 158-164. [Crossref]
130. Christine L. Exley, Ragan Petrie. 2016. Finding Excuses to Decline the Ask. *SSRN Electronic Journal* . [Crossref]
131. Pertti Ahonen. 2015. Institutionalizing Big Data methods in social and political research. *Big Data & Society* **2**:2. . [Crossref]
132. RYAN D. ENOS, EITAN D. HERSH. 2015. Party Activists as Campaign Advertisers: The Ground Campaign as a Principal-Agent Problem. *American Political Science Review* **109**:2, 252-278. [Crossref]
133. Ramona McNeal, Lisa Dotterweich Bryan. The Mobile Presidential Election 992-1002. [Crossref]
134. Todd Rogers, Don A. Moore. 2014. The Motivating Power of Under-Confidence: 'The Race is Close But We're Losing'. *SSRN Electronic Journal* . [Crossref]
135. Stephan Whitaker. 2014. Big Data versus a Survey. *SSRN Electronic Journal* . [Crossref]
136. Marta Vidal, Javier Vidal-García, Rafael Hernandez Barros. Big Data and Business Decision Making 140-157. [Crossref]
137. Javier Vidal-García, Marta Vidal. Big Data Management in Financial Services 217-230. [Crossref]
138. Kees Zeelenberg, Barteld Braaksma. Big Data in Official Statistics 274-296. [Crossref]